2007

Fraudulent Behavior Forecast in Telecom Industry Based on Data Mining Technology

Sen Wu
School of Economics and Management University of Science and Technology Beijing

Naidong Kang
School of Economics and Management University of Science and Technology Beijing

Liu Yang
School of Economics and Management University of Science and Technology Beijing

Follow this and additional works at: https://scholarworks.lib.csusb.edu/ciima

Fraudulent Behavior Forecast in Telecom Industry
Based on Data Mining Technology[*]

Sen Wu
Naidong Kang
Liu Yang
School of Economics and Management
University of Science and Technology Beijing

ABSTRACT

Outlier analysis in data mining is to find the deviation and abnormal data in the databases. By finding and explaining the outliers, we can apply them to detecting the business frauds effectively. This paper analyzes the common characteristics of fraudulent behavior of customers in telecom industry systematically. Based on the outlier-finding by clustering in data mining, we propose an effective solution to forecast the customers who are maliciously in arrears. Coupled with the actual application of forecasting the customers who are maliciously in arrears in telecom industry, we propose the specific method to forecast this kind of customers by using Kohonen neural network clustering algorithm.

INTRODUCTION

The behavior of malicious arrearage in telecom industry in China is common and has lasted for a long time. The malicious arrearage increases the bad account ratio in telecom companies and makes the profit untrue and the state-owed assets lost.

Telecom industry is data-intensive. After many years of development, telecom companies have the detail records of both the customers and the real-time calls. Therefore, by analyzing the data on hand, companies can know the purchase habits and the nature attributes of customers. Based on the analysis of customers' behavior, credit and risk, operators can build a system to avoid frauds. Through this system, operators can manage customers on-line at any time. Once a customer has abnormal phone call, operator gives alarm to service staff to take necessary actions, in case the customer leaves in arrears maliciously.

Based on the clustering technique in data mining, this paper focuses on the problem of finding abnormal outliers in large data set by using Kohonen neural network clustering algorithm, and gives the specific method to forecast the behavior of malicious arrearage.

Clustering Technique in Data Mining

(1) Knowledge discovery by clustering

From business aspect, data mining is a new method to process business information. Its main characteristic is that it selects, transforms and analyzes transaction data in business databases, extracts crucial information to help making business decisions. A common problem all companies face is that valuable information is limited though the data is rich. Therefore it is just like to collect gold from mine for company to gain information from the large data set, so data mining gets its name(Han and Kamber, 2001). Company can enhance its competency by deep analysis of the information.

Outlier detection in data mining is to find the deviation in databases, this kind of exploration of deviation or abnormal pattern has important meaning in related fields such as forecasting the customers who are maliciously in arrears in telecom industry.

We could use clustering to accomplish the outlier detection. Knowledge discovery by clustering is one of the important functions of data mining. From the aspect of data mining, clustering research extracts valuable knowledge from large data sets intelligently and automatically. Knowledge discovery by clustering was proposed along with the development of databases and the emergence of data mining and Knowledge discovery technology. Knowledge discovery by clustering is applied in many areas, such as: forecast of bankruptcy, pattern recognition, marketing, market segmentation and so on. The abnormal data implies outliers in clustering. Through the finding and explaining of the outliers, we can extract the features of abnormal data, so apply them to detect business frauds effectively.

(2) Kohonen neural network clustering algorithm

We compared three kinds of clustering algorithms by experiment, Kohonen neural network algorithm, two-step clustering algorithm, and K-means algorithm. We figure out that Kohonen neural network algorithm is the most effective algorithm to find outliers.

Kohonen neural network was proposed in 1981. It is also called self-organizing feature maps neural network clustering method. It gets the clusters of data set by using iterative algorithm to optimize the objective function (Li, Deng and Li, 2004; Yuan, 2000; Chen, 1996).

Kohonen neural network is one of the most popular neural network methods for cluster analysis. Its goal is to represent all points in a high-dimensional source space by points in a low-dimensional (usually 2-D or 3-D) target space, such that the distance and proximity relationships are preserved as much as possible. The method is particularly useful when a nonlinear mapping is inherent in the problem itself (Han and Kamber, 2001).

Kohonen neural network is a two layers feed-forward neural network, it has input layer and output layer. The number of neurons of input layer is M, which amounts to the dimension of input sample vector. The neurons of output layer are competitive output neurons, whose values are in {0, 1}. After self-organizing learning, Kohonen neural network makes the density of the connection weight vectors consistent with the probability distribution of input model, which means the density of the connection weight vectors could reflect the statistical features of input model. The layout of neurons in output layer has many forms, the typical one is two-dimension plane matrix (Yan and Zhang, 2000; Malone, etc., 2005).

BASIC IDEAS TO FORECAST THE MALICIOUS ARREARAGE

(1) Outlier finding

To forecast the behavior of malicious arrearage in telecom industry based on outlier detection in data mining, is to find the abnormal customers from the large customer data set who have high-level call fees in short period of time and leave in arrears at last. In order to find the outliers, we first capture the features of the customers who leave network in arrears, then build model and apply it to analyze the current customers by comparing the features of current calling behavior of customers with those of malicious arrearage finally, detect the customers who will be possibly in arrears, warn them and suspend service to them (Wang, 2004).

Classification is widely used for prediction in data mining, but when the ratio of target customer is small, it will miss many important rules by using classification, so the forecast result is poor. Outliers are the seldom abnormal case in data set. Outlier finding is a special and crucial application in data mining. We view the customers who are maliciously in arrears as outliers in the entire customer set. Based on Kohonen neural network clustering, we can find the outliers before classification in order to improve the quality of classification and forecast. The process of clustering can extract the outliers from the entire data set, but cannot state the features of this group, so we still have to use the decision tree of classification to extract the classification rules of this special group, and apply them to the process of business forecast.

(2) Main steps to forecast the malicious arrearage

We use a data mining tool Clementine (a data mining tool of SPSS) and Kohonen neural network clustering algorithm to analyze the data of training set, in order to find customer groups which are maliciously in arrears and

the basic features of these groups, then we will use these features to forecast the customers who may be maliciously in arrears and the probability of malicious arrearage happens so that we can ask for calling fees before this kind of customers leave the network. In order to evaluate the effectiveness of the model, we will test the model by test set different from training set. We follow the standard process of data mining CRISP-DM (Wu, Gao and Bastian, 2003), the specific steps are as follows:

Step1: Business understanding: Data source is prepared in order to build the model for analysis; according to the business rules and requirements of telecom companies, clear definition is given of customers who are maliciously in arrears, and the entire customers are divided into target customers and unrelated customers.

Step 2: Data understanding: Relevant data is selected based on the business understanding. By processing the bills and detail records of customers in a regular time period, we connect the fees of calling, time of calling, duration of calling, type of calling with information table and arrears table of customer and build feature records set of target customers.

Step 3: Data preprocessing: it is the most time-consuming process in data mining. Its goal is to transform the original data into reliable data for mining (Corinne, etc., 2001). In the subject of forecasting and analysis of the malicious cheating, data preprocessing includes sampling, setting the label of cheat, the feature field calculation and so on.

Step 4: Modeling: The sample customers are cluttered by Kohonen neural network algorithm in order to find the target customers which have features of malicious arrearage. We can extract the classification rules of target customers and build the model to forecast malicious arrearage.

Step 5: Evaluation: adjusting and optimizing the model, evaluating the effectiveness of forecast model by test data set.

Step 6: Implement: we forecast the behavior of malicious arrearage in the coming period by the forecast model, give the probability of malicious arrearage, and show it to telecom supervisor.

DATA PREPARATION BEFORE MODELING

(1) Target customer definition

In order to forecast cheat, we should define the target customers beforehand. The target customers are the ones who are in debt and leave the network with high-level calling fees not paid. According to their information of consuming, account, state-changing, we forecast the probability of them to become the target customers in the later months.

Customers may be halted out of services for arrearage for some reasons by the telecom companies, but they will pay for the bill in time. These customers are not target customers. We discover that 96.6% of the customers, who are halted out of services for more than four months, will not pay for the bill in the fifth month. And we should focus on the customers whose amounts of debts are high. If we could make them pay for the bill, we can recover major losses of the operators.

We assume that the forecast period is one month. The customers who are maliciously in arrears, or the target customers, are described as follows: In forecast period, their states are normal and the bills for each customer are no less than 200 yuan, but they always halt with bills not paid or leave network in debt in the next four months.

(2) Setting the time window

The goal of forecast is to find the target customers before they leave the network. If the alarming point is too early, the accuracy is lower; on the contrary, if the alarming point is too late, we have no time to take action. Therefore our goal is to find the inclination of customers to leave network 2 or 3 months ahead. Based on the definition of target customers given above, we use the data in July, 2006 and the states of these customers in November, 2006, to build the model. That means if we want to forecast the customers who leave network in arrears in November, we can know it in August, the forecast depends on the data before August.

Indexes of the time window are as follows: (1) data window: July, 2006; (2) forecast window: November, 2006; (3) delay period: three months.

(3) Data understanding

Given the specific business goal, data understanding collects all the internal and external data relevant to the object of business. Then we check the quality of data and find subsets of data which are interesting and contain information.

In the subject of forecasting the customers who are maliciously in arrears, we extract customers' basic information table, bill of month, detailed records of calling, table of arrears and table of state-changing as our original data tables. From these tables, we can get the original data to forecast the customers who are maliciously in arrears. The original data includes: customer's ID, gender, age, time in network, state, fees of long-distant phone call, fees of roaming, lasting time of phone call, times of phone call, fees in arrears, service change type and so on.

(4) Setting the label of cheat

According to the standard of time window, we correlate the customer information tables from July to November. We choose the customer who joined the network before July and his/her state was normal and he/she spent more the 200 yuan in calling in July. We choose their states from August to November using customer's ID as foreign key. We add a field of flag as attribute of cheat in the tables. If the customer whose state is out of service and leave network in arrears, the value of flag is 1, which identifies the target customers; otherwise the flag is 0, which denotes that the customer is not our target.

(5) Feature field selection

We analyzed the amount of customer's spending in consecutive months and found that the range of cost fluctuates, but the range is small. Before the behavior of malicious arrearage conducted, the amount of customer's spending would rise dramatically compared with before. Therefore we create a derived variable, using the ratio of rising range of spending to the standard deviation of the amount of spending to measure the raised spending is normal or not. We call it index of spending. The formula is as below:

Index of spending of the month $n+1$

$$S_{n+1} = \frac{M_{n+1} - \overline{M}}{\sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2}} \qquad (1)$$

M_i is the amount of spending in the month i, $\overline{M}$ is average amount of spending by month,

$$\overline{M} = \frac{\sum_{i=1}^{n} M_i}{n} \qquad (2)$$

IMPLEMENT OF THE MODEL TO FORECAST MALICIOUS ARREARAGE

(1) Clustering analysis

The customers who are maliciously in arrears are outliers. Therefore when we set the parameters of the model, we should set enough classes to reflect the features of outliers so that the outliers will not submerged in the classes. We set 70 classes (the two-dimension matrix, length is 10, width is 7). The result of clustering analysis is show as figure 1. The black spots represent cheating customers, the gray spots represent the normal customers. A couple of

horizontal and vertical coordinates represent one class. In the picture, the black spots are mainly in the cluster (0, 6) and cluster (9, 6).

Figure 1:   Distribution map of the result of clustering analysis.
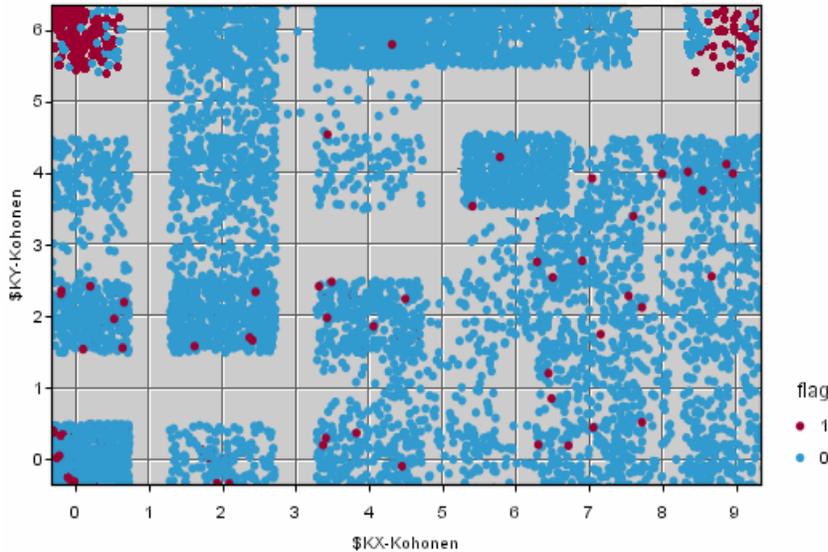
(2) Extracting the rule of feature

We get the cluster information of customers who are maliciously in arrears by using clustering analysis. If we want to acquire the main features of customers in that cluster, we should use decision tree of classification algorithm to attain the features of the target customer group. For example, a customer is target customer, if he/she has long-distant phone call with a particular number for a lot of times, the total time of calling is more than 12420 seconds or about 3.27 hours, he/she was in the network less than 6 months, and the index of spending is more than 4.164409.

(3) Evaluating the effectiveness of the model

We need to evaluate the forecasting effectiveness of the model by using test set. We can calculate the confusion matrix based on the test set.

Table 1: Forecasting Effectiveness

	Forecast not in arrears 0	Forecast in arrears 1
Actually not in arrears 0	A	B
Actually in arrears 1	C	D

Accuracy rate of forecasting malicious arrearage = D/ (B+D)
Accuracy rate of forecasting not in arrears maliciously = A/ (A+C)
Hitting rate of forecasting malicious arrearage = D/(C+D)
Hitting rate of forecasting not in arrears maliciously = A/ (A+B)

(4) Application of the model

The model is a classifier. We can forecast the target customers by it. If the probability of a customer to be our target customer is higher than a given value, we can suspect that the customer inclines to cheat and take necessary action in order to reduce the probability of malicious arrearage and to effectively protect the profits of telecom companies. Because the parameters of the model are retained by historical data the model will be "old" after a few months. We can calculate the accuracy rate and hitting rate by comparing the target customers we forecasted 2 or 3 months ahead with the customers who actually leave the network. When the two rates decline, we should retrain and adjust the model.

CONCLUSION

In this paper, we apply Kohonen neural network clustering algorithm to the analysis of forecasting the customers who are maliciously in arrears in telecom industry. The main contents and results of the research are:

(1) Data mining technology is used to forecast the cheat in telecom industry based on Kohonen neural network clustering algorithm. And the main steps are given to accomplish the task of forecasting the customers who are maliciously in arrears.

(2) Following the CRISP-DM data mining model, we give the definition of target customers who are maliciously in arrears and the model of data extracting. And the specific methods are proposed to build, evaluate, and apply the model.

(3) We find the outliers by using Kohonen neural network clustering algorithm, and furthermore extract the consumption features of target customers. A forecasting system is developed to find the customers who are maliciously in arrears for a telecom company.

REFERENCES

Chen, M. (1996). Neural Network Model. Dalian University of Technology Press. Dalian.

Corinne, B., etc.. (2001). Mining Your Own Business in Telecoms Using DB2 Intelligent Miner for Data. IBM RedBooks,

Malone, J., etc.. (2005). Data mining Using Rule Extraction from Kohonen Self-organising Maps. Neural Comput & Applic 15: 9-17.

Han, J., Kamber, M. (2001). Data Mining Concepts and Techniques. Academic Press, New York.

Li, Z., Deng, Q. and Li, H. (2004). Kohonen SOFM Neural Network Evolution and Research. Computer Engineering and Design. 1729-1830.

Wang, Y. (2004). The Study on Data Warehouse and Data Mining in Telecom industry Management Analysis and Application. College of Computer Science, Chongqing University. Chongqing.

Wu, S., Gao, X. and Bastian, M. (2003). Data Warehousing and Data Mining. Metallurgical Industry Press. Beijing.

Yan, P., Zhang, Ch. (2000). Artificial Neuron Network and Simulated Evolution Computing. Tsinghua University Press. Beijing.

Yuan, C. (2000). Artificial Neuron Network and Application. Tsinghua University Press. Beijing.