International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.9 (2014), pp.281-298
http://dx.doi.org/10.14257/ijmue.2014.9.9.30

Modeling and Application Research on Customer Churn Warning System Based in Big Data Era

Yihua Zhang, Yuan Wang*, Chunfang He and TingTing Yang

School of Business Administration, Jimei University, Xiamen 361021, China
*yward@jmu.edu.cn, *wangyuan@jmu.edu.cn, hcf051@163.com,*
1363762128@qq.com

Abstract

Customer churn warning prediction is one of the most important problems in customer relationship management (CRM). Its aim is to monitor customer loss conditions, explore for internal loss rules of mass transaction data, give early warning about customer loss inclination and retain valuable customers to maximize the profit of a company. Under the circumstance of rapidly emerging customer churn warning research under the big data background, we attempt to explore for mass security customer data and to focus on analysis and design on customer churn warning model based on data mining technologies and the theory of customer churn management process.

Keywords: *Big Data, Customer Churn Warning, Data Mining*

1. Introduction

For the concept of big data, no acknowledged precise definition has been reached among enterprises and the academic world at present. Although there are different statements, it is universally acknowledged that the concept of big data is consistent with that of mass data and large-scale data. However, big data exceeds traditional data forms in terms of data volume, data complexity and generation speed. Moreover, big data also exceeds processing capacity of existing technical means and brings about great opportunities of industrial innovation. How to make better decisions on the huge amount of data is an urgent task [1]. In the field of customer relation management, relevant practices and researches present some change trends with great significance with increasing emergence and all-around development of big data. In 2014 coming to past, China's economy was confronted with unfavorable external environment. However, China persisted in seeking improvement in stability and had realized improvement in stability. Various industrial markets suffered from challenges in different degrees. As an integral part of the financial market, the securities market certainly attracted more and more international security companies to occupy the market. This indicates that China's security companies are confronted with more and more fierce competitions from share broking and other businesses. Such competitions impose more and higher requirements for security companies. Hence, customer relation management under the large number background is gradually lifted to the core business level of various security companies. Based on real demands of customers, security companies devote themselves to satisfy demands of different customers. However, these companies are usually faced with another major difficulty while positioning customer demands, *i.e.*, customer churn.

According to American survey data, each company universally loses half of customers every five years. Such continuous and considerable customer churn takes place continuously. It is difficult for enterprises to make any response. Moreover, they hardly have time to retain

ISSN: 1975-0080 IJMUE

specific customers. For a security company, customer churn is just as friction to a mechanical system. Friction consumes capability of the mechanical system. Then, customer churn continuously wastes manpower, materials and financial resources of the enterprise. Simply from the perspective of financial resources, customer churn causes direct decrease of enterprise sales and profits. Moreover, the enterprise must increase relevant expenditures to obtain new customers. Relevant researches ever showed that a company's profits would increase by 25% - 85% if its customer churn was reduced by 5%. This is a fascinating figure. From a long-term perspective, continuous customer churn also gives a message of deteriorating values provided by the enterprise, resulting in extremely unfavorable impacts on the enterprise reputation.

The profit resulting from customer relationship is essential to ensure companies viability, so an improvement in customer retention is crucial for competitiveness. As such, companies have recognized the importance of customer centered strategies and consequently customer relationship management (CRM) is often at the core of their strategic plans [2]. Customer churn doesn't deny customer relations. Instead, it demonstrates again urgency and necessity of its implementation. . To predict whether a customer will be a churner or non-churner, there are a number of data mining techniques applied for churn prediction, such as artificial neural networks, decision trees, and support vector machines [3]. Therefore, the main commercial target of security customer churn is to monitor customer churn conditions, explore for internal loss rules of mass transaction data and give early warning about customer churn inclination. Causes for customer churn should be analyzed and relevant customers should be selectively retained, so as to reduce the customer churn rate. Thus, it is helpful for enterprises to find links necessary for improvement in their operation and management processes. As a result, lost customers can be even attracted back. Moreover, firmer customer relations will be established. Under the circumstance of rapidly emerging customer churn warning research under the big data background, we attempt to explore for mass security customer data and to lay emphasis on analysis and design on customer churn warning model based on data mining technologies and the theory of customer churn management process.

In recent years, considerable analysis is made onto mass data by numerous scholars and employees through data mining. Upon customer classification, customer value analysis, customer churn analysis and other customer analysis activities, such scholars and employees help securities traders to implement strategic management of customer relations in a better way and to improve their personalized service levels.

Su Qian, Shao PeiJi and Zou Tao (2010) analyzed the features of CRBT (Coloring Ring Back Tone), compared various data mining techniques that could assign a 'propensity to churn' to each CRBT subscriber through empirical evaluation. The results indicate d that the models could achieve satisfactory prediction effectiveness by using customer demographics, billing and service usage information [4]. Reza Allahyari Soeini and Keyvan Vahidy Rodpysh (2012) used the K-Means algorithm to classify security customers and then used decision-making trees respectively to establish customer loss warning model. After evaluation, Reza found that the model established with decision-making tree CART had relatively high precision [5]. Kerdprasop Nittaya1, Kongchai Phaichayon and Kerdprasop, Kittisak (2013) designed the framework of the proposed BI system to predict customer churn in the telecommunication industry. The logic-based implementation and performance testing results of the constraint-based pattern mining were also illustrated in their paper [6]. Tang Leilei, Thomas Lyn, Fletcher Mary, Pan Jiazhu and Marshall Andrew (2014) applied an orthogonal polynomial approximation analysis to derive unobservable information, which was then used as explanatory variables in a probit-hazard rate model. The results showed that derived information could help our understanding of customer attrition behavior and gave better predictions [7].

Based on considerable references, most scholars and employees are engaged in the research on customer churn warning and customer classification. Up to now, no research is found onto the analysis on customer churn warning by combining two models (i.e. customer churn warning and customer classification). Based on numerous researches made by scholars and employees, therefore, a new customer churn warning analysis thought is proposed in this paper. The design is of feature selection optimization and dual model (customer churn warning model and lost customer classification model) structure. An attempt is made to effectively find out customers with loss inclination. At the same time, lost customer categories are recognized by combining results of lost customer classification, so as to rapidly find out causes for loss of relevant customers and selectively retain those customers about to lose.

2. Background Knowledge

The interest for data mining techniques has increased tremendously during the past decades, and numerous classification techniques have been applied in a wide range of business applications [8].

2.1. C5.0 Algorithm

C5.0 algorithm develops on the basis of ID3 algorithm of decision-making tree. Since J.R.Quinlan proposes it in 1797, C4.5 algorithm with decision-making tree milestone significance forms through continuous improvement. C5.0 algorithm is commercial version of C4.5 algorithm. Its core is the same with C4.5 algorithm. It just improves in the aspect of execution efficiency and execution efficiency. One of the core problems of decision-making tree is to confirm branch norms of decision-making tree. C5.0 algorithm deems information gain as the standard to confirm the best grouping variable and segmentation point. The core concept is information entropy.

-P(U$_i$) which is the probability of occurrence of information $U_i(i=1,2,......,r)$ constitutes the mathematical model pf information source and

$$\sum P(U_i)=1 \qquad (i=1,2,...,r)$$

Units of the amount of information are bit, and its mathematical definition is as follows.

$$I(U_i)=\log_2\frac{1}{P(U_i)}=-\log_2 P(U_i) \tag{1}$$

Information entropy is mathematical expectation of information quantity and average uncertainty before information source sends messages. The definition of information entropy is as follows.

$$Ent(U)=\sum_i P(U_i)\log_2\frac{1}{P(U_i)}=-\sum_i P(U_i)\log_2 P(U_i) \tag{2}$$

When the concept distribution P (U) of information U is known and after the signal V=v$_j$ is received, the probability distribution of sent signal is shown in the following formula:

$$Ent(U|v_j)=\sum_i P(u_i|v_j)\log\frac{1}{{}_2P(u_i|v_j)}=-\sum_i P(u_i|v_j)\log_2 P(u_i|v_j) \tag{3}$$

As the receipt signal V is a random variable, expectations to the posterior entropy are shown the following formula:

$$Ent(U|V)=\sum_j P(v_j)\sum_i P(u_i|v_j)\log_2\frac{1}{P(u_i|v_j)}=-\sum_j P(v_j)\sum_i P(u_i|v_j)\log_2 P(u_i|v_j) \tag{4}$$

Conditional entropy or channel equivocation refers to uncertainty of information sink still existing for sent signal U after receiving all Vs. This is caused by random disturbance. In general, $Ent(U|V)<Ent(U)$.

$$Gains(U,V)=Ent(U)-Ent(U|V) \tag{5}$$

Information gain Gains(U, V) reflects random uncertainty of information elimination.

C5.0 algorithm regards information gain rate as the standard. In other words, it not just considers the size of information gain, but also takes into account of the cost paid in order to gain information gain. The definition of information gain rate is as follows.

$$GainsR(U,V)=Gains(U,V)/Ent(V) \tag{6}$$

2.2. K-Means Algorithm

Primary idea of K-means is taking mean(distance) of all samples points in sub-class as clustering centers, and through multiple iterations, separates data samples into different sub-classes, and makes clustering performance evaluation reaching optimization, finally makes sub-class internal compact and mutually independent among sub-classes. In algorithm bellow, when calculating distance between data samples, Euclidian distance algorithm is adopted [9,10].

The steps are as followed:

(1) Firstly select an initial clustering center for each sub-class, if there are K cluster centers;

(2) Distribute individuals in samples into the nearest subclass;

(3) Re-calculate mean values of each subclass, and take them as new cluster centers.

(4) Repeat step (2),(3), until lastly all the cluster centers do not move;

(5) Finally K cluster centers are obtained.

It can be found: (1) when initial cluster centers selected randomly are inappropriate, it will lead to low classification efficiency, then decrease the reliability of classification. (2) Each repeat requires re-calculate the distance of each data to each cluster center, if the amount of data is large, the iteration count is quite terrible.

Adopting traditional K-Means cluster analysis method may cause instability in cluster result, consequently decreases classification reliability. This thesis adopts improved K-Means cluster method, firstly searches data set N times, selects K initial cluster centers, then uses triangle scalene rule, decreases calculation amount of each iteration gradually, simplifies calculation and comparison procedure.

3. Data Pre-Processing

3.1. Data Description

The original data are from Xiamen securities companies in 2011, including 22151 pieces of information in the first quarter, 13792 pieces of information in the second quarter, 14783 pieces of information in the third quarter.

Before pre-processing the data, it is required to be familiar with the data, identify their quality problems, describe the data, generate data property report and understand the significance or calculation formula of each field in the original data. Important fields includes the account opening date, Customer status, the beginning market value, closing market value, the total commission, total amount of transactions, the average turnover, ending total assets, accumulate assets, amount of profit or loss, profit or loss rate, number of transactions, beginning total assets, ending total assets, turnover rate, total amount of commission, number of days of commission, times of commission and average traction amount.

3.2. Data Cleaning

3.2.1. Data Cleaning in the First Quarter of 2011: There were 22151 pieces of initial data in the first quarter of 2011. Firstly, delete 9620 pieces of repeated customer records, delete 2858 pieces of the records of the customers with account cancellation on the account cancellation date and fill all NULL values in numerical values to 0. Secondly, carry out deletion operation in allusion to overdue customer data with the account cancellation date earlier than the quarter. A total of 380 pieces of data are deleted. Finally, mark out quarter time through adding the field “transaction time”. The final data handling result is 9293 pieces.

3.2.2. Data Cleaning in the Second Quarter of 2011: There were 13792 pieces of initial data in the second quarter of 2011. Firstly, delete 5110 pieces of the records of the customers with account cancellation on the account cancellation date and fill all NULL values in numerical values to 0. Secondly, carry out deletion operation in allusion to overdue customer data with the account cancellation date earlier than the quarter. A total of 346 pieces of data are deleted. Finally, mark out quarter time through adding the field “transaction time”. The final data handling result is 8336 pieces.

3.2.3. Data Cleaning in the Third Quarter of 2011: There were 14783 pieces of initial data in the third quarter of 2011. Firstly, delete 5110 pieces of the records of the customers with account cancellation on the account cancellation date and fill all NULL values in numerical values to 0. Secondly, carry out deletion operation in allusion to overdue customer data with the account cancellation date earlier than the quarter. A total of 380 pieces of data are deleted. Finally, mark out quarter time through adding the field “transaction time”. The final data handling result is 9293 pieces.

3.3. Data Combination

Longitudinal merger of data means a process of adding records from multiple inputs. The data in each quarter collect the transactions of each customer in the quarter. Longitudinally merge data according to the data in the first 3 quarters in 2011 stored in 3 different excel tables. This brings convenience for overall statistics of transactions of the same customer in 3 quarters and for modeling.

3.4. Variable Management

Longitudinal merger of data means a process of adding records from multiple inputs. The data in each quarter collect the transactions of each customer in the quarter. Longitudinally merge data according to the data in the first 3 quarters in 2011 stored in 3 different excel tables. This brings convenience for overall statistics of transactions of the same customer in 3 quarters and for modeling.

3.4.1. Variable Description: Variable declaration defines, checks and modifies effectiveness of variable values in read-in data flow and meanwhile indicates the roles of each variable in future modeling. Set customer state to the output variable and dichotomy typ, set transaction time and customer number to multi-section type, set customer risk level to an ordered set; set other attributes to continuous values.

3.4.2. Generate Sample Set Segmentation Variable: For data mining modeling, the function of sample set segmentation is significant. In order to improve feasibility and operability of the model, in predictability mining process, a model often forms based on training set. The new model is used to classify new data or future data. In the end, verification sample set is adopted to carry out error statistics of the model.

Before customer churn warning model is established, divide the sample set to training set and test set at random according to 65% and 35% proportions. In model establishment process, the training set is mainly used for model selection. Different models are selected according to estimation accuracy of the model. The test set is used to further optimize the model. The specific parameters of the decision-making tree are set in accordance with accuracy of the test set. The error is directly estimated in the training sample set by use of statistics confidence interval estimation method in C5.0 algorithm. Thus, segmentation verification set used for model error calculation is cancelled.

3.5. Sample Management

3.5.1. Classification and Summary of Samples: The original data are 3 datasets of a security company from the first quarter to the third quarter in 2011. The same customer may have transaction records in the three quarters. After Longitudinal merger of data, it is required to collect the datasets. Customer number and customer state are selected as key words for summarization, while the remaining numeric data serve as summarization fields. Based on different requirements of field data, we select summation, average value, minimum, maximum and standard deviation mode for data summarization. For example, summarizing rule of transaction times is to synthesize transaction times of 3 quarters.

3.5.2. Conditional Filtering of Samples: In the process of setting up customer churn warning model, it is found that all transaction information is null apart from customer's basic information. Delete the customers without transaction record. During lost customer segmentation, it is necessary to extract lost customers in samples and set the node input condition to customer state=account cancellation.

3.5.3. Equalization treatment of Samples: The proportion of the quantity of lost customers is relatively small, relative to normal customers. It is found through statistics that the customers canceling their accounts only account for 1.66% of total sample size. Data mining is conducted on the samples with such proportion. Based on the principle of 0 model error rate, the model gained will be obviously partial to normal customers. Although such model has high accuracy for normal customers, it fails to effectively distinguish lost customers and existing customers, thus leading to objective errors for data modeling.

Select over-sampling method, *i.e.* sample more in rare samples(churners) and sample less in common samples(normal customers) so as to adjust the distribution proportion of the two kinds of samples and make then balanced. In modeling process, balance the customer sample set through sampling method in allusion to customer churn problem. There are 120 pieces of account cancellation records and 127 pieces of normal records.

3.6. Importance of Variables

The data mining involves enormous data size. 49 fields are left after filtering. Try to find out the variables and samples making great contributions to customer churn warning, keep them and get rid of those unimportant variables and samples. The objective of security customer churn warning analysis is to set customer state to output variable, set other variables to input variables and give the importance sequence of input variables relative to output variables. Input variables are numeric-type variables. Output variables are classification-type variables. Through variance analysis, set input variables to observational variables and set output variables to control variables to analyze whether the mean value of input variables under different output variables and classification levels has significant difference.

4. Model of Customer Churn Warning

We need to choose suitable model tools according to the analysis objective, set up the model through samples and evaluate the model. Certain logic relationship exists from customers' transaction situation to customer churn result. Through setting up churn warning model, it is necessary to find out the generality of churn problem and mine loss law so as to reach the purpose of predicting customer churn intention.

4.1. Set the Model Parameter

4.1.1. Interactive Verification: the sample data are divided into 10 equal data randomly on average, 10 models are established respectively, and the error is the average value of 10 models in the remaining 1/10 samples. The forecast result is the multi-model voting result. The 10 models in the decision-making tree model has an average forecast precision of 99.1, standard deviation of the forecast precision is 0.1.

4.1.2. Expert Model: Automatically adjust confidence degree and the minimum number of records of each sub-branch. Larger confidence degree means simpler decision-making tree. Within the 100%-25% default, try repeatedly at the interval of 5%. When the confidence degree is higher than 80%, the tress is relatively simple, but the prediction precision is low. After repeated debugging, 75% confidence degree is finally selected. Through gradually adding the minimum number of records of each sub-branch, the input field of the decision-making tree increases one when the number of records is 3 and the relative number of records is 2. Although the accuracy rate of the training set increases by 0.07%, the accuracy rate of the test set remains unchanged. When the number of testing records us 10, the precision of the decision-making tree is just like the situation when the number of records is 5. The decision-making tree constructed when the number of records is 4 and 5 is consistent with the situation when the minimum number of records is 2. Through overall consideration, the minimum number of records of each sub-branch is set to 5.

4.1.3. Misclassification Loss: Since this model takes large pseudo loss, the confidence degree of the model giving customer normal judgment is very high. It is relatively cautious. Compared with original value, when trueness abandonment loss is set to 1 and pseudo taking loss is set to 2, the effect will be better.

4.2. Data Flow

After data preprocessing of initial data, there are a total of 55 fields of inflow data flow and 7488 pieces of customer transaction records. The data flow of C5.0 model is generated.

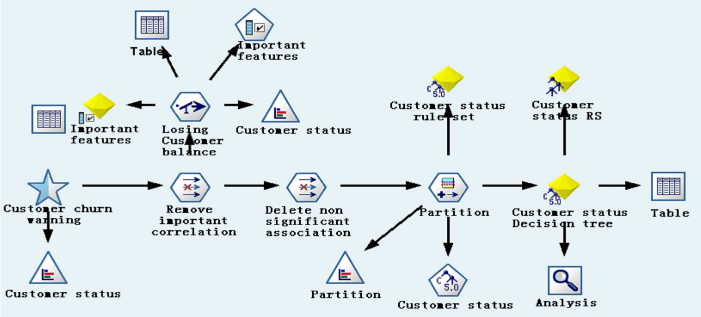


Figure 1. Data Flow of Customer Churn Warning Model

4.2.1. Delete Repeated Relevance Quantity: There are 4 fields for commission parts of Share A, including Share A commission, Share A transaction volume, Share A exchange fees and Share A net commission. In practice, security companies generally set Share A commission and Share A exchange fees to relevant linear quantity of Share A transaction volume. Here, we simply suppose they are percentage a and b. Share A commission=a*Share A transaction volume, Share A exchange fees=b* A transaction volume, Share A net commission= Share A commission- Share A exchange fees=(a-b)* Share A transaction volume. Through selecting the best correlated variables, measure relevant stability of data. Now that these variables present linear correlation, repeated relevance quantity can be deleted. Select to keep Share A net commission. Similarly, the 4 fields of warrant commission include warrant commission, warrant transaction volume, warrant exchange fees and warrant net commission. Select to keep warrant net commission. Here, 6 variables are filtered.

4.2.2. Delete Unimportant Features: Based on equalization treatment, use features to select the fields of the first 30 important transactions according to the importance sequence of field relevance. After adding the output field - customer state, there are a total of 31 fields and 7488 pieces of customer transaction records used for modeling.

4.3. Results

The model owns 7488 samples involved for analysis, but the samples of the training set approximate 65%. 4 841 samples are used to set up the model and the remaining samples serve as the test set. The analysis conclusion of the loss law is as follows.

Customer churn will happen and the confidence degree is 0.979 in the following conditions: the negative value of total amount of balance entry and market value transfer in the first 3 quarters is greater than-316.940(61 samples) and loss rate in the first 3 quarters is less than or equal to -0.496(110 samples); the negative value of total amount of bank transfer in the first 3 quarters is greater than -0.336 (47 samples).

Customer churn will happen and the confidence degree is 1.0 in the following conditions: the negative value of total amount of balance entry and market value transfer in the first 3 quarters is greater than-316.940(4780samples); the negative value of total amount of market value transfer in the first 3 quarters is greater than -47736(4751 samples) and the mean value of profit and loss rate in the first 3 quarters is less than or equal to -0.233(8 samples).

Customer churn will happen and the confidence degree is 1.0 in the following conditions: the negative value of total amount of balance entry and market value transfer in the first 3 quarters is greater than-316.940(4780samples); the negative value of total amount of market value transfer in the first 3 quarters is greater than -47736(4751 samples); the mean value of profit and loss rate in the first 3 quarters is less than or equal to -0.496(110 samples); the negative value of total amount of bank transfer in the first 3 quarters is greater than -383.500(11 samples); the negative value of total amount of bank entry in the first 3 quarters is less than 0.010(8 samples).

5. Segmentation Model of Customer Churn

The customers with potential loss tendency are mined through setting up the loss warning model. Why the customers are lost and how should the securities traders keep these customers? Then, we subdivide lost customers, deeply investigate the relations among basic customer features, customer behavior and customer churn according to different types of lost customers and analyze customer features and causes for customer churn. When potential lost customers are recognized, it is necessary to judge the cause fast according to lost customer segmentation and then carry out effective customer maintenance strategy.

5.1. Data Flow of Segmentation Model

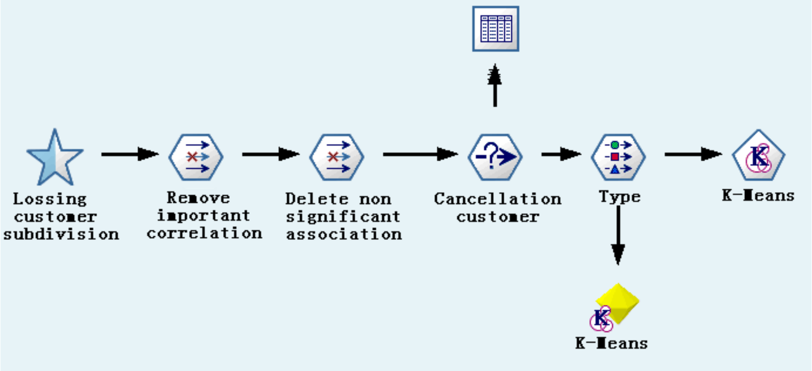

Lossing customer subdivision
Remove important correlation
Delete non significant association
Cancellation customer
Type
K-Means
K-Means

Figure 2. Data Flow of Segmentation Model

Consistent with C5.0, filter six repeated relevance fields, select the first 30 important relevance variables and finally import 31 fields in the data flow. Screen out the customers canceling the accounts, conduct clustering analysis with K-Means clustering algorithm and gain the subdivided model of lost customers (Figure 2). Through the established model, subdivide all customers, find out customer group with large loss tendency and put forward corresponding CRM strategy.

5.2. Results of Customer Churn Segmentation

The pie chart shows sample proportions of all kinds. The number of samples in the second kind is the minimum (1 sample). The number of samples in the first kind is the maximum (74 samples). The final row means whether the mean value of cluster variables has significant difference. Different icons shoe the degree of difference significance. Since various cluster variables are of numeric type, significance ranking still adopts F test method of variance analysis. According to significance ranking, find out features of these customers and subdivide lost customers, as shown in Figure 3 and Table 1.

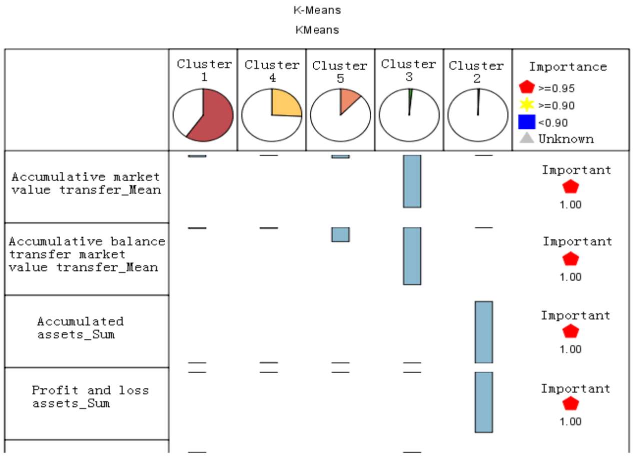


Figure 3. Results of Customer Churn Segmentation

Table 1. Important Indicator Comparison of Classification Results

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Average
AQAMVT	-13635.574	0.0	-411011.465	-3064.401	-22764.371	-90095.2
AQTABMVT	-8437.276	0.0	-449827.125	-7339.476	-112486.198	-115618
TAA	124962.269	29104862.52	58165.083	291246.619	171731.28	5950194
TPALA	-5359.301	-28464042.99	-58165.082	-62170.12	-1728.597	-5718293
Age	60.905	44.0	61.0	33.531	53.6	50.6072
BEATTMV	-40022.574	0.0	-504206.498	-46031.381	-275390.114	-173130
TMVT	-44670.912	0.0	-469449.433	-23060.812	-34558.513	-114348
TBT	-157806.407	-443742.59	-130779.28	-1048320.948	-16902.313	-359510
TF	23.0	27.0	2.5	18.406	14.867	17.1546
TTC	617.467	599.75	164.575	1031.814	391.457	561.0126

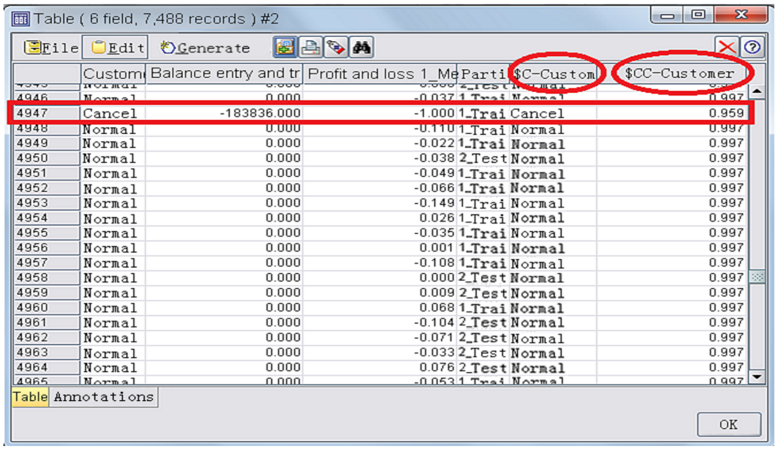
In the Table 1, AQAMVT represents average quarter time accumulative market value transfer. AQTABMVT represents average quarter time accumulative balance market value transfer. TAA represents total accumulative assets. TPALA represents total profit and loss assets. BEATTMV represents balance entry and transfer total market value. TMV represents total market value. TMVT represents total market value transfer. TBT represents Total bank transfer. TF represents transaction frequency. TTC represents the total commission.

6. Model Evaluation

Evaluate whether the model mined is rational and suitable for realistic problems. So, we need evaluate on the basis of training set mining model. In the modeling process, the process from modeling to pruning is based on the training set. Precision evaluation on the test set will impose direct influence on successful modeling. Thus, it is especially important to observe the effects of prediction results. During optimizing the prediction result, we need analyze the rationality of the generated result, review the whole data mining process, consider whether each link is treated rigorously, fine adjust data flow or change parameter setting of decision-making tree for model optimization.

6.1. Model of Customer Churn Warning

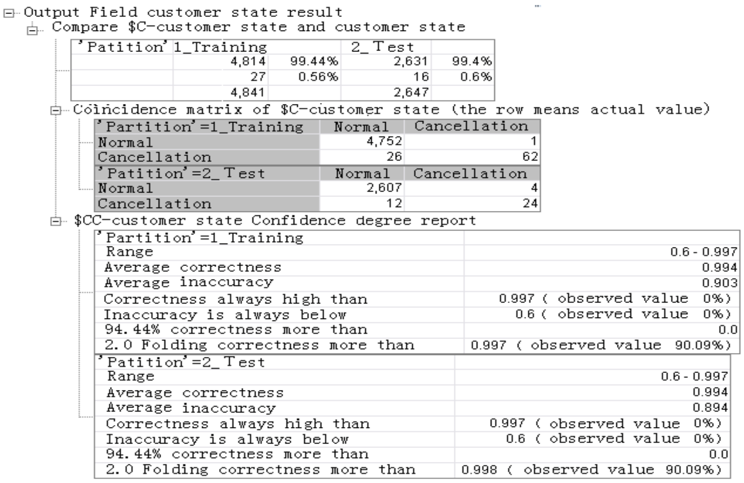
6.1.1. Predicted Results: In the prediction result table, the variables starting with character strings $C and $CC mean prediction classification value and prediction confidence degree of each sample and display the prediction situation of daily records. It is found through calculations that if the negative value of total amount of balance entry and market value transfer in the first 3 quarters is less than or equal to -316.940 and the mean value of profit and loss rate in the first 3 quarters is less than or equal to -0.336, customer churn appears and the confidence degree is 0.979. Correspondingly, 47 samples are included in nodes of the decision-making tree. 46 samples are predicted correctly. Laplace adjustment result of rule confidence degree is (46+1)/(47+2)=0.959. This value is the confidence degree of corresponding samples, as shown in Figure 4.



	Custom	Balance entry and tr	Profit and loss 1_Me	Parti	$C-Custom	$CC-Customer
4946	Normal	0.000	-0.037	1_Trai	Normal	0.997
4947	Cancel	-183836.000	-1.000	1_Trai	Cancel	0.959
4948	Normal	0.000	-0.110	1_Trai	Normal	0.997
4949	Normal	0.000	-0.022	1_Trai	Normal	0.997
4950	Normal	0.000	-0.038	2_Test	Normal	0.997
4951	Normal	0.000	-0.049	1_Trai	Normal	0.997
4952	Normal	0.000	-0.066	1_Trai	Normal	0.997
4953	Normal	0.000	-0.149	1_Trai	Normal	0.997
4954	Normal	0.000	0.026	1_Trai	Normal	0.997
4955	Normal	0.000	-0.035	1_Trai	Normal	0.997
4956	Normal	0.000	0.001	1_Trai	Normal	0.997
4957	Normal	0.000	-0.108	1_Trai	Normal	0.997
4958	Normal	0.000	0.000	2_Test	Normal	0.997
4959	Normal	0.000	0.009	2_Test	Normal	0.997
4960	Normal	0.000	0.068	1_Trai	Normal	0.997
4961	Normal	0.000	-0.104	2_Test	Normal	0.997
4962	Normal	0.000	-0.071	2_Test	Normal	0.997
4963	Normal	0.000	-0.033	2_Test	Normal	0.997
4964	Normal	0.000	0.076	2_Test	Normal	0.997
4965	Normal	0.000	-0.053	1_Trai	Normal	0.997

Figure 4. Predicted Results

6.1.2. Evaluation Results: The predicted value and confidence degree of the training set and the test set samples will be calculated according to analysis nodes and exhibited through comparison as an evaluation result to visually embody model precision and predictability, as shown in Figure 5.



Output Field customer state result

Compare $C-customer state and customer state

'Patition'	1_Training		2_Test	
	4,814	99.44%	2,631	99.4%
	27	0.56%	16	0.6%
	4,841		2,647	

Coincidence matrix of $C-customer state (the row means actual value)

'Partition'=1_Training	Normal	Cancellation
Normal	4,752	1
Cancellation	26	62
'Patition'=2_Test	Normal	Cancellation
Normal	2,607	4
Cancellation	12	24

$CC-customer state Confidence degree report

'Partition'=1_Training	
Range	0.6 - 0.997
Average correctness	0.994
Average inaccuracy	0.903
Correctness always high than	0.997 (observed value 0%)
Inaccuracy is always below	0.6 (observed value 0%)
94.44% correctness more than	0.0
2.0 Folding correctness more than	0.997 (observed value 90.09%)

'Patition'=2_Test	
Range	0.6 - 0.997
Average correctness	0.994
Average inaccuracy	0.894
Correctness always high than	0.997 (observed value 0%)
Inaccuracy is always below	0.6 (observed value 0%)
94.44% correctness more than	0.0
2.0 Folding correctness more than	0.998 (observed value 90.09%)

Figure 5. Evaluation of C5.0 Model

(1) **Compare $C-customer State and Customer State:** Overall accuracy rate and error rate of the model in the training sample set and test sample set are respectively displayed. Here, the training sample set includes 4 841 samples, where actual value and predicted value of 4814 samples coincide. The accuracy rate is 99.44%. 27 samples are predicted wrongly. The error rate is 0.56%. The test sample set contains 2641 samples. The prediction accuracy rate of the model is 99.4%, down compared with the training sample set. 16 samples are predicted wrongly. The error rate is 0.6%.

(2) **Coincidence Matrix of $C-customer State** (the row means actual value): Confusion matrixes of the training sample set and test sample set are respectively shown here. In the test sample set, there are 24 samples where the actual value of the variables is lost customers and the model prediction is also the lost customers. There are 4 samples where the actual value is normal customers and the predicted value is lost customers. Others are in the same way. In allusion to customer churn warning, evaluate the model with customer churn evaluation matrix, calculate the prediction accuracy, hit rate, coverage and upgrade degree and evaluate the model property, as shown in Table2. Prediction accuracy refers to the proportion of lost customers and non-lost customers predicted. Prediction hit rate means the proportion of actual lost customers to the predicted lost customers. It is an index to describe model accuracy. Prediction coverage refers to the proportion of predicted lost customers to actual lost customers. It is an index to describe model universality.

Table 2. Customer Churn Evaluation Matrix-1

	Normal Customers Predicted	Customers with Account Cancellation Predicted	Total
Actual Normal Customers	TP	FP	TP+FP
Actual Customers with Account Cancellation	FN	TN	FN+TN
Total	TP+FN	FP+TN	TP+FP+FN+TN

In the Table, TP represents predicted normal customers, *i.e.*, actual non-lost customers. FN represents predicted non-lost customers, *i.e.*, actual lost customers. FP represents predicted lost customers, *i.e.*, actual non-lost customers. TN represents predicted lost customers, *i.e.*, actual lost customers.

The prediction accuracy=(TP+TN)/(TP+FP+FN+TN)

Hit rate=TN/(TN+FP). Coverage=TN/(TN+FN)

Upgrade degree= hit rate*(TP+FP+FN+TN)/(FN+TN)

Through evaluating 2647 pieces of test data and customer churn warning model, we gain the following customer churn evaluation matrix, as shown in Table3. The accuracy of model is 99.39%; its hit rate is 85.71%; its coverage is 66.67%; it upgrade degree is 63.02. The results show good warning effects.

Table 3. Customer Churn Evaluation Matrix-2

	Normal Customers Predicted	Customers with Account Cancellation Predicted	Total
Actual Normal Customers	2607	4	2611
Actual Customers with Account Cancellation	12	24	36
Total	2619	28	2647

(3) **Confidence Degree Report of $CC-customer State**: The report shows assessment results of prediction confidence degree of the training sample set and test sample set. In the test sample set, the minimum confidence degree of sample prediction is 0.6 and the maximum is 0.997. For correctly predicted samples, the mean value of prediction confidence degree is 0.994. For wrongly predicted samples, the mean value of prediction confidence degree is 0.813. The correctness is always above 0.997, which means the samples with prediction confidence degree above 0.997 are usually normal. Since 0.997 is the maximum of prediction confidence degree, only 0% samples comply. Similarly, inaccuracy is always below 0.6, which means the samples with prediction confidence degree below 0.6 are usually wrong. At least 99.4% accuracy 0.0 means: among the samples with prediction confidence degree above 0.0, 99.4% samples are predicted correctly. But since it is very hard to find an accurate value, it approximates 0.0. Folding correctness 0.998(90% of observed value) above 2.0 means, Clementine gives the samples with prediction confidence degree above 0.998 (including 90.09% samples), prediction accuracy is twice of all samples. In general, the accuracy of the model in the trading set and test set is high. In repeated tests, the confidence degree of the model is also high. So, its prediction ability is excellent.

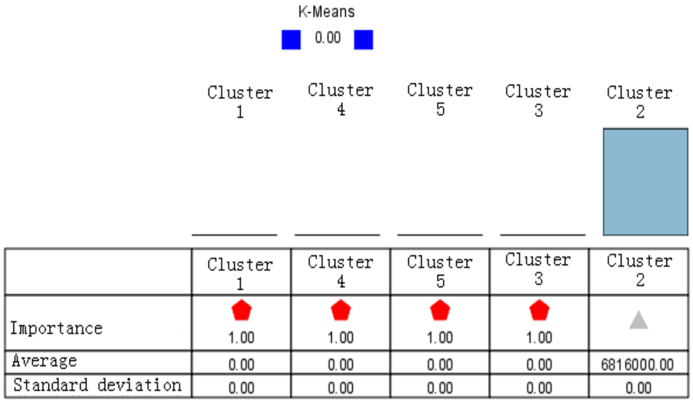
6.2. Model of Customer Churn Segmentation

6.2.1. Iteration Results: The following figure displays iteration of each step in the clustering process. A total of 4 iterations happen here. Relative to initial class center, the first iteration makes 5 class centers deviate and the largest offset value is 0.256. The second iteration also deviates and the largest offset value is 0.037. The largest offset value of the third iteration is 0.021. Almost no offset happens in the fourth iteration. After the iteration is over, the class center has been basically stable.

6.2.2. Reason of Customer Churn: Deeply investigate the relations among basic customer features, customer behavior and customer churn according to lost customer segmentation model. According to different situations of customer churn, customers are classified into the following five types: sudden account cancellation, natural loss, instability, heavy loss and small loss. The five types of customers have obvious behavior characteristics.

Cluster 1 customers belong to the type of sudden account cancellation. Such customers belong to the customer group with advanced age. Their transactions are frequent. Capital inflow and outflow are in the normal level. Besides, the loss amount is low. The commission they provide is also in the medium level. Such customers have no obvious loss tendency. Thus, it can be predicted that they will continue to carry out security trading activities in the future. If other security companies attract such customers with various favorable conditions, temporary account cancellation of such customers may be caused.

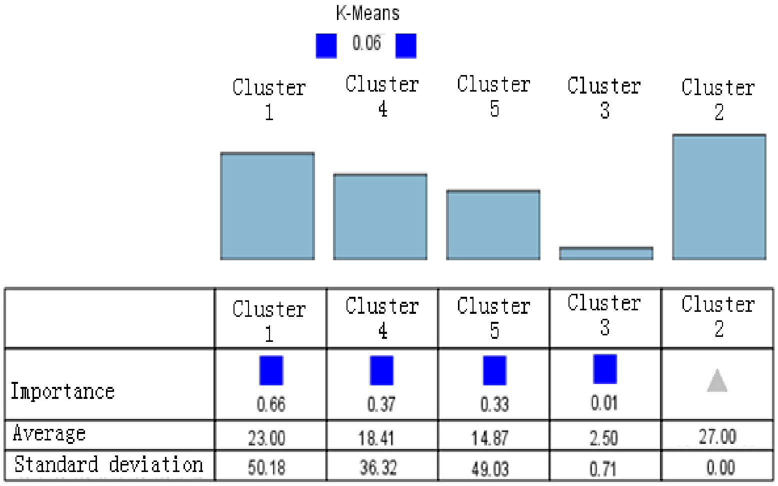
Cluster 2 customers belong to the type of heavy loss. Based on the important indexes in the Table1, the loss amount of such customers is much greater than that of other customers. Through carefully searching features of such customers, we can discover that although the transaction situation at Shanghai Exchange is almost blank, they transact at Shenzhen Exchange very frequently. This indicates such customers may prefer to invest small and medium-sized enterprises and stock market of growth enterprises, as shown in Figure 6. There are large quantities of market value inflow, but there is almost no market value outflow. Such conclusion can be drawn that such customers cancel their accounts due to heavy loss caused by operation errors.



	Cluster 1	Cluster 4	Cluster 5	Cluster 3	Cluster 2
Importance	⬟ 1.00	⬟ 1.00	⬟ 1.00	⬟ 1.00	▲
Average	0.00	0.00	0.00	0.00	6816000.00
Standard deviation	0.00	0.00	0.00	0.00	0.00

Figure 6. Result of Average Quarter Time Accumulative Market Value To

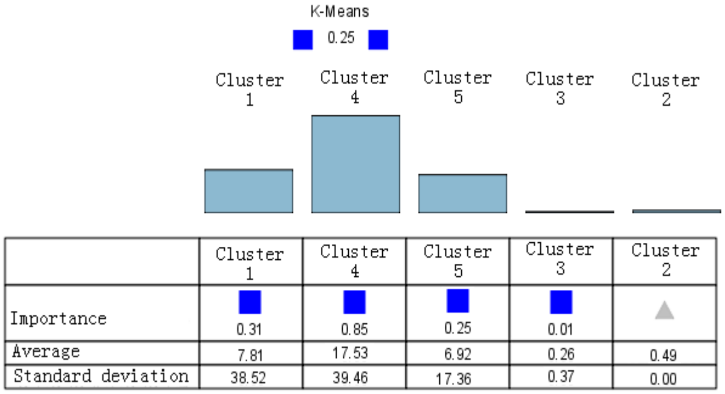
Cluster 3 customers belong to the type of natural loss. The average transaction volume of such customers in 3 quarters is only 2.5. Besides, the standard deviation value is small, as shown in Figure7 and Table1. Almost no transaction happens. Meanwhile, sudden high total market value of balance inflow and outflow means such customers undersell stocks in quantity and have obvious tendency of account cancellation. Such customers may be dissatisfied with the services of the security company and will be lost very easily.



	Cluster 1	Cluster 4	Cluster 5	Cluster 3	Cluster 2
Importance	■ 0.66	■ 0.37	■ 0.33	■ 0.01	▲
Average	23.00	18.41	14.87	2.50	27.00
Standard deviation	50.18	36.32	49.03	0.71	0.00

Figure 7. Result of Total Transaction Frequency

Cluster 4 customers belong to the type of instability. Such customers are relatively young and show continuous and stable transactions. The largest feature is that bank inflow amount is large. They subscribe new stocks in quantity. The turnover rate is relatively high. Meanwhile, there are numerous cancellation records, as shown in Figure8. Such customers are very willing to participate in security trading, but their transaction objective may be not too clear. They show hesitation during selection transaction process. Thus, relative loss amount is high.



	Cluster 1	Cluster 4	Cluster 5	Cluster 3	Cluster 2
Importance	■ 0.31	■ 0.85	■ 0.25	■ 0.01	▲
Average	7.81	17.53	6.92	0.26	0.49
Standard deviation	38.52	39.46	17.36	0.37	0.00

Figure 8. Result of the Quarterly Average Exchange Rate

Cluster 5: customers belong to the type of small loss. Such customers have the smallest customers belong to the type of loss amount. Their transaction situations are stable. Their transaction situations, turnover rate and commission contribution situations are in the medium level it can be inferred that such customers cancel their accounts mainly out of their own reasons, such as being busy in their work or having no interest in security investment.

6.2.3. Strategies of CRM: Corresponding customer maintenance strategies are proposed through predicting the customers who will be about to be lost, setting up lost customer segmentation model according to features of lost customers and main causes for customer churn and special conditions of different types of customers.

Customers in Cluster 1 are the customer group with advanced age. The causes for customer loss may include the following: dissatisfaction with the services, interpersonal relationship and commission competition etc. For the competitions of other security traders, it is required not only to display brand advantages to customers, but also to implement effective favorable marketing according to customer's needs. Effective profit point of security companies is to maintain customers' security, meanwhile help customers choose effective investment modes for investment operation, attract a large number of customers through high-quality services and stable operation and make the company and customers achieve win-win. Therefore, firstly, it is required to establish brand advantage, display steady business conditions of the company to customers and promote win-win with "customer-entered" service idea. It is necessary to analyze customers' demands and meet their needs. For example, in the face of commission competition, Guotai Junan Securities Co., Ltd. adopts flexible solutions for commission contrast calculation. For example, if the transaction volume in one year is 100000 and the commission is 0.25% or 0.2%, commission competition balance is 50. Or bundle commission method is used. Since customers' commission requirement cannot reach the existing standard, some standard privileges can be provided such as introducing new customers and transaction mode bundle.

The customers in Cluster 2 have a set of unique logics for their capital operation and may have fixed investment preference, but they fail to seize the investment opportunities and often miss profit opportunities. Such customers transfer large quantities of funds and take active part in securities transactions, but they suffer heavy loss. They are the customers deserving maintenance. Aiming at such situation, security traders should take quick reactions and answer doubts of customers to reduce customer churn risk. Such customers can be invited to participate in some regular investment solons or product mix can be recommended to such customers in allusion to their investment preference.

The customers in Cluster 3 are the customers without transaction wish attracted by large quantities of expanded business. The maintenance cost is large and the benefit gained is low. Too many marketing measures are not needed. The SMS or email can be adopted to provide investment portfolio information for customers. If customers are lost due to dissatisfaction with security traders, security traders need to know situations, clarify some facts to customers, explain the causes, meanwhile provide solutions with considerate service attitude and persuade customers to change the original intention. This is not just responsible for customers, but also beneficial to continuous improvement of security traders.

The customers in Cluster 4 are willing to participate in security transactions. They may not know security market very well and be not clear about transaction targets. They hesitate to select transaction products. Thus, relative loss amount is high. Aiming at such customers, such marketing modes as security investment suggestions and product mix recommendation should be taken. During daily transaction period, SMS or email can be adopted to enrich customers' security knowledge and investment information. High-frequency and high-quality investment solons can be organized regularly. The best analyzers can be invited on the site to analyze the whole investment. Consultation with senior consultants can be provided to answer customers' doubts so as to reduce customer churn risk.

It is difficult to predict the loss of cluster customers. The loss causes are wide. So, it is hard to formulate maintenance strategy. The loss causes can be investigated for the customers with large commission contribution.

7. Conclusions

This topic adopts feature selection optimization and dual model (customer churn warning model and lost customer segmentation model) structure. Mining the customers with potential loss tendency through setting up loss warning model can contribute to security traders finding lost customers. On this basis, combining the result of lost customer segmentation model and rapidly knowing customer churn causes in allusion to types of lost customers can contribute to enterprises to carry out targeted marketing, reduce customer churn rate and decrease loss. It is necessary to establish customer churn warning model, find out the generality of loss problem and mine the loss law to reach the purpose of predicting customer churn tendency. Through setting up lost customer segmentation model, we deeply investigate the relations among basic customer features, customer behavior and customer churn according to different types of lost customers and analyze customer features and causes for customer churn. When potential lost customers are recognized, lost customer segmentation can be rapidly utilized to judge the loss causes and carry out effective customer maintenance strategies.

Due to customer data confidentiality, customers' actual dynamic data cannot be gained. During studying customer churn warning, there is lack of estimation of the trend of transaction changes with time. Customer churn warning model has certain fundamentality function in customer churn management theory. Much work in the fields of theoretical basis and practical application is worth doing.

References

[1] X. Li, H. Zhang, Z. Zhu, Z. Xiang, Z. Chen and Y. Shi, "An Intelligent Transformation Knowledge Mining Method Based on Extenics", Journal of Internet Technology, vol. 14, no. 2, (**2013**), pp. 315-325.

[2] V. L. Migueis, A. Camanho and F. e Cunha Joao, "Customer attrition in retailing: An application of Multivariate Adaptive Regression Splines", Expert System With Applications, vol. 40, no. 16, (**2013**), pp. 6225-6232.

[3] C.-F. Tsai and Y.-H. Lu, "Data mining techniques in customer churn prediction", Recent Patents on Computer Science, vol. 3, no. 1, (**2010**), pp. 28-32.

[4] Q. Su, P. Shao and T. Zou, "CRBT customer churn prediction: Can data mining techniques work?", International Journal of Networking and Virtual Organisations, vol. 7, no. 4, (**2010**), pp. 353-365.

[5] R. A. Soeini and K. Vahidy, "Rodpysh Evaluations of Data Mining Methods in Order to Provide the Optimum Method for Customer Churn Prediction, Case Study Insurance Industry, vol. 24, (**2012)**, pp. 290–297.

[6] K. Nittaya1, K. Phaichayon and K. Kittisak, "Constraint mining in business intelligence: A case study of customer churn prediction.", International Journal of Multimedia and Ubiquitous Engineering, vol. 8, no. 3, (2013), pp. 11-20.

[7] T. Leilei, T. Lyn, F. Mary, P. Jiazhu and M. Andrew, "Assessing the impact of derived behavior information on customer attrition in the financial service industry", European Journal of Operational Research, vol. 236, no. 2, (**2014**), pp. 624-633.

[8] V. Thomas, V. Woute and B. Bart, "A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer ChurnPrediction Models", IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 5, (**2013**), pp. 961-973.

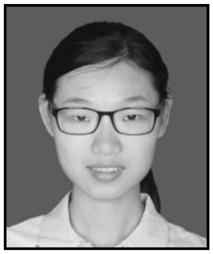
[9] J. Tang, "Improved K-means Clustering Algorithm Based on User Tag", Journal of Convergence Information Technology (JCIT), vol. 5, no. 10, (**2010**), pp. 124-130.

[10] Z. Yong and S. H. Haibin, "Adaptive K-means clustering for Color Image Segmentation", Advances in information Sciences and Service Sciences (AISS) , vol. 3, no. 10, (**2011**), pp. 216-223.

Authors

Yihua Zhang, he is an associate professor at the School of Business Administration, Jimei University. He received his bachelor's degree in Economic Information Management (1994) from Southwestern University of Finance and Economics and M.S.E. in Software Engineering (2007) from Huazhong University of Science and Technology. His current research interest includes Data Mining, Big Data, System Engineering and Business Intelligence.

Yuan Wang, she received her M.S.E. in System Engineering (2005) and PhD in System Engineering (2013) from Xiamen University. Now she is full researcher of School of Business Administration, Jimei University. Her current research interest includes Data Mining, System Engineering, Business Intelligence and Theory and Technology of Decision-Making.



Chunfang He, she is an undergraduate student who major in E-commerce at Jimei University. Her current research interest includes Data Mining, Data Warehouse, System Modeling and Business Intelligence.

TingTing Yang, she is an undergraduate student who major in E-commerce at Jimei University. Her current research interest includes Data Mining, System Analysis and Design, Software Engineering and Network Programming.