Machine Learning (2018) 107:1647–1649
https://doi.org/10.1007/s10994-018-5749-6

Introduction to the special issue on discovery science

Michelangelo Ceci[1,2] · **Toon Calders**[3]

Received: 4 July 2018 / Accepted: 18 July 2018 / Published online: 31 July 2018

Welcome to the Discovery Science special issue. This issue contains both extended papers from the Discovery Science 2016 conference, held in Bari, Italy (19–21 October 2016), as well as new contributions solicited by an open call. Discovery science is a research discipline spanning multiple areas including advances in the development and analysis of methods for discovering scientific knowledge coming from machine learning, data mining, and intelligent data analysis, as well as their application in various scientific domains including, but not limited to, biomedical, astronomical, physics and social sciences. Applications to massive, heterogeneous, complex, continuous or imprecise data sets are of particular interest for the discipline.

We received twenty-five diverse submissions showing the liveliness and the breadth of this field. Of the received submissions, we eventually selected eight for inclusion in this special issue. The accepted articles have undergone two or three rounds of rigorous peer-reviewing according to the journals high standards. The accepted contributions encompass a wide range of research topics, from purely methodological to highly applied, such as hierarchical multi-label classification, graph mining, anomaly detection, learning from unbalanced data streams, learning autoencoders and graph classification, thereby appealing to both the experts in the respective fields and those who want a snapshot of the current breadth of topics covered by discovery science. The accepted articles are briefly summarized below.

The paper "A comparison of hierarchical multi-output recognition approaches for anuran classification" by Colonna et al. presents a bioacoustic recognition framework based on hierarchical classification to recognize several species of anurans. The paper encompasses the different steps of the task: signal decomposition, feature extraction, learning and evaluation. Experiments show significant classification rates for the specific task at hand.

Breskvar et al. present in their paper "Ensembles for multi-target regression with random output selections" a novel multi-target regression algorithm based on ensembles of predictive clustering trees (PCTs), including bagging and random forests of PCTs and extremely ran-

✉ Michelangelo Ceci
michelangelo.ceci@uniba.it

Toon Calders
toon.calders@uantwerp.be

[1] Department of Computer Science, University of Bari Aldo Moro, Bari, Italy

[2] CINI - National Interuniversity Consortium for Informatics, Rome, Italy

[3] Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium

domized PCTs. Through extensive experimentation and comparisons with several competing methods, the performance of the ensembles of PCTs is thoroughly tested.[1]

The work "Reservoir of Diverse Adaptive Learners and Stacking Fast Hoeffding Drift Detection Methods for Evolving Data Streams" by Pesaranghader et al. focuses on adaptive learning from evolving data streams. The paper introduces the TORNADA framework, in which multiple diverse classifiers and drift detection algorithms are executed in parallel. The intuition behind this work is that the best (classifier, detector) pairs are not only heavily dependent on the characteristics of a stream, but also that this selection evolves as the stream flows. The TORNADO framework further incorporates the novel FHDDMS drift detection methods, where different sized sliding windows and fading factors are utilized to detect heterogeneous drifts in a timely fashion.

In the paper "On Analyzing User Preference Dynamics with Temporal Social Networks" by Pereira et al., the authors investigate the interplay between user preferences and social networks over time. Specifically, they use temporal networks concepts to analyze the evolution of social relationships and propose strategies to detect changes in the network structure based on node centrality. The proposed method also identifies correlations between preference change events and node centrality change events.

In their paper "Discovering a Taste for the Unusual: Exceptional Models for Preference Mining", de Sa et al. show empirically how Exceptional Preference Mining (EPM) can be used in problems where the target concept can be represented as a preference of a set of labels (such as in rankings or pairwise comparisons). EPM is a cross-over between local pattern mining and preference mining, aimed at finding subsets of observations where some preference relations between labels significantly deviate from the norm. Experiments show that EPM can find interesting new knowledge.

The paper "Targeted and contextual redescription set exploration" by Mihelčić and Šmuc tackles the problem of exploring sets of redescriptions. Redescriptions are multiple descriptions of the same, or similar subsets of entities in the data which are useful for understanding target labels and their connection to different subsets of features, feature selection, model selection, and ultimately feature engineering. The paper presents a redescription set exploration methodology and tool that use various types of information derived from the available redescription set as a whole to enhance the exploration process. Evaluation, performed on three use-case datasets, show that deriving additional information from the redescription set truly helps in selecting the subsets of redescriptions connected with some research hypothesis.

"Probabilistic Frequent Subtrees for Efficient Graph Classification and Retrieval" by Welke et al. contains several efficient techniques for using frequent subtrees as features in applications such as graph classification. Sampling of spanning trees is used to alleviate the intractable subtree isomorphism problem in a regular graph. Also the related problem of computing embeddings of graphs in feature spaces spanned by tree patterns is discussed. Several novel and elegant ways to achieve efficiency gains, including exploiting subtree relationships and using a minhashing approximation technique are introduced. Empirical evaluations show that these techniques can tremendously reduce the number of evaluations of subtree isomorphism as compared to a standard brute-force algorithm.

Nolle et al. show in their paper "Analyzing Business Process Anomalies Using Autoencoders" an interesting application of deep learning to business process management. They propose a method, using autoencoders, for detecting and analyzing anomalies occurring in the execution of a business process. The method is resilient to noise and does not require

[1] Due to a conflict of interest with one of the guest editors, this paper was handled by one of the editors only (Toon Calders).

background knowledge of the process to be applied. Extensive experimentation and comparisons with several state-of-the-art methods show excellent performance of the proposed autoencoder-based technique.

We believe this special issue presents a diverse and interesting set of papers which we hope you will enjoy reading. We would like to thank all authors for their contributions. Also the reviewers deserve a big thank you for the high-quality reviewing and the many suggestions that were made during the reviewing process which improved the quality of many of the papers significantly.

Acknowledgements Funding was provided by European Commission (Grant Nos. ICT-2013-612944 MAESTRA and H2020-ICT-688797 TOREADOR).