Data Min Knowl Disc (2014) 28:1129–1133
DOI 10.1007/s10618-014-0374-x

EDITORIAL

Guest editors' introduction: special issue of the ECML/PKDD 2014 journal track

Toon Calders · Floriana Esposito · Eyke Hüllermeier · Rosa Meo

Received: 9 July 2014 / Accepted: 14 July 2014 / Published online: 2 August 2014

This special issue is a collection of papers that were submitted to the ECML/PKDD 2014 journal track and have been accepted for publication in "Data Mining and Knowledge Discovery".

The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML/PKDD, launched its journal track last year in 2013. In order to cover the full scope of the conference, which is a merger of the formerly independent conferences ECML and PKDD, two journals were involved: "Machine Learning" and "Data Mining and Knowledge Discovery". Apart from being published in the respective journal, papers accepted in the journal track are also presented at the conference, just like the contributions to the regular conference

Responsible editor: Geoffrey I. Webb.

T. Calders (✉)
Department of Computer and Decision Engineering, Faculty of Applied Sciences, Université Libre de Bruxelles, 50, Av. F. Roosevelt, CP 165/15, B-1050 Brussels, Belgium
e-mail: toon.calders@ulb.ac.be

F. Esposito
Dipartimento di Informatica, Universitá di Bari, Via Orabona 4, 70126 Bari, Italy
e-mail: floriana.esposito@uniba.it

E. Hüllermeier
Department of Computer Science, University of Paderborn, Paderborn, Germany
e-mail: eyke@upb.de

R. Meo
Dipartimento di Informatica, Facoltà di Scienze della Natura, Università di Torino, corso Svizzera 185, I-10149 Torino, Italy
e-mail: meo@di.unito.it

proceedings. Thus, all papers of this special issue will also be presented by their authors at the ECML/PKDD 2014 conference in Nancy, France, from September 15–19th.

Given the special nature of the ECML/PKDD journal track, submissions are supposed to meet specific criteria. First and foremost, like any other submission, they are of course expected to comply with the high scientific standards of the two journals. Not less importantly, however, they should naturally lend themselves to conference presentations. In other words, they are supposed to be conference and journal papers at the same time: as novel and intriguing as the former, and as substantial and mature as the latter. These requirements exclude, for example, journal versions of previously published conference papers as well as survey papers, which were not considered for the special issue.

In total, 67 original manuscripts were submitted to the ECML/PKDD 2014 special issue of "Data Mining and Knowledge Discovery" in the course of the year 2014, out of which 11 were eventually accepted in time. In addition, 5 papers from the last year's edition of the journal track were included; due to a delay in the reviewing and revision process, these papers, despite being accepted, were too late for the 2013 special issue. In what follows, we briefly summarize the contents of the accepted papers.

Been Kim and Cynthia Rudin. *Learning about meetings*. The goal of this paper is highly original. The authors discover social signals that allow to study what happens during meetings from an annotated corpus. For this purpose they use a data-driven approach, descriptive statistics, and machine learning methods. The paper is the first to give answers to questions regarding recognition of when decisions are taken during a meeting, discovery of interaction patterns in terms of dialogue acts, prediction of meeting duration from time spent, and prediction if a proposal will be accepted on the basis of the language adopted by the speaker.

Sara Hajian, Josep Domingo–Ferrer, and Oriol Farras. *Generalization-based Privacy Preservation and Discrimination Prevention in Data Publishing and Mining*. The authors of this paper show how to transform a microdata set by means of data suppression and generalization such that this transformation minimizes both the privacy invasion, occurring as a consequence of the linking between a sensitive piece of information and the identity of an individual, and the discrimination threats that result from an unfair treatment of the individual because of his/her membership to a category. At the same time, the proposed data transformation maximizes the usefulness of the original data for learning models and finding patterns. The goal is to find a good trade-off between privacy protection against discrimination and the quality of the resulting training datasets. In the end the work evaluates data quality loss incurred as a consequence of the data transformation.

Seyda Ertekin, Cynthia Rudin, and Haym Hirsh. *Approximating the Crowd*. In this work the goal is to estimate a crowd's majority opinion by querying only a subset of its members. The presented algorithm works in an on-line fashion and samples the crowd with an exploration/exploitation trade-off coming from the desire to reduce the costs but also increase the accuracy of the estimates. To this purpose the algorithm seeks to identify the best labelers whose opinions best approximate the majority. The work presents two probabilistic versions of the algorithm: one that assumes independence of the labelers' opinions and one that assumes a binomial distribution.

Pance Panov, Larisa Soldatova, and Saso Dzeroski. *Ontology of Core Data Mining Entities* This paper presents a step forward in the development of a data and knowledge exchange standard for the area of data mining. It consists in an ontology of data mining entities composed of a specification, an implementation and an application level. The entities correspond to data, tasks and algorithms. In addition the ontology includes a definition and a taxonomy of constraints, constraint-based algorithms, scenarios and work-flows. The paper discusses the connections and a mapping among the different ontologies and presents lessons learned while developing and using the presented one.

Annalisa Appice and Donato Malerba. *Leveraging the power of local spatial autocorrelation in geophysical interpolative clustering*. A time-evolving, hierarchical clustering model applied to geophysical sensors applications is presented. The clustering model accounts for both the need for data summarization on the one hand, and for the estimation of unknown data in locations of interest. For this purpose, interpolation techniques as well as the property of spatial autocorrelation of geophysical variables between nearby locations are taken into accout. The spatial autocorrelation is coupled with a multivariate analysis by maximizing the variance reduction of local indicators of the spatial autocorrelation itself.

Hiroshi Kajino, Hiromi Arai, and Hisashi Kashima. *Preserving Worker Privacy in Crowdsourcing*. This paper proposes a quality control method on crowdsourcing, a recently popular methodology that consists in outsourcing a task to a set of workers. Due to the different degrees of skill of the workers, crowdsourcing needs to estimate reliability measures from low-quality results. This process, however, might infer personal information of the workers that can lead to leaking the workers' privacy. One of the contributions of this work is the definition of the problem of worker private quality control and the introduction of a worker private latent class protocol with which it is possible to reliably estimate the crowdsourcing results while maintianing the privacy of the workers. The method applies the EM algorithm and uses decentralised secure computation. The security of the computation is theoretically guaranteed and experimentally verified.

Hoai An Le Thi and Manh Cuong Nguyen. *Self-Organizing Maps by Difference of Convex functions optimization*. Self-Organizing Maps have become a popular method for data-driven dimensionality reduction and data compression that transform an incoming signal pattern of arbitrary dimensionality into a low dimensional discrete map. One of the training techniques consists in optimizing an energy function that yields the parameters that correspond to the desired topographic mapping. In particular the optimization function is given by the sum of all the distances from an observation to the best matching neuron. This paper proposes to solve the optimization problem by the use of an optimization approach in a nonconvex programming framework, already applied with success in many domains of applied science. The results show a good trade-off between the speed and scalability of the computations achieved by a novel training algorithm that applies a cooling schedule on the one hand, and the quality of the resulting map and the classification on the other.

Hoang-Vu Nguyen, Emmanuel Müller, Jilles Vreeken, and Klemens Bohm. *Unsupervised Interaction-Preserving Discretization of Multivariate Data*. This paper proposes an information-theoretic framework for unsupervised descretization that aims at preserving interactions in multivariate data. According to the proposed technique, con-

secutive multivariate regions are combined if they are statistically similar and reduce the MDL encoding cost. Similarity is assessed by a new interaction distance and can be computed efficiently. An extensive empirical evaluation in many fields such as pattern-based compression, classification, and outlier detection shows superior performance in time and quality w.r.t. the state of the art.

Hao Wu, Jilles Vreeken, Nikolaj Tatti, and Naren Ramakrishnan. *Uncovering the Plot: Detecting Surprising Coalitions of Entities in Multi-Relational Schemas*. This paper deals with uncovering unusual coalitions of entities such as telephone numbers, people, places, events in multi-relational data. A new type of patterns, bicluster chains, that chain together strong clusters of entities over different dimensions are mined. To assess the surprisingness of the bicluster chains, the data is modeled using the maximal-entropy principle and this model is updated with each pattern found, thus assuring that the next patterns found are non-redundant with respect to the already found patterns. The authors illustrate how their method could for instance discover hidden plots in multi-relational intelligence analysis datasets.

Nikolaj Tatti. *Discovering Bands from Graphs*. Graphs often have structures in which nodes can be ordered in such a way that it is more likely that a node is related to another if they are not far apart in the order. This is for instance the case if the nodes have been created over time. The author of this paper proposes a technique to find such an order on the vertices in a graph which translates to bands around the diagonal that are more dense than the rest of the adjacency graph. For a fixed number of bands K, the most optimal order and segmentation into K bands is searched for. The goodness of a segmentation is measured using the log-likelihood of a log-linear model and the problem is divided into two subproblems: finding the order and finding the bands.

Josif Grabocka and Lars Schmidt-Thieme. *Invariant Time-Series Factorization*. This paper introduces a new representation of time series, which can capture patterns that are invariant to shifts and scales. It is assumed that time series are generated by a set of latent (hidden) patterns. The paper introduces a method which detects a set of latent patterns for a time series dataset together with a convolutional degree of membership weights. Such a decomposition is a tailored dimensionality reduction for time-series that could for instance be used in time series classification.

Aditya Telang, Deepak P, Salil Joshi, Prasad Deshpande, and Ranjana Rajendran. *Detecting Localized Homogeneous Anomalies over Spatio-Temporal Data*. An automated, domain-independent approach of exploring and discovering anomalous spatio-temporal patterns is proposed. The approach consists in the combination of two phases: discovering homogeneous regions, and evaluating these regions as anomalies based on their statistical difference from a generalized neighborhood. In contrast to existing works that analyze spatial and temporal anomalies in isolation, the approach in the paper focusses on detecting spatio-temporal anomalies within a single setting. Experiments on a climate and an "ocean-bed topography" dataset as well as a small user study shows that the approach performs better than existing state-of-the-art approaches.

Andreas Henelius, Kai Puolamäki, Henrik Boström, Lars Asker, and Panagiotis Papapetrou. *A Peek into the Black Box: Exploring Classifiers by Randomization*. In many prediction problems users do not only want to have good classification performance, but they also want to understand why certain predictions are made. Classifiers, however, are often opaque and cannot easily be inspected to gain understand-

ing of which factors are of importance. Therefore, the authors propose an efficient randomization-based algorithm for discovering how classifiers exploit associations between attributes in a dataset. The structure of the exploited attributes is presented in the form of groupings, which may provide insights into the structure of the dataset.

Jussi Korpela, Kai Puolamäki, and Aristides Gionis. *Confidence bands for time series data*. Confidence intervals are typically used to describe a univariate distribution, but the concept can be extended to multivariate time series data. The authors introduce a minimum width envelope method that can be used to compute confidence bands when several observations of a time series are available. The method is non-parametric and adjusts automatically to varying degrees of correlation within the data. A procedure to ensure that the confidence bands are such that the family-wise error rate remains controlled is introduced as well.

Orestis Kostakis. *Classy: Fast Clustering Streams of Call-Graphs*. The paper presents Classy, a scalable distributed system that clusters streams of large call-graphs for purposes including automated malware classification and facilitating malware analysts. A call-graph of an executable file is its representation as a directed graph with labeled vertices, where the vertices correspond to functions and the edges to function calls. In Classy, graph similarity is determined by a graph edit distance which is computed using an adapted version of Simulated Annealing. To speed up computation, a novel lower bound for the graph edit distance is employed. The author presents results and statistics from a real production-side system with more than 0.8 million graphs.

Esther Galbrun, Aristides Gionis, and Nikolaj Tatti. *Overlapping community detection in labeled graphs*. In this paper the authors develop a method to detect overlapping communities in labeled graphs. Such labels could for instance be occupation, hobbies, or preferences of the users of a social network. The objective in the paper is to discover a set of k communities, maximizing the total edge density over all k communities. Furthermore, each community must be succinctly described by a label set S. The authors propose and empirically test a solution inspired by a greedy algorithm for the generalized maximum-coverage problem.

The double-track publication model was introduced by the program chairs of 2013 and continued by us in 2014, in an attempt to bring the thorough and efficient reviewing process of journals to the conference context, while safeguarding the possibility to have innovative work presented at the conference at an early stage. It enables authors to immediately publish their newest results in a journal, without giving up the opportunity of presenting them at a conference. We believe that this model can result in a faster, more efficient and higher-quality review process, of which all benefit: journals, conferences, authors, reviewers, and ultimately the reader.

This special issue would not have been possible without the help of many people. We would like to especially thank the members of the ECML/PKDD 2014 "guest editorial board", as well as the additional reviewers for their hard work for timely reviewing the papers of the special issue.