Mach Learn (2017) 106:741–743
DOI 10.1007/s10994-016-5625-1

EDITORIAL

Special issue on discovery science

Nathalie Japkowicz[1] · **Stan Matwin**[2,3]

Received: 17 November 2016 / Accepted: 30 December 2016 / Published online: 27 April 2017

Discovery Science is a discipline located at the cusp of Machine Learning, Data Mining and Data Analysis, which is concerned with the discovery of scientific knowledge embedded in complex data. This special issue takes the reader through the entire Discovery Science Process, from data collection, to algorithm design, learning theory considerations, and performance evaluation. It surveys a large variety of diverse application domains including insurance, recommender systems, web-based news, gene sequencing, currency movement prediction, and music.

This special issue was organized following the 18th International Conference on Discovery Science, held in Banff, Alberta (Canada) from October 4th to October 6th, 2015. We received nineteen submissions and accepted seven, three of which are extensions of papers presented at the conference: "Multi-label Classification via Multi-target Regression on Data Streams", "Stream-based Semi-supervised Learning for Recommender Systems", and "An Evaluation of Linear and Non-Linear Models of Expressive Dynamics in Classical Piano and Symphonic Music". As a reflection of the prominence of data streams in today's problems, four of the seven papers focus on that issue, creating new methods for multi-label classification, semi-supervised learning, high utility sequential pattern (HUSP) mining, and clustering in the streamed data context. One of these papers also includes a fair amount of methodology as it presents, among other themes, the very pressing issue of fusing heterogeneous sources of data together and discusses the problems that typically emerge from such fusion. Two other papers discuss techniques that do not operate in the data stream context. In particular, one of them is a theoretical study that introduces a careful approach for binary classification that aims at minimising an upper bound on the future misclassification rate. The second one is

✉ Nathalie Japkowicz
nathalie.japkowicz@american.edu

Stan Matwin
stan@cs.dal.ca

[1] Department of Computer Science, American University, Washington, DC, USA

[2] Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

[3] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

an applied paper which proposes a series of linear, nonlinear and recurrent non-linear basis models to capture musical expression. The last paper is fully methodological and makes the interesting argument that many statistical tests, such as the now widely used Friedman test, are not appropriate for Machine Learning, and proposes an alternative to such tests. We now summarize each of these papers in more detail.

In the article entitled "Multi-label Classification via Multi-target Regression on Data Streams" by Aljaž Osojnik, Panče Panov and Sašo Džeroski, the authors follow an approach different from the traditional one used in both batch and streaming data, which consists of transforming the multi-label problem into a number of independent binary classification problems. Instead, in order to preserve the label inter-correlation, they transform the multi-label classification problem into a multi-target regression problem which they found is closely related. Their framework allows them to handle multi-label classification as well as hierarchical multi-label classification.

In the article entitled "Stream-based Semi-supervised Learning for Recommender Systems" by Pawel Matuszyk and Myra Spiliopoulou, in order to respond to an extreme data label sparsity, the authors introduce a co-training approach in which several learning algorithms run in parallel and share their reliable predictions with each other so that these predictions can be used as labels. This research is conducted in the framework of recommender systems and continuously adapts to changes in preferences and new information. A new evaluation protocol is proposed that splits data sets in an unbiased way and uses corrections for multiple statistical tests.

In the article entitled "Memory-Adaptive High Utility Sequential Pattern Mining over Data Streams" by Morteza Zihayat, Yan Chen and Aijun An, the authors look at the problem of high utility sequential pattern (HUSP) mining in the data stream context. HUSP are frequent sequential patterns of particular interest such as, for example, patterns that include very profitable items. There are many domains in which data comes as a data stream and HUSPs need to be discovered in a single scan. These include web clickstream mining, network traffic analysis and intrusion detection. The authors propose an efficient tree-based structure to store candidate HUSPs. They use that memory adaptive structure in a stream-mining algorithm. They apply their approach to a web clickstream dataset obtained from a major digital news portal, and to a microarray dataset.

In the article entitled "Big Data: From Collection to Visualization" by Mohammed Ghesmoune, Hanene Azzag, Salima Benbernou, Tarn Duong, Mustapha Lebbah and Mourad Ouziri, the authors focus on two problems: the methodological problem of how to collect, represent, merge and enrich heterogeneous data sets and the problem of clustering and visualizing streaming data. To do so, they develop a big data platform and they adapt a growing neural gas algorithm (an algorithm related to Self-Organizing Maps) to enable it to handle data streams. They apply and test their approach on a massive set of insurance data.

In the article entitled "High-Probability Minimax Probability Machines" by Simon Cousins and John Shawe-Taylor, the authors focus on building a classifier that minimizes the upper bound on the expected misclassification rate on future instances. In particular, they seek to improve an algorithm named the Minimax Probability Machine which makes the strong assumption that the means and covariance matrices are known in advance. The algorithm they propose corrects for that assumption by adding a measure of uncertainty. Therefore, instead of looking at the worst case as in the original Minimax Probability Machine, they look at the probable worst-case. They test their new algorithm on a large variety of UCI domains as well as in the domain of currency movement prediction, showing that their approach is competitive, especially in the case where small training sets are used.

In the article entitled “An Evaluation of Linear and Non-Linear Models of Expressive Dynamics in Classical Piano and Symphonic Music” by Carlos Eduardo Cancino Chacon, Thassilo Gadermaier, Gerhard Widmer and Maarten Grachten, the authors propose an approach to model emotional expression in music. In addition to being interesting in and of itself, the practical aim of such research is to improve a computer’s rendition of musical scores. Since expressive dynamics is created from a number of different combined factors, a previously proposed approach consisted of modelling these factors using different basis functions and combining them together. The previously proposed combination, however, was linear. In this work, the authors propose to expand this combination to a nonlinear combination. Their experiments reveal that the nonlinear combination allows their system to capture interactions between the various components of expression better than the linear model.

In the article entitled “Confidence Curves: An Alternative to Null Hypothesis Significance Testing for the Comparison of Classifiers” by Daniel Berrar, the author contends that Null Hypothesis Significance Testing by tests commonly accepted by the research community such as the Friedman Test are not appropriate for comparing multiple classifiers over multiple domains. Instead, he argues, confidence curves which put a greater emphasis on how large a difference in performance has been observed are better suited. The article presents an extensive critical review of the way in which statistical tests are currently used and interpreted in the field of Discovery Science, and argues for an alternative approach. He demonstrates how confidence curves can be used in a series of experiments performed on UCI domains.

To summarize, this Special Issue gives the reader a spectrum of papers, from purely methodological to highly applied, with an emphasis on stream data—the type of data in which the Machine Learning community is currently very interested. Three of the papers originated in the Discovery Science 2015 conference, the others are the original submission for this Special Issue. We believe they all bring new content and results to the field, and that they will meet with the interest of the readers.