Introduction to Data Mining and Knowledge Discovery

H. Michael Chung[*], Paul Gray [**], Michael Mannino [***]

[*] California State University, Long Beach, hmchung@csulb.edu
[**] Claremont Graduate University, grayp@cgs.edu
[***] University of Colorado, Denver, mmannino@carbon.cudenver.edu

Overview

Data mining is a recent popular theme in reflecting the effort of discovering knowledge from data. It provides the techniques that allow managers to obtain managerial information from their legacy systems. Its objective is to identify valid, novel, potentially useful, and understandable correlation and patterns in data. Data mining is made possible by the very presence of the large databases.

While knowledge discovery often refers to the process of discovering useful knowledge from data, data mining focuses on the application of algorithms for extracting patterns from data. Knowledge discovery seeks to find patterns in data and to infer rules (that is, to discover new information) that queries and reports do not reveal effectively. Thus, knowledge discovery has a R&D flavor and data mining an operational process one. Data mining is a basis of knowledge discovery.

Research Issues

Starting from a simple regression analysis, information theoretic methods (Quinlan, 1979, 1983; Michalski and Chilauski, 1980), genetic algorithms (Holland, 1975; Forsyth and Rada, 1986), and neural networks (Rumelhart and McCelland, 1986; Lippman, 1987) have been applied to numerous occasions. They differ in the ways the models are generated.

Inductive machine learning approaches have been employed to build the models of human decision making with rules/patterns (Braun and Chandler, 1987; Messier and Hansen, 1988; Remus and Hill, 1990; Chung and Silver, 1992; Chung and Tam, 1994). When human decision-makers are involved, decision-making modeling often involves more complicated issues. It is because human decision-makers are under the influence of various cognitive factors (Dawes and Corrigan, 1974; Payne, 1976, 1982; Kim, Chung, and Paradice, 1997).

In addition, certain domain characteristics could affect the model performance (Chandrasekaran, 1989; 1996). Furthermore, the way the model performance is measured is an important consideration. The same model performs differently at different domains due to the quality of data values, normative criteria, and the decision-maker factors involved. Noise tolerance and sensitivity analysis is another important issue (Mookerjee and Mannino, 1995). An example of such issue is a cost minimizing inverse problem in classification systems (Mannino and Kousik, 1995).

A particularly important problem is scalability. It refers to the ability to maintain performance as the size of the data base being mined increases. Scalable mining tools take

1060-3425/98 $10.00 (c) 1998 IEEE

advantage of hardware parallelism, including better parallel algorithms as well as direct access to parallel data base management systems.

In its operation, data mining can be either bottom up (explore raw facts to find connections) or top down (search to test hypotheses). Data mining typically seeks following relationships (Gray, 1998): *Classes* in which stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order; *Clusters* in which data items are grouped by logical relationships. For example, data can be mined to identify market segments or consumer affinities; *Associations* in which data can be mined to identify associations. The beer-diaper example is typical of associative mining; *Sequential* patterns in which data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack purchase based on sleeping bag or hiking shoes sale.

Since data mining and knowledge discovery often deal with very large databases, or high dimensionality that increases the size of the search space, it is necessary to exercise caution on creating spurious patterns, over-fitting data, and handling missing data as well as changes in data over time.

Applications

The following are some examples of early successes with data mining and knowledge discovery: The size of the data problem can be seen from the explosion of retail sales data from groceries due to of the availability of bar code information. Data mining can reveal the effects of 'cents off' coupons both by the company and its competitors. Another marketing application is targeting direct advertising by determining who is more likely to become a customer. The result is a profile which is matched to determine whom to target for the new campaign.

Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express suggests products to its cardholders based on analysis of their monthly expenditures. WalMart analyzes massive data to discover its supplier relationships by capturing point-of-sale transactions from over 2,900 stores in six countries and more than 3,500 suppliers.
The National Basketball Association is exploring a data mining application that can be used in conjunction with image recordings of basketball games. Data mining analyzes the movements of players to help coaches orchestrate plays and strategies. For example, an analysis of the play-by-play sheet of the game played between the New York Knicks and the Cleveland Cavaliers reveals that when Mark Price played the Guard position, John Williams attempted four jump shots and made each one.

In addition, data mining and knowledge discovery have been applied to fraud detection as well as consumer loan analysis with significant results.

Papers Presented

The minitrack covers the theoretical issues related to data mining, learning-by-examples, knowledge acquisition, knowledge discovery, and inductive decision making. A goal for the minitrack is to build a foundation for the application of data mining and knowledge discovery from an interdisciplinary perspective including artificial intelligence, psychology, computer science, statistics, and management.

Some of the relevant topics are new data mining algorithms, analysis of existing algorithms or applications, comparison of inductive learning and data mining concepts, new ways of thinking about statistics, data mining measures, knowledge re-use, acquisition of qualitative knowledge, data visualization technique to further understand the model/data structure, maintenance and adaptation of algorithms, human factors in decision modeling, task characteristics, and economics of decision making.

This year, the mini track has a total of seven papers presented: four of them are related to foundation and theoretical issues. Three are application oriented papers. Gelman investigates the comprehensibility of the discovered patterns, which are often domain and context dependent. Gala, Cook, and Holder address an approach for scaling a knowledge discovery system using parallel and distributed resources. Piramuthu considers feature selection methods in preprocessing input data for training. Bolloju proposes an approach to discover decision making styles and decision models. In applications, Spenceley and Warren explore the role of temporal information in predictive performance of a model in an medical record area. Kawano discusses a text data mining technique for Web search. Finally, Lee and Kim describe a casual knowledge acquisition with a prototype.

Selected References

Chung, H.M. and Tam, K.Y. "A Comparative Analysis of Inductive Learning Algorithms," International Journal of Intelligent Systems in Accounting, Finance, and Management, Vol. 2, No.1, 1994, pp. 3-18.

Gray, P., "The New DSS; Data Warehousing, OLAP, MDD, and KDD," Tutorial, Thirty First Hawaii International Conference on Systems Sciences, Kohala Coast, Hawaii, January 6-9, 1998.

Kim, C., Chung, H.M., and Paradice, D. "Inductive Modeling of Expert Decision Making in Loan Evaluation: A Decision Strategy Perspective," Decision Support Systems, Vol. 20, No. 4, 1997.

Mookerjee, V. and Mannino,M., "Improving the Performance and Stability of Inductive Expert Systems Under Input Noise," Information Systems Research Vol. 6, No.4, 1995, pp 328-356.

Mannino, M. and Koushik, M., "The Cost of Minimizing Inverse Classification Problem: A Genetic Algorithm Approach," Working Paper #7-95, Department of Management Science, University of Washington, 1995.

1060-3425/98 $10.00 (c) 1998 IEEE