Introduction : knowledge quality measures in a data-mining process

Vipin Kumar*, **

* Director, Army High Performance Computing Research Center
** Professor, Department of Computer Science & Eng, University of Minnesota
kumar@cs.umn.edu
http://www.cs.umn.edu/~kumar

The author. Vipin Kumar is currently the Director of Army High Performance Computing Research Center and Professor of Computer Science and Engineering at the University of Minnesota. His research interests include High Performance computing, data mining, and their applications to information assurance. He has authored over 200 research articles, and co-edited or co-authored 9 books including the widely used text book ``Introduction to Parallel Computing", and an upcoming edited collection, "Managing Cyber Threats: Issues, Approaches and Challenges" to be published by Kluwer Academic Publishers.
Kumar has served as chair/co-chair for many conferences/workshops in the area of data mining and parallel computing, including the IEEE International Conference on Data Mining (2002) and the 15th International Parallel and Distributed Processing Symposium (2001). Kumar serves as the chair of the steering committee of the SIAM International Conference on Data Mining, and serves on the editorial boards of Knowledge and Information Systems, IEEE Computational Intelligence Bulletin, Annual Review of Intelligent Informatics, Parallel Computing, the Journal of Parallel and Distributed Computing, and has served on the editorial boards of IEEE Transactions of Data and Knowledge Engineering (93-97), IEEE Concurrency (1997-2000), and IEEE Parallel and Distributed Technology (1995-1997). He is a Fellow of IEEE, a member of SIAM, and ACM, and a Fellow of the Minnesota Supercomputer Institute.

Advances in information technology and data collection methods have led to the availability of large data sets in commercial enterprises and in a wide variety of scientific and engineering disciplines. Examples of large data sets are genomic data, climate data, and market basket data collected at commercial outlets. There is an unprecedented opportunity to analyze such data and extract intelligent and useful information. The potential for return to the society from this analysis is huge. For example, the understanding of genome expression data can lead to the development of better and cheaper drugs that have fewer side effects. The information collected from the analysis of market-basket data can improve the profitability of a corporation. Indeed, researchers in the fast growing discipline of knowledge discovery and data mining are developing automated techniques for discovering novel and useful information and patterns from large amounts of data.

The analysis of relationships among variables is a fundamental task at the heart of many data mining problems. For example, the central task of association analysis is to discover sets of binary variables (called items) that co-occur together frequently in a transaction database, while the goal of feature selection is to identify groups of variables that are highly correlated with each other, or with respect to a special target variable. Regardless of how the relationships are defined, such analyses often require a suitable measure to evaluate the dependencies between variables.

Numerous measures have been developed in diverse fields such as statistics, social science, machine learning, and data mining. For example, objective measures such as support, confidence, interest factor, correlation, and entropy are often used to evaluate the interestingness of association patterns - the stronger is the dependence relationship, the more interesting is the pattern. Similarly, measures such information gain and gini index are used for measuring the quality of rules and predictive models. A measure helps in identifying a small set of potentially interesting association patterns, rules, or predictive models out of a very large number of possibilities that are typically available for most non-trivial data sets.

However, measures do not always agree with each other. In fact, in many situations, different measures may provide conflicting information about the interestingness of a pattern. The applicability and suitability of a measure is often dependent upon the nature of the data and the application domain. The understanding of this dependence is key to selecting the right measure for the application at hand.

This special issue focuses on the critical topic of knowledge quality measures in the data mining process. Papers included in this issue present new methods for evaluating association patterns and fuzzy rules, new criteria to evaluate the goodness of measures, and systematic evaluations of several measures commonly used. It is hoped that this timely collection of papers on a very important topic will be found valuable by researchers and practitioners in data mining.