Editorial: Third special issue on Knowledge Discovery and Business Intelligence

Paulo Cortez[1] Manuel Filipe Santos[1]

[1] ALGORITMI Research Centre, Department of Information Systems, University of Minho, 4800-058 Guimarães, Portugal
Email: pcortez@dsi.uminho.pt, mfs@dsi.uminho.pt

1 Introduction

Expert Systems (ES) were proposed in the mid 1970s [Arnott and Pervan, 2014] with the goal of building computerized systems that mimic human behavior to solve real-world tasks. Such systems were based on Artificial Intelligence (AI) techniques, typically by adopting explicit (human understandable) knowledge, extracted from domain experts (e.g. by using interviews), and that was stored in a knowledge base [Buchanan, 1986].

In the last decades, the world as changed due to advances in information and communication technology (e.g. massive usage of computers and personal mobile devices, Internet and social media, usage of digital cameras and other sensors). In effect, we are now in the age of data, where a large portion of organizational, societal or personal activities is captured digitally [Mojsilovic, 2014]. Following this change, ES have evolved to include data-driven models, either solely or complemented by expert-driven knowledge. Such change is reflected in the ES journal, which currently publishes several articles related with data analysis fields (e.g. analytics and business intelligence, data mining and knowledge discovery, big data, data science).

In this special issue, we highlight two of data related terms: Knowledge Discovery (KD) and Business Intelligence (BI). KD is often used as a synonym of data mining and it consists in an AI subfield that uses machine learning algorithms to extract high-level interesting knowledge from raw data [Fayyad et al., 1996]. BI is a popular management term [Arnott and Pervan, 2014] and it represents several technologies (e.g., data warehouses, KD, dashboards) that store and process organizational data in order to support managerial decision-making [Delen et al., 2014].

The 'Knowledge Discovery and Business Intelligence' (KDBI) thematic track was proposed for the EPIA conference on AI in 2009 with the goal of promoting the interaction between the KD and BI areas. Since then, the track has been included in all EPIA biennial conferences. After 2011, the KDBI track has been associated with special ES journal issues, which included extended versions of the best KDBI papers. The first special issue was published in 2013 and it included the best KDBI 2011 track papers [Cortez and Santos, 2013], while the second special issue appeared in 2015 and it encompassed the best KDBI 2013 track papers [Cortez and Santos, 2015].

This issue, entitled 'Third special issue on Knowledge Discovery and Business Intelligence', contains recent KD and BI contributions that can be used in ES to produce a valuable impact in real-world applications. It includes extended versions of papers from the 4th KDBI thematic track, of the 17th EPIA conference on AI (EPIA 2015), held in Coimbra, Portugal. The track received 18 paper submissions and the authors of the best papers were invited to extend their works for this special issue. After two rounds of reviews, which included reviewers from the KDBI track and also ES journal, the best six papers were accepted, corresponding to an overall acceptance rate of 33%.

Due to the interest in data-driven models, in the last years there has been several interesting developments in the KDBI area. Despite this progress, there are still many challenges and opportunities. For instance, most KD algorithms were targeted for single label classification tasks and often these algorithms cannot deal adequately with label ranking, which is useful in several real-world applications (e.g. modeling user preferences). Also, the financial domain still rises many challenges: it is not clear what is the best approach (regression or classification) to forecast trading actions; and easy to interpret tools are needed to better disclose the relationships among price variations from distinct financial products. Moreover, most of the real-world data has a temporal dimension and there is still room for proposing specialized algorithms that use this dimension in the KD or BI process (e.g. time series retrieval, visual representation of temporal changes). Furthermore, the 'Extract, Transform, and Load' (ETL) is a vital component of BI systems but it often requires a substantial manual effort in terms of its design and implementation, even when there are several common ETL processes that are repeated through distinct BI projects. All these challenges and opportunities are addressed in the six papers accepted in this special issue, which we will briefly detail in the next section.

2 Contents of the special issue

In the first paper 'Label Ranking Forests', de Sá et al (2016) propose a novel KD algorithm for Label Ranking (LR), which is a classification variant task. The goal is to learn the implicit function that performs a mapping between a set of inputs, which characterize an item, and a ranking of labels (instead of a single label, as in standard classification). LR is used in several real-world applications, such as microarray analysis, image categorization or modeling user preferences. The proposed Label Ranking Forests (LRF) algorithm uses an ensemble of decision tree methods for LR and it is considered a natural adaptation of the popular random forest algorithm for LR. Several experiments were held using 16 datasets from the KEBI Data Repository. Overall, competitive LR results were achieved by the LRF algorithm when compared with LR decision tree methods.

In 'A comparative study of approaches to forecast the correct trading actions', Baía and Torgo (2016) perform an extensive comparison of two main approaches to forecast trading actions of financial markets. The first approach uses standard regression models to predict the daily variation on prices and then uses pre-defined decision rules in order to transform the numeric predictions into trading actions. The second approach uses classification models that directly forecast the trading decision

(hold, buy, sell). The data analyzed included assets prices of 12 companies, ranging from 7 to 30 years of closing prices daily data. Several machine learning models were tested (e.g. neural networks, support vector machines, random forests). The main conclusion of the paper is that there is no significant difference between the two main approaches: numeric price variation prediction and and direct classification of trading actions. The study contains two additional recommendations: re-sampling strategies (for regression or classification) are not recommended in this financial domain, even if the data is imbalanced; also, the usage of cost-benefit matrices is promising to enhance the classification models.

Also approaching the financial context, in 'Ramex-Forum: a tool for displaying and analysing complex sequential patterns of financial products', Tiple et al (2016) present an improved version of the Ramex-Forum algorithm, leading to a KD method that is capable of extracting knowledge from multivariate time series in a visual way and that is easy to interpret. The algorithm was applied in two real-world applications: petroleum production prices and risk analysis of European financial institutions. The obtained results attest the algorithm capability to show relevant price variations in financial markets.

In 'Aggressive pruning strategy for time series retrieval using a multi-resolution representation based on vector quantization coupled with discrete wavelet transform', Muhammad Fuad (2016) proposes a novel KD algorithm for time series indexing and retrieval. The algorithm adopts a multi-resolution representation of the series based on Haar wavelets and vector quantization. The experiments were conducted using 10 time series with distinct sizes and from different repositories. The proposed algorithm obtained competitive results when compared with a two other representation methods (single-resolution and other multi-resolution algorithm).

Temporal data was also studied in the paper of Gérk (2016), entitled 'Visual analytics of educational time-dependent data using interactive dynamic visualization'. The paper presents a new web-based visualization framework for BI and that is targeted for an interactive exploration by the decision maker. In particular, the framework is based on motion charts and clustering techniques, allowing to reveal changes over time in terms of two-dimensional space animations. The framework was demonstrated using real-world educational data. A total of 16 human participants evaluated the quality of the framework, confirming the proposed animations as an interesting contribution to the analytic process.

In the last paper, Oliveira and Belo (2016) approach ETL processes, which play a key role in BI by extracting data from several sources into a data warehousing system. The paper, entitled 'On the specification of ETL patterns behavior, a domain-specific language approach', proposes the use of patterns to represent common ETL tasks (e.g. surrogate key pipelining or intensive data loading). Using the business process modeling notation (BPMN), it is shown how a typical and crucial ETL process (data quality enhancement) can be improved in terms of its design and implementation.

Acknowledgments

We would like to thank the other KDBI 2015 track (of EPIA) co-organizers, Luís Cavique, João Gama and Nuno Marques. Also, we thank the authors, who contributed with their papers, and the reviewers (from the KDBI 2015 program committee and the ES journal). This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT - *Fundação para a Ciência e Tecnologia* within the Project Scope: UID/CEC/00319/2013.

References

[Arnott and Pervan, 2014] Arnott, D. and Pervan, G. (2014). A critical analysis of decision support systems research revisited: the rise of design science. *Journal of Information Technology*, 29(4):269–293.

[Baía and Torgo, 2016] Baía, L. and Torgo, L. (2016). A comparative study of approaches to forecast the correct trading actions. *Expert Systems*, pages –.

[Buchanan, 1986] Buchanan, B. G. (1986). Expert systems: working systems and the research literature. *Expert Systems*, 3(1):32–50.

[Cortez and Santos, 2013] Cortez, P. and Santos, M. F. (2013). Knowledge Discovery and Business Intelligence. *Expert Systems*, 30(4):283–284.

[Cortez and Santos, 2015] Cortez, P. and Santos, M. F. (2015). Recent advances on knowledge discovery and business intelligence. *Expert Systems*, 32(3):433–434.

[Delen et al., 2014] Delen, D., Turban, E. and Sharda, D. R. (2014). *Business Intelligence: A Managerial Perspective on Analytics*. Pearson, 3rd edition.

[de Sá et al, 2016] de Sá, C. R., Soares, C., Knobbe, A. and Cortez, P. (2016). Label Ranking Forests. *Expert Systems*, pages –.

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). *Advances in Knowledge Discovery and Data Mining*. MIT Press.

[Muhammad Fuad, 2016] Muhammad Fuad, M. M. (2016). Aggressive pruning strategy for time series retrieval using a multi-resolution representation based on vector quantization coupled with discrete wavelet transform. *Expert Systems*, pages –.

[Géryk, 2016] Géryk, J. (2016). Visual analytics of educational time-dependent data using interactive dynamic visualization. *Expert Systems*, pages –.

[Mojsilovic, 2014] Mojsilovic, A. (2014). The age of data and opportunities. IBM Big Data & Analytics Hub, `http://www.ibmbigdatahub.com/blog/age-data-and-opportunities`.

[Oliveira and Belo, 2016] Oliveira, B. and Belo, O. (2016). On the specification of ETL patterns behavior, a domain-specific language approach. *Expert Systems*, pages –.

[Tiple et al, 2016] Tiple, P., Cavique, L. and Cavalheiro Marques, N. (2016). Ramex-Forum: a tool for displaying and analysing complex sequential patterns of financial products. *Expert Systems*, pages –.

The authors

Paulo Cortez

Paulo Cortez is Associate Professor with Habilitation at the Department of Information Systems, University of Minho, Portugal. He is also coordinator of Information Systems and Technologies (IST) research group of ALGORITMI Centre with 48 PhD researchers. His research interests include: Business Intelligence (Decision Support, Data Mining and Forecasting); and Artificial Intelligence (Computational Intelligence, Neural Networks, Evolutionary Computation and Applications). Currently, he is associate editor of the *Expert Systems* journal and participated in 13 R&D projects (principal investigator in 3). He is co-author of more than one hundred indexed (ISI, Scopus) publications in international journals (e.g., Decision Support Systems, Applied Soft Computing) and conferences (e.g., IEEE IJCNN). Web-page: `http://www3.dsi.uminho.pt/pcortez`

Manuel Filipe Santos

Manuel Filipe Santos is Associate Professor at the Department of Information Systems, University of Minho, Portugal, and leader of the Intelligent Data Systems group of Centro Algoritmi, with interests in the fields of: Business Intelligence, Data Mining and Learning Classifier Systems. He participated in several R&D projects (principal investigator in 3). He is also co-author of several publications in international journals (e.g., Artificial Intelligence in Medicine) and conferences (e.g., GECCO).