DATA MINING: AN INDUSTRIAL RESEARCH PERSPECTIVE

Chidanand Apté

Just what exactly is data mining? At a broad level, it is the process by which accurate and previously unknown information is extracted from large volumes of data. This information should be in a form that can be understood, acted upon, and used for improving decision processes. Obviously, with this definition, data mining encompasses a broad set of technologies, including data warehousing, database management, data analysis algorithms, and visualization. The crux of the appeal for this new technology lies in the data analysis algorithms, since they provide automated mechanisms for sifting through data and extracting useful information.

The crux of this technology lies in the data analysis algorithms, which provide automated ways to sift through data and extract useful information.

The analysis capability of these algorithms, coupled with today’s data warehousing and database management technology, make corporate and industrial data mining possible.

The data representation model for such algorithms is quite straightforward. Data is considered to be a collection of records, where each record is a collection of fields. Using this tabular data model, data mining algorithms are designed to operate on the contents, under differing assumptions, and delivering results in differing formats.

The data analysis algorithms (or data mining algorithms, as they are more popularly known nowadays) can be divided into three major categories based on the nature of their information extraction: predictive modeling (also called classification or supervised learning), clustering (also called segmentation or unsupervised learning), and frequent pattern extraction.

Predictive modeling

Predictive modeling is based on techniques used for classification and regression modeling. One field in the tabular data set is pre-identified as the response or class variable. These algorithms produce a model for that variable as a function of the other fields in the data set, pre-identified as the features or explanatory variables. If the response variable is discrete-valued, classification modeling is employed; if the response variable is continuous-valued, regression modeling is employed. The principal problem being addressed by this family of algorithms is to produce a predictively accurate function approximation for the response variable by using the data set as examples of the relations between instances of explanatory variables and the response variable, in the presence of noise. Once produced, the model can be used to predict the value of a response variable, given the specifications for the explanatory variables.

This modeling work has its roots in classical statistics, although many recent advances have come from other areas, including pattern recognition, information theory, and machine learning. The important shift in the modeling paradigm that has taken place here is the shift toward nonparametric techniques such as neural networks, decision trees, and decision rules, where no assumptions are made about any underlying distributions in the data. The two key technical aspects to all predictive modeling algorithms is their ability to generate models in the presence of noise in the data, and their emphasis on producing an accurate error estimate on the model that is produced. Many techniques have been developed for noise handling and error estimation, and provide the foundational basis for most modern predictive modeling methods.

A further level of interest has been created by data mining applications based on decision-tree and rule modeling, due to the nature of their modeling representation. Both decision-tree- and rule-modeling algorithms produce outputs in symbolic notation, which is very amenable to inspection and interpretation. This characteristic allows business end users and analysts to understand the underlying decision boundaries in data, and take actions based upon them. Although alternate techniques such as neural networks may produce predictively accurate models, their outputs are highly quantitative in nature, and not easily understandable. However, when evaluating
and determining a modeling technique to use from a given set of alternatives, users have to weigh the compromise between predictive accuracy, level of understandability, and computational demand. Often these alternatives need to be traded off against one another, because algorithms often compromise one to gain performance in another.

Clustering

Clustering is another major class of data mining algorithms. Using the same tabular data model described earlier, these algorithms attempt to automatically partition the data space into a set of regions or clusters, to which each of the examples in the table are assigned, either deterministically or probabilistically. The goal of the search process used by these algorithms is to identify all sets of similar examples in the data, in some optimal fashion.

Obviously, the notion of similarity is highly subjective, and one of the more popular criteria that has been used for identifying similarity has been Euclidean distances (k-means, hierarchical), based upon early research in pattern recognition and, more recently, in Bayesian statistics (AutoClass) and neural networks (Self Organizing Map). In many cases, clustering results can be judged only by the value perceived by end users; no strict evaluation guidelines exist, as in predictive modeling.

While the early pattern recognition work produced clustering methods that were adequate for continuous-valued space, data mining requirements have generated interest in newer techniques that can operate on variables that need not be continuous valued. Of these, techniques based on relational voting are gaining in popularity. The idea here is to determine similarity between examples based on how many features the examples agree on, and not how close they are in Euclidean space. It is obvious that this approach will work well on data sets that have discrete-valued variables in them. However, real-world data has both continuous-valued as well as discrete data, and combinations of Euclidean techniques and relational techniques need to be used for business and industrial applications.

Frequent pattern extraction

The final class of data mining algorithms is frequent pattern extraction. The goal here is to extract from the tabular data model all combinations of variable instantiations that exist in the data with some predefined level of regularity. Typically the basic kind of pattern to be extracted is an association: a tuple of two sets, with a unidirectional causal implication between the two sets, \( A \rightarrow B \). Attached with this tuple are two statistical measures, confidence and support.

Confidence measures the fraction of times \( B \) exists in the data when \( A \) is present. Support measures the number of times \( A \) exists as a fraction of the total data. Thus, an association with a very high support and confidence is a pattern that occurs so often in the data that it should be obvious to the end user. Patterns with extremely low support and confidence should be regarded as outliers with no significance. It is patterns with combinations of intermediate values for confidence and support that provide the user with interesting and previously unknown information. Many variations of this basic association pattern have been formulated, with algorithms to extract them. Some patterns have time stamps attached; the temporal relations that are extracted may hold significance, in terms of either frequent occurrence or frequent matching between groups of temporal patterns.

And more...

While this characterization of data mining algorithms is certainly not all-inclusive, it does represent a major por-

Methods for data selection, cleaning, and transformation play a necessary and critical role.

...tion of the analysis. Many other types of data mining algorithms are also making their impact upon applications—deviation detection and attribute focusing being two of them. Many of these algorithms are statistical algorithms geared toward computing normative behavior of the tabular data model, and then using that to focus into either examples or variables that tend to deviate from the derived norm. This identification is provided to the end user in the form of detected exceptions, which could be valuable information to the user for taking corrective actions.

Finally, for all these data mining algorithms to work, methodologies for data selection, cleaning, and transformation play a necessary and critical role. For data selection, data needs to be extracted from different databases and joined, and perhaps sampled. Once selected, the data may need to be cleaned. If the data is not derived from a warehouse but from disparate databases, values may be represented using different notations in the different databases. Also, certain values may require special handling, since they may imply missing or unknown information. After
the selection and cleaning process, certain transformations may be necessary. These range from conversions from one type of data to another, to deriving new variables using mathematical or logical formulae. Many times, if a data warehouse does not already exist, this step of selection, cleaning, and transformation may take up to 80 percent of a data mining analysis job for a large business data set.

Once the mining is done, visualization plays an important role in providing adequate bandwidth between the results of the data mining and the end user. There are a few generic tools derived from statistics and computer-based 3D modeling that can help users examine the results of data mining in useful and informative ways. Figure 1 illustrates a “big picture” view of data mining.

**Key applications**

Using the techniques described earlier, a growing body of emerging applications is changing the landscape of business decision support. Predictive modeling techniques are best used when a large body of historical data is available. This data is used to model a variable of interest, so that this variable may be forecast in future scenarios, and effective actions taken based on that forecast. Some examples are as follows:

- **Risk analysis.** Given a set of current customers and an assessment of their risk worthiness, develop descriptions for these classes. Use these descriptions to classify a new customer into one of the risk categories.
- **Targeted marketing.** Given a database of potential customers and how they have responded to a solicitation, develop a model of customers most likely to respond positively. Use the model for more focused new-customer solicitation.
- **Customer retention.** Given a database of past customers and their behavior prior to attrition, develop a model of customers most likely to leave. Use the model for determining the best course of action for these customers.
- **Portfolio management.** Given a particular financial asset, predict the return on investment to determine whether to include the asset in a portfolio.
- **Brand loyalty.** Given a particular customer and a product she uses, predict whether the customer will switch brands.

Using frequent patterns extracted from data, one can build link analysis and item set analysis applications. These are used for determining business values of promotional effectiveness, analyzing subscriber services, forecasting demand, etc. One of the more powerful applications of this technique is market basket analysis, in which databases of sales transactions are examined to extract patterns that identify what items sell together, what items sell better when relocated to new areas, and what customer groupings improve department sales. Potential benefits include improved store layout and organization, better inventory management, more effective tie-in promotions, and eventually increased sales.

Clustering approaches are one of the more pervasive applications of data mining. As databases grow, it is often necessary to partition them into collections of related records to obtain better summaries of the subpopulations present in the data. These applications are most appropriate in marketing planning and promotions, where one wants to constantly monitor and identify the segments of the marketplace served. For example, a bank may want to segment all its retail customers to get a better feel for the demographic and psychographic breakdown (rather than use just a one-dimensional breakdown,
Vendors will offer their systems and algorithms as Internet servers; clients will mine their data on these servers via electronic fee-based access.

Chidamard Apte is manager of the Data Abstraction Research group at IBM’s T.J. Watson Research Center; apte@watson.ibm.com. His research focuses on classification systems, knowledge representation, knowledge discovery, and data mining. He is a senior member of the IEEE, and a member of the AAAI.

References