Mining the (data) bank

PEDRAM ATAEE

"Where is the knowledge we have lost in information?" said T.S. Eliot. In recent years, we have been confronted with a growing amount of data, which has caused our ability to analyze, interpret, understand, visualize, and make sense of our data to decrease. So a new field called "data mining" was created. Data mining, according to Fabian, is the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories, using pattern recognition as well as statistical and mathematical techniques.

Various types of data such as transaction, demography, lifestyle, finance, economy, and government can be used in the process of turning data into strategy. Many tasks can be accomplished with data mining: classification, forecasting, clustering, deviation detection, description, visualization, and link analysis. In banking, there are several applications for data mining: customer segmentation/profitability, predicting payment default/attrition, branch network optimization, fraud containment, optimizing stock portfolios, ranking investments, and economic forecasting. Various techniques of data mining exist such as exploratory analysis, segmentation, and modeling, each of which has several methods:

- exploratory analysis: correlation analysis, univariate, CHAID/CART, and quality control review
- segmentation: RFM, clustering
- modeling: regression-based models, neural networks, and genetic algorithms.

Data mining can be divided into two main categories, supervised and unsupervised. In supervised methods, the goals are clear, and the researcher must show the goals. In unsupervised methods, there are no clear-cut goals, and the researcher has to extract patterns and knowledge from this raw data.

The unsupervised method was used in this article. The analysis was done by different methods such as windowing, time series, and clustering, which were subsets of the segmentation family. The results of using pure statistical functions are shown in the end of this article.

Extracting pattern

Our database contains "amount of transaction" and "date of transaction," which were done at the automatic teller machine (ATM). This database includes transaction data of three financial years (1998, 1999, and 2001). Because of security regulations, I was unable to observe or survey their customers for more information. Available data from customers' bank accounts were limited; therefore descriptive statistical methods and an empirical approach were chosen as the data mining technique.

The data had to be cleaned before the processing was started. The cleansing stage was done in both SQL-Server and MATLAB. Each record that had wrong data was deleted from the database. For example, if account information from 1999 was considered, each record that was related to 2001 would have to be deleted. If a word was observed in date place, this record would be deleted from the database. If any record had missing data, it would also be deleted The date was transformed to number of days, starting from the beginning of the year. For example 5 February was transformed to 36 (31 + 5).

After the data had been cleaned, the data mining technique was chosen. Because of being just two types of data, "amount of transaction" and "date of transaction," the statistical method could analyze the data more accurately and fast than other methods that needed different type of data. The patterns that I expected were related to the days of the year. Therefore, it was guessed that weekly, monthly, or seasonal windows could be shown the periodical fact that existed in the raw data. This method was called windowing. If profit-making of bank customers were divided into several groups, it would follow the "80/20 rule." This method was called clustering. When the number of ATM users or amount of transactions according to the relevant day was drawn, the time series model had been used. The time series model is a model that forecasts future values of a time series based on past values. A pattern can be a complicated nonlinear relationship between two variables. Each

0278-6648/05/$20.00 © 2005 IEEE

data-mining technique that could detect this pattern was named as the pattern recognition method.

The bank database had been prepared in SQL-Server format. Therefore the banking data was transmitted from SQL-Server to MATLAB. This transition could be done with the interface of a program like Notepad or a direct link from both MATLAB and SQL-Server. All the MATLAB calculations were in the matrix format, so the records were saved as a matrix in MATLAB for future analysis. To discover knowledge from raw data, the analyzer program used mathematical functions such as mean, sum, smooth, and normalize. To attain a more profitable program, the computer program was designed to be applicable to other similar data banks.

Because of disturbances in such data which was influenced by so many factors, smoothed data was defined in (1). To reduce mentioned disturbances, transaction amount of a day (Y_n) was affected by transaction amount of yesterday (Y_{n-1}), tomorrow (Y_{n+1}), the day before yesterday (Y_{n-2}) and the day after tomorrow (Y_{n+2})

$$Y_{\text{smoothed}} = \frac{1}{2}Y_n + \frac{1}{5}Y_{n-1} + \frac{1}{5}Y_{n+1} + \frac{1}{20}Y_{n-2} + \frac{1}{20}Y_{n+2}. \quad (1)$$

For comparing the same kind of parameters with different scale, it was needed to adjust all of them on the same scale. Therefore, normalized parameters, which were between 0 and 1, were used. In (2), Y was a transaction amount index of each day while $Y_{\max}$ and $Y_{\min}$ were the maximum and minimum of that index in all 365 days of the year.

$$Y_{\text{normalized}} = \frac{Y - Y_{\min}}{Y_{\max} - Y_{\min}}. \quad (2)$$

After the calculations had been done, related figures were drawn. Each type of graph has three samples in each year (1998, 1999, and 2001). Because of the extreme similarity between these three graphs, only one year is shown in this article. Figure 1 shows the number of ATM users on all days on the basis of relevant day. It shows that many people use the ATM on the first day of each month as enormous peaks in the graph are observed then. Figure 2 shows the daily sum of transaction amount on the basis of the relevant day. An increase of the daily sum of transaction amount was observed 10–20 February as that was a national Islamic holiday. During that time, stores lower their prices, and people withdraw money from the ATM to go shopping. Figure 3 is a sample of windowing. The data was considered in a weekly window. This figure shows the number of customers who use the ATM according to relevant days. This pattern shows that people use ATMs on Saturday more than the other days of the week. We show the clear pattern in all three years which it was shown in the following equations:

$$Y_{\text{Saturday}} > Y_{\text{Sunday}} > Y_{\text{Monday}}, Y_{\text{Tuesday}}$$
$$Y_{\text{Wednesday}} > Y_{\text{Monday}}, Y_{\text{Tuesday}}$$
$$Y_{\text{Wednesday}} > Y_{\text{Thursday}} > Y_{\text{Friday}}. \quad (3)$$

Some rules were attained with computation. The "80/20 rule," which exists in many fields, was observed here. It means that 80% of the total transaction

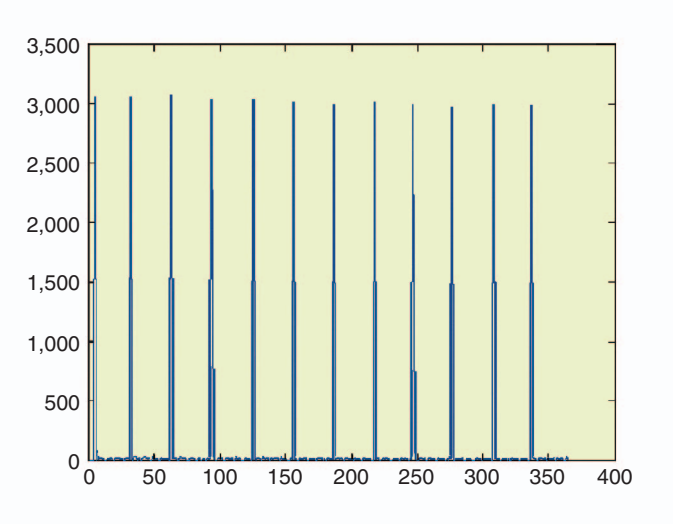

Fig. 1 Number of ATM user on the basis of the relevant day

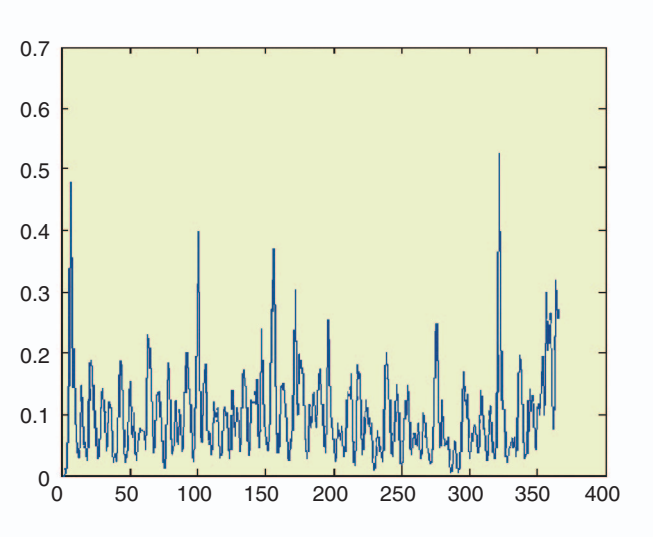

Fig. 2 Sum of transaction amount on the basis of the relevant day

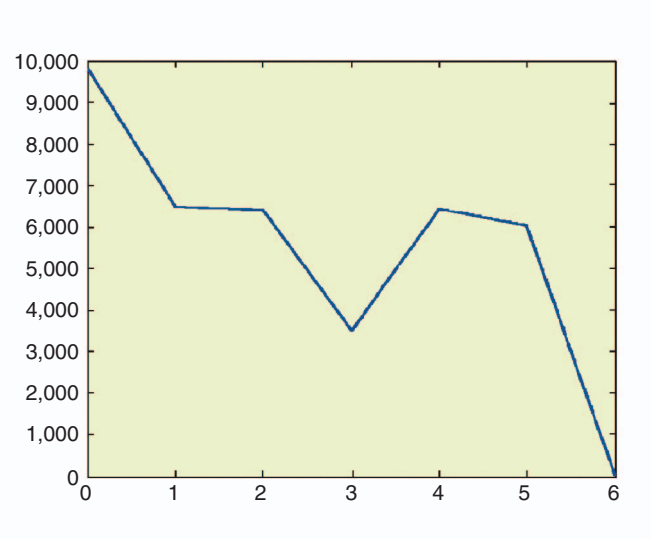

Fig. 3 Number of ATM user on the basis of the relevant day

amount was done by 2% of the greatest amount of the transaction, and the rest was done by 98% of the lowest amount of transaction. If the amount of transaction was considered as a statistical variable, the variance of this data would be approximately constant. Therefore it showed that the majority of the transaction amounts didn't show much change in 1998, 1999 and 2001.

Conclusions

According to the results of this article, bank officers could make decisions based on mined knowledge. For example, bank employees who are responsible for ATMs should consider that these machines not be out of service on the first day of each month and on Saturdays. Also machines should be full of cash so it doesn't run empty on these days. Therefore, if bank employees consider the mined knowledge, they can increase income, promote offered facilities, and prevent fraud.

On the other hand, some results can help bank customers. For instance, if they know that Saturday is the most crowded day at the ATM, they should go on other days to save time waiting in line.

Acknowledgments

I was very grateful for the support from the Research Centre of Mellat Bank. And special thanks to Prof. Caro Lucas for advising me in this article.

Read more about it

• M. Fabian, "Data mining to drive business intelligence and effective decision strategy," *Ottawa Congress Centre*, May 2002.

• M.L. Barja and J. Cerquides, "Application of data mining in banking," *Ubilab IT Laboratory*, Zurich, Switzerland, 1999.

• M. Spiliopoulou, "Intensive course for the SAS Data Mining challenge," *Data-Mining-Cup*, 2003.

• Data mining glossary [Online]. Available: <http://www. twocrows.com/glossary.htm>

• M.J.A. Berry and G. Linoff, *Data Mining Techniques for Marketing, Sales and Customer Support*. New York: Wiley, 1997.

• D.C. Montgomery and L.A. Johnson, *Forecasting and Time Series Analysis*. New York: McGraw-Hill, 1976.

About the author

Pedram Ataee graduated from the University of Tehran with a B.Sc. in electrical engineering in 2005. He is a Student Member of the IEEE. For more information, you can contact him at Ataee@ieee.org.

2006 WISE Program

Application Deadline: 16 December 2005 (postmarked)

The Washington Internships for Students of Engineering (WISE) program offers eligible engineering students the opportunity to spend a summer in D.C. learning how government officials make decisions on complex technological issues and how engineers can contribute to legislative and regulatory policy decisions. IEEE is a sponsor of this program.

Throughout the nine-week internship, students meet with leaders in the Congress and the Administration, prominent nongovernmental organizations, and industry. Applications for WISE are sought from outstanding engineering students who display leadership skills and have a keen interest in public policy. For application forms and more information, go to <http://www.wise-intern.org>.

WISE Program c/o IEEE-USA
1828 L Street, N.W.
Suite 1202
Washington, DC
20036-5104
Tel: + 1 202 785 0017
Fax: +1 202 785 0835
E-mail: info@wise-intern.org

The Piled Higher & Deeper

Paper Review Worksheet

Stuck reviewing papers for your advisor? Just add up the points using this helpful grade sheet to determine your recommendation.

No reading necessary!

Paper title uses witty pun, colon or begins with "On..." (+10 pt)	
Paper has pretty graphics and/or 3D plots (+10 pt)	
Paper has lots of equations (+10 pt) (add +5 if they look like gibberish to you)	
Author is a labmate (+10 pt)	
Author is on your thesis committee (+60 pt)	
Paper is on same topic as your thesis (-30 pt)	
Paper cites your work (+20 pt)	
Paper scooped your results (-1000 pt)	
TOTAL	

Points	Recommendation
< 0	Recommend, but write scathing review that'll take them months to rebuff.
0-120	Recommend, but insist your work be cited more prominently.
>120	Recommended and deserving of an award

JORGE CHAM © 2005 www.phdcomics.com