Customer Churn Prediction Using Improved FCM Algorithm

Shaoying Cui
Transportation Management College
Dalian Maritime University
Dalian, P.R.China
e-mail: cherry_apple8232@icloud.com

Ning Ding
Transportation Management College
Dalian Maritime University
Dalian, P.R.China
e-mail: dingning860711@163.com

Abstract—**Data mining can provide support for bank managers to effectively analyze and predict customer churn in the era of big data. After analyzing the reasons for the bank customer churn and the defects of FCM algorithm as a data mining algorithm, a new method of calculating the effectiveness function to improve the FCM algorithm was raised. At the same time, it has been applied to predict bank customer churn. Through data mining experiments of customer information conducted on a commercial bank, it's found out the clients have been lost and will be lost. Contrast of confusion matrixes shows that the improved FCM algorithm has high accuracy, which can provide new ideas and new methods for the analysis and prediction of bank customer churn.**

Keywords: Bank customers churn; Empirical analysis; FCM algorithm; Validity function

I. INTRODUCTION

At present, the customer information of most banks is in an idle state. The customer relationship management system in bank is weak in the application of data mining and analysis and lacks the quantitative analysis of massive customer information. The management and development of clients is still restricted in the subjective judgment of decision makers. As an ever increasing competition of the market, the homogenization of products and services offered by banks are getting worse, customer loyalty becomes lower and lower too. The 5% customer loyalty decline will lead to a reduction of 2% of the bank profits. The cost of retaining the customers is only the 1/5 of developing them. In addition, the advent of the era of big data sound alarm clock for bank operators and convey to them the information cannot be ignored; data mining era of big data can provide valid data to support the bank in the process of customer relationship management. Applying the data mining to customer churn analysis model has become a top priority. Therefore, the churn management in big data era has become the key to existence and development of banks. While developing new customer resource, banks should continue to predict the customers' churn behavior, propose appropriate measures to retain customers and strengthen customer relationship management.

Extracting hidden and valuable customer information from massive，noisy and fuzzy data core idea of data mining in big data era is a new method. Compared with OLAP, it can identify potential relationships and features between things automatically and using these feature patterns for effective predictive analysis. OLAP is a validation of the data analysis, it need a question or hypothesis. Therefore, data mining in big data era can ensure the authenticity and objectivity of mining results. How to select and improve data mining methods has become a hot issue in this field. Existing research most focuses on methods: T.Sato[1] put forward a customer churn prediction model based on principal component analysis, which is more accurate in predicting results compared with decision tree C5.0. Wojewnik P[2] combined the K-mean algorithm and traditional classification algorithms, and then put up with new hybrid algorithm, proved its higher accuracy in customer churn prediction. Heitz C et al.[3] put forward the method of using data mining techniques to excavate the customer's consumption behavior and basic information. On this basis, they predicted customer churn rate and analyzed the reasons. Shi Yang and Yue Jiajia[4] put forward a decision tree algorithm to applied it into churn crisis model. Zhu et al.[5] used Bayesian algorithm to study customer churn. Yu Lu[6] has analyzed and predicted the telecom customer churn using a model combining decision tree, neural network and logistic regression three algorithms. Fan Chunmei[7] has been using improved CRISP-DM data mining procedure, based on the China Mobile CRM system business records, analyzed and established the customer churn prediction model using decision tree algorithm and Clementine visual tools. On the basis of analyzing the characteristics of customer churn problem, Ying et al.[8] used vector institutions to build customer churn models.

The existing customer churn prediction method in domestic was mostly still restricted in the subjective judgment of decision makers. This method is based on experience. In this study, the authors synthesis the research papers of Li Zhaoying's "Science of customer management methods – how to reduce customer churn"[9], He Qingzhe et.al's "Customer churn general research" [10], as well as Wu Jingjing's "Applying data mining technique in customer churn applied research" [11], then exploit both qualitative and quantitative methods to improve the traditional fuzzy C-means clustering algorithm, and apply it into the data mining of customer

978-1-5090-6306-2/17/$31.00 ©2017 IEEE

churn prediction. C++ program language was used to implement the algorithm on a commercial bank data. Analysis and comparison of the result shows that this algorithm has higher accuracy.

II. BANK CHURN ANALYSIS

A. Commercial banks churn concept

Customer churn is defined as the propensity of customers to cease doing business with a company or turn to the services provided by other banks [4]. In order to survive in an increasingly competitive marketplace, banks are trying to tap more customers, which resulting in many customers switching the capital from a bank to another to achieve lower prices and better quality of service. Customer churn has become a common problem many banks worldwide having to face. What the bank most concerned about is the number of customers as well as changes in customer account balances, especially the loss of customer account balances. In many cases, customers transfer its assets without writing off. Banks must predict and prepare for this kind of behavior difficult to detect.

B. Analysis of the reasons for customer churn

There are two forms in bank customer churn, one is writing off the account; the other is turning the account into hibernation. When the consolidated assets of customer are less than a certain amount and stay in this state for a long term, it can be called dormant accounts. There will be many signs before dormant accounts, such as the decline of transaction frequency. Bank customer churn is divided into two types: one is called involuntary churn, which is referred to the bank abolishing the service initiatively, including the cancellation of a service and stopping serving for some kind of customer. That is customer natural demise. The other type is voluntary loss, which refers the customer doesn't use services provided by one bank any longer, even transfer the asset to another bank. Such loss is most concerned by the bank. The state of bank customer churn can also be divided into two types: One is customers have lost out already; the other is the customer on the decreasing, that the customers are transferring the property to other place. For the customers on the churn, the bank should find out the cause of churn and take appropriate measures to retain them in time.

The various uncertainties in bank customer relationship management process are the fundamental reason for customer churn. They can be divided into external factors and internal factors according to different root cause of customer churn produced. External factors refer to the external conditions of bank's management, including mutations and deterioration of policies, market demand, technological development and competition in the state. For example, the balance treasure introduced by the alipay platform caused a serious impact on the banking sector. Internal factors are errors in customer relationship management due to poor management of the bank managers, strategic decision-making mistakes, concept lag and technological innovation capacity decrease. Customer churn crisis is caused by both external and internal factors. Banks need to constantly adapt to the pressure of the external environment and continuous improvement of its own deficiencies and establish their own advantages.

III. IMPROVED FCM ALGORITHM

Clustering refers extracted pattern data of things to be clustering from the real world, seize the main features of different models by simplifying the number of dimensions and divide things into different categories with clustering decision. For each core point, an objective function was used to describe the extent of the merits of the candidate clusters, the most commonly used algorithms is means clustering method. This algorithm defines a cluster center. By means of repeated iteration and minimizing the overall degree of dispersion around the point, ultimately determine the category element belongs. C-means clustering method is divided into C-clear means clustering method and C-fuzzy means clustering method (FCM). Among them, FCM algorithm is applicable to a certain fuzziness of clustering analysis, the case that the factors affecting cluster analysis and its degree of influence are not entirely certain. As there is a certain degree of ambiguity and uncertainty in the collection, description and processing of bank customer information, thus, the paper selects FCM algorithm based on the objective function to research bank customer churn analysis model.

There are some drawbacks in FCM algorithm. First one is to determine the clustering number of datasets. Clustering number c must be clear before using, it greatly affects the validity of the objective function. Second, the objective function of the algorithm is non-convex problem. Its implementation is based on hill climbing method, and is also sensitive to the initial solution. So the algorithm is easy to fall into local optimal solution. Third, clustering trend and validity analysis of the algorithm are isolated. Additionally, there are other problems in FCM algorithm, such as sensitivity of fuzzy weighting exponent and the construction of dataset structure.

Bezdek[12] used "separation" and "compactness" to verify the validity of the clustering results. "Separation" refers to the degree of separation between sample classes; "compact" refers to the degree of dispersion or deterioration of the sample within one class. Bezdek and Dave[13]、Zhang M R et al.[14] Respectively constructed validity function based on datasets or membership degree

and found out the optimal clustering number by constructed validity function. This paper follows the idea that is constructing validity function to present a new and improved fuzzy clustering, the function is composed of separation measure and compactness measure, which performs better to find the optimal clustering number C* in a dataset.

A. Separation measure

The degree of separation was measured by calculating the distance between fuzzy sets. $X=\{x_1,...,x_n\}$ is the dataset, $L(x)$ is the set composed of datasets, $L=\{L_1,L_2,...L_c\}$ the fuzzy partition on X. In this paper, the similarity was used to measure the degree of separation.

$$S(L_i,L_j)=\max_{x\in X}\min\left(u_{L_i}(x),u_{L_j}(x)\right) \quad (1)$$

$$Sep(c,U_c)=1-\max_{i\neq j}S(L_i,L_j) \quad (2)$$

Where, $S(L_i,L_j)$ is the similarity degree of fuzzy sets and $Sep(c,U_c)$ is the degree of isolation and irrelevance. When $L_i=L_j$, which means that two fuzzy sets are the same then $Sep(c,U_c)=0$; when $Sep(c,U_c)$ is comparatively large, it means that fuzzy partition got is good in separation.

B. Compactness measure

Compactness measure can be defined as follows

$$Z(U,V)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}d^2(x_j,v_i)}{e(i)}\times\sqrt{\frac{c+1}{c-1}} \quad (3)$$

$$d(x,y)=\sqrt{1-exp\left(-\alpha\|x-y\|\right)} \quad (4)$$

$$a=n\Big/\left(\sum_{j=1}^{n}\left\|x_j-\bar{x}\right\|^2\right),\bar{x}=\left(\sum_{j=1}^{n}x_j\right)\Big/r \quad (5)$$

Where, $e(i)$ is the quantity of data in class i; α the reciprocal of the sample covariance; $Z(U,V)$ refers to illustrating average compactness between within-cluster by calculating the error intra-class. Because of the exponential measure distances with better noise robustness, the exponential function to calculate the error between cluster centers and sample points is used. Different weights coefficient for different categories with $1/e(i)$ were endowed; what's more, the coefficient $\sqrt{(c+1)/(c-1)}$ also has minor adjustments to the overall effect, which is better for division. When the clustering number c approaches the total number of samples n, $\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}d^2(x_j,v_i)$ is monotonically decreasing. In this case, $1/e(i)$ and $\sqrt{(c+1)/(c-1)}$ increased with the increase of c, which limits the decreasing of metrics, but also reach the purpose that compactness metrics should be as large and clustering number be as small as possible.

C. Validity function and optimal clustering number

The validity function in use of maximum method is stated as follows.

$$Q(U_c,V_c)=\frac{Z'(U_c,V_c)}{Sep'(c,U_c)} \quad (6)$$

Where,

$$Sep'(c,U_c)=\frac{Sep(c,U_c)}{max\,Sep(c,U_c)},c=2,3,...c_{max} \quad (7)$$

$$Z'(U_c,V_c)=\frac{Z(U_c,V_c)}{max\,Z(U_c,V_c)},c=2,3,...c_{max} \quad (8)$$

$Q(U_c,V_c)$ is the ratio between compactness and separability. The smaller the compactness metrics intra-class, the smaller the degree of dispersion; the larger the separation metrics between samples, the higher the degree of separation, and the results are better. Therefore, the minimum value of $Q(U_c,V_c)$ corresponds to the optimal clustering number C^* of the sample.

The $Q(U_c,V_c)$ is used as validity function and calculate the optimal clustering number C^* of sample set based on FCM algorithm. Implementation is illustrated in Figure 1.

IV. Empirical Analyses

3000 customers' basic information and historical business data were obtained in three months (January9, 10, 11 in 2013) of a commercial bank, and predict the customer churn based on the improved FCM algorithm proposed. The results show the accuracy of the algorithm.

A. Selections of indexes

By analyzing the factors that influence the bank customer churn, five indicators were finally selected as follows:

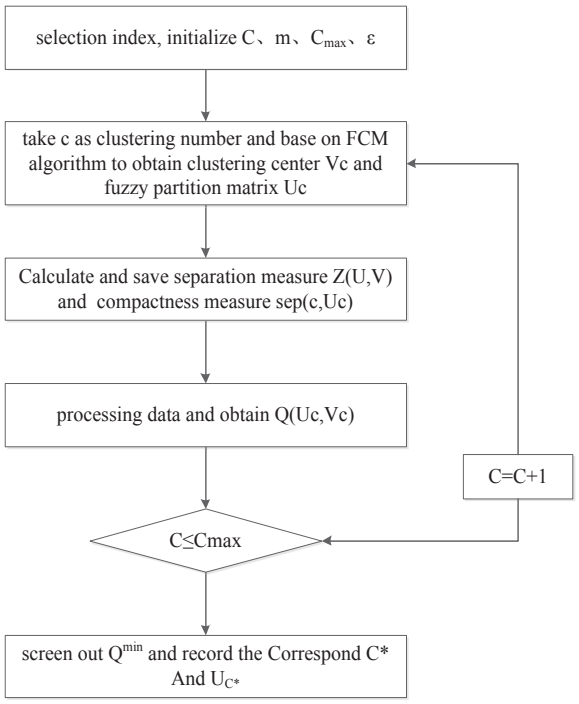

Figure 1. Implementation of improved FCM algorithm

1. Integrated contributions: integrated contributions of customer are the embodiment of customer value, the standard of measuring the customer value creation for banks. Integrated contribution value is determined by the full value brought from deposits, loans and investment of customer in bank, the calculate model is as follows:

$$V_i = Vl_i + Va_i + Vm_i \tag{9}$$

$$Vl_i = El_i - Cl_i \tag{10}$$

$$Va_i = Ea_i - Ca_i \tag{11}$$

$$Vm_i = Em_i - Cm_i \tag{12}$$

Where, V is the comprehensive contribution of customer i; Vl_i is the deposit value contribution of customer i; Va_i is the loan value contribution of customer i; Vm_i is the intermediary business product value contribution of customer i; El_i is the deposit funds transfer revenue of customer i d; Cl_i is accrued interest expense on deposits of customer i; Ea_i is loan products revenue of customer i; Ca_i is loan products cost of customer i, which is the sum of loan risk reserve, loan business taxes and internal funds transfer spending; Em_i is the intermediary business fee income of customer i; Cm_i is the intermediary business product cost of customer i, which is the difference between service fee expense and sales tax revenue of intermediary business.

The model takes into account not only the opportunity benefit and cost of internal capital, but also the cost of capital and tax cost. It is characterized by a comprehensive measure of the benefits and cost-sharing customers make to bank, which can measure the intrinsic value of customer more comprehensively and objectively. According to time properties feature, this paper selects integrated contribution a year, decreased percentage of integrated contribution in this month, and decreased percentage of integrated contribution in last month, which is respectively set to X_1, X_2, and X_3 as indexes.

①integrated contribution a year (X_1): As provided by the bank, if the integrated contribution of a customer to the bank is 0 in a year, it indicates that the customer has been lost.

②decreased percentage of integrated contribution in this month (X_2): That is the ratio of the difference between the integrated contribution in last month and this month and last month's overall contribution.

③decreased percentage of integrated contribution in last month (X_3): Calculation method is same to X_2, which are both the key to customer churn prediction.

2. The other two indicators: aggregate contribution of this month is set to X_4 and total client assets X_5.

B. Data processing and results analysis

The data was processed by cleaning, integrating and standardization. The decreased percentage of integrated contribution in this month is -1, if the difference between last month and this month's integrated contribution appears negative and the integrated contribution in last month is zero, and 0 if the difference between the integrated contribution of last month and this month is zero and the integrated contribution in last month is zero too. Similarly, the same rule of value assignment is used to calculate the decreased percentage of integrated contribution in last month in the two cases above.

In the process of parameter initialization for FCM algorithm and validity function, set $c = 2$, $m = 2$, $c_{max} = \sqrt{n}$, and $\varepsilon = 0.00001$, and implement the algorithms for every clustering number $c\left(c = 2,3,...,c_{max}\right)$, the results are showed in tab.1. The optimal clustering number C^* finally got is 5.

Cluster 1: decreased percentage of integrated contribution in last month is negative and close to zero, just as the decreased percentage of integrated contribution in last month. This indicate that customer of this kind is relatively stable. Although the integrated contribution and total assets of clients are not high, it is largest in customer number, accounting for 54.83% of the total sample, which can be defined as a potential customer base.

Cluster 2: the decreased percentage of integrated contribution in last and this month, and the integrated contribution in this month are all zero, integrated

contribution a year and total assets are particularly small, the customers in this kind are defined as the customers lost. The proportion of customer churn is as large as 24.77% of total number of samples, which indicate that the bank do not pay attention to the customer churn prediction and customer detainment.

TABLE 1. FINAL CLUSTER RESULTS BASED ON IMPROVED FMC **Algorithm**

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
X_1	0.16	0	0.59	0.23	0.12
X_2	0.09	0	-0.59	0.35	0.13
X_3	-0.06	0	-0.42	0.31	0.19
X_4	0.28	0.08	0.63	0.39	0.2
X_5	0.26	0.01	0.55	0.39	0.15
Number of clients included	1645	743	146	97	469

Cluster 3: the decreased percentage of integrated contribution in last and this month are both negative, and the amplitude is relatively large, the integrated contribution and total assets of clients are high. Such customers are defined as gold customers, who are expending the business volume in the bank continually and are the group who create most profit for banks. Of course, such a small number of customers, accounting for 4.87% of the total sample, are the key objects of bank customer detainment.

Cluster 4: the decreased percentage of integrated contribution in last and this month are both positive, integrated contribution and total assets of clients are relatively high, and the number of customer group is small, accounting for 3.23% of the total sample. High value of their customers means also higher risk of loss, and total client assets are transferred largely and fast, which can be defined as a group with high value customers on the churn. Some measures should be taken for banks to maintain and detain the customers in this kind.

Cluster 5: the decreased percentage of integrated contribution in last and this month are both positive, the values are all smaller than cluster 4, except the customer number, which can be defined as a group with sensitive value customers on the churn, accounting for 15.63% of the total sample.

C. Comparative analysis

Accuracy and coverage rate is used to evaluate the customer churn prediction model. Where the accurate rate is proportion of the actual number of customer churn in the number predicted by model; coverage rate is the proportion of the number predicted by model in the actual customer churn number. The number of customers with high value and low value receptively is counted and the results are shown in Table 2 and Table 3.

The accuracy and coverage of predicting customer churn in this algorithm is calculated and shown in the tab.4.

TABLE 2. CONFUSION MATRIX OF CUSTOMERS WITH HIGH VALUE

high-value customer	Predicted churn customer	Predicted not churn customer
actual churn customer	78	27
actual not churn customer	19	119

TABLE 3. CONFUSION MATRIX OF CUSTOMERS WITH LOW VALUE

low-value customer	Predicted churn customer	Predicted not churn customer
actual churn customer	1006	224
actual not churn customer	206	1421

TABLE 4. PERFORMANCE EVALUATION FOR PREDICTING IN HIGH/LOW VALUE CUSTOMER CHURN

	accuracy	coverage
high-value customer	80.4%	74.3%
low-value customer	83.0%	81.8%

As can be seen, the low coverage of high-value customers is caused by many bumps in the data selected. The other three indicators are all higher than 80%, which indicate the high accuracy of improved FCM algorithm in predicting bank customer churn. By comparing the accuracy and coverage of customers with different clustering numbers, it shows that when the clustering number is 5, all the indicators are in the best state, that 5 is the optimal clustering number. This result is consistent with the experiment, as shown in Figure 2, which fully proved that the algorithm has high validity and has certain significance for bank managers.

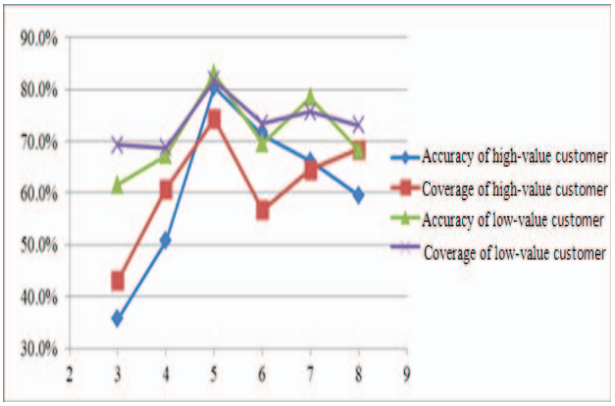

Figure 2. Comparison of accuracy and coverage in different clustering number

V. Conclusion and prospect

Big data has accelerated the research in data collection, storage, mining and application, among which, the improvement of data mining technology is vital to the entire era of big data. In this paper, an improved method based on FCM and reconstruct validity function on the use of compactness and separation was presented. The algorithm was applied to the bank customer churn prediction and verifies its feasibility and effectiveness, which provides new ideas for mining bank customer information and predicting customer churn. However, the validity of exploring clustering is a complex and difficult task. It needs to be further explored in future research and application process.

References

[1] Sato T, Huang B Q, Huang Y, et al. Using PCA to Predict Customer Loss in Telecommunication Dataset [M]. BERLIN: SPRINGER-VERLAG BERLIN, 2010: 6441, 326-335.

[2] Wojewnik P, Kaminski B, Zawisza M, et al. Social-Network Influence on Telecommunication Customer Attrition [M]. BERLIN: SPRINGER-VERLAG BERLIN, 2011: 6682, 64-73.

[3] Heitz C, Ruckstuhl A, Dettling M. Customer Lifetime Value under Complex Contract Structures [M]. BERLIN: SPRINGER-VERLAG BERLIN, 2010: 53, 276-281.

[4] Shi Yang, Yue Jiajia, Using data mining technique for bank customer churn with decision tree prediction algorithm [J]. Computer Knowledge and Technology, 2014 (11): 2533-2536.

[5]ZHU Zhi-yong, XU Chang-mei, LIU Zhi-bing, HU Chen-gang. Customers churn analysis based on bayesian network [J]. Computer Engineering & Science, 2013(03): 155-158. (in Chinese)

[6] YU Lu, Telecom customer churn integrated prediction model [J], Journal of Huaqiao University (Natural Science), 2016(05) :637-640. (in Chinese)

[7] FAN Chunmei, Forewarning modeling of mobile communication churn based on CRM [J], China Training, 2017(06):283-284 (in Chinese)

[8]YING Wei-yun, QIN Zheng, ZHAO Yu, LI Bing, LI Xiu. SVM method and its application in the prediction of customer churn [J]. Systems Engineering-Theory & Practice, 2007(07): 105-110. (in Chinese)

[9]LI Zhaoying, Science of customer management methods – how to reduce customer churn [J], China Management Informationization, 2016(01):109-110. (in Chinese)

[10]HE Qingzhe, XIA Guoen, Customer churn general research [J], Oriental Enterprise Culture, 2015(19):189. (in Chinse)

[11]WU Jingjing, Applying data mining technique in customer churn applied research [J], Ability and Wisdom, 2017(02):280. (in Chinese)

[12] Bezdek. Pattern recognition with fuzzy objective function algorithms [J]. Pattern recognition with fuzzy objective function algorithms, 1981.

[13] Dave R N. Validating fuzzy partitions obtained through c-shells clustering[J]. PATTERN RECOGNITION LETTERS, 1996, 17(6): 613-623.

[14] Zhang M R, Zhang W, Sicotte H, et al. A New Validity Measure for a Correlation-Based Fuzzy C-means Clustering Algorithm[M]. NEW YORK: IEEE, 2009, 3865-3868.