A Step Towards Demand Sensing: Employing EDI 852 Product Activity Data in Demand Forecasting

Van Belle, Jente Emiel; Verbeke, Wouter

Published in:
Third Conference on Business Analytics in Finance and Industry

Publication date:
2018

BAFI 2018

Jan17th-19th
Santiago - Chile

Jan 17th - 19th 2018

THIRD CONFERENCE ON BUSINESS ANALYTICS IN FINANCE AND INDUSTRY

PROCEEDINGS

www.baficonference.cl

ORGANIZERS

ISCI Instituto Sistemas Complejos de Ingeniería

Universidad de Chile

SPONSORS

Facultad Economía y Negocios Universidad de Chile | DCS Departamento de Control de Gestión y Sistemas de Información

fcfm Facultad de Ciencias Físicas y Matemáticas Universidad de Chile

Ingeniería Industrial Universidad de Chile

Universidad de los Andes > Ingeniería 20 años

Unholster

MetricArts

Centro de Finanzas

KDnuggets

IEEE Computational Intelligence Society

SANTIAGO, CHILE.

CONFERENCE PRESENTATION

BAFI 2018
Jan17th-19th
Santiago - Chile

The aim of the BAFI 2018 conference is to bring together researchers and developers from data science and related areas with practitioners and consultants applying the respective techniques in different business-related domains. Our goal is to stimulate an academic exchange of recent developments as well as to encourage the mutual influence between academics and practitioners.

At BAFI 2018, the focus will be on methodological developments aimed at uncovering information contained in large data sets, as well as on business applications in various sectors, among them finance, retail, and telecommunications.

INDEX

Page **Session**

TIMETABLE

Jan 17th - 19th 2018

THIRD CONFERENCE ON BUSINESS ANALYTICS IN FINANCE AND INDUSTRY

	WEDNESDAY 17TH		THURSDAY 18TH		FRIDAY 19TH		SATURDAY 20TH
08:30 - 09:00	*Registration & Welcome*						
09:00 - 10:00	**Keynote W1:** Witold Pedrycz		**Keynote T1:** Peter Flach		**Keynote F1:** Dominik Slezak		09:00 - 18:00 **EXCURSION**
10:00 - 11:00	**Keynote W2:** Francisco Herrera		**Keynote T2:** Gianluca Bontempi		**Keynote F2:** Usama Fayyad		
11:00 - 11:30	*Coffee Break*		*Coffee Break*		*Coffee Break*		
11:30 - 13:00	**WA1.1**	**WB1.1**	**TA1.1**	**TB1.1**	Panel	**FB1.1**	
	WA1.2	**WB1.2**	**TA1.2**	**TB1.2**		-	
	WA1.3	**WB1.3**	**TA1.3**	**TB1.3**		**FB1.3**	
	WA1.4	**WB1.4**	**TA1.4**	**TB1.4**		**FB1.4**	
13:00 - 14:00	*Lunch*		*Lunch*		*Lunch*		
14:00 - 15:30	**WA2.1**	**WB2.1**	**TA2.1**	**TB2.1**	**FA2.1**	**FB2.1**	
	WA2.2	**WB2.2**	**TA2.2**	**TB2.2**	**FA2.2**	**FB2.2**	
	WA2.3	**WB2.3**	**TA2.3**	**TB2.3**	**FA2.3**	**FB2.3**	
	WA2.4	**WB2.4**	**TA2.4**	**TB2.4**			
15:30 -16:30	**Keynote W3:** Enrique Herrera-Viedma		**Keynote T3:** Richard Weber		**Keynote F3:** Pablo Zegers		
16:30 - 17:00	*Coffee Break*		*Coffee Break*		Closing Session		
17:00 - 18:30	**WA3.1**	**WB3.1**	**TA3.1**	**TB3.1**	*Farewell Drink*		
	WA3.2	**WB3.2**	**TA3.2**	**TB3.2**			
	WA3.3	**WB3.3**	**TA3.3**	**TB3.3**			
	WA3.4						
18:30	*Welcome Reception*						
20:00			*Dinner*				

CONFERENCE PROGRAM

WEDNESDAY 17

9:00 - 10:00
Lecture room: P309

Keynote **W1** "Linkage Discovery: Bidirectional and Multidirectional Associative Memories In Data Analysis"
Witold Pedrycz

10:00 - 11:00
Lecture room: P309

Keynote **W2** "A tour on Imbalanced big data classification and applications"
Francisco Herrera

SESSION **WA1** 11:30 - 13:00

Lecture room: P309

WA1.1
Skewness measures based on OWA operator and applications to decision making
Rajkumar Verma and José M. Merigó

WA1.2
Divisions with the induced ordered weighted average
José M. Merigó and Sigifredo Laengle

WA1.3
The induced heavy moving average volatility
Ernesto Leon-Castro, Ezequiel Aviles-Ochoa and Jose Maria Merigo-Lindahl

WA1.4
Formulating the Weighted Average - Experton
Salvador Linares-Mustarós, Dolors Corominas-Coll, Joan Carles Ferrer-Comalat and Jose Merigo

SESSION **WA2** 14:00 - 15:30

Lecture room: P309

WA2.1
Combining Support Vector Machine classification and profit measures in credit scoring
Sebastián Maldonado, Cristian Bravo, Julio López and Juan Perez.

WA2.2
Thirty years of the Journal of Business & Industrial Marketing: a bibliometric analysis
Leslier Valenzuela, José M. Merigó, Wesley Johnston, Carolina Nicolas and Jorge F Jaramillo

WA2.3
An overview of the most cited papers in computer science
Gustavo Zurita, José M. Merigó, Valeria Lobos-Ossandón and Carles Mulet-Forteza

WA2.4
Academic research on support vector machines: A bibliometric overview
Jaime Miranda, José M. Merigó and Sebastian Maldonado

SESSION **WB1** 11:30 - 13:00

Lecture room: P301

WB1.1
Spectral Mixture Kernels for Multi-Output Gaussian Processes
Gabriel Parra and Felipe Tobar

WB1.2
Reproducing kernel Hilbert space approach to Stochastic Frontier Analysis (SFA)
Carlos Felipe Valencia, Daniel Arocha and Ader Villar

WB1.3
An Evaluation of Missing Data Handling Mechanisms of Evidential Reasoning Rule for Data Classification
Shao Cong Lim, Dong-Ling Xu and Jian-Bo Yang

WB1.4
Reparameterizing the Birkhoff Polytope for Variational Permutation Inference
Scott Linderman, Gonzalo Mena, John Cunningham, Liam Paninski and Hal Cooper

SESSION **WB2** 14:00 - 15:30

Lecture room: P301

WB2.1
Accounting Fraud Detection Through Forensic Analytics
Maria Jofre

WB2.2
Theoretical and practical aspects of various measures of portfolio diversification
Tomáš Tichý

WB2.3
Animal movement in Mato Grosso do Sul and its implications for economic impacts of potential outbreaks of foot-and-mouth disease
Tais Cristina de Menezes, Silvia Helena Galvão de Miranda and Ivette Luna

WB2.4
Patterns and Insights of an Asset Order Book
Andrew Day and Matt Davison

15:30 - 16:30
Lecture room: P309

Keynote **W3** "Bibliometric Tools for Discovering Information in Science"
Enrique Herrera-Viedma

SESSION **WA3** 17:00 - 18:30

Lecture room: P309

WA3.1
Churn Prediction in Telco using Adapted node2vec on CDR graphs enriched with RFM information
Sandra Mitrovic, Bart Baesens, Wilfried Lemahieu and Jochen De Weerdt

WA3.2
Privacy preserving Customer Churn Models using Support Vector Machines
Abelino Jimenez and Bhiksha Raj

WA3.3
Churn Prediction through Customer Feedback Analytics
Carolina Martinez, Babis Theodoulidis and David Diaz

WA3.4
Effect of Sample Representativeness in Multivariate Symmetrical Uncertainty for Categorical Attributes
Gustavo Sosa-Cabrera, Miguel García-Torres, Santiago Gómez-Guerrero, Christian Schaerer and Federico Divina

SESSION **WB3** 17:00 - 18:30

Lecture room: P301

WB3.1
The (pseudo-)social behavior of products in offline retail stores: Predicting increase in product interpurchase time
Jasmien Lismont, Sudha Ram, Bart Baesens, Wilfried Lemahieu and Jan Vanthienen

WB3.2
Factors affecting banking efficiency scores in Network SBM DEA model: Dealing with heterogeneity
Skarleth Carrales Escobedo and Jamal Ouenniche

WB3.3
Analysis of UK and US SME Platform Markets using Business Model Theory: An emphasis on Web 2.0 technology sophistication
María Manuela Gutiérrez-Leefmans

THURSDAY 18

9:00 - 10:00
Lecture room: P309

Keynote T1 "The value of evaluation: towards trustworthy machine learning"
Peter Flach

10:00 - 11:00
Lecture room: P309

Keynote T2 "Machine Learning for Predicting in a Big Data World"
Gianluca Bontempi

SESSION TA1

11:30 - 13:00

Lecture room: P309

TA1.1
Forecasting blood donations with neural networks
Tine Van Calster, Michael Reusens, Bart Baesens and Wilfried Lemahieu

TA1.2
Predicting dwell times of import containers in a container terminal: case study of the port of Arica in Chile
Francisca Quijada, Sebastian Maldonado and Rosa Guadalupe Gonzalez Ramirez

TA1.3
A Step Towards Demand Sensing: Employing EDI 852 Product Activity Data in Demand Forecasting
Jente Van Belle and Wouter Verbeke

TA1.4
Data-driven inventory management: A random forest-based joint estimation and optimization model for the newsvendor problem
Fabian Taigel and Jan Meller

SESSION TB1

11:30 - 13:00

Lecture room: P301

TB1.1
Behavior based time-to-default predictions
María Óskarsdóttir, Cristian Bravo, Bart Baesens and Jan Vanthienen

TB1.2
Leveraging PD Models for Bayesian Inference of Default Correlations
Miguel Biron and Victor Medina

TB1.3
Bias-Free Text Evaluations in Micro and SME Credit Scoring using Deep Learning
Cristian Bravo and Andrés Medina

TB1.4
Psychometric Credit Scoring Model for Microloan based on HEXACO Personality Inventory
Bo Kyeong Lee, Dong Ha Kim and So Young Sohn

SESSION TA2

14:00 - 15:30

Lecture room: P309

TA2.1
Exploring the relationship between online activity and achievement across two universities and learning management systems
Sergio Celis, Joaquín Muñoz, Dany Lopez and Augusto Sandoval

TA2.2
Combining student learning research and learning analytics to understand student's learning process
Maximiliano Montenegro and Carlos Gonzalez

TA2.3
Learning analytics: traps for the unwary
Carolina Guzmán

TA2.4
Blended analytics: Capturing and visualizing physical and digital learning
Sarah Howard, Jie Yang and Jun Ma

SESSION TB2

14:00 - 15:30

Lecture room: P301

TB2.1
Outlining New Product Development Research through Bibliometrics: Analyzing Journals, Articles and Researchers
Nelson Andrade-Valbuena and Jose Merigo-Lindahl

TB2.2
Featurization Methods and Predictors for Income Inference based on Communication Patterns
Martin Fixman, Martin Minnoni, Matias Travizano and Carlos Sarraute

TB2.3
New Item Recommendation Method Based on Latent Topic Extraction
Maria Emilia Charnelli, Laura Lanzarini, Aurelio Fernández and Javier Díaz.

TB2.4
A comparative assessment of machine learning techniques using payroll issuers data
Hugo Pérez, Jonas Velasco and Ramiro Navarro

15:30 - 16:30
Lecture room: P309

Keynote T3 "Dynamic Data Mining"
Richard Weber

SESSION TA3

17:00 - 18:30

Lecture room: P309

TA3.1
Predictive model for selection of undergraduate applicants
Felipe Bugueno and Jaime Miranda

TA3.2
Forecasting individual course demand in higher education institutions using Machine Learning
Giancarlo A. Acevedo, Jorge Amaya, Salvador Flores, Pablo Huentelemu

TA3.3
Educational Retention Program: Data Mining Techniques for improving performance
Jonathan Vasquez, Jaime Miranda and Sebastián Maldonado

SESSION TB3

17:00 - 18:30

Lecture room: P301

TB3.1
Collecting and Analyzing Customers Experiences From Trip Advisor Social Media
Sebastian Maldonado Alarcon, Carla Marina Vairetti and Guillermo Armelini Wilde

TB3.2
An Approach to Identify Cohesive Subgroups of Banks in Bank-Firm Networks
Samrat Gupta and Pradeep Kumar

TB3.3
Detection of suppliers communities in e-commerce through graph analytics
Fabiola Herrera, Romina Torres and Rodrigo Salas

FRIDAY 19

9:00 - 10:00
Lecture room: P309

Keynote **F1** "Toward Approximate Analytics – Approximate Query Engines & Approximate Data Exploration"
Dominik Slezak

10:00 - 11:00
Lecture room: P309

Keynote **F2** TBA
Usama Fayyad

SESSION **FA1** 11:30 - 13:00

Lecture room: Aula Magna

PANEL

(in Spanish)
"Los desafíos para la colaboración Universidad-Empresa en el área de Data Science"

El panel será en castellano y pondrá un énfasis especial en la igualdad de género en el área de data science, analizando oportunidades y desafíos.

Panelistas:
Patricio Cofré. *Metric Arts, Chile*
Antonio Díaz-Araujo. *Unholster, Chile*
Nuria Oliver. *Vodafone, Spain*
Marina Tannenbaum. *Easybots, Chile*
Carla Vairetti. *Universidad de Los Andes, Chile*

SESSION **FA2** 14:00 - 15:30

Lecture room: P309

FA2.1
Real-time Pedestrian Detection and Tracking in a Multicamera System
Roberto Muñoz, Roberto Gonzalez, Alejandro Sazo and Patricio Cofre

FA2.2
Optimal Selling of a Commodity via Forward Markets in a Cash-and-Carry Trade
Behzad Ghafouri and Matt Davison

FA2.3
Visualizing and Analyzing the 2017 Elections in Chile from a Public Perspective
Sebastian Acuña and Cristóbal Huneeus

SESSION **FB1** 11:30 - 13:00

Lecture room: P301

FB1.1
Identifying successful search patterns for improved job recommendation
Michael Reusens, Wilfried Lemahieu, Bart Baesens and Luc Sels

FB1.2 -

FB1.3
Beyond clickthrough rate: measuring the true impact of personalized e-mail product recommendations
Stijn Geuens, Koen W. De Bock and Kristof Coussement

FB1.4
Exploring online travel reviews using data analytics: an exploratory study
Vera Migueis

SESSION **FB2** 14:00 - 15:30

Lecture room: P301

FB2.1
Finding vehicle theft patterns using association rules and text mining
Cristian Aguayo and Richard Weber

FB2.2
Obtaining and evaluation of extractive summaries from stored text documents
Augusto Villa-Monte, Laura Lanzarini, Aurelio Fernández-Bariviera and José A. Olivas

FB2.3
Market Basket Analysis Insights To Support Category Management
Luis Aburto and Andrés Musalem

15:30 - 16:30
Lecture room: P309

Keynote **F3** "Artificial Intelligence, a Revolution in Latin America As Well"
Pablo Zegers

CONFERENCE PROCEEDINGS

BAFI2018 KEYNOTES

Keynote **W1**

"Linkage Discovery: Bidirectional and Multidirectional Associative Memories In Data Analysis"

Witold Pedrycz

Associative memories are representative examples of associative structures, which have been studied intensively in the literature and have resulted in a plethora of applications in areas of control, classification, and data analysis. The underlying idea is to realize associative mapping so that the recall processes (both one-directional and bidirectional) are characterized by a minimal recall error.

We carefully revisit and augment the concept of associative memories by proposing some new design directions. We focus on the essence of structural dependencies in the data and make the corresponding associative mappings spanned over a related collection of landmarks (prototypes). We show that a construction of such landmarks is supported by mechanisms of collaborative fuzzy clustering. A logic-based characterization of the developed associations established in the framework of relational computing is discussed as well.

Structural augmentations of the discussed architectures to multisource and multi-directional memories involving associative mappings among various data spaces are proposed and their design is discussed.

Furthermore we generalize associative mappings into their granular counterparts in which the originally formed numeric prototypes are made granular so that the quality of the associative recall can be quantified. Several scenarios of allocation of information granularity aimed at the optimization of the characteristics of recalled results (information granules) quantified in terms of coverage and specificity criteria are proposed.

Keynote **W2**

"A tour on Imbalanced big data classification and applications"

Francisco Herrera (Universidad de Granada, Spain)

Big Data applications are emerging during the last years, and researchers from many disciplines are aware of the high advantages related to the knowledge extraction from this type of problem.

The topic of imbalanced classification has gathered a wide attention of researchers during the last several years. It occurs when the classes represented in a problem show a skewed distribution, i.e., there is a minority (or positive) class, and a majority (or negative) one. This case study may be due to rarity of occurrence of a given concept, or even because of some restrictions during the gathering of data for a particular class. In this sense, class imbalance is ubiquitous and prevalent in several applications. The emergence of Big Data brings new problems and challenges for the class imbalance problem.

In this lecture we focus on learning from imbalanced data problems in the context of Big Data, especially when faced with the challenge of Volume. We will analyze the strengths and weaknesses of various MapReduce-based algorithms that address imbalanced data. We will present the current approaches presenting real cases of study and applications, and some research challenges.

Keynote **W3**

"Bibliometric Tools for Discovering Information in Science"

Enrique Herrera-Viedma (Universidad de Granada, Spain)

In bibliometrics, there are two main procedures to explore a research field: performance analysis and science mapping.Performance analysis aims at evaluating groups of scientific actors (countries, universities, departments, researchers) and the impact of their activity on the basis of bibliographic data. Science mapping aims at displaying the structural and dynamic aspects of scientific research, delimiting a research field, and quantifying and visualizing the detected subfields by means of co-word analysis or documents co-citation analysis. In this talk we present two bibliometric tools that we have developed in our research laboratory SECABA: H-Classics to develop performance analysis by based on Highly Cited Papers and SciMAT to develop science mapping guided by performance bibliometric indicators.

Keynote **T1**

"The value of evaluation: towards trustworthy machine learning"

Peter Flach (University of Bristol, UK)

Machine learning, broadly defined as data-driven technology to enhance human decision making, is already in widespread use and will soon be ubiquitous and indispensable in all areas of human endeavour. Data is collected routinely in all areas of significant societal relevance including law, policy, national security, education and healthcare, and machine learning informs decision making by detecting patterns in the data. Achieving transparency, robustness and trustworthiness of these machine learning applications is hence of paramount importance, and evaluation procedures and metrics play a key role in this.

In this talk I will review current issues in theory and practice of evaluating predictive machine learning models. Many

issues arise from a limited appreciation of the importance of the scale on which metrics are expressed. I will discuss why it is OK to use the arithmetic average for aggregating accuracies achieved over different test sets but not for aggregating F-scores. I will also discuss why it is OK to use logistic scaling to calibrate the scores of a support vector machine but not to calibrate naive Bayes. More generally, I will discuss the need for a dedicated measurement theory for machine learning that would use latent-variable models such as item-response theory from psychometrics in order to estimate latent skills and capabilities from observable traits.

Keynote **T2**

"Machine Learning for Predicting in a Big Data World"

Gianluca Bontempi

The increasing availability of massive amounts of data and the need of performing accurate forecasting of future behavior in several scientific and applied domains demands the definition of robust and efficient techniques able to infer from observations the stochastic dependency between past and future. The forecasting domain has been influenced, from the 1960s on, by linear statistical methods such as ARIMA models. More recently, machine learning models have drawn attention and have established themselves as serious contenders to classical statistical models in the forecasting community.

This talk will present an overview of machine learning techniques in time series forecasting and will focus on machine learning strategies to address three important tasks: univariate one-step-ahead prediction, univariate multi-step-ahead prediction and multivariate multi-step-ahed forecasting. Also, it will present DFML, a machine learning version of the Dynamic Factor Model (DFM), a successful forecasting methodology well-known in econometrics. The DFML strategy is based on a out-of-sample selection of the nonlinear forecaster, the number of latent components and the multi-step-ahead strategy. We will show that DFML can consistently outperform state-of-the-art methods in a number of synthetic and real forecasting tasks.

Keynote **T3**

“Dynamic Data Mining”

Richard Weber (Universidad de Chile, Chile)

We are witnessing a tremendous interest in data mining and related topics in research, industry, and public organizations. Virtually all areas of our daily life are affected, such as health, security, business, education, transportation, to name just a few. But most of the systems that are being used today are static in the sense that they consider snapshots of the respective phenomena under study. In this talk, we will present situations where dynamics play a crucial role in order to better understand the analyzed behavior. By reviewing some techniques for dynamic data mining we will provide the current state as well as challenges of this area. The talk ends with some ideas regarding future developments.

Keynote **F1**

“Toward Approximate Analytics – Approximate Query Engines & Approximate Data Exploration”

Dominik Slezak (Institute of Informatics, University of Warsaw, Poland)

Corresponding Author(s):

Artificial Intelligence (AI) methods are regaining a lot of attention in the areas of data analytics and decision support.

Given the increasing amount of information and computational resources available, it is now possible for intelligent algorithms to learn from the data and assist humans more efficiently. Still, there is a question about the goals of learning and a form of the resulting data-driven knowledge. It is evident that humans do not operate with precise information in decision-making and, thus, it might be unnecessary to provide them with complete outcomes of analytical processes. Consequently, the next question arises whether approximate results of computations or results derived from the approximate data could be delivered more efficiently than their standard counterparts. Such questions are analogous to the ones about precision of calculations conducted by machine learning and KDD methods, whereby various heuristic algorithms could be boosted by letting them rely on approximate computations. This leads us toward discussion of the importance of approximations in the areas of machine intelligence and business intelligence and, more broadly, the meaning of approximate derivations and representations for various aspects of AI. In this talk, we refer to this discussion using three industry-related case studies: 1) The case of approximate analytical database software based on the paradigms of rough-granular computing applied in the area of cyber-security; 2) The case of rough-set-based feature subset ensemble selection / approximation methodology applied in the area of online health support services; and 3) The case of approximate generation of the training data used for tuning an online eSports coaching platform.

Keynote **F2**

TBA

Usama Fayyad

Keynote **F3**

"Artificial Intelligence, a Revolution in Latin America As Well"

Pablo Zegers (Sortbox, Chile)

The impact caused by Artificial Intelligence everywhere follows two main drivers: (i) the extent to which Artificial Intelligence (AI) can be applied in human society, a discipline that basically aims at building machines that mimic all human behavior, hence with the potential of affecting all human activity, (ii) the lucky discovery of mathematical rules that facilitate building incredible complex learning machines. These two drivers merged some 5 years ago and started what is now called the AI revolution. In this presentation these two driving components will be explained, followed by a review of the current state of the art.

Given that AI has been present in games, movies, and books, it is important to gain a precise knowledge of what can currently be achieved by this field, and to be conscious of its limitations, in order to avoid unrealistic expectations.

Also, in order to understand how serious is what is happening, a review of the reactions of the private sectors, the government in many countries, and the societies all around the planet, is presented as well. The presentation continuous with an analysis of the expected scenarios, and how the job market is expected to adapt to the changes induced by the introduction of AI systems. The hybrid scenario, also called the centaur scenario, where man and machine work together, allowing for important quality and productivity increases, is analyzed.

Finally, the Latin American view is presented, where local opportunities are pointed out. In general, these opportunities are where: (i) processes are composed of very structured subprocesses, ripe for replacement by an AI system, (ii) private silos with private data exist, out of the reach of the big AI companies, where local companies can gain the trust of their local peers and offer AI services, and (iii) things need to run out of the internet, thus no SaaS system can

compete with them (massive real-time video processing, AIs coordinating robots, etc.). The presentation continues pointing out the challenges in Latin America, mostly related to the lack of prepared professionals in the field, impeding a fast deployment of the technologies, deterring a fast product assembly, henceforth weakening the local capacity to compete with companies from abroad. The presentation ends with an analysis of the ethical implications of applying AI technologies in a society.

BAFI2018 PAPERS

WEDNESDAY 17

Sessions WA 1

WA 1.1

Skewness measures based on OWA operator and applications to decision making

Rajkumar Verma (University of Chile, Chile), José M. Merigó (University of Chile, Chile)
Corresponding Author(s): Rajkumar Verma (rverma@fen.uchile.cl)

Abstract

In statistical theory, skewness is a basic notion which used to describe the distributional asymmetry of data about its mean. The skewness measure can be positive, negative or zero (for normal distribution/ symmetric data), depending on whether data points are skewed to the left or to the right of the data average. In the literature, a number of measures associated with skewness are proposed by different researchers.

This paper introduces a new class of skewness measures based on OWA operator. We will call it Skewness Ordered Weighted Average (Skew-OWA) operator. The main advantage of Skew-OWA is that it provides a parameterized family of skewness measures between the maximum and the minimum skewness based on a complex reordering process according to the attitudinal character of the decision-maker. A number of properties of the new aggregation operator are studied and their special cases are discussed. In addition, some other generalizations are also introduced by using generalized and quasi-arithmetic means with Skew-OWA. Based on the Skew-OWA operator, a decision-making approach is developed to deal

with multiple attribute group decision making problems. The paper ends with a practical example focused on money investment to illustrate the decision-making process and flexibility of the approach.

WA 1.2

Divisions with the induced ordered weighted average

José M. Merigó (University of Chile, Chile), Sigifredo Laengle (Facultad de Economía y Negocios, Universidad de Chile, Chile)
Corresponding Author(s): José M. Merigó (jmerigo@fen.uchile.cl)

Abstract

The induced ordered weighted average (IOWA) is an aggregation operator that provides a parameterized family of aggregation operators between the minimum and the maximum (Yager and Filev, 1999). The main difference with the ordered weighted average (OWA) (Yager, 1988) is that the IOWA operator uses a reordering of the data based on order inducing variables while the OWA operator uses a decreasing or increasing numerical order. Recently, Merigó et al. (2016) introduced the OWA division (OWAD) as an aggregation operator that aggregates a set of divisions providing summarized information from the minimum to the maximum division of the set.

The aim of this study is to present the induced OWAD (IOWAD) operator. It is a new aggregation operator that aggregates a set of divisions from the minimum to the maximum by using complex reordering processes based on order inducing variables. The work considers a wide range of particular cases including the average division, the OWAD operator and the IOWA operator. The article also develops several generalizations by using generalized and quasi-arithmetic means (Merigó and Gil-Lafuente, 2009) forming the induced generalized OWAD (IGOWAD) and the Quasi-IOWAD operator.

The paper ends analyzing the applicability of the IOWAD operator. The focus is on the aggregation of a set of divisions in a business decision making problem based on a multi-person aggregation process.

References

Merigó, J.M., & Gil-Lafuente, A.M. (2009). The induced generalized OWA operator. Information Sciences, 179, 729–741.

Merigó, J.M., Laengle, S., & Yager, R.R. (2016). The ordered weighted average division, INFORMS 2016, Nashville, USA.

Yager, R.R. (1988). On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Transactions on Systems, Man and Cybernetics B, 18, 183–190.

Yager, R.R., & Filev, D.P. (1999). Induced ordered weighted averaging operators. IEEE Transactions on Systems, Man and Cybernetics B, 29, 141–150.

WA 1.3

The induced heavy moving average volatility

Ernesto Leon-Castro (Universidad de Occidente, Mexico), Ezequiel Aviles-Ochoa (Universidad de Occidente, Mexico), Jose Maria Merigo-Lindahl (Universidad de Chile, Chile)
Corresponding Author(s): Ernesto Leon-Castro (ernesto134@hotmail.com)

Abstract

Volatility is a basic concept in economics for measuring the variance of some variables like exchange rate, stock prices and some other. To calculate the volatility, the coefficient of variation is used in historical data (Garman & Klass, 1980; Minton & Schrand, 1999). It is important to note, that volatility is not only influenced by the historical data, but can also be influenced by some macroeconomics variables like GDP, interest rate, foreign reserves and others (Grossmann *et al*., 2014; Rabbani *et al*., 2017), but

also there are some other information that can be added to the results, such as, knowledge of the decision maker about the future scenarios that it is important to add when there is uncertainty in the problem (Yager, 2006).

One way to add information to the volatility formula, is changing the usual average by adding weights and other tools. In this sense, we can use the ordered weighted average (OWA) operator, developed by Yager (1988) to generate new scenarios between the minimum and the maximum operator.

The aim of this abstract is to analyze the use of the OWA operator and some of its extensions in the volatility formula. The main advantage of doing this is that it is possible to generate new scenarios. In this sense, we introduce new concepts of volatility like the OWMA-volatility, IOWMA volatility and HOWMA-volatility. These definitions are as follows

Definition 1. An OWMA-Volatility operator of dimension m is a mapping $OWMA - Volatility: R^m \to R$ that has an associated weighting vector W of dimension m with $W = \sum_{j=1+t}^{m+t} w_j = 1$ and $w_j \in [0,1]$, such that

$$OWMA - Volatility\ (a_{1+t}, a_{2+t}, \dots, a_{m+t}) = \frac{\sigma - OWMA}{\mu - OWMA}, \quad (1)$$

where $\sigma - OWMA$ is the OWMA standard deviation, $\mu - OWMA$ is the OWMA average.

Definition 2. An IOWMA-Volatility of dimension m is a mapping $IOWMA - Volatility: R^M \times R^M \to R$ that has an associated weighting vector W of dimension m with $W = \sum_{j=1+t}^{m+t} w_j = 1$ and $w_j \in [0,1]$, such that

$$IOWMA - Volatility(\langle u_{1+t}, a_{1+t}\rangle, \dots, \langle u_{m+t}, a_{m+t}\rangle) = \frac{\sigma - IOWMA}{\mu - IOWMA} \quad (2)$$

where $\sigma - IOWA$ is IOWMA standard deviation and $\mu - IOWMA$ is IOWMA average.

Definition 3. A HOWMA-Volatility is defined as a sequence given $\{a_i\}_{i=1}^{N}$, where you get a new sequence $\{s_i\}_{i=1}^{N-n+1}$ which is associated with a weight vector w with $w_j \in [0,1]$ and $1 \leq \sum_{j=1}^{n} w_j \leq n$, so that

$$HOWMA - Volatility(s_i) = \frac{\sigma - HOWMA}{\mu - HOWMA}, \quad (3)$$

where $\sigma - HOWMA$ is HOWMA standard deviation and $\mu - HOWMA$ is HOWMA average.

An application of this new formulas in foreign exchange market is also developed. We use exchange rate USD/MXN, EUR/MXN and EUR/USD information from 2015-2016 to forecast the volatility for all the months for the year 2016. Finally, it is important to take in account that with all these new scenarios it is possible to increase the knowledge of the financial market and explain in a better way how they will be in the future.

References

Garman, M. B., & Klass, M. J. (1980). On the estimation of security price volatilities from historical data. Journal of Business, 53(1), 67-78.

Grossmann, A., Love, I., & Orlov, A. G. (2014). The dynamics of exchange rate volatility: A panel VAR approach. Journal of International Financial Markets, Institutions and Money, 33, 1-27.

Minton, B. A., & Schrand, C. (1999). The impact of cash flow volatility on discretionary investment and the costs of debt and equity financing. Journal of Financial Economics, 54(3), 423-460.

Rabbani, A. G., Grable, J. E., Heo, W., Nobre, L., & Kuzniak, S. (2017). Stock market volatility and changes in financial risk tolerance during the great recession. Journal of Financial Counseling and Planning, 28(1), 140-154.

Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Transactions on Systems, Man, and Cybernetics, 18(1), 183-190.

Yager, R. R. (2006). Generalizing variance to allow the inclusion of decision attitude in decision making under uncertainty. International Journal of Approximate Reasoning, 42, 137–158.

WA 1.4

Formulating the Weighted Average – Experton

Salvador Linares-Mustarós (University of Girona, Spain), Dolors Corominas-Coll (University of Girona, Spain), Joan Carles Ferrer-Comalat (University of Girona, Spain), Jose Merigo (Universidad de Chile, Chile)

Corresponding Author(s): Salvador Linares-Mustarós (salvador.linares@udg.edu), Jose Merigo (jmerigo@fen.uchile.cl)

Abstract

This work presents a formulation of a new data-fusion mathematical object. The new object is constructed by extending the weighted arithmetic mean operator in the process of creating an experton (Kaufmann, 1988).

The main advantage of this approach is that it can represent an attitudinal character of the decision maker in the construction of the experton. Therefore, this approach represents a new method for dealing with multi-person problems by using data elements with the idea that some data elements contribute more than do others elements (Merigó et al. 2014). The work presents different practical examples and software for the calculation of this type of expertons.

Finally, the work ends with an application in business decision making regarding the calculation of the expected benefits (Linares-Mustarós et al. 2015).

References

Kaufmann, A. (1988). Theory of expertons and fuzzy logic. Fuzzy Sets and Systems, 28, 295–304.

Linares-Mustarós, S., Merigó, J.M. & Ferrer-Comalat, J.C. (2015). Processing extreme values in sales forecasting, Cybernetics and Systems, 46 (3-4), 207-229.

Merigó, J.M., Casanovas, M. & Yang, J.B. (2014). Group decision making with expertons and uncertain generalized probabilistic weighted aggregation operators, European Journal of Operational Research, 235, 215-224.

Sessions WB 1

WB 1.1

Spectral Mixture Kernels for Multi-Output Gaussian Processes

Gabriel Parra (Universidad de Chile, Chile), Felipe Tobar (Universidad de Chile, Chile)

Corresponding Author(s): Gabriel Parra (gparra@dim.uchile.cl), Felipe Tobar (ftobar@dim.uchile.cl)

WB 1.2

Reproducing kernel Hilbert space approach to Stochastic Frontier Analysis (SFA)

Carlos Felipe Valencia (University of los Andes, Colombia), Daniel Arocha (University of los Andes, Colombia), Ader Villar (University of los Andes, Colombia)
Corresponding Author(s): Carlos Felipe Valencia (cf.valencia@uniandes.edu.co)

Abstract

Production frontier is defined as the maximal output that can be obtained for a particular product given a set of production factors (inputs). Productive units (firms) are efficient if given their level of input can produce the output determined by this frontier. Historically, there are two main methodologies for measuring productivity and technical inefficiency: (i) deterministic non-parametric methods (e.g. DEA and FDH) and (ii) statistical estimation of stochastic models (e.g. SFA). The first set of methodologies do not suggest a statistical model for the data generation process, which translates to an approximation process without statistical inference. However, the resulting efficient frontiers are very flexible being able to capture the patterns in the data of inputs and outputs. On the other hand, the latter statistical methodologies such as SFA assumes that data points are generated according to statistical model. This allows to estimate the fundamental parameters that define the frontier and measure efficiency. The price to pay, however, is that the specification of the model is not exact given that the frontier function parametric space is not flexible enough (Parmeter and Kumbhakar, 2014).

In this study we propose and evaluate a new estimator for the stochastic frontier in a SFA framework using a penalized likelihood approach on a reproducing kernel Hilbert space. The frontier function itself is estimated non-parametrically to gain flexibility and better specification. The error term is defined as the convolution of two independent variables: the measurement error and the inefficiency stochastic sampling variation. Those errors are modeled as parametric random variables (normal and half normal respectively). The resulting statistical model is semiparametric in the sense that the frontier function is not constrained in a finite dimensional space, but given the frontier the model is represented on the distribution parameters of the error terms.

Computational implementation: The production frontier function is assumed to be in a (infinite dimensional) reproducing kernel Hilbert space with associated kernel function K. The estimator of the frontier is defined as the minimizer the negative log-likelihood in this space. With the use of the kernel trick, the problem is computable with (n*p + 2) decision variables, where n is the sample size and p the number of inputs. The non-linear optimization, however, is computationally demanding to solve the problem efficiently in real applications. We propose an approximation using pseudo-splines by representing the frontier on a truncated projection using the Demmler-Reinsch functional basis. To account for several input variables, we model an additive frontier that is estimated using the back-fitting algorithm.

Theoretical properties: The frontier estimator achieves minimax optimal rate of convergence, as usual on these non-parametric smoothers. In addition, the two parameter estimators that specified the errors distributions are semi-parametrically efficient and asymptotically normal.

Simulation study and general results: We perform a simulation study and show the benefits of the proposed estimator. In general, the MSE are smaller than the ones produced by alternative semi-parametric methods (e.g. Fan et al., 1996 and Martins-Filho and Yao, 2015). Computationally, the Demmler-Reinsch basis truncation seems to produce fast solutions with no compromise on the estimations performance.

References

1. Fan, Y., Li, Q., & Weersink, A. (1996). Semiparametric Estimation of Stochastic Production Frontier Model. Journal of Business & Economic Statistics, 14, 460 - 468.

2. Martins-Filho, C., & Yao, F. (2015). Nonparametric Stochastic Frontier Estimation Via Profile Likelihood. Econometric Reviews, 34(4), 413-451.

3. Parmeter, C. F., and Kumbhakar, S. C. (2014). Efficiency analysis: a primer on recent advances. Foundations and Trends® in Econometrics, 7(3–4), 191-385.

WB 1.3

An Evaluation of Missing Data Handling Mechanisms of Evidential Reasoning Rule for Data Classification

Shao Cong Lim (Singapore Armed Forces (SAF), Singapore), Dong-Ling Xu (Manchester Business School, University of Manchester, United Kingdom), Jian-Bo Yang (The University of Manchester, United Kingdom)
Corresponding Author(s): Dong-Ling Xu (L.Xu@mbs.ac.uk), Jian-Bo Yang (Jian-Bo.Yang@manchester.ac.uk)

Abstract

The Evidential Reasoning (ER) rule (Yang and Xu 2013) has recently been applied to data classification. The literature on the applications has been focused on comparing classification accuracy of the ER rule against other well-regarded classifiers (e.g. Xu et al. 2017). However, discussions of the inherent features of the ER classifier including its strengths and weaknesses as compared to other classifiers are noticeably absent.

In this paper, the performance of the ER classifier in the context of supervised learning in the presence of missing data are evaluated and discussed through the use of numerical studies to evaluate the prediction accuracy of the ER classifier against naïve Bayes and C5.0 decision-tree algorithms. These two algorithms are deliberately chosen as a benchmark for their resilience to missing data.

Although it is known that the ER classifier is able to handle missing data, there exist different interpretations and implementations. Two of such implementations named as ER0 and ER1 in this paper are discussed. In ER0, missing data are ignored, which is equivalent to pairwise deletion (Marsh 1998). In ER1, missing values of a variable are treated as a special value (unknown) of the variable (Yang and Xu 2014).

The numerical studies use 11 data sets from the UCI Machine Learning Repository. Missing data corresponding to two of the missing data mechanisms (Rubin 1976), namely missing completely at random (MCAR) and missing at random under certain conditions (MAR) with different levels of missing volumes are introduced artificially into multiple copies of the 11 data sets.

The results show that the ER0 and ER1 compare competitively against the two benchmark algorithms in terms of classification accuracy and resilience to the presence of missing data. Here the resilience is measured by the loss of classification accuracy at the presence of different levels of missing data volumes. The competitive prediction accuracy and prediction loss exhibited by both ER0 and ER1 show that the internal mechanism for missing data handling is effective.

The majority of the results suggest the absence of any differences between ER0 and ER1 (57%). However, ER0 was found to give superior prediction accuracy to ER1 in 37% of the results. Both ER0 and ER1 classifiers were also found to be insensitive to the missing data mechanisms and various levels of missing data volumes included in this study.

References

Marsh, H. W., 1998. Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. Structural Equation Modeling: A Multidisciplinary Journal, 5, p.22-36.

Rubin, D. B., 1976. Inference and missing data. Biometrika, 63(3), p.581.

Xu, X. et al., 2017. Data classification using evidential reasoning rule. Knowledge-Based Systems, 116, p.144–151.

Yang, J. B. & Xu, D. L., 2014. A Study on generalising Bayesian inference to evidential reasoning. In Belief Functions: Theory and Applications SE - 20. pp. 180–189.

Yang, J. B. & Xu, D. L., 2013. Evidential reasoning rule for evidence combination. Artificial Intelligence, 205, pp.1–29.

WB 1.4

Reparameterizing the Birkhoff Polytope for Variational Permutation Inference

Scott Linderman (Columbia University, United States), Gonzalo Mena (Columbia University, United States), John Cunningham (Columbia University, United States), Liam Paninski (Columbia University, United States), Hal Cooper (Columbia University, United States)

Corresponding Author(s): Gonzalo Mena(gem2131@columbia.edu)

Sessions WA 2

WA 2.1

Combining Support Vector Machine classification and profit measures in credit scoring

Sebastián Maldonado (Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Chile), Cristian Bravo (University of Southampton, United Kingdom), Julio López (Universidad Diego Portales, Chile), Juan Pérez (Universidad de Los Andes, Chile)

Corresponding Author(s): Sebastián Maldonado (smaldonado@uandes.cl)

Abstract

Support Vector Machine (SVM) is an effective supervised method with appealing advantages such as adequate generalization to new instances thanks to structural risk minimization principle, a single global optimum, and a representation that depends on few data points (Vapnik, 1998).

Feature selection is an important Analytics task, especially in domains where interpretability is a key issue for a classifier. Selecting the relevant attributes improves the model's generalization ability and reduces the risk of overfitting, improving predictive performance (Guyon et al., 2006). In addition to interpretability and predictive power, feature selection reduces the variable collection costs because it allows the elimination of irrelevant variables from expensive data sources (Maldonado et al., 2017).

Unlike methods such as decision trees, SVM cannot perform feature selection automatically. Modifications to the original quadratic optimization problem can be done in order to identify the relevant variables together with the classifier construction. Unfortunately, most strategies lead to a less efficient optimization strategy, such as nonconvex or integer optimization. (Bradley and Mangasarian, 1998).

The choice of the best classification approach is usually made using traditional, statistically grounded techniques. However, a new research line proposes using business-oriented measures for model selection and validation (Verbeke et al., 2012). This is particularly important in credit scoring since the decision whether to accept an applicant or not is made based on financial criteria.

In this work, we introduce a novel SVM method for profit-based classification and feature selection. The idea is to balance the profit of granting credit with the variable acquisition costs in order to construct the most effective classifier in terms of predictive performance but using few data sources. A group penalty function, namely the L-infinity norm, is included to penalize the use of groups of related features. The advantage of this strategy is that the convexity of SVM is not affected, leading to an efficient optimization process.

A case-study of a Chilean bank is presented. Credits are granted to microentrepreneurs based on information from five different data sources. Our proposal concludes that best solutions in terms of profit are achieved using one or two cheap data sources, without the need of expensive interviews. Additionally, important managerial insights are gained into the application

thanks to the identification of the relevant variables.

References

- Bradley, P., Mangasarian, O., 1998. Feature selection via concave minimization and support vector machines. In: Machine Learning proceedings of the fifteenth International Conference (ICML'98) 82-90, San Francisco, California, Morgan Kaufmann.

- Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. A., 2006. Feature extraction, foundations and applications. Springer, Berlin.

- Maldonado, S., Pérez, J., Bravo, C., 2017. Cost-based feature selection for svm classification - an application in credit scoring. European Journal of Operational Research 261 (2), 656–665.

- Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B., 2012. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research 218 (1), 211–229.

- V. Vapnik, 1998. Statistical Learning Theory. John Wiley and Sons.

WA 2.2

Thirty years of the Journal of Business & Industrial Marketing: a bibliometric analysis

Leslier Valenzuela (University of Chile, Chile), José M. Merigó (University of Chile, Chile), Wesley Johnston (Georgia State University, United States), Carolina Nicolas (Universidad Santo Tomas, Chile), Jorge F Jaramillo (University of Texas at Arlington, United States)

Corresponding Author(s): Leslier Valenzuela (lvalenzu@fen.uchile.cl), José M. Merigó (jmerigo@fen.uchile.cl), Carolina Nicolas (cnicolas@santotomas.cl)

Abstract

In commemoration of the 30th anniversary of the Journal of Business & Industrial Marketing this study presents an overview of the journal through a bibliometric analysis of scientific content during the period of 1986-2015. The analysis is concentrated on the most cited papers and authors, the h-index (Hirsch, 2005), and publications per year, among others. The article begins with a qualitative introduction referring to the emergence of the magazine, its origins, editorial and positioning followed by bibliometric quantitative analyses. The study also investigates the distribution of annual publications, citations and keywords as well as authorship and institutions. Additionally, the work develops a graphical visualization of the bibliographic material by using the visualization of similarities (VOS) viewer software (Van Eck and Waltman, 2010).

Research findings reveal that the journal covers a wide variety of topics with business-to-business marketing, relationship marketing, buyer-seller relationships, innovation, and industrial marketing as the most representative. The international scope of the journal is also highlighted with authorship from countries distributed all over the world. A significant portion of the contributions to the journal comes from top tier universities hosted in the United States and Europe. Note that the analyses are limited to data derived from the Scopus database. Therefore, the study excludes publications not contained in Scopus and may underestimate the full impact of JBIM to the marketing field. It is part of the trend that several journals (Journal of Marketing, Journal of Public Policy & Marketing, Journal of Marketing Research, and Journal of Business Research) made special sections to show progress and contribution of these journals to scientific research (Merigó et al. 2015; Sprott and Miyazaki, 2002).

Finally, note that the full version of this paper has been recently published in the Journal of Business & Industrial Marketing (Valenzuela et al. 2017).

References

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102, 16569–16572.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84, 523–538.

WA 2.3

An overview of the most cited papers in computer science

Gustavo Zurita (University of Chile, Chile), José M. Merigó (University of Chile, Chile), Valeria Lobos-Ossandón (University of Chile, Chile), Carles Mulet-Forteza (University of Chile, Chile)
Corresponding Author(s): José M. Merigó (jmerigo@fen.uchile.cl)

Abstract
Computer Science arises as one of the most influential sciences of the last decades. Its influence reflects in many areas and in different ways. This research area is in a constant development, becoming more and more specialized. Traditional views of Computer Science are not enough to describe the topics that are now developing. Thus, this leaves us to question us what are the hot topics in Computer Science today. This work analyzes the most cited papers of Computer Science over the last 25 years (1990-2014), according to Web of Science (WoS) database. WoS categorizes Computer Science into seven categories: Artificial Intelligence, Cybernetics, Hardware and Architecture, Information Systems, Interdisciplinary Applications, Software Engineering and Theory and Methods. Are they enough to describe the most popular subjects in Computer Science? Results show us that there are many topics that are not fully represented by the categories mentioned above and the need for making a distinction between topics. Hot topics in Computer Science today includes Machine Learning, Social Networking and Biochemistry and Technology. How do we categorize these topics in the traditional ones? This paper aims to respond to the questions mentioned above and to propose new categories for Computer Science journals.

WA 2.4

Academic research on support vector machines: A bibliometric overview

Jaime Miranda (Facultad de Economía y Negocios, Universidad de Chile, Chile), José M. Merigó (University of Chile, Chile), Sebastian Maldonado (Universidad de los Andes, Chile)
Corresponding Author(s): Jaime Miranda (jmirandap@fen.uchile.cl), José M. Merigó (jmerigo@fen.uchile.cl)

Abstract
Bibliometrics is the research field that analyzes the bibliographic information by using quantitative techniques (Pritchard, 1969). It is very useful for providing a general picture of a research field (Merigó et al. 2015). This study presents a bibliometric overview of academic research in support vector machines. The search uses the Web of Science Core Collection database identifying the most cited papers, the most productive and influential authors, institutions and countries, and the publication and citation structure. The paper uses several bibliometric indicators including the number of publications and citations, the cites per paper, the h-index (Hirsch, 2005) and citation thresholds.

Additionally, the work also develops a graphical analysis by using the VOS viewer software (Van Eck and Waltman, 2010). In this part of the study, the work considers bibliographic coupling, co-citation and citation analysis, co-authorship and co-occurrence of author keywords (Laengle et al. 2017). Recall that bibliographic coupling measures the documents that cite the same third document. Co-citation occurs when two documents receive a citation from the same third work. Citation analysis measures how the documents cite each other and co-authorship focuses on the co-authors of the documents. Co-occurrence of author keywords analyses the most frequent keywords and those that appear often in the same papers.

References

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102, 16569–16572.

Laengle, S., Merigó, J.M., Miranda, J., Slowinski, R., Bomze, I., Borgonovo, E., Dyson, R. G., Oliveira, J.F., Teunter, R. (2017). Forty years of the European Journal of Operational Research: A bibliometric overview. European Journal of Operational Research, 262 (2017) 803–816.

Merigó, J.M., Gil-Lafuente, A.M., & Yager, R.R. (2015b). An overview of fuzzy research with bibliometric indicators. Applied Soft Computing, 27, 420–433.

Pritchard, A. (1969). Statistical bibliography or bibliometrics? Journal of Documentation, 25, 348–349.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84, 523–538.

Sessions WB 2

WB 2.1

Accounting Fraud Detection Through Forensic Analytics

Maria Jofre (The University of Sydney, Australia)
Corresponding Author(s): Maria Jofre (maria.jofre@sydney.edu.au)

Abstract

Accounting Fraud (AF) is one of the most harmful financial crimes as it often results in massive corporate collapses, commonly silenced by powerful high-status executives and managers. AF represents a significant threat to the financial system stability due to the resulting diminishing of the market confidence and trust of regulatory authorities. The catastrophic consequences of accounting fraud expose how vulnerable and unprotected the community is in regards to this matter, since most damage is inflicted to investors, employees, customers and government.

Accounting fraud is defined as the calculated misrepresentation of the financial statement information disclosed by a company in order to mislead stakeholders regarding the firm's true financial position. Different fraudulent tricks can be used to commit AF, either direct manipulation of financial items or creative methods of accounting, hence the need for non-static regulatory interventions that take into account different fraudulent patterns.

Accordingly, this study aims to identify signs of accounting fraud occurrence to be used to, first, identify companies that are more likely to be manipulating financial statement reports, and second, assist the task of examination within the riskier firms by evaluating relevant financial red-flags, as to efficiently recognise irregular accounting malpractices. To achieve this, a thorough forensic data analytic approach is proposed that includes all pertinent steps of a data-driven methodology.

First, data collection and preparation is required to present pertinent information related to fraud offences and financial statements. The compiled sample of known fraudulent companies is identified considering all Accounting Series Releases and Accounting and Auditing Enforcement Releases issued by the U.S. Securities and Exchange Commission between 1990 and 2012, procedure that resulted in 1,594 fraud-year observations.

Then, an in-depth financial ratio analysis is performed in order to evaluate publicly available financial statement data and to preserve only meaningful predictors of accounting fraud. In particular, two commonly used statistical approaches are proposed to assess significant differences between corrupted and genuine reports as well as to identify associations between the considered ratios. Results obtained from non-parametric hypothesis testing and correlation analysis support the selection of a smaller subset of explanatory variables, later reinforced by the implementation of a complete subset logistic regression methodology.

Finally, statistical modelling of fraudulent and non-fraudulent instances is performed by implementing several machine learning methods. Classical classifiers are considered first as benchmark frameworks, including logistic regression and discriminant

analysis. More complex techniques are implemented later based on decision trees bagging and boosting, including bagged trees, AdaBoost and random forests.

Generally, out-of-sample results suggest there is a great potential in detecting falsified accounting records through statistical modelling and analysis of publicly available accounting information. It has been shown good performance of classic models used as benchmark and better performance of more advanced methods.

WB 2.2

Theoretical and practical aspects of various measures of portfolio diversification

Tomáš Tichý (VSB-TU Ostrava, Czech Republic)
Corresponding Author(s): Tomáš Tichý (tomas.tichy@vsb.cz)

Abstract
It is well known that the returns of financial assets generally do not follow the Gaussian law, which also implies that the Pearson measure of linear correlation is not suitable to correctly describe dependencies among random variables. We first focus on possible usage of different correlation measures in portfolio problems. We characterize especially semidefinite positive correlation measures consistent with the choices of risk-averse investors. Moreover, we propose a new approach to portfolio selection problem, which optimizes the correlation between the portfolio and one or two market benchmarks. We also discuss why one should use correlation measures to reduce the dimensionality of large scale portfolio problems and study an impact on such decisions. Next, a so called stochastic alarm, which should allow us to predict market periods of systemic risk and price drawdowns, is utilized. Finally, through an empirical analysis using US data, we show the impact of different correlation measures on portfolio selection problems and on dimensionality reduction problems.

WB 2.3

Animal movement in Mato Grosso do Sul and its implications for economic impacts of potential outbreaks of foot-and-mouth disease

Tais Cristina de Menezes (University of São Paulo, Brazil), Silvia Helena Galvão de Miranda (University of São Paulo, Brazil), Ivette Luna (Universidade Estadual de Campinas, Brazil)
Corresponding Author(s): Tais Cristina de Menezes (taismenezes@usp.br)

Abstract
Brazil has the second largest cattle herd (22.5%) and is the second largest beef producer in the world. Since 2004, Brazil hanks first in the ranking of the world's largest beef exporters. Due to the importance of the livestock sector in the Brazilian economy, for years Brazil has promoted control and eradication of foot-and-mouth disease (FMD), always pursuing the maintenance of its FMD disease-free status with vaccination. The state of Mato Grosso do Sul (MS), due to its location – borders with Paraguay, Bolivia and five other Brazilian states – and its significant weight in the national herd (10% of the Brazilian herd), assumes a prominent role in maintaining this status. Since sanitary barriers are currently some of the main impediments to international trade, knowledge about the dynamics of FMD spread in MS in an eventual outbreak and the identification of the central counties on the livestock production is a subject of national relevance. In this sense, this work analyzes the economic dynamics of the livestock flows in MS and identifies its geographic distribution within the state. Areas of greatest risk of occurrence and greater economic impacts are also identified. The analysis is based on socioeconomic networks. Quarterly networks were built using data composed by 432,457 State Animal Transit Guides registered in 2014, which provide the movement of more than 12 million animals within the state and the supply of more than 400,000 animals to other Brazilian states during the year. Data processing and networks construction were performed using softwares R and Pajek. The resulting networks presented a

strongly connected nature, with 79 counties and 20 others states as nodes, and with links representing the flow of animals transported between counties within the state or even to the other 20 states. Centrality measures (degree, closeness and betweenness centrality) allowed the identification of the counties with the highest probability of incidence of FMD. The counties of Campo Grande and Corumbá were the most central ones is this process, playing both the roles of livestock suppliers and receivers, further increasing the risk faced by them in the case of a potential outbreak. For comparison purposes, the dynamic of the diffusion of FMD based on the networks considered both central and peripheral counties as possible starting points of the diffusion process. In both cases, we considered that the direct neighbors of an infected county were infected within a first stage and that the contagion process continued until every county was infected. Therefore, taking Corumbá as the outbreak focus of FMD (similar results were observed for other central counties), the simulation showed an infection rate of 50% within the first stage and an infection rate of 100% in three stages on average. We identify in this way a potentially very fast diffusion process with logistic diffusion curves (S-shape) and steeper than the observed when a peripheral county was considered as starting point of the contagion process. Due to the supply of animals to other states, the spread also occurs at the national level, raising the economic impacts of the disease. Hence, based on this preliminary analysis, we conclude that because of its accelerated dynamics, in face of a potential FMD outbreak in MS with origin in central counties, the diffusion process would imply strong direct and indirect impacts for Brazilian livestock and other economic sectors as well. Also, the infection rate at each stage is higher in the third quarter, when fattening prevails within the livestock cycle also identified in our analysis. The next phase of this work will improve the analysis with simulations from daily networks and extend it to the whole country.

WB 2.4

Patterns and Insights of an Asset Order Book

Andrew Day (University of Western Ontario, Canada), Matt Davison (University of Western Ontario, Canada)
Corresponding Author(s): Andrew Day (aday46@uwo.ca)

Abstract

Today many brokerage firms use computer algorithms to make trade decisions, submit orders, and manage orders after submission. This is known as algorithmic (or algo) trading. The reason for using computers is to maximize execution speed, minimize the cost, market impact and risk associated with trading large volumes of securities.

Participants provide liquidity to the market by placing buy or sell orders to an exchange. These show the intention to buy or sell a given amount of a security for a specific price. The buy order with the highest price is called the best bid, while the sell order with the lowest price is called the best offer. The difference between the best bid and best offer is the spread. These buy and sell orders accumulate in what is called the order book until they find a counter-party for execution or are canceled. All participants can also issue market orders to buy or sell at the best available prices and is immediately executed. This removes liquidity. This tends to be done on a `first come first serve' basis.

We were given a data set of all buy/sell orders, market orders, and cancellations made for hundreds of stocks between April 17 and 28, 2017 on the Toronto stock exchange. This amounted to roughly 35 GB of data which had to be separated by stock ticker and day in order to construct the order book at a given time

First we analyse the dynamics of the order book data using functional data analysis. The key assumption is that the volume of stock offered is a smooth process of the price, but you make noisy observations of it. The goal is to then use our data to approximate the functional equivalent of a mean and variance for this random distribution of functions. The methodology shares many similarities to independent component analysis. The results are then interpreted as how the order book responds, on average, to an interaction with a broker for a given day.

Second we model the dynamics of one of the predictors, the order imbalance ratio, used by our classifiers. This is done by matching the interday data of a given stock to a Vasicek model which we can solve exactly. The results capture the qualitative aspects of the interday order book dynamics. We then present the results of order book classification following [1].

In this talk we describe approaches taken to reconstruct the order book for individual stocks from data provided by the TMX group, the company that operates the Toronto Stock Exchange. We describe modelling and statistical methods to study the dynamics of the resulting order books, and approaches taken to classify them. We also investigate applying these results to study the asset order book in an optimal control setting - can one use information gathered from the order book to make better decisions?

References

[1] Sirignano, J. (2016). Deep Learning for Limit Order Books. arXiv:1601.01987.

Sessions WA 3

WA 3.1

Churn Prediction in Telco using Adapted node2vec on CDR graphs enriched with RFM information

Sandra Mitrovic (Katholieke Universiteit Leuven, Belgium), Bart Baesens (Katholieke Universiteit Leuven, Belgium), Wilfried Lemahieu (Katholieke Universiteit Leuven, Belgium), Jochen De Weerdt (Katholieke Universiteit Leuven, Belgium)
Corresponding Author(s): Sandra Mitrovic (sandra.mitrovic@kuleuven.be)

Abstract

A Call Detailed Record (CDR) in Telco is a structured log containing information about customer calls, e.g. in the form of: caller, callee, call date/time, call duration. This format allows for transformation into networked data (by representing every customer by a node and connecting those nodes/customers among which calls were recorded). Given the strong uptake of social network analytics in recent years, using CDRs i.e. networks generated from CDRs, has become a predominant strategy for solving various telco-related data mining problems, including churn prediction. While the construction of the network itself is pretty standardized in the literature (as already explained), the way of featurizing, that is, deriving informative features from these networks, remains highly versatile and non-systematic. The reason for this is twofold: (1) the complex structure of CDR networks and (2) the absence of an encompassing methodology for tackling the feature extraction. The complexity of featurizing CDR networks is caused by the fact that two types of information are conveyed: first, the structural connections between nodes which can be derived from underlying graph topology (we refer to these as structural features) and second, the characteristics of customer interactions (calls) provided in the CDR, expressed by measures like number of calls, call duration (we refer to these as interaction features). In the current literature, the former are mainly operationalized by different centrality measures, while for the latter the RFM (Recency-Frequency-Monetary) model is frequently used. The main problem, however, is the fact that most of the centrality measures (and especially for large graphs such as those derived from CDR data) become computationally intractable. This is the reason why many studies either do not include structural features at all, or simply restrict themselves to only degree-related measures. On the other hand, RFM features are fairly simple to compute but are usually handcrafted based on expert knowledge due to the many possible variations.

In order to fully exploit the potential of the CDR networks and to overcome the identified problems, in this work, we propose a novel approach which can be perceived as a four-layered approach. In the first layer, we propose four different RFM operationalizations. Based on these, we propose novel network constructions where we enrich original CDR graphs with RFM

information. We devise two different directions for network extensions. One is based on retaining the original topology while using RFM information to characterize the weights of the edges in the original graph. The other direction consists on extending the original topology by adding artificial Ri, Fi, Mi nodes based on different ways of categorizing node-level RFM information. These enriched networks represent the second layer of our approach. In the third layer, we use the recently proposed deep learning architecture approach node2vec to learn unsupervised node representations from RFM-enriched networks. We perform necessary adaptations of node2vec to make it scalable for our RFM-enriched CDR graphs. In the final, fourth layer, we use the learnt representations as input features for different predictive models and evaluate results using well-known measures: AUC and lift.

Obtained results on one postpaid and one prepaid dataset demonstrate that our method outperforms the classical approach based on RFM features, both in terms of AUC and lift.

As such, the contribution of this work is three-fold. First, we design novel network constructions, which enable the integration of structural and behavioral information and therefore allow for exploiting the full potential of CDR data. Second, by adapting and applying the unsupervised learning approach, we outsmart current studies, which mostly derive features in an ad-hoc manner based on handcrafting. Third, we prove the benefits of our approach in terms of predictive performance.

WA 3.2

Privacy preserving Customer Churn Models using Support Vector Machines

Abelino Jimenez (Carnegie Mellon University, United States), Bhiksha Raj (Carnegie Mellon University, United States)
Corresponding Author(s): Abelino Jimenez (abelinoj@andrew.cmu.edu)

Abstract
Advances in cloud computing and machine learning have allowed to spread the use of client-server models, where a client provides its data and a server has a machine learning model to evaluate the client's data. However, for various privacy considerations, the client may not want to reveal the required information to the server or let the server know the outcome of the model. This is a common situation when the kind of information involved corresponds to financial data. For instance, a Bank A would like to apply a machine learning model provided by the Company B for getting a churn prediction of its customers. Despite the valuable results given by the model, the Bank is concerned about revealing customers' information to this external company; the data can be compromised and misused raising strong legal problems.

In order to protect the client's information, we may consider encrypting the data prior to sending it to the server. However, if any form of computation must be performed on the data, they must be decrypted by the server. This implies that at the very least, the server has access to the raw data to evaluate the machine learning model, as does any hacker or other malicious entity who has managed to gain access to the server while it is performing computations on the decrypted data.

Another solution is to send the model directly to the client and allow them to perform the computation locally, avoiding the need for the client to send its data to an external party. Nevertheless, there are many reasons from the server side for not sharing its model. The most common ones are related to intellectual property issues; model construction represents value for the provider and part of its worth comes from its secrecy. Hence, we have a conflict between two parties; in one side the client, which has data but does not want to reveal it to get a desirable outcome, and in the other side a server, which has a model and offers the service of providing its evaluation, without revealing the model.

Several attempts to resolve this problem can be found in the literature, with tools such as homomorphic encryption and secure

multiparty computation schemes being the most popular approaches. They have important theoretical guaranties but are not efficient in real contexts, and are only practical just for simple and small models.

In this paper, we propose a scheme which satisfies the privacy guarantees desired above, for the use of Support Vector Machine (SVM) models on untrusted server platforms. SVMs have been very popular in the Machine Learning with important results in different fields, such as Health and Financial systems. We consider the client poses a real valued vector, and the server has a trained SVM model. The client wants to obtain the model prediction without revealing its vector to the server side. We study the combination of homomorphic encryption techniques with hashing methods to decrease the complexity of the computations involved. We analyze both linear and non-linear SVMs discussing different approaches depending on the kernel in use.

We present experimental results for Customer Churn models showing that the computation can be done in a reasonable amount of time under the described constrains. Implications to Business Analytics and Machine Learning applied to Financial systems as well as extension to other kind of models are discussed.

WA 3.3

Churn Prediction through Customer Feedback Analytics

Carolina Martinez (Alliance Manchester Business School, United Kingdom), Babis Theodoulidis (Alliance Manchester Business School, United Kingdom), David Diaz (Universidad de Chile, Facultad de Economia y Negocios Departamento de Administracion, Chile)
Corresponding Author(s): David Diaz (ddiaz@unegocios.cl)

Abstract
The aim of this paper is to explore the determinants of switching behavior embedded into unstructured customer feedback in order to predict churn. To achieve this purpose, case study methodology is applied in the retail banking sector. Retail banking industry shows a competitive environment with high switching rates despite of it operates under a contractual setting. We used one-year direct requests and complaints as a customer feedback data. An integrative approach of frameworks was applied to deploy text mining. The determinants of switching behavior was extracted through a generic theory developed by Keaveney in 1995 within the activities, resources and contexts involved, i.e. ARC model in order to resolve the uni-linear perspective of Keaveney model. Based on the text mining results, classification algorithms, such as, logistic regressions, decision trees and support vector classifiers are later used to predict churn.
The main findings show that the customers tend to articulate determinants of switching behavior within their requests and complaints. These determinants might be used as predictors of churn to generate actionable information to develop customer engagement strategies in real time. In addition, the extended ARC model provided fruitful insights to comprehend the co-creation of value process in this particular domain.
Customer feedback management has become a critical success factor to improve customer engagement in service businesses today. In a network economy context, in which all participants are connected, customer feedback must be understood as one particular type of interaction triggered by the service experience. At the same time, the dialogical interaction between company and customer is one of the main drivers of the value co-creation process. In this sense, there is a consensus among service researchers that the co-creation of value involves complexities that are not entirely well understood, which might be resolved if companies generate a deeper understanding of how customers articulate their own service experiences. In parallel, the current social media rising as a new platform of interaction has facilitated the development of large amounts of textual and unstructured customer feedback data. Hence, the main challenge for companies not only rely on a good data-driven consumer understanding, but also involves the development of big data and text analytics approaches in order to improve customer

engagement.
This research contains theoretical, methodological and managerial implications. Theoretically contributes to expand the ARC model in relation to switching behavior to deal with churn. In terms of methodology, the application of text analytic techniques to automate the analysis of textual and dynamical data contributes to enhance the service experience. In the development of managerial implications this research contributes to operationalize customer feedback information to optimize service performance in the banking sector.

Finally, this research restates that the model of Keaveney is appropriate for explaining retail banking churn behavior. The development of an extended churn framework contributes to adapting the Keaveney model in online feedback context. According to previous literature, between 6 and 8% of respondents indicated that they had spoken to the bank staff about their churn intentions before exit. This paper shows that customers refer to churn intentions 12.5% before exit. The online feedback context can explain this rise.

* ARC model is developed in the article "Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach" (Ordenes, Theodoulidis, Burton, Gruber & Zaki, 2014)

WA 3.4

Effect of Sample Representativeness in Multivariate Symmetrical Uncertainty for Categorical Attributes

Gustavo Sosa-Cabrera (Polytechnic School, Universidad Nacional de Asuncion, Paraguay), Miguel García-Torres (Computer Science, Universidad Pablo de Olavide, Spain), Santiago Gómez-Guerrero (Polytechnic School, Universidad Nacional de Asunción, Paraguay), Christian Schaerer (CIMA, Centro de Investigación en Matemática, Paraguay), Federico Divina (Computer Science, Universidad Pablo de Olavide, Spain)
Corresponding Author(s): Santiago Gómez-Guerrero (sgomezpy@gmail.com)

Abstract
Symmetrical Uncertainty (SU) is an entropy-based non-linear correlation measure between two categorical random variables. SU is a normalization of Mutual Information to compensate overestimation in the presence of features with many values [1]. However, SU is limited to two variables at a time and it does not take into account the possible interactions among three or more features.

We propose a generalization of SU, called Multivariate Symmetrical Uncertainty (MSU), for three or more categorical features [2]. Synthetic data were generated simulating a classification problem with n input attributes under different combinations of four factors: number of features, cardinality of the features, informativeness of the features, and sample size. From experiments with these data, sample size emerges as a factor that can decrease bias and stabilize the behavior of MSU. Also, we obtain an empirical expression that relates the sample size with the cardinality of all features including the class.

While in numerical variables it is implicitly assumed that sample spread resembles the population spread, in categorical variables there is no measurable dispersion. However, the higher the variety of sample values, the better the sample entropy will estimate its real population value since more terms will be taken into the summation. This inspires the concept of "total representativeness" as a desirable property in a single-attribute sample.

For the multivariate case, the joint distribution of n categorical features with finite individual cardinalities can be seen as a multinomial variable whose cardinality is the product of the individual cardinalities. Frequency distributions of this multinomial show totally representative samples displaying histograms where no bin is 0. Samples that are not totally representative have

one or more zero frequencies, and we call them extreme samples.

Using confidence intervals, we express the condition for total representativeness with probability 1 − α. This probability is a decreasing function of m, hence it is possible to find m = m^* such that the probability of an extreme sample becomes lower than a pre-set level (say, α=0.05). Thus m^* is the smallest m that guarantees a totally representative sample with probability 1 - α, and it agrees with the empirical values found.

On datasets that have known dependencies among attributes, the MSU computed using the calculated m^* correctly detects correlations among features when the sample size gives high assurance of total representativeness. One of the tested types of dependencies, the collective correlation or interaction of several features having zero pairwise correlation, is also detected by the MSU.

This extension of the correlation concept to n dimensions is of importance by itself. Moreover, the generalization to the MSU is very significant considering the efficacy to also detect interactions between several categorical variables (as opposed to numerical variable measures that only capture pairwise correlations).

Summarizing our contributions: (1) MSU, based on information theory concepts, is studied under four factors affecting its behavior. This measure of multivariate correlation and collective interaction among categorical attributes is now completely tested and ready for use. (2) Total representativeness is introduced as a worthy condition in a sample of one categorical attribute, and extended to the multivariate case. (3) A procedure is derived to calculate the minimum sample size guaranteeing a totally representative sample with desired probability level. The empirical expression separately found to approximate m using cardinalities of features, is in agreement with the procedure.

References

[1] M. A. Hall, Correlation-based feature subset selection for machine learning, Ph.D. thesis, University of Waikato, Hamilton, New Zealand (1998).

[2] R. Arias-Michel, M. García-Torres, C. Schaerer, F. Divina, Feature selection using approximate multivariate markov blankets, in: Hybrid Artificial Intelligent Systems 11th International Conference, HAIS 2016, Seville, Spain.

Sessions WB 3

WB 3.1

The (pseudo-)social behavior of products in offline retail stores: Predicting increase in product interpurchase time

Jasmien Lismont (Katholieke Universiteit Leuven, Belgium), Sudha Ram (University of Arizona, United States), Bart Baesens (KU Leuven; University of Southampton, Belgium), Wilfried Lemahieu (Katholieke Universiteit Leuven, Belgium), Jan Vanthienen (Katholieke Universiteit Leuven, Belgium)
Corresponding Author(s): Jasmien Lismont (jasmien.lismont@kuleuven.be)

Abstract
Within marketing, the importance of the interpurchase time of products may not be underestimated. An increasing average interpurchase time per customer can forebode some issues with a particular product, and eventually even product churn. This means that a certain product will (almost) not be sold anymore. Especially in offline retail, experiencing heavy competition from online sales, this information can be used by marketing and operations experts in order to undertake further actions.

Our goal is to predict product churn, defined as a significant increase in interpurchase time. As such, we can identify those products which deserve additional marketing attention. Additionally, production and logistics can use this information to adjust their operations. We work with data from a European low-cost food and non-food retailer, from which we extract product characteristics based on one year of transaction, product, customer and payment data. The resulting feature set, based on existing literature, consists of recency, frequency and monetary (RFM) values, price details, payment details, etc.

Consecutively, we make a contribution to existing literature by applying social network analytics (SNA) techniques. However, we work with 'pseudo-social networks' of which previous work has already proven the potential. These networks are called 'pseudo-social', because they do not represent actual social behavior. Instead, we create a network of products which are connected by the customers who purchased them. This leads to a bipartite customer-product network -also referred to as a graph- with two types of nodes, namely products and customers. There are multiple ways to include SNA into your models. We will focus on featurization, which means that we will extract network features from the graph. Furthermore, we can map this bipartite network onto a unipartite graph which contains only product nodes. In total, we extract four types of network features. Firstly, we include a propagation of RFM values which is based on an adaption of Google's PageRank algorithm. Secondly, we take bipartite RFM features into account linked to the customer-product connections. Then, we focus on the unipartite network of products and extract variables related to the immediate neighborhood of products, i.e. similar products. As such, we can use similarities in customer preferences. Lastly, we calculate centrality measures. Centrality measures express how a product is situated in the graph of products and how closely it is related to other products. Network information can be valuable in certain situations. For example, imagine that customers stop buying bread at the retailer due to environment (e.g. new bakery in town) or product quality reasons. This behavior might spread through the customer network and affect other products which are often bought together with bread.

An advantage of using featurization, is that we can apply known classification techniques and –depending on the technique- analyze the effect of the product and network characteristics. Thus, by means of applying basic and advanced machine learning techniques such as logistic regression and random forests, we create product churn prediction models. For this purpose, we split our data in three samples distinctive in time. Our models are trained on five months of data and tested on the consecutive five months. Performance is measured in terms of area under the receiver operating characteristic curve (AUC), sensitivity and specificity using five-fold cross-validation. Initial results are promising, indicating a significant increase in AUC of 6% if we include network features.

WB 3.2

Factors affecting banking efficiency scores in Network SBM DEA model: Dealing with heterogeneity

Skarleth Carrales Escobedo (The University of Edinburgh, United Kingdom), Jamal Ouenniche (The University of Edinburgh, United Kingdom)
Corresponding Author(s): Skarleth Carrales Escobedo (skarleth.carrales@ed.ac.uk)

Abstract
Data Envelopment Analysis as a field has substantially evolved both methodologically and in terms of applications. So far, efficiency and productivity studies in the banking sector proved to be amongst the most popular application areas (e.g., Liu's et al., 2013). The popularity of DEA in this field, amongst others, is due to its unique features such as its non-parametric nature, it benchmarks against the best practice performers rather than the average performers, it allows one to identify targets for improvement, it does not need any functional specification of the relationship between inputs and outputs, it provides a variety of efficiency measures most suitable for a variety of applications, it provides a wide range of models to perform analyses at the aggregate level and the detailed level, on one hand, as well as models to perform static analyses and dynamic analyses.

There is evidence that Units within a group of evaluation always present heterogeneity characteristics. Several authors have investigated methods to deal with non-homogeneous DMUs that affect the efficiency scores that are not related with management inefficiency. In Data Envelopment Analysis, the homogeneity of the Decision Making Units (DMUs) is a fundamental assumption. However, the DMUs are most of the time non-homogeneous for different factors such as the difference in the availability of inputs and outputs or the difference of the environmental context where the DMUs operate.

Extreme low levels of efficiency can be an indicator of heterogeneity in the DMUs. DEA was originally developed by Charnes et al. (1978). DEA was used as a tool to measure the efficiency of a DMU as a whole unit, without considering its internal structure. Fare and Grosskopf (1996) introduced the Network DEA model, used to explore the processes inside the so-called "black box". Therefore, this model looks into the internal structure connecting each stage or process with intermediate products. In Network DEA, the stage or DMSUs are interdependent and do not share the same objective (indicated by the output). In fact, they can even have different approaches, and for this reason, the outputs and inputs are different between them, making naturally non-homogeneous.

In the Network Slack Based Measure (SBM) model, we are dealing with interdependent stages and therefore, the intensity vectors are different among the stages. In other words, a specific intensity vector is assigned to each stage. When we measure the Overall Efficiency of the DMU, the results suggest very low efficiency scores. In this research, we will deal with heterogeneity in Network SBM DEA model by categorizing the main issues of heterogeneity in the DMUs efficiency as a whole and their stages efficiency of the internal process. Therefore, we will compare the impact of those adjustments and discuss which ones have a bigger effect on the efficiency scores.

Empirical results suggest that by overcoming the heterogeneity in the DMUs, the low levels of efficiency in the Network SBM model with different intensity vectors in each stage can be improved meanwhile the discriminant power of the model is kept.

WB 3.3

Analysis of UK and US SME Platform Markets using Business Model Theory: An emphasis on Web 2.0 technology sophistication

María Manuela Gutiérrez-Leefmans (Universidad de la Américas Puebla, Mexico)
Corresponding Author(s): María Manuela Gutiérrez-Leefmans (maria.gutierrez@udlap.mx)

Abstract
Consumer use of popular Web 2.0 and social media platforms such as Facebook and Twitter is well documented. However, the use of such technologies by Small and Medium Sized Enterprises (SMEs) has received relatively little attention. This study focuses on platforms for SMEs. These are websites designed specifically for SMEs to gain information, network with each other and in some cases conduct sales through an electronic marketplace. Data indicates that there is extensive interest in these platforms, which offer relevant content and networking opportunities to small businesses.

The competitive landscape for these platforms is mapped out using business model theory. In total, 144 platforms in the UK and the US were identified. Using a mixed method approach of online panel data, cluster analysis and website content analysis, 32 were analyzed in detail. For such purpose, a scale of Web 2.0 technology sophistication is proposed using inter-coder reliability. The differences between both markets were identified and, based on the results, a taxonomy is proposed based on value proposition, Web 2.0 sophistication and revenue model maturity that defines five distinct strategic groups: information laggards, basic networking, advanced networking, advanced networking mature and social media markets.

Over the last decade, business model literature has identified different approaches, but few studies have provided the empirical evidence to test these. In its second phase, this study applies Zott and Amit's activity-system concept of business models by using a synthesis of online panel data and case studies of the leading SME platforms in the UK (and one example from the US). The model that emerges as a result is one where the platform has different revenue models and exploits the network effects that increase the user base; it offers content and the opportunity to network; and has a broad product-market scope. A theoretical framework is proposed, that explains the interplay between the business strategy, value proposition, end-user and Web 2.0 sophistication, emphasizing the relevance of the user as the center of the business model.

Unlike most business model research that focuses on the firm level, this study presents a synthesis of the market level and the firm level, which brings insights into business model innovation and strategic group transition. The benefits from doing a study of two whole markets and detailed case studies are therefore, confirmed. Finally, the implications for SMEs, platform owners, banks and government agencies, are outlined.

THURSDAY 18

Sessions TA 1

TA 1.1

Forecasting blood donations with neural networks

Tine Van Calster (Katholieke Universiteit Leuven, Belgium), Michael Reusens (Katholieke Universiteit Leuven, Belgium), Bart Baesens (Katholieke Universiteit Leuven, Belgium), Wilfried Lemahieu (Katholieke Universiteit Leuven, Belgium)
Corresponding Author(s): Tine Van Calster (tine.vancalster@kuleuven.be)

Abstract
Machine learning applications in healthcare have become manifold in recent years, ranging from medical image and text processing to predicting individual patient outcome. In this paper, we focus on a novel use of ML techniques in healthcare: blood donation forecasting. Predicting the turn-out for blood donations is essential for organizations such as the Red Cross, who are responsible for maintaining the blood supply for local hospitals. If the supply for a certain blood type is in danger, potential donors need to be contacted, either personally or through more extensive marketing campaigns. Furthermore, this forecast has another purpose, as it is also used for the day-to-day logistics of the blood donations centres, such as the amount of personnel that needs to be present and the number of donation spots that will be available. This paper will focus on this second application, which entails a daily forecast of the number of blood donations. This turn-out is known to be influenced by external factors, such as newspaper articles and (inter)national events, and is therefore susceptible to last-minute changes. These sudden fluctuations make the forecast even more challenging, as under-forecasts may lead to missing out on valuable blood donations and over-forecasts cause unnecessary employment costs. The model in this paper aims to forecast the daily turn-out for blood donations for the Red Cross in Flanders by means of neural networks, which try to capture the influence of last-minute events by turning to website data.

Neural networks are an integral part of forecasting applications with frequent time series, such as load forecasting and traffic forecasting, and are therefore a perfect fit with the problem setting of this paper. This type of applications generally has a lot more training data than traditional time series problems, which is why they can truly benefit from neural networks in general. Some network architectures were even especially designed to deal with the long-distance dependencies that are characteristic for time series, i.e. Long Short Term Memory neural networks (LSTM). In this paper, we will take a look at three different network structures in order to see which one generates the most accurate forecasts: Multilayer Perceptron (MLP), Simple Recurrent

Neural Network (RNN) and Long Short Term Memory neural network (LSTM). These models take previous time points as their input, as well as website data and seasonal data, such as the day of the week and holidays. Furthermore, we also investigate the hierarchical aspect of the data by verifying whether there is an improvement in performance if we train a separate model for each blood type and sum up the forecasts. The performance of these models is compared to three benchmarks: a naïve seasonal model – in order to demonstrate the improvement in accuracy that neural networks provide – , an MLP neural network without the website data – in order to ascertain the added value of this external factor –, and an MLP network that only takes the total time series into account – in order to investigate the benefit of the hierarchy in the time series.

The results of this analysis lead to three main conclusions. Firstly, all neural network architectures outperform the seasonal naïve model in this application. Secondly, website data greatly improves the accuracy of the models, which is especially noticeable for days with an exceptionally large amount of blood donations. Thirdly, the accuracy of the neural networks that sum up the forecasts for the different blood types, is higher than the accuracy of the neural network that is trained on the total time series.

TA 1.2

Predicting dwell times of import containers in a container terminal: case study of the port of Arica in Chile

Francisca Quijada (Universidad de Los Andes, Chile), Sebastian Maldonado (Universidad de Los Andes, Chile), Rosa Guadalupe Gonzalez Ramirez (Universidad de Los Andes, Chile)
Corresponding Author(s): Rosa Guadalupe Gonzalez Ramirez (rgonzalez@uandes.cl)

Abstract
Along with the globalization, international trade has shown an increasing trend, where maritime transport mode accounts for the biggest participation. As such, maritime ports are achieving a significant role and this has motivated the need of new operational strategies to efficiently handle cargo in the port terminals. During the transfer services of a port, containers (or cargo in general) need to be temporarily stored in the yard. In the case of export operations, there is a time window in which containers are received and stacked at the yard to later be loaded in the corresponding vessel. Same situation occurs with the import containers. When unloaded from the vessel, are stored in the yard to be later dispatched to their consignees. The time spent by a container in the port terminal is denoted as "dwell time". Dwell times are a performance indicator monitored by container terminals with the aim to reduce it as this allows better utilization of the space, which is a scarce resource.

In this work, we consider as case study the port terminal of Arica (TPA), located in the city of Arica, Northern Chile. The port terminal has significant logistics challenges. Approximately 70% of the cargo corresponds to in-transit cargo from Bolivia that has special conditions due to the Peace Agreements between Chile and Bolivia. Particularly, no storage fee to the cargo in long periods of time are charged, reason for which high dwell times are observed.

Motivated by this situation and the fact that yard managers do not have a criterion to stack import containers in the yard, we propose a methodology to predict dwell times of containers that can be used to segregate cargo into classes of containers with same range of dwell times. This extends the work proposed by Gaete et al., (2017) by considering the same problem but employing different techniques.
The dwell time prediction model considers as variables the container´s weight, size, type, port of origin and the corresponding consignee. A data set of 165,848 observations between September 2013 and May 2016 is considered. Three models were implemented with a subset of 20,000 observations that correspond to the year 2016: Multiple Linear Regression, Decision Trees and Random Forest. The models were implemented in the R programming language. With the predictions obtained, three classes were generated for comparison purposes: "Less than 7 days", "Between 7 and 14 days" and "Over 14 days". The results obtained with the models were evaluated using the Balanced Accuracy as a performance measure.

The results obtained for each algorithm, using only data of the year 2016 are summarized as follows: Multiple Linear Regression obtained a MAPE (Mean Absolute Percentage Error) of 44.55% and Balanced Accuracy of 0.61, while Decision Trees a MAPE of 40.10% and Balanced Accuracy of 0.60. Random Forest is the algorithm performing best, with a MAPE of 38.70% and Balanced Accuracy of 0.61. Therefor the latter model is chosen to conduct the final prediction.

In the final prediction analysis, a MAPE of 50.28% and Balanced Accuracy of 0.54 are achieved. Motivated by these findings, Random Forest is fed with the last 50,000 observations, obtaining a Balanced Accuracy of 0.58, inferior to the performance obtained using data of 2016. This result indicates the possible existence of a strong time component present in the data set. In comparison to the results of related literature, it is important to point out that our results overcome those found by Gaete et al., (2017) and other related work in the literature.

TA 1.3

A Step Towards Demand Sensing: Employing EDI 852 Product Activity Data in Demand Forecasting

Jente Van Belle (Vrije Universiteit Brussel, Belgium), Wouter Verbeke (Vrije Universiteit Brussel, Belgium)
Corresponding Author(s): Jente Van Belle (Jente.Van.Belle@vub.be)

Abstract
This paper deals with short-term demand forecasting of medicines in a US drugs factory based on historical sales and EDI 852 product activity data. Traditionally, demand forecasting relies on statistical methods such as ARIMA, and smoothing methods such as exponential smoothing, to extrapolate the series of historical sales. Although these methods produce reasonably accurate forecasts in many cases, companies are looking for an increasingly higher level of forecast accuracy to further enhance the efficiency of their supply chains. With more and more diverse data becoming available at a higher velocity, it becomes possible to employ non-traditional data sources and generate forecasts for increasingly shorter time periods. From this point of view, in recent years, the concept of demand sensing emerged. However, in the scientific literature and empirical studies, the concept has gained only very little attention to date. Essentially, demand sensing comes down to leveraging a combination of near real-time (i.e. orders), downstream (ideally point-of-sale) and external data to improve the accuracy of short-term forecasts. In their purest form, however, the traditional models do not allow for the inclusion of covariates. In this paper a particular type of downstream data, EDI 852 product activity data (a data standard used for exchanging product related information between the supplier and the (end)customer), is employed in producing short-term weekly demand forecasts obtained from linear, dynamic (i.e. with ARIMA errors) and lasso regression and an ETS-X, artificial neural network and support vector regression model. In order to produce the forecasts, three years of historical weekly factory sales data and weekly wholesaler sales, ending inventory, on order quantities and receipts for the same time period are used. As a benchmark, also forecasts from an ARIMA and ETS model are produced. To evaluate the forecast accuracy, we adopt a sliding window approach and measure out-of-sample RMSE, MAE, MAPE and MASE, so that both comparison between the methods as well as comparison across different series is facilitated. Our results clearly indicate that forecast accuracy can effectively be improved by leveraging downstream data from wholesalers. The extent of improvement over traditional forecasting techniques varies between the forecast items included in the study, as well as between the various models considered. Although linear models can already produce considerable gains in accuracy levels, nonlinear methods tend to outperform.

TA 1.4

Data-driven inventory management: A random forest-based joint estimation and optimization model for the newsvendor problem

Fabian Taigel (University of Würzburg, Germany), Jan Meller (University of Würzburg, Germany)
Corresponding Author(s): Jan Meller (jan.meller@uni-wuerzburg.de)

Abstract

The problem of how to determine inventory targets facing uncertain demand has been at the center of attention in operations management research over decades. When demand is nonstationary, that is, demand patterns reveal seasonality, follow a trend or are influenced by a variety of other factors, a common approach is to explain demand variations with the help of predictive auxiliary data. The classical methodology follows a two-step logic: First, we fit a forecasting model to a training data set consisting of historical demand observations and data features, i.e., summarized representations of the auxiliary data per demand observation. Then we evaluate the forecast error of the fully calibrated model in order to characterize the remaining uncertainty, i.e., the distribution of forecast errors. Based on this estimated "uncertainty distribution", an additional safety stock is calculated to account for forecasting errors.

Despite its intuitiveness, this approach, which we further refer to as separate estimation and optimization (SEO), possesses a major drawback. We solve two separate optimization problems (as the calibration of the forecasting model can be considered as an optimization problem) which are not necessarily aligned: Since the forecast aims at predicting the conditional mean, a typical objective function is the squared error. However, in the subsequent inventory problem, costs for underage respectively overage are typically asymmetric and as a consequence, here we are optimizing a different loss function. Consequently, we lose information about relationships between feature data and demand that are not relevant for the conditional mean, but might be for the quantile of the conditional demand distribution – the quantity we are actually interested in. To account for this problem, recently, the operations management community has proposed approaches that integrate the feature data directly into the decision support model and hence only deal with one single optimization problem. We further refer to these approaches as joint estimation and optimization (JEO) approaches. To the best of our knowledge, up to now, no thorough examination exists about the different behavior of SEO versus JEO in different settings. In addition, the factors that drive the ultimate performance of both approaches have not been examined yet.

In our work, we analytically compare SEO and JEO approaches and carve out their structural differences. We study the impact of different feature-demand structures on the performance of the two approaches and find that particularly the relationship between features and the demand uncertainty has an important effect on the respective performances. Furthermore, we propose a new JEO approach for Newsvendor problems based on the random forest machine learning algorithm. With our approach, random forests are learned from the data to make the best inventory decision – instead of a prediction of demand.

To this end, we account for the (typically) asymmetric costs of overage and underage quantities in the learning algorithm. Finally, we examine the performance of our approach and its SEO counterpart both in a controlled simulation setting as well as on a real-world data set. Our first results are very promising: While in the setting with stationary demand uncertainty we could not identify major performance differences between SEO and JEO, our new JEO approach is able to identify and exploit settings where the features have predictive information about nonstationary demand uncertainty. In such settings with feature dependent demand uncertainty, we are able to beat the SEO benchmark significantly.

Sessions TB 1

TB 1.1

Behavior based time-to-default predictions

María Óskarsdóttir (Katholieke Universiteit Leuven, Belgium), Cristian Bravo (University of Southampton, United Kingdom), Bart Baesens (Katholieke Universiteit Leuven, Belgium), Jan Vanthienen (Katholieke Universiteit Leuven, Belgium)
Corresponding Author(s): María Óskarsdóttir (maria.oskarsdottir@kuleuven.be)

Abstract

One of the oldest applications in analytics is credit scoring where, traditionally, people's banking history is used to assess their creditworthiness. However, as data is continuously being generated in more volume and variety than ever before, new credit assessment methods are emerging. In particular, new variables to capture borrower behavior going beyond simple repayment history have been shown to be good predictors of whether or not people will default on their loans. In industry, this is being utilized by the means of smartphone applications which facilitate microlending under the assumption that the way people use their phones is a proxy for how they lead their lives. These applications analyze the data that is generated when the phone is used –thus logging various aspects of people's behavior–
to decide whether the person should be granted the loan. The impact of these applications is especially important in developing countries where large portions of the population do not have any banking history, and therefore no means of receiving a loan in the traditional way. In contrast, since most people have cell phones, they have the opportunity to request credits of variable amounts with flexible payments, and the decision is based on data stored in the phone, i.e. their behavior.

In addition to suitable data, the assessment of creditworthiness depends on the models that are used. In credit risk modelling the goal is to predict whether a loan applicant will default on his loan or not and to this end binary classifiers have been used to a great extent (Lessmann et al., 2015). Survival analysis techniques have also been deployed with great success. They have a clear advantage in this regard, because they facilitate modelling when the default will occur. In particular, mixture cure models have been shown to excel when predicting time to default in credit risk models (Dirick et al., 2017). When it comes to credit risk, it is not certain that a customer will eventually default. In fact, the opposite is more often true, because there is always a fraction of people that do pay back their loans eventually, which mixture cure models take into account.

In this study we analyze the dataset of a microlending smartphone application to create score cards. We build credit scoring models in the regular way, using binary classifiers such as logistic regression, decision trees and random forests, thus including a selection of both popular and easily interpretable models and powerful black box models. In addition we apply survival analysis techniques to predict the time of default, which is very fitting for this kind of data, since the loan period is typically not very long and the user has freedom to choose how he pays off the loan. We pay special attention to mixture cure models which assume that a fraction of the customers never default, as is the case in reality. We compare the performance of the models in terms of AUC and Expected Maximum Profit.

The benefits of smartphone facilitated microloans are substantial, since they can help individuals as well as small business owners ensure the success of their operations without being too risky, for both parties. Their success, however, highly depends on using appropriate credit scoring models that are able to accurately assess people's creditworthiness.

References

Dirick, L., Claeskens, G., Baesens, B. (2017): Time to default in credit scoring using survival analysis: a benchmark study. Journal of the Operational Research

Society 68, Issue 6, 652–665

Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L. C. (2015): Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research 247, Issue 1, 124-136

TB 1.2

Leveraging PD Models for Bayesian Inference of Default Correlations

Miguel Biron (Superintendency of Banks and Financial Institutions of Chile, Chile), Victor Medina (Superintendency of Banks and Financial Institutions of Chile, Chile)

Corresponding Author(s): Miguel Biron (mbiron@sbif.cl)

Abstract

We present a method to infer the default correlation between loans in a single risk factor (SRF) model for a loan portfolio. For each period, this model assumes a portfolio composed of a given number of defaultable loans. The event of default of each loan depends both on an idiosyncratic and a common systemic risk factor. The magnitude of dependence on the systemic factor is related to the correlation parameter. Our objective is to infer this parameter from historical data.

Unlike existing approaches, we propose a two-stage procedure that takes advantage of models that estimate the Probability of Default (PD) for each of the loans in a portfolio. These models are a standard component of credit risk management frameworks at banks and other financial institutions. By leveraging existing PD models, we aim to facilitate the inclusion of the analysis of correlations to these frameworks. Moreover, since we only require knowing the output of the PD model, we do not need to impose any structure on it, which allows for the use of PD models built using more advanced statistical learning techniques. This is an improvement over existing approaches, which require jointly estimating the default correlation and a generalized linear model (GLM) for the PDs.

The drawback of our formulation is that the uncertainty on the parameters will be underestimated, because we omit variation from the first stage by assuming that the PDs are fixed. We believe that this downside is offset by the gains described above (although more research is needed to quantify this trade-off). Furthermore, we take a Bayesian approach because it allows for better accounting of uncertainty.

We start with a probabilistic definition of the SRF model, which allows us to present a simple theoretical justification for our approach. Then, we propose two variants for our procedure. The first, which we refer to as "naive", assumes the PDs passed to it are exactly true. The second one, which we call "robust", jointly fits the correlation and a simple linear regression in which the only predictor is the PD from the model. This additional degree of freedom could correct biases introduced in the PD model trained assuming independence.

We implement the methods in the probabilistic programming language Stan (using its interface for R). In order to evaluate our methods, we first apply them to a series of simulated portfolios with known correlations, and find that they can recover the parameters correctly. Then, we study their performance on a real dataset extracted from a portfolio of mortgages, for which a PD model exists. We find that our methods produce results which are consistent with the literature. Furthermore, we find that the robust approach performs better than the naive method under various measures.

Finally, we empirically assess the adequacy of the assumption of constant correlation through the business cycle. We find a pattern that is again consistent with the literature, in which the correlation is found to be higher before a period of high default rates. Based on these findings, we discuss various possible extensions for the model.

We hope that the two-stage procedure presented in this work will motivate more research on the subject of inferring default correlations from data, given the ease with which it can be incorporated to existing credit risk evaluation frameworks.

TB 1.3

Bias-Free Text Evaluations in Micro and SME Credit Scoring using Deep Learning

Cristian Bravo (University of Southampton, United Kingdom), Andrés Medina (Instituto Sistemas Complejos de Ingeniería, Chile)
Corresponding Author(s): Cristian Bravo (c.bravo@soton.ac.uk)

TB 1.4

Psychometric Credit Scoring Model for Microloan based on HEXACO Personality Inventory

Bo Kyeong Lee (Department of Industrial Engineering, Yonsei University, 134 Shinchon-dong, Seoul 120-749, South Korea, South Korea), Dong Ha Kim (Department of Industrial Engineering, Yonsei University, 134 Shinchon-dong, Seoul 120-749, South Korea, South Korea), So Young Sohn (Department of Industrial Engineering, Yonsei University, 134 Shinchon-dong, Seoul 120-749, South Korea, South Korea)
Corresponding Author(s): So Young Sohn (sohns@yonsei.ac.kr)

Abstract
Microfinance institutions (MFIs) have been established to provide loans to the poor and low-income earners, who have been excluded from the formal banking system. Following the rapid growth of microfinance industry across the world, the issue of trade-off between financial sustainability and outreach goal has been raised in both academic and practical fields. Financial sustainability of MFIs and outreach are seemingly conflicting with each other. To improve both outreach and sustainability, a tool for selecting proper borrowers is necessary. The past 20 years have witnessed an increased interest in the usefulness of a credit scoring model for microfinance.
The conventional credit scoring models used in finance industry including commercial banks and MFIs were designed to predict the loan default based on the borrower's ability to repay (ATR). ATR consists of financial ability and willingness to repay (WTR), and loan can be retrieved when both financial ability and WTR are satisfied. However, existing credit scoring models have not distinguished these two aspects so far, although each can be fulfilled independently.
Some studies found that WTR is a more critical factor for loan default than financial ability in MFIs because MFIs do not require collaterals and there is no sanction for the loan default. Thus, repayment of microloan highly relies on individual borrower's decision. Although the necessity of distinguishing the backgrounds of loan default has been recognized, most previous studies on the micro-credit scoring model have not separated the WTR from the ATR when performing credit scoring.
WTR depends largely on the borrower's personality. Klinger et al. [1] first implied the relationship between the borrower's personality and loan default. The authors suggested a credit scoring model based on a Big-Five personality questionnaire. The Big-Five personality inventory is a popular way of measuring individual entrepreneurship. They considered that a borrower with a personality demonstrating good entrepreneurship presented a low risk of loan default. This concept assumes that borrowers who can succeed in their business will have the financial ability to pay back the loan. However, even when the borrower's business is profitable, loan default can occur in the absence of a WTR.
The aim of this study is to develop a psychometric credit scoring model predicting default risk related to WTR. Thus, we employ the data from MFIs in developed country because financial ability of the borrowers in developed countries is expected to be less volatile than that in developing countries generally. Further, we categorize default cases into two groups: default group due to financial difficulties and default group due to lack of WTR. We then utilize their personality measured using the HEXACO model [2], consisting of six personality dimensions: Honesty-Humility (H), Emotionality (E), eXtraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O). HEXACO model includes an additional dimension,

"Honesty-Humility" compared to the Big Five Inventory which has been widely used to measure the personality. Finally, we conduct a logistic regression using the respondents' default records to investigate which personality corresponds to a high WTR. Based on the results of the logistic regression, we develop a credit scoring model that is expected to prevent microloan default.

References

[1] Klinger, B., Khwaja, A. and Del Carpio, C. (2013). Enterprising psychometrics and poverty reduction. Springer.

[2] Ashton, M. and Lee, K. (2007). Empirical, Theoretical, and Practical Advantages of the HEXACO Model of Personality Structure. Personality and Social Psychology Review, 11(2), pp.150-166.

Sessions TA 2

TA 2.1

Exploring the relationship between online activity and achievement across two universities and learning management systems

Sergio Celis (Universidad de Chile, Chile), Joaquín Muñoz (Universidad de Chile, Chile), Dany Lopez (Pontifica Universidad Católica de Chile, Chile), Augusto Sandoval (Pontifica Universidad Católica de Chile, Chile)
Corresponding Author(s): Sergio Celis (scelis@ing.uchile.cl)

Abstract

The growing field of learning analytics has already contributed to both institutional decision-making and theory development (Ferguson, 2012). The potential of this field is yet to be defined and many challenges need to be addressed. One of them is the generalization of outcomes across institutions and different platforms. So far, most of the studies work with a single institutional dataset (e.g., Ellis, Han, & Pardo, 2017). This study explores the relationship between learning management system (LMS) online activity and academic achievement across two different undergraduate degrees, in two different institutions and in two different learning management systems, in Chile

Literature Review

Learning analytics is a fairly new area of research that uses mathematical and computing tools to analyze educational data generated by the interaction between students and online platforms that universities use for supporting learning processes (Larusson & White, 2014; Romero et al., 2008). Among the learning analytics' common tasks are classification, clustering, text mining, and visualization. Their common tools are Bayesians networks, neural networks, and decisional trees, which are often complemented with correlations and regressions (Romero & Ventura, 2010). These tools are used for data visualization (e.g., Leony et al., 2012), feedback for instructors (e.g., Mazza & Ditrova, 2004), prediction models for students' academic achievement, and institutional decision-making (Macfadyen & Dawson, 2010; Treaster, 2017). Recent studies are intending to connect learning analytics to established and emerging learning theories. For example, Ellis et al. (2017) found relationships between student's approaches to learning (i.e., deep and surface approach to study), interaction with online tools, and final course grades. The relationships were in the expected sign, students reporting deep approaches to study tend to interact more with online platforms and to obtain better grades than their counterparts declaring surface approaches. Authors (2016) found similar results with the students' use of print and electronic resources from the university library. These two studies have used data from a single institution. This is a restriction to draw general conclusions. In this study, we explore how learning analytics models and tools perform within and across different learning and institutional environments.

Methods

The data for this study come from academic and LMS information in relation to first year students enrolled at a School of Engineering and at a School of Education. Both schools, each of which belongs to a different research university in Chile, have constructed statistical models for understanding and predicting academic achievement. At both schools, we collected data from the 2013 and 2014 entry cohorts. Generally speaking, we can classify the data into three categories: Pre-college data, academic performance, and LMS interaction. Pre-college data include categories such as gender, high school GPA, socio-economic background, and admission test scores. Academic performance consists of the final grade of each first-year course, and pass rates in the first semester. LMS represented the most challenging type of data to process and include in the final dataset. In both schools, students interact in multiple ways with LMS systems associated to each course. In the LMS platform, students usually read, download, and upload course content, upload homework, participate in course forums, and consult course calendars.

As activities in LMS are encoded differently, we performed a homogenization of data to conduct comparisons and spot patterns that are inherent to each LMS and would not otherwise emerge. We created five categories in this process: Read Comment, Write Comment, Academic Content, Administrative Content and Test.

With the results, we seek to predict student success and gain insights about the learning processes across different disciplinary and institutional environments. Thus, this article tackles key methodological and practical implications.

TA 2.2

Combining student learning research and learning analytics to understand student's learning process

Maximiliano Montenegro (Pontifica Universidad Católica de Chile, Chile), Carlos Gonzalez (Pontifica Universidad Católica de Chile, Chile)

Corresponding Author(s): Maximiliano Montenegro (maximiliano.montenegro@uc.cl)

Abstract

Student learning research (e.g., Parpala & Lindblom-Ylänne, 2012) is a well-established line of research in higher education. It has established that students adopt deep approaches to studying in situations in which they have positive perceptions of the learning context (in terms of good teaching, clear goals and standards, appropriate assessment, and appropriate workload) and in which their teachers adopt student-focused approaches to teaching. In this manner, it has helped to support the improvement of teaching, providing better learning opportunities for students. On the other hand, learning analytics (e.g., Larusson & White, 2014) is a relatively recent area of investigation, which uses mathematical and computing tools to analyze educational data generated by the interaction between students and online platforms that universities use for supporting learning processes. Among the learning analytics' common tasks are classification, clustering, text mining, and visualization. Their common tools are Bayesians networks, neural networks, and decisional trees, which are often complemented with correlations and regressions. These tools are used for data visualization, feedback for instructors, prediction models for students' academic achievement and institutional decision-making. We consider both areas of research have a common interest on understanding and supporting students' learning and therefore they may provide a more holistic and integrated perspective on students' learning. Recently, authors have started to work combining data from LMS with traditional students' questionnaires according to that perspective, like the one from Ellis et al. (2017) that conducted a study with engineering students in one class using the well-known Revised Study Process Questionnaire (R-SPQ) (Biggs et al., 2001) and combined the results with LMS frequencies of use.

This study aims to extend the previous study by relating students learning approaches to their library loans, use's patterns for the Learning Management System, and average score attending a college of Education of a research oriented university. It was applied LEARN questionnaire (Parpala & Lindblom-Ylänne, 2012) at the end of the academic year to a sample of 147 college students and later, students' data were obtained from the library loan system, the university's LMS and registrar for the same

academic period. Our results show some expected patterns of learning processes: deep learning students using more the library; and also, some unexpected ones: surface students using more the communication features of the LMS and the digital library. These patterns may suggest that, effectively, deep learners are good library users, thus providing data on observable behaviors to evidence based on perceptions captured by a questionnaire. Similarly, it raises questions on the aims of surface students use of LMS, particularly its communicative features. This is, somehow, contrary to what may be expected from surface learners, usually characterized as unengaged. Therefore, it open questions for further research which may have not emerged using questionnaires only.

Using student learning research and learning analytics in conjunction, as demonstrated in this study, may be useful to answer questions such as the one raised here or to describe patterns of learning not captured using only one perspective. Thus, we see that combining them will contribute to a better understanding of student learning by 1) integrating data from cognitive processes (beliefs and perceptions) with empirically identifiable behaviours (students' interaction with digital systems), bringing forward a longstanding debate in social science; 2) providing a strong theoretical framework (students approaches to learning theory) to interpret data generated through learning analytics techniques and 3) generating and analysing evidence from different sources that allows data triangulation to provide a stronger model of students' learning. Our future research will advance in this line.

TA 2.3

Learning analytics: traps for the unwary

Carolina Guzmán (Centre for Advanced Research in Education. University of Chile., Chile)
Corresponding Author(s): Carolina Guzmán (carolina.guzman@ciae.uchile.cl)

Abstract
Global trends such as the privatisation of higher education systems, the reduction of public funds and the competition for resources are putting higher education institutions under pressure to produce the best possible outcomes and compete among each other (Naidoo, 2016). Universities have to attract/produce research incomes/outcomes and they need to improve the graduation rates while diminishing students' non-completion rates. Complementarily, massification processes, especially in non-selective universities, have allowed more and diverse students to access universities. Because of this diversity, universities have been compelled to develop a series of mechanisms and practices to help undergraduate students to succeed in their academic journey. Learning analytics can be considered as one of these mechanisms.

In this paper, learning analytics from a critical perspective is going to be examined. In particular, issues related to the over-simplification of teaching-learning processes and ethical challenges will be discussed.

Learning analytics in higher education: critical points

Generally speaking, learning analytics is a data-driven approach which allows the gathering of large amounts of data produced by students in order to predict their individual learning pattern (Fynn, 2016; Siemens, 2010). It also includes intervention programmes for students targeted as at risk (Fynn, 2016). Usually, learning analytics involve the collection of data about the interaction between students and institutional learning management systems but this might also be complemented with socio-demographic information, grades of entrance tests, library usage, among others (Ifenthaler & Schumacher, 2016).

Learning analytics, though, poses a series of challenges and questions that call for caution:
a) It might underplay the complexity of teaching-learning processes. Multiple and complex factors interact and intervene to

shape learning in higher education, especially when students' and institutional profiles are diverse. As a result, 'Only a relatively low proportion of student success variation can be explained by traditional statistical modelling techniques such as multiple linear regression analyses. These techniques simply establish valid and reliable relationships between relatively few variables relevant to a specific context' (Subotzky & Prinsloo, 2011:183). Disciplinary issues, the relationship between students and teachers, individual factors and institutional cultures might also be affecting learning processes. Additionally, the use of learning analytics fails to acknowledge what occurs with students' learning when they use other informal learning spaces including those in social media, where students today spend much time (Wintrup, 2017).
b) Learning analytics might be seen as a tool of surveillance through which students are permanently spied upon (Wintrup, 2017). Complementarily, it might be conceived as a limitation of freedom of students (Wintrup, 2017). Through machine-driven algorithms, governments and institutions might steer students' learning choices (Fynn, 2016) in ways that they are not aware of.
c) Also, questions arise about who collects the data, where it is stored, whether or not it has an expiration date, who is accountable for it, the extent to which it is secured, and what is going to be done with it (Ifenthaler & Schumacher, 2016; Slade & Prinsloo, 2013).
d) That students are identified as at risk might promote their labelling (Wintrup, 2017; Scholes, 2016) and act as a set of self-fulfilling prophecies. Academic staff might also be trapped in these processes, coming to hold unduly limited expectations about students' academic success.

In this paper, these concerns about learning analytics are going to be expanded in the light of our own experiences as academics using a learning analytic approach. Also, recommendations for better practices that mitigate these risks are going to be offered.

TA 2.4

Blended analytics: Capturing and visualizing physical and digital learning

Sarah Howard (University of Wollongong, Australia), Jie Yang (University of Wollongong, Australia), Jun Ma (University of Wollongong, Australia)
Corresponding Author(s): Sarah Howard (scelis@ing.uchile.cl)

Abstract
Current methods of research have struggled to meaningfully capture the longitudinal and complex nature of human interactions, specifically learning and teaching processes. To further complicate this, individuals are increasingly "blending" traditional face-to-face interactions with digital places. This has been a particular problem in schools and higher education, where it has been a struggle to capture how students engage in blended learning, combining work in the face-to-face classroom and through their computers. Multi-modal methods are needed to capture a wider range of classroom data to understand the range of teaching and learning practices and use of digital technologies to support blended learning (Blikstein, 2013). In this paper, we present results from two classroom data collections, which include observations in both physical and digital spaces. Video and data mining techniques are used for analysis and visualization of the multi-modal data. Implications of the findings in relation to blended learning design, as well as combining physical and digital observation data to inform learning and teaching are discussed.

In both studies, data are collected through components of a combined physical and digital observation system. This system is designed to gather naturalistic multi-modal data on learning and teaching, using a low-disturbance classroom video kit (physical) and embedded computer agents (digital). The aim of both the computer agents and the low-disturbance observation system is to capture physical classroom activity for extended periods of time with as little disruption as possible. Two studies will be addressed in the paper, one illustrating the physical data collection and the other the digital data collection.

The first study was an observation of the physical classroom is presented through a pilot study in an Australian Year 1 classroom, including 25 students, one teacher and one student-teacher. The physical movements of five students were tracked over a 12-minute period. Object tracking/motion using Optical Flow was the main approach to analyzing the video. Results showed three different patterns of movement and interaction among the five students. Students' behaviors were not necessarily categorically successful or unsuccessful, but showed use of different classroom resources and student interactions. Findings suggest different learning processes, and have implications for the physical classroom layout and task design.

The second study, which addressed observation of the digital space, is drawn from a dataset of students and teachers' real-time digital-device usage behaviors, from 50,000 Android tablet devices used in Year 1-3 classrooms. Data span a period of four months, at a high level of fine granularity. Usage patterns were clustered in relation to scores on the Australian National Assessment Program – Literacy and Numeracy, to explore associations among usage and performance. Patterns revealed associations among complex combinations of tablet app use and performance. Findings suggest that particular combinations of app use have an effect on learning.

Being able to naturalistically observe in physical and digital spaces increases the possibility of extracting patterns among student and teacher practices in blended learning. These relations can them be analysed to identify overarching principles of learning in the blended space, and combined with other classroom data (Nagy, 2016). The next step is linking data from the two spaces together. This is currently being piloted in a secondary Science classroom observation, looking at students' laptop and OneNote use. With access to this kind of classroom information, teachers can adjust their learning designs to accommodate what may be observed as a successful process in the physical and/or digital space.

References

Blikstein, P. (2013). Multimodal Learning Analytics. In LAK '13 (pp. 102–106).

Nagy, R. (2016). Tracking and visualizing student effort: Evolution of a practical analytics tool for staff and student engagement. Journal of Learning Analytics, 3(2), 165–193. https://doi.org/10.18608/jla.2016.32.8

Sessions TB 2

TB 2.1

Outlining New Product Development Research through Bibliometrics: Analyzing Journals, Articles and Researchers

Nelson Andrade-Valbuena (Universidad de Chile, Chile), Jose Merigo-Lindahl (Universidad de Chile, Chile)
Corresponding Author(s): Nelson Andrade-Valbuena (nandradev@fen.uchile.cl)

Abstract

New Product Development (NPD) is a noteworthy field that has attracted the attention of scholars for its relevance for firm success. Based on bibliometric indicators and Visualization of Similarities (VOS) network analysis, this papers outlines a general perspective on NPD research from last 40 years, detecting the most prominent papers and journals and the most prolific and relevant authors. The Journal of Product Innovation Management was found to be the most prominent journal for new product development, followed by Management Science and Industrial Marketing Management. Several researchers are highlighted as central contributors, including Gary Lynn, Roger Calantone, Michael Song, Robert Cooper and Abbie Griffin. This paper is informative and contributes to the NPD literature by offering a global perspective on the field.

TB 2.2

Featurization Methods and Predictors for Income Inference based on Communication Patterns

Martin Fixman (Universidad de Buenos Aires, Argentina), Martin Minnoni (Grandata Labs, United States), Matias Travizano (Grandata Labs, United States), Carlos Sarraute (Grandata Labs, Argentina)
Corresponding Author(s): Martin Fixman (martinfixman@gmail.com), Carlos Sarraute (charles@grandata.com)

Abstract
Patterns of mobile phone communications, coupled with the information of the social network graph and financial behavior, can be leveraged to make inferences of users' socio-economic attributes such as their income level. We present in this work several methods to extract features from mobile phone usage, and compare different combinations of supervised machine learning techniques and feature sets used for the inference of users' income.

For this study, we had access to anonymized Call Detail Records (CDR), composed of voice calls and text messages, from a telecommunication company in a Latin American country for a period of 3 consecutive months. Additionally, we used a set of account balances of millions of clients of a bank for a period of 6 consecutive months in the same country. The data of each bank client contains the phone number anonymized with the same cryptographic function as the telco dataset, allowing us to join the datasets, along with the average income of this person over 6 months.

We represent the network as a directed graph G = < V, E >, where the nodes represent users and the edges represent their communications. When we analyzed the communications between bank clients, we observed a strong homophily in the graph respect to the users' income, which is tied to the social stratification between populations of different purchasing power [1]. For the classification task, we divided the ground truth set in two groups of equal sizes: High Income and Low Income.

Each edge in E contains the information: total number of calls; total time (in seconds) of all the calls; and total number of text messages exchanged. We now describe several ways of transforming data from the graph G = < V, E > into individual features. First, we aggregate the number of calls, total time and SMS for every node, separated on whether those features are incoming or outgoing. Then we extend this feature generation to nodes in the ego network of order n of a node v, which is the subgraph composed of all the nodes which are at distance at most n from v, obtaining in this way the user data of order n, for n = 1, 2, 3.

For our experimental setting, we consider the set of nodes which have at least one neighbor with income information, called the Inner Graph. In this set, each edge feature is further separated in 2 features, depending on the category of the other endpoint (high or low income), obtaining a total of 72 features per node.

The inferences based on features aggregated by node were performed using Logistic Regression and Random Forest classifiers. Finally, we compared the results with the Bayesian Method presented in [1], which only uses the amount of High Income and Low Income users in the egonetwork.

Our experiments show that the Bayesian Method obtains an AUC = 0.746, and thus makes a better prediction than the machine learning methods Logistic Regression (AUC = 0.693) and Random Forest (AUC = 0.714) using the most comprehensive set of features. We can reach the conclusion that, in this case, smaller is better. The machine learning methods which use many features (despite these features being informative) are not better at predicting the socioeconomic level of a user than the Bayesian Method which uses only 2 simple features of the communication graph.

References

[1] Martin Fixman, Ariel Berenstein, Jorge Brea, Martin Minnoni, Matias Travizano, and Carlos Sarraute. A Bayesian approach to income inference in a communication network. In 2016 ASONAM, pages 579–582. IEEE, Aug 2016.

TB 2.3

New Item Recommendation Method Based on Latent Topic Extraction

Maria Emilia Charnelli (LINTI - New Information Technologies Research Laboratory. School of Computer Science, National University of La Plata, Argentina., Argentina), Laura Lanzarini (III LIDI - Computer Science Research Institute III-LIDI. School of Computer Science, National University of La Plata, Argentina., Argentina), Aurelio Fernández (Department of Economics, Rovira i Virgili University, Reus, Spain., Spain), Javier Díaz (LINTI - New Information Technologies Research Laboratory. School of Computer Science, National University of La Plata, Argentina., Argentina)

Corresponding Author(s): Maria Emilia Charnelli (mcharnelli@linti.unlp.edu.ar), Laura Lanzarini (laural@lidi.info.unlp.edu.ar), Aurelio Fernández (aurelio.fernandez@urv.net), Javier Díaz (jdiaz@unlp.edu.ar)

Abstract

Recommender systems are widely known for their ability to automatically assist in the decision-making process. Their operation is based on user interest patterns. The better they adapt to user preferences, the higher the quality of the recommendations will be, which will result in higher user satisfaction.

The highly increased connectivity between individuals has put these applications in the center of attention for research, since their usefulness is not limited to major decisions, but they can also be used for simple recommendations such as what items to buy or what movies to watch. These uses are not minor for business entrepreneurs.

The term "item" is used to indicate what the system recommends to its users. Nowadays, especially due to the massive use of social media, these items are extracted by analyzing and modeling text information. This is a complex task that has not been fully solved yet.

Latent topic analysis has emerged as one of the most effective methods for classifying, grouping, and recovering text data. Being able to identify the underlying topics in short texts is essential for a wide range of tasks, such as characterizing content or modeling user interests profiles. In this sense, the biterm topic model (BTM) [1] allows efficiently extracting topics that characterize a set of short texts. With BTM, through a biterm generation analysis, the underlying topics in a set of documents can be extracted, and the global distribution of each topic in each document can be established.

On the other hand, it is well known that the most commonly used approach in recommendation systems is the collaborative filtering technique based on neighborhood models. Their original form is based on similarities among users. These user-user methods calculate unknown scores based on scores recorded by users with similar ideas. Then, a similar approach that considered item similarity gained popularity [2]. In these methods, a rate or score is calculated using assessments made by users themselves in relation to similar items. An improved scalability and enhanced accuracy make the item approach better in many cases [3]. Additionally, item-item methods are more likely to explain the reasoning behind predictions. This is because users are familiar with the items previously preferred by them, but they do not know supposedly similar users. Most of the item-item approaches use a similarity measurement between item ratings.

In this article, we propose a method based on the item-item approach that uses latent topics to model the items to be recommended and establishes similarities between these items that improve recommendation performance. The method

proposed here is assessed using a dataset of movies from MovieLens and a dataset of books from Amazon.

The results obtained indicate that our method has been successful in modeling a set of items using latent topic detection through their descriptions. This allowed identifying the topics that describe the items and how they relate to each other. The methodology used in the method proposed, as well as the validation metrics applied, present successful preliminary results that are competitive versus traditional methods.

References

1. X. Cheng, X. Yan, Y. Lan, and J. Guo, "Btm: Topic modeling over short texts," IEEE Transactions on Knowledge and Data Engineering, vol. 26, pp. 2928–2941, 2014.

2. B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in Proceedings of the 10th international conference on World Wide Web. ACM, 2001, pp. 285–295.

3. R. M. Bell and Y. Koren, "Scalable collaborative filtering with jointly derived neighborhood interpolation weights," in Data Mining. ICDM. 7th IEEE International Conference on, 2007, pp. 43–52.

TB 2.4

A comparative assessment of machine learning techniques using payroll issuers data

Hugo Pérez (Universidad Iberoamericana Ciudad de México, Mexico), Jonas Velasco (CONACYT Research Felow -Center For Research in Mathematics, Mexico), Ramiro Navarro (Universidad Iberoamericana Ciudad de México, Mexico)
Corresponding Author(s): Hugo Pérez (hugo.perez@ibero.mx)

Abstract

In our previous work, we found that data mining techniques are suitable when data transaction provided by payroll data issuers from the Mexican banking system is studied. In this particular problem, loan is given to people who receive their salary payment on a payroll account. The bank retains a specific amount of customer's money, e.g. every month in order to pay his/her credit or loan; previously a contract is signed and the customer 's credit record has been evaluated as well as his credit payment capacity. If an issuer or company stops paying the payroll, every creditor anchored to that issuer will default regardless of the customer's credit quality. This relationship with the company is likely be lost and the payment omitted increases the risk of default. In fact there are many reasons a payroll issuer (company) can end the commercial relationship with the bank, for example when the bank provides bad payroll services or when the issuer goes bankrupt. Usually, many researchers interested in credit risk models are focused on studying data based on transactions of consumers or persons. However, in this work, our customers are payroll issuers and that means that the company controls the dispersion of money to the final issuer. Thus, these potential risks are studied using predictive modeling techniques or algorithms in order to understand this phenomenon.

Most popular data mining techniques like decision trees, logistic regression and neural networks, including ensemble models are still used to model this behavior in practical issues; in our previous investigation, we proposed two approaches, one based on predictive modeling and other based on credit scoring models. We found that decision trees are better than both logistic regression models and ensemble models, and can improve depending on business objectives. In recent years many methods from machine learning regularly improve the performance, particularly in classification tasks. So, in this work we test machine learning classifiers in order to continue the prediction if a payroll issuer will abandon the relationship in the next six months; this allows the decision maker to determine the appropriate business retention actions in order to avoid future payment loan losses. We compare some classification algorithms when they use individual classifier like a Support Vector Machine (SVM) with radial basis kernel function and multilayer perceptron Artificial Neural Network (ANN), classification models from homogeneous ensembles like Random Forest (RF) and Gradient Boosting (GB) algorithms as a different way of combining models. These methods were selected based on Lessmann's research (2015) and on their availability in the software in this work.

In this study we concluded that recent machine learning algorithms are adequate. The Gradient Boosting method was better than RF, SVM and ANN with a low misclassification rate. However; when we used this model in industry we discovered decision makers expect an easy-to-use tool that is not necessarily highly accurate. Therefore, the business value is still, at least in this case, both a highly precise and an easy- to- use tool for the decision maker similar to the traditional credit scoring methodology.

Sessions TA 3

TA 3.1

Predictive model for selection of undergraduate applicants

Felipe Bugueno (Facultad de Economía y Negocios - Universidad de Chile, Chile), Jaime Miranda (Facultad de Economía y Negocios - Universidad de Chile, Chile)
Corresponding Author(s): Felipe Bugueno (fbugueno.arcos@gmail.com), Jaime Miranda (jmirandap@fen.uchile.cl)

Abstract
In Chile, there were changes in legislation such as the Education Financing Law, a sustained increase in university admission by secondary students, as well as the phenomenon of university dropout, which, in our country is around 22% according to the Chilean Ministry of Education. These facts have generated a high level of competition between universities for those students with good scores in their application test results.

Therefore, the selection tools for secondary students that show a good academic performance in the future play an important role. The objective of this paper is to develop a predictive model using Data Mining, to determine the profile of outstanding secondary students who apply to the Faculty Business and Economics of the Universidad de Chile.

Literature provides several studies on the variables that influence academic performance, the first variables used were the IQ, personality and context, secondary scores, emotional intelligence, interpersonal relationships and the level of adaptation to the environment, and finally parents' concern. From Data Mining, there are large studies that predict certain results in education-related situations; for example, the acceptance of candidates or the classification according to performance in university students.

The methodology of this paper is the Knowledge Discovery in Database (KDD). Four solution approaches are used with prediction models with different levels of complexity, using variable selection techniques, clustering and model combination techniques. The selection techniques used are Sequential Backward Elimination and Decision Tree relevance. The machine learning techniques are support vector machine, logistic regression and decision Tree. The results found by this paper have implications on the variables used to identify good students.

Classification models reach up to 80.4% of accuracy, with an identified rate of interest class up to 100% or 75%. The best model is called Combined Sequential approach, and in particular it uses techniques of variable selection and a combination of techniques to obtain accuracy of 80.4% with an accuracy rate in outstanding applicants of 89.4 %. As for the variables that stand out because of their influence on the prediction are: school grades, female sex, 19 years old or less, the language score and the GPA score, as well as if the student works, if he graduated from high school the same year that took the University Selection Test, if use a financial credit to fund the university, and the differences of scores between the students with the average of its own school. The implications of these results question the current importance of certain selection variables, of which, the most important is the change of relevance of math score because it is the current focus for the best secondary applicants nowadays.

The results indicate that only the math score should be high enough to enter the University, and after that, other variables become relevant such as the language score. The results show that it is possible to make an effective prediction using KDD, as well as to integrate this predictive model into the current selection process of the Faculty of Economics and Business to support and improve the current process of attracting secondary applicants to the university. In turn, new variables are recommended for the selection of outstanding applicants and the attraction of them. Many studies with important results are observed when using additional variables in the selection stage, such as university pre - placement tests, talent tests or psychological tests. The best mechanisms for selecting students as a country should continue to be explored, as the development of tools by universities to enrol the best students.

TA 3.2

Forecasting individual course demand in higher education institutions using Machine Learning

Giancarlo A. Acevedo (Universidad de Chile -- Center for Mathematical Modelling, Chile), Salvador Flores (Universidad de Chile -- Center for Mathematical Modelling, Chile), Jorge Amaya (Universidad de Chile -- Center for Mathematical Modelling, Chile), Pablo Huentelemu (Universidad de Chile -- Center for Mathematical Modelling, Chile)
Corresponding Author(s): Giancarlo A. Acevedo (gacevedo@dim.uchile.cl), Salvador Flores (sflores@uchile.cl), Jorge Amaya (jamaya@dim.uchile.cl), Pablo Huentelemu (pablo.huentelemu@cmm.uchile.cl)

Abstract
Course schedules in Higher Education Institutions (HEI) are prepared well in advance of the start of the academic period. For this reason, an accurate prediction of enrollments is a central issue for it can make a big difference in terms of efficiency in the use of scarce resources such as classrooms or labs and in the assignment of limited highly specialized human resources.
Despite the apparent need for academic demand forecasting in HEIs, there exists a very limited bibliography on the subject, mostly out-of-date, and there are not adequate software systems to cope with those needs.

Presently, each HEI solves this problem locally, using custom procedures that are often costly and require lots of manual intervention to make key decisions during the forecasting process, making the procedure non-reproducible and extremely dependent on the expert person with a trained eye.

The goal of this work is to model this multifactorial forecasting problem, for then devising algorithms to be implemented in a predictive computational system. It seeks to give a holistic view applicable for any HEI, considering not only deterministic factors, but also inherently random ones to provide a solution adapted to the heterogeneity of the HEIs and their idiosyncrasy.

Predicting individual course enrollments is a difficult problem because of high variability among institutions. Even inside institutions, each department or major has freedom to construct its own academic program, which can include courses of different type, such as compulsory, elective, optional and complex combinations of requisites of pretty different nature such as number of credits and previous courses of a particular type approved; other complexities include internships requisites that can span over more than one academic period and frequent changes of the academic programs, just to mention some. The uncertainty in student retention, course pass rates and the inclusion of elective and optional courses introduces the inherent randomness of student's choices and performance.

From a methodological viewpoint, we work under the assumption that students have freedom to choose the courses to take as far as they fulfil the requirements, and that they do that independently of other students. The first modelling step consists in identifying, looking at actual databases, the relevant variables driving the phenomenon. We retained as relevant information individual student information along with their academic trajectories, that is, the list of the courses taken, approved, and

reproved with their respective grades since they entered the HEI.

The proposed algorithm is an ensemble method combining Random Forests, Support Vector Machines and Logistic Regression, in a stratified way depending on the course characteristics and the student type and academic program. In this way, we estimate, for each student/course pair the probability of that student to take that course. Then we use these probabilities to simulate future registration processes and predict the aggregated demand for each course as well as measures of variability and sensitivity.

We present results obtained on real instances of the problem, which are very accurate. Indeed, most mismatches correspond to database inaccuracies such as mislabeled courses or ambiguity in the determination of requisites.

Finally, we comment on the applicability of this type of models to analyze patterns of behavior for other relevant variables in a HEI, such as desertion and reiterated reproval.

TA 3.3

Educational Retention Program: Data Mining Techniques for improving performance

Jonathan Vasquez (Department of Management Control and Information Systems, School of Economics and Business, University of Chile, Av. Diagonal Paraguay 257, 8330015 Santiago, Chile, Chile), Jaime Miranda (Department of Management Control and Information Systems, School of Economics and Business, University of Chile, Av. Diagonal Paraguay 257, 8330015 Santiago, Chile), Sebastián Maldonado (Faculty of Engineering and Applied Sciences, University of Andes, Chile, Chile)
Corresponding Author(s): Jonathan Vasquez (jovasque@fen.uchile.cl)

Abstract
Student attrition is understood as the failure or early dropout of an individual to complete a program. This could be voluntary or involuntary and implies a set of cost, such as financial, personal and familial frustration, and reduction of future revenues [1]. From a social perspective, when an individual does not graduate, the country loses a professional that might generate benefits to society [2]. Education Ministry of Chile has published a set of reports made by its Research Center where they estimated a 50% of student attrition in the higher education systems [3].

Chile becomes involved in a structural change of its educational system. The government has pushed to full-funding at all educational levels and making student-attrition reduction a more important objective for educational institutes. In fact, these have included initiatives to their educational projects that increase retention rate, and most investigators have showed interest on doing research on this topic [5, 6]. However, few data has raised about the effectiveness and efficiency reducing student dropout of these initiatives. We implemented data mining techniques in a Chilean university's business school for identifying those students that participate in a program which objective is increase retention for the first year.

Academic Support Program is an initiative of Undergraduate School, giving academic support to student who are historically susceptible to dropout mainly for academic reasons. The program started in 2011 with 278 students invited to participate (241 accepted and most of them were freshmen), and last year, 2016, the program sent 390 invitations and 215 students accepted. Gradually the number of invitations has increased, but the accepting rate has decreased. The group of students that participated in the program has a lower rate of first-year dropout than those did not (5.7% in case of participants and 10% for non-participants). These figures, and the annual cost per individual participating in the program (around US$ 600) suggest the implementation of tools that would improve the effectiveness and efficiency of the support program.

We built models that combined balancing data techniques, classification models, and thresholding of classifiers, and it was applied a process of optimization parameters for each combination of these techniques. According to accuracy performance, the best model is composed by RUS balancing data techniques, SVM, and 0.2 of thresholding. This reached an accuracy of 85% and True Positive Rate of 53%, and the most important factors are: enrolling in tutorials of math and economic subjects, familiar background (who is the family boss, number of members working, number of members, family income, number of parents alive), and whether spent time working before getting higher school. As was expected, these factors are different of previous works [7] due that most of dropouts in this research are because of academic reasons and the studied group has different features, mainly low score in PSU (Higher Education Selection Test, PSU by its name in Spanish).

Sessions TB 3

TB 3.1

Collecting and Analyzing Customers Experiences from Trip Advisor Social Media

Sebastian Maldonado Alarcon (Universidad de los Andes, Chile), Carla Marina Vairetti (Universidad de los Andes, Chile), Guillermo Armelini Wilde (ESE Universidad de los Andes, Chile)
Corresponding Author(s): Sebastian Maldonado Alarcon (smaldonado@uandes.cl), Carla Marina Vairetti (cvairetti@uandes.cl), Guillermo Armelini Wilde (garmelini.ese@uandes.cl)

Abstract
Sentiment analysis or opinion mining aims to determine the attitude of people with respect to some topic. For example, businesses always want to find consumer opinions about their products (or services). Likewise, potential customers also want to know the prior users' opinions before they purchase a product (or use a service).

Social networks remain under a rapidly growing, facilitating the creation and exchange of information, ideas, comments and interests through virtual communities and networks. However, finding and monitoring opinion sites on the Web is still a complicate task due to the number of sources of information available, the incorporation of euphemisms and also the human analysis of text information is subject to different ways of interpreting words with their own preferences.

Limited research has been conducted applying social media analysis in hospitality research. In particular, TripAdvisor is a travel website that provides rich travel-related information from the consumer experience.

The purpose of this paper is to use TripAdvisor analysis to explore diner perceptions of Chilean restaurants. Helping, for example, to improve the quality of service and avoid the leakage of potential customers.

Using 40,212 comments referring to Chilean restaurants, this research present a corpus of client experiences using natural language processing tools, which are used to divide comments into sentences, extract nouns and adjectives, and generate analysis trees and dependency trees for each sentence.

In order to calculate, analyze and interpret popular words and emotional states of each comment, techniques of text mining, machine learning and sentiment analysis were used. The corpus can be used in future studies to extract meaningful knowledge of clients' experiences from their perspectives. Conclusions and recommendations, to improve the quality of service, are provided in this work.

TB 3.2

An Approach to Identify Cohesive Subgroups of Banks in Bank-Firm Networks

Samrat Gupta (Indian Institute of Management, India), Pradeep Kumar (IIM Lucknow, India)
Corresponding Author(s): Samrat Gupta (fpm14008@iiml.ac.in)

Abstract

In financial systems, the interactions manifested due to credit linkages are supposed to be potential sources of systemic risk. For instance, the financial market turmoil in 2007 followed by insolvency of several global investment banks such as Lehmann Brothers, Bear Sterns, and Merrill Lynch exposed the entwined nature of banking systems. Due to these catastrophic incidents, governments and policy makers have put a lot of effort at understanding the hidden risk in complex interconnected banking systems [1].

The links between banks are mainly formed through interbank lines of credit [1]. Heterogeneity in bank credit network arises due to business loans from banks to firms/companies. These credit relationships between banks and firms fuels the business growth of firms, and garners interest margin as profit for banks. For firms, multiple borrowing relationships ensures them against liquidation risk. For banks, multiple lending relationships hedges them against firms' risk of failure. However, the tendency to form single or multiple relationships varies with external and internal conditions [2]. On the other side, decrease in credit supply by banks affects firms unfavorably, and firm failure deteriorates banks' lending capacity. Since the probability of contagion within a subgroup is high in credit networks [1,3], detection of cohesive subgroups (generally known as "communities") can be used to investigate, intervene and prevent the transmission of contagious shocks in sub-structures to the global financial system.

In this work, a bipartite bank-firm credit network is first modelled using the concepts of rough set theory-a robust mathematical approach for modelling uncertainty and vagueness [4]. Then a topological operation named neighborhood upper approximation (NUA) is used to expand the granules and a measure based on relative linkage is used for constraining the subsets of NUA iteratively. The main contributions of this work are two-fold:

1) We propose a rough set based community detection approach for complex networks. The approach is illustrated on a two-mode benchmark network. Experiments and comparison with state-of-the-art methods reveal the superiority of proposed approach.
2) We scrape data from the publicly available annual reports of twenty heavily indebted Indian firms and map the credit network of these firms and their bankers. The proposed community detection method is applied to this network of 20 firms and 65 banks to expose the cohesive subgroups of banks in Indian banking sector

Experiment on bank-firm credit network revealed seven cohesive subgroups of banks in Indian banking sector. Results clearly indicate the extent of connectedness among banks in India. ICICI, SBI, Union and Corporation banks are the most critical banks for Indian banking system and if a firm defaults to any of these banks, credit burden will spread to almost all the banks. Some other banks with multiple memberships likes BNP, United Bank, LIC and BoI will also lead to significant contagious shocks in case of default. Banks such as RBL and Bank of America are relatively isolated and won't propagate the adverse effects in case of default. They will also be relatively unaffected if default occurs in other bank(s). This work has important policy implications for managing crisis and designing prudential regulations.

References

[1] L. Bargigli and M. Gallegati, "Finding communities in credit networks," Economics, vol. 7, no. 1, 2013.

[2] G. De Masi, Y. Fujiwara, M. Gallegati, B. Greenwald and J.E. Stiglitz, "An analysis of the Japanese credit network," Evolutionary and Institutional Economics

Review, vol. 7, pp. 209-232, 2011.

[3] R. Iyer and J.-L. Peydro, "Interbank contagion at work: Evidence from a natural experiment," Review of Financial Studies, vol. 24, 2011.

[4] Z. Pawlak, "Rough sets," International Journal of Computer and Information Sciences, vol. 11, pp. 341–356, 1982.

TB 3.3

Detection of suppliers' communities in e-commerce through graph analytics

Fabiola Herrera (Universidad Andrés Bello, Chile), Romina Torres (Universidad Andrés Bello, Chile), Rodrigo Salas (Universidad de Valparaíso, Chile)

Corresponding Author(s): Fabiola Herrera (fabiolaherrera@gmail.com), Romina Torres (romina.torres@unab.cl), Rodrigo Salas (rod.salas@gmail.com)

Abstract

ChileCompra is the largest e-commerce Chilean platform. Through this platform, most of the goods and services of public entities are traded, where contracts are generated from pre-agreements made in this system. One of the objectives of this platform is to make transparent and to make more efficient the contracting processes of products and services between buyers and tenderer of the state.

A tender can be awarded to more than one supplier and the supplier's participation in the tender may vary depending on the amount of the purchase order accepted for that supplier. The aim of this research is to detect communities of suppliers that participate together to win a tender whose amount is greater than 1,000 UTM (Chilean monthly tax unit). Moreover, we identify categories within the detected communities. The data for the analysis were obtained from the https://www.mercadopublico.cl/. The study will be limited to tenders whose amount is greater than 1,000 UTM and year of closure is equal to 2015. In addition, only bids will be considered, whose associated purchase orders will be accepted, obtaining a sub-set of data of 8,087 Tenders.

To obtain some insight about the data, we are going to do data analytics for visualization and graph mining. We use the Tableau Software for data exploration and understanding, and we use the Gephi for graph mining and to detect the communities.

For both the quantity and the amount of the purchase of those tenders which values bigger than 1000UTM, the exploratory analysis shows that are distributed almost evenly among the companies of different sizes with the exception of year 2013, where large companies have larger sales.

We have built a network considering the suppliers and the tenders as nodes. The tender node and the supplier node are linked whenever the company wins the tender, and we have considered the amount of the purchase as the weight on this edge. The resulting graph is analyzed with Gephi analytics toolbox. First, we have applied the multi-mode network transformation to obtain a suppliers' graph where the edges link those companies that have awarded together a tender.

We have applied a modularity class to participate the graph and to detect the communities, we have applied the Reingold Fruchterman distribution and Force Atlas to expand the graph and improve the visualization. As a result, we were able to detect 3397 communities. This communities were clustered and categorized. As a result, we were able to recognize the following categories: Medical Supplies, Pharmaceutical, Personal recruitment, Engineering, Automotive and fuel, Communications, Mining machinery, carriers among other.

For further analysis, we have subtracted the large companies for the analysis, and we knew categories are identified as food, building materials and passenger transportation. Moreover, in the visualization we can observe that the resulting communities

has a stronger modularity and are more compact than the previous case.

We can make further analyses by studying each community in specific.

FRIDAY 19

Sessions FB 1

FB 1.1

Identifying successful search patterns for improved job recommendation

Michael Reusens (Katholieke Universiteit Leuven, Belgium), Wilfried Lemahieu (Katholieke Universiteit Leuven, Belgium), Bart Baesens (Katholieke Universiteit Leuven, Belgium), Luc Sels (Katholieke Universiteit Leuven, Belgium)
Corresponding Author(s): Michael Reusens (michael.reusens@kuleuven.be)

Abstract
Job recommender systems help job seekers find a job by recommending vacancies the system believes relevant for them. This work presents the first research done on the impact of the temporal aspect of online job search on job recommendation. Our starting hypothesis is that job search is a process, during which job seekers can have evolving ideas of what vacancies are relevant for them. E.g. a specific job seeker might have no clue what jobs he is interested in at first, preferring a wider set of job recommendations that can inspire him/her. Over time, the job seeker will hopefully get an idea of the type/location/… of jobs (s)he is interested in, preferring a narrower set of job recommendations that fall within the range of that idea. Still, if a suitable job is not found over a longer period of time, the job seeker might be open to a wider range of alternatives. Intuitively there likely exist several of such search strategies. By modelling these job search strategies, we can use them to predict what the current stage of a job seeker's job search process is, and adapt job recommendations accordingly. Similar temporal dynamics in user preferences have been studied in the context of online music recommendation: in one listening session someone might be open to discovering new music, whereas in another session (s)he could only be interested in known favorites, making novel recommendations a poor choice. For online music listening, this has shown to improve recommendation.
Since the job search process is quite complex, and hard to compare to listening to online music, the scope of this research goes beyond predicting the novelty for a single session. The research goals of this work are threefold: First, we model the search process of job seekers, with focus on job diversity, novelty and reciprocity. Second, we investigate the link between search process and unemployment duration, to see which search strategies are more effective than others. Third, we propose changes to existing recommendation strategies that take into account the personalized search process of job seekers.
We perform our experiments on data gathered via a job search engine provided by the Flemish public employment services during the period 2015 – 2016. Our data contains approximately 33,000,000 clicks covering 250,000 job seekers and 950,000 vacancies. We aggregate this data into sessions, storing for each session the diversity (job type, location, etc.), novelty, and reciprocity of vacancies looked at. Furthermore, we link demographic information and unemployment duration to the search behavior. We employ a two-step approach to model the existing job search processes. First, we perform a clustering on the sessions to define a state-space of session types (an example session-type could be "session during which vacancies are looked at that have a highly diverse job type, but in a very small location radius"). Secondly, we look at how job seekers transition between session-clusters over time. In order to model common transition paths, we use various techniques, such as hidden Markov models, sequence- and process mining techniques.
Our results show that there is a clear difference in job search processes between individual job seekers, indicating there is much to be gained from tailoring recommender systems towards the job search process. We present a strategy on how to tune existing job recommender systems to best consider each job seeker's personal evolution. We believe that this work is not only

valuable w.r.t. improved job recommendation, but that its results can also be used to improve job seeker counseling, potentially leading to a more efficient job search experience and a reduced unemployment duration.

FB 1.3

Beyond clickthrough rate: measuring the true impact of personalized e-mail product recommendations

Stijn Geuens (Digipolis, Belgium), Koen W. De Bock (Audencia Business School, France), Kristof Coussement (IESEG School of Management, France)
Corresponding Author(s): Koen W. De Bock (kdebock@audencia.com)

Abstract
Recommendation systems constitute a nowadays widely-applied personalisation instrument for purposes of e-commerce and digital marketing. While their merits are universally assumed, empirical evidence on the true impact of personalized product recommendations on business outcomes is scarce. This study presents an elaborate field test of the impact of product recommendations, obtained through a variety of alternative algorithms, on e-commerce performance. Experiments are set up in a context of e-mail marketing, in close collaboration with a large European online retailer. The study distinguishes itself from prior research in multiple ways. While many studies evaluate recommenders in function of theoretical measures like F1 measure, in this study business outcomes are measured throughout the entire purchase funnel. A large set of well-known web analytics and e-commerce metrics with high practical adoption are considered, including clickthrough rate, conversion metrics, e-commerce outcomes such as value per visit, and average order value. Second, multiple algorithms are compared. Special attention is given to alternative hybridization strategies, a practice now commonplace in recommendation system approaches. Hybrid techniques accommodate multiple types of input-data, including customer, product and behavioural data. Two common hybridization methods are compared: algorithms are adopted and compared: factorization machines and SVM-based a-posteriori weighting. Further, value-conscious techniques that optimize expected revenue and purchase probability, are considered as well. Finally, the study presents convincing evidence on the value of implementing alternative recommendation systems on global company revenue. The main findings of the study are the following: (i) personalized product recommendations have a substantial, positive impact on metrics throughout the purchase funnel, (ii) hybridization matters, (iii) different techniques optimize different purchase funnel phases. This study offers a framework for practitioners for choosing a recommendation system in function of relevant business outcomes.

FB 1.4

Exploring online travel reviews using data analytics: an exploratory study

Vera Migueis (FEUP, Portugal)
Corresponding Author(s): Vera Migueis (vera.migueis@fe.up.pt)

Abstract
The information provided by online traveler reviews is becoming a key element in the decision-making process of hotel customers, reducing the uncertainty and the perceived risk of a traveler. A typical online customer review briefly describes the most memorable features of the overall experience, either positive or negative, as well as displays an aggregated rating, an easy-to-process information. As such, a careful analysis of the content provided by online customer's reviews might give invaluable information concerning the key determinants, from a user's perspective, of the quality of the service provided, justifying the attributed service's rating.
The objectives of this study are two-fold: (1) use text mining techniques to analyze the user's generated content automatically collected from hotels in Porto in a certain period of time, and, from this analysis, derive the most frequent terms used to describe the service; (2) understand if it is possible to predict the aggregated rating assigned by reviewers based on the terms used, and,

at the same time, identify the terms showing high predictive capacity.
Our study attempts to use innovative text and data mining tools to explore in an easy and timely way the wealth of information provided by user generated content, opening up new opportunities for re-design and improvement of hotel services.
Text mining tools are used to compute the frequency of the words used in online traveler reviews and to create a word cloud synthesizing the topics mostly expressed. A data mining classification technique, i.e. Random Forests is used to support the prediction model.
The data set used to test the validity of the model consisted roughly in 3 thousand online guests reviews taken in September 2015 from TripAdvisor, referring to 132 hotels in Porto. From this study we could conclude that aspects linked to the facilities (such as bed and rooms), staff, price and location are the ones most frequently addressed by TripAdvisor users in their online reviews. It was curious to note that the set of most informative words regarding the link with customers' rating does not actually correspond to this set of most frequent words, as was initially anticipated. Nevertheless, the underlying topics appear to be similar, i.e. facilities, staff and location. It is important to highlight that the prediction model proposed is able to predict the rating class assigned by guests with an average accuracy of about 60%, showing the potential of this method in determining critical terms linked with the end customer's overall satisfaction, when comparing with a random prediction.
The promising results obtained demonstrate the potential of the user generated data to support service management in the hotel industry. In fact, customers' increasing use of social media and digital technology may deeply change the way service providers evaluate and operate to improve their service levels. Given the possibility of easily deriving key topics/concerns behind customer perceptions, managers can promptly respond to negative feedbacks by overcoming the service limitations emphasized. The topics linked to positive reviews may also be used by hotel managers to gain competitive advantage, either by working them more intensively or as a way to leverage hotels' image towards potential guests.

Sessions FA 2

FA 2.1

Real-time Pedestrian Detection and Tracking in a Multicamera System

Roberto Muñoz (Metric Arts, Chile), Roberto Gonzalez (Metric Arts, Chile), Alejandro Sazo (Metric Arts, Chile), Patricio Cofre (Metric Arts, Chile)
Corresponding Author(s): Patricio Cofre (pcofre@metricarts.com)

Abstract
I. Introduction
The recognition and tracking of pedestrians using computer vision methods have greatly improved during the last decade. The state-of-the-art detectors have missrates lower than 10% and there is no sign of saturation. Although major improvements in object recognition, there are sill challenges related to the development of accurate real-time detections, the use of uncalibrated cameras and the detection of pedestrians with severe occlusions. In this work, we propose and implement an hybrid CPU/GPU architecture for developing a real-time surveillance system.

II. Methods
Most of the pedestrian detection and tracking methods can not be applied to real situations such as surveillance systems due to the high computational cost involved. A typical surveillance camera has a framerate of 20 FPS and a resolution of 2 Mpx, and its real-time processing is challenging for any of the modern detectors. According to several authors, a minimum of 8 FPS is necessary for doing a reliable tracking of pedestrians.

A typical modern CCTV system consists of more than 10 cameras, which can not be processed in real-time using a

state-of-the-art deep Learning system because hardware limitations. Furthermore, most the current detectors were not developed to optimise the combined use of CPU and GPU, quite common in modern architectures, which restricts even more the raised problem.

III. Software architecture

The proposed architecture consists of optimising and parallelising the computing intensive stages of the workflow. Our approach allows to transform any camera into a sensor and store structured information about pedestrians in a relational database. In most of the cases, demographic and biometric information can not be computed in real-time because of hardware limitations, so the raw data is transmitted to a remote server and computed in the cloud.

The main components of the system are,

1) Stream Decoding: The capture process is made using GPU acceleration.

2) Detection: Several methods were tested, among which stand out LBP, HOG and SSD. A parallel version of HOG was developed in CUDA in order to speed up the detection.

3) Tracking: A novel hierarchical framework is proposed for tracking, which includes track repair, classic Kalman filter, color information through RGI colorspace, histograms and track speed.

4) Data Management: Detections and tracks are structured data and can be stored in a relational database. Data structure is optimized for minimal storage usage.

IV. Discussion

The tracking results strongly depend on the detection stage. The main advantage of our modular architecture approach is that many detection methods can be easily tested and deployed. Camera location and orientation strongly affects the performance of the system, so hardware and software should be jointly designed.

The architecture allows linear scaling with the number of cameras by distributing the load in multiple GPUs. A crucial limitation to increase the number of cameras is the network bandwidth (each camera stream 4Mbps of data), so a Gigabit network is mandatory for any large application.

The main results are presented to the final user using dashboards and graphs. We are currently developing a responsive web platform to improve the user experience (visit http://video.metricarts.com/en.html).

V. Conclusion

It is feasible to perform computation with multiple sources simultaneously at great speed by prioritizing the essential tasks and leaving more complex tasks working asynchronously. The GPU as a processing unit allows us to optimize the use of resources and adding processing capability by a fair cost. Database design allows connection with external applications in order to extract even more information in an easy and transparent way using well developed systems for this offline tasks.

Acknowledgements: This project is supported by CORFO projects No. 15COTE-46317 and 15DESF-48635-11

FA 2.2

Optimal Selling of a Commodity via Forward Markets in a Cash-and-Carry Trade

Behzad Ghafouri (University of Western Ontario, Canada), Matt Davison (University of Western Ontario, Canada)
Corresponding Author(s): Behzad Ghafouri (b.ghafouri@gmail.com)

Abstract

We are motivated by studying and optimizing the off-shore oil storage trade observed in contango markets (i.e. upward sloping futures curve), where crude oil is bought cheap on the spot, sold using a forward contract at a higher price, and stored in a tanker until delivery. This type of strategy is known as Cash-and-Carry arbitrage in general, and Contango-and-Carry trade in the case of crude oil. The trade can be profitable if the gap between the forward and spot price is higher than the costs associated with the storage. We are mainly interested in the dynamic decisions the trader makes to sell the oil by taking optimal short positions in forward contracts. At each time step, the trader has the option to sell the oil on the spot, or adjust the maturity of his short forward contract, or do both on partial quantities. Undertaking these actions successively until the oil is sold or the problem deadline is reached generates a sequence of cash flows, the sum of which will define the total profit from this trade. According to the assumptions of the Cash-and-Carry trade, the quantities of the short forward contract must always be equal to the existing inventory.

The storage trade was prevalent during the super contango of late 2008 to early 2009 when oil prices hit a low point. For instance, on Feb 12, 2009, there was a very steep $21.97/barrel 12-month contango between March-2009 and March-2010 futures contracts. In the recent decade, there has been a shift towards an upward sloping curve compared to the previous decade, which makes the subject trade more attractive. The slope of the forward curve is essentially a measure which captures the change in value between two delivery points in time. If the slope is greater than the storage cost, there is incentive to long the front-end by buying oil on the spot and short the far-end by selling a forward contract. So, the spread between, rather than the absolute value of, the forward prices plays the critical role in this argument. To simulate the crude oil forward curve, the Schwartz-Smith model is used here, where the logarithm of the spot price is represented as the sum of a short-term mean-reverting factor, and a long-term one.

The optimal sale of a commodity (crude oil) stored in a rental tanker over a finite horizon is studied. The possible trading strategies are studied via two different approaches. One method is Forward Dynamic Optimization, also known as Rolling Intrinsic method, which is a simple heuristic (suboptimal) approach. The second approach formulates the problem based on the Dynamic Programming (DP) framework, which leads to an optimal solution. At each time step, the quantity of inventory sold at the spot, and the forward contract maturity to sell the remaining inventory are decided. Using the short/long-term Schwartz-Smith price model, the problem is formulated and analyzed as a Markov Decision problem. Two different algorithmic strategies are used to solve the resulted DP equations; one algorithm uses an exact method, which is based on discretization on the state space. Another algorithm uses an Approximate Dynamic Programming (ADP) technique, which is based on the estimation of continuation values by the least-squares Monte Carlo regression. Both exact and ADP optimal policies are investigated. The value of dynamic selling in the forward market and adaptively adjusting the forward maturities are compared to the initial spot selling. The future work is proposed by relaxing some of the underlying assumptions, and incorporating a dynamically changing tanker rent.

FA 2.3

Visualizing and Analyzing the 2017 Elections from a Public Perspective

Cristóbal Huneeus (Ministerio del Trabajo, Chile), Sebastián Acuña (Unholster, Chile)
Corresponding Author(s): Cristóbal Huneeus (chuneeus@gmail.com), Sebastián Acuña (sebastian@unholster.com)

Abstract

Visualizing the Chilean election results 2017 offers several challenges. On the one hand, the country's elongated geography and its concentrated population distribution mean that simply overlaying the results on a map presents readability and perceptual issues. Additionally, the 2017 elections include methodological changes to the parliament elections as well as a change in the number of Congressmen and Senators, were preceded by volatile polls, and included voting abroad for the first time.
All these factors present an enormous number of questions that any non-expert user would like to answer: Which second-round candidates are Sánchez' first-round voters inclined towards? Does participation increase from round one to two? Does the candidate Guillier enthuse the center-left? What are the key territories in the election? How do these results compare with other years? Who did not vote? How did crossover voting take place? How did voters of third-place political forces vote in the second round?

All this would have been impossible to know instantly without the use of software specially designed to understand the voter registry. These are just some of the questions that we answered in seconds and then brought to print [1]. The common practice in Chile is to show results in tabular format (http://www.servelelecciones.cl/) or otherwise as graphs that do not offer additional insight (https://goo.gl/yemjif , https://goo.gl/J69dDX, https://goo.gl/8ekAD1).
These approaches do not allow users to explore, compare and obtain their own conclusions. They are designed for simple consumption of simple information and not for analysis.
We chose to put our technological expertise at the service of the community by creating DecideChile.
Our idea was to create a digital platform that would inform in real time the results of the presidential elections of Chile and at the same time keep a historical database online of the elections of our country since the return to democracy. This is how DecideChile was born: the first electoral intelligence platform in Chile.
DecideChile is a web platform that provides several visualizations that allow analysis, understanding, and summarization of election results. We present visualizations for first and second round, as well as the design decisions behind these and future works.

References

[1] Huneeus, Lagos, Diaz (2015): Los Dos Chiles. Editorial Catalonia.

Sessions FB 2

FB 2.1

Finding vehicle theft patterns using association rules and text mining

Cristian Aguayo (Universidad de Chile, Chile), Richard Weber (Universidad de Chile, Chile)
Corresponding Author(s): Cristian Aguayo (cristian.aguayo.q@gmail.com), Richard Weber (rweber@dii.uchile.cl)

Abstract

Vehicle theft is a major problem worldwide, both for car owners and insurance companies. Furthermore, stolen cars are often used to commit other criminal acts. Different kinds of modus operandi can be observed, ranging from stealing parked cars to violent assaults.

It is known that crime in general is not random, but follows certain patterns. The same holds for car theft, where these patterns are not static in the temporal dimension, nor in the spatial dimension. When a theft pattern is discovered in a certain location, it may "stay" for some time in that location, and then, it may "move" to another location. Emerging or disappearing patterns are further expressions of such a dynamic behavior.

For car owners, insurance companies, and police it would be important to receive the most current information on such theft patterns as soon as possible in order to take the respective preventive decisions. Unfortunately, the information currently received by car insurers has several drawbacks. It covers only those vehicles that are insured, i.e. it is not representative for all car thefts. It contains unstructured information, such as the victim's narration. It arrives late to the insurer's database, in some cases days after the theft occurred.

We present a data mining model that is part of a car theft observatory where relevant information from crime reports, social networks, and news from the press are loaded and prepared for an advanced car theft analysis tool.

We used a database provided by an insurance company, with reports from about 7,200 cases of vehicle theft from 2015. For each case we had, among others, the vehicle's model, date, time, and location where the theft occurred, as well as a narrative by the victim, providing details of the theft.

To detect crime patterns, we used association rules, where each theft is considered as a "transaction", and the "item set" is a set of keywords associated to the robbery. These words are vehicle's model, date and place of the robbery, and keywords from the victim´s narration. In order to avoid finding uninformative association rules, we created a set of keywords which subsumed different communes, the vehicles´ models, and other kinds of relevant words such as "weapons", "knives", and "parked", among others.

Applying this model to the available cases, we found several interesting patterns that have been confirmed by the insurance company. Among the most relevant results of this analysis, we revealed a robbing pattern that had a lot of media coverage during the second half of 2015 in Chile: the so-called "portonazos", which are thefts with intimidation and/or violence executed at the entrance of domiciles. Other interesting association rules also appeared, e.g. that violent assaults are performed in groups, and other theft patterns such as intercepting the driver while the car is stopped at a traffic light.

As future work we propose more advanced text mining, such as topic modeling; applying social network analysis, state-of-the-art data mining techniques, and advanced visualization tools.

FB 2.2

Obtaining and evaluating extractive summaries from stored text documents

Augusto Villa-Monte (Institute of Research in Computer Science LIDI, Faculty of Computer Science, National University of La Plata, Argentina), Laura Lanzarini (Institute of Research in Computer Science LIDI, Faculty of Computer Science, National University of La Plata, Argentina), Aurelio Fernández-Bariviera (Department of Economics, Rovira i Virgili University, Spain), José A. Olivas (Department of Information Technologies and Systems, University of Castilla-La Mancha, Spain)
Corresponding Author(s): Augusto Villa-Monte (avillamonte@lidi.info.unlp.edu.ar)

Abstract

Current technology allows recording and storing all types of information. For years, text data has grown exponentially and represents one of the most valuable resources in all fields. In the scientific field, a myriad of articles is constantly produced and hosted on the cloud. These documents encompass all types of topic areas and represent the human knowledge. Today, much of these documents are available, but it is impossible for a human analyze all of them. This requires application of automatic text processing techniques that reduce this huge volume of non-structured information to its most important content, separating what is essential from what is not to facilitate its treatment and exploitation. This allows obtaining the core contents of a document in less time than of what it would take to do it manually. This task is known as automatic summarization generation and can be of two types: abstractive and extractive.

An extractive summary is formed by the set of sentences of the document appropriately selected and the abstractive by the ideas developed in the document without using the sentences exactly as they appear in the original document. In both cases, the goal is reducing the size document while keeping its information without human intervention. The first summarization approach is quite used in literature since it does not necessarily require a complex semantic analysis of the text. Thus, this can be use as previous step to the second one, making a first reduction of the document size that allowing to approach the expected human summary. However, not for this reason, this approach is more trivial. The extractive approach not only requires select parts of a document but also to assign a specific score to each one allowing to order them from highest to lowest in a ranking where the top positions are more relevant.

This relevance value or score is obtained by calculate a metric according its rating criterion. The metrics range from identifying certain expressions within the text to more complex calculations. There are many works on extractive summaries in which documents are modeled as n-dimensional vectors of numerical features obtained by calculating n metrics. This feature vectors are used to obtain a automatic summary by applying to these vectors other algorithms more sophisticated. However, in very few works the way in which those features were calculated is developed in depth. Design a program that selects representative phrases from documents automatically requires precise instructions.

This work aims to show the impact of the pre-processing performed on the documents when calculating the metrics. To this end, 30 medical documents in XML format from the free access journal PLOS were used to calculate different variants of a metric. Python was used for document download and their processing, as well as for calculate of metrics and comparing the obtained summaries to the author abstracts. The Porter stemming algorithm and English stopword list of NLTK package were used. To assess summary quality, ROUGE-1 and ROUGE-2 measures was used because they are frequently used in literature. For the tests, a database was designed that store the text documents facilitating calculations through SQL queries. This design saves processing time, clearly expresses how the metrics is calculated, and facilitates experimentation in this research area.

The tests carried out showed that with variations in the calculations of each metric, differences are obtained in the values of ROUGE if they are analyzed document by document or all together. It is expected that incorporating a semantic analysis to the calculation of the metrics, can be corrected these variations in the results.

FB 2.3

Market Basket Analysis Insights to Support Category Management

Luis Aburto (Universidad de Chile, Chile), Andrés Musalem (Universidad de Chile, Chile)
Corresponding Author(s): Luis Aburto (luaburto@gmail.com)

Abstract

This research presents an approach to detect interrelations among product categories, which are then used to produce a partition of the retailer's business into subsets of categories. As the number of possible relationships among them can be very large, we introduce an approach that generates an intuitive graphical representation of these interrelationships using data analysis techniques. The proposed methodology is a fast and wide-range approach to study the shopping behavior of customers, detect cross-category interrelations and segment the retailer's business and customers based on information about their shopping baskets. Using transactional data, we estimated a Jaccard ratio to measure similarity or the join purchase probability of two product categories. This Jaccard ratio is used as similarity input for Multidimensional Scaling (MDS) and k-means clustering to map and group the categories. The methodology also yields a segmentation of the entire set of shopping trips based on the composition of each shopping basket. In particular, the proposed methodology was applied using one month of transactional data of a supermarket store to analyze 33 different product categories. As a result, we discover four groups of products categories that are often jointly purchased: Immediate consumption, Nonperishable, Hygiene and hedonic cluster of product categories. The study of each of these groups allowed us to conceive the retail store as a small set of sub-businesses. We make a linear projection into the map of several transaction attributes (i.e. average number of categories in tickets including a specific category, amount spend, number of transactions, among others). The regression analysis can help us

to describe and explain the positioning in the map of each category using these transaction descriptive variables. Conclusions reinforce the strategic need for proactive coordination of marketing activities across interrelated product categories. These insights can support several tactic decisions in the retail supermarket such us: role definition, promotions and pack definitions, promotional checklist, store layout and performance analysis. Compared to existing approaches, its simplicity should facilitate its implementation by practitioners. Our results suggest that retailers could potentially benefit if they transition from the traditional category management approach where retailers manage product categories in isolation into a customer management approach where retailers identify, acknowledge and leverage interrelations among product categories. It is important to note that our approach may be used beyond the particular industry under study in this paper. In this regard, this approach should be valuable for any business or activity where customers buy, hire or consume bundles of products or services from the same provider (e.g. department stores, financial institutions). This is also relevant for internet sites where visitors navigate through different pages of the same portal (e.g., espn.com) or purchase products from different categories (e.g., amazon.com).

BAFI 2018

Jan17th-19th
Santiago - Chile

ISSN 0719-8981

www.baficonference.cl

ORGANIZERS

ISCI Instituto Sistemas Complejos de Ingeniería

Universidad de Chile

SPONSORS

Facultad Economía y Negocios Universidad de Chile | DCS Departamento de Control de Gestión y Sistemas de Información

fcfm Facultad de Ciencias Físicas y Matemáticas Universidad de Chile

Ingeniería Industrial Universidad de Chile

Universidad de los Andes > Ingeniería 20 años

Unholster

METRICARTS

CF Centro de Finanzas

KDnuggets

IEEE Computational Intelligence Society