SESSION

REAL-WORLD DATA MINING APPLICATIONS, CHALLENGES, AND PERSPECTIVES

Chair(s)

Drs. Mahmoud Abou-Nasr
Robert Stahlbock
Gary M. Weiss
Diego Galar

The need for Big Data collection and analyses to support the development of an advanced maintenance strategy

Dr David Baglee, Dr Salla Marttonen, and Professor Diego Galar

***Abstract*— Data mining applications are becoming increasingly important for the wide range of manufacturing processes. During daily manufacturing operations large amounts of data is generated. The abundance of data however, often impedes the ability to extract useful knowledge. In addition, the large amount of data stored in often unconnected databases makes it impractical to manually analyse for valuable decision-making information. New intelligent Data Mining tools and techniques are required which can intelligently analyse data and produce useful knowledge for manufacturing. This is an important issue with regard to the development of an advanced maintenance strategy. Maintenance optimization is critical for enhancing the productivity of assets within an organisation. Maintenance effectiveness depends on the quality, timeliness, accuracy and completeness of the information related to asset optimization based on which decisions are made. Recently developed Condition Monitoring Systems (CMS) generate and collect large amount of data during daily operations. These systems contain hundreds of attributes, which need to be simultaneously considered in order to accurately model the system's behaviour and provide operators and senior management with the necessary data required to ensure production levels are met.**

This paper will present an overview of the big data tools and techniques required to collect and analyse a range of data to support the development of an advanced maintenance strategy. The challenges of big data in maintenance including capturing, accessing, and processing information will be analysed. To achieve e-maintenance, how to integrate information and communication technologies into maintenance and the corresponding requirements and constraints will be identified.

I. Introduction

Effective use of leading edge Information and Communication Technologies (ICT) is seen as important, and possibly critical, to the future competitiveness of European Industry. In particular, manufacturing organisations are frequently characterised by high staff turnover, lack of knowledge and training, and a lack of appropriate asset management strategies. This has resulted in poor manufacturing efficiency and large amounts of waste. The implementation of structured maintenance methods has made possible the development of ICT including software and hardware systems.

The production and process industry are passing through a continuous transformation and improvement for the last couple of decades, due to the global competition coupled with advances in ICT. Manufacturing organisations are focusing more on big data collection and analyses to support e-business intelligence. The data should also be used to support other functions within the organisation which could impact asset management such as marketing and customer relations. The aim is to remain competitive and efficient by improving equipment performance and reliability by introducing an asset management strategy based upon accurate data collection and analyses tools and techniques.

Maintenance effectiveness depends on the quality, timeliness, accuracy and completeness of information related to machine degradation state, based on which decisions are made. This translates into two key requirements: (i) preventing data overload, ability to differentiate and prioritize data (during collection as well as reporting) and (ii) to prevent, as far as possible, the occurrence of information islands. With the emergence of intelligent sensors to measure and monitor the state of health of the component and gradual implementation of ICT in organizations, conceptualization and implementation of e-maintenance is turning into a reality [1]. While e-maintenance has a number of benefits seamless integration of ICT into the industrial environment remains a challenge. A variety of techniques are available to enable the above goals. Different data mining techniques serve different purposes, each offering its own advantages and disadvantages. The most commonly used techniques can be categorized in the following groups: Statistical methods, Artificial Neural Networks, Decision Trees, Rule Induction, Case-Based Reasoning, Bayesian Belief Networks, and Genetic Algorithms and Evolutionary Programming. It is very critical to understand and address the requirements and constraints from the maintenance as well as the ICT standpoints in parallel in order to identify and understand which information is required and when.

II. Big Data Benefits and Challenges for Maintenance

Big data is a revolutionary advanced methodology where big data sets which are collected at an unprecedented scale, are often complicated and difficult to process using traditional data processing tools such as relational and object-relational database management systems. Big data refers to the datasets that could not be perceived, acquired, managed, and

Dr David Baglee is with the Institute for Automotive & Manufacturing Advanced Practice (AMAP), University of Sunderland, Sunderland, SR5 3XB, UK (e-mail: David.baglee@sunderland.ac.uk,

Dr Salla Marttonen is with the School of Business and Management, Lappeenranta University of Technology, FIN-53851 Lappeenranta, Finland. (e-mail: salla.marttonen@lut.fi).

Professor Diego Galar is with the Division of Operations and Maintenance, University of Lulea, Sweden(e-mail: diego.galar@ltu.se)

processed by traditional Information Technology (IT) and software/hardware tools within a tolerable time [6].

Regarding the adoption of Big Data technologies by industrial sectors, following a pattern typical in technology transference between sectors, there are important differences. For example, in sectors not very fragmented where most of the information is already structured and comes from the same source, the use of big data analytics is nowadays a standard (e.g. bank sector or pharmaceutical sector). For these sectors, there is also a great number of SW tools and IT services that cover most of the end user needs.

However, these examples are only the exception, since massive business, susceptible to incorporate the Big Data concept, have not adopted Big Data yet, either for the lack of specific tools or the excessive cost to involve all the required stakeholders. One of these sectors is maintenance of assets. Within this field, big data has become a new specialization for monitoring, maintaining, and optimizing assets for better quality and performance. Kurtz [2] states that big data helps to solve complex technical and operational issues in maintenance, such as:

- Lack of visibility into asset health;
- Unexpected costs for unscheduled maintenance and unexpected failure;
- Not capable to accurately predict asset downtime and maintenance costs;
- Lack of analytical insights and tools for maintenance optimization.

Therefore potential benefits of Big Data technologies in the field of maintenance will require predictive algorithms using heterogeneous data sources, scalable data structures, real-time communications and visualizations techniques. These technologies and methodologies applied to such a challenging industry relevant sector will provide the expected system component degradation prediction modelling, maintenance cost prediction modelling, and asset condition monitoring. This should lead to boost the efficiency and maintenance cost reduction.

As an advanced predictive analytics methodology, big data is tailored to meet the needs of optimizing maintenance tasks in order to reduce operational expense and increase equipment reliability. For example, big data can be used to improve the production line continuity: A sensor network can be applied to collect the real-time production line data. The data is then used to analyse the asset health and predict failure or the mean time to failure (MTTF) and suggest possible solutions to minimize disruptive and unscheduled downtime. Big data is a multi-stage process, including data acquisition, information extraction, data modelling and analysis, decision making. Big Data can also be used to influence the next generation of products by identifying the issues that cause unnecessary and unplanned downtime. An analysis of the data could provide an insight to known and unknown issues and by feeding the results back into the design process the aim is to improve the manufacturing process and product quality based upon accurate data.

According to authors including [3] and [4], big data intelligent mining techniques should be applied within manufacturing organisations to support a number of processes including (1) Manufacturing knowledge acquisition by examining relevant and accurate data, which implicitly contains most of the required expert knowledge (2) adaptive or intelligent manufacturing system which are capable of learning from previous situations (3) quality control systems which with monitor standard operating procedures and identify deviation from the norm. New intelligent data mining tools and techniques are required which can intelligently analyse data and produce useful knowledge for manufacturing. This is an important issue with regard to maintenance strategy development.

The following section will explain the process of big data in maintenance and discuss the specific challenges to each step.

A. Data Collection, Storage and Integration

Acquiring and storing such large and rapidly increasing volumes of data has often been challenging. With the deployments of mobile networking, cloud computing has become the best solution for big data in data collection, storage, integration, and distribution. However, a widely accepted solution for data management in cloud computing still has not been designed [5]. Cloud computing still encounters unsolved problems related to e.g. data heterogeneity, data redundancy, assessing the value of data (to decide which data should be discarded and which stored), and data confidentiality [6].

Also moving big data to and from the cloud has presented a challenge because the capacity of the network bandwidth has proven to be a bottleneck [7]. Traditional wide area network (WAN) based data transfer methods use a fraction of available bandwidth for transmission; they cannot move such large amounts of data at a suitable speed, which may introduce unacceptable delays in data collection. IBM Aspera [8] had created an innovative data transport technology to solve this issue: Without using traditional transmission control protocol (TCP), IBM Aspera [8] designed FASP (Fast, Adaptive, and Secure Protocol) for transferring files over public and private internet protocol (IP) networks which is independent of network delay and packet loss [8].

Data integration can be seen as a process including data extraction, transformation and loading. It aims for uniform data despite the numerous data sources used [6]. Comprehensive solutions for integrating big data do not exist at the moment, and this poses a challenge for developing advanced data-based maintenance strategies.

B. Data Modelling and Analysis

The use of big data, to support maintenance task selection, could be described as (i) the use of data to detect and predict product failures and (ii) to increase equipment effectiveness i.e. increase quality, reduce costs and improve up-time. Generating user-friendly predictive models and conducting

cause analysis are therefore very important actions which need to be supported by the use of big data. To fully realize the potential benefits of big data, there are two technical challenges that need to be addressed:

a) Data uncertainty and inconsistency: Besides volumes of data, the inconsistency, uncertainty, and incompleteness of data makes modelling and analysis more challenging. Different from small samples, big data is always noisy, dynamic, diverse, inter-correlated, and sometime inaccurate. In fact, with generating suitable statistics, one can use approximate analysis to expose some reliable knowledge hidden in the data [9].

b) Analysis timeliness: As data grow rapidly in volume and it is not economical to store all raw data, real-time analysis techniques are needed to perform data processing. Some examples of general platforms designed for real-time analysis are EMC Greenplum and SAP HANA [6]. Regarding maintenance, one possible solution is to find elements that meet a specified maintenance criterion. And in this case, index structures to support various criteria need to be designed.

There are already a number of commercial and open source software systems available for mining and analysing big data [6]. However, according to Begoli and Horey [10] specialized data management systems are needed to support the range of analysis methods and environments. The software architecture should not extensively limit the tools available for the user because the data needs are very different depending on the decision-making situation in question. It can also be stated that the currently applied analysis methods are based on data mining from the 1980s and statistical methods from the 1970s [11]. Currently there are no ground breaking modern approaches available for analysing big data.

C. Decision Making and Actions Recommendation

Ultimately, provided with the result of analysis, decision will be made and maintenance actions will be recommended. As reported in [12], there exist many challenges during this process, including getting functional managers to make decisions rather than based on intuition, putting analysis of big data in a presentable form for making decisions, determining actions with the insights created from big data, etc. These challenges hold the manufacturing and maintenance managers back from seizing the benefits offered by big data. It is important to address this. Chen and Zhang [7] state that the weaknesses of the existing visualisation tools for big data focus on response time, functionalities, and scalability. In addition to visualisation, also mobile interfaces and human-computer interaction have been identified as major topics for future research [11].

III. E-MAINTENANCE EXPECTATIONS AND INTEGRATION

Condition based maintenance (CBM) is the first step toward e-Maintenance practice. It is important to note that e-maintenance is more than a collection of tools and techniques joined to enhance maintenance, it must be seen as a complete system which must be dynamic and flexible and able to interact with CBM technologies. Companies are moving from traditional corrective and preventive maintenance program to CBM to reduce the maintenance cost and unnecessary maintenance schedules. A CBM program consists of three key steps [13]:

1. Data acquisition, to obtain data relevant to the system health
2. Signal processing, to handle the data or signals collected in step 1 for better understanding and interpretation of the data,
3. Maintenance decision making, to recommend efficient maintenance policies based on diagnosis and prognosis extracted from the data.

A CBM programme essentially forms part of the e-maintenance system, as the assessment of machine's performance information requires an integration of different components health status and the performance requirements. For achieving near zero down time, near zero defects, instantaneous response, decision-making and world-class OEE performance prognostics and diagnostics are used through embedded sensors and device to business tool. All these needs have led to e-health card for equipment's degradation assessment, which forms part of e-maintenance.

For an integrated e-maintenance improvement programme, the information logistic as described below needs to be streamlined [1]

- Right information (in right quantity and quality),
- In right formats and form, as per stakeholders requirement,
- To right person,
- In right time,
- At right place

The plant and or equipment health management system (HMS) could consist of condition monitoring (CM) diagnostics and prognostics, and condition based operation and support, to improve the dependability and safety of the technical systems, besides decreasing life cycle cost of operation and support [14, 15, 16]. This system delivers data and information, which indicates the health condition of the system. The stakeholders of the system are the receivers of the data and information [17, 18, and 19]. The problem today in a health management system is the existing information islands, i.e., the different specialized systems, within an organization speaking different data and information languages. In order to destroy these information barriers some objectives have to be accomplished.

A stakeholders requirements based health management system (HMS) framework is given at Fig. 1[1]. With increasing use of condition monitoring, data collection, and internet in management of maintenance process, the information logistic is required to be streamlined. Condition

monitoring uses various intelligent health monitoring techniques to monitor and control the health status of plant and machineries by analysing the data after it has been collected. The identification of effective and efficient strategies for the maintenance of a plant and machineries is of a major importance from global competition, safety and financial point of view. Today, most of the organizations are trying to follow the condition based preventive maintenance, based on the state of component degradation. However, in reality, the relevant parameters behind the degradation process are very complex, and needs to be undertaken analytically.

Other aspects of enhanced maintenance effectiveness are to integrate the ICT with the strategy and objectives of the organization with that of the maintenance division. This will facilitate the management with effective decision making.

ICT is changing the way we communicate; it is not only connecting us to new people, but developing a global network for conversation and facilitating the mechanism of feedback. ICT with its communication capacity can dramatically improve the standard of information and can create a new social and economic network. ICT is global; as it creates a global network, applied to the whole range of human activities, encourages the dissemination of information and knowledge regardless of geographic boundaries, and is low cost, can therefore lead to substantial efficiency gains.

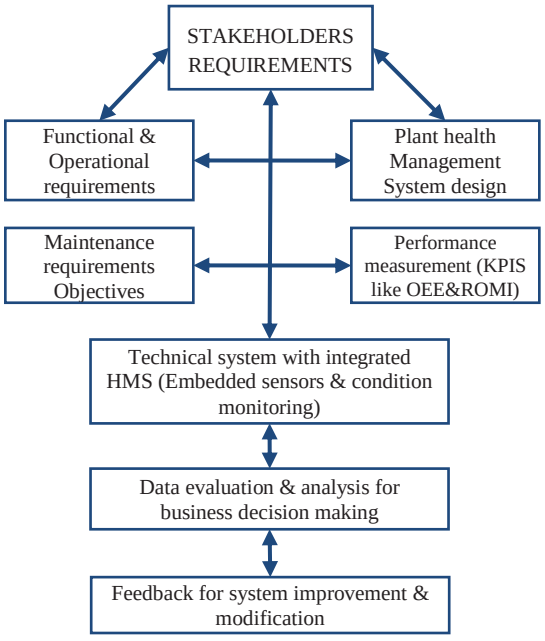


Fig. 1. A stakeholder based health management system (HMS) framework. (Adapted from Health management of Complex technical systems [20])

Integration has been addressed this far largely from the view point of representing the collected information to the end-user (operator or manager) in an effective manner, i.e. bridging the gap between information collected from plants and equipment and the enterprise resource planning (ERP) platforms. According to [21], initiatives have been developed which integrate open, industry-driven, integrated solutions using big data analyses tools for asset management. Such systems provide an information schema at the application-level and an application programming interface (API) to communicate with the underlying protocol stack (e.g., the TCPIIP suite). To our knowledge, existing communication technologies are not well-suited for reliable and timely delivery of appropriate data between distributed end-systems in industrial environments; this, in our opinion, remains a critical missing link in the seamless integration vision.

A. Integration of data sources

The main function of CBM is to monitor the operation of equipment by condition monitoring (CM), and to analyse the sensor data by comparing with normal state parameters based on historical knowledge of the equipment. If failures are detected, CBM will determine the fault location and fault type via its diagnostic function and then make maintenance implementation according to the maintenance strategy.

The modules of system architecture of CBM are presented as follows:

- Physical layer: It consists of a variety of equipment and component parts.
- Information acquisition layer: It acquires running state of equipment by setting up various sensors, filtering and amplifying the sensor data, and submitting these data to the information processing layer. It consists of various sensors, information acquisition terminals, direct numerical control and other intelligent devices.
- Information processing layer: It is to process information provided by the information acquisition layer and to support the function of application layer. The processing includes identification, transformation, classification, feature extraction, feature fusion, etc.
- Data layer: It consists of a variety of database such as maintenance database, knowledge database and equipment information database. It stores maintenance operations, maintenance plans, maintenance events, reference values, etc.
- Application layer: It consists of online monitoring module, troubleshooting module, failure prediction module and maintenance management module. Its function is to display the running state to users, perform fault diagnosis and prediction，and implement the maintenance management.
- User layer: It can be divided into three types: administrator, operator and serviceman.

This modules listed above are based upon a system developed for the United States military. The framework for the next generation machinery monitoring and diagnostic systems, named Open System Architecture for Condition Based Maintenance (OSA-CBM). This comprised of 7

functions which would request data directly from any other layer as needed. The functions are: Data Acquisition, Data Manipulation, Condition Monitor, Health Assessment, Prognostics, Decision support and Presentation [22]. However, implementing the standard is often a difficult task data processing aspects, such as Fast Fourier Transform (FFT) algorithm, k-means clustering, Bayesian reasoned [23]. Therefore a more simplified system with the key functions is required.

The complete workflow of data integration in CBM is shown in Fig. 2. As shown in this figure, fault diagnostics and prognostics are two important steps. Fault diagnostics includes fault detection, classification and identification, where fault detection is a task to indicate faults, fault classification is to locate the faulty component or the parts of equipment, and finally fault identification determines the nature and causes. Fault prognostics deals with fault prediction, in order to determine whether a fault is impending and estimate how quickly and how likely a fault will occur. That is, diagnostics is posterior event analysis and a prognostic is prior event analysis. Diagnostics will be combined with prognostics to achieve an almost zero-downtime performance [24].

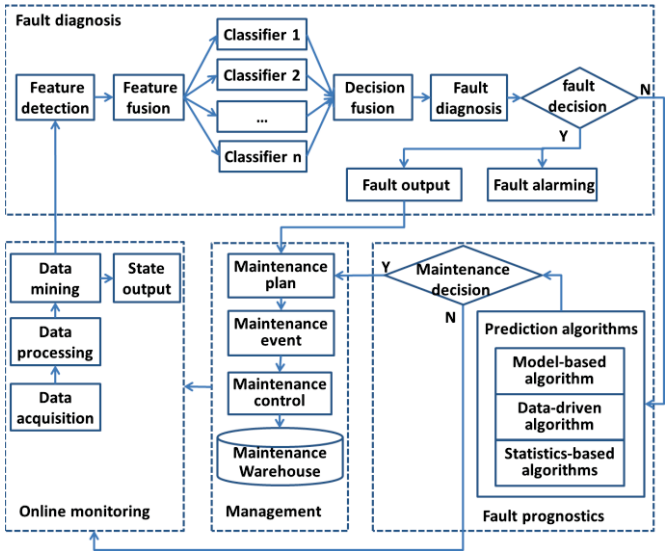


Fig. 2. A workflow of data integration in CBM.

B. Analysis and Correlation

Preventing data overload and closing the gap between information islands are the key requirements. Data overload is one of the common drawbacks noticed in most of the organization and this creates serious problem for the maintenance manager for analysis of the right information. If the data is not specified or prioritized for the decision making, it creates a tremendous work overload, missing the right information. The other issue of Information Island speaks of having individual excellence in isolation. Lack of integration of these excellences, in any organization is a waste of resources, not only from maintenance point of view, but from organizational point.

We tie these maintenance requirements in the light of ICT system constraints. We foresee the following constraints and challenges in the design, development and deployment of an e-maintenance system from an ICT perspective.

a) In an industrial environment, the ICT deployment is seen as heterogeneous – the types of plants and equipment being monitored, the types of computing devices involved the physical media of communication and the nature of access. Hence, the one size fits all paradigms is often seen (incorrectly) as inapplicable, e.g., standard commercial-off-the-shelf communication equipment (such as Bluetooth and IEEE 802.11) and standard Internet protocols (such as the TCP/IP suite) are not suitable for implementing the entire system. For seamless operation, it is important that e-maintenance platforms account for this heterogeneity and operate on standard as well as proprietary protocol stacks. Further, heterogeneity in terms of network capacity should be addressed. Wireless networks are resource-constrained (in terms of limited bandwidth, battery-powered devices with limited processing and storage capacity) as compared to the wired counterparts. Avoiding data overload becomes significant in such networks with scarce resources.

b) Given the challenging, hostile environment in which computation and communication will be carried out, network survivability - the robustness of a communication network in preventing failures, and in case of failure, its ability to gracefully degrade to a state where it can still operate optimally within the constraints of available resources is of primary importance. Network survivability can be viewed as comprising two complementary mechanisms: prevention methods that minimise the probability of a communication network being disrupted by failure, and mitigation methods that limit the damage when a failure occurs. Mitigation can be implemented via network design approaches based on redundancy, e.g., using route redundancy to reroute data flows in case of failure of one or more routes. Redundancy, of course, can be expensive and hence is limited in its extent. More importantly, both methods call for innovation in development of new protocols and mechanisms, both at the application level as well as underlying networking layers that are fault tolerant and provide feedback options to the e-maintenance system. The idea is to allow the e-maintenance system to recover and sustain itself with minimal human intervention, in the wake of failures.

c) In harsh industry environments, we anticipate intermittent connectivity among network devices as being the rule rather than the exception. This clearly rules out the use of traditional TCP/IP based protocols to transport and route data in these networks, since they assume connectivity between end-points of a data flow.

Given the ad hoc nature of communication, an opportunistic communication architecture is necessary - one that takes advantage of existing connectivity to optimize data transfer among network devices.

d) Cognitive radios are set to define wireless access in the future [25]. Cognitive radios are intelligent and flexible in that they can adapt their spectrum use in response to the operating environment, identify spectrum that is unusable under current conditions and enable efficient spectrum utilization. These radios have the potential to radically improve the way wireless networks operate, and are a promising choice for building future industrial ICT infrastructure. Given that sensor networking will be a crucial part of such infrastructure, one important task is to explore if the new generation of cognitive, smart radios can be integrated onto miniature sensors for facilitating robust, energy-efficient sensor networking.

e) The existence of ambience intelligence (sensors) in the environment should be used for ubiquitous, pervasive, context-aware (e.g., location of equipment and personnel, situation: normal vs. emergency, etc.) computing. To our knowledge, most existing e-maintenance platforms lack context-awareness. The need is to develop new generation of e-maintenance platforms that fully utilize context-awareness to pre-process gathered data and to disseminate proper information, in proper amount and at proper time, in tum alleviating problems such as data overload and occurrence of "information islands".

f) Given that the data being monitored and transmitted could be of varying levels of importance (from mission critical to casual), there is a need to provide differentiated service while collecting and transferring such data. This term is called Quality of Service (QoS) which means to classify various applications into different service classes, assign different priority levels to the service classes and allocate different amount of resources to those classes. In computer networks, QoS mechanisms have been primarily used for providing performance assurances to the different types of applications. These mechanisms can be tailored to specific needs of e-maintenance; they can also be used to address the data overload problem as well as to enhance network survivability via graceful degradation when network resources become scarce.

IV. Conclusion

Competitive pressures found within manufacturing has forced organisations to examine systems, strategies, tools and techniques to increase asset efficiency and effectiveness, Management are now aware that for decades manufacturing and maintenance data had been collected yet rarely utilised due to the large amounts of data and the uncertainty of what to analyse and how to decipher the data to ensure the data is supporting new approaches to manufacturing and maintenance. However, organisations are aware that computing resources have increased in capacity and computational speed while decreasing in cost. This has allowed Big Data collection and analyses techniques to improve asset monitoring and management. Indeed big data algorithms will be directed to Data Analytics, Data Based Models and Decision making algorithms. These algorithms will aid to the asset maintenance and wearing cost assignment since current traditional methods are not able to handle all the data captured from infrastructure due to its volume, velocity and variety.

In order to adapt an approach to using big data tools and techniques to support an advanced maintenance strategy development it is important to take full advantage of recent advances in information technologies related to CBM, software and semantic information to develop an effective information and communication infrastructure. While implementing an e-maintenance system, a thorough understanding of the requirements and constraints in conjunction from maintenance and ICT perspectives is necessary.

In this paper, benefits and challenges of big data implementation have been identified in order to achieve optimized maintenance. The main benefits include detecting and predicting product failures, reducing operation expenses, and improving maintenance reliability. However, the challenge is not to collect as much data as possible but to collect, store and analyse the necessary data to make informed decisions based upon accurate and up-to-date data.

References

[1] A. Parida and U. Kumar, "Managing Information is key to Maintenance Effectiveness," in *Proceedings of lntelligent Maintenance System*, Arles, France, 15-17 July, 2004.

[2] J. Kurtz, P. Hoy, L. McHargue, and J. Ward, "Improving Operational and Financial Results through Predictive Maintenance," *Smarter Analytics Leadership Summit*, New York, USA, February 21, 2013.

[3] F. Niedermann, and H. Schwarz., "Deep Business Optimization: Making Business Process Optimization Theory Work in Practice," in *Proceedings of the 12th International Conference, BPMDS 2011, and 16th International Conference, EMMSAD 2011, CAiSE 2011*. Berlin: Springer, 2011, pp. 88–102

[4] K. Wang, S. Tong, and B. Eynard, "Review on Application Data Mining in Product Design and Manufacturing," in *Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery : FSKD 2007*. Los Alamitos: IEEE Computer Society, 2007, pp. 613–618.

[5] D. Agrawal, S. Das, and A. Abbadi, "Big data and cloud computing: Current state and future opportunities." *In Proceedings of the 14th International Conference on Extending Database Technology*, Uppsala, Sweden, March 2011 pp. 530–533.

[6] M. Chen, S. Mao, Y. Zhang, and V.C.M. Leung, "Big data. Related technologies, challenges and future prospects," *Springer*, 2014, 89 p., ISBN 978-3-319-06245-7

[7] C.L.P. Chen, and C.-Y. Zhang, "Data-intensive applications, challenges, techniques, and technologies: A survey on Big Data," *Information Sciences*, http://dx.doi.org/10.1016/j.ins.2014.01.015, 2014.

[8] Aspera, "Solution Brief: IBM Aspera High-Speed File Transfer," *IBM Software*, Sept 2014.

[9] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinoum, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big data and its technical challenges," *COMMUNICATIONS OF THE ACM*, vol.57, no.7, pp. 86-94, July 2014.

[10] E. Begoli, and J. Horey, "Design principles for effective knowledge discovery from big data," In *Proceedings of the Software Architecture (WICSA) and European Conference on Software Architecture (ECSA)*, Helsinki, Finland, August 20-24,2012 pp. 215–218.

[11] H. Chen, R.H.L. Chiang, and V.C. Storey, "Business intelligence and analytics: From big data to big impact. MIS Quarterly," Vol. 36, Iss. 4, pp. 1165–1188, 2012.

[12] B. Nedelcu, "About Big Data and its Challenges and Benefits in Manufacturing," *Database Systems Journal*, vol IV, no. 3/2013, pp. 10-19.

[13] J. Lee, A. Ramji, K. S. J. Andrews, L. Darning and B. Dragan, "An integrated platform for diagnostics, prognostics and maintenance optimization," in *e-Proceedings of Intelligent Maintenance System*, Arles, France, 15-17 July, 2004

[14] R. K. Mobley, *An introduction to predictive maintenance*, Van Nostrand Reinhold, New York, 1990

[15] J. D. Campbell and A. K. S. Jardine, *Maintenance excellence: optimizing equipment life-cycle decisions*, Marcel Dekker, New York, 2001

[16] P. Soderholm and P. A. Akersten, "Aerospace Diagnostics and Prognostics in a TQM Perspective," in *Proceedings of the 15th International Congress of COMADEM 2002*, Birmingham, UK, 2-4 September, pp. 80-89

[17] K. Lyytinnen and R. Hirschheim, "Information system failures: a survey and classification of the empirical literature," Oxford surveys in information technology, Oxford University Press, Oxford, 1987, pp. 257-309

[18] ISOIIEC 15288, "Systems Engineering: System Life Cycle Processes," International Organization for Standardization, Geneva Commission Electrotechnique Internationale,Geneva, 2002.

[19] P. Soderholm, "Continuous Improvement of Complex Technical System: Aspects of Stakeholder Requirements and System Functions," Licentiate Thesis, Division of Quality and Environmental Management, Lulea University of Technology, Lulea, 2003.

[20] P. Soderholm, and A. Parida, "Health management of complex technical systems," in Proceedings of *the 17th International Congress of COMADEM*, 23-25 Aug, 2004, Cambridge, UK, pp. 214-221.

[21] J. Lee, E. Lapira, S. Yang, and HA. Kao, "Predictive manufacturing system trends of next generation predictive systems," in *Proceedings of the 11th IFAC workshop on intelligenrt manufacturing systems*. vol 11(1): 2013 Pp. 150-156

[22] M. Lebold, and K. Reichard, "OSA-CBM Architecture Development with Emphasis on XML Implementations," in *Maintenance and Reliability Conference (MARCON)*, May 6-8, 2002.

[23] A. Helsinger, and T. Wright, "Cougaar: A Robust Configurable Mulit-Agent Platform," in *Proceedings of the IEEE Aerospace Conference*, Big Sky, MT, 2005.

[24] M. Polczynski, and A. Kochanski, "Knowledge Discovery and Analysis in Manufacturing," *Quality Engineering*, vol. 22 (3), pp. 169–181, 2010.

[25] FCC, Cognitive Radio Technologies Proceedings, *Federal Communications Commission (FCC)*, May 2003.

Financial Footnote Analysis: Developing a Text Mining Approach

Maryam Heidari[1] **and Carsten Felden**[1]
[1]Information system department, University of Freiberg, Freiberg, Saxony, Germany

Abstract - *Financial footnotes analysis provides an opportunity to communicate with stakeholders beyond the numbers in the main body of financial statements. The combination of values in financial reports and their disclosure in footnote parts supports financial decisions in a wisely manner. Nevertheless, the unstructured nature of footnotes poses a barrier for an accurate, automatic, and real-time financial analysis. To address this issue, this paper implements a text classification procedure to evaluate the benefits of text mining deployment to react to the manual financial footnote analysis. This supports the classification of textual parts of financial footnotes automatically into related financial categories, which are relevant for financial analysts, in order to avoid reading entire textual parts manually. This research provides useful insights about the impact of using text mining for an automatic financial footnote analysis in terms of time saving and increasing accuracy.*

Keywords: financial footnotes, text mining, text classification, income tax

1 Introduction

Analyzing financial disclosures is a key mechanism, which facilitates communication of financial analysts, auditors, internal and external decision makers, and other stakeholders gaining benefits from analyzing financial reports [1]. Using solely financial statements in this context does not represent the comprehensive financial story of a company. Financial footnotes provide useful information about company's financial performance [19], [25], [28]. However, the unstructured format of financial footnotes makes it difficult to analyze them automatically. A manual analysis is still a time-consuming issue, restricted in terms of accuracy and real time analysis. Based on 45 financial analysis research papers, Heidari and Felden show that there is no identified method to integrate financial analysis methods of structured values with unstructured footnotes automatically [14]. Furthermore, the widely usage of XBRL as a standard platform for financial data exchange, even with the detailed tagging process, has no impact on an automatic financial footnotes analysis [14], [31]. To overcome the identified bottleneck and to facilitate financial footnote analysis, our study suggests a text-mining approach and applies text classification procedures. The paper's goal is to assess to what extent and how text mining application in footnotes can support financial analytical tasks.

Existing approaches in financial analysis literature shows the variety usage of text mining in financial market prediction such as economic crises, stock price prediction, and risk management [5], [13], [12], and [20]. However, no identified article addresses a text mining application for financial footnotes analysis. Most related research in this area suffers from the automatic solution for financial footnotes and rely on manual information extraction [14]. There exist a number of attempts based on coding schemas to analyze financial disclosures [3]. But it seems to be obvious that a manual code assignment has no advantage for an automatic footnotes analysis [16]. This study tries to make a contribution regarding financial analysts and any stakeholders who benefit from financial analysis by applying a text mining approach based on pre-defined classes in order to assist them in required financial information extraction from footnotes in a real time and automatic fashion.

The paper proceeds as follows: Section 2 discusses related work in terms of existing solutions in literature to deal with unstructured financial information. Section 3 presents the details of the applied research method and the proposed text mining framework. Section 4 demonstrates the empirical implementation of text classification techniques and the results of applying the developed framework including the validation results. Section 5 concludes with implications of the research results and further research directions.

2 State of the art

Today, due to existing various financial analysis software, which provides companies with the financial information to make better decisions, an automatic analysis of financial figures is straightforward. However, it is difficult to analyze textual parts of financial reports automatically to explicate company's financial behaviors and to identify valuable information about the current and future financial status of a company [17], [18].

In the financial footnotes analysis domain, a literature review by Heidari and Felden identified the importance and the effect of financial footnote information on financial analysis processes [14]. However, existing methods for financial footnotes analysis rely on manual processes and there is a strong need to develop an appropriate solution to support footnotes analysis automatically within an integrated financial analysis process [14].

Since financial footnotes have textual format and consist of soft financial information, we reviewed existing literature

regarding text mining in financial analysis area as well. This means, according to Miner et al., particular characteristics and the purpose of text mining, information retrieval, concept extraction, clustering, and classification as typical text mining approaches [21]. As an example in terms of financial analysis, Beattie et al. introduced a comprehensive four-dimensional framework for content analysis of accounting narratives [3]. It uses a coding system based on four attributes in order to give structure to accounting texts. They performed this framework by qualitative research software in order to support the coding procedures via an index system. The advantage of this coding procedure is that it focuses not only on the topic, but also considers the different kinds of attributes. However, this does not have benefit for analyzing financial footnotes automatically, because of manual code assignment to the texts [16].

Among different researches in the financial analysis area, some articles concentrate on text classification techniques, which can be trained for recognizing and differentiating significant categories in documents automatically. Most related research in this area comprises text mining processes, which translate unstructured financial information from numerous text references into a machine readable format for predictive purposes. For example, Brent and Atkisson built a coding scheme by training pre-defined categories assigning code to text documents automatically in order to analyze economic crises through newspaper articles [5]. Neumann et al. designed a text classification approach to process financial news to automate stock price prediction [13]. It bases on three main text processing step: Dataset as a basis for the classification, feature processing to extract different features and generating machine readable information, and finally the machine learning step using a subset of data to train a classification algorithm to be able to response to the stock market trends. Another related study performed by Li, who employs several pre-defined dictionaries to predict stock feedbacks based on US corporate filings [20]. In another attempt, Groth et al. focus on German announcement to evaluate stock price effects [12]. It can be recognized that in the most related research in financial analysis area text mining approaches are used to classify financial news into positive and negative categories to have verifiable market trend [13].

Regarding reviewed articles, it can be mentioned that text mining methods are widely used in financial market trend prediction. Nevertheless, in financial footnote analysis area, the most used method bases on manual procedures. Due to human interference and probability of neglecting relevant content, the manual procedures is extremely time-consuming and error-prone. Therefore, it seems to be appropriate to identify an automatic-based solution to overcome this gap and to compose financial analysis processes reliable as well as accurate. Besides, text mining research shows that using classification techniques is an appropriate solution for analyzing information in textual formats. However, it is not enough to apply just a text classification algorithm, but there is a need to use the classified results regarding the main body of financial reports.

Regarding the identified gap, we demonstrate the utility of a text mining application in automatic financial footnote analysis through implementing a text classification workflow in the next section.

3 Research method

Concerning text mining application for financial footnotes analysis, we applied the Cross-Industry Standard Process (CRISP) process flow to define a complete lifecycle of the text mining workflow [7]. CRISP methodology is based on six phases providing a comprehensive coverage of all activities involved in data/text mining projects [21]. Fig. 1 illustrates the cyclic form of the CRISP process flow.

Concerning reviewed literature and regarding the purpose of this research, Fig. 2 shows the application of the proposed financial footnote analysis. After data preparation, the financial patterns in the narrative part of financial reports are recognized. Later on, well-known classification algorithms are applied and evaluated according to accuracy performance and other obtained results. We assess to what extent text classification method assists and facilitates financial footnote analysis considerably.

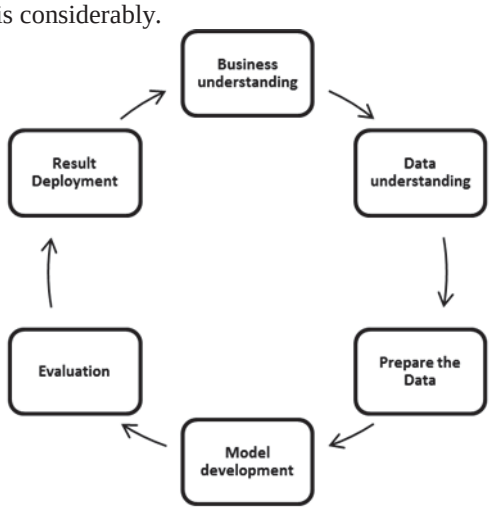


Fig. 1. CRISP process flow [7]

3.1 Data collection

The preliminary step is the collection of financial footnotes. We focused on one footnote item in financial reports: the income tax footnote, which is a material component of most financial statements. Income tax accounting requires the use of estimates, judgments, and other subjective information that cannot be fully discovered in the financial statement reports [11]. According to Graham et al. the income tax footnote can enable users to gain a better understanding of the income tax status of a company [11]. The used database to get financial footnotes of companies' filings was Edgar online of U.S. SEC[1]. This database provides free public access to corporate

[1] U.S. Securities and Exchange Commission

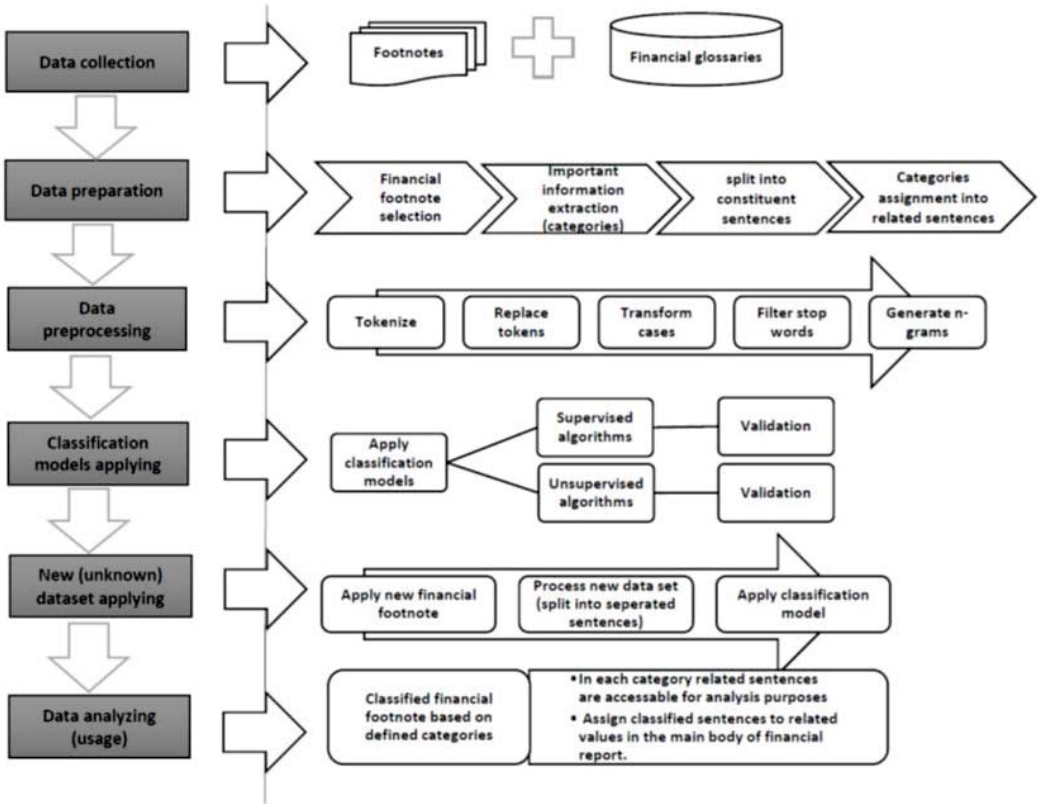


Fig. 2. Proposed text classification procedures for analyzing financial footnotes

information, allowing quick research on a company's financial information by reviewing registration statements and periodic reports filed on forms like 10-K (annual reports) and 10-Q (quarterly reports). Furthermore, existing financial glossaries and references are required to recognize important terms and issues not appearing in financial statement reports and which can be hidden in footnote lines.

3.2 Data preparation

We observed income tax footnotes from 120 different companies in different industries. First, we read some of the extracted income tax footnotes and compared them with existing financial glossaries and financial and accounting audit references in order to recognize financial terms which cannot be found in the main body of financial reports. The six identified categories related to income tax footnote are demonstrated in Table 1.

Table 1. Identified categories in income tax footnote

Income tax footnote	Deferred tax
	Effective tax rate
	Net operating loss
	Unrecognized tax benefit
	Taxation authority
	Valuation allowance

Afterwards, we split extracted footnotes into constituent sentences in order to assign all sentences to their related categories. This process supports not only recognizing rates and trends in footnotes' line, but also identifying soft information through sentences and their relations to financial figures in the main report body. Accordingly, we have extracted totally 1,290 sentences. Each sentence belongs to one particular pre-defined income tax category. It should be noted that alongside these six categories, other financial issues might appear in income tax footnotes, which can be categorized under one of these main categories. As an example, financial terms such as permanent and temporary differences, litigation charge, and net deferred tax asset are some income tax issues which can be classified under the deferred tax category.

3.3 Data pre-processing

The next step encompasses ordinary text mining preprocesses such as lower case transformation, stop words filtering, and tokenization. During the tokenization process, each sentence splits into a sequence of tokens in order to build a word vector for each sentence. The replace tokens operator allows replacing substrings within each token. To that end, the user can specify arbitrary patterns in the replace dictionary parameter. Table 2 shows some replacement examples, which have been applied to income tax footnotes analysis process.

Table 2. Sample replacement list in income tax footnote

Replace what	Replace by
jurisdiction	authority
exemption	deductibility
tax gain	tax benefit
unremitted	undistributed
uncertain tax position	unrecognized tax benefit
U.S.	united states
…	…

Afterwards, English stop words are removed from a document by checking every token that equals a stop word from the built-in stop word list. Finally, in the last pre-processing step, n-grams of tokens in a document are generated. A term n-gram is defined as a series of consecutive tokens of length n.; we defined n=3, because some financial terms consist of three relative words.

3.4 Apply classification model

Due to two main existing training concepts in the area of text classification methods, we applied both supervised and unsupervised machine learning algorithms. Although, in both methods, textual parts of financial footnotes are classified based on particular similarity criteria; in supervised algorithms we utilize training data sets (last two phases) including pre-defined classes in order to train the classification model and to classify new and unlabeled documents. As opposed to supervised algorithms, in unsupervised algorithms similar clusters are discovered without using training or labeled documents [9], [22], [29], and [30]. Both algorithms should be validated to observe the performance results, which assist users to employ more appropriate techniques for financial footnote analysis [15].

3.5 Apply new dataset

In this phase, the identified algorithm will be used for new financial footnotes. We processed this phase by splitting the financial text into constituent sentences. As an example, if our new income tax footnote is a document text file including 50 sentences, it will be split into 50 separated text files. Each one consists of one sentence and serves as input data set to our process. Later on, the approved classification model will be applied in order to classify income tax footnotes based on the recognized categories.

3.6 Data analyzing (usage)

It should be mentioned that the final purpose behind text classification processes is to facilitate text analyzing in order to recognize hidden textual patterns with reduced manual efforts [4]. To do this, we focus on utilization of classified financial footnotes in the last phase of the research, which consists of representation of classified output in each category. Related sentences in each category are accessible for further analysis, which helps analysts to extract required terms and related sentences without reading the whole footnote. Another usage of financial footnotes classification is the assignment of classified sentences to related values in the main body of financial statements like balance sheet or income statements.

4 Implementation and results

The implementation has been done by using the tool Rapid Miner[2], which is an open source tool for data mining and predictive analytics. All the above-discussed steps have been performed in this tool. As mentioned earlier, in order to obtain the best results, we applied both supervised and unsupervised algorithms to recognize the more appropriate one for our research goal.

The basic objective of all supervised classifiers is to recognize the degree of similarity between pre-classified training data and a new, unlabeled data set [6], [10], [23]. To do this, we preprocessed our training documents and trained the supervised model based on defined steps. We tested various supervised classification algorithms such as K-NN, Naïve Bayes, Support Vector Machine (SVM), and decision tree. After checking performance measures like accuracy, run time, absolute error, and Root Mean Square Deviation (RMSE)[3], we ended up to Naïve Bayes as the most appropriate supervised classifier for financial footnote analysis with 82.86% accuracy and the most minimum run time (Table 3).

Table 3. Results of supervised algorithm

Supervised algorithm	Run time	Accuracy	Absolute error	RMSE
K-NN	7s	81.82%	0.183	0.362
Naïve Bayes	4s	82.86%	0.171	0.414
SVM	28s	79.22%	0.784	0.786
Decision tree	1m45s	90.65%	0.136	0.280

In terms of unsupervised algorithms, we performed K-means as a clustering technique, which is used for extracting information from unlabeled data. According to an experimental study by Steinbach et al., among clustering algorithms, K-means method has better performance for text clustering. They compared two main approaches in clustering techniques: agglomerative hierarchical clustering and different types of K-means [27]. They argued that the hierarchical clustering does not work well because most of the time (like our case) it cannot be fixed by the hierarchical scheme. In contrast, K-means groups objects together that are similar to each other and dissimilar to the objects belonging to other clusters [22]. For this method of clustering we start by

[2] http://www.rapidminer.com

[3] Root Mean Square Error is a frequently used measure of the differences between value predicted by a model or an estimator and the values actually observed

deciding how many clusters (K) we would like to form based on our data. We set K to 6 based on our classes in supervised classification. Different performance criteria such as run time, Davies-Bouldin and average within centroid distance are supervised (Table 4).

Table 4. Results of k-means unsupervised algorithm

Unsupervised algorithm	K	Run time	Ave. within centroid distance	Davies-Bouldin
K-means	6	7min 20s	-0.827	-4.763

According to the obtain results, applying unsupervised algorithms can avoid relatively large amount of supervision and manual tasks [9], [29]. Nonetheless, regarding to the financial analysis purposes where searching particular financial terms in specified footnotes is significant, classification through supervised algorithms yield more reliable results. Furthermore, applying unsupervised methods are more appropriate n case of working with not very clean texts such as large amount of text data created by dynamic applications such as social networks [2]. Fig. 3 summarizes strengths and weaknesses of each machine learning algorithm based on this research criteria.

	Strengths	**Weaknesses**
Supervised algorithm	Classification based on pre-defined and specific required categories High accuracy Short run time	Complex data training process
Unsupervised algorithm	Not required training process Acceptable accuracy in related clusters	Clustering process happens not based on desired categories Long run time

Fig. 3. Comparison between supervised and unsupervised algorithms in this research

Regarding the analyzing step in terms of facilitating the usage of financial footnotes, two main benefits are notable. The first contribution of performing the text classification process is that analysts or any stakeholders can access financial footnotes in a regular and ordered fashion. Table 5 shows an example of the output report for one of the major income tax category so called unrecognized tax benefit which consists of extracted sentences from income tax footnote of a company. Another benefit of the text mining approach to facilitate financial footnote analysis is the mapping of extracted sentences to related values in the main body of financial statements in order to perform a fully automatic process. In our case, all extracted sentences are related to "deferred income tax" (sometimes appears separately as deferred tax asset and deferred tax liability), which is one of the balance sheet report's term. As a result, the proposed text mining approach can be determined as an appropriate semi-automatic solution to overcome time-consuming manual analysis of financial footnotes by classifying footnotes based on pre-defined categories automatically. Thus, analysts can directly access required sentences of related categories instead of reading the entire text document in a real-time. The proposed text classification method can be seen as a starting point for further research to evaluate practically the influence of this solution on analysts' workflow in financial analysis process in a real world. We acknowledge that proposed text mining approach could still be improved by identifying other relevant footnotes items and by mapping extracted sentences to values in the main body of financial statements in order to implement fully automatic process.

Table 5. Extracted sentences related to unrecognized tax benefit from income tax footnote of a company

Income tax footnote	Number of sentences	Related sentence
Unrecognized tax benefit	1	Differences between tax positions taken or expected to be taken in a tax return and the net benefit recognized and measured pursuant to the interpretation are referred to as "unrecognized benefits.
	2	A liability is recognized for unrecognized tax benefit because it represents an enterprise's potential future obligation to the taxing authority for a tax position that was not recognized as a result of applying the provisions of ASC 740.
	3	If applicable, interest costs related to the unrecognized tax benefits are required to be calculated and would be classified as "Other expenses – Interest" in the statement of operations.
	4	As of October 31, 2014 and October 31, 2013, no liability for unrecognized tax benefits was required to be reported.
	5	The Company does not expect any significant changes in its unrecognized tax benefits in the next year.

5 Conclusion

Despite significant developments in fields of financial analysis e.g. based on XBRL-formatted data, the textual parts of financial reports, which are critical for comprehensive financial analysis, are still dependent on time-consuming manual procedures.

We addressed this challenge in this paper by proposing a text mining approach to be able to automate financial footnotes analysis and facilitate the usage of footnotes information by applying text classification methods. We have chosen income tax as a footnote example. We classified it sentence by sentence with supervised and unsupervised classification algorithms and implemented the proposed solution using an analytics tool. In terms of accuracy and run time, a supervised algorithm gains better results, however it requires a precise and careful data training process.

Comparing to existing approaches in terms of financial footnote analysis, our preliminary results show that text mining could be an appropriate semi-automatic solution to facilitate manual analysis of unstructured parts of financial reports. As a matter of fact, the text mining approach helps users to access required soft information as a separate sentence based on each financial pre-defined category.

It is also of interest in this research to develop this solution by adding more capabilities thereby to map extracted sentences into related figures in financial statements.

However, due to some limitations, it is not practically implemented, yet. One of the limitations is that footnote parts are normally received by analysts or auditors as separate documents and are not attached to main financial statements. This makes it difficult to map financial sentences into figures in financial statements. Another limitation is that some financial footnote sentences carry only some informative and explanatory data about financial terms and status of the organization and therefore cannot be connected to any financial values in the main body.

There is still room for an improvement concerning the automatic financial footnote analysis by evaluating this approach thorough different case studies and expert interviews to demonstrate the usefulness of this approach in accordance to analysts' processes. This is one of the future research areas of this research.

6 Reference

[1] Abahoonie, E. et al. (2013, December) Tax accounting services: Income tax disclosure, available at: http://www.pwc.com

[2] Aggarwal, Charu C., and ChengXiang Zhai. "A survey of text clustering algorithms." Mining Text Data. Springer US, 2012. 77-128.

[3] Beattie, V., McInnes, B. & Fearnley, S., 2004. A methodology for analyzing and evaluating narratives in annual reports: a comprehensive descriptive profile and metrics for disclosure quality attributes. Accounting Forum, 28(3), pp.205–236.

[4] Botzenhardt, A., Witt, A., & Maedche, A. (2011). A Text Mining Application for Exploring the Voice of the Customer. AMCIS 2011 Proceedings.

[5] Brent, E.; Atkisson, C., 2011 " A standard-based automated coding program for unstructured text" Veyor ® Survey presentation, University of Surrey, USA

[6] Chaovalit, Pimwadee, and Lina Zhou. (2005) "Movie review mining: A comparison between supervised and unsupervised classification approaches." System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference.

[7] Chapman, P.; Clinton, J.; Kerber, R.; Khabanza, T.; Reinartz, T.; Shearer, C.; Wirth, R. (2000) " CRISP-DM- step by step data mining guide." SPSS, Chicago, IL.

[8] Crammer, K.; Dredze, M.; Ganchev, K.; Talukdar, P. P.; Carroll, S., 2007, "Automatic code assignment to medical text" In Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, pp. 129-136. Association for Computational Linguistics.

[9] Dharmadhikari, S. C., & Kulkarni, P. (2011). Empirical Studies on Machine Learning Based Text Classification Algorithms. Advanced Computing: An International Journal, 2(6), 161–169.

[10] Ghosh, S., Roy, S. & Bandyopadhyay, P.S.K. (2012) A tutorial review on Text Mining Algorithms. , 1(4), pp.223–233.

[11] Graham, John R. & Raedy, Jana S. & Shackelford, Douglas A., 2012. "Research in accounting for income taxes," Journal of Accounting and Economics, Elsevier, vol. 53(1), pages 412-434

[12] Groth, S.S.; Muntermann, J., Supporting investment management processes with machine learning techniques, in: H.R. Hansen, D. Karagiannis, H.-G. Fill (Eds.), Proceedings of the 9. Internationale Tagung Wirtschaftsinformatik, Österreichische Computer Gesellschaft, Wien, Austria, 2009.

[13] Hagenau, M.; Liebmann, M.; Neumann, D.: Automated news reading: Stock price prediction based on financial news using context-specific features. System Science (HICSS), 2012 45th Hawaii International Conference on IEEE. (2012)

[14] Heidari, M.; Felden, C.: Toward Supporting Analytical Tasks in Financial Footnotes Analysis- A State of the Art, In: Multikonferenz Wirtschaftsinformatik MKWI 2014, Paderborn, Deutschland, 26-28, Februar, 2014.
[15] Hotho, A., Andreas, N., Paaß, G., & Augustin, S. (2005). A Brief Survey of Text Mining, 1–37.
[16] Hussainey, K. S. M. (2004). A study of the ability of (partially) automated disclosure scores to explain the information content of annual report narratives for future earnings, Doctoral dissertation, University of Manchester.
[17] Kloptchenko, Antonina; (2004) "Toward Automatic Analysis of Financial Reports- Readability of Quarterly reports & company's financial performance", AMCIS 2004 Proceedings, paper 412.
[18] Kloptchenko, Antonina; Eklund, Tomas; Back, Barbro; Karlsson, Jonas; Vanharanta, Hannu; and Visa, Ari (2002) "COMBINING DATA AND TEXT MINING TECHNIQUES FOR ANALYZING FINANCIAL REPORTS", AMCIS 2002 Proceedings. Paper 4.
[19] Leder, Michele; (2003), "Financial Fine Print: Uncovering a Company's True Value", John Wiely & Sons Inc., New Jersey, 17p.
[20] Li F., The information content of forward-looking statements in corporate filings — a Naïve Bayesian machine learning approach, Journal of Accounting Research 48 (5) (2010) 49–102.
[21] Miner, G.; Elder, J.; Nisbet, B.; Delen, D.; Fast, A.; Hill, T. (2012) Practical text mining and statistical analysis for non-structured text data applications. Massachusetts, USA: Elsevier
[22] Ozgür, Arzucan. (2004) Supervised and unsupervised machine learning techniques for text document categorization. Diss. Bogaziçi University.
[23] Padhye, Apurva. (2006) Comparing Supervised and Unsupervised Classification of Messages in the Enron Email Corpus. Diss. UNIVERSITY OF MINNESOTA, 2006.
[24] Pakhomov, Serguei VS, James D. Buntrock, and Christopher G. Chute. (2006) "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques." Journal of the American Medical Informatics Association 13(5), pp. 516-525.
[25] Putra, Lie Dharma (2008) Understanding footnotes to financial statements, http://accounting-financial-tax.com/2008/08/understanding-footnotes-to-financial-statement
[26] Rapid miner, http://www.rapidminer.com
[27] Steinbach, Michael, George Karypis, and Vipin Kumar. "A comparison of document clustering techniques." KDD workshop on text mining. Vol. 400. No. 1. (2000)
[28] Tergesen, Anne (2002) "Getting to the bottom of a company's Debt", Business Week, 10/14/2002, Issue 3803, p156-158.
[29] Tsarev, Dmitry, Mikhail Petrovskiy, and Igor Mashechkin. (2013) "Supervised and Unsupervised Text Classification via Generic Summarization."
[30] Wagstaff, Kiri Lou. (2002) intelligent clustering with instance-level constraints. Diss. Cornell University.
[31] Weglarz, Geoffrey (2004), "Two worlds of data: unstructured and structured", DM Review, September2004

Product's Quality Prediction with respect to equipments data

M. Melhem[1], B. Ananou[1], M. Djeziri[1], M. Ouladsine[1], and J. Pinaton[2]
[1]LSIS, Aix-Marseille University, Marseille, Provence-Alpes-Cote-d'Azur, France
[2]ST-microelectronics, Rousset, Provence-Alpes-Cote D'Azur, France

Abstract – *The semiconductor manufacturing process is a complex process that consists in a big number of equipments and enormous data. This paper presents a Least Absolute Shrinkage and Selection Operator (LASSO) based method for predicting the product's quality with respect to data of many equipments. The ability of the prediction model allows the product's quality to be estimated in real-time instead of a sampling inspection. An application to data provided by semiconductor manufacturing is presented and the results show the ability of the proposed method to predict the product quality efficiently and effectively with an improvement of more than 90% compared to the multivariate linear regression*.

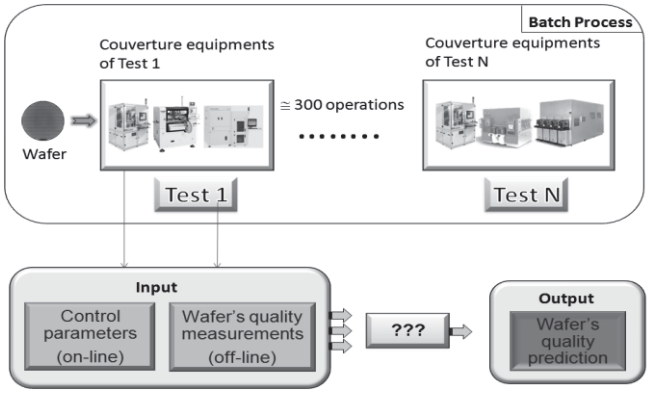
Keywords: Prognosis, RUL Prediction, Semiconductor Manufacturing

1 Introduction

Semiconductor manufacturing is a complex process in which a Wafer goes through hundreds of sequential process steps with different recipes to produce a collection of chips. This process consists mainly of seven steps: Lithography, etching, deposition, chemical mechanical planarization, oxidation, ion implantation, and diffusion. This process is characterized by different types of equipments which are associated with FDC (*Fault Detection and Classification*) databases collected by real-time measurements. For monitoring the production quality, product quality parameters are gathered by sampling testing after manufacturing steps and are applied to evaluate the quality. Thus only the quality of monitored wafers is grasped. Therefore, a failure occurred during the non-sampling periods where other wafers are processed may severely deteriorate the final product's quality, which results in a large number of scrapped products and thus a huge loss of Yield of fabrication. To overcome this problem, an efficient way is to predict the product quality with respect to process parameters and sensor data (Figure 1).

Since final product's quality depends on how it was processed, value series of sensor measurements recorded at each processing step might contain quality-related patterns. Therefore, it is useful to identify quality deviations as early as possible and in real-time by data mining tools on distributed sensor measurements along the process chain [1]. As the number of sensors is enormous, the product's quality prediction is very complicated. Furthermore, the lack of knowledge about the relevant variables that affect the quality makes the problem more difficult. Different types of methods for the prediction of the product's quality are presented in the literature, which can be divided in three categories: expertise-based methods, model-based methods and data-based methods. Although different methods have established huge popularity in the industry, they have some limitations for the semiconductor manufacturing. Engineering knowledge is not always sufficient for building prediction models in this domain, and physical models can't be constructed due to the complexity of the process. So, preferred methods for the wafer's quality prediction are the data-based methods. Data-based methods can be divided in two categories: Statistical based methods and artificial intelligent methods.

Fig 1 : Description of semiconductor manufacturing process and wafer's quality prediction



A number of previous works have been proposed for modelling the manufacturing processes and predicting the associated product quality prediction. A DPNN-based process management system is proposed [2] to predict four quality parameters associated to the ingot fabrication corresponding to control parameters. Multiple regression models and a Bootstrap algorithm were applied to generate sufficient data for ingot prediction. A polynomial neural network is applied

in [3] to construct a predictive model of Plasma Etch process. Two Chemical Vapor Deposition (CVD) predictive models are constructed by a Radial Basis Function Neural Network [4] and a Support Vector Machine [5]. Bayesian Networks were used in [6] to generate causal relations between process variables and wafer quality. Regression methods are applied in [7] [8] to predict the CVD thickness. A quality prognostic scheme is developed in [9] to estimate the sputtering thickness as a processing quality with respect to processing parameters and sensor data in the TFT-LCD manufacturing process. For this purpose, neural networks and Weighted Moving Average algorithms are applied.

The works cited above construct a prediction model of the product's quality at any stage of the process without taking into account the cumulative effect of previous stages. A method is developed in [10] for continual prediction of manufactured microprocessor quality with respect to sparsely sampled control measurements prior to final testing by using an average prediction of linear regression and boosted trees. But this work doesn't consider any data characterizing the fab for prediction. A useful idea is to consider FDC data as additional powerful predictors.

In this paper, a regularized regression model (LASSO) is applied to investigate the influence of equipments FDC data on the measured quality parameters while taking into account the relationship between the quality specifications of previous stages. This method can be considered as a combination of a multiple regression model and a variable selection method and thus, it can construct a prediction model that take into account the cumulative effect of many equipments and avoid overfitting caused by the complex models.

The remaining of this paper is organized as follows: Section 2 summarizes the methods used in the literature for the product's quality prediction with respect to many equipments, and it particularly explains the LASSO-based regression. Section 3 presents the proposed method for online prediction based on the LASSO-based regression. Section 4 provides an example with application in semi conductor manufacturing process to illustrate the feasibility of the proposed method. Finally, section 5 concludes the paper and identifies future work for improvements.

2 Model description

2.1 Literature review

In the semi-conductor manufacturing, a huge amount of high-dimensional and correlated data are collected through many equipments and requires a reduction. Multivariate statistical techniques can be used for feature extraction, like Principal Component Analysis (PCA) [11] and Partial Least Squares (PLS). PCA is used to develop a prediction model from a historical database when product quality data are not available [12]. However, it is able to analyze the correlation between variables in a particular manufacturing stage and thus, it consider the whole manufacturing as happened in a single stage. A solution is recommended in [13] to estimate the effect of each stage on output quality of the next stage by a regression model, and it is applied to mobile phone production line. However, the semiconductor process is complicated and the quality measurements are not always available. For considering the correlation between manufacturing stages, a Cascade Quality Prediction Model is developed in [12] based on the PCA and decision trees. But this requires a significant expert knowledge.

To overcome the shortcomings of the existing methods, a sequential feature extraction method based on the regularized least-squares regression algorithm so called LASSO is proposed in this paper which will improve the prediction accuracy. This algorithm is well suitable to control the large number of variables, it reduces the observable variables to fewer numbers of factors by shrinking the non pertinent variables to zero.

2.2 LASSO regression

Standard linear regression models formulated as (1) work by identifying a set of regression coefficients that minimize the Residual Squared Error between the observed values and the fitted values from the model (equation 2) to obtain the Ordinary Least Square (OLS) estimate.

$$y = X\beta + \varepsilon \tag{1}$$

$$\min_{\beta} \|y - X\beta\|_2^2 \tag{2}$$

Where X(N×P) is the matrix of process variables, y(N×L) is the matrix of quality measurements, β(P×L) is the vector of regression coefficients, and ε is the residual vector. Multiple linear regressions are a particular case where a combination of the predictors that best fit the response is identified.

Given the problem of data correlation and the fact that the number of process variables is very large in many manufacturing processes, many techniques has been developed that deals with such problems. Partial Least Squares regression is used to deals with correlated predictor variables by constructing new components as linear combination of them. This method is usually used when the columns of X are highly correlated and their number is very large. The idea is to decompose the matrices X and Y like in Principal Component Analysis:

$$\begin{cases} X = TP^T + E \\ Y = UQ^T + F \\ T = XW^* \end{cases} \tag{3}$$

Where T and U are the component or factor matrices, P and Q are the orthogonal loading matrices, and E and F are the error terms.

In this way, this technique appears as a mixture of Multiple Linear Regression and Principal Component Analysis, and thus, it can be considered as a way of features dimension reduction.

Many extensions of the PCA/PLS methods were used in the literature for the end-product quality prediction. The Multi-way Partial Least Squares is the most famous method with good applications. However, the MPLS takes all the process as happened in a single stage, and involves all process variables in the model no matter they are critical to the end-product-quality or not. A Least absolute Shrinkage and Selection Operator (LASSO) type regularization were developed in [14] to predict the end-product quality and it overcomes the problems of the MPLS by selecting the critical-to-quality phases.

Least absolute Shrinkage and Selection Operator (LASSO) is a regularization method originally proposed for variable selection and it is demonstrated to be the best subset selection method [15]. It introduces an additional term to the minimization problem, which is the L^1-norm of the regression coefficients vector multiplied by a weight parameter between zero and one, which tends to produce sparse models, which verifies its use as a variable selection tool.

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \qquad (4)$$

The tuning parameter λ controls the strength of the penalty. A value of zero is equivalent to a standard linear regression, and as it increases, regression model coefficients are shrunk toward zero. To find the optimal model, regression models for various values of λ are evaluated and the best model is chosen by a Cross Validation as having the smallest Mean Squared Error.

3 Proposed method

3.1 Data description

As explained above, a huge amount of FDC data is collected from the semiconductor manufacturing process. FDC data are usually stored in a three-way matrix X(I×J×K), where I is the number of monitored wafers stored in the FDC database, J is the number of process variables, and K is the number of observations of each variable for each wafer. At first, X is unfolded into a two-dimensional matrix with I rows and P= J×K columns before applying the regression model.

Quality parameters Y(I×L) are obtained by periodically testing a sample of products with measurement equipments after the completion of critical stages of manufacturing for monitoring the production. They contain various items such layer thickness mean or uniformity, etch rate... However, measurement steps are performed on randomly selected lots, on at least one wafer within each sampled lot. Thus, most of the data items are missing in the quality database. A drift happening between the scheduled measurements cannot be detected, and the quality of other processed wafers is unknown and need to be estimated for maintaining high yield of production.

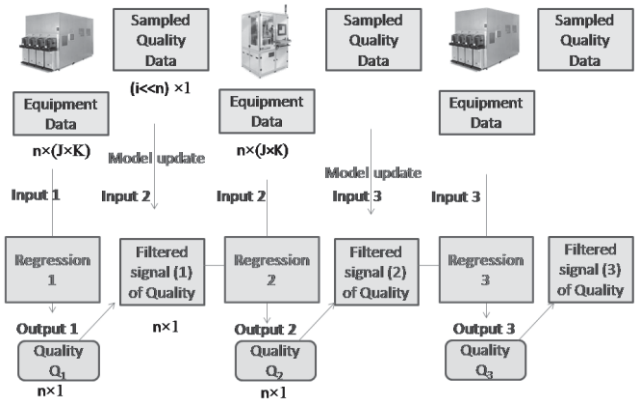
The purpose of our study is to predict the quality parameters corresponding to FDC data of many types of equipments. As unnecessary inputs can affect the prediction results, selection of critical parameters is necessary to improve model performance. The restriction to considerably less but the most pertinent FDC parameters improves substantially the performance. This can be achieved by using a LASSO regression model.

3.2 Method description

At every production stage, a LASSO-based model is used to estimate the missing quality data for each no-measured wafer. In this way, the quality parameters data are completed.

The equipment FDC data and the estimated quality already obtained from the previous equipments can be considered as inputs in modelling. Meanwhile, one quality parameter is used as output. The measured sampled quality parameters are used to evaluate the model quality.

Fig. 2 : Method Description



A Low Pass Filter is used to remove noise from the modelling signals, and then the filtered signal at a particular manufacturing step is used instead of the original signal as additional input to the prediction model which will be constructed at the next stage. An overview of the applied methodology is shown in (2).

The Mean Squared Error (MSE) is adopted here as the evolution criterion to evaluate the prediction accuracy and it is described by the following expression:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (5)$$

Where n is the number of sampled measured products, y_i and $\hat{y}_i$ are the measured and the estimated values, respectively. Closer is the MSE value to zero, better the prediction accuracy is. But as the length of the measured sample is small, a standard procedure for estimating the performance of the model is the n-Cross-Validation. The original sample is randomly partitioned into n equally sized subsamples and each subset is used once as a hold-out set for testing the model, and the remaining n-1 samples are used for training. The total accuracy is calculated as the average of the accuracies on all the hold-out subsets.

As already said, the quality measurements are sparsely sampled, our objective is to estimate the product's quality where it is processed in a particular equipment with respect to FDC data and the measured and estimated quality corresponding to the previous equipments. Whenever a quality measurement is available for one of the previous equipment, the model is updated and evaluated for improving outcomes.

4 Application results

4.1 Results of the proposed methods

The developed method is applied on data provided by three equipments (A, B and C) in the semiconductor manufacturing process. A LASSO model is applied for each equipment where the FDC parameters are considered as input and a quality parameter is considered as output of the model. Regression parameters optimization is performed and evaluated via Cross Validation, using the Mean Squared Error. The figure (3) displays the relationship between the tuning parameter λ and the Cross Validated Mean Squared Error of the LASSO model for the equipment A. The dots show the MSE of the corresponding model. The vertical line segments stretching out from each dot are error bars for each estimate. The line on the right is drawn at the minimum CV error, the other is drawn at the maximum value of λ within 1 SE of the minimum. Vertical bars depict 1 standard error.

For model construction, 1500 wafers are used to construct the model, while the remaining 500 wafers are used for testing. Figure 4 shows the estimated model signal and the filtered signal for the training and testing set. Figure 5 represent the difference between the measured and the predicted quality data for equipment A. The filtering of the model signal allows obtaining the estimated quality data as shown in figure 6. We can notice that the error after filtering data is slightly larger than the one before filtering, and therefore, we do not lose a lot of information by using the Low pass filter.

Fig. 3. Cross-Validated MSE for different values of Lambda of LASSO fit for equipment A.

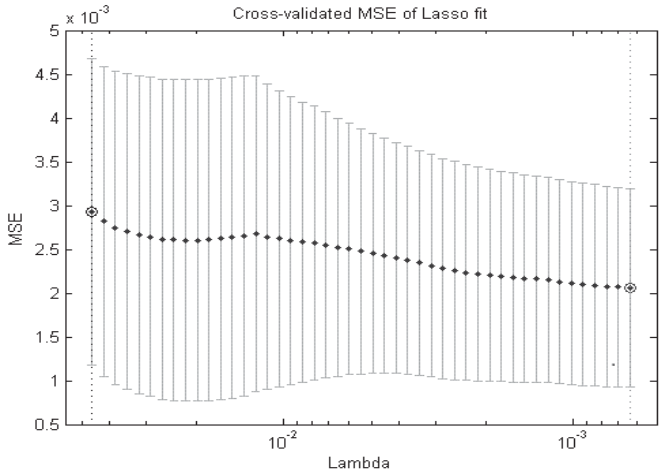
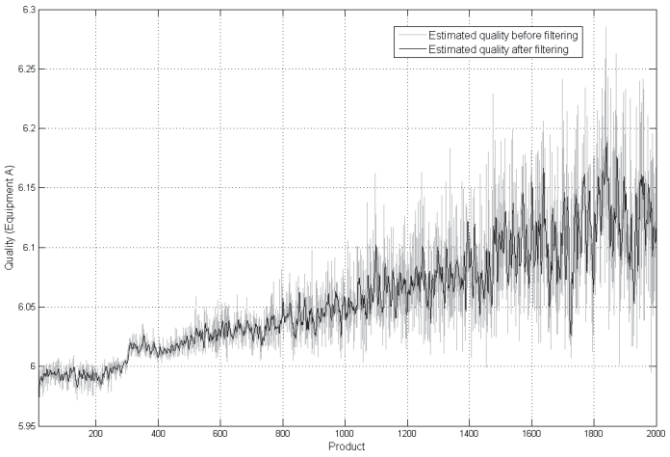


Fig. 4. Estimated signals for quality data obtained by LASSO for equipment A before and after filtering. The first 1500 wafers represent the training set, while the remaining of the signal represents the predicted quality data for the testing set.



The obtained filtered signal shown in figure 4 has been used as input with the FDC data of the equipment B to construct the signal shown in figure 7. This figure shows also the filtered signal. The figure 8 shows the sampled and the predicted quality measurements.

This procedure has been repeated for the equipment C where the results shown in figures 9 and 10 are obtained. And thus, the LASSO model can perform a variable selection and a quality prediction model for each equipment with respect to equipments data and the output of the previous equipments.

Fig. 5. Measured and estimated quality data by LASSO for equipment A before filtering.

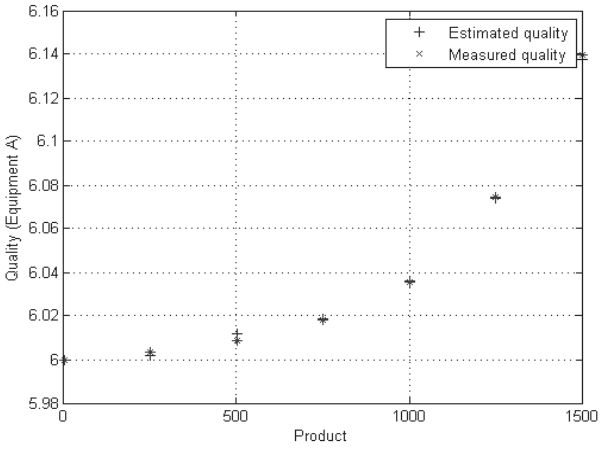


Fig. 6. Measured and predicted quality data after filtering by LASSO for equipment A.

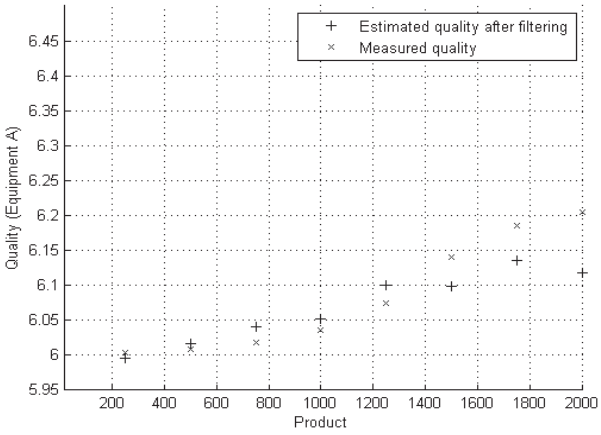


Fig. 7. Estimated signals for quality data obtained by LASSO for equipment B before and after filtering.

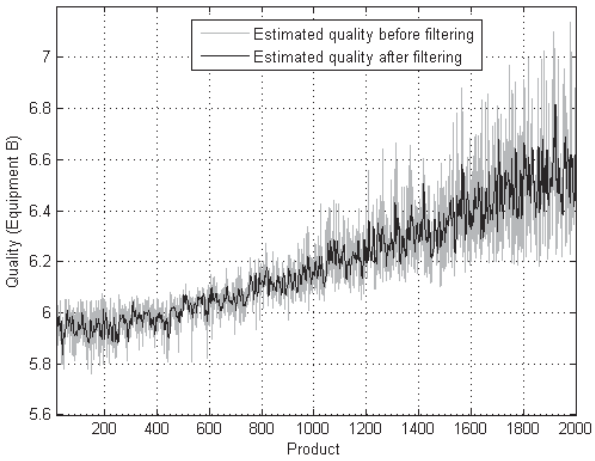


Fig. 8. Measured and predicted quality data by LASSO for equipment B.

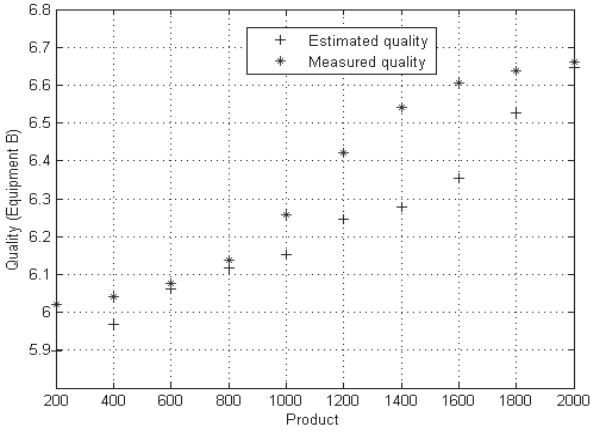


Fig. 9. Estimated signals for quality data obtained by LASSO for equipment C before and after filtering.

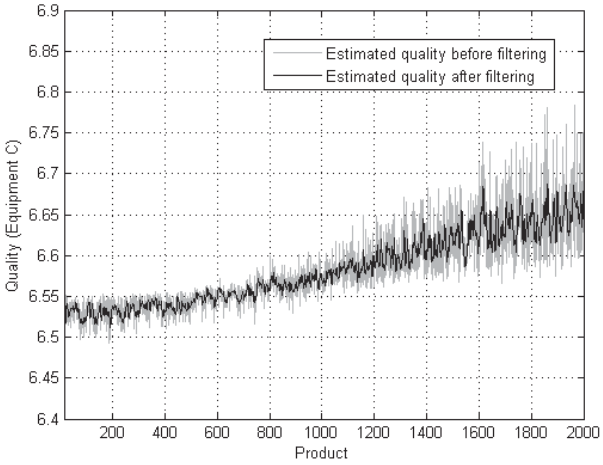
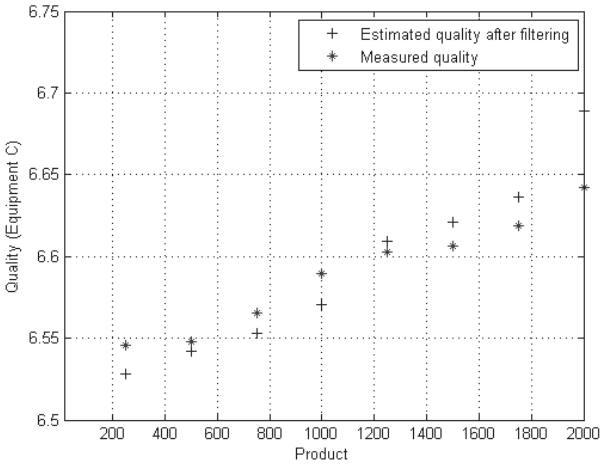


Fig. 10. Measured and predicted quality data by LASSO for equipment C.



The model was evaluated for 5 cases according to the availability of the quality data. The first case consists of estimating the product's quality for the equipment C if the first 1500 wafers are used for training and the remaining wafers are used for validating the model. A quality measure is available at the 1600th wafer for example for the equipment C, the model is updated by adding this measure in the training data. This procedure is repeated at each point where one quality measure is available. The evolution of the MSE is represented in (11) for the training data and in (12) for the testing data according to the available quality data. For the 1750th product, two patterns were added in the training data in estimating the product's quality for equipment C that are two measured products for the equipments A and C. By taking into account these measurements in addition to the new available FDC data, the quality prediction shows an improvement of 99% for the testing data as shown in the table 3 and the figure 12. The table 1 and 2 show the MSE of the training and test data, and the figures 11 and 12 show the evolution of the MSE.

Fig. 11. MSE evolution for the training data with the LASSO model

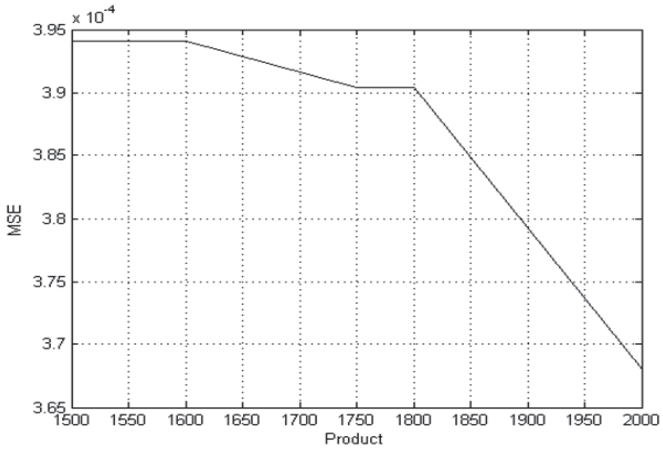


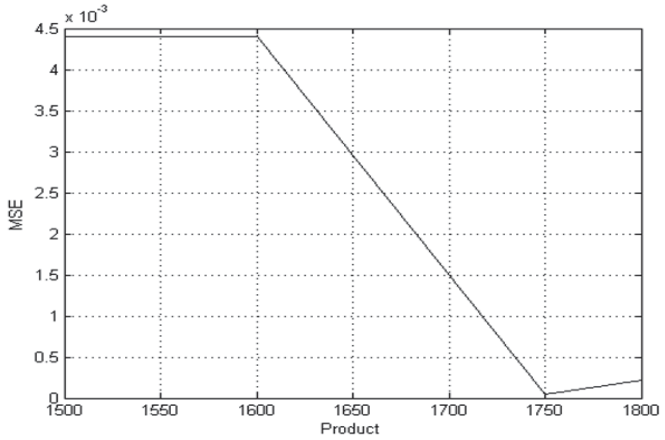
TABLE 1
MEAN SQUARED ERROR OF THE TRAINING DATA OF THE LASSO METHOD

Product	Training data
1600	3.9406*10-4
1750	3.9037*10-4
2000	3.6804*10-4

TABLE 2
MEAN SQUARED ERROR OF THE TEST DATA OF LASSO METHOD

Product	Test data
1600	0.0044
1750	5.1116*10-5
1800	2.1797*10-4

Fig. 12. MSE evolution for the validation data with the LASSO model.



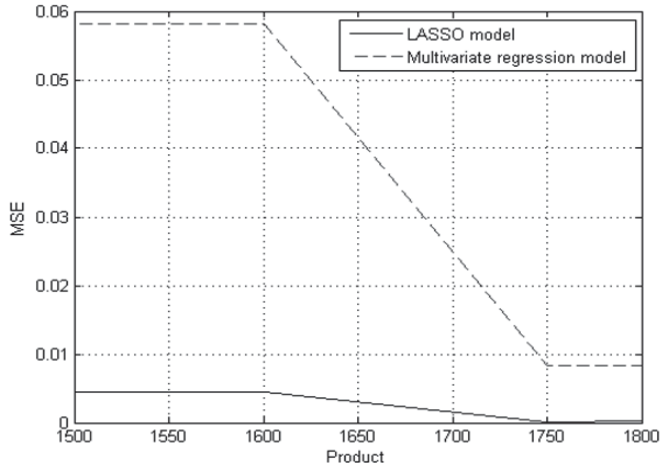
4.2 Comparison with multivariate linear regression

The developed LASSO method is compared with a multivariate linear regression where we obtain the results shown in (13) and Table 3. The LASSO model presents an improvement of more than 90% compared to the multivariate regression model, which implies the great importance of variable selection in case of multivariate and correlated data.

TABLE 3
COMPARISON BETWEEN THE MSE OF THE LASSO MODEL AND THE MULTIVARIATE REGRESSION MODEL FOR THE TESTING DATA

Product	LASSO	Multivariate Regression model
1600	0.0044	0.0582
1750	5.1116*10-5	0.0083
1800	2.1797*10-4	0.0083

Fig. 13. Comparison between the MSE of the testing data of the LASSO and the multivariate regression model



5 Conclusions

In this paper, we proposed a LASSO-based method for predicting the product's quality in a manufacturing process composed of many equipments, and an application of semiconductor manufacturing process is given. This method works as an interpolating method that is updated every time a new quality measurement is available. Results of our work have validated the effectiveness of using the LASSO regression method and show an improvement of more than 80% compared to a multivariate regression model. This method provides benefits for our application, but in the other hand, it has limitations that need to address in future. Firstly, as the prediction model is constructed by a statistical regression model, it lacks a physical significance, so the relationships between process parameters and quality data need to be evaluated by process engineer. Furthermore, the lack of measured quality data may cause an over-fitted model.

6 References

[1] D. Lieber, M. Stolpe, B. Konrad, J. Deuse, and K. Morik, "Quality Prediction in Interlinked Manufacturing Processes based on Supervised & Unsupervised Machine Learning," Forty Sixth CIRP Conference on Manufacturing Systems, 2013.

[2] H. Bae, S. Kim, K.-B. Woo G. S. May, and D.-K. Lee. "Fault Detection, Diagnosis, and Optimization of Wafer Manufacturing Processes utilizing Knowledge Creation"; International Journal of Control, Automation, and Systems, vol. 4, no.3, pp.372-381, June 2006.

[3] B. Kim, D. W. Kim, and G.T. Park. "Prediction of Plasma Etching Using a Polynomial Neural Network"; IEEE Transactions on Plasma Science, vol. 31,no.6,December 2003.

[4] M. H. Hung, T.-H. Lin, F.-T. Cheng, and R.-C. Lin. "A Novel Virtual Metrology Schemes for Predicting CVD Thickness in SemiConductor Manufacturing"; IEEE/ASME Transactions on mechatronics, vol. 12, no.3, June 2007.

[5] P.-H. Chou, M.-J. Wu, and K.-K. Chen. "Integrating support vector machine and genetic algorithm to implement dynamic wafer quality prediction system"; Expert Systems with Applications 37, pp. 4413-4424,2010.

[6] L. Yang, and J. Lee. "Bayesian Belief Network-based approach for diagnosis and prognostics of semiconductor manufacturing systems"; Robotics and Computer-Integrated Manufacturing 28, pp. 66-74, 2012.

[7] H. Purwins, A. Nagi, B. Barak U. Höckele, A. Kyek, B. Lenz, G. Pfeifer, and K. Weinzierl. "Regression Methods for Prediction of PECVD Silicon Nitride Layer Thickness"; 7th Annual IEEE Conference on Automation Science and Engineering (CASE), 2012.

[8] M.-H. Hung, T.-H. Lin, F.-T. Cheng, and R.-C. Lin. "A Novel Virtual Metrology Scheme for Predicting CVD Thickness in Semiconductor Manufacturing"; IEEE/AMSE Transactions on Mechatronics, vol. 12, no. 3, June 2007.

[9] Y.-C. Su, M.-H. Hung, F.-T. Cheng, and Y.-T. Chen, "A Processing Quality Prognostics Scheme for Plasma Sputtering in TFT-LCD Manufacturing," IEEE Transactions on semiconductor manufacturing, vol. 19, no. 2, May 2006.

[10] S. M. Weiss, A. Dhurandhar, R. J. Baseman, B. F. White, R. Logan, J. K. Winslow, and D. Poindexter. "Improving Quality Control by Early Prediction of Manufacturing Outcomes"; Journal of Intelligent Manufacturing, 1-13.

[11] T.-B.-L Nguyen, M. Djeziri, B. Ananou, M. Ouladsine, and Jacques Pinaton. "The International Federation of Automatic Control"; Cape Town, South Africa, August 24-29, 2014.

[12] F. Arif, N. Suryana, and B. Hussin. "A Data Mining Approach for Developping Quality Prediction Model in Multi-Stage Manufacturing"; International Journal of Computer Applications (0975-8887), vol. 69, no. 22, May 2013.

[13] M. M. Movahedi, M.B. Moghadam, H. Saiedi, and S. Eftekhari H. "A Solution for Statistical Control of Correlated Processes"; Middle-East Journal of Scientific Research 8, (6), pp. 1034-1045, 2011.

[14] Z.-Yan , C.-C. Chiu, W. Dong, and Y. Yao. "A LASSO-based batch process modeling and end-quality prediction"; The International Federation of Automatic Control, Cape Town, South Africa, August 24-29, 2014.

[15] R. Tibshirani. "Regression shrinkage and selection via the Lasso"; Journal of the Royal Statistical Society, Series B (Methodological), 58, pp. 267-288.

eMaintenance Platform for Performing Data Fusion Mutation on Machine Tools

V. Simón[1], D. Galar[1] and D. Baglee[2]
[1]Division of Operation and Maintenance Engineering, Luleå University of Technology, Luleå, Sweden
[2] The Institute for Automotive & Manufacturing Advanced Practice, University of Sunderland, Sunderland , United Kingdom

Abstract - *Condition monitoring (CM) plays a relevant role in production systems, for example, with machine tools. To obtain an accurate result when analyzing the condition of a machine tool and its components, it is necessary to integrate data from different sources. The types of data include: internal data from the Computer Numerical Control (CNC), external sensors, and value-added information coming from the study of the system's behavior.*

Data from disparate sources can be integrated using several pre- and post-processing methods that provide partial or total results in different formats. The use of an eMaintenance platform seems a reasonable and easy solution when faced with a challenge of such dimensions. This paper proposes an architecture able to cope with the challenge.

Keywords: eMaintenance platform, Maintenance 4.0, data taxonomy, SOA, Web Services

1 Introduction

The application of preventative maintenance techniques is an appropriate strategy to reduce the impact of malfunctions or machine breakdowns on the productivity, cost and quality of production systems. More specifically, the deployment of intelligent predictive tools and technologies can help detect potential failures and provide a solution.

Condition Based Maintenance (CBM) activities can be based on data obtained from sensors on a machine. The subsequent analysis of these data helps to measure and understand the machine's performance. This approach facilitates the computation of indicators at the local level to monitor the machine's local health; in addition, the information can be sent to an eMaintenance platform for more detailed analysis.

It should be noted that there is no single definition of eMaintenance. The term eMaintenance emerged in the early 2000s and is now a common term in maintenance related literature. eMaintenance is sometimes [1] considered a maintenance strategy, a maintenance plan, or a maintenance support: "e-Maintenance is a multidisciplinary domain based on maintenance and information and communication technologies (ICT) ensuring that the e-Maintenance services are aligned with the needs and business objectives of both customers and suppliers during the whole product lifecycle" [2].

eMaintenance can also be considered a philosophy supporting the move from ''fail and fix'' maintenance practices, to ''predict and prevent'' strategies (proactive approach), maintenance as a process (holistic approach), and an integrated concept to optimize performance [3]. Some well-known eMaintenance platforms are ICAS-AME [4], PROTEUS [5], TELMA [6], CASIP and its up- graded version KASEM [7] or DYNAMITE [8]; a more thorough classification appears in [9].

The current analysis assesses the nowcasting of a machine in a preliminary attempt to achieve proactive condition monitoring using non-intrusive monitoring techniques, affordable in terms of cost and effectiveness. In the proposed approach, the health index of the machine can be computed from the results of the signature analysis and associated with the degradation modes of the various components (e.g. gears, missing teeth, etc.).

An original feature of the proposed approach is the use of a remote level, whereby data from several machines are sent to an eMaintenance platform able to store data from different machines. This enables fleet-level performance management and monitoring, across the fleet and over time.

The paper breaks down the proposed process by explaining the data transformation from the initial data collection and warehousing, to the final data management procedures.

2 Data collection

Condition monitoring methods such as vibration or acoustic monitoring usually require expensive sensors. Electrical Signature Analysis primary application includes the diagnostics of electrical machines. Several authors have applied this technique to detect induction motor failures [10]. Others [11] have detected other failures using the induction motor current signature analysis. The controlled values, for example, of a gearbox failure, can be compared in the stator current spectrum, because diverse picks are related to shaft and gear speed. Characteristic gearbox frequencies can be detected in the stator current spectrum. Current-based diagnosis of mechanical faults such as unbalance and misalignment can be performed in the same way.

2.1 From testing to data collection

Good maintenance policies lead to less energy consumption by assets, as stated by [12]. However, the relationship between an electric signal and wear for any complex electro-mechanical system, for instance, a machine-tool spindle, is less evident. The potential correlation has to be learned based on experimental research. The use of test benches allows us to identify a machine´s operating condition, to analyze and describe its various failure modes, to pinpoint the most significant signal to be used in tests for failure, and to design and execute a test plan for fault detection and prognosis.

Laboratory research gives us the ability to run components to failure, working in a controlled failure environment. This helps us relate current and power signal analysis to the selection of features for failure diagnosis.

To achieve statistical consistency, during the first phase of testing, failure diagnosis, various faults should be tested along with the nominal one [13].

A local CBM module may consist of two main components based on Condition Monitoring (CM) techniques: first, the fingerprint to be used for the health assessment of the critical elements of the machine and second, operational data to infer the use of the machine.

A health monitoring system helps avoid component defects; consequently, it can prevent poor performance or even breakdowns. As an example of component defects, spindle defects include bearing damage, defects in rotary transmission, clamping malfunction, imbalance, stator error or alignment error. Operational data (i.e. feed, speed) can be used for energy and reliability management. An example is the different usage ratios of the machine: loads, speeds, etc. Note that the collection of operational data (real- and non-real-time) and fingerprint collection do not need to be performed simultaneously.

A fingerprint executed on a periodic basis (weekly, monthly etc...) generates raw data. These data are integrated with available inputs from the operational data to give information on the usage of the machine. These mixed data are pre-processed to obtain a set of relevant features that will be further analyzed for the nowcasting process. In parallel, data obtained from the machine are pre-processed to register the usage of the machine. The three main components of this process are operational data, fingerprints, and health assessments (the latter belongs to data management).

2.1.1 Monitoring working conditions (operational data)

Determining the usage of the machine by the end user yields a more holistic understanding of the real status of a machine's critical components. The historical use of the machine is found in the operational data. The main reason to collect operational data is to determine the operating environment of the machine with the purpose of finding possible reasons for malfunction or failure and optimizing reliability through the proper selection of component or machining parameters. Depending on the already installed or optional sensors, the solution may vary, but in any case, the required data rate should not be high (tens of Hertz). In modern Computer Numerical Control (CNC) systems, several configurations are available: sensors can be connected to the CNC or digital drive system or to a specialized hardware (for accelerometers or main power monitoring). In any case, there are two options to obtain operational data from the machine. The first is dialoguing with the CNC using specific hardware; this facilitates higher acquisition speed and detailed data, enabling some pre-processing. The second procedure implies the use of CNC internal data accessible through different links, like OPC servers, libraries, etc.; this limits the information available on how the machine is being used to showing only its acquisition rate.

Some processing is done to extract all the information from the data, using it to build a historical register of the use of the machine and obtain the data required for further service implementation.

Co-relating operational data and machine condition data using the correct algorithms can guide component maintenance, help to change working conditions to extend component life or even to select a different component, more appropriate for the real machine use.

2.1.2 Machine fingerprint

The term fingerprint has been coined to denote the electrical signature of a machine in a specific time domain.

To obtain the main fingerprint features, machines are run in a pre-defined test cycle in no-load condition to achieve better failure detection and to remove any noise that could affect the normal machine process load. Condition monitoring data are based on the fingerprints obtained from the machine. In the first stage, data analyzed during the experimentation phase may help in the selection of the type of sensors, acquisition rates and tests to be performed on the machine in the production plant. The idea behind the fingerprint is that any load and speed variation within an electro-mechanical system produces correlated variations in current and voltage. The resulting time and frequency signatures reflect loads, stresses, and wear throughout the system, but identifying them requires a mapping process or pattern recognition. Comparing the electric signature of equipment in good condition and equipment under monitoring supports fault identification. Note that Signature Analysis is only applicable to cases where the principal cause-effect is verified and modeled.

3 Data taxonomy

3.1 Asset technical information

Asset data should be collected in an organized and structured way. The two major data categories for equipment are: classification data, including system, location, plant and industry; and equipment attributes as technical features or design characteristics. These data categories are common to all equipment classes, although some specific data for a specific equipment class (e.g. number of stages for a compressor) could be needed.

Finally, the classification of equipment into technical, operational, safety related and environmental parameters is the basis for the collection of asset data, given the different nature of different devices. This information is also necessary to determine if the data are suitable or valid for various applications. Some data are common to all equipment classes, and some data are specific to a particular equipment class.

3.2 Events

Fingerprint trajectory tracking provides a solid study of the evolution of machine components [14]. This evolution, along with context mapping and past situation-based feedback, causes various triggers that define different events. The appropriate maintenance policy is selected based on the appearance of certain events.

3.3 Maintenance policies

Maintenance information includes unit life plans, job cataloguing, etc. for each unit in two different categories: preventive and corrective maintenance. These data are characterized by their identification (record numbers), by the parameters characterizing them (category, activities involved, impact, date), by the resources that imply their deployment (man/hours, equipment), and by outputs in terms of active maintenance time and downtime.

Recording maintenance actions is crucial for successful knowledge extraction at some later date. Preventive maintenance (PM) records are mainly useful for the maintenance engineer to estimate equipment availability; lifetime analysis is not only based on failures, but also on maintenance actions intended to restore the failed item to "as-good-as-new" condition. During the execution of preventive actions, impending failures may be discovered and corrected as part of the preventive activities.

A final option is to record the planned PM program as well. In this case, it is possible to record the differences between the planned and the actual performed preventive maintenance (i.e., the backlog). An increasing backlog will be an indication that the control of the conditions of the plant is jeopardized, possibly leading to equipment damage, pollution or personnel injury.

For corrective maintenance, failure records are especially relevant to knowledge extraction; therefore, failure data should be recorded in such a way as to allow further computation. A uniform definition of failure and a method of classifying failures are both essential when data from different sources (plants and operators) need to be fused in a common maintenance database.

Finally the combination of plant inventory and maintenance based information produces a maintenance schedule which is a mixture of available techniques to meet constraints and achieve company goals. This mixture is usually composed of scheduled maintenance and CM to perform CBM.

The maintenance schedule includes preventive maintenance jobs (over a year and longer) listed against each of the units in the life plans. The CM schedule is a schedule of the condition monitoring tasks listed against each of the units in the life plans.

The system must plan and schedule preventive jobs (arising from the maintenance schedule), corrective jobs (of all priorities) and, where necessary, modification jobs. Information coming back from the work orders (and other documents) is used to update the planning system; this provides information useful for maintenance control [15].

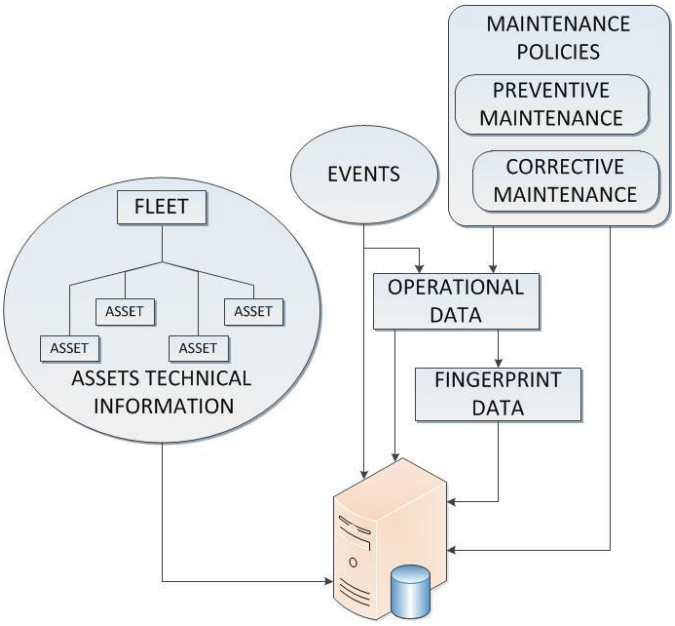


Figure 1 Inputs for the database

4 Data warehousing

4.1 Local module

As mentioned, data must be stored, because continuous connectivity is not assured. It is crucial to find a structured way to save the data; having the data in an organized state facilitates better treatment and sharing. An important issue is the limited storing capacities of some devices, as this could potentially lead to a conflict.

Various general-purpose database management systems (DBMSs) allow the definition, creation, querying, updating and administration of databases. Databases can be divided into two groups: embedded or client/server databases. Those belonging to the first group have an easy installation procedure. Few resources are needed, normally access for a single user, and they are tightly integrated with the application software requiring access to stored data; the database engine runs in the same process as the application. In the second group, the client/server database architecture is oriented to run in a server, running the database in a differentiated process from the software application and supporting multiuser access. It needs administrative privileges for installation.

Two mechanisms are used in the local level data model to support interoperability between local and remote servers: the

Machinery Information Management Open Systems Alliance (MIMOSA) and Open System Architecture for Condition-Based Maintenance (OSA-CBM).
MIMOSA focuses on the standardization of data models in the maintenance and condition monitoring domains. The Common Conceptual Object Model (CCOM) builds a foundation for all MIMOSA standards, while the Common Relational Information Schema (CRIS) provides a means to store enterprise Operation and Maintenance (O&M) information. The standard defines the various types of information that should be gathered to share information among different processes, systems and people.
OSA-CBM is an industry led team for standardizing the interoperability of systems participating in machine health assessment; as such, it is an implementation of the ISO-13374 functional specification. OSA-CBM adds data structures and defines interface methods for the functionality blocks defined by the ISO standard. Whereas the ISO-13374 standard provides guidelines on how to develop a condition monitoring system, OSA-CBM provides additional tools for implementation. Its goal is to promote the adoption of the CBM paradigm.

4.2 eMaintenance Cloud (eMC)

Scattered and fragmented databases cannot be replicated; this encourages the use of cloud computing in asset management. In such cases, a remote server enables us to receive data coming from different machines located in several places (the fleet), aggregate these data and make them semantically comparable, while considering their different contexts: i.e., technical differences (the machines may not be exactly the same), operational conditions, historical failures, etc. The goal of fleet management is to balance acquisition, recapitalization, reset, sustainment, and divestiture decisions across systems' life cycles in order to meet equipping and operating requirements, achieve optimized budgets, and communicate critical knowledge to stakeholders.
eMaintenance solutions offer a mechanism that supports organizations in their transfer of data to handle risk-based decisions through system overview. Decisions should be based on the understanding of data patterns and relationships. Materialized as a set of inter-operable, independent and loosely coupled information services, a framework with its own inherent infrastructure (i.e. eMaintenance Cloud, eMC) can provide fleet-wide, continuous, coordinated service support and service delivery functions for operation and maintenance.
In order to baseline the fleet and assess technical feasibility, fleet managers must have visibility into global equipment inventory and readiness status. This includes having knowledge of current configurations, systems, and block upgrade information, along with access to real-time asset information by system, component, and other customer distribution requirements. It also requires the ability to cross-check the accuracy of the data retrieved from data sources or other data management systems accessible to fleet managers. Other data needed to baseline fleets and determine technical feasibility include planned acquisition fielding, past fielding, system losses, system asset position, new or replacement systems, joint service requirements, divestiture requirements, data interchange requirements, system modifications, and funding requirements. By establishing a baseline, fleet managers will gain an accurate, common operational picture of the fleet, define areas of risk, and develop appropriate risk mitigation by recommending courses of action while achieving an optimized budget. Each is an integral and significant outcome of the fleet management process [16].
Beyond implementing a data warehouse and an application server, the eMaintenance platform should also support infrastructure grounded on Service Oriented Architecture (SOA) and Enterprise Service Bus (ESB) architecture based on Web Services, to bring together a set of company applications in an XML-based engine. The flexible infrastructure makes it possible for different parties to develop different modules in different development environments.

4.3 Module interoperability

As stated above, local and remote levels should exchange information for users to be adequately served. The remote service and the eMaintenance platform need data from different machines to provide added value services. Therefore, service must be defined in the context of end-users' needs; following this, the data that machines should exchange with the platform must be defined.
In terms of connectivity between local machines and remote servers, the availability of full permanent remote connection cannot be assumed.

4.3.1 Local-Remote module connectivity

Communication between machine tools and a remote data warehouse server can be handled in several ways:

- Machine directly connected to the Internet: VPN connection; direct transmission via Internet (HTTPS, FTPS); mail server; GSM card connection…
- Machine not directly connected: the user must periodically retrieve the data exports, and transmit them by creating an email attachment or using FTP. Another option is to upload the files to the server manually.

Various standards of reference guide the information exchange between different systems. Standardization is crucial, as it facilitates data interchange between different applications. A common language is a key component of standardization, as it facilitates the collaboration of various agents and the integration of information systems.
O&M activities are standardized by different organizations on different levels of abstraction. Although there is no single standard, the activity models, information exchange patterns and data models can be standardized. Formal data descriptions, such as XML schemas, can be extracted from the data models and used in content based information system integration, while activity models help in the analysis of business processes.

Operators, maintenance personnel, original equipment manufacturers, part suppliers and engineers have always wanted to have information about the condition of equipment assets at their fingertips when they need it. However, this information is not shared because it is split on different information systems which are not interconnected; these systems include manufacturer's data, operational data, condition monitoring data, diagnostic, reliability data, etc.
Interconnectivity of the islands of engineering, i.e., maintenance, operations, and reliability information, is embodied in MIMOSA's Open Systems Architecture for Enterprise Application Integration (OSA-EAI) specifications. Adopting these specifications offers advantages to maintenance and reliability users; it facilitates the integration of the asset management information and saves money by reducing integration and software maintenance costs. It is equally advantageous for technology developers and suppliers, because it stimulates and broadens the market, allows concentration of resources on core high-value activities rather than low-value platform and custom interface requirements, and provides an overall reduction in development costs.

5 Data management

The evolution of the data throughout the eMaintenance process is governed by the platform. Here, two crucial aspects must be contemplated: the representation of the data in a formalized and standardized way that allows sharing the data easily, and the sharing process itself.

5.1 Data mutation

As explained previously, for a local CBM module, one of the most important issues is defining the component fingerprint and, thus, the component health.
To obtain the component fingerprint, data must pass through several stages to find the most important features of the collected signal, i.e. Knowledge Discovery in Databases (also known as Data Mining). The first step requires processing the data to obtain a prepared and reduced dataset using various techniques: time synchronous averaging or windowing, for instance.

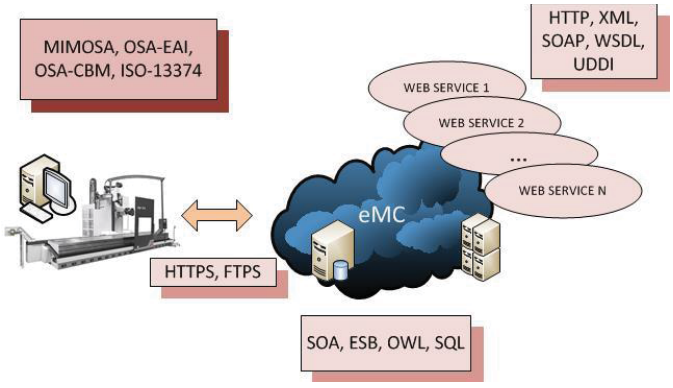


Figure 2 Overall Architecture

At a second stage, features are selected and extracted from the prepared data. This selection it is not yet a minimum feature set, however. Feature extraction is a common term used in pattern recognition and image processing. To classify a fingerprint, some characteristics are extracted for future identification and comparison. Features extracted are used to characterize properties of a component's condition. Implementation techniques are commonly used for vibration analysis [17] and motor current signal analysis [18]. Most methodologies are applied to the signal in the time and frequency domains.
In the time domain, a number of statistical parameters are used: root-mean-squared (RMS), peak value, crest factor, kurtosis, skewness, clearance, impulse and shape factor, average, median, minimum, maximum, variance and deviation. These parameters attempt to capture unusual behavior and/or impacts associated with early degradation stages and faults [14].
Frequency domain analysis refers to the mathematical functions or signals related to frequency, rather than time. Fast Fourier Transform (FFT) or wavelet transform (see for example [19]) are techniques to consider. In the frequency domain, frequency bands may differ based on the design of the machine.
At this stage, condition assessment and forecasting are done with the selected features, using such techniques as support vector machine, self-organizing map, artificial neural network, or regression methods.
A reduced set of relevant features is derived and compared to the fingerprints' time values. This provides the health assessment of the machine.

5.2 Data depiction

Ontology involves formal specification of knowledge in a domain by defining the terms (vocabulary) and relations among them [20]. Ontologies are composed of classes, descriptive concepts, class properties, and classes' relationships and instances.
Ontologies represent a suitable modeling method to provide common semantics and to query heterogeneous databases. Reference [21] states an ontology process enables: sharing common understandings of the structure of information among people or software agents, making domain assumptions explicit, defining concepts and knowledge and making domain inferences to obtain non-explicit knowledge.
Web Ontology Language (OWL) facilitates the definition of generic conceptualizations that can be used in multiple domains; it enables the creation of Web-based applications, such as a module of an eMaintenance platform [22].
OWL provides inference capabilities with plugged reasoners which perform consistency checking. Hence, there is the need to ensure that ontology is built correctly, in the sense that no syntactic error or inconsistency remains in it. In addition, explicit and manually constructed classes that belong to taxonomy constitute an asserted hierarchy, but thanks to OWL reasoners, an inferred hierarchy is automatically computed, allowing us to infer new knowledge.

5.3 Data sharing

A remote platform aims to provide openness and connectivity of the components installed in each machine and to guarantee the added value remote services. Modern day information technologies have resulted in a set of principles for designing and developing software as interoperable services, also known as SOA. SOA consists of the implementation of a platform containing several services. These services are business process oriented resources representing the possibility of performing tasks that ensure coherent functionality from both the provider and the consumer points of view. Within the operation and maintenance context, eMaintenance solutions provide a mechanism that supports organizations in transferring data to enable decision-making from a system perspective, and facilitates their understanding of data relationships and patterns.

SOA software architecture aims to implement an information system comprising independent but interconnected services. In that sense, the objective of SOA is to decompose functionalities in a set of services and describe their interactions.

From a technical point of view, SOA defines software in terms of discrete services, implemented using components that can be called upon to perform a specified operation for a specific business task. The SOA concept changes the existing software concept of a function—a specific piece of code that performs one particular task—to include the notion of a contract, a technology-neutral but business-specific representation of the function [17].

In that sense, SOA considers different elements such as the service concept and the service provider, consumer and broker. These elements interact to perform business tasks. The roles of SOA can be described as follows:

- Service: self-contained business function that accepts requests and returns responses through a well-defined standard interface.
- Service Provider: the function which performs a service, i.e., the owner of the web services. It registers the services with a Service Broker (registry), and it publishes information about the service to the service broker in standard format.
- Service Consumer: the function which uses the result of the services supplied by a provider, i.e., the user of the web services. It searches the registry provided by Service Broker and gets the information about the service. It builds the request message and sends to the Service Provider and gets the response back.
- -Service Broker: provides a registry of available services, i.e., the registry of the web services. The requester builds the request message, sends it to the service and gets a response.

Benefits of using this type of architecture are: reusability promotion, modular programming, better flexibility and easy to maintain services. SOA promotes the goal of separating users (consumers) from the service implementations. Services can, therefore, be run on various distributed platforms and accessed across networks.

SOA is implemented using new technologies and is principally based on XML and Web Services. Web services consist of exposing one or more applications (i.e. services) to an Internet network. These services can propose simple functions or a set of tools to compose a complete application. The following open standards are regularly used:

- SOAP (Simple Object Access Protocol): an XML based protocol specifying envelope information, contents and processing information for a message.
- WSDL (Web Services Description Language): an XML-based language used to describe the attributes, interfaces and other properties of a Web service. A WSDL document can be read by a potential client to learn about the service.
- UDDI (Universal Description Discovery and Integration): a specification for creating an XML-based registry that lists information about businesses and the Web services they offer. Though implementations vary, UDDI often describes services using WSDL and communicates via SOAP messaging.

As explained previously, SOA defines a method to design applicative interactions between different distributed components. This method is based on the services which are executed by a supplier component for a consumer. One of the properties of such a method is that it allows a component to be on different systems and distributed over various networks. SOA is based on services invoked through interfaces and vocabulary common to all agents (supplier and consumer). The more advanced these elements are in terms of modeling, the more advanced the services are in terms of different treatments and larger evolutions. SOA allows the architecture to be flexible and adaptable to many situations.

As a result, the eMaintenance platform enables the integration of other applications by acting as a hub of technologies. As it includes a Service Oriented Architecture foundation and web-based technologies, the platform offers openness and integration, with (web) services sharing data and results with other applications to help users cope with various business organizations and models (e.g. fleet monitoring application with expertise center, multi-site applications with expertise center, multi-client and multi-site, etc.) within an integrated enterprise architecture. The openness and flexibility of the SOA platform offers many possibilities, supporting data acquisition, storage, and transportation and contributing to service implementation at the remote level.

Finally, the service-based principle of the SOA design offers the possibility of using a methodology based on service composition and service reuse.

6 Conclusions

As the paper shows, the suggested architecture deals satisfactorily with the integration of different data formats through a combination of local and remote modules, in spite of

their disparate nature and granularity. The paper also suggests the possibilities inherent in an architecture that allows the monitoring of machine tool performance to support proactive degradation detection through the analysis of the current signal, enhanced by the application of several data transformations extracting the required information.

The sharing capabilities of the proposed platform provide an excellent opportunity to improve work already done in the area by facilitating communication with the most up-to-date and powerful methods currently used in the maintenance arena.

In summary, eMaintenance, or most recently Maintenance 4.0, provides forecasting capabilities to determine machine health in order to optimize maintenance actions and maximize the productive capacity of assets, expanding their lifespan.

7 References

[1] B. Iung, E. Levrat, A.Crespo, and H. Erbe. "Conceptual Framework for eMaintenance: Illustration by e-Maintenance Technologies and Platforms", *Annual Reviews* in Control 33, no. 2 : 220–229, 2009.

[2] M. Kajko-Mattsson, R. Karim, and A. Mirijamdotter, "Essential components of emaintenance", *International Journal of Performability Engineering*, 7(6), 555-571, 2011.

[3] B. Iung, G. Morel, J.B. Léger, " Proactive maintenance strategy for harbour crane operation improvement", *In Erbe, H. (Ed.), Robotica. Special Issue on Cost Effective Automation*, 21(3), 313–324, 2003.

[4] M. DiUlio, B. Finley, C. Savage and K. Krooner, "Revolutionizing Maintenance Through Remote Monitoring via ICAS & Distance Support", *DOD Maintenance Symposium*, Oct 2002.

[5] T. Bangemann, X. Rebeuf, D. Reboul, A. Schulze, J. Szymanski, J. P. Thomesse, *et al,* "Proteus-Creating distributed maintenance systems through an integration platform", in *Computers in Industry*, Vol. 57(6), 2006.

[6] E. Levrat and B. Iung. "TELMA: A full e-maintenance platform" in Centre de Recherche en Automatique de Nancy Université, 2007.

[7] www.predict.fr.

[8] IST. Dynamic Decisions in Maintenance (DYNAMITE). Information Society Technology (IST) 2008, available athttp://www.ist-world.org/ProjectDetails.aspx?ProjectId=2ae617135ecd4cb7bcefcaf8f05577c5 accessed on: 08 April 2015.

[9] J. Campos, "Development in the application of ICT in condition monitoring and maintenance", *Computers in Industry*, 60(1), 1–20, 2009.

[10] M. Messaoudi, and L. Sbita, "Multiple Faults Diagnosis in Induction Motor Using the MCSA Method", *International Journal of Signal and Image Processing*, Vol.1, Iss.3 pp. 190-195, 2010.

[11] E.L. Bonaldi, L.E.L. Oliveira, J.G. Borges da Silva, G. Lambert-Torres, and L.E. Borges da Silva, "Detecting Load Failures using the Induction Motor as a Transducer", *10th International Conference on Control, Automation, Robotics and Vision*; Hanoi, Vietnam 2008 pp. 196-199, 2008.

[12] EN 15341

[13] I. Bravo-Imaz, A. Garcia-Arribas, S. Ferreiro, S. Fernandez, and A. Arnaiz, "Motor current signature analysis for gearbox health monitoring: Experiment, signal analysis and classification", *Second European Conference of the Prognostics and Health Management Society (PHM 2014)*; Nantes, France 2014

[14] D. Galar, U. Kumar, J. Lee, and W. Zhao, "Remaining useful life estimation using time trajectory tracking and support vector machines", *International Journal of C O M A D E M*, 15(3), 2-8, 2012.

[15] D. Galar, A. Gustafson, B. Tormos, B and L. Berges, "Maintenance Decision Making based on different types of Data Fusion", *Maintenance and Reliability*, 14(2), 135-144, 2012.

[16] D. Pack, "Enabling fleet management with CBM+", *Army Sustainment,* April, 2014.

[17] S. Ferreiro, A. Arnaiz, B. Sierra, and I .Irigoien, "Application of Bayesian networks in prognostics for a new Integrated Vehicle Health Management concept". *Expert Systems with Applications*, vol. 39(7), pp. 6402-6418, 2012.

[18] G. Medina-Oliva, A. Voisin, M. Monnin, M., F. Peysson, JB. Leger, "Prognostics Assessment Using Fleet-wide Ontology". *PHM Conference 2012*, Minneapolis, Minnesota, USA, 2012.

[19] N. Bieberstein, S. Bose, M. Fiammante, K. Jones, K., and R. Shah, "Service-Oriented Architecture Compass: Business Value, Planning, and Enterprise Roadmap", *Prentice Hall PTR*, Upper Saddle River, NJ, 2005.

[20] T. Gruber, "Ontology", *Encyclopedia of Databases Systems*, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2009.

[21] N. Noy and D.L. McGuinnes, "Ontology development 101: A guide to creating your first ontology", *Technical Report SMI-2001-0880*, Stanford Medical Informatics, 2001.

[22] A. García-Crespo, B. Ruiz-Mezcua, J.L. López-Cuadrado and I. González-Carrasco, "Semantic model for knowledge representation in e-business", Knowledge-Based Systems, Vol. 24, pp. 282-296, 2011.

Selecting a Classification Ensemble and Detecting Process Drift in an Evolving Data Stream

Alejandro Heredia-Langner[1], Luke R. Rodriguez[1], Andy Lin[1], Jennifer B. Webster[1]
[1]Pacific Northwest National Laboratory, 902 Battelle Boulevard, PO Box 999 Richland, Washington 99352

***Abstract*— We characterize the commercial behavior of a group of companies in a common line of business, or network, by applying small ensembles of classifiers to a stream of records containing commercial activity information. This approach is able to effectively find a subset of classifiers that can be used to predict company labels with reasonable accuracy. Performance of the ensemble, its error rate under stable conditions, can be characterized using an exponentially weighted moving average (EWMA) statistic. The behavior of the EWMA statistic can be used to monitor a record stream from the commercial network and determine when significant changes have occurred. Results indicate that larger classification ensembles may not necessarily be optimal, pointing to the need to search the combinatorial space spanned by the classifiers in a systematic way. Results also show that current and past performance of an ensemble can be used to detect when statistically significant changes in the activity of the commercial network have occurred. The dataset used in this work contains tens of thousands of high level commercial activity records with continuous and categorical variables and hundreds of labels, making classification challenging.**
***Key Words:* Ensemble classifiers, EWMA, optimization**.

1. Introduction

Approaches to mine and analyze streaming data that use a single classifier or a fixed ensemble assume that the classifiers at hand are, as a set, optimal for the problem under consideration. Under evolving conditions and at the rate at which data streams are generated in today's commercial, scientific, and security environments, it is unlikely that a single classifier, or even a fixed ensemble, can reliably provide an acceptable level of performance for a prolonged period of time. In addition to making reliable predictions in high volume data streams, researchers and analysts may also be interested in using classifiers to detect when the behavior of the data has changed significantly`. Significant changes in the predictive performance of a classification system signal the need for an analyst to get involved, determine the cause, and decide if the classifiers need to be updated. This situation is difficult when multivariate and complex data streams are involved, such as those considered here, containing large scale business activity of companies in the private sector.

Working with classification ensembles can be challenging because of the size of the combinatorial space that needs to be explored when searching for an optimal set for the current operating conditions. Exhaustive search is only possible if the number of distinct classifiers available is relatively small, while larger spaces can only be partially explored. Researchers in [1] investigate the performance of 15 classifiers on a variety of datasets using several search methods and optimality criteria and find that, in general, the best results are produced when using a direct search approach for the selection of an optimal ensemble. They also find that using a criterion that correlates strongly with the overall classification error to determine performance produces better results than using other measures of classifier diversity.

Even when an optimal ensemble can be found, it may remain so only for a narrow period, since high volume data rates usually mean that only a relatively small window of records is available to characterize the data stream and train the classifiers. The selection of an adequate window of training data can in itself be a difficult problem. Results in [2] indicate that using a fixed number of records can be problematic, since a wide window may make classifiers insensitive to trends and a narrow one may result in classifiers that simply chase the noise in the data. For this reason, it is important to detect when significant changes in the underlying distribution of the data stream have occurred.

An additional difficulty arising when applying a classification ensemble to a data stream is determining how and whether the ensemble should be modified. In [3] examples are presented of a dynamic weighted majority ensemble method where individual classifiers (also called base learners or experts) are selected based on the performance of the ensemble. In that approach, new individual classifiers are added to the ensemble when a threshold of poor performance by the current ensemble has been crossed, while the influence of some base learners currently present may be down-weighted if their individual performance is poor. Results in [3] are encouraging, but the size of the ensembles considered can become large, unless this number is explicitly restricted.

To address the issue of the changing nature of streaming data, also known as concept drift, and the detection of a point where new records should be obtained for training and selecting a new ensemble, we propose the use of an exponentially weighted moving average statistic to detect significant concept drift and the use of a small population of classifiers to build a classification system. In our approach, different combinations of individual classifiers are used to find an ensemble that is optimal for the current conditions and that can help estimate the effect that each individual classifier has on the overall classification rate. The ensemble selected can then be

applied to a stream of new records for as long as a stable and acceptable level of performance is maintained. In this way, a new window for re-training and finding a potentially new classification ensemble can occur only when necessary.

We demonstrate this approach though an example using a set of commercial records from nearly 400 companies in the automotive field. Characterizing this set is challenging because it contains continuous and categorical features and because the records can be relatively vague and prone to contain erroneous entries caused by humans entering the data. This means that perfect classification may not be achievable.

2. Materials and Methods

The PIERS records (Port Import/Export Reporting Service, [4]) database contains international trade information from vessels arriving to or departing ports in the U.S. The database contains millions of records with information such as port of entry/departure, estimated value of the shipment, tonnage, and brief descriptions of the contents of the shipment, among many other fields. The information in the PIERS records can be used to research whether and what kinds of relationships exist between certain commercial entities, and, as shown in [5] and [6], is a useful source of data in business analytics.

The data available in the PIERS database can be challenging to analyze because the number of features, or fields, available for each shipment is large and the fields include numerical, categorical and text data. The information available in PIERS can also be fairly ambiguous, such as when a single tariff code is used to describe the contents of a shipment, and the code may cover a wide variety of items. In spite of this, PIERS records contain valuable information about the activity of commercial actors, and this information can be aggregated and analyzed in ways that are meaningful for describing the behavior of companies operating in a business group or network.

For the present work, we employed 52353 records for the year 2013 for 396 companies in the automotive industry. These records are of interest because they contain transactional information between commercial actors in what can be considered a fairly well defined line of business. The first objective of this work was to determine if the records can be used by machine learning algorithms to adequately classify the companies they belong to. The features selected as inputs for the classifiers are shown in Table 1.

Table 1. Name, description and type of the features used to build the classifiers

Feature Name	Description	Type
YRMTH	Combined year and month date of record	Continuous
CTRYCODE	Code for country of origin/destination	Categorical
FCODE	Foreign port code	Categorical
USCODE	US port code	Categorical
HSCODE	Harmonized System Code, for tariff purposes	Categorical
QTY	Quantity shipped, integer	Continuous
MTONS	Metric Tons shipped	Continuous
TEUS	Twenty-Foot Equivalent Units, integer	Continuous
VALUE	Estimated value, USD	Continuous
CONVOL	Container Volume, m^3	Continuous

This set of records is challenging for classifiers because it is highly unbalanced (some companies have thousands of records to train on while others have only a few), there are more than 80 different countries involved in the trades and hundreds of tariff codes used for the items shipped by the companies in this network.

The classifiers employed include a Naïve Bayes (NB) classifier, a k-nearest neighbor (k-NN) classifier and two classification trees. The classification trees employ different split criteria: Gini's diversity index (GDI) and maximum deviance reduction. All results in this document were obtained using MatLab [7]. Because only four individual classifiers are involved, it was possible to explore the performance of different ensembles using a factorial approach, where all possible combinations of the four classifiers are applied to the same training/testing partitions of the data.

3. Results

There are 16 classifier combinations, including one where none of the four individual classifiers is used. The case with no classifiers represents a truly naïve predictor, used to establish a lower bound on performance. The truly naïve predictor produces labels for new records randomly, but in the same proportions as those found in the training set. For example, if 50% of the records in the training set belong to a single company, the truly naïve classifier will predict, with probability of 0.5, that any given record in the test set will belong to that particular company.

The dataset was divided repeatedly and independently into training and test sets using a fixed number of records for training and testing, and the names of the 396 companies as labels. This resulted in a training/testing partition of roughly 74%/26%. Predicted company labels for the test records were obtained directly when a single classifier was used, or by averaging scores when more than one classifier was involved, breaking ties randomly. The process of training and testing was carried out repeatedly and independently to assess the performance of the classifiers. Boxplots of the fraction of mislabeled test records in 10 independent trials are shown in Figure 1. The labels on the x-axis of Figure 1 indicate which classifier combination was used on the test sets, and the individual classifiers are identified in the plot.

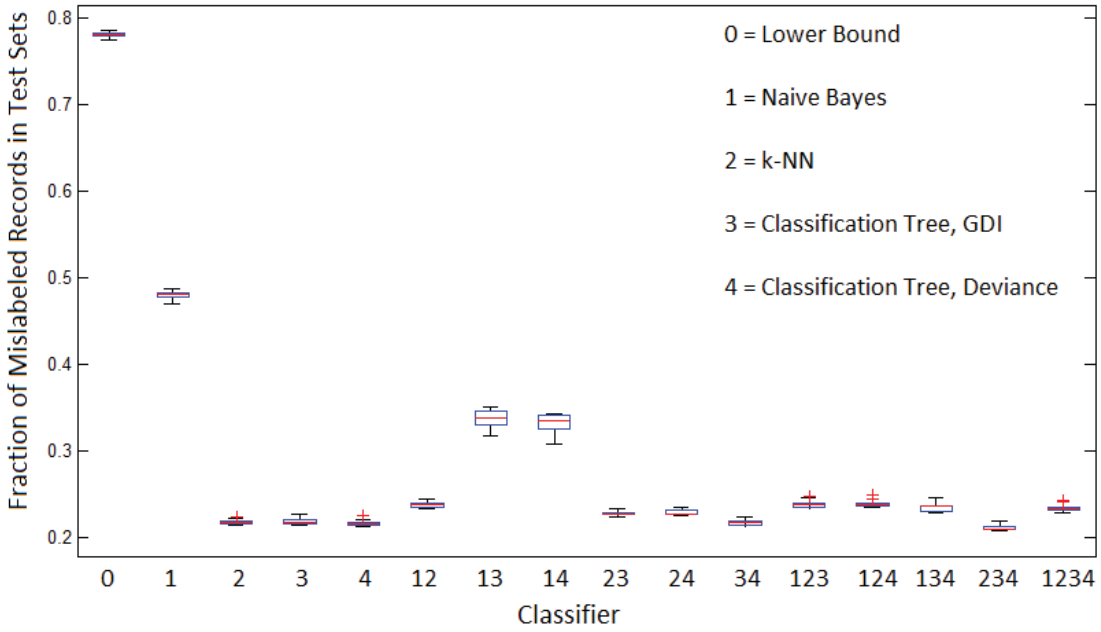


Figure 1. Boxplots of the fraction of mislabeled records in the test sets over ten independent train/test (76%/ 24%) partitions of 52353 records. The fraction of mislabeled records in 10 test sets is shown in the y-axis. The x-axis shows the coded values of the individual classifier or combination used. The boxes in the plot encompass the first, second and third quartiles, with whiskers denoting the most extreme points not considered outliers, and any outliers are marked with crosses.

Figure 1 shows that all the classifiers, alone or in combination, perform better than the truly naïve classifier used to establish a lower bound. It is also interesting to notice that the performance of the Naïve Bayes classifier can, in this case, be dramatically improved by combining its predictions with those from any other classifier available.

Figure 1 shows that there are several good options to choose from when it comes to selecting a classifier to predict the labels in this business network. The boxplots in Figure 1 can be used to choose the ensemble that minimizes the fraction of mislabeled records in a majority of test sets. However, the information gathered can also be used to estimate the impact that each classifier has on the observed error rate. This type of modeling may provide insights into how classifier diversity affects the results. The error rates obtained in the ten train/test trials that produced the results shown in Figure 1 were used as input for a generalized linear model, employing the percent error as the response. The model for the mean predicted percent error rate is:

$$\widehat{Error\ Rate} = \beta_0 + \beta_1 NB + \beta_2 kNN + \beta_3 GDI + \beta_4 DEV + \beta_{12} NB \cdot kNN + \beta_{13} NB \cdot GDI + \beta_{14} NB \cdot DEV + \beta_{23} kNN \cdot GDI + \beta_{24} kNN \cdot DEV + \beta_{34} GDI \cdot DEV + \beta_{123} NB \cdot kNN \cdot GDI + \beta_{124} NB \cdot kNN \cdot DEV + \beta_{134} NB \cdot GDI \cdot DEV + \beta_{234} kNN \cdot GDI \cdot DEV + \beta_{1234} NB \cdot kNN \cdot GDI \cdot DEV$$

where NB represents use of the Naïve Bayes classifier, kNN represents use of the k-NN classifier, GDI and DEV represent use of the respective classification tree, and the β_i are model parameters, which are estimated using iteratively reweighted least squares [8]. Model parameter estimates and related statistics are shown in Table 2.

Table 2. Percent error rate model parameter estimates and their corresponding standard errors, t- and p-values

Model Parameter	Estimate	Std. Error	t-value	p-value
β_0	78.10	0.1887	413.78	<0.0001
β_1	-30.20	0.2669	-113.14	<0.0001
β_2	-56.30	0.2669	-210.92	<0.0001
β_3	-56.10	0.2669	-210.17	<0.0001
β_4	-56.10	0.2669	-210.17	<0.0001
β_{12}	32.20	0.3775	85.30	<0.0001
β_{13}	41.90	0.3775	111.00	<0.0001
β_{14}	41.30	0.3775	109.41	<0.0001
β_{23}	57.20	0.3775	151.53	<0.0001
β_{24}	57.50	0.3775	152.32	<0.0001
β_{34}	55.80	0.3775	147.82	<0.0001
β_{123}	-42.70	0.5339	-79.98	<0.0001
β_{124}	-42.50	0.5339	-79.61	<0.0001
β_{134}	-51.00	0.5339	-95.53	<0.0001
β_{234}	-58.90	0.5339	-110.33	<0.0001
β_{1234}	53.10	0.7550	70.33	<0.0001

The model with the parameter estimates in Table 2 has an R^2 of 0.9984, an adjusted R^2 of 0.9983 and all of the parameters have highly significant p-values. Analysis of other performance statistics did not reveal major anomalies with the model. The model parameters in Table 2 show how the presence or absence of each classifier affects the predicted percentage error rate, and how individual classifiers interact with each other. The model can be useful because it can be employed to determine what level of improvement can be expected when adding or removing a particular subset of classifiers to a stream of data that retains the characteristics of the training sets.

Analysis of the model developed indicates that an ensemble that includes the k-NN classifier and the two classification trees produces the best predicted mean percentage error rate. However, the analysis also indicates that using any one of those three classifiers alone would produce results that are nearly as good. The model and parameter estimates also provide information of how the diversity in this set of classifiers affects the accuracy of the ensemble. As shown in Figure 1, the ensemble that contains all four classifiers does not result in the best rate of correct predictions for this particular dataset.

After selecting an ensemble with good performance, a key question that remains when applying the ensemble to streaming data is how to detect concept drift. A classifying ensemble can be expected to maintain a stable level of performance only as long as the characteristics of the new records remain more or less the same as those in the training data. For this reason, it is important to know when the behavior of the data stream has changed significantly, so that an analyst or monitoring system can be alerted and a new set of classifiers trained under the new conditions.

Selecting and training a new classification ensemble involves not only additional time and effort, but it also means that, during this time, classification of currently available data has to be put on hold. If the streaming data has not changed in meaningful ways and an ensemble with good performance is available, stopping to acquire new training data and selecting a potentially new ensemble is wasteful and could result in a process that may simply chase the natural noise in the data, increasing the variability of the predictions.

Monitoring the performance of a classification ensemble involves assessing when, and whether, a significant change in the data stream has occurred. Because the performance of a classification ensemble can be measured by its error distribution [9], it is important to find a way to detect when significant changes in the misclassification rate have occurred.

For this work, the performance of an ensemble is measured using an Exponentially Weighted Moving Average (EWMA) applied to a measure of classification error (see [10] for an excellent introduction to the EWMA). An EWMA is a weighted average of current and past observations, and it has been used extensively to monitor performance in industrial and scientific settings [11], [12].

Performance of a classification ensemble applied to streaming data can be characterized and monitored by the number of misclassified observations in a fixed number of records. In this work, the number of misclassified observations in every ten records was used. If c_t is the number of misclassified observations at period t, that is, a period that involves ten consecutive records, then the EWMA statistic at period t is given by:

$$z_t = \lambda c_t + (1 - \lambda) z_{t-1}$$

where the value for z_0, needed for the first set of ten records (t=1), is computed as the average number of misclassified observations per ten records in the data used to train the ensemble. The EWMA statistic can be monitored using control limits given by:

$$UCL = \bar{c} + k\sqrt{\frac{\lambda\bar{c}}{2-\lambda}} \quad \text{and} \quad LCL = \bar{c} - k\sqrt{\frac{\lambda\bar{c}}{2-\lambda}}$$

where UCL is the Upper Control Limit, LCL is the Lower Control Limit, $\bar{c}$ is the average rate of misclassification in every ten records in the training data, and k and λ are constants chosen so that, if no concept drift is present, the EWMA statistic remains relatively stable and within the control limits. Values of $k = 3$ and $0.05 \leq \lambda \leq 0.25$ are common in practice, but other values can also be used [10]. The LCL and UCL need to be calculated using a training set that is representative and stable, that is, data where no concept drift has occurred.

To test the usefulness of the EWMA in detecting concept drift, the records used to generate the results shown in Figure 1 and Table 2 were used to compute LCL, UCL and EWMA values to monitor the error produced by the optimal classification ensemble. As stated previously, the number of misclassified observations in every ten records in the training data was used to compute the EWMA statistic and calculate the LCL and UCL values. Misclassification rates were obtained using predictions from an ensemble containing the k-NN classifier and the two classification trees. A plot of the EWMA statistic for 12,000 test set records, k =2.7 and λ = 0.08 is shown in Figure 2.

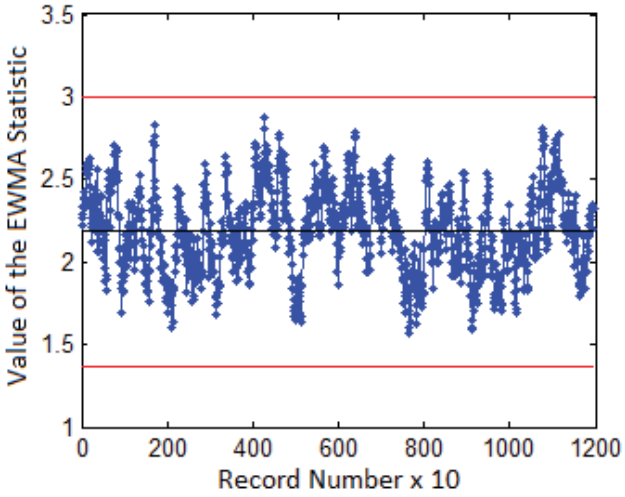


Figure 2. EWMA plot for 12000 test records. Training data for the number of misclassified observations in every ten records and a classification ensemble that includes the k-NN classifier and the classification trees with the GDI and deviance split criteria were used to compute the centerline, LCL (bottom horizontal line) and UCL (top horizontal line) shown in the plot.

Figure 2 shows the behavior of the EWMA for the optimal ensemble applied to test records, that is, records that were not used in training the classifiers. The statistic plotted in Figure 2 shows that the ensemble remains stable around the center line, with fluctuations showing the natural variability of the classification process. Figure 2 indicates that the error rate for the ensemble selected remains close to two observations mislabeled in every ten records, which is consistent with the behavior for this ensemble in Figure 1.

It is of interest to determine if the EWMA shown in Figure 2 can be used to detect changes in the behavior of companies in this particular business network. These changes may come about if, for example, one or more companies in the set start producing shipment records that are more commonly associated with other companies in the network. This type of change in behavior could be the result of individual companies making incursions into new markets or entering new lines of business, information that would be of interest to business analysts.

To investigate if this type of change would result in a significantly different behavior of the EWMA statistic, test sets were produced where the labels for records from a pair of companies were exchanged. This swap impacts around 6% of the total number of records in the test set, leaving the majority of the records unchanged. Using the same EWMA parameter values shown in Figure 2 on the test set with changed records for two companies produces the results in Figure 3.

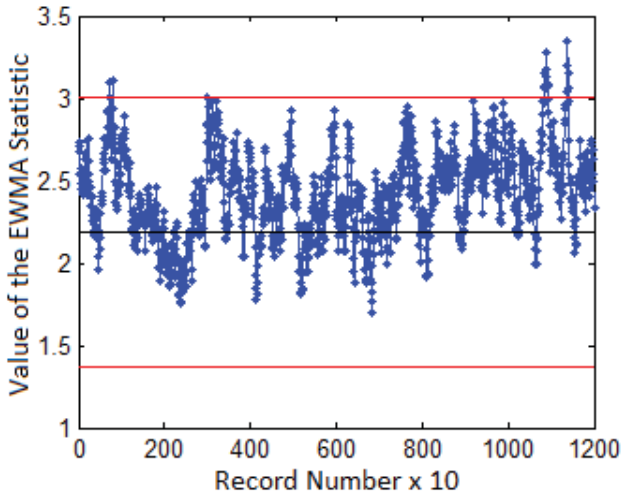


Figure 3. EWMA plot obtained using test data where the labels for the records of two out of the 396 companies (involving around 6% of all test records) were exchanged. The classifier ensemble, center line, LCL and UCL are the same as those in Figure 2.

Figure 3 shows that the EWMA statistic crosses the UCL early on, signaling that the process has drifted significantly. Figure 3 also shows a very clear upward shift in the level of the EWMA, with a large majority of the points in the plot falling above the centerline, providing more evidence that a significantly larger than expected number of misclassifications per set of ten records is occurring.

In practice, predictions by the optimal ensemble would be stopped immediately after the UCL has been crossed, since this is an indication that the process has drifted. At that point, an analyst would determine if a cause for the signal can be found (erroneous record keeping, for example), or if this behavior represents the new state of the commercial network. Only if the latter is true, a new ensemble of classifiers would need to be trained under the new conditions, from which new LCL and UCL values would be calculated.

4. Conclusions and Future Work

We have presented an approach for modeling the performance of a classification ensemble and used a measure of that performance to detect process drift. The example presented involves application of a classification ensemble to a set of commercial records involving a group of companies in a common line of business. The case considered is challenging because of the relatively large number of records involved, the variety and coarseness of the predictors and the relative lack of information available for some of the companies in the network.

Performance of four different classifiers, alone and in combination, was investigated and it was found that, in this case, the most complex classification ensemble is not optimal. Several choices of classifiers, including use of some single classifiers, produce optimal or nearly optimal predictions. This is an indication that the combinatorial space available when multiple classifiers are used should be explored in a systematic way, and that practical considerations, such as the time needed to train and evaluate different ensemble combinations, should be considered as part of the overall ensemble design strategy.

The approach presented to evaluate classification performance involves monitoring a meaningful measure of the misclassification rate, in this case errors in every ten new consecutive records in the data stream. We have shown that the use of an exponentially weighted moving average statistic measuring this proportion of misclassifications is an effective and relatively simple way to detect when significant changes in the behavior of the data stream have occurred. In the example presented, the EWMA is able to detect when a change impacting fewer than 10% of the records in the test set has occurred, suggesting this as a promising tool for detecting concept drift, minimizing the number of interruptions and effort involved in re-training a new classification ensemble.

In the near future, we plan to apply this methodology to data streams from other technical and business areas with the goal of developing a general approach for detecting concept drift in the context of small classification ensembles.

Acknowledgment

The research described in this paper is part of the Analysis In Motion Initiative and the Signature Discovery Initiative at Pacific Northwest National Laboratory. It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy.

5. References

[1] Ruta, D. and Gabrys, B. (2005). Classifier selection for majority voting. Information fusion 6(1), pp. 63-81.

[2] Wang, H., Fan, W., Yu, P.S., Han, J. (2003). Mining concept-drifting data streams using ensemble classifiers. Proceedings of the ninth ACM SIGKDD internationalconference on Knowledge discovery and data mining. pp. 226-235.

[3] Kolter, J. Zico, and Marcus A. Maloof (2007). Dynamic weighted majority: An ensemble method for drifting concepts. The Journal of Machine Learning Research 8, pp. 2755-2790.

[4] PIERS Trade Database (2014). https://www.piers.com/ JOC Group Inc. Newark, NJ.

[5] Pagell, R.A., and Halperin, M (1998). International business information: how to find it, how to use it. Greenwood Publishing Group.

[6] Kennedy, A. (1994). International Business Information Sources and Their Utilization in Export/Import Research. Journal of Teaching in International Business, Vol. 6(2), 83-101.

[7] MatLab.R2014a (2014). The Mathworks Inc., Natick, MA.

[8] Myers, R.H., Montgomery, D.C., Vining, G.G. (2002). Generalized Linear Models. John Wiley & Sons, Inc. New York, NY.

[9] Tulyakov, S., Jaeger, S., Govindaraju, V. and Doermann. (2008). Review of Classifier Combination Methods in Machine Learning in Document Analysis and Recognition. Springer Berlin Heidelberg. pp. 361-386.

[10] Montgomery, D.C. (1991). Introduction to Statistical Quality Control, 2nd Ed. John Wiley & Sons, Inc. New York, NY.

[11] Testik, M. and Borror, C. (2004). Design strategies for the multivariate exponentially weighted moving average control chart. Quality and Reliability Engineering International 20(6), pp. 571-577.

[12] Qin, Qin et al. (2014). Application of EWMA and CUSUM Models to School Absenteeism Surveillance for Detecting Infectious Disease Outbreaks in Rural China. Online Journal of Public Health Informatics 6(1):e14.

Reliability Evaluation of Underground Power Cables with Probabilistic Models

Hassan M. Nemati, Anita Sant'Anna, Sławomir Nowaczyk

Abstract—**Underground power cables are one of the fundamental elements in power grids, but also one of the more difficult ones to monitor. Those cables are heavily affected by ionization, as well as thermal and mechanical stresses. At the same time, both pinpointing and repairing faults is very costly and time consuming. This has caused many power distribution companies to search for ways of predicting cable failures based on available historical data.**

In this paper, we investigate five different models estimating the probability of failures for in-service underground cables. In particular, we focus on a methodology for evaluating how well different models fit the historical data. In many practical cases, the amount of data available is very limited, and it is difficult to know how much confidence should one have in the goodness-of-fit results.

We use two goodness-of-fit measures, a commonly used one based on mean square error and a new one based on calculating the probability of generating the data from a given model. The corresponding results for a real data set can then be interpreted by comparing against confidence intervals obtained from synthetic data generated according to different models.

Our results show that the goodness-of-fit of several commonly used failure rate models, such as linear, piecewise linear and exponential, are virtually identical. In addition, they do not explain the data as well as a new model we introduce: piecewise constant.

I. Introduction

Electric power transmission and distribution networks consist of different types of cables, some of which have been installed more than 50 years ago, and some are newly added to the network. The major problem with these power cables is the lack of efficient condition monitoring methods [14].

Power outages, i.e. the unavailability of electricity supply due to faults, have many undesirable effects and are a high cost to the society as a whole. Loss of production, cost of repair, and customers' dissatisfaction are some of the important factors to be considered when analyzing the impact of outages. For institutions like hospitals, airports, and train stations, power outages can be disastrous.

There are many different reasons for power outages. According to a study by the Edison Electric Institute [10], 70 percent of power outages in the USA are weather related phenomena such as lightning, rain, snow, ice, etc. Another 11 percent of outages are caused by animals, such as birds, coming into contact with power lines. To reduce the impact of such incidents, many power electric companies are moving towards underground transmission and distribution lines. However, underground cables may also cause outages, most commonly due to insulation degradation and ruptures in conductors.

One drawback of underground cables is that the procedure for finding the exact place of failure is harder, since no visual inspection can be performed. In addition, even when a fault is localized, the process of digging the ground to reach the cable, and also repairing the cable, is more difficult and requires more skill than for aerial cables.

Many governing bodies are continuously increasing requirements put on distribution companies concerning the acceptable number and duration of power outages. In addition, in many areas of life, society is more and more relying on electrical power. Consequently, there is a great need for better methods to determine the condition of the in-service underground cables and their remaining useful life. In particular, it is important that those methods are cost effective.

In this paper, we analyze five different models to estimate the relationship between the age and failure rate in underground high voltage cables. In addition to commonly used models (linear, piecewise linear, and exponential), we also consider constant and piecewise constant models. In particular, we focus on the methodology for evaluating how well different models fit the data. As is common in this domain, the amount of data we have available is very limited, and it is difficult to know how much confidence should one have in the goodness-of-fit results.

We calculate the empirical failure rates based on real data of over fifty years of historical faults from a small European city. The data comes from historical databases at Halmstad Energi och Miljö (HEM Nät), one of the Swedish electricity distribution companies.

The remaining of this paper is structured as follows. Background and related work is presented in section 2. In section 3 we explain the proposed model evaluation methodology, and we describe our experiments and results in section 4. We summarize our contribution and discuss future work in section 5.

II. Background and Related Works

A mathematical model that represents the current condition of a cable is known as the state of the cable [13]. The state represents the condition of the cable at a given point in time. Owing to the fact that the cables are laid under the ground, their current state is not directly observable. Depending on the amount of available information, one can estimate the state in different ways, using different models. Clearly, if the information about the cables increases, the

Authors are with Center for Applied Intelligent Systems Research, Halmstad University, Sweden. Email addresses: hassan.nemati@hh.se, anita.santanna@hh.se and slawomir.nowaczyk@hh.se

representing model becomes more precise. However, there is a tradeoff between the cost of collecting additional data and the benefits such data would provide.

There are mainly two methods for condition assessment of underground cables. The first is measuring the cables' condition by using different types of diagnostic and stress test analysis such as partial discharge (PD) and dielectric losses. The second is mining historical information such as age of the cables, and previous failures.

The condition of power cables can be measured in two ways: using on-site testing [6], [7], [9] or laboratory testing [16]. On-site testing is performed directly on the in-service cables. In the laboratory testing, first, a new cable undergoes accelerated aging processes to simulate the condition of aged cables, which are then analyzed. In both of these methods the amount of PD, oil analysis, and bulk properties of insulation, e.g. tang δ measurement, are used to determine the cables condition. The tang δ measurement is a diagnostic test conducted on cables' insulation to measure their deterioration. In fact, the tang δ measurement is used as the loss factor of the insulation material which will increase during the aging process. The assessment of the in-service cables should be performed every 3-5 years and the results classify the investigated cables into different categories based on which future maintenance can be performed. Both of these measurements are very costly and complex processes.

The historical data analysis is usually performed in one of the two ways. The first is based on Crow-AMSAA and reliability growth model [1], [2], [8], [14]. Based on the time duration between each recorded failure in the system, historical failures are modeled using a Weibull distribution. This Weibull model is then used to estimate the time to the next failure, usually in the whole system, i.e., for all underground cables, without any distinction between aged and new cables. In other words, all the cables are considered to be in the same condition, regardless of their age, type, and other factors.

In the second historical data analysis method, in addition to the previous failures, other information such as age, and insulation condition are used to model failure rate [11], [12], [18]. Bloom et al. [3], [4] used historical data for age and number of previous failures as "observable condition"; and experts' judgment for insulation degradation condition, environmental stressor, and effect of the previous failures as "unobservable conditions". By using the historical data and the experts' knowledge they modeled the changes in cables' condition probabilistically, i.e., given the current state of a cable, what is the probability of different cable states in the future. Of all the factors used in their work, only age and historical failure rate are extracted from actual data, and all the rest of the information is based on the experts' judgment.

The failure rate model is usually used for estimating the expected number of future failures. One important aspect is that future failures are influenced by the replacement strategy employed, which is one of the possible solutions for electric power companies to reduce the number of outages. Replacement actions, also known as rejuvenation, is the procedure of replacing the old and faulty parts with new cables. There has been some research analyzing how the replacement of old cables reduces the number of expected failures and improves reliability, for example [11] and [12], however, the majority of work in the field does not take rejuvenation into account.

In general, there are three types of underground cables widely used in distribution power grids [5]:

- Oil-Filled cable
- Paper Insulated Lead Cover cable (PILC)
- Cross-linked Polyethylene cable (XLPE)

Before development of XLPE cables in 1993, PILC cables were the most common installed underground power cables [15]. Their estimated expected lifetime is declared to be around 40 years [17], but they have been used for more than that in many transmission and distribution grids. In these grids, the problem of degradation of underground cables due to aging is becoming more and more severe.

The old Paper Insulated Lead Cover (PILC) cables, which are of main concern in this study, are heavily affected by a number of factors such as ionization, thermal breakdown as well as electrical and mechanical stresses [5]. Since the paper insulation is made of cellulose, the quality of the insulation degrades over time and causes more frequent breakdowns. One way to decrease the corrosion speed and cable fragility is to fill the paper insulation with oil.

There are several important factors that accelerate the aging process in PILC cables. The ones most commonly mentioned in the literature are cyclic overloading, thermal breakdown, PD, irregular load pattern, direct or indirect spiking, inadequate depth in the ground, and very low temperature.

Cable joints, which are part of the underground cables, can also cause outages in the network. The jointing is the act of reconstructing two cables to become one. It is used when a longer cable is needed or when a part of an old cable is replaced with a new cable. A joint is usually the weakest part of an underground cable and it is affected by three types of stressors: thermal, electrical, and mechanical stress. Mechanical stress and water ingress are the main causes of failures in cable joints [5]. The fault in the joints might affect the conductor, insulation, or sheath. The sheath of the joints get corroded due to overloading and the chemicals present in the soil over a period of time. This increases the chance of moisture seepage into the joint, which subsequently causes failure.

In this work, we only use available historical data to compute failure rate. This approach is not as accurate as performing direct measurements on individual cables, but is often preferred in practice since mining the available data to find a model is significantly cheaper than performing laboratory or field tests.

III. Methodology

It is well known that by analyzing historical information of cables inventory, it is possible to predict the future failures in

cables with some degree of accuracy. One common example is modeling the failure rate for a particular type of cables. We use historical data from a small European city to estimate the parameters of the model. This model can then be used to predict future faults for different cables.

In particular, in this paper we focus on the failure rate for PILC underground cables at a certain age. Note that there are several other factors affecting failure rate variation in cables, such as number of joints, history of previous failures, environmental stressors, usage patterns, manufacturer and cable type, etc. Here, however, we only consider the age and the number of historical faults to estimate failure rate.

To estimate failure rate we need to have access to historical databases containing information such as installation year, date of previous failures, and the age of the cable at the time of failure. Furthermore, to calculate the proportion of faulty cables over total cables, we need to know the total length of all the in-service cables during each year.

The process of calculating the failure rate, estimating model parameters, and finally, evaluation of the results is described below, as shown in Figure 1.

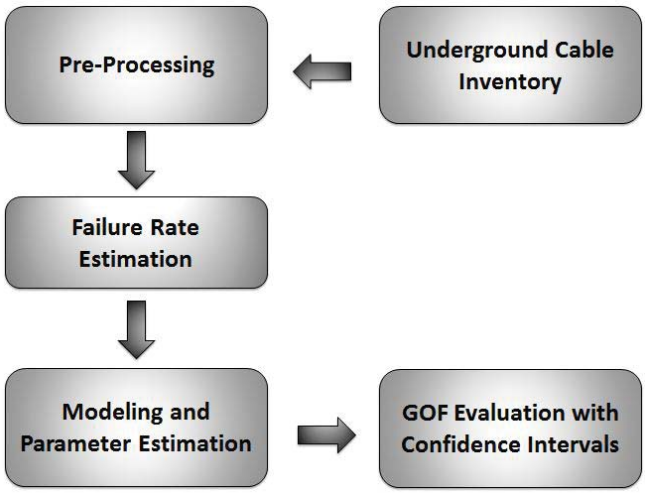


Fig. 1. Overview of the model creation and evaluation process.

A. Pre-processing

Due to the requirements explained above and the available databases, we have selected cable inventory data set. This data set contains historical information about both the in-service and destroyed cables that have been installed since 1908 in Halmstad power distribution grid. Each cable is described with a unique ID and the transmission line to which it belongs, as well as additional information such as insulation type, conductor size, installation year, length, etc. In this work, we only analyze in-service high-voltage PILC cables.

A transmission line between two cable boxes consists of a number of cables. According to the data set, the total number of high-voltage transmission lines containing PILC cable is about 500.

The cables in a line may have different installation years. We assume that the initial installation year of the line is the earliest installation year among all cables in the group.

In addition to length of in-service cables, we require information about past failures. In our case the historical failure database could not be directly linked to the cable information, since the two use different asset identifiers. Therefore, to identify past failures, we use the assumption that *short* cables in any given line are artifacts of previous repairs. Therefore, we consider each cable of length smaller than 20 meters to correspond to a failure in the line. The failure is assumed to have taken place in the year of the installation of the short cable, and to take place in the oldest cable within this line. Those assumptions are not fully accurate, but we have confirmed, through discussions with domain experts, that they are realistic.

B. Failure rate estimation

Failure rate is the frequency with which a system or component fails within a given unit of time. This definition can be naturally extended to a population of systems, for example a network of cables. In this work we consider the number of failures per year per kilometer. The general equation for the empirical failure rate is:

$$FR = \frac{N}{L},$$

where N is the number of failures in a year and L is the total length of in-service cables.

There are many factors that influence the failure rate, however, in this work we only focus on cable age (understood as the number of years between installation of the cable and the time of the failure). It is a well-known fact that the likelihood of failure changes with age. Therefore, we express the empirical failure rate for underground cables at age α as the total number of failure that happened to cables at age α, denoted $N(\alpha)$, divided by the total length of cables that were in-service at age α, denoted $L(\alpha)$:

$$FR(\alpha) = \frac{N(\alpha)}{L(\alpha)}.$$

Among several factors affecting failure rate, we only consider the factors that can be estimated from the historical databases we have access to: installation year of each cable (age), length, voltage class (high voltage or low voltage), and failure history: number of failures, and age at time of failure.

C. Modeling and parameter estimation

A failure function, $\lambda(\alpha)$, is a function that describes changes in failure rate depending on age. Figure 2 shows a commonly used model that represents the failure function known as the bathtub curve [11]. The model begins with a high failure rate (infant mortality), followed by fairly constant failure rate (useful life). Finally, the failure rate increases again as the component reaches the end of its life (wear-out).

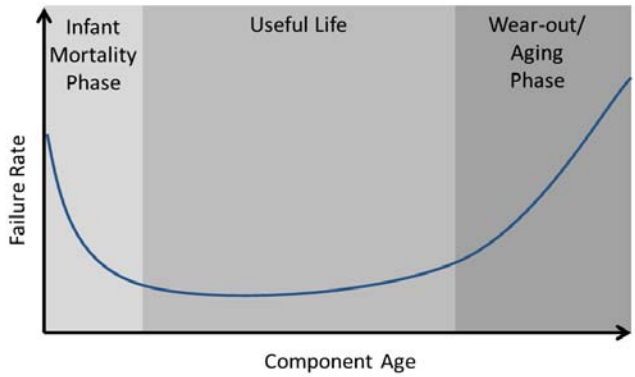


Fig. 2. Bathtub curve of typical failure rate for components.

When discussing power cables, we are particularly interested in modeling the "wear-out" time and the effects of aging process on failure rate. It is commonly believed that the failure rate increases as cables get older.

Our goal is to find an appropriate failure function $\lambda(\alpha)$. To this end, we investigate five different models and evaluate how well, the empirical failure rates fit each model. Observe that we do not specifically consider the "infant mortality" period in this analysis.

We have decided to perform experiments using five different failure functions. The three commonly used models in statistical analysis are linear, piecewise linear, and exponential. In addition to these, we have also investigated constant and piecewise constant models.

Constant: This model is described by a constant line with the failure rate equal to $\lambda(\alpha) = \mu$, where μ is the mean failure rate value of all the empirical data points $FR(\alpha)$.

Piecewise constant: This model is constructed by two constant lines at different values, μ_1 and μ_2, where μ_1 is the mean failure rate before T_{pwc} and μ_2 is the mean failure rate after T_{pwc}.

$$\lambda(\alpha) = \begin{cases} \mu_1 & \text{if } T_{pwc} \leq \alpha \\ \mu_2 & \text{if } T_{pwc} > \alpha \end{cases}$$

Linear: The linear model is specified by a linear function with two parameters: slope m_l and intercept b_l. In this model, the increment of failure rate between two consecutive time points is constant.

$$\lambda(\alpha) = m_l(\alpha) + b_l$$

Piecewise linear: This model represent the failure rate to be constant at the beginning up to age T_{pwl}, and then failure rate grows linearly with slop m_{pwl}. Therefore, the function is specified by three parameters, the constant failure rate b_{pwl}, the time which failure rate starts to increase linearly T_{pwl}, and the slop m_{pwl} of the line.

$$\lambda(\alpha) = \begin{cases} b_{pwl} & \text{if } T_{pwl} \leq \alpha \\ m_{pwl} \cdot (\alpha - T_{pwl}) + b_{pwl} & \text{if } T_{pwl} > \alpha \end{cases}$$

Exponential: this distribution is described by the function:

$$\lambda(\alpha) = \beta \cdot e^{\beta \cdot \alpha}$$

For each model, the corresponding parameters are calculated by Levenberg-Marquardt optimization algorithm implemented in Python `scipy` library, minimizing the mean square error.

After parameter estimation, we need to evaluate how well do the empirical data points fit each model. This can be done by using different goodness-of-fit measures.

D. GOF evaluation

In this study we employ two goodness-of-fit measures; the first is based on calculating the *probability of generating the data* from a given model (PGD); the second is based on *mean square error* between the data and the model (MSE).

In the PGD measure, for each age, the value of the failure function $\lambda(\alpha)$ at that age is considered to be the mean value of a normal distribution. The variance of this normal distribution is computed from the empirical data points. At each age, the cumulative probability function is used to calculate the probability that a given data point belongs to the normal distribution centered around the failure function. Finally, the calculated probabilities for each age are multiplied together to give the value of GOF for that model. The higher this probability is, the better the data points fit the model.

However, the resulting numbers are very small and difficult to analyze, and thus we use the negative logarithm (base 10) of those values to make them easier to interpret. Therefore, the lower the value of the GOF, the better the empirical failure rates fit the model under consideration.

$$GOF_{PGD} = \\ - log_{10} \prod_{\alpha} P(x \leq FR_\alpha | FR_\alpha \in X_i \sim (\mu = \lambda(\alpha), \sigma^2))$$

The MSE GOF measure is the sum of squared differences between each data point and the value of the failure function at corresponding age. Also in this case, the lower the GOF value the better the data points fit the model under consideration.

$$GOF_{MSE} = \frac{1}{n} \sum_{\alpha} (FR(\alpha) - \lambda(\alpha))^2$$

where *n* is the number of data points.

Finally, it is important to note that, while GOF results can be compared directly, it is often difficult to properly interpret the results, especially when the data is of limited quantity (and also quality) and it does not fit any of the models perfectly. Therefore, we propose a way to interpret the results by comparing the obtained GOF measures with expected GOF and confidence intervals, estimated using synthetic data.

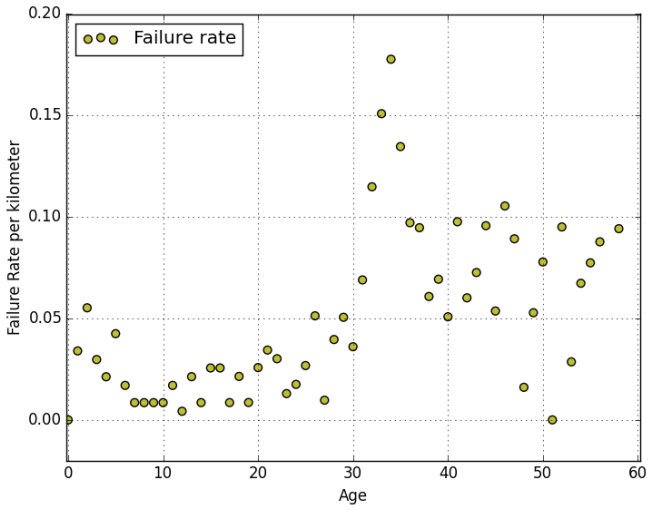


Fig. 3. Empirical failure rate per kilometer as function of age, for high voltage PILC cables.

For each model, a number of synthetic data sets are generated by drawing random points from a normal distributions with mean equal to the failure function at each age and variance computed from the empirical data points. The synthetic data sets should have the same number of points as the empirical data points. The PGD and MSE GOF are computed between each synthetic data set and the corresponding model, confidence intervals are derived based on the variance of the GOF values. The GOF of the synthetic data sets generated by one model are also compared to all other models in order to determine how well a data set generated from model A fits model B. These comparisons will help us draw conclusions about the how well the empirical data points fit each of the proposed models.

IV. Results and Discussion

The result of calculating empirical failure rates at each age, for high voltage PILC cables, is shown in Figure 3. The horizontal axis represents the cable age at time of failure and the vertical axis represents the failure rate λ (per kilometer).

By comparing this result with Figure 2 it is possible to extract three lifetime phases. The empirical data starts with higher failure rates at ages 1-6, the "infant mortality" period. It then continues with a period of low and fairly constant rates during ages 7-19, the "useful life". And finally, the higher failure rates start again from age 20, the "wear-out" phase. However, there are also some differences from the bathtub curve, the most clear ones being the peak at ages around 30 years, and the shape of the wear-out phase.

The parameters for constant, piecewise constant, linear, piecewise linear, and exponential models were estimated from the empirical data. Each resulting model is shown in Figure 4. The resulting parameter for the constant model is $\mu = 0.052$, and for the piecewise constant model are $\mu 1 = 0.023$, $\mu 2 = 0.082$, and $T_{pwc} = 30$. The parameters for the linear model are $m_l = 0.0013$, and $b_l = 0.0128$. For the piecewise linear $b_{pwl} = 0.0231$, $m_{pwl} = 0.00147$, and $T_{pwl} = 0.00695$. For the exponential model, the parameter β is equal to 0.0254.

To compare the results of GOF between different models, first we generated 100 synthetic data sets based on each model, and then measured the PGD and MSE between each randomly generated data set and all the models. In Figure 5, one randomly generated data set is shown for each model. Then, for each group of 100 generated data sets, we found the mean value of all calculated GOF to all models and the corresponding 95 percent confidence interval.

We performed the PGD and MSE tests for all combination of synthetic data sets and models. In this case, data sets A, B, C, D, and E are the 100 randomly generated data sets from constant, piecewise constant, linear, piecewise linear, and exponential models respectively. The results of GOF tests based on PGD and MSE are presented in Table I and Table II. For example, the result of PGD GOF test of the data generated from constant model (A) with respect to the linear model (C) is 43.8960 ± 0.6276.

From the GOF results presented in Table I and Table II, several observations can be made, as follows.

As expected, the best GOF results are obtained when the data set is compared to the model which generated it. For example, Data A fits model A better than any other model. These correspond to the diagonal entries in Table I and Table II.

The results of GOF measurements from fitting each generated synthetic data with the same model (diagonal of the tables) does not show any statistically significant differences. This verifies that performing this type of comparison between synthetic data and models is systematically correct, i.e., the result of comparing model A with synthetic data A is as good as comparing model B with synthetic data B.

Except the constant model (model A) which is statistically very different from the rest of the models, the result of pairwise comparison between a GOF test in synthetic data generated by a model but fitting with another model, and a result of GOF test in the other combination of this two models, is not significantly different. For example, GOF between data B and model D, is not significantly different than the GOF between data D and model B.

There is no statistically significant difference between GOF of the data points, neither the empirical nor synthetic, between linear, piecewise linear, and exponential models. That is, the GOF results are within the respective confidence interval obtained from the synthetic data. This indicates that those three models are virtually identical.

Nonetheless, the real data seems to fit the piecewise constant model better than the other models. This suggests that the failure rate could be modeled by two constant lines; low failure rate up to age 30 and higher failure rate after that. This does not confirm the assumption that the failure rate increases monotonically as a function of age. This observation is quite surprising, and we believe it deserves

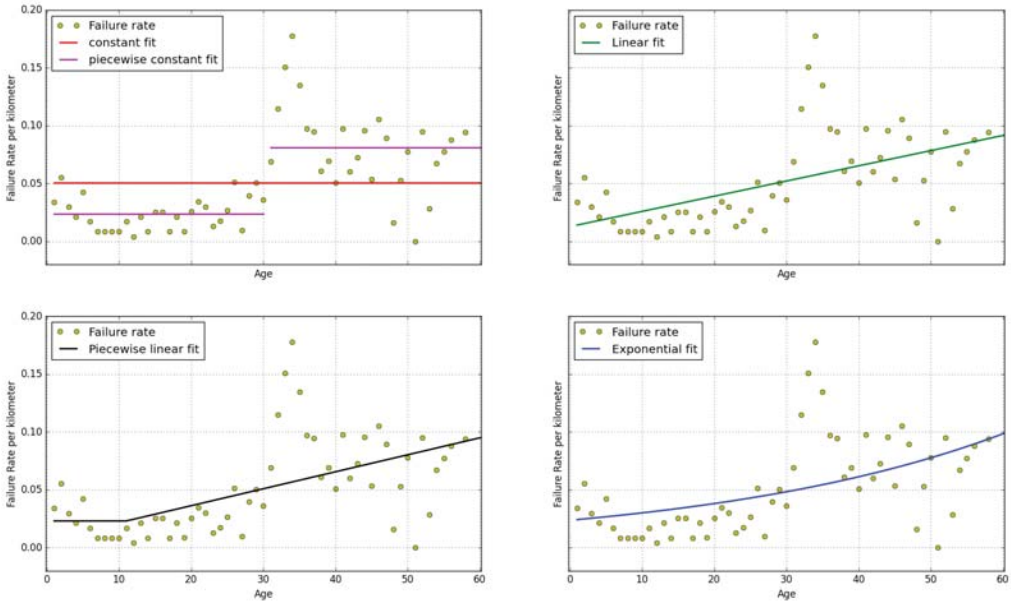


Fig. 4. Five different failure rate models, fitted to the empirical data.

further analysis in the future.

This result might be caused by several factors. First, the very high values of failure rate at ages between 32 and 35 years affect other models more than the piecewise constant model. Second, the input data set and the in-use information is not enough to uniquely and with confidence identify the best model. Therefore, other information should also be taken into the account. Third, we did not considered the affect of repair and replacement of cables on failure rate estimation. In fact, the process of rejuvenation of the underground cables prevents the failure rates from becoming too high, especially after experiencing number of failures (in our case after age of 40 years or so).

V. Conclusion and Future Work

In this paper we have presented some of the characteristics of power grid cables, especially PILC underground cables, which are used in many power transmission and distribution networks. We have also discussed the main challenges regarding fault prediction for these cables.

We have introduced five different probabilistic models for predicting failure rate depending on cable age, and evaluated how well does each of these models fit the real-world, historical fault data. We have employed two different goodness-of-fit measurements, one based on mean square error and one based on probability of generating the data.

In order to compare the GOF measures between various models, a new methodology is presented. The GOF test results are interpreted by generating 100 synthetic data sets for each model, and estimating the corresponding confidence intervals. Then, pairwise comparisons are performed between each model and synthetic data sets.

According to the result of GOF from PGD and MSE tests, the linear, piecewise linear, and exponential models do not show significant difference. On the other hand, the piecewise constant model fits the failure rates better, in a statistically significant way, than other models.

This result was quite surprising, since we expected that the failure rate to be an increasing function of age. This could be explained by the fact that the faulty cable sections are

TABLE I
Goodness-of-fit measurement by using PGD test

(-Log10 ∏ cdf)	Data A	Data B	Data C	Data D	Data E	Real Data
Constant (Model: A)	38.8769 ±0.5209	43.4234 ±0.6310	42.0862 ±0.6638	42.9481 ±0.6564	42.7505 ±0.6278	43.6205
P.W. Constant (Model: B)	46.9932 ±0.6820	38.4123 ±0.5535	39.8806 ±0.6393	39.4594 ±0.5502	40.8079 ±0.5812	32.9166
Linear (Model: C)	43.8960 ±0.6276	39.2477 ±0.4468	37.8395 ±0.5524	37.0250 ±0.4749	38.0640 ±0.5256	37.0885
P.W. Linear (Model: D)	44.4949 ±0.6098	39.0726 ±0.5864	37.7932 ±0.5307	37.3312 ±0.5280	38.2944 ±0.5902	36.4592
Exponential (Model: E)	42.7155 ±0.6604	39.7050 ±0.5764	37.9634 ±0.4783	37.3717 ±0.4866	37.7189 ±0.4820	37.8377

TABLE II
Goodness-of-fit measurement by using MSE test

(1+e3 mse)	Data A	Data B	Data C	Data D	Data E	Real Data
Constant (Model: A)	1.2409 ±0.0455	1.6574 ±0.0585	1.5392 ±0.0623	1.6216 ±0.0613	1.5937 ±0.0590	1.6408
P.W. Constant (Model: B)	1.9951 ±0.0632	1.2550 ±0.0486	1.3805 ±0.0601	1.3414 ±0.0502	1.4503 ±0.0523	0.8647
Linear (Model: C)	1.7007 ±0.0577	1.3124 ±0.0407	1.1822 ±0.0483	1.1088 ±0.0423	1.1880 ±0.0457	1.2345
P.W. Linear (Model: D)	1.7557 ±0.0570	1.2952 ±0.0520	1.1784 ±0.0460	1.1429 ±0.0467	1.2106 ±0.0517	1.2184
Exponential (Model: E)	1.6032 ±0.0608	1.3525 ±0.0532	1.1819 ±0.0427	1.1278 ±0.0435	1.1592 ±0.0421	1.3008

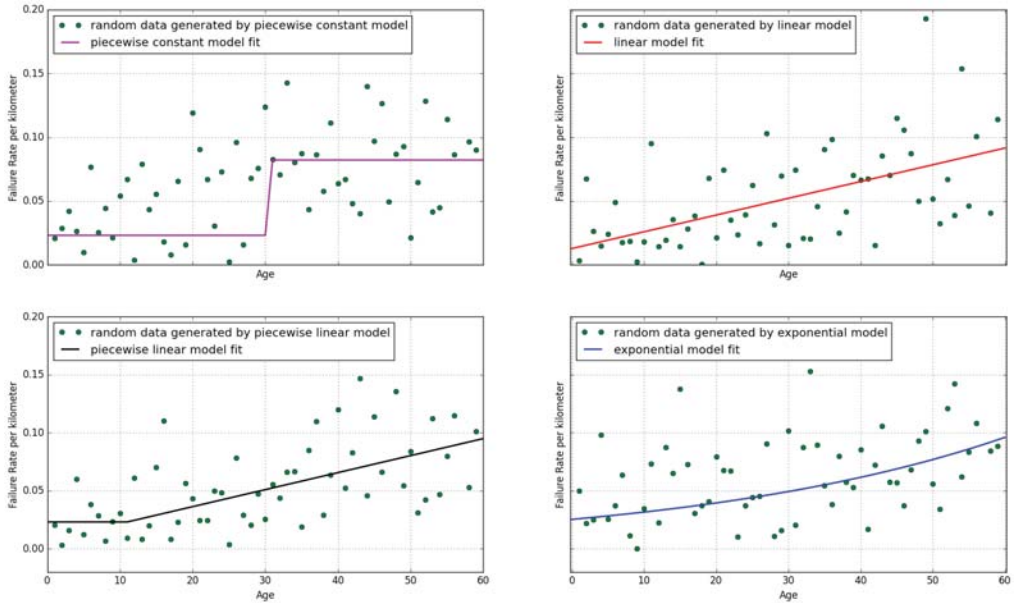


Fig. 5. Synthetic data generated based on different models

continuously replaced by new cables. In fact, the replacement strategy in the underground cables is something we plan to look into in the future in more detail.

In this work we have only considered the failure rate based on the age and the total number of previous failures. However, from the available data set, we can obtain the effects of failure rate based on other factors such as the number of joints, history of previous failures, geographical location, etc. For example, we can cluster cables based on the number of joints per kilometer, and then calculate the failure rate for cables at each cluster. Therefore, by adding more information to the failure rate estimation we can have a better interpretation of the cables failure rate variation over age.

The probabilistic model can also be updated by considering additional information such as load patterns, temperature, and effects of replacement. By exploiting useful information one can determine the condition of in-service equipment, and better plan the scheduling maintenance. Consequently, instead of unplanned outages, power distribution companies can have planned outages, which are shorter and less disruptive.

REFERENCES

[1] H. P. Barringer. Use Crow-AMSAA reliability growth plots to forecast future system failures. *Second Annual International Maintenance Excellence*, 2006.

[2] Paul Barringer. Predict failures: Crow-AMSAA 101 and Weibull 101. *Proceedings of IMEC*, 2004.

[3] J. A. Bloom et al. Guidelines for intelligent asset replacement, vol.3, underground distribution cables. Technical report, EPRI, Palo Alto, CA, 2005.

[4] Jeremy A. Bloom, Charles Feinstein, and Peter Morris. Optimal replacement of underground distribution cables. In *Power Systems Conference and Exposition*, pages 389–393. IEEE, 2006.

[5] R. M. Bucci, R. V. Rebbapragada, A. J. McElroy, E. A. Chebli, and S. Driller. Failure prediction of underground distribution feeder cables. *IEEE Transactions on Power Delivery*, 9(4):1943–1955, 1994.

[6] P Cichecki, E Gulski, JJ Smit, and RA Jongen. Statistical analysis of transmission power cables condition data. *Proceedings of the 16th International Symposium on High Voltage Engineering*, pages 851–856, 2009.

[7] Piotr Cichecki, R Jongen, Edward Gulski, Johan J Smit, Ben Quak, Frank Petzold, and F Vries. Statistical approach in power cables diagnostic data analysis. *IEEE Transactions on Dielectrics and Electrical Insulation*, 15(6):1559–1569, 2008.

[8] J. Y. Gill. *Forecasting Underground Electric Cable Faults Using the Crow-AMSAA Model*. Engineering Information Transfer, 2011.

[9] E. Gulski, J. J. Smit, F. J. Wester, and J. W. van Doeland. Condition assessment of high voltage power cables. In *International Conference on Power System Technology*, volume 2, pages 1661–1666. IEEE, 2004.

[10] Edison Electric Institute. The many causes of power failures. Available online, 2013.

[11] RA Jongen, PHF Morshuis, E Gulski, JJ Smit, J Maksymiuk, and ALJ Janssen. Application of statistical methods for making maintenance decisions within power utilities. *IEEE electrical insulation magazine, 22,(6)*, 2006.

[12] RA Jongen, PHF Morshuis, JJ Smit, ALJ Janssen, and E Gulski. Failure analysis of in service failed resin cable joints by means of a statistical approach. In *Electrical Insulation and Dielectric Phenomena, 2006 IEEE Conference on*, pages 517–520. IEEE, 2006.

[13] R. Keefe. Cable reliability management strategies. Technical Report CA:2004.1002257, EPRI, 2004.

[14] E. L. Mariut and E. Helerea. Enhancing reliability for medium voltage underground power lines. *7th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, pages 1–6, 2011.

[15] Harry Orton. History of underground power cables. *Electrical Insulation Magazine, IEEE*, 29(4):52–57, 2013.

[16] Cristina Stancu, Petru V Notingher, and Mihai Gabriel Plopeanu. Electrical lifetime estimation of underground power cables. *Journal of International Scientific Publications: Materials, Methods & Technologies*, 6(part 1):165–178, 2012.

[17] Zairul Aida Abu Zarim and Tashia Marie Anthony. HV cable diagnostics. In *International Conference on Condition Monitoring and Diagnosis*, pages 1151–1155. IEEE, 2012.

[18] Yujia Zhou and Richard E Brown. A practical method for cable failure rate modeling. In *Transmission and Distribution Conference and Exhibition, 2005/2006 IEEE PES*, pages 794–798. IEEE, 2006.

Use of Social Networks Sites (SNSs) as A Collaborative Learning Technique: Survey Analysis and Mining Approach

Nevine M. Labib[1], Ahmed E. Sabry[1], Rasha H. A. Mostafa[2], Edward W. Morcos[1]

[1]*Computer and Information Systems Department, Sadat Academy for Management Sciences, Cairo, Egypt.*

[2]*Business Administration Department, Faculty of Commerce, Ain-Shams University, Cairo, Egypt.*

Abstract— *This study adopts a multi-disciplinary approach, relating social psychology and information sciences. It aims at measuring the significance of social networks usage in collaborative learning using different information science techniques. After extensive review for relevant literature it has been noticed that the implementation context namely Middle East and North Africa (MENA) region is starving for such stream of researches.*

A number of studies underscored various aspects of the relationship between Social Network Sites (SNS) and collaborative learning such as perception, satisfaction, collaboration, engagement, integration, innovation, performance, interaction, problem solving, motivation, knowledge sharing and discovery, information sharing, and communication.

A survey targeting about 300 students as a sample of relevant stakeholders (users) was conducted over a period of one-year. Three data mining models are implemented using the transformation methods, clustering techniques, and decision tree classification methods. They are all included as part of the triangulation of methods for providing the research analysis higher credibility, reliability and validity.

The originality of this research stems from the following: first, applying novel methodological techniques in social networks domain. Second, improved validity and reliability of the results through triangulation of methods applied.

Keywords: *Social Network Sites (SNS), Data Mining (DM), Decision Tree (DT), Triangulation of techniques, K-means, Clustering, Association Rules, MENA (Middle East and North Africa), Collaborative Learning (CL)*

I. INTRODUCTION

Nowadays, Social Network Sites (SNS) are being used by students, not only for social interactions but also for learning activities since they increase student engagement. Hence, SNS can lead to the creation of virtual communities of learners, which eventually increase the overall learning [1].

Learning activities may include sharing information, doing assignments, discussing issues and other activities that fall under the umbrella of collaborative learning. Nevertheless, there are still more activities to be explored in order to benefit from SNS in the domain of education. This study explores the dimensions of SNS use for collaborative learning, among undergraduate and graduate University students, by means of data mining techniques.

1.1 Problem Definition and Objectives

This study discusses the different dimensions of Social Networks (SNs) in Collaborative Learning among University students using different data mining techniques.

As for the objectives, they consist of the following:

- Analyzing SNS dimensions in the domain of Collaborative Learning among University students in Egypt (based on data collected via structured questionnaire).
- Conducting a comparative study between three different data mining techniques in this domain.
- Comparing the results with statistical outcomes previously revealed.

1.2 Originality and Value

The originality of the study stems from applying different data mining techniques for social networks' use in education in Egypt [2]. Moreover, results' validity and reliability is improved through triangulation of the methods applied.

Results drawn out of this study may help educators to foster student learning by incorporating social media

into taught modules. In addition, they will be able to deal with the negative effects of social media on different types of learners.

1.3 Structure of the paper

It starts by reviewing similar researches that make use of data mining techniques related to the use of Social Networks Sites (SNS's) in collaborative learning among University students. Second, it provides a detailed description of the survey used in the study. Then it discusses the data-mining framework along with the data collection details and relevant results. Finally, conclusions, recommendations, as well as future work are drawn.

II. LITERATURE REVIEW

In order to analyze and discover the role of SNS in education and the different interactions between students, several techniques may be used, whether linear or non-linear. In this paper, we explore the use of data mining techniques to analyze the role of SNS in collaborative learning among undergraduate and graduate students in the Egyptian Universities as part of MENA region.

A survey paper [3] studied several data mining techniques, such as graph theoretic, clustering, recommender system, semantic web, and opinion analysis and classification were used in order to analyze different aspects of SNS. This study showed that data mining techniques are very useful when it comes to retrieving information from a huge amount of data. The selected technique should be based on the kind of data to be analyzed [3].

Another study discusses the usefulness of social media for collaborative learning in higher education by using a social media platform, Graasp. It is implemented in a project-based course and evaluated from different perspectives, such as collaboration and knowledge management. It was found that students were satisfied with using Graaspas it was able to enhance knowledge management and collaboration [4].

A study used Spectral clustering as a data mining method to discover students' behavioral patterns performed in an e-learning system. In order to do so a software was developed. It allowed the tutor to define the data dimensions and input values to obtain appropriate graphs with behavioral patterns that meet his/her needs. Then, the discovered behavioral patterns were compared with students' study performance and evaluation with relation to their possible usage in collaborative learning [5].

A study tackled some SNA techniques, namely community mining, in order to discover relevant structures in social networks. Using new ideas in a toolbox, named "Meerkat-ED", which automatically discovers relevant network structures, visualizes overall snapshots of interactions between the participants aiming to facilitate fair evaluation of students' participation in online courses [6].

III. DESCRIPTION OF CONDUCTED SURVEY

A random sample of three Egyptian public universities students was drawn. The usable sample consisted of 300 students divided evenly between undergraduate and postgraduate. It is a common practice to rely on students' sample, specifically that they are considered heavy users of SNSs [7]. The characteristics of the sampled students are provided in Table1.

Following an extensive review of relevant literature in the areas of Technology Acceptance Model (TAM) and information technology, a self-administered multi-item structured questionnaire was developed to collect data in relation to the research problem. Moreover, seven constructs, including: Perceived Usefulness (PU) measured by four items, attitude measured by three items, and intention to use SNSs measured by four items. All multi-item scales were adapted from Davis [8]. Whereas, Perceived Enjoyment (PE) assessed by four multi-item scale, Perceived Connectedness (PC) measured by three items, and Perceived Involvement (PI) measured by three items were adopted from Nysveen et al., [9]. Further, actual use of SNSs multi item scales for collaborative learning (7 items) and socializing (9 items) were adopted and modified from Saw et al., [10] and Li [11]. All research constructs were assessed on five-point Likert-type scales. In addition, some demographic items were included in the questionnaire.

Table 1: Survey sample characteristics

Characteristics	%
Gender	
Male	54.6 %
Female	45.4 %
Age	
18-22	44 %
22-30	35 %
30 and above	21 %
Favorite SNs	
Facebook	96.6%
YouTube	48.6%
Twitter	26%
Length of Usage of SNs	
Less than a year	6.8 %
1-2 year	14.6 %
2 year+	78.6 %
Frequency of Using SNs	
Several time a day	74 %
Once a day	15.6 %
Every few days/Once a week	10.4 %

IV. MINING FRAMEWORK

Data mining is a term coined to describe the process of shifting through large databases in search of interesting and previously unknown patterns. The accessibility and abundance of data today makes data mining a matter of considerable importance and necessity. The field of data mining provides the techniques and tools by which large quantities of data can be automatically analyzed. Data mining is a part of the overall process of Knowledge Discovery in Databases (KDD).

The following techniques used throughout this research as part of the triangulation of techniques used for validating the results.

1. Unsupervised Clustering Using K-Means.
2. Mining Supervised Classification using Decision Tree.
3. Rules induction using association rules.

A cluster analysis is a type of classification and analysis phase techniques within data mining frameworks. A major issue with cluster analysis is identifying the appropriate number of clusters. Following Lehmann (1979), initial guidelines and the given sample size of 158, the appropriate number of clusters for the available data falls in the range of two to five clusters. Hence, Hierarchical clustering was used to derive solutions within these ranges. "Ward's method was chosen to minimize the within-cluster differences and to avoid problems with "chaining" of the observations found in linkage methods" [12].

One of the most important phases of a Data Mining process (and one that is usually neglected) is that of data exploration through visualization methods.

Visualization feature is considered as one of the important tools for disseminating results in order to discover valid, novel and potentially useful patterns from this relatively highly dimensional and large amounts of data and make use of those patterns to come up with some rules, interpretation, and prediction. The analyzed data cover metrical scales for the computations in addition to nominal scales in the classification process to cover non-numerical values

A classification and clustering computation characteristics are analyzed and described. These characteristics are taken from different prospective cover size, shape, and average density. In addition to these unary features, also binary features or relations between the clusters used. These characteristics then help to identify clusters with similar characteristics, or even to identify objects. The concluded patterns may provide useful input for model-based interpretation.

Researchers mainly used RapidMiner Studio [13] as a data mining modeling and analysis tool. RapidMiner is a code-free modern analytics platform for data ingestion, data blending, predictive modeling, and deployment.

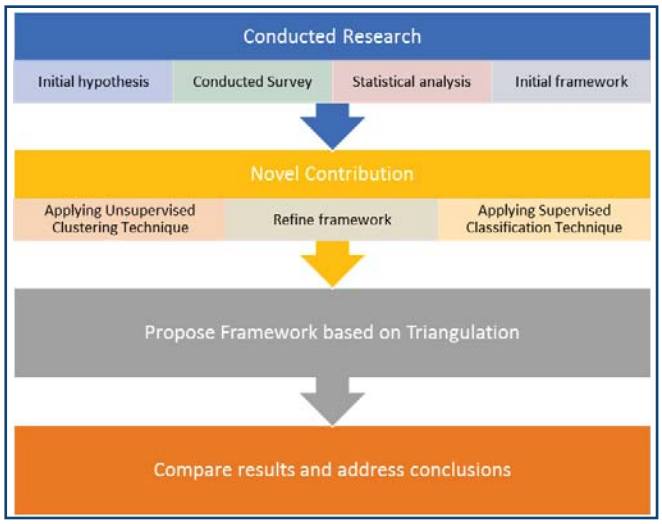


Figure 1: Research Stream Process with Triangulation

V. ANALYSIS AND RESULTS

An unsupervised clustering conducted as a first exploratory mining technique with four and five clusters.

Table 2: Cluster center of gravity (COG) – Four Clusters

Dimension	C1 (n=62)	C2 (n=78)	C3 (n=84)	C4 (n=76)
Socializing	3.37	4.15	3.58	4.36
Usefulness	2.88	3.91	3.44	4.45
Enjoyment	3.06	4.08	3.44	4.54
Attitude	2.99	4.19	3.79	4.46
Intention	2.99	4.13	3.61	4.57
Involvement	2.85	3.25	3.04	3.74
Connectedness	2.99	3.44	3.42	4.24
Learning	2.49	2.00	3.86	3.92

The clustering technique showed that cluster 4 respondents are considered as best users. The findings characterized them as the ones who have high level of perceived usefulness, perceived enjoyment, perceived connectedness and perceived involvement with SNS's. Further, this high level of perceptions has led to positive attitude followed by high level of intentions towards using SNSs in both collaborative learning as well as socializing. It is worthwhile noting that the clustering technique applied did not differentiate between undergraduate and postgraduate respondents.

Table 3: Cluster center of gravity (COG) – Five Clusters

Dimension	C1 (n=59)	C2 (n=51)	C3 (n=43)	C4 (n=70)	C5 (n=77)
Socializing	4.35	3.27	4.42	3.85	3.67
Usefulness	4.44	2.79	4.47	3.50	3.49
Enjoyment	4.52	3.03	4.62	3.60	3.50
Attitude	4.49	2.86	4.50	3.83	3.86
Intention	4.53	2.91	4.57	3.71	3.71
Involvement	3.88	2.78	3.26	3.20	3.05
Connection	4.41	2.91	3.81	3.20	3.46
Learning	4.08	2.93	2.49	1.90	3.93

As revealed no significant clusters centroids data raised from increasing number of clusters.

As shown in Figure 2 the resulted decision tree shows that the high-level using SNS's in collaborative learning mainly by learning construct, BSc, with high enjoyment, intention, while the low level using SNS's in collaborative learning Usefulness and Attitude.

To finalize the knowledge discovery process, another model is developed. It aims at identifying the extent of correlation of these features. It has as input the features extracted from the previous clustering model.

The Rules display the qualified association rules. The rule grid displays all qualified rules and their probabilities (correctness).

Table 4: Clusters general description summary

Cluster	Size of Evaluated Data	Percentage (Density)	Given Name	Description	CL Rank
Cluster 1	2,294	21%	Periodical User	Periodically use SNS's mainly for socializing (responding to others)	4
Cluster 2	2,886	26%	Socializing User	Usually use SNS's and initiating conversations mainly for socialization purposes	3
Cluster 3	3,108	28%	Frequent user	Use SNS's in common purposes including collaborative learning	2
Cluster 4	2,812	25%	High Frequent User	Use SNS's often efficiently with highest enjoyment and collaborative learning best candidate	1
Total	**11,100**	**100%**			

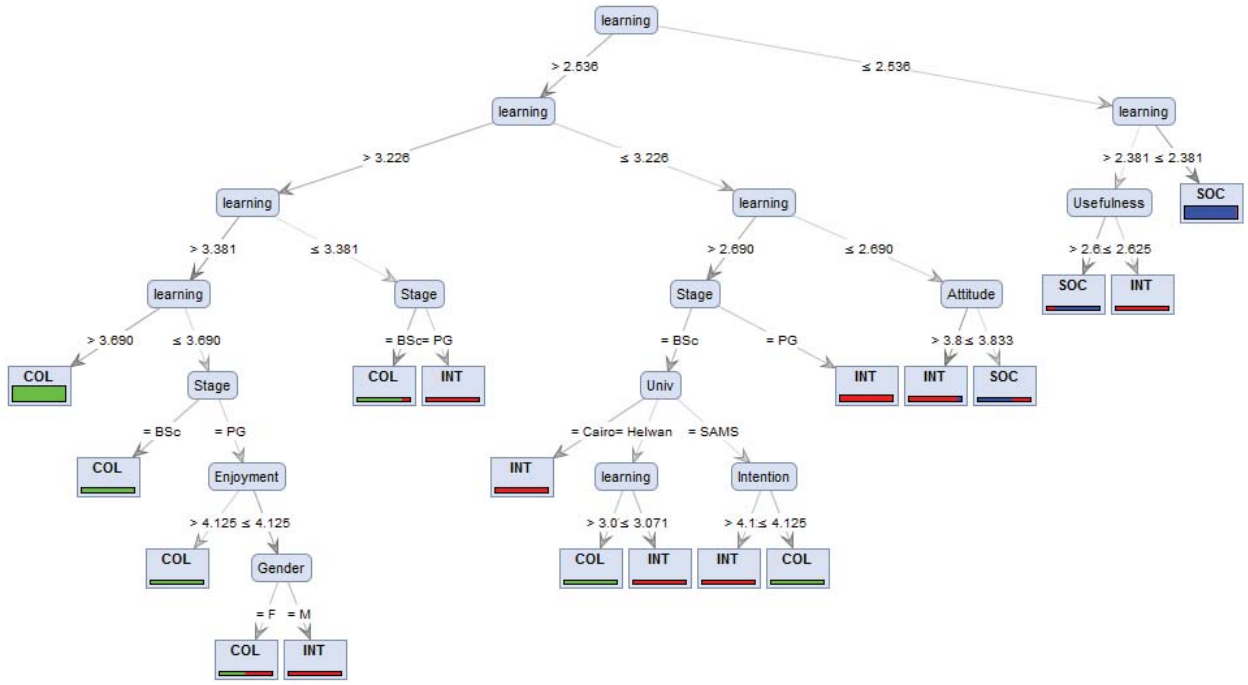


Figure 2: Decision tree induction for the classified data

The following Table 5 shows the set of rules produced by the model, which explains the power of relationship between different attributes.

The main contributors for socializing within inducted rules were mainly measured by the learning group with low values (lower than 2.38) then connection group, while the collaborative learning identified by mainly learning group with high values (higher than 3.2) then the intention group.

Table 5: Rules produced by the applied model.

	Classified as:		
Rules	Socializing	Collaborative	Neutral
if learning > 3.226 then COL	1	129	9
if learning ≤ 2.381 then SOC	83	0	1
if Univ = Cairo then INT	2	0	38
if Usefulness ≤ 3.250 then INT	1	1	10
if Usefulness > 4.375 and Intention ≤ 4.875 then INT	0	0	9
if Univ = Helwan then INT	0	2	5
if Connect > 3.167 then SOC	5	0	0
if Intention ≤ 4.375 then COL	0	3	0
else INT (0 / 0 / 0)	0	0	0

These results conforming to the decision tree results within the learning part while slightly changed in the socializing part.

VI. DISCUSSION

The clustering technique resulted in four clusters. The findings identified cluster 4 respondents as "the best user" of SNSs in both collaborative learning and socializing. Moreover, no differences were identified between undergraduate and postgraduate students.

The results drawn from the association rules technique underscored that the respondents have high level of intentions towards SNSs, which is reflected in their usage in collaborative learning. Yet, the findings did not show any significance between undergraduate and postgraduate with respect to the abovementioned results. Whereas, the Decision Tree technique emphasized that undergraduate students usage of SNS's in collaborative learning was dependent on their level of perceived enjoyment as well as their intentions to use SNS's.

In conclusion, the results of the three data mining techniques applied emphasized the following:

1) All three techniques underscored that intention towards SNS's usage is positively associated with its use in collaborative learning. Likewise, partial support for this result was emphasized by Labib and Mostafa [2] that underscored statistical significant association between intention towards SNSs and collaborative learning among postgraduate students only.
2) 2) Both Decision Tree and clustering techniques identified those respondents Perceived Enjoyment is significantly related to collaborative learning.
3) Clustering and association techniques show insignificant differentiation between undergraduate and postgraduate respondents. Such result is consistent with the same study previously mentioned [2] where statistical results revealed insignificant differences between under and post graduate students in terms of collaborative learning and socializing.

VI. CONCLUSION AND FUTURE WORK

The triangulation of techniques led to reliable results.

The findings did not show significance for any of demographic attributes of respondents, or between undergraduate and postgraduate.

All three techniques addresses that SNSs' usage is positively associated with its use in collaborative learning. The perceived enjoyment, learning, and intentions were the most significantly related constructs to collaborative learning.

Based on the previous results and discussion, a number of issues may be considered as future opportunities to be explored by interested researchers. They are the following

1. Compare between linear and non-linear techniques in the use of SNs in Collaborative learning.
2. Extend the sample used to cover larger demographic scale and to include more dimensions.
3. Use the output of the model to improve the e-learning practices used in education in MENA region.

REFERENCES

[1] K. Tarantino, J. McDonough and M. Hua, "Effects of Student Engagement with Social Media on Student Learning: A Review of Literature," *The Journal of Technology in Student Affairs,* 2013.

[2] N. Labib and R. Mostafa, "Determinants of Social Networks Usage in Collaborative Learning," in *International Conference on Communication, Management and Information Technology*, Prague, 2015.

[3] M. Adedoyin-Olowe, M. Gaber and F. Stahl, "A Survey of Data Mining Techniques for Social Network Analysis," 2014.

[4] S. E. H. D. G. Na Li, "Using Social Media for Collaborative Learning in Higher Education: A Case Study," in *5th International Conference on Advances in Computer-Human Interactions*, Valencia, Spain, 2012.

[5] P. D. J. M. K. S. V. S. Gamila Obadi, "Using Spectral Clustering for Finding Students' Patterns of Behavior in Social Networks," in *DATESO*, 2010.

[6] M. T. a. O. R. Z. Reihaneh Rabbany k., "Social Network Analysis and Mining to Support the Assessment of Online Student Participation," vol. 13, no. 2, 2011.

[7] D. Shin, "Analysis of online social networks: A cross national study," *Online Information Review,* vol. 34, no. 3, pp. 473-495, 2010.

[8] F. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly,* vol. 13, no. 3, pp. 319-40, 1989.

[9] H. Nysveen, P. Pedersen and H. Thorbjornsen, "Intentions to use mobile services: antecedents and cross-service comparisons," *Journal of the Academy of Marketing Science,* vol. 33, no. 3, pp. 330-46, 2005.

[10] G. Saw, W. Abbott, J. Donaghey and C. McDonald, "Social media for international students, it's not all about Facebook," *Library Management,* vol. 34, no. 3, pp. 156-174, 2013.

[11] D. Li, "Online social network acceptance: a social perspective," *Internet Research,* vol. 21, no. 5, pp. 562-580, 2011.

[12] J. J. ,. A. R. T. R. B. W. Hair, "Multivariate Data Analysis," 1998.

[13] RapidMiner, "RapidMiner Studio Manual," RapidMiner, London, 2014.

Wavelet-Coupled Machine Learning Methods for Drought Forecast Utilizing Hybrid Meteorological and Remotely-Sensed Data

R. Tan, and M. Perkowski
Department of Electrical and Computer Engineering, Portland State University, Portland, Oregon, USA

Abstract - *In this study, a statistical drought early warning method is proposed using novel machine learning algorithms, with the inclusion of multiple drought-related attributes from precipitation, satellite-derived land cover vegetation indices, and surface discharge. The forecast is made for the long-term hydrological drought in the region of Central Valley, California. The wavelet transform analysis is employed in combination with support vector regression and artificial neural network algorithms for improving the drought prediction effectiveness. The performance of the drought prediction is evaluated using three statistical metrics: Coefficient of Determination (R^2), Root-Mean-Square Error (RMSE), and Mean-Absolute-Error (MAE). The results clearly indicate that using hybrid precipitation and satellite remotely-sensed data, the proposed wavelet-coupled machine learning method can effectively predict long-term drought in the area of Central Valley California, over a lead time of 3 to 6 months, which is crucial for agricultural planning, reservoir management, and authorities' allocation of water resources.*

Keywords: Drought Forecast, SPI, NDVI, NDWI, Machine Learning, Wavelet Transform

1 Introduction

Among all natural disasters, drought is the most costly environmental catastrophe [1]. As drought is the consequence of precipitation deficiency over an extended period of time, mitigating the detrimental effects of droughts fundamentally lies in the ability to forecast droughts accurately ahead of time to enable the effective planning of water resources.

Due to complex atmospheric processes, accurate drought prediction over a large time span has been one of the biggest challenges in hydrology. In the United States, the U.S. Drought Monitor [2] provides a weekly update of current drought conditions at the national and state levels by publishing an interactive colored-map. However, few uniform drought early warning systems exist globally due to the complexity of the drought process and expensive operational land surface models.

Over the last decade, there has been growing scientific interest in using statistical data-driven methods for drought forecast. Various machine learning algorithms have been investigated, including Autoregressive Integrated Moving Average [3, 4], Artificial Neural Network (ANN) and Support Vector Regression (SVR) [5-8], and Adaptive Neuro-Fuzzy Inference System (ANFIS) [9]. More recently, wavelet analysis has been introduced for analyzing the time series data at different frequency bands, which demonstrates positive impacts in solving a number of problems in water resources when combined with different machine learning algorithms [10-13].

The above studies, however, have been carried out using only precipitation data for drought prediction. Meanwhile, there has been an increasing popularity of using satellite remote sensors for drought condition monitoring, as satellite data are consistently available and nearly continuous in space and time. For example, Y. Gu performed 5-year grassland drought assessment using Normalized Difference Vegetation Index (NDVI) and Normalized Difference Water Index (NDWI) [14], and L. Wang introduced Normalized Multi-band Drought Index (NMDI) for monitoring soil and vegetation moisture with remotely-sensed satellite data [15].

Considering the strong correlation between satellite-derived vegetation indices and drought conditions, this study focuses on combining multiple drought-related attributes, including precipitation, satellite-derived land cover vegetation indices, and surface discharge, for forecasting long-term drought in the region of Central Valley, California. The research first investigates how to characterize drought behavior from different monitoring sources, and how to properly combine those attributes for statistical analysis. The wavelet transform (WT) is incorporated with both the ANN and SVR algorithms for drought forecasting. The results are presented for a forecast lead time of up to six months. The performance is analyzed and future improvement is discussed in the end.

2 Study area and methodology

This study aims to forecast drought in the region of Central Valley, California. Over the last decade, the California State has suffered severe drought over consecutive months, resulting in significant reduction in groundwater level, lake water capacity, stream-flow, and reservoir storage that were all reflected in the long-term precipitation anomalies. Therefore, this study focuses on the long-term drought forecast associated to the hydrological system. The methodology used in this study is shown in Figure 1.

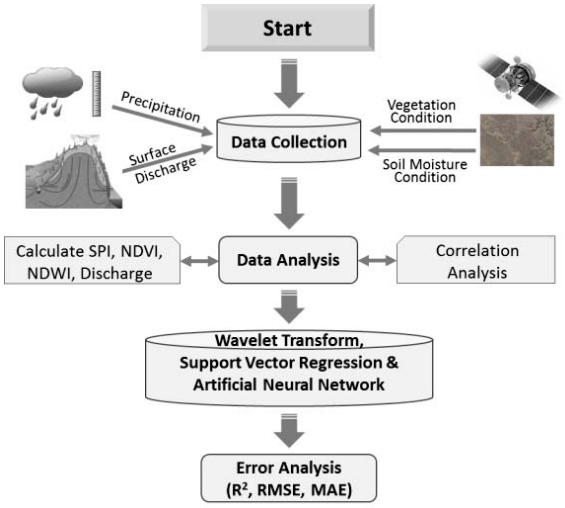


Figure 1: Methodology

The following sections discuss the methods used to characterize drought behavior using the data collected from precipitation, satellite remote sensor and surface discharge. A correlation analysis among those drought attributes is studied for determining the effective inputs to the machine learning algorithms.

2.1 Standardized precipitation index calculation

In this study, the Standardized Precipitation Index (SPI) is used for forecasting drought in California, because it allows for temporal flexibility in evaluation of precipitation conditions [16]. Based on the number of months over which the statistical precipitation is calculated, the SPI is defined as SPI3 and SPI6 that represent short-term agricultural drought, and SPI12 and SPI24 that represent long-term hydrological drought. Here SPI3, SPI6, SPI12 and SPI 24 are the SPI for a period of 3, 6, 12, and 24 months respectively. This study focuses on hydrological drought impact so the forecast will be made based on SPI12 and SPI24.

To calculate the SPI, the monthly precipitation from 1948 to 2014 is collected at the two in-situ weather stations close to the cities of Stockton and Sacramento, CA, from the National Oceanic and Atmospheric Administration (NOAA). The concept of SPI12 and SPI24 calculation is described below:

1) Calculate the cumulative precipitation value for SPI12 and SPI24 for each month from 1948 to 2014.

2) Fit the precipitation data for the same month of each year into a Gamma distribution.

3) Convert the Gamma distribution into a standard Gaussian distribution based on an equal probability transformation.

4) The SPI is a z-score and represents an event away from the mean value in Gaussian distribution.

Based on the calculated SPI values at each time scale, the drought can be classified as given in Table I [16].

Table I: SPI value and drought conditions

SPI Value	Drought Class
SPI ≥ 2.0	Extremely wet
1.5 ≤ SPI < 2.0	Very wet
1.0 ≤ SPI < 1.5	Moderate wet
-1.0 ≤ SPI < 1.0	Normal
-1.5 ≤ SPI < -1.0	Moderate drought
-2.0 ≤ SPI < -1.5	Severe drought
SPI < -2.0	Extreme drought

In this study, the SPI12 and SPI24 are calculated using SPI_SL_6 program developed by the National Drought Mitigation Center, University of Nebraska-Lincoln.

2.2 Grassland vegetation indices from satellite remote sensor

Satellite remotely-sensed data are a promising source of drought condition monitoring as it is possible to measure every component of the hydrological cycle at the land surface and the status of natural vegetation and agriculture, at a very high spatial resolution and nearly real time. To calculate the regional satellite-derived indices for surface land vegetation condition assessment, a study area is chosen near Stockton in the Valley, with a grassland cover that allows the influence of drought to be isolated from other human effects. The satellite remotely-sensed data are acquired from the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard the Aqua and Terra satellites.

Two satellite-derived indices are used to assess vegetation and soil moisture conditions. The Normalized Difference Vegetation Index (NDVI), as defined in Equation (1), is an indication of live green vegetation conditions by detecting the reflection to sunlight at two optical wavelength bands as illustrated in Figure 2. Here ρ_{645nm} and ρ_{860nm} are the reflection detected at 645nm and 860nm respectively. A larger NDVI value means a higher density of green vegetation.

$$NDVI = \frac{\rho_{860nm} - \rho_{645nm}}{\rho_{860nm} + \rho_{645nm}} \quad (1)$$

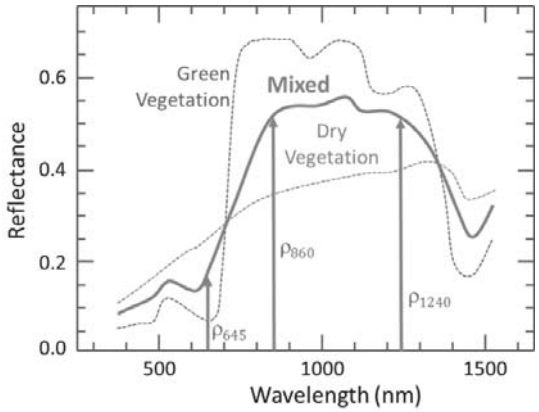


Figure 2: Sunlight reflection vs. wavelength

The Normalized Difference Water Index (NDWI), as defined in Equation (2), uses a similar principle to that for the NDVI to assess soil moisture by detecting the reflection to sunlight at other two optical wavelength bands:

$$NDWI = \frac{\rho_{860nm} - \rho_{1240nm}}{\rho_{860nm} + \rho_{1240nm}} \quad (2)$$

where ρ_{860nm} and ρ_{1240nm} denote the reflection detected at 860nm and 1240nm respectively. The satellite data acquired in this study are 8-day composite of 500-meter surface reflectance data from 2000 to 2014. Data from different optical bands are extracted for multiple pixels. The NDVI and NDWI for each month are calculated according to Equations (1) and (2).

2.3 NDVI12/24 and NDWI12/24 calculation

To utilize satellite-derived indices for drought forecasting, the monthly NDVI and NDWI are converted in the same time scale as SPI12 and SPI24 so that they effectively represent the drought conditions to be predicted. First, the moving average is applied to both NDVI and NDWI at each month over its previous 12-month or 24-month separately. Due to the fact that the satellite data are only available for 14 years, which is not long enough to represent an effective statistical distribution, a simple deviation from the mean value of the overall 14 years is calculated for each month. The newly calculated time series are referred to as NDVI12, NDVI24, NDWI12 and NDWI24 respectively. The time series data for SPI12, NDVI12 and NDWI12 are depicted in Figure 3, and SPI24, NDVI24, NDWI24 are plotted in Figure 4.

The surface discharge or stream-flow data from 2000 to 2014 are obtained from the United States Geological Survey (USGS) gage station located in a natural stream near Stockton. A similar method as described above is applied to the discharge data for calculating Discharge12 and Discharge24, which are newly introduced indices based on surface discharges of 12 months and 24 months.

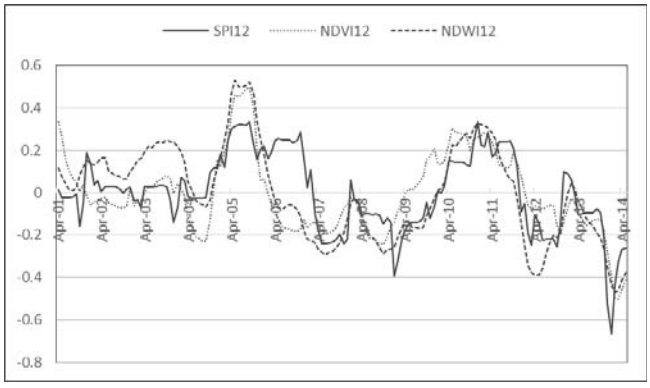


Figure 3: SPI12, NDVI12 and NDWI12 time series plot

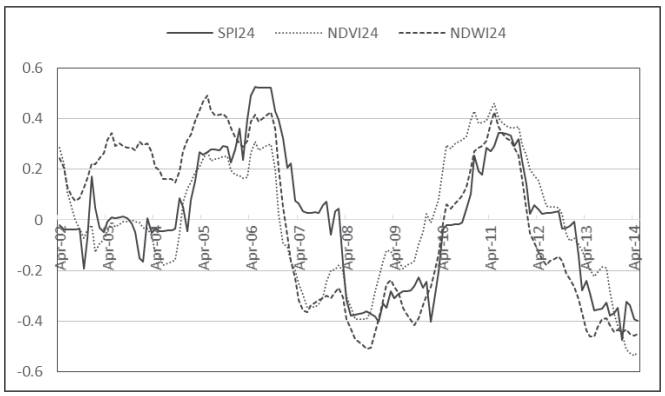


Figure 4: SPI24, NDVI24 and NDWI24 time series plot

2.4 Correlation analysis

Before applying NDVI, NDWI and Discharge for augmenting SPI forecast, a correlation analysis is made to ensure that the data taken for the study area are well correlated with the SPI observation data such that they can be helpful for drought prediction. The correlation results are shown in Table II and Table III respectively.

Table II: Correlation coefficient (R) between SPI12 and NDVI12, NDMI12, and Discharge12

R	NDVI12	NDWI12	Discharge12
SPI12	0.62	0.72	0.57

Table III: Correlation coefficient (R) between SPI24 and NDVI24, NDMI24, and Discharge24

R	NDVI24	NDWI24	Discharge24
SPI24	0.73	0.78	0.69

Table II and Table III clearly indicate that NDVI, NDWI and Discharge all have a reasonably good correlation with the SPI, therefore are useful for augmenting the SPI forecast.

3 Wavelet-ANN and wavelet-SVR

This section discusses the model development using the ANN and SVR, as well as the method used to apply wavelet transform for data pre-processing. The complete data sets from 2000 to 2014 are divided into two sections: the data from 2000 to 2010 are used for training in each machine learning algorithm; the data from 2011 to 2014 are used for validating the performance of the models.

3.1 Artificial neural network

The ANN is a machine learning method which was inspired by how neurons communicate in the human brain. The ANN architecture used in the present study is a feed-forward hierarchical structure that consists of an input layer with multiple input elements, a hidden layer with multiple neurons, and an output layer called the target layer. The ANN used in this study can be represented by [8]:

$$\hat{Y}(t) = f_o\left[\sum_{j=1}^{M} W_j \cdot f_n\left(\sum_{i=1}^{N} W_{ji} \cdot X_i(t) + W_{jO}\right) + W_O\right] \quad (3)$$

where M is the number of neurons in the hidden layer, N is the number of input attributes, i is the input element, j is the hidden neuron, and t is the function of time. X_i, W_{ji}, f_n, f_o and $\hat{Y}$ represent the input viable, the weight operators, the activation function of hidden neuron, the activation function of output neuron, and the forecast output respectively. A recursive, multi-step neural network approach is chosen such that the forecast performance for each leading month can be optimized, as illustrated in Figure 5.

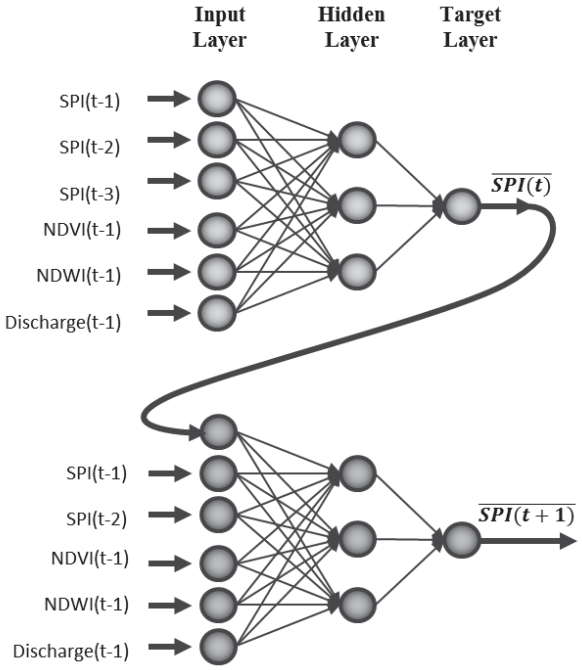


Figure 5: ANN with recursive architecture

In Figure 5, the ANN is trained using Levenberg–Marquardt (LM) algorithm with back propagation [7]. For each leading month forecast, a program is developed using Matlab to optimize the following parameters:

- Number of input combinations
- Number of neurons in the hidden layer
- Activation function in the hidden layer

The optimized ANN model is chosen, which gives the highest determination coefficient (R^2) for the validation data set.

3.2 Support vector regression

The SVR is a novel machine learning algorithm characterized by the usage of the kernel function, ε-insensitive loss function, and capacity control obtained by a trade-off between the margin maximization and the smoothness of the function f(x) [13]. To solve a non-linear regression problem, the input data space is firstly mapped onto an m-dimensional kernel-induced feature space where linear regression can be applied:

$$f(x) = \sum_{j=1}^{m} w_j \times \phi_j(x) + b \quad (4)$$

where w_j is the weight factor, b is the bias term, ϕ_j denotes a set of non-linear transformation functions in the feature space. The goal of the SVR algorithm is to estimate the regression function f(x) that minimizes the following:

$$C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(f(x_i), y_i) + \frac{1}{2}\|w\|^2 \quad (5)$$

$$L_\varepsilon(f(x), y) = \begin{cases} 0, & |f(x) - y| < \varepsilon \\ |f(x) - y| - \varepsilon, & |f(x) - y| \geq \varepsilon \end{cases} \quad (6)$$

Equations 5 and 6 describe the trade-off between the empirical risk and the flatness. The term $L_\varepsilon(f(x), y)$ is called ε-insensitive loss function, where y denotes the observed data with a total set number of n, and ε controls the width of the ε-insensitive zone, which can affect the number of support vectors used to construct the regression function. $C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(f(x), y)$ is called an empirical error, which measures the deviation of the training samples outside of the ε-insensitive zone. C is the control capacity that determines the trade-off between the tolerated empirical error and the flatness of the model. The smoothness of the function is measured by $\frac{1}{2}\|w\|^2$.

In this study the SVR model is developed using ε-SVR function from LibSVM Matlab toolbox (Chang & Lin, 2014). The kernel function uses Radial Basis Function (RBF) characterized by γ [17]. For each leading month forecast, the parameters C, γ, and ε are optimized through a trial-and-error method for getting the best R^2 for the validation data set.

3.3 Wavelet analysis for data pre-processing

Despite the fact that both the ANN and SVR are powerful in dealing with non-linear hydrologic problems, they both have limitations that input data must be stationary for a reliable operation. By analyzing the SPI12 and SPI24 time series data and their autocorrelation function (ACF), it can be found that SPI12 and SPI24 are not highly stationary, with the ACF showing a slowly decaying sinusoidal behavior over a number of lags. To mitigate this shortcoming, the wavelet analysis is applied for data pre-processing that allows the use of larger time intervals for more precise low-frequency information and shorter time intervals for extracting high-frequency information, thus generating a time-frequency representation of the time series signal. Using the wavelet analysis algorithm, the original time series data can be hierarchically decomposed into N-level sub-series at different frequency bands for noise reduction or peak detection. This technique is especially useful for analyzing time domain waveforms where sharp spikes need to be localized, which appear to be the case for data plotted in Figure 3 and Figure 4.

In this study, the discrete wavelet transform (DWT) is used to decompose each of the input attributes to the ANN or

SVR models into a number of sub-series at different resolution levels. The Daubechies mother wavelet is used, which provides a family of wavelets called dbN, where N is the order of wavelets. The data decomposition is done by passing the input data through a series of high-pass and low-pass filters, to obtain the detailed series (D) and approximation series (A), as illustrated in Figure 6.

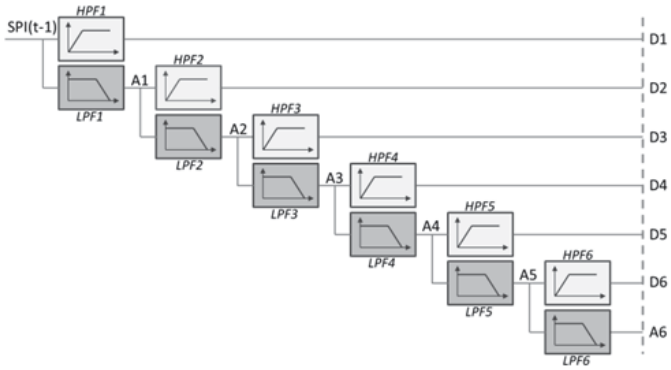


Figure 6: 6-level WT decomposition for SPI(t-1)

As an example, Figure 7 shows the SPI12(t-1) decomposed waveform for D1-D6 and A6, using Daubechies db1 as the master wavelet.

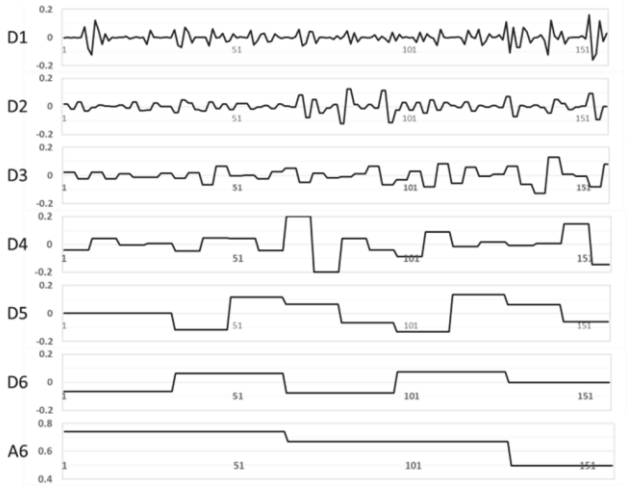


Figure 7: SPI12 decomposed waveforms using DWT

After data decomposition, a portion of the signal associated with certain frequency bands will be eliminated if there is a poor correlation between the decomposed signal and the observation data. Only the decomposed signals that have significant correlation with the observation signal will be used in the forecast model. Figure 8 shows the methodology of using the WT-ANN (WT based ANN) and the WT-VSR (WT based VSR) for forecasting SPI12 and SPI24.

After a correlation analysis, it is found that the newly formed time series data using the DWT algorithm appear to have a better correlation with the observation data than the original time series. Therefore using the DWT for data pre-processing should be helpful for improving SPI12 and SPI24 prediction accuracy. For each leading time prediction using the WT-ANN and WT-SVR, a program is developed using Matlab that optimizes drought forecast performance by using different Daubechies wavelets for data decomposition. The wavelet that gives the highest R^2 and the lowest MAE and RMSE will be selected.

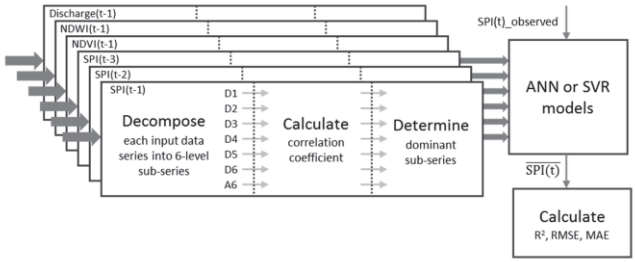


Figure 8: WT-ANN and WT-SVR configuration

4 Performance evaluation

To assess the performance of the WT-SVR and WT-ANN models, three statistical performance evaluation criteria are used: R^2, RMSE and MAE. The R^2 measures the degree of linear correlation between the predicted data and the observed data, the RMSE gives the variant of the total errors, while the MAE provides the absolute error information. The models with the highest R^2 and the lowest RMSE and MAE indicate the best performance. The R^2, RMSE and MAE are defined in following equations:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (8)$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (9)$$

where n is the number of data points, y_i is the observed value, $\hat{y}_i$ is the predicted value, and $\bar{y}_i$ is the mean value of the observation data.

5 Results and discussion

Using the algorithms described above, drought forecast based on SPI12 for a lead time of up to 3 months and SPI24 for a lead time of up to 6 months have been investigated using the statistical methods described above. The forecast performances are evaluated on the validation data sets from January 2011 to May 2014, and the results are presented in Table IV and Table V. Figure 9 and Figure 11 illustrate the time series plots for SPI12 and SPI24 using the WT-ANN, including observation, training and validation data sets; Figure 10 and Figure 12 show the corresponding scatter plots for the validation data set, at a 1-month lead time.

Table IV: SPI12 prediction using ANN and SVR

Leading Time	WT-ANN			WT-SVR		
	R^2	RMSE	MAE	R^2	RMSE	MAE
1-month	0.91	0.29	0.21	0.91	0.29	0.21
2-month	0.74	0.52	0.34	0.73	0.50	0.32
3-month	0.69	0.59	0.41	0.60	0.62	0.40

Table V: SPI24 prediction using ANN and SVR

Leading Time	WT-ANN			WT-SVR		
	R^2	RMSE	MAE	R^2	RMSE	MAE
1-month	0.97	0.13	0.10	0.97	0.14	0.11
2-month	0.95	0.19	0.12	0.95	0.19	0.15
3-month	0.95	0.21	0.16	0.95	0.19	0.15
4-month	0.95	0.20	0.17	0.94	0.21	0.17
5-month	0.93	0.24	0.18	0.93	0.27	0.21
6-month	0.94	0.21	0.17	0.90	0.34	0.27

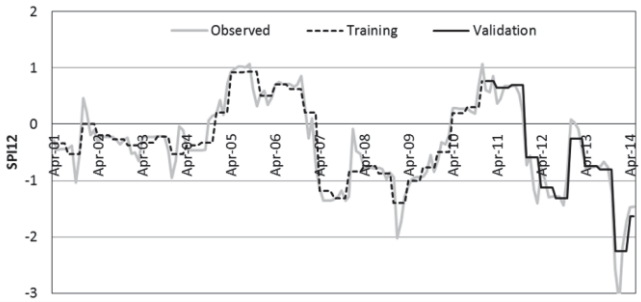


Figure 9: SPI12 time series for observation, training and validation using WT-ANN

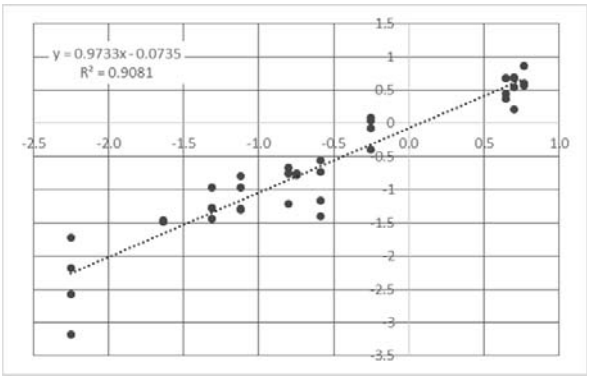


Figure 10: SPI12 scatter plot using WT-ANN

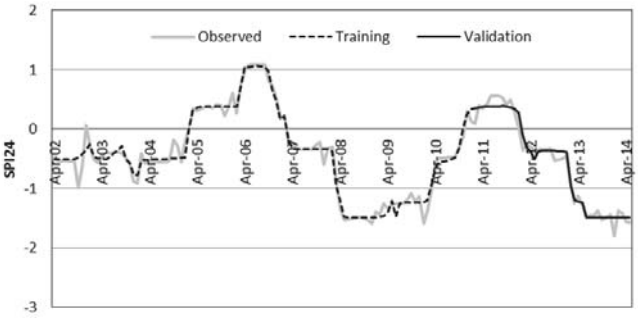


Figure 11: SPI24 time series for observation, training and validation using WT-ANN

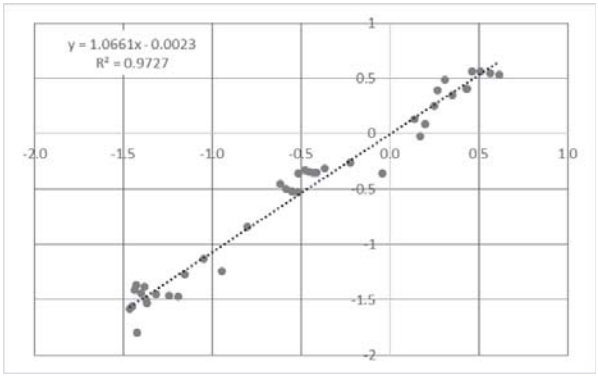


Figure 12: SPI24 scatter plot using WT-ANN

To demonstrate the impact of using satellite data on drought forecast, Figure 13 and Figure 14 illustrate the results of comparison between analyses with and without satellite data, when the same WT-ANN model is used.

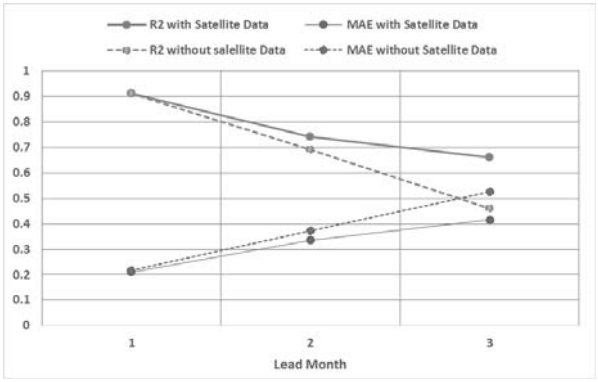


Figure 13: SPI12 forecast analyses with and without satellite data

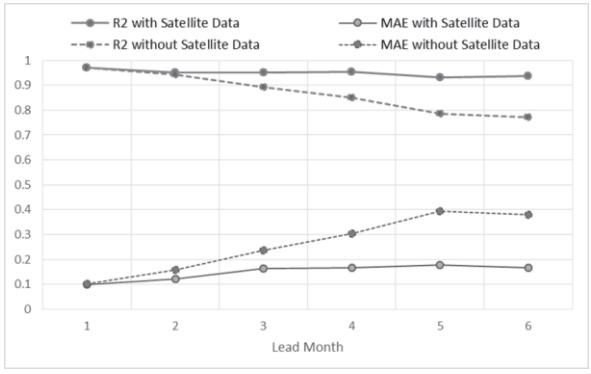


Figure 14: SPI24 forecast analyses with and without satellite data

The aforementioned results indicate the following:

1) Using both the WT-ANN and WT-SVR models, the inclusion of satellite data provides a better accuracy than the results obtained with precipitation data only. The effect of hybrid precipitation and satellite data becomes more obvious as forecast leading time increases. This can be explained by the partial autocorrelation function analysis which shows a strong correlation between SPI(t) and SPI(t-1), therefore, for

one-month lead time prediction, the SPI(t) will be dominated by SPI(t-1).

2) For the SPI24 forecast, the obtained R^2 is greater than 0.9 using both the WT-ANN and WT-SVR methods for a lead time of 1-6 months ahead. The superior forecast accuracy is partially due to the fact that SPI24 represents the average precipitation over a period of 24 months, therefore is less sensitive to the monthly variation in precipitation.

3) For the SPI12 forecast, both the WT-ANN and WT-SVR indicate R^2 greater than 0.9 for a one-month lead time, but the results are getting worse as lead time increases. This can be explained by the SPI12 observation, training, and validation plots in Figure 9, which shows sharp spikes in multiple time slots that are not well predicted even for one-month lead time. This error will be accumulated as lead time increases. In fact, using Daubechies wavelet for data pre-processing has helped greatly in reducing the sensitivity to big changes in monthly precipitation within the SPI12. Further improvement of SPI12 forecast can be investigated by using other types of wavelet transform or parameter optimization of machine learning algorithms.

6 Conclusions

This study aims to investigate the ability of machine learning methods for long-term hydrological drought forecast in the region of Central Valley, CA, based on SPI12 and SPI24. A solution is proposed, for the first time, using hybrid precipitation and satellite remotely-sensed data for drought forecast. The results indicate that integrating precipitation and remotely-sensed data is a promising solution for an effective drought forecast. It is also demonstrated that combining wavelet analysis with novel ANN or SVR is a powerful method for improving the drought forecast accuracy, especially when data is not stationary. To enhance the drought prediction capability, especially for SPI12, further work can be carried out including an investigation of other wavelet functions, as well as taking other drought-related attributes into account, such as ground water level, snowpack and soil moisture. A hybrid geographic and statistical method can also be considered for future study [18].

7 References

[1] D. A. Wilhite and M. D. Svoboda, "Drought Early Warning Systems in the Context of Drought Preparedness and Mitigation", National Drought Mitigation Center, Lincoln, Nebraska, U.S.A.

[2] "United States Drought Monitor", http://droughtmonitor.unl.edu/

[3] K. Shatanawi, "Characterizing, Monitoring and Forecasting of Drought in Jordan River Rasin", Journal of Water Resource and Protection, pp. 1192-1202, May 2013

[4] J. Adamowski, H. F. Chan, S. O. Prasher, B. Ozga-Zielinski and A. Sliusarieva, "Comparison of Multiple Linear and Nonlinear Regression, Autoregressive Integrated Moving Average, Artificial Neural Network, and Wavelet Artificial Neural Network Methods for Urban Water Demand Forecasting in Montreal, Canada", Water Resources Research, Vol. 48, W01528, 2012

[5] A. K. Mishra, and V. R. Desai, "Drought Forecasting Using Stochastic Models", Stoch Environ Res Risk Assess, pp.326-339, June 2005

[6] A. K. Mishra, V. R. Desai, and V. P. Singh, "Drought Forecasting Using a Hybrid Stochastic and Neural Network Model", Journal of Hydrologic Engineering, pp. 626-638, December 2007

[7] S. Morid, V. Smakhtin, and K. Baghertadeh, "Drought Forecasting using Artificial Neural Networks and Time Series of Drought Indices", International Journal of Climatology, pp. 2103-2111, April 2007

[8] A. Belayneh, and J. Adamowski, "Drought Forecasting using New Machine Learning Methods", Journal of Water and Land Development, pp. 3-12, No. 18, 2013

[9] B. Shirmohammadi, H. Moradi, V. Moosavi, M. T. Semiromi, and A. Zeinali, "Forecasting of meteorological Drought using Wavelet-ANFIS Hybrid Model for Different Time Steps", Springer Science + Business Media Dordrecht, May 2013

[10] J. R. Mohammed, and H. M. Ibrahim, "Hybrid Wavelet Artificial Neural Network Model for Municipal Water Demand Forecasting", ARPN Journal of Engineering and Applied Sciences, pp. 1047-1065, Vol. 7, No. 8, August 2012

[11] A. Belaynch, and J. Adamowski, "Standard Precipitation Index Drought Forecasting Using Neural Networks, Wavelet Neural Networks, and Support Vector Regression", Applied Computational Intelligence and Soft Computing, 2012

[12] A. Belaynch, J. Adamowski, B. Khalil, and B. Ozga-Zielinski, "Long-term SPI Drought Forecasting in the Awash River Basin in Ethiopia Using Wavelet Neural Network and Wavelet Support Vector Regression Models", Journal of Hydrology, pp. 418-429, 2013

[13] Q. Feng, X. Wen and J. Li, "Wavelet Analysis-Support Vector Machine Coupled Models for Monthly Rainfall Forecasting in Arid Regions", Water Resour Manage, # 29, pp1049 – 1065, 2015

[14] Y. Gu, J. Brown, J. P. Verdin, and B. Wardlow, "A Five-year Analysis of MODIS NDVI and NDWI for Grassland Drought Assessment over the Central Great Plains of the United States", Geophysical Research Letters, Vol. 34, L06407, 2007

[15] L. Wang and J. J. Qu, "NMDI: "A Normalized Multi-band Drought Index for Monitoring Soil and Vegetation Moisture with Satellite Remote Sensing", Geophysical Research Letters, Vol. 34, L20405, 2007

[16] "The Standard Precipitation Index (SPI)", http://drought.unl.edu/ranchplan/DroughtBsics/WeatherDrought/MeasuringDrought.aspx

[17] A. J. Smola and B. Scholkopf, "A Tutorial on Support Vector Regression", NeuroCOLT2 Technical Report Series, 1998

[18] H. Yan and F. Edwards, "Effects of Land Use Change on Hydrologic Response at a Watershed Scale, Arkansas", J Hydrol Eng., Vol. 18, #12, pp. 1779–1785, December 2013

A Hierarchical Clustering Approach to Analyze Similarities between Sea Surface Temperature Patterns in the Caribbean

Marc Boumedine
Computational and Computer Sciences Department
College of Science and Mathematics
University of the Virgin Islands
U.S. Virgin Islands, St. Thomas, 00802
mboumed@uvi.edu

***Abstract*—This study presents a clustering approach to analyze similarities between sea surface temperature patterns in the Caribbean. Our goal is to supplement existing predictive systems with data mining techniques by automatically extracting new patterns and ultimately increase the predictive accuracy of coral reef monitoring systems. The approach presented analyze times series sea temperature data from 2000 until 2014 collected the National Oceanographic and Atmospheric Administration National Environmental Satellite, Data and Information Science (NOAA/NESDIS) Coral Reef Watch(CRW) monitoring system. Unsupervised techniques (cluster analysis) are performed on Twenty three virtual stations in the Caribbean in an attempt to discover similarities and stressing patterns that might affect coral reef health in the region. The approach follows main three steps: (1) raw data selection and processing, (2) discretization and dimensionality reduction of time series using the Symbolic Aggregate Approximation (SAX), and (3) determination of sequence similarities/dissimilarities and hierarchical clustering of virtual weather stations according to sea surface temperature patterns.**

Keywords—Times Series Similarity; Discretization; SAX; Data Mining; Hierarchical Clustering; Coral Reel Ecological Systems (key words)

I. Introduction

Sea warming acts as an environmental stressor on coral health and may have a devastating impact on Caribbean economies. Environmental agencies and decision makers are strongly committed in assessing the impact of climate change and land use on coral health [1-3]. In order to support this type of assessment, intensive data analysis is required to better understand this phenomena. Due to the vast amount of data to process, this analysis can be aided by automated or semi-automated computational tools in order to discover interesting patterns, anomalies or trends in various data sources available such as atmospheric, oceanographic, biologic etc. Coral reef organisms are very extremely sensitive to changes (increase or decrease) in water temperatures. Ocean excessive warming causes coral polyps to expel the symbiotic algae (called zooxanthellae), essential for its survival. Once the algae is expelled, coral polyps look white or bleached. If stressing conditions persist, the coral will likely die. In order to monitor ocean warming and impacts on ecosystems, complex sensor networks and satellite imaging systems have been deployed. These systems collect large temporal data sets that can be exploited for discovering any hidden structure or useful patterns leading to stressing conditions.

This work investigates relationships between sea surface temperature patterns and stressing episodes that might threaten coral reef ecosystems. We specifically focus on similarities (dissimilarities) between SST patterns occurring in the Caribbean. Data sets have been obtained from the National Oceanographic and Atmospheric Administration National Environmental Satellite, Data and Information Science (NOAA/NESDIS) Coral Reef Watch(CRW) monitoring system. We analyze times series sea surface temperatures (SST) from January 2000 until October 2014 sampled from twenty three virtual stations in the Caribbean. We use unsupervised techniques (cluster analysis) of time series in an attempt to discover stressing patterns and trends that might affect coral reef health in the region. Data transformations are carried out to allow a lower dimensionality representation, analysis and visualization.

The major contribution of this paper is the application of SAX algorithm in an attempt to discover association pattern between thermal stress and coral bleaching alerts in the Caribbean. This new approach offers a new way of analyzing data collected from (NOAA/NESDIS) Coral Reef Watch(CRW) monitoring system and providing new insights on analyzing globally the effect of SST on fragile ecosystems such as coral reefs.

The approach described in this study presents the three steps: (1) raw data selection and processing, (2) using the Symbolic Aggregate Approximation (SAX), (3) determination of sequence similarities using hierarchical clustering and (4) Discussion and validation of the results. The remainder of this paper is organized as follows: section II describes the background and previous work, section III describes the methodology, section IV reviews SAX method , section V presents the clustering approach, and finally section VI presents the resulting clusters .

II. Background and Previous Work

The past two decades, intense efforts have been developed to monitor sea surface temperatures via remote sensing and in situ technologies. As a result, an increasing number of applications and opportunities are becoming available to drill into these data sets and contribute to developing ecological forecasting system in the Caribbean and globally . Building empirical models is time consuming and requires very specific knowledge of the domain. As the number of environmental variables increases, it is imperative to derive models assisted with machine learning and data mining techniques [4]. Machine learning techniques have been successfully used in many knowledge discovery applications [5-6]. However, despite the availability of NOAA's products and services, there has been very little time series data mining research conducted on NOAA produced SST data sets [7]. NOAA/NESDIS Coral Reef Watch(CRW) monitoring system provides five thermal stress alert levels: *no stress*, *bleaching watch*, *bleaching warning*, *bleaching alert 1* and *bleaching alert 2 (see Fig. 3)*. These levels are calculated based on NOAA's cumulative sum Degree Heating Week model [8]. Using unsupervised techniques (cluster analysis) on SST time series our ultimate goal is to discover association patterns between SST and thermal stress that might affect coral reef health in the Caribbean region.

Table 1. Partial SST St. Croix U.S.Virgin Islands- http://coralreefwatch.noaa.gov/satellite/index.php

Date	SST	SST ANO.	HoT SPOT	DHW	Lat	Lon
11/28/2000	27.1	-0.5	0	0	18	-65
12/2/2000	27	-0.6	0	0	18	-65
12/5/2000	26.9	-0.6	0	0	18	-65
12/9/2000	26.9	-0.5	0	0	18	-65
12/12/2000	26.8	-0.5	0	0	18	-65
…						

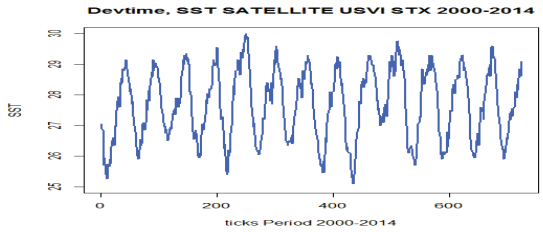


Fig. 1: SST times series for the US Virgin Islands(USVI) St. Croix Station January 2000-October 2014

Most algorithms reduce time series dimensionality using different representation in order to manage computational cost [9]. This is usually accomplished by preserving the general trends of the data using techniques such as single value decomposition, discrete Fourier transformation, piecewise aggregate approximation, adaptive piecewise constant approximation and symbolic aggregate approximation (SAX).

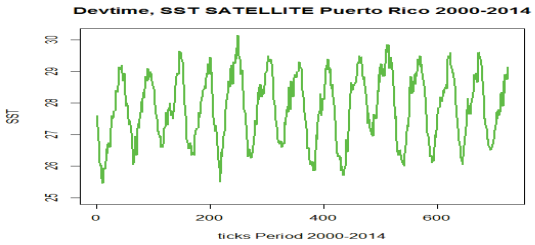


Fig. 2: SST times series for Puerto Rico Station January 2000-October 2014

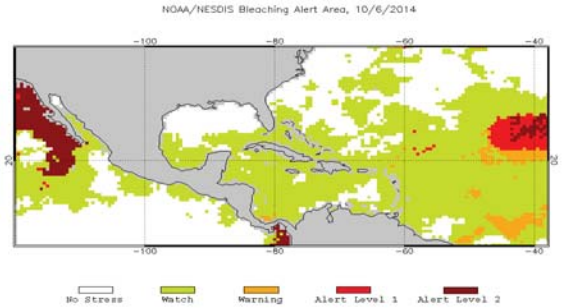


Fig. 3: SST times series for US Virgin Islands St. Croix Station January 2000-October 2014

Clustering algorithms seek to group object that are the most similar in the same cluster while minimizing similarity between clusters. Hierarchical clustering algorithms produce a nested representation represented graphically as a dendrogram which is easier to interpret and validate by domain experts.

Similarity (dissimilarity) measures can be expressed using a variety of approaches such as Euclidean distance, Dynamic Time Warping, distance based on Longest Common Subsequence.

III. Methology

The overall methodology is shown on Fig. 4. As mentioned previously, our goal is to supplement existing models by automatically extracting new knowledge from NOAA data sets. In particular, this study focuses on finding similarities/dissimilarities between virtual stations in order to detect SST patterns as precursor of thermal stress leading to coral bleaching episodes in the Caribbean. The methodology is summarized below.

- We analyze SST times series obtained from (NOAA/NESDIS) Coral Reef Watch Monitoring systems. from January 2000-October 2014 sampled from virtual stations in the Caribbean (see Figures 1 and 2). Sea Surface Temperatures (SST) observations are sampled twice weekly at night-time at 0.5-degree (50-km) resolution by infrared radiometers**.** SST Times series are accessible from NOAA Coral Reef Watch portal (see Table 1).

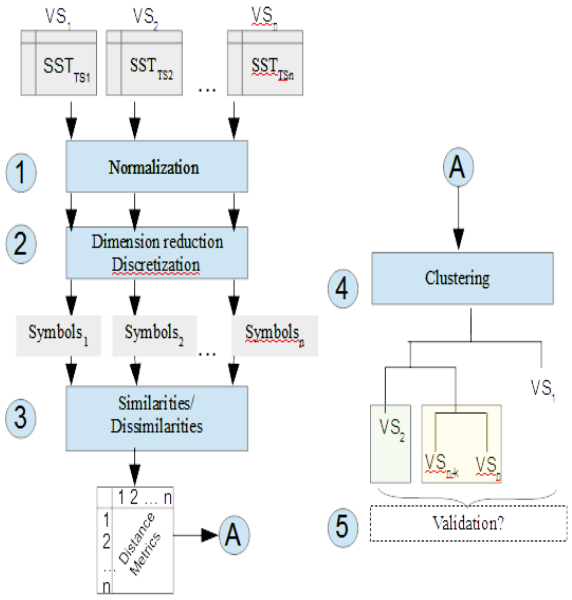


Fig.4 : Overall approach for constructing the clusters between virtual stations

- Time series are then normalized and transformations are carried out using Symbolic Aggregate Approximation [10] and to allow a lower dimensionality representation for analysis efficiency and visualization purposes (Fig. 4 steps 1 & 2). Partial samples observations are shown in Table I. Figs. 2 and 3 show NOAA/NESDIS SSTs observed at nighttime for two stations (U.S. Virgin Islands and Puerto Rico) in order to reduce variability due to solar glare.
- Hierarchical clustering is applied based on a pairwise distance matrix (see Fig.4,steps 3&4).
- The resulting dendograms are validated the the approximately unbiased and bootstrap probability Value [12]

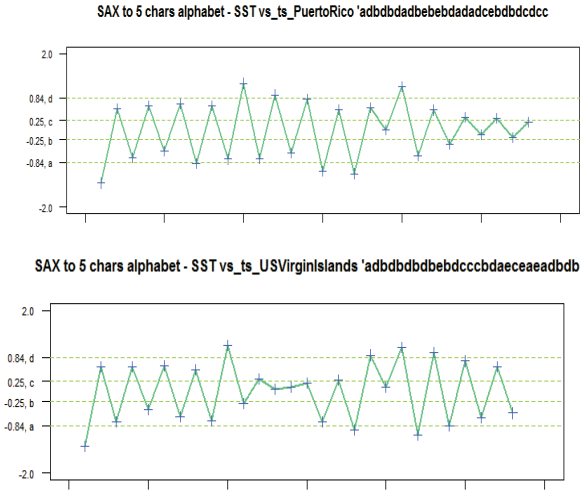


Fig. 5: SAX symbols representing USVI and Puerto Rico SST

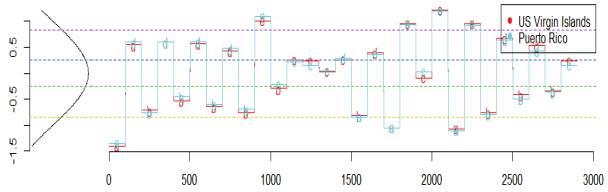


Fig. 6: SAX transformation of SST for USVI and Puerto Rico

IV. Symbolic Aggregate Approximation (SAX)

Sea Surface Temperatures (SST) observations are sampled twice weekly at night-time at 0.5-degree (50-km) resolution by infrared radiometers**.** SST Times series are accessible from NOAA Coral Reef Watch portal. Since our approach focus on SST trends (increase or decrease), we are particularly interested in high-level representation representation of the data. The piecewise aggregate approximation (PAA) and Symbolic Aggregate Approximation (SAX) have been widely used to compare sequence similarities for their interesting properties. SAX is briefly reviewed in the following section.

SAX is used to transform original time series into a symbolic representation [10-11] while preserving essential trends (see Fig. 5). This approach is based PAA representation and required to normalize SST observations in order to take advantage of Gaussian distribution properties. Normalized SST vectors are reduced using PAA which produces equal sized segments. The segments are converted into a symbolic representation (or a string) using an alphabet of symbols. Because of the five coral reef stressing level we chose to represent the size of the alphabet with the set {a,b,c,d,e}. Each symbols is then assigned to an equal sized interval under the Gaussian curve (see Fig. 6).

Lin and Keogh have shown that the distance measure between two symbolic strings created by SAX is a lower bound of the true distance between the two original time-series [10]. Since the times series for all the virtual stations have the same length the Euclidean distance applied to derive the distance matrix used in the hierarchical clustering process.

V-Clustering Approach

The purpose of the hierarchical clustering process is to reveal any similarities/dissimilarities between SSTs at various Caribbean virtual stations in an attempt to discover relationships between SST and coral reef thermal stress episodes. These episodes are determined using the Degree Heating Weeks (DHW) variable. DHW represents the weekly accumulation of heat exceeding the coral bleaching threshold. DHW values are available for each virtual stations for the period 2000-2014. In order to clusters the set of virtual stations represented by the sequences of symbols (or strings) we are using hierarchical clustering algorithm for both SST and DHW SAX symbols. Basically, the algorithm attempts to group the N

virtual stations (N=23) based on the *N*N* similarity/dissimilarity matrix between all sequences of symbols. Fig. 5 shows SAX symbols obtained after PAA transformation for Puerto Rico and US Virgin Islands stations. The length of the original time series have been reduced to twenty-nine (w=29). The alphabet size is five (alpha=5).

Clustering process

1.Initially, each object (sequence of symbols) is assigned to a cluster.

2.The closest (most similar) couple of clusters are identified and merged.

3. The similarity matrix is updated by computing the distance between the new merged cluster and other clusters.

4.Steps 2 and 3 are repeated until all objects are grouped into a single cluster.

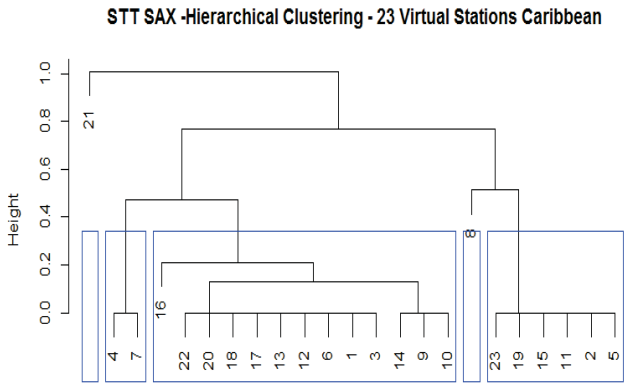


Fig.7: Dendogram for SST obtained with average method, Euclidean distance

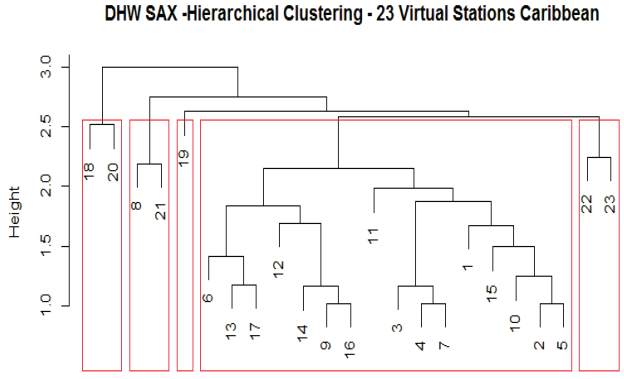


Fig. 8- Dendogram for 23 stations based on Degree Heating Week

7.VI- Preliminary Results

We run experiments with the single, complete and average linkage hierarchical clustering. In [17] Kaufman and Rousseeuw have shown that average linkage is a good technique performs with Euclidean distances but others metrics as well such . The results presented on Figs. 7 & 8 show clusters that were generated by the average-linkage clustering algorithm from the Tclust R package [16] The distance between each pair of clusters is the average distance from any items of one cluster to any items of the other clusters [13]. Table 2 list the names of the virtual stations with their associated numbers used on the dendograms shown on Figs. 7-10.

Table 2: Caribbean Virtual Stations

1) Banco Chinchorro, Mexico	9)Flower Garden Banks, Texas	17)Negril, Jamaica
2)Barbados	10)Glovers Reef, Belize	18) Puerto Morelos, Mexico
3)Bay Islands, Honduras	11) Guadeloupe	19) Puerto Rico
4) Bocas del Toro, Panama	12) Isla de la Juventud, Cuba	20)San Bernardo, Colombia
5)Bucco Reef, Tobago	13)Jardines de la Reina,	21) Santa Marta, Colombia
6) Cayman Islands	14)Lee Stocking Island, Bahamas	22)Turks and Caicos
7)Cayos Miskitos, Nicaragua	15) Los Roques, Venezuela	23) US Virgin Islands
8)Curacao and Aruba	16)Montecristi, Dominican Republic	

In order to assess results, the validation consists of analyzing the results and determine if the partitioning best fits the patterns represented by the SAX strings. Various cluster validity approaches have been proposed in literature such as the Adjusted Rand Index, the Silhouette Width, the Dunn Index [14-16]. Our approach assess the validity through the stability criteria. This criteria will guarentee that the clustering output will be similar for two different time series. In order to measure stability and reduce the uncertainty of the clustering process, *p* values are calculated based on the multiscale bootstrap resampling technique developed by Suzuki and Shimodaira [12]. Two indicators, the approximately unbiased (AU) and Bootstrap Probability Value (BP), provide a level of support into the reliability of the structure obtained with the hierarchical clustering algorithm. Fig. 9 shows p-values greater than 95% which suggest that the clusters were not generated by chance.

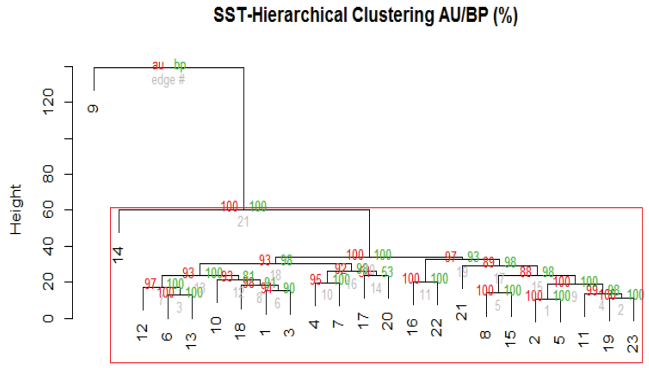


Fig. 9. Dendogram for SST with AU/AP percentage for Caribbean Stations

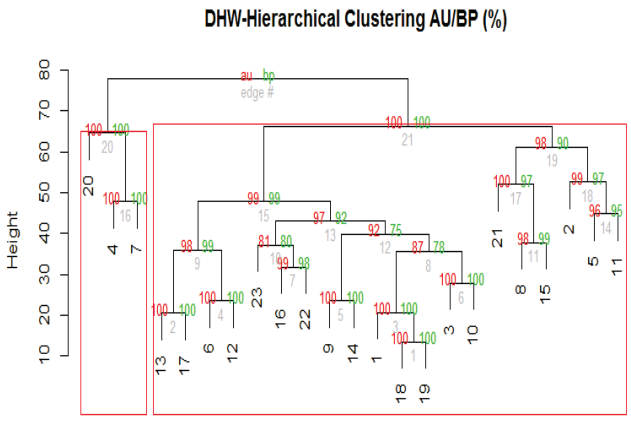


Fig. 10. Dendogram for DHW with AU/AP percentage for Caribbean Stations

VII- Conclusions and Future Work

In this work, we proposed a hierarchical clustering approach in an attempt to reveal relationships between common SST patterns and thermal stress patterns (DHW) that might affect coral reef health. Observations from twenty-three virtual stations from 2000 until 2014 have been transformed using SAX and agglomerated into dendograms for further analysis. Before mapping SST patterns and DHW patterns the stability of the results have been assessed. The mapping between SST and DHW clusters will be presented in future work.

References

[1] C. M. Eakin, J.M. Lough and S.F. Heron (2009). Climate Variability and Change: Monitoring data and evidence for increased coral bleaching stress. In M. Van Oppen & J.M. Lough [Eds.], Coral Bleaching: Patterns, Processes, Causes and Consequences. Ecological Studies 205, Springer, Berlin. 178 pp.Amigó G., Gonzalo, J. Artiles J., and Verdejo F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints," Inf. Retr. Boston., vol. 12, no. 4, pp. 461–486, 2009

[2] G.M. Wellington, P.W. Glynn, A.E. Strong, S.A. Navarrete, E. Wieters and D. Hubbard (2001). Crisis on coral reefs linked to climate change.EOS 82(1): 1,5.

[3] J. P McWilliams., I.M. Côté, J.H. Gill., W.J. Sutherland, and A. R. Watkinson (2005). Accelerating impacts of temperature-induced coral bleaching in the caribbean. Ecology 86:2055–2060. http://dx.doi.org/10.1890/04-1657.

[4] M. Boumedine. 2008. Mining ICON/CREWS Data Sets for Discovering Relationships Between Environmental Factors and Coral Bleaching. 11th International Coral Reefs Symposium (2008), Fort Lauderdale, USA, unpublished.

[5] Böttcher M, Höppner F, Spiliopoulou M (2008) On exploiting the power of time in data mining. ACM SIGKDD Explorations 10(2):3–11.

[6] Duda, R.O.,Hart, P.E.and Stork, D.G. (2001) Pattern Classification, 2nd ed. John Wiley and Sons Ltd.

[7] NOAA Coral Reef Watch, updated twice-weekly. NOAA Coral Reef Watch Operational 50-km Satellite Coral Bleaching Degree Heating Weeks Product, Jan. 1, 2001-Dec.31, 2010. Silver Spring, Maryland, USA: NOAA Coral Reef Watch. Data set accessed 2014-04-15 at http://coralreefwatch.noaa.gov/satellite/hdf/index.php

[8] Liu, G., A.E. Strong, W.J. Skirving and L.F. Arzayus (2006). Overview of NOAA Coral Reef Watch Program's Near-Real-Time Satellite Global Coral Bleaching Monitoring Activities. Proceedings of the 10th International Coral Reef Symposium, Okinawa: 1783-1793.

[9] Keogh, E., Chakrabarti, K., Pazzani, M., & Mehrotra, S. (2001). Dimensionality reduction for fast similarity search in large time series databases. Knowledge and information Systems, 3(3),263-286.

[10] Lin, J., Keogh, E., Lonardi, S. & Chiu, B. (2003) A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.

[11] Lin, J. Keogh, E. Wei, L. and Lonardi, S. "Experiencing SAX: a novel symbolic representation of time series," Data Mining and Knowledge Discovery, vol. 15, pp. 107–144, 2007.

[12] R. Suzuki, and H. Shimodaira (2004) "An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters?", The Fifteenth International Conference on Genome Informatics 2004, P034

[13] Liao T. W, (2005) Clustering of time series data: a survey. Pattern Recognigtion 38(11):1857–1874

[14] Ben-Hur, A., A. Elisseeff, and I. Guyon (2002). A stability based method for discovering structure in clustered data. In Aetman, R.B. et al. (eds), Pacific Symposium on Biocomputing World Scientific Publishing Co., New Jersey.

[15] Dunn, J. C. (1974). Well separated clusters and fuzzy partitions. J.Cybernet.,4, 95–104.

[16] H. Fritz, L. A. Garcia-Escudero, A. Mayo-Iscar A. (2012). tclust: An R Package for a Trimming Approach to Cluster Analysis. Journal of Statistical Software, 47(12), 1-26.

[17] L. Kaufman and P. Rousseeuw (1990). Finding groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, Inc.

Interactive Data Quality Assistance – An Approach for Min(d)ing the Quality of Data

Nadia El Bekri, Elisabeth Peinsipp-Byma
Fraunhofer Institute of Optronics System Technologies and Image Exploitation (IOSB)
Karlsruhe, Germany

Abstract - *In this paper we introduce the concept of an interactive data quality system. Today one of the most challenging goals of data processing is to achieve and to maintain high data quality. Especially when the data is added manually by multiple users. The idea was originated out of the need to analyze the data set from a recognition assistance system. The system supports aerial image analysts in the task of the object recognition. Depending on which object features the aerial image analyst selects the solution set of the object types gets more precise. Especially in this field mechanisms that improve high data quality are important. Therefore we developed an interactive data quality system that helps experts generating potential rules through correlation that describe the whole data set. More precisely, we search for rules within the data set that are in general valid for a certain group of object types.*

Keywords: data quality, interactive data analysis

1 Introduction

The main idea was to develop a system that analyzes a given set of data with the help of an interactive data quality system and thereby derive rules that are valid for the entire data set. The underlying data set was taken from the WDI (World Development Indicators). This data set is a collection of development indicators from international resources. It presents the most current and accurate global development data available, and includes national, regional and global estimates. The database contains more than 900 indicators for over 210 countries [1]. For example the country "Switzerland" contains the indicators "GDP per capita" and many other indicators. In this case, the countries represent our objects and the indicators are the features of the objects that describe every country in a particular way. What we did was the first introductory step in a whole quality assurance process. The idea was originated out of the need to analyze the data set from a recognition assistance system. The system supports aerial image analysts in the task of object recognition by allowing them to describe single object features. Thereby the aerial image analyst can interactively classify the objects by selecting the visually extracted object features. The solution set contains only the amount of objects that match the selected features. In a previous step the objects are added manually by multiple users to the database or multiple databases are fused. Obviously, this is a critical issue. Why? First, the user can assign the features to the wrong objects or can forget to assign a potential feature to an object. Second, the feature values can be out of range for a certain group of objects. This means that probably for certain object types only a specific range is right. All this causes can lead to wrong or incomplete solution sets. The potential benefit for the user, in this case to recognize a specific object, is getting lost. We took the data from the WDI as underlying data set because the military data set is confidential and cannot be released for publication purposes. In addition it is irrelevant which data set is used as long there is an object-feature relation between them.

2 Quality assurance process

In order to be able to achieve and maintain a high quality data set we build up a quality assurance process. From analyzing the data set with algorithms, we receive certain rules that are probably generally valid for a certain group of objects containing the same object features and correlate very strongly. First, the data needs to be analyzed and rules have to be derived. The second step is then to apply the potential rules on the data set in order to prevent misentries in advance.

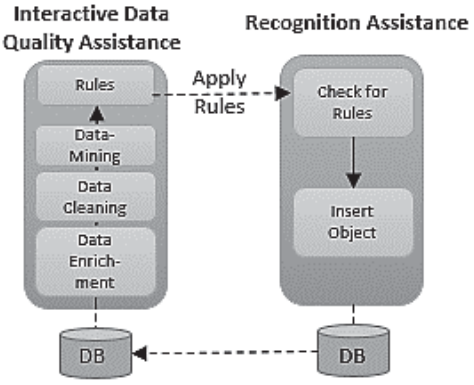


Fig. 1. Data quality assurance process

Figure 1 illustrates the general process of the data quality assurance. The data set we regard to analyze is the data set that already contains objects from the assistance system. The objects in this case are the countries, the features of the objects are the indicators that describe them. First, the data needs to be enriched. The data set of WDI e.g. did not contain the continents within the data point of a single country.

Therefore we enriched every country with the corresponding continent to be able to group them afterwards also by their continents. The Data Cleaning includes that empty data points are deleted from the data set that are considered into the data mining analysis. The next stage is to apply the data mining algorithm to find specific rules. The last step is to visualize the found rules for the analyst in a structured way. A strong correlation does not always imply causality and therefore the found rules within the data set need to be reviewed by the analyst before they can be applied to the whole data set as guideline.

2.1 Structure of the underlying data set

Figure 2 illustrates the structure of the underlying WDI data set in the interactive data quality system. In this case the countries are our objects symbolized with flags and the indicators listed below are the features of the objects.

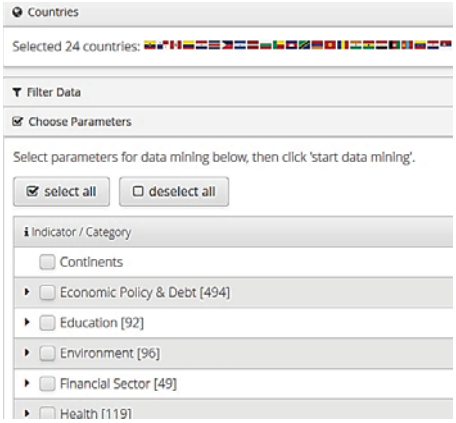


Fig. 2. Structure of the WDI data (Screenshot of the interactive data quality assistance system)

For example the country Colombia contains data points for the indicator "Education".

2.2 User interface

The user interface is split up into four different sections. The first section "Filter countries" offers the analyst the possibility to filter the data regarding a specific indicator. In case of the WDI data set it is possible to filter between specific ranges of numbers, e.g., within the "GDP per capita". This function is quite useful to find rules only within a specific filtered data set. The second section "Choose parameters for data mining" offers the analyst the possibility to only consider the relevant indicators for the data mining analysis. Although it is possible to consider all indicators for data mining. The section "Navigation and Results" provides the opportunity to navigate through the decision tree by selecting the provided intervals. In decision trees, the data is represented by a hierarchal tree, where each leaf refers to a concept [2]. Furthermore the section shows the diagrams of the selected rule and the corresponding countries. The intervals are generated automatically on the basis of information gain. Furthermore it displays all rules that are found within the data set. The section "Countries" illustrates which countries match with the selected rule symbolized by the national flags.

2.3 Applied algorithm

The underlying algorithm we used is k-means clustering with the application of the Lloyd algorithm. Clustering in general means partitioning a group of data points into different arrays. K-means clustering is a method to automatically partition a data set into k groups. The application of the Lloyd algorithm is efficient and resulted in the optimal solution for our specific problem to find rules that serve as guidelines within the data set. The algorithm has four major steps that need to be done in order to cluster each data point [3] [4]:

1. Initialize the centroids of the clusters
2. Search for every data point the closest cluster
3. Set the position of each cluster to the mean of all data points belonging to that cluster
4. Repeat the steps 2 and 3 until convergence

After performing the algorithm on the specified data set each object, in this case the countries, are assigned to a specific cluster. In order to navigate through the decision tree containing the different indicators, we split them automatically on the basis of the information gain. The information gain is predicated up on the reduced entropy after a data set is split up on an indicator. The major task during building up a decision tree is to find the attribute that returns the highest information gain. So in a first instance the entropy needs to be calculated. The second step is to split the data set on the different indicators. After this, the entropy for each branch is computed and added proportionally to get the entire entropy for the split. This entropy is then subtracted from the entropy before the split. The outcome is the information gain.

2.4 Example

In this section we want to illustrate an example rule generated by the interactive data quality assistance system. To control the variety of rules in advance we chose the parameters "GDP per capita" and „ Mortality rate, under -5 (per 1,000 live births) "for the data mining algorithm. Under five years mortality rate is the probability per 1000 that a newborn baby will die before reaching age five [1]. After starting the data mining algorithm, the highest information gain is found at „Mortality rate, under -5 (per 1,000 live births)". Figure 3 illustrates the separation for „Mortality rate, under -5 (per 1,000 live births)".

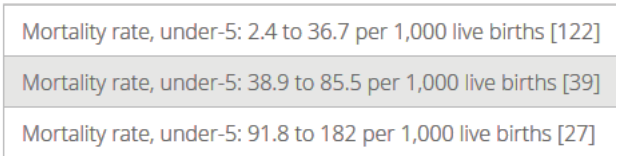


Fig. 3. Separation for „Mortality rate, under -5 (per 1,000 live births)".

In the next step we chose the interval with the highest mortality rate. After choosing this interval the next recommended separation is at "GDP per capita". Figure 4 illustrates the separation for "GDP per capita".

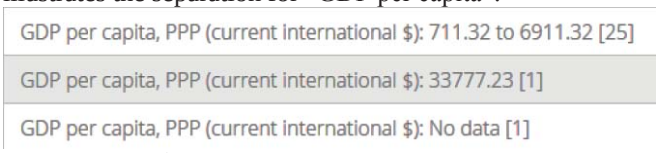


Fig. 4. Separation for „GDP per capita".

Countries with a high "Mortality rate, under -5 (per 1,000 live births)" seem to have a lower "GDP per capita". The derived rule is then:

'Mortality rate, under-5 (per 1,000 live births): 91.8 to 182.4' implies
'GDP per capita, PPP (current international $): 711.3 to 6911.3': 25/27 (92%) matching

Fig. 5. Derived rule from data mining algorithm.

Figure 5 illustrates for this derived rule a 92 per cent match for the countries within.

3 Conclusions

At the first stage we build up the interactive data quality assistance system with the underlying data set of the WDI. The system delivers rules that are supposedly general valid for the data. The next step will be to substitute the data set with the military data set and to apply the data mining algorithms on it. We need this step to be able to perform pilot studies with experts to improve the correctness of the derived rules and to compare different algorithms and the results. Furthermore this derived rules after being checked by the experts then will be applied as a guideline to the recognition assistance system while inserting a new object.

4 Acknowledgment

The underlying project to this article is funded by the WTD 81 of the German Federal Ministry of Defense. The authors are responsible for the content of this article.

5 References

[1] http://data.worldbank.org/data-catalog/world-development-indicators.

[2] Maimon, O., Rokach L. Data Mining and Knowledge Discovery Handbook. Springer Sciene + Business Media, 2010, pp. 284.

[3] Faber, V. Clustering and the continuous k-means algorithm. Los Alamos Science, 1994, pp. 140–142.

[4] MacQueen, J. B. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Symposium on Math, Statistics, and Probability, Berkeley, CA: University of California Press, 1967, pp. 281-297.

SESSION

SEGMENTATION, CLUSTERING, ASSOCIATION + WEB / TEXT / MULTIMEDIA MINING + SOFTWARE

Chair(s)

Drs. Robert Stahlbock
Peter Geczy
Gary M. Weiss

AHC based Word Clustering considering Feature Similarity

Taeho Jo
Department of Computer and Information Engineering, Inha University, Incheon, South Korea

Abstract— *In this research, we propose that the AHC (Agglomerative Hierarchical Clustering) algorithm should be used for clustering words, considering the feature similarities. Among the features, their dependencies and relations are available in the reality; texts which are features for encoding words into numerical vectors have own similarities with others. In this research, we define the similarity measure which considers both the features and the feature values, and use it for modifying the AHC algorithm as the approach to the word clustering. As the benefits from this research, we may obtain the potential possibility of more compact word representations and the more tolerance to the sparse distributions of numerical vectors. Therefore, the goal of this research is to implement the word clustering systems with the benefits.*

Keywords: Word Clustering, Feature Similarity

1. Introduction

The word clustering refers to the process of segmenting a group of words into subgroups of content based similar words. The group of words is encoded into their structured forms and a similarity measure between them is defined. The words are arranged into their closet clusters based on the similarity measure, as the clustering proceeds. The results from clustering the words are unnamed clusters and cluster naming and cluster prototype definition are regarded as other tasks in this research. The scope of this research is restricted to cluster words by their meanings.

Let us mention some challenges which this research tries to solve. The strong dependency among features exists especially in the text mining tasks, so the Bayesian networks which considers it was proposed as the approach, but it requires very much complicated analysis for using it [1]. If the independences among features are assumed, it requires many features for encoding words or texts into numerical vectors. Since each feature has very little coverage in the domain of text mining, we cannot avoid the sparse distribution of numerical vectors which represent words or texts[3]. Therefore, this research is intended to solve the problems by considering the feature similarity as well as the feature value one.

Let us mention what we propose in this research as its idea. In this research, we consider the both similarity measures, feature similarity and feature value similarity, for computing the similarity between numerical vectors. The AHC (Agglomerate Hierarchical Clustering) algorithm is modified into the version which accommodates the both similarity measures. The modified version was applied to the word clustering task. Therefore, the goal of this research is to improve the word clustering performance by solving the above problems.

Let us mention the benefits which we expect from this research. The consideration of both the feature similarity and the feature value similarity provides the way of reducing the dimensionality of numerical vectors, potentially. We discover semantic relations among words through this research for performing other text mining tasks. The improvement of discriminations among even sparse numerical vectors is caused by computing the similarity between numerical vectors using the two measures. Therefore, the goal of this research is to pursue the benefits for implementing the text clustering systems.

This article is organized into the four sections. In Section **??**, we survey the relevant previous works. In Section 3, we describe in detail what we propose in this research. In Section 4, we mention the remaining tasks for doing the further research.

2. Previous Works

Let us survey the previous cases of encoding texts into structured forms for using the machine learning algorithms to text mining tasks. The three main problems, huge dimensionality, sparse distribution, and poor transparency, have existed inherently in encoding them into numerical vectors. In previous works, various schemes of preprocessing texts have been proposed, in order to solve the problems. In this survey, we focus on the process of encoding texts into alternative structured forms to numerical vectors. In other words, this section is intended to explore previous works on solutions to the problems.

Let us mention the popularity of encoding texts into numerical vectors, and the proposal and the application of string kernels as the solution to the above problems. In 2002, Sebastiani presented the numerical vectors are the standard representations of texts in applying the machine learning algorithms to the text classifications [4]. In 2002, Lodhi et al. proposed the string kernel as a kernel function of raw texts in using the SVM (Support Vector Machine) to the text classification [5]. In 2004, Lesile et al. used the version of SVM which proposed by Lodhi et al. to the protein classification [6]. In 2004, Kate and Mooney used also the SVM version for classifying sentences by their meanings [7].

It was proposed that texts are encoded into tables instead of numerical vectors, as the solutions to the above problems. In 2008, Jo and Cho proposed the table matching algorithm as the approach to text classification [8]. In 2008, Jo applied also his proposed approach to the text clustering, as well as the text

categorization [12]. In 2011, Jo described as the technique of automatic text classification in his patent document [10]. In 2015, Jo improved the table matching algorithm into its more stable version [11].

Previously, it was proposed that texts should be encoded into string vectors as other structured forms. In 2008, Jo modified the k means algorithm into the version which processes string vectors as the approach to the text clustering[12]. In 2010, Jo modified the two supervised learning algorithms, the KNN and the SVM, into the version as the improved approaches to the text classification [13]. In 2010, Jo proposed the unsupervised neural networks, called Neural Text Self Organizer, which receives the string vector as its input data [14]. In 2010, Jo applied the supervised neural networks, called Neural Text Categorizer, which gets a string vector as its input, as the approach to the text classification [15].

The above previous works proposed the string kernel as the kernel function of raw texts in the SVM, and tables and string vectors as representations of texts, in order to solve the problems. Because the string kernel takes very much computation time for computing their values, it was used for processing short strings or sentences rather than texts. In the previous works on encoding texts into tables, only table matching algorithm was proposed; there is no attempt to modify the machine algorithms into their table based version. In the previous works on encoding texts into string vectors, only frequency was considered for defining features of string vectors. Texts which are used as features of numerical vectors which represent words have their semantic similarities among them, so the similarities will be used for processing sparse numerical vectors, in this research.

3. Proposed Approach

This section is concerned with modifying the AHC (Agglomerative Hierarchical Clustering) algorithm into the version which considers the similarities among features as well as feature values, and it consists of the three sections. In Section 3.1, we describe the process of encoding words into numerical vectors. In Section 3.2, we do formally the proposed scheme of computing the similarity between two numerical vectors. In Section 3.3, we mention the proposed version of AHC algorithm which considers the similarity among features as the approach to word clustering. Therefore, this article is intended to describe in detail the modified version of KNN algorithm and its application to the word clustering.

3.1 Word Encoding

This subsection is concerned with the process of encoding words into numerical vectors. Previously, texts each of which is consists of paragraphs were encoded into numerical vectors whose attributes are words. In this research, we attempt to encode words into numerical vectors whose attributes are text identifiers which include them. Encoding of words and texts into numerical vectors looks reverse to each other. In this Section, we describe in detail the process of mapping words into numerical vectors, instead of texts.

In the first step of word encoding, a word-document matrix is constructed automatically from a text collection called corpus. In the corpus, each text is indexed into a list of words. For each word, we compute and assign its weight which is called TF-IDF (Term Frequency-Inverse Document Frequency) weight [2], by equation (1),

$$w_i = TF_i(\log_2 N - \log_2 DF_i + 1) \tag{1}$$

where TF_i is the total frequency in the given text, DF_i is the total number of documents including the word, and N is the total number of documents in the corpus. The word-document matrix consists of TF-IDF weights as relations between a word and a document computed by equation (1). Note that the matrix is a very huge one which consists at least of several thousands of words and documents.

Let us consider the criterion of selecting text identifiers as features, given labeled sampled words and a text collection. We may set a portion of each text in the given sample words as a criteria for selecting features. We may use the total frequency of the sample words in each text as a selection criterion. However, in this research, we decided the total TF-IDF (Term Frequency and Inverse Document Frequency) which is computed by equation (1) as the criterion. We may combine more than two criteria with each other for selecting features.

Once some texts are selected as attributes, we need to consider the schemes of defining a value to each attribute. To each attribute, we may assign a binary value indicating whether the word present in the text which is given as the attribute, or not. We may use the relative frequency of the word in each text which is an attribute as a feature value. The weight of word to each attribute which is computed by equation (1) may be used as a feature value. Therefore, the attributes values of a numerical vector which represent a word are relationships between the word and the texts which are selected as features.

The feature selection and the feature value assignment for encoding words into numerical vectors depend strongly on the given corpus. When changing the corpus, different texts are selected by different values of the selection criterion as features. Even if same features are selected, different feature values are assigned. Only addition or deletion of texts in the given corpus may influence on the feature selection and the assignment of feature values. In order to avoid the dependency, we may consider the word net or the dictionary as alternatives to the corpus.

3.2 Feature Similarity

This subsection is concerned with the scheme of computing the similarity between numerical vectors as illustrated in Figure 1. In this research, we call the traditional similarity measures such as cosine similarity and Euclidean distance feature value similarities where consider only feature values for computing it. In this research, we consider the feature similarity as well as the feature value similarity for computing

it as the similarity measure which is specialized for text mining tasks. The numerical vectors which represent texts or words tend to be strongly sparse; only feature value similarity becomes easily fragile to the tendency. Therefore, in this subsection, as the solution to the problem, we describe the proposed scheme of computing the similarity between numerical vectors.

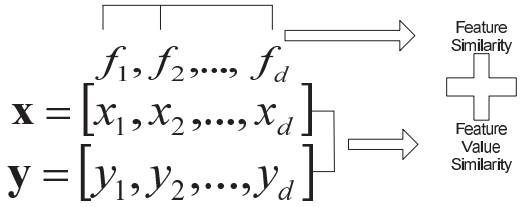


Fig. 1
THE COMBINATION OF FEATURE AND FEATURE VALUE SIMILARITY

Text identifiers are given as features for encoding words into numerical vectors. Texts are dependent on others rather than independent ones which are assumed in the traditional classifiers, especially in Naive Bayes [1]. Previously, various schemes of computing the semantic similarity between texts were developed [2]. We need to assign nonzero similarity between two numerical vectors where non-zero elements are given to different features with their high similarity. It is expected to improve the discriminations among sparse vectors by considering the similarity among features.

We may build the similarity matrix among features automatically from a corpus. From the corpus, we extract easily a list of text identifiers. We compute the similarity between two texts by equation (2),

$$s_{ij} = sim(d_i, d_j) = \frac{2 \times tf(d_i, d_j)}{tf(d_i) + tf(d_j)} \tag{2}$$

where $tf(d_i, d_j)$ is the number of words which are shared by both texts, d_i and d_j, and $tf(d_i)$ is the number of words which are included in the text, d_i. We build the similarity matrix which is consists of similarities between text identifiers given as features as follows:

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1d} \\ s_{21} & s_{22} & \dots & s_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ s_{d1} & s_{d2} & \dots & s_{dd} \end{pmatrix}.$$

The rows and columns in the above matrix,S, correspond to the d text identifiers which are selected as the features.

The texts, $d_1, d_2, ..., d_d$ are given as the features, and the two words, t_1 and t_2 are encoded into the two numerical vectors as follows:

$$t_1 = [w_{11}, w_{12}, ..., w_{1d}]$$
$$t_2 = [w_{21}, w_{22}, ..., w_{2d}].$$

The features,$d_1, d_2, ..., d_d$ are defined through the process which was described in Section 3.1. We construct the d by d matrix as the similarity matrix of features by the process mentioned above. The similarity between the two vectors are computed with the assumption of availability of the feature similarities, by equation (3),

$$sim(t_1, t_2) = \frac{\sum_{i=1}^{d} \sum_{j=1}^{d} s_{ij} w_{1i} w_{2j}}{d \cdot \|t_1\| \cdot \|t_2\|} \tag{3}$$

where $\|t_1\| = \sqrt{\sum_{i=1}^{d} w_{1i}^2}$ and $\|t_2\| = \sqrt{\sum_{i=1}^{d} w_{2i}^2}$. We get the value of s_{ij} by equation (2).

The proposed scheme of computing the similarity by equation (3) has the higher complexity as payment for obtaining the more discrimination among sparse vectors. Let us assume that two d dimensional numerical vectors are given as the input for computing the similarity between them. It takes only linear complexity, $O(d)$, to compute the cosine similarity as the traditional one. However, in the proposed scheme takes the quadratic complexity, $O(d^2)$. We may reduce the complexity by computing similarities of some pairs of features, instead of all.

3.3 Proposed Version of AHC Algorithm

This section is concerned with the modified version of AHC algorithm which considers both the feature similarity and the feature value one. The words which are given as clustering targets are encoded into numerical vectors whose features are texts by the scheme which was described in section **??**. The numerical vectors which represent words or texts tend to be sparse, inherently; zero values in each vector tend to be dominant over 90%. In the proposed version, the similarities among the numerical vectors are computed by equation (3). In order to provide the detail explanation, we describe the proposed AHC version, together with the traditional one.

The traditional version of AHC algorithm is illustrated in Figure 2. Words are encoded into numerical vectors, and it begins with unit clusters each of which has only single item. The similarity of every pairs of clusters is computed using the Euclidean distance or the cosine similarity, and the pair with its maximum similarity is merged into a cluster. The clustering by the ACH algorithm proceeds by merging cluster pairs and decrementing number of clusters by one. If the similarities among the sparse numerical vectors are computed, the traditional version becomes very fragile from the poor discriminations among them.

The proposed AHC version is illustrated in Figure 3. Words are encoded into numerical vectors, and the clustering begins with individual items. The similarities among numerical vectors are computed by equation (3) which was presented in section 3.2. Clustering proceeds by merging the pair with its maximum similarity. By replacing the Euclidean distance or the cosine similarity by equation (3), it is expected to improve the discriminations among even sparse numerical vectors.

We may consider several schemes of computing a similarity between clusters. We may compute similarities of all possible pairs of items between two clusters and average over them as the cluster similarity. The maximum or the minimum among

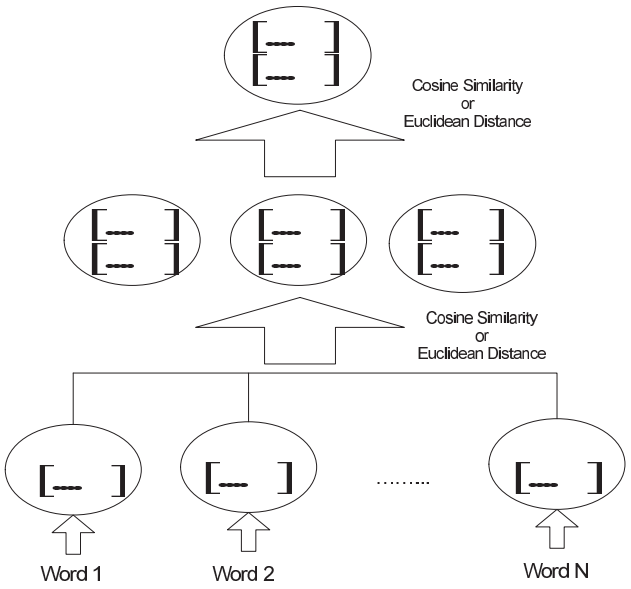


Fig. 2
THE TRADITIONAL VERSION OF AHC ALGORITHM

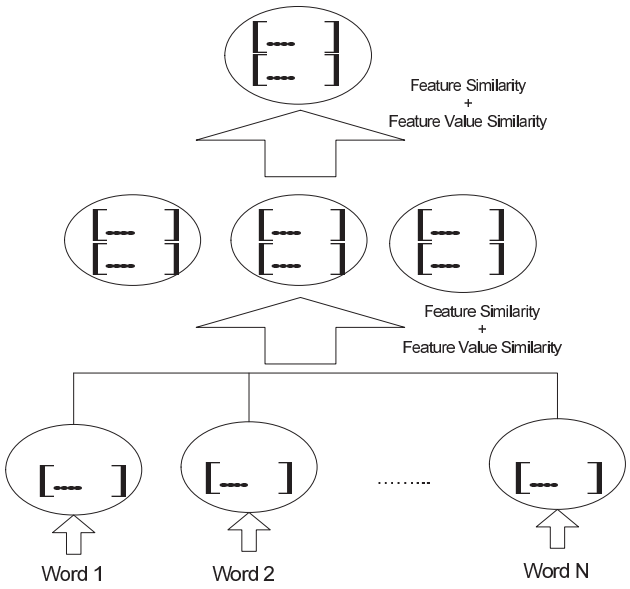


Fig. 3
THE PROPOSED VERSION OF AHC ALGORITHM

similarities of all possible pairs is set as the cluster similarity. In another scheme, we may select representative members of two clusters and the similarity between the selected members is regarded as the cluster similarity. In this research, we adopt the first scheme for computing the similarity between two clusters in using the AHC algorithm; other schemes will be considered in next research.

Let us compare the both AHC versions with each other. Both versions begin the clustering process with individual numerical vectors. In computing the similarity between two numerical vectors, the traditional version uses the Euclidean distance or cosine similarity mainly, whereas the proposed one uses the equation (3). Like the traditional version, the pair of clusters with its maximum similarity is merged into a single cluster in doing the clustering process. However, the proposed version is more tolerant to sparse numerical vectors in computing the similarities among them than the traditional version.

4. Conclusion

Let us mention the remaining tasks for doing the further research. We need to validate the proposed approach in specific domains such as medicine, engineering, and economics, as well as in generic domains such as ones of news articles. We may consider the computation of similarities among some main features rather than among all features for reducing the computation time. We try to modify other machine learning algorithms such as Naive Bayes, Perceptrons, and SVM (Support Vector Machine) based on both kinds of similarities. By adopting the proposed approach, we may implement the word clustering system as a real program.

References

[1] T. Mitchell, *Machine Learning*, McGraw-Hill, 1997.
[2] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search*, Addison-Wesley, 2011.
[3] T. Jo, "The Implementation of Dynamic Document Organization using Text Categorization and Text Clustering" PhD Dissertation, University of Ottawa, Ottawa, Canada, 2006.
[4] F. Sebastiani, "Machine Learning in Automated Text Categorization", *ACM Computing Survey*, Vol. 34, pp. 1-47, 2002.
[5] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification with String Kernels", *Journal of Machine Learning Research*, Vol. 2, pp. 419-444, 2002.
[6] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, "Mismatch String Kernels for Discriminative Protein Classification", *Bioinformatics*, Vol. 20, pp. 467-476, 2004.
[7] R. J. Kate and R. J. Mooney, "Using String Kernels for Learning Semantic Parsers", in *Proc. ICCL '06*, 2006, pp. 913-920.
[8] T. Jo and D. Cho, "Index based Approach for Text Categorization", *International Journal of Mathematics and Computers in Simulation*, Vol. 2, 2008, pp. 127-132.
[9] T. Jo, "Single Pass Algorithm for Text Clustering by Encoding Documents into Tables", *Journal of Korea Multimedia Society*, Vol. 11, 2008, pp. 1749-1757.
[10] T. Jo, "Device and Method for Categorizing Electronic Document Automatically", South Korean Patent 10-1071495, 2011.
[11] T. Jo, "Normalized Table Matching Algorithm as Approach to Text Categorization", *Soft Computing*, Vol. 19, 2015, pp. 849-849.
[12] T. Jo, "Inverted Index based Modified Version of K-Means Algorithm for Text Clustering", *Journal of Information Processing Systems*, Vol. 4, 2008, pp. 67-76.
[13] T. Jo, "Representationof Texts into String Vectors for Text Categorization", *Journal of Computing Science and Engineering*, Vol. 4, 2010, pp. 110-127.
[14] T. Jo, "NTSO (Neural Text Self Organizer): A New Neural Network for Text Clustering", *Journal of Network Technology*, Vol. 1, 2010, pp. 31-43.
[15] T. Jo, "NTC (Neural Text Categorizer): Neural Network for Text Categorization", *International Journal of Information Studies*, Vol 2, 2010, pp. 83-96.

Extraction of Relevant Entities in Textual Documents. Modeling Intelligence Maps

Isnard Thomas Martins[1] **and Edgard T. Martins**[2]
[1]Administração, Universidade Estácio de Sá, Rio de Janeiro, R.J, Brazil; 55 21 98748873
[2]Coordenação Ergonomia, UFPE, Recife, Pernambuco, Brazil ; 5581 41413131

Abstract - *Police investigation activities are conducted on historical, records and occurrences reports in informations data bases structured and not structured, where are extracted knowledge to elucidate authorship, interests, crime dynamics and objects involved in criminal activities. The complexity inherent in not structured informations sources and police records, the restrictions associated time and resources available to authorship analysis assume a critical condition in the elucidation of the crimes. As a result, the automatic extraction of knowledge in criminal databases assume great importance in generating intelligence maps and research activities. Criminal reports, source of the research and criminal knowledge, usually present themselves in not structured format, inaccurate in content and difficult to analyze*

Keywords: Datamining, Shortest-path Algorithms, Law Enforcement Analysys

1 INTRODUCTION

In a criminal report are transcribed all the facts, people, circumstances and relationships that characterized the police report. Because the need to capture the occurrence with maximum accuracy of reality, investigative studies devoted to authorship analysis are based on historical police, whose transcription, aims to provide freedom for communication through narratives, often developed by people of different features, both comunicante and witnesses, as police responsible for the occurrence record

The discovery of traces who carrying the criminal to the crime scene or their activities to achieve the offense becomes one of the fundamental problems encountered in the authorship's investigations of criminal analisys [1].

The police reports are dense, complex and characterized by scattered data. The not structured police report, done on free text style, offers better support for research evidence and relationships, imposing, however complexity, time and personnel allocated in analysis.

Gradually criminal activities are growing in sophistication, technology and planning. Employing resources and technologically advanced methods, criminals are connecting in social networks and using modern communication systems such as Internet, wifi, telephone and radio. Crimes fraud related, drug trafficking, money laundering and gangs's investigation, an activity of complex and sophisticated police intelligence [1].

Analysts and researchers produce criminal knowledge from information sought as a bulky bases of police reports [2] . The biggest challenge for analysts, researchers and police intelligence departments is to achieve efficiency and accuracy in the face of growing demand for raw data found in criminal intelligence bases [3] . The potential demand of police reports without authorship identified in countries with high volumes of criminal incidents, as verified in Brazil, is the characterization of the impunity.

Xu and Chen [4] reported that the successful construction of concept maps, based on documents and police reports analyzed in the criminal investigation, depends on extensive use of techniques to automate the most of the data mining operations and identification of useful entities such as people, places, events, organizations or objects involved in the researched historical, whose identification will contribute to clarifying the facts and understanding of the relationships investigated. However, the efficiency to be obtained in the extraction of useful entities for analysis depends essentially on the cleaning of the data entered in the extraction process

Oriented systems to support the extraction of relevant entities and intelligence activities in unstructured documents are intended to increase the speed in the analysis and reduction of researchers's time in the preparation activities of relational maps extracted from criminal reports.

Expert systems, however, require many hours in the preparation of information for procedural support, such as dictionaries, references and linguistic rules [5].

According Baluja et al [5], specialized systems in data mining are specific in their goals, such as tax evasion, money laundering, census or commercial research, those systems require many hours in the preparation of information for procedural support, and its operation have restricted rules. Dictionaries are originally built in language English, Japanese, Portuguese, etc, for this reason, they are restricted, starting to understand only a particular branch of knowledge [6] .It would not be possible, for example, migrate directly rules and special specifications used by an extractor system of criminal organizations operated in the United States for the extraction of entities in criminal bulletins in Brazil [5]. An extensive list of previous procedures for preparing the extractor systems should be deployed, to meet the restrictions and formalities language as a basis for nominal entities extraction in any language. Some authors mention the use of automated tools

to accelerate the preparation of these lists in support of expert systems.

No matter the method used, the expert system development will require a training method in order to obtain the best efficiency in the analysis and recovery of entities that can meet the principles and necessary rules for entities extraction in textual documents.

2 NETWORK RELATIONSHIPS AND CLASSIFICATION OF ENTITIES

Actors, objects, events and relationships captured in the police reports can provide valuable historical evidence and provide crime patterns, usually hidden in the reports and police reports. A synthesis drawing format and simulated scenarios is known in police and criminal context map, research map or intelligence map, topological representations of the crime scene [1]. The represented social network in the intelligence maps must integrate all parts of research activities and identify possible connections between actors and potentially involved events [4]. The map or criminal relationship tree is often treated as a network [7], provides valuable evidence of extracted crime patterns in investigations, resulting in accumulated knowledge in the analysis of relationships between entities involved in the criminal offense.

The intelligence maps allow inter-relate several useful entities extracted from texts treated, establishing an association value between these various entities. These associations are of great relevance to the criminal investigation. The resulting structure aims to provide aid for research and pattern recognition in criminal offenses [8]. Systems for intelligence analysis are applied for tracking of individuals and organizations involved in criminal activities such trafficking, terrorism and fraud. [9]. The collect of entities is based on previous patterns, becoming simplified for not requiring the understanding of the text by the system entities extraction operator [10]. Extractors Systems are also employed to identify patterns such as dates, times, numeric expressions and email addresses.

The arc value associated between entities in an intelligence map expresses the intensity, on which the entities are closer or distant from each other. The value assigned to the mapped relationships helps the visibility of existing links, identifies involvement of the actors present in the scene and produces knowledge to generate conclusions and reports on the facts of the cases analyzed [1].

Various distance types are used to calculate specific measures of distance between entities in the vector space. Some measures are used applying simple Euclidean distance while others are used applying the square Euclidean distance or absolute Euclidean distance, where the distance is the sum of the square of distances, avoiding the square root calculation, which offers advantages for computational speed in applied calculations [6].

According to textual entities standardization procedures developed in the MUC-7, Seventh Message Understanding Conference and Second Multilingual Entity Task [11], nominal entities are defined as proper names, numbers, people, local references, schedules dates, percentages and monetary values . The scenario selected for extraction site is built according with the events in which the entities are participating, whose definition of domain and importance depend on the purpose of the analysis and presence of the entity on the analyzed text. Entities assist the police investigation and provide the necessary allowance for identifying patterns related to the "modus operandi" of the crime [2].

For each extracted entity must be associated attributes residing in a specific parameter table representing the properties and characteristics. This table is called Elements Table (TE), whose purpose is to qualify the identification of each entity, beyond the simple name reference.

Selecting the domain of entities and structure of the model elements table depends on the size of each entity in the specific scenario in which it is inserted. The MUC-7 provides that such definitions depend subjectively system of the author, however its accuracy is linked to the wealth of the parameters associated with the entities, serving to increase its effectiveness with the users [11].

Chen and Lynch [12] cite knowledge bases specialized on automatic creation of thematic dictionaries and algorithms for generating statistical coefficients related to frequency ratios between concepts extracted from text documents . Furthermore, the available literature provides in various academic segments developed studies in both fields, information science and cognitive studies, confirming the creation of specific areas for scientific dictionaries, such as medicine, engineering and business that resulted in the creation of efficient thesaurus, robust potentially available as basis for information retrieval applications [13] . Chen and Lynch [12] cite the specific steps for the implementation of preparatory processing for recovery of useful entities :

- Development of the list of objects and documents
- Filter the objects
- Indexing
- Analysis of co -occurrence (frequency studies)
- Recovery of associations

The crime often involves organized gangs, whose members are connected by various associations such as common interests, friendship, neighborhood or criminal association . This relationship, similarly can be treated as a network in which such criminals can perform various activities and illegal actions . Textual documents such as police reports and others are rich in information, from which you can extract entities, converting them into a topological representation connected by their criminal relationship and their criminal activities. The base of knowledge, represented by a semantic network in which nodes are words, phrases and concepts and connections represent the semantic relationship between nodes [12]. The system for capturing concepts consists of rules or procedures operated, on according the knowledge base, similar to the decisions rules from experts patterns

3- MODELING INTELLIGENCE MAPS

Martins [1] presented an expert system to capture entities from free texts and intelligence maps modelling, called Anaphora that apply as example for illustration of an automatic extractor system model. Used by some intelligence agencies in Brazil, the Anaphora system integrates the major phases of an extractor project: construction of thematic dictionary, training the useful entities and network construction. The final output provides network representation on a Graph format that examines the strongest connections between nodes of entities network, using one shortest path algorithm.

The first phase of Anaphora system involves the construction of a specialized dictionary in radicals, using policial languagem, which will serve as keywords for later extraction of knowledge.

Cognates are words derived from the same root. Is the irreducible element common to all the words of the same family [14], also called lexical family [15]. The element is irreducible when it can no longer be reduced.

Some examples of families who have the same root:

- Moon, moonlit, moonligth
- Sea, salt, sailor
- Crime, criminal, criminology
- Love, loving
- friendly, friend, friendship

Monteiro [14] mentions that the internal structure of word consists of words with associated elements which represent the minimum elements of language emissions containing individual significance. The radical in its original form is the root, the minimum element of a family of words and irreducible and common element of this family of words. The root is the element from where the first morphological operation, so their root shall be different from the radical. The radicals may have one or more affixes derivative. Thus the same word can have several radicals. The neighbor word can offer three degrees radical :

I. neig
II . neighbor
III. neighborhood

The meaning is essential in the root concept that carries the semantic word load . The suffixe particularizes the generic meaning of the root (smaller part of the word) in a series of derivatives. The more affixes (words derived by prefixes and suffixes), less general will be the meaning of the word. The roots are minimal morphological construction of a core, which may be free or attached [15]. The high degree of radical will include all derived words. We can conclude the following assertions, mutually inverse :

- A higher volume of derived words requires a higher degree of extracted radicals, as close as possible to its original form, that is, the smallest format .
- It is important to keep the meaning of the radical in its minimal primitive form, keeping the association of meaning with the derived word [14], avoiding multiple interpretations or derivations with the family of other extracted derivatives [16].

The specialist pre-processing dictionary is based on the principle of extracting radicals, derived from training sets. The extracted lists are then used as keywords, in an interactive way for obtaining derived words. The resulting structure is refined, obtaining also stop words, which are words with little meaning in the analyzed text. Dictionaries are generated from classified information, contained in documents belonging to the application domain, which are converted into specialized structures through continuous training [12]. Research in dictionaries is an important source for retrieval information systems [16]. Dictionaries include selected information in documents, databases or manually, generated by experts who provide guidance for extraction algorithms such as keywords and critical debugging routines. The resulting structures, prepared in automatic form or manually allow extracting keywords from textual documents with little or no manual interference [1].

Automatic dictionaries or semi- automatic dictionaries can be generated from processed radicals, by algorithms that serve as keywords for knowledge extraction. The initial structures are subsequently processed through specialized training performed by learning machine [18]. Dictionaries generated manually can be obtained by combining public domain words (geographic information, professions, usual acronyms, common names, titles etc). The resulting structures, prepared in automatic form or manually allow extracting keywords, from textual documents with little or no manual interference [1].

The construction of specialist dictionary is based on linguistic studies, which provides the basic guidelines for the learning algorithm of extraction model. An ordered set of phonemes is considered a word when has a meaning . The words include the names (nouns, adjectives and adverbs and verbs [14].

Table 1 shows an example of extracted key from historical words that belong to the domain of a collection of documents that were investigated.

TABLE 1.
Example of extracted keywords from police texts in portuguese language [29]

KEYWORDS	FREQUENCY
dp	56
vulgo	47
inq	44
favela	36
traficante	28
policia	25
dinamica	22
dre	22
mandado	21
traficantes	20
cv	19
prisão	16
comando vermelho	15
preso	15
policiais	15

The second stage of the extractor system involves the extraction of useful entities, modeling of relationships and calculations of co-occurrences between the extracted entities. We used the Hauck algorithm [19] adapted from the method developed by Chen and Lynch [12], an algorithm for treatment of co-occurrences on data mining routines. The algorithm sets relative levels of importance between the extracted entities on researched documents, calculating weights for relationships between each pair of extracted entity. The weights are calculated based on statistics frequencies corresponding to a value for the co- related associations. The Hauck algorithm calculates the relative weight of each entity, on each document of collection.

Originally, the co-occurrences analysis approach was devoted to the automatic generation of dictionaries based on textual documents, reflecting the frequency with which two sentences appeared together in the same document . The modern statistical approach defines co-occurrence as the frequency among entities, based on lexical statistics. Assuming that two entities appear together in a same document, there may be an association and involvement between these entities . A co-occurrence with non-zero value indicates the weight of the rapprochement between entities, so strongly associated so higher be the value represented by their co -occurrence [20] .

Statistics co-occurrence are related to the useful words found in the analyzed text. The co-occurrence concept is based on the proposition of Chen and Lynch [12] for calculating the statistics co-occurrences between extracted words on text documents. Xu and Chen [4] define co-occurrence or associative relationship as the relationship between a pair of entities, when they are found together on one document

Step 1.1

The Hauck algorithm [2] calculates the relative weight of each entity, in each document of the collection (D_{ij}; entity - document).

Equation [1] shows the calculation of the co-occurrence D_{ij} in each document of the collection, given by :

$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j} \times w_j\right) \qquad (1)$$

Where:

i - represents each document of the collection

j - represents each entity found on document i

N - number of collection's documents

Df_j - Number of documents in which j is present

Tf_{ij}. number of occurrences of J entity in each document in which j entity was located

W_j - factor of importance of j entity on extraction process (relative value that can assume, greater or lesser degree, according to importance of entity on extraction process)

Step 1.2

The algorithm calculates the co-occurrence between each pair of entities found together in documents in the collection (W_{jk} and W_{kj}), using a asymmetric function, shown in (2) and (3)

$$W_{jk} = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \times WeightingFactor(k) \qquad (2)$$

$$W_{kj} = \frac{\sum_{i=1}^{n} d_{ikj}}{\sum_{i=1}^{n} d_{ik}} \times WeightingFactor(j) \qquad (3)$$

Where:

j - represents the first entity of each examined pair in document i

k - represents the seond entity of each examined pair in document i

W_{ij} - represents the final weight among entity j and k entity

W_{kj} - represents the final weight calculated between the entity k and entity j

d_{ij} - weight of the entity j, calculated as shown in step 2 of this topic

Df_{jk} - represents the number of documents in the collection N, where the entities j and k are revealed together.

D_{ijk} - Hauck algorithm calculates the combined weight of each pair of entities found together in each document in the collection.

Equation (4) shows the calculation of the combined weight of pair jk on document i and (5) shows the calculation of the combined weight of the kj on document i. The difference between these functions is the factor of relative importance (Wi / Wi) in the calculation of the function

$$d_{ijk} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_j\right) \qquad (4)$$

$$d_{ikj} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_k\right) \qquad (5)$$

Where:

WeightingFactor j e WeightingFactor j - Influence factor that reduces the value of the very common generic instances, reducing the value of their respective influences. WeightingFactor is obtained through the calculation shown in (6) and (7):

$$WeightingFactor(j) = \frac{\log\frac{N}{df_j}}{\log N} \qquad (6)$$

$$WeightingFactor(k) = \frac{\log \frac{N}{df_k}}{\log N} \qquad (7)$$

Where:

WeightingFactor $_k$ - reduction factor for the entity k (6)

WeightingFactor $_j$ - reduction factor for the entity k j (7)

The algorithm proposed by Hauck et al. [2] produces asymmetric values for associations between entities , however penalizes with a final reduction factor the value of words most often found in the studied texts . This reduction factor is used in order to minimize the importance of extracted generic terms.

Figure 1 shows an example of useful entities extracted from the analyzed domain. The columns present the results of calculations processed by each stage of the frequency and approximation algorithm.

	Total TF i j	DF j	D i j	W j	WeightFac j
duque de caxias	22	17	60,687	0,9	5,924
comando vermel	17	7	61,979	0,9	4,779
centro	11	8	38,635	0,9	4,477
eder gonca	10	5	39,823	0,9	3,912
rio de janeiro	10	9	33,945	0,9	4,500
mangueira	10	6	38,000	0,9	4,094
ramos	9	7	32,812	0,9	4,143
naldo medeiro	9	3	40,438	0,9	3,296
acari	9	2	44,087	0,9	2,890
campo grande	8	6	30,400	0,9	3,871
rocinha	8	6	30,400	0,9	3,871
luiz costa	8	4	33,644	0,9	3,466
madureira	8	5	31,858	0,9	3,689
nova iguaçu	7	5	27,876	0,9	3,555

Fig 1. Função WeightingFactor [29]

In the third phase of construction, the entities are extracted from textual documents and organized in a structure indexed, by document , keeping data available for access of the algorithm . Each pair of extracted entities is analyzed, according to frequency computed in each document , subsequently consolidated in accordance with the totals processed throughout collection. The product obtained by this method comprises a weighted array of relationships where each array element represents a entity and the extracted weights computed represents the importance of these relationships . The depiction of a network in matrix format provides a means to describe a graph , eliminating the existence of a list of nodes and arcs to build or a representative drawing of a network [21] .

Let N be a weighted matrix with m rows and n columns, corresponding to each of the extracted entities (vertices) . Let n_{ij} the representation of the element in the ith row and column jth [21] . Each element n_{ij} of the array corresponds to an arc (i, j) and refers to a association value between entities i and j if these entities are present in the extracted relationship. The resulting structure is so called matriz of Criminal Relationships . As a result of extraction, the Anaphora system produces the following results:

Step 2.1

Construction of a temporary structure containing the totalization of frequencies and the temporary variables, such as strengthening's factor for each pair of entity extracted

Step 2.2

Construction of a temporary array containing consolidated frequencies for each pair of extracted entities calculated by the co-occurrences algorithm. The raw results processed by the algorithm will be in accordance with the frequencies computed, between each pair of entities.

Step 2.3

Constructing a normalized final results matrix containing co-occurrences for each pair of entities.
The resulting structure of the normalized matrix corresponds to a directed graph , whose vertices are represented by nominal extracted entities and their arcs are represented by the results of the entity - entity
The structure is stored in an auxiliary file (setting file) generated for further analysis , completing the cycle developed by Anaphora System [29].

The file for analysis consists of three types of information, corresponding to each pair of entities associates:

- Numeric code of connected vertices;
- Association value, calculated by Anaphora system [29];
- Reference name of connected entities.

4 - ANALYSIS OF THE STRONGEST LINKS BETWEEN ENTITIES IN THE INTELLIGENCE MAP

From associations's matrix, is then constructed a second array that will contain the reverse tracking of possible paths between pairs of entities present on the graph . This structure called Reach matrix is based on the reverse access tracking of Dijkstra algorithm, optimizes the use of the intelligence's Map because provides the pre-calculation of all possible paths between related entities , thus avoiding the time spent processing the strongest associations in research activity [1] .

Each cell of the Reach Matrix represents an entity of reference identified by column number where it is located, indicating the associations of the reverse path between pairs of entities line / column of the matrix .

The Dijkstra algorithm [22] is the classic method for minimum cost of path calculation from a source node to all other nodes of a weighted graph [4] , assuming that the graph contain no negative arcs [23] . Dijkstra lent his studies for the more efficient algorithms and solutions to shortest path, the

principles of which were based on the original structures of Dijkstra algorithm. Xu and Chen [4] mention that in a criminal network represented by a directed graph, the value of a connection, which can assume a number between zero and one, can be treated as an probability measure for approximation calculation between two directly connected entities. As a general rule, the joint probability of occurrence of a group of mutually independent events is equal to the product of the individual probabilities of occurrence of these same events. If two nodes in a graph are only connected through a sequence of intermediate connections , the association value between the two nodes is equal to the product of intermediate weights. The strongest association between a pair of nodes is represented by the largest product of the weights between the nodes.

Since the shortest path algorithms recognizes the shorter distances between graph nodes, where the value of the arcs indicates the weight of the associations , the representation of the strongest connections, after application of the shortest path algorithm will not guarantee that the strongest associations will be identified [1] .

Xu and Chen [4] proposes a heuristic search for transformation by the shortest path to the location of the strongest connections in a directed graph , using the logarithmic transformation : $l = -\ln(w) \quad 0 < w \leq 1$

Where:

l is the weight of the connection in the new transformed graph

w is the corresponding weight in the original graph

With the proposed transformation are obtained the following axioms [4] :

- All the connections in the transformed graph are non negative numbers.:
- Since: $0 < w \leq l$, thus $\ln(w) \leq 0$, wich sugest that: $-\ln(w) \geq 0$;
- The lowest values of the arches in the transformed graph correspond to higher values in the original graph
- If $l1 < l2$, then $-\ln(w1) < -\ln(w2)$ or $\ln(w1) > \ln(w2)$.
- Since ln (w) is a monotonic increasing, is follows that $w1 > w2$;
- The shortest paths using the sum of the weights values of the transformed graph, correspond to larger Arches products using the original network.

After the modeling of a associations matrix , represented by a directed graph and constructed as a relationship between the product extracted entities from text files , the stronger links between the graph's entities are calculated. The associated weights with arcs provide calculation probabilities between each pair of entities , with the possible path and the representative value of the strongest chances of approximation between entities [1].

The calculated results are presented in a matrix of associations, as shown in Figure 2.

	luiz fernando d	comando verm	marcos marinh	marcos antoni
luiz fernando	0	100,0%	81,3%	43,8%
comando ver	86,9%	0	71,0%	38,0%
marcos mari	66,0%	66,0%	0	35,5%
marcos anto	31,2%	31,0%	31,2%	0
ederson jose	8,0%	8,0%	7,0%	4,0%
amigos dos a	22,1%	22,1%	18,0%	10,0%
celso luiz rod	34,0%	39,2%	28,0%	15,0%

Fig 2. Relationships matrix containing precalculated associations [29]

5 CONCLUSIONS AND PROBLEMS IN THE EXTRACTION OF ENTITIES FROM THE POLICE REPORT

Various errors may occur during data mining, particularly when treating extractions in textual documents, which can make inconsistent modeling, contribute for distortions or inconclusive results [1] .

Kohonen [24] quotes that are frequently occurring errors when converting text to entities, producing inaccuracies in the calculation of distances.

Goldberg & Senator [25] reported that several information bases have inconsistencies, incomplete data or multiple identifications for the same extracted references.

Han & Kamber [26] reported that many inconsistencies may occur in information bases such as violation of restrictions or cases of redundancies that can be removed by integrating of data routines. Some attributes can take different names in heterogeneous information bases. The errors and inconsistencies can be deleted manually through external references, imposing dependencies between attributes, correction parameters or creation of criticism against violation of restrictions .

May occur with some functional inadequacies applied in entities extraction model in the surveyed bases. The data handlers must identify, debug or discard incompatible documents and routines to reduce errors and deviations that would minimize the expected results in the extraction routines [27] .

Xu and Chen [28] point to problems observed in the extraction of entities related to incomplete data, incorrect or inconsistent in searched data records.

Incomplete data - criminal networks operate in stealth mode or hidden . Criminals minimize interactions, in order to not attract police attention. The data captured may become incompletes, causing the loss of connections between the nodes and loss of integrity on the the network structure.

Incorrect data - inaccuracies relating to identification, physical or addresses can result in errors in the transcription of information that are generated intentionally by the criminals themselves, aiming to confuse the police investigations . Criminals lie about their addresses and their identities when captured or investigated, which can

introduces ambiguities and inaccuracies in the bases of police records.

Inconsistency - information on criminals can come from multiple sources entries simultaneously, feeding the police records inputs, not necessarily consistently. The criminal may appear in historical police records with multiple identifications, presenting itself as different individuals, causing inaccuracies in the processed queries.

Working with criminal historic records and other intelligence policial sources used for investigation, the policial activity lacks intelligence tools to aid and elucidate crimes and discover knowledge in databases occurrences of policials records.

The shared effort between preventive police action and investigative policing is a complex scenario for operational planning decisions. Prevent and investigate deal with the same variables represented by the police force and should be shared as cooperative resources, but competitors in the search for the final results

The efficiency and effectiveness of police investigation request the use of automated tools to cover the entire research cycle, comprising the record of the occurrence, analysis and extraction of the police historical knowledge.

REFERENCES

[1] MARTINS, I. Descoberta de Conhecimento em Históricos Criminais: Algoritmos e Sistemas. Tese de Doutorado PUC-Rio Dep Engenharia Industrial, 2009

[2] HAUCK R.V., H. Atabakhsh, P. Ongvasith, H. Gupta, H. Chen, Using coplink to analyze criminal-justice data, IEEE Computer 35 (3) 30– 37, 2002.

[3] CHEN, H., Chung, W., Xu. J., Wang, G., Qin, Y., and Chau, M., Crime Data Mining: A General Framework and Some Examples, IEEE Computer, 37(4), 50-56, 2004

[4] XU Jennifer, Chen H., Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks . Decision Support System 38 (2004) 473-487

[5] BALUJA, V. Mittal, and R. Sukthankar, Applying Machine Learning for High Performance Named-Entity Extraction. Pacific Association for Computational Linguistics, 1999.

[6] VIDAL L. A. Carvalho. DataMining, a mineração de Dados no Marketink, Medicina, Economia e Administração. Editora Ciência Moderna, Rio de Janeiro, 2005.

[7] McANDREW D, The structural analysis of criminal networks, in: D. Canter, L. Alison (Eds.), The Social Psychology of Crime: Groups, Teams, and Networks, Offender Profiling Series, Aldershot, Dartmouth, vol. III, 1999.]

[8] CHAU M., J. Xu, H. Chen, Extracting meaningful entities from police narrative reports, Proceedings of the National Conference on Digital Government Research (Los Angeles, CA), 2002, pp. 271– 275.

[9] LEE R, Automatic information extraction from documents: a tool for intelligence and law enforcement analysts, Proceedings of 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis, AAAI Press, Menlo Park, CA, 1998.

[10] WITTEN I. H., Zane Bray, Malika Mahoui, W.J. Teahan. Using language models for generic entity extraction. Teahan Computer Science University of Waikato Hamilton, New Zealand, 1999.

[11] CHINCHOR Nancy MUC-7 Overview, seventh Message Understanding Conference and the Second Multilingual Entity Task, CA, EEUU, 1999. Search in august, 2007, http://www.muc.saic.com/proceedings/muc_7_proceedings/overview.html

[12] CHEN, H., and K. J. Lynch. Automatic construction of networks of concepts characterizing document databases. IEEE Transactions on Systems, Man and Cybernetics, 22(5):885-902, September/October 1992.

[13] CHEN, H., Martinez, J., Tobun D. Ng, and Bruce R. Schatz. A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System. Journal of the American Society for Information Science, 1997

[14] MONTEIRO J. Lemos. Morfologia Portuguesa. Pontes Editora. São Paulo 2002

[15] LAROCA M. N. C., Manual de Morfologia do Português - 4a Edição. Editora Pontes, São Paulo, 2005.

[16] HULL D.A. Stemming Algorithms: A Case Study for Detailed Evaluation. In: Journal of the American Society for Information Science 47(1), 1996, p. 70-84.

[17] ANTIQUEIRA L, Nunes M. Oliveira Jr, Costa L. F. Modelando Textos como Redes Complexas. Encontro para o Processamento Computacional da Língua Portuguesa. PROPOR, MG, 2003.

[18] PORTER M.F. The Porter Stemming Algorithm, Computer Laboratory, Cambridge (England) 1997, revisado em Jan 2006. Disponível em http://tartarus.org/~martin/PorterStemmer/, Consulta em setembro 2007.

[19] HAUCK R.V., H. ATABAKHSH, P. ONGVASITH, H. GUPTA, H. CHEN, Using coplink to analyze criminal-justice data, IEEE Computer 35 (1.5.3) 30– 37, 2002.

[20] SCHROEDER J, J. XU, H. CHEN, M, CHAU. Automated criminal link analysis based on domain knowledge. Journal of the American Society for Information Science and Technology Volume 58, # 6 , 2007.

[21] EVANS J, E.MINIEKA. Optimization Algorithms for Networks and Graphs, Marcel Dekker, New York, 1992.

[22] DIJKSTRA E. A note on two problems in connection with graphs, Numerische Mathematik 1 269– 271, 1959.

[23] BOAVENTURA Netto. Teoria e Modelos de Grafos. Editora Edgard Blücher Ltda, São Paulo, 1979.

[24] KOHONEN, T. Self-Organization Maps, Springer-Verlag, Berlin. 1997.

[25] GOLDBERG, H.G., SENATOR, T.E. Restructuring databases for knowledge discovery by consolidation and link formation, Proceedings of the First International Conference on Knowledge Discovery in Databases, AAAI Press, Menlo Park, CA, 1995.

[26] HAN J., M, KAMBER, Data Mining. Concepts and Techniques. Morgan Kaufman San Francisco, USA, 2001.

[27] LIFSCHITZ S.,CÔRTES S., PORCARO R. Mineração de Dados, Funcionalidades, Técnicas e Abordagens. ISSN 0103-9741, PUC-Rio 2002

[28] XU J., CHEN H. Criminal Network Analysis and Visualization: A Data Mining Perspective . Available in http://ai.bpa.arizona.edu/COPLINK/publications/crimenet/Xu_CACM.doc Search in March , 2008

[29] Sistema ANAPHORA, Projeto para Extração e Análise em Histõricos Policiais. Isnard Martins, PUC-Rio, 2008

Graph-based Link Prediction in Cross-session Task Identification

Chao Xu[1], Mingzhu Zhu[1], Wei Xiong[2], and Yi-fang Wu[1]
[1]Information Systems, New Jersey Institute of Technology, Newark, NJ, USA
[2]Mathematics, Computer Science and Information Systems, Northwest Missouri State University, Maryville, MO, USA
Email: [1]{cx26, mz59, wu}@njit.edu, [2]xiong@nwmissouri.edu

***Abstract* - The information needs of search engine users vary in complexity. Some simple needs can be satisfied by using a single query, while complicated ones require a series of queries spanning a long period of time. The search task, consisting of a sequence of search queries serving the same information need, can be treated as an atomic unit for modeling user search preferences and has been well applied in information retrieval to improve the accuracy of search results. Most existing studies have focused on over-session based task identification and heavily relied on human annotations for supervised classification model learning, which are not ideal in large, real time search applications where users have long-term interests spanning over multiple search sessions. In this study, a cross-session based method is proposed for discovering search tasks by modeling the latent structure of task information in the search log dataset, without needing human annotations. Experimental results show that the proposed cross-session based method contributes to an increased accuracy of task identification.**

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms

Algorithms, Performance, Experimentation

Keywords

Search Session, Search Task, Search Log

1. Introduction

The information needs of search engine users span a broad spectrum. Some simple information needs, such as finding a person's homepage or navigating a social networking site, can be accomplished in a single search session. Yet addressing complex information needs, such as planning a vacation, organizing a wedding, or repairing a laptop, requires a user to issue a series of queries, spanning a long period of time and over multiple search sessions. For example, if a user's laptop is broken and he wants to find the solution on the internet, usually, he will search a query first, such as "macbook pro broken", and then go through search results. If the user fails to find relevant information, he would most likely revise his query. Moreover, a user may open multiple web browsers and work on several search tasks at the same time. In this study, the user's search activity is examined at the task level based on the session information, where a search task is defined as a unit of representing one distinct information need.

In most of the existing studies [1, 2], a search task is defined as one or multiple sessions that correspond(s) to a distinct information need. The task is extracted based on the segmented session information, which is also used as the unit for extracting user's interests. These methods are referred to as over-session based task identification, because the task information is constructed over the session units. One obvious problem is that it oversimplifies the user's search activity by assuming that users only work on the same search task within a short period of time. Yet people might work on different search tasks at the same time. Thus, it is needed to examine the search task both within and cross session boundaries to improve the performance of task identification.

Recently, several studies have been conducted on identifying tasks within search sessions. For example, some studies [3, 4] adopt supervised methods to label search tasks using a pairwise classification methods. However, pairwise prediction might not be consistent. For example, two pairs: (query q_i and q_j), (query q_i and q_k) are predicted to be in the same task, while query q_j and q_k are not. Meanwhile, some studies [5, 6] use an external dataset such as the Open Directory Project or Wikipedia. Because the labels and categories of search tasks are generated from an external dataset, the total number of labels or categories is

fixed rather than adaptive to the user's search activities. However, it is usually the case that most users have multiple information needs and they are dynamically changing [7]. To solve these problems, in this study, a cross-session based query analysis method with a best-link model is proposed to improve the performance of task identification. Specifically, search queries within a search session are segmented into sub-tasks by using the best-link model to learn query connections from users' search activities. And a graph-based representation method is utilized to calculate the contextual pairwise similarity of queries. Then, search tasks are identified by grouping similar sub-tasks from all search sessions together.

This paper makes the following contributions: 1) a cross-session based task identification method; 2) a best-link prediction method for identifying the structural dependencies of queries; and 3) a graph-based representation method for determining the link relation between a pair of queries.

The rest of the paper is organized as follows. Section 2 summarizes related studies. Section 3 presents the proposed cross-session based task identification method. Section 4 introduces the dataset, experimental design, evaluation methods, and performance comparison between the proposed method and baselines. Section 5 summarizes the main conclusions of this study.

2. Related Work

A search session, as defined by Boldi et al. [4], is a sequence of queries issued by a single user within a specific time limit. The related queries of the same session often refer to the same search goal or search activity. Based on this assumption, He et al. [5] propose to group queries into search sessions through detecting the topic shifts among queries. Hassan et al. [6] adopt topic models to extract session-level search goals. It is concluded that the method of examining user search activities through search sessions outperforms the traditional approaches that are based on only relevance feedbacks. Piwowarski et al. [7] model a hierarchy of users' search activities through a layered Bayesian network to identify distinct patterns of users' search behaviors. They use classification methods to learn the connection of latent states for a clicked document to the relevance assessment of that document without considering the document content. Mei et al. [8] propose a framework of studying the sequences of users' search activities, in which an algorithm is introduced to segment the query stream into goals and missions.

Recently, several studies have noticed the necessity of going beyond the session boundary and examining the user's information needs in a task. For example, Spink et al. [17] indicate that multi-tasking behavior occurs frequently in which users switch search tasks within a short period of time. Lucchese et al. [14] model task-based sessions to extract multiple tasks from the search session. Meanwhile, Hassan and White [9] indicate that a search task can be complex and span a number of search sessions. To tackle this, they propose a method to generate a task tour which comprises a set of related search tasks. Kotov et al. [11] explicitly define the cross-session task as the one extending over multiple sessions and corresponding to a certain high-level search intent. To extract cross-session tasks, Jones et al. [18] have built classifiers to identify task boundaries and pairs of queries belonging to the same task. Agichtein et al. [19] have examined the cross-session task identification by using a binary classification method and have found that different types of tasks have different life spans. Besides, a few studies [11, 20, 21, 22, 23, 24, 25] have proven the effectiveness of classifying queries and web pages into search tasks on improving the search performance. Although they prove that the search task information contributes to the improvement of search performance, all of them have two main issues. The first issue is that they define the search task manually. The fixed number of search tasks is not suited to predict the user's future search activities – since it will be an incomplete representation, if the number is too small; and noises will occur, if the number is too large. The second issue is that existing classification-based methods rely on human annotated dataset for training models, which is not applicable when only few manual annotations are available.

The main difference between this study and existing cross-session based task identification studies is that we model this problem as a link prediction problem rather than a binary classification problem. The advantage of this study is that the latent dependencies between queries within each task are modeled explicitly.

3. Methods

3.1 Task analysis

Search logs are proven as a valuable data resource for analyzing user's search activities and information needs. In this study, the AOL search log dataset is examined to extract users' search tasks. A search log is a dataset that records users' search activities, which can be denoted by the vector $< a_i, q_i, t_i, c_i, r_i >$, where a_i is the identifier of the user, q_i is the query submitted by the user a_i, t_i is the time of the user activity, c_i is the click on the relevant result returned for q_i, and r_i is the rank position of c_i [10].

The primary mechanisms for segmenting the logged query streams are session-based. A search session is usually considered as the basic unit of information in search log analysis [1]. In a search engine which works in the session mode, the user's search activities are recorded and earlier search data, i.e. queries and results clicked, in the same session is used to update user's current search actions. A search session is defined as a sequence of search activities $S = \{< a_j , q_j , t_j , c_j , r_j > \ldots < a_k , q_k , t_k , c_k , r_k >\}$ issued by a single user within a specific time limit.

Table 3.1 Sample of Session Segmentation

User_ID	Query	QueryTime	Clicked_URL	Rank
382351	apple warranty	2006-04-24 22:00:21	http://www.superwarehouse.com	6
382351	ipod questions	2006-04-24 22:17:42	http://www.maclink.co.uk	1
382351	dogwood festival	2006-04-29 21:46:30	http://www.fayettevilledogwoodfestival.com	5
382351	myrtle beach map	2006-05-29 22:58:09	http://travel.yahoo.com	3
382351	cherry grove south carolina	2006-05-29 23:03:03	http://www.tripadvisor.com	4
382351	cherry grove south carolina	2006-05-29 23:03:03	http://www.cherrygrovebeachhouses.com	9
382351	body kits for civic	2006-05-30 20:03:12	http://www.modacar.com	2
382351	motley crue jackets	2006-03-01 17:41:26	http://www.motley.com	9
382351	ticketmaster	2006-03-16 14:40:40	http://www.ticketmaster.com	1

Methods of extracting relevant sessions from search logs should examine all queries issued by a user. Short inactivity timeouts between user actions are applied as a means of demarcating session boundaries [4]. In the field of session segmentation, the relations between queries are categorized as Topic Continuation and Topic Shift. In Figure 3.1, query q_1 and q_2 are semantically related, so they should be grouped in the same session and the relation between them is Topic Continuation. On the contrary, q_2 and q_3 have no semantic relation, so the relation between them is Topic Shift, which generates a session boundary. In this study, user inactivity periods are adopted to segment the search session. The time interval within a search session should be less than a threshold σ (where σ is set at 25 minutes according to an empirical study). Table 3.1 shows a sample of segmented sessions.

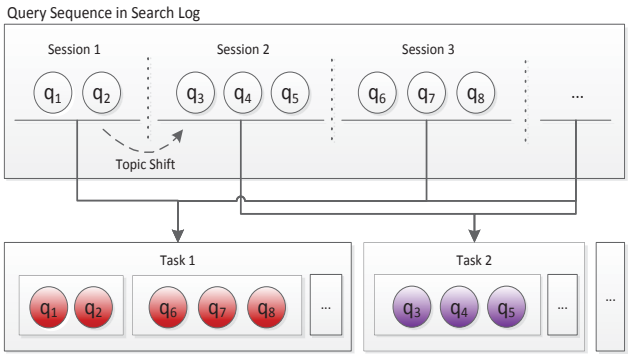


Figure 3.1 Task identification by grouping similar search sessions.

Meanwhile, search engine users have various search intentions. Addressing complex information needs usually requires a user to issue a series of queries, spanning across multiple search sessions. To tackle this problem, a fine-grained task identification method, which is also called the cross-session based task identification method, is proposed in this study. As shown in Figure 3.2, search queries within a search session are segmented into sets of queries which are formed to achieve specific search tasks. Each set of queries is called a sub-task. For example, in the first session, predicting q_2, q_4 and q_5 belonging to the same task would immediately lead to the conclusion that all these three queries are in the same task, even though q_2 and q_5 are not directly connected to each other. Then, after examining all search sessions of the user, search queries related to a particular search task are identified by grouping similar sub-tasks together.

To generate these sub-tasks for each search session, an unsupervised best-link model is proposed. The main idea is that the best-link defines a hierarchical tree structure of "strong" connections among the queries: rooted in the fake query q_0, and each sub-tree of q_0 corresponds to one specific search sub-task in a search session. For a new query, it can only belong to a previous search task or be the first query of a new task. Therefore, the temporal order provides a helpful signal to explore the dependency between queries.

Specifically, given a query sequence $Q = \{q_1, q_2, \ldots, q_m\}$ within a search session, f is introduced to refer the latent best-link structure. $f(q_i, q_j)$ indicates the existence of a link between q_i and q_j as following:

$$f(q_i, q_j) = \begin{cases} 1, & \varphi(q_i, q_j) > \gamma \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

where $f(q_i, q_j) = 1$, if query q_i and q_j are directly connected; and otherwise, $f(q_i, q_j) = 0$. $\varphi(q_i, q_j)$ indicates the similarity between query q_i and q_j. To model the first query of a new search session, i.e., the query that does not have a strong connection with any previous queries, a fake query q_0 is added at the beginning of each search session. All the queries connecting to q_0 would be treated as the initial query of a new search sub-task. Besides, it is enforced so that a query can only link to another query in the past, or formally,

$$\sum_{i=0}^{j-1} f(q_i, q_j) = 1, \forall j \geq 1 \tag{3.2}$$

Note that the best-link method is conducted within each search session to generate a list of sub-tasks. Similar sub-tasks are grouped together as a search task using the hierarchical clustering [8].

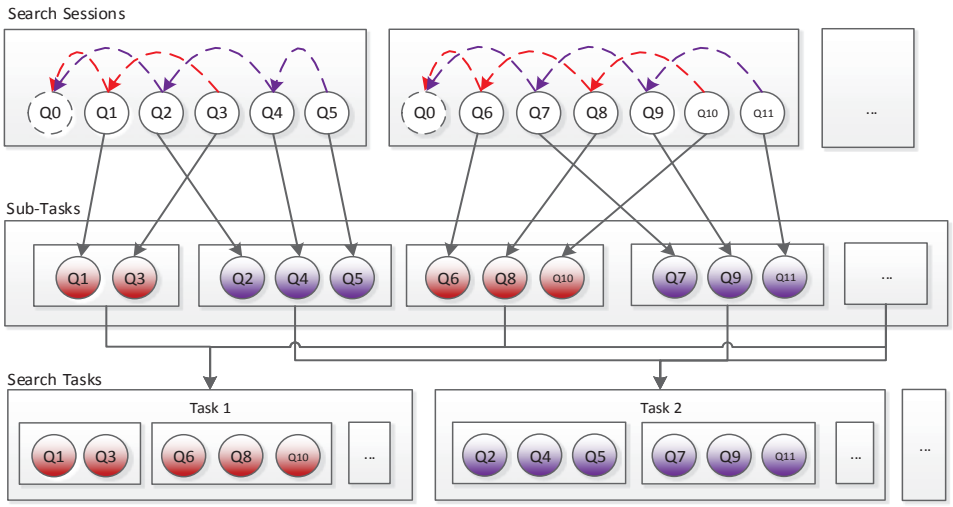


Figure 3.2 Task identification by grouping similar sub-tasks.

3.2 Graph-based Link Prediction

To achieve the latent structure $f(q_i, q_j)$ as defined in formula 3.1, $\varphi(q_i, q_j)$ should be determined first. As shown in Figure 3.3, the pairwise similarity between relevant feedback documents of q_i and q_j is adopted for determining the link relation between two queries. Specifically, the queries resulting in none click action are defined as invalid queries, such as q_3, q_4 and q_6. By contrast, the queries resulting in at least one clicked result are defined as valid queries, such as q_2 and q_5. All invalid queries are ignored in this study as did in one existing study [16]. For example, to determine if q_2 and q_5 belong to the same task, two similarities between the relevant feedback documents of these two queries are calculated, including $sim(d_{2,1}, d_{5,3})$ and $sim(d_{2,1}, d_{5,5})$, where $d_{2,1}$ denotes the first retrieved document of q_2, sim() represents the similarity between a pair of queries. Then q_2 and q_5 are segmented into the same task if $sim(d_{2,1}, d_{5,3})$ or/and $sim(d_{2,1}, d_{5,5})$ is/are bigger than the γ as indicated in formula 3.1.

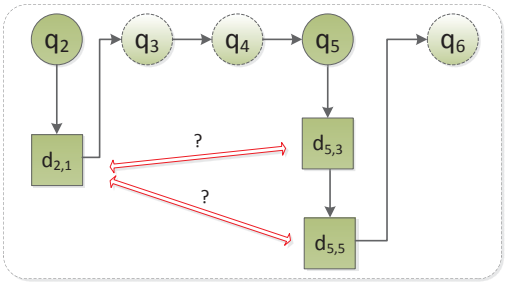


Figure 3.3 Example of the pairwise similarity.

However, there are two problems of calculating the above pairwise similarity using the original page contents, including data noise and data scarcity [12]. On one hand, many relevant documents contain other non-pertinent information such as advertisements and navigations, causing difficulty in summarizing their latent meanings. On the other hand, for a search log dataset, such as AOL, it does not contain snippets, but URLs that might not point to a live site anymore, or for which the content might have been changed after the dataset was created. To tackle these problems, a two-step graph-based representation method is proposed for predicting the pairwise similarity between the relevance feedback documents from two different search queries.

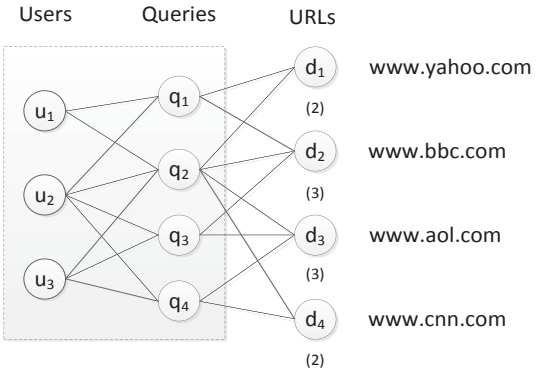


Figure 3.4 Example of a click graph.

First, a click graph is constructed for generating the pseudo-document of each clicked URL. An example of a click graph with four queries and four URLs is shown in Figure 3.4. The edges of the graph capture the relationships between the queries and the URLs. Based on the observation that different users may use different queries to describe their latent topics of interests within a particular web page, it is proposed to generate a pseudo-document for each URL by combining all its connected queries in this graph. For example, two different queries (q_1 and q_2) from different users (u_1, u_2 and u_3) are connected to the same URL, "www.yahoo.com". The queries (q_1 and q_2) are then

combined to represent the pseudo-content of "www.yahoo.com".

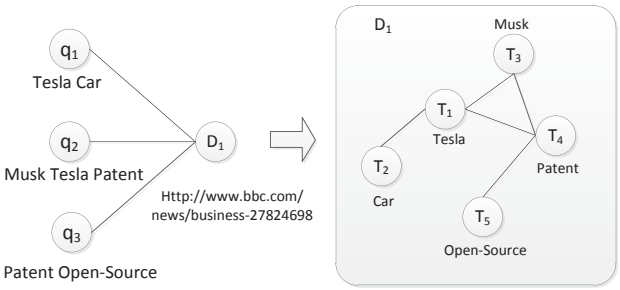


Figure 3.5 Graph-based representation of a relevance feedback document.

Second, simply adopting a bag-of-word to represent the content of a document will lose the structural semantic information. To tackle it, a graph-based representation of the pseudo-document is proposed. Specifically, the unique terms, denoted as $\{T_i\}$, are extracted from the pseudo-document. For example, as shown in Figure 3.5, there are five unique terms within the pseudo-content of D_1, including T_1: "Tesla", T_2: "Car", T_3: "Musk", T_4: "Patent", and T_5: "Open-Source". Afterwards, a pair-wise examination is automatically conducted within each query string to determine the existence of a binary non-directional edge between two terms. For example, T_1 and T_2 are connected with an edge because they are in the same query q_1; T_2 and T_3 are not connected because no query in D_1 contains both of them. Then each pseudo-document is represented as a graph G = (N, E), where N denotes the nodes (unique terms) and E denotes the edges. Finally, given two semantic graphs $G_1 = (N_1, E_1)$ and $G_2 = (N_2, E_2)$ constructed for two relevance feedback documents, a graph similarity measure is adopted to estimate their semantic relatedness. Specifically, the metric called "p-homomorphism" [13] is adopted as the underlying graph matching method, because the p-homomorphism concept extends the traditional graph homomorphism and sub-graph isomorphism concepts by additionally mapping edges from one graph to their corresponding edge paths in another graph.

4. Experimental Design

4.1 Data Sets and Evaluation Methods

Lucchese et al. [14] develop a Web application that helps human assessors manually identify the optimal set of user tasks from the AOL query log. They produce a ground truth that can be used for evaluating any automatic user task discovery method, which is also publically available at "http://miles.isti.cnr.it/~tolomei/downloads/aol-task-ground-truth.tar.gz". It contains a total of 554 search tasks with average 2.57 queries per task. And 143 cross-session tasks are contained in this dataset. In this experiment, this dataset was adopted as the ground truth for comparing the performance of the proposed task identification method and the baselines.

To evaluate the performance of the proposed task identification method, it is necessary to measure the degree of consistency between manually-extracted user tasks of the ground truth and search tasks generated by our algorithms. Specifically, both classification- and similarity-oriented measures [14] were adopted in this experiment. Predicted task indicates the user task where a query is assigned by a specific algorithm, while true task indicates the user task where the same query is in the ground truth.

Classification-oriented approaches measure how closely predicted tasks match true tasks. F1 is one of the most popular measures in this category, as it combines both precision and recall. In this study, precision measures the fraction of queries that were assigned to a user task and that were actually part of that user task. Instead, recall measures how many queries were assigned to a user task among all the queries that were really contained in that user task. Globally, F1 evaluates the extent to which a user task contains only the queries that were actually part of it. Two notations, $p_{i,j}$ and $r_{i,j}$, are introduced to represent the precision and recall of predicted task i with respect to true task j, then F1 corresponds to the following weighted harmonic mean of $p_{i,j}$ and $r_{i,j}$.

$$F1 = 2 \times p_{i,j} \times r_{i,j} / (p_{i,j} + r_{i,j}) \quad (4.1)$$

Similarity-oriented measures consider pairs of objects instead of single objects. Let T be the sets of predicted tasks, four values were computed, including: 1) t_n - number of query pairs that are in different true tasks and in different predicted tasks (true negatives); 2) t_p - number of query pairs that are in the same true task and in the same predicted tasks (true positives); 3) f_n - number of query pairs that are in the same true task but in different predicted tasks (false negatives); 4) f_p - number of query pairs that are in different true tasks but in the same predicted task (false positives). Then, two different measures were adopted as following:

Rand index: $R(T) = (t_n + t_p) / (t_n + f_p + f_n + t_p)$ (4.2)

Jaccard index: $J(T) = t_p / (f_p + f_n + t_p)$ (4.3)

4.2 Experimental Setup and Results

The experiment analyzed the contributions of the proposed cross-session based task identification methods including best-link method (BL) and best-link with graph-based representation method (BL-G). The difference is that BL adopts the bag-of-word method for representing the features of the pseudo-document while BL-G uses the proposed graph-based representation method for modeling rich semantic features.

Three baselines were adopted in this experiment, including one over-session based method and two cross-session based methods. The best performing over-session based

method (OS) is proposed by Luxenburger et al. [15] who adopt a hierarchical clustering method in which the atomic units to be clustered are past sessions. The two best performing cross-session based methods, QC_wcc and QC_htc, are proposed by Lucchese et al. [14]. Specifically, QC_wcc performs clustering by dropping "weak edges" among queries and extracting the connected components as tasks. QC_htc assumes that a cluster of queries can be well represented by only the chronologically first and last queries in the cluster; therefore only the similarity among the first and last queries of two clusters is considered in the agglomerative clustering.

The annotated log dataset was randomly split into a training set with 270 annotated search tasks, and a test set with the other 270 annotated tasks. The parameters in each model were tuned by a 5-fold cross-validation on the training set. All baselines and our methods were trained on the same training set.

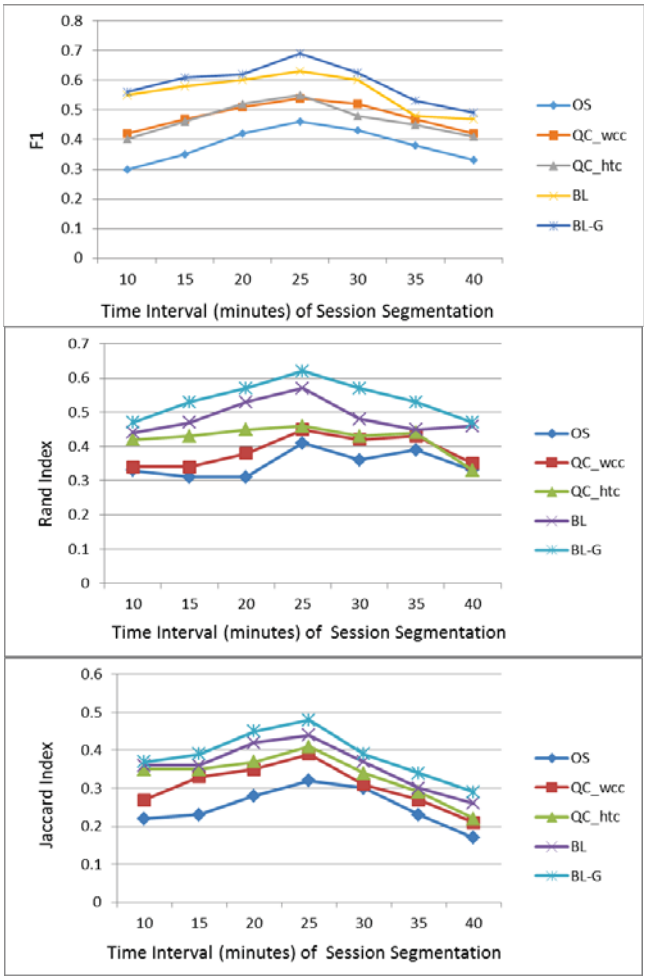


Figure 4.1 Performance comparisons of proposed methods with baselines.

Figure 4.1 shows the performance comparisons between proposed methods and baselines. It was first observed that the session boundary does impact the performance of all compared task identification methods. Most of them achieve the highest performance on these three evaluation metrics when the time interval is set at 25 minutes, which is consistent with existing studies [7, 15]. The proposed methods BL and BL-G outperformed QC_wcc and QC_htc significantly in all three metrics. The reason is that both QC_wcc and QC_htc target on predicting whether two queries represent the same task. However, the pairwise prediction cannot directly generate the task information and post-processing is required to obtain the tasks. Such a post-processing is independent from the classifier training therefore is not necessarily optimal.

Also, the OS baseline, as the over-session based method, performed much worse than the others especially on Rand Index and Jaccard Index metrics. The possible reason is that it assumes that users work on the same task within each period of a search session which results in a high f_p value. Finally, BL-G performs better than BL, because BL-G utilizes the proposed graph-based representation while BL adopts the bag-of-word representation in which the semantic structure is lost.

Table 4.1 Performance Comparisons between Session-based and Non-session based Task Identification Methods

Task Identification Methods		Evaluation Metrics		
		F1	Rand Index	Jaccard Index
Non-session based	BL-NoSS	0.560	0.478	0.422
	BL-G-NoSS	0.603	0.539	0.439
Session-based	BL	0.628	0.571	0.446
	BL-G	**0.695**	**0.619**	**0.483**

So far, the proposed best-link model for task identification is conducted within a session scope. One interesting question is whether the session information is contributive in the proposed best-link method. Table 4.1 illustrates the performance comparisons between the best-link methods using the search session and the ones without using the session data (denoted as BL-NoSS and BL-G-NoSS respectively). Note that both BL and BL-G were optimized by setting session interval at 25 minutes. It was observed that the proposed methods, BL and BL-G, using session information performed much better than the ones without using the session data, i.e., BL-NoSS and BL-G-NoSS. For example, the F1 scores of BL and BL-G were 0.628 and 0.695, whereas those of BL-NoSS and BL-G-NoSS were 0.560 and 0.603. The major reason for these performance differences is that the session plays the role of setting a temporal boundary for identifying the latent link structure of queries from the same search task. And this boundary prevents the incorrectly predicted link information from spanning so that the prediction error made in previous session will not affect the prediction accuracy in the current session. Furthermore, the fact that BL-G and BL-G-NoSS outperformed BL and BL-NoSS respectively, indicates that the proposed graph-based representation for query similarity computation is more effective.

5. Conclusions

Users switch search tasks frequently during their search activities, thus developing methods to extract these tasks from historical data is central to understanding longitudinal search behaviors and developing search systems to support users' long running tasks. In this study, a new cross-session based method is presented for extracting search tasks from users' historic search activities. Specifically, a best-link model is introduced which is capable of learning query connections from users' searching activities. Then a graph-based representation method is utilized to estimate the contextual pairwise similarity of queries. Finally, an experiment using a publically available annotated dataset from AOL log is conducted to demonstrate the superior performance of our method in identifying search tasks versus a number of state-of-the-art algorithms.

6. References

[1] B. Tan, X. Shen, and C. Zhai, "Mining long-term search history to improve search accuracy," Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, Philadelphia, PA, USA, 2006.

[2] M. Ji, J. Yan, S. Gu, J. Han, X. He, W. V. Zhang, "Learning search tasks in queries and web pages via graph regularization," Proceedings of the 34th international ACM SIGIR conference on research and development in Information Retrieval, Beijing, China, 2011.

[3] D. H. Widyantoro, T. R. Ioerger, and J. Yen, "An adaptive algorithm for learning changes in user interests," Proceedings of the eighth international conference on information and knowledge management, Kansas City, Missouri, United States, 1999.

[4] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna, "The query-flow graph: model and applications," Proceedings of the 17th ACM conference on information and knowledge management, Napa Valley, California, USA, 2008.

[5] D. He, A. Göker, and D. J. Harper, "Combining evidence for automatic web session identification," information process of management , 2002.

[6] A. Hassan, R. Jones, and K. L. Klinkner, "Beyond DCG: user behavior as a predictor of a successful search," Proceedings of the third ACM international conference on Web search and data mining, New York, New York, USA, 2010.

[7] B. Piwowarski, G. Dupret, and R. Jones, "Mining user web search activity with layered bayesian networks or how to capture a click in its context," Proceedings of the second ACM international conference on Web search and data mining, Barcelona, Spain, 2009.

[8] Q. Mei, K. Klinkner, R. Kumar, and A. Tomkins, "An analysis framework for search sequences," Proceedings of the 18th ACM conference on information and knowledge management, Hong Kong, China, 2009.

[9] A. Hassan and R. W. White, "Task tours: helping users tackle complex search tasks," Proceedings of the 21st ACM international conference on information and knowledge management, Maui, Hawaii, USA, 2012.

[10] B. Zhou, D. Jiang, J. Pei, and H. Li, "OLAP on search logs: an infrastructure supporting data-driven applications in search engines," Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, Paris, France, 2009.

[11] A. Kotov, P. N. Bennett, R. W. White, S. T. Dumais, and J. Teevan, "Modeling and analysis of cross-session search tasks," Proceedings of the 34th international ACM SIGIR conference on research and development in Information Retrieval, Beijing, China, 2011.

[12] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Probabilistic document-context based relevance feedback with limited relevance judgments," Proceedings of the 15th ACM international conference on information and knowledge management, Arlington, Virginia, USA, 2006.

[13] W. Fan, J. Li, S. Ma, H, Wang., and Y. Wu, "Graph homomorphism revisited for graph matching," Proceeding of the VLDB endowment, 2010.

[14] C. Lucchese, S. Orlando, R. Perego, F. Silvestri, and G. Tolomei, "Identifying task-based sessions in search engine query logs," Proceedings of the fourth ACM international conference on Web search and data mining, 2011.

[15] J. Luxenburger, S. Elbassuoni, and G. Weikum, "Matching task profiles and user needs in personalized web search," Proceedings of the 17th ACM conference on information and knowledge management, California, USA, 2008

[16] L. Bing, W. Lam, and T.-L. Wong, "Using query log and social tagging to refine queries based on latent topics," Proceedings of the 20th ACM international conference on information and knowledge management, Glasgow, Scotland, UK, 2011.

[17] A. Spink, M. Park, B. J. Jansen, and J. Pedersen, "Multitasking during web search sessions," Information processing and management, 2006.

[18] R. Jones and K. L. Klinkner, "Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs," Proceedings of the 17th ACM conference on information and knowledge management, 2008

[19] E. Agichtein, R. W. White, S. T. Dumais, and P. N. Bennet, "Search, interrupted: understanding and predicting search task continuation," Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval, 2012

[20] C. Xu, M. Zhu, Y. Liu, and Y. B. Wu. "User Profiling for Query Refinement", 20th Americas Conference on Information Systems, 2014

[21] D. Shen, J. Sun, Q. Yang, and Z. Chen. "Building bridges for web query classification," Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, 2006

[22] Z. Liao, Y. Song, L. He, and Y. Huang, "Evaluating the effectiveness of search task trails," Proceedings of the 21st international conference on World Wide Web, 2012

[23] C. Xu, M. Zhu, Y. Liu, Y. B. Wu, "Personalizing Query Refinement Based on Latent Tasks", 2014 International Conference on Data Mining, 2014

[24] M. Zhu, C. Xu and Y. B. Wu, "Topic Model based Query Intent Prediction for Search by Multiple Examples". 2014 International Conference on Artificial Intelligence, 2014

[25] P. S. Dhillon, S. Sellamanickam, and S. K. Selvaraj, "Semi-supervised multi-task learning of structured prediction models for web information extraction," Proceedings of the 20th ACM international conference on information and knowledge management, 2011

Learning Temporal Regression Models and Voronoi Tessellation for Job Offers Recommendation.

Sidahmed Benabderrahmane (a), Nedra Mellouli (a), Myriam Lamolle (a), Jean Baptiste Gabriel (b).
(a): Paris 8 University, IUT Montreuil, 140 Rue de la Nouvelle France, 93100 France.
(b): Multiposting, 3 rue Moncey, 75009 Paris - France.
s.benabderrahmane@iut.univ-paris8.fr Tel: 0033(01)48703700.

***Abstract*— Nowadays, the best ways to attract job candidates is through dedicated web-based portals, and therefore match their related data automatically using optimized algorithms. In this perspective, with the goal of sharing, at best, the job offers, many online job boards have been created, the choice of which can be sometimes very hard for the recruiters that aim at attracting the best possible candidates in the shortest amount of time. Based on these considerations, in this paper, we propose a novel jobboard recommendation system that aims at estimating the best potential job-boards for a given text job offer. Our efficient predictive model for job boards recommendation, is based on a hybrid representation, that combines semantic knowledge and time series forecasting. The semantic classification of job boards requires a textual analysis using domain knowledge. The time series analysis module is to predict the best job board for a given offer. The proposed system has been evaluated on real data, and preliminary results seem very promising.**

***Index Terms*— Recommendation; Time series; Clustering; Forecasting; Data Mining; Big Data.**

I. Introduction

Since the last two decades, the use of Internet for recruitment purposes has grown considerably. This recruitment process, also known as "*e-recruitment*" is based on the use of information technology and communications. In this context, expansion of the Web has led to an increase in the number of job diffusion web sites (also called *job board*) and consequently the number of candidates that can be contacted through these intermediary tools. However, despite the wide dissemination of existing platforms of e-recruitment, the main concern of recruiters rest of "finding" the best profiles (i.e., the most talented potential candidates) for a given position. To better target potential candidates, some problems are to be solved such as clarity of the offer and its relevance to potential candidate profile, or adequacy of the job board in relation to the core business of the offer itself even. The search for the best candidates for a given offer returns among others target the most appropriate job board, and after that the most relevant profiles among the mass of available profiles. Current recommendation systems process only a part of the recruitment process, concentrating on matching offers with CVs. However, the selection of the most appropriate job board regarding an offer is also very important for the optimization of this fully digital recruitment process. For this reason, various questions arise concerning the criteria for relevance of a job board over a given offer. For example, is a job board considered relevant if the number of offers are increasing? Or, if the number of visits and / or the number of clicks to view the offers by potential candidates tend to grow compared to those observed in the past?

Our main goal is to provide a tool to help recruiters to i) select the most relevant job board for a given position, ii) diffuse more effectively job ads, that is to say at the right place at the right time, iii) provide tools to connect candidates and offers automatically.

To meet the above objectives, we are also faced with problems related to the specificity of the data to be processed. We dispose from our industrial partner, a history of job advertisements on web sites, and the quantity of their visits (clicks), that are stored in a big database. The recorded data also concern the number of candidates obtained through various job boards and social networks. In this context, we propose a recommendation system of *job boards* based on a hybrid model combining modular semantic classification approaches, and time series forecasting [4]. The semantic classification of job boards and job ads requires a textual analysis of the content on the basis of business description that is given by a public french organization (ROME code [1]). The time series analysis module, aims to predict the best *job board* for a given offer, combined with textual analysis module.

The rest of this article is organized as follows: a state of the art will be discussed in section 2. The proposed model will be presented in section 3. Finally, in section 4 we discuss preliminary results and conclude the paper in section 5.

II. State of the Art

Nowadays, few automatic recommendation systems of job offers to particular users exist. These systems are generally classified into three main categories namely textual recommendation systems [1], recommendation systems based on collaborative filtering, and hybrid recommendation systems [5].

Textual-based recommendation systems analyze the content of job descriptions as well as information provided by users to identify the semantic content. To that aim, two types of semantic analysis approaches exist: approaches based on ontologies [8] and text mining approaches [10]. Whatever the approach used in the purely textual recommendation systems, weaknesses may occur. Indeed, the existing approaches require manual annotation by the recruiter and the

[1] www.pole-emploi.fr/candidat/le-code-rome-et-les-fiches-metiers-@/suarticle.jspz ? Id = 15734

candidate to describe both job offers and CVs. Therefore, the volume of processed data is quite large and require the use of highly optimized algorithms. Collaborative filtering systems are based on the analysis of the opinions of a group of users. Their opinions are considered similar to that of an active identified user. These recommendation systems can target CVs only from items related information (such as the title). The use of items certainly reduces the mass of processed data but with a loss of precision. As for hybrid systems, they combine the two previously mentioned categories.

In parallel of the semantic approaches, more formal approaches based on vectors and probabilistic models have been proposed [7] for profiling applications according to a specific offer. Although these approaches seem to be transposed to our problem since they concern the profiling of a job boards for a given job, they are unusable in our context because they only deal with text data. Finally, another approach is essentially based on the predictive linear models was proposed in [10]. This work considers the problem of recommendation as a prediction of the performance of offers on a job board regardless of the behavior of this job board in the past. The linear model was proposed in that work assumes that the model parameters are independent for simplicity. But the real data do not always check their working hypothesis since the dependencies are spatial order (depending on the job board) and also temporal (dependent of the past).

Most of these recommendation systems could be improved if the temporal dimension of information related to the job board was more taken into account in the models. We wish in this work, to consider the information relative to the temporal aspect of the dissemination of offers in the different job boards, to create a predictive model based on the values observed in the past. We propose a representation based on time series, for highlighting the trend and seasonality in the recruitment data. With these informations, decision-making and recommendation of relevant job boards for job opportunities, could contribute to long-term automating the assignment of these bidding job offers to one or more the most appropriate job board. It is this approach that we favor in our study and that we will detail starting with the presentation of our model in the following sections.

III. Description of the temporal recommendation system architecture

As reported in the precedent section, our interest in this study is to characterize the best job boards for a given job offer to automatize the the process of diffusion of vacancies (postings). Job boards of vacancies available on the web are multiple. Some may be specialized to broadcast certain categories of business or certain types of deals, for example, internships, PhD, or fixed-term contracts. Defining the intrinsic characteristics common to all job boards is a very important step for the analysis of their behavior in order to compare and evaluate them. These characteristics are related to the properties of vacancies advertised on these job boards. In particular, in this study we consider initially the business description of the offer, and number of its clicks accrued per job board over a given period. We first present the classification used for the characterization of offers based on their type of business. This classification is used later in the formalization of data relating to job boards.

A. Data formalization

For the formalization of offers, we will use two different types of textual content: jobs and job categories (business classes). An *offer* can be defined as a structured text document used to formalize an offer of an employer. It is divided (or can be divided theoretically) in various fields such as title, business description, skills, education level, etc., organized according to the publisher and / or some high standards level. Given this definition, we formalize the contents of a job offer j as a set of vectors, each representing a text field, as follows:

$$o_j = (\vec{v_{j,1}}, \vec{v_{j,2}}, ..., \vec{v_{j,k}}) \quad (1)$$

where k is the number of text fields in the offer j and v is a vector of weighted keywords representing the frequencies of words of each field (according to the method described by [9]). Specifically, considering a generic text field i (with i ¡k) of job j, we formalize its contents with a vector $v_{j,i}$

$$\vec{v_{j,i}} = \{w_{j,i,1}, w_{j,i,2}, ..., w_{j,i,n}\} \quad (2)$$

where n is the size of vocabulary of the field and $w_{j,i,k}$ is the inverse frequency (TF-IDF) of k terms in the i field in the j offer. Similarly, a job category (also known generically category in this article) can be defined as a textual description of a specific category of occupations. Its definition is generally provided by a domain expert (or any authority) and can be used effectively for classification and indexing of job offers. In our case, we used the French public ROME code categories. Thus, according to the same principle, we formalize the contents of a job class c as

$$c_i = (\vec{v_{i,1}}, \vec{v_{i,2}}, ..., \vec{v_{i,l}}) \quad (3)$$

where l is the number of terms describing the job class c, v is the key words vector representing the frequency of terms. Then we used a vector-based distance and an SVM classifier to annotate each job offer to a semantic business category.

Each day, a set of offers is deposited on one or more job board on a given date. An offer made on a job board has a finite life cycle. In this period, the number of clicks associated with each bid is incremented. Therefore, the daily number of clicks associated with an offer and job boards is available. This number is shown on other time scales: weekly, monthly, and annual midyear. We denote by T the period or the time scale associated with the number of clicks considered. To formulate such data, in particular the number of clicks, we consider a job board, noted JB, as a set of offers on a given period T:

$$JB_T = \cup_j o_j \text{ for } \ j = 1, ..., p \quad (4)$$

Using the vectorial classification of offers described previously, each JB become a class of offers on a period T:

$$JB_T = \cup_k c_k \text{ for } \ k = 1, ..., m \quad (5)$$

and a class c_j as the union of offers contained in :

$$c_j = \cup_i o_i \text{ for } \ i = 1, ..., n \tag{6}$$

For each class c_j, we introduce a ratio X^{c_j} calculated as the total number of relative clicks of offers in this class in a period T :

$$X^{c_j} = \frac{nb.clic}{|c_j|} \tag{7}$$

in particular, when j=1 in a JB then $X^{JB} = \frac{nb.clic}{|JB|}$.
In the following of this paper, we consider T a discrete interval $[1, N]$. Having a series of observations $X_1^{c_j}, X_2^{c_j}, .., X_N^{c_j}$ on a fixed period T, we propose the definition of previsions on a date N with a time series of observations, to estimate $\hat{X}^{c_j}(N, h)$ on future dates within a given horizon h. The objectives of temporal analysis in our study are multiple. Firstly, it concerns the prevision of future realization of a random variable X^{c_j} using the previously observed values $X_1^{c_j}, X_2^{c_j}, X_N^{c_j}$ for each class c_j and for each JB. Secondly, we are interested by estimating the trend of time series. In addition, we are interested in analyzing the variations; for example, one may ask whether an observed change in the number of visitors to a JB is the result of a seasonal fluctuation or is a reflection of a trend. Finally, evaluation of the impact of an event on a variable will measure the JB sensitivity to potential disturbances and noise. For these reasons, we will use univariate time series only, and we notice the variable X^{c_j} by x_t observed at time t. We chose to use a regression model applied to the time series using the information on the number of clicks on the offers. The figure 4 give an example of a time series of job board, where values x_t are the clicks ratios between 2008 till 2014 (1716 days).

1) Statistical Data Analysis: To analyze a time series (x_1, x_2, ..., x_n), it is useful to have statistical indexes to summarize the series. As an indicator of central tendency, we calculated firstly the average as follows: $\bar{x_n} = \frac{1}{n}\sum_{j=1}^{n} x_j$. We also calculated the index of empirical variance (or standard deviation), to rise up comprehensive information on the dispersion of temporal observations with respect to their central tendency. The variance of a set of values in a time series is defined by: $\hat{\sigma}(0) = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x_n})^2$. Later we calculated dependencies between two successive observations by the empirical auto covariance (of order h): $\hat{\sigma}_n(h) = \frac{1}{n-h}\sum_{t=1}^{n-h}(x_t - \bar{x_n})(x_{t+h} - \bar{x_n})$. After that we calculate for each series the empirical autocorrelation that is given by the ratio of auto-covariance and the empirical variance: $\hat{\rho}_n(h) = \frac{\hat{\sigma}_n(h)}{\hat{\sigma}_n(1)}$. These autocorrelations characterize dependencies between series values (x_1, x_2, ..., x_{n-1}) et (x_1, x_2, ..., x_n). This value can provide an overall idea about the implicit regression in the data set. Indeed, linear regression can be observed if the value of $\hat{\rho}_n(h)$ is close to +1 or -1. More auto-correlation tends to 1 in absolute value, the higher the series has a trend. The slope of the linear regression line follows the sign of $\hat{\rho}_n(h)$, while the cloud of the series values is more rounded when it is close to zero.

2) Trend and seasonality in time series: As part of our study, we seek to identify seasonality and trends in job boards time series. Seasonality is an important factor that indicates the repetition frequency of a phenomenon in a periodic manner. The trend is an attendance indicator of a stationary, increasing or decreasing job board. Therefore, it is useful to decompose a time series in order to separate the content of the trend, irregular components and the seasonal component, if it is present. In the case where the series is non-seasonal, that means it is composed of a trend and an irregular residual component. Thus, the decomposition generally requires the separation of the three (or two) sources and estimation of the trend. There are several decomposition models present in the state of the art [11]. Among these models, we opted for the special Holt-Winters (HW) probabilistic model [3], [6]. This method has the advantage of being simple to implement and can take into account both the trend and seasonality. Given the probabilistic aspect of this model, it is robust to noise and provides a single model to extract the three components. Two scenarios of decomposition by HW model are possible: additive and multiplicative. As we have no information *a priori*, we consider the two scenarios using a sliding window throughout the series. Figure 7 shows an example of decomposition. In the first part, we have the observed series of a job board (top), followed by trend (2nd), the seasonal component (3rd) and finally the noise fluctuations. This is so important because if we have a job board where the trend is decreasing hence the probability of recommending offers in it will be weak.

B. The Global Architecture of the System

The proposed recommendation system of job offers in the job boards is described in Figure 1. It is based on two main phases namely learning predictive model step and the recommendation step. During the first phase (left part of the figure), we used the previously described vectorial model and SVM classifier to annotate each job offer with the French job offers semantic vocabulary (ROME code). After that, we performed a clustering process, to regroup similar job offers on the basis of their shared semantic annotations. The clustering step is based on GLA (Generalized Lloyd Algorithm) that generates a Voronoi tessellation on the input data, by separating the job offers into convex regions (see Step 1 in Figure 1). These regions represent the classes $\zeta^{posting}$ (or clusters) of job offers (postings) obtained with GLA. Since we have more than 1 Million offers, we implemented a *Hadoop MapReduce* version of the clustering algorithm on R using *rmr2* and *rhdfs* packages, on *Hadoop 2.5 CloudEra* version. Results of the produced Voronoi tessellation with GLA clustering algorithm are displayed in figure 2.

Similarly on the same DB, we have several thousands of job boards. For each Job board and for each class of postings $\zeta^{posting}$, we construct a time series vector (see step 2 in Figure 1). Hence a time series will represent the temporal behavior of a job board by considering only job offers belonging to a similar semantic class $\zeta^{posting}$. Then, for each time series, a regression model is learned using

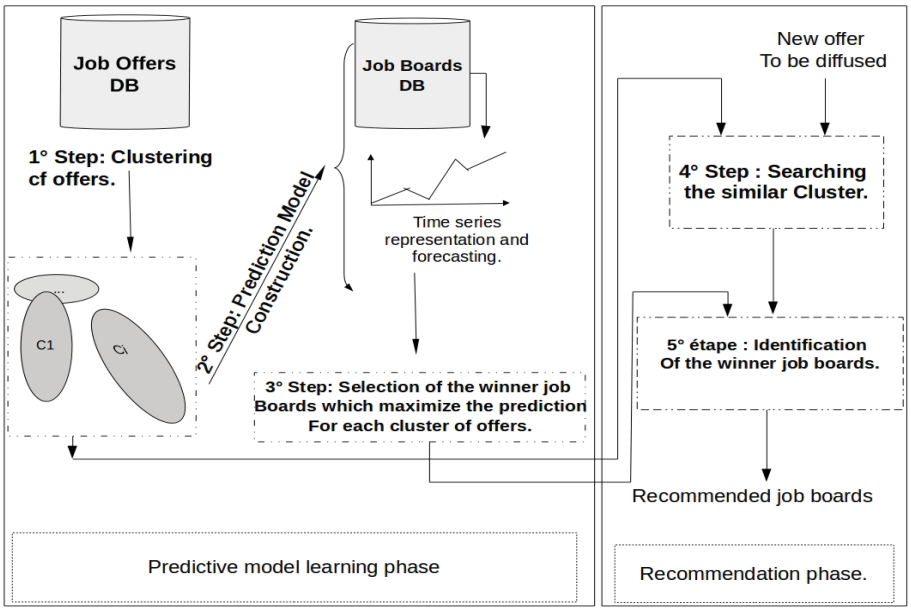


Fig. 1. Global architecture of the recommendation system.

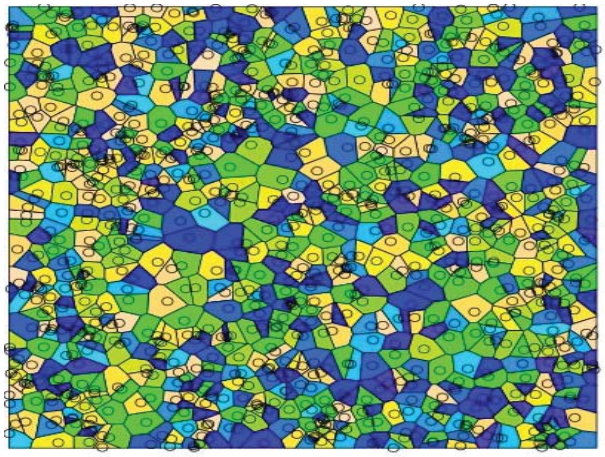


Fig. 2. Voironoi tessellation with GLA clustering algorithm.

Holt Winters probabilistic model, and future values of clicks ratios are predicted within an horizon h. After that, the job board(s) maximizing the different predicted ratio values are considered the most appropriate for the dissemination of the offers belonging to the considered class (see Step 3 in Figure). A hash table is created, containing key / values as class of offers, winner job boards.

In the second phase (the online stage), we seek to identify the job boards adapted to receive the diffusion of a new incoming job offer. The recommendation comes in two stages. Firstly, the system seeks to identify the nearest pre-generated class of offers compared to this new offer in term of its similarity with the centroid of the class (see Step 4 in the Figure). Since we already have an association for each class, the corresponding winner job board, thus we recommend the diffusion of this new offer on this job board(the fifth step on the figure).

C. The prediction algorithm

The Algorithm 1 illustrates the great steps described previously of our recommendation system. As we have shown in the previous section, this algorithm requires a list of jobs classes, regrouped by supervised manner (classification) or not (clustering). For a cluster of offers, the algorithm generates for each job board portal, a representation in time series of ratios.

IV. Experimental Results

A. Description of the dataset

To evaluate the proposed model, we used a big database of job offers and job boards. This complex DB, large and with heterogeneous information has been provided by an industrial partner (Multi Posting) in Sonar Project (http://sonar-project.com/). By considering confidentiality issues we can not present the architecture of the Data Base. However, we can attest that it represents more than a six-year follow-up containing about ten thousand of job boards and millions of job offers.

B. Results discussion

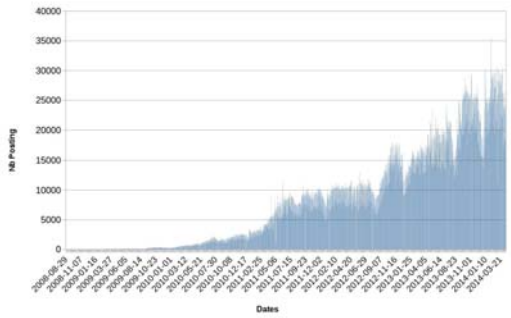
1) Quantitative analysis of the dataset: In order to quantitatively explore our data set, we firstly analyzed the variation of diffusing job offers in the different job boards, at different times. Figures 3.(a) and 3.(b) respectively show the distributions of job vacancies advertised and consulted on channels with different dates. For example, in Figure 3.(a), we can see that between 2008 and early 2011, the amount of vacancies advertised in the job boards was low. By cons, in the interval 2011-2012, offers are constant and distribution deals are around 10000 offers per day (to a maximum of about 15,000 offers). Between 2013 and early 2014, the increase is of the

Algorithm 1 Clicks forecasting in each job board.

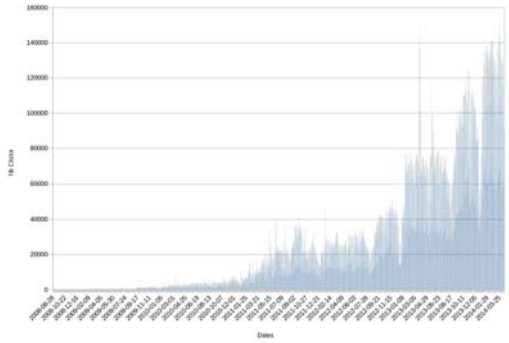
Require: A cluster C_{off} of similar offers, a list of job boards JB, and an horizon value h.

Ensure: The appropriate job board which maximizes the predicted value of clicks ratio.

Begin
$Maxclic = 0$
By considering all the offers $posting_j$ in C_{off}
for Each job board JB_i in DB **do**
  for each Instant $t \in \Delta_t$ **do**
    Calculate the ratio: $x(t) = \frac{|clics|}{|C_{off}|}$.
  end for
  Construct time series $X(JB_i)$=$\{x(t)|t \in \Delta_t\}$.
  Apply moving average filter on $X(JB_i)$ to reduce noises.
  Learn Holt-Winters model on $X(JB_i)$.
  Calculate $forecast(JB_i)_h$ to estimate future values of clicks in an horizon h.
  if $Maxclic \leq forecast(JB_i)_h$ **then**
    $Maxclic = forecast(JB_i)_h$
  end if
end for
return the winner JB_i having $Maxclic$.
End



(a) Job offers distribution.



(b) Clicks on job offers distribution.

Fig. 3. (a) : Distribution of the diffusion of Job offers on different job boards of our DB at different dates. (b) : Distribution of the number of clicks of the offers on job boards at different dates.

order of 25,000 to 30,000 offers per day with two points of visible change between late 2012 and early 2013. Scattering peaks are visible in the months of January, June or September which could be interpreted by the fact that certain offers are Seasonal. Referring to Figure 3.(b), we can see that there is a correlation between the number of diffusion of offers in Figure ref dist2.(a) and the number of clicks in the second figure 3.(b). The number of clicks is constant for a number of offers between July 2011 and November 2012. We also note that from November 2012 to August 2013, a sudden increase in clicks is notable for a constant number of offers. Finally, from August 2013, we can observe an extreme increase in clicks of job offers (average 120,000 hits per day). Slopes changes corroborate these observations.

2) Quantitative analysis using statistical indicators: Qualitative analysis is to explore, through statistical indicators, the properties of a series representing a job board. Figure 4 illustrates the variation of the ratio as a time series of a job board in our DB. Peak values can be seen in 2008, early 2010 and late 2012. Visually analyze a series seems difficult; for this reason, it is useful to use descriptive statistical tools to extract hidden information. For example, Figures 5 and 6 represent variations of calculating co-variances and correlations between the observed time series on different neighborhoods. We can see that for smoothing sliding windows (*Lag*) ranging from 1 to 10, the auto-correlation indexes are positive and close to 1.

3) Analysis of trends and seasonality: The information generated in the time series can look very noisy. This is due to random fluctuations intrinsic to the measures. To remedy this, and to visualize potential trends in the series, we used a moving average filter with different neighborhood sizes. After filtering the input time series with moving averages, it is now possible to build a predictive model. We decided initially to use the method of Holt-Winters [3], [6]. Figure 8 shows the result of Holt-Winters prediction; in Black, the original series and dark, we have a prediction of the number of clicks in a five-day horizon. We can therefore see the execution of our algorithm which, for a set of job board, calculates the prediction with the exponential model, and offers the portal that maximizes the prediction in terms of number of clicks.

4) Evaluating the predictive model: To evaluate the performances of Algorithm 1, we followed a test protocol that is to cut each time series of a job board channel into two parts. The first part of the series is used to create the regressive model with Holt-Winters. Then, we use the second part of the time series to compare the predicted values with the rest of real values. The difference is calculated using the mean square error. In Table IV-B.4 we illustrate the results obtained for a set of 22 job boards only due to space limit. We can observe the number of seasonality obtained with the additive and multiplicative decomposition, the trend and the prediction errors. A total of 11 JBs have an upward

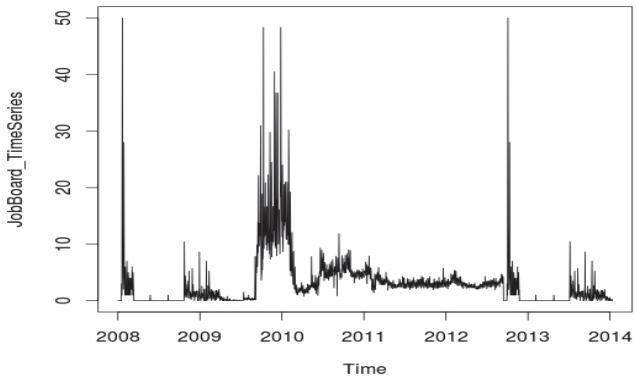


Fig. 4. Time series representation of a job board where clicks ratios values are calculated between 2008 and 2014.

Fig. 6. Correlogram of the same series with different lag values.

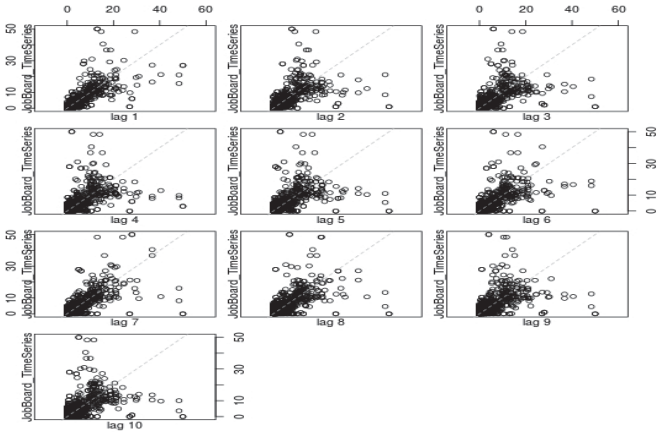


Fig. 5. Auto-correlation Analysis between observed values of the series on different neighborhoods (Lag 1 to 10).

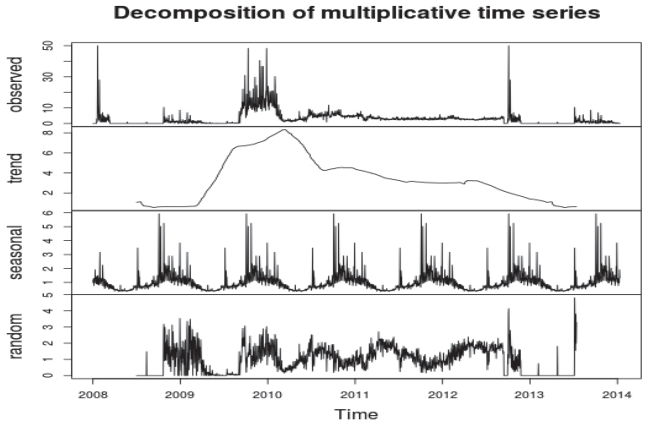


Fig. 7. Example of an additive decomposition of the original time series data with it seasonality, trends and residues.

trend (U), and we can observe that the number of additives and multiplicative seasonality are almost similar with higher values during each year. It means that these JBs not only have increasing clicks values, but also offers high diffusion frequencies in the year by their attractiveness. In addition, the prediction error in these job boards remains on average low, and thus they are highly recommendable to disseminate job offers. We can also see that there are 8 job boards with downward trends (D). These channels have a variable number of the two kind of seasonality models, averaging 5 to 6. For stationary job boards, there are 3 JBs with low seasonality.

V. Conclusion and perspectives

In this paper, we have presented a recommendation system based on semantic knowledge and temporal representation by time series. The objective is to diffuse a job offer to one or more adequate job boards. The system is based on two main stages namely learning the predictive model step and the recommendation one. We have shown the need of taking into account the seasonality for finer predictive studies, particularly in the context of the diffusion of job offers. We have integrated the probabilistic model of Holt-Winters to decompose the time series in order to identify trends, seasonality and possibly the residual noises.

Recommendation phase takes place in two stages: the identification of the most similar class of job offers regarding a new offer; then, the recommendation of job boards that maximize the ratio of the prediction of the clicks.

We presented some results of our experiments that revealed potential interesting job boards for certain class of job offers. The perspectives of this work will concern mainly taking into account other domain knowledge. On the other hand, the success of social networks has also contributed significantly to the evolution of the recruitment market on the Internet. Indeed, personal information posted by users of a social network such as LinkedIn can identify with more or less precisely their profile (academic curriculum, professional background, passion, etc). The integration of this information into the process of recommendation could refine the relevance of the proposed job board.

Processing numerical time series could represent a hard task and has the inconvenient of manipulating complex and

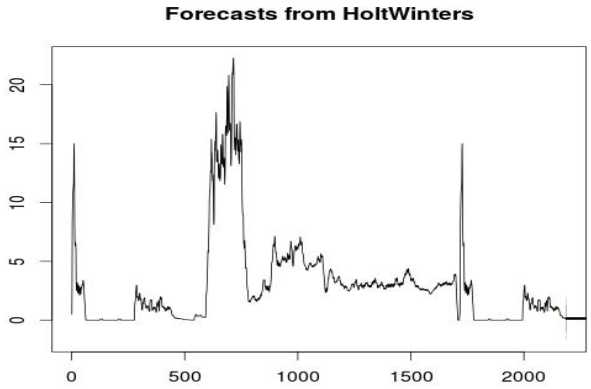


Fig. 8. Time series of the job board represented in Figure 4 with an exponential model generated by the Holt-Winters approach.

possible noisy data. We would like to use symbolic time series representation to avoid such problems and to transform the time series into symbolic sequences. Thereafter it will be possible to use symbolic data mining methods such as motifs discovery or similarity search [2]. We want also, in this context, to use other prediction algorithms on the the symbolic sequences such as Markov Models.

ACKNOWLEDGMENT

This work was supported by French Government and *Ile de France region* under a grant for FUI SONAR Project for automatic recruitment tasks. The authors would like to thank Multi Posting for data sharing.

Job board Id	Nb Seasonality (Additive model)	Nb Seasonality (Multiplicative model)	Trend	Error of prediction
1	5	5	D	0.1
4570	9	9	S	0.0022
4962	36	36	S	0.4
4630	15	16	U	0.016
6922	20	24	U	0.26
6605	10	10	U	0.4
919	5	5	U	0.1
4499	10	11	U	0.1
757	5	5	S	0.03
4497	5	5	D	0.16
5638	5	5	D	0.1
139	5	5	U	0.041
434	5	5	U	0.2
446	6	5	D	0.3
9	6	5	U	0.08
2769	6	6	D	0.1
32	6	6	U	0.031
16	6	6	U	0.11
2843	6	6	U	0.044
14	5	5	D	0.07
5	5	5	D	0.16
8	12	11	U	0.05
7	6	6	D	0.7
6	6	6	D	0.02
18	6	7	D	0.22
2843	6	6	S	0.2
447	11	11	U	0.007
25	6	6	U	0.48
11	6	7	U	0.8
12	12	12	D	0.12
13	6	6	S	0.18
15	6	6	D	0.043
17	7	7	S	0.05
47	6	6	S	0.2
125	6	7	U	0.03
136	6	6	S	0.1
166	7	7	S	0.046
169	5	6	S	0.083
174	6	6	U	0.026
263	20	23	S	0.1
272	7	7	S	0.05
505	18	18	S	0.1
422	6	6	S	0.25
594	6	6	S	0.01
667	5	5	S	0.01
675	5	5	S	0.1
680	12	12	S	0.04
30	6	6	U	0
798	6	6	U	0.001
815	6	6	S	0.65
1375	12	12	S	0.02
773	6	6	D	0.2

TABLE I

EXAMPLE OF A SOME JOB BOARDS AVAILABLE IN THE DB. FOR EACH RANDOMLY SELECTED JOB BOARD, WE HAVE THE NUMBER OF SEASONALITY OBTAINED BY ADDITIVE AND MULTIPLICATIVE DECOMPOSITION, AND THE TREND (U: UPWARD, D: DOWNWARD, S: STATIONARY), AND THE PREDICTION ERROR OF THE PREDICTIVE MODEL FOR TIME SERIES.

REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, June 2005.

[2] Sidahmed Benabderrahmane, Rene Quiniou, and Thomas Guyet. Evaluating distance measures and times series clustering for temporal patterns retrieval. In James Joshi, Elisa Bertino, Bhavani M. Thuraisingham, and Ling Liu, editors, *Proceedings of the 15th IEEE International Conference on Information Reuse and Integration, IRI 2014, Redwood City, CA, USA, August 13-15, 2014*, pages 434–441. IEEE, 2014.

[3] Chris Chatfield and Mohammad Yar. Holt-winters forecasting: Some practical issues. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 37(2):pp. 129–140, 1988.

[4] Tak chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164 – 181, 2011.

[5] Mamadou Diaby and Emmanuel Viennet. Développement d'une application de recommandation d'offres demploi aux utilisateurs de facebook et linkedin. In *Atelier Fouille de Données Complexes de la 14e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'14)*, Rennes, jan 2014.

[6] Sarah Gelper, Roland Fried, and Christophe Croux. Robust forecasting with exponential and Holt-Winters smoothing. *Journal of Forecasting*, 29(3):285–300, 2010.

[7] Joseph S. Kong, Kyle Teague, and Justin Kessler. The love-hate square counting method for recommender systems. In Gideon Dror, Yehuda Koren, and Markus Weimer, editors, *KDD Cup*, volume 18 of *JMLR Proceedings*, pages 249–261. JMLR.org, 2012.

[8] V. Radevski, Z. Dika, and F. Trichet. Common: A framework for developing knowledge-based systems dedicated to competency-based management. In *Information Technology Interfaces, 2006. 28th International Conference on*, pages 419–424, 2006.

[9] Ian Ruthven and Mounia Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, June 2003.

[10] J. Sgula. *Fouille de donnes textuelles et systmes de recommandation appliqus aux offres d'emploi diffuses sur le web*. PhD thesis, CEDRIC Laboratory, Paris, France, 2012. Directeur: Gilbert Saporta Rapporteur 1: Ludovic Lebart Rapporteur 2: Emmanuel Viennet Examinateur 1: Yves Lechevallier Examinateur 2: Thierry Artieres Examinateur 3: Michel Crucianu Examinateur 4: Yannick Fondeur Examinateur 5: Stphane Le Viet .

[11] Robert A. Yaffee and Monnie McGee. *Introduction to Time Series Analysis and Forecasting: With Applications of SAS and SPSS*. Academic Press, Inc., Orlando, FL, USA, 1st edition, 2000.

Using Text Mining of Amazon Reviews to Explore User-Defined Product Highlights and Issues

L. Jack and Y.D. Tsai
Intel Corporation, Santa Clara, CA, USA

Abstract- *Advances in technology have made user-generated content ubiquitous. This includes user reviews of products which are publicly available on the internet and has led to an increase in the use of text mining to analyze consumer behavior. This paper presents a framework for using text mining to gather customer feedback. Text mining techniques are used to aggregate the top attributes associated with groups of devices, laptops and tablets, as well as individual devices. A case study comparison of three devices compares and contrasts positive and negative aspects mentioned by the users, which is useful to improve future generations of products. Manufacturers can incorporate and review product attributes when a product is launched and over time correct product issues, understand customer requirements, and maintain customer satisfaction.*

Keywords: text mining, customer, reviews, R

1. Introduction

User reviews on E-commerce websites like Amazon.com have a large influence on product reputation as they are heavily viewed by prospective buyers before they decide to make purchases. Text mining tools and algorithms can help uncover customer attitudes and sentiments on products they have purchased and used. This paper reviews a method of applying text mining techniques to compare and highlight top customer opinions of a product (in this case, laptops and tablets) as a means to provide feedback to enhance future products. Understanding the overall positive and negative perceptions of a product enables manufacturers to be in tune with the reception of their products. It also enables them to identify, fix, and resolve issues uncovered in user reviews.

Flanagin and colleagues [1] found that product ratings are used as a barometer of product quality, where higher perceived quality is associated with greater purchasing intentions. User ratings are considered a credible source of information about products consumers are intending to buy though users may only attend to average product ratings when making purchasing decisions. When a product reached a certain level (4.4 stars), a ceiling effect was found and ratings above the level did not result in perceptions of enhanced product quality. Potential buyers may only look at top level information to make purchasing judgments, however, the review text gives insight as to what contributes to their overall rating.

Several studies on review helpfulness suggest that extreme reviews are the most helpful. Chen and Tseng [2] found that high-quality reviews are those that subjectively comment on several product features. There is greater ambiguity in positive product assessments than in negative product assessments when comparing extreme reviews (4 and 5-star reviews vs. 1 and 2-star reviews). Mudambi, Schuff, and Zhang [3] call attention to the text of the review to get accurate details on the user's view of product quality. They compared the rating differences between feature-based goods and experience goods. Users interested in buying feature-based goods, such as music players, prefer reviews that outline pros and cons of the product and contain mainly objective inform mation with only few subjective statements on the product. In contrast, users rely on personalized, sentimental reviews, not captured by the product description, for experience goods like a movie DVD [5]. Review characteristics such as subjectivity and readability were also key feature categories that determine review helpfulness to the potential buyer [6]. These studies emphasize the need to not only pay attention to the summary characteristics of the product reviews, but also the detailed nuances of a user's likes and dislikes of the product.

2. Research Questions

There are many possible ways to collect feedback from users about products they have purchased and used. One typical way to do so is to ask users to complete surveys. Another is to do experimental research or an observational study of users interacting with devices. In this study, text analyses were completed using Amazon review data. Amazon reviews are considered a good source of data for capturing consumer perceptions primarily because of the large number of data points. In addition, customers are able to post their reviews after they have used the product and know its pros and cons, and while they are in a low-to-no-pressure setting, usually sharing their thoughts and feelings at home versus being in a timed lab setting.

These research and analyses were conducted as part of a larger project to identify, understand, and evaluate the basics of system performance, specifically in tablets and laptops. Essentially, the research question was, what basic features are most important to users of the product and most

influential in molding their perceptions? Similar to collecting feedback, there are several ways one could go about assessing "importance." Some possible ways are to set a threshold for how many people share the same perception or what percentage have similar views. The determinant of importance in this research was frequency; the topics that came up the most were prioritized and used as a standard to assess how important all other opinions of a product were when considering the group of users as a whole.

The end goal of the project was to have assessments of both individual products (e.g. a specific laptop) and groups of products (e.g. laptops in general) that provided information about what features were important to excel at or improve upon to create a better customer experience with the product. Lastly, these assessments would become recommendations to the manufacturers for future product development and improvement. The focus of this paper is to detail the methodology and results of said investigation of Amazon reviews.

3. Methods

This analysis of Amazon product reviews focused on evaluating a variety of user reviews of tablet and laptops. "Two-in-one" devices were not included because they straddle between the two device types and could make category comparisons more difficult. To cover an assortment of products, 40 devices were chosen; 20 laptops and 20 tablets. The devices were chosen to create a diverse set of products, to ensure variety in operating systems (OS), price, brand, popularity, and to be representative of what was available and purchased by consumers on the (Amazon) marketplace.

Amazon review data was web scraped and text mined using R, a statistical software. R is an open-source programming language commonly used for statistical computing. R has both data mining (web scraping) and data analyses (statistical and text analysis) capabilities and the analyses are scripted, customizable, and repeatable. An R script was developed in this research to pull Amazon reviews of the devices of interest. In total, the number of reviews collected across all devices was 19,080.

The number of reviews per device ranged from 50 - 4,100. The prices ranged from $46.99 - $2,249.99. Table 1 (a) lists the price and number of reviews associated to the laptops; Table 1 (b) lists the tablets used in the study. Thirty-seven out of the 40 products had over 100 reviews. Using calculations based on power analysis, 100 reviews was estimated to be a large enough sample size to begin finding significant relationships [7] [8]. The three products with less than 100 reviews were chosen because of their low ratings. This was a potential confound as consumers do not often choose to buy poorly rated devices. Therefore, less people purchased these devices and they had less reviews,

TABLE I (a)

LAPTOP PRICES AND NUMBER OF REVIEWS

Laptop ID	Product Price	Number of Reviews
A	249.99	230
B	199.99	2,470
C	249	280
P	1779	130
D	247	550
V	265	100
F	299	720
U	736.59	100
G	283.38	390
H	387.99	370
J	1139.99	240
K	1229	150
S	1299	90
T	2249.99	50
Q	1099.99	100
E	378	110
R	588.96	50
L	439	220
M	294	140
N	159.94	110

TABLE I (b)

TABLET PRICES AND NUMBER OF REVIEWS

Tablet ID	Product Price	Number of Reviews
UU	92.99	140
ZZ	47.99	570
EE	169	100
FF	49.95	270
GG	149.99	550
HH	51.99	640
JJ	89.5	100
LL	353	2140
XX	559.99	1770
WW	309	260
MM	379	160
NN	306.83	170
PP	374.95	250
QQ	79.95	260
RR	198.99	4100
YY	299	190
SS	46.99	120
TT	299	380
VV	124	170
KK	59.95	140

which results in a skew of the total product ratings toward higher scores. Also, the character counts of lowly rated reviews of 1 and 2 was significantly higher (Median = 251.5, SD = 699.16) than those of highly rated reviews (Median = 163 SD = 786.07), $p < 0.05$. All of these reviews were included in the analyses, however, as they aid in understanding what elements of the product caused the users to rate the device well or poorly.

4. Analyses

The research was primarily focused on understanding what was really important to users, what positively or negatively affected product reviews, and what specifically users choose as highlights or pain points when reviewing laptop and tablets. This then, in terms of ratings, translated into which reviews were the best and the worst. The analyses focused on subsets of the data: the lowest ratings (ratings of 1 and 2) and the highest ratings (ratings of 5). Ratings of 1s and 2s were grouped together because the positive ratings outweighed the negative ones. As mentioned previously, products that are highly rated (and therefore have a good reputation), have more positive reviews than negative ones as potential buyers are unlikely to purchase a product that many people rated badly and has a poor reputation. Therefore, in order to get a large enough sample of lowly rated reviews, both 1 and 2 rated products were combined into a single category. In the end, there were 11,730 5-rated reviews and 1,678 1- or 2-rated reviews (Figure 1).

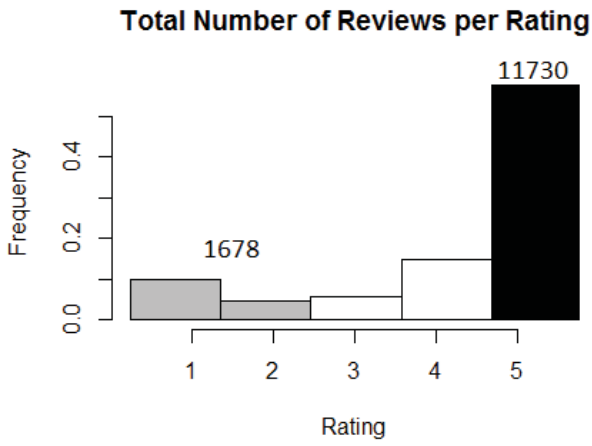


Fig 1. Histogram of the total number of reviews by rating

After collecting all the data, common text mining techniques [4] [5] were performed to prepare the data for analysis. Another R script was developed to analyze and understand patterns of discussion across groups of reviews. The scripts developed for the following analyses used the R libraries plyr, tm, RStem, stringr, ggplot2, and xml. The reviews were treated as documents and were aggregated into larger corpuses depending on what subset of data was being analyzed - all reviews, laptops, tablets, reviews from a single device, and review groupings that mentioned common key terms (for example, "battery life").

Regular expressions were used to break the data into desired groups based on a word or words with similar meanings. Examples include words about the screen, touch, and resolution, and adjectives that describe the goodness of a product and its features - great, good, excellent, amazing, etc. Some of the words and general product features that were examined included battery life, display, touchscreen, touchpad, keyboard, and price. These product features were defined from previous work on the overall project which aggregated common attributes mentioned from technical press reviews. The key terms were also used to parse out specific sentences for further analyses.

After being turned into a corpus, punctuation and extra symbols were removed. Words were converted to lowercase, and stopwords were removed. Plots were created to ascertain correlations between numeric variables and significance testing was conducted on the proportions of times words and phrases appeared in different groups to see if there was a real difference in how many times one group mentioned a topic compared to another. Lastly, n grams were performed, which is a text mining technique for assessing the frequency of words and phrases (with n being the length of the phrase) in a corpus. Lastly, the three devices with the most reviews are presented in this paper as a case study for individual analysis.

5. Results

5.1 Overall user feelings

Assessing all 19,080 reviews across the 40 devices in aggregate, the most common aspects that were addressed overall by users were the: 1) battery life, 2) (touch) screen, 3) value/price and 4) generally (positive) feelings about the device. In reviews rated 1 or 2, reviewers used negative adjectives like bad, poor, worse, worst, and horrible most often when they were describing the quality of the product or the customer service provided. In reviews rated as 5 stars, reviewers used the positive adjectives "nice", "best", "good", "great", "perfect", "excellent", and "amazing" most often when they were describing the battery life, screen, and keyboard. These positive adjectives were also often associated with describing the sound (audio quality), value, and physical traits of the device, but to a lesser extent. Positive emotive words such as like, love, and happy were most often used by reviewers to describe a general love for the device and to lesser extent, to describe the screen and backlit keyboard. In an interesting contrast, the words "like" or "love" were some of the most common words whereas the word "hate" rarely if ever appeared, even in the reviews of poorly rated devices.

5.2 Frequent words and phrases

When analyzing the reviews, some words and phrases were more popular and mentioned more frequently than others. These words and phrases will be addressed in general and a few have been selected to be addressed in detail to provide more context of the analyses conducted.

Reviewers often mentioned the purpose of buying the product and wrote about key features that were surprising when encountered. Only a small percentage of reviewers (3.2%, 609 reviewers) mentioned that a product was "bought for" someone else, not for themselves. For example, this could be a parent purchasing the device for their children. Thus, the majority of reviewers have personally used their devices and are capable of talking about its specific features, their likes, and dislikes. In addition, users revealed that their "time" is valuable to them by negatively reviewing the length of time it took to accomplish tasks for setup or troubleshooting a new device. Some key words also highlighted specific features that were important to customers. For example, SD card slots were an unexpected feature that customers seemed to appreciate and miss when not present. The power button differentially appeared in low versus high ratings for two reasons: 1) In low ratings, the power button received negative reviews for being broken, and 2) In high ratings, reviewers mostly disliked the placement of power and volume buttons. Lastly, one operating system (OS) present in several devices received many mentions for varied reasons. For example, reviewers voiced frustration with the difficulty of interacting with their device due to the software, the incompatibility of the software version loaded on their device with another similarsoftware version, and the limited content of the app store. On a positive note, for those who mentioned this OS, the ability to multi-task on tablet devices was also mentioned in 10.3% of the user reviews as an appreciated feature. Example text from two reviewers, one that gave a low rating and one that gave a high rating, is shown with some of the most common concepts bolded in Figures 2 &3.

1 star rating

"I sent two of these back ...one decided to work off and on, and the other ones ***screen cracked*** *without any impact involved. I don't know if anyone else noticed, but there is a paper that comes in the box that basically says call the original seller, not Amazon. Fortunately I did not see it. I am only dealing with Amazon. Now, I am going to go ahead and* ***spend the extra money*** *to get [another device]... Do I want to* ***pay that much*** *for these, no, but I know what [it] does, and how much more sturdy they are. If I had one way to describe these ..., it would be flimsy, with poor technology. By the way, they take pictures, and they are poor quality. The* ***battery life is too short****... I guess I have learned, once again,* ***you get what you pay for****, with the exception of items you know are good, and only get a good sale."*

Fig. 2 Example low rating review

4 star rating

"I bought this device mainly because the ***retail price was very low*** *compared to competitors and the device looked attractive and had a* ***nice feel****... if you don't always have an internet connection and use spreadsheets a lot then this device and OS is not for you. ..... that I saw was missing here was the speech input.... The build quality of this machine is one of the* ***best I've seen at this price point****! I find it* ***visually appealing*** *.... The soft touch material being used for the outer body of this laptop* ***feels nice to the hands****... This device is also* ***very thin****. The device also has a nice selection of ports providing USB 2.0, USB 3.0, full size HDMI,* ***SD card reader****, headphone/microphone combo jack, lock slot and an indicator light for sleep/use.:... but for $300 I guess you can't expect to get a* ***1080p IPS display****...Some* ***back-lit keys would be a nice*** *addition in future models ... the trackpad had a silky smooth texture to it ...I* ***love*** *the addition of trackpad gestures ...So far from my testing the performance has been great.... I was able to get 4 days of use out of it before having to plug it in and combined use time was 7 hours. 1.75 hours of this was watching videos so* ***battery life*** *could have been better if just used for web browsing and document editing... I'm confident 9.5 hours is an accurate claim.... this machine should be able to* ***last close to 6 hours*** *for viewing back-to-back movies ....****Charging time*** *from 5% took exactly 2 hours to fully charge back to 100% which is great...Overall I am pleased with the device but* ***$300 might still be a little high*** *for something that only serves as an internet browsing machine... It simply doesn't outperform ... has more offline functionality on-the-go and the battery life of this machine comes up a little short....should allow add-ons or upgrades on their website for a* ***backlit keyboard****..."*

Fig. 3 Example high rating review

5.3 Battery life

Battery life was mentioned most often across all devices, in 2,665 1-5 star reviews, across device types (laptop, tablet) and individual devices. Additionally, reviewers mentioned battery life statistically more often when they were giving high ratings than low ratings (162 out of 1,678 1-2 star reviews, 9.6%; 1,539 out of 11,730 5 star reviews, 13%, $p = 0.0$). The reviews mentioning battery life were mined to further understand the context of why battery life was so prominent and to analyze battery life expectations in hours. For laptops, reviewers considered four hours of battery life with moderate use satisfactory, around five hours to be good or standard, and above eight hours to be noteworthy. A further exploration was conducted of highly rated reviews that explicitly mentioned both "hours" and "battery [life]". Seven percent of those reviewers said that the battery lasted around five hours, and gave the product a rating of 5 (the other 93% did not have a unified voice about battery life duration). As for tablets usage, many users spoke of watching TV and video streaming from their device. Reviewers assumed that tablets had a longer battery life compared to laptops. Over seven hours of battery life was considered satisfactory for a tablet, and 10 or more hours was considered excellent. In addition, how long the device took to charge seemed just as important as how long the charge lasted; this was especially true for tablets.

5.4 Screens

For reviewers who gave low ratings, tablet (touch) screens were most often mentioned in reviews, 21% of the time. Some reasons users cited were responsiveness issues with the touchscreen or the screen being cracked or broken (sometimes found upon unboxing the device). Example responsiveness issues include slow or lagging performance and inaccurate registration of touched screen locations. For laptop screens, some were said to have resolution mismatch issues, which meant that the capability of the device to display at a higher resolution made viewing lower resolution content that was not able to scale and adapt to a higher resolution unfavorable. In these cases, the product feature detracted from the overall experience of the product.

5.5 Case study comparison between individual devices

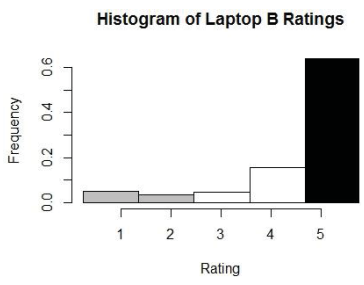
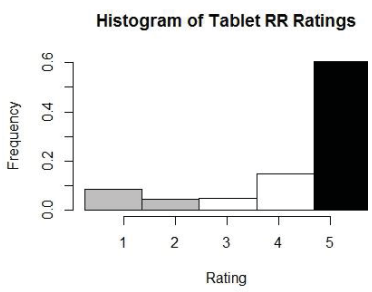
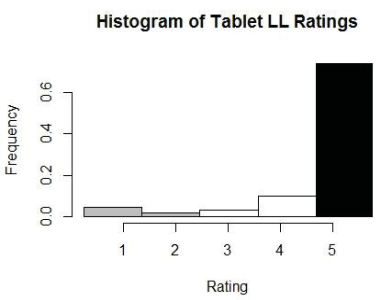
This analysis delved into the three devices with the most reviews: Laptop B, Tablet RR, and Tablet LL (Table 2). All product names and some specifications have been changed to protect brand privacy. For each of the three devices, analyses were done to find the positive aspects and negative aspects of the device. For example, one of the negative aspects mentioned about Tablet LL was a general warning from past customers to not buy the product from a well-known website. Delving more deeply into the reviews, surfaced the reason: customers experienced differences in device quality and customer returns when interacting with the original supplier of the device compared to a third party vendor; discouraging reviews were posted as a result. The comparison also highlighted interesting trends such as, people use "love" twice as often in reviews of Tablet LL than Tablet RR.

In addition to the results listed in the table, some data mining was done to ascertain user defined problems with the devices. Specifically, reviews of rating 1 or 2 that mentioned the words "issue" or "problem" were scrutinized. The results from this exercise were then compared to published articles (e.g. technical press articles) on device issues for verification of issues data mined from the reviews. When Tablet RR, for example, was compared to the top web search results for issues pertaining to that device, all items that arose from data mining (battery charging, freezes, random reboots, touchscreen responsiveness, and Wi-Fi connectivity) were addressed and verified as indeed being widespread problems.

6. Discussion

There are some key takeaways from this research. First, if one can create a feeling of love for a product, that could improve product ratings. The Amazon review data showed that people use "love" very often when describing their product interaction but rarely use "hate" when talking about the product, even for devices that received a poor rating. In addition, love was used to describe a general feeling toward

TABLE II
EXAMINING THE THREE DEVICES WITH THE MOST REVIEWS

Device ID, number of reviews	**Laptop B** 2470	**Tablet RR** 4100	**Tablet LL** 2140
Distribution			
Popular topics in reviews rated 5	•Specific preloaded software and operating system •Everything they need •Battery life •Love it •Fast	•Comparing it to a similar tablet •App availability •Screen •Price •Fast	•Would recommend it •App availability •Easy to use •Love it •Price
Popular topics in reviews rated 1 & 2	•Reboots and deaths •Keyboard and trackpad •Operating system and its applications •Printer & Wi-Fi connectivity	•Touchscreen •Battery life •Customer service	•Do not buy it from a specific dealer •Wi-Fi Connectivity

the device, not any particular aspects of it, but the device holistically. It is possible that customers are very susceptible to product branding, which indicates that product marketers stand to benefit from understanding why people love a product. Other studies have examined this in detail and have cited physical attributes (e.g. the object is beautiful, ultimate) and significance in personal value have been correlated with users having engrossing or transcending experiences with the object that leads to love [10]. The association of love for the product may be a key barometer of its success.

Second, users definitely take note of battery life. They have expectations around how good the battery needs to be and the data shows it is the most talked about topic in both positive and negative reviews. For tablets, users expected battery life to last over seven hours, for laptops, over 5 hours. Improving battery life can improve product ratings of average products to stand apart from other similar devices, but not when there are other glaring issues with the device. Thus, battery life is a feature that can be improved to get a better rating, but is not the only contributing feature for devices that have high ratings. Battery issues should not be overlooked as they directly affect how long users can use their devices.

Last but not least, manufactures need to make sure all devices are functional before shipping them out to customers. This includes making the touchscreen responsive and out of a good material. It is apparent that the level of quality control or functionality in some products needs to be addressed as many reviewers cited their devices breaking within the first few uses, or worse, already being broken upon arrival. Too many reviewers complained that their devices simply did not meet basic expectations, exhibiting issues such as being unable to turn the device on, connect to Wi-Fi, have working trackpads, have functioning power buttons, etc. Having a reliably functioning device or a means to quickly address these problems for the customer may help alleviate aggrieved customers with device issues. This should be an aim for all manufacturers.

6.1 Limitations & Next Steps

One limitation of this research is that the text mining was only conducted on Amazon reviews and it is possible that Amazon attracts a unique group of customers. By comparing the results of this research to a few technical review articles written by experts, the validity of the results was confirmed for one of the devices. However, there may still be some misjudgments across other devices. Future iterations of this research could include reviews from other databases and review websites. An advantage of this text mining methodology is that it could be repeatedly conducted to gain insight about how customers see products and how opinions may change over time with improvements to technology and to the products.

One major question that resurged repeatedly during this research was, "where does one draw the line?" When conducting surveys and questionnaires, or asking for opinions, deciding the point at which to start considering a person's opinion or problem as an overall issue or problem is quite difficult. For example, when deciding to take action on customer feedback, does one consider: What are their top 10 concerns and highlights? Is the same issue brought up by 10% or more of the customers? Do certain highlights or concerns appear more often than would be "expected"? How does one define how often to expect a word, phrase, or concept to appear? So far, the literature in this area is either lacking or difficult to find. In this research, the decision was made to delve more deeply into the top (defined as the most frequent) concerns and opinions of the reviewers. However, other ways of assessing where the line is are equally valid.

In future iterations of this type of research, deeper machine learning type techniques could be applied to the conduct a predictive analysis. For example, one could attempt to predict what a product rating would be solely based on a review. That exercise in and of itself would help tease out what is important to customers and why they may rate a product in a particular way.

7. Conclusion

This paper outlines a method to apply text mining to understand consumer feedback about purchased products. Any person or business can use this framework to quickly gain insights about what customers in their field are saying about their products and customize the methodology to fit their needs. Information about customer preferences, key features, and encountered issues can then be used to improve upon the product. When this method is applied to review one product, the top features that are important to the user can be gleaned, as well as the main problem areas of the product. When applied to review a group of products, comparisons can be made across product types, comparing the overall features of importance and from there, generalizing to determine what areas need refinement across the entire product group. This method can also be used in a cyclical manner, to keep track of changes in opinion and product specifications over time as new products emerge on the market. Tracking of this process could essentially be delivered as reports, directly from the consumers to those who need this information - designers, product developers, etc., and ultimately result in the delivery of a better product. This is a practical way to use crowdsourced data in the form of online reviews to inform a company on how customers think about and react to products and what is most important to them and urgent to fix; it is a method of feedback to manufacturers.

8. References

[1] Flanigin, A.J., Metzger, M.J., Pure, R., Markov, A. and E. Hartsell. "Mitigating risk in ecommerce transactions: perceptions of

information credibility and the role of user-generated ratings in product quality and purchase intention," Electronic Commercial Research, 14, 2014, pp. 1-23.

[2] Chen, C.C., and Y.-D. Tseng. "Quality evaluation of product reviews using an information quality framework", *Decision Support Systems*, 5, 2011, pp. 755-768.

[3] Mudambi, S.M., Schuff, D., and Z. Zhang. "Why aren't the stars aligned? An analysis of online review content and star ratings", In *Proceedings of 47th Hawaii International Conference on System Science*, 2014, pp. 3139-3147.

[4] Hu, M. and B. Liu, "Mining and summarizing customer reviews," *Proceedings of the tenth ACM SIGKDD International conference on Knowledge discovery and data mining*, Aug. 22–25, 2004, Seattle, Washington, pp. 168-177.

[5] Dellarocas, C. and R. Narayan, "What motivates consumers to review a product online? A study of the product-specific antecedents of online movie reviews," In: Aberer, K., Peng, Z., Rundensteiner, E.A., Zhang, Y., Li, X. (eds.) *WISE 2006*. LNCS, vol. 4255. Springer, Heidelberg (2006).

[6] Ghose, A. and P.G. Iperiotis. "Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics," *IEEE Transactions on Knowledge and Data Engineering*, 23(10), 2011.

[7] Mooney, R. J. and R. Bunescu, "Mining knowledge from text using information extraction," *ACM SIGKDD Explorations Newsletter - Natural language processing and text mining*, 7(1), June 2005, pp. 3-10.

[8] Suresh, K. P. and S. Chandrashekara, "Sample size estimation and power analysis for clinical research studies," *Journal of Human Reproductive Sciences*, Jan-Apr 2012, 5(1), pp. 7-14.

[9] J.S. Tanaka, "How big is big enough?": sample size and goodness of fit in structural equation models with latent variables, *Child Development*, 1987, 58, pp. 134-146.

[10] Ahuvia, " For the love of money: materialism and product love," in *SV - Meaning, Measure, and Morality of Materialism*, eds. Floyd W. Rudmin and Marsha Richins, Provo, UT: Association for Consumer Research, 1992, pp. 188-198.

How Can We Measure the Similarity Between Résumés of Selected Candidates for a Job?

Luis Adrián Cabrera-Diego[1,2], **Barthélémy Durette**[2], **Matthieu Lafon**[2],
Juan-Manuel Torres-Moreno[1,3] **and Marc El-Bèze**[1]
[1]LIA, Université d'Avignon et des Pays de Vaucluse, Avignon, France
[2]Adoc Talent Management, Paris, France
[3]École Polytechnique de Montréal, Montréal, Canada

Abstract— *Several researches in Natural Language Processing (NLP) have developed e-Recruitment systems. Despite these researches, none of them has been interested in the way the similarity and distance measures, using different vector weights, behave when they have to determine the likeness of résumés. Therefore, in this paper we present a comparative analysis of 5 measures using different vector weights done over a large set of French résumés. The aim is to know how these measures behave and whether they validate the idea that selected résumés have more in common with themselves than with the rejected résumés. We make use of NLP techniques and ANOVAs to do the comparative analysis. The results show that the selection of measures and vector weights must not be considered negligible in e-Recruitment projects, specially in those where the résumés' likeness is measured. Otherwise, the results may not be reliable or with the expected performance.*

Keywords: e-Recruitment, résumé analysis, similarity measures, matching systems, data mining

1. Introduction

During the last 15 years, the massification of computers and the Internet have had an impact on the way humans search for jobs and employees [1], [2], [3]. Internet has become the main medium to select and recruit candidates [4]. The use of information and communication technologies to recruit and select candidates for a job offer is what has been called *e-Recruitment* [4], [5], [6].

E-Recruitment has brought several benefits to Human Resources Managers (HRM), employers and job seekers. Nowadays, employers reach larger audiences [7], [8], HRM reduce their operating costs [7] while job seekers can search easily a job offer [9]. Although the e-Recruitment has brought the aforementioned benefits, some undesirable consequences have also arisen for HRM: an important increment in the number of unqualified applications [10] and the recruiters' difficulty to manage correctly and rapidly the great amount of received data [11], [12].

Many researches, usually in Natural Language Processing (NLP), have developed systems in order to increase the performance of HRM. These systems can be classified in three types: systems that extract specific résumé[1] data [14], [15], [16], systems that extend the information of job offers and/or résumés [14], [16] and systems that try to find the best candidate(s) for a job offer using ontology matching [17], semantic similarity [6], [8], automatic learning [18], [19] or relevance feedback [3]. Nevertheless, even if the aim of these researches is to create tools to assist HRM, to our knowledge, none of them have analyzed the role that similarity and distance measures, with different vectors weights, play in the likeness calculation of résumés. Furthermore, these researches have been developed and tested mainly using small datasets.

For these reasons, we present in this paper a comparative analysis of 5 measures applied with at least 3 different vector weights. The aim of this paper is to determine experimentally how the likeness of résumés behaves using these measures and vector weights; the results may be of help to determine in the future the best methodology to create e-Recruitment tools. The analysis is done over a large set of French résumés organized by job offer and which has been used and annotated by expert recruiters. We make use of NLP techniques in order to preprocess the data (i.e. language and résumé detection), and of statistical tests of Analysis of Variance (ANOVA) to asses each hypothesis.

The structure of this paper is the following: first, in Section 2, we present the data used in this project and their preprocessing. Then, in Section 3, we describe the experimental method. Later, we present and discuss the results in Section 4 and Section 5, respectively. Finally, in Section 6, we present the conclusions and future work.

2. Data

The corpus used in this paper comes from a HRM firm and is a collection of résumés, motivation and recommendation letters, diplomas, interview minutes and social networks invitations (LinkedIn, Twitter, Facebook, among others). The corpus is organized by job positions, which in turn are divided into candidates. It is annotated with meta-data that

[1]In some researches and books it is possible to find the Latin locution *curriculum vitæ (CV)* instead of the term *résumé*. However, for [13] both expressions are synonyms. Therefore, in this paper we will use the résumé as common term.

allow us to determine, for each job applicant, its unique ID, the applied position, the receiving date of each application and the last recruitment phase reached by the applicant (*Analyzed*, *Contacted*, *Interviewed* or *Hired*). French is the main language in the corpus although it is possible to find documents in English, Spanish and German. Table 1 shows the number of files, job offers, and job applications in the corpus.

Table 1: Number of job offers, job applications and files in the corpus.

Job offers	Job applications	Files
296	29,368	47,388

The four recruitment phases were classified into two classes: *Selected* and *Rejected*. The first class, corresponding to the phases *Contacted*, *Interviewed* or *Hired*, represents the candidates that are approached by a recruiter. The second class, contains the candidates that are only *Analyzed* but not contacted after reading their résumé. Figure 1 presents the histogram of Selected candidates, measured by percentage, in the corpus; the median is 40.94% and the mean $42.93\% \pm 1.44$.

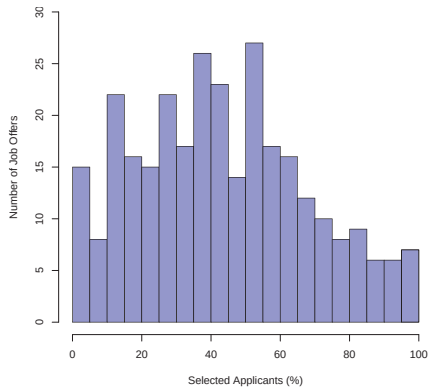


Fig. 1: Percentage of Selected candidates by the number of job offers.

2.1 Document Conversion and Language Recognition

Four types of documents are analyzed in this work: PDF (*.pdf*), Microsoft Word (*.doc* and *.docx*), OpenDocument Text (*.odt*) and Rich Format Text (*.rtf*). They represent 80.52% (38,161 files) of the corpus; the rest belongs mainly to HTML (social networks invitations) and image files. To be able to apply NLP techniques we first converted these files into plain UTF-8 text. For this we used:

- *Calibre Ebook Management*[2] for files having a *.pdf*, *.docx*, *.odt* or *.rtf* extension. The accentuated letters of PDF files are verified to know if they were correctly coded (see [20] for a discussion).
- *Catdoc*[3] is used only for files with *.doc* extension.

In order to detect only the French documents we used the Google's *Compact-Language-Detector (CLD2)*[4] through its Perl module[5]. The CLD2 is a tool that makes use of probabilities and 4-grams of letters to predict the language of documents[6]. In the corpus, 32,845 documents are in French (69.31% of the total corpus and 86.06% of the analyzed file formats).

2.2 Résumé Detection

To sort out the résumés from other types of documents, like motivation letters, interview minutes and publications lists, we developed a Résumé Detector based on a Support Vector Machine (SVM) [21] through LIBSVM [22].

2.2.1 Training

The training corpus was established through a manual classification of résumés (699) and other documents (635), all in French, from a collection of spontaneous applications[7]. We tested 2 different SVM kernels (linear and radial) following the procedure proposed by [23]. They were tuned up through a grid-search and a five-fold cross-validation.[8] Table 2 presents the results of the cross validation of the SVM using the best parameters for the linear and radial kernel. These results are presented in terms of precision, recall and F-score.

Table 2: Parameters, precision, recall and F-score for the linear and radial kernel.

Linear ($C = 0$)						
Subcorpora	L_1	L_2	L_3	L_4	L_5	Mean[9]
Precision	0.972	1.00	0.950	0.971	0.971	0.979
Recall	0.986	0.971	0.971	0.992	0.992	0.982
F-score	0.979	**0.985**	0.960	**0.982**	**0.978**	**0.977**
Radial ($C = 0$; $\gamma = 1 \times 10^{-4}$)						
Subcorpora	L_1	L_2	L_3	L_4	L_5	Mean[10]
Precision	0.979	0.963	0.964	0.951	0.932	0.952
Recall	0.986	0.949	0.971	0.985	0.992	0.977
F-score	**0.982**	0.956	**0.967**	0.968	0.961	0.964

[2]http://calibre-ebook.com/

[3]http://site.n.ml.org/info/catdoc/

[4]https://code.google.com/p/cld2/

[5]"Lingua::Identify::CLD" https://github.com/ambs/Lingua-Identify-CLD

[6]Documentation available at: https://code.google.com/p/cld2/wiki

[7]Job applications that are not related to any job offer and in consequence they do not belong to the experimental corpus.

[8]For the different models, all the documents from the training subcorpora passed through a basic preprocessing: stopwords suppression and stemming (Porter's Algorithm).

[9]Mean F-score is obtained from the average Precision and Recall.

[10]Idem.

As seen in Table 2, the best results are obtained with the use of the linear kernel, with an average F-score of 0.977; this performance was expected, as the number of features (in this case words) is much greater than the 2 possible classes (Résumés and Other documents) [23]. The Résumé Detector was implemented using a SVM with a linear kernel and the complete training corpus.

2.2.2 Evaluation

The evaluation corpus is a multilingual and heterogeneous set of 240 documents (résumés, motivation letters, publications lists, diplomas, etc.), divided into 4 groups of 60 documents: *French Résumés*, *Résumés in other languages*, *Other French documents* and *Other documents in other language*. It was generated manually by a non-expert recruiter who classified documents randomly chosen from the corpus. Two expert recruiters were asked separately to classify the files from the evaluation corpus into the same groups. The agreement between both expert recruiters was calculated with *Cohen's Kappa* ($\kappa = 93\% \pm 0.04$) and *Kendall's W* ($W = 0.905$ $p\text{-}value = 2.58 \times 10^{-13}$). In order to evaluate the Résumé Detector, each evaluation corpus (*Recruiter 1* and *Recruiter 2*) passed though the document conversion and language detection. Then, the Résumé Detector was utilized to determine which French documents, from both processed corpora, were résumés. Table 3 shows the results from this evaluation in terms of Precision, Recall and F-score; a mean for each measures is presented as well.

Table 3: Evaluation of the Résumé Detector in terms of Precision, Recall and F-score.

Corpus	Precision	Recall	F-score
Recruiter 1	0.964	0.916	0.939
Recruiter 2	**0.982**	**0.965**	**0.973**
Mean	0.973	0.940	0.956

The Résumé Detector reaches a good performance over the evaluation corpora with an average F-score of 0.956. Some cases where the Résumé Detector and the recruiters did not agree are the documents which are bilingual résumés or motivation letters that have short résumé attached.

We applied the Résumé Detector over the documents of the corpus that were detected previously in French. From this task the module detected 22,439 French résumés (47.35% of the total corpus and 68.31% of the converted French documents).

2.3 Résumé Uniqueness

We found that in the corpus there are candidates which have more than one résumé for the same job offer. This happens either because the applicants have sent several résumés for one application or because the applications have been forwarded more than once. The information inside the multiple résumés may or not be exactly the same.

To avoid false or biased results from these cases, we validated the résumé uniqueness in each job offer. The validation is done using 3 tests applied sequentially over all the possible couples of résumés in a job offer[11]:

1) One résumé by candidate: both résumés must come from two different applicants.
2) Résumés with different content: the Linux tool *Diff*[12] is used to validate if both résumés are equal[13].
3) Not equal e-mails: e-mails addresses[14] from the two résumés must be different.

After the verification, for each existing problematic couple, the oldest résumé, according to the receiving date, is deleted.[15] Nevertheless, if a *Rejected* résumé is identical to a *Selected* one, the former will be the deleted one. This exception only applies when a candidate has sent two applications to the same job offer and the first one was *Selected* and the second one, in consequence, *Rejected*.

3. Methodology

We inferred that the résumés of *Selected* candidates are more alike with themselves than with the rest of résumés sent to a job offer. This is because the candidates with résumés fulfilling the characteristics of a job offer are the only contacted by a recruiter. In this paper, we would like to know how the use of certain measures and data weighting affects this inference.

If we consider a *Likeness Score* (LS) as the measure where the higher the value the more alike are the résumés and a set J as all the possible couples of résumés for a job offer, our inference can be represented mathematically with Equation 1.

$$\overline{LS}(J_S^c) < \overline{LS}(J_S) \tag{1}$$

where $\overline{LS}$ is the average Likeness Score, J_S is the subset of J that contains all the possible couples of *Selected* résumés and J_S^c is the complement of J_S. Figure 2 shows an example of possible couples of subset J_S and J_S^c with 3 *Selected* résumés and 3 *Rejected* ones.

3.1 Data Representation

To use a Vector Space Model (VSM) [24] as data representation, we converted each résumé into 3 different vectors. These vectors are constructed from unigrams, bigrams and skip bigrams (SU4) [25], [26] of words. It should be noted

[11]The number of possible couples for a certain job offer is given by all the possible combinations of two résumés (C_2^n) taken from the total number of résumés (n).

[12]http://www.gnu.org/software/diffutils/

[13]The *Diff* tool has been configured to ignore the multiple white spaces and lines but also to be case-insensitive.

[14]The e-mails were detected using a regular expression.

[15]Since the tests are applied in pairs, one résumé can be deleted due to several reasons.

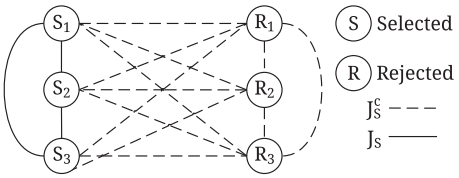


Fig. 2: Example of the possibles couples of the subset J_S and J_S^c with 3 *Selected* candidates and 3 *Rejected* ones.

that before the résumés' n-grams extraction, we lowercased all documents. Also, we removed all the punctuations marks, numbers and stop-words[16] from each document. And we reduced the documents' lexicon through Porter's algorithm for French (stemming).[17] These tasks were done to reduce the possible noise in the text, the size of the VSM and the curse of dimensionality [27].

In addition, the 3 resulting vectors have been represented by 3 types of weights: *absolute frequency*, *relative frequency* and *TF-IDF*. In the case of the TF-IDF, the values are calculated with respect to each job offer. The relative frequencies are obtained résumé by résumé.

3.2 Similarity and Distance Measures

To calculate the Likeness Score of résumés, we implemented 3 similarity measures (*Cosine Similarity*, *Jaccard's Index*, *Dice's Coefficient*) and 2 distance measures (*Euclidean Distance*, *Manhattan Distance*). Table 4 recalls the formula of each measure.

Table 4: Formulæ of the similarity and distance measures.

Measure	Formula	Measure	Formula
Cosine Similarity	$\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$	Jaccard's Index	$\frac{\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert}$
Manhattan Distance	$\sum \lvert x_i - y_i \rvert$	Dice's Coefficient	$2\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$
Euclidean Distance	$\sqrt{\sum \lvert x_i - y_i \rvert^2}$		

Each measure was applied by type of n-grams (unigrams, bigrams and skip-bigrams) and type of weight (absolute frequency, relative frequency or TF-IDF values). In the case of Jaccard's Index and Dice's Coefficient, we only make use of the absolute frequency as weight. The reason is that we implemented their binary version, which only takes into account the existence or absence of elements. As well, we only applied Cosine Similarity to absolute frequency and TF-IDF values as the results using absolute or relative frequencies will be always the same [28, Page 111].

In order to have only one Likeness Score by type of weight, we decided to merge the 3 n-grams' Likeness Scores into one by a simple combination. This combination consists in multiplying each Likeness Score by an *influence factor* and making the addition of the resulting values. The influence factor of the 3 types of n-grams has been settle to $1/3$, giving the same leverage to all of them. However, the influence factor can be changed independently on the condition that the sum of them is equal to one. This merge is quite *naïve* but our purpose is to follow an *a fortiori* principle. If this combination method leads us to good results, the use of more sophisticated merging methods or influence factor setting may lead us to better results.

The calculation of each Likeness Score was parallelized using GNU Parallel [29].

3.3 Statistical Test

To know whether the Likeness Scores of the groups, J_S and J_S^c are statistically different, we performed a two-tailed Analysis of Variance (ANOVA) for independent groups.

Owing to the characteristics of the corpus, not all the job offers are analyzable with our methodology. We found that there are 63 job offers where it does not exist at least one French résumés in J_S or J_S^c, making impossible to calculate any measure. These cases represent the job offers without résumés from a Selected or Rejected applicant, see Figure 1, and the job offers ($\approx$ 25) where non-French résumés are dominant. We found as well 9 cases where the number of résumés of group J_S prevent us to verify whether the measure distribution is normal. These 72 job offers were deleted from the analysis.

Before doing any ANOVA we suppressed the outliers from the groups J and J_S^c of each analyzable job offer (224). We defined the outliers as the values that are 1.5 times the interquartile range ($IQR = Q_3 - Q_1$) below the first quartile (Q_1) or above the third quartile (Q_3) [30, Page 208]. After deleting the outliers, we verify that both groups fulfill the following two assumptions, which are necessary to do an ANOVA:

- Normal distribution: a Shapiro-Wilk Test ($\alpha = 0.05$) is applied to verify data normality. As this test can only be used in groups that contain between 3 and 5,000 elements [31], the groups with less than 3 elements are not considered as Gaussian. The groups with more than 5,000 elements are considered as normal even if it may be a violation of the assumption. However, the ANOVA is a robust test where the normality assumption can be discarded with minor effects [32, Page 424].
- Variance equality: The homogeneity of variances is tested with a Bartlett's Test ($\alpha = 0.05$). In case of heterogeneous variances, the ANOVA is only done if the biggest variance is not greater than 4 times the smallest one [32, Page 354].

In case one of the groups of a job offer does not surpass the previous conditions, the job offer is considered not analyzable (NA).

[16] List taken from the Perl's module "Lingua::StopWords".

[17] We used the Perl's module "Lingua::Stem::Snowball", an interface for the stemmers of the Snowball project (http://snowball.tartarus.org/).

Once the ANOVA of a job offer has been calculated, we consider that the averages of the Likeness Score for both groups, $\overline{LS}(J_S)$ and $\overline{LS}(J_S^c)$, are statistically different only if the ANOVA's $p\text{-}value < 0.05$.

4. Results

The results for Cosine Similarity, Manhattan Distance, Euclidean Distance, Dice's Coefficient and Jaccard's Index are shown in Table 5. The outcomes for the first 3 measures are divided by vector component weight: absolute frequency, relative frequency and TF-IDF.

For all the results, we present the number of job offers where the average Likeness Score ($\overline{LS}$) for both groups is statistically different ($p\text{-}value < 0.05$) and statistically equal ($p\text{-}value \geq 0.05$). As well, we present the number of cases that did not surpass the ANOVA's conditions (NA), the number of job offers where the elements of J_S have more in common with themselves (J_S+) and the cases where the elements of J_S^c have more in common with themselves (J_S-). The total number of analyzable job offers in the corpus was 224.

Table 5: Results for Cosine Similarity, Manhattan distance, Euclidean distance, Dice's Coefficient and Jaccard's Index.

		$p\text{-}value$			
		< 0.05		≥ 0.05	NA
		J_S+	J_S-		
Cosine Similarity	Absolute/Relative Frequency	163	11	44	6
	TF-IDF	158	10	55	1
Manhattan distance	Absolute Frequency	53	90	63	18
	Relative Frequency	164	7	50	3
	TF-IDF	59	86	62	17
Euclidean distance	Absolute Frequency	69	83	58	14
	Relative Frequency	124	30	62	8
	TF-IDF	69	78	61	16
Jaccard's Index		164	1	58	1
Dice's Coefficient		164	2	56	2

As seen in Table 5, there are seven cases that clearly follow our inferred behavior (Cosine Similarity; Manhattan and Euclidean Distances with relative frequencies; Jaccard's Index and Dice's Coefficient) and 4 cases where it is clear that the inference does not behave as inferred (Euclidean Distance with absolute frequency and TF-IDF; Manhattan Distance with TF-IDF and absolute frequency).

With the purpose of comparing easily the results between the different measures and vectors' weights, 3 rates and one score have been established:

$$SR = \frac{\text{Total}_{\text{Significant}}}{\text{Total}_{\text{Analyzable job offers}}} \quad (2)$$

$$TR = \frac{\text{Total}_{\text{Significant}} + \text{Total}_{\text{Not Significant}}}{\text{Total}_{\text{Analyzable job offers}}} \quad (3)$$

$$IR = \frac{\text{Total}_{J_S+}}{\text{Total}_{\text{Statiscally different}}} \quad (4)$$

$$RS = \sqrt[3]{SR * TR * IR} \quad (5)$$

The *Significant rate* (Equation 2) indicates the ratio between the number of job offers with a significant ANOVA test and the total number of analyzable job offers in the corpus. The *Testing rate* (Equation 3) expresses the proportion of job offers tested with an ANOVA regarding the total number of analyzable job offers. Our inference about the résumés' Likeness Scores ($J_S^c < J_S$) is measured with the *Inference rate* (Equation 4), which is the number of job offers following the expected behavior per the number of job offers with an ANOVA $p\text{-}value < 0.05$. The *Ranking Score* (Equation 5) is a value which allow us to rank the measures according to their Significant, Testing and Inference rates. For the three rates and the score the higher the value, the better (the maximum value is 1 while the minimum is zero). Table 6 shows the values of the 3 rates and the score for each measure.

Table 6: Significance rate (SR), Testing rate (TR), Inference rate (IR) and Ranking Score (RS) for each measure.

		SR	TR	IR	RS
Cosine Similarity	Absolute/Relative Frequency	**0.776**	0.973	0.936	0.890
	TF-IDF	0.750	**0.995**	0.940	0.888
Manhattan distance	Absolute Frequency	0.638	0.919	0.370	0.600
	Relative Frequency	0.763	0.986	0.959	0.896
	TF-IDF	0.647	0.924	0.406	0.623
Euclidean distance	Absolute Frequency	0.678	0.937	0.453	0.660
	Relative Frequency	0.687	0.964	0.805	0.810
	TF-IDF	0.656	0.928	0.469	0.658
Jaccard's Index		0.736	**0.995**	**0.993**	**0.899**
Dice's Coefficient		0.741	0.991	0.987	0.898

Taking into account the results presented in Table 6, we can see that in terms of the Significant rate, the highest score is the one of Cosine Similarity using frequencies (0.776); in terms of Testing rate, the highest rates are obtained by Cosine Similarity with TF-IDF and Jaccard's Index (0.995). Regarding the Inference Rate, the highest score is for Jaccard's Index (0.993). And with respect to the Ranking Score the leading one (0.889) is also for Jaccard's Index. The overall lowest scores are those obtained by Manhattan Distance using absolute frequency with a Significant Rate of 0.638, a Testing Rate of 0.919, an Inference rate of 0.370 and Ranking Score of 0.600.

Finally, from the results of the 11 analysis we can point out three points:

- Seven analysis clearly present results that follow the expected behavior.
- Relative frequencies as vector weight improve the performance of distances measures.
- Binary measures have comparable performance to Cosine Similarity.

5. Discussion

The performance of the Manhattan and Euclidean distances are not the expected one. We attended to have similar results, for all types of weights, like Cosine Similarity. However, only the use of relative frequencies, in both measures, give the expected outcome. Moreover, the performance of Manhattan distance can be greatly improved, passing from the worst Ranking Score (0.600) to the third best (0.896), better than Cosine Similarity using frequencies.

It can be thought that the disagreement behavior obtained by Manhattan and Euclidean distances using TF-IDF or absolute frequency is linked to the Gaussian assumption. We considered as normal the groups J_S^c or J_S having more than 5,000 elements, as their size exceed the superior limit of the Shapiro-Wilk Test. Nonetheless, the effect of this decision may not be relevant if it is taken into account that only 45 cases (20.08%) of the analyzable job offers (224) have at least one group with more than 5,000 elements. Moreover, 44 of these cases have always a homogeneous variance and one of them, depending on the measure and the vector weight, can have a homogeneous variance or not.

Actually, we think that the disagreement behavior is related to the intrinsic characteristics of Manhattan and Euclidean distances. Unlike Cosine similarity, Dice's Index and Jaccard's Coefficient, which are measures always delimited by the interval of values $[0, 1]$, Manhattan and Euclidean distances are measures that can have a $[0, \infty)$ interval. This means that the superior interval limit is not defined and that it is restricted to the size and lexical richness of the documents. Therefore, two measures of the same distance may have different interval limits and comparing them may not be equivalent. For example, for two completely different documents, their distance X means 0% in common, while for two documents half different, their distance X means 50% in common; both distances have the same value X but different scale, making their comparison incompatible.

In our case, the résumés are not limited neither in size nor in vocabulary, hence the use of not normalized versions of Manhattan and Euclidean distances, i.e. with a closed interval, are not reliable in most cases. The exception is Manhattan Distance with relative frequencies, in this case this type of weight works like a distance normalization as it close the interval[18] into $[0, 2]$. It may be thought that the use of Euclidean Distance with relatives frequencies would be an exception as well, however, the relative frequency does not close the interval.

In order to understand better our results, we analyzed the job offers marked as NA. We found that all the NA cases are job offers with a heterogeneous variance. This means, that all the analyzed job offers have more than 3 elements and those between 3 and 5,000 elements have a normal distribution. In addition, we can see that the number of cases with heterogeneous variances arises when we make use of distance measures without relative frequencies.

The performance of Cosine Similarity with TF-IDF did not turn out as anticipated. We assumed that the use of TF-IDF would boost the performance of Cosine Similarity, as the components of the vectors would be weighted by their importance [33]. Nevertheless, the difference of the Ranking Score between the use of frequencies and TF-IDF values is about -0.225%; for the others rates the difference are: Significant Rate -3.35%, Testing Rate $+2.26\%$ and Inference Rate $+0.42\%$.

Finally, the performance of Dice's Index and Jaccard's Coefficient exceeded our expectation. We never thought that only the presence or absence of n-grams could be enough to determine the inferred behavior; however, we found that binary measures are sensible enough to determine résumés' likeness. Thus, we can infer that Selected résumés have a specific vocabulary which is not present in the Rejected résumés. Moreover, this means that the frequency of "terms" is not relevant for recruiters, instead of it, the most important thing is the appearance or not of "terms" related to the job offer requirements.

6. Conclusions and Future Work

In this paper, we made a comparative analysis of 3 similarity measures (Cosine Similarity, Dice's Coefficient, Jaccard's Index) and 2 distance measures (Euclidean and Manhattan distances). All the measures, excepting Dice's Coefficient, Jaccard's Index, were compared with at least 3 types of vector weights (absolute frequency, relative frequency and TF-IDF values). The objective was to determine how the use of different measures and vector weight affects the likeness detection of Selected résumés, i.e. résumés from applicants contacted by a recruiter.

This work was done over a large annotated recruitment corpus coming from a HRM firm. We made use of NLP techniques in order to detect the French résumés from the corpus. As well, we utilized an Analysis of Variance (ANOVA) to determine how the 5 measures considered the likeness of the résumés. And therefore, to compare with our inference: whether Selected résumés have more in common with themselves than the rest of résumés do.

Results varied according to type of vector weight and to measure. The use of Manhattan or Euclidean Distances may not be reliable to measure the likeness of résumés if some

[18] If two vectors X and Y using relative frequencies are completely different, their Manhattan Distance becomes $\sum X_i + \sum Y_i = 1 + 1 = 2$. Therefore, the maximum value possible in this case is 2.

considerations are not taken. The document size and lexical richness affects strongly these measures. Therefore, it may be better to use their normalized versions.

Cosine Similarity has shown the best results when it is used with frequencies, relative or absolute. Nonetheless, their performance was reduced when we used TF-IDF.

The use of Jaccard's Index and Dice's Coefficient, presented a good performance and exceeded our expectations. Moreover, we think that the use of these measures to find likeness between résumés, for example in matching systems, may give good results.

And we believe, according to the results obtained from Jaccard's Index and Dice's Coefficient, that there must be a specific vocabulary that could lead us to detect easily the résumés from candidates that should be Selected by a recruiter or not.

As future work, we are going to analyze other languages, like English, in order to determine whether our methodology can be defined as language independent. In addition, we will implement new procedures to reduce the lexicon, like lemmatizers or other stemmers. As well, we will analyze how other types of vector weight affect the tests, for example other TF-IDF methods. We will improve the method utilized to merge the likeness score of n-grams, for example using a weighted mean, calculating the influence factor of each type of n-gram in the likeness score or creating one vector with all the n-grams. New distances will be also tested, like normalized versions of Manhattan and Euclidean distances, or non-binary versions of Jaccard's Index and Dice's Coefficient. The use of non parametric ANOVAs and/or robust ANOVAs is expected. Finally, we will do more inferences about the Rejected résumés, and about the job offer and the résumés.

Acknowledgments

This project has received the support from the Agence National de la Recherche et de la Technologie (ANRT) from the CIFRE convention 2012/0293b and from the Consejo Nacional de Ciencia y Tecnología (CONACyT) grant 327165.

The authors would like to thank Amandine Bugnicourt, Clémence Chardon and Elodie Chevalier from Adoc Talent Management. They helped us to understand better the HRM tasks and gave us several ideas in order to develop some tools used in this project.

References

[1] R. Rafter, K. Bradley, and B. Smyth, "Automated Collaborative Filtering Applications for Online Recruitment Services," in *Adaptive Hypermedia and Adaptive Web-Based Systems*, ser. Lecture Notes in Computer Science, P. Brusilovsky, O. Stock, and C. Strapparava, Eds. Springer Berlin Heidelberg, 2000, vol. 1892, pp. 363–368.

[2] P. De Meo, G. Quattrone, G. Terracina, and D. Ursino, "An XML-Based Multiagent System for Supporting Online Recruitment Services," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 37, no. 4, pp. 464–480, July 2007.

[3] R. Kessler, N. Béchet, M. Roche, J.-M. Torres-Moreno, and M. El-Bèze, "A hybrid approach to managing job offers and candidates," *Information Processing & Management*, vol. 48, no. 6, pp. 1124–1135, 2012.

[4] C. Bizer, R. Heese, M. Mochol, R. Oldakowski, R. Tolksdorf, and R. Eckstein, "The impact of semantic web technologies on job recruitment processes," in *Wirtschaftsinformatik 2005*. Springer, 2005, pp. 1367–1381.

[5] V. Radevski and F. Trichet, "Ontology-based systems dedicated to human resources management: An application in e-recruitment," in *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, ser. Lecture Notes in Computer Science, R. Meersman, Z. Tari, and P. Herrero, Eds. Springer Berlin Heidelberg, 2006, vol. 4278, pp. 1068–1077.

[6] L. Yahiaoui, Z. Boufaïda, and Y. Prié, "Semantic Annotation of Documents Applied to E-Recruitment." in *Proceedings of SWAP 2006, the 3rd Italian Semantic Web Workshop*, 2006, pp. 1–6.

[7] D. S. Chapman and J. Webster, "The use of technologies in the recruiting, screening, and selection processes for job candidates," *International Journal of Selection and Assessment*, vol. 11, no. 2-3, pp. 113–120, 2003.

[8] P. Montuschi, V. Gatteschi, F. Lamberti, A. Sanna, and C. Demartini, "Job recruitment and job seeking processes: how technology can help," *IT Professional*, vol. 16, no. 5, pp. 41–49, 2014.

[9] D. Looser, H. Ma, and K.-D. Schewe, "Using formal concept analysis for ontology maintenance in human resource recruitment," in *Proceedings of the Ninth Asia-Pacific Conference on Conceptual Modelling-Volume 143*. Australian Computer Society, Inc., 2013, pp. 61–68.

[10] E. Faliagka, L. Kozanidis, S. Stamou, A. Tsakalidis, and G. Tzimas, "A personality mining system for automated applicant ranking in online recruitment systems," in *Web Engineering*. Springer, 2011, pp. 379–382.

[11] R. Rafter, B. Smyth, and K. Bradley, "Inferring relevance feedback from server logs: A case study in online recruitment," in *The 11th Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2000)*, 2000.

[12] F. Trichet, M. Bourse, M. Leclère, and E. Morin, "Human resource management and semantic web technologies," in *Information and Communication Technologies: From Theory to Applications, 2004. Proceedings. 2004 International Conference on*. IEEE, 2004, pp. 641–642.

[13] M. A. Thompson, *The global resume and CV guide*. Chichester, New York: Wiley, 2000.

[14] A. Singh, C. Rose, K. Visweswariah, V. Chenthamarakshan, and N. Kambhatla, "PROSPECT: a system for screening candidates for recruitment," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 659–668.

[15] W. B. A. Karaa and N. Mhimdi, "Using ontology for resume annotation," *International Journal of Metadata, Semantics and Ontologies*, vol. 6, no. 3, pp. 166–174, 2011.

[16] D. Çelik, A. Karakas, G. Bal, C. Gultunca, A. Elçi, B. Buluz, and M. C. Alevli, "Towards an Information Extraction System Based on Ontology to Match Resumes and Jobs," in *Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual*. IEEE, 2013, pp. 333–338.

[17] V. Senthil Kumaran and A. Sankar, "Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping (EXPERT)," *International Journal of Metadata, Semantics and Ontologies*, vol. 8, no. 1, pp. 56–64, 2013.

[18] E. Faliagka, L. Iliadis, I. Karydis, M. Rigou, S. Sioutas, A. Tsakalidis, and G. Tzimas, "On-line consistent ranking on e-recruitment: seeking the truth behind a well-formed CV," *Artificial Intelligence Review*, pp. 1–14, 2013.

[19] R. Kessler, J. M. Torres-Moreno, and M. El-Bèze, "E-gen: Profilage automatique de candidatures," *TALN 2008, Avignon, France*, pp. 370–379, 2008.

[20] L. A. Cabrera-Diego, J.-M. Torres-Moreno, and M. El-Bèze, "SegCV : traitement efficace de CV avec analyse et correction d'erreurs," in *Actes de TALN 2013*, 2013, pp. 707–714.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
[22] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
[23] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Department of Computer Science, National Taiwan University, Tech. Rep., 2003.
[24] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.
[25] X. Huang, F. Alleva, H.-w. Hon, M.-y. Hwang, and R. Rosenfeld, "The SPHINX-II Speech Recognition System: An Overview," *Computer, Speech and Language*, vol. 7, pp. 137–148, 1992.
[26] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Association for Computational Linguistics, 2004, pp. 74–81.
[27] R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton, New Jersey: Princeton University Press, 1961.
[28] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.
[29] O. Tange, "GNU Parallel - The Command-Line Power Tool," *;login: The USENIX Magazine*, vol. 36, no. 1, pp. 42–47, Feb. 2011.
[30] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, 5th ed. Hoboken, New Jersey: John Wiley & Sons, 2010.
[31] P. Royston, "Remark AS R94: A remark on algorithm AS 181: The W-test for normality," *Applied Statistics*, pp. 547–551, 1995.
[32] D. Howell, *Fundamental statistics for the behavioral sciences*, 8th ed. Wadsworth, California: Cengage Learning, 2013.
[33] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

A Language-Based and Process-Oriented Approach for Supporting the Knowledge Discovery Processes

Hesham A. Mansour[1], **Daniel Duchamp**[1], **and Carl-Arndt Krapp**[2]
[1]Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ 07030, USA
[2]FJA-US, Inc., 1040 Avenue of the Americas, 4th Floor, New York, NY 10018, USA

Abstract - *Knowledge Discovery in Databases (KDD) processes are complex, highly interactive and iterative. The similarities between KDD processes and software development processes suggest that approaches used to manage the development of software processes are also applicable and in fact advantageous to KDD processes. In this paper, we examine the current approaches for supporting KDD and note to their limitations in providing comprehensive and effective process support. We propose a language-based and process-oriented approach for supporting KDD processes that is based on explicitly representing KDD processes as process programs that can be analyzed, validated, and enacted. We illustrate the proposed approach using a novel process programming language that is designed to describe general process concepts as well as specific KDD concepts. Along with the KDD process language, an IDE-style development environment is proposed to assist in modeling and enacting KDD processes. The overall approach is evaluated and illustrated by modeling and enacting a traditional KDD process.*

Keywords: Data Mining, Knowledge Discovery in Databases, KDD Process, Process Programming.

1 Introduction

Today, KDD projects are typically approached in an unstructured, ad hoc manner [6]. The lack of systematic approaches for managing and keeping track of the different parts of KDD projects means that some steps may unintentionally be repeated, adding overhead to the knowledge discovery task. Rudiger et al. [6] have noted major problems during the development of many KDD projects at Daimler-Benz due to the lack of a methodology and lack of a usable process model with proper tool support. Marban et al. [28], [29] have noted that the number and complexity of data mining projects has increased in recent years, that nowadays there isn't a formal process model for this kind of project, and that existing approaches are not correct or complete enough. They also noted that not all projects end successfully. The failure rate is actually as high as 60%.

In this paper, we proposes the Knowledge Discovery Process Modeling and Enacting Language (KDPMEL) along with its Process-Centered Software Environment (PCSE-KDD) that can be used to develop KDD processes in a way that is similar to developing software processes: KDD processes as process programs written in KDPMEL and exploited by PCSE-KDD to provide execution support and management for KDD processes.

Considerable value can be gained from materializing KDD processes via process programming. Novice participants, in particular, can benefit the most from knowing and learning their roles in the process and how their work and contributions would be coordinated and fit with others' work and contributions. It has been observed by many KDD practitioners [22] that the results of KDD projects are often highly dependent on the experience of the persons doing the work. This phenomenon would likely be mitigated by having the work explicitly defined in a way that allows sharing the experience among the persons doing the work.

2 Current Approaches for Supporting KDD Processes and Their Limitations

We distinguish three major KDD support approaches found in the literature: activity-oriented support, KDD support environments, and process-oriented support.

2.1 KDD Activity-Oriented Support

This approach provides support only for individual activities such as data preprocessing or algorithm selection and settings. Examples of such support are proposed by [14]-[18], [42]. In this approach, the process concept, if used at all, is only represented in the form of documentation and guidelines. Also, the tools supporting the process tasks are isolated without any means of integration that would facilitate their usage and can enforce consistency conditions among the produced artifacts.

2.2 KDD Support Environments

The development of software environments supporting the overall KDD process has been identified by Padhraic [5] as a grand challenge for KDD. The architectures proposed for such environments are mainly based on a hardwired process model such as the traditional KDD process model [1] or the CRISP-DM [2] model. Some of the research efforts that fall under this category can be found in [8], [19]-[27], [30], [41].

Although these environments can be used to define and execute KDD processes, the provided process support is mainly derived from the hardwired process model, which includes major process phases along with their generic tasks and simple interactions. This sort of guidance is too generic and clearly insufficient for effectively supporting KDD processes, where specialized guidance is needed to assist in selecting valid, desirable, and effective process configurations. Moreover, the guidance provided by these systems is limited to a few standard KDD techniques and prescribed set of supporting tools that are mandated by the environments. This generic guidance and limited support is insufficient for a dynamic field such as KDD where scores of new techniques, guidelines, and tools for data manipulation and analysis are added on regular basis.

2.3 KDD Process-Oriented Support

KDD process-oriented support assures that activities performed within KDD processes are properly controlled and data analysis and manipulation techniques are used appropriately. Very few researchers [7], [22], [32]-[35] have addressed the issue of providing process-oriented support.

The project CITRUS [22] has extended a commercial knowledge discovery tool, CLEMENTINE, to enhance its user support capabilities by providing a process support interface. The main limitations of CITRUS are its dependency on CLEMENTINE and its supported process model and offered techniques; and its high-level process guidance.

Osterweil et al. [7] were the first to propose the use of process programming to address the coordination of KDD techniques. This process-oriented approach is illustrated using the Little-JIL language. Little-JIL is a visual language derived from a subset of JIL, a "process language" originally developed for software development processes [9]. Although the use of Little-JIL to specify a representative bivariate regression process has shown that many coordination aspects of the process can be easily expressed, it has uncovered some deficiencies in the language. Although this attempt is very promising, it concentrates on supporting only the coordination aspect of the process. In addition to the discovered deficiencies in Little-JIL, only the simplest processes can be modeled visually using Little-JIL.

Collaboration is another process-oriented aspect that has been recently adopted by some KDD support proposals [32]-[35], which are based on the paradigm of Service-Oriented Architecture (SOA). Collaborative KDD (CKDD) is an emerging field that seeks to cope with the distributed structure of modern organizations and the consequent increased complexity of the knowledge discovery process [31]. Although CKDD is beyond the scope of our work, it's worth noting that our process-oriented approach, which employs a process-centered environment, inherently promotes the collaboration aspect.

2.4 The Need for a more Comprehensive Approach for Supporting KDD Processes

As discussed previously, the current state of KDD support is that the first approach (activity-oriented) supports only fragments of the KDD process, the second approach (KDD support environments) supports only a particular KDD process model, and the third approach (process-oriented) supports only certain process aspects of the KDD process.

Each of these half measures is inadequate. The support needed for a KDD process varies greatly based on the specifications of the concrete KDD process, and cannot be based purely on a generic process model. A KDD process might have many different configurations and can be instantiated in a number of ways, and each configuration might require different support.

Highly specialized KDD process support presently takes the form of technical documentation that specifies desirable and effective configurations for the process steps in an informal way [7]. This requires KDD practitioners to learn and apply these specifications manually. While this may be acceptable for experienced KDD practitioners who can cope with only high level guidance, it is not suitable for the less sophisticated KDD practitioners who participate in the development and enactment of the majority of KDD processes. We believe that the guidance necessary for the typical user can be achieved by explicitly representing the concrete KDD process using a flexible and rigorous formalism provided by the language-based approach of process programming. Further, explicit representation of the KDD process can be exploited by a process-aware environment to support process execution and guide users in carrying out their duties.

3 The Knowledge Discovery Process Modeling and Enacting Language (KDPMEL)

KDPMEL is a novel process programming language for modeling and enacting KDD and data analysis activities along with their resources, interactions, and coordination.

The process aspects of the language such as process structuring and task ordering are similar to those found in general process languages, such as JIL [9], Little-JIL [7], and PML [11]. The KDD aspects of the language are specific features for modeling KDD artifacts, tools, and tasks.

KDPMEL supports modeling KDD tasks at different levels of detail and abstraction in order to specify both generic and specialized tasks. Specialized KDD tasks are defined in KDPMEL through external commands that are modeled in the program and executed through a flexible plug-in mechanism for the tools of these commands.

KDPMEL provides special control constructs to explicitly model task dependencies on other tasks. This

feature is particularly important for KDD and is intended to effectively manage the dependencies between KDD techniques. Having these dependencies explicitly defined can assure that they are appropriately handled.

A process program written in KDPMEL is organized into three major sections that specify the resources (*artifacts*, *roles*, and *tools*) of the process, general information about the process (*goal*, *input*, *outcome*, and *assessment*), and the steps (*activity*, *action*, and *command*) of the process along with their sequencing (*sequence*, *parallel*, *choice*, and *loop*) and dependencies (*disallow*, *require*, and *enable*).

3.1 Language Goals

The major goals of the KDPMEL are the *simplicity*, *flexibility*, *expressiveness*, and *generality*.

3.2 Language Approaches

KDPMEL combines both the graphical and process language modeling approaches. It employs a number of graphical modeling editors on top of a textual process programming language designed specifically for KDD. The goal of the modeling editors is to facilitate the construction and presentation of certain process components, as represented by four types of graphs that display the overall process (process graph), the resources of the process (resources graph), a graph for each activity (activity graph), and a graph for each action (action graph). Furthermore, a read-only graph that shows the progress of the artifacts within the process (artifact flow graph) is provided. The process graph indicates process information such as goal, input, outcome, etc. The resources graph shows the artifacts, tools, and roles/actors of the process. Each activity graph shows the activity's constituent actions while each action graph provides information such as tools utilized by the action, the artifacts consumed and produced by the action, and the actor assigned to the action. In addition to the graphical editors, a number of form-based editors exist (process form, activity form, artifact form, role/actor form, and tool form) to present and update certain information that is best shown and updated in a form-based style.

Combining different types of editors and views in source-based, graph-based, and form-based styles is novel and allows both technical and non-technical users to participate in the modeling phase of the process. This hybrid modeling approach combines the benefits of the underlying approaches and enables specification of high-level process models as well as more complex ones in a manner that is convenient for both technical and non-technical users.

3.3 KDPMEL Meta-Model (Abstract Syntax)

KDPMEL is defined in terms of a meta-model [38] based on the OMG's SPEM [13] and CWM [37] meta-models, which represents the abstract syntax and static semantics of the language. The KDPMEL meta-model consists of a Core meta-model upon which the other meta-models depend, a Process meta-model representing the process aspects, and a KDD meta-model representing the KDD aspects.

3.4 KDPMEL Concrete Syntax

The concrete syntax of KDPMEL is provided in two flavors, textual and graphical, to serve different purposes. The textual concrete syntax is useful when specific complex details must be specified. The graphical concrete syntax is easy to understand and use, and is useful for communicating structural and higher level views of the process. Also, a form-based interface is provided for process components to allow for presenting and updating their properties.

3.4.1 KDPMEL Textual Concrete Syntax (Grammar)

The Process meta-model provides process specific entities such as *Process*, *Activity*, and *Action*. The syntax for defining these constructs is given by the following rules:

<process> ::= "**process**" <IDENTIFIER> "{"…"}"
<activity> ::= "**activity**" <IDENTIFIER> "{"…"}"
<action> ::= "**action**" <IDENTIFIER> "{"…"}"

Process Syntax

A process can be decomposed into an ordered collection of activities. The activities can be grouped using one of the control constructs *sequence*, *parallel*, *choice*, or *loop*. The process syntax is defined as follows:

<process> ::= "**process**" <IDENTIFIER> "{" …
(<activitySequencing> | <activity>)*
"}"

Activity Syntax

An activity represents a composite task and it is mainly intended to represent the phases of the KDD process. An activity may have pre-conditions and post-conditions to guard entry into and exit from the activity. An activity may consume and produce some artifacts during its performance, which is monitored by an actor. An activity can be decomposed into smaller units of sub-activities and/or actions. The activity syntax is defined as follows:

<activity> ::= "**activity**" <IDENTIFIER> "{"
["**preconds**" <constraint> ("," <constraint>)*]
["**postconds**" <constraint> ("," <constraint>)*]
[<consumedArtifacts>] [<producedArtifacts>]
[<performer>]
["**sub-activities**" "{" (<activity>)+ "}"]
(<actionSequencing> | <action>)*
"}"

Actions within an activity can be grouped using one of the control constructs *sequence*, *parallel*, *choice*, or *loop*. The following defines an activity that has two actions grouped by the *parallel* construct:

```
activity DataMining {
```

```
    parallel predict {
      action buildDecisionTreeModel {…}
      action buildNeuralNetModel {…}
   }
}
```

The decomposition of an activity allows for defining the tasks of the activity. The activity→sub-activity decomposition provides a strict control, whereas the activity→action decomposition provides both strict (e.g., *sequence*/*loop*) and loose control (e.g., *choice*/*parallel*).

Action Syntax

An action represents a primitive task and it is intended to represent the generic tasks of the KDD process. An action may have pre-conditions and post-conditions. An action is performed by an actor with the help of some tools. An action may consume and produce some artifacts. To help perform an action, guidance information for the actor may be associated with the action. Finally, an action may have dependencies with other actions. The action syntax is defined as follows:

<action> ::= "**action**" <IDENTIFIER> "{"
["**preconds**" <constraint> (",” <constraint>)*]
["**postconds**" <constraint> (",” <constraint>)*]
[<consumedArtifacts>] [<producedArtifacts>]
[<performer>] [<utilizedTools>]
[<dependencyDecl>] [<guidanceDecl>]
"}"

The following example is a KDD task for building a decision tree model that specifies that the task is performed by a data mining analyst with the help of a particular mining tool over a specific dataset:

```
action buildDecisionTreeModel {
    consume sampleDataset;
    produce sampleDecisionTreeModel;
    performer dmAnalyst; utilize { call miningTool }
}
```

3.4.2 KDPMEL Graphical Syntax

The *Process*, *Activity*, and *Action* constructs, in addition to the process resources are represented in the graphical syntax. In addition to the graph-based notation, a form-based interface is provided. The source-based, graph-based, and form-based notations of the process program share a common object model for the process that is updated by and translated into the various representations of the process program.

3.5 KDPMEL Semantics

3.5.1 Control Flow and Ordering

The activities within a process and the actions within an activity can be grouped using one of the control constructs *sequence*, *parallel*, *choice*, or *loop*. The default grouping is *sequence*. Additionally, activities may be decomposed into a hierarchy of sub-activities and actions.

3.5.2 Dependency Control Constructs

The states of KDPMEL tasks and their dependency requirements are recorded during the execution of the tasks. Upon beginning the execution of a task, a test is performed against the completed actions to check whether their *disallow* dependency prohibit execution of the task. The *enable* dependency is checked only for the *choice* control construct to determine if one of the choices has been enabled by a completed action. If that was the case, the actor making the choice will be notified. Another test is performed upon beginning the execution of an action to check its *require* dependency against the completed tasks to determine if the action can be executed. An action can only be executed if its required tasks are completed.

We believe that this is a novel approach for managing KDD task dependencies that are dynamically reflected at runtime, as opposed to statically structuring these tasks according to their dependencies at modeling time [7], which provides more flexibility not only in modeling time but in execution time also. In addition, it also leads to much shorter programs.

3.5.3 Action Specialized KDD Tasks

KDPMEL models specialized KDD tasks through the use of external commands that can be associated with an action and a tool. Each tool that is associated with an external command is represented by a plug-in module that is invoked to execute the command.

3.5.4 Task States and Transitions

The states and transitions of KDPMEL tasks are implemented using the State Pattern [40]. Tasks within a KDPMEL program go through several states during the execution of the program. The state of a task changes based on the control flow of the program, the availability of the resources needed by a task, and the explicit response from human actors. KDPMEL adopts states similar to those of Little-JIL--*posted*, *started*, *completed*, *terminated*, and *retracted*--- and adds the two new states *suspended* and *resumed* that have been suggested by Lee [12]. Figure 1 illustrates KDPMEL task states and their transitions as suggested in [12].

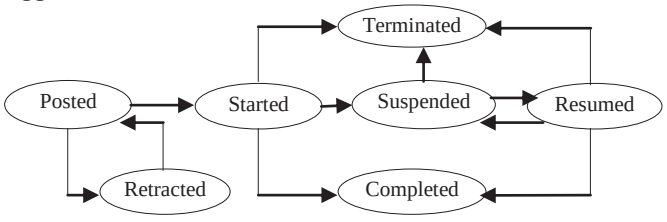


Fig. 1. Task State Transitions in KDPMEL [12]

When a KDPMEL task becomes available for execution, a task instance is created and its state is set to *posted*. A *posted* task instance is started by an explicit *start* action from the task performer (actor), which sends a start event to the task controller to change the state of the task instance to *started*. A *posted* task instance can also be temporarily

retracted by a *retract* action. A *started* task instance is completed by an explicit *complete* action. A *completed* task indicates that the task has successfully finished execution. This causes the enactment engine to continue executing the rest of the program by finding and posting the next available task. A *started* task instance can also be terminated by an explicit *terminate* action. A *terminated* task indicates an exception that caused the task not to be finished successfully. The handling of a terminated task varies depending on the type of control construct governing the task. For instance, while a *terminated* task in a *sequence* or *loop* control construct causes the termination of the other tasks in the construct, a *terminated* task in a *choice* control construct causes the enactment engine to offer the construct alternatives to the actor to select a task.

4 The KDD Process-Centered Support Environment (PCSE-KDD)

PCSE-KDD is an Integrated Development Environment that is built around KDPMEL, with an IDE-style approach to facilitate the development, execution, and management of KDPMEL programs. The environment offers a variety of services, similar to those offered by PCSEEs, but directed toward KDD processes.

4.1 Architecture

The environment implements the process definition/instantiation/enactment paradigm, found in PCSEEs, and includes a number of modeling editors for modeling KDD processes, an Enactment Engine for providing runtime process execution support, and a Repository for providing persistency support to both process artifacts and process execution states. Figure 2 illustrates the high level architecture of the environment.

The PUI exposes the various services offered by the environment. Through the PUI, users are able to define, update, and persist process models during the modeling phase, instantiate a process model for enactment, participate in the enactment phase by performing interactive tasks in the process, are notified by the enactment engine about the status of the process being enacted, and are guided by the enactment engine about what to do next.

The Enactment Engine is responsible for executing KDPMEL programs. It guides and supports users in performing their assigned tasks, controls the invocation of tools, accesses the process artifacts, and maintains the process execution states. It includes three significant components: KDPMEL Interpreter, the Repository Management Unit (RMU), and the Tool Invocation Unit (TIU). The KDPMEL Interpreter implements the semantics of the language. The RMU maintains the process data. The TIU manages the invocation of tools specified in the process program.

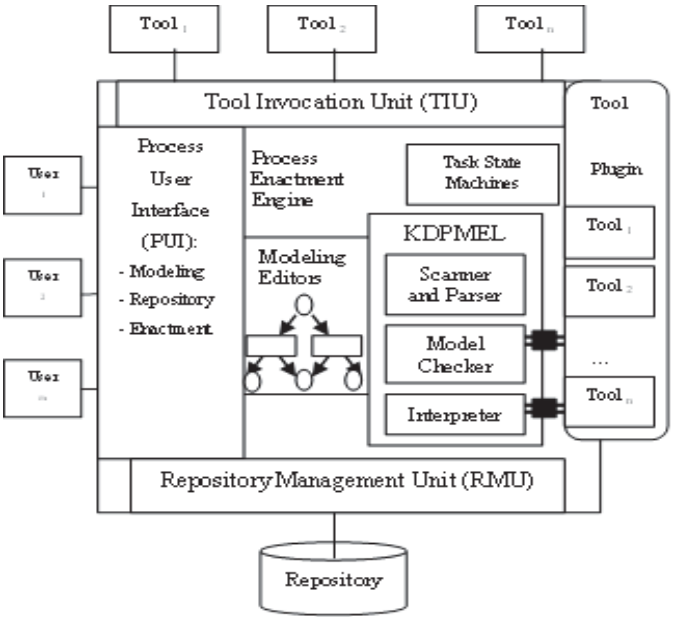


Fig. 2. The high level architecture of the PCSE-KDD

5 An Example for Developing a Traditional KDD Process in PCSE-KDD

The example process is used for predicting the likelihood that bank customers will reply to a mailing campaign for buying a Personal Equity Plan (PEP) [36].

5.1 The Example Process Specification

Data Selection

The data is available in a comma-separated value (CSV) file.

Data Preparation

This stage includes steps to transform the selected dataset file into its WEKA dataset representation, remove unnecessary attributes, and construct the training and test datasets.

Data Mining

A decision-tree technique using the C4.5 (J48) WEKA classifier [39] is used to predict the PEP value (YES/NO).

Interpretation/Evaluation

The results are evaluated using a tree visualization technique to display the decision-tree graph model in addition to inspecting the detailed results using a text editor.

5.2 The KDPMEL Prediction Process Program

5.2.1 Process Resources

Figure 3 illustrates the process resources as they are depicted in the Resources Graph.

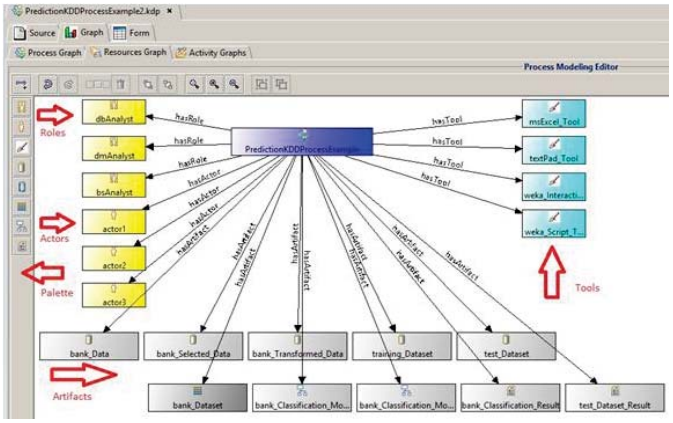


Fig. 3. The Example Process Resources Graph

Artifacts

Figure 4 illustrates the process artifacts as they are depicted in the Artifact Flow Graph.

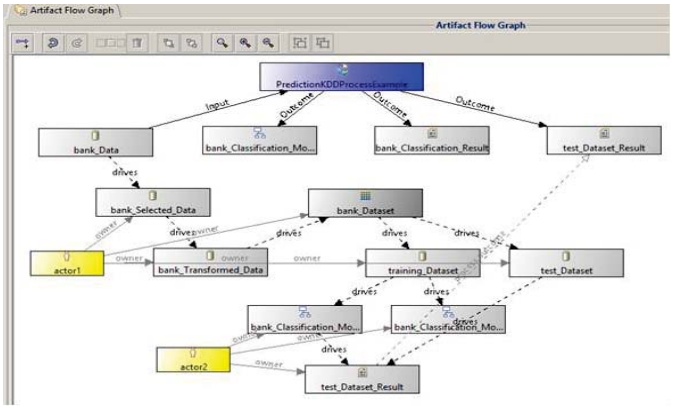


Fig. 4. The Example Process Artifact Flow Graph

Tools

The tools utilized by the process are Microsoft Excel for the data selection tasks, the *WEKA* data mining framework [39] in both the interactive and command-line modes for the Data Preparation, Data Mining, and Interpretation tasks, and the text editor *TextPad* for some of the Interpretation tasks. Figure 5 illustrates the utilized tools.

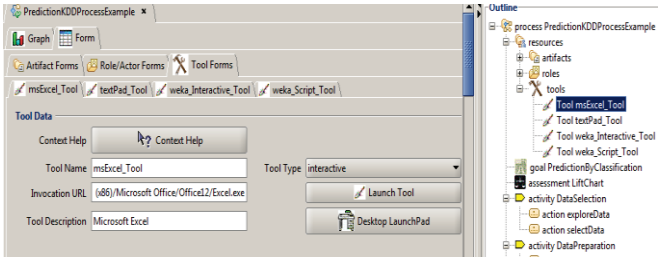


Fig. 5. The Example Process Utilized Tools

Roles/Actors

Three different roles--database, data mining, and business analysts--are fulfilled by three different actors are defined in the program. Figure 6 illustrates the process roles/actors.

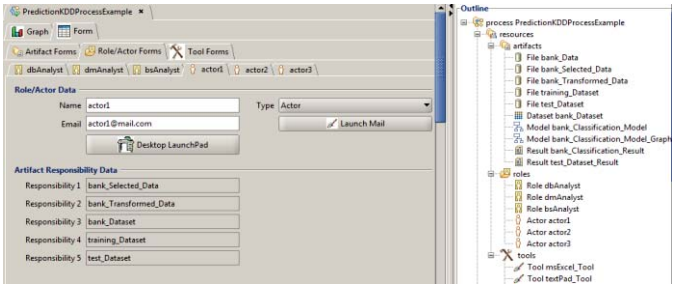


Fig. 6. The Example Process Roles/Actors

5.2.2 The Process Phases

The process includes four phases for *data selection*, *data preparation*, *data mining*, and *interpretation*. Each phase is defined using a KDPMEL *activity* construct as follows:

```
process PredictionKDDProcessExample { …
  activity DataSelection {...}
  activity DataPreparation {...}
  activity DataMining {...}
  activity Interpretation {...}
}
```

Figure 7 illustrates the phases of the process as they are depicted in the Process Graph.

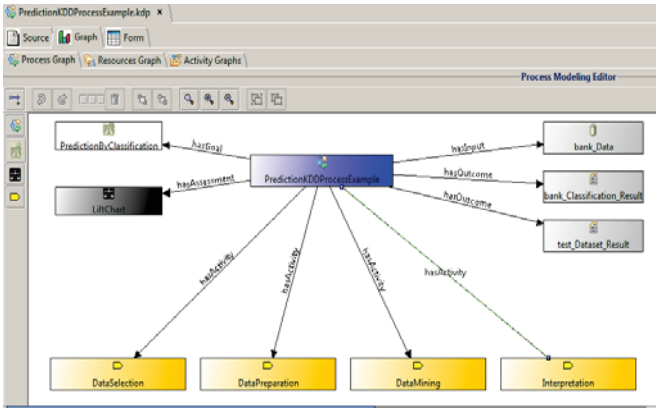


Fig. 7. The Process Graph of the Example Process

Each phase includes a number of generic and specialized tasks that are defined using KDPMEL *action* and *command* constructs. For instance, the *DataPreparation* phase is defined in KDPMEL as follows:

```
activity DataPreparation {
  action transformCsvData { ...
    command transformCSVDataCommand {...}  ...
  }
  action viewTransformedCsvData {...}
  action constructMainDataset { ...
    command transformCSVDataCommand {...} ...
  }
  action viewConstructedDataset {...}
  action buildTrainingAndTestDatasets { ...
    command buildTrainingDatasetCommand {...}
    command buildTestDatasetCommand {...}  ...
  }
}
```

The *DataPreparation* activity (Figure 8) includes a sequence of five tasks for transforming the CSV data to its *WEKA* representation, viewing the transformed data, constructing the main dataset, viewing the constructed dataset, and constructing the training and test datasets.

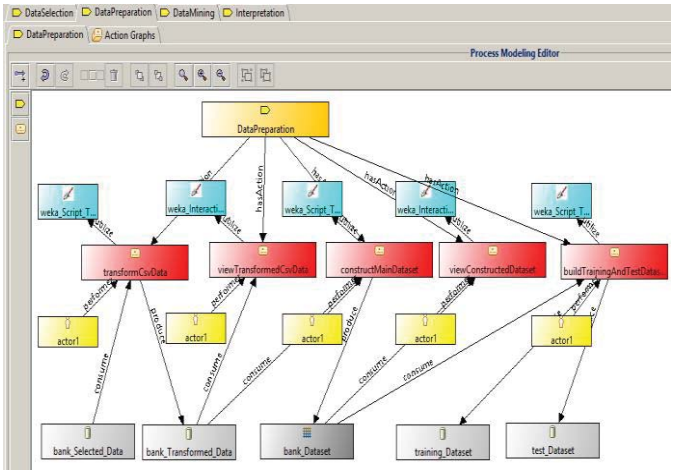


Fig. 8. The Example Process *DataPreparation* Activity Graph

5.2.3 The Process Actions and Commands

The process includes a number of generic and specialized tasks. For instance, the *DataPreparation* activity includes a specialized task that is used to transform the CSV data file to its *WEKA* format. This is defined as follows:

```
action transformCsvData { …
  utilize {
    call weka_Script_Tool {
      command transformCSVDataCommand { …
        input bank_Selected_Data; output …
        operation "weka.core.converters.CSVLoader";
        parameters "";
      }
    }
  }
}
```

5.3 Enacting the Example Process Program

The PCSE-KDD Enactment Engine establishes a process instance and presents it in the Enactment Perspective. The Enactment Perspective displays the overall process execution flow organized by the actors. Each actor is presented with its assigned tasks in a tree view. The description of each task is presented in a form view. A GUI mechanism to apply appropriate transition states (e.g., a start command to execute a posted task) for each task is provided to the actor.

Figure 9 illustrates the enactment of the *transformCsvData* action. When selecting the start command, the Enactment Engine starts the action and identifies its utilized tool and specialized command and offers to invoke them through dialogs. The actor confirms the invocation.

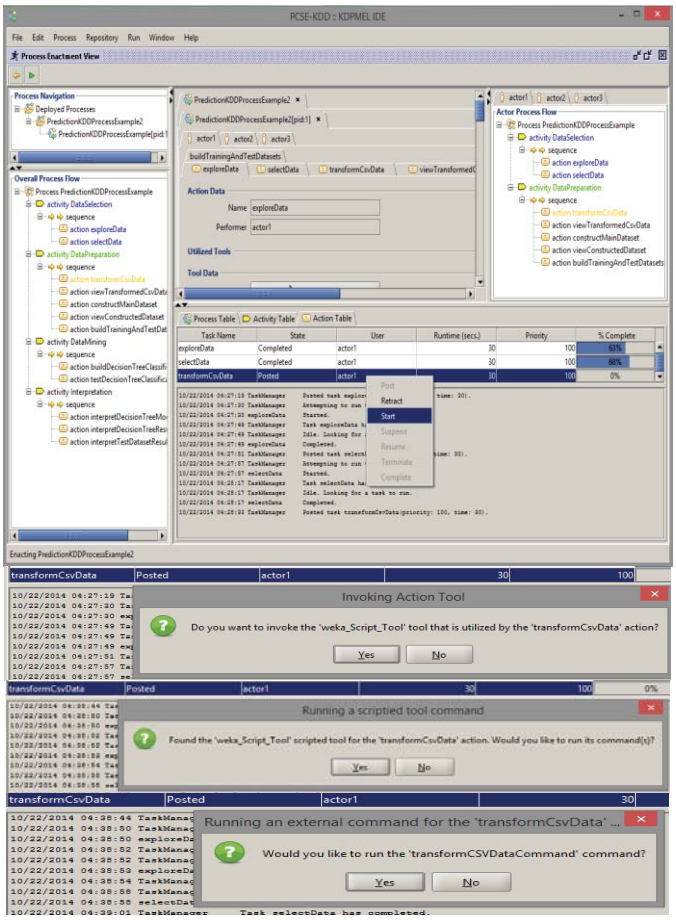


Fig. 9. The *transformCsvData* Action execution dialogs

5.4 Evaluation and Lessons Learned

Our experience using KDPMEL to specify the example KDD process, as well as other KDD processes, supports our first hypothesis that a language-based and process-oriented approach is a flexible and effective approach to precisely and explicitly specify KDD processes as process programs that can be manipulated by programming techniques to reason about the process and support its correct execution.

KDPMEL simplicity goal is achieved by having simple syntax. KDPMEL flexibility goal is mainly achieved by providing various levels for representing tasks at different levels of detail. The generality goal is mainly achieved by not making KDPMEL and PCSE-KDD bound to any particular process models, techniques, or tools. KDPMEL expressiveness goal is achieved by providing constructs to represent both generic and specialized tasks along with their sequencing and dependencies, consumed and/or produced artifacts, utilized tools, and performing actors.

Our experience using KDPMEL and PCSE-KDD to represent and execute the example process supports our second hypothesis that effective support and customized guidance, which depend on the concrete process itself rather than its generic process model, can be achieved by manipulating the explicit representation of the process in

order to manage its various components and support its performance.

6 Conclusions

KDPMEL provides a hybrid modeling approach for specifying KDD processes, mixing different types of editors and views in source-based, graph-based, and form-based styles to allow both technical and non-technical users to participate in the development of KDD processes.

PCSE-KDD is an Integrated Development Environment for KDPMEL. It has been prototyped in Java plus a number of open source libraries and tools. It has the look and feel of Eclipse IDE and has a similar Workbench that includes three different Perspectives, similar to Eclipse's Perspectives, for its Modeling, Enactment, and Management functionalities.

In PCSE-KDD/KDPMEL the process concept is supported and enforced according to a specialized KDD process that includes specific tasks organized according to their sequencing, dependencies, and alternatives. In PCSE-KDD, tools are loosely integrated through a flexible and expandable plug-in mechanism. They are launched automatically and dynamically according to the execution order of the process tasks. PCSE-KDD provides a centralized repository for maintaining and managing the process artifacts. PCSE-KDD employs an engineering approach to develop KDD processes. It is a language-based and process-driven approach. The process is explicitly defined as a program in KDPMEL and manipulated by PCSE-KDD to support its analysis and execution. Specialized user guidance during execution is extracted from the interpretation of the process program. Users are offered their assigned tasks and supported in executing them. In this language-based approach, KDD processes are managed. Their specifications can evolve and executions can be repeated. Moreover, they are validated according to standard programming techniques.

Our future work includes modeling a wide range of KDD processes and increase the level of sophistication of those processes, expanding the support for more detailed KDD artifacts, and continuing the development of KDPMEL and PCSE-KDD to provide more enhanced and expanded graphical modeling to cover the entire process, better user interaction during process enactment, and to expand the integration with a wider range of external process resources.

7 References

[1] Fayyad U. M., Piatetsky-Shapiro, G., and P. Smyth. "Knowledge Discovery and Data Mining: Towards a Unifying Framework". MIT press, Cambridge, Mass., 1996.

[2] Colin Shearer. "The CRISP-DM Model: The New Blueprint for Data Mining". Journal of Data Warehousing, Volume 5, Number 4, 2000.

[3] Graham J. Williams and Zhexue Huang. "Modeling the KDD Process". 1996.

[4] SK Gupta, Vasudha Bhatnagar, and SK Wasan. "A Proposal for Data Mining Management System". Dept. of Computer Science and Engineering, Indian Institute of Technology, 2001.

[5] Padhraic Smyth. "Breaking Out of the Black-Box: Research Challenges in Data Mining". The 2001 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2001.

[6] Rudiger Wirth and Jochen Hipp. "CRISP-DM: Towards a Standard Process Model for Data Mining". DaimlerChrysler Research & Technology.

[7] David Jensen, Yulin Dong, Barbara S. Lerner, Eric K. McCall, Leon J. Osterweil, Stanley M. Sutton, Jr., and Alexander Wise. "Coordinating Agent Activities in Knowledge Discovery Processes". Department of Computer Science, University of Massachusetts Amherst, 1999.

[8] C. Zeng, Y. Jiang, L. Zheng, J. Li, L. Li, H. Li, C. Shen, W. Zhou, T. Li, B. Duan, M. Lei, and P. Wang. "FIU-Miner: A Fast, Integrated, and User-Friendly System for Data Mining in Distributed Environment". In Proc. of the Nineteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013.

[9] Sutton Jr, Stanley M., and Leon J. Osterweil. "The design of a next-generation process language". In Software Engineering—ESEC/FSE'97, pp. 142-158. Springer Berlin Heidelberg, 1997.

[10] Fayyad U. M., Piatetsky-Shapiro, G., and Uthurusamy, R. "Summary from the KDD-03 Panel – Data Mining: The Next 10 Years". The 9th International Conference on Data Mining and Knowledge Discovery (KDD-03), 2003.

[11] Noll, J. and Scacchi, W. "Specifying process-oriented hypertext for organizational computing". Journal of Network and Computer Applications 24, 39-61, 2001.

[12] Lee, H. "Evaluation of Little-JIL 1.0 with ISPW-6 Software Process Example". Department of Computer Science, University of Massachusetts, Amherst, March 1999.

[13] OMG, Inc. "Software Process Engineering Metamodel Specification".URL:http://www.omg.org/technology/documents/formal/spem.htm, Version 1.1, January, 2005.

[14] Abraham Bernstein, Foster Provost, Shawndra Hill. "Toward Intelligent Assistance for a Data Mining Process: An Ontology-Based Approach for Cost-Sensitive Classification". IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, pp. 503-518, April, 2005.

[15] Robert Engles, Guido Linder, and Rudi Studer. "A Methodology for Providing User Support for Developing Knowledge Discovery Applications". URL:http://www.aifb.uni-karlsruhe.de/WBS/publications/

[16] Blake, M. B. and Williams, A. B. "Development and Operational Processes for Agent-Oriented Database Navigation for Knowledge Discovery". In Proc. of the 15th International Conference on Software Engineering & Knowledge Engineering (SEKE '2003), 2003.

[17] Petr Aubrecht, Petr Miksovsky, and Lubos Kral. "SumatraTT: a Generic Data Pre-processing System". 14th

International Workshop on Database and Expert Systems Applications (DEXA'03), 2004.

[18] Kamil Matousek and Petr Aubrecht. "Data Modeling and Pre-processing for Efficient Data Mining in Cardiology". IEEE ITAB'06, Ioannina, October 28, 2006.

[19] S.K. Gupta, V.Bhatnagar, S.K. Wasan, and DVLN Somayajulu. "Intension Mining: A New Paradigm in Knowledge Discovery". Dept. of Computer Science and Engineering, Indian Institute of Technology, 2000.

[20] M.C. Fernandex, O. Delgado, J. I. Lopez, M. A. Luna, et al. "DAMISYS: An Overview". In Proc. of 1st Int'l Conf. on Data Warehousing and Knowledge Discovery, Aug 1999.

[21] R. Meo, G. Psaila, and S. Ceri. "A Tightly-Coupled Architecture for Data Mining". In Proc. of 1st Int'l Conf. on Data Warehousing and Knowledge Discovery, Aug 1999.

[22] Rudiger Wirth et al. "Towards Process-Oriented Tool Support for Knowledge Discovery in Databases". DaimlerChrysler Research & Technology, 1997.

[23] S.K. Gupta, V.Bhatnagar, and SK Wasan. "Architecture for knowledge discovery and knowledge management". Knowledge and Information Systems, Volume 7, Issue 3, pp. 310-336, 2005.

[24] Martin Spott and Detelf Nauck. "Intelligent Data Analysis: Developing New Methodologies Through Pattern Discovery and Recovery". Chapter I: Automatic Intelligent Data Analysis, Copyright ©, 2009, IGI Global, ISBN: 978-1-59904-982-3, 2008.

[25] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. "YALE: Rapid Prototyping for Complex Data Mining Tasks". In Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006.

[26] J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, and F. Herrera. "KEEL: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems". Soft Computing 13:3 (2009) 307-318, 2009.

[27] A. Romei, S. Ruggieri, and F. Turini. "KDDML: a middleware language and system for knowledge discovery in databases". In Data and Knowledge Engineering. Vol 57, Issue 2, pages 179-220, May 2006.

[28] Marban, O., Mariscal, G., Menasalvas, E., and Segovia, J. "An Engineering Approach to Data Mining Projects". Intelligent Data Engineering and Automated Learning – IDEAL 2007, LNCS 4881, pp. 578-588, 2007.

[29] Marban, O., Segovia, J., Menasalvas, E., and Fernndex-Baizn, C. "Toward data mining engineering: A software engineering approach". Information Systems 34 (1), 2009.

[30] Vincenzo Cannella, Giuseppe Russo, Daniele Peri, Roberto Pirrone, and Edoardo Ardizzone. "Towards MKDA: A Knowledge Discovery Assistant for Researches in Medicine". In Proc. of the 10th Conf. of the Italian Association for Artificial Intelligence, pp. 773-780, 2007.

[31] Diamantini, C., Potena, D., and Smari, W. "Collaborative Knowledge Discovery in Databases: A Knowledge Exchange Perspective". AAAI 2006 Fall Symposium on Semantic Web for Collaborative Knowledge Acquisition, 2006.

[32] Diamantini, C., Potena, D, and Panti, M. "Developing an open knowledge discovery support system for a network environment". In Proc. of the IEEE International Symposium on Collaborative Technologies and Systems, pages 274-281, Saint Louis, MO, USA, May 18-19, 2005.

[33] Diamantini, C. and Potena, D. "Representing Service Information in a Collaborative KDD Environment". In Proc. of the International Symposium on Collaborative Technologies and Systems, pages 331-338, Irvine, CA, USA, May 19-23, 2008.

[34] Diamantini, C., Potena, D., and Storti, E. "Collaborative management of a repository of KDD processes". International Journal of Metadata, Semantics and Ontologies, 9(4), 299-311, 2014.

[35] Esmin, A. A., Pereira, D. A., Pereira, M. R., & Araújo, D. L. "SMINER–a platform for data mining based on service-oriented architecture". International Journal of Business Intelligence and Data Mining, 8(1), 1-18, 2013.

[36] DePaul University, Chicago, IL, Classification via Decision Trees in WEKA. URL: http://maya.cs.depaul.edu/classes/ect584/WEKA/classify.html

[37] OMG, Inc., Common Warehouse Metamodel (CWM) Specification, URL: http://www.omg.org/technology/documents/formal/cwm.htm, Version 1.1, March 2003.

[38] Greg Nordstrom, Janos Sztipanovits, Gabor Karsai, and Akos Ledeczi. "Metamodeling - Rapid design and evolution of domain-specific modeling environments". IEEE Engineering of Computer Based Systems (ECBS), Nashville, TN, pp. 68-74, April 1999.

[39] University of Waikato, New Zealand. "Weka 3: Data Mining Software in Java". URL: http://www.cs.waikato.ac.nz/ml/weka/. Version 3.6, 2010.

[40] Open Source, The State Machine Compiler (SMC) Framework, URL: http://smc.sourceforge.net/

[41] Nurdatillah Hasim and Norhaidah Abu Haris. "A study of open-source data mining tools for forecasting". In Proc. of the 9th International Conference on Ubiquitous Information Management and Communication (IMCOM '15), 2015.

[42] Serban, F., Vanschoren, J., Kietz, J.-U., and Bernstein, A. "A survey of intelligent assistants for data analysis". ACM Computing Surveys (CSUR) v.45 n.3, p.1-35, June 2013.

Exploiting temporal patterns of hot events in Weibo

Jiakun Huang[1], **Kai Niu**[2], **and Zhiqiang He**[2]
[1]Department of Information and Communication Engineering,
Beijing University of Posts and Telecommunications, Beijing, China 100876
huangjiakun1991@126.com, {niukai,hezq}@bupt.edu.cn

Abstract— *With explosive growth of the Internet, microblog has become the largest source of public opinion. The propagation of hot events in microblog has drawn much concern. In this study, we extract 218 time series of hot events in 240 million tweets crawled from Sina-Weibo, the biggest Twitter-like microblog in China, and find that the diffusion process is divided into two step. Furthermore, the patterns can be clustered to several centroids by applying the K-Spectral Centroid (K-SC) clustering algorithm. The centroids are quite qualified for demonstrating the different information propagation features in weibo. We also introduce a modified SpikeM model to fit the centroids. Our results demonstrate that the new model describes all the rise and fall centroids with high accuracy, while SpikeM is only capable of fitting the first spike.*

Keywords: time series analysis, information propagation

1. Introduction

The emergence of microblog has dramatically changed the way people access to information. Due to its convenience and real-time property, people are increasingly engaged in sharing and consuming information in microblog services, which turns microblog to a form of online word of mouth branding [1]. In the case of Sina-Weibo, the biggest Twitter-like microblog of China, there exists 1.3 billion registered users and over 150 million monthly active users. So to some extent, weibo has become the dominate source of public opinion in the new media age.

Hot events reflect social opinion and impact the society both positively and negatively in return. The propagation of hot events has been a hot research topic. However, most of the researches focus on modeling propagation process over graph transmitting information from one node to another [2], [3], which are not suitable for large-scale social networks. Few researchers study the temporal dynamics of hot events. Yang et al. [4] propose a time series cluster algorithm K-Spectral Centroid, and discover six patterns of twitter topics. Yasuko et al. [5] introduce SpikeM, which is based on the so-called 'Susceptible-Infected' (SI) [6] model, performing well on fitting the six patterns. It shows that the temporal dynamics of hot events start with an exponential rise and a power-law decay, which is consistent to our observation in real data. But SpikeM is only applicable for the patterns with one spike or additional periodic tails, since it assumes there exists no 'revive' state in the social network.

As far as we know, the previous literature concentrates on modeling the time series of topic mentions. People participating in the discussion of online topics doesn't mean that they are unknown of the information. The periodicity of temporal dynamics directly owns to users' repeatedly participation of discussion. So topic mentions reflect the popularity of hot events, which is not directly related to the information propagation process. On the other hand, the reposting behavior correctly reflect the dynamics of public awareness over time. When a message is published, all the user's followers will have access to it. Secondary reposting behavior transmits the message to user's followers' followers, forming information cascade between disconnected nodes, which will spread to much more audiences. We focus on modeling reposting behavior to figure out the temporal patterns of information propagation.

The main goal of this paper is to discover how the diffusion process of hot events evolves over time, what kinds of temporal patterns are exhibited by weibo, and how to fit the patterns with high accuracy. First of all, our data set and basic statistical findings are introduced in Section 2. Then in Section 3, we use K-Spectral Centroid algorithm to cluster the time series of hot events, revealing that there exists three representative patterns in Sina-Weibo. In Section 4, a modified SpikeM model is introduced, which performs well as for modeling the diffusion process in Sina-Weibo.

2. Statistical Regularities

2.1 Dataset description

To obtain time series of hot events, we crawled more than 250 million tweets during a three-year period from 2012 to 2014. All the tweets are obtained through Sina Open API. Then 218 hot events are manually extracted from the dataset, according to the monthly reported hot events of Sina Weibo Data Center. Each hot event corresponds to an original tweet, with a retweet list filtered from the whole dataset. For the sake of simplicity, we use symbol consisting of a character "#" and a number to represent specified hot event, such as "#1" which is short for "the disappearing of MH370 on Saturday, 8 March 2014".

Table 1 gives several simples of hot events. Every retweet list is sorted by retweet time, but it needs to be quantized to create a time series of the amount of retweets per

Table 1: Four hottest events of Weibo in 2014

Symbol	Description
#1	The disappearing of MH370 on 8 March, 2014
#2	The 2014 Kunming terrorist attack
#3	The famous apology of Wen Zhang over affair
#4	The first Memorial Day of China

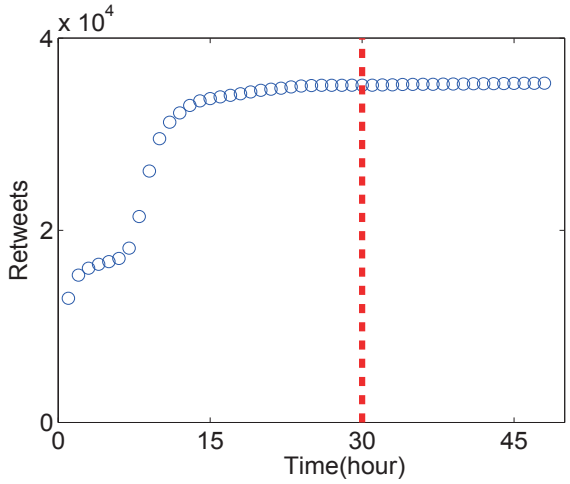


Fig. 1: CDF of #2

unit time interval. In Fig. 1, it shows that the shape of Cumulative Distribution Function has almost no increment after 30 hours, which means the spreading is completed in 30 hours and the subsequent points can be abandoned so as to concentrate on the analysis of preceding variable shapes. Further more, in order to get more variations of the time series shape, we specify the time unit to 10 minutes.

2.2 Findings

Two-stage process. Among the 218 time series, we observe that all these patterns contain two rise and fall spikes peaked at different time points, in which the first one is often much higher than the second one. Fig. 2 shows 4 hottest events in 2013 and 2014, and this unexpected phenomenon is quite different from the finding of six patterns in twitter. In fact, two spikes indicates that the information propagation process is divided to two stages in the life-cycle of hot event. The varying parameter of different hot events is spiking time and the proportion of the first peak and the second peak, indicating that similar diffusion process shares the same temporal pattern.

Causes of the two spikes. For the purpose of figuring out what actually gives rise to the two spikes, we turn to analyze the number of followers in each spike. In weibo social network, the so-called opinion leaders have a major impact on the public opinion, in most cases, and the number of followers is generally a convincing indicator for measuring their significance. In fact, the more followers they have, the more possibility that more people have access to the original message at one point in time. So it is sufficient to just focus on calculating the proportion of users with significant followers. More specially, this proportion is generally very small since the degree distribution social network has a power-law tail, indicating that small changes of the proportion might have huge consequences.

According to the official description, opinion leaders are those who have more than 5 million followers. In order to analysis the different influence of opinion leaders and grass roots, we first divide the number of followers into four levels, 4 to 7, which takes the logarithm base 10, and then calculate the occupation of each level in different spikes. For the most part, as we can see from Table 2, the occupation of users with large number of followers in the first spike, is significantly higher than the second one. As for users with follower count greater than 10^7, who are absolutely authoritative celebrities in Sina-Weibo, there always exists a small proportion in the first stage of diffusion, while in the second one, the proportion is generally zero.

Another special event #4, whose second spike possesses much higher peak value than the first one, is just presenting the opposite case. Furthermore, users with larger number of followers are correspondingly distributed in the higher spike. The above observations are consistent with other hot events in the dataset, strongly suggesting that the first stage of most information propagation process in Sina-Weibo is directly triggered by opinion leader, while the second long-lasting stage is generally caused by the crowd.

This observation is exactly consistent with the so-called Multistep Flow Model [7], which says that most people form their opinions under the influence of opinion leaders, who in turn are influenced by the mass idea. A small fraction of the hot events are exactly the opposite, representing that the information is first introduced by grass roots and propagated in a small scale of the social network, then it is detected by opinion leader which lead to widely spread of the information after several hours. Moreover, the consistency of t_F and t_P also shows that the time of peak point is quite related to the retweet time of users with the largest number of followers, which means opinion leader plays a very important role in the diffusion process.

From the last line of Table 2 we find that for social security events like #1 and #2, the overall retweets in peak 1 is far more than peak2, while entertainment events like #3 and #4 tend to have more retweets in the second peak. This interesting phenomenon indicates that people are more sensitive to events involving social security, and as for entertainment events they tend to have a delayed response.

3. Clustering

In order to figure out typical temporal patterns of hot events in Weibo, we implement the K-Spectral Centroid (K-SC) clustering algorithm to find the clusters.

Table 2: Statistics of the four patterns in Fig. 2. The number of followers is in log scale. $F_1 > 4$:The proportion of users with followers more than 10^4, and so on. P_1:The overall retweets in the first stage. P_2:The overall retweets in the second stage. t_F:The time point when user get the most retweets. t_P:The time point of the maximum peak.

	$F_1 > 4$	$F_2 > 4$	$F_1 > 5$	$F_2 > 5$	$F_1 > 6$	$F_2 > 6$	$F_1 > 7$	$F_2 > 7$	P_1	P_2	t_F	t_P
#1	22.9%	19.9%	4.02%	3.74%	0.71%	0.19%	0.36%	0	63.9%	28.9%	8	8
#2	6.27%	3.58%	1.47%	0.82%	0.35%	0.23%	0.07%	0	62.1%	34.3%	5	5
#3	17.0%	16.5%	2.01%	2.10%	0.16%	0.14%	0.04%	0	41.7%	52.7%	2	3
#4	10.1%	8.57%	1.56%	2.02%	0%	0.48%	0%	0	10.2%	82.8%	57	58

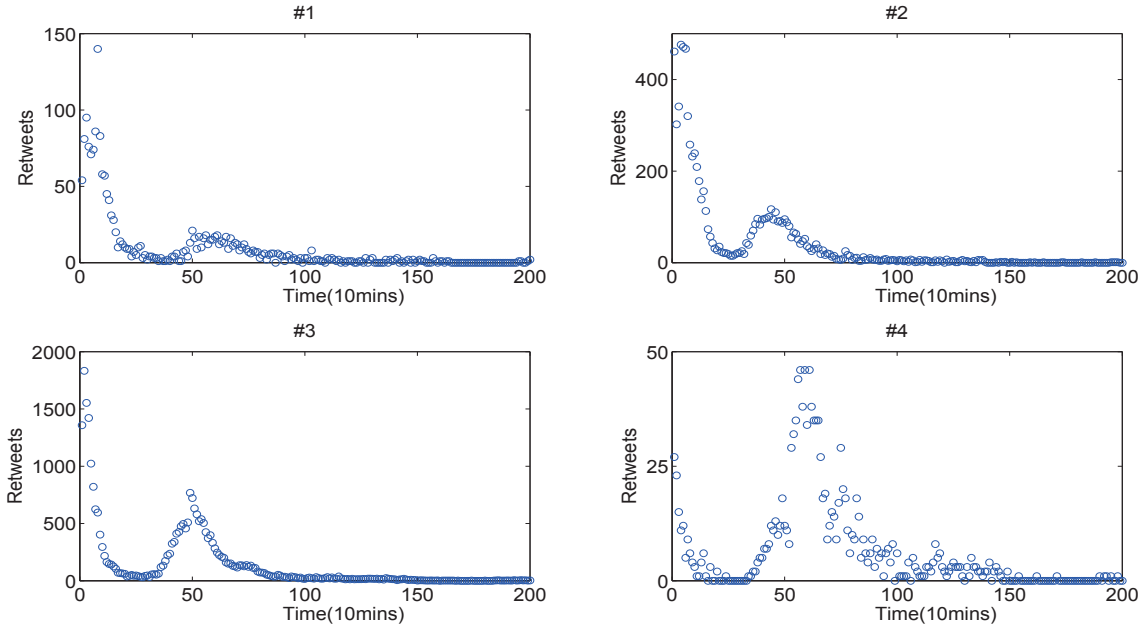


Fig. 2: PDF of the four events in Table 1.

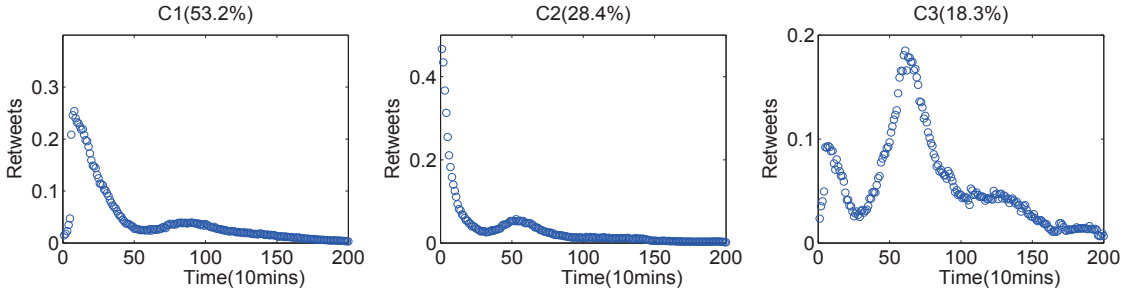


Fig. 3: Clustering results of K-SC. On the top are symbols of each cluster plus percentages of all the time series.

3.1 K-SC

K-SC is an algorithm similar to the classical K-means clustering algorithm, which is mainly comprised of similarity metric and calculation of clustering center. The basic idea of K-SC is iterating a two step procedure, the assignment and the refinement step. In the assignment step, every time series is assigned to the closest cluster by computing the distance between presenting time series and cluster center. In the refinement step, the cluster centroids are then updated. The similarity metric is only related to the shapes of time series by applying scaling and translation. Given two time series x and y, the similarity metric $d(x, y)$ is defined as follows:

$$d(x,y) = \min_{\alpha,q} \frac{\left|x - \alpha y_{(q)}\right|}{|x|} \tag{1}$$

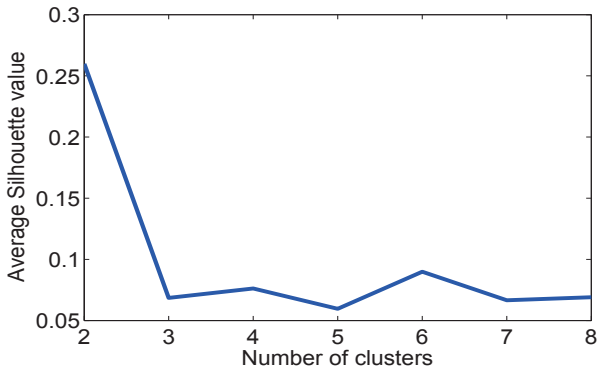


Fig. 4: Average Silhouette of different number of clusters.

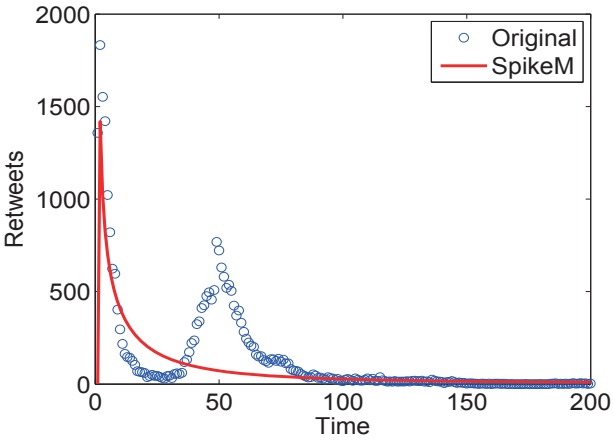


Fig. 5: SpikeM fitting result of #1 with *RMSE=189*.

where $y_{(q)}$ is the variation of time series y by shifting q time units and $|\cdot|$ is the l_2 norm. On the other hand, the new cluster center μ_k^* is updated by calculating virtual center of the cluster C_k, rather than simply averaging every member. It should be the minimizer of the sum of $d(x_i, \mu_k)^2$ over all $x_i \in C_k$:

$$\mu_k^* = \arg\min_{\mu} \sum_{x_i \in C_k} d(x_i, \mu_k)^2 \tag{2}$$

3.2 Experimental Results

As other variants of K-means algorithm, K-SC is also sensitive to the initially specified cluster centers. We use evaluation method Average Silhouette to determine the best number of cluster. Fig. 4 suggests that the Average Silhouette value keeps fluctuating across a fixed value when cluster count is bigger than 3. Empirically we find that the cluster centers are quite stable while setting the number of clusters from 3 to 8. Hence we choose 3 as the best number of clusters.

Fig. 3 shows the three cluster centers, which are represented by C_1 to C_3. As discussed in Section I, the patterns have no periodic trailing, which is totally different from the six patterns in tweeter. Note that the ordinate values are normalized by scaling. Occupying almost half(53.2%) of all the time series, C_1 is supposed to be the most common pattern. Its shape is also a compromised of the three cluster centers, confirming that C_1 is a very typical temporal pattern of hot events in Weibo. It has a brief rising period before reaching the peak, and then follows a pow law decay after peak point. The overall period around the second peak is very long, which means that it takes much long time to get the information widely adopted by the crowd. This matches the reality because in most cases the hot event is initially exposed to a small slice of users, then it is well adopted by the general public through the opinion leaders' significant influence which is corresponding to the rapidly rising period. Soon the propagation process experiences a descending period after the effect of opinion leader, stepping into the second stage. It rises and falls much more gently in the second stage since most users are grass roots with few followers.

C_2 is quite different from C_1 both in the first stage and in the second stage. It doesn't experience a rising period in the first stage. This significantly indicates that the information is directly published by opinion leader. When confronting celebrity gossips, the general public seems to be much more sensitive than usual. So the second peak reaches more quickly and greatly than C_1, and the second stage lasts a shorter period implying high-volatility property.

C_3 is entirely different from the above cluster centers. It represents rare circumstances of diffusion process, possessing only a proportion of 18.3%. According to C_3, the second stage plays a leading role. It has a much higher peak and a much longer lasting period than the first stage. In this case, the original source of information is generally grass roots. The information is initially spread in their small social network, soon it is adopted by opinion leaders due to the increasingly popularity, which in turn creates trend in the entire network.

4. Modeling the Shapes

4.1 SpikeM

SpikeM is a variation of 'Susceptible-Infected' (SI) model, which is the most basic epidemic model. On one hand, it assumes that the infectivity f of a node decays with pow-law distribution:

$$f(\tau) = \beta * \tau^{-1.5} \tag{3}$$

where τ corresponds to the time. Our observation is concordant with this assumption. In Fig. 4 we can see that every pattern has two power-law fall periods. On the other hand, it conditions that the total population of the social network is finite so as to avoid the divergence to infinity. The base

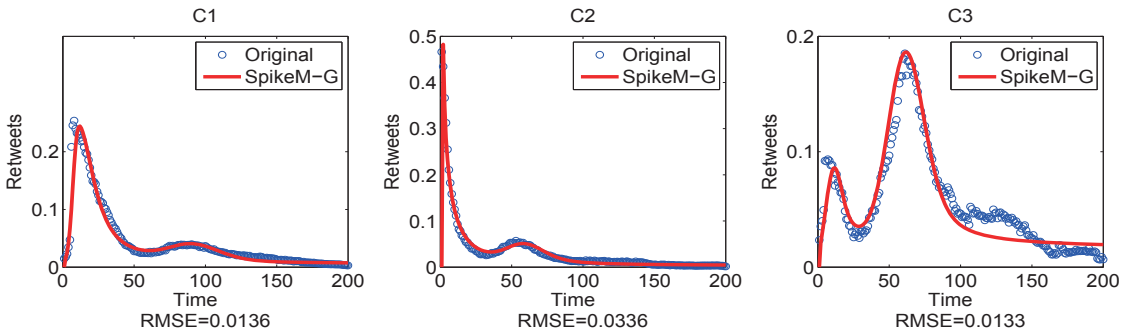


Fig. 6: Fitting results of SpikeM-G. On the bottom of each figure is the RMSE of fitting result.

model is defined by the following equations:

$$\Delta B(n+1) = U(n) \cdot \sum_{t=n_b}^{n} (\Delta B(t) + S(t)) \cdot f(n+1-t) + \epsilon \quad (4)$$

$$\Delta U(n+1) = U(n) - \Delta B(n+1) \quad (5)$$

where $\Delta B(n)$ is the number of retweets at time n, $U(n)$ is the count of un-informed nodes, $S(n)$ is an external shock generated at birth-time n_b.

Although SpikeM model correctly captures the exponential rising period and the power-law decay period, it is only appropriate for the patterns with one spike. Because SpikeM assumes that the diffusion process is consist of only one stage. According to SpikeM, the number of un-informed nodes in the social network keeps dropping after the first peak, neglecting that the general public are unresponsive which will generate the second gently "shock" after several hours. Fig. 5 provides the fitting result of #1. Note that SpikeM model successfully captures the first rise and fall pattern, while the fitting result keep descending at the second period. We also evaluate the fitting accuracy by using the root mean square error (RMSE) metric between estimated values and real values:

$$RMSE = \sqrt{\frac{1}{n} \sum_{1}^{n} (X_{model} - X_{real})^2} \quad (6)$$

As expected, the RMSE of SpikeM is 189, indicating a poor fitting.

4.2 Modified SpikeM

As for the above shortcomings of SpikeM, we propose a modified SpikeM model named SpikeM-G(short for SpikeM with Gaussian Function) based on the following assumptions:

The information propagation of hot events in Weibo is consist of two stages. The first stage generally experiences a short and rapidly spreading period, resulting in a much spiky pattern. Then after a peak-to-trough period, the "sleeping" nodes of the social network begin to "wake up", stepping into the second propagation stage. The accumulation of these nodes relatively generates another external shock. But in this stage the propagation process is much more gently and has a long-lasting period since the "waking up" time of each node is usually not the same.

Macroscopically speaking, there are two cases of the two-stage diffusion process of in Weibo. In the first case, which is more generally, information is propagated from the opinion leader to the crowd. The former plays an important role, while the latter act as audiences. The second case is just the opposite, where information is first published by the general public, and then it is spread to the whole network under the leadership of opinion leaders.

SpikeM only models the first stage of the diffusion process with external shock $S(n)$, so it needs another external shock at the second stage. We use gaussian function here since the rise and fall pattern of around the second peak is much gentle. Our modified model SpikeM-G is governed by the equations:

$$\Delta B(n+1) = U(n) \cdot \sum_{t=n_b}^{n} (\Delta B(t) + S(t) + G(t)) \cdot f(n+1-t) + \epsilon \quad (7)$$

$$\Delta U(n+1) = U(n) - \Delta B(n+1) \quad (8)$$

and $G(t)$ is defined as:

$$G(t) = a \cdot e^{-w(t-t_p)} \quad (9)$$

where a is the volume of the second peak, t_p is time point of the second peak. The term $G(t)$ is very important. It models both the overall volume and the lasting period of the second stage. It also ensures the power-law decaying pattern, since it is multiplied by the infectivity function $f(\tau)$ outside the brackets.

Fig. 6 describes the results of SpikeM-G fitting on the three typical clustered temporal patterns. On the bottom of Fig. 6 displays the RMSE of each fitting result. In this figure, we can see that SpikeM-G is quite consistent with the previous two assumptions. Firstly, it successfully characterizes the two stages of information propagation process in Weibo where SpikeM model fails. On the other hand, whether information is propageted from opinion leader to

the crowd or the opposite, SpikeM-G is capable of correctly capturing the temporal pattern.

5. Conclusions

In this paper, we study the temporal patterns of hot events in three steps. Firstly, we analysis the statistics of all the temporal patterns, figuring out two basic fundamentals. On one hand, the information propagation of hot events in Weibo is comprised of two stages. On the other hand, we find out who actually contributes to the spike of each stage by analyzing the number of followers. Then the three typical temporal patterns of hot events are uncovered by implementing the KSC clustering algorithm. Finally, we introduce SpikeM-G which is based on SpikeM to get better fittings of the patterns. The experimental results show that our method performs well as for capturing the shape and achieving high accuracy.

This study helps to figure out who actually promotes information diffusion process in social network, which will contribute to the effectiveness of viral marketing. In order to produce increases in brand awareness, the viral campaign can be divided to opinion leader advertising stage and grass roots advertising stage. What's more, the study also provides a new access to public opinion monitoring since the spreading process is predictable.

References

[1] B.J.Jansen, M.Zhang, K.Sobel and A.Chowdury, *QMicro-blogging as Online Word of Mouth Branding*, CHI EA '09, pp. 3859–3864, Apr. 1977.

[2] T.Lou and J.Tang, *Mining Structural Hole Spanners Through Information Diffusion in Social Networks*, WWW'13, pp. 825–836, May. 2013.

[3] Z.Yin and W.Chen, *Discovering Patterns of Advertisement Propagation in Sina-Microblog*, ADKDD'12, Aug. 2012.

[4] J.Yang and J.Leskovec, *Patterns of Temporal Variation in Online Media*, WSDM'11, pp. 177–186, Feb. 2011.

[5] Y.Matsubara and Y.Sakurai and B.A.Prakash, *Rise and Fall Patterns of Information Diffusion: Model and Implications*, KDD'12, pp. 6–14, Aug. 2012.

[6] M.E.J.Newman, *The structure and function of complex networks*, KDD'13, pp. 6–14, Mar. 2013.

[7] Elihu Katz, *The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis*, Public Opinion Quarterly, vol. 21(1), pp. 61–78, 1957.

SESSION

REGRESSION AND CLASSIFICATION

Chair(s)

Drs. Robert Stahlbock
Gary M. Weiss

Efficient Classifier over Stream Sliding Window using Associative Classification

K.Prasanna Lakshmi[1], **Dr.C.Ramesh Kumar Reddy**[2]
[1]Department of Information Technology, GRIET, Hyderabad, Telangana, India
[2]Department of Computer Science, CBIT, Hyderabad, Telangana, India

Abstract- *Prominence of data streams has dragged the interest of many researchers in the recent past. Research is going in the direction of formulating association rules on data streams for the purpose of prediction. From among the classification techniques the associative classification mining stands out with better performance over former classification techniques. A new technique is introduced through this paper, which takes the advantage of the associative classification for mining of data streams. To the best of our knowledge there are only a few techniques existing in the domain of data streams. We have designed a compact data structure to maintain data stream efficiently without losing important information. We present a PSToSW for mining rules from the tree. Subsequently, an optimized algorithm called PSToSWMine is proposed for mining a classifier which contains set of high qualified classification rules. We have conducted experiments both on synthetic and real data sets for the purpose of assessing the performance of our approach. The results that we arrived at prove that our approach is superior to existing algorithms in terms of accuracy of prediction and run time efficiency.*

Keywords: Data Streams, Associative Classification, Frequent Item sets, Prediction

1 Introduction

Data stream mining deals with gaining knowledge from the stream of data. Most of the recent applications involve processing of large volumes of data, flowing in continuously [11]. To take a few examples, web click streams, financial transaction, science surveillance data etc., Given the nature of data, mining of these data streams necessitate a real time response after analysis. This also means that the technique applied should be capable enough to process the data quickly as it should be read in a single pass and produce the results [11]. Sequential access methods for stream mining are cost effective and better than random access methods. As mentioned earlier, stream mining is applicable for applications of large data sets, this would be impractical to store on main memory and hence stored on secondary storage devices. Data sets such as sensor data, router packet statistics are temporal and need not be stored in disk; these must be processed and discarded. And as the size of these data sets increases far beyond the space available to an algorithm, it is not possible for the streaming algorithm to remember too much of data scanned in the past. In order to mitigate the challenges posed by the situation mentioned above, there is a need to design algorithms that store summary of past data, so as to make memory available for processing future data. The quintessence of the algorithm for data streams would be to examine each data element at most once in least possible time and occupy minimum memory space for storage.

Association and Classification are two useful and ubiquitous tools in data analysis. Mining based on association is concerned with extracting correlated features shared among transactions of data streams. These algorithms give the statistical relationship between items without giving significance of items [4]. On the other hand, classification uses class attribute in construction of classifier. The classifier needs the significance of items for predicting the class label. Integration of these two methods will provide efficient associative classifier [13], [1]. We study the associative classification in the stream context and provide a streaming algorithm with performance guarantees. Associative classifier predicts class from rules generated using association for unseen stream of data. Compared with existing classification techniques, classification based on association gives more accurate results due to better classifier. Rules containing class information are stored in classifier. These rules are generated from frequent pattern mining concept of association. So, frequent pattern mining plays an important role in associative classification. Many frequent pattern mining techniques exist currently and many more efficient techniques will evolve in future as these have direct impact on performance of associative classification.

Moreover, classifying data streams using this technique is a newly explored area of research [12]. Due to inimitable features of streaming data, it is not possible to simply apply the algorithms designed for static datasets to data streams. Challenges posed on associative classification of data streams include working with limited memory, processing

data at a glance, concept drift and improving accuracy of classification. Many researchers have devoted their efforts to frequent item set mining on data streams as this method is used for feature selection and classifier construction.

Time windows are commonly used for handling data streams [6], [2], [7]. Based on application, landmark, sliding and damped windows can be used. A landmark window is divided into many windows and the data in these windows is used as updating units. As the name suggests, in sliding window only a fixed number of data elements present in recent window can be used for mining. In applications where all historical data is needed with more weight age on recent data than older data then damped windows are preferred.

Algorithm for mining classifier containing associative rules over sliding window for data streams will be very useful for classifying unseen data. In this paper, we propose a new algorithm PSToSWMine, for associative classification mining over data streams from sliding window. A new storage structure called PSTree [10] which is already proposed by us is adopted. It dynamically restructures to reflect the growth of item sets frequencies over time. Intensive study shows that our proposal is efficient and attains high classification accuracy.

The work we carried out is briefed out below:

- We created a compact data structure called PSTree [10] to maintain the relevant and current information.
- We devise an algorithm for mining frequent item sets with class labels from a streaming data within sliding window.
- We defined an algorithm PSToSW, to directly extract rules for classification on sliding window.
- Performed experiments and found PSToSWMine to have achieved better accuracy than other algorithms designed for the similar task.

Rest of paper is organized as follows. We discuss related work in the next section and give problem definition in Problem Statement Section. Proposed algorithm PSToSWMine along with PSToSW is discussed in next section. The empirical results are shown in Experimental Analysis section and finally we conclude in last section.

2 Related Work

The problem of Associative classification is to find a subset of rules which satisfy supports and confidence. An Associative Classification approach called HARMONY algorithm [8] directly mines k best rules for each transaction and uses these for building a classifier. HARMONY uses an instance-centric rule generation to discover the highest confidence discovering rules.

Another algorithm called DDPMine [5] uses sequential covering paradigm for constructing classifier. DDPMine tries to find the best discriminative rules from those transactions which have not been covered and removed and finds locally optimal rules.

STREAMGEN algorithm [3] constructs an enumeration tree for each sliding window and mines a set of item set generators for classification. This algorithm directly mine a set of high quality classification rules over stream sliding windows while keeping high performance When compared to DDPMine and Moment, the accuracy of prediction of StreamGen algorithm is on the higher side.

Classifying a data stream with an associative classifier is a newly explored area of research. There is no algorithm which accurately mines a set of frequently generated rules for classification by taking less amount of time.

Recently another algorithm called AC-DS [12] is proposed as an associative classification algorithm for data streams which works by using support threshold and land mark window model. AC-DS uses single rule for predicting a new data stream. This is biased on general rule, and not appropriate for streams that are slowly changing from time to time. This algorithm works well with single concept. If the concept function is a concept drift one then the algorithm will not output an accurate result.

Motivated by these, we proposed *PSToSWMine* which improves the efficiency of mining in terms of accuracy of prediction and time consumed for prediction. A new data structure called *PSTree* is developed for online incremental maintenance of data. Because the focus of the paper is on building classifier using associative classification over data streams with sliding window, we mainly compare our approach with StreamGen and DDPMine algorithms.

3 Problem statement

Let data stream D_S be a set of instances I which are grouped under batches. Each batch contains equal number of instances. Each instance in a batch B contains set of values for attributes and class label value. Instance i is represented as $< id, A, y >$ where $i \in I$, id being instance identification number, A is set of normal attributes present in instance i and y is class label. An item set S is present in I if $S \subseteq I$ holds. The number of instances containing item set S is support count of S denoted as $supCount_S$. A common rule is shown as $A \rightarrow y$ where A is set of normal attributes and y is class label attribute. The quality of a rule generated is measured using minimum support denoted as sup_{min} and $conf_{min}$. The rules which do not satisfy these thresholds are called infrequent rules which are rejected and rules satisfying these are used for constructing classifier.

Given sup_{min} we have following definitions.

Definition 1. Current length of data stream is given as $DSL = \mid B_1 \mid + \mid B_2 \mid + \ldots\ldots + \mid B_m \mid$ where $Bj = \{I\}$ in which I is set of instances of stream and j is batch number

Definition 2. Item set S is frequent if it satisfies minimum support. That is, $\text{supCount}_S \geq \text{sup}_{min}$.

Definition 3. A rule is of form $A \rightarrow y$ where $A \subseteq I$ and $A \cap \{y\} = \phi$.

TABLE I
TRAINING DATASET

ID	Attribute A_1	Attribute A_2	Attribute A_3	Class
1	a_1	a_2	b_3	y_1
2	a_1	a_2	c_3	y_2
3	a_1	b_2	b_3	y_1
4	a_1	b_2	b_3	y_1
5	b_1	b_2	a_3	y_2
6	b_1	a_2	b_3	y_1
7	a_1	b_2	b_3	y_1
8	a_1	a_2	b_3	y_1
9	c_1	c_2	c_3	y_2
10	a_1	a_2	b_3	y_1
.	.	.	.	.
.	.	.	.	.

W—Sliding Window

Compared with traditional associative classification, associative classification mining over data streams must be an incremental task. The important task of our work is to find complete set of frequent rules from most current sliding window of *I* instances of data stream. Algorithm for gaining knowledge quickly and accurately is also needed. Table 1 depicts an example of data stream with sliding window size of 2 batches where each batch contains 2 instances of data stream for associative classification.

4 Sliding Window

Stream data processing is done using landmark window [11], damped window [11] or with sliding window [11]. In a Sliding window model, discovery of knowledge is achieved using a fixed number of recently generated data stream. For example, given a window of size N over data streams, only latest |N| transactions of stream or all transactions in the last |N| time are used for knowledge gaining. As a new transaction arrives the oldest transaction in the window expires [16]. Representation of sliding window is shown in Fig. 1.

Fig. 1. Sliding Window.

Advantages of using sliding window model are

- It is well defined and easily understood
- It is deterministic
- It emphasizes on recent data

There are two types of sliding windows

1. Transaction sensitive sliding window
2. Time sensitive sliding window

The sliding window model is therefore widely used to find recent frequent patterns in data streams [3], [6], [7].
The proposed PSToSWMine works over transaction sensitive sliding window.

5 Prefix streaming tree over sliding window mining framework

PSToSWMine is a learning classification model based on frequent pattern mining.

The framework for PSToSWMine contains three phases:

1. Representation of stream in a compact data structure called PSTree [10].
2. Frequent item set mining and feature selection called frequent rules using PSToSW and
3. Model learning phase.

Framework for associative classification is built in two phases.

- In the first phase, classification rules are discovered from training dataset using frequent item set concept of association. The right–hand-side of the rules is restricted to class label. Rules are represented as $X \rightarrow C$ where X is an item set and C is a class label.
- In the second phase, pruning techniques are applied for generating high quality rules for building accurate classifier. Pruned association rules were used to form classifier based on confidence.

The methodology used is illustrated in Fig. 2.

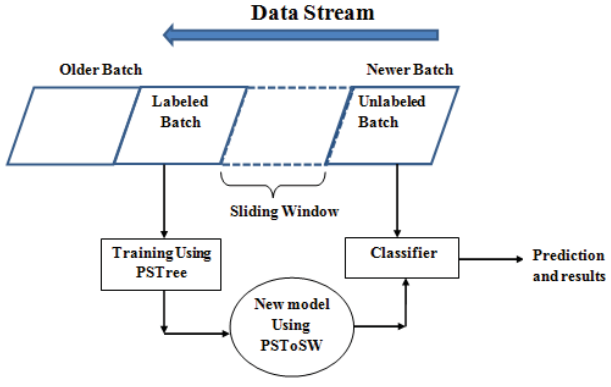


Fig. 2. Stream Mining approach using Associative classification over sliding window.

First, we present some common properties which are used in algorithm design. Then, we introduce the compact tree *PSTree*. Later, we show the construction of model for

learning called classifier. This is build based on conditional pattern base used in FP-Growth mining.

Some properties used in this paper are

Property 1. A frequent item set S is used in classifier if it meets the minimum confidence threshold.
Property 2. Given a classifier M, any subset of M would also be a classifier.
Property 3. Given an unpromising item set S, any superset of S must be either unpromising or infrequent.
Property 4. For a new instance of data stream the state of set of frequent item set S many change depending on frequency of new item sets in the instance.

5.1 PSTree

PSTree, the structure used to store data, is based on principles of prefix tree [16] which is an ordered tree. It represents instances flowing through sliding window in a highly compact form. Each read instance is inserted into the tree in a path. As it is possible for multiple instances to have the same items, their path in the tree is overlapped. Due to the overlapping, the compactness of the tree is enhanced. To facilitate the concept of sliding window and tree updating with new instances, each window *W* is decomposed into number of fixed size batches of instances called a batch *B*. Window slides batch by batch.
PSTree is constructed using FP-tree concept for inserting instance into the tree. Creation of this compact tree happens with the help of three stages.

1. Insertion stage
2. Restructuring stage
3. Refreshing stage

Initially the PSTree is empty. After receiving a new instance from a batch of data stream it is inserted into PSTree according to an order which is maintained in *I-List*. The order is based on support count of items. Later, after complete insertion of instances present in current window, the tree is restructured to maintain compactness. It is done based on sorted list called I_{sort}-*List*. This I_{sort}-*List* is created by sorting the items present in *I-List* based on their support count. For restructuring PSTree, we used Branch sorting method [9], [10], [16].
The window slides if the size of current window exceeds the user specified value for window size. Before sliding, algorithm performs refreshing stage by extraction of older batch information to maintain current information of data streams. During next insertion of instances of second window, the item details are maintained using I_{sort}-*List*. Batch information of instances is maintained in tree by using *batch-counter*. This information is stored in leaf node of every path along with class label information of the tree.
The methodology of constructing PSTree for data streams through sliding window is illustrated in Fig.3. Fig.3 (a) shows the initial tree which is empty. Fig.3 (b) depicts the insertion of two instances represented as first batch in data streams. Fig.3 (c) illustrates the restructuring step using I_{sort}-*List*. Fig.3 (d), (e) shows the same for second batch of instances. This is repeated till the last stream of data.

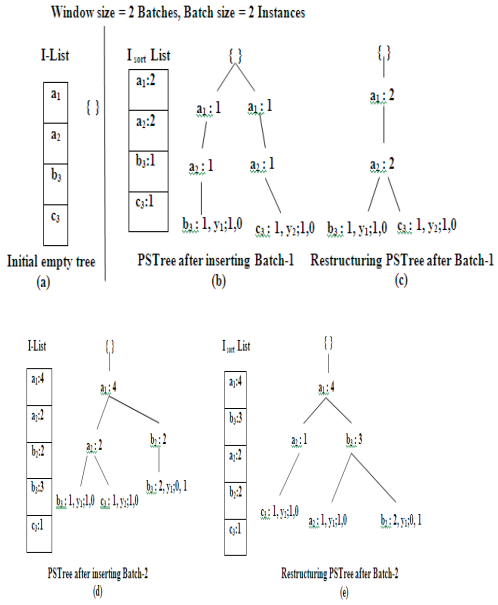


Fig. 3. PSTree Construction

The algorithm used for constructing and maintaining *PSTree* is depicted below along with two methods used for insertion and restructuring the tree.

ALGORITHM 1. Construction and Maintenance of PSTree

Algorithm **Construct**
Input: Data Stream *DS* where each record contains *N* items, *W*-window-size, *B*-Batch Size, *I*-List
Output: PSTree for the window

Begin
P ← *0;*
T ← *tree with null as initial value;*
CI-List ← *I-List;*
while*(P ≠ W)* **do**
call *Insert_Batch(T);* *// Insertion stage*
CI-List ← *Sort_Order;*
call *Restructure(T, CI-List);* *// Restructuring stage*
P=P+1;
end while
end

Algorithms for insertion and restructuring are presented in [17]. Insertion and Restructuring steps are repeated sequentially for all successive batches till the end of data stream. If the batches B_{i-1}, B_i are currently present in window W_j then first insertion step followed with restructuring step for these batches is performed. Later, when window W_j slides to W_{j+1} containing batches B_i, B_{i+1} the same two steps are repeated. While inserting the new batch B_{i+1}, the oldest batch B_{i-1} is deleted by changing the batch number. Time

complexity of insertion step is *O(mn)* and for restructuring step it would be $O(n\log_2 m)$ where *m* is number of items in a transaction and *n* is number of transactions.
PSTree is refreshed before every slide of window in order to provide an environment which helps to mine exact content from the current window. Upon sliding a window the first value in *Batch-counter* in each *leaf node* and same value from *support count* value of each node up to the root in the path are removed, and the remaining values in the list are moved left by one position which shows that the earlier batch is expired.

5.2 Generation of Classifier

The technique used for generation of classifier uses two thresholds given by user as input called minimum support sup_{min} and minimum confidence $conf_{min}$. The generated rules contain item sets and class label which are denoted as $X \rightarrow c$, where $X \subseteq$ frequent item sets and $c \subset$ class label. The generated rules are first arranged in an order based on confidence, support and length of generated rule. Ordered rules are pruned using statistical method called chi-square testing (χ^2). This measure helps in testing correlation among rules [15]. The rules which satisfy this testing are used in construction of classifier.

5.3 Learning Model

For the purpose of predicting unknown data, the classifier acts as a model. The mining operation is efficient due to frequency descending prefix structure. The rules found in model are globally optimal. The classifier build using *PSToSW* tend to have better accuracy in classification. For predicting test data *t* exactly only those rules $X \rightarrow c$ matching *t* i.e., $X \subseteq t$ are selected**.** This algorithm maintains recent information from the data streams. For predicting a new tuple for class label recent information is not sufficient. For doing this the *PSToSW* must be converted into an incremental algorithm. As the insertion and refreshing stages are independent it is very easy to convert the *PSToSW* into an incremental algorithm. As these two stages are not related, they can be easily combined depending on the specific type of application in *PSToSW*.

Incremental *PSToSW* contains only two stages.

1. Insertion stage
2. Restructuring stage

PSToSW without refreshing stage generates all frequent rules of recent window for classifier. *PSToSW* with refreshing stage generates a classifier containing all frequent rules collected from entire data stream. The algorithm used for mining data streams using *PSToSW* is shown below

ALGORITHM 2. PSToSW mining for a window

Input: *min_sup, min_conf, Data Stream DS where each record contains N items, W-window-size, B-Batch Size, I-List*

Output: *Classifier for the window*

Begin
call Construct *for constructing and restructuring PSTree*
Generate frequentPatterns containing Class label which satisfy min_sup
Build classifier with Rules satisfying min_conf
end

6 Experimental Analysis

In this section we compare the performance and classification accuracy of our incremental *PSToSWMine* algorithm against several existing algorithms. Incremental *PSToSWMine* mines rules from real datasets and synthetic datasets by considering window size as two batches where each batch is half the size of data stream. All programs are written in Java and run on windows XP on a 2.53GHz Intel PC with 1.0GB of main memory.

Real Datasets. Real datasets are from UCI Machine Learning Repository [14] and Intel Berkeley Research Lab. The important characteristics of these datasets are listed in Table 2.
Sensor Stream : The data set contains information collected from 54 sensors. It contains information about temperature, humidity, light and sensor voltage. Sensor ID is used as class label, so the task of mining this stream is to correctly identify the sensor ID.

TABLE II
REAL DATA SETS CHARACTERISTICS

Dataset	Number of Transactions	Number of Attributes	Number of Classes	Number of Items
adult	48842	14	2	128
breast-w	699	10	2	29
horse	368	28	2	61
hepatitis	155	19	2	33
mushroom	8124	22	2	116
pima	768	8	2	15
Sensor stream	2,219,803	5	58	--

Synthetic Data Streams. We generated synthetic data streams using MOA (Massive Online Analysis) whose characteristics are listed in Table 3. These streams are approximately 80MB in size, consisting of 1 Lakh to 100 Lakhs transactions. All these datasets are very widely used for evaluation of associative classification.

TABLE III
SYNTHETIC DATA SETS CHARACTERISTICS

Dataset	Number of Transactions	Number of Attributes	Number of Classes
Stagger Generator Stream	100,00,000	4	2
Hyper Plane Generator Stream	1,00,000	10	5
Agarwal Generator Stream	1,00,000	10	2
Random Tree Generator Stream	1,00,000	5	2
Sea Generator Stream	1,00,000	4	2

6.1 Accuracy

To our knowledge, currently there are only few existing algorithms which mine classifier for classification over a data stream using sliding window. Table 4 shows the accuracy comparison of *PSToSWMine* with *StreamGen* Rules [3] and *DDPMine* [5] with minimum support threshold of 1 percent, minimum confidence threshold of 50 percent. These two methodologies perform similar tasks as *PSToSWMine* does.

Comparison was done using six datasets. It is seen that the *PSToSWMine* gives better accuracy than *StreamGen* by an average percent of 5.63. It even excels *DDPMine* by an average accuracy of 7.11. Methodology which attains highest accuracy is shown in bold font. Fig.4. depicts the accuracy comparisons of these algorithms for various datasets. The entire study shows that *PSToSWMine* outperforms both the classifiers in terms of accuracy.

TABLE IV
ACCURACY COMPARISON

Dataset	StreamGen	DDPMine	PSToSW
Adult	82.1	81.29	79
breast-w	96.7	95.28	96.15
Horse	81.51	81.24	92
Hepatitis	82.0	76.98	100
mushroom	98.91	97.18	100
Pima	74.81	75.12	80.31
sensor stream	---	---	100
Average Accuracy	86.0	84.51	91.24

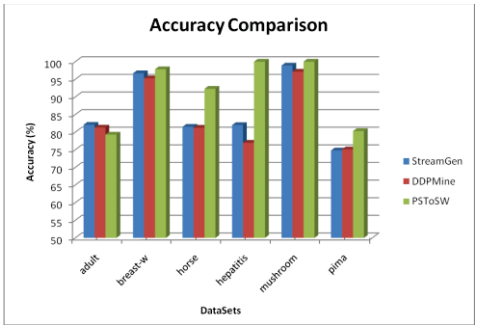


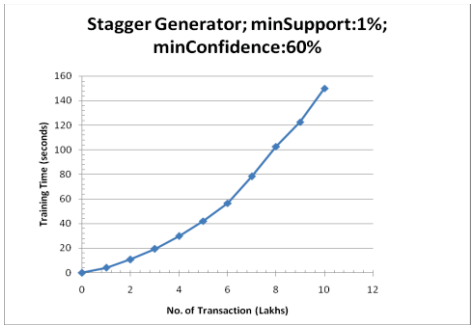
Fig.4. Accuracy Comparison

6.2 Runtime Efficiency

We have conducted experiments for evaluating the runtime efficiency of *PSToSW* with *STREAMGEN* [3] and with *DPMine* [5]. Table 5 shows the time taken for construction, restructuring and prediction in *PSToSWMine*. Fig.5 (a) depicts the plot between training time and number of transactions for Stagger generator. Fig.5 (b) plots the prediction time against number of transactions for Stagger data stream.

TABLE V
RUNTIME DISTRIBUTION IN SECONDS

Data streams with minimum support and confidence	Number of Transactions	Tree Construction Time	Tree Restructuring Time	Prediction Time	Total Time
Real Time Datasets					
Adult min_sup=1, min_conf=50%	48842	49	9	54	112
mushroom min_sup=10, min_conf=50%	8124	770	769	61	1600
Sensor Stream min_sup=0.1, min_conf=50%	1,00,000	20	11	62	93
Synthetic Datasets					
StaggerGenerator min_sup=1, min_conf=50%	100,00,000	60	0.001	16	76.001
Hyper Plane Generator min_sup=1, min_conf=50%	1,00,000	15	6	43	64
Agarwal Generator min_sup=1, min_conf=50%	1,00,000	4795	42	26	4863
Random Tree Generator min_sup=1, min_conf=50%	1,00,000	714	1.9	0.8	716.7
Sea Generator min_sup=1, min_conf=50%	1,00,000	99	50	12	161



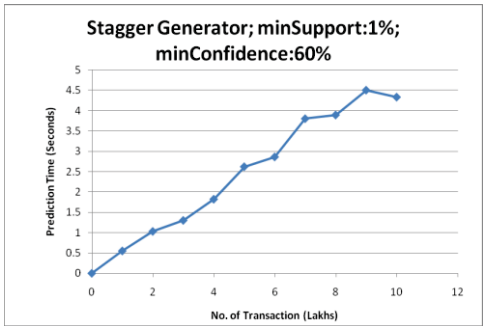


Fig.5 (a) Training Time when varying number of transactions
(b) Prediction Time when varying number of transactions.

Fig.6. shows that *PSToSW* takes less time for generating frequent item sets when compared with StreamGen. The plot is between various support thresholds and time consumption.

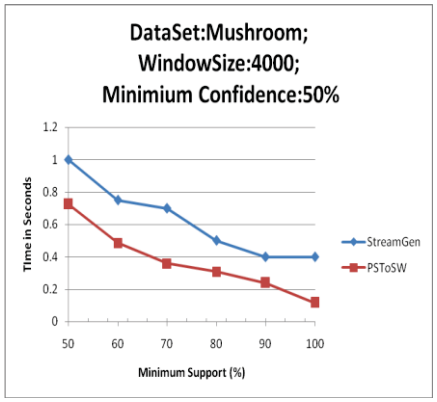


Fig.6. Runtime Comparison of PSToSW with StreamGen

7 Conclusion and Future Work

In this paper we introduced an associative classification algorithm called *PSToSWMine* for data streams. Dynamic tree restructuring is a novel concept that we used to handle streaming data. *PSTree* construction algorithm uses this technique for achieving a highly compact prefix structure within a single pass on a sliding stream. As it is fast in updating, *PSToSWMine* is the apt algorithm for mining data streams. Despite of restructuring cost, *PSToSWMine*'s overall runtime cost is much less than any one of the existing algorithms. Experimental results show that the proposed mining technique increases the classification accuracy due to the availability of large rule sets. By implementing a statistical technique, chi-square testing, the process of rule generation for classifier has been greatly enhanced. This technique shuns information loss and generates the complete non-redundant rule set needed by the classifier. As a future work, we plan to improve the performance of *PSToSWMine* by reducing the number of rules generated without affecting the accuracy of mining.

8 Acknowledgment

The authors would like to thank the reviewers for helpful comments

9 References

[1] B. Liu, W. Hsu, and Y. Ma, "Integrating Classification and Association Rule Mining," Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining (KDD '98), Aug. 1998.

[2] C.K.S. Leung, Q.I. Khan, DSTree: a tree structure for the mining of frequent sets from data streams, in: Proc. ICDM, 2006, pp. 928– 932.

[3] Chuancong Gao, Jianyong Wang, "Efficient item set generator discovery over a stream sliding window" in C IKM'09, November 2009, Hong Kong, China, ACM 978-1-60558-512-3/09/11

[4] Hong Yao, H.J Hamilton (2006)," Mining item set utilities from transaction data bases", IEEE Transactions on Data and Knowledge Engineering, volume 59, issue 3, pp.603-626.

[5] H. Cheng, X. Yan, J. Han, and P. S. Yu. Direct discriminative pattern mining for effective classification. In Proceedings of the 24th International Conference on Data Engineering, pages 169–178, Cancun, Mexico, 2008. IEEE.

[6] J.H. Chang, W.S. Lee, estWin: Online data stream mining of recent frequent item sets by sliding window method, Journal of Information Science 31 (2) (2005) 76–90.

[7] J. Li, D. Maier, K. Tuftel, V. Papadimos, P.A. Tucker, No pane, no gain: efficient evaluation of sliding-window aggregates over data streams, SIGMOD Record 34 (1) (2005) 39–44.

[8] J. Wang and G. Karypis. On mining instance-centric classification rules. IEEE Trans. Knowledge Data Engineering 18(11):1497–1511, 2006

[9] Koh, and Shieh, 2004. An efficient approach for maintaining association rules based on adjusting FP-tree structures. In Proc. of DASFAA 2004. Springer-Verlag, Berlin Heidelberg New York, 417–424.

[10] K.Prasanna Lakshmi, Dr.C.R.K.Reddy, "Compact Tree for Associative Classification of Data Stream Mining", IJCSI International Journal of Computer Science Issues, Vol 9, Issue 2, No 2, March 2012, ISSN(online) : 1694-0814

[11] K.Prasanna Lakshmi, Dr.C.R.K.Reddy, "A Survey on Different Trends in Data Streams " pp.451-455, In Proc of 2010 IEEE International Conference on Networking and Information Technology, (ICNIT'10), 2010. ISBN : 978-1-4244-7577-3.

[12] L. Su, H. Liu and Z. Song, "A New Classification Algorithm for data stream". I.J.Modern Education and Computer Science, 4, 32-39, 2011.

[13] W. Li, J. Han, and J. Pei, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules," Proc. IEEE Int'l Conf. Data Mining (ICDM '01), Nov. 2001.

[14] R. C. Agarwal, C. C. Agarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent item sets. Journal of Parallel and Distributed Computing, 61(3):350– 371, 2001.

[15] Snedecor. W, and Cochran. W(1989) Statistical Methods, Eighth Edition, Iowa State University Press.

[16] Tanbeer, S. K., Ahmed, C. F., Jeong, B.-S., and Lee, 2008. CP-tree: a tree structure for single-pass frequent pattern mining. In Proc. of PAKDD, Lect Notes Artif Int, 1022-1027.

[17] Prasanna K Lakshmi and C.R.K.Reddy. Article: Efficient Classifier Generation over Stream Sliding Window using Associative Classification Approach. International Journal of Computer Applications 115(22):1-9, April 2015.

Identifying Causes of Neonatal Mortality from Observational Data: A Bayesian Network Approach

K. A. Wilson[1,3], D. D. Wallace[1], S. S. Goudar[2], D. Theriaque[1], E. M. McClure[1]
[1]RTI International, Biostatistics and Epidemiology Division, Durham, NC, USA
[2]KLE University's Jawaharlal Nehru Medical College, Belgaum, Karnataka, India
[3]University of Liverpool, School of Medicine, Liverpool, UK

Abstract - *Despite improvements in access to birth facilities, neonatal mortality remains a critical health issue in many developing countries and causes are not fully understood. The Global Network Maternal Newborn Health Registry provides a rich source of data of neonatal mortality risk factors and outcomes to identify direct causes and higher-level determinants, however performing causal inference using observational data is difficult and remains an open problem in epidemiology. In this paper we sought to determine whether Bayesian networks can be used to identify the complex causal pathways leading to neonatal mortality outcomes and to quantify the effect of each cause on mortality. Our analysis identified a complex network of causes that contribute to neonatal mortality, including maternal death, pre-term birth, movement and breathing at birth. For variables identified as direct causes we estimated the average causal effect using logistic regression models that controlled for known confounders.*

Keywords: Causal inference, Bayesian network, neonatal mortality.

1 Introduction

While there has been a significant reduction in neonatal deaths from 5.6 million per year in 1990 to 4.0 million per year in 2000, neonatal mortality remains a major global public health issue [1]. Of the 130 million children born annually, approximately 4 million will die in the first month of life, 75% within the first 7 days and 25% in the first 24 hours [2]. According to the UN, the 3.6 million neonatal deaths that occurred in 2008 comprised 41% of all deaths under age 5. This underscores the importance of reducing neonatal mortality, and has been formalized as the fourth UN Millennium Development Goal [3]. To meet this goal of a two-thirds reduction in mortality of children under age 5 by 2015, the current rate of improvement must be increased 6-fold.

In India, while the overall neonatal mortality is 31 deaths per 1000 live births, the rate varies widely by region and birth facility [4]-[5]. And, despite improvements in access to birth facilities, neonatal mortality remains high, suggesting that the causes of neonatal death may be more complex than previously thought [6]. The Global Network Maternal and Child Health Registry provides a unique source of data relating to maternal and neonatal risk factors that can potentially explain these complex causal relationships. More than 70 variables were collected from pregnant women enrolled at 20 geographic clusters in Belgaum, India, including demographics, antenatal care, maternal and neonatal health conditions, delivery characteristics, and medical treatments. Although this data set is comprehensive, identifying the complex causal pathways between risk factors and outcomes is challenging due to it being observational in nature [5].

Observational studies are particularly susceptible to selection bias and confounding, which can result in biased estimates of effect [7]. Under certain assumptions, Bayesian networks (BNs) have shown promise in performing causal inference using observational data. So-called causal BNs can be used to model relationships between random variables, where the direction of the edges in the graph signifies a direct causal relationship [8]. Algorithms exist to identify the graph structure directly from data in the presence of confounding and selection bias [9]-[10]. Once the causal structure has been identified, the BN can be used to estimate the effect of manipulating key variables on a specified outcome variable, such as neonatal mortality [11]. Thus the BN approach promises to be a useful technique for identifying the causes of neonatal mortality given a rich observational data set.

The goal of this work is to extend and enhance existing Bayesian network methods to perform causal inference and to estimate causal effects of neonatal mortality using observational data from the Global Network Maternal and Child Health Registry. The remainder of this article is organized as follows. In section 2 we discuss the challenges of causal inference and approaches to overcome some of these challenges to obtain valid inferences based on analyses of observational data. In section 3 we describe our Bayesian network-based methods of identifying causal factors and estimating effects. Our results are presented and discussed in sections 4 and 5. In section 6, we present our conclusions and ideas for future work.

2 Background and Related Work

2.1 Causal Inference

The fundamental problem of causal inference is that it is not possible to measure the difference in outcome for an

individual for different levels of a variable of interest [12]. As a result, estimating a causal effect can only be accomplished by comparing groups of similar individuals at different levels of a given variable. To ensure that the true causal effect is estimated, this comparison requires both the *manipulation* of a variable and measurement of the change in the outcome variable while *accounting for clinical and environmental variables that could confound the conclusions about the variable of interest.* Valid causal inference is often achieved in randomized controlled trials through the use of an intervention, with the randomized assignment to this intervention, which theoretically balances known and unknown confounders across treatment groups.

In comparison, observational studies are problematic because of non-random group assignment and the absence of manipulation [13]. As a result, measures of effect can easily biased due to confounding, and thus estimating the average causal effect of changes in one variable on an outcome of interest requires controlling for potential confounders, some of which may be unobserved [14]. Most analytical methods used with observational data focus on ensuring that comparison groups are as similar as possible with respect to measured and unmeasured confounders [13]-[16]. However, the absence of a true manipulation or intervention, at best, results in unbiased estimates of association and not causal inference. Bayesian networks, and in particular, *causal Bayesian networks* can potentially address this weakness. The reader is referred to the seminal work by Pearl for a more complete discussion of causal inference algorithms [8].

2.2 Bayesian Networks

A Bayesian network (BN) is a probabilistic graphical model, in which the nodes in a directed acyclic graph represent random variables and the edges represent probabilistic associations between the variables. A BN models the joint distribution over all the variables in the graph, factored into a series of conditional probability distributions, resulting in a compact and efficient representation [17], [11].

Spirtes' *PC-algorithm* can be used to learn a causal BN [14]. The algorithm performs a series of conditional independence tests to determine directed relationships between the variables [18]. Kalisch and Bühlmann achieve a true positive rate of over 80% and false positive rate of less than 1 percent [18]. Nguefack-Tsangue and Zucchini argue that in the absence of unmeasured confounders, causal BNs are able to identify all causal relationships up to sampling error [19]. Shrier and Platt confirmed that this approach does not introduce additional conditional associations or bias [20]. Li, Shi and Satz used the *PC-algorithm* to successfully estimate the causal relationship between risk factors and disease using case-control data [7]. Kalisch et al. provide an efficient implementation that supports both categorical and continuous variables in *R* [21].

2.3 Estimating Causal Effects

Estimating causal effects from observational data can be achieved by simulating an intervention on a variable, a process known as manipulation. Pearl provides a theoretical background for estimating causal effects through the *do()* operator, which performs a manipulation on the variable of interest while accounting for clinical and environmental variables not on the causal pathway that could confound conclusions about the variable of interest. [12]. With this approach, parent nodes of the manipulated variable are included as covariates, a process known as adjusting for the direct causes, which captures the prior state of the probability distribution. Applying a manipulation to a variable removes the influence of any other variables and sets the value of that variable for all members of the sample. Maathuis, Kalisch and Bülmann show that the average causal effect can be estimated using a linear regression model, and that this approach is equivalent to Pearl's *do()* operator [22]. This method is implemented as the *ida* algorithm by Kalisch et al. in the *pcalg R* package [21]. One limitation of this implementation is its use of linear regression to estimate causal effects, which prevents it from being used to estimate the casual effect on a dichotomous variable.

2.4 Bayesian Network Assumptions

In standard BNs the directions of the edges do not imply any specific causal direction and probabilistic inference is agnostic to the directions of the edges. For a BN to be causal, additional assumptions are required, including the Causal Markov Assumption and the Causal Faithfulness Assumption. The Causal Markov Assumption attributes a direct causal relationship when two variables are connected by a directed edge, and states that each variable is independent of its non-effects given its causes [23]. The Causal Faithfulness Assumption states that the graph structure and the independence relationships in the data are isomorphic [10]. Additional assumptions include the absence of hidden common causes, causal feedback loops, and selection bias [24]. While methods exist to accommodate the existence of unmeasured confounders, causal effect estimates are undefined in these latent confounder models. As a consequence, most methods assume that all potential confounders are included in the graph. In this case, un-confounded estimates of causal effects can be determined [22].

3 Methods

Our methods consist of three steps: data processing, learning the optimal causal Bayesian network, and estimating the causal effects for direct causes of neonatal mortality.

3.1 Data Processing

Data were collected on all mothers and nenonates at three time points. At enrollment, basic demographic information was collected for all eligible and consented women. Maternal and neonatal outcomes were collected at the time of delivery and subjects were followed up at 42 days after birth to collect the 28-day neonatal mortality outcome.

Data from these time points were combined into a single analysis dataset using SAS 9.3, with one observation for each birth outcome. Data that were missing due to skip

patterns in the data collection forms were coded as "not collected." All variables were categorical except for hemoglobin level and BMI. These continuous variables were discretized using standard categories

Missing data analysis was performed for all variables included in the model. We assumed data were missing at random. Where the amount of missing data was significant and could potentially introduce bias, multiple imputation was performed prior to the estimation of causal effects. Imputation was performed in R using the MICE package, using polytomous regression with 20 imputed datasets [25]. To test our missing at random assumption and to ensure that the imputation did not introduce bias into the causal effect estimates, we built models based on the original un-imputed data and performed a sensitivity analysis.

The final analysis dataset contained 70 variables and 60,985 observations.

3.2 Learning the Causal Bayesian Network

The causal Bayesian network was learned using the *PC-Algorithm*, which was initially developed by Spirtes et al. and implemented in the *R* package, *pcalg*, by Maathuis, Kalisch and Bülmann [26], [22]. Stacked output from multiple imputation was used as the training dataset.

The PC-Algorithm is a constraint-based algorithm that estimates the conditional probability distribution over all variables using a series of conditional independence tests. One problem with this approach is the use of multiple comparisons, which can result in false positives. In the context of a causal Bayesian network, a false positive implies that two nodes are not independent given a set of conditioning nodes. Residual dependence after conditioning results in a graph that less sparse, where spurious causal relationships are uncovered. To address this issue, we treated the P-value used in the conditional independence tests as a tuning parameter and built a series of models using different P-values. We optimized the P-value using the Bayesian Information Criterion (BIC), and selected the model with the lowest BIC. A P-value of 0.0005 was used to learn the final model.

3.3 Estimating Causal Effects

One limitation of the PC-Algorithm, and constraint-based methods in general, is an inability to learn a unique Bayesian network. This problem arises from a failure to uniquely identify a network's structure using only conditional independence tests, as multiple graphs can encode the same conditional independencies. We addressed this issue through the development of an enhanced method for estimating causal effects, which we have named *ida+*. Our method is an extension of the *ida* method developed by Maathuis, Kalisch and Bülmann [22].

The *ida* algorithm uses the Markov blanket of a specific variable to build a multiple linear regression model and estimate the causal effect of predictor variable on an outcome using the direct parents of the predictor as covariates in the model. Because the PC-Algorithm is often unable to identify a single causal Bayesian network, the *ida* algorithm returns a multi-set of possible causal effects. An additional limitation of this method is that the estimation of causal effect may not be valid when the outcome is dichotomous or multinomial. Thus, our *ida+* algorithm incorporates several key enhancements:

- Logistic regression is used to estimate causal effects for dichotomous outcomes;
- Polytomous regression is used to estimate causal effects for categorical variables with more than two outcomes;
- The Cox goodness of fit test for non-nested models is used to determine which of the multiset of possible causal effect estimates is most likely correct;
- Confidence intervals, standard errors, and p-values are returned to quantify the precision of the estimates.

Logistic and polytomous regression models enable the estimation of odds ratios for categorical outcomes. The Cox goodness of fit test for non-nested models determines the best set of covariates for a model on the principle that if a given model contains the correct covariates then fitting a second model to these covariates should add no explanatory value [27]. Because calculated odds ratios are estimates subject to sampling error, quantifying their precision is essential.

The *ida+* algorithm is shown in Fig. 1.

```
Input: Set of Causal BNs (G),   Predictor (x),
  Outcome (y), Outcome type {linear | logistic
                  | polytomous}
Output: Causal Effect of Predictor on Outcome with
95% confidence intervals and p-value

for each graph in G {
  if y in parents(x)
     model ← null
  else
  {
      if length(parents(x)) > 0
      {
         model <- glm(y ~
           x + parents(x))
      }
      else
      {
         model ← glm(y ~ x)
      }
   }
   model_array ← model
}

lowest_p_val ←1
correct_model ← null

for each model in model_array {
  if p_value(model) < lowest_p_val
  {
     lowest_p_val <-
             p_value(model)
     correct_model ← model
   }
  return correct_model
```

Fig. 1. The *ida+* algorithm.

4 Results

Fig. 2 provides a simplified view of the Bayesian network generated by the PC-Algorithm with P=0.0005 and neonatal (28-day) mortality as the outcome. The summarized view is presented for clarity and includes only the outcome variable, its direct causes, and the parents of the direct causes. The algorithm identified 9 direct causal factors of neonatal mortality: maternal mortality, gender, pre-term birth, multiple birth, whether the baby moved upon birth, whether the baby was breathing when born, the presence of one or more neonatal conditions, whether transport was available if a hospital referral was needed, and whether the neonate was seen at a facility. For each of these causal factors, direct upstream causes were also identified and the relationship between all variables in the model can be seen.

For each of the direct causes, the *ida+* algorithm estimated the average causal effect using a logistic regression model with neonatal mortality as the dependent variable, the direct causes as the primary independent variable, and the parents of the direct causes as covariates in the model. The overall causal effect of the 9 direct causes is displayed in Table 1 along with 95% confidence intervals and P-values. The effect estimate is an odds ratio calculated as the exponent of the beta coefficient of the primary dependent variable in each model. Included covariates for each model are summarized in Table 2. The addition of some covariates introduced multicollinearity into the models. Multicollinearity generally occurred as a result of the structure of the questionnaires that generated the dataset. For example, the Bayesian network model shows Maternal Cause of Death as a direct cause of Maternal Mortality. Multicollinearity was likely introduced because cause of death was not collected for mothers that did not die. A similar phenomenon occurred with hospital referral and admission variables. As a result of multicollinearity, the estimates produced were not deemed reliable, and these covariates were dropped from the model.

Maternal mortality (OR: 7.972), gender (OR: 1.264), pre-term birth (OR: 1.247), neonatal conditions (OR: 21.704), breathing (OR: 9.974) and movement of the baby (OR: 30.139) all exhibit a substantial and significant effect on neonatal mortality, with baby movement and neonatal conditions having the largest effects. Multiple births appears to have a protective effect with an odds ratio of 0.774. Nenonate Seen at Facility is uninformative, likely due to the multicollinearity issues described above.

5 Discussion

The main finding of our research is that constraint-based methods of learning Bayesian networks can be used to identify direct and in-direct causes of neonatal mortality from an observational data source, and that the effects of these causes can be estimated using logistic regression models that control for appropriate confounders. The constraint-based PC-Algorithm identified 9 direct causes of neonatal mortality: maternal mortality, gender, pre-term birth, movement at birth, breathing at birth, presence of neonatal conditions, transport to facility, and neonate seen at facility. The use of the Cox goodness of fit test for non-nested models employed by our *ida+* algorithm was able to disambiguate multiple possible Bayesian networks, identify the single most likely graph, and estimate the causal effect of the 9 component causes on the mortality outcome. In contrast to standard associational approaches, such as linear or logistic regression modeling, the Bayesian network was able to identify more complex relationships between variables.

The causal effects shown in the results tables represent the odds of a neonate dying within 28 days of birth when the given variable is manipulated with an intervention and all other variables are held constant. In contrast to standard observational approaches, these estimates give greater insight into the impact of these direct causes on the mortality outcome in a situation where direct, real intervention with a randomized controlled trial is infeasible, primarily for ethical reasons. The causes identified by the algorithm can be considered component causes that contribute to the overall cause of morality. The largest causal factors are maternal death (8 times increase in odds), movement at birth (24 times increase in odds), breathing at birth (10 times increase in odds), and the presence of one of a number of neonatal health conditions (baby stopped feeding, high fever, hypothermia, difficulty breathing, bleeding from umbilicus – 22 times increase in odds). While the correctness of the graph cannot be determined formally, in general, the algorithm was able to identify several major causes of neonatal mortality. Developing public health interventions aimed at prevention or treatment of these causes should result in reduced mortality.

Fig. 2. Simplified Bayesian network for identified causes of 28-day mortality.

Table 1. Causal estimates for direct causes of 28-day mortality.

Variable (Reference Value)	*Causal Effect*	*Lower 95% CI*	*Upper 95% CI*	*P-Value*
Maternal Mortality (No)				
Yes	7.972	3.736	17.010	0.000
Gender (Female)				
Male	1.264	1.129	1.414	0.000
Pre-term Birth (Term (≥ 37 wks))				
Preterm (< 37 wks)	1.247	0.951	1.636	0.110
Multiple Birth (No)				
Yes	0.774	0.585	1.025	0.074
Movement at Birth (Yes)				
No	30.139	24.285	37.404	0.000
Breathing at Birth (Yes)				
No	9.974	7.995	12.443	0.000
Neonatal Conditions Present (No)	21.704			
Yes		17.879	26.347	0.000
Transport to Facility (Yes)				
No	0.984	0.259	3.738	0.981
Neonate Seen at Facility (Baby dead at arrival)				
Did not reach facility	0.000	0.000	Inf	0.972
No	0.000	0.000	Inf	0.970
Yes	0.000	0.000	Inf	0.966

The validity of the estimates of causal effects relies on the ability of the PC-Algorithm to correctly identify the causal relationships in the data and to reflect these relationships in the structure of the Bayesian network. In the absence of test data, it is impossible to formally validate the correctness of the resultant network, although Maathuis, Kalisch and Bülmann argue that the PC-Algorithm is guaranteed to uncover the correct causal graph up to sampling error [22]. It is difficult to determine whether this statement is true and the degree to which it is necessary to adhere to the underlying assumptions of the model. Nevertheless, the fact that the model identified pre-term birth and neonatal health conditions is consistent with the literature, particularly Bassani et al. who argue that pre-term birth, low-birth weight and neonatal health conditions account for 78% of all neonatal deaths in India. Although low birth weight is not identified in the Bayesian networks as a direct cause, it clearly defines several causal pathways that lead to neonatal mortality: it is shown to cause neonatal health conditions, which in turn causes neonatal mortality, and it also appears to be strongly associated with facility referral and pre-term birth, although the directionality of the pathways in these cases is questionable [28]. Additional factors, such as antenatal care and the administration of cost-effective interventions discussed by Bhaumik are reflected in the Bayesian network as higher-level determinants [29]. In fact, lack of antenatal care is on the causal pathway for neonatal mortality and has a direct effect on the presence of neonatal health conditions, which in turn affects mortality. There are also a number of spurious causal relationships, such as the association of antenatal care with hemoglobin level. Although this relationship is present in the data from a probabilistic perspective, hemoglobin is likely only collected during antenatal visits, and as a result, this pathway introduces bias into the model.

There are a number of limitations to our research. While the direct causes of mortality identified are consistent with the literature, some of the indirect causes appear to be problematic. For example, Transport to Facility is identified as a cause of Pre-term Birth. While there is clearly an association between these two variables, it is more likely that Pre-term birth is a cause of Transport to Facility. Thus, the model was unable to correctly orient the edge between these variables. Another example is the identification of Bag and Mask Resuscitation as a cause of Neonatal Conditions, which is also likely to be reversed.

These errors in identification could be attributed to a number of factors, including lower sample sizes of these higher-level causes resulting in a lack of power to detect the true relationships, an absence of temporal information (e.g., the fact that Neonatal Conditions must occur before Bag and Mask Resuscitation is used), and the inability of current methods to extract this information from the conditional probability distribution. In addition, the lack of formal evaluation of the Bayesian network or the causal effect estimates is a weakness. The best solution to this problem would be the use of an independent validation dataset; this approach would also assess the generalizability of the model. A more viable approach, however, would to use cross-validation techniques to assess the fit of the model to held-out data using an objective metric, such as Bayesian Information Criterion. One additional limitation is the lack of validation of causal estimates, through traditional approaches, such as cross

validation. However, a comparison of causal effects and associations estimated using standard regression models provides some useful insight. For example, for neonatal mortality, an intervention on maternal mortality has an estimated effect of 7.972 (95% CI: 3.3736 – 17.010), which is substantially larger than the association odds ratio of 2.299 (95% CI: 0.554 – 9.538). Therefore, an intervention on maternal mortality results in an 8-times increased risk of neonatal mortality, whereas the association when controlling for other factors, results in only a 2 times increase in risk. Similar differences in effect (including for protective factors) are noted for the other direct causes. One additional limitation of our methodology is the relatively strong assumption of no unmeasured confounders. In practice, unmeasured confounders very likely exist, and our inability to account for these may limit our ability to identify uncounfounded causes and higher level determinants of neonatal mortality.

Table 2. Covariates included in each logistic regression model.

Cause	*Covariates*
Maternal mortality	None
Gender	Parity
	Antenatal Location
	Birth Location
	Fetal Heartrate
Pre-term Birth	Cluster Resident
	Birth Weight
	Transport to Facility
Multiple Birth	Age of Mother
	Antenatal Location
	Maternal Conditions
	Maternal Mortality
	Birth Attendant
	Birth Location
	Birth Weight
	Pre-term Birth
Baby Move	Born in Cluster
Baby Breathe	Baby Move
	Baby has Heartbeat
Neonatal Conditions	Age of Mother
	Antenatal Location
	Maternal Conditions
	Birth Weight
	Pre-term Birth
	Multiple Birth
	Baby Breathe
	Bag and Mask Resuscitation
Transport to Facility	Birth Weight
	Multiple Birth
	Neonatal Conditions
	Oxygen Treatment
Neonate Seen at Facility	Prenatal Vitamins
	Multiple Birth

These limitations point to a number of paths for future research. Improved methods of learning and evaluating causal Bayesian networks are needed, along with methods to evaluate the accuracy of causal effect estimates. One possible approach is to compare results of these methods with results from a RCT where an actual intervention was performed. Theoretically, odds ratios obtained from an RCT should be equivalent to those generated by these methods. Further work with this dataset should include validation of the graph and estimates with cross-validation approaches.

6 Conclusions

Although formal validation of results is needed, this research has demonstrated the promise of Bayesian networks as a method for identifying causal factors from observational data. The methods described, and the specific application to neonatal mortality, are of strong public health relevance for several reasons. First, the ability to perform causal inference from observational data is a critical issue in epidemiology where a major emphasis in any study is the identification and control of confounding factors, particularly when conduct of a randomized controlled trial is not possible. Second, the inductive nature of the Bayesian network learning algorithms provides an opportunity to uncover previously unknown causal factors and pathways. Although these methods are imperfect, their use in exploratory data analysis can augment traditional research hypothesis generation. Application of these tools can thus inform future research studies, increasing our ability to the identify causes of, and develop effective interventions for, critical public health issues.

7 Acknowledgements

Data were originally collected by the Global Network for Women's and Children's Health funded by grants U01HD042372 and U01HD040636 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the US National Institutes of Health. All participants signed informed consent prior to study participation. This secondary analysis was conducted under the auspices of the University of Liverpool and RTI International. Institutional Review Board approval was obtained from both organizations prior to gaining access to the data. The authors acknowledge the support of Belgaum site of the Global Network in conducting this research.

8 References

[1] Lawn, J. E., Kerber, K., Enweronu-Laryea, C. & Cousens, S. (2010). '3.6 million neonatal deaths — what is progressing and what is not', Semin Perinatol, 34, pp.371-386, [Online]. Available from: http://www.healthynewbornnetwork.org/sites/default/files/resources/Epidemiology_Lawn.pdf (Accessed: 26 November 2013).

[2] Jehan, I. et al. (2009). 'Neonatal mortality, risk factors and causes: a prospective population-based cohort study in urban Pakistan', Bulletin of the World Health Organization, 87, pp.130-138, [Online]. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636189/ (Accessed: 30 November 2013).

[3] United Nations (2013). 'Goal 4: Reduce Child Mortality', Millennium Development Goals, [Online]. Available from: http://www.un.org/millenniumgoals/childhealth.shtml (Accessed: 7 September 2013).

[4] World Bank (2013). 'Data: Mortality Rate, Neonatal (per 1,000 live births)', [Online]. Available from: http://data.worldbank.org/indicator/SH.DYN.NMRT (Accessed: 9 November 2013).

[5] Goudar, S.S. et al. (2012a). 'The maternal and newborn health registry study of the Global Network for Women's and Children's Health research', International Journal of Gynecology & Obstetrics, 118 (3), pp.190-193, [Online]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22738806 (Accessed: 30 November 2013).

[6] World Health Organization (2013). 'Newborn health epidemiology', Maternal, newborn, child and adolescent health, [Online]. Available from: http://www.who.int/maternal_child_adolescent/epidemiology/newborn/en/index.html (Accessed: 7 September 2013).

[7] Li, J., Shi, J. & Satz, D. (2008). 'Modeling and analysis of disease and risk factors through learning bayesian networks from observational data', Qual Reliab Engng Int, 24, pp.291-302, [Online]. Available from: http://141.213.232.243/bitstream/handle/2027.42/58076/893_ftp.pdf?sequence=1 (Accessed: 30 November 2013).

[8] Pearl, J. (2009). Causality. Cambridge University Press.

[9] Kleinberg, S. & Hripcsak, G. (2011). "A review of causal inference for biomedical informatics', J Biomed Inform, 44 (6), pp.1102-1112, [Online]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21782035 (Accessed: 30 November 2013).

[10] Cooper, G. F. (1999). An overview of the representation and discovery of causal relationships using Bayesian networks. In: Glymour, C and Cooper, G. F. Computation, Causation, and Discovery. AAAI Press.

[11] Darwiche, A. (2010). 'Bayesian Networks', Communications of the ACM, 53 (12), pp.80-90, [Online]. Available from: http://cacm.acm.org/magazines/2010/12/102122-bayesian-networks/abstract (Accessed: 30 November 2013).

[12] Höfler, M. (2005). 'Causal inference based on counterfactuals', BMC Medical Research Methodology, 5 (28), [Online]. Available from: http://www.biomedcentral.com/1471-2288/5/28 (Accessed: 3 January 2014).

[13] Trojano, M. et al. (2009). 'Observational studies: propensity score analysis of non-randomized data', The International MS Journal, 16, pp.90-97, [Online]. Available from: http://www.msforum.net/journal/download/20091690.pdf (Accessed: 3 January 2014).

[14] Spirtes, P. (2010). 'Introduction to causal inference', Journal of Machine Learning Research, 11, pp.1643-1662, [Online]. Available from: http://jmlr.org/papers/volume11/spirtes10a/spirtes10a.pdf (Accessed: 3 January 2013).

[15] Hernán, M. A. & Robins, J. M. (2006). 'Estimating causal effects from epidemiological data', J Epidemiol Community Health, 60 (7), pp.578-586, [Online]. Available from: http://jech.bmj.com/content/60/7/578.abstract (Accessed: 3 January 2014).

[16] Winship, C. & Morgan, S. L. (1999). 'The estimation of causal effects from observational data', Annual Review of Sociology, 25, pp.659-706, [Online]. Available from: http://dash.harvard.edu/bitstream/handle/1/3200609/Winship_EstimationCausal.pdf?sequence=1 (Accessed: 18 October 2013).

[17] Ben-Gal, I. (2007). 'Bayesian networks', In: Ruggeri, F., Failtin, F. & Kennet, R. Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons (2007), [Online]. Available from: http://www.eng.tau.ac.il/~bengal/BN.pdf (Accessed: 11 January 2014).

[18] Kalisch, M. & Bühlmann, P. (2007). 'Estimating high-dimensional directed acyclic graphs with the PC-algorithm', Journal of Machine Learning Research, 8, pp.613-636, [Online]. Available from: http://jmlr.org/papers/volume8/kalisch07a/kalisch07a.pdf (Accessed: 11 January 2013).

[19] Nguefack-Tsague, G. & Zucchini, W. (2011). 'Modeling hierarchical relationships in epidemiologic studies: a Bayesian networks approach', Epidemiol Health, 33, [Online]. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3132659/ (Accessed: 11 January 2013).

[20] Shrier, I. & Platt, R. W. (2008). 'Reducing bias through directed acyclic graphs', BMC Medical Research Methodology, 8 (70), [Online]. Available from: http://www.biomedcentral.com/1471-2288/8/70 (Accessed: 11 January 2014).

[21] Kalisch, M. et al. (2012). 'Causal inference using graphical models with the r package pcalg', Journal of Statistical Software, 47 (11), [Online]. Available from: http://www.jstatsoft.org/v47/i11 (Accessed: 11 January 2014).

[22] Maathuis, M. H., Kalisch, M. & Bühlmann, P. (2009). 'Estimating high-dimensional intervention effects from observational data', The Annals of Statistics, 37 (6A), [Online]. Available from: http://arxiv.org/pdf/0810.4214.pdf (Accessed: 11 January 2014).

[23] Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000). 'Using Bayesian networks to analyze expression data', Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, [Online]. Available from: http://www.cs.huji.ac.il/~nirf/Papers/FLNP1Full.pdf (Accessed: 11 January 2014).

[24] Neapolitan, R. E. (2009). Probabilistic Methods for Bioinformatics with an Introduction to Bayesian Networks. Elesvier.

[25] van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL http://www.jstatsoft.org/v45/i03/.

[26] Spirtes, P., Glymour, C. & Scheines, R. (2000). Causation, prediction, and search. MIT Press, Cambridge, MA.

[27] R. Davidson & J. MacKinnon (1981). Several Tests for Model Specification in the Presence of Alternative Hypotheses. Econometrica, 49, 781-793.

[28] Bassani, D. et al. (2010). 'Causes of neonatal and child mortality in India: a nationally representative mortality survey', The Lancet, 376 (9755), pp.1853-1860, [Online]. Available from: http://search.ebscohost.com.ezproxy.liv.ac.uk/login.aspx?direct=true&db=cmedm&AN=21075444&site=eds-live&scope=site (Accessed: 7 September 2013).

[29] Bhaumik, S. (2013b), 'India tops world table for number of babies who die on day of birth', BMJ, 346, p.f3123, [Online]. Available from: http://www.bmj.com/content/346/bmj.f3123 (Accessed: 7 September 2013).

Learning Decision Trees From Histogram Data

Ram B. Gurung
Dept. of Computer and Systems Sciences
Stockholm University, Sweden
Email: gurung@dsv.su.se

Tony Lindgren
Dept. of Computer and Systems Sciences
Stockholm University, Sweden
Email: tony@dsv.su.se

Henrik Boström
Dept. of Computer and Systems Sciences
Stockholm University, Sweden
Email: henrik.bostrom@dsv.su.se

***Abstract*—When applying learning algorithms to histogram data, bins of such variables are normally treated as separate independent variables. However, this may lead to a loss of information as the underlying dependencies may not be fully exploited. In this paper, we adapt the standard decision tree learning algorithm to handle histogram data by proposing a novel method for partitioning examples using binned variables. Results from employing the algorithm to both synthetic and real-world data sets demonstrate that exploiting dependencies in histogram data may have positive effects on both predictive performance and model size, as measured by number of nodes in the decision tree. These gains are however associated with an increased computational cost and more complex split conditions. To address the former issue, an approximate method is proposed, which speeds up the learning process substantially while retaining the predictive performance.**

***Index Terms*—histogram learning, histogram tree**

I. Introduction

Standard machine learning algorithms are designed for handling data represented by numeric and categorical variables. Even in cases when it is known that the data does have some structure, e.g., some groups of variables are related, such information is lost when the data is encoded as ordinary numeric and categorical variables and provided as input to the standard learning algorithms. One particular type of structure that we focus in this paper is histogram data, i.e., sets of variables representing the frequency distributions of some (implicit) variables. For example, we may use three variables (bins) to represent the relative frequency distribution of days during a month with average temperature lower than zero degrees, between zero and twenty degrees, and above twenty degrees. Histogram data is frequently encountered in domains where multiple observations are aggregated. One reason for aggregating data can simply be to save storage space, e.g., when dealing with big data, while in other cases the aggregation is necessary for being able to represent all data points (observations) on the same format, i.e., with the same number of variables. For example, if each customer in a database corresponds to one data point, where information on the purchase amounts should somehow be represented, then since the number of purchases may vary from customer to customer, each single purchase cannot be represented by a unique variable without introducing problems with missing variables and undesired ordering effects. Instead, the information can readily be represented by a histogram, e.g., where the different bins correspond to intervals for the purchase amounts. Histograms are also widely used to aggregate data streams where data are collected over time, e.g., readings in sensor networks.

Research on complex data structures, such as histograms, has been undertaken within the field of symbolic data analysis (SDA) [1]. Symbolic data represents complex data types which do not fall under the traditional categories of numeric and categorical variables. One specific type of histograms that have been studied are categorical in nature with a relative frequency assigned to each bin. Such histograms are classified as modal multi-valued variables in the terminology of the SDA framework, while Diday [2] refers to such histograms as categorical histogram data. More formally, for observations with n categorical histogram variables $X_i, i = 1...n$, with m_i bins $x_{ij}, j = 1...m_i$ each bin is assigned a relative frequency r_{ij} such that $\sum_{j=1}^{m_i} r_{ij} = 1$ and each observation is associated with a class label Y. For all observations, bin descriptions of a histogram variable are identical. This is the type of histogram we will be considering in this study.

Research on learning from histogram data is still at an early stage. To the best of our knowledge, no studies have been published on learning classifiers from histogram data. However, there have been some studies on applying linear regression [3], [4], PCA [5] and clustering [6] to histogram data. While most of the considered approaches take into account the actual bin boundaries, the work on adapting PCA for categorical histogram data [5] deals with data of the same type as considered here. It should be noted, however, that the approach in [5] is aimed for dimensionality reduction and not for performing classification. The type of histogram data considered in this study and in [5] is closely related to "compositional" variables within compositional data analysis [7], where weights associated with each variable represent distributions over possible values. However, the research in compositional data analysis has not been on learning classifiers.

In this paper, we will propose an adaptation of the standard decision tree algorithm [8] to allow for learning from categorical histogram variables. We will compare the performance of the adapted learning algorithm to using the standard learning algorithm with histogram data represented by ordinary variables, i.e., with no structural information. The main contributions of the paper are:

- A novel approach for learning decision trees from histogram data, including an approximation to allow for substantial speedup
- An empirical evaluation comparing the new approach to the standard decision tree learning algorithm on both synthetic and real-world data sets
- Findings concerning the utility of exploiting the structure in histogram data when learning decision trees

In the next section, the novel approach for learning decision trees from histogram data is presented. In Section III, the experimental setup and results are presented. The empirical findings and limitations of the proposed approach are discussed in Section IV. Finally, in Section V, we summarize the main conclusions and point out directions for future research.

II. Method

The standard decision tree algorithm [8] was adapted to learn from histogram data. Therefore, the approach can be viewed as a generalization of the standard algorithm where the bins of each histogram is handled as a vector and partitioning takes place by finding a separating hyperplane in the corresponding space. Fig. 1 provides an illustration of the approach. The best split plane for each histogram variable is obtained and then compared to the best splits of the other histogram variables, as well as to splits obtained from the regular numeric and categorical variables. The split with the highest information gain is finally selected for partitioning the examples in the node. Similar approaches to employing multivariate splits have been proposed in the past, e.g., using linear combination of multiple features to perform splits at each intermediate nodes [9], [10]. In these approaches, all the features are considered simultaneously for splitting, while in our case, multiple variables considered for a split are bins of same histogram variables with unit sum constraint. Hence, there can be more than one histogram variable in a dataset that would require evaluation of multiple multivariate splits.

In this section, we first provide a formalization of the node-splitting part of the adapted decision tree learning algorithm and then illustrate its workings with two very simple examples. We proceed by providing an analysis of the computational complexity of the algorithm and end the section by proposing an approximation of the original method for speeding up the node-splitting process.

A. The node-splitting algorithm

The aim of the algorithm is to find the optimal node splitting hyperplane. Because of the unit sum constraint, a histogram with m bins is a vector point that lies in a hyperplane, which is represented by

$$x_1 + x_2 + ... + x_m = 1 \tag{1}$$

Let the equation for the linear splitting hyperplane be

$$c_1x_1 + c_2x_2 + c_3x_3 + ... + c_mx_m = 1 \tag{2}$$

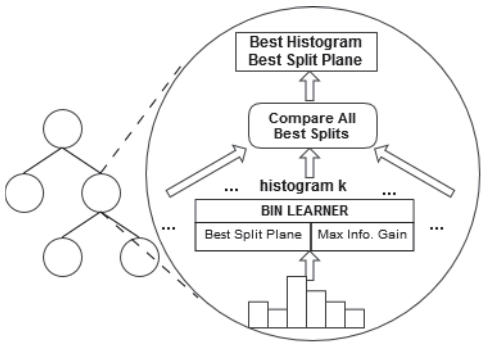


Fig. 1. Overview of the node splitting approach

where $C = (c_1, c_2, c_3, ..., c_m)$ are the unknown coefficients to be solved. Hyperplanes represented by equation 1 and 2 are assumed to be orthogonal, which results in

$$\begin{bmatrix} 1 & 1 & ... & 1 \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ | \\ c_m \end{bmatrix} = 0$$

$$c_1 + c_2 + c_3 + ... + c_m = 0 \tag{3}$$

Solving for m unknowns in C requires m linear equations. In addition to equation 3, substituting m-1 points for $X = (x_1, x_2, ..., x_m)$ in equation 2 would give sufficient number of equations to solve for C. Selection of m-1 points out of n data points can be done in $\binom{n}{m-1}$ ways. The resulting system of linear equations can be solved as follows.

$$\begin{bmatrix} 1 & 1 & ... & 1 \\ x_{1,1} & x_{1,2} & ... & x_{1,m} \\ | & | & | & | \\ x_{m-1,1} & x_{m-1,2} & ... & x_{m-1,m} \end{bmatrix} \times \begin{bmatrix} c_1 \\ c_2 \\ | \\ c_m \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ | \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} c_1 \\ c_2 \\ | \\ c_m \end{bmatrix} = \begin{bmatrix} 1 & 1 & ... & 1 \\ x_{1,1} & x_{1,2} & ... & x_{1,m} \\ | & | & | & | \\ x_{m-1,1} & x_{m-1,2} & ... & x_{m-1,m} \end{bmatrix}^{-1} \times \begin{bmatrix} 0 \\ 1 \\ | \\ 1 \end{bmatrix}$$

Algorithm 1 specifies the node splitting process. For an m binned histogram, m-1 vector points are chosen in $\binom{n}{m-1}$ ways. Each combination is examined for a split plane (lines 5 to 24). The left hand sides of m linear equations are captured in mXm square matrix A (line 6). The right hand sides of these linear equations are represented as column vector B of size m (line 8). If the inverse of A exists, the product of the inverse of A and B results in coefficients of the split plane (line 9). For all the points in the node, the scalar product of the point and coefficient vector gives a value that determines whether to assign the point to the left or the right node (lines 10 to 17). The information gain obtained from the split can be calculated and compared with the previous best gain (lines 18 to 22).

Algorithm 1 Finding best split plane in a node

Input: obs: observations in a node
$histogram\_variables$: names of histogram variables
Output: $best\_split\_plane$: coefficients of best split plane

for all $histogram$ in $histogram\_variables$ **do**
  $m \leftarrow$ number of bins in $histogram$
  $h\_points \leftarrow$ $histogram$ values in obs
  $combinations \leftarrow$ ways of choosing m-1 points from h_points
  for all $combn$ in $combinations$ **do**
    $A \leftarrow$ matrix of m-1 points in $combn$ with all elements of first row 1.
    if A^{-1} exists **then**
      $B \leftarrow$ column vector of m-1 ones, first element 0.
      $split\_plane\_coefs \leftarrow$ multiply A^{-1} and B
      for all $point$ in h_points **do**
        $value \leftarrow$ multiply $point$ and $split\_plane\_coefs$
        if $value < 1$ **then**
          $l\_obs \leftarrow$ assign $point$ to left node
        else
          $r\_obs \leftarrow$ assign $point$ to right node
        end if
      end for
      $info\_gain \leftarrow$ get information gain of the split
      if $info\_gain$ is greater than previous best **then**
        $best\_info\_gain \leftarrow info\_gain$
        $best\_split\_plane \leftarrow split\_plane\_coefs$
      end if
    end if
  end for
end for

B. Examples

We illustrate the workings of the algorithm using histogram variables with two and three bins respectively. The left graph in Fig. 2 shows the splitting process when the histogram variable has two bins x_1 and x_2. All the points (x_1, x_2) lie on the line AB. The splitting line CD is orthogonal to AB and passes through a point in AB. The coefficients of CD, a and b, can be determined by solving two linear equations. The process is repeated allowing CD to pass through all the points and choosing the one that gives the highest information gain. The process is similar for a histogram variable with three bins, as illustrated by the right graph in Fig. 2, in which all vector points are spread in the 3-D plane ABC. A three-dimensional splitting plane DEFG can be defined by the equation $ax_1 + bx_2 + cx_3 = 1$. DEFG is orthogonal to ABC and passes through two vector points. Three linear equations on a, b and c can be formed to solve for these unknowns.

Figure 3 shows a 3-D scatter plot for a small sample set of 100 observations that has a histogram variable with three bins x_1, x_2 and x_3. A simple pattern was injected in the data, if $x_1 + x_3 < 0.3$ then class label $y = 1$ else $y = 0$. Green

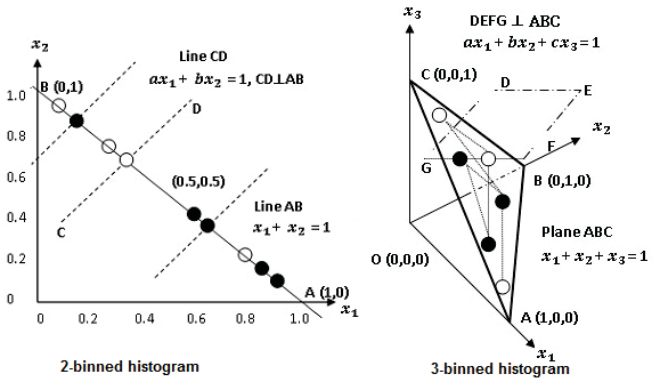


Fig. 2. Splitting in two and three binned histogram

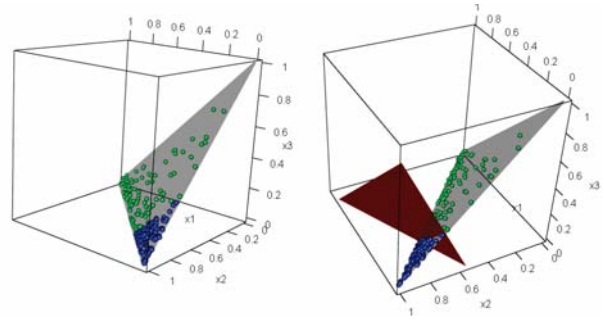


Fig. 3. Split plane in 3-binned histogram variable

and blue points correspond to the negative and positive cases respectively. The red plane is the splitting plane discovered by the algorithm. The equation of this splitting plane turns out to be: $0.885x_1 - 1.769x_2 + 0.885x_3 + 1 = 0$. Projecting the plane intothe 2-D plane of x_1 and x_3 would result in $x_1 + x_3 = 0.29$ which is approximately the same pattern injected in the data set. More experimental results will be presented in Section III.

C. Computational complexity

The computational cost of the proposed approach increases as the number of observations n and number of bins m increase. The cost is due to the combinatorial explosion of having to evaluate $\binom{n}{m-1}$ combinations while searching for the best separating hyperplane. Therefore, the computational complexity of the algorithm is proportional to $O(n^m)$.

D. Speeding up the node-splitting process

The computational cost can be reduced by either limiting the number of bins m or the number of observations n or both. The former was in this study handled in a straightforward manner, by merging bins for details see *Real-world Data* in the Experiments section. The latter, i.e., limiting the number of observations, was addressed using a more elaborate approximation method.

1) Approximation Approach: Each observations in a node for a histogram variable can be considered as a point in a m-dimension space, m being the number of bins. So, for convenience, observations in a node shall be referred as points

henceforth. In this approach, instead of using all the points in a node to build and evaluate splitting planes as described in algorithm 1, we generate small number of candidate split points and then build and evaluate split planes out of those newly generated points. It should be noted that new points are synthetically generated from original ones as will be described later in algorithm 2. The parameter num_points is the number of such candidate split points, as a consequence the algorithm only needs to consider $\binom{num\_points}{m-1}$ combinations. By choosing $num\_points < n$, the computational cost can be significantly reduced. In order for such approximate plane to make a good split, the new points that the plane passes through should be carefully generated.

As shown in the right half of figure 4, if there exist an optimal decision boundary, we want $best\_split\_plane$ to pass along this decision boundary as shown by red line. This is only possible when newly created synthetic split points lie close to the optimal decision boundary as shown by asterisks (*). Therefore, the first step in the approximation approach is to generate new candidate split points that are likely to fall around optimal decision boundary. Algorithm 2 describes the process of generating new synthetic candidate split points.

The algorithm first tries to locate the boundary regions since new synthetic points should come from such regions. This is done by examining the neighborhood around each point. For each point, a certain number e.g. 10 nearest neighbors are taken to form a group, which we shall simply refer as a cluster. The size of a cluster i.e. the number of nearest neighbors around the point is treated as parameter (N_c). Basically, the algorithm builds a cluster of N_c neighboring points around each point in a node. If the cluster lies in boundary region, its member points will be of different classes. As shown in the left half of figure 4, cluster A and C have all its members of same class whereas cluster B has a mix of both classes. So, cluster B is an ideal type of clusters that the algorithm prioritizes. The entropy value of the cluster determines how ideal the cluster is. Higher entropy values are preferred. Once the entropy of the clusters is calculated, they are prioritized according to the entropy value. A certain number of best clusters (num_points) e.g. 15 are selected and their centers are obtained. The cluster centers are used as new synthetic candidate split points to build and evaluate an approximate split plane.

If we were to use a standard clustering algorithm e.g. K-means it probably would result in new split points from regions that are not useful in finding optimal splits, as K-means do not focus on finding clusters with high class entropy (actually it would penalize different classes). As shown in the left half of figure 4, split points around the region of cluster A and C will not contribute much in finding optimal split. However, evaluating these points to search for an optimal split plane would consume valuable search time. Therefore, the tailored approach of obtaining relevant split points as explained in the algorithm 2 was preferred.

A scatter plot of a small sample training set with three binned histogram is shown in figure 4. Blue and green points

Algorithm 2 Finding split points around decision boundary

Input: h_points: observations for a histogram variable
$class$: class label of h_points
num_points: number of split points required
N_c: number of nearest neighbors to consider
Output: $split\_points$: candidate split points

```
if |h_points| > num_points then
    for all point in h_points do
        cluster ← find N_c nearest points around point
        center ← find center of cluster
        entropy ← get entropy of cluster using class
        list ← save center and entropy in a list
    end for
    d_list ← sort list in descending order of entropy
    new_points ← get top num_points centers in d_list
    split_points ← new_points
else
    split_points ← sample_points
end if
```

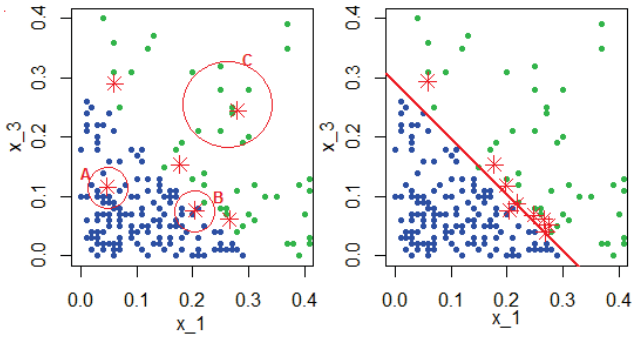


Fig. 4. Left: Generating split points, Right: Forming splitting plane

correspond to examples from each of the two classes, respectively. In the right part, the points marked with an asterisk (*) are the candidate split points generated by using algorithm 2. The approximate split plane obtained by using the algorithm is shown as red line which passes through two of these split points. In the left part, three clusters A, B and C are shown just for illustration with some cluster centers marked with asterisks (*).

III. Experiments

The proposed approach has been implemented in R[1]. Experiments were performed on a synthetic, a semi-synthetic and a real-world data. The bin values in the synthetic data set were obtained using uniform random sampling. The bins of a histogram were later normalized to satisfy the unit sum constraint. Synthetic dependencies among the bins were then injected by labeling the examples according to a set of formulated rules. The purpose of synthetic data set is to show that the algorithm can exploit the dependencies in the bins better by treating them together compared to when bins are treated individually.

[1]http://www.r-project.org/

The semi-synthetic data set was derived from the publicly available 'iris' data set [11] where original numeric variables were converted into histogram variables. The purpose of the semi-synthetic experiment was to investigate the robustness of the algorithm when bins have no inherent dependencies, i.e., the class labels are not dependent on interactions among the bins. The real-world data was provided by the heavy truck manufacturing company Scania CV AB and consists of histogram variables that describe operational profiles of trucks. Certain pattern among the bins of a histogram might reveal information about the truck's usage that could be associated to the breakdown of various component in the truck. The goal of the algorithm therefore is to discover useful pattern among the histogram bins by treating them together. The predictive performance of both standard decision tree learning algorithm and the proposed approach are compared with respect to classification accuracy measured using cross validation. In addition to accuracy, the tree size, i.e., the number of nodes, is also presented for each method. We here report only the results of applying the histogram approach using the approximation method. Using the exact approach, i.e., using all samples for generating split planes, was practically infeasible due to the excessive computational cost. Brief descriptions of the data sets, the experimental settings and the results observed from the experiment with each data sets are provided in the following sub-sections.

TABLE I
SYNTHETIC DATASET: FOUR BINS

Split Points	Cluster Size	Tree Nodes	Accuracy
7	10	8.2	99.6
7	15	11.6	99.8
7	20	11	99.7
11	10	7	99.4
11	15	9.4	99.2
11	20	8.8	99.2
15	10	5.6	99.6
15	15	8.2	99.6
15	20	8.6	99.4
Bins Treated Individually (Standard Tree Algorithm)			
—	Decision Tree	38.8	98.2

TABLE II
SYNTHETIC DATASET: FOUR AND FIVE BINS

Split Points	Cluster Size	Tree Nodes	Accuracy
7	10	29.2	97.3
7	15	41.6	95.7
7	20	38.2	96.3
11	10	13.6	99.3
11	15	16.2	98.1
11	20	17	97.9
15	10	11.4	98.7
15	15	11.2	98
15	20	11	98.6
Bins Treated Individually (Standard Tree Algorithm)			
—	Decision Tree	50.8	95.9

A. Synthetic Data

Two synthetic datasets, each consisting 1000 observations with equal proportions of observations labeled as positive and negative, were considered separately in two different experiments. First dataset consists of single histogram variable X with 4 bins. Simple pattern was injected in this dataset: if $X_1 + X_3 < 0.3$ then target class variable $Y = 1$, otherwise $Y = 0$. X_1, X_2, X_3 and X_4 are the bins of histogram X. Similarly, second dataset has two histogram variables; $X1$ with four bins and $X2$ with five bins. A more complex pattern was injected in the data which involve both histogram variable: if ($X1\_1 + X1\_3 < 0.3$ and $0.3 < X2\_1 + X2\_3 < 0.7$) then $Y = 1$ else $Y = 0$, where Y is the target (output) variable. For both experiments parameter settings were identical. The termination condition for the tree building algorithm i.e., stop expanding the current node, is when the number of observations in a node drops below 5 or the split does not provide any information gain. Three values for the number of new split points (num_points in algorithm 2) to be used for forming splitting plane, were examined: 7, 11 and 15. This can be any value higher than number of bins but higher value result in longer model training time. In order to cover wider range, these 3 values were chosen. Three different cluster sizes i.e. number of points considered to form a cluster (N_c in algorithm 2), were examined: 10, 15 and 20. 10-fold cross-validation was performed. The outcome of the experiment with the first dataset is presented in table I whereas the results of the second dataset is presented in table II. The columns of tables respectively show the number of points used for approximating optimal split, the size of the cluster considered for generating new split point, the tree size and the accuracy, where the latter two correspond to averages over the ten folds.

B. Semi-synthetic Data

A semi-synthetic dataset was generated from the publicly available Iris dataset [11]. The dataset has four numeric variables: petal length, petal width, sepal length and sepal width. Each observation belongs to one of three classes: Iris-versicolor, Iris-setosa and Iris-virginica. The data set contains 150 observations, 50 from each class. Each of the four variables was used to generate a synthetic histogram variable by including two additional variables such that they satisfy unit sum constraint. For example, in order to transform numeric variable petal length into histogram variable, it was first normalized to lie between 0 and 1. Let it be X_1. For each X_1, two integers in the range of 1 to 100 were uniform randomly selected. These two integers, X_2 and X_3 were then scaled down as:

$X_2 => X_2 * (1 - X_1)/(X_2 + X_3)$

$X_3 => X_3 * (1 - X_1)/(X_2 + X_3)$

The new dataset hence has four histogram variables, each with three bins. 5-fold cross validation was performed. The same termination condition for the tree growing as applied in the previous experiment was used. Number of split points used for searching split planes were: 5, 7 and 9. The chosen number should be higher than number of bins in the histogram variable. Model training time would increase as we select higher numbers, so for simplicity only three values were

TABLE III
SEMI-SYNTHETIC: IRIS DATASET

Split Points	Cluster Size	Tree Nodes	Accuracy
5	10	13	91.67
5	15	12.6	92.33
5	20	13.8	89
7	10	12.2	91
7	15	10.6	89
7	20	11.4	85
9	5	11.4	91
9	10	10.6	92.33
9	15	10.6	88.33
Bins Treated Individually (Standard Tree Algorithm)			
—	Decision Tree	9.4	92.67

TABLE IV
OPERATIONAL DATA SET

Split Points	Cluster Size	Tree Nodes	Accuracy
13	10	82.4	59.17
13	15	68.8	59.33
13	20	77.8	57.83
15	10	73	57.17
15	15	65.8	56.17
15	20	68	58.67
17	10	67.8	61.67
17	15	64	59.83
17	20	58.2	57.33
Bins Treated Individually (Standard Tree Algorithm)			
—	Decision Tree	106.8	59

examined. Three different cluster sizes i.e. number of points used to form a cluster, were arbitrarily chosen: 5, 10 and 15. The results of the experiment are shown in table III.

C. Real-world Data

Each observation in operational data is a snapshot of operational profile of a truck. Histogram variables in operational data holds information about how often the truck had operated under a particular feature range. The histogram variable, for example temperature, has 10 bins. Each bin measures the number of times the truck had operated within certain ambient temperature range. Temporal information about the truck's operation at various ambient temperature is transformed into relative frequency count as histogram over time. The histogram transformed data is extracted from the truck when it visits workshop for maintenance. Certain patterns within the bins of a histogram might reveal useful information about the truck's usage that are related to breakdown of certain components. The objective of the experiment is to distinguish trucks with battery failure from those without failure by using histogram feature variables. Bins of the histograms are normalized. The original data set had 33603 observations spread along 308 variables (counting bins as independent variables). There are 17 histogram variables of various length. The data set is very sparse and skewed in terms of battery failure as class label. Out of the 33603 observations, only 720 had battery problems.

For experimental purposes, a smaller data set was extracted. Given the computational cost that would incur, it was a necessary step. Out of these, only four histogram variables which were deemed as important were selected. Only histogram variables were selected for the experiment because the purpose here was to compare the performance when training tree as histogram against training by treating the bins individually. So the influence from any other variables either numeric or categorical was not desired. Issues related to missing values and skewed class distribution were set aside by including observations that had no missing values and selecting an equal number of positive and negative cases. Finally, a data set with 300 positive and 300 negative cases was extracted for the experiment. For confidentiality purpose, original variables are anonymised.

First histogram variable has 9 bins. Originally this variable was a 6X6 matrix. In order to handle computational complexity, adjoining 4 cells were merged resulting in 3X3 matrix. These 9 cells were then treated as bins. Second histogram variable also has 9 bins which was similarly transformed from 6X6 matrix. Third and fourth histogram variables have 10 bins. 10-fold cross validation was performed. The stop criteria was set to 15 or less observations in a node or if the split did not ensure any information gain. Three values for the number of split point used for forming split plane were examined:13, 15 and 17. Three different sizes of cluster (i.e. number of neighboring points used to form a cluster) were examined: 10, 15 and 20. The result of the experiment is shown in table IV and the description of the table is same as those in table II. Results are discussed further in *Discussion* section.

IV. DISCUSSION

Results in the tables I, II, III, IV show the performance of the algorithm at various parameter settings for four data sets. Performance of the standard decision tree approach has been shown at the bottom of the table as a base line. As observed from the synthetic data experiment results in table I and table II, the proposed method is better or at least as good as standard decision tree algorithm for all the parameter settings examined. Size of the tree in terms of number of nodes is significantly smaller than that of standard tree. When the number of approximate split points (cluster centers) increases, number of tree nodes decreases while accuracy increases as expected. Influence of the size of the cluster is however not clearly evident in the results.

Purpose of the experiment on semi-synthetic iris data set was to examine the robustness of the algorithm when the bins of histogram do not have any inherent dependencies. The result as shown in table III suggest that the performance of the proposed method does not suffer heavily because of non-informative bins. Accuracy performance was almost comparable with base line performance except in some cases where accuracy drops by around 7%. One reason for this high variation in performance could probably be attributed to the small size of data set which is only 150. On the best parameter settings, accuracy raised up to 92.33% which was very close to base line performance of 92.67%.

Experiment results on operational data presented in table IV could not decide clear winner. With some parameter settings, average accuracy of proposed method exceeded base line performance while at other instances performance dropped well below the base line. It should however be noted that since the purpose of the experiment is the comparison of proposed approach against standard tree method, poor performance in both approach should not be a concern. This poor performance could be attributed to the smaller size of dataset or insignificance of variables selected. Since, some of the variables were heavily reduced in size by merging the bins together, we might have lost the information about the patterns inherent in those bins. Presence of such inherent dependency among bins is where the proposed method thrives on. This probably could be one reason why proposed method could not perform better all the time.

Considering all four experiments as a whole, out of 36 occasions, histogram approach was better in 21 occasions although most of the wins were from synthetic data experiments. Although promising, there are some downsides to this approach. The histogram approach has some inherent limitations. The first limitation is due to number of bins, the higher the number of bins the higher the computational complexity. So, somehow, higher number of bins have to be merged to get fewer bins which will result in information loss. Another inherent limitation of the approach lies in the least number of observations required at each node for making a split decision. Since, solving the system of linear equations lies at the heart of the model, at least as many observations are needed as there are number of bins in order to be able to solve such a system. One of the inherent limitation of the proposed method is that it assumes linear separation in the data and tries to approximate linear separation when the decision boundary is nonlinear. This linear approximation of possibly nonlinear pattern in histogram variables in operational data could be another reason why the method was not always the winner.

Probably the most prominent limitation of the approach lies in interpretation of split condition. Unlike in standard decision tree where nodes store information about best split variable and split point, here each node stores the information about the best split plane. It is usually difficult to interpret the split plane in the context of the variable.

V. Concluding Remarks

Standard learning algorithms are designed to learn from numeric and categorical variables. However, practitioners in both academia and industry often have to deal with more complex variables. Histograms in the form of frequency distributions is one of such example. To the best of our knowledge, no previous learning algorithms have been designed to generate classifiers from histogram variables. Instead, bins of histograms are commonly handled as separate variables, ignoring the underlying structure. In this paper, we adapted the standard decision tree learning algorithm to learn from histogram variables. Experimental results from both synthetic and real-world datasets would suggest that gains in terms of predictive performance and a reduction of the number of nodes might be obtained by exploiting the underlying structure of histogram variables.

Although encouraging, the proposed approach could be refined in various ways. Some directions for improvement include investigating techniques for efficiently handling histogram variables with large numbers of bins which at the moment is addressed simply by merging some of them thereby losing information. Comprehensive study of comparing the performance of the proposed method against existing multivariate split methods can be done in future. This study has shown one of the ways to extend the basic decision tree learning algorithm to handle histogram data; directions for future research include similar extensions of the numerous other standard learning algorithms, e.g., SVMs, random forests etc.

Acknowledgment

This work has been funded by Scania CV AB and the Vinnova program for Strategic Vehicle Research and Innovation (FFI)-Transport Efficiency.

References

[1] L. Billard and E. Diday, *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, ser. Applied Optimization. Chichester, England ; Hoboken, NJ: John Wiley and Sons Inc., 2006.

[2] E. Diday, "Principal component analysis for categorical histogram data: Some open directions of research," in *Classification and Multivariate Analysis for Complex Data Structures*, ser. Studies in Classification, Data Analysis, and Knowledge Organization, B. Fichet, D. Piccolo, R. Verde, and M. Vichi, Eds. Springer Berlin Heidelberg, 2011, pp. 3–15. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-13312-1_1

[3] A. Irpino and R. Verde, "Linear regression for numeric symbolic variables: an ordinary least squares approach based on wasserstein distance," 2012.

[4] S. Dias and P. Brito, "Distribution and Symmetric Distribution Regression Model for Histogram-Valued Variables," *ArXiv e-prints*, Mar. 2013.

[5] P. Nagabhushan and R. Pradeep Kumar, "Histogram pca," in *Advances in Neural Networks ISNN 2007*, ser. Lecture Notes in Computer Science, D. Liu, S. Fei, Z. Hou, H. Zhang, and C. Sun, Eds. Springer Berlin Heidelberg, 2007, vol. 4492, pp. 1012–1021. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-72393-6_120

[6] A. Irpino and R. Verde, "A new wasserstein based distance for the hierarchical clustering of histogram symbolic data," in *Data Science and Classification*, ser. Studies in Classification, Data Analysis, and Knowledge Organization, V. Batagelj, H.-H. Bock, A. Ferligoj, and A. iberna, Eds. Springer Berlin Heidelberg, 2006, pp. 185–192. [Online]. Available: http://dx.doi.org/10.1007/3-540-34416-0_20

[7] J. Aitchison, *The Statistical Analysis of Compositional Data*. London: Chapman and Hall, 1986.

[8] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks, 1984.

[9] P. E. Utgoff and C. E. Brodley, "An incremental method for finding multivariate splits for decision trees," in *In Proceedings of the Seventh International Conference on Machine Learning*. Morgan Kaufmann, 1990, pp. 58–65.

[10] I. Sethi and J. Yoo, "Design of multicategory multifeature split decision trees using perceptron learning," in *Pattern Recognition*, vol. 27, pp. 939–947.

[11] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

Comparison Between Random Forest Algorithm and J48 Decision Trees Applied to the Classification of Power Quality Disturbances

Fábbio A. S. Borges, Ricardo A. S. Fernandes, Member, IEEE, Lucas A. M. and Ivan N. Silva, Member, IEEE

***Abstract*— This paper presents a methodology for the classification of disorders related to the area of Power Quality. Therefore, we used the Random Forest algorithm, which corresponds to an effective data mining technique, especially when dealing with large amounts of data. This algorithm uses a set of classifiers based on decision trees. In this sense, the performance of the proposed methodology was evaluated in a comparative way between the Random Forest and the type J48 Decision Tree. For this analysis to be possible, synthetic electrical signals were generated, where this disturbances were modeled through parametric equations. After the performance analysis, it was observed that the results were promising, since the Random Forest algorithm has the best performance.**

Index Terms— Random Forest, power quality disturbances.

I. INTRODUCTION

Disturbances related to the area of Power Quality (PQ) are characterized by changing the waveforms of sinusoidal voltage and current, which can affect the operation of certain equipment [1]. Among these disturbances, there are sags, swells, interruption, harmonic distortion, oscillatory transients. Such disturbances are becoming a problem for both the power utilities as well as consumers, making it necessary to eliminate or mitigate the cause of their occurrence in order to ensure good power quality.

Thus, the detection and classification of disturbances is a primary task so that measures to control and mitigate disturbances can be adopted. However, this is no easy task, because the identification of these disturbances often require the analysis of a large amount of data measured by equipment installed on the network, besides the fact that many of the disturbances have similar features.

In this context, it is desirable to employ data mining tools, because they can identify these PQ disturbances in a fast and automated manner. Additionally, it is desirable that these tools are able to analyze a large volume of data and to acknowledge a pattern in the data in order to relate it to a possible disturbance.

The area of disturbance detection and classification has been the subject of several studies in recent years [2]. These studies utilize techniques to extract relevant signal characteristics, they reduce the dimensionality of the input data and remove redundant features of the original vector. The extracted features are then used as inputs to a method of pattern classification responsible for relating an input vector with a disturbance. Among the most used methods we highlight Fuzzy Logic, Artificial Neural Network and Support Vector Machine (SVM).

Following the above context, this work proposes the Random Forest algorithm with the interest to hold a review/classification for a database composed of waveforms that contain power quality disturbance. Random Forest is developed by Leo Breiman [3]. RF fits many classification trees to a data set and then combines the prediction from all the correlated trees. Each tree depends on the value of a separately sampled random vector.

During the feature extraction step, time calculations on time domain that have low computational effort are used. Following, the feature vector is used as the RF input so that the final response is defined by the class that has the highest number of outputs, that is, by the account of the outputs presented by each of the decision trees that compose the algorithm. Finally, the classification results are obtained and compared with the response of a Decision Tree (DT) that uses the training algorithm J48.

II. DATABASE COMPOSED OF SYNTHETIC SIGNALS

The objective of the database modeling is to store the maximum number of signals with different characteristics of the disturbance. These signs will be used to test the proposed methodology. In this work, the occurrence of the following disturbances was considered: voltage sags, swells, flickers, harmonic distortion, voltage interruptions, oscillatory transients, voltage sags in conjunction with harmonic distortion and swells together with harmonic distortions. A database was created consisting of windows obtained for synthetically modeled disturbances, based on mathematical models proposed in [4].

Therefore, the windows that make up the database were derived from 100 case studies for each disturbance, and each of the disturbances has a total of 10 cycles at nominal frequency of 60 Hz and sampled rate of 128 points per cycle. This windowing occurs through the shifting of the data window (which is the size of one cycle of the signal) in steps of 1 point until it covers the entire length of the signal.

Fábbio A. S. Borges, Lucas A. M and Ivan N. Silva are with the Department of Electrical and Computing Engineering, University of São Paulo, São Carlos, 13566-590, Brazil (e-mail: fabbioanderson@gmail.com, lucas.moraes@usp.br, insilva@sc.usp.br). Ricardo A. S. Fernandes is with the Department of Electrical Engineering , Federal University of São Carlos, 13565-905, São Carlos, Brazil (e-mail: ricardo.asf@ufscar.br)

The result of the process is the construction of a database comprised of approximately: 14864 sag windows, 12671 swell windows, 19706 flicker windows, 34084 harmonic distortion windows, 13277 harmonic distortion with windows, 12769 harmonic distortion with swell windows, 10366 interruption windows and 5836 transient windows

III. FEATURE EXTRACTION FROM THE WINDOWED SIGNALS

As soon as a disturbance is detected, the classification step is activated, which uses a stage of extraction of features combined to a decision tree. Thus, a set of 11 features is extracted with the purpose of reducing the dimension of data and hence reducing the computational effort. This set consists of the following features: standard deviation, entropy, Rényi entropy, Shannon entropy, mean deviation, Kurtosis, RMS value, crest factor, the balance between the maximum and minimum amplitude and peak value. Thus, for each dj data a Ck vector is extracted, where j represents the index of each element contained in the window and that varies in the range $\{1 \rightarrow N\}$ N is the size window; k represents each characteristic in the range $\{1 \rightarrow 11\}$.

IV. RANDOM FOREST

Random Forest corresponds to a collection of combined Decision Tree {hk(x,Tk)}, where k = 1,2,...,L where L is number the tree and Tk is the training set built at random and identically distributed, hk represents the tree created from the vector Tk and is responsible for producing an output x.

Decision Trees are tools that use divide-and-conquer strategies as a form of learning by induction [5], Thus, this tool uses a tree representation, which helps in pattern classification in data sets, being hierarchically structured in a set of interconnected nodes. The internal nodes test an input attribute/feature in relation to a decision constant and, this way, determine what will be the next descending node. Therefore, the nodes considered as leaves classify the instances that reach them according to the associated label.

The trees that make up the Random Forest are built randomly selecting m (value fixed for all nodes) attributes in each node of the tree; where the best attribute is chosen to divide the node. The vector used for training each tree is obtained using random selection of the instances. Thus, to determine the class of an instance, all of the trees indicate an output, where the most voted is selected as the final result. So the classification error depends on the strength of individual trees of the forest and the correlation between any two trees in the forest.

V. RESULTS

As previously mentioned, the decision trees and the Random Forest were trained and validated using a set of data consisting of the signals windows acquired from the database. Thus, the training set is composed of 70% of the windows and the test/validation suite corresponds to the 30% remaining windows. The random forest is formed by 10 decision tree and the number of attributes selected in each node is equal to 4. This made it possible to obtain and evaluate the success rate for each disturbance, as well as the average accuracy of classifiers. The comparison of the classification results is presented in Table I.

TABLE I. RESULTS OBTAINED FOR SYNTHETIC SIGNALS .

Power Quality Disturbances	**DT**	**RFs**
Voltage Sags	83,0%	99,4%
Voltage Swells	94,4%	99,9%
Flickers	97,9%	99,9%
Harmonic Distortions	96,9%	99,6%
Voltage Sags with Harmonic Distortions	78,9%	98,5%
Voltage Swells with Harmonic Distortions	88,8%	98,9%
Voltage Interruptions	89,4%	99,3%
Oscillatory Transients	87,5%	99,2%
Mean Precision	**89,6%**	**99,3%**

Through the results presented in Table I it is found that the performance of the two used classifiers is satisfactory, however, it can be seen that the approach based on Random Forest presents better results when compared with the approach based on type J48 Decision Trees. The RF had a precision 10% higher than the DT, additionally the proposed algorithm can identify large part of the disturbances with an accuracy greater than 99%.

VI. CONCLUSIONS

The paper presents a performance comparison between type J48 Decision Trees and the Random Forest algorithm for classification of power quality disturbances. According to the results, we note that the worst performances were obtained for windowss containing combinations of voltage sags with harmonic distortion and swells with harmonic distortion (respectively, 98.5% and 98.9%). Therefore, in general, the results may be considered satisfactory for electric power systems.

REFERENCES

[1] R. C. Dugan, M. F. Mc Granagham, S. Santoso, and H. W. Beaty, Electrical Power Systems Quality, 3rd edition. New York, 2002.

[2] D. Granados-Lieberman, R. J. Romero-Troncoso, R. A. Osornio-Rios, A. Garcia-Perez, and E. Cabal-Yepez, "Techniques and Methodologies for Power Quality Analysis and Disturbances Classification in Power Systems: A Review," IET Generation, Transmission & Distribution, vol. 5, no. 4, pp. 519-529, 2011.

[3] L. Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[4] H. Erişti, A. Uçar and Y. Demir, "Wavelet-based feature extraction and selection for classification of power system disturbances using support vector machines," Electric Power Systems Research, vol. 80, pp. 743-752, 2010.

[5] I. H. Witten and E. Frank, Data Mining: Pratical Machine Learning Tools and Techniques. Morgan Kaufmann, 2005G.

Modelling Ground-Level Ozone Concentration using Ensemble Learning Algorithms

Eman S. Al Abri[1], Eran A. Edirisinghe[1], and Amin Nawadha[2]
[1]Department of Computer Science, Loughborough University, United Kingdom
[2] Environment Research Centre, Sohar University, Sultanate of Oman.

Abstract— *Environmental risks caused by exposure to ground level ozone have significantly increased during recent years. One main producer of ozone is the photochemical reaction between volatile organic components and the anthropogenic nitrogen oxides created by vehicular traffic. Therefore the measurement and monitoring of atmospheric ozone concentration levels is important. In this paper we propose a study of the use of state-of-the-art machine learning approaches in modelling the concentration of ground level ozone. The prediction is based on concentrations of seven gases (NO_2, SO_2, and BTX (Benzene, Toluene, o-,m-,p-Xylene) and six meteorological parameters (ambient temperature, air pressure, wind speed, wind direction, global radiation, and relative humidity). The analysis of the results indicates that accurate models for the concentration of ground level ozone can be derived with the best performance accuracies indicated by the Ensemble Learning Algorithms. The investigation carried out compares the use of different machine learning classifiers and show that the Ensemble-classifier Bagging performs superior to standard single classifiers, such as Artificial Neural Networks and Support Vector Machines, popularly used in literature. In addition, we study the performance of the meta-classifier Bagging when different base classifiers are used in optimised configurations and compare the results thus obtained. The research conducted bridges an existing research gap in big-data analytics related to environment pollution prediction, where present research is largely limited to using standard learning algorithms such as Neural Networks and Support Vector Machines often available within popular commercial software packages.*

Keywords: Ozone, Atmospheric pollution, machine learning, Environment Science, Ensemble classifiers

1 Introduction

Ozone is a trans-boundary air pollutant that can be formed by photochemical reactions between anthropogenic nitrogen oxides(NO_x) and Volatile Organic Compounds (COVs) in the presence of sunlight[1]. When O_3 is formed, it remains suspended in the lower atmosphere (ground level ozone) for hours to days depending on the meteorological conditions and can endanger local and regional receptors.

In recent years, the environmental risks caused by exposure to ground level ozone (O_3) from both stationary and mobile sources have increased annually[2]. Several studies that analyse the effects of meteorological conditions on the formation and transport of O_3 have been listed in the work of [3]. Further, statistically significant relationships have been identified between elevated concentrations of O_3 and environmental risks in [4],[5].

A number of studies in the field of environmental science and engineering have focused their interest on constructing models to predict the concentrations of gases that result in air pollution. The majority of environmental researchers tend to use Artificial Neural Networks (ANN) and Support Vector machines (SVM) to predict ozone concentration [6]-[11]. Although there are more developed data mining / machine learning techniques, such as Ensemble learning approaches [12], only two attempts have investigated their use in predicting the ozone concentration; they are the work of [13] and [14]. This research showed that improvements in predictions can be obtained using bagging[15] as against using the popular single classifiers such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM). However these investigations were limited in the fact that Bagging was used only with the default single classifier RepTree[12] in WEKA (Waikato Environment for Knowledge Analysis) toolkit [16] as the base classifier. In the field of air pollution monitoring, no attempt has been made to test other ensemble classifier, select the best base classifier or to optimise the performance of the base classifier based on various possible parameter selections, all of which can lead to significant improvements in prediction accuracies. On the other hand, several attempts have been made in areas beyond air quality prediction in the use of ensemble classifiers, such as in bioinformatics, medicine and marketing, to build predictive models [17]-[21]. This work has shown that ensemble classifiers outperform the corresponding single classifiers and that the ultimate answer to the question, which classifier works best, depends on the dataset. It is clear that different datasets, in particular from different application domains, are statistically different and this has a high impact on the variability of results obtainable from different classifiers.

From the review of literature conducted and summarised above, a lack of research into effectively utilising Ensemble learning to predict ozone concentration was identified. Therefore the research proposed in this paper aims to find

accurate models that can be used to predict ground level ozone concentrations, given a multitude of environmental parameters. An investigation was carried out comparing the performance of several machine learning techniques. Multiple predictive models were built using popular single classifiers namely Multilayer Perceptron (MLP) and Support Vector Machines and two ensemble learning algorithms, namely Bagging and Random Forests[22], using the WEKA toolkit. In addition, comparative analyses were performed to determine the algorithm that produced the best performance and to optimize the performance of each selected approach. The dataset considered in this work was obtained from Sohar University, Oman, which used a DOAS instrument [23] to gather the environmental data. The dataset includes concentrations of eight gases (O_3, NO_2, SO_2, and BTX (Benzene, Toluene, o-,m-,p-Xylene)) and six meteorological parameters (ambient temperature, air pressure, wind speed and direction, solar radiation, and relative humidity).

As implantations of the machine learning algorithms used for pre-processing/data-cleaning, feature selection, optimizing classifier parameters, modelling and performance analysis, WEKA has been used throughout this paper. Initially, training phases based on different classification algorithms for predicting O_3 concentration were performed. Subsequently, the prediction performance of different algorithms, were examined using ten-fold cross validation as implemented in WEKA. Various evaluation metrics have been utilised to analyse the results. It should be noted the key focus of the research conducted is not time-series analysis of O_3 concentration (i.e. predicting how O_3 concentration changes with time) but how to predict O_3 concentration based on the concentrations of the primary pollutant gases and the environmental parameters that can have an impact. In particular when O_3 creation is assumed to be due to the production of primary pollutant Nitrogen Dioxide, generated by vehicular traffic in this area, the time dependent analysis is not essentially useful.

For clarity of presentation this paper is divided into several sections. Apart from this section that provided the reader with an insight to the research context and identified research gaps, section-2 provides the background to data collection and presentation. Section-3 details the experimental procedure followed and section-4 provides the experimental results and a detailed analysis of the results. Finally section-5 concludes with an insight into future research.

2 Data collection and representation

This section provides details of the data collection approach used and how this data was represented for subsequent analysis.

2.1 The sampling site

Measurements were recorded across the Sohar Highway (SHW), Oman, in front of the main entrance to the Sohar University (SU) with a Differential Optical Absorption System (DOAS) instrument that was professionally installed (see Fig.1. for an aerial view of the system). The light beam travels a round-trip of 477 meters from A, which is located on the roof of the main administrative office building of SU, to B, where a reflector (or receiver) is installed on the top of another building situated across the road, as illustrated in Fig.1. The SHW has two lanes in each direction and an additional two single carriageway roads, in parallel, on both sides, bringing the total number of lanes to eight. Additionally, there is the SU car park, marked as C, where vehicular traffic may be present and thus would result in higher levels of O_3 concentrations. In order to capture the rapid variations of the concentrations of gases present in the space of the monitoring path, evaluations of light captured by the DOAS instrument is performed every 30 seconds for the measurement of the concentrations of O_3, NO_2, and SO_2 gases and every one minute for measurement of the concentrations of BTX. Additionally, the meteorological parameters, including wind speed and direction, relative humidity, pressure, temperature, precipitation, global solar radiation etc., are separately measured by sensors located on the roof of the SU building at A. The height of the instruments from ground level was approximately 12 metres.

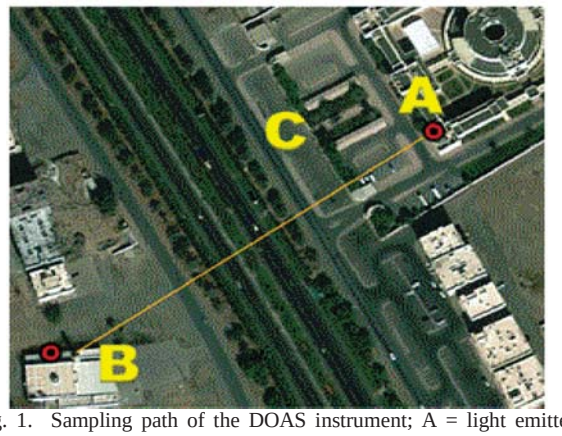


Fig. 1. Sampling path of the DOAS instrument; A = light emitter location, B = reflector location and C = car park.

2.2 Ozone dataset

The dataset used for the experiment was captured by the Sohar University DOAS system during 2013/2014. However, due to a technical fault in the system, the dataset collected during the specified period is not continuous. Nonetheless, a sufficiently large dataset was gathered to make the experiments statistically relevant. This dataset was analysed to investigate the modelling algorithm that gives the best prediction accuracy.

In the dataset used so far, there are a total of 6,744 instances spread across the years 2013-2014, as detailed in Table I.

TABLE I
DATASET DESCRIPTION

Dataset	2013			2014			Total number of records
	Start Date	End Date	No. of Rec.	Start Date	End Date	No. of Rec.	
	1st April 2013	23rd Aug. 2013	3480	1st March 2014	14th July 2014	3264	6744

2.3 Dataset representation

The target dataset is a sequence of measurements presented in a time series. The measurements are concentration values of eight gases measured in μgm^{-3} and readings of six environmental parameters. Table II lists the 14 attributes of each measured data value with their descriptive statistics.

TABLE II
ATTRIBUTES OF THE DATASET

2013-2014	Unit	Min	Max	Standard deviation	Mean
Sulphur Dioxide (SO_2)	μgm^{-3}	1.61	15.11	2.33	4.96
Nitrogen Dioxide (NO_2)	μgm^{-3}	0.02	83.99	16.65	18.24
Ozone (O_3)	μgm^{-3}	0.85	139.50	24.25	43.25
Benzene (C_6H_6)	μgm^{-3}	0.05	19.56	4.17	6.13
Toluene (C_7H_8)	μgm^{-3}	0.73	47.14	7.77	15.16
p-Xylene (C_8H_{10}(p))	μgm^{-3}	0.10	8.75	1.18	3.30
m-Xylene (C_8H_{10}(m))	μgm^{-3}	0.69	5.44	0.52	2.44
o-Xylene (C_8H_{10}(o))	μgm^{-3}	0.80	58.15	6.91	29.56
Temperature	ºC	16.19	45.06	3.53	31.10
Relative Humidity	%	8.47	93.57	19.33	64.38
Pressure	kPa	98.94	102.89	0.56	100.19
Global Radiation	W/m^2	-2.75	1120.24	247.95	201.13
Wind speed	m/s	0.31	6.266	1.02	1.77
Wind Direction	degree	0.11	359.99	91.50	137.52

Having collected the above dataset section-3 presents the method adopted for its analysis and detailed investigation.

3 Proposed method

The proposed approach adopts standard data mining procedure that involves data pre-processing prior to data modelling using machine learning. WEKA (version 3.7.11) is a toolkit that supports open source software implementation and operation of a large number of options for both data pre-processing and modelling. In this section we introduce the reader to the specific data pre-processing and modelling algorithms that have be adopted within the research context of the proposed work, as implemented by WEKA. Note that for our data analysis and method evaluation comparison purposes both Explorer and Experimenter software environments have been used, as appropriate.

3.1 Data pre-processing

Outlier Removal: In the data captured by the DOAS, missing values are recorded as -999.00. A careful analysis of the captured data also revealed that there are data measurement outliers, which could have resulted from instances of temporary sensor malfunctioning due to dust, high temperatures and overheating. Therefore a data cleaning operation within WEKA (listed under pre-processing) was utilised for the removal of outliers. The two filters interquartileRange (filters -> unsupervised -> attribute -> interquartileRange) and removeWithValues (filters -> unsupervised -> instances -> removeWithValues) were used respectively to clean the data in hand. Note that the first filter adds two extra columns to the data to indicate instances which contains the outliers and extreme values and the second filter removes such data by refereeing to the extra columns added by the first filter. After this cleaning process, only approximately 62% (4,173 out of 6,744 instances) of the original dataset were utilised for the next stage (modelling phase).

Data transformations: Since the wind direction is originally measured as an angle from the north in a clockwise direction, with values ranging from 0-360 degrees, the originally recorded with related data will have to be re-represented to avoid 0 and 360 degree directions being considered as different. The Wind Speed (WS) and Wind Direction (WD) have been combined and divided into two orthogonal compenents u = WS×cos(WD) and v = WS×sin(WD). (u,v) parameters will replace (WS, WD) in order to compensate for the above issue with regards to the original value of WD.

Attribute selection: Reduction of the attributes by eliminating the msot insignificant attributes can lead to both improved accuracy and speed of data modelling. The use of three popular feature selection filters have been investigated in the proposed research, namely, CFS Subset Evaluator with Best First and Greedy Stepwise Search methods, ReliefFAttributeEval with attribute ranking (removed last three attributes), and Principal Components. In the

experiments conducted it was revealed that none of these filters uenhanced the accuracy of modelling although in the case of using the RelieFAttributeEval filter three of the most insignificant features were removed from used in modelling thus impacting positvely on speed.

3.2 Modelling the ozone concentration

As previously stated WEKA consists of implementations of a large number of classifiers that includes all state-of-the art and the popular traditional classifiers, such as, the Artificial Neural Networks and the Support Vector Machines. Our detailed experiments were designed to test all possible classifiers as single classifiers and as combined approaches (as appropriate). The purpose of this exhaustive investigation was to find the best classifiers / classifier combinations that outperformed traditional approaches for the prediction of air (Ozone) pollution thus generating new and useful knowledge for the community involved in environmental science and engineering research.

The initial exhaustive list was reduced to investigating 16 learning algorithms in detail from WEKA classifier categories, namely, Functions (4 different functions), Lazy (3), Meta (2), Rule (2) and Tree (5). The two meta-classifiers included the two popular Ensemble learning approaches Bagging and Random Forests. Furthermore, more detailed investigations were conducted with the Bagging ensemble classifier due to the initial indication of its superiority of performance. Within the detailed experiments thus conducted all the single learner classifiers initially experimented, were utilised as the base classifier of the ensemble classifier, Bagging.

Within the experimental context of this paper only six classification algorithms are analysed and discussed in detail. These include the two most popular single learning algorithms used in research that focus on air pollution analysis, Artificial Neural Networks [ANN] (implemented in WEKA as Multi-Layer Perceptron [MLP]) and Support Vector Machines [SVM] (implemented in WEKA as SMOreg) and the basic Ensemble Classifier, Random Forest [RF]. In conducting more detailed performance analysis of Bagging, the above three experiments are complemented with using them within Bagging as a base-classifier, namely Bagging with MLP, Bagging with SMOreg and Bagging with Random Forest. Although a large number of other classifiers and classifier combinations were evaluated, the detailed analysis of only these algorithms is presented in section-4. The accuracy of the algorithms are evaluated using two widely used evaluation metrics: Correlation Coefficient, Mean Absolute Error.

To present a fair performance comparison between the prediction models presented, optimal parameters for each classifier has been examined prior to conducting detailed modelling. The CVParameterSelection optimisation algorithm of WEKA has been used for this purpose.

The Explorer GUI environment of WEKA has been used to construct individual classifier models using their optimal parameters settings. Hence, the performance of different classifiers have been analysed and compared, using the same dataset (see section 2) using the Explorer. Since the Explorer does not provide the statistical significance of the improvements achievable by different classifiers, WEKA's Experimenter GUI environment was utilised to obtain additional information. A statistical test (Paired T-Tester corrected) was used to calculate the statistical significance between the different predictive models. The performance of the classifiers were examined using 10 fold cross validation and was compared using the Correlation Coefficient.

4 Experimental results and analyses

Experiments were conducted to analyse and compare the performance of six different classifiers: MLP (WEKA's ANN implementation), SMOreg (WEKA's SVM implementation), Random Forest (RF), Bagged-MLP, Bagged-SMOreg and Bagged-RF. Further detailed experiments were also conducted to determine the potential impact of feature reduction / selection and in the selection of classifier parameters in optimising classifiers, in the overall accuracy obtainable from each of the six evaluated classifiers. Further the original readings recorded for wind direction was a measure in the range 0-360 degrees. In order to compensate for the fact that 0 and 360 degree readings mean the same, we have combined wind direction (WD) with wind speed (WS) to replace them with two orthogonal components WS×cos(WD) and WS×sin(WD).

It is noted that all of the classifiers investigated (i.e. regardless of whether the classifier is of the single classifier type or the ensemble classifier type) consist of a number of input parameters that may have a vital impact on the accuracy of predictions obtainable. Although WEKA provides default parameter values for each classifier, our preliminary experiments suggested that these values do not result in optimised prediction. Therefore it was vital to select a set of parameters which provide optimal prediction accuracy. For this purpose we made use of WEKA's CVParameterSelection filter. Table III tabulates the prediction accuracy obtainable via each approach in terms of correlation coefficient. The results indicate that the optimal parameter selection has a positive impact only when use the single classifiers MLP (i.e. ANN) and SMOreg (i.e. SVM for regression). When using ensemble classifiers Random Forest and Bagging, the optimal parameter selection algorithm has no impact, indicated by the accuracy figures that remain unchanged. It is noted that even though the CVParameterSelection filter changes some parameters in its attempt to optimise the accuracy, no change is indicated in comparison to the accuracy obtainable using default settings.

For clarity of comparison Table IV tabulates overall prediction accuracies obtainable by each classifier presented in terms of the Co-relation Coefficient and Mean

Absolute Error with both using the default parameter settings of WEKA and with optimised parameter settings.

Fig.2 illustrates graphs representing the actual Ozone concentration vs the predicted Ozone concentrations. The graphs illustrate the better prediction capability of Bagged Random Forest classification approach as compared to the others. Data points lie closer to the line of approximation (less spread) than in the other graphs.

TABLE III
EXPERIMENTS TO OPTIMISE THE CLASSIFIERS

Classifier Name	Default settings	Correlation Coefficient	Optimal Parameters	Correlation Coefficient
MLP	Learning Rate (L)=0.3 Momentum /(M)=0.2 Hidden layer= a (attribute/class)/2	0.85	Learning Rate(L)=0.1 Momentum (M)=0.1 Hidden layer= 5	0.88
Bagged MLP	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 MLP: Learning Rate (L)=0.3 Momentum /(M)=0.2 Hidden layer= a (attribute/class)/2	0.90	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 MLP: Learning Rate(L)=0.1 Momentum (M)=0.1 Hidden layer= 5	0.90
Random Forest	NumTree (I)=10 NumFeature (K)=0	0.92	NumTree (I)=20 NumFeature (K)=0	0.92
Bagged RandomForest	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 Random Forest: NumTree (I)=10 NumFeature (K)=0	0.92	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 Random Forest: NumTree (I)=20 NumFeature (K)=0	0.92
SMOreg	C:1.0 Kernal: polyKernel	0.84	C:1.0 Kernel: NormalizedPolyKernel	0.89
Bagged SMOreg	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 SMOreg: C:1.0 Kernal: polyKernel	0.84	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 SMOreg: C:1.0 Kernal: NormalizedPolyKernel	0.89

TABLE IV
RESULTS OF THE PREDICTION MODELS

	Default Parameters		Optimal Parameters	
Classifier	**Correlation Coefficient**	**Mean Absolute Error**	**Correlation Coefficient**	**Mean Absolute Error**
MLP	0.85	9.81	0.88	8.51
SMOreg	0.84	9.54	0.89	8.05
RandomForest	0.91	7.52	0.92	7.16
Bagged MLP	0.90	7.64	0.91	7.27
Bagged RandomForest	0.92	7.08	0.92	7.05
Bagged SMOreg	0.84	9.54	0.89	8.04

Table V tabulates the accuracy values obtained when using four different attribute filtering approaches implemented within WEKA, namely, CFS Subset Evaluator, with Best First and Greedy Stepwise search, ReliefF Attribute Evaluator and Principle Component Analysis. The results indicate that no improvement of accuracy is achieved in comparison with using all attributes. We also investigated the impact of removing wind direction from being considered, taking only the wind speed into account (from the original data recorded). It was seen that the wind direction has negligible impact on the Ozone concentration prediction accuracy. This is justifiable as the measurements for Ozone was done across the road, i.e. at its source, as it was vehicular traffic that was suspected to create the Ozone from the Nitrogen Dioxide emissions.

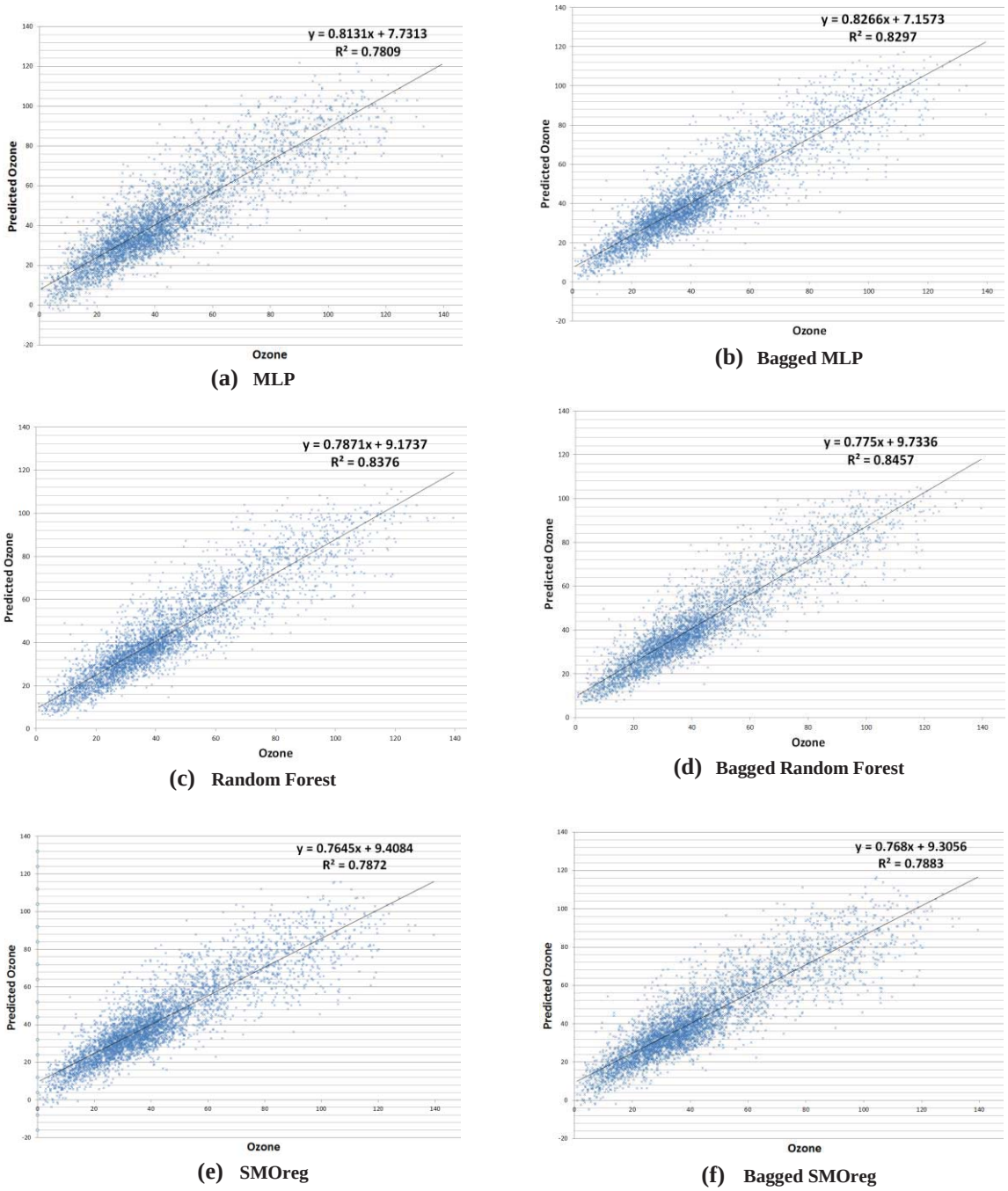


Fig. 2. Scatter Plots of the actual and predicted Ozone for 6 Models

TABLE V
RESULTS OF APPLYING FEATURE SELECTION

	MLP	SMOreg	Random Forest	Bagged MLP	Bagged SMOreg	Bagged RandomForest
CFS-Best First	0.82 (-3)	0.82 (-2)	0.89 (-3)	087 (-3)	0.82 (-2)	0.90 (-2)
CFS-Greedy Stepwise	0.81 (-4)	0.82 (-2)	0.88 (-4)	0.86 (-4)	0.82 (-2)	0.90 (-2)
ReliefF Att. Eval.	0.83 (-2)	0.83 (-1)	0.91 (-1)	0.89 (-1)	0.83 (-1)	0.92 (0)
PCA	0.84 (-1)	0.83 (-1)	0.87 (-5)	0.89 (-1)	0.83 (-1)	0.89 (-3)
Using All Attributes	**0.85**	**0.84**	**0.92**	**0.90**	**0.84**	**0.92**

Ensemble learning is an approach that uses different classification techniques to build up a single model. Proposed by Breiman, 1996, Bootstrap Aggregation (Bagging) is a common type of an ensemble learning approach. Bagging resamples the original data, by using the bootstrap method, randomly, but with replacement (some can be selected repeatedly while other may not). The data produced are different from each other, however, the size of these samples are equal. Subsequently, a tree is built up from each sample. Later a classification model is developed from each sample using a single learning algorithm. Subsequently the outputs of different models are integrated into a single predication model. It uses either the weighted vote or average vote, depending on the type of task (i.e., a classification task or regression task, respectively). Due to the above process adopted by Bagging it resolves the data over-fitting problem associated with most classifiers, in this case with MLP and SVM in particular.

This is the reason for the significantly better prediction accuracies obtainable from using the Ensemble Classifier Bagging as against the accuracies obtainable from the traditional single classifiers commonly used in predicting Ozone, ANN and SVM.

5 Conclusion and future works

In this paper we have compared the performance of six machine learning algorithms in predicting the ground level atmospheric ozone concentrations. The prediction was based on concentrations of seven gases (NO2, SO2, and BTX (Benzene, Toluene, o-,m-,p-Xylene) and six meteorological parameters (ambient temperature, air pressure, wind speed, wind direction, global radiation, and relative humidity). Results prove the ability of ensemble learning algorithms, Random Forests and Bagging to perform significantly better than the traditional single classifier based learning algorithms, Artificial Neural Networks and Support Vector Machines.

We are currently extending the research presented within this paper to predict Ozone concentration variations over long periods of time, extending beyond a five year period, attempting to identify patterns and trends.

6 Acknowledgment

Authors would like to acknowledge the support of the research Council and the Ministry of Manpower of the Sultanate of Oman for providing the funding for this research and Loughborough University UK for providing research support. The research leading to these results has received research project grant funding from the Research Council of the Sultanate of Oman, research grant agreement no [ORG SU EBR 12 013].

7 References

[1] U.S. Environmental Protection Agency, "Guidelines for Developing an Air Quality (ozone and PM2.5) Forecasting Program," 2003.

[2] Region 7 Air Program, "Health Effects of Air Pollution," *EPA*. [Online]. Available: http://www.epa.gov/region07/air/quality/health.htm. [Accessed: 28-Feb-2015].

[3] D. M. Agudelo–Castaneda, E. C. Teixeira, and F. N. Pereira, "Time–series analysis of surface ozone and nitrogen oxides concentrations in an urban area at Brazil," *Atmos. Pollut. Res.*, vol. 5, pp. 411–420, 2014.

[4] M. Jerrett, R. T. Burnett, A. P. I. C, K. Ito, G. Thurston, D. Krewski, Y. Shi, E. Calle, and M. Thun, "Ozone exposure and mortality.," *N. Engl. J. Med.*, vol. 360, p. 2788; author reply 2788–2789, 2009.

[5] WHO Regional Office for Europe, "Health risks of ozone from long-range transboundary air pollution," p. 111, 2008

[6] M. S. Baawain and A. S. Al-Serihi, "Systematic approach for the prediction of ground-level air pollution (around an industrial port) using an artificial neural network," *Aerosol Air Qual. Res.*, vol. 14, pp. 124–134, 2014.

[7] N. Loya, I. Olmos Pineda, D. Pinto, H. Gómez-Adorno, and Y. Alemán, "Forecast of air quality based on ozone by decision trees and neural networks," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7629 LNAI, pp. 97–106, 2013.

[8] B. Mileva-Boshkoska and M. Stankovski, "Prediction of missing data for ozone concentrations using support vector machines and radial basis neural networks," 2007.

[9] A. S. Luna, M. L. L. Paredes, G. C. G. de Oliveira, and S. M. Corrêa, "Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil," *Atmos. Environ.*, vol. 98, pp. 98–104, Dec. 2014.

[10] S. . Abdul-Wahab and S. . Al-Alawi, "Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks," *Environ. Model. Softw.*, vol. 17, no. 3, pp. 219–228, Jan. 2002.

[11] A. Coman, A. Ionescu, and Y. Candau, "Hourly ozone prediction for a 24-h horizon using neural networks," *Environ. Model. Softw.*, vol. 23, no. 12, pp. 1407–1421, Dec. 2008.

[12] I. H. Witten, E. Frank, and M. Hall, *Data Mining Practical Machine Learning Tools and Techniques*, 3rd ed. Elsevier, 2011.

[13] K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," *Atmos. Environ.*, vol. 80, pp. 426–437, Dec. 2013.

[14] A. J. Cannon and E. R. Lord, "Forecasting Summertime Surface-Level Ozone Concentrations in the Lower Fraser Valley of British Columbia: An Ensemble Neural Network Approach," *J. Air Waste Manage. Assoc.*, vol. 50, no. 3, pp. 322–339, Mar. 2000.

[15] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[16] WEKA; the University of Waikato, "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/index.html. [Accessed: 27-Feb-2015].

[17] P. Yang, Y. Hwa Yang, B. B. Zhou, and A. Y. Zomaya, "A Review of Ensemble Methods in Bioinformatics," vol. 5, pp. pp.296–308, 2010.

[18] E. Alfaro, N. García, M. Gámez, and D. Elizondo, "Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks," *Decis. Support Syst.*, vol. 45, no. 1, pp. 110–122, Apr. 2008.

[19] L. A. Gabralla and A. Abraham, "Prediction of Oil Prices Using Bagging and Random Subspace," Advances in Intelligent Systems and Computing, Proceedings of the Fifth International Conference on Innovations in Bio-Inspired Computing and Applications IBICA 2014 P. Kömer, et al. (eds)., Volume 303, pp. 343–354, 2014.

[20] A. Fathima, J. A. Mangai, and B. B. Gulyani, "An ensemble method for predicting biochemical oxygen demand in river water using data mining techniques," *Int. J. River Basin Manag.*, vol. 12, no. 4, pp. 357–366, Oct. 2014.

[21] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," *Int. J. Electr. Power Energy Syst.*, vol. 60, pp. 126–140, Sep. 2014.

[22] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[23] OPSIS, "UV DOAS Technique," 2014. [Online]. Available: http://opsis.se/Techniques/UVDOASTechnique/tabid/632/Default.aspx. [Accessed: 09-Feb-2015].

Automatic Detection of Small Groups of Persons, Influential Members, Relations and Hierarchy in Written Conversations Using Fuzzy Logic

French Pope III [1], Rouzbeh A. Shirvani [1], Mugizi Robert Rwebangira[2], Mohamed Chouikha[1], Ayo Taylor[2], Andres Alarcon Ramirez[1], Amirsina Torfi[1], french.pope@bison.howard.edu, Rouzbeh.asghari@gmail.com, rweba@scs.howard.edu, mchouikha@howard.edu, ayo.taylor@gmail.com, alarcon27@hotmail.com, amirsina.torfi@ bison.howard.edu

[1]Electrical and Computer Engineering, Howard University, Washington, D.C. 20059 USA
[2]Systems and Computer Science, Howard University, Washington, D.C. 20059 USA

***Abstract*— Nowadays a lot of data is collected in online forums. One of the key tasks is to determine the social structure of these online groups, for example the identification of subgroups within a larger group. We will approach the grouping of individual as a classification problem. The classifier will be based on fuzzy logic. The input to the classifier will be linguistic features and degree of relationships (among individuals). The output of the classifiers are the groupings of individuals. We also incorporate a method that ranks the members of the detected subgroup to identify the hierarchies in each subgroup. Data from the HBO television show *The Wire* is used to analyze the efficacy and usefulness of fuzzy logic based methods as alternative methods to classical statistical methods usually used for these problems. The proposed methodology could detect automatically the most influential members of each organization *The Wire* with 90% accuracy.**

Keywords- Fuzzy Logic; Text Conversations; subgroup Identification; hierarchy

I. INTRODUCTION

In the last decade there has been increasing use of online platforms such as opinion forums, chat groups, and social networks because of broad access to the internet and people's communication needs. This new way of communicating has allowed people with different customs, cultures, and locations to get together virtually to interact and sometimes cooperate around common interests. On the other hand, the motivation of many e-commerce companies for understanding the behavior of internet users as well as the interest of some security agencies for detecting security threats has created the need for analyzing the data generated by online communities. In addition, because of the massive use of online communication tools and large amount of information generated by their users, it is almost impossible to manually analyze all of the generated information. Therefore, there have lately been important efforts that seek to automatically analyze and extract relevant information from written data corresponding to dialogues among several persons. One of the active areas of research is to detect associations among the members of an online community by subgroup identification in written conversations. The idea of subgroup identification is to identify members from a community who have similar ways of thinking or have the same affiliation and may cooperate each other. Yessenalina et al. [1] proposed a methodology that classifies the speaker's side in a corpus of congressional floor debates, using the speaker's final vote on the bill as a labeling for side. This work infers agreement between speakers based on cases where one speaker mentions another by name, and a simple algorithm for determining the polarity of the sentence in which the mention occurs. Gupte et al. [2] address the problem of segmenting small group meetings in order to detect different group configurations in an intelligent environment. They propose an unsupervised method based on the calculation of the Jeffrey divergence between histograms of speech activity observations. These histograms are generated from adjacent windows of variable size slid from the beginning to the end of a meeting recording. Elson et al. [3] proposed a method for detecting social networks from nineteenth-century British novels and serials. They linked two characters based on whether or not they conversed.

Tan et al. [4] proposed an algorithm that seeks to detect groups of people in Twitter with the same affiliation. To do this, it assumes that connected users are more likely to hold similar opinions. Finally, the discussants were classified in groups based on how often they reply to each other. Kunegis et al. [5] studied user relationships in the Slashdot technology news site. Slashdot gives users the option of tagging other users as friends or foes, providing positive and negative endorsements. Abu et al. [6] identified subgroups in ideological discussions. To do this, they identified the discussion participants, comments, and the reply structure of the thread (i.e. who replies to whom). Then, they used sentiment analysis to determine the polarity of the comment (positive or negative) made by a particular participant. Finally, to identify the subgroup membership of each discussant, they use the fact that the attitude profiles of discussants who share the same opinion are more likely to be similar to each other than to the attitude profiles of discussants with opposing opinions. Hassan et al. [7] take into consideration the posts exchanged between participants and sentiment analysis to build a signed network representation of the discussion. After building the signed network representation of discussions, the large group of discussants split into many subgroups with coherent opinions.

Also of great interest is the identification of the hierarchy of the members from a particular subgroup. The hierarchy of a group is important because it allows us to detect the most

influential members from a group as well as the role and importance of each member in a group. In the case of identifying influential members, Rienks [8] proposed an algorithm for detecting influencers in a corpus of conversations. He focuses entirely on non-linguistic behavior and looks at (verbal) interruptions and topic initiations. Brdiczka et al. [9] proposed a method for deciding for each participant in a thread whether or not he or she is an influencer in that particular thread, this approach relies on identification of three types of conversational behavior: persuasion, agreement/disagreement, and dialog patterns. In the same way, Clauset et al. [10] used Markov Chain Monte Carlo sampling to estimate the hierarchical structure in a network. Gupte et al. [2] proposed a measure of hierarchy in a directed online social network, and proposed an algorithm to compute this measure.

The fuzzy logic approach to determining groupings of individuals in written conversation extracts features from conversations and determines through fuzzy logic the likelihood that individuals are in the same group. The approach then considers those who communicate with each other coupled with the previous results to increase the accuracy of the grouping. The fuzzy logic method allows for weight and values to be assigned to features displayed through written speech as well as taking into account who is in a conversation. The approach relies on empirical data that is extracted from the written conversations to determine the grouping of each individual. The counting of features provides a way to move from qualitative space to quantitative space which enables the measurement of the distance between characters and assigning them to different groups. Features are used as input into the fuzzy logic algorithm which groups individuals in conversation based on the empirically used features.

This paper is organized as follows. Section 2 describes the proposed methodology to extract influential members from a diverse group of persons. Additionally, it also describes a method to identify small subgroups of people as well as close relations of the members of small subgroups. In the same way, Section 3 describes how to find the relationship matrix between characters. Section 4 tries to describe the data used in the experiments and the results obtained. Finally, Section 5 shows the conclusions and future works.

II. Feature Extraction

Extracting features in this context amounts to identifying linguistic characteristics of individuals under consideration. This is performed in two steps, using the LightSide tool [12] for identifying the common characteristics of a group using feature vectors, and then reducing the feature vectors to a minimum of independent characteristics. LightSide is an open source text mining and machine learning tool that can extract frequency of word usage and parts of speech features to predict membership in certain groups. LightSide is used in this case to extract frequently displayed features by multiple individuals within the text. "The Wire" text was used and the goal was to classify individuals by their place in the hierarchy, whether they are Gang, Police, or Informant. By informant, we mean that person is connected to both gang and police, but he/she is a gang member or used to be a gang member. In order to do that, we should have some initial knowledge about each group so that we can extend our knowledge with a learning algorithm. So we first, marked four characters in each group that we already know that whether they are police or gang members so we could have some initial knowledge. This is where LightSide comes in, this software is able to extract the features of each labeled character. It goes through all of the characters that are labeled (as police or gang members) and extracts the part of speech (POS) features that we are going to deal with for them. As a definition, each part of speech refers to a category to which a word is assigned in accordance with its syntactic function. In English the main parts of speech are noun, pronoun, adjective, determiner, verb, adverb, preposition, conjunction, and interjection. For instance, the part of speech distribution in the sentence below is as follows:

This (determiner) student (noun) is (auxiliary verb) working (verb) on (preposition) an (determiner) interesting (adjective) project (noun).

The way the features are dealt with is general and they can be extended to more general cases, because the method is not based on some specific words and based on the structure of the text. Now two categories of feature are gathered, the first one is the initial police POS features (F_{1P}) and the second one is the initial gang POS features (F_{1G}). Both groups can have some features in common, but some filters are applied to and reduce the features to the ones that are specific to each group making the features independent. Independence by the two groups not having features in common and the two feature spaces not having features in common. It is shown in equation (1) that F_{1P} is the initial gang POS features, and F_{1G} is the initial gang POS features, after reduction the feature space will change to F_P and F_G. The two feature spaces, F_{1P} and F_{1G}, may have some features in common, but F_P and F_G are fully independent feature vector space and orthogonal to each other.

$$\exists\, f_{1P} \in F_{1P} : f_{1p} \in F_{1G} \quad \text{or} \quad F_{1P} \cap F_{1G} \neq \emptyset \tag{1}$$

$$\exists\, f_{1G} \in F_{1G} : f_{1G} \in F_{1P} \quad \text{or} \quad F_{1P} \cap F_{1G} \neq \emptyset \tag{2}$$

$$\text{If } f_G \in F_G \longrightarrow f_G \notin F_P \quad \text{or} \quad F_P \cap F_G = \emptyset \tag{3}$$

$$\text{If } f_P \in F_P \longrightarrow f_P \notin F_G \quad \text{or} \quad F_P \cap F_G = \emptyset \tag{4}$$

Now, the police POS features and gang POS features are characterized by separate groups. It is the time to go through the other characters that have no information about them. In other words, extend the algorithm throughout the whole text and get some information about the unknown characters. The unknown characters are the ones with undetermined group affiliation. The labels are removed from the characters that

were labeled previously and the text is run through LightSide. LightSide will extract all possible POS feature for each characters. Each character might have thousands of POS features, but among them only the ones common with F_P and F_G are needed. As a result, the feature space is reduced.

At the next step, we assigned two values for each character, one for the number of times they show F_P and the other for the number of times they show F_G. Each person is assigned two values of A and B in which,

A=Number of times character shows F_G

B=Number of times character shows F_P

Equation (5) and (6) show the ratio that each character shows F_G and F_P.

$$a=\frac{A}{A+B} \tag{5}$$

$$b=\frac{B}{A+B} \tag{6}$$

III. Relationship Determination

The objective of our approach is to explore the relationships of individuals in conversation and use this as another factors in finding characters of the same group. The approach used to examine relationships is developed out of the need to determine who is communicating with whom and assign values to those in conversations together. Values are assigned to each individual in a conversation by taking the previous individuals in the conversation as well as those who follow. There are two methods to determine the values which are to assign a true or false value to those in conversation or to simply count the number of times that there is a back and forth in the conversation. A vector was created for each individual in the conversations. The vector consists of N columns for each of the N individuals in the total text. The approached takes the individual currently talking and assigns a 1 to the individuals directly before and after. This would mean that the characters are in direct conversation according to the text. The algorithm also assigns the value 0.5 to person that proceeds and follows the individual by two for indirect contact. Suppose the part of the conversation is in the order of Table I. Consider Person 1, as connected to person 2 directly one time and indirectly one time so a value of 1.5 is assigned for the relation between Person 1 and Person 2. The same can be done for the relation between Person 1 and 3, they are connected three times directly which means they are repeated right after each other three times and they are connected indirectly one time which will assign the value of 0.5 to their connection. The total value of relation between Person 1 and Person 3 would be 3.5. Without seeing the individuals who are in conversation this algorithm yields insight as to those who are related simply by when they speak. The next step is to aggregate the results of the line by line conversation and create a vector for each person that represents the relationships developed throughout the text.

Fuzzy logic is the tool used to deal with the crisp numbers of equation (5) and (6), by that the numbers are made into some interpretable values.

TABLE I. Part of a Conversation

Conversation
Person 1
Person 2
Person 3
Person 1
Person3
Person 1
Person 4
Person 1

In Table II the relation matrix related to Table I is shown, the higher the value in the matrix is considered as a higher weight in the relationship between the two individuals. Relation matrix is named R and is a N×N matrix which is a symmetric matrix.

TABLE II. Relation matrix of Table 1

Convers.	Person 1	Person 2	Person 3	Person 4
Person 1		1+0.5=1.5	0.5+1+1+1= 3.5	1+1=1
Person 2	1+.5=1.5		1	
Person 3	0.5+1+1+1= 3.5	1		.5
Person 4	1+1=2		.5	

IV. Fuzzification

Fuzzy logic is the tool used to deal with the crisp numbers of equation (5) and (6), by that the numbers are made into some interpretable values. The fuzzy logic algorithm used is based on c-mean fuzzy logic clustering algorithm [13]. With three different groups (police, gang, and informant), fuzzy logic approach helps to assign membership values for each person to each group. In this paper we are trying to use the fuzzy C-Means algorithm, but it has a little difference with C-Means method. The difference is that it is not iterative and the center of each cluster (group) is known, so there is no need to iterate. The C-Means algorithm is a method to calculate the degree of membership for each person to each group. The basic and challenging part in fuzzy logic is the rules. This algorithm helps us to reduce the rules in a great deal which leads to a faster process. We have three major rules in our two fuzzy logic boxes that are as follows;

1. If a=1 and b=0 and d=A-B=15 character is gang.
2. If a=0 and b=1 and d=A-B=-15 character is police.
3. If a=0.5 and b=0.5 and d=A-B=0 character is informant.

Different characters have different a, b, and d value, we compare their closeness to each group based on these three values. If their values are closer to each of the above rules, they will have a higher membership value to that group. According to equation (7) algorithm developed allows to calculate the degree of membership of each character to each group (See Figure 1).

$$\mu_{G_i}(x)=\frac{1}{\sum_{j=1}^{3}\frac{\|x-F_i\|^2}{\|x-F_j\|^2}} \quad 1\le i\le 3,\ x\in X \tag{7}$$

G_i is one of the three groups, in this study G_1 is considered for gang, G_2 for police, and G_3 for the informant case. Vector F_i is the center of each group in which the values are assigned based on empirical observation from dataset, knowing that there are three groups implies that there will be three centers of the groups F_1, F_2, and F_3. In other words F_1, F_2, and F_3 show ideal values for a person to be gang, police, and informant respectively. X is the dataset, and x is one member in the dataset which is the normalized feature vector that yields some values for each character in order to compare them together and see how close they are to each other or how far they are from the center of each group (F_1, F_2, F_3). Note that each character has a normalized vector x. As a result, we are able to calculate how close each character is to the ideal gang, police, and informant. There will be three values (μ_{G_1}, μ_{G_2}, μ_{G_3}) for each person that gives information about the membership of each person to each of the groups. Indeed, these three numbers are stored in each row of M_1 for each character. Each row shows the membership of each character in gang, police, and informant group. The same approach can be done for the second fuzzy logic box and the membership values are named $\mu_{G_{i2}}$ (stored in M_2) in which i=1, 2, 3 refers to gang, police, and informant respectively. This algorithm reduces the overlearning process and CPU time since this clustering (grouping) method will reduce our rules drastically.

V. RESULTS

Up to now, the algorithm that is used in this paper has been described, now to implement the algorithm on the dataset. The overall flow diagram of the method is shown in Figure 1. Parts that are inside the dashed line is our black box that processes the input data and gives the output as the degree of membership of being gang, police, and informant for each character. As it was said in previous section, for each character there will be three values for $\mu_{G_{i1}}$ and three values for $\mu_{G_{i2}}$ in which;

$$\mu_{G_{11}}+\mu_{G_{21}}+\mu_{G_{31}}=1 \quad , \quad \mu_{G_{12}}+\mu_{G_{22}}+\mu_{G_{32}}=1 \tag{8}$$

The $\mu_{G_{i1}}$ values for all characters are collected in one matrix and called M_1, also the same for $\mu_{G_{i2}}$ and call it M_2. By considering the total number of characters as N, M_1 and M_2 are shown in equation (9).

$$M_1=\begin{bmatrix} \mu^1_{G_{11}} & \mu^1_{G_{21}} & \mu^1_{G_{31}} \\ \mu^2_{G_{11}} & \mu^2_{G_{21}} & \mu^2_{G_{31}} \\ \vdots & \vdots & \vdots \\ \mu^N_{G_{11}} & \mu^N_{G_{21}} & \mu^N_{G_{31}} \end{bmatrix} \quad M_2=\begin{bmatrix} \mu^1_{G_{12}} & \mu^1_{G_{22}} & \mu^1_{G_{32}} \\ \mu^2_{G_{12}} & \mu^2_{G_{22}} & \mu^2_{G_{32}} \\ \vdots & \vdots & \vdots \\ \mu^N_{G_{12}} & \mu^N_{G_{22}} & \mu^N_{G_{32}} \end{bmatrix} \tag{9}$$

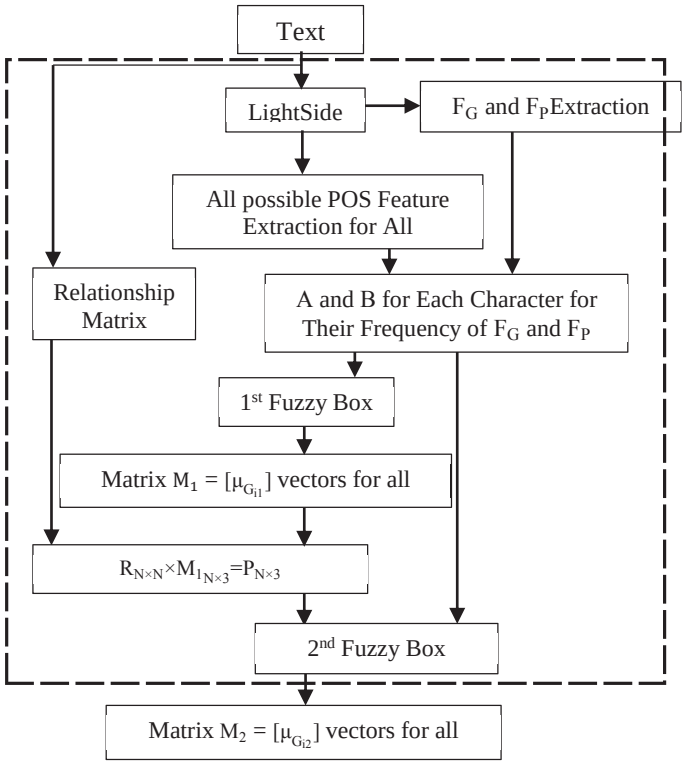


Figure 1. Flow diagram of the method

The characters that are going to be dealt with are the main characters in "The Wire". The output result after applying the first and second fuzzy box to the characters and their text are according to the Table IV, V. As it can be seen in the, the first fuzzy box was not successful to extract the exact rule of some characters. It was after using relation matrix and second fuzzy box that we were able to extract the rule of each character as a gang, police, and informant with high accuracy. Table III compares our output results with the k-mean statistical based method. It can be seen that k-mean statistical based method is able to do so with 85 percent accuracy, but we were able to extract the rule of all characters with high accuracy. The initial selection of the most influential people among a group of persons based on the number of comments made by the participants in a conversation yielded the results shown in the Table VI.

TABLE III. COMPARISON BETWEEN THE RESULT OF STATISTICAL BASED METHOD AND FUZZY LOGIC BASED METHOD

	K-Mean Statistical Based Method		Fuzzy Logic Based Method	
Character	Accuracy	Overall accuracy	Accuracy	Overall accuracy
Avon	✓	85%	✓	100%
Stringer	✓		✓	
McNulty	✓		✓	
Carver	✓		✓	
Freamon	✓		✓	
Greggs	×		✓	
Dee	✓		✓	
Omar	✓		✓	
Bunk	✓		✓	
Daniels	✓		✓	
Russel	✓		✓	
Nick	✓		✓	
Sobotka	×		✓	
Ziggy	✓		✓	

Table VI shows the members who make the highest number of comments in a conversation where 251 persons participates. The Table VI also shows the affiliation of each selected person, and it is marked with red color the persons involved in criminal activities, whereas it is highlighted in green the members with no criminal activity. The accumulated value of all comments made by the members shown in the Table VI accounts for the 90% of the total number of comments in the 10 episodes of the TV show, The Wire. Finally, the methodology based on the number of comments recovers the most important characters of the TV show. The first group shown in the Table VI is constituted mainly by police officers or persons who enforce the law. The only character classified in this group who is not a police officer is Mr. Sobotka. During the TV show, he makes arrangements with European gangsters to smuggle illegal goods through the Baltimore's port. On the other hand, the second group is constituted mostly by persons involved with criminal activities, that is, drug dealers, smugglers, etc. However, a total of 4 characters who are police officers were misclassified in the second group primarily related with criminal activities. In order to identify the hierarchy or the grade of importance of the members that constitute a group of persons, it is analyzed three distinct groups of people present in the TV show, The Wire.

TABLE IV. THE RESULT FIRST FUZZY LOGIC BOX

	Results After First Fuzzy Box			
Characters	μ_{G11}	μ_{G21}	μ_{G31}	Accuracy
Bey	0.14	0.03	0.81	✓
Avon	1	0	0	✓
Stringer	1	0	0	✓
Phelan	0.00	0.99	0.00	✓
McNulty	0	1	0	✓
Pearlman	0.00	0.99	0.00	✓
Carver	0.00	0.99	0.00	✓
Freamon	0.00	0.99	0.00	✓
Greggs	0.00	0.99	0.00	✓
Dee	0.17	0.17	0.65	✓
Omar	0.00	0.76	0.22	×
Bunk	0	1	0	✓
Norris	0.00	0.01	0.98	×
Daniels	0	1	0	✓
Landsman	0.00	0.99	0.00	✓
Prez	0.00	0.99	0.00	✓
Burrel	0.00	0.99	0.00	✓
Russel	0	1	0	✓
Nick	1	0	0	✓
Sobotka	1	0	0	✓
Ziggy	0.99	0.00	0.00	✓

TABLE V. THE RESULT SECOND FUZZY LOGIC BOX

	Results After First Fuzzy Box			
Characters	μ_{G11}	μ_{G21}	μ_{G31}	Accuracy
Bey	0.81	0.07	0.10	✓
Avon	0.98	0.00	0.00	✓
Stringer	0.96	0.01	0.01	✓
Phelan	0.02	0.82	0.15	✓
McNulty	0.01	0.96	0.02	✓
Pearlman	0.00	0.96	0.02	✓
Carver	0.02	0.82	0.15	✓
Freamon	0.01	0.94	0.04	✓
Greggs	0.02	0.83	0.14	✓
Dee	0.00	0.00	0.99	✓
Omar	0.03	0.33	0.63	✓
Bunk	0.00	0.96	0.02	✓
Norris	0.00	0.98	0.01	✓
Daniels	0.00	0.96	0.02	✓
Landsman	0.01	0.94	0.03	✓
Prez	0.01	0.90	0.07	✓
Burrel	0.01	0.90	0.07	✓
Russel	0.00	0.98	0.00	✓
Nick	0.97	0.01	0.01	✓
Sobotka	0.89	0.04	0.06	✓
Ziggy	0.96	0.01	0.01	✓

TABLE VI. IDENTIFICATION OF THE MOST INFLUENTIAL MEMBERS AMONG A GROUP OF PERSONS

Members	Affiliation	# of comments
McNulty	Police Officer	373
BUNK	Police Officer	238
NICK	Involved in Criminals Activities	228
SOBOTKA	Involved in Criminals Activities	191
FREAMON	Police Officer	175
STRINGER	Top Drug Dealer	164
DANIELS	Police Officer	148
ZIGGY	Involved in Criminals Activities	146
AVON	Top Drug Dealer	126
RUSSELL	Port Authority Officer	120
GREGGS	Police Officer	114
DEE	Drug Dealer	92
OMAR	Involved in Criminals Activities	82
CARVER	Police Officer	77
VALCHEK	Police Commander	70
HERC	Police Officer	56
PREZ	Police Officer	53
SPIROS	Involved in Criminals Activities	52
LANDSMAN	Police Officer	50
ELENA	Police Officer	50
BODIE	Drug Dealer	49
LEVY	Attorney	47
PEARLMAN	leading Assistant State's Attorney	44
RAWLS	Police Officer	41
HORSEFACE	Involved in Criminals Activities	40

That is, the police officers, and two organizations related to criminal activities. One of the criminal groups is the Barksdale organization which is led by Avon Barksdale and Stringer Bell. This criminal organization is responsible for multiples crimes and is the most powerful and violent crew in the Baltimore area. The other criminal group is the Sobotka family that is a Polish American Baltimore family. The head of the family is Frank Sobotka, a treasurer for the local union at the Baltimore docks. However, He is also involved along with his family in arrangements with criminals to smuggle illegal goods through the port. Thus, the Sobotka family not only has extensive connections to the Baltimore port, but also links to the criminal underworld. The hierarchies of the main members that constitute the three existing groups are shown in Table VII. The Table VII shows the main members of the distinct organizations present in TV show, The Wire. Additionally, the heads of each organization are highlighted in green color; in the same way, the mid-level and low-level members are colored in yellow and red respectively. On the other hand, the proposed methodology also seeks to automatically detect the hierarchy of the members that constitute a particular group of persons.

To do that, it takes into account the following features: average value of coordination, number of formulated questions, use of modal verbs, number of hedge, use of profanity, and number of terms of address. In order to compare the rank made by the proposed methodology with actual hierarchy of the members that belong to a particular group, it is used the squared difference of the obtained ranking and the actual ranking such as follows.

$$e=\frac{3}{n^3-n}\sum_{k=1}^{n}(R_k-E_k)^2 \quad , \quad n>1 \tag{10}$$

Where, n, is the number of members of the group being analyzed, the parameters R_k and E_k are the actual ranking and the obtained ranking respectively. The parameter e, called ranking error, takes values that range from 0 to 1, where a value of 0 means no error in the ranking, and a value of 1 is the maximum error.

TABLE VII. HIERARCHIES OF THE THREE EXISTING GROUPS

Police Officers	Barksdale organization (Criminals)	Sobotka family (Docks)
1.Daniels (Deputy Commissioner) 2.Freamon (Detective) 3.McNulty (Detective) 4.Bunk (Detective) 5.Greggs (Detective) 6.Carver (Detective) 7.Russell (Port Authority Police Officer)	1.Avon (Kingpin) 2.Stringer (Kingpin) 3.Bey (Soldier) 4.D'angelo (Dealer) 5.Bodie (Dealer) 6.Poot (Dealer)	1.Sobotka (Head of the family) 2.Nick (Sobotka's Nephew) 3.Ziggy (Sobotka's Son)

TABLE VIII. OVERALL OBTAINED RANKING OF THE SEASON 2 OF THE WIRE

Police	Criminals	Docks
Freamon	Avon	Ziggy
Daniels	Stringer	Sobotka
McNulty	Bodie	Nick
Bunk	D'angelo	
Russell	Bey	
Carver	Poot	
Greggs		

Table VIII shows the overall ranking of the members in the three existing groups after averaging the positions of each member.

VI. CONCLUSION

We have explored the possibility of using fuzzy logic in computational linguistics to determine characteristics from text. For the particular case that we studied we found that indeed fuzzy logic can be a powerful method with high accuracy that out performs other methods in clustering and subgroup identification. One important aspect for future work is more extensive testing on different corpuses of data.

REFERENCES

[1] Ainur Yessenalina, Yisong Yue, and Claire Cardie, "Multi-level structured models for document-level sentiment classification," In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2010.

[2] Mangesh Gupte, Pravin Shankar, Jing Li, Muthukrishnan, S.,Liviu Lftode, "Finding hierarchy in directed online social networks," In Proceedings of the 20th International Conference on World Wide Web, pp. 557–566, 2011.

[3] David Elson, Nicholas Dames, and Kathleen McKeown, "Extracting social networks from literary fiction," In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 138-147, 2010.

[4] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li, "User-level sentiment analysis incorporating social networks," In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '11, pp. 1397-1405, 2011.

[5] Jerome Kunegis, Andreas Lommatzsch, and Christian Bauckhage, "Extracting social networks from literary fiction," The slashdot zoo: mining a social network with negative edges, pp. 741-750, New York, NY, USA, 2009.

[6] Amjad Abu-Jbara, Pradeep Dasigi, Mona T. Diab, and Dragomir R. Radev, "Subgroup Detection in Ideological Discussions," In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 1), pp. 399-409, 2012.

[7] Ahmed Hassan, Amjad Abu-Jbara, Dragomir Radev, "Detecting subgroups in online discussions by modeling positive and negative relations among participants," In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 59-70, 2012.

[8] Rutger Joeri Rienks, "Meetings in smart environments," implications of progressing technology. Ph.D, 2007.

[9] Oliver Brdiczka, Patrick Reignier, Jérôme Maisonnasse, "Detecting Influencers in Written Online Conversations," In Proceeding of the Second Workshop on Language in Social Media, 2012.

[10] Aaron Clauset, Cristopher Moore, and Mark Newman, "Structural inference of hierarchies in networks," In International Conference on Machine Learning,", June 2006.

[11] Oliver Brdiczka, Dominique Vaufreydaz, Jérôme Maisonnasse, and Patrick Reignier, "Unsupervised Segmentation of Meeting Configurations and Activities using Speech Activity Detection," 3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI), 2006.

[12] E. Mayfield, D. Adamson, R. Enand (2014, May), *Computational Linguistics*, Available: http://www.lighsidelabs.com.

[13] John Yen and Reza Langari, "Fuzzy Logic: Intelligence, Control, and Information," *Center for Fuzzy Logic, Robotics, and Intelligent Systems,* T. Robbins, 1st ed. New Jersey: Prentice-Hall, 1999, pp. 351-360.

A Machine Learning Approach for Business Intelligence Analysis using Commercial Shipping Transaction Data

Lisa Bramer, Samrat Chatterjee, Aimee Holmes, Sean Robinson, Steven Bradley, and Bobbie-Jo Webb-Robertson

***Abstract*— Business intelligence problems are particularly challenging due to the use of large volume and high velocity data in attempts to model and explain complex underlying phenomena. Incremental machine learning based approaches for summarizing trends and identifying anomalous behavior are often desirable in such conditions to assist domain experts in characterizing their data. The overall goal of this research is to develop a machine learning algorithm that enables predictive analysis on streaming data, detects changes and anomalies in the data, and can evolve based on the dynamic behavior of the data. Commercial shipping transaction data for the U.S. is used to develop and test a Naïve Bayes model that classifies several companies into lines of businesses and demonstrates an ability to predict when the behavior of these companies changes by venturing into other lines of businesses.**

***Keywords*- Incremental machine learning; Naïve Bayes model; Business intelligence; Commercial shipping data**

I. Introduction

MANY "intelligence" problems are particularly challenging because of the complexity of the underlying phenomenon and the lack of consensus on "ground truth" that drives the need to have a team of expert analysts apply their collective knowledge. In some cases, the volume and velocity of data to be analyzed makes the application of machine-based reasoning desirable to assist these domain experts in their analysis, but many new analytic advances are needed to realize such an operational capability.

This study utilizes the Port Import/Export Reporting Service (PIERS) data [1]—a comprehensive database of U.S. international trade—to drive the research for developing advanced intelligence capabilities.

The PIERS data consists of commercially available U.S. import and export shipping transactions, which are typically used for competitive business intelligence. In this paper, this data is utilized specifically to: 1) characterize the lines of business (LOB) to which a particular company belongs based on their procurement activity, and 2) detect possible dynamic changes in LOB as a company's procurement behavior varies. From a business intelligence perspective, it is important to understand when competitors make significant changes to their business operations, especially expansions into new lines of business. While the use of PIERS data is focused on a business intelligence problem, it serves as a proxy to address analytic challenges that may be applicable to other domains.

We begin by discussing key analytic challenges and past work. This is followed by a description of the PIERS dataset and our machine learning based methodology for LOB classification. The results of our algorithms are presented next. We conclude with a discussion of possible extensions of this work.

II. Analytic Challenges and Past Work

Our research approach is driven by commercial shipping transactions for a set of companies over a ten-year period, and produces hypotheses about whether these companies are changing their LOB. While, at first glance, this may seem straight forward, there are analytic challenges that are discussed below; along with a summary of past work using the PIERS data.

A. Dynamic Models

Standard supervised machine learning techniques may be applied to build models that classify a company to a LOB based on features extracted from the PIERS shipping records. However, a company changing its procurement behaviors does not necessarily indicate that it is expanding into a new line of business. If a majority of companies within an LOB happen to adopt similar new procurement behaviors, then one could just as accurately infer that these companies are not expanding into new LOBs, but are simply reacting to a dynamic business environment that is having an impact on the LOB as a whole. Rather than inundating analysts with inaccurate hypotheses, we would want the models to detect this LOB-wide behavior change and evolve accordingly.

The research described in this paper is part of the Analysis in Motion Initiative at Pacific Northwest National Laboratory (PNNL). It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy.

L. Bramer is with the Applied Statistics and Computational Modeling Group, Pacific Northwest National Laboratory, Richland, WA 99352 USA (phone: 509-375-4553; fax: 509-375-2522; e-mail: lisa.bramer @pnnl.gov).

S. Chatterjee is with the Applied Statistics and Computational Modeling Group, Pacific Northwest National Laboratory, Richland, WA 99352 USA (e-mail: samrat.chatterjee@pnnl.gov).

A. Holmes is with the Applied Statistics and Computational Modeling Group, Pacific Northwest National Laboratory, Richland, WA 99352 USA (e-mail: aimee.holmes@pnnl.gov).

S. Robinson is with the Human Centered Analytics Group, Pacific Northwest National Laboratory, Richland, WA 99352 USA (e-mail: aimee.holmes@pnnl.gov).

S. Bradley is with the Cyber Innovation & Operations Center, Pacific Northwest National Laboratory, Richland, WA 99352 USA (e-mail: steven.bradley@pnnl.gov).

B-J. Webb-Robertson is with the Applied Statistics and Computational Modeling Group, Pacific Northwest National Laboratory, Richland, WA 99352 USA (e-mail: bobbie-jo.webb-robertson@pnnl.gov)

B. Hypothesis Rationale

Models that generate inductive as well as deductive hypotheses could be useful for domain experts. For example, it may be helpful for a user to be alerted that a company is suddenly behaving in ways that are no longer consistent with its previously classified LOB, it appears to be also useful for expert analysts to know *why* the models have reached that conclusion. For example, a classification model may compute over the last 90 days that the likelihood has dropped from 98% to 90% that Ford Motor Company is an automobile manufacturer; however, this doesn't provide the analyst with the insight required to assess whether the models took into account observations that she missed or whether she believes that the models are flawed, which is critical for model steering.

C. Machine Learning with Streaming Data

Desirable features of machine learning models from streaming data involve: 1) accounting for *recent history* when making predictions, and 2) allowing the models to *evolve* or *update* with the data streams. Conditioning predictions based on history, with moving training windows, is an approach that addresses the first case above. For the second case, a machine-learning algorithm that incrementally learns over the data and updates the model with new training instances appears to be appropriate. Giraud-Carrier [2] describes incremental learning as applied to tasks and algorithms. An incremental learning task involves the availability of training examples over time; and an incremental learning algorithm, also referred to as a *memoryless online* algorithm, produces hypotheses that depend on past hypothesis and the current training example.

D. Past Work using PIERS Data

Limited applications were found in the open source literature that involved the use of PIERS records for data mining. Jeske et al. [3] describe a platform for generating synthetic data for testing data mining tools. They implemented a resampling data generation algorithm using the PIERS data.

Das and Schneider [4] describe an anomaly detection problem and discuss the use of unsupervised methods applied to categorical datasets, including: association rule; likelihood; and bayesian network based approaches. The authors implemented a likelihood-based approach using the PIERS data to detect unusual shipments among all imports into the country. The focus was on detecting unusual combinations of attribute values in the data.

III. DATASETS

Our study analyzed PIERS import data records [1], from January 2005 to December 2014. The PIERS database contains records for every company importing or exporting goods in the U.S. For this study, we selected a subset of these companies, in particular 17 companies that could be categorized within one of three lines of businesses. These companies were selected because they had a large number of records available and had a well-defined LOB. Future analysis will incorporate other lines of business and companies. PIERS data is rich with shipment related information and at times is noisy with possibly inconsistent data entries. Access to the PIERS data records was made possible due to the establishment of a strategic goods testbed (or data library) at PNNL [5]. The PNNL testbed team has created a centralized data location and with a single agreement allows access to the PIERS data for research purposes. The lines of businesses include: 1) Automotive, 2) Clothing, and 3) Appliance. The Automotive companies chosen were BMW, Ford, Honda, Hyundai, Nissan, Toyota, and Volkswagen. Clothing companies were Guess, Gymboree, Hennes & Mauritz, J Crew, Levi, and Ralph Lauren. Finally, the appliance companies considered were Bosch, Electrolux, General Electric, and LG Electronics. The 10-year shipping record counts associated with these companies ranged from 108,828 for LG Electronics to 7,572 for Gymboree.

In addition to the companies mentioned above, we also merged records for several pairs of companies belonging to different lines of businesses (where over time, the record counts from the starting LOB company incrementally decreases and the other LOB company increases). The motivation behind this merge was to test whether our classification algorithms can detect changing LOB over time. Several hybrid companies were formed with several different rates of change. For illustration purposes, we examine one such hybrid, which started with records from Ford and slowly injected records from Old Navy into the data over time.

Categorical	Quantitative
Date	Weight (lb, kg, etc.)
Shipper	Measure (cubic ft, etc.)
Shipper Address	Quantity (bags, pkgs, etc.)
Consignee	Estimated Value
Consignee Address	
Carrier	
Country of Origin	
Port of Arrival	
Port of Departure	
U.S. Destination	
HS Code	
Short Commodity Description	

Table 1. Examples of PIERS record attributes by variable class.

Every record in the dataset has as many as 54 different attributes. These attributes contain information about the shipper, shipment, and arrival/departure locations. Table 1 presents a selection of these attributes separated by variable class: quantitative or categorical. A challenge working with this dataset was the identification of attributes that characterize and classify companies into a LOB and can help detect deviations with dynamic changes in procurement behavior. One such challenge is that limited quantitative

variables are available, and the variables that are available are recorded with many different units of measurement. Additionally, in some records, no units of measurement are recorded. Many categorical variables are available including but not limited to: the final destination of the shipment, the departure port, the harmonized system (HS) code for tariff purposes, and a commodity short description.

IV. Methodological Approach

Our modeling methodology is comprised of five steps: 1) identification of key data attributes, 2) creation of a data-driven library of attribute values, 3) selection of a machine learning model, 4) training and testing strategy, and 5) model evolution plan. A description of each methodological step follows.

A. Selected Data Attributes

The choice of data attributes was driven by their potential to characterize a company within a LOB. We explored the evolution of several attributes over time for various companies, and our attribute set for further analysis included: 1) *Commodity Description*, 2) *U.S. Destination*, and 3) *Port of Departure*. All three selected attributes contain text information. Commodity description contains blocks of text associated with the shipment and/or company. Since we only consider import data, U.S. Destination is listed a city within the U.S. where the shipment is headed, and Port of Departure is a foreign port where the shipment began its journey. Figure 1 presents an example frequency plot of words that are contained in the commodity description field of Hennes & Mauritz, over a subset of time. In this example, *Ladies* is the most frequently occurring word.

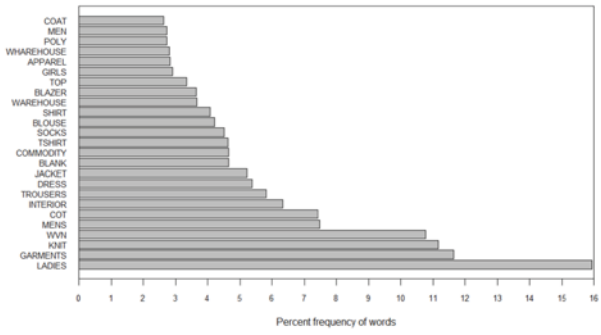


Fig. 1. Example frequency plot of words contained within commodity description attribute for Hennes & Mauritz.

Similarly, example frequency plots of shipment counts by U.S. Destination over time were prepared to assess variability of shipment location characteristics (see Figure 2 for data associated with Hennes & Mauritz). Each plot in Figure 2 corresponds to a different U.S. Destination city and each bar in a given plot corresponds to shipment counts for a chosen time block. In this example, the location with the most frequent spikes/bars (i.e. count of arriving shipments) is *New York City*.

B. Library of Attribute Values

Attribute values (or text strings) were first split to create a list of unique keywords, final U.S. destination cities, and departure ports for each company within the three lines of businesses. The percent occurrence frequencies of these unique attribute values were then computed for different blocks of time; and an overall mean percent frequency was evaluated. Table 2 presents an example of percent frequencies of keywords for Hennes & Mauritz. Similar frequency tables were created for the cities and departure ports. A minimum mean occurrence threshold level of 5% was chosen to select unique keywords, and a threshold of 2% was chosen for selecting cities and departure ports; leading to the creation of the attribute value library.

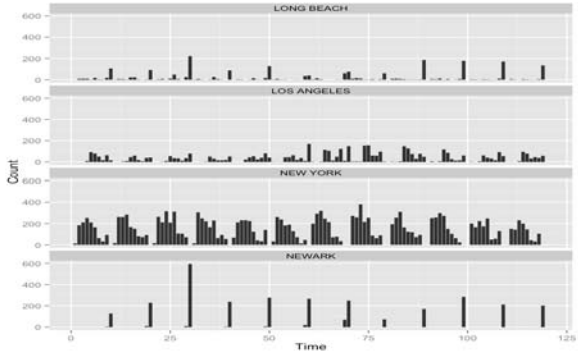


Fig. 2. Example frequency plot of shipment counts by U.S. destination for Hennes & Mauritz. Each plot corresponds to a different U.S. Destination city and each bar in a given plot corresponds to shipment counts for a chosen time block.

Three attribute value libraries were developed thereafter; one each for the keywords, cities, and departure ports covering information from all companies across all lines of businesses. Each library contains a list of primary attribute values along with their spelling and parts of speech variations found within the data records. For example, the keyword library list item *Auto* along with *Automobile*, *Automotive*, and *Autos*. These data libraries were key inputs for training the machine learning algorithms.

Sample Attribute Value Percent Frequencies

Keyword	Time Block 1	Time Block 2	…	Overall Mean
Cot	0.185	0.185	…	0.207
Ladies	0.140	0.145	…	0.162
Knit	0.115	0.150	…	0.149
⋮	⋮	⋮	⋮	⋮

Table 2. Attribute frequency as a proportion of records.

C. Machine Learning Algorithms

A large number of features could possibly be extracted from the three selected attributes within the PIERS data.

Moreover, dependencies may also exist among these features. As a starting approach, a Naïve Bayes classification technique [6] was adopted for the LOB classification problem. The Naïve Bayes approach is based on Bayes theorem and assumes that conditional probabilities of independent variables are statistically independent. Three independent Naïve Bayes models, one for each LOB, were fit to training data from companies from all three LOB's. Thus, the conditional probabilities of a company being in each LOB do not necessarily have to sum to one.

The nodes or explanatory variables in the Naïve Bayes model were the proportion of records in a given timeframe that contained each of the items listed in the data libraries for keywords, cities, and departure ports. As a result, we had 163 total nodes (83 keyword types, 26 cities, and 54 departure ports). A probabilistic expression for the Naïve Bayes algorithm LOB classifier can be expressed as:

$$P(B_i|W_1, W_2, \dots, C_1, C_2, \dots\ D_1, D_2, \dots) \propto \prod_{j=1}^{83} P(W_j|B_i) \cdot \prod_{k=1}^{26} P(C_k|B_i) \cdot \prod_{q=1}^{54} P(D_q|B_i) \cdot P(B_i) \qquad (1)$$

where *B* refers to a LOB, *W*, *C*, and *D* refer to the proportion of records that contained a keyword K, a destination city *C*, and a departure port *D*, respectively. $P(B)$ is the prior probability of a LOB, and $P(x|y)$ is the conditional probability of event *x* given event *y* is observed.

D. Training and Testing

The first 5,000 records of each of the 17 companies were used as training data for each of the Naïve Bayes models. The explanatory variables were calculated for windows of training records of 150 records. Training cases were computed for moving windows of 150 records with a step size of 50 records. Each rolling window was evaluated on the attribute values of interest (for keywords, U.S. Destination, and Port of Departure) and a proportion of occurrence was calculated. The step between different windows was of size 50 records, and this rolling window process was repeated over all of the training records. Additionally, each training set summary record was assigned a response variable of one or zero (for each Naïve Bayes model: Appliance, Automotive, and Clothing) indicating the company's true LOB during the training period.

A separate Naïve Bayes model was fit from the training data for each LOB: clothing, automotive, and appliances, resulting in a total of three models. Fitting models for each LOB independently allows for the possibility of an individual company behaving in a manner similar to several LOBs and does not force predicted probabilities to sum to one. Predictions of a company's LOB can be generated for any reasonable moving block size at a true streaming level (i.e. a new predicted probability of each LOB can be generated with each new incoming record). However, for the purpose of demonstration here, the testing data that was then evaluated on these models was again created by a similar method to that described above for the training data (window width = 150, sliding windows), except that the step between different windows was of size 15 records. The testing data for each company was comprised of the remaining records for each company (after the first 5,000 records were removed for training purposes) over a ten year period as described previously.

V. Results and Discussion

We proceed by evaluating the predictive capability of the Naïve Bayes models with the 17 companies previously discussed. We then investigate the models' capability to detect changes in company behavior, by examining model performance for the aforementioned hybrid company.

The accuracy of each model for each company was assessed for the testing data. Figure 3 summarizes the accuracy of each model by company. Most clothing companies had a near perfect accuracy across all three models. The accuracy of the clothing LOB model is very accurate (greater than 95% accuracy) for all companies. However, the accuracy of the auto and appliance LOB models performed less accurately in the case of a few companies. For example, the auto LOB model incorrectly identified Bosch as an automobile company in more than half of the testing data cases. This behavior is not entirely unexpected as both the automobile and appliance industries involve electronics and other similar products. Upon further inspection, Bosch contained many records with keywords that were also seen in the automobile companies (e.g. parts, motor, etc.).

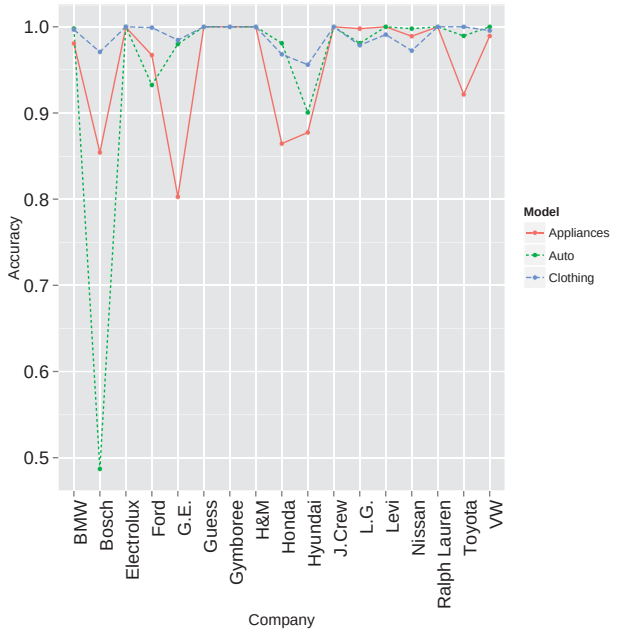


Fig. 3. Naïve Bayes model accuracy by company for three LOB models.

The overall model performance was assessed taking into account all companies. Because the number of records, and thus testing data points, varied from one company to

another, we consider the first 250 summarized window testing data points. Table 3 summarizes the accuracy, false positive rate, and false negative rate for each of the LOB models. Overall, the three models are able to discriminate between different LOB's. Additionally, Figure 4 gives a receiver operating characteristic curve (ROC) for the clothing LOB model.

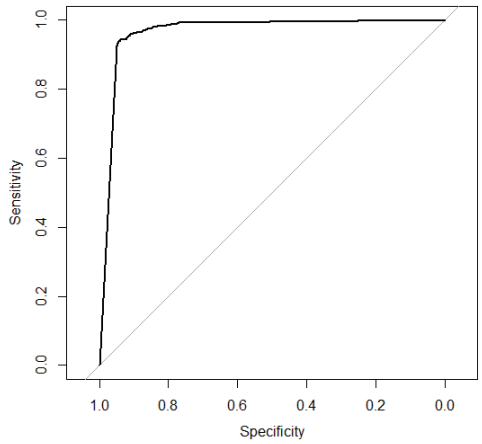


Fig. 4. ROC curve for clothing LOB model over all companies.

We further evaluate the models' ability to identify changes in company procurement/LOB behavior by generating predictions for a hybrid data set that transitions from a purely automotive company to adding a partial LOB in the clothing industry.

Model	Accuracy	FPR	FNR
Auto	0.9877	0.0204	0.0009
Clothing	0.9971	0.0045	0.0001
Appliances	0.9789	0.0100	0.0571

Table 3. Overall performance metrics for each LOB model.

Figure 5 shows the predicted probability of the hybrid company belonging to each LOB. In the beginning periods of the testing data when the testing data is comprised of just automobile records, the models classify the pure records correctly. Additionally it can be seen that the models pick up on the injection of clothing records into the testing data. However, the predicted probabilities tend to switch between the two models in a dichotomous manner. This behavior is due to some highly discriminate explanatory variables (e.g. keywords of auto or seatbelt). When these words appear in the dataset in any proportion, the records get classified as being from the automobile LOB. This dichotomous behavior continues, because the model is never updated to reflect changes in company procurement habits and entry into a new LOB.

VI. Conclusions and Future Work

We have demonstrated that Naïve Bayes classification models using keywords, destination cities, and ports of departures are able to effectively classify a businesses LOB, based on past procurement behavior. Additionally, these models are able to detect changes in a company's procurement behavior. However, the ability to model a company going into a second LOB and accurately model the company still participating in the original LOB was unsuccessful with only dichotomous training examples.

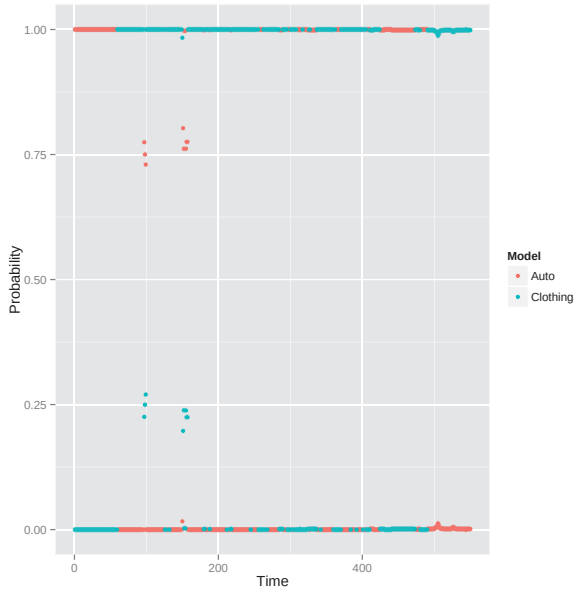


Fig. 5. Probabilities of LOB's for records in the hybrid auto and clothing company.

A class of algorithms that may naturally support predictive analysis on this streaming data may be found in the vicinity of incremental machine learning. Traditional machine learning approaches assume that a good training set is always available a priori and contains all the required knowledge to construct sufficient models that may applied to new examples or problems, which is not the case when changes in data dynamics are present. A wide variety of incremental learning algorithms have been developed in machine learning areas such as Bayesian networks [7-9], neural networks [6-7], support vector machines [10-12], and decision trees [13]. These methods should be adapted to automatically generate or retrain the incremental models to automatically evolve as drifts in company behavior and procurement features emerge in the data streams. Additionally, metrics for model evolution and the evolution of model features should be developed to help in eliciting domain expert feedback.

References

[1] JOC Group. (2015, March). PIERS Data. Available: https://www.piers.com/.

[2] C. Giraud-Carrier, "A note on the utility of incremental learning," *AI Communications*, vol. 13, pp. 215–223, 2000.

[3] D.R. Jeske, P.J. Lin, C. Rendon, R. Xiao, and B. Samadi, "Synthetic data generation capabilities for testing data mining tools," in *Proc. MILCOM'06 – 2006 IEEE Conference on Military Communications*, Washington, DC, October 23–25, 2006, pp. 3449–3454.

[4] K. Das and J. Schneider, "Detecting anomalous records in categorical datasets," in *Proc. KDD'07 – The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, CA, August 12–15, 2007, pp. 220–229.

[5] J.B. Webster, L.E. Erikson, C. Toomey, and V.A. Lewis, "PNNL strategic goods testbed: a data library for illicit nuclear trafficking," Pacific Northwest National Laboratory, Richland, WA Technical Rep. PNNL-SA-102611, 2014.

[6] StatSoft. (2015, March). Naïve Bayes Classifier. Available: http://www.statsoft.com/textbook/naive-bayes-classifier.

[7] Daly, R., Shen, Q., and Aitken, S. (2011). Learning Bayesian Networks: Approaches and Issues. *The Knowledge Engineering Review*, 26(2), pp. 99-157.

[8] Samet, S., Miri, A., and Granger, E. (2013). Incremental Learning of Privacy-Preserving Bayesian Networks. *Applied Soft Computing*, 13(2013), pp. 3657-3667.

[9] Cauwenberghs, G. and Poggio, T. (2000). Incremental and Decremental Support Vector Machine Learning. In *Proc. of NIPS*, pp. 409-415.

[10] Diehl, C.P. and Cauwenberghs, G. (2003). SVM Incremental Learning, Adaptation and Optimization. In *Proceedings of the 2003 International Joint Conference on Neural Networks*, pp. 2685-2690.

[11] Ralaivola, L. and d'Alche-Buc, F. (2001). Incremental Support Vector Machine Learning: A Local Approach. In *Proceedings of ICANN*, pp. 322-329.

[12] Chao, S. and Wong, F. (2009). An Incremental Decision Tree Learning Methodology Regarding Attributes in Medical Data Mining. In *Proc. of the 8th International Conference on Machine Learning and Cybernetics*, pp. 1694-1699.

SESSION

DATA MINING: LATE BREAKING PAPERS

Chair(s)

TBA

Supervised Machine Learning Approach for Gender Disambiguation from User Generated Unstructured (Text) Documents

Amit Choudhary, Praveen Kumar, and Sridhar Jeyaraman,
Analytics, Annik Technology Services Pvt. Ltd., Gurgaon, Haryana, India

Abstract *- The rise of social media has led to an explosive growth in the size of data generated, data growth has undergone a renaissance. The pervasive espousal of social media into our daily lives has opened many opportunities for researchers to deep dive into human behavior. Many aspects of human behaviors have been explored using media data, example, detecting and monitoring mood state, forecasting sentiment analysis etc. Another important aspect of human behavior where a significant interest lies is identification of author identity. Predicting author characteristics, preferences and opinions helps answer many social science questions and support many commercial applications especially in e-commerce business.*

Identifying gender using author names and profile names by Twitter and Google are some examples of many advances in this area.

Our work in this research takes it to the fore with ability to even classify anonymous users or authors. It is engrossed towards disambiguating author gender through lexical choice, choice of syntactic structure, capitalizing on linguistic nuances and textual meaning.

The results of our research are quite promising and endure witness to the validity of approach.

1. INTRODUCTION

Electronically available texts are sometime casual and sometime topic driven. Sociolinguistic variation is found among gender in articulation of their thoughts through words. The language-gender relationship has long been point of discussion within Sociolinguistics, Linguistic Anthropology, Linguistic Psychology and related disciplines. More so, since men and women have different language styles with different pattern of communication, lexical pattern and paralinguistic cues. Therefore, even in e-text there is latent difference in trace & trend of writing and word selection.

Most of the players in the social media space are predicting gender using either the name of the authors or user account name/ followers. Twitter says, "*We're able to understand gender by taking public signals users offer on Twitter, such as user profile names or the accounts she or he follows*". However, Google determines user's gender when a user provide his name to a Google account or to a Google partner. With these approaches classifying gender is easy to discover using a service like Dots Name Validation however in many scenarios this information is not available or is in a tangled state. In this paper, we have explored statistical and machine learning techniques that can help us in identifying the gender of authors based on text flowing around the social media and blogs. Our main premise in this paper is based on that men and women use different lexicon. Statistical methods can substantially reduce noise and predict the speaker's gender guided by the signals derived from the words they use.

There are diverse applications to this methodology across computational linguistics, such as identifying the target potential user groups, targeting gender specific campaigns, building customers for remarketing and diversity of preferences. For the ecommerce business, this will help in segmenting their customers by product, categories, brands and other demographic composition. This can also help in pursuing changes to gender specific styles in lexicon over a period of time.

To address these concerns, we used machine learning techniques and algorithms for gender identifications after limiting the gender bias in concepts while extracting the data. The goal of this paper is to answer *four* specific questions, they are as follows:

a) *Does any gender specific patterns exists in the documents*

b) *Can these patterns be captured from the word corpus and converted to predictors*

c) *Are these patterns able to statistically predict gender*

d) *Which supervised learning approaches are able to provide optimal results*

The novelty of our approach lies in leveraging sociolinguistics, gender specific characteristics on social platforms, stylometric analysis techniques and vocabulary indicators to disambiguate user gender on social media. *Our approach is independent of deterministic information revealed by the user.* Even if the deterministic indicators are not available which is more so the case on social platforms our framework is able to automatically infer gender from user generated content on social websites. Our experiment results have confirmed that personal traits can be predicted with higher level of accuracy by analyzing linguistic variation.

Textual meaning includes savor of affect, genre and personality these characteristics are broadly captured by writing style and classify the 'how' of text

The key steps in our framework includes:

a) *Data extraction and pre-processing*
b) *Language processing – tokenization, lemmatization and part of speech*
c) *Lexicon formation*
d) *Supervised learning and modelling*

The further details of each of these steps are outlined below in sections 3 to 7.3

2. Related work

There exists a quantum of research in the area of gender prediction using textual data from social media and it is established that there are significant linguistic differences between women and men. Majority of the previous studies have their links to social roles and explored the forms of these linguistic distinctions. Numerous linguistic features and patterns have been determined, primarily, functional words, writing syntax, character usage, frequency of words. Robin Lakoff (1973) introduces the research to the field of sociolinguistic based on the idea about women's language, noted mainly to social justice and power in conjunction to gender. As per Lakoff, women's talk has some key characteristics:

1. Use of Hedges, e.g. "it seems to be, "sort of"
2. Tag questions and rising intonation, e.g. "You don't mind reading this, do you?"
3. Use of intensive, e.g. "so"
4. Very polite, e.g. "It would be much appreciated…"
5. Empty adjectives, e.g. "divine", "lovely", "cute"
6. Set of words specific to their interests, e.g. "dart", "shirr".

Modern time researcher's view point is that, the main drawback of Lakoff's studies is lack of any empirical foundation. Instead of collecting the corpora of male and female speech, Lakoff made claims based on observations around her own social circle and intuitions.

Numerous studies from previous to most recent research have similar conclusions regarding which of these linguistic patterns that best differentiate female and male authors. Largely based on these researches that women tend to use more emotionally charged language as well as more adverbs and adjectives, and they apologize more often than men.

3. Data Extraction And Pre-Processing

One primary challenge in data extraction was selection of gender neutral topics and getting an un-biased mix for both genders. For that purpose, we extracted data from various blogs on politics, education, travel and tourism, science & technology, entertainment, sports. Post extraction data was integrated and pre-processed merged all the data sets on different topics in randomized order. At the same time kept note that we have balanced representation of all the topics as well as both the genders. We extracted data in form of blogs, sub-blogs and respective gender of the bloggers. To simplify things, we appended all the blogs and sub-blogs and randomized it again. Now, we had data with sample size of around 23K (Male – 48% and Female – 52%) with balanced representation of both genders and all the relevant topics.

As part of pre-processing we eliminated spam posts, eliminated duplicate posts to avoid any attempts to obfuscate lexicon words.

4. Language Processing

Basic intent behind this step was laying the foundation for statistical modelling. Therefore, we parsed our data into concepts (i.e. relevant words) and frequency of concepts for each case data /document in our data set. We used whitespace to tokenize and extract all possible concepts from the text data. For simplicity, our implementation was focused on single-word terms. However, for group of words we took reference of our dictionary and tokenized those group of words as one concept (e.g. as soon as, thank you, instead of, etc.) Testing point in this step was tokenization of misspelled words, gibberish characters, and multiple lines in same cell of our data. With this process we come across 16000 plus concepts & for each we had to have 1 column. Thus, we had 16000 plus variables in totality to create this model. These many numbers of variables were itself a challenge. So as a next step, we went for categorization of concepts (i.e. grouping words into different categories). At the one end it was helpful in reducing the concepts but at the same time it was the most important step towards capturing the linguistic patterns and variations for statistical modeling. As an inherent step for statistical modelling we required an extensive dictionary which should have all the possible concepts and categories for grouping those concepts.

5. Dictionary grounding and maturing

Since, one of the most important concept behind disambiguating gender is difference in stylometry and use of lexicon by different genders. Therefore, this step required special attention and holistic approach such that, we could capture maximum possible linguistic variations. In this step we took reference of several renowned word categorization dictionaries like ROGET, POS, etc. (as represented in Table-1). During implementation of this step we came across a well-known scenario, where some of the concepts were falling into more than one category. In lieu of capturing the optimal stylometric pattern without overlay of information we took out the duplicity. After going through many approaches we finally choose conventional approach (i.e. putting those words into most appropriate categories with simple English grammar) to resolve this problem of duplicity. At the end of this step we

accomplished our dictionary with grouping all the possible words/concepts into 101 categories.

There are many state of the art techniques for text taxonomy like Rule Based, Decision Trees, SVM and Variable Clustering. To statistically cross-validate our corpus we tried several of the above given techniques. Results from all the tested techniques were more or less in line with our corpus, however variable clustering results were the most closest to our dictionary leading to 101 categories.

It was not possible to mature our dictionary in the first go. Thus, we repeated above mentioned steps several times to mature our dictionary to include most of the possible words and make it vast & enriched.

POS	ROGERT	Stylometry	Sentiments	Other word type
Adj.All, Adj.Pert, Adj.Ppl, Adv.All	Existence, Relation, Quantity, Order, Time, Change, Causation	Number of Sentence, Number of Words	Negative Positive	Tone, Mood and Emoticons
Noun	Space in General, Dimensions, Form, Motion	Mean Sentence	Uncertainty Litigious	Causal, Certainty, Cognitive, Exclusionary, Factual, Function Words, Hedge Words, I Words
Verb	Matter in General, Inorganic Matter, Organic Matter	Mean Words	Constraining Superfluous Interesting	Taboo
Preposition	Formation of Ideas, Communication of Ideas	Word 1 To Word >= 10		Personal , Family
Number	General, Personal, Sympathetic, Moral and Religious Affections			Physical Function

Table 1: Key features used in lexicon formation

6. DATA PREPARATION

Data preparation was an integral part for this work. We delineated two approaches towards this important step to consider both dimensions i.e. accuracy and computational overhead. In the first instance, aggregated sum of frequency of all the concepts falling into same categories was taken into consideration. Advantage of this approach was that it was taking the holistic pattern and variability from data.

Secondly, we formulated flag variables as the predictors for all the concepts falling into the categories. The primary intent to this method was to check if we can make it less computational intensive without losing too much accuracy.

7. MODELING

In this segment, we describe the complete statistical modeling framework that learn patterns and features at different depth. Based on the enriched dictionary that we prepared, we classified them in categories for processing them for our analysis.

The figure below illustrates the framework for gender disambiguation.

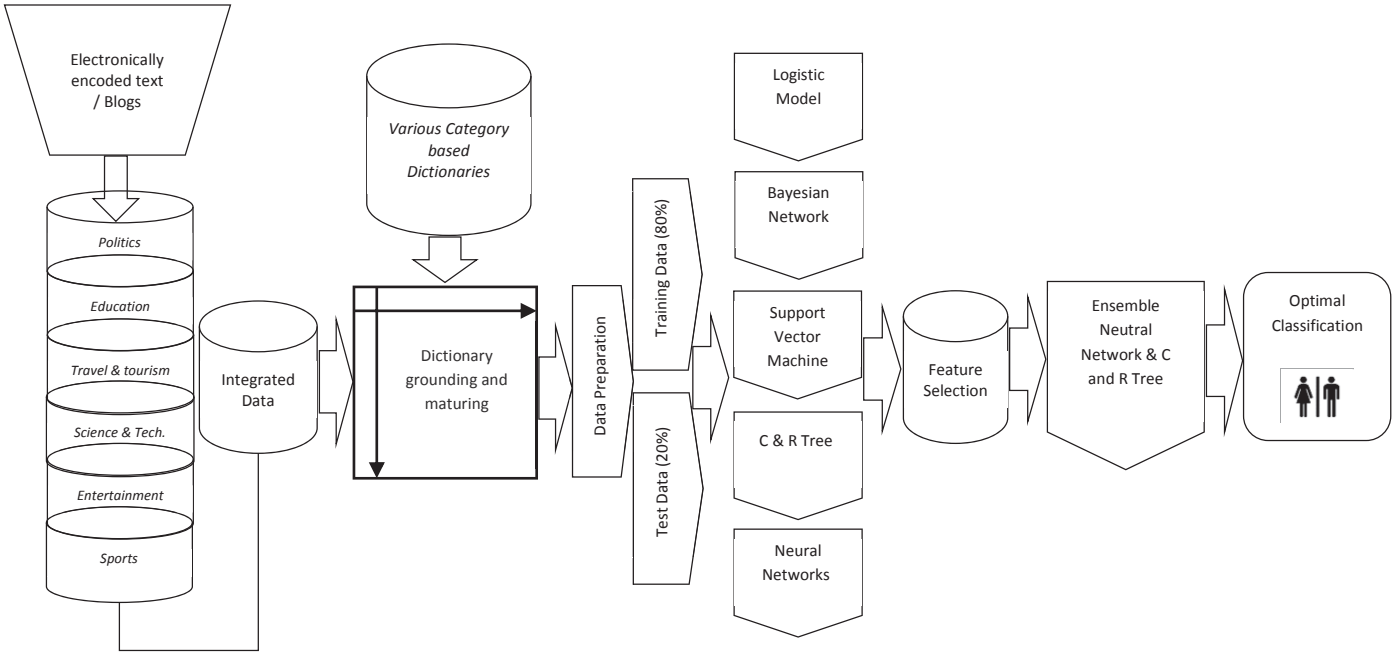


Figure 1: Framework for gender disambiguation

7.1 FEATURE EXTRACTION AND SELECTION

In order to identify the most important categories that have the ability to classify the gender we used feature selection. This also helped us in speeding up the modeling and removing the categories that has little or no predictive importance. We have applied different techniques to address this problem, ranging from classifying the categories based on significant to the target at different confidence interval levels (95% and 90%) to using random forest for calculating the variable importance and also the dimension reduction techniques like principal component analysis. By reducing the dimensionality of the predictors correlated information is eliminated at the cost of loss of accuracy.

So there is always a trade-off between accuracy and the computational overheads of retaining all the predictors without application of feature selection. Thus, broadly we had two feature selection objectives: (a) finding significant variables highly related to the target to magnify all the important variables; and (b) to find a small number of predictors sufficient for good prediction of the target in order to have a parsimonious set of important variables. We appropriated the best subset of predictors from both the techniques (i.e. random forest & predictors significant to target) which had the higher discriminating power on gender. The tables below illustrates the variable importance and ranked predictor significance for top predictors from both the techniques.

Top 11 Predictors	*Mean Decrease Accuracy*	*Mean Decrease Gini*
S_IWORDS	39.92019002	54.90933723
S_ARTICLE	21.90926807	38.11648143
S_FAMILY_REFERENCE	18.07971937	15.71815225
S_NOUNSWORD	16.76391246	33.02705219
S_COMMUNICATION_OF_IDEAS	15.68705501	21.85298504
S_NOUNFOOD	14.64962223	9.46747482
S_CAUSATION	12.36346868	25.34362742
S_PREPOSITION	12.22683828	34.29444727
S_DIMENSIONS	12.16362851	27.67310641
S_ADVALL	11.71435787	27.82333671
S_NOUNGROUP	11.69029254	11.70854028

Table 2: Variable Importance from RandomForest Algorithm

Rank	*Top 10 Predictors*	*Value*
1	S_NOUNGROUP	0.999985537
2	S_IWORDS	0.99989695
3	S_FAMILY_REFERENCE	0.996534627
4	S_ADJPERT	0.996268832
5	S_SOUNDADJECTIVES	0.993602358
6	S_NEGATIVE	0.992823307
7	S_NOUNATTRIBUTE	0.992798283
8	S_VERBCOMPETITION	0.991902935
9	S_POSITIVE_TONE	0.989867443
10	S_FORM	0.98637848

Table 3: Ranked Predictor Significance against target

7.2 STATISTICAL TECHNIQUES

In order to select the most efficient classification method based on accuracy and stability, the experiments were performed with different statistical models using the matching combination of features and patterns to ascertain the most optimal algorithm. We applied varied classification techniques ranging from Logistic Regression, Bayesian Network Classifier, Support Vector Machine, Classification & Regression Tree and Neural Networks with the features set described earlier. The individual model results established that Neural Network algorithm was most accurate in determining the gender. To gauge the accuracy level and reduce the error, we ensemble the best performing individual models. The results were quite interesting, we were able to increase the overall ensemble model performance better than the individual models. Gender classification was performed on the extracted web blog dataset with balanced topics. In the overall dataset, 80% of the documents are used for training and the remaining were used for testing.

		Training dataset			**Testing dataset**		
		Percent Correct	**AUC**	**Gini**	**Percent Correct**	**AUC**	**Gini**
Overall	Logistic Regression	65.5%	0.721	0.442	60.5%	0.632	0.661
	Bayesian Network Classifier	63.0%	0.691	0.381	59.2%	0.622	0.653
	Support Vector Machine	63.6%	0.678	0.356	59.3%	0.626	0.661
	C & R Tree	62.6%	0.682	0.363	58.5%	0.643	0.686
	Neural Networks	**72.5%**	**0.634**	**0.268**	**69.3%**	**0.747**	**0.805**
Male	Logistic Regression	64.2%	0.688	0.376	54.7%	0.572	0.373
	Bayesian Network Classifier	62.3%	0.719	0.437	57.0%	0.588	0.348

		Training dataset			Testing dataset		
		Percent Correct	**AUC**	**Gini**	**Percent Correct**	**AUC**	**Gini**
	Support Vector Machine	62.7%	0.691	0.382	57.3%	0.604	0.401
	C & R Tree	61.6%	0.698	0.395	56.6%	0.623	0.302
	Neural Networks	**71.3%**	**0.650**	**0.299**	**67.4%**	**0.711**	**0.276**
Female	Logistic Regression	66.8%	0.724	0.448	66.3%	0.677	0.338
	Bayesian Network Classifier	63.7%	0.695	0.389	61.3%	0.633	0.416
	Support Vector Machine	64.5%	0.665	0.329	61.2%	0.653	0.435
	C & R Tree	63.5%	0.665	0.330	60.3%	0.647	0.375
	Neural Networks	**73.7%**	**0.618**	**0.236**	**71.1%**	**0.757**	**0.306**

Table 4: Overall and Gender wise accuracy (best-fit model in bold) with 95% CI

The results of the different statistical techniques were applied on the sum of frequency dataset. The predictive power for overall data and across the gender was in sync. The model results on the training and testing datasets were profoundly stable with (±) 5% deviation in accuracy.

As expected at individual level, the best performing algorithm was found as *neural network* (NN) with the deviation of (±) 3% in train and test datasets. Rational behind our expectation of NN being the best performer was that, since this kind of data has large interaction and considering it NN would be the best statistical technique for it. In Neural Network algorithm we used *multilayer perceptron* and *radial basis function* separately with different hidden layers ranging from 2 to 16, the model was more or less stabilized at 8 hidden layers. At the same time Multilayer perceptron was performing better than radial basis function in terms of accuracy. For several methods, we used bagging technique to control over-fitting. At the individual model stage, we have not used any feature extraction or selection methodology.

To enumerate the strength of relationship between gender and the category corpus in our blog data, we initially trained a logistic regression classifier. Our approach was somewhat different from the traditional approach, we considered gender as the dependent variable, and the independent variables are the 101 grouped categories developed around our dictionary. For evaluation of our model we relied mostly on percentage correct (i.e. accuracy percentage), AUC (i.e. area under the curve) and Gini coefficient (i.e. measure of statistical dispersion between observed and predicted values). Although Table 4 above is self-explanatory, however, we would like to highlight on some of the key results. Looking at the comparison of models NN was providing the highest prediction accuracy percentage of 72.5% in training dataset and 69.3% test dataset.

It was indicating stability in model but there was greater scope of work on increasing the accuracy. Thus, we went ahead to ensemble different models.

We tried several combination of algorithms to ensemble and achieve the optimal accuracy and stability in and out of sample. The best ensemble algorithm was achieved by the combination of NN and Classification & regression tree (C&R). Combining multiple classifiers generally performs better than the single classifier, however the challenge remains in the construction progression (parameter tuning, predictor diversity and the technique of combining individual models). Boosting and bagging were used for engendering the individual networks, followed by bootstrapped aggregation to construct ensembles. Since the base classifiers were unstable, bagging worked well for us even reducing the variance of the individual classifier. With the use of NN algorithm in the ensembles models we were able to approximate the complex non-linear mappings. Also since it does not make priori assumptions on the distribution of the data and minimizing the interactions between the factors, NN algorithm was more stable on our training sets in the ensemble models.

We have combined the individual predictions of NN by using majority voting, as that worked well across the bootstrapped sample. The C &R tree algorithm we examined the best split among the classifiers by reducing the impurity index, further the tree was prune based on the cost complexity algorithm which adjusts the risk estimate based on the number of terminal nodes in our model. For our ensemble models we have used the feature selection as discussed in the above section.

		Training			Testing		
		Accuracy	**AUC**	**Gini**	**Accuracy**	**AUC**	**Gini**
Overall	Ensemble (Neural Network and C & R Tree)	87.3%	0.948	0.895	85.97%	0.936	0.872
Male		85.8%	0.932	0.880	84.45%	0.919	0.857
Female		88.3%	0.959	0.905	87.06%	0.948	0.883

Table 5: Overall and Gender wise accuracy with 95% CI

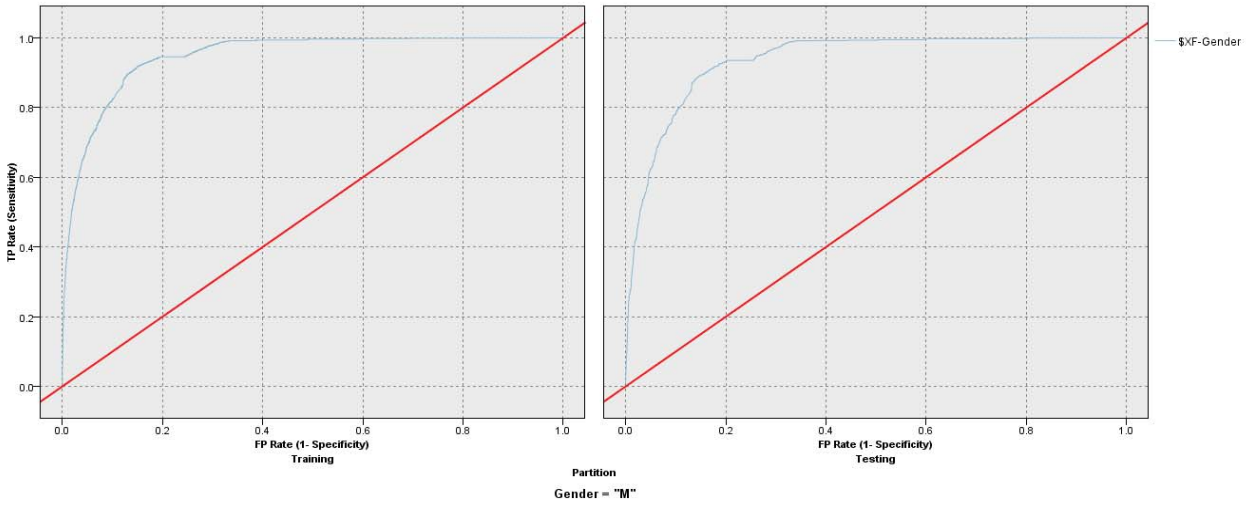


Figure 2: ROC Curve

7.3 EVALUATION OF MODELS

Assessment of the ensemble model performance was done by a method that was independent of any threshold, the area under the receiver operating characteristics (ROC) curve. This is mostly recommended for comparing two class classifiers, since it does not merely summarize performance at a single arbitrarily selected decision threshold, but across all the possible threshold. Furthermore, ROC curves are unvarying under changing distribution of the binary classes (like Gender).

Lastly, in order to have a more reliable estimate of the ensemble models, we used five–fold cross validation. The original data was randomly divided into five mutually exclusive subsets. One subset of the data was removed and remaining cases was used to build the ensemble models. This model was then applied to the removed section and its empirical error was calculated. This process is repeated on the other four subsets and the mean empirical error was considered as the final estimate of the model. We achieved the same level of accuracy in five-fold cross validation as the one achieved in the test dataset.

8. CONCLUSIONS AND FUTURE WORK

As described earlier, the drive of this work was to capture stylometric difference between Male/Female and developing an optimal algorithm to predict the gender for same. We believe, with the level of accuracy we have got in this venture, it would be reasonable to conclude that we have succeeded in our efforts. At the same time we would like to honestly accept that we are still finding some scope of improvement in it, which would be pivotal for the industry & our clients.

We hope that our efforts will be valuable since our dictionary enrichment process is an ongoing process. This pattern mining from e-text and various other sources of text will be statistically viable and would be expedient in business use cases. We have tried to simplify the process of Gender prediction, so that it is easy to cognize both at theoretical and deployment level.

Future work would include expanding the categories and features to include other style markers as well. We also plan to combine three approaches:

1. Our approach through sociolinguistics
2. Discovering gender through names
3. Any cues given in the content like 'today is my wife's birthday' etc.

To reap the benefits out of all and plug-in any gaps of individual approaches.

9. CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the authors at:

Amit Choudhary, Amit.Choudhary@anniksystems.com
Praveen Kumar, Praveen.Kumar@anniksystems.com
Sridhar Jeyaraman, Sridhar.Jeyaraman@anniksystems.com

10. REFERENCES

[1] Robin T. Lakoff. 1973. Language and woman's place. In Language in Society, Vol. 2, No. 1, pages 45 – 80.

[2] Eckert, P. (1997). Gender and sociolinguistic variation, in J.Coates ed., Readings in Language and Gender, Blackwell, Oxford 1997, pages. 64-75.

[3] Juola, Patrick (2006). "Authorship Attribution" (PDF). Foundations and Trends in Information Retrieval 1: 3. doi:10.1561/1500000005

[4] Janet Holmes and Miriam Meyerhoff (2003). "The handbook of language and gender", pages 161-177

[5] Meyerhoff, M. (2006) Introducing Sociolinguistics. London and New York: Routledge

[6] Carsten Peterson and Thorsteinn Rögnvaldsson, An introduction to Artificial Neural Networks, pages 123 – 139

[7] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, Zheng Chen, Demographic Prediction Based on User's Browsing Behavior

[8] Cristian Bucila, Rich Caruana, Alexandru Niculescu-Mizil, Model Compression

[9] Mittal C. Patel, Prof. Mahesh Panchal, Ensemble of Diverse Artificial Neural Networks

[10] Deborah Cameron, Fiona McAlinden & Kathy O'Leary, Lakoff in context: the social and linguistic functions of tag questions

[10] Deborah Cameron, Fiona McAlinden & Kathy O'Leary, Lakoff in context: the social and linguistic functions of tag questions

Scalable Mining of Frequent and Significant Sequential Patterns

Zs.T. Kardkovács[1], and G. Kovács[1]
[1]U1 Research, Budapest, Hungary

Abstract— *In this paper, we argue on and we prove constructively that considering significance and support in a hand to find sequential patterns can be solved more efficiently than the original problem in terms of using less time and computer resources, and even finding new interesting patterns of low support items. Our approach helps to define patterns and anti-patterns as well, which provide us to do a search for frequent patterns and null hypothesis testing simultaneously. Experiments in health care databases show that our method not only reduces the search space for new candidates, but it scales much better than well-known algorithms for large databases.*

Keywords: sequence mining, interestingness, frequent sequential patterns, tf-idf, health care database

1. Introduction

Finding patterns in sequential data sets has become important in many applications such as fraud detection, personalized health care, and recommendation systems. In these applications, there are known observations/items/events/evaluations (items in short), and the task is to find a statically relevant connection between prior and current items, assuming that the correlation between these items are static, i.e. the variance of correlations do not change in time. A sequence of items called pattern is relevant if it is frequent enough and unexpected in the sense that measured frequencies are notably different from null hypothesis.

Data mining solutions suggest finding frequent patterns first based on absolute and/or relative frequencies called support as a sole criterion. Unexpectedness/significance/interestingness (significance in short) is being addressed separately, later in the process. This approach implicitly states that support overweights significance hence the latter is taken into account if and only if a pattern is frequent enough. Figuratively speaking, well-known pattern mining techniques reveals frequent patterns about e.g. beers, cheeses, and dumpers, but they eventually miss the most profitable high-end market, e.g. old whiskeys, caviars in general because customers purchase expensive products far less often.

The two properties strongly correlate:

- if the minimum threshold for support is set too high then significant patterns might be omitted,
- if it is set too low then most of the found patterns are trivial, particular cases of others, or they are not significant.

There is no known golden rule for setting the minimum support threshold properly. Can significance help setting minimum support threshold?

Definition 1: A data mining pattern P is said to be relatively significant regarding an item I if $I \in P$, P is previously unknown, and there exists a null model $\mathcal{Z}$ for which $\Pr(P|I)$ is statistically relevant. P is said to be absolutely significant if $\Pr(P)$ is statistically relevant regarding a null model $\mathcal{Z}$.

It is easy to prove that absolute significance can be expressed by relative ones without loss of generality. The only difference between these notions is that relative significance defines a "fix point" from which relevance can be measured. Note that significance is a vague notion in general: how to calculate $\Pr(P)$ is not defined, it can be adapted for problem specific needs, see e.g. [1], [2] for details. Later in this paper, we use the term significance for relative significance where it is not confusing.

In this paper, we argue on that relative significance based algorithms can handle significance and frequency in a hand. We propose a novel sequence mining algorithm called REVIEW (RElevance from a fix point of VIEW) which finds all relevant patterns in linear time. In addition, we show that REVIEW also finds the most likely anti-patterns as well, i.e. those items which tend to mutually exclude each other. This property is beyond the capabilities of state-of-the-art sequence mining algorithms.

The paper is organized as follows. Section 2 gives the elementary definitions of sequence mining: items, itemsets, sequences, sequence databases, and gives illustrative examples for the definitions. The algorithm we propose for finding frequent and significant patterns is defined in Section 3. In Section 4, we review the most important sequence mining algorithms and show an example of their scalability issues. Empirical comparison is given in Section 5, our algorithm is tested on a real-life health care database against PrefixSpan and SPADE, the two fastest algorithms in the literature.

2. Preliminaries

This section gives preliminary definitions necessary for the formalization of the problem statement. Throughout this paper, we use the following conventions:

- sets and elements of sets are denoted by capital letters and lower case letters, respectively,
- itemsets and items are taken from the beginning of the latin alphabet,
- $|X|$ denotes the size of X where X is a set of attributes or itemsets,

Patient #1			Patient #2			Patient #3			Patient #4		
I	T	A	I	T	A	I	T	A	I	T	A
1	234	a	2	57	f	3	186	h	4	33	a
1	234	b	2	63	g	3	186	i	4	93	k
1	234	c	2	74	g	3	186	a			
1	234	d	2	78	e	3	186	j			
1	237	e	2	78	g	3	186	e			
						3	199	a			
						3	199	e			

Table 1: A small anonymized piece of the database

- we use R, S symbols for database relations defined as subsets of a Cartesian product, and $r, s \ldots$ for tuples, records or elements of a relation. If $r \in R$ and $R \subseteq A \times B$ then $r(a, b)$ is a short form to say $a \in A$, $b \in B$, and $r[A] = a$, $r[B] = b$ where $r[X]$ stands for the attribute values of r on attribute set X (projection in relational database theory),
- identifiers are denoted by letters near I,
- $\top$ and $\bot$ stand for logical values `true` and `false`, respectively,
- T, t are used for time related sets and variables, respectively,
- we also introduce the symbol $\mathcal{D}_R(X)$, which denotes the domain of an attribute set X in a relation R, i.e. $\mathcal{D}_R(X) = \{r[X] | r \in R\}$.

Let the input database be defined as follows, the definition is analogous to the one in [3].

Definition 2: Let $A = \{a_1, a_2, \ldots, a_n\}$ be a finite itemset, where $n \geq 1$, T is a non-empty set of timestamps, and I is a non-empty set of unique identifiers. Let a relation R be defined over $I \times T \times A$, i.e. $R \subseteq I \times T \times A$, then R is a sequence database. For simplicity and better understanding, we use the notion $R(ITA)$ to express R is determined by sets I, T, and A, i.e. $R \subseteq I \times T \times A$ in that order. If $|I| = 1$ in a sequence database $R(ITA)$, then R is called a sequence database.

Example 3: Table 1 shows a small set of records from the anonymized health care database we use in this paper. The columns of the table reflect relation R. The twelve columns of the table are organized into four groups of three column. Each column group represents a patient. The first column in a group represents anonymized patient identifiers. The second column contains timestamps replaced by integer numbers because of anonymity reasons. The third column in a group contains the items: treatment codes are replaced by letters.

Let a binary total ordering $\leq$: $R \times R \rightarrow \{\top, \bot\}$ be defined on sequence databases such that the ordering of elements of R is determined by the natural ordering over T. For simplicity, we also use $\leq$ on relations such that if $S_1, S_2 \subseteq R$ and $S_1, S_2 \neq \emptyset$, then $S_1 \leq S_2$ iff $\forall s_1 \in S_1 \forall s_2 \in S_2 : s_1 \leq s_2$. In other words, the ordering of items in sequence databases is based on time related attributes. If time related attributes do not differ, then we assume they are simultaneous events.

Definition 4: Let $R(ITA)$ be a sequence database, and $\mathcal{S} =< S_1, S_2, \ldots S_n >$ be defined as an ordered set of relations where $\forall i : 1 \leq i \leq n \implies S_i \subseteq R$ such that

$$S_i, S_j \in \mathcal{S} : 1 \leq i < j \leq n \implies S_i \leq S_j, \neg S_j \leq S_i.$$

We say $\mathcal{S}$ is a sequence if and only if

$$\forall r, s : \; r \in S_i, s \in S_j \implies r[I] = s[I]$$
$$\forall r, s : \; (r \in S_i, s \in S_j \implies r[T] = s[T]) \iff i = j$$

for all $S_i, S_j \in \mathcal{S}$. Since identifiers are the same in the sequence, and there are sets of items that share the same timestamps, we use the representation $\mathcal{S} =< A_{t_k}, \ldots A_{t_l} >$ for better readability, where $t_i \in T$ and $A_{t_i} \subseteq A$ is a set of elements indexed by their shared timestamps. We also introduce the following notions:

- $|\mathcal{S}| = n$ denotes the length (number of relations) of the sequence,
- $U(\mathcal{S})$ stands for the shared identifier in $\mathcal{S}$,
- and $\tau(S_i)$ (or $\tau(A_i)$) for the shared timestamp in S_i where $1 \leq i \leq n$.

We also define the following terms in order to introduce different kinds of supports.

Definition 5: Let $\mathcal{S}_1 =< A_{t_k}, \ldots A_{t_l} >$ and $\mathcal{S}_2 =< A_{t_m}, \ldots A_{t_n} >$ be two sequences defined on $R(ITA)$. We say

- $\mathcal{S}_1$ is a proper subsequence of $\mathcal{S}_2$ denoted by $\mathcal{S}_1 \sqsubseteq \mathcal{S}_2$ if and only if $U(\mathcal{S}_1) = U(\mathcal{S}_2)$ and $\forall A_{t1} \exists A_{t2} : A_{t1} \in \mathcal{S}_1, A_{t2} \in \mathcal{S}_2 \implies A_{t1} \subseteq A_{t2}, \tau(A_{t1}) = \tau(A_{t2})$,
- $\mathcal{S}_1$ is a subsequence of $\mathcal{S}_2$ denoted by $\mathcal{S}_1 \preceq \mathcal{S}_2$ if and only if for all a_1, a_2, and $t1, t2 \in T$ there exist $t3, t4 \in T$ such that $a_1 \in A_{t1}, a_2 \in A_{t2}, t1 \leq t2 \implies a_1 \in A_{t3}, a_2 \in A_{t4}, t3 \leq t4$ where $A_{t1}, A_{t2} \in \mathcal{S}_1$, and $A_{t3}, A_{t4} \in \mathcal{S}_2$.
- the union of sequences for which $U(\mathcal{S}_1) = U(\mathcal{S}_2)$ denoted by $\mathcal{S}_1 \cup \mathcal{S}_2$ is defined as an $\leq$-ordering preserving merge of these sets such that if $A_{t1} \in \mathcal{S}_1$, $A_{t2} \in \mathcal{S}_2$ and $\tau(A_{t1}) = \tau(A_{t2})$ then the resulting $A_t = A_{t1} \cup A_{t2}$,
- $\mathcal{S}_2$ is the prefix cut of $\mathcal{S}_1$ by an item $a \in A$ if and only if $\mathcal{S}_2 \sqsubseteq \mathcal{S}_1$ and if there exists $a \in A_t$, $A_t \in \mathcal{S}_1$ then $\max_{A_i \in \mathcal{S}_2}(\tau(A_i)) \leq \tau(A_t)$. In this paper, we denote by $\Phi(\mathcal{S}_1, a)$ the union of all possible prefix cuts of $\mathcal{S}_1$ by $a \in A$.

Definition 6: Let a sequence $\mathcal{S} =< S_1, S_2, \ldots S_n >$ be defined on a sequence database $R(ITA)$. We say $\mathcal{S}$ is a closed sequence and it is denoted by $\overline{\mathcal{S}}$ if and only if

$$\forall r \exists S_i \; r \in R, S_i \in \mathcal{S}, r[I] = U(\mathcal{S}) \implies r \in S_i$$

for some $1 \leq i \leq n$. The largest possible set of closed sequences in $R(ITA)$ is denoted by Σ.

Example 7: Table 2 shows the records of Table 1 transformed into the form used for representing sequence databases in the literature. The first column is the sequence

identifier which comes from the patient identifying I attribute of Table 1. The second column contains the sequences, where each sequence is a comma separated list of sets of items shown in braces. The ordering of the sets of items is determined by attribute T. If the T value is identical for two A items, then those appear in the same itemset. Sequence

I	Σ
1	$<(a,b,c,d),(e)>$
2	$<(f),(g),(g),(e,g)>$
3	$<(h,i,a),(j,e)>$
4	$<(a),(k)>$

Table 2: Relation R transformed to a sequence database

$<(a),(e)>$ is a subsequence of both the sequence identified by $I = 1$ and $I = 3$. In both sequences itemset (a) is a subset of the first itemset, and (e) is a subset of the second itemset.

In our example, a specific $S_i \in \Sigma$ represents a patient's anamnesis. If a patient was diagnosed with two or more symptoms/diseases/conditions then they are considered to have occured simultaneously. $A_i \leq A_j$ in a sequence (patient anamnesis) means all diagnoses of A_{t_i} precede any diagnoses in A_{t_j}. There can be many different sequences with the same identifier according to Definition 4, and there is no sequence with the length of 0. It is easy to prove that the maximal number of closed sequences in a sequence database $R(ITA)$ equals to $\mathcal{D}_R(I)$ hence every closed sequence has a natural identifier: the elements of set I in our database.

Definition 8: We introduce the following support metrics for a sequence $\mathcal{S} =< S_1, S_2, \dots S_n >$ defined on $R(ITA)$:

- the support of $\mathcal{S}$ denoted by $\text{supp} : \mathcal{S} \to [0,1]$:
$$\text{supp}(\mathcal{S}) = \frac{||\{\mathcal{S}_i \mid \mathcal{S}_i \in \Sigma, \mathcal{S} \preceq \mathcal{S}_i\}||}{||\Sigma||},$$
where $||S||$ stands for the number of elements in set S,
- the conditional support of $\mathcal{S}$ assuming there is a $\mathcal{S}_i \in \Sigma$, which has an item $a \in A$ is
$$\text{supp}(\mathcal{S}|a) = \frac{||\{\mathcal{S}_i \mid \mathcal{S}_i \in \Sigma, \mathcal{S} \preceq \Phi(\mathcal{S}_i, a)\}||}{||\{\mathcal{S}_i \mid \mathcal{S}_i \in \Sigma, \exists A_t \in \mathcal{S}_i, a \in A_t\}||}.$$
If there is no sequence in Σ that contains a, then let $\text{supp}(\mathcal{S}|a) = 0$,
- the conditional unsupport of $\mathcal{S}$ assuming there is a $\mathcal{S}_i \in \Sigma$ that contains no item $a \in A$ is
$$\text{supp}(\mathcal{S}|\neg a) = \frac{||\{\mathcal{S}_i | \mathcal{S} \preceq \mathcal{S}_i \in \Sigma, \neg(\mathcal{S} \preceq \Phi(\mathcal{S}_i, a))||}{||\{\mathcal{S}_i \mid \mathcal{S}_i \in \Sigma, \neg(\mathcal{S} \preceq \Phi(\mathcal{S}_i, a))\}||}.$$
If all sequences in Σ contain a, then let $\text{supp}(\mathcal{S}|\neg a) = 0$.

Example 9: Table 3 shows the support of sequences with length of one based on Table 2. Items a and e occur in three different sequences. Though g has three occurrences as well, those are limited to a single sequence.

Theorem 10: Let $\mathcal{S}_1 =< A_{t_1}, \dots, A_{t_n} >$, and $\mathcal{S}_2 =< A_{t_1}, \dots, A_{t_{n-1}} >$ be two sequences defined on $R(ITA)$, then $\forall a \in A_{t_n}$: $\text{supp}(\mathcal{S}_1) \leq \text{supp}(\mathcal{S}_2|a)$.

	$\mathcal{S}$					
	a	b	c	d	e	g
$\text{supp}(\mathcal{S})$	0.75	0.25	0.25	0.25	0.75	0.25
$\text{supp}(\mathcal{S}\|e)$	0.66	0.33	0.33	0.33	0	0.33
$\text{supp}(\mathcal{S}\|\neg e)$	0.66	0	0	0	0	0.33

Table 3: Support of some items in Table 2 database

We omit the proof here.

3. Problem Statement

The original problem was defined in [4] as given a database of sequential items (transactions), the problem of mining sequential patterns is to find the maximal sequences among all sequences that have a certain user-specified minimum support (min_support), i.e. to find maximum length frequent sequences (FS). In [4], [5], the well-known Apriori algorithm and its first clones were introduced based on the hypothesis that all subsequences of a frequent sequence are frequent sequences themselves, formally if $\mu \leq \text{supp}(\mathcal{S})$ then $\forall \mathcal{S}' \preceq \mathcal{S} \implies \mu \leq \text{supp}(\mathcal{S}')$ where $\mu \in [0,1]$ stands for min_support.

Apriori-like algorithms are quasi linear whenever the size of frequent sets of items are small; polynomial in the sum of the size of the input (transactions) plus output (frequent patterns)[6]. According to [7], however, the overall operation time is close to the size of the input multiplied by the length of sequences. In other words, Apriori-like sequence mining algorithms are exponential for long sequences or large inputs.

If a standalone item is frequent independently from others then it will also appear in every frequent enough patterns as well. In addition, low support and significant sequences cannot be discovered in FS problem space which begs the question: can we determine frequent and significant sequential patterns in a hand?

Example 11: Consider the sequence database presented in Table 2 where we are analyzing a disease represented by j. Data indicate that there are three potential preconditions: a, h, and i that could lead to j. Nevertheless, apriori frequencies of items a and e are relatively high (see Table 3), so diagnoses frequent in the general population such as flu or hypertension are not necessarily relevant. Such items should not be considered when building frequent sequences and provide a basis for prefiltering.

Let us reformulate the FS problem in the following way: the problem of mining sequential patterns is to find the maximal *significant* sequences among all sequences that have a certain user-specified minimum support and minimum significance. Each such sequence represents a sequential pattern, or frequent-and-significant sequence (FASS). In this section, we outline a new algorithm which solves the FASS problem, and we prove that it can be done more effeciently than other well-known solutions discussed in Section 4.

The idea behind prefiltering in our support calculation method comes from text mining where the TF-IDF [8] metric has been successfully used to connect different documents based on their contents. The sequence metric is similar to the SIF-IDF metric defined in [9] for protecting sensitive data in databases. Importance metric of a pattern is derived from two values associated with two sequences generated by the pattern. For a pattern we maintain two sets of key-value pairs: one in which support values are calculated on closed sequences of customers who favor the pattern, and another one set of those who does not. The former one is called *frequent set*, the latter is called *inverse set*. The normalized rate of relative occurrences of the frequent and inverse sets is a suitable parameter the item to be used to grow the sequence.

3.1 Definitions

Definition 12: Let $\mathcal{S}$ be a pattern over a relation $R(ITA)$, and F be a function which maps sequences to a set of itemset-number pairs such that

$$F(\mathcal{S}) = \{(a, \text{supp}(\mathcal{S}|a))|\ a \in A\}.$$

Definition 13: Let $\mathcal{S}$ be a pattern over a relation $R(ITA)$, and $\overline{F}$ be a function which maps sequences to a set of an itemset-number pairs such that

$$\overline{F}(\mathcal{S}) = \{(a, \text{supp}(\mathcal{S}|\neg a))\ |\ a \in A\}.$$

$F(\mathcal{S})$ and $\overline{F}(\mathcal{S})$ contain information on each item, e.g. $(a, n) \in F(\mathcal{S})$, and $(a, m) \in \overline{F}(\mathcal{S})$. Note that a correlation between values n, m might indicate relevance. If $m \simeq 0$ and $n \geq \mu$ then $\mathcal{S} \rightarrow a$ (i.e. $\mathcal{S}$ is followed by a) show high correlation which means $\mathcal{S}$ as a series of events highly suggest a to be happening. If $m \geq \mu$ and $n \simeq 0$ then $\mathcal{S}$ and a show high inverse correlation, i.e. the pattern of $\mathcal{S}$ almost always inhibits the event a to be happen.

Let $F(\mathcal{S})[a] = n$ be a shorthand for the fact that $(a, n) \in F(\mathcal{S})$. Let $E(F(\mathcal{S}))$, $\text{Var}(F(\mathcal{S}))$, and $\text{Sum}(F(\mathcal{S}))$ be the mean, devation, and the sum of $F(\mathcal{S})[a]$ values, respectively, for all $a \in A$.

Definition 14: Let Imp be defined as a measure on a sequence $\mathcal{S}$, and $a \in A$ item of a relation $R(ITA)$ such that $\text{Imp}(\mathcal{S}, a) = 0$ iff $\text{Sum}(F(\mathcal{S}))\ \text{Sum}(\overline{F}(\mathcal{S})) = 0$, and

$$\text{Imp}(\mathcal{S}, a) = \frac{|\frac{F(\mathcal{S})[a]}{\text{Sum}(F(\mathcal{S}))} - \frac{\overline{F}(\mathcal{S})[a]}{\text{Sum}(\overline{F}(\mathcal{S}))}|}{\max(\frac{F(\mathcal{S})[a]}{\text{Sum}(F(\mathcal{S}))}; \frac{\overline{F}(\mathcal{S})[a]}{\text{Sum}(\overline{F}(\mathcal{S}))})}$$

otherwise, where $|n|$ stands for the absolute value of a number n. We say $\mathcal{S}$ is a relevant antecedent of a if $\text{Imp}(\mathcal{S}, a)$ is greater or equal to a certain threshold.

Relevance measures both frequencies and infrequences, which leads to a re-formulation on how important an item a regarding a preliminary series of items $\mathcal{S}$. That is, it is a potential relative significance measure (see Definition 1).

Definition 15: Let Ind be defined as a measure on $\mathcal{S}$ sequences, and $a \in A$ items of a relation $R(ITA)$ such that

$$\text{Ind}(\mathcal{S}, a) = \frac{\text{Imp}(\mathcal{S}, a) - E_{a \in A}(\text{Imp}(\mathcal{S}, a))}{\text{Var}_{a \in A}(\text{Imp}(\mathcal{S}, a))}$$

We say $\mathcal{S}$ is an import antecedent of a if $|\text{Ind}(\mathcal{S}, a)|$ is greater or equal to a certain threshold.

Importance is a normalized value of relevance, and it measures how much $\mathcal{S} \rightarrow a$ is unusual. In most the cases, mean value of relavances are being around 0, i.e. common illnesses are independent from others in general. Statistically, if absolute value of the $\text{Ind}(\mathcal{S}, a) \geq 3$ (the triple of the variance) then it is an outlier value which is usually a strong indication for a deep connection between variables.

3.2 REVIEW Algorithm

We propose Algorithm 1 for identifying important sequences in the sequence database R. The inputs of the algorithm are the database R itself, a μ minimum support threshold, and a ν relevance threshold. The output is the set of frequent-and-relevant (important) sequences Σ_f. In the body of the algorithm, a loop variable k, the frequent set $F(\mathcal{S})$, the inverse set $\overline{F}(\mathcal{S})$, and the set of new sequences Σ_c are used locally.

The algorithm works as follows. In the initialization phase (line 1), we add items as sequences of length 1 to the Σ_f set, if their support is over the minimum threshold μ. The main loop iterates over the sequences of maximum length. First, it removes all elements from the Σ_f new important sequence set (line 8). It computes the frequent $F(\mathcal{S})$ and the inverse $\overline{F}(\mathcal{S})$ sets for the current $\mathcal{S}$ sequence (line 10). If there are candidate postfix items, then we iterate over it, and filter the sequences with the formula of Definition 15. We utilize an significance threshold ν input parameter (line 11). If the significance is over that threshold, then that item c is appended to the end of $\mathcal{S}$ (line 12), and the new sequence is added to the important set of sequences (line 13). The main loop is repeated until the generated set Σ_c is not an empty set. If no further candidate sequences can be generated, the algorithm returns the Σ_f set (line 17), otherwise the elements if Σ_c are added to Σ_f (line 18). If all sequences are processed, the maximum length loop variable k is increased (line 20), and the main loop is restarted.

Lemma 16: Algorithm 1 identifies all frequent patterns which have a support greater or equal to a min_support according to Definition 8, but only those of important ones are returned.

According to Theorem 10, conditional support $\text{supp}(\mathcal{S}|a)$ is greater or equal to the $\text{supp}(\mathcal{S} \rightarrow a)$. It means that by generating $F(\mathcal{S})$ Algorithm 1 finds all candidates for which support is greater or equal to a certain threshold. However, if either a or $\mathcal{S}$ is independent, or too frequent in general, it entails $\overline{F}(\mathcal{S}, a) \simeq F(\mathcal{S}, a)$ and as such is omitted from the

input : $R(ITA)$ database, μ minimum support threshold, ν significance threshold
output: Σ_f set of important sequences
data : k cycle variable, Σ_c set of new sequences, $F(\mathcal{S})$ frequent next item set, $\overline{F}(\mathcal{S})$

```
1  /* Initialization                                   */
2  k := 1;
3  Σ_f = {S | a ∈ A, S =< {a}_{-∞} >, supp(S) ≥ μ };
4  /* Main loop                                        */
5  while true :
6  do
7      foreach S ∈ Σ_f where len(S) = k do
8          Σ_c := ∅;
9          foreach a ∈ A do
10             Compute the sets F(S, a) and F̄(S, a);
11             if |Ind(R, S, c)| ≥ ν then
12                 S' := concat(S, c);
13                 Σ_c := Σ_c ∪ {S'};
14             end
15         end
16         if Σ_c = ∅ then return Σ_f;
17         ;
18         Σ_f := Σ_f ∪ Σ_c;
19     end
20     k := k + 1;
21 end
```

Algorithm 1: Importance based frequent sequential pattern generation

output. As a consequence, Algorithm 1 is a one-step method to find frequent *and* relevant patterns.

Lemma 17: Algorithm 1 can identify important sequences with regard to the significance threshold ν that are not frequent regarding a threshold μ.

Definition 15 is independent from the minimum support threshold μ, so it is possible to construct an example, where this statement holds. If $\Sigma = \{< \{a\},\{b\},\{c\} >, < \{c\},\{d\} > \}$, and $\mu = 60\%$, then subsequence $< \{a\},\{b\} >$ can not be frequent as it occurs only in the first closed sequence. However, it is important because $|Ind(< \{a\}, b >)| \geq \nu$ for an appropriate ν because b is always preceeded by a.

Algorithm 1 builds important sequences on the pattern of GSP. The number of database scans is twice the number of important sequences identified: the computation of sets $F(\mathcal{S})$ and $\overline{F}(\mathcal{S})$ requires a scan each. With regard to candidate genereation and data structure efficiency, there is a lot of room for improvement.

Lemma 18: All subsequences of important sequences generated by Algorithm 1 are important sequences.

Lemma 18 gives a property similar to the one existing in the case of sequential pattern generation algorithms, and this way patterns can not only be grown, but joined as well.

4. Related works

In this section, we give an overview of the most important frequent sequential pattern mining algorithms: GSP, PrefixSpan, SPADE and SPAM. In the literature several other algorithms exist as well, however, those can be considered as the variants, extensions of the ones presented here. We also discuss the performance problems that arise when the size of the input database grows.

4.1 GSP

The GSP algorithm proposed by Srikant and Agrawal in [5] is based on the pattern of the Apriori algorithm. First, it scans the database and counts the support of each item, detects all frequent single item sequences. Then in each subsequent pass a candidate generation and candidate counting takes place. Candidate generation uses the frequent sequential patterns of the previous pass: if removing the first element of a frequent sequence and removing the last element of another frequent sequence are the same, then the two sequences are joined and a new sequence with one more item is created. The candidate counting scans for each new sequence in the database counting the occurrences, and the ones with support greater than the user defined minimum support are retained as frequent sequences of the pass. The candidate generation and candidate counting are repeated until no frequent sequences are found.

4.2 PrefixSpan

PrefixSpan proposed by Pei et al. [10] is also based on the frequent pattern growth principle like GSP, however, it does not perform the search on the entire database for each candidate sequence, but on smaller projected databases. The sequence database is partitioned based on the itemsets of each frequent sequence of previous passes in a way that all sequences that support the frequent sequence are within the partition and the sequences not supporting it are not. If several frequent sequences share the same itemset, then they use the same database partition. The hypothesis is that the support of a frequent sequence that is one item longer can be calculated on that partition as outside of that partition it is not supported. New candidate sequences are generated only locally by combining sequences that use the same partition. This method is a significant speed improvement over GSP as database partitions are smaller and because of the shared itemsets candidate counting does not need to be performed for each frequent sequence, but only for shared prefixes.

4.3 SPADE

SPADE (Sequential PAttern Discovery using Equivalence Classes) proposed by Zaki [3] aims to reduce the number of database scans and minimize computational costs. During the database scans frequent sequences of length one and two are searched for and their support is counted. The algorithm maintains an id-list for each item where each element of

the id-list is a pointer to a sequence id and an itemset the item occurs in. Candidate sequences with one more item are generated with temporal joins or intersections on the id-lists of frequent sequences of maximum length, the support is calculated in the memory, and the new sequence is frequent if the cardinality of the resulting id-list is greater than the minimum support value. Frequent sequences are clustered into smaller sub-lattices based on common prefixes that enables independent processing.

4.4 SPAM

SPAM (Sequential PAttern Mining) proposed by Ayres et al. [11] assumes that the entire sequential database can completely fit in the memory and no sequences are longer than 64. The hypothesis is that frequent sequences can be found in the lexicographic tree with a simple depth-first search. Each sequence is represented with a vertical bitmap, if an item appears in a sequence then the corresponding element of the bitmap is set to one. Itemsets are generated with a bitwise and operation on the vectors of the items. Candidate sequences are generated with depth-first search from bitvectors of previous sequences and the vector of a next item in the lexicographic tree such that a bitwise and operation is performed on the two vectors, the candidate is frequent if it has more ones in its bitvector than the minimum support. The algorithm is fast but very limited regarding the input database.

4.5 Performance issues

In [12], Gouda and Hassaan argue that typical sequential pattern mining algorithms tend to lose their efficiency when applied to a dense database. Their experiments confirm that the execution time increases exponentially as the number of frequent sequences increases even when the execution times in their experiments remain in the order of a few hundred seconds.

5. Comparison: Experimental Results

In this section, we present the experiments we conducted on real-life clinical data. The clinical database was anonymized [13] before use.

We defined one sequence for each unique patient identifier, i.e. the I set comprises patient identifiers. Treatments and diagnoses have unique medical codes that define the A itemset. Treatment and diagnosis timestamps are aggregated on daily level, that is, two treatments that happened on the same day are considered to be simultaneous and have the same $t \in T$ element associated with them.

The properties of the data set are shown in Table 4.

The experiments we conducted on an Oracle Sun Server X3-2 with 256GB RAM and 32 cores of 4 Intel Xeon E5-2660 CPUs. The mining processes were allowed to use up to 48GB of RAM and 200GB of disk space. We used the reference implementations available in the SPMF library [14].

Number of records in the database	66870306
Number of natural identifiers	455514
Average length of maximum sequences	146.8
Average size of maximum sequences	6.18
Number of items	9291

Table 4: Properties of the data set

Since the preliminary experiments with PrefixSpan and SPADE have shown that these algorithms are not able to process this amount of data within reasonable time, once again we have used a random sample and limited the maximum length of sequences to 5, 10, 20, 40, 60, 80, 100, 120 and 140. The highest value used is still below the average length of sequences in the whole database. Table 5 shows the properties of these samples.

Max. length	Sequences	Length	Itemsets
5	2150	5048	3441
10	4082	18064	28255
20	6416	50781	17379
40	8979	124289	36448
60	10499	197817	53894
80	11571	271818	71685
100	12263	333065	86099
120	12789	390307	100058
140	13186	441514	112012

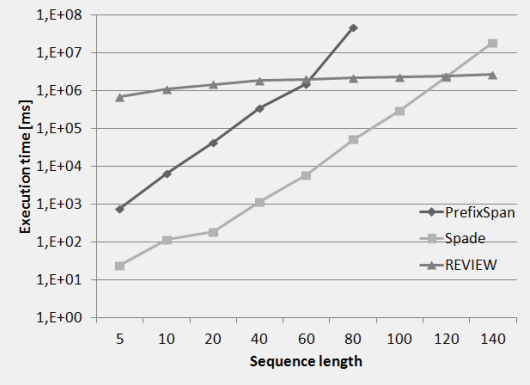
Table 5: Properties of the sample data sets

Figure 1 compares REVIEW with PrefixSpan and SPADE over the same dataset with the same minimum support threshold settings. Algorithms was forced to end after 10^7 ms execution time. The figures show how the execution time requirements and the disk space usage scale as the maximum length of input sequences grows. The number shown is the average of three runs. The algorithms are deterministic, so the number of frequent sequences does not vary. We represent the output sequences in the same form on the disk, so we consider that the number of output sequences is proportional to the disk usage.

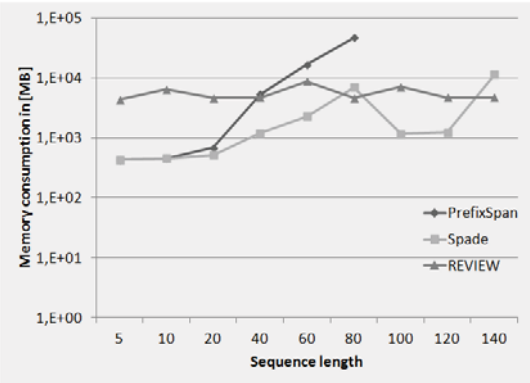
REVIEW was found to scale better as the length and number of input sequences grow than Preview and PrefixSpan. The chart on the left shows that though REVIEW has a high initial time requirement, it does have a much lower gradient on the log scale than the other two. Around the sequence length of 60 REVIEW becomes quicker than PrefixSpan, and around the sequence length of 120 it surpasses SPADE in speed. Though REVIEW works over the same search space as shown in Theorem 10, it is more effective in filtering frequent sequence candidates than the other algorithms, and yet it retains the relevant information.

6. Conclusion

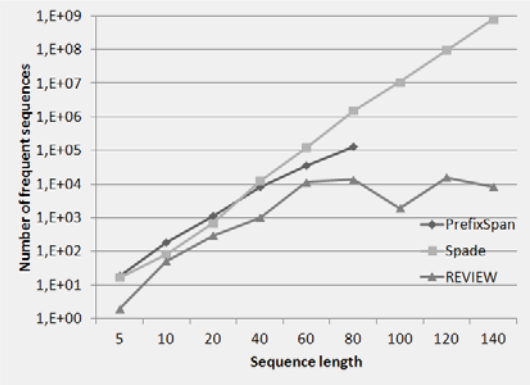
In this paper, we generalized the well-known sequential pattern mining problem in order to find frequent and signif-



(a) Sequence length vs. execution time



(b) Sequence length vs. memory consumption



(c) Sequence length vs. frequent sequences

Fig. 1: Performance of REVIEW against PrefixSpan and SPADE. The maximum length in a sequence database vs. the execution time, the number of frequent sequences and memory consumption

icant patterns in a hand. Usint the REVIEW algorithm we demonstrated how to deal with frequent closed sequences in linear time which significantly improves the performance of those known from literature. Moreover, the re-formulation of the sequential pattern mining problem enables us to find significant patterns for low support items as well, which would be omitted by other methods.

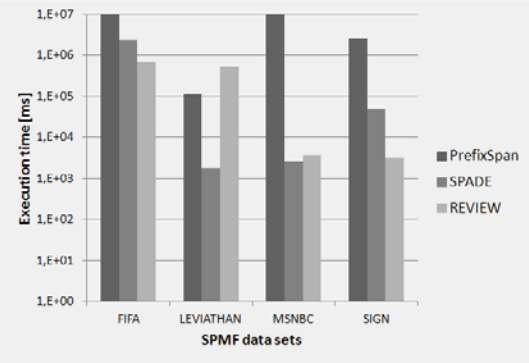


Fig. 2: Execution times of PrefixSpan, SPADE, and REVIEW for data sets [14] ($\mu = 0.05$)

Acknowledgements

Publishing this work were funded by the European Union and co-financed by the European Social Fund under the grant number TÁMOP-4.2.2.D-15/1/KONV-2015-0006.

References

[1] K. McGarry, "A survey of interestingness measures for knowledge discovery," *Knowledge Eng. Review*, vol. 20, no. 1, pp. 39–61, 2005.
[2] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Computing Surveys (CSUR)*, vol. 38, no. 3, p. 9, 2006.
[3] M. J. Zaki, "Spade: An efficient algorithm for mining frequent sequences," *Machine Learning*, vol. 42, no. 1-2, pp. 31–60, Jan. 2001.
[4] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*. IEEE, 1995, pp. 3–14.
[5] R. Srikant and R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements," in *Proceedings of the 5th EDBT*. London, UK, UK: Springer-Verlag, 1996, pp. 3–17.
[6] P. W. Purdom, D. V. Gucht, and D. P. Groth, "Average-case performance of the apriori algorithm," *SIAM Journal on Computing*, vol. 33, no. 5, pp. 1223–1260, 2012.
[7] L. M. Aouad, N.-A. Le-Khac, and T. M. Kechadi, "Performance study of distributed apriori-like frequent itemsets mining," *Knowledge and Information Systems*, vol. 23, no. 1, pp. 55–72, 2010.
[8] G. Salton, E. A. Fox, and H. Wu, "Extended boolean information retrieval," *Communications of ACM*, vol. 26, no. 11, pp. 1022–1036, 1983.
[9] T.-P. Hong, C.-W. Lin, K.-T. Yang, and S.-L. Wang, "A heuristic data-sanitization approach based on tf-idf," in *Proceedings of the IEA/AIE'11*. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 156–164.
[10] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *Proceedings of the 2000 ACM SIGMOD ICMD*. New York, NY, USA: ACM, 2000, pp. 1–12.
[11] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential pattern mining using a bitmap representation," in *Proceedings of the Eighth ACM SIGKDD*, ser. KDD'02. New York, NY, USA: ACM, 2002, pp. 429–435.
[12] K. Gouda and M. Hassaan, "Mining sequential patterns in dense databases," *International Journal of Database Management Systems*, vol. 3, no. 1, pp. 179–194, 2011.
[13] T. Z. Gál, G. Kovács, and Z. T. Kardkovács, "Survey on privacy preserving data mining techniques in health care databases," *Acta Universitatis Sapientiae, Informatica*, vol. 6, no. 1, pp. 33–55, 2014.
[14] P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C.-W. Wu, and V. S. Tseng, "Spmf: a java open-source pattern mining library," *Journal of Machine Learning Research*, vol. 15, pp. 3389–3393, 2014.

Raise regression: selection of the raise parameter

Catalina García, José García, Román Salmerón and María del Mar López

***Abstract*— Collinearity exists when there is a linear quasi-dependence between the explanatory variables of a regression model and in this case the Ordinary Least Square estimator is unstable. Different methodologies have been developed to estimate under collinearity. The raise estimator faces the problem from a geometric point of view by separating both vectors with a raise parameter λ, see [1]. The higher the raise parameter is, the greater the separation between both vector and, consequently, the correlation will be lower. For this reason, the selection of the parameter λ is very important. In this paper, we propose two criteria based on select the value of λ that provides the lowest Mean Square Error analogously to the method proposed to select k in ridge regression. We present an empirical application to compare the results of both criteria and we summarize the most important conclusions in the final section.**

I. Introduction

Given the following linear model with n observations and two variables

$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \mathbf{u} \tag{1}$$

it is assumed the lineal independence between the explanatory variables. When this assumption is not verified it is said that the model has collinearity and in this case the Ordinary Least Square estimator is unstable and frequently it provides misleading results.

Different estimation methods have been proposed in the scientific literature to estimate a model with collinearity. We can distinguish two kinds: methods that solve the problem from an algebraic point of view (such as the ridge estimator [2], [3]) and other methods focused on the sample that propose increase it or eliminate variables, etc. ([4], [5], [6], [7], [8], [9], [10]).

Within this last group we can find the raise method presented by [1]. This method is focused on the sample but instead of deleting data, which can content relevant information, it faces the problem from a geometric point of view taking into account that the collinearity appears due to the vector $\mathbf{x}_1$ and the vector $\mathbf{x}_2$ are geometrically very close. See Figure 1.

Catalina García is with the Department of Quantitative Methods for economics and Business, University of Granada, Spain, (email: cbgarcia@ugr.es).

José García is with the Department of Economic and Business, University of Almería, Spain, (email: jgarcia@ual.es).

Román Salmerón is with the Department of Quantitative Methods for economics and Business, University of Granada, Spain, (email: romansg@ugr.es).

María del Mar López is with the Department of Didactics of Mathematics, University of Granada, Granada, Spain, (email: mariadelmarlopez@ugr.es).

To correct this problem before proceeding to the estimation, we will separate them through the following regression:

$$\mathbf{x}_1 = \alpha \mathbf{x}_2 + \epsilon, \tag{2}$$

whose estimation by OLS is:

$$\hat{\alpha} = (\mathbf{x}_2' \mathbf{x}_2)^{-1} \mathbf{x}_2' \mathbf{x}_1, \tag{3}$$

so it is verified that $\hat{\alpha} = \rho$.

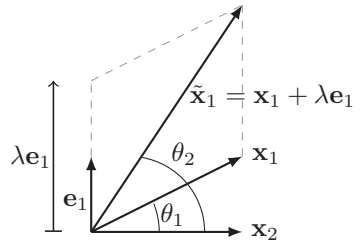


Fig. 1. Representation of raise method

Thus, we can say that $\mathbf{x}_1 = \rho \mathbf{x}_2 + \mathbf{e}_1$ with $\mathbf{e}_1 \perp \mathbf{x}_2$ being $\mathbf{e}_1$ the residual obtained from regression (2). Then, the raise vector, denoted by $\tilde{\mathbf{x}}_1$, is defined as:

$$\tilde{\mathbf{x}}_1 = \mathbf{x}_1 + \lambda \mathbf{e}_1. \tag{4}$$

The raise method will be obtained by substituting in the model (1) the vector $\mathbf{x}_1$ by the raise vector $\tilde{\mathbf{x}}_1$, it is to say, the raise method will be the OLS regression with the vectors $\tilde{\mathbf{x}}_1$ and $\mathbf{x}_2$ instead of $\mathbf{x}_1$ and $\mathbf{x}_2$. Then, the model to estimate will be given by:

$$\mathbf{y} = \beta_1(\lambda) \tilde{\mathbf{x}}_1 + \beta_2(\lambda) \mathbf{x}_2 + \mathbf{w}, \tag{5}$$

where the estimated parameters depend on λ and they will be called raise estimators and denoted as $\hat{\beta}_1(\lambda)$ and $\hat{\beta}_2(\lambda)$.

From Figure 1, it is evident that the angle θ_2 between $\tilde{\mathbf{x}}_1$ and $\mathbf{x}_2$, will be bigger than the angle θ_1 between $\mathbf{x}_1$ and $\mathbf{x}_2$. Thus, the correlation between both vectors will be lesser and the correlation problem has been mitigated. The higher the parameter λ (raise parameter) the greater the angle between the vectors and the lower the correlation. For this reason, it will be very relevant to correctly select the value of λ.

In this paper we focus on how to select the raise parameter λ and we propose to select the value of λ that minimizes the Mean Square Error (MSE) analogously to the criterion applied in ridge regression to obtain k. With this purpose, in Section II we briefly review the mean characteristics of raise estimator and obtain the Mean Square Error of raise

estimator which will be the basis of the criteria to select the raise parameter. We also review some measure to diagnose the existence of collinearity in raise regression. Section III exposes the different criteria that we propose to select the value of λ. Section IV presents an empirical application and, finally, Section V resumes the mean conclusions.

II. Mean characteristics of raise estimator

Since all variables of the raise model (5) are centered, it is evident that $\text{ESS}(\lambda) = \hat{\boldsymbol{\beta}}(\lambda)'\tilde{\mathbf{X}}'\mathbf{y}$. García [11] showed that $\text{ESS}(\lambda)$ will be equal to the ESS in OLS and, evidently, the explained variable in raise model (5) will coincide also with the one in the model (1) estimated by OLS. And then, the Total and the Residual Square Sum (TSS and RSS, respectively) of both models will be also equal. In consequence, the estimated variances will be also similar:

$$\hat{\sigma}^2(\lambda) = \frac{\text{RSS}(\lambda)}{n-2} = \frac{\text{RSS}}{n-2} = \hat{\sigma}^2. \tag{6}$$

An important characteristic of raise estimation is that it keeps the value of the coefficient of determination, R^2, whatever the value of the raise parameter, λ, will be.

Now we will obtain the Mean Square Error (MSE) of raise estimator that will be the basis of the selection criteria that will be proposed.

A. The Mean Square Error of raise estimator

Given an estimator $\boldsymbol{\beta}^*$ of $\boldsymbol{\beta}$, the Mean Square Error (MSE) is calculated as:

$$\begin{aligned}\text{MSE}(\boldsymbol{\beta}^*) &= E[(\boldsymbol{\beta}^* - \boldsymbol{\beta})'(\boldsymbol{\beta}^* - \boldsymbol{\beta})] \\ &= \text{tr}\left(Var\left(\boldsymbol{\beta}^*\right)\right) + \left(E[\boldsymbol{\beta}^*] - \boldsymbol{\beta}\right)'\left(E[\boldsymbol{\beta}^*] - \boldsymbol{\beta}\right).\end{aligned} \tag{7}$$

It is known that the MSE for OLS (unbiased) estimator is given by:

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = \frac{2\sigma^2}{1-\rho^2}, \tag{8}$$

In the case of the raise estimator we obtain the following expression:

$$\text{MSE}(\hat{\boldsymbol{\beta}}(\lambda)) = \frac{2 + (2\lambda + \lambda^2)(1-\rho^2)}{(1+\lambda)^2(1-\rho^2)}\sigma^2 + \frac{\lambda^2(1+\rho^2)\beta_1^2}{(1+\lambda)^2}. \tag{9}$$

Note that for $\lambda = 0$ expression (9) coincides with (8). Also:

$$\lim_{\lambda \to +\infty} \text{MSE}(\hat{\boldsymbol{\beta}}(\lambda)) = \sigma^2 + (1+\rho^2)\beta_1^2. \tag{10}$$

The MSE of raise estimator depends only on the unknown parameter β_1. Specifically, the raise MSE function is a parabola with a minimum at the point $\beta_1 = 0$ and it will be decreasing for negatives values of β_1 and increasing for positive values. In addition, it is verified that:

$$\text{MSE}(\hat{\boldsymbol{\beta}}(\lambda)) \leq \text{MSE}(\hat{\boldsymbol{\beta}}) \Leftrightarrow |\beta_1| \leq \sigma\sqrt{\frac{\lambda+2}{\lambda(1-\rho^2)}}. \tag{11}$$

Consequently, if $\beta_1 = 0$ then it will be always verified that $\text{MSE}(\hat{\boldsymbol{\beta}}(\lambda)) \leq \text{MSE}(\hat{\boldsymbol{\beta}})$.

TABLE I
LIMITS FOR $|\beta_1|$ TO GET THAT $\text{MSE}(\hat{\boldsymbol{\beta}}(\lambda)) \leq \text{MSE}(\hat{\boldsymbol{\beta}})$

λ	$\sigma\sqrt{\frac{\lambda+2}{\lambda(1-\rho^2)}}$
0	∞
0.1	32.4850
0.2	23.5109
0.3	19.6280
0.4	17.3640
0.5	15.8511
0.6	14.7565
0.7	13.9222
0.8	13.2620
0.9	12.7248
1	12.2782

These characteristics can be observed in Figure 2 which displays the $\text{MSE}(\hat{\boldsymbol{\beta}}(\lambda))$ and $\text{MSE}(\hat{\boldsymbol{\beta}})$ for $\sigma^2 = 1$, $\rho = 0.99$, $\lambda = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$ and β_1 taking 2001 values equally distributed in the interval $[-100, 100]$, see Table I. The dark area around $\beta_1 = 0$ are the values where it is verified that $\text{MSE}(\hat{\boldsymbol{\beta}}(\lambda)) \leq \text{MSE}(\hat{\boldsymbol{\beta}})$.

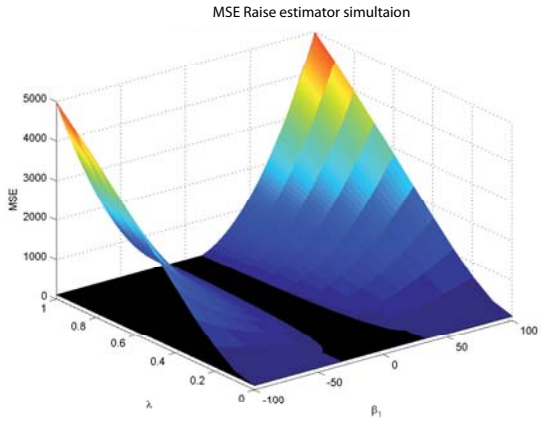


Fig. 2. MSE Raise estimator simulation

B. The VIF and the CVIF in raise regression

The Variance Inflator Factor (VIF) is a widely applied measure to diagnose the existence of collinearity. Garcia et al. [12] extended it to be applied in ridge regression and now we present the extension to raise regression. From the traditional expression

$$\text{VIF}(\lambda) = \frac{1}{1 - R^2_{aux}}, \tag{12}$$

where R^2_{aux} is the coefficient of determination of the regression of $\mathbf{x}_2$ from $\tilde{\mathbf{x}}_1 = (1+\lambda)\mathbf{x}_1 - \lambda\rho\mathbf{x}_2$ (or vice versa):

$$R^2_{aux} = \frac{\rho^2}{(1+\lambda)^2 - (\lambda^2 + 2\lambda)\rho^2}. \tag{13}$$

it is obtained that:

$$\text{VIF}(\lambda) = \frac{(1+\lambda)^2(1-\rho^2) + \rho^2}{(1+\lambda)^2(1-\rho^2)}. \tag{14}$$

Alternatively to the traditional VIF, Curto and Pinto [13] showed that the real impact on variance can be overestimated by the traditional VIF when $R^2 > R_0^2$ being R_0^2 the sum of all square of the correlation coefficients between the dependent variable and each one of the explanatory variables and presented the Corrected VIF (CVIF) in OLS. This measure is obtained by dividing the estimated variance of the estimated coefficient $\beta_j, j = 1, 2$, by the corresponding variance if the variables were orthogonal ($R_j^2 = 0$, with coefficient $\beta_{j0}, j = 1, 2$):

$$\text{CVIF} = \frac{\widehat{var}\left(\widehat{\beta}_j\right)}{\widehat{var}\left(\widehat{\beta}_{j0}\right)} = \text{VIF} \cdot \frac{1 - R^2}{1 - R_0^2} = \text{VIF} \cdot C. \quad (15)$$

In this case, $R^2 > R_0^2$, it is verified that $0 < C < 1$, and then CVIF $<$ VIF. It is to say, the CVIF corrects the overestimation given by the traditional VIF. The authors preserve the rule of thumb CVIF ≥ 10 to decide when the variance magnification effect is serious.

We have obtained the following expression to the CVIF in raise regression:

$$\text{CVIF}(\lambda) = \text{VIF}(\lambda) \cdot \frac{1 - R^2(\lambda)}{1 - R_0^2(\lambda)} = \text{VIF}(\lambda) \cdot C(\lambda), \quad (16)$$

where VIF(λ) is given by (14) and $R^2(\lambda) = R^2$. Indeed, $R_0^2(\lambda) = r^2_{\mathbf{y}\widetilde{\mathbf{x}}_1} + r^2_{\mathbf{y}\widetilde{\mathbf{x}}_2}$, that is, the sum of square of correlation coefficients between $\mathbf{y}$ and $\mathbf{x}_i$, $i = 1, 2$, where $\widetilde{\mathbf{x}}_1 = (1 + \lambda)\mathbf{x}_1 - \lambda\rho\mathbf{x}_2$ and $\widetilde{\mathbf{x}}_2 = \mathbf{x}_2$. Thus, it is possible to obtain the following expression:

$$R_0^2(\lambda) = \frac{\lambda^2(\gamma_1 - \rho\gamma_2)^2 + 2\lambda(\gamma_1 - \rho\gamma_2)\gamma_1 + \gamma_1^2}{(1 - \rho^2)(\lambda^2 + 2\lambda) + 1} + \gamma_2^2. \quad (17)$$

III. Selection of the raise parameter

Similarly to what happen with the parameter k in ridge regression, one of the mean questions in the raise regression is how to select the parameter λ. In this section we propose two criteria to select the parameter λ but both are focused on select a value of λ that allows to solve the collinearity and, simultaneously, presents the lowest MSE.

The first criterion, that we call criterion A, presents the following steps:

1) Firstly, we propose to estimate MSE($\hat{\boldsymbol{\beta}}(\lambda)$) by considering the different values of β_1 obtained from the OLS confidence interval.
2) Next, we select the value of β_1 with the lowest MSE.
3) For the selected value, we represent the value of the MSE for $\lambda \geq 0$.
4) Finally, we select the value of λ that solves the collinearity (VIF < 10) and, simultaneously, presents the lowest MSE.

Some considerations to take into account is that in the second step it will be selected the lowest absolute value of β_1 due to the characteristics of MSE($\hat{\boldsymbol{\beta}}(\lambda)$). Indeed, if the confidence interval used in step 1 contents the zero, then it will be selected the $\beta_1 = 0$. In that case:

$$\text{MSE}(\hat{\boldsymbol{\beta}}(\lambda)) = \frac{2 + (2\lambda + \lambda^2)(1 - \rho^2)}{(1 + \lambda)^2(1 - \rho^2)}\sigma^2,$$

being decreasing in λ and verifying that

$$\lim_{\lambda \to +\infty} \text{MSE}(\hat{\boldsymbol{\beta}}(\lambda)) = \sigma^2.$$

A second criterion, that we call criterion B, is composed by the following steps:

1) To estimate MSE($\hat{\boldsymbol{\beta}}(\lambda)$) by considering the values of β_1 for each λ as the values obtained for $\hat{\beta}_1(\lambda)$.
2) To represent the MSE for $\lambda \geq 0$ for the selected value.
3) Finally, we select the value of λ that solves the collinearity (VIF < 10) and, simultaneously, presents the least MSE.

In this criterion it is verified that (see expression (10))

$$\lim_{\lambda \to +\infty} \text{MSE}(\hat{\boldsymbol{\beta}}(\lambda)) = \sigma^2,$$

since $\lim_{\lambda \to +\infty} \hat{\beta}_1(\lambda) = 0$. And, in this case:

$$\lim_{\lambda \to +\infty} \text{MSE}(\hat{\boldsymbol{\beta}}(\lambda)) < \text{MSE}(\hat{\boldsymbol{\beta}}).$$

In addition, the second member of (9) tends to zero as λ increases, while the first member is decreasing in λ. Consequently, in this case the MSE decreases as λ increases and we will select as ideal value of λ the superior extreme of the considered possible value set of λ.

Note that in both criteria we substitute in MSE($\hat{\boldsymbol{\beta}}(\lambda)$), ρ for its value obtained from the data and σ^2 for its estimation (expression (6)) and the difference between them is the way to estimate MSE($\hat{\boldsymbol{\beta}}(\lambda)$).

IV. Empirical application

To illustrate the contribution of this paper, we will use the empirical application previously applied by [14] and [15]. In this example the total mortality rate, $\mathbf{y}$, is related to the nitrogen oxide pollution potential, $\mathbf{x}_1$, and the hydrocarbon pollution potential, $\mathbf{x}_2$, for 60 cities where $\rho = 0.984$, $\gamma_1 = -0.077$ and $\gamma_2 = -0.177$.

Figure 3 displays the VIF values for raise and ridge estimation. Note that for values of k to 0.02, the ridge VIF is equal to 6.4199, lesser than 10. However, the raise VIF will be lesser than 10 only from values of λ equal to or higher than 0.85. Remember that the value 10 is usually applied as the limit to consider the collinearity problem mitigated. Thus, by using the ridge regression the collinearity will be mitigated for values of k equal or higher than 0.02 while in the raise regression it is necessary values of λ equal or higher than 0.85.

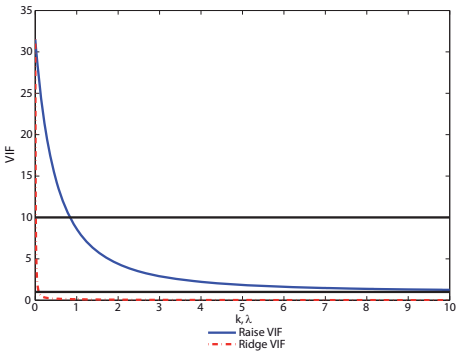


Fig. 3. Representing the VIF of raise and ridge estimation

Next, we will apply the two criteria presented in Section III for the selection of λ. Firstly, by applying the Criterion A, we display in Figure 4 the MSE for the raise and the OLS estimators for the confidence interval $[1.8775,\ 4.1775]$ obtained by OLS for β_1. In this case, all values are positive and then the lowest MSE is obtained by the inferior extreme of the interval, it is to say 1.8775.

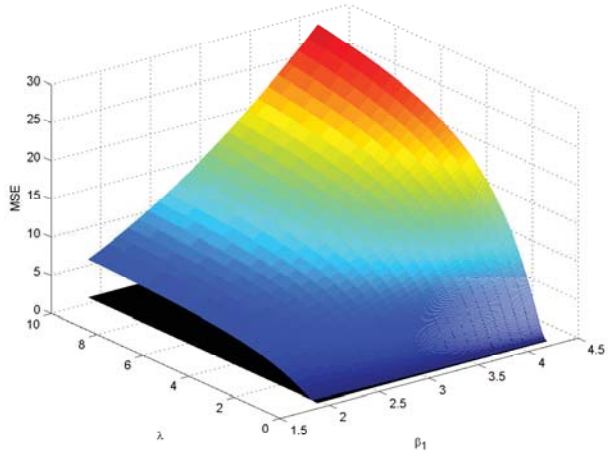


Fig. 4. Raise and OLS mean square error

Then, taking $\beta_1 = 1.8775$ in (9) we will select the value of $\lambda = 0.85$ since from this value we can consider that the collinearity is mitigated (the raise VIF < 10) and it presents the lowest MSE, 1.686, (see Figure 5). Note that for $\lambda > 0.24$ the values of MSE$(\hat{\boldsymbol{\beta}}(\lambda))$ are always higher than MSE$(\hat{\boldsymbol{\beta}})$, which in this case is equal to 0.7292.

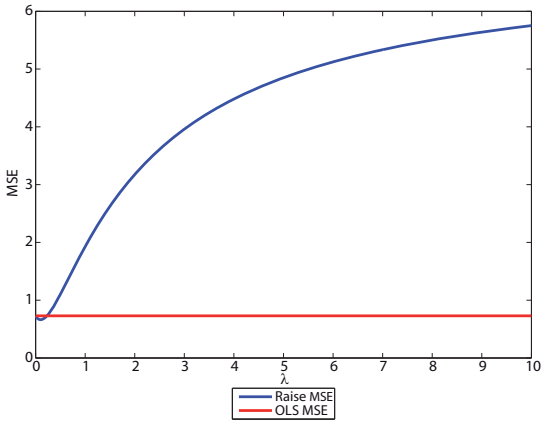


Fig. 5. Raise and OLS mean square error for $\beta_1 = 1.8775$

Then, for $\lambda = 0.85$ it is possible to obtain the following model:

$$\begin{array}{lccc} \hat{\mathbf{y}} & = & 1.6546\ \tilde{\mathbf{x}}_1 & - & 1.8051\ \mathbf{x}_2, \\ & & (0.3264) & & (0.3387) \\ t_{exp} & & 5.0695 & & -5.3269 \end{array}$$

with $\hat{\sigma} = 0.1076$. Both coefficients are significantly different to zero and, consequently, we can state that the nitroxen oxide pollution has a positive influence on the total mortality rate while the hydrocarbon pollution potential has a negative influence. On the other hand, the model is globally significant due to the experimental value, $F_{exp} = 28.4071$, is higher than the theoretical one, 4.0069.

Secondly, we will apply the Criterion B where the ideal value of the raise parameter will be $\lambda = 10$, due to for this value, the collinearity is mitigated (see Figure 3) and it presents the lowest MSE (see Figure 6). Then:

$$\begin{array}{lccc} \hat{\mathbf{y}} & = & 0.2783\ \tilde{\mathbf{x}}_1 & - & 0.4508\ \mathbf{x}_2. \\ & & (0.0549) & & (0.1204) \\ t_{exp} & & 5.0695 & & -3.7451 \end{array}$$

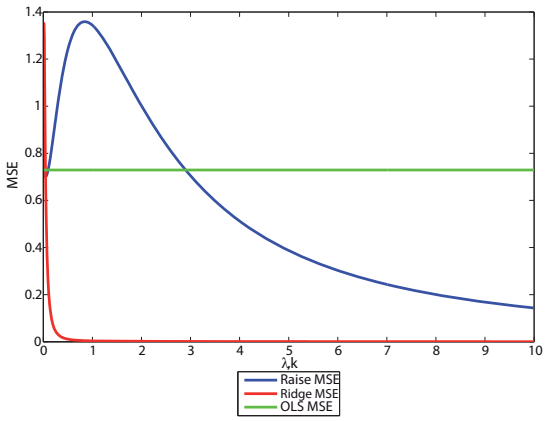


Fig. 6. Raised, Ridge and OLS mean square error

In this case the second variable is not individually significant although the sense of the relation is the same than the one obtained for the model estimated with $\lambda = 0.85$. From the comparison of both models we can highlight that:

- For $\lambda = 10$ if the nitrogen oxide pollution increases one unit, the total mortality rate increases 0.2783 units while for $\lambda = 0.85$ increases 1.6546 units which is almost six times more.
- For $\lambda = 10$ if the hydrocarbon pollution potential increases one unit, the total mortality rate decreases 0.4508 units while for $\lambda = 0.85$ decreases 1.8051 units which is approximately four times more.

In addition, in Criterion B we appreciate a dependence to the interval where λ takes value, which is arbitrary, since the MSE decreases as λ increases. That is, the value of λ with lowest MSE will be the superior extreme of the considered interval. This does not occur in Criterion A, see Figure 7.

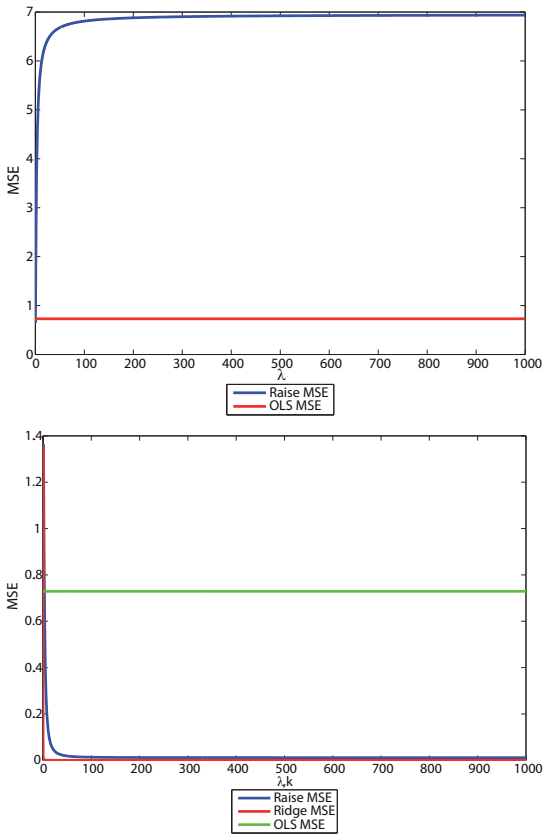


Fig. 7. Mean square error for Criterion A (top) and Criterion B (bottom) with $\lambda \in [0,\ 1000]$

Now, we analyze the behaviour of both criteria with the CVIF. In this case, $R_0^2(\lambda)$ is represented in Figure 8. Indeed, it is possible to note that $R_0^2(\lambda) < R^2$ for all λ, and then CVIF(λ) < VIF(λ) for all λ (see Figure 9). Then, in this case we could state that the CVIF corrects the overestimation of the VIF. It is also possible to prove that CVIF$(\lambda) > 1$ and decreasing for all λ.

For criterion A, and by using the CVIF as measure to diagnose the existence the collinearity, we select the value $\lambda > 0.49$ since from this value it is verified that CVIF$(\lambda) < 10$. Then, for $\lambda = 0.49$ it is obtained the following estimated model:

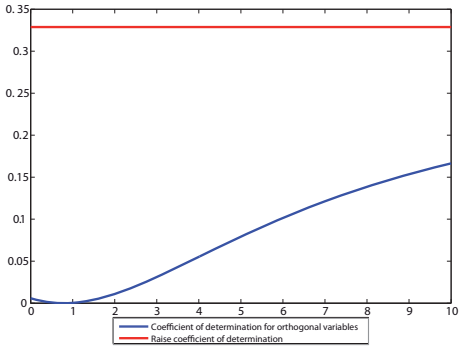


Fig. 8. $R_0^2(\lambda)$

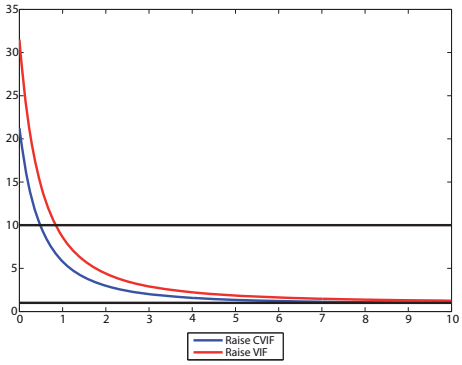


Fig. 9. VIF and CVIF in raise regression)

$$\begin{array}{lccc} \hat{\mathbf{y}} & = & 2.0544\ \mathbf{x}_1 & - & 2.1985\ \mathbf{x}_2. \\ & & (0.4052) & & (0.413) \\ t_{exp} & & 5.0695 & & -5.3231 \end{array}$$

Both coefficients are individually significant and the sense of the influence is the same than in the other presented cases. In the case of criterion B, the selected value of λ will be again 10.

V. Conclusions

One of the main issues in raise regression is how to select the parameter λ as occurs in the ridge estimation with the parameter k. In this paper we propose two criteria based on selecting the value of λ that allows to mitigated the collinearity and simultaneously presents the lowest Mean Square Error. With this purpose we have developed the expression of the Mean Square Error (MSE) for the raise regression and also the expression of the VIF and the CVIF that have been applied as measure to diagnose the collinearity. We present an empirical application to compare the results of both criteria and conclude with the selection of the parameter λ and the estimation and interpretation of the raise regression. This criteria can be applied in many different applications within a great diversity of fields where collinearity exists.

References

[1] C. B. García, J. García, and J. Soto, "The raise method: An alternative procedure to estimate the parameters in presence of collinearity," *Quality and Quantity*, vol. 45, no. 2, pp. 403–423, 2010.

[2] A. E. Hoerl and R. W. Kennard, "Ridge regression: applications to nonorthogonal problems," *Technometrics*, vol. 12, pp. 69–82, 1970.
[3] A. E. Hoerl, R. W. Kennard, and K. F. Baldwin, "Ridge regression. some simulations," *Communications in Statistics*, vol. 4, no. 2, pp. 105–123, 1975.
[4] M. G. Kendall, *A course in Multivariate Analysis*. London: Griffin, 1957.
[5] C. R. Rao, "A note on a generalized inverse of a matrix with applications to problems in mathematical statistics," *Journal of the Royal Statistical Society*, vol. B, no. 24, pp. 152–158, 1962.
[6] W. F. Massy, "Principal components regression in exploratory statistical research," *Journal of the American Statistical Association*, vol. 60, pp. 234–256, 1965.
[7] D. E. Farrar and R. R. Glaubert, "Multicollinearity in regression analysis. the problem revisited," *Review of Economics and Statistics*, vol. 49, pp. 92–107, 1967.
[8] S. D. Silvey, "Multicollinearity and imprecise estimation," *Journal of Statistical Society*, vol. 31, no. B, pp. 539–552, 1969.
[9] D. W. Marquardt, "Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation," *Technometrics*, vol. 12, no. 3, pp. 591–612, 1970.
[10] J. Johnston, *Métodos de Econometría*. Barcelona: Vicens-Vives, 1989.
[11] C. García, J. García, M. M. López Martín, and R. Salmerón, "Raise estimator: inference and properties," *Communications in Statistics-Theory and Methods*, vol. Under review, 2016.
[12] ——, "Collinearity: Revisiting the variance inflation factor in ridge regression," *Journal of Applied Statistics*, vol. 42, no. 3, pp. 648–661, 2015.
[13] J. Curto and J. Pinto, "The corrected vif," *Journal of Applied Statistics*, vol. 38, no. 7, pp. 1499–1507, 2011.
[14] G. C. McDonald and R. C. Schwing, "Instabilities of regression estimates relating air pollution to mortality," *Technometrics*, vol. 15, pp. 463–481, 1973.
[15] G. C. McDonald, "Tracing ridge regression coefficients," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, pp. 695–703, 2010.