Guest editor's introduction: special issue on discovery science 2012

Philippe Lenca · Jean-Marc Petit

With the considerable amount of automatically generated scientific data it becomes more and more essential to automate hypothesis generation and tests in many domains of the scientific activity, e.g., in molecular biology, medicine, astronomy, physics, social sciences, etc. It's not only a new fancy that could lead to mechanize discoveries and creative processes, but a necessity for almost all the contemporary scientists who have to manage big volumes of data stored in huge databases or datawarehouses. Streams data increase this necessity. In this context –the era of huge data–, the sciences are facing an *epistemological break*: a new experimental approach is emerging, which makes an extensive use of automated knowledge extraction techniques to deal with huge data sets in order to induce and to validate new scientific theories, while in the traditional experimental approach, the scientists built physical devices to confirm – or dis-confirm – the newly generated theories by just comparing observations to anticipations. The field of *Discovery Science* aims at inducing and validating new scientific hypothesis from data. This brings together many scientific disciplines and technical fields that contribute to this new method of scientific investigation toward automatic knowledge discovery tools. It goes without saying that it plays a key role in the *data science* that opens many exciting, promising and stimulating perspectives.

In this setting, the main objective of the Discovery Science conference series is to provide an open forum for intensive discussions and the exchange of new ideas and information among researchers working in the area of Discovery Science. The scope of the conference includes the development and analysis of methods for automatic scientific knowledge discovery, machine learning, data mining, intelligent data analysis, theory of learning, tools for supporting the human process of discovery in science, as well as their application to knowledge discovery.

This special issue is focused on some of the issues characterizing Discovery Science. It contains five revised and extended papers from the best ones published in the "15th International Conference on Discovery Science (DS 2012)" which held in October 2012, Lyon, France.

P. Lenca (✉)
Institut Mines Telecom, Telecom Bretagne, UMR 6285 CNRS, Lab-STICC, 29238 Brest Cedex 3, France
e-mail: philippe.lenca@telecom-bretagne.eu

J.-M. Petit
INSA Lyon, Université de Lyon, UMR 4205 CNRS, LIRIS, Lyon, France
e-mail: jean-marc.petit@insa-lyon.fr

The paper "Soft constraints for pattern mining" by Willy Ugarte, Patrice Boizumault, Samir Loudni, Bruno Crémilleux and Alban Lepailleur deals with Constraint-based pattern discovery by using Constraint Programming techniques. The authors point out several advantages related to *softness* in the discovery process, i.e., to relax threshold constraints or constraints involved in top-k patterns or skypatterns. Experiments on chemo-informatics for discovering toxicophores are conducted to validate the propositions.

The paper "Fast progressive training of mixture models for model selection" by Prem Raj Adhikari and Jaakko Hollmén is about finite mixture models and the Expectation Maximization algorithm. The authors propose a fast training of a series of mixture models using progressive merging of mixture components to facilitate model selection algorithm to make appropriate choice of the model. Several experiments on synthetic and real-life datasets show the interest of the approach.

The paper "Change point detection for burst analysis from an observed information diffusion sequence of tweets" by Kazumi Saito, Kouzou Ohara, Masahiro Kimura and Hiroshi Motoda proposes to detect the period in which a burst of information diffusion took place from an observed diffusion sequence data over a social network. The approach is experimented on both synthetic and real life tweeter data.

The paper "Efficient redundancy reduced subgroup discovery via quadratic programming" by Rui Li, Robert Perneczky, Alexander Drzezga and Stefan Kramer addresses the crucial redundancy problem in subgroup discovery. Many experiments conducted on several datasets assess the interest of the reduction.

The paper "A scalable approach to spectral clustering with SDD solvers" by Nguyen Lu Dang Khoa and Sanjay Chawla concerns spectral clustering in large and high dimensional spaces. It proposes to bypass the eigen decomposition of the original Laplacian matrix by leveraging the near-linear time solver for symmetric diagonally dominant (SDD) linear systems and random projection. Experiments on synthetic and real-life datasets point out that the method is faster and deliver better quality clusters than the state-of-art competitive methods.

Last, we would like firstly, to thank the authors and reviewers for their hard work and secondly, Zbyszek Ras, co-editor in chief of JIIS, and the Springer team who contributed to prepare this special issue.

Philippe Lenca and Jean-Marc Petit
Special Issue Guest Editors