A Conversation with Professor Ruqian Lu et al.

Shuigeng Zhou
Fudan University
sgzhou@fudan.edu.cn

Fei Wang
Fudan University
wangfei@fudan.edu.cn

Shanfeng Zhu
Fudan University
zhusf@fudan.edu.cn

Caiyan Jia
Fudan University, Beijing Jiaotong University
cyjia@bjtu.edu.cn

Qiwen Dong
Fudan University
qwdong@fudan.eud.cn

Ruqian Lu
Fudan University, Academy of Systems Science and Mathematics
rqlu@math.ac.cn

1. Please share with us your view on the history and important milestones of the Chinese KDD research and application areas.

To the best of our knowledge, the research on automated knowledge discovery in China started in the early 1990s when Yihua Wu, a PhD student supervised by Prof. Shulin Wang, who finished his thesis on discovering scientific rules from a large set of experimental data. Wu's research followed and extended the work of Professor Pat Langley in scientific knowledge discovery. At almost the same time, Prof. Zhongzhi Shi published his book 'Principles of Machine Learning' where knowledge discovery is extensively discussed. The first three papers on data mining appeared in Chinese academic journals and were published in 1997.

Two events played very important role in pushing KDD research forward in China. The first one was the 1998 IBM Summer School on Data Mining and Data Warehousing. Prof. Jiawei Han and Prof. Hongjun Lu were the two only lecturers of the summer school, which attracted almost 100 attendees from the major universities of China. The second event is PAKDD 1999 held in Beijing, which was the first international conference of KDD organized in China (the second such event PAKDD 2007 was held 8 years later in Nanjing, and the third event PAKDD 2011 was held 12 years later in Shenzhen). This event provided a valuable opportunity for Chinese researchers and students to show their early achievements on KDD.

The most recent and significant research on data mining and knowledge discovery is knowledge mining from the Web and the Internet. The Internet is a huge, dynamic and open reservoir of data and knowledge. An important work is Prof. Cungen Cao's research on concept acquisition from Internet data. The result of his research is a knowledge base with more than 3 million concepts, more than 4 million concept relationships, and more than 80,000 concept attributes.

We are very happy that KDD 2012 will be held in China, and we believe that this event will surely inject new impetus to KDD research in China, and will play an essential role in pushing the internationalization and applications of China's KDD research.

2. Please describe your expertise and contribution to KDD.

Our contributions to KDD are mainly in biological data mining area.

- **Microarray data mining**

A 3D cluster over gene-sample-time (simply GST) microarray data may contain information on useful phenotypes, their related potential genes and expression rules. Our 3D cluster-mining algorithm gTRICLUSTER is based on a more general model and can find more biologically meaningful coherent gene clusters than the existing one. It also outperforms TRICLUSTER [1] in robustness in noisy data.

Bayesian Networks (BN) can be used to predict gene regulatory networks using microarray data, but they suffer from insufficient samples and huge search space. Our new method of learning BN introduces fuzzy clustering algorithm to reduce the search space. From the view of systems biology, modularity and hierarchy are key features of biological networks. This allows us to gain insight into global biological networks by assembling local components.

- **Promoter prediction based on pattern mining**

Core promoters are a key clue in understanding gene regulations. We devised a uniform framework by using pattern-based nearest neighbor search to predict promoters based on the structural features of promoters. We first mine structural patterns from sequences, and then evaluate the similarity between sequences based on the pattern similarity. The proposed pattern-based approach can significantly improve performance of promoter prediction.

- **Protein domain, location structure and function prediction**

In this area, our research focused on domain prediction system (KemaDom) based on ab initio method, established by assembling three kernel machines with local context information among neighboring amino acids; prediction of protein subcellular localization based on Gene Ontology Transfer; prediction of protein subnuclear organelles based on K-spectrum kernel

method; taxonomic fold recognition using SVM with autocross-covariance (ACC) transformation; prediction of 3D structures of proteins from amino acid sequences based on effective interaction potentials; prediction of protein-protein interaction sites based on effective ensemble learning method; iterative semi-supervised algorithm (SemiHS) improving the accuracy of hot spots prediction.

- **miRNA prediction and classification**

An ensemble SVM classifier was established to deal with the imbalance issue of miRNA gene mining where multi-loop features are included for identifying the pre-miRNAs with multi-loops. Besides, a multiclass SVM employing n-grams to extract features from known precursor sequences was proposed to classify new miRNAs.

- **Biomedical text mining**

A new strategy for clustering MEDLINE documents that incorporates MeSH (Medical Subject Heading) similarity was proposed. We also proposed a new finite mixture model FICM for clustering multiple-field documents. Experimental results show that FICM outperformed the classical multinomial model and the multivariate Bernoulli model at a statistically significant level.

- **Immunoinformatics**

Another focus of our research is predicting antigenic peptides binding to major histocompatibility complex (MHC) molecules, whichis the prerequisite of cellular immune responses. We developed a new Web server, MetaMHC, to integrate the outputs of leading predictors by several popular ensemble strategies. MetaMHCI achieved good prediction results in an international competition, and outperformed two well-known predictors in all six categories and was awarded the winner in one of the categories.

- **Generalized center string problem is fixed parameter tractable**

This is a problem in which one is asked to find all substrings of length *l* coinciding with a (hidden) motif up to *d* mismatches in at least *q* of the N strings of length L. We proved that this problem is fixed parameter tractable with our Bpriori algorithm.

3. Please share with us your view on the future of KDD both in China and the world.

Effectiveness and efficiency will still be the focus of KDD algorithm research. On one hand, to develop highly effective KDD algorithms and tools, we need essential break-through in KDD theories and techniques, which greatly rely on the advances in statistics, machine learning and other data analysis disciplines. On the other hand, for boosting the efficiency of KDD processes, there are two possible research directions: 1) developing high-performance computing paradigms or platforms to support KDD tasks, 2) exploring novel data access methods to enable efficiently massive data mining.

Generally speaking, KDD techniques are domain-specific. We believe that the most promising KDD research areas in the future are:

- KDD in networked data such as Web data and social networks, where graph mining is an important and challenging research direction.
- KDD in biological data/texts to perform genome-wide tasks such as sequence analysis, genes finding, protein structure and function as well as interaction prediction, and drug discovery etc.

KDD in financial data. In particular, real time (non NP-complete) stream data mining algorithms are needed to detect fraud behaviors and potential risks in financial markets.

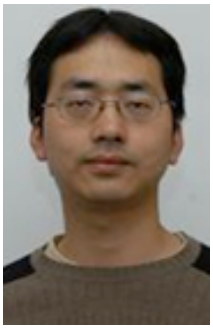
4. REFERENCES

[1] Lizhuang Zhao and Mohammed J. Zaki. 2005. TRICLUSTER: an effective algorithm for mining coherent clusters in 3D microarray data. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* (SIGMOD '05). ACM, New York, NY, USA, 694-705.

About the authors:

Shuigeng Zhou is a professor of School of Computer Science, and the Standing Vice director of Shanghai Key Lab of Intelligent Information Processing at Fudan University. He received his Bachelor degree from Huazhong University of Science and Technology (HUST) in 1988, his Master degree from University of Electronic Science and Technology of China (UESTC) in 1991, and his PhD of Computer Science from Fudan University in 2000. He served in Shanghai Academy of Spaceflight Technology from 1991 to 1997, as an engineer and a senior engineer (since August 1995) respectively. He was a post-doctoral researcher in the State Key Lab of Software Engineering, Wuhan University from 2000 to 2002. His research interests include data management and search in network environments, mining and learning over/from massive datasets, and bioinformatics. He has published more than 100 papers in domestic and international journals (including IEEE TKDE, IEEE TPDS, IEEE TCBB, Bioinformatics, BMC Bioinformatics, DKE, PRE, EPL and EPJB etc.) and conferences (including SIGMOD, VLDB, ICDE, SIGKDD, ICDCS, SIGIR, and BIBE etc.). Currently he is member of IEEE, ACM and IEICE.

Fei Wang is associate professor at Shanghai Key Lab of Intelligent Information Processing and School of Computer Science, Fudan University. She obtained her bachelor degree and Ph.D degree of Computer Science from Jilin University in 1996 and 2001, respectively. She joined Fudan University and Fudan Open Lab of Intelligent Information Processing in 2001, which became Shanghai Key Lab of Intelligent Information Processing in 2005. Her research interests include machine learning, Bayes nets, inexact reasoning and bioinformatics. She published about 30 papers and has been principal investigator of projects titled gene data warehouse, intelligent multimedia expert system, intelligent agriculture management platform, spatial reasoning and implicative methodology of inexact reasoning.

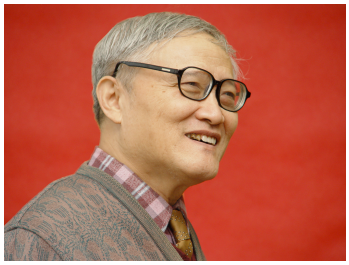


Shanfeng Zhu is an associate professor at Shanghai Key Lab of Intelligent Information Processing and School of Computer Science, Fudan University. He obtained his bachelor degree and Master degree of Computer Science from Wuhan University in 1996 and 1999, respectively. In 2003, he was awarded Ph.D. degree on Computer Science at City University of Hong Kong. Before joining Fudan University in July 2008, he was a JSPS postdoctoral fellow at Bioinformatics Center, Institute for Chemical Research, Kyoto University. His research interests focus on the development of machine learning and data mining algorithms, as well as their applications in Bioinformatics and information retrieval.

Caiyan Jia is an associate professor at the School of Computer and Information Technology, Beijing Jiaotong University since 2006. She received her PhD degree from Institute of Computing Technology, Chinese Academy of Sciences in 2004, followed by a. postdoctoral fellowship in Shanghai Key Lab of Intelligent Information Processing, Fudan University. Her research interests include data mining, bioinformatics, social computing and complex network analysis.

Qiwen Dong is now an associate professor at Shanghai Key Lab of Intelligent Information Processing and School of Computer Science, Fudan University. He received the BE, ME, and PhD degrees from Harbin Institute of Technology.Currently, he is a postdoctor at Fudan University. His current research interests include computational investigation of sequence-structure function relationships in proteins and language model of biological sequence.

http://www.iipl.fudan.edu.cn/staff/dongqw/en.html



Ruqian Lu got his Diplom-Mathematiker degree from Jena University, Department of Mathematics, Germany, in 1959. He worked in the area of function theory of several complex variables and moved to computer science in 1972. Now he is a professor of computer science of the Institute of Mathematics, Academy of Mathematics and Systems Science, at the same time an adjunct professor of Institute of Computing Technology, Chinese Academy of Sciences, Fudan University, Beijing University of Technology and Peking University. He is Chair of Academic Committee of numerous institutions, including Shanghai Key Lab of Intelligent Information Processing, Beijing Key Lab of Multimedia and Intelligent Software, Ministry of Education's Key Lab of Data and Knowledge Engineering and Jiangsu Province's Key Lab of e-Commerce. He was elected to fellow of Chinese Academy of Sciences in 1999. He is the Editor-in-Chief of the International Journal of Software and Informatics, and the Executive Editor-in-Chief of the Journal Frontiers of Computer Science. His research interests include artificial intelligence, knowledge engineering, knowledge based software engineering, formal semantics of programming languages and quantum information processing. He has published about 150 papers and 10 books. He has won two first class awards (in 1981 and 1992 resp.) from the Chinese Academy of Sciences and a National second-class prize (1993) from the Ministry of Science and Technology. He has also won the Hua Loo-keng Mathematics Prize from the Chinese Mathematics Society in 2003.

http://www.math.ac.cn/index_e/Personal_Web/luruqian_E.htm