

**Data Mining: Guest Editorial**

**S.J. Hong**

**Future Generation Computer Systems**  
**November 1997**

# GUEST EDITORIAL

## Data Mining

Se June Hong

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598

The ever-increasing quantity of data in every computing environment presents both an opportunity to extract useful information and a challenge to process the massive volume of data effectively. Analyzing and generating models from data used to be in the domain of classical statistics. During the past few decades the pattern-recognition and machine-learning communities have greatly expanded their areas of application and the kind of information to be extracted, as well as the variety of models. The database community joined the endeavor in the early 90s and a new multidisciplinary field began, which we now call data mining. The term KDD (Knowledge Discovery in Databases) refers to a broader process of collecting and cleansing the data, extracting the useful information (data mining), and presenting and embedding the information in a decision support application.

This new field is growing vigorously, due in large part to the increasing awareness of the potential competitive business advantage of using such information. Important knowledge has been extracted from massive scientific data as well. Numerous conferences and journals are addressing data mining issues or specializing in them. And since data mining emphasizes the ability to deal with massive data, high performance algorithms, parallel computation and effective access to disk resident data (a concern of large database systems) all become more relevant and essential: it is timely to introduce this field to the readership of FGCS.

What is useful information depends on the application. Of course, each record in a data warehouse full of data is useful for daily operations, as in on-line transaction business, and for traditional database queries. Data mining is concerned with extracting more global information that is generally the property of the data as a whole. Thus the diverse goals of data mining algorithms include clustering the data items into groups of “similar” items, finding an explanatory or predictive model for a target attribute in terms of other attributes, finding frequent patterns and sub-patterns that co-occur with an associated sub-pattern, and finding trends, deviations, and “interesting” relations between the attributes. In this special issue, we address the three most common data mining tasks: Clustering, modelling, and finding frequent association patterns of items. These are also the areas that are most readily used in decision support applications.

The first paper on the promise and challenges by Fayyad and Stolorz is a perspective intro-

duction to this special issue based on their pioneering personal experience. The last paper by Uthurusamy, Soparkar, Szaro and Dunkel on the systems aspects of data mining concludes this special issue with a reality check based on considerations for the practical use of data mining techniques.

In the second paper, Hosking, Pednault and Sudan discuss the evolution of statistical insights on modelling from the classical approaches (mostly parameter fitting to a given model family) to the new statistical learning theory (based on VC dimension) and computational learning theory. Statistical learning theory deals with the trade-off between the complexity of the model and the defined loss function of the prediction such that selection of an appropriate model family can be an integral part of model construction. Computational learning theory identifies the learning tasks that can be PAC (probably approximately correct) learnable with given computational complexity. These new insights are beginning to be adapted to common model families such as rules, trees and neural networks.

The next two papers deal with clustering problems, also known as unsupervised learning. Customer segmentation is a widely recognized application area in business. Since grouping "similar" data elements together begs the question of the purpose for which they are similar, there are many clustering approaches, which depend on various notions of the similarity between data elements expressed in terms of the attribute values associated with the data elements. Michaud discusses these techniques and argues for a relatively new approach based on the theory of voting (i.e. each attribute votes that same valued elements belong in the same cluster). Evaluating the resultant clusters is difficult in the absence of a formally defined purpose of the application. Zait and Mesatfa present a benchmark-style comparison of some major clustering techniques using artificially generated data, which gives some idea as to what computing resources they require and how they behave in different clustering situations.

In the next paper, Srikant and Agrawal introduce association rules and present a new technique for finding generalized association rules. Given a large quantity of point-of-sale (POS) data, for instance, one would naturally like to know what items are frequently sold together (frequent item set) and, among them, what subset implies the remaining subset with high confidence level (association rules). This kind of information is even more useful when the POS data is augmented with a taxonomy hierarchy (e.g.; jackets and ski pants are outerwear; outerwear and shirts are clothes; shoes, sneakers and boots are footwear). One can then automatically find relations between classes of items, e.g. that most of the time clothes are sold some footwear is also sold. This is an example of a generalized association rule. In practice, the problem of finding such frequently occurring patterns requires that gigabytes of data can be processed efficiently.

The next set of three papers address classification and regression problems. Kononenko and Hong discuss the need for selecting essential attributes, various measures of the strength of an attribute for modelling purposes, and ways to select attributes for a given modelling situation. Apte and Weiss discuss key ideas for generating rules and trees, perhaps the most popular model family for classification and regression. Craven and Shavlik present another popular model family, neural networks, with particular emphasis on generating “understandable” rules using a neural-network approach. Neural networks are well established for predictive modelling in many areas of application, but lack of human comprehensibility of the networks made them unsuitable in some, and these rules may complement and give an insight into the underlying neural network model.

Although the impetus for the birth of data mining came mainly from the rapidly increasing size of current databases and commercial interest in utilizing the information hidden in them, data mining is concerned with issues broader than just dealing with large volumes of data. In the short history of data mining, it has already been shown that synergy between different disciplines has been fruitful in advancing the art of extracting useful information from data. Data mining algorithms must scale up to handle large quantities of data, but that is not the same as insisting that we throw away sampling techniques and algorithms that are not linear in the number of examples, if the utility of the results can be improved by using them. There are some applications where the comprehensibility of the extracted model is of prime importance, but this does not preclude the usefulness of more accurate models that may not be easily “understandable” by an end user. Data mining is all-of-the-above in these respects as well. It is an important core area within the larger framework of the KDD process, and it in turn challenges the future generation of computing techniques and computer systems. Further background on data mining can be found in *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy Editors, AAAI Press / The MIT Press, 1996. The KDNUGGETS web site, at <http://www.kdnuggets.com>, is an excellent source of news on data mining and KDD.

My editorial goal for this special issue was to cover and introduce the key ideas of data mining in a balanced perspective. These are not review papers; accordingly, authors were strongly urged to restrict the references to those that are essential for the ideas conveyed and also those that point to further references. I solicited authors who have contributed new techniques in their respective areas, and asked for papers that offer new insights rather than new techniques. I underestimated the time needed to prepare such a paper by very active and busy authors by more than six months. I thank the authors for the quality they delivered. And I thank the FGCS editor-in-chief, Prof. L.O. Hertzberger, and the editorial staff for their encouragement for the idea of a special issue on data mining and for their patience. Finally, I would like to express my gratitude to Dr. J. Hosking for the cheerful help he provided with many editorial tasks.