Editorial

Xingquan Zhu[1], Chengqi Zhang[2], and David L. Olson[3]

[1](Dept. of Computer Science & Engineering, Florida Atlantic University,

Boca Raton, FL, USA, xqzhu@cse.fau.edu)

[2](Faculty of Engineering and Information Technology, University of Technology,

Sydney, Australia, chengqi@it.uts.edu.au)

[3](College of Business Administration, University of Nebraska-Lincoln,

Lincoln, NE, USA, dolson3@unl.edu)

1   Introduction

Data mining is dedicated to exploring and analyzing large volumes of data, by automatic or semi-automatic means, to discover meaningful patterns in describing the given data[1,2]. Boosted by the industrial needs and supported by numerous developments in machine learning, statistics, artificial intelligence, and database systems, data mining research has now spanned various fields from computational science, business intelligence, life science, scientific computing, to ecology and geology. The data mining community also has rapidly grown from an international workshop (co-located with the AAAI-1994[3]) to a research field with three major international conferences (KDD, ICDM, and SDM) and numerous journals such as TKDD, DMKD and KAIS.

Behind the prosperity of the data mining research, the driving force is typically twofold: data and applications. First, the progress we made in hardware and networking technology has made it easier than ever to collect and aggregate large volumes of data for rapid decision making, which were previously impossible or difficult to realize. For example, a desktop computer nowadays can easily afford to have Terabyte storage whereas a decade ago, this required a dedicated data center with numerous costly supporting devices. For many other domains, such as sensor networking and RFID, hardware devices are becoming more widely available to generate an enormous amount of data on a daily basis. With these large volumes of data, the data owners have an immediate need of turning data into useful knowledge. Such needs are becoming an extreme for commercial applications in business intelligence, where data users are always trying to gain more but pay less for their business activities[4]. Consequently, data mining research is not only challenged by the data volumes, but is also severely challenged by the speed and efficiency in digesting the data. Second, for some applications, data volumes might not be a major concern whereas the data modality and data relationships are difficult to characterize or are subject to complex correlations. For example, the data mining research for multimedia data (audio, video, and images) is severely challenged by the reality that the relationships between

multimedia objects (e.g., picture-picture, frame-frame, and video-video) are difficult to characterize by simple measures such as Euclidean distances. In addition, recent research on graphs, social networks, information networks, and multiple information sources[5,6], has shown that many real-world data are subject to complex relationships. As a result, existing statistical or machine learning methods must be customized in order to harness such complex data.

2   Special Issue Thrust Areas

The rapid evolving of the data mining research motivates the guest editors to compile this special issue, which intends to report recent research results in broader areas of data mining and further promotes the data mining research in the community. Notice that the data mining research is rather broad[7], but we have put primary focus on the following five thrust areas:

(1) Classical Data Mining Algorithms

Including supervised and semi-supervised learning, active learning, cost-sensitive learning, multi-task and transfer learning, rough sets, fuzzy modeling, clustering, and association rule mining.

(2) Mining Web, Multimedia, Graph, and Complex Data

Including web data mining, recommendation systems, multimedia data mining, graph mining, and information network mining.

(3) Data Quality, Access Control, and Privacy Issues in Data Mining

Data quality assessment and control, privacy preserving data mining, data pre-processing and enhancement.

(4) Business Intelligence and Risk Assessment

Including OLAP, decision support systems, business planning, sales intelligence, business intelligence tools and systems, and customer relationship management.

(5) Computational Biology and Bioinformatics

Including biological sequence analysis, alignment, and assembling, microarry expression data analysis, pathway, network, and system biology and modeling.

3   Special Issue Papers

Our call for papers attracted a total number of 25 submissions worldwide, out of which only 7 papers were included in this special issue. In order to accelerate the publishing circle, the guest editors practiced a two-stage review process. In the first stage, the guest editors ranked all 25 submissions based on a number of criteria, such as the topics, the content, and the significance of the paper. The first 13 papers ranked at the top of the list were forwarded to the second round review. In the second round review, some authors were invited to participate in the review process. To avoid conflict of interests, each paper is reviewed by at least one reviewer who did not submit any paper to this special issue. The guest editors also carefully read all 13 papers and their review reports before the final decisions were made. The list of the accepted seven papers follows, where the bold text following the title indicates the category the paper may fit into the five main thrust areas of this special issue.

Mohamed Farouk Abdel Hady and Friedhelm Schwenker, Co-Training by Committee: A Generalized Framework for Semi-Supervised Learning with Committees

(**Classical Data Mining Algorithms**)

Taufik Djatna and Yasuhiko Morimoto, Attribute Selection for Numerical Databases that Contain Correlations (**Classical Data Mining Algorithms**)

Manjeet Rege and Qi Yu, Efficient Mining of Heterogeneous Star-structured Data (**Mining Complex Data**)

Mathieu Roche and Violaine Prince, Managing the Acronym/Expansion Identification Process for Text-Mining Applications (**Web and Text Mining**)

Peng Zhang, Zhiwang Zhang, Aihua Li, and Yong Shi, Global and Local (Glocal) Bagging Approach for Classifying Noisy Dataset (**Data Quality Issue in Data Mining**)

Anthony Quinn, Andrew Stranieri, John Yearwood, Gaudenz Hafen, and Herbert Jelinek, AWSum – Combining Classification with Knowledge Aquisition (**Business Intelligence and Risk Assessment**)

Nicolas Pasquier, Claude Pasquier, Laurent Brisson, and Martine Collard, Mining Gene Expression Data using Domain Knowledge (**Computational Biology and Bioinformatics**)

The first two papers address classical data mining problems from semi-supervised learning and attribute selection perspectives. The main contribution of the first paper is that the authors propose a co-train by committee framework. This simple design has shown to be effective on a number of selected benchmark datasets (including image data). In the second paper, the authors consider interaction between attribute pairs to evaluate the importance of each attribute for classification. This is shown to be more effective than simple measures which traditionally ignore the attribute correlations.

The third and the fourth papers propose solutions for mining complex structure data and mining text data. Among them, Rege and Yu's paper proposes a graph theoretical framework for addressing star-structured co-clustering problems in which a central data type is connected to all the other data types. The guest editors believe that the authors' work will significantly advance our view for mining multi-source and complex data. The fourth paper focuses on the problem of extracting acronym/definition from textual data, which is, in guest editors' view, an important task for text and web mining.

In the fifth paper, Zhang and Shi address data quality in data mining research and propose a bagging framework, which takes the existing global and local bagging approaches into consideration, for classifying noisy data. As real-world data are usually imprecise and contain a significant amount of errors, the guest editors believe that the solutions proposed in the paper will advance our skills and help put data mining to the frontier of handing real-world noisy data.

The sixth paper, by Quinn *et al.*, represents another set of problems in data mining, typically raised in the medical and business intelligence domains, where the mining is required to produce accurate results with self-interpretability, such as what is the influence of the feature values on a class value? In the paper, the authors propose a simple approach which has shown to be able to deliver accurate classification results and provide insightful interpretation for the data.

The last paper by Pasquier *et al.* provides a survey on mining gene expression data using domain knowledge. The authors use gene expression data to demonstrate

the entire process from a priori knowledge based gene expression data analysis, including data preparation, filtering, clustering, and association rule mining, to knowledge based post-processing and data integration. The guest editors believe that this paper provides a good starting point to demonstrate the applications of the data mining in bioinformatics research.

4   Conclusions and Acknowledgement

All together the topics of the selected seven papers span the five thrust areas the guest editors intended to address in this special issue. The guest editors wish to thank all the authors and reviewers who have participated in the review process. The support of the National Natural Science Foundation of China (#60674109) is acknowledged! Finally, we hope the reader will enjoy this special issue and find it useful. Special Issue Reviewers:

Gong Chen	University of California, Los Angles, USA
Taufik Djatna	Hiroshima University, Japan
Mohamed F.A. Hady	Institute of Neural Information Processing, Germany
Dan He	University of California, Los Angles, USA
Mathieu Roche	LIRMM, France
Anthony Scrime	State University of New York College at Brockport, USA
Andrew Stranieri	University of Ballarat, Australia
Huaifeng Zhang	University of Technology, Sydney, Australia
Peng Zhang	Chinese Academy of Sciences, China
Yan Zhang	University of Vermont, USA
Yanchang Zhao	University of Technology, Sydney, Australia
Zhenfeng Zhu	Fudan University, China

References

[1] Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamyj R. Advances in Knowledge Discovery and Data Mining. The MIT Press, 1996.
[2] Zhu XQ, Davidson I. Knowledge Discovery and Data Mining: Challenges and Realities. IGI Global, 2007.
[3] Fayyad UM, Uthurusamy R. Proc. of the First AAAI International Workshop on Knowledge Discovery in Databases. AAAI Press, 1994.
[4] Olson DL, Shi Y. Introduction to Business Data Mining. McGraw-Hill/Irwin, 2006.
[5] Zhang SC, Zhang CQ, Wu XD. Knowledge Discovery in Multiple Databases. Springer, 2004.
[6] Zhu XQ, Jin RM, Agrawal G. Prof. of the First ACM International Workshop on Mining Multiple Information Sources (MMIS). San Jose, 2007.
[7] Wu XD, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH, Steinbach M, Hand DJ, Steinberg D. Top 10 Algorithms in Data Mining, Knowledge and Information Systems, 2008, 14(1): 1–37.

Xingquan Zhu is an Assistant Professor in the Department of Computer Science and Engineering, Florida Atlantic University, Boca Raton, FL. He received his Ph.D. degree in Computer Science from Fudan University, Shanghai, China, in 2001. From February 2001 to October 2002, he was a Postdoctoral Associate in the Department of Computer Science, Purdue University, W. Lafayette,

IN. From October 2002 to July 2006, He was a Research Assistant Professor in the Department of Computer Science, University of Vermont, Burlington, VT. His research interests include data mining, machine learning, multimedia systems, and information retrieval. He is the Associate Editor of the IEEE Transactions on Knowledge and Data Engineering (2009-). He is also serving on the editorial boards of the Knowledge and Information System Journal (2003-), Advances in Data Warehousing and Mining (2007-), and World Scientific and Engineering Academy and Society: WSEAS Transactions on Signal Processing (2005-). He has served on the program committees for many international conferences, including SIGKDD, ICDM, and SDM. He is the founder and co-chair of the ACM KDD Workshop on Mining Multiple Information Sources (MMIS). His work on Aggressive Classifier Ensemble won the Best Paper Award at the IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2005). Since 2000, he has published over 80 technical papers in referred journals and conference proceedings.

Chengqi Zhang is a Research Professor of Information Technology at University of Technology, Sydney (UTS) since December 2001. He is currently the Director of UTS flagship Research Centre for Quantum Computation and Intelligent Systems which is one of five research flagships at UTS. In addition, he is the Leader of Data Mining program of Australian Capital Market Cooperative Research Centre and he is the Chairperson of Australian Computer Society's National Committee for Artificial Intelligence. Prof. Zhang obtained his PhD degree from Queensland University in 1991 and Doctor of Science (DSc) from Deakin University which is the Higher Doctorate in 2002. His research interests include Multi-Agent Systems, Data Mining, and their integrations. He has published more than 200 research papers in the first class international journals including Artificial Intelligence, IEEE and ACM Transactions. Prof. Zhang is the Associate Editor for IEEE Transactions on Knowledge and Data Engineering. He has been selected as the Chairperson of the Steering Committee for Knowledge Science, Engineering, and Management. He is also the member of PRICAI and PAKDD steering committee members since 2004. He has been invited to deliver six keynote speeches in international conferences. He is a senior member of the IEEE Computer Society (IEEE).

David L. Olson is the James & H.K. Stuart Professor in MIS and Chancellor's Professor at the University of Nebraska. He has published research in over 100 refereed journal articles, primarily on the topic of multiple objective decision-making and information technology. He teaches in the management information systems, management science, and operations management areas. He has authored the books Decision Aids for Selection Problems, Introduction to Information Systems Project Management, and Managerial Issues of Enterprise Resource Planning Systems and co-authored the books Decision Support Models and Expert Systems; Introduction to Management Science; Introduction to Simulation and Risk Analysis; Business Statistics: Quality Information for Decision Analysis; Statistics, Decision Analysis, and Decision Modeling; Multiple Criteria Analysis in Strategic Siting Problems, Introduction to Business Data Mining, Enterprise Risk Management, Advanced Data Mining Techniques, and New Frontiers in Enterprise Risk Management. He is associate editor of Service Business and co-editor in chief of International Journal of Services Sciences. He has made over 100 presentations at international and national conferences on research topics. He is a member of the Association for Information Systems, the Decision Sciences Institute, the Institute for Operations Research and Management Sciences, and the Multiple Criteria Decision Making Society. He was named Best Enterprise Information Systems Educator by IFIP. He is a Fellow of the Decision Sciences Institute.