

Mining Large Databases – A Case Study

Herb Edelstein

Two Crows Corporation

Remember when a megabyte was a lot of data? And 100 megabytes was an enormous database?

Today however, companies are dealing with databases of multiple terabytes – over 1,000,000,000,000 bytes of data. Getting useful information from this much data is a challenge. When there are millions of trees, how can one draw meaningful conclusions about the forest?

This is exactly the situation facing a very large provider of consumer information in the United States. Among their many databases is a credit information database containing information on 190 million consumers. The database is used to offer a wide variety of products to financial institutions who want information that will help them identify new customers, meet the needs of their existing customers, allocate their resources most efficiently, and minimize expenses.

Two key challenges this data provider faces are using their data to build new products and to enhance existing products. Consultants from IBM's Global Business Intelligence Consulting Practice and computer scientists from IBM Research worked with the client to attack these problems using IBM database management and data mining technology and the resources of the IBM RS/6000 Teraplex Integration Center in Poughkeepsie, New York. The IBM Teraplex Integration Centers use three large parallel servers – RS/6000 SP, S/390, and AS/400 – capable of handling databases in excess of a terabyte. The Centers provide a unique test bed for companies to explore solutions to huge business intelligence problems that can only be addressed with parallel DBMSs such as DB2 and parallel data mining tools such as Intelligent Miner.

The results they achieved were truly impressive, especially in light of the difficulty of mining such an enormous database. For example, they found an untapped market of good credit risks hidden among households that traditional methods had identified as bad credit risks. They could also more precisely identify potential bankruptcies, which not only allows lenders to avoid losses, but more importantly, helps consumers by offering them financial planning services and not giving them more credit than they can afford.

It was the combination of high performance parallel tools (DB2 and Intelligent Miner) on a large parallel computer (RS/6000 SP) coupled with insightful analyses that enabled them to effectively mine this data. How IBM achieved these results is the focus of this white paper. It will describe what IBM accomplished, while showing you how data mining in conjunction with parallel computers and DBMSs can be used to achieve highly valuable business results.

Data Mining

Within the masses of information in the consumer information database lies hidden information of strategic importance. Data mining is a key element in finding the particular patterns and relationships that can help their business.

Data mining finds these patterns and relationships using sophisticated data analysis tools and techniques to build models. Models, like maps, are abstract representations of reality. While you should never confuse your model with reality, a good model is a useful guide to understanding your business and suggests actions you can take to help you succeed.

There are two main kinds of models in data mining. The first kind, predictive models, use data with known results to develop a model that can be used to explicitly predict values. For example, based on the customers who have responded to an offer, the model will predict prospects likely to respond to the same offer. The second kind, descriptive models, describe patterns in *existing* data which may be used to guide decisions as opposed to making explicit predictions. For example, the model might identify different customer segments in a database. IBM used both kinds of models in mining the data.

The Business Problem

The target database contains detailed credit information on about 200 million households. This data includes information about what type of credit they have, such as mortgages, bank cards, or auto loans. It also includes transaction histories of their payments. The total volume of this data is about 4 terabytes.

IBM was presented with two important problems to solve.

The first problem was to identify new products that would help credit-granting organizations, such as credit card companies and banks, find good candidates for different credit products. Over two billion offers for credit cards are mailed each year. By precisely identifying good candidates, financial institutions can focus their marketing efforts, which not only reduces their costs, but also keeps them from bothering consumers who are not interested in these new credit instruments.

Second, could IBM improve the accuracy of predicting who would declare bankruptcy? The number of bankruptcies has grown from under one million in 1995 to 1.4 million in 1997. The estimates of the losses to credit card companies range from \$5 billion to \$40 billion. And the credit industry is also criticized for offering credit to people who may not be able to make appropriate use of it, thus contributing to the rise of bankruptcies. By more precisely targeting credit offers to consumers who can benefit from them, credit-granting organizations save billions of dollars a year, while helping consumers in the process.

IBM's client was already quite good at bankruptcy predictions. But they wanted to do an even better job because of the enormous sums at stake. Could IBM's technology make a difference?

The Data

IBM was given a random sample of the credit data, including transaction data for an 18-month period. This is a massive amount of data, with a total of over 900 million records taking up 360 gigabytes of storage, but still just a fraction of the total data. Random sampling can often be used to reduce the size of a problem, but when you have so much raw data, even the sample is enormous. Most data mining problems and traditional statistical analysis rarely exceed a few tens of gigabytes.

Both the amount of data and the degree of necessary manipulation demanded a large computer and a fast database management system. The original data was converted from a proprietary file structure containing 30 tables into a DB2 UDB Version 5 parallel database on an RS/6000 SP parallel computer with four SMP (symmetric multiprocessing) nodes, each consisting of eight CPUs with two gigabytes of memory per node (Figure 1). The total storage was about one terabyte, including 240 gigabytes of data, indexes, mirror tables (to ensure data availability), and the transformed data tables.

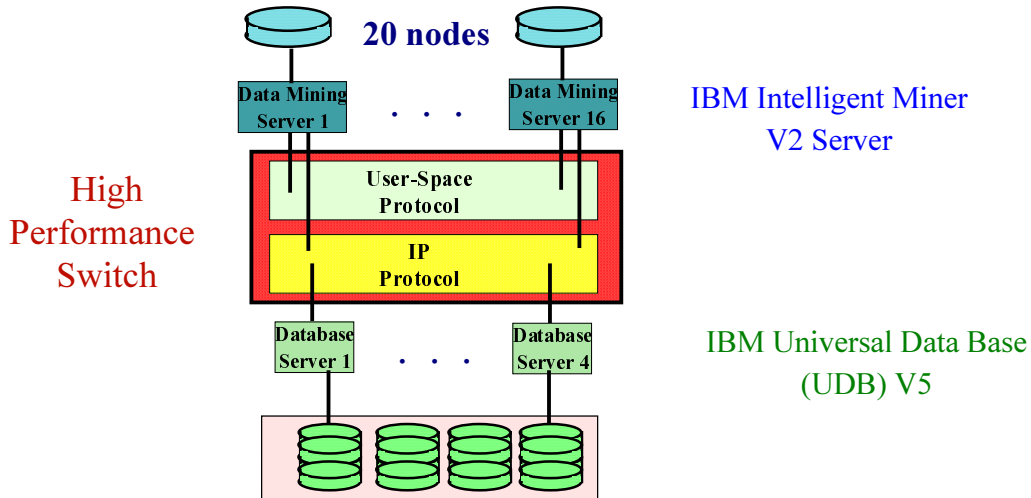


Figure 1. Hardware configuration for data manipulation and data mining

In addition to the original data, IBM added new variables based on transformations of the supplied variables. From the enriched data IBM needed to choose an appropriate subset of variables from which to build the models. The data transformations and selection of variables are the keys to success in data mining. The most successful models are those built with the “right” data.

The process of creating and selecting data from the raw data is an iterative process that may be quite time-consuming. Analysts need to add new variables, build some models and then go back and work with the data some more.

Even with a four node, 32-CPU RS/6000 SP, some of the transformation queries were so complex they took quite a while to complete. For example, there were SQL queries which required joining 12 to 15 tables including several 400-million row tables. Only in a parallel environment such as parallel DB2 and RS/6000 SP (the environment in the Teraplex Integration Center) could these data transformations have been successfully carried out. Serial processing, even on a large computer, would have extended the time necessary to transform the data. Since the RS6000/SP scales close to linearly, a query that took many hours to complete on it would take months to complete on a serial computer equal in power to one CPU of the RS/6000 SP. Clearly, this would have greatly reduced the amount of experimentation and analysis the data miners could perform in the time available to them, and made it impossible to duplicate the high quality results they achieved using parallel computing.

The first major transformation was to aggregate the data about individuals into data about the household – a process called “householding.” Rather than look at how individuals behave with credit, it is important to look at how the family or household deals with credit. Figuring out how to gather individual records into a household is an interesting problem. For example, a family may have multiple credit cards and different individuals will use each of them at different times. Mortgages and auto loans are part of a household’s debt rather than belonging to an individual, and if two names are on the mortgage, the debt would be attributed to each of them. Consequently, the 18 months of consumer transaction data were transformed into a set of household tables while eliminating duplicate debts and performing other data cleansing actions. The result was 75 gigabytes of data on over 1 million households.

The next step was to create the mining databases for bankruptcy prediction and new product identification. Because two different problems were being addressed, different mining databases needed to be created from the original customer data. The input table for bankruptcy prediction consisted of 280 columns for each of the million households, totaling about one gigabyte, while the new products model input table contained 482 columns totaling about four gigabytes.

At this point new variables were created from the existing data and added to it to see if they would be better predictors or descriptors than the raw data by itself. The best variables were then selected to build the models. IBM created and carefully examined over 250 variables, including how many accounts a household opened in the previous year, how many times a household was late in its payment, how much it spent per month, etc. The result was that 36 input fields were selected for bankruptcy prediction, while 41 input fields were selected and 21 supplementary fields crafted for the new product identification.

Building the Models

The Computer Environment

Intelligent Miner has parallel algorithms that allow it to scale to handle large amounts of data on a parallel computer. It was run on a 16-node RS/6000, each with one CPU and one gigabyte of memory (Figure 1). Intelligent Miner can build its algorithms directly from the DB2 database or from a binary format that allows fast I/O and a compact representation so that as much of the data as possible (and preferably all) can be stored in memory. This results in much faster model building so that the analyst can explore more options.

New Product Identification

In order to identify possible new products that could be sold to financial institutions, the households were divided into segments using clustering. Segmentation is a business problem that requires identifying groups with common characteristics while clustering is an unsupervised learning technique that groups similar records.

The clustering in IBM's Intelligent Miner uses either a method based on a type of neural net called a Kohonen feature map or else a method based on demographic clustering. After selecting clustering as the type of modeling, the neural net option was selected. The data flow for neural clustering is shown in Figure 2, which is a screen shot from Intelligent Miner.

IBM's analysts specified 100 clusters, far more than the more typical 10 to 20 clusters. Finding this many clusters is a computationally intense process, requiring many passes through the data, that can only be done effectively with parallel algorithms such as those used by Intelligent Miner.

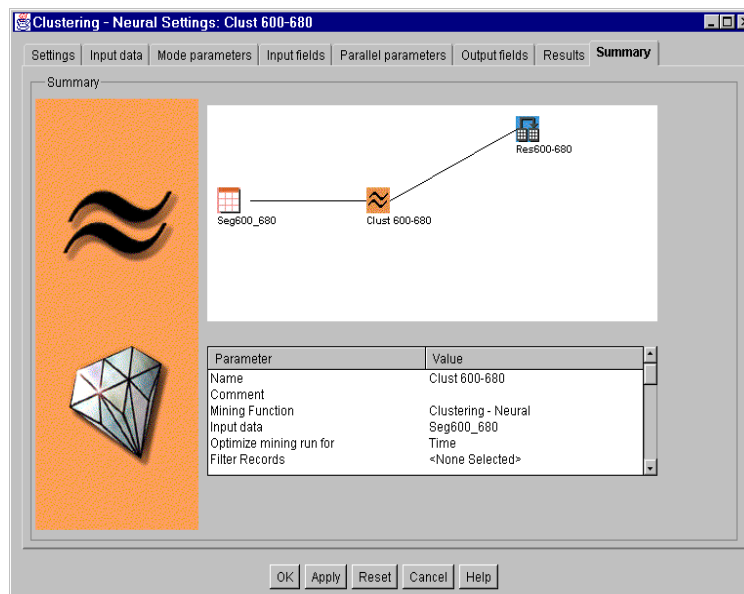


Figure 2. Data flow for neural clustering

The real difficulty with clustering is not in creating the clusters but in figuring out what they mean. The algorithm has determined that certain rows are very similar, but it is up to the analyst to interpret the clusters. Although choosing 100 clusters consumes a lot of computer resources, a finer grained structure emerges enabling the analyst to more precisely identify customer segments and group different clusters into the correct segments.

Visualization tools can be a big help to the analyst in understanding what the clusters mean and translating those insights into meaningful customer segments.

Intelligent Miner comes with a cluster visualization tool that shows the distribution of values for each variable in the cluster. Figure 3 shows an example of this clustering visualization for a sample database. Each cluster is shown as a horizontal band, with the largest cluster at the top and the smallest cluster at the bottom. A vertical bar at the left shows relative cluster size, and also labels each cluster with a percentage of total data.

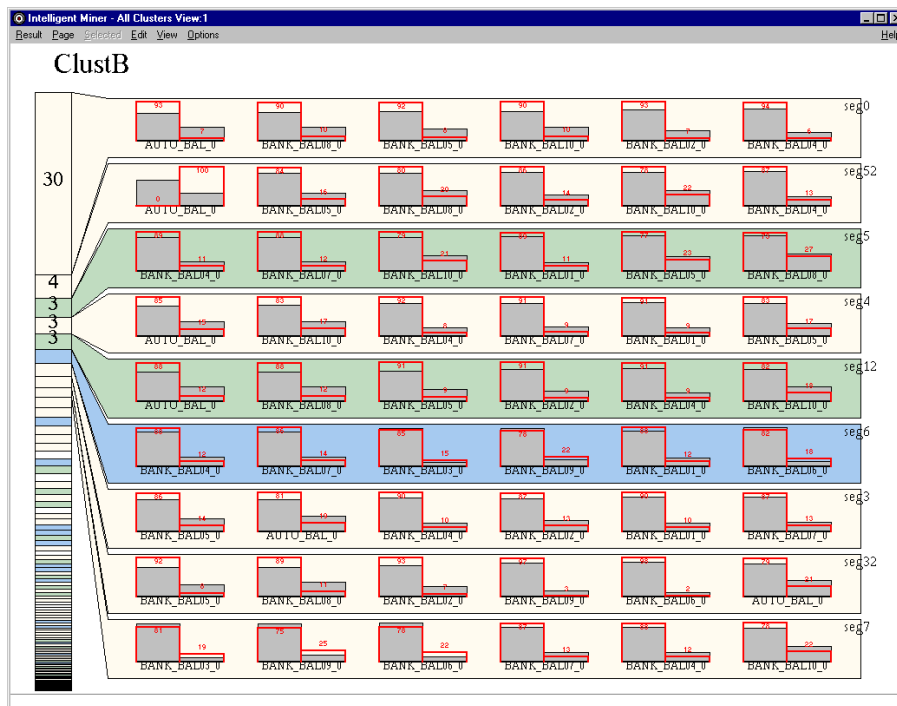


Figure 3. Intelligent Miner cluster visualization

Within each cluster are graphs for each of the six fields that was most important in determining the cluster. A pie chart, histogram or line chart (depending on the type of variable) is displayed. The distribution of values for both the whole data set and the cluster is shown.

IBM also developed a visualization tool called Kmap to help with this application. Although not yet a released product, it is one of the most useful visualization aids for clustering. In Figure 4, you can see the bankruptcy data clustered into 36 groups. The outer coloring of each group shows the population size, while the inner coloring of each group can be chosen by the user to show different attributes (in this case, total spending). The actual cluster population and spending are shown at the top of each cluster while the top three variables (items purchased, in this case) are shown in the body.

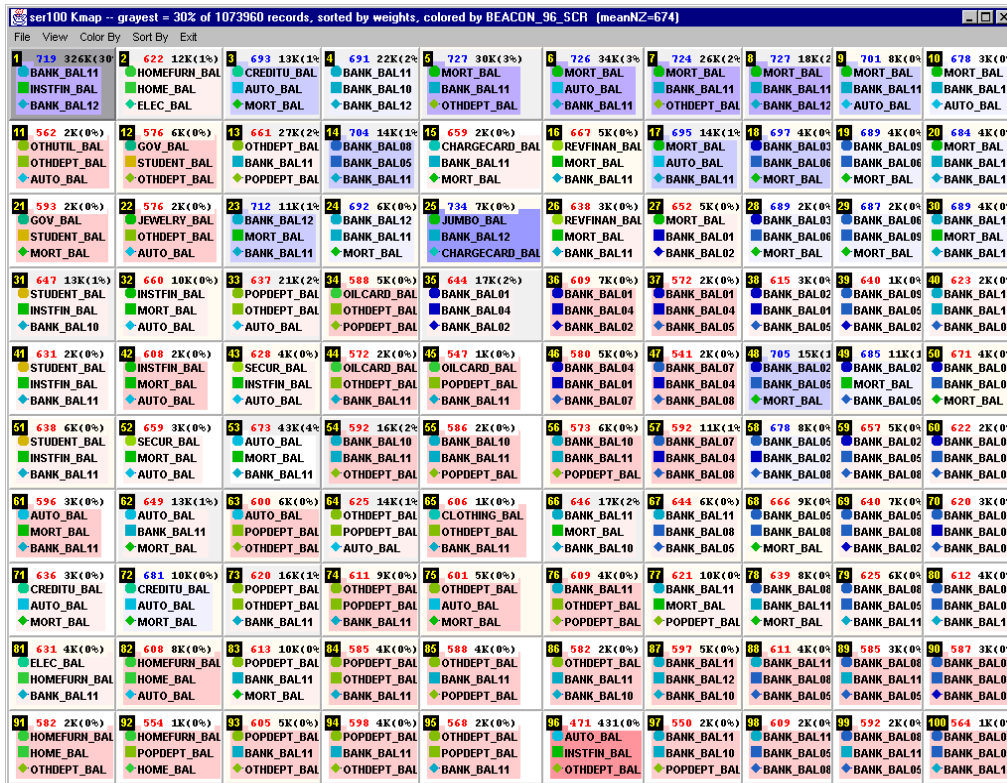


Figure 4. Kmap cluster visualization

To better understand each cluster in the data, the top contributors can be shown in a table (Figure 5). In this example, the columns have been sorted to show the average credit balance. Another important feature is that the percentage of an attribute in each cluster can be easily displayed. These charts form an indispensable aid to interpreting clusters and aggregating them into sensible business segments.

Field	Wgt	mean	R	meanNZ	R	partcp	R
STUDENT_BAL	0.914	66,434.94	66.27	66,434.94	8.51	1.00	7.79
INSTFIN_BAL	0.335	42,003.10	16.85	46,700.18	5.12	0.90	3.29
MORT_BAL	0.100	39,285.70	1.50	86,997.49	1.22	0.45	1.23
JUMBO_BAL	0.013	16,616.15	2.66	246,285.99	0.95	0.07	2.80
AUTO_BAL	0.092	7,648.11	1.88	14,515.20	1.21	0.53	1.56
BANK_BAL12	0.067	3,456.31	2.13	6,675.02	1.53	0.52	1.39
BANK_BAL11	0.151	3,148.16	2.21	3,797.28	1.50	0.83	1.47
CREDITU_BAL	0.031	1,804.55	1.87	9,908.60	1.24	0.18	1.51
GOV_BAL	0.027	1,577.69	5.36	14,947.84	2.64	0.11	2.03
SECUR_BAL	0.015	1,499.21	1.60	10,621.94	1.23	0.14	1.31
BANK_BAL05	0.090	909.13	2.05	1,617.42	1.12	0.56	1.82
REVFINAN_BAL	0.021	891.76	1.48	3,095.53	1.06	0.29	1.39

Figure 5. Top ten contributors to a cluster

Bankruptcy Prediction

The approach selected for bankruptcy prediction was to build scoring models, in which a score indicating the likelihood of bankruptcy would be calculated.

As in the new product identification problem, creating and selecting variables was one of the keys to success. In particular, IBM's insightful encoding of a household's credit limits over time helped yield good results.

The models were built using both neural nets and decision trees. After selecting the input data sets, selecting the variables for input and the class variable to be predicted (Figure 6) and setting the

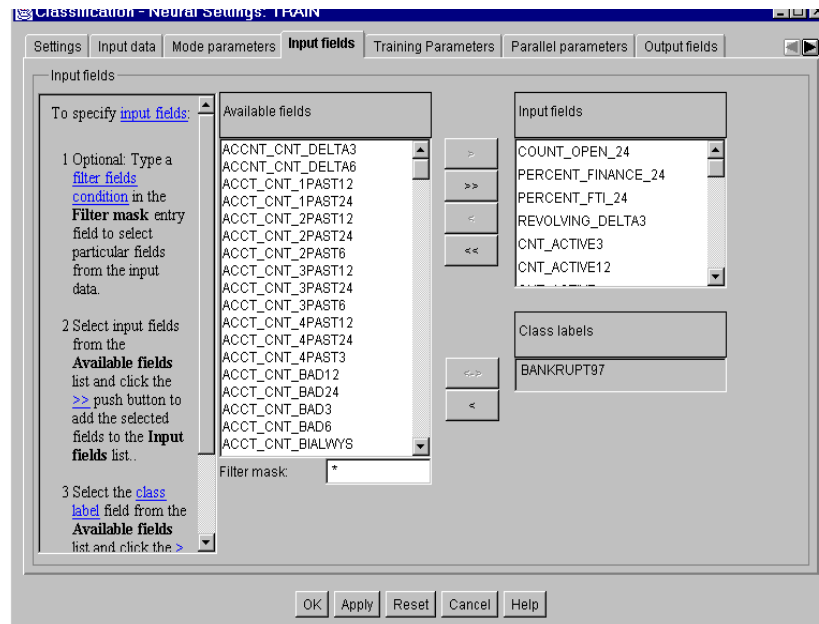


Figure 6. Selecting the input variables and class variables

parameters (including for the neural net the number of hidden layers and nodes, stopping conditions, learning rate and momentum), the neural net and decision tree were built. Figure 7 shows the data flow for training and testing the neural net.

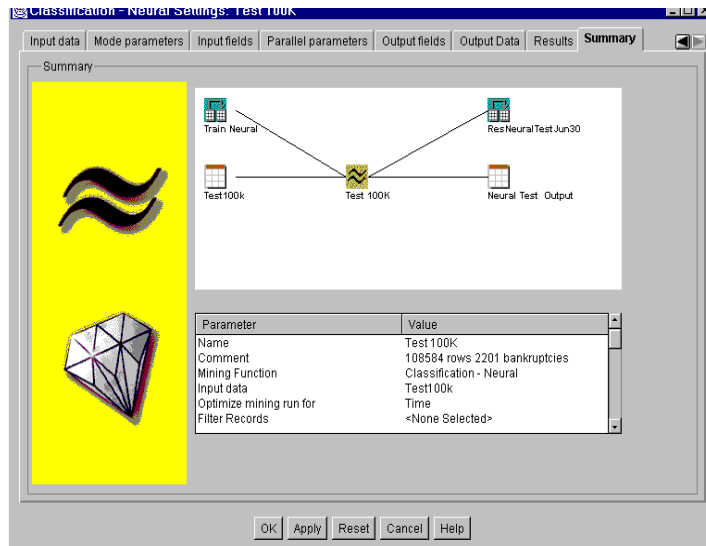


Figure 7. Data flow for training and testing the neural net classification

Results

The results were very good. To protect client confidentiality, however, the following discussion must be somewhat general.

New Product Identification

A careful analysis of the one hundred clusters identified over a dozen interesting customer segments. Furthermore, over 75% of the clusters could be identified as belonging to one of the segments. As can be seen in the diagram (Figure 8), the clusters didn't form neatly into segments, but clusters from different parts of the diagram were grouped together to form segments. The extra computing resources necessary for fine grain clustering as IBM did here paid off in the usefulness of the segments identified and the inclusion of most of the data in the segments.

LOW			Old	Low rollers	Low rollers	Low rollers	Low rollers	Low rollers	Low rollers
	STUDENT	Occasional Shoppers	Medium	Low rollers	Low rollers	AUTO	BIG CHURN	Low rollers	Low rollers
STUDENT				JUMBO	Stodgy Mortgage	Little Churn	BIG CHURN	Small Business	Small Business
GOVMT			OIL	Little Churn	Little Churn	Little Churn	MEDIUM CHURN		
GOVMT		Popular	OIL	OIL	Little Churn		MEDIUM CHURN	MEDIUM CHURN	MEDIUM CHURN
GOVMT		AUTO					MEDIUM CHURN	MEDIUM CHURN	MEDIUM CHURN
AUTO	AUTO	AUTO	Occasional Shoppers		Old	Old	Medium	Medium	Medium
AUTO	AUTO	Popular	Occasional Shoppers	Occasional Shoppers	Occasional Shoppers	Old	Medium	Medium	Medium
		Popular	Occasional Shoppers	Occasional Shoppers	Occasional Shoppers	Old	Medium OLD	Medium	Medium
		Popular Dept		Occasional Shoppers	AUTO	Old	Old	Medium	Medium

Figure 8. Segments found from combining clusters

These segments suggest business opportunities in that they represent clearly distinguished populations with characteristics that can be used in selling credit products. For example, many financial institutions are trying to market to small businesses, but it can be hard to identify these prospects. Some reasons for this difficulty include the facts that they are not always incorporated or that they may operate out of a home. However, the segmentation study identified small businesses managing cash flow using credit cards. A list of such potential customers is clearly very valuable.

Another interesting example was a segment that identified households with low credit usage. Marketers can avoid this group as one which is not particularly profitable, or they can study it in more detail to understand the various reasons for low credit usage in hopes of finding untapped sub-populations for whom they can create customized offerings that meet their individual needs.

One of the most useful results was identifying a subset of a group that is usually rejected for credit but who in fact are good credit risks. This list undoubtedly contains numerous households who want credit, but haven't been able to get it – in other words, a virtually untapped market.

Bankruptcy Prediction

The bankruptcy prediction models worked extremely well on the test data — so well in fact that a great deal of analysis was done to ensure that the results were not the result of a subtle error. Remember that in the test data, we already know who declared bankruptcy and we want to test how good our model's prediction is. Getting a result that is too good is cause for suspicion. For example, analysts occasionally make the mistake of using a predictor variable that includes the result in a disguised fashion.

In this case, a small number of households with the highest scores (the likeliest to declare bankruptcy) accounted for over 25% of the bankruptcies. In fact, as the scores dropped and analysts looked at more and more households, the model was consistently better than the old method was in predicting the bankruptcies in that group of households. This difference is called lift, and a model that is consistently superior to another is a highly desirable outcome.

However, a careful evaluation showed no error. Rather, the most likely explanation for this great improvement in performance was the incorporation of the 18 months' transaction history which the older methodology did not use.

Performance and Scalability

This case study is a good illustration of how scalability is vital to successfully mining large databases.

We saw earlier how important the scalability of DB2 was to allowing the analysts to explore the data and create the database. It is clear that given the huge amount of data that needed to be househanded, the new variables added, the different data sets created for different modeling runs, and the multiple iterations of the data transformations, all the time available could be soaked up in data preparation were it not for parallel DBMSs and computers. This is why twice as many CPUs were devoted to data preprocessing as to model building.

The scalability of Intelligent Miner was equally crucial. In performing the clustering study, a comparison was made between a one-CPU and 16-CPU RS/6000 SP. The difference in time was over eight hours for one CPU as compared to about half hour for 16 CPUs. (It is interesting to note that the speedup was approximately linear: 16 CPUs took one-sixteenth of the time.)

This difference would be of minor significance if you anticipated doing only a single clustering run. You could start it when you left work, and examine the results when you returned. However, data mining is an iterative process: you need to look at multiple models to see if you can improve the result. Perhaps you didn't get the right variables on the first shot, or some parameters needed be changed. In such an environment, an eight hour turn-around is unacceptable given deadlines and the finite resources most organizations are able to devote to a problem.

Another issue to explore in assessing the importance of scalability was the necessity of using so much data. Did the usefulness of the results improve as a function of the amount of data used?

The one million households analyzed represented about a 0.5% sample of the entire database. If a smaller sample could have been used, then the necessity for parallel computation would have been reduced.

IBM analysts investigated this for the clustering models by building models using one-sixteenth, one-eighth, one-quarter, and one-half the data. Four different metrics were used to compare the segmentation of these subsets with the segmentation of the full data set. As the amount of data used to build the model increased, the segmentation continually improved and approached the segmentation of the full data set.

Conclusion

Many companies have enormous amounts of data containing valuable information for running and building a business. Extracting the value of that data is a challenge that takes first and foremost an understanding of the business problems to be addressed and of the data itself. Using that understanding as the launch point of a data analysis effort, data mining can provide valuable insights and guidance for such diverse tasks as product creation and operational decisions. For cracking open large databases, parallel computers and software are key components that enable analysts to deliver accurate results in a timely fashion.