Editorial: Recent Advances 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

Information Technology (IT) is advancing rapidly. In the last years, IT costs have reduced while data storage, communication and processing capabilities have increased. In effect, vast datasets are becoming commonplace. All this data hold valuable knowledge such as trends and patterns, which can be used to improve decision making and optimize chances of success. These datasets are often highly complex for a manual analysis due to the volume, variety and velocity properties of data. Following these advances, there has been a growing interest in using IT to provide useful knowledge, extracted from humans or raw data, for decision support. This interest was approached under different perspectives through the last decades [Carbonell et al., 1983, Ein-Dor and Segev, 1993, Turban et al., 2010, Piatetsky-Shapiro, 2011]:

- Machine Learning (ML), since the 1960s;
- Decision Support Systems (DSS), since the 1970s;
- Expert Systems (ES), since the 1980s;
- Analytics, Business Intelligence (BI), Data Mining (DM), Enterprise Information Systems (EIS), Knowledge Discovery (KD) in data, since the 1990s; and
- Big Data and Data Science, since the 2000s.

In particular, ES were originally focused on using symbolic information and inference in order to mimic human specialists [Buchanan, 1986]. In most cases, these ES were based on expert-driven knowledge that was extracted from the specialists (e.g., by using interviews). However, since the 1990s, there has been a shift towards the use of data-driven models and increase of DM, KD and BI fields. These data-driven models were built from past data in order to produce informed predictions or descriptive knowledge that could be useful for supporting decisions. Following

this trend, data-driven knowledge, used either solely or complemented by expert-driven knowledge, turned into a key element of modern ES, highlighting the importance of the KD and BI areas. KD is an Artificial Intelligence subfield that is related with the extraction of useful and understandable knowledge from raw data [Fayyad et al., 1996], while BI is an umbrella term that represents several technologies (e.g., Data Warehouses, KD and Dashboards) to access past data and support decision-making [Turban et al., 2010].

Since the 1990s, there has been a rapid expansion of the KD and BI fields in terms of both research contributions and their use in real-world applications. This special issue, entitled 'Recent Advances on Knowledge Discovery and Business Intelligence' focuses on novel KD contributions that present a valuable impact in several BI domain applications. It consists of extended versions of papers from the 3rd Knowledge Discovery and Business Intelligence (KDBI) thematic track of the 16th Portuguese Conference on Artificial Intelligence (EPIA 2013) that was held in Angra do Heroísmo, Açores, Portugal. A total of 18 papers were submitted to the 3rd KDBI thematic track, from which the best 9 papers were invited for this special issue. Each extended paper was reviewed by a minimum of three reviewers (related with both the 3rd KDBI thematic task and Expert Systems journal), and passed through two rounds of reviews. Finally, the best four papers were accepted, corresponding to an acceptance rate of 22%, when considering the initial 3rd KDBI submitted papers, and 44%, when considering the extended invited papers. The next section briefly introduces this special issue accepted papers.

2 Contents of the special issue

In the first paper 'Contrast set mining in temporal databases', Magalhães and Azevedo (2014) extend the Rules for Contrast Sets (RCS) technique for handling temporal data mining tasks. The goal of a Contrast Set is to discover attribute-value pairs that differ in two or more groups. The presented extension proposes a set of temporal patterns that can capture significant differences in Contrast Sets discovered through time. Two distinct real-world problem domains, related with Portuguese labor and NBA basketball data, were used to demonstrate the capabilities of the RCS extension for a temporal analysis, revealing interesting patterns.

In 'Feature selection for clustering categorical data with an embedded modeling approach', Silvestre *et al.* (2014) present a novel approach that simultaneously clusters categorical data and selects relevant features. The approach is based on a Gaussian mixture model, where the Minimum Message Length (MML) criterion is used to guide the selection of the relevant features and a modified Expectation-Maximization (EM) algorithm estimates the model parameters. The usefulness of the proposed approach was illustrated on synthetic datasets and real-world data related with the perceived quality of life from the European Official Statistics (EOS).

In the work 'Eigenspace method for spatiotemporal hotspot detection', Fanaee-T and Gama (2014) deal with hotspot detection, which aims at the identification of unexpected subgroups, in space-time data. The proposed EigenSpot method tracks changes in space-time occurrences and it is potentially useful for surveillance systems (e.g. bio-surveillance). The experiments held adopted synthetic and real data. The

latter data was related with brain tumor cases in 32 subregions of the New Mexico State, United States, from 1973 to 1991. The obtained results have shown that EigenSpot compares favorably against the state-of-the-art STScan method.

In the last paper, entitled 'Re-sampling strategies for regression', Torgo *et al.* (2014) present a general re-sampling approach for regression tasks, which can be used for forecasting rare values of continuous target variables. Their approach is based on modifications of two popular re-sampling classification techniques (under-sampling and SMOTE). An extensive set of experiments was conducted using 18 datasets from several domains (e.g., automotive industry), highlighting the advantages of the proposed re-sampling methods.

All these papers represent the best contributions to the 3rd 'Knowledge Discovery and Business Intelligence' track of EPIA 2013. We hope that the contributions of this special issue enrich both KD and BI, promoting the interaction between both fields.

Acknowledgments

Many individuals contributed to this special issue. We would like to thank Dr. Jon G. Hall, Editor-in-Chief of Expert Systems, for his interest and support. We also wish to thank the other KDBI 2013 track (of EPIA) co-organizers, Luís Cavique, João Gama and Nuno Marques, without their help this issue would not have been possible. Finally, we thank the authors, who contributed with their papers, and the reviewers (from the KDBI 2013 program committee and also others), who contributed with their detailed comments and valuable opinions. This work was supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope: PEst-OE/EEI/UI0319/2014.

References

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

[Carbonell et al., 1983] Carbonell, J., Michalski, R. and Mitchell, T. (1983). Machine Learning: A Historical and Methodological Analysis. *Artificial Intelligence Magazine*, 4(3):69–79.

[Ein-Dor and Segev, 1993] Ein-Dor, P. and Segev, E. (1993). A classification of information systems: Analysis and interpretation, Information Systems Research, 4(2):166–204.

[Fanaee-T and Gama, 2014] Fanaee-T, H. and Gama, J. (2014). Eigenspace method for spatiotemporal hotspot detection. *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.

[Magalhães and Azevedo, 2014] Magalhães, A. and Azevedo, P. (2014). Contrast set mining in temporal databases. *Expert systems*, pages –.

[Piatetsky-Shapiro, 2011] Piatetsky-Shapiro, G. (2011). Poll Results: What do you call analyzing data? `http://www.kdnuggets.com/2011/09/the-top-name-for-data-mining.html`.

[Silvestre et al., 2014] Silvestre, C., Cardoso, M. and Figueiredo, M. (2014). Feature selection for clustering categorical data with an embedded modeling approach. *Expert systems*, pages –.

[Torgo et al., 2014] Torgo, L., Branco, P., Ribeiro, R. and Pfahringer, B. (2014). Re-sampling strategies for regression. *Expert systems*, pages –.

[Turban et al., 2010] Turban, E., Sharda, R., Aronson, J., and King, D. (2010). *Business Intelligence, A Managerial Approach*. Prentice-Hall, 2nd edition.

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. 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* and *Neural Processing Letters* journals and participated in 9 R&D projects (principal investigator in 2). He is co-author of more than eighty 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).