Machine Learning (2020) 109:1993–1995
https://doi.org/10.1007/s10994-020-05922-3

Guest editors' introduction to the special issue on Discovery Science

Larisa Soldatova[1] · **Joaquin Vanschoren**[2]

Published online: 20 October 2020

1 Introduction

Over the years, the Discovery Science conferences have provided a forum for discussion of the latest and most innovative research in machine learning, data mining, knowledge discovery in databases, and their use in a wide range of applications. This special issue on Discovery Science is focusing on the development and analysis of methods for discovering scientific knowledge, coming from machine learning, data mining, intelligent data analysis, semantic web, big data analysis as well as their application in various scientific domains.

The special issue on Discovery Science includes several papers that were presented in preliminary form at the 21st International Conference on Discovery Science in October 2018. It also includes papers that were submitted following the open call for papers for this special issue. The Editors received a total of 25 submissions, out of which 9 have been selected for this special issue. All papers have been thoroughly reviewed following the rigorous scholarly standards of the *Machine Learning* journal. Papers in this special issue report results of the analysis of different types of massive and complex data, including structured, spatio-temporal and network data.

1.1 Summaries of the selected papers

The paper titled '**Evaluating time series forecasting models: an empirical study on performance estimation methods**' by Cerqueira et al. provides an extensive comparative study, evaluating the performance of several univariate time series forecasting models. Similar comparisons have been made before, but the authors argue that the previous studies have been biased towards cross-validation and also towards stationary time series. Cerqueira et al. apply a wide set of methods to real-world scenarios and synthetic case studies. More specifically, they consider 174 real-world time series from different domains of application, including finance, physics, economy, energy, and meteorology, 97 of which are stationary and 77 are non-stationary. The results of their study suggest that blocked cross-validation can be applied to stationary time series. However, when the time series

✉ Larisa Soldatova
l.soldatova@gold.ac.uk

[1] Goldsmith, The University of London, London, UK

[2] Eindhoven University of Technology, Eindhoven, The Netherlands

are non-stationary, the most accurate estimates are produced by out-of-sample methods, particularly by the holdout approach repeated in multiple testing periods. While this study focused on univariate numeric time series with a high sampling frequency (for example, hourly or daily data), the authors believe that the conclusions of the study should extend to other types of time series.

Z.S. Abdallah and M.M. Gaber in their paper '**Co-eye: a multi-resolution ensemble classifier for symbolically approximated time series**' present a new nature-inspired classification technique for time series. The technique, named Co-eye, draws biological inspiration from the compound eyes of flies, made up of thousands of lenses that together create a broad field of vision. Co-eye employs different lenses created through hyper-parameterisation of symbolic representations to look at the time series, and then combines them for a broader vision. The authors present the combination of random forest ensembles built using different representations of time series data at different resolutions. Co-eye has the advantage of combining various lenses together using hyper-parameterisation in order to decide the best lenses for accurate classification based on cross-validation over training data. The proposed technique has been verified on 114 publicly available time series datasets. The experimental results show the benefits of Co-eye in bringing together different perspectives. The authors illustrate the performance of Co-eye on a case study addressing the problem of distinguishing between an outline of a beetle and an outline of a fly, where it achieves very high accuracy.

M. H. Chehreghani, in his work titled '**Unsupervised representation learning with minimax distance measures',** proposes a general-purpose computationally efficient framework for employing minimax distances suitable for a wide range of machine learning methods working with numerical data. Typically, data is described by a set of objects and a corresponding representation with an accompanying distance, like Euclidean distance, Mahalanobis distance, and Pearson correlation. However, in real-world applications, the data is often so complex that such representations might fail to correctly capture the underlying patterns and structures. Minimax measures are better in capturing patterns in complex arbitrarily shaped data. A minimax measure selects the minimum largest gap among all possible paths between the objects. The advantages of the proposed method are shown on a variety of synthetic and real-world datasets. Real-world experiments were carried out on twelve datasets from different domains, and the proposed approach provided results that are often better than those shown by traditional methodologies. M. H. Chehreghani concludes that the minimax variant is more appropriate for low dimensional data, whereas the dimension-specific minimax variant performs better on high-dimensional data.

Khandagale et al., in their work **Bonsai: diverse and shallow trees for extreme multi-label classification,** introduce a suite of algorithms, named Bonsai, generalizing the notion of label representation in extreme multi-label classification, involving hundreds of thousand or even millions of labels. Bonsai partitions the labels in the representation space to learn shallow trees. Learning shallow trees improves the prediction accuracy by preventing error propagation in the tree cascade. The main advantages of the approach reported by Khandagale et al. is in the ability of Bonsai to retain the training speed, while achieving better prediction accuracy and better tail-label coverage. The authors demonstrate that, on a benchmark dataset with 3 million labels, Bonsai outperforms a state-of-the-art one-vs-rest method in terms of prediction accuracy, while being approximately 200 times faster to train.

A. Osojnik et al. contribute to this issue their work on **Incremental predictive clustering trees for online semi-supervised multi-target regression.** Since labeling data examples is costly and particularly problematic in an online setting, they turn to a semi-supervised approach to leverage unlabeled examples. They propose a method for online semi-supervised

multi-target regression, based on incremental trees for multi-target regression and the predictive clustering framework. Their proposed method, named iSOUP-PCT, is compared against several alternative state-of-the-art methods, fully-supervised tree methods, and an oracle. It outperforms state-of-the-art methods in scenarios with very few labeled examples, while achieving comparable performance when the labeled examples are more common.

M. Petkovic et al. present their work on **Multi-label feature ranking with ensemble methods.** They propose three ensemble-based feature ranking scores for multi-label classification, and demonstrate empirically that the proposed ranking scores outperform current state-of-the-art methods in the quality of the produced rankings as well as in their time efficiency. They also show that one of these scores, Genie3, proves to be the best performing score overall, based on the quality of the rankings first and—in the case of ties—time efficiency.

B. Škrlj et al. propose a novel method for **Embedding-based silhouette community detection.** This is an approach for detecting communities in networks, based on clustering of network node embeddings, i.e. real valued representations of nodes derived from their neighborhoods. Extensive experiments on synthetic and real-world networks show that this method performs comparably or better than state-of-the-art community detection algorithms, such as the InfoMap and Louvain algorithms. The authors also demonstrate that the method's outputs can be used along with domain ontologies in semantic subgroup discovery, yielding human-understandable explanations of communities detected in a real-life protein interaction network.

O. Orhobor et al. contribute novel work on **Predicting rice phenotypes with meta and multi-target learning.** This work leverages the existence of inherent groupings of dataset features, such as genomic data, where features can be grouped by chromosome. The authors present a meta-learning framework in which a series of base-learners is used to produce new features which are then integrated by a meta-model. They also compare this approach to multi-target learning, given that one is typically interested in predicting multiple phenotypes. Empirical results show that both the meta and multi-target approaches significantly outperform the base learners, yielding better predictions for rice phenotypes.

Finally, D. Kocev et al. in their paper titled '**Ensembles of extremely randomized predictive clustering trees for predicting structured outputs'** extend the Extra-Tree ensemble learning method to address three structured output prediction tasks (multi-target regression, multilabel classification, and hierarchical multi-label classification) by using predictive clustering trees (PCTs) as base models in the ensemble. The experimental evaluation reveals that this method, called Extra-PCTs, outperforms or ties with other ensemble methods in terms of predictive power and computational cost, and can also be used to learn good feature rankings for all of the tasks.

Acknowledgements As editors of this special issue, we thank all the authors who submitted papers, the Program Committee members of the Discovery Science 2018 conference, and all of the reviewers of the submitted papers, for their excellent work. We are grateful for help in putting this special issue together received from Melissa Fearon and Subhashini Gopal from Springer, as well as the Machine Learning journal editors, i.e. the special issue editor, Dragos Margineantu, and the editors-in-chief, Prof. Peter Flach and Prof. Hendrik Blockeel.