J Intell Inf Syst (2015) 45:295–297
DOI 10.1007/s10844-015-0370-7

Guest editor's introduction: special issue on quality issues, measures of interestingness and evaluation of data mining models

Philippe Lenca[1] · **Stéphane Lallich**[2]

Received: 17 June 2015 / Accepted: 18 June 2015 /
Published online: 3 July 2015

There are many data mining algorithms and methodologies for various fields and various problematic. Each data mining researcher/practitioner should describe the intrinsic quality of the discovered models and patterns. In addition he/she is faced with assessing the performance of his/her own solution(s) in order to make –fair– comparisons with state of the art approaches. Which methodology, which benchmarks, which measures of performance, which tools, which measures of interest, which scenarios, etc., should be used, and why? Every one should answer these questions, and assessing the quality and the performance of a data mining process is a critical issue. Assessing the quality and the performance is a critical issue for classical situations and even more in the Big Data era where we can not manage this issue with classical and current approaches. Clearly, large scale data, complex data and streaming data bring new challenges (e.g. large scale inference, fake correlations, necessity of perpetual validation).

In this setting, the *Quality Issues, Measures of Interestingness and Evaluation of data mining model*s Workshop series provide an open forum for intensive discussions and the exchange of new ideas and best practices on quality issues among data mining researchers. QIMIE focuses on the theory, the techniques, the links between the discovery stage and the quality assessment stage, and the practices that can ensure that the discovered knowledge of a data mining process is of quality. It is organized with the IEEE Task Force on *Evaluation and Quality issues* (Data Mining and Big Data Analytics Technical Committee, IEEE Computational Intelligence Society).

✉ Philippe Lenca
philippe.lenca@telecom-bretagne.eu

Stéphane Lallich
stephane.lallich@univ-lyon2.fr

[1] Institut Mines Telecom, Telecom Bretagne, UMR 6285 CNRS, Lab-STICC,
29238 Brest Cedex 3, France

[2] ERIC Labotary, Université Lyon 2, 5 avenue Pierre Mendès France, 69676 Bron Cedex, France

This special issue contains seven revised and extended papers from QIMIE'13, the third edition of QIMIE, organized in association with PAKDD'13 conference (Pacific-Asia Conference on Knowledge Discovery and Data Mining, Gold Coast, Australia, April 14–17, 2013).

How effective are condensed representations (frequent, closed, maximal) and interestingness measures in recovering the underlying patterns? Albrecht Zimmermann proposes an original framework to answer this question. He uses an extension of the Alamaden Quest data generator which allows constructing data from explicit patterns. So, he can thoroughly compare the mined patterns against the source itemsets used to construct the data and evaluate the performance of the different pattern representations and the different intestingness measures.

Marwan Hassani, Yunsu Kim, Seungjin Choi and Thomas Seidl propose a new effective measure for subspace stream clustering. Subspace stream clustering aims at finding evolving clusters within subgroups of dimensions and is thus attractive to take into account the increasing dimensionality of data streams. The authors also present a method for designing new stream subspace and projected clustering algorithms, and a novel method for using available offline subspace clustering measures for data streams within the Subspace MOA framework.

Rob Martinus Konijn, Wouter Duivesteijn, Marvin Meeng and Arno Knobbe deal with data where examples are labeled with a classical binary target but also have costs associated with them. They propose new measures taking into account both the binary and the cost target in order to detect clearly defined subsets of the data where the values of these two targets have an unusual distribution. Each subset can be judged with respect to the entire dataset (Subgroup Discovery) or with a local reference group (Local Subgroup Discovery). Two real-life health care applications illustrate the proposed approach.

To address the rare item problem, Uday Kiran and Masaru Kitsuregawa propose to consider multiple minimum all-confidence thresholds. This approach extend the existing correlated pattern model and facilitates the user to specify a different threshold for each pattern depending upon its items' frequencies. An efficient pattern-growth algorithm is designed to discover interesting patterns involving both frequent and rare items without generating a huge number of meaningless correlated patterns.

Jean-Charles Lamirel, Pascal Cuxac, Aneesh Sreevallabh Chivukula and Kafil Hajlaoui propose an efficient adaptation of the feature maximization metric for clustering. They provide a new feature selection and feature contrasting model in the context of supervised classification. Experiences on textual data show that the new method is very efficient for highly unbalanced, highly multidimensional and noisy textual data.

Dang Bach Bui, Fedja Hadzic, Andrea Tagarelli, Michael Hecker improve the study of a tree-structured classification technique based on association rules generated from a structure-preserving flat representation of the data. Corresponding subtrees are constrained by the position in the original trees, leading to a drastic reduction in the number of rules generated. The authors provide an extensive evaluation of this approach in terms of coverage rate and accuracy, compared with a state-of-the-art structural classifier without positional constraint on three realworld data sets.

Gowtham Srinivas Parupalli, Krishna Reddy Polepalli, Venkata Trinath Atmakuri, Bhargav Sripada and Uday Kiran Rage first propose a model of coverage patterns (CPs). The CPs are specified by defining two measures: coverage support and overlap ratio. To extract CP's, the authors discuss level-wise pruning approach which exploits the sorted closure property of overlap ratio. They also propose two algorithms based on projected pattern growth approach by exploiting the notions of non-overlap projection and minimal CPs.

Experiment results show that the proposed model and methodology can effectively discover CPs.

Last, we would like firstly, to thank the authors and reviewers for their hard work, the IEEE Data Mining and Big Data Analytics Technical Committee (IEEE Computational Intelligence Society) for hosting the Task Force on *Evaluation and Quality issues* and secondly, Zbyszek Ras, co-editor in chief of JIIS, and the Springer team who contributed to prepare this special issue.

Philippe Lenca and Stéphane Lallich

Special Issue Guest Editors