Editorial: Fourth 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) are a core element of human decision making. Initially, in the 70s and 80s, ES were focused on extracting explicit knowledge from human experts. With the availability of big data, after the 2000s, ES incorporated data-driven models, thus being associated with business intelligence, big data, data science and machine learning systems [Cortez and Santos, 2017]. The importance of data-driven models in the ES area is confirmed by the recent Wiley's Expert Systems (EXSY) literature survey that analyzed all journal research articles published from 2000 to 2016 [Cortez et al., 2018]. The survey revealed data-driven as the most prevalent ES method type, corresponding to around 35% of all recently published EXSY papers.

The first 'Knowledge Discovery and Business Intelligence' (KDBI) track was held at the EPIA conference on Artificial Intelligence in 2009, with the goal of strengthening the interaction between Knowledge Discovery (KD) and Business Intelligence (BI). Both are important data related topics. KD is the subfield of Artificial Intelligence that is focused on the extraction of human interesting knowledge from raw data [Fayyad et al., 1996]. BI is an umbrella term that encompasses methodologies and technologies (e.g., data warehousing, dashboards) to support managerial decision-making [Delen et al., 2014]. Following the success of its first 2009 edition, the KDBI workshop become a regular track of the biannual EPIA conference, with its fifth edition taking place at the 18th EPIA conference, held at Porto, Portugal, in September 2017. Since 2011 the track had a dedicated special EXSY issue [Cortez and Santos, 2013, Cortez and Santos, 2015, Cortez and Santos, 2017].

This is the Fourth special issue on Knowledge Discovery and Business Intelligence and it includes extended versions of the best papers presented at the 5th KDBI thematic track of EPIA 2017, which received 20 paper submissions. The EXSY special issue included two rounds of reviews, which involved reviewers from the 5th KDBI track of EPIA2017 and also the EXSY journal. After the revision stage, a total of five papers were accepted to be published in this issue, which corresponds to an acceptance rate of 25%.

While in the last decade there have been remarkable developments in the KD and BI areas, there are still challenges and opportunities. For instance, most KDBI

research has been focused on processing structured data but little has been devoted to Mining Software Repositories (MSR), which is quite valuable for the software engineering industry. Also, feature selection is a challenging task in the context of large genomics datasets. Some domain applications, such food truck recommendation, produce difficult multi-label classification tasks. Another challenge comes from the development of online recommendation systems, which are relevant for high velocity data and big data streams. Moreover, imbalanced domains are important in several real-world applications, such as medicine and finance, often prejudicing the predictive performance of the data-driven models. These challenges are approached in the five accepted papers published here. In the next section, we summarize the main contributions of these papers.

2 Contents of the special issue

The first paper, entitled 'Analysis of a Token Density Metric for Concern Detection in Matlab Sources using UbiSOM' is related with the MSR field. Marques et al. (2018) perform an exploratory data analysis, based on the Ubiquitous Self Organising Map (UbiSOM), to detect software concerns (i.e., cohesive set of functionalities enclosed in a module) in the Matlab language. The UbiSOM technique was illustrated using a repository with 35,000 Matlab files and the visual analytics results revealed interesting concern co-occorrence patterns.

In 'A Bi-objective Feature Selection Algorithm for Large Omics Datasets', Cavique et al. (2018) propose a bi-objective version of the Logical Analysis of Inconsistent Data (LAID) algorithm. The bi-objective approach considers both the predictive accuracy and the result comprehensibility. The approach was tested using omics datasets with genome-like charateristics of patients with rare diseases, which are quite relevant for the medicine domain and that contain a high number of genetic features and very small number of samples. A huge reduction was achieved by the bi-objective approach, from one million features to just three sub-datasets with 19 features each.

The third paper, 'Enhancing multi-label classification for food truck recommendation', by Rivolli et al. (2018), addresses a popular Brazilian application, food truck recommendation, which is a complex multi-label classification task. Using data from a market research and that involved hundreds of consumers, the paper explored several data-driven methods, including the proposed Ensemble of Single Label method, which reduced misclassification of under-represented labels.

Vinagre et al. (2018) propose an 'Online bagging for recommender systems'. The paper presents three online bagging variants for recommendation tasks in the context of data streams. Several experiments were conducted using four distinct datasets (e.g., Movielens-1M, last.fm). The obtained results show that the user-based bagging variant produces the best overall accuracy.

The last paper 'Resampling with Neighborhood Bias on Imbalanced Domains', by Branco et al. (2018), addresses imbalanced classification and regression tasks. In particular, the authors propose resampling techniques that use information from the neighborhood of the data samples, reinforcing some regions of the training sets. An extensive set of experiments, using 16 classification and 18 regression tasks,

confirmed the value of the proposed resampling neighborhood bias strategies.

Acknowledgments

We would like to thank the other KDBI 2017 track (of EPIA) co-organizers: Albert Bifet, Luís Cavique and Nuno Marques. Also, we thank the authors, who contributed with their papers, and the reviewers (from the KDBI 2017 program committee and the EXSY 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

[Branco et al., 2018] Branco, P. and Torgo, L. and Ribeiro, R.P. (2018). Resampling with Neighborhood Bias on Imbalanced Domains. *Expert Systems*, pages –.

[Cavique et al., 2018] Cavique, L. and Mendes, A.B. and Martiniano, H. and Correia, L. (2018). A Bi-objective Feature Selection Algorithm for Large Omics Datasets. *Expert Systems*, pages –.

[Cortez and Santos, 2013] Cortez, P. and Santos, M. F. (2013). Knowledge Discovery and Business Intelligence. *Expert Systems*, 30(4):283–284. doi: 10.1111/exsy.12042.

[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. doi: 10.1111/exsy.12087.

[Cortez and Santos, 2017] Cortez, P. and Santos, M. F. (2017). Third special issue on knowledge discovery and business intelligence. *Expert Systems*, 34: e12188. doi: 10.1111/exsy.12188.

[Cortez et al., 2018] Cortez, P. and Moro, S. and Rita, P. and King, D. and Hall, J. (2018). Insights from a text mining survey on Expert Systems research from 2000 to 2016. *Expert Systems*, e12280. doi: 10.1111/exsy.12280.

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

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

[Marques et al., 2018] Marques, N. and Monteiro, M.P. and Silva, B. (2018). Analysis of a Token Density Metric for Concern Detection in Matlab Sources using UbiSOM. *Expert Systems*, pages –.

[Rivolli et al, 2018] Rivolli, A. and Soares, C. and de Carvalho, A.C. (2018). Enhancing multi-label classification for food truck recommendation. *Expert Systems*, pages –.

[Vinagre et al., 2018] Vinagre, J. and Jorge, A.M. and Gama, J. (2018). Online bagging for recommender systems. *Expert Systems*, pages –.

The editors

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* and *Decision Support Systems* journals and participated in 16 R&D projects (principal investigator in 8). He is co-author of more than 140 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 with Habilitation 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., ACM GECCO).