Exploiting the Strategic Potential of Data Mining

Chandra S. Amaravadi
Western Illinois University, USA

INTRODUCTION

Data mining is a new and exciting technology to emerge in the last decade. It uses techniques from artificial intelligence and statistics to detect interesting and useful patterns in historical data. Thus traditional techniques such as regression, Bayesian analysis, discriminant analysis, and clustering have been combined with newer techniques such as association, neural nets, machine learning, and classification (Jackson, 2002).

Applications of data mining have ranged from predicting ingredient usage in fast food restaurants (Liu, Bhattacharyya, Sclove, Chen, & Lattyak, 2001) to predicting the length of stay for hospital patients (Hogl, Muller, Stoyan & Stuhlinger 2001). See Table 1 for other representative examples. Some of the important findings are: (1) corporate bankruptcies can be predicted from variables such as the "ratio of cash flow to total assets" and "return on assets," (2) gas station transactions in the U.K. average £20 with a tendency for customers to round the purchase to the nearest £5 (Hand & Blunt, 2001), (3) 69% of dissatisfied airline customers did not contact the airline about their problem, (4) sales in fast food restaurants are seasonal and tend to peak during holidays and special events (Liu et al., 2001), (5) patients in the age group over 75 are 100% likely to exceed the standard upper limit for hospital stays (Hogl et al., 2001). In recent years, data mining has been extended to mine temporal, text, and video data as well. The study reported by Back, Toivonen, Vanharanta, and Visa (2001) analyzed the textual and quantitative content of annual reports and found that there is a difference between them and that poor organizational performance is often couched in positive terms such as "improving," "strong demand," and so forth. Both applications and algorithms are rapidly expanding. The technology is very promising for decision support in organizations. However, extracting knowledge from a warehouse is still considered somewhat of an art. This article is concerned with identifying issues relating to this problem.

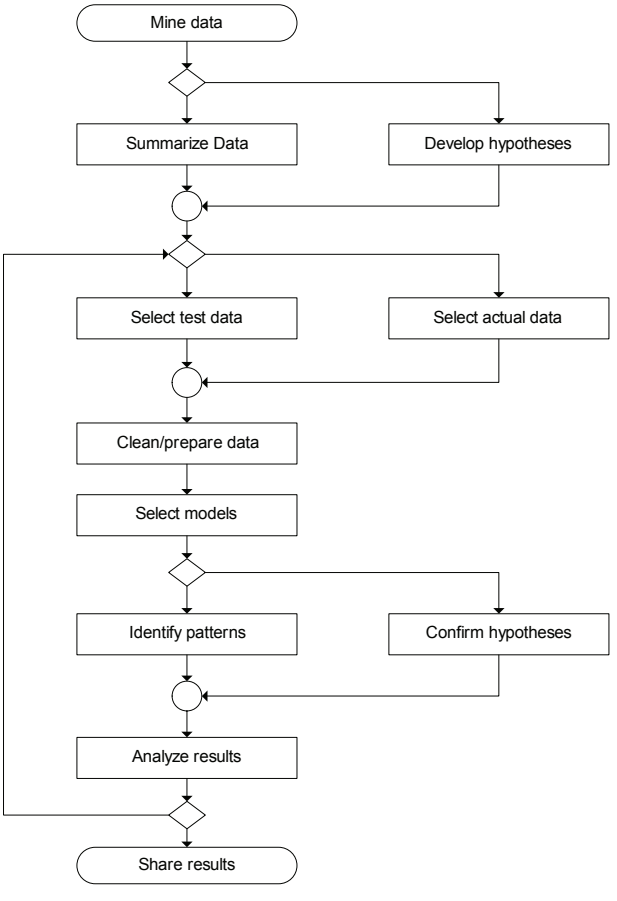
Table 1. Examples of data mining applications

• Predicting supplies in fast food restaurants (Liu, Bhattacharyya, Sclove, Chen & Lattyak 2001).
• Quality of health care (Hogl, Muller, Stoyan & Stuhlinger 2001).
• Analyzing Franchisee sales (Chen, Justis & Chong 2003).
• Predicting customer loyalty (Ng & Liu 2001).
• Mining credit card data (Hand & Blunt 2001).
• Analyzing annual reports (Back et al. 2001).

BACKGROUND

The mining process is often labeled as knowledge discovery in databases (KDD). An extended KDD model is presented in Figure 1. As illustrated in the figure, mining is carried out in two modes: "data-driven" or "hypothesis-driven" ("question-driven"). The *data-driven* approach is also referred

Figure 1. The extended KDD process



to as exploratory data mining and is often carried out to develop a preliminary understanding of the data. As a first step, data are summarized to identify their characteristics. Typical measures such as frequency distribution, mean, variance, and so forth are computed. For credit card data, what is the average monthly spending for customers? What credit card products are most frequently purchased? Based on results from this step, models are selected and the mining is performed.

Hypothesis-driven methods attempt to verify whether or not a particular pattern exists (Hogl et al., 2001). In this mode, the process starts with the identification of hypotheses that are motivated by business concerns. Each hypothesis can be thought of as a micro theory (MT) about the domain or an assumption to be verified.

The space of patterns that can be explored in data mining is very large and is restricted by computational constraints. Large data sets with high dimensionality will preclude data driven methods. Hypothesis driven approaches are more tractable and therefore more appealing. But even with these approaches, organizations still lack the resources to mine all available data. Formulation of good hypotheses is therefore important and will influence data selection.

The mining data are partitioned into a test set and the evaluation set (Jackson, 2002). The test set is generally 10-20% of the actual data. Neural nets and machine learning algorithms typically require more test data. Models developed with the test set are validated with the actual data. In the question driven approach, the required columns are selected depending on the hypothesis to be tested. This is a difficult problem ("curse of dimensionality") in data mining and sometimes referred to as *feature selection*. The large number of attributes such as demographic variables in a warehouse makes data selection a challenging process because the variables contributing to a pattern cannot be known a priori. Selection of irrelevant attributes adds to computational complexity and perhaps even misleading results. At the present time, the best method is trial and error.

Mining starts with data that are often integrated from several sources. Integration can present challenges due to differences in formats or attribute definitions. The data are cleaned and transformed by (Peacock, 1998): (a) checking for file transfer errors, where some of the records are not properly loaded or some of the columns are missing, (b) standardizing formats or codes such as converting from text to numeric codes or vice versa (1-married, 2-single, etc.), (c) dealing with sparse records; missing values are either filled or the record is deleted, (d) performing required calculations such as debt/equity, number of faculty/student, student credit hour per faculty, and so forth, (e) identifying and deleting outliers. Outliers are detected during the process of data summarization.

Data selection is followed by *model selection*. As shown in Table 2, mining techniques are broadly classified into summarization/visualization, clustering, classification, association, and sequence with a choice of algorithms in each. The limitations and conditions of each algorithm have to be borne in mind when selecting a suitable algorithm. For example, the decision tree approach may produce erroneous results for data with small training sets. Similarly, time series analyses are computationally expensive and cannot be effectively carried out on large data sets (Kumar, 2002).

Testing with the model will result in initial results. The test data are iteratively used to refine the model, and the analysis is run on the evaluation set. Results can take the form of cluster plots, dependency rules, regression coefficients, bar graphs, and so forth. These are analyzed to identify patterns and discussed with functional area specialists to ensure that they are meaningful in the organizational context. Findings are then shared.

THE KDD PROCESS AND DOMAIN UNDERSTANDING

Data mining results in the identification of patterns. Identifying those that are relevant and interesting to the organization is dependent on the analyst's skill, experience, and understanding of the *organizational context* of the data. To develop a general understanding of the organization, the analyst can look through the company's annual reports, Web site, and news articles. The analyst can also meet with managers working in the area to become familiar with its goals, objectives, and critical issues. For example, a manager may be concerned about how to improve catalog mailings. Another may be concerned about increasing catalog sales. While this general understanding is critical in KDD activities, the analyst must also develop a detailed understanding of the data, that is, the relationships among variables. We will discuss the relationship between important KDD activities and this understanding.

Data Selection/Hypothesis Formation

Knowledge of the domain is essential in pre-mining as well as post-mining activities. Hypothesis formation, data selection, and transformation require conceptualizing relationships among attributes (dimensions) and their impact. For example, converting numerical income data in a bank warehouse to categorical attributes ("hi," "mid," "lo") requires knowledge of expected income levels. Generally miners anticipate two types of patterns: (a) attributes concerning an entity or issue of interest such as computer defects, user complaints, employees/suppliers, and so forth; (b) attributes influencing organizationally relevant behavior such as hiring/firing, bankruptcy, and hospital stay. Some questions that might be asked are: