Special issue on selected papers from IEEE DMF 2008

Keith C. C. Chan · Xindong Wu

Published online: 7 September 2010
© Springer-Verlag London Limited 2010

Data mining has emerged as one of the most dynamic and lively areas in IT research and development. Many data mining techniques and algorithms have been used in a diversity of application areas in banking, finance, medicine, education, business, and marketing analysis, etc. They have been found to be able to effectively discover meaningful patterns in large volumes of data in these application areas, and these achievements are due directly and indirectly to the results of many years of effort in the R&D of new data mining methodologies, new processes in knowledge discovery, new theory on data mining foundations and new insights into the use of data mining techniques in emerging domains.

By promoting novel, high-quality research and by developing innovative solutions to challenging real-world problems, we expect data mining to continue to advance and create a big impact in the world. It is this belief that prompted us to organize a two-day IEEE Data Mining Forum in May, 2008 at The Hong Kong Polytechnic University, Hung Hom, Hong Kong. For the Forum, we invited a select number of the world's leading experts in data mining from North America, Europe, Australia, and Asia to give invited talks and participate in panel discussions. The forum was attended by academics and practitioners alike and the presentations on original, cutting-edge, and state-of-the-art progress in data mining have provided insights for the audience and the speakers to further advance the area. This special issue grows out of the presentations at the Forum. Some speakers were invited to submit a paper on the topic they presented. After a careful review process, four papers were finally selected. They cover the research areas in peer-to-peer networks, optimization-based data mining, graph clustering, and semi-supervised learning.

The first paper in this special issue is “A local asynchronous distributed privacy preserving feature selection algorithm for large peer-to-peer networks” by Kamalika Das, Kanishka Bhaduri and Hillol Kargupta. The paper describes a local-distributed privacy-preserving

---

K. C. C. Chan
Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

X. Wu (✉)
Department of Computer Science, University of Vermont, Burlington, VT 05405, USA
e-mail: kais@cs.uvm.edu

algorithm for feature selection in a large peer-to-peer environment. It focuses on the role of feature selection in distributed data to avoid the costs involved to centralize the entire dataset or portions of it obtained from a large number of data sources. The problem, in the presence of the asynchronous nature of the peer-to-peer networks, the dynamic nature of the data/network, and the privacy concerns, is very challenging to solve. The authors propose a privacy-preserving asynchronous algorithm for feature selection in a large P2P network. The proposed solution allows people to perform feature selection in an asynchronous fashion with a low communication overhead where each peer can specify its own privacy constraints. The algorithm performs its tasks based on local interactions among participating nodes, and its performance is tested on real-world data.

The second paper is “Multiple criteria optimization-based data mining methods and applications: a systematic survey” by Yong Shi. It provides a systematic survey on a series of optimization-based classification models. Multiple Criterion Linear Programming (MCLP) and Multiple Criterion Quadratic Programming (MCQP) have been used to develop and extend such models in the last 10 years. The author reviews basic concepts and frameworks of these methods and intends to promote research interests in them in the data mining community. According to the evolution of multiple criterion programming, the paper starts with the bases of MCLP. Then, it further discusses penalized MCLP, MCQP, Multiple Criterion Fuzzy Linear Programming (MCFLP), Multi-Class Multiple Criterion Programming (MCMCP), and the kernel-based Multiple Criterion Linear Programming, as well as MCLP-based regression. This paper also outlines several applications of Multiple Criterion optimization-based data mining methods, such as Credit Card Risk Analysis, Classification of HIV-1-Mediated Neuronal Dendritic and Synaptic Damage, Network Intrusion Detection, Firm Bankruptcy Prediction, and VIP E-Mail Behavior Analysis.

In the third paper “A general framework for relation graph clustering” by Bo Long, Zhongfei Zhang and Philip S. Yu addresses a classical data mining problem to obtain cluster patterns from relation graphs. The authors propose a general framework to derive a family of clustering algorithms including both hard and soft versions, which are capable of learning cluster patterns from relation graphs with various structures and statistical properties. This work presents a number of classic approaches on special cases of relation graphs, such as traditional graphs with single-type nodes and bi-type relation graphs with two types of nodes, and they can be viewed as special cases of the proposed framework. Both theoretic analysis and experiments demonstrate the effectiveness of the proposed framework and algorithm.

The last paper “Semi-Supervised Learning by Disagreement” by Zhi-Hua Zhou and Ming Li proposes solutions for semi-supervised learning. Due to the facts that the number of labeled training examples is limited in many real-world tasks and the labeling task requires human efforts and expertise, the authors aim at improving the learning performance with semi-supervised learning to exploit the unlabeled examples. Disagreement-based semi-supervised learning is proposed in this paper, where multiple learners are trained for the task and the disagreements among the learners are exploited during the semi-supervised learning process. This survey article provides an introduction to research advances in this paradigm.

Finally, we would like to thank all authors who submitted their manuscripts to this special issue and the external reviewers for their invaluable contributions to the reviewing process for giving us this great opportunity of organizing this special issue. We believe that this collection of contributions will play a key role in the future development of data mining research and applications.