Business Applications of Data Mining

Chidanand Apte, Bing Liu, Edwin P.D. Pednault, Padhraic Smyth

May 22, 2002

The traditional approach to data analysis for decision making has been to couple business and scientific expertise with statistical modeling techniques in order to develop hand-crafted solutions for specific problems. In recent years, a number of trends have emerged that have started to challenge this traditional approach. One trend is the increasing availability of large volumes of high-dimensional data, occupying database tables with many millions of rows and many thousands of columns. Another trend is the increasing competitive demand for rapidly building and deploying data-driven analytics. A third trend is the increasing need to present analysis results to end-users in a form that can be readily understood and assimilated so that end-users can gain the insights they need to improve the decisions they make.

KDD techniques that emphasize scalable, reliable, fully-automated, and explanatory structures are demonstrating that such techniques can step up to the data analysis challenge. They are supplementing, and sometimes supplanting, existing human-expert-intensive analytical techniques for achieving significant improvements in the quality of decisions.

1 Examples of KDD Applications

Exemplary KDD applications are ones that deliver measurable benefits, such as reduced costs of doing business, improved profitability, or enhanced quality of service. Areas in which such benefits have been demonstrated include insurance, direct mail marketing, telecommunications, retail, and medicine. Selected examples of applications in these areas are presented in this section. However, neither the examples nor the application areas are meant to be exhaustive. They are merely illustrative of the tremendous potential of KDD technology.

1.1 Risk Management and Targeted Marketing

Insurance and direct-mail retail are examples of businesses that rely on effective data analysis in order to make profitable business decisions. For example, insurers must be able to accurately assess the risks posed by policyholders in order to set insurance premiums at competitive levels. Overcharging low-risk policyholders would motivate such policyholders to seek lower premiums elsewhere. Undercharging high-risk policyholders would attract more high-risk policyholders because of the lower premiums. In both cases, costs would increase and profits would decrease. Effective data analysis leading to the creation of accurate predictive models is essential in order to address these issues.

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

Businesses that rely on data-driven analysis for decision making typically construct data warehouses of one form or another in order to capture as much information as they can about their customers. Examples of such information include details on all past customer transactions, as well as additional information obtained from third-party data providers, such as credit scores and demographic information for targeted marketing purposes and motor vehicle records for insurance purposes.

To aid decision making, predictive models are constructed using warehouse data in order to predict the outcomes of different decision alternatives. For example, in order to set policy premiums, insurers need to predict the expect cost of claims filed by policyholders per year given what is known about each policyholder. In order to select customers for a targeted marketing campaign, retailers need to predict the expected revenues or gross profits that would be generated for the customers that are mailed.

A popular approach to predictive modeling that is used by many data analysts and applied statisticians involves partitioning the data records for a population of customers (or other entities) into segments and developing separate predictive models for each segment. Typically, data is partitioned using a combination of domain knowledge, simple heuristics, and/or the use of clustering algorithms. Predictive models are then constructed once segments have been identified. The drawback of this sequential approach is that it ignores the strong influence that 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 perform segmentation and predictive modeling within each segment simultaneously, and to optimize the segmentation so as to maximize the overall predictive accuracy of the resulting model. The benefit of this optimizing approach is that it can produce better models than might otherwise be obtained.

This latter approach is the one taken in the IBM Probe (Probabilistic Estimation) data mining server. The ProbE server provides capabilities to automatically construct high-quality segmentation-based predictive models from very large high-dimensional data sets. A top-down tree-based algorithm is currently used to construct segmentations. A collection of other algorithms is also incorporated for constructing segment models. The latter includes 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 that it can be 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 that utilize the ProbE data mining server. One is the IBM ATM-SE (Advanced Targeted Marketing for Single Events) application built jointly with the Business Intelligence group at Fingerhut Inc. for constructing customer-profitability and response-likelihood models for targeted marketing in the retail industry [1]. Another ProbE client is the IBM UPA (Underwriting Profitability Analysis) application, which was co-developed with Farmers Group for discovering homogeneous insurance risk groups [2].

Fingerhut's evaluation of the ATM-SE application demonstrated that the application produced segmentation-based response models that either equaled or slightly outperformed Fingerhut's own proprietary models. The outcome of this evaluation is significant because numerous vendors and consultants have been unsuccessful in beating Fingerhut's in-house modeling capability. If these results hold across all of Fingerhut's models, the ATM-SE models would yield an estimated increase in yearly profits of over one million dollars. Moreover, the ProbE data mining server achieved this 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 in order to simultaneously model both the frequency with which insurance claims are filed by policyholders and the amounts (i.e., severities) of those claims for each segment. Using this class of segment models, the segments that are identified then correspond to distinct risk groups whose loss characteristics (i.e., claim frequency and severity) are estimated in accordance with standard actuarial practices.

Farmers Group evaluated the ability of the UPA application to analyze on insurance policy and claims data for all policyholders in a particular state. Mining runs were conducted for 18 unique combinations of books of business, explanatory variables, and coverages. Each run generated about 40 rules. From this collection of rules, 43 were identified as "nuggets", i.e. previously unknown rules with sizable potential impact. Six of these nuggets were selected by the insurer for a detailed benefits assessment. The benefits assessment study indicated that implementing just these 6 nuggets in a single state could potentially realize a net profit gain of several million dollars.

One of the six nuggets has already been widely publicized in the media. While it is well-known among insurers that drivers of high-performance sports cars are more likely to have accidents than are other motorists, the UPA discovered that if the sports car was not the only vehicle in the household, then the accident rate is not much greater than that of a regular car. In one estimate, "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 relates to experienced drivers, who tend to have relatively low claim frequencies. However, the UPA turned up a particular segment of experienced drivers who are unusually accident prone.

ProbE's segmentation-based predictive modeling capability permits the construction of mining applications that are optimized to specific problems. Indications are that the ProbE server is capable of consistently producing 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 be key enablers in making data mining attractive to medium-sized businesses.

1.2 Customer Profiles and Feature Construction

An important ingredient needed to obtain highly predictive models is to have highly predictive features (i.e., attributes, variables, etc.) that can be used as inputs to the models. 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 in order to obtain accurate models.

Transaction data is notorious for requiring transformation before it can be used. Such data consists of records of pairs of individuals and events. An example is sets of retail items purchased by customers and grouped into market baskets. Another example is sets of Web pages requested from a Web-site by a particular surfer and grouped by session. Our ability to collect such transaction data has far outpaced our ability to analyze it in recent years.

Transaction data is quite challenging from a data mining perspective due to a number of factors: (a) Massive numbers of records; large retail chains can generate millions of transactions per day. (b) Sparseness; a typical basket only contains a small fraction of the total possible number of items, and for individual customers we may have very few baskets, perhaps just one. (c) Heterogeneity; there is considerable variability in purchasing behavior across different individuals, as well as in the purchasing patterns of the same individual over time.

All of the above characteristics combine to make transaction data highly non-trivial to deal with using traditional data analysis techniques. In fact these challenges, and transaction data in particular, motivated much of the early work in data mining, such as the development of association rule algorithms to efficiently search 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 that are often purchased together, it is not as well-suited for the problem of *predicting* customer behavior at the individual level.

In recent work [3], a framework called *predictive profiling* has been developed for directly handling transaction data in predictive modeling. A predictive profile is a model that predicts future purchasing behavior of a customer given historical transaction data both for that individual customer and for the larger population of all customers.

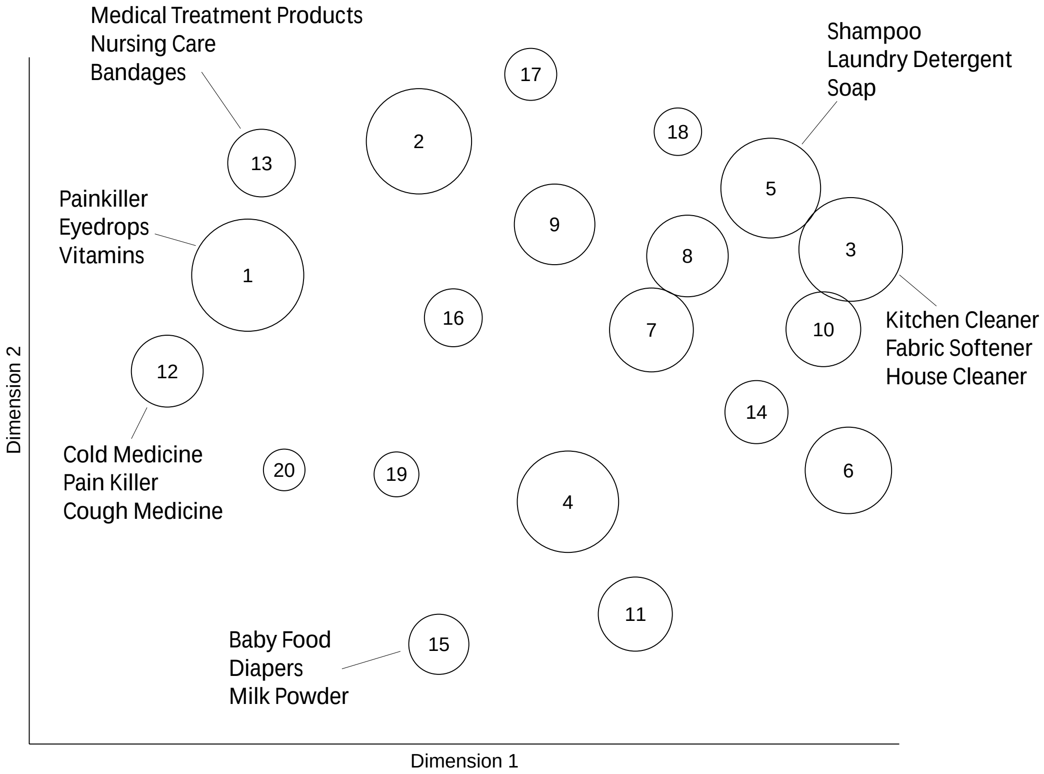

Figure 1: A set of $K = 20$ prototypes are represented here in a two-dimensional space by using multidimensional scaling. The prototype baskets were learned from a set of about 6 million baskets from a chain of popular drugstores in Japan. The numbers in each circle refer to different prototypes and the area of each circle represents how likely a randomly chosen basket is to belong to that prototype. The names of the 3 items with the highest lift (defined as $p(item|prototype)/p(item)$) are also displayed for some of the prototypes. Prototypes that are close together in this 2-dimensional space are also statistically close in the data.

The approach is based on a flexible probabilistic model that works as follows. Let $\mathbf{y}$ be a randomly chosen market basket, where $\mathbf{y}$ 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, $p(\mathbf{y})$, 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 using a well-known expectation maximization procedure for statistical estimation. In the second phase of modeling, each customer is then "mapped" onto the product space represented by the K prototypes, where the mapping is based on their 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. The transformation is not defined prior to data mining, but instead is inferred by the mining algorithm.

This type of model is not intended to capture all aspects of customer behavior in detail, but instead is designed to extract useful "first-order" characteristics of customers in terms of how they shop. Figure 1 illustrates how the prototypes can be 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 method was tested on two large real-world transaction data sets collected over several years. The data sets involve several million baskets and about 500,000 customers. Models were trained using historical data from the first few years of each data set and then tested on data from later time-periods, typically using from $K = 20$ to $K = 100$ prototypes. The models demonstrated systematic improvements in out-of-sample predictive performance compared to more standard alternatives. The time taken to fit the models was found empirically to scale linearly with both the number of baskets and the number of fitted prototypes K. The wall-clock time to learn all of the prototypes and customer profiles took only a few hours on a standard PC.

Methods such as this for handling transaction data are likely to prove increasingly useful across a variety of business applications, including customer segmentation, personalization, forecasting, and change detection. This should particularly be true in e-commerce environments, where real-time modeling of a customer and personalized feedback can be very valuable. Providing scalable, robust, and accurate solutions to these problems has the potential to have significant economic impact in the business world.

1.3 Medical Applications: Diabetic Screening

Pre-processing and post-processing steps can often be the most critical elements in determining the success of real-life data-mining applications. This fact was particularly evident in a recent medical application for diabetic patient screening.

In Singapore, about 10 percent of the population is diabetic. This disease has many side effects such as higher risk of eye disease, higher risk of kidney failure, and other complications. However, early detection of the disease and proper care management can make a difference.

To combat this disease, Singapore introduced a regular screening program for the diabetic patients in 1992. Patient information, clinical symptoms, eye-disease diagnosis and treatments, etc., are captured in a database. After eight years of data collection, a whole wealth of information has been gathered.

The availability of historical data lead naturally to the application of data mining techniques to discover interesting patterns. The objective was to find rules that can be used by medical doctors to improve their daily work; that is, to understand more about the diabetic disease and/or to find out how the disease might be associated with different segments of the population [4].

Two major problems were encountered in this application. First, the data captured by health clinics are very noisy. Many of the patient records in the database contain typographical errors, missing values, or incorrect information, such as street names, date of birth, etc. Worse still, many records are in fact duplicate records. Cleaning these data takes tremendous amount of effort and time. In addition, many of the data records collected are not in the forms that are suitable for data mining. They need to be transformed to more meaningful attributes before mining can proceed.

The second major problem encountered is that some state-of-the-art association rule algorithms generate too many rules from the data. Because health doctors are too busy seeing patients each day, they cannot afford the time or the energy to sieve through large numbers of rules. It therefore became important to present discovered rules in some easy-to-understand fashion.

To overcome the problem of noisy data, a semi-automatic data cleaning system was developed. The system reconciles database format differences by allowing doctors to specify the mapping between attributes in different format styles and/or different encoding schemes. Once the format differences were reconciled, the problem of identifying and removing duplicate records was addressed.

To resolve the problem of too many rules generated by the mining algorithms, a user-oriented approach was applied that provides step-by-step exploration of the data and the discovered patterns. Data visualization is used as an integral part of the process to provide users a general picture of the findings.

During rule mining, an effective pruning method was employed to remove insignificant rules. The final rules were also organized into general rules and exceptions in order to facilitate browsing and analysis [5].

This approach to organizing mining results is useful because it allows users to view first the general patterns that are discovered followed by the detailed patterns. Because this approach is also a common strategy that people employ in everyday learning, the mining results are then much easier to interpret.

Rules were also identified that represent possible causal relationships. The doctors involved with the project confirmed that many of the rules and causal relationships discovered conform to the trends that they had observed in their practices. They were also surprised by many of the exceptions that they did not know before. As a result of data mining, the doctors gained a much better understanding of how diabetes progresses over time and how different treatments affect its progress.

2 Conclusion

Data mining applications have been shown to be highly effective in addressing many important business problems. We expect to see a continuing trend in the building and deployment of KDD applications for crucial business and scientific decision support systems. Exemplary applications that employ data mining will require the KDD technical community to continue making advances in techniques for model building and model understanding. The emphasis in model building will be on developing mining techniques that are highly automated, scalable, and reliable. For domain understanding, the challenge is to continue developing more sophisticated techniques that can assist users in analyzing discovered knowledge easily and quickly. The next article elaborates on some of the challenges that will need to be addressed to enable a whole new set of exemplary applications.

References

[1] C. Apte, E. Bibelnieks, R. Natarajan, E.P.D. Pednault, F. Tipu, D. Campbell, and B. Nelson. Segmentation-Based Modeling for Advanced Targeted Marketing. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*, pages 408–413, August 2001.

[2] C. Apte, E. Grossman, E.P.D. Pednault, B. Rosen, F. Tipu, and B. White. Probabilistic Estimation Based Data Mining for Discovering Insurance Risks. *IEEE Intelligent Systems*, 14(6):49–58, November/December 1999.

[3] I.V. Cadez, P. Smyth, and H. Mannila. Probabilistic Modeling of Transaction Data with Applications to Profiling, Visualization, and Prediction. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*, pages 37–46. ACM Press, New York, NY, 2001.

[4] W. Hsu, M.L. Lee, B. Liu and T.W. Ling. Exploration mining in diabetic patients databases: findings and conclusions. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000)*, pages 430–436. ACM Press, New York, NY, 2000.

[5] B. Liu, M. Hu, and W. Hsu. Multi-level organization and summarization of the discovered rules. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000)*, pages 208–217. ACM Press, New York, NY, 2000.