


BUSINESS APPLICATIONS OF DATA MINING

They help identify and predict individual, as well as aggregate, behavior, as illustrated by four application domains: direct mail, retail, automobile insurance, and health care.

The traditional approach to data analysis for decision support has been to couple domain expertise with statistical modeling techniques to develop handcrafted solutions for specific problems. More recently, several trends have emerged to challenge this approach. One is the increasing availability of large volumes of high-dimensional data occupying database tables with millions of rows and thousands of columns. Another is the competitive demand for the rapid construction and deployment of data-driven analytics. A third is the need to give end users analysis results in a form they readily understand and assimilate, helping them gain the insights they need to make critical business decisions. Moreover, knowledge discovery in databases (KDD) techniques emphasizing scalable, reliable, fully automated, explanatory structures have shown that in data analysis, such structures supplement, and sometimes supplant, existing human-expert-intensive analytical techniques for improving decision quality.

Measurable Benefits

KDD applications deliver measurable benefits, including reduced cost of doing business, improved profitability, and enhanced quality of service. Industries in which such benefits have been demonstrated include insurance, direct-mail marketing, telecommunications, retail, and health care.

Risk management and targeted marketing. Insurance and direct mail are two industries that rely on data analysis to make profitable business decisions. For example, insurers must be able to accurately assess the risks posed by their policyholders to set insurance premiums at competitive levels. For example, overcharging low-risk policyholders would motivate them to seek lower premiums elsewhere; undercharging high-risk policyholders would attract more of them due to the lower premiums. In either case, costs would increase and profits inevitably decrease. Effective data analysis leading to the creation of accurate predictive models is essential for addressing these issues.

In direct-mail targeted marketing, retailers must be able to identify subsets of the population likely to respond to promotions in order to offset mailing and printing costs. Profits are maximized by mailing only to those potential customers most likely to generate net income to a retailer in excess of the retailer's mailing and printing costs.

Businesses relying on data-driven analysis for decision making typically construct data warehouses to capture as much information as possible about their customers. Examples of such information include details of past customer transactions, as well as additional information obtained from third-party data providers, including credit scores and demographics, for targeted marketing purposes and motor vehicle

BY CHIDANAND APTE, BING LIU, EDWIN P.D. PEDNAULT, AND PADHRAIC SMYTH

While the
association-rule approach can be
useful for exploratory analysis of
transaction data, it is **LESS WELL-SUITED FOR
PREDICTING INDIVIDUAL
CUSTOMER BEHAVIOR.**

records for insurance purposes.

To aid decision making, analysts construct predictive models using warehouse data to predict the outcomes of a variety of decision alternatives. For example, in order to set policy premiums, insurers need to predict the cost of claims filed by policyholders annually, given what is known about each policyholder. In order to select customers for a targeted marketing campaign, retailers need to predict revenues or gross profits that would be generated for the customers receiving the mailings.

A popular approach to predictive modeling used by many data analysts and applied statisticians involves partitioning the data records for a population of customers (or other entities) into segments, then developing separate predictive models for each segment. Typically, data is partitioned through a combination of domain knowledge, simple heuristics, and clustering algorithms. Predictive models are constructed once segments are identified. The drawback is that this sequential approach ignores the strong influence segmentation exerts on the predictive accuracies of the models within each segment. Good segmentations tend to be obtained only through trial and error by varying the segmentation criteria.

A better approach is to simultaneously perform segmentation and predictive modeling within each segment, optimizing the segmentation so as to maximize the overall predictive accuracy of the resulting model. This approach is built into the IBM Probabilistic Estimation (ProbE) data mining server, making it possible to automatically construct high-quality segmentation-based predictive models from very large high-dimensional data sets. A top-down tree-based algorithm is used to construct the segmentations. A collection of other algorithms is incorporated for constructing segment models, including stepwise linear regression and stepwise naive Bayes algorithms for general-purpose modeling and a joint Poisson/log-normal algorithm for insurance risk modeling. A key feature of the ProbE server is it is readily extended to incorporate different types of predictive modeling algorithms for the segments,

as well as different types of segmentation algorithms.

Two different client applications have been developed by IBM's Data Abstraction Research Group that utilize the ProbE data mining server. One is called IBM Advanced Targeted Marketing for Single Events (ATM-SE), built jointly with the Business Intelligence group at Fingerhut, Inc., a large U.S. catalog and Internet retailer based in Minnetonka, MN, for constructing customer-profitability and response-likelihood models for targeted marketing in the retail industry [1]. The other is the IBM Underwriting Profitability Analysis (UPA) application, co-developed with Farmers Insurance Group, a large automobile and home insurance company based in Los Angeles, for discovering homogeneous insurance risk groups [2].

Fingerhut's 2000 evaluation of the ATM-SE application for direct-mail response modeling demonstrated the application produced segmentation-based response models that either equaled or slightly outperformed Fingerhut's own proprietary models. This evaluation was significant because numerous vendors and consultants had previously failed to beat Fingerhut's in-house modeling capability. If these results ultimately hold across all of Fingerhut's models, the ATM-SE models would yield an estimated increase in annual profits directly to Fingerhut of more than \$1 million. Moreover, the ProbE server achieved its result in a fully automated mode of operation, with no manual intervention.

The UPA application configures the ProbE server so as to use a joint Poisson/log-normal statistical model within each segment to simultaneously model both the frequency with which insurance claims are filed by policyholders and the amounts, or severities, of these claims for each segment. Using this class of segment model, the identified segments correspond to distinct risk groups whose loss characteristics, such as claim frequency and severity, are estimated in accordance with standard actuarial practices.

The Farmers Group's 1997 evaluation of the application's ability to analyze insurance policy and claims data for all policyholders in one state involved mining runs for 18 unique combinations of cus-

tomers with specific insurance products and coverage, including explanatory variables. Each run generated about 40 rules, from which 43 combinations were identified as “nuggets,” or previously unknown rules with significant potential value. Six nuggets were selected for a detailed benefits assessment, which indicated that implementing just these six in a single state could potentially yield a net profit gain of several million dollars in the first year alone.

Although insurers know that drivers of high-performance sports cars are more likely to have accidents than drivers of other types of cars, the UPA found that if a sports car was not the only vehicle in the household, the accident rate is not much greater than that of a regular car. One estimate determined that “just letting Corvettes and Porsches into [the insurer’s] ‘preferred premium’ plan could bring in an additional \$4.5 million in premium revenue over the next two years without a significant rise in claims.” Another publicly disclosed nugget related to experienced drivers, who tend to have relatively low claim frequencies. However, the UPA also turned up one particular segment of experienced drivers who are unusually accident prone.

ProbE’s segmentation-based predictive modeling capability permits construction of mining applications optimized for specific problems. Indications are that the ProbE server can consistently produce high-quality models on a fully automated basis without requiring costly manual adjustments of the models or the mining parameters by data mining experts. These characteristics will make data mining increasingly attractive to mid-size businesses, as well as to their much larger counterparts.

Customer profiles and feature construction. An important ingredient for obtaining highly predictive models is to use highly predictive features, or attributes and variables, as model input. Although a database might contain sufficient information to construct highly predictive models, it is not always stored in a form that permits the data to be used directly as input to a model. In such cases, the data must be transformed to obtain accurate models.

Transaction data is notorious for requiring transformation before it can be used for data mining applications. Such data consists of records of pairs of individuals and events. An example is a set of retail items purchased by a customer and grouped into a “market basket.” Another is a set of Web pages requested from a Web site by a particular surfer and grouped by session. The ability of companies worldwide to collect vast amounts of such transaction data has far outpaced their ability to analyze it. Transaction data is especially challenging from a data mining

perspective due to several factors:

Massive numbers of records. Large retail chains generate millions of transactions per day.

Sparseness. A typical basket contains only a small fraction of the total possible number of items; individual customers may have few baskets, perhaps only one.

Heterogeneity. Purchasing behavior varies considerably, depending on individual tastes and means, along with individual purchasing patterns over time.

These factors combine to make transaction data highly nontrivial when using traditional data analysis techniques. The related challenges, along with the transaction data itself, motivated much of the early work in data mining, including development of association-rule algorithms for efficiently searching for correlations among items in retail transaction data. While the association-rule approach can be useful for exploratory analysis of transaction data, such as discovering combinations of products purchased together, it is less well-suited for predicting individual customer behavior.

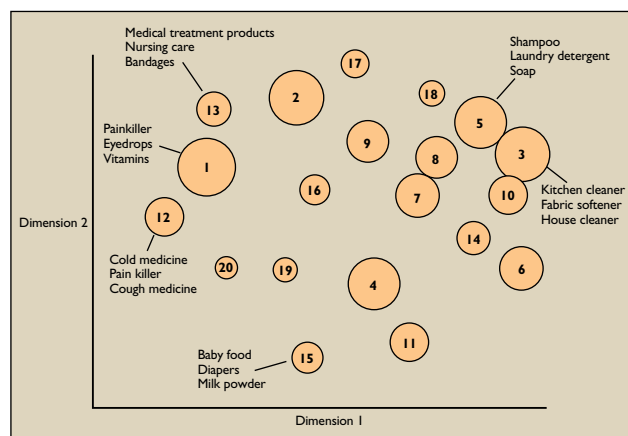
A recently developed framework called predictive profiling handles transaction data in predictive modeling [3]. A predictive profile is a model that predicts future purchasing behavior of an individual customer, given historical transaction data for both the individual and for the larger population of all of a particular company’s customers. The predictive profiling approach is based on a flexible probabilistic model that works in the following way: Let y be a randomly chosen market basket, where y is a d -dimensional vector describing how many of each of the d items were purchased in the basket. The high-dimensional joint distribution on baskets, $p(y)$, is approximated by a linear combination of K simpler models. Each of the K simpler models in effect captures “prototype combinations” of products in baskets.

In the first phase of modeling, the K prototype combinations are learned from the data through a well-known expectation-maximization procedure for statistical estimation. In the second phase, each customer is “mapped” onto the product space represented by the K prototypes, where the mapping is based on individual past purchasing patterns. The mapping effectively transforms the transaction data for each customer into a set of feature values that are then used to make predictions about future purchasing behavior. The transformation is not defined prior to data mining but is inferred by the mining algorithm.

These methods
for handling
transaction data are as likely to prove useful across a variety of
business applications, including forecasting, where **REAL-TIME**
modeling of individual customer and personalized
feedback is valuable.

This model is not designed to capture all aspects of individual customer behavior but to extract useful “first-order” characteristics in terms of how customers shop. The figure outlines how the prototypes are used to support exploratory visualization of the data, providing an interpretable description of the heterogeneity of customer behavior as reflected by different basket prototypes.

The predictive profiling method was tested by a University of California, Irvine, research team on two large real-world transaction data sets collected



A set of $K = 20$ prototypes represented here in a 2D space by using multidimensional scaling. The prototype baskets were learned from a set of about six million baskets from a chain of drugstores in Japan. The numbers in each circle refer to different prototypes; the area of each circle represents how likely a randomly chosen basket belongs to that prototype. The names of the three items with the greatest lift (defined as $p(\text{item} \setminus \text{prototype}) / p(\text{item})$) are also displayed for some of the prototypes. Prototypes close together are also statistically close in the data.

over several years. The data sets involved several million baskets and about 500,000 customers. Models were trained using historical data from the early years of each data set, then tested on data from later years, typically using from $K = 20$ to $K = 100$ prototypes. The models demonstrated systematic improvement in out-of-sample predictive performance compared

to more standard alternatives. Empirically, the time taken to fit the models was found to scale linearly with both number of baskets and number of fitted prototypes K . The wall-clock time to learn all the prototypes and customer profiles took only a few hours on a standard PC.

Such methods for handling transaction data are likely to prove useful across a variety of business applications, including customer segmentation, personalization, forecasting, and change detection, especially in e-commerce environments, where real-time modeling of an individual customer and personalized feedback is valuable. Scalable, robust, and accurate solutions to these problems promise significant economic payoff in the business world.

Medical applications (diabetic screening). Preprocessing and postprocessing steps are often the most critical elements determining the effectiveness of real-life data-mining applications, as illustrated by the following recent medical application in diabetic patient screening. In the 1990s in Singapore, about 10% of the population was diabetic, a disease with many side effects, including increased risk of eye disease, kidney failure, and other complications. However, early detection and proper care management can make a difference in the health and longevity of individual sufferers. For example, to combat the disease, the government of Singapore introduced a regular screening program for diabetic patients in its public hospitals in 1992. Patient information, clinical symptoms, eye-disease diagnosis, treatments, and other details, were captured in a database maintained by government medical authorities. Today, after almost 10 years of collecting data, a wealth of medical information is available. This vast store of historical data leads naturally to the application of data mining techniques to discover interesting patterns. The objective is to find rules physicians can use to understand more about diabetes and how it might be associated with different segments of the population [4].

However, the data miners encountered two major problems. First, the data captured by health clinics turned out to be very noisy; for example,

many patient records in the database contained typographical errors, missing values, and incorrect information, including street names and date of birth. Worse, many records contained duplicate data. Cleaning data takes a great deal of effort and time. In addition, many of the records were not in a form suitable for data mining; they had to be transformed to more meaningful attributes before mining could proceed. The second problem was that some state-of-the-art association-rule algorithms generate too many rules from the data, no matter how clean it is. Because physicians are busy seeing patients, they cannot take the time to sift through large numbers of rules. It was therefore important to present the discovered rules in some easy-to-understand form.

To overcome the problem of noisy data, a data mining team at the National University of Singapore developed a semiautomatic data-cleaning system to reconcile database format differences by allowing physicians to specify the mapping between attributes in different format styles and/or different encoding schemes. Reconciling the format differences addressed the problem of identifying and removing duplicate records.

To resolve the problem of too many rules generated by the mining algorithms, the same team developed a user-oriented approach providing step-by-step exploration of both the data and the discovered patterns. Data visualization is used as an integral part of the process to give users a general view of the findings. During rule mining, the mining algorithm employs a pruning method to remove insignificant rules. The final rules were also organized into general rules and exceptions to facilitate browsing and analysis [5]. This rule-mining approach to organizing mining results is useful to Singapore's medical authorities because it allows them to view the general patterns that are discovered, as well as the detailed patterns. Because it is also a common strategy people employ in everyday learning, the mining results are easy to interpret.

The physicians confirmed that many of the rules and causal relationships the data mining algorithms discovered conformed to the trends they observed in their practices. However, they were surprised by many of the exceptions they did not know before. As a result of data mining, they gained a much better understanding of how diabetes progresses over time and how various treatments affect its progress.

Conclusion

Data mining applications have proved highly effective in addressing many important business prob-

lems. We expect to see the continued construction and deployment of KDD applications for crucial business decision support systems. Exemplary applications employing data mining analytical techniques will require the KDD technical community to keep improving the underlying techniques for model building and model understanding. The emphasis in model building will be on developing mining techniques that are automated, scalable, and reliable. For domain understanding, the challenge is to keep developing sophisticated techniques that assist users in analyzing discovered knowledge easily and quickly. ■

REFERENCES

1. Apte, C., Bibelnicks, E., Natarajan, R., Pednault, E., Tipu, F., Campbell, D., and Nelson, B. Segmentation-based modeling for advanced targeted marketing. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)* (San Francisco, Aug. 26–29). ACM Press, New York, 2001, 408–413.
2. Apte, E., Grossman, E., Pednault, E., Rosen, B., Tipu, F., and White, B. Probabilistic estimation-based data mining for discovering insurance risks. *IEEE Intell. Syst.* 14, 6 (Nov./Dec. 1999), 49–58.
3. Cadez, I., Smyth, P., and Mannila, H. Probabilistic modeling of transaction data with applications to profiling, visualization, and prediction. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)* (San Francisco, Aug. 26–29). ACM Press, New York, 2001, 37–46.
4. Hsu, W., Lee, M., Liu, B., and Ling, T. Exploration mining in diabetic patient databases: Findings and conclusions. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000)* (Boston, Aug. 20–23). ACM Press, New York, 2000, 430–436.
5. Liu, B., Hu, M., and Hsu, W. Multi-level organization and summarization of the discovered rules. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000)* (Boston, Aug. 20–23). ACM Press, New York, 2000, 208–217.

CHIDANAND APTE (apte@us.ibm.com) is Manager of the Data Abstraction Research Group at the IBM T.J. Watson Research Center, Yorktown Heights, NY.

BING LIU (liub@cs.uic.edu) is an associate professor in the Department of Computer Science at the University of Illinois at Chicago.

EDWIN PEDNAULT (pednault@watson.ibm.com) is a research staff member in the Data Abstraction Research Group at the IBM T.J. Watson Research Center, Yorktown, NY.

PADHRAIC SMYTH (smyth@ics.uci.edu) is an associate professor in the Information and Computer Science Department at the University of California, Irvine.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.