Expert Knowledge in Data Mining

D

Anthony Scime
The College at Brockport, State University of New York, USA

INTRODUCTION

Data mining involves the analysis of data to find interesting patterns and previously unknown relationships in data. Data mining not only predicts the results of a future event, but it also can provide knowledge about the structure and interrelationships among the data. These predictions and relationships are expressed as decision trees, classification rules, association rules, or clusters.

But, data mining occurs in a domain. The data mining algorithms operate on data that was collected, or is now being used for, a specific purpose. The data is being used to study a domain question. Both the question and the data selected to answer the question need to be identified by a domain expert. This expert must drive the data cleaning process and act as a participant in the data mining process.

BACKGROUND

Data mining (also known as Knowledge Discovery from Data or KDD) is a term used to describe a number of analytical techniques that can be used to identify meaningful relationships in data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Data mining models can make predictions for individual records using complex sets of rules found in the data. Additionally, data mining defines relationships in the data (Scime, Murray, Huang, & Brownstein-Evans, 2008; Chang, 2006). "In contrast to more conventional multivariate statistical methods such as factor analysis, principal component analysis, and multidimensional scaling, they [data mining techniques] tend to be less bound by a priori assumptions" (Spielman & Thill, 2008, p. 111).

Data mining is a data-intense analytical technique that is designed to exploit large data sets. It involves the analysis of data to find interesting patterns, confirm and probe previously known relationships, and detect previously unknown relationships in the data. Data mining models not only predict the results of a future event, but they also can provide knowledge about the structure and interrelationships among the data. It is these interrelationships that can lead to a better understanding of the data. As a discipline, data mining has its origins in artificial intelligence, machine learning, and statistics.

There are many data mining techniques. Three of the major techniques are classification, association, and clustering. Classification analysis constructs a decision tree model, finding a path to a predetermined dependent or target variable for each data record. A classification decision tree contains branches that can be converted to rules unique to the dataset, but applicable to future similar datasets. Research in data classification evolved from two sources. In statistics, CHAID (Chi-Squared Automatic Interaction Detection) (Kass, 1980) is a well known classification method that uses the chi-squared statistic to determine model structure. Machine learning research produced a number of classification methods, the best known of which is the C4.5 algorithm (Quinlan, 1993), which uses information gain to define the model's structure. Both of these techniques produce a classification decision tree from which rules can be easily derived.

Association mining, which is a product of machine learning research, is used to find patterns of data that show conditions where sets of variables and their values occur frequently in the data set. With association mining, there is no predetermination of a target variable. Apriori (Agrawal, Imieliński, & Swami, 1993) is the predominant association mining algorithm. It is an algorithm that produces many rules, and domain expertise and special techniques are needed to reduce the rule set to those that are interesting and actionable.

Clustering is used to find groupings of data that show where data records occur in the multidimensional problem space, where each variable is represented as a dimension. It is often used to determine relationships between the data records. The most popular cluster-

DOI: 10.4018/978-1-4666-5888-2.ch171

ing algorithm is k-means (MacQueen, 1967). Again, analysis of the clusters needs special techniques and domain expertise.

Data mining techniques produce models of the data and domain often in the form of rules. There may be a large number of rules, from which the most useful ones must be identified. Rules can be selected and reduced mechanically and with various levels of guidance from the domain expert. There has been extensive work on the mechanical selection and reduction of classification and association rules, in which the use of domain expertise is limited (Jaroszewicz & Simovici, 2004; Deshpande & Karypis, 2002; Freitas, 2000; Padmanabhan & Tuzhilin, 2000). But the domain expert often plays the primary role. For instance, the Perception-based Classification (PBC; Ankerst, Ester, & Kriegel, 2000) system presents to the expert a classification tree node, an attribute. Based on domain knowledge, the expert approves this node, manually selects another node, or asks the system to look ahead for another node. Once the expert approves a node, using domain knowledge, the expert can assign a class label, manually select a split point, instruct the system to select a split point, direct the system to expand the node, or remove the node from the tree. The process then continues with the next node. The user and the computer together create the classification tree, which is then converted into rules.

Finally, the evaluation of multiple models is common in data mining (Osei-Bryson, 2004), and domain expertise is needed to set the criteria for and identify the most promising model (Ali & Smith, 2005; Andoh-Baidoo & Osei-Bryson, 2007; Scime & Murray, 2007).

Expert domain knowledge has been applied to the American National Election Studies (ANES, 2005) and National Survey on Child and Adolescent Well-Being (NSCAW, 2006) data sets and to a data set constructed from the Global Terrorism Database (2009), the Correlates of War (2009), the Database of Political Institutions (2006), and the World Development Indicators (2008).

The American National Election Studies (2005) is an ongoing, long-term series of public opinion surveys intended to produce research-quality data for experts who study the theoretical and empirical bases of American national election outcomes using voting behavior, public attitudes, and measures of political participation.

The ANES data set is used primarily in the field of political science and contains a large number of records (more than 47,000) and attributes (more than 900). Data mining combined with domain expertise reduced the number attributes necessary to predict the presidential vote choice to 13 attributes (Scime & Murray, 2007). This model yields correct results 65.6% of the time. Previous studies using statistical techniques have shown only 50.6% accuracy.

Another important issue in political science is the likelihood of a citizen voting in an election. Again using the ANES, but selecting a different set of attributes and records the domain expert data mined to identify two survey questions that together can be used to categorize citizens as voters or non-voters. These results met or surpassed the accuracy rates of previous non-data mining models (Murray, Riley, & Scime, 2009).

The National Survey on Child and Adolescent Well-Being (2006) is a rich collection of data designed to represent children and families who enter the child welfare system. These data are appropriate for analysis of child welfare outcomes such as the safety, permanence of care, and well-being of children.

The NSCAW data set is used in the fields of social work and child welfare and contains 5,501 records and more than 20,000 attributes. Using classification data mining, the domain expert reduced the number of attributes necessary to understand child placement conditions in homes to eight with 84.3% accuracy. As significant, the resulting rules identify important relationships regarding the living arrangements of maltreated children; suggesting vital inflection points pertaining to the living arrangements for these children (Scime, Murray, Huang, & Brownstein-Evans, 2008).

Terrorism analysts have related a number of social, political, and economic conditions at the national level with the likelihood that a nation will fall victim to a terrorist event. The domain expert constructed a unique data set comprised of terrorism events and measures of social, political, and economic contexts in 185 countries worldwide between the years 1970 and 2004. The domain expert selected attributes and instances from the Global Terrorism Database (2009), Correlates of War (2009), Database of Political Institutions (2006), and the World Development Indicators (2008) to construct a data set. This data set contained 126 attributes and 5431 records. Analysis using association mining reduced the number of attributes to