Special issue on

Quality Issues, Measures of Interestingness and Evaluation of Data Mining Models

for the

Vietnam Journal of Computer Science

Introduction

There are many data mining algorithms and methodologies for various fields and various problems. Each data mining researcher/practitioner is faced with assessing the performance of his own solution(s) in order to make comparisons with state-of-the-art approaches. He should also describe the intrinsic quality of the discovered patterns. Everyone should answer the following questions: which methodology, which benchmarks, which measures of performance, which tools, which measures of interest, etc. should be used, and why? Assessing the quality and the performance is a critical issue for classical situations and even more in the Big Data era where we cannot 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, the need for perpetual validation).

This special issue will focus on the theory, techniques and practices that can ensure that discovered knowledge is of appropriate quality. It will thus cover the problem of measuring the quality of patterns, the evaluation of data mining models, and the links between the discovery stage and the quality assessment stage. It follows the fourth QIMIE workshop which was organized in association with the PAKDD 2015 conference (19th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Ho Chi Minh City, Vietnam, May 19-22, 2015). Original submissions and extended and revised papers presented at QIMIE 2015 and at other conferences are welcomed. Papers that extend conference contributions must include at least 25% new material, and the authors should explain how such papers have been extended with a mention of the conference name and a reference to the original paper.

Topics

Major topics will include but are not limited to the following:

- objective measures of interest (for individual rules or rules basis, patterns, graphs, data streams, clusters, etc.);
- subjective measures of interest and quality based on human knowledge, quality of ontologies, actionable rules;
- algorithmic properties of measures of interest especially in the context of Big Data;
- comparison of algorithms: issues with benchmarks, experiments and parameter tuning, questioning also the reproducibility of data mining results, the need for new data sets which match new problems, methodologies, statistical tests, etc.;
- robustness evaluation and statistical evaluation;
- graphical tools like ROC, cost curves, user-friendly visualization tools;
- special issues: unbalanced data, very large data sets, very-high-dimensional data, changing environments and dynamic data, data streams, Big Data, a lack of training data, graph data, annotated data and semi-supervised learning, methods for performance evaluation with no ground
truth data, etc.;
- special issues in specialized domains: bioinformatics, security, information retrieval, sequential and time series data, social networks, geolocalized data, etc.

From the previous list of key topics, although not exhaustive, one can identify **four major themes:**

- Properties of objective measures of interest (for individual rules, for rules basis) which leads to the problem of how to choose, depending on the user's goal and other factors, an appropriate interestingness measure in order to filter the huge amount of individual rules or to evaluate a set of rules. [This theme is of great interest to the machine learning and data mining community.] Properties of subjective measures of interest, integration of domain knowledge, quality based on human knowledge, quality of ontologies, actionable rules.
- Algorithmic properties of interestingness measures which leads to the problem of how to mine efficiently interesting patterns, i.e., can we use interestingness measures as soon as possible (as well as the well-known support) to reduce both the time to mine databases and the number of found patterns? This question has attracted a lot of work but is still a very challenging problem.
- Challenges with new data and new problems. These are generally the subject of fruitful specialized workshops (e.g., very large and very-high-dimensional data, unbalanced data, etc., and related specialized domains like bioinformatics, life sciences in a broad manner, etc.).
- Evaluation and comparison of algorithms which lead to debate on how an algorithm should be evaluated, on properties (e.g., accuracy, conciseness, specificity, sensitivity, etc.), on trade-offs between the different types of errors for multiple simultaneous hypothesis testing, on how to construct new evaluation measures? etc.

**Important Dates**

- First Decision: July 20, 2016
- Revision Submission: August 30, 2016
- Second Decision: October 15, 2016
- Publication Dates: end of 2016

**Guest Editors**

- Prof. Philippe Lenca (Institut Mines-Télécom, Télécom Bretagne, UMR 6285, Lab-STICC, France; philippe.lenca@telecom-bretagne.eu)
- Prof. Stéphane Lallich (Université Lyon 2, Laboratory ERIC, France; stephane.lallich@univ-lyon2.fr)
- Dr. Thanh-Nghi Do (Can Tho University, College of Information & Communication Technology, Vietnam; dtnghi@cit.ctu.edu.vn)