J Intell Inf Syst (2013) 41:1–4
DOI 10.1007/s10844-013-0261-8

Special issue on challenges in knowledge discovery and data mining

Shusaku Tsumoto

Published online: 27 July 2013

Guest editor introduction

Ten years have passed since Jan Zytkow passed away in the midst of rapid progress in data mining. Jan Zytkow started his research career in physics, where he became interested in the philosophical nature of scientific discovery. Then, he posed a question on how a computational machine could reason and discover knowledge, and performed research in Artificial Intelligence, more specifically on "scientific discovery" as proposed by Herbert Simon. He was involved in starting the area of knowledge discovery and data mining, now simply called "data mining". However, what he had in mind were the sophisticated relationships among philosophy, science and data; he envisioned computer-based scientific discovery, which would contribute to science and other disciplines in the 21st century. We dedicate this Special Issue to Dr. Jan Zytkow, so as to honor his pioneering work.

In this special issue, we reflect on the questions posed by Jan Zytkow, focusing on recent challenges in data mining, and envisioning the future of data mining from his perspective. Let us look at the status of data mining research in 2012.

The Guest Editor has examined data mining manuscripts submitted to the IEEE International Conference on Data Mining (ICDM) since 2002; he tracked the trends in data mining research by using several techniques (IEEE ICDM 2002). The results obtained by trend-detection text-mining methods (Abe and Tsumoto 2012) are shown below. The following 20 keywords of abstracts in accepted papers are shown in increasing order of number of appearances: social, learning, matrix, network(s), prediction, factorization, graphs, sparse, robust, feature, topic, graph, online, detection, novel, anomaly detection, multiple applications, multi-task, and influence maximization. On the other hand, 20 decreasing keywords (in number of appearances) are: patterns, data mining, mining, databases, techniques, query,

S. Tsumoto (✉)
Faculty of Medicine, Shimane University, Izumo, Japan
e-mail: tsumoto@med.shimane-u.ac.jp

comparison, data sets, semantic, knowledge, tasks, algorithm, attributes, accuracy, optimization, microarray data, data clustering, study, time series clustering and common. Ten emerging keywords which appear from 2010 onwards are: multi-task, map reduce, diffusion, Gaussian, influence maximization, sparse representation, hashing, label, cascade and risk, all of which focus on scalability issues in data mining.

Figure 1 shows the results of multidimensional scaling (MDS) by using similarities among temporal sequences of Term Frequency–Inverse Document Frequency (TFIDF) values of keywords. Most of the keywords are concentrated in the central region, but increasing and decreasing keywords are assigned around the center. Keywords shown in red are well-known keywords in the data mining community. Interestingly, the horizontal and vertical axes can be interpreted as frequency and trend, respectively in an reverse fashion. For example, clustering is high frequency with decreasing trend and social is low frequency with increasing trend.

Based on these empirical observations, the guest editor selected the following five topics which are trending since Jan Zytkow passed away, except for high performance computing: network mining, transfer learning, clustering, mining on multi-granularity level (application), and sensor data mining.

The first paper, entitled *Learning to predict opinion share and detect anti-majority opinionists in social networks* written by Masahiro Kimura, Kazumi Saito, Kouzou Ohara, and Hiroshi Motoda addresses the problem of detecting anti-majority opinionists using the value weighted mixture voter (VwMV) model. This problem is motivated by the fact that 1) each opinion has its own value and an opinion with a higher value propagates more easily/rapidly and 2) there are always people who have a tendency to disagree with any opinion expressed by the majority. The authors extend the basic voter model to include these two factors with the value of each opinion and the anti-majoritarian tendency of each node as new parameters, and learn these parameters from a sequence of observed opinion data over a social network. The authors show both theoretically and experimentally that the proposed method

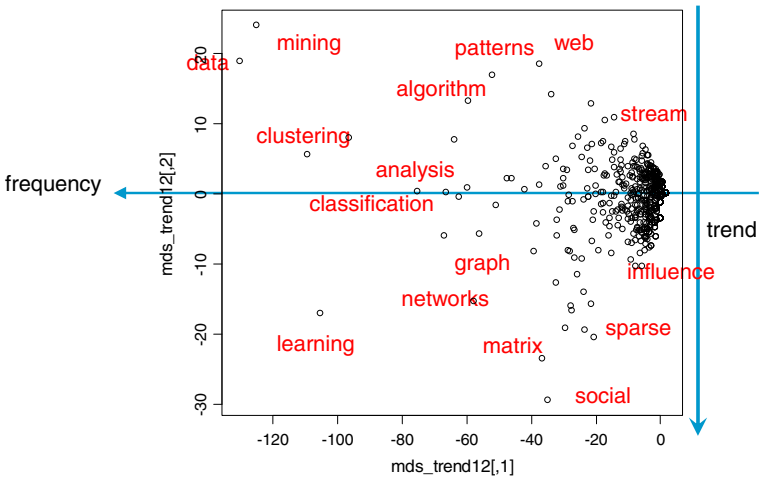

Fig. 1 Visualization of trends in keywords by multidimensional scaling

can help to understand the short- and long-term behavior of opinion propagation. The paper discusses important aspects of *network mining*, which is one of the most popular field in data mining from this century.

The second paper, entitled *Transfer Learning by Centroid Pivoted Mapping in Noisy Environment*, written by Thach Nguyen Huy, Bin Tong, Hao Shao, and Einoshin Suzuki proposes a new approach to transfer learning, which is a widely-investigated learning paradigm that is initially proposed to reuse informative knowledge from related domains, given that supervised information in the target domain is scarce, while it is sufficiently available in the multiple source domains. One of the challenging issues in transfer learning is how to handle the distribution differences between the source domains and the target domain. The authors propose a robust framework against noise in the transfer learning setting by proposing a novel centroid pivoted mapping. They assume the differences on both conditional distributions and marginal distributions among multiple source domains and the target domain. Motivated by the observation that local structure is important when a transfer strategy is designed, the authors present a variant of a density-based clustering algorithm which can remove instances corrupted by feature noise and label noise, while taking great care to preserve the local structure. Experimental validation shows that their framework works effectively even in high noisy environments up to 20 % of noise level.

The third paper, entitled *Clustering of non-metric proximity data based on bi-links with ε-indiscernibility* written by Shoji Hirano and Shusaku Tsumoto proposes a hierarchical grouping method for non-metric proximity data based on bi-links and ε-indiscernibility. It hierarchically forms directional links among objects according their directional proximities. A new cluster can be formed when objects in two clusters are connected with bi-directional links (bi-links). The concept of ε-indiscernibility is incorporated into the process of establishing bi-links. This scheme enables users to control the level of asymmetry that can be ignored in merging a pair of objects. Experimental results on the soft drink brand switching data showed that this approach is capable of producing better clusters compared to the straightforward use of bi-links. Clustering is one of the major topics in data mining: however, there are few studies on clustering for non-metric data.

The fourth paper, entitled *Improving Customer Acquisition Models by Incorporating Spatial Autocorrelation at Different Levels of Granularity,* written by Philippe Baecke and Dirk Van den Poel, discusses the problems with Customer Relationship Management (CRM). Traditional customer acquisition models often ignore the spatial correlation that may exist between the purchasing behaviors of neighboring customers and treats this as nuisance in the error term. Based on data of a Japanese automobile brand, this study shows that, even in a model that already includes a large number of socio-demographic and lifestyle variables typically used for customer acquisition, extra predictive value can still be obtained by taking spatial interdependence into account using a generalized linear autologistic regression model. Further, this study indicates that the marketing decision maker should carefully choose the granularity level on which the neighborhoods are composed, because this can have an important impact on the model's accuracy. In this research, the best predictive performance was obtained at granularity level 3. The author discusses that granularity plays a central role in knowledge discovery: if the marketing decision maker has sufficient resourses, it is advisable to obtain data that divides customers

into neighborhoods at multiple granularity levels. Although the paper is application-based, the authors empirically show that dealing with granularity even for statistical modeling is important for discovery. This was one of the most important issues for knowledge discovery on which the late Jan Zytkow focused.

The final paper, entitled *A Framework for Analysis of the Effect of Time on Shopping Behavior* written by Keiji Takai, and Katsutoshi Yada applies statistical methods to data collected from sensors. The authors propose a framework that considers heterogeneity in the number of items a customer buys. The first step of our framework is based on the Poisson mixture regression model using a stationary time in the department where the items are sold as its independent variable. This model finds latent homogeneous groups of customers and gives the regression models within each group. It simultaneously classifies the customers into the homogeneous groups. In the second step of their framework, a method is presented to investigate whether another factor (variable) influences the classification into homogeneous groups. This proposed framework was applied to real data collected from the customers, including RFID data and POS data, and the effectiveness of the framework is shown. The managerial implications were drawn from the result of the analysis. Mining sensor data will be a new future trend of data mining in 21st century.

References

Abe, H., & Tsumoto, S. (2012). Detection of research trends from bibliographical data. *IJDMMM 4*(3), 255–266.

IEEE International Conference on Data Mining (2002). IEEE Computer Society. http://www.cs.uvm.edu/~icdm/main.shtml. Accessed 10 July 2013.