Data Min Knowl Disc (2010) 21:221–223
DOI 10.1007/s10618-010-0192-8

Guest editors' introduction: special issue of selected papers from ECML PKDD 2010

José L. Balcázar · Francesco Bonchi ·
Aristides Gionis · Michèle Sebag

Received: 30 April 2010 / Accepted: 20 June 2010 / Published online: 24 July 2010

This special issue of the Data Mining and Knowledge Discovery Journal contains seven selected papers presented in the 2010 edition of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD). The ECML PKDD conference is an annual gathering that has been established as a premium forum for presenting research advances in machine learning and data mining. Combining the two sibling disciplines in one conference promotes synergy and allows researchers and practitioners of one area to benefit from and contribute to progress in the other.

This year ECML PKDD 2010 attracted 658 full-paper submissions. The submitted papers underwent a rigorous reviewing process. Each paper was assigned to three reviewers and one Area Chair. Reviewers evaluated the work on the basis of novelty, creativity, technical quality, intellectual depth, experimental evaluation, and overall impact of the work to the community. Based on the initial reviews and the author feedback collected in a rebuttal phase, the area chairs initiated discussions

J. L. Balcázar
Dept Matemáticas, Estadística y Computación, Universidad de Cantabria Santander,
Santander, Spain
e-mail: joseluis.balcazar@unican.es

F. Bonchi (✉) · A. Gionis
Yahoo! Research Barcelona, Avinguda Diagonal 177, 08018 Barcelona, Spain
e-mail: bonchi@yahoo-inc.com

A. Gionis
e-mail: gionis@yahoo-inc.corp

M. Sebag
TAO, CNRS–INRIA–LRI, Université Paris-Sud, 91405 Orsay, France
e-mail: sebag@lri.fr

among reviewers, reconciled scores, and provided a recommendation for each paper. The reviewing phase resulted in selecting 120 papers to be presented in the conference.

The articles presented in this special issue were selected for their strong theoretical contribution, their applicability in real-world problems, and their potential to spur new research in the field. The selected articles address a broad spectrum of topics, ranging from the traditional areas of itemset mining and clustering, to the newest trends of social-network analysis and game theory. The selection also reflects nicely the fusion of the borders between data mining and machine learning.

The first paper entitled "A Game-Theoretic Framework to Identify Overlapping Communities in Social Networks" by Wei Chen, Zhenming Liu, Xiaorui Sun, and Ya jun Wang, provides a fresh look in the widely-studied problem of finding communities in social networks. The community-detection problem is approached from a game-theoretic perspective, in which users are considered strategic agents who decide which communities to join. The proposed framework allows finding overlapping communities without the need to specify the number of communities or their size.

The second paper "Accelerating Spectral Clustering with Partial Supervision" by Dimitrios Mavroeidis demonstrates the use of partial supervision in order to speed up spectral clustering, but also in order to improve the quality of the solution obtained. The method is suitable for large and sparse matrices, which makes it especially attractive for practical scenarios, since large and sparse data are ubiquitous in many applications domains.

The next paper "Maximal Exceptions with Minimal Descriptions" by Matthijs van Leeuwen introduces exception maximization/description minimization as a powerful and novel algorithmic paradigm for exceptional model mining. The paper introduces an iterative scheme that alternates between two phases: local maximization of the exceptionality of the current pattern (EM—exception maximization) versus local minimization of the description complexity of the current solution (DM—description minimization). As a result, the proposed approach delivers maximally exceptional models with minimal descriptions.

With the paper "Three Naive Bayes Approaches for Discrimination-Free Classification", Toon Calders and Sicco Verwer investigate how to modify the naive-Bayes classifier in order to perform classification that is restricted to be independent with respect to a given sensitive attribute. The important task of preventing discrimination when building a classification model is intriguing and well-justified. The authors make an excellent job in motivating their work, communicating their methods, and interpreting their results.

As data mining algorithms are designed to produce results that presumably reveal structure in the data, independently on whether such structure exists or not, being able to assess the significance of data mining results is a vital problem. The next paper of our collection "Using Background Knowledge to Rank Itemsets" by Nikolaj Tatti and Michael Mampaey addresses the problem of assessing data-mining results in the context of itemset mining. The novelty of the work is to incorporate various count statistics in order build a maximum-entropy model that estimates the expected frequency of itemsets. The model is computed by the iterative-scaling algorithm. A nice property of the algorithm is that even though finding a maximum-entropy model is in general infeasible, in this particular setting the problem can be solved in polynomial time.

The next paper entitled "Mining Top-*k* Frequent Itemsets Through Progressive Sampling" by Andrea Pietracaprina, Matteo Riondato, Eli Upfal, and Fabio Vandin brings ideas from sampling and randomized algorithms to the problem of summarizing collections of frequent itemsets. Extracting top-*k* frequent itemsets is considered a viable alternative in the classic paradigm of computing all frequent itemsets. The contribution of this work is to provide an upper bound on the number of samples required in order to estimate the top-*k* frequent itemsets within a pre-specified accuracy. An algorithm based on progressive sampling algorithm is presented and analyzed. A nice feature is that the proposed algorithm not only provides a good approximation of top-*k* frequent itemsets, but it can also be used to estimate their frequencies.

The last paper in our collection "Predicting labels for dyadic data" by Aditya Menon and Charles Elkan introduces a new and interesting problem called the dyadic-prediction problem. We believe that this nice problem will receive further attention in the community. Two well-studied problems, collaborative filtering and link prediction, are shown to be instances of the dyadic-prediction problem. To solve this new prediction problem, the authors present a general approach that uses components of supervised and unsupervised learning. The method users latent features that help explain both the dyad observations as well as the item labels. The strength of the method is that it seamlessly fills in missing observations as well as predicts missing labels.

We hope that the reader will find this fine collection of articles as exhilarating and inspiring as we did. We wish to warmly thank all authors whose hard work is distilled in this special issue. The contribution of reviewers and area chairs has also been invaluable.

Enjoy reading.