

Guste Editors' Notes on the special issue

“Data Mining in Chemistry and Traditional Chinese Medicine”

The main objective of this special issue is to address some problems in data mining on the huge amount of data from chemistry and traditional Chinese medicine. From our experience based on a large joint research project by groups in both statistics and chemistry over several years, we developed strong feeling that there might be a great opportunity for scientists to discover meaningful results.

The idea for a special issue on this was formed two years ago, when we discussed some topics represented in “The Gordon Conference on Statistics in Chemistry and Chemical Engineering” held in Williamstown in 2001. After several workshops the motivation to publish a special volume on data mining in chemistry and traditional Chinese medicine became stronger and stronger. Thanks to the enthusiastic encouragement and support from Prof. Min-Te Chao, it comes now in this form in the *Journal of Data Science*.

Chemistry has essentially been a strongly experience-dependent scientific discipline, which studies constituents, properties, structure and variety of matter. As the number of known chemical compounds is more than twenty million, certainly chemical experiments and measurements contain a great deal of knowledge. Unfortunately, we have not witnessed corresponding advances in computational techniques to help us to analyze the accumulated data. As pointed out by the book “The perspective on chemistry in 21st century” compiled by the National Natural Science Foundation of China (NSFC) and the Chinese Academy of Sciences, “After accumulation for a long time, ever-growing mountains of chemical data now contain a large amount of information, and how to discover the chemical knowledge lying dormant in huge databases is a great task for chemists.” Prof. G. X. Xu, a famous Chinese chemist and a member of Chinese Academy of Sciences, also stressed the point that the large accumulation of measurement data often leads to great discoveries, such as Mendeleev’s element periodical table and Pauling’s theory of the chemical bond. Today, large-scale chemical data will give us another great opportunity, and we should not miss it. Such words really give us great inspiration to do the research project.

Traditional Chinese medicine (TCM) has a long therapeutic history of thousands of years but its chemical background and formula synergic effects are still a mystery because of its complex phytochemicals. As pointed out in “General Guidelines for Methodolo-

gies on Research and Evaluation of Traditional Medicines” (World Health Organization 2000), “Despite its existence and continued use over many centuries, and its popularity and extensive use during the last decade, [traditional] medicine has not been officially recognized in most countries. Consequently, education, training and research in this area have not been accorded due attention and support. The quantity and quality of the safety and efficacy data on traditional medicine are far from sufficient to meet the criteria needed to support its use worldwide. The reasons for the lack of research data are due to not only to health care policies, but also to a lack of adequate or accepted research methodology for evaluating traditional medicine.” It seems to be the consensus that multivariate technique may be barely satisfactory for quality control of TCM and TCM products with multiple chemical components, such as chromatographic fingerprinting of the TCM. However, how to quantitatively and qualitatively analyze the chromatographic fingerprint of the TCM and, how to use multiple chemical components to address the quantity and quality of the safety and efficacy data on traditional medicine are still open questions.

The goal in data mining in chemistry and TCM is to try to extract useful information from databases, and then classify and recognize the compounds or medicines by their related molecular structure, topological index or chemical fingerprints. Newly developed techniques, such as data mining or knowledge discovery in database (KDD), might provide us with a very good opportunity to do research on the large-scale chemical and TCM data. However, in order for fruitful achievement, it is necessary and urgent for chemists and statisticians to work together to find the hidden knowledge and information in chemical and TCM data.

In this issue, we have collected some work mainly done in our joint research group on the subject. We hope that more people will be interested in this challenging subject, and that more excellent papers will come along in this direction.

K.-T. Fang
Hong Kong Baptist University
September 2, 2003

Y.-Z. Liang
Central South University