Intelligent Information Management, **2011, 3, 160-165**
doi:10.4236/iim.2011.34019 Published Online July 2011 (http://www.SciRP.org/journal/iim)

Churn Forecast Based on Two-step Classification in Security Industry

Ying Li[1], Ziyang Deng[1], Qiang Qian[1], Rui Xu[2]
[1]*East China University of Science and Technology, Shanghai, China*
[2]*Anhui University of Architecture, Hefei, China*
E-mail: {*liying*, falcondzy}@ecust.edu.cn, *elvaqq*2003*@yahoo.com.cn*

Received May 15, 2011; *revised May* 30, 2011; *accepted June* 8, 2011

Abstract

Customer is a determinant factor that decides whether a security company will be alive. As a result, the competition for customers is more and more intense between security companies. In order to avoid profit decrease caused by churn, security companies must find those customers who have the loss risk and make measures to maintain loyal customers. Now it is the question that how to find and analyze those customers. In this paper, a two-step classification method about churn Analysis is proposed and the problem of churn in security is analyzed.

Keywords: Customer Segmentation, Churn, Self-Organize Map Neural Network, Decision Tree

1. Introduction

In recent years, the scale of securities market has become larger and larger. Over 1700 firms and companies have listed in Shanghai and Shenzhen Market, and the Second-board Market has also opened. Meanwhile, the number and scale of securities traders keep increasing, with the result that competition is getting more intense day by day. To make the situation even more complicated, the number of investors who owns an account in Shanghai and Shenzhen has been increasing so rapidly that it has already exceeded 100,000,000.

However, our country's securities market is imperfect in the system. As a result, the market fluctuation range is great, and most of the time, retail investors suffer losses rather than make profits. Some of them are even trapped in the securities market or have to cut loss to exit the market. In securities market, while a large number of new customers open an account, a batch of customers outflow. An extreme example occurred in 2008 when the market index plunged suddenly. There was an obvious increase in the number of closed accounts as compared with 2007. Meanwhile, a great many non-transaction accounts emerged. The total business income of security companies grew relatively slow, hence the phenomenon “The increase in the amount of customers does not result in the increase in business income.” Therefore, analyzing the cause of churn, attracting potential customers to conduct transactions, enhancing the existing customer satisfaction, reducing the possibility of churn, raising the customer’s transaction level and taking a full share of the market is the key for Security Companies to win in the fierce market competition [1,2].

The paper proposes a two-step classification method about churn analysis that combines SOM network clustering method with decision tree classification method. It can provide a focused churn forecast analysis based on securities customer information that consists of a large amount of data and multi-attribute dimensions.

2. A Two-Step Classification Method about Churn Analysis

Usually, the effect of classification algorithm has much to do with characteristics of sampled data. Different data have different characteristics. For example, some data have much noise, some data’s value is vacant, some distribute sparsely and some sampled data’s attributes are strongly correlated. It is widely believed that there does not exist a kind of classification algorithm that can be fit for each kind of characteristic data at present.

Along with the use of database system after long times, the data accumulated in the database are also getting bigger and bigger. If we directly use some kind of classification algorithm to carry on the classification for the data that consists of a large amount of data and multi-attribute dimensions, the pertinence of classification will reduce and the classified effect will possibly sell at a discount. Therefore, In order to achieve the target-oriented classification and enhance the classified effect, we have the necessity to carry on the preprocessing that to discover the common characteristic from the sampled data and reduce choice of classified attribute prior to carrying on the classification analysis for the large-scale database.

The two-step classification method divides the entire classified process into two stages. The first stage is Customer Clustering which fulfills customer segmentation, firstly by mapping the data into a two dimensional plane by using the SOM neural network, then putting the similar characteristic data object together to form the clusters which are output from different output level unit. The second stage is customer classification. According to the result of segmentation, we carry on churn forecast by using decision tree classification method. The model realizes the process as shown in **Figure 1**.

2.1. SOM Clustering

Clustering analysis is an active research area in which the massive, the classic and the popular algorithms emerged, for instance, k-means, DBSCAN, STING, Wave Cluster, artificial neural networks and so on[1]. We may get the different result in various clustering method even regarding the same data set. Because the data in the paper is quite big, the dimension is high, and the amount of divided clusters is unknown in advance, clustering algorithm based on the division does not fit it. Moreover, the level clustering also has the request for the amount of cluster at the beginning that if the amount is too large, the complexity of the method is too high, which may cause the low-quality result of clustering. Therefore, the paper will use the SOM to carry on the clustering [3].

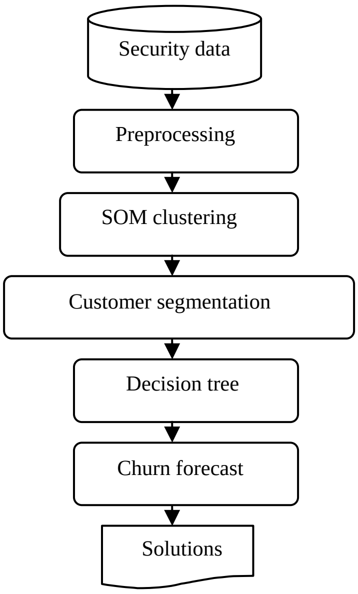

Figure 1. the model of two-step classification.

SOM (Self-Organize Map) is a kind of special neural network which maps the data from high dimensions input space to lower dimensions output space through non-linear transformation, and simultaneously maintains the adjacency relationship in the topology. Compared with other clustering algorithms, it simultaneously has the function of data synthesis and data mapping[4].

SOM, composed of input layer and competition layer, is a two structure neural network. Input layer is composed of N input neurons. Competition layer is composed of $m \times m = M$ output neurons which forms a 2-dimension plane array. It realizes all mutually connects between various neurons of input layer and various neurons of competition layer. According to repeated study of input model as its study rule, the network catches the model characteristic in which each input model contains, and according to self-organize, the result of clustering is displayed in competition layer, carrying on the automatic clustering. Any neuron of competition layer may represent the result of clustering [5].

2.2. Decision Tree Classification

In many classification methods, decision tree is a commonly used and intuitive fast method which consists of a commonly used structure that could guide study and rules of classifying inferred from a group of disordered instances [6].

There are usually two steps building a decision tree: making use of the training set to build decision tree; pruning the decision tree. Building a decision tree, started from the root node, is a recursive process from top to bottom in which we usually take a policy of divide and rule to construct the decision tree unceasingly through dividing the training sample into the subset. Decision tree's pruning is a process that to trim the tree structure and delete the unnecessary branching. While classifying the new sample with decision tree, we start testing the attribute of sample from root node and determine the next node according to test result until arriving at leaf node. The category of leaf node is the predicted category of new sample.

In the course of building decision tree, the node split is the key. Only by splitting the nodes according to the various attributes can it generates categories. Therefore, the core of whole question is how to choose splits. The procedure generally is to test all attributes, make the quantification evaluation of each attribute, then choose a best split way. The feature selection strategy has provided the quantification indices which mainly include information gain, the ratio of information gain, the distance measure, the G statistics, the smallest description length, the degree of correlation and so on[7].

3. Process of the Two-Step Classification Method

3.1. Data Resource

The data source used in the paper is from a security company's business hall in Shanghai from 2006 to 2008. The quantity of data surpasses 5 G, and the data recorded reaches more than 6,000,000. The database contains 18 data tables at all, 3 of which are considered to be the primary sources according to the purpose and the content of the paper shown as Customer Information Table (TCL_CUST_BASE_INFO), Customer Capital Table (TCL_FUND) and Historical Transaction Table (TCL_HIS_DONE).

Customer Information Table totally has 65531 records which records customers' basic information data of the business hall, including customer capital account, name, sex, certification type, ID Identification, nationality, authority scope, address, telephone, educational level and so on.

Customer Capital Table totally has 67255 records which records customers' capital account of the business hall, including security market value, assets and so on.

Historical Transaction Table, records from October 9th, 2006 to April 30, 2008, is the primary data table for data mining which belongs to a big multi-dimension data set, totally having 1621697 records and 71 dimensions. Each transaction data records much information, such as transaction date, customer ID, trading market, FRS, turnover, transaction volume, commission mode and so on.

3.2. Customer Segmentation Analysis

Based on the data table of choice, all indices, needed by SOM clustering analysis and generated from basic field[8,9], are as follows: (1) stocks speculation years; (2) capital scale; (3) customer holding ratio; (4) customer transaction times; (5) average commission; (6) average turnover; (7) average transaction volume; (8) average transaction price; (9) state. Customer clustering table is created by mixing these indices and fields of primary table. In addition, we add a set of fields into it: capital type, customer type, invest style, customer value, trade frequency and position. We carry on discretization according to the fields: capital scale, stocks speculation years, customer holding ratio, average commission and customer transaction times. Exact discrete standard is as follows:

According to capital scale, customer can be divided into small scale customer (capital scale < 50000), medium scale customer (50000 < = capital scale < 300000), major account (300000 < = capital scale < 1000000) and super scale customer (capital scale > = 1000000).

According to stocks speculation years, customer can be divided into new customer (stocks speculation years < = 1), steady customer (2 < = stocks speculation years < 6) and loyal customer (stocks speculation years> = 6).

According to customer holding ratio, customer can be divided into prudent customer (customer holding ratio < 0.34), steady customer (0.34 < = customer holding ratio < 0.67) and radical customer (customer holding ratio > = 0.67).

According to average commission, customer can be divided into less-valued customer (average commission < 30), medium-valued customer (30 < = average commission < 100) and high-valued customer (average commission > = 100).

According to customer transaction times, customer can be divided into occasional-transaction customer (customer transaction times < 20), active-transaction customer (20 < = customer transaction times < 60), positive-transaction customer (60< = customer transaction times < 100) and frequent-transaction customer (customer transaction times > = 100).

Position have 4 values: bear position(customer holding ratio < = 0.05), light position(0.05 < customer holding ratio < = 0.5) ,lager position(0.5 < customer holding ratio < = 0.9), full position(0.9 < customer holding ratio).

After data preprocessing, we carry on clustering analysis by the use of Data Mining tool SPSS Clementine. The result of customer segmentation is as shown in **Figure 2**:

Seen from the figure, there are totally 9 clusters. In the indices, trade frequency means the frequency of transaction in a period of time. The more frequently customer trades, the more commission securities trader takes. Position means the ratio of capital that used for stocks speculation. The higher the ratio, the higher the risk which customer faced will be. Customer value is an index that evaluates customer's importance for securities traders. The higher the value, the greater the contribution to securities trader will be. Customer type is standard by the date of opening in the security companies. The earlier customer opened and always made deals, the more loyal customer will be. Invest style means customer's style of stocks speculation. Some customers may be appetite for risk, some may be risk avoidance. Capital scale means the amount of customer's capital contribution. As a rule retail investors may be small scale of capital, while institutional investors may be large. We can also see from

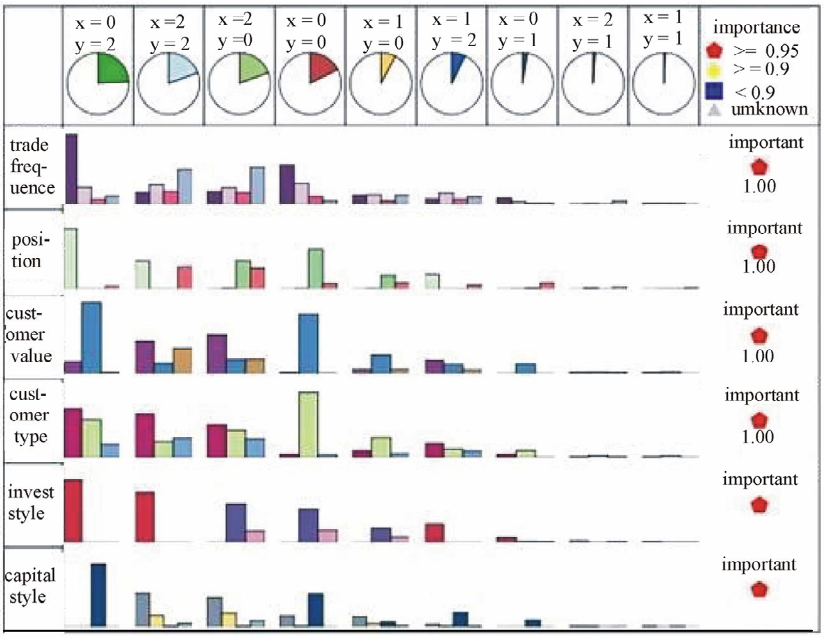

Figure 2. The result of customer segmentation.

the figure that the ratio of each cluster is various. 6 clusters that listed ahead are considered to be strong cell, in contrast 3 clusters that listed behind are considered to be weak cell.

Among the nine clusters, cluster(X = 0, Y = 2) take the biggest ratio that arrives at 24.28% in which the ratio of full position reaches 95.53%, the ratio of less-valued customer reaches 84.45%, invest style fully belongs to radical customer, and basically to be small scale customer. Cluster(X = 2, Y = 2) take the ratio of 19.39% in which customer also belongs to radical customer, medium scale customer take the domination position that arrived at 67.58%, the ratio of medium-valued customer is 47.85%, and there are also many high-valued customers. Cluster(X = 2, Y = 0) totally contain 2680 customers at the ratio of 19.28% in which customers belong to steady customer and the ratio of medium scale customer reaches 58.25%. cluster(X = 0, Y = 1), cluster(X = 1, Y = 1) and cluster(X = 2, Y = 1) , whose ratio separately are 2.72%,0.73%,1.14%, are the smallest cluster in which customers are radical and new, but have various trade frequency which are occasional, active and frequent.

Based on the analysis as before, 9 clusters can be summarized as shown in **Table 1**:

3.3. Churn Analysis

According to statistics by SQL statement, there are totally 13904 active customers which had ever carried on stocks speculation. Among them, 723 customers are lost and the ratio of churn reaches 5.2%.

Table 1. summarizing.

Cluster	summarizing
1 (X=0, Y=0)	less-valued, steady, occasional-transaction or active-transaction, new customer, small scale
2 (X=0, Y=1)	less-valued, radical, occasional-transaction, new customer, small scale
3 (X=0, Y=2)	less-valued, radical, occasional-transaction, new customer, small scale
4 (X=1, Y=0)	less-valued, steady, active-transaction, new customer, medium scale
5 (X=1, Y=1)	less-valued, radical, positive-transaction, new customer, medium scale
6 (X=1, Y=2)	less-valued or medium-valued, radical, active-transaction, mass customer, small scale
7 (X=2, Y=0)	medium-valued, steady, frequent-transaction, mass customer, medium scale
8 (X=2, Y=1)	high-valued, radical, frequent-transaction, new customer, major account
9 (X=2, Y=2)	medium-valued or high-valued, radical, frequent-transaction, loyal customer, medium scale

We make a count on the ratio of churn about each cluster according to the result of customer segmentation shown in chapter 3.2. The statistics is shown as **Table 2:**

Seen from above table, the ratio of the cluster(X = 0, Y = 2) is the highest, which reaches 12.38%. It is also the only cluster whose churn ratio surpasses 10%, in which the amount of churn reaches 418 in the proportion of 64.7% in total churn. It means that the cluster has the most loss risk that worth making further churn analysis. Therefore, the paper take the cluster(X = 0, Y = 2) as a data sample of churn analysis.

The model of churn analysis has totally 9 indices:

Table2. the statistics of churn.

cluster	churn	Total customers	ratio
X = 0, Y = 0	36	2431	1.48%
X = 0, Y = 1	5	378	1.32%
X = 0, Y = 2	418	3376	12.38%
X = 1, Y = 0	12	1075	1.11%
X = 1, Y = 1	0	102	0%
X = 1, Y = 2	55	1007	5.46%
X = 2, Y = 0	60	2680	2.24%
X = 2, Y = 1	9	159	5.66%
X = 2, Y = 2	128	2696	4.75%
Sum	723	13904	5.2%

capital account, recency, trade frequency, turnover, capital profit ratio, capital turnover ratio, average position ratio and state[10,11].

Because several indices include transaction information and capital news for a year, we should get rid of these customers who do not satisfy statistic requirement. Firstly, we should get rid of the account that closed before September, 2007, and then get rid of the account that opened after June, 2007. After that, there are totally 2714 customers in which 320 customers are lost.

We make use of Data Mining tool SPSS Clementine to set up a task of churn data flow, set the primary parameter of connected components, carry out the task and then generate decision tree of churn shown as **Figure 3**.

Rules of classification are set up according to **Figure 3**:

1) if trade frequency < = 9 and capital profit ratio < = 3%, state of customer is labeled as "churn";

2) if trade frequency < = 9 and capital profit ratio > 3%, state of customer is labeled as "not churn";

3) if trade frequency > 9 and recency < = 65, state of customer is labeled as "not churn";

4) if trade frequency > 9, recency > 65 and capital profit ratio <= –6%, state of customer is labeled as "churn";

5) if trade frequency > 9, recency > 65 and capital profit ratio > –6%, state of customer is labeled as "not churn";

These classification rules created by decision tree have important directive for security companies to carry on daily customer management and maintenance. Obviously, trade frequency, capital profit ratio and recency are important indices. For example, security companies should pay attention to small scale customer whose trade frequency is low, because if customer's capital profit is lower than 3% or be in a loss, the customer is very likely to be lost and measures should be made to maintain the customer.

Security companies should make proper marketing measures to avoid churn after identifying the customer segment of high churn trends. We can see from above analysis that trade frequency is the most important index, capital profit ratio and recency the second, which correspond to general rule seen from the experience of security investment. The reasons of customer lacking of transaction mainly are: (1) invest style of customer is in long-term hold; (2) While customer spends more capital on stocks speculation and is fastened by market, If he is not ready to sell the stocks in low price, recency becomes longer. In this situation, it is obviously that capital profit ratio is the key factor. Capital profit ratio being low or loss for a long time, investment incentives of the customer will decrease, which make the customer drain. Security companies should establish corresponding strategy to maintain the customers in the face of the phenomenon.

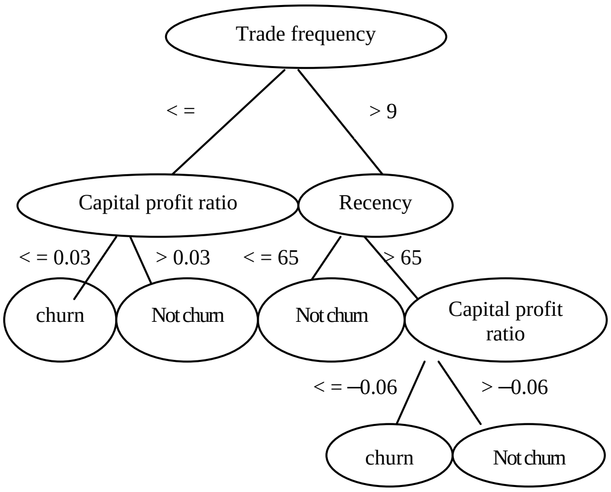

Figure 3. Decision tree of churn about cluster(X = 0, Y = 2).

3.4. Evaluation of Churn Model

The performance of models is a key factor of evaluating classification algorithm. We may get the different classification result in various classification algorithms even regarding the same data. To decide which is the best depends on user's explanation about the problem. The performance of classification model, as a rule, is evaluated by the accuracy of model.

The accuracy of classification, expressed as the percentage of tuple that is classified correctly to all tuple in the class, neglects the expense of classifying the tuple that not belongs to into the class, which should be considered. The paper evaluates the classification model of cluster(X = 0, Y = 2) through adopting the accuracy to explain the performance of models.

The method of accuracy calculation is: In the software SPSS Clementine, firstly, put the model nodes as middle nodes and append it to the type nodes; then append analysis nodes to model nodes; at last, execute it.

As a result, we can see from **Table 3** that 1499 customers that inferred from model are accordant with actu-

Table 3. comparison inferring with actuality.

	Not churn (inferred from model)	Churn (inferred from model)
Not churn (Actuality)	1345	44
Churn (Actuality)	41	154

ality, while 85 customers are not. The accuracy of classification reaches 94.63%, which means that the model can be considered as the classifier of churn analysis because of its high performance.

4. Conclusions

Since recent years, with the reducing of securities transaction commission, the appearing of sub-owned shares policy and the continuously depression of securities market, market competition in security industry is increasingly fierce and the traditional extensive operation is suffering a substantive change. At the same time, as a result of rapidly developing of Database Techniques and widely using of database management system in security industry, security companies have already maintained the massive data. Under such a background, extracting the latent and valuable model or the rule, revealing concealment commercial rule and constructing CRM in security industry according to mining and analyzing the exchange data by data mining technology, could effectively raise the security companies' service and decision level, and improve company's core competition.

The paper proposes a two-steps classification method and does the churn analysis of some security companies' business hall in Shanghai by the use of data mining tool SPSS Clementine. The rule of churn has been discovered according to the decision tree, which could certainly help security companies to avoid churn.

5. References

[1] H. H. Xie, "Application of Data Mining in Customer Segmentation of Stockjobber's CRM," *Computer Engineering*, Vol. 30, No. 10, 2002, pp. 553-554.

[2] W. H. Chen, "Application of Data Mining in CRM," *Microcomputer Applications*, Vol. 17, No. 10, 2001, pp. 25-28.

[3] G. J. Mao, L. J. Duan, S. Wang and Y. Shi, "Principle and Algorithm of Data Mining," Tsinghua University Press, Beijing, 2007.

[4] Y. C. Cai and L. C. Chen, "Summarizing the Researches on the Clustering Algorithm," *Science and Techenology Information Development & Economy*, Vol. 17, No. 1, 2007, pp. 145-146.

[5] F. Murtagh, "Interpreting the Kohonen Self-Organizing Feature Map Using Contiguity-Constrained Clustering," *Pattern Recognition Letters*, No.16, 1995, pp. 99-408.

[6] A. Chen, N. Chen and L. X. Zhou, "The Data Mining Technology and Application," Science Press, Shanghai, 2006.

[7] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, Vol. 1, 1986, pp. 81-106. doi:10.1007/BF00116251

[8] Y. Liu, D. Wan and P. Zhang, "Comparative Research of Solutions for the Customer Segmentation Based on the Purchase Behave," *Management Science in China*, Vol. 16, No. 1, 2003, pp. 69-72.

[9] L. Liu and G. P. Yu, "Application of Discovering the Potential Customers Based on Self-Organizing Feature Map Neural Network," *Journal of Nanchang University(Natural Science)*, Vol. 30, No. 5, 2006, pp. 508-510.

[10] L. Wang, S. L. Chen and X. D. Gu, "Analysis and Application for National Telecoms of Customer Churn Alarm Models," *Telecommunications Science*, No. 9, 2006, pp. 47-5l.

[11] S. R. Combating, "The Churn Phenomenon of Telecommunications," *International edition*, Vol. 31, No. 10, 1997, pp. 77-79.