Process-Based Data Mining

Karim K. Hirji
AGF Management Ltd, Canada

INTRODUCTION

In contrast to the Industrial Revolution, the Digital Revolution is happening much more quickly. For example, in 1946, the world's first programmable computer, the Electronic Numerical Integrator and Computer (ENIAC), stood 10 feet tall, stretched 150 feet wide, cost millions of dollars, and could execute up to 5,000 operations per second. Twenty-five years later, Intel packed 12 times ENIAC's processing power into a 12–square-millimeter chip. Today's personal computers with Pentium processors perform in excess of 400 million instructions per second. Database systems, a subfield of computer science, has also met with notable accelerated advances. A major strength of database systems is their ability to store volumes of complex, hierarchical, heterogeneous, and time-variant data and to provide rapid access to information while correctly capturing and reflecting database updates.

Together with the advances in database systems, our relationship with data has evolved from the prerelational and relational period to the data-warehouse period. Today, we are in the knowledge-discovery and data-mining (KDDM) period where the emphasis is not so much on identifying ways to store data or on consolidating and aggregating data to provide a single, unified perspective. Rather, the emphasis of KDDM is on sifting through large volumes of historical data for new and valuable information that will lead to competitive advantage. The evolution to KDDM is natural since our capabilities to produce, collect, and store information have grown exponentially. Debit cards, electronic banking, e-commerce transactions, the widespread introduction of bar codes for commercial products, and advances in both mobile technology and remote sensing data-capture devices have all contributed to the mountains of data stored in business, government, and academic databases. Traditional analytical techniques, especially standard query and reporting and online analytical processing, are ineffective in situations involving large amounts of data and where the exact nature of information one wishes to extract is uncertain.

Data mining has thus emerged as a class of analytical techniques that go beyond statistics and that aim at examining large quantities of data; data mining is clearly relevant for the current KDDM period. According to Hirji (2001), data mining is the analysis and nontrivial extraction of data from databases for the purpose of discovering new and valuable information, in the form of patterns and rules, from relationships between data elements. Data mining is receiving widespread attention in the academic and public press literature (Berry & Linoff, 2000; Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Kohavi, Rothleder, & Simoudis, 2002; Newton, Kendziorski, Richmond, & Blattner, 2001; Venter, Adams, & Myers, 2001; Zhang, Wang, Ravindranathan, & Miles, 2002), and case studies and anecdotal evidence to date suggest that organizations are increasingly investigating the potential of data-mining technology to deliver competitive advantage.

As a multidisciplinary field, data mining draws from many diverse areas such as artificial intelligence, database theory, data visualization, marketing, mathematics, operations research, pattern recognition, and statistics. Research into data mining has thus far focused on developing new algorithms and tools (Dehaspe & Toivonen, 1999; Deutsch, 2003; Jiang, Pei, & Zhang, 2003; Lee, Stolfo, & Mok, 2000; Washio & Motoda, 2003) and on identifying future application areas (Alizadeh et al., 2000; Li, Li, Zhu, & Ogihara, 2002; Page & Craven, 2003; Spangler, May, & Vargas, 1999). As a relatively new field of study, it is not surprising that data-mining research is not equally well developed in all areas. To date, no theory-based process model of data mining has emerged. The lack of a formal process model to guide the data-mining effort as well as identification of relevant factors that contribute to effectiveness is becoming more critical as data-mining interest and deployment intensifies. The emphasis of this article is to present a process for executing data-mining projects.

BACKGROUND

The fields of machine learning, pattern recognition, and statistics have formed the basis for much of the developments in data-mining algorithms. The field of statistics is one of the oldest disciplines concerned with automatically finding structure in examples. Discriminant analysis (Fisher, 1936), for example, is the oldest mathematical classification technique used to separate data into classes

by generating lines, planes, or hyperplanes. Through the pioneering work on classification and regression trees (CART) by Breiman, Friedman, Olshen, and Stone (1984), the statistical community has made an important contribution in legitimizing the use of decision trees, in data mining, for classification and regression. Pattern-recognition research emphasizes the creation of machines that can perform tasks more accurately, faster, and cheaper than humans (Fukunaga, 1972; Ripley, 1993), and has made an important contribution to data mining by popularizing the use of neural networks. A feed-forward neural network is a network in which the nodes (or processing units) are numbered so that all connections go from a node to one with a higher number. In practice, the nodes are arranged in layers with connections only to higher layers. Back propagation is an implementation for a feed-forward neural network in which error terms, from the output layer, are propagated back to the input layer so that the resulting connection weights at each node adjusted can be adjusted by means of an error-minimization method called gradient descent.

The multitude of data-mining algorithms can be linked to three main data-mining-problem approaches: clustering, association and sequential pattern discovery, and predictive modeling. Clustering (or segmentation) is concerned with partitioning data records into subsets. The *K*-means clustering algorithm is used for demographic clustering because categorical data are predominant. This algorithm, which is efficient for large databases, clusters a data set by determining the cluster to which a record fits best. Once clusters have been found in a data set, they can be used to classify new data. To uncover affinities among transaction records consisting of several variables, association algorithms are used. These algorithms are used to solve problems where it is important to understand the extent to which the presence of some variables implies the existence of other variables and the prevalence of this particular pattern across all data records. Sequential-pattern-discovery algorithms are related to association algorithms except that the related items are spread over time. Finally, the predictive-modeling data-mining-problem approach involves the use of a number of algorithms (e.g., binary decision tree, linear discriminant function analysis, radial basis function, back-propagation neural network, logistic regression, and standard linear regression) to classify data into one of several predefined categorical classes or to use selected fields from historical data to predict target fields.

The initial implementation of data-mining applications has been in the banking, consumer marketing, insurance, and telecommunications industries. Credit scoring, direct-mail target marketing, policy-holder risk assessment, and call graph analysis are but a few of the "killer" applications of data mining in these respective industries. As a result of some of the realized benefits of data mining, new applications are emerging in a number of areas including biomedicine where molecular data are combined with clinical medical data to achieve a deeper understanding of the causes for and treatment of disease, national security where unusual patterns and fraudulent behavior play a role in identifying and tracking activities that undermine security, pharmaceuticals where interest in understanding the 3D substructure of a molecule and how it interacts with the target is a crucial step in the design of new drug molecules, and ecology where large amounts of climate data, terrestrial observations, and ecosystem models offer an unprecedented opportunity for predicting and possibly preventing future ecological problems. Although the frontiers of data-mining applications continue to expand, focus on developing a data-mining process has not met with similar enthusiasm.

DATA-MINING PROCESS OVERVIEW

New product development (NPD) is a well-researched area (e.g., Hauptman & Hirji, 1999) and, thus, it is the foundation for the data-mining process model because NPD projects, by their very nature, are the most complex as they include systems, subsystems, components, and modules, as well as physical product and software aspects. Focusing on the NPD literature and synthesizing the elements of the various process models allows for the development of an information-centric process model for performing data-mining projects. Table 1 provides a baseline of what an inclusive process for performing data-mining projects might look like.

The phases in the baseline data-mining process include Phase 0, Phase 1, Phase 2, and Phase 3. Phase 0 is the *discovery* phase that supports the subsequent three phases. The set of proposed activities in this phase include (a) assessment of the organization's orientation toward data-centricity, (b) assessment of the capability of the organization to apply a portfolio of analytical techniques, and (c) strategy development for the use of analytics throughout the department or organization. Phase 1 is the *entry* phase. The underlying intent of this phase is to define the candidate business problem that is solvable and that can at least partially use existing data resident in the organization's databases. Prospecting and domain analysis, business problem generation and preliminary assessment, and data sensing are the proposed set of activities in this phase. Data sensing in particular is concerned with the representational faithfulness of the data set in question. Phase 2 is the *launch* phase. In this phase the data-mining project becomes a formal project with an associated capital and operational budget. The set of proposed activities in this phase include (a) secure