Data Mining in Practice

Sherry Y. Chen
Brunel University, UK

Xiaohui Liu
Brunel University, UK

INTRODUCTION

There is an explosion in the amount of data that organizations generate, collect, and store. Organizations are gradually relying more on new technologies to access, analyze, summarize, and interpret information intelligently and automatically. Data mining, therefore, has become a research area with increased importance (Amaratunga & Cabrera, 2004). Data mining is the search for valuable information in large volumes of data (Hand, Mannila, & Smyth, 2001). It can discover hidden relationships, patterns, and interdependencies, and generate rules to predict the correlations, which can help the organizations make critical decisions faster or with a greater degree of confidence (Gargano & Raggad, 1999).

There is a wide range of data mining techniques that has been successfully used in many applications. This paper is an attempt to provide an overview of existing data-mining applications. The paper begins by building a theoretical background to explain key tasks that data mining can achieve. It then moves to discuss applications domains that data mining can support. The paper identifies three common application domains, including bioinformatics data, electronic commerce, and search engines. For each domain, how data mining can enhance the functions will be described. Subsequently, the limitations of current research will be addressed, followed by a discussion of directions for future research.

BACKGROUND

Data mining can be used to achieve many types of tasks. Based on the types of knowledge to be discovered, it can be broadly divided into supervised learning and unsupervised learning. The former requires the data to be preclassified. Each item is associated with a unique label, signifying the class in which the item belongs. In contrast, the latter does not require preclassification of the data and can form groups that share common characteristics (Nolan, 2002). To achieve these two main tasks, four data mining approaches are commonly used: classification, clustering, association rules, and visualization.

Classifications

Classification, which is a process of supervised learning, is an important issue in data mining. It refers to discovering predictive patterns where a predicted attribute is nominal or categorical. The predicted attribute is called the class. Subsequently, a data item is assigned to one of predefined sets of classes by examining its attributes (Changchien & Lu, 2001). One example of classification applications is to analyze the functions of genes on the basis of predefined classes that biologists set (see "Classifying Gene Functions").

Clustering

Clustering is also known as *Exploratory Data Analysis* (EDA; Kohonen, 2000). This approach is used in those situations where a training set of preclassified records is unavailable. Objects are divided into groups based on their similarities. Each group, called a cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups (Roussinov & Zhao, 2003). From a data mining perspective, clustering is unsupervised learning of a hidden data concept. One of the major applications of clustering is the management of customers' relationships, which is described in "Customer Management."

Association Rules

Association rules, which were first proposed by Agrawal and Srikant (1994), are mainly used to find out the meaningful relationships between items or features that occur synchronously in databases (Wang, Chuang, Hsu, & Keh, 2004). This approach is useful when one has an idea of the different associations that are being sought out. This is because one can find all kinds of correlations in a large data set. It has been widely applied to extract knowledge from Web log data (Lee, Kim, Chung, & Kwon, 2002). In particular, it is very popular among marketing managers and retailers in electronic commerce who want to find associative patterns among products (see "Market Basket Analysis").

Visualization

The visualization approach to data mining is based on an assumption that human beings are very good at perceiving structure in visual forms. The basic idea is to present the data in some visual form, allowing the human to gain insight from the data, draw conclusions, and directly interact with the data (Ankerst, 2001). Since the user is directly involved in the exploration process, shifting and adjusting the exploration goals is automatically done if necessary (Keim, 2002). This approach is especially useful when little is known about the data and the exploration goals are vague. One example of using visualization is author co-citation analysis (see "Author Cocitation Analyses").

DATA-MINING APPLICATIONS

As mentioned in the aforementioned discussion, data mining can be used to achieve various types of tasks, such as classification, clustering, association rules, and visualization. These tasks have been implemented in many application domains. The main application domains that data mining can support include bioinformatics data, electronic commerce, and search engines.

Bioinformatics Data

In the past years, bioinformatics has been overwhelmed with increasing floods of data gathered by the Human Genome Project. Consequently, a major challenge in bioinformatics is extracting useful information from these data. To face this challenge, it is necessary to develop an advanced computational approach for data analysis. Data mining provides such potentials. Three application areas, which are commonly presented in the literature, are described below.

Clustering Microarray Data

Unsupervised learning produces clustering algorithms, which are being applied to DNA microarray data sets. The clustering algorithms are often incorporated into the analysis software of microarray images and are therefore frequently used to visualize local and global relationships among hybridization signals captured by the array. Currently, hierarchical clustering is the most popular technique employed for microarray data analysis. Basically, this approach is based on similarity or dissimilarity to proceed successively by either merging smaller clusters into larger ones or by splitting larger clusters (Moller-Levet, Cho, Yin, & Wolkenhauer, 2003). The results of the algorithm are a tree of clusters called a *dendrogram* (Liu & Kellam, 2003). One disadvantage of this approach is that clustering algorithms are very sensitive to noisy data and errors in the data set.

Classifying Gene Functions

Biologists often know a subset of genes involved in a biological pathway of interest and wish to discover other genes that can be assigned to the same pathway (Ng & Tan, 2003). Unlike clustering, which processes genes based on their similarity, classification can learn to classify new genes based on predefined classes, taking advantage of the domain knowledge already possessed by the biologists. Therefore, the classification approach seems more suitable than clustering for the classification of gene functions. To conduct the classification, we need supervised learning to assign pathway memberships that correspond well to the true underlying biological pathways.

Identifying Phenotype Data

In the two aforementioned approaches, the genes are treated as objects while the samples are the attributes. Conversely, the samples can be considered as the objects and the genes as the attributes. In this approach, the samples can be partitioned into homogeneous groups. Each group may correspond to some particular phenotype (Golub, et al., 1999). A phenotype is the observable and physical characteristics of an organism. Over the past decade, growing interest has surfaced in recognizing relationships between the genotypes and phenotypes. Tracing a phenotype over time may provide a longitudinal record for the evolution of a disease and the response to a therapeutic intervention. This approach is analogous to removing components from a machine and then attempting to operate the machine under different conditions to diagnose the role of the missing component. Functions of genes can be determined by removing the gene and observing the resulting effect on the organism's phenotype.

Electronic Commerce

The widespread use of the Web has tremendous impact on the way organizations interact with their partners and customers. Many organizations consider analyzing customers' behavior, developing marketing strategies to create new consuming markets, and discovering hidden loyal customers as the key factors of success. Therefore, new techniques to promote electronic business become essential, and data mining is one of the most popular