International Journal of u- and e- Service, Science and Technology
Vol.8, No. 10 (2015), pp.31-40
http://dx.doi.org/10.14257/ijunesst.2015.8.10.04

Customer Churn Warning System based on Business Intelligence

Yuan Wang[a] and Yihua Zhang[b*]

[1,2]*Jimei University, Xiamen 361021, China*
Email:[a]*wangyuan@jmu.edu.cn,* [b]*yward@jmu.edu.cn*

Abstract

Business intelligence approach is adopted for data mining analysis on the customer transaction data and basic customer information of securities companies. Decision tree algorithm is used to create the customer loss warning model, which is then used to analyze and design the securities data mining system. This system based on Weka source code, realized by JSP language and takes the classification analysis in the business intelligent core technology as the core algorithm. Analyzing this system can assist to mine the customers with potential loss trend and help establish specific detainment strategy to prevent great loss due to customer loss.

Keywords: *Business Intelligence, Customer Loss, Decision tree*

1. Introduction

Under the context of aggravated market competition, cost strategy has become the first choice of brokers. However, the compression of cost can only solve the short-term problems. In the long and sustainable terms, it is required to control the loss of customers and develop new customers by starting from the customer relation management. For a securities company, the loss of customers is like the action of fraction system to a mechanical system. Like the fraction force consumes the capacity of mechanical system, loss of customers continuously consumes the enterprise's manpower, material and finance. Loss of customers does not mean to deny the customer relation, but is the reproof to the urgency and necessity to implement it. Some researches have shown that if a company can reduce its customer loss rate by 5%, its profit would increase by 25% to 85%, and this is a very considerable income. In the long run, the continuous customer relation also conveys the signal of worsened enterprise provision value and will have an extremely adverse influence on the enterprise's reputation. Therefore, enterprises can monitor the condition of customer loss, obtain the loss rules from masses of transaction data, warn the customer loss trend and detain the corresponding customers selectively so as to reduce the customer loss rate, and in this way, enterprises can find the links urgent to be improved in their operation and management, and even attract the lost customers again and establish a stronger customer relation [1].

In 2011, Xie Fang employed RFM-ROI method to analyze and forecast customer loss, warn the potential customer loss and find out the customer groups with a great trend for targeted public relation in combination with the rules of loss warning on this basis [2]. 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 [3]. 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 [4]. Tang Leilei, Thomas Lyn, Fletcher Mary, Pan Jiazhu and Marshall Andrew (2014) applied an

ISSN: 2005-4246 IJUNESST

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 [5].

Through the review of a lot of literatures, most predecessors researched the customer loss warning and customer segmentation based on the data mining tool. Based on the business intelligence technology, a loss warning model of securities customers is established to forecast the trend of customer loss, and analyze, design and realize the analysis data mining system of B/S model securities customers. Functions such as property selection, data preprocessing, data mining, model evaluation and model storage management etc. is realized in this system.

2. Background Knowledge

2.1. Decision Tree

Decision tree is popular and powerful for both classification and prediction [6]. A decision tree is a classifier which conducts recursive partition over the instance space [7-8]. The attractiveness of tree-based methods is due largely to the fact that decision tree represent rules (Berry and Linoff, 2004). A decision tree is based on the methodology of tree graphs and can be considered one of the more simple inductive study methods (Quinlan, 1986, 1993; Russell and Norving, 1995). Even if the user lacks any statistical knowledge, he or she can use a decision tree to analyze specific behavior and it can be converted into rules easily. However, if it becomes too complicated or too huge for decision-making, trimming some of its leaves or branches may become necessary in order to improve its effectiveness. Of all the calculative methods, ID3, C4.5 (Quinlan, 1993; Cheng, *et al.*, 1998), CART (Breiman, *et al.*, 1984) and CHAID (Magidson and Vermunt, 2004) are the most well known [9].

2.2. Main Algorithm of ID3

The steps of main algorithm are shown as followings:

Step1: To randomly select a subset with positive samples and negative samples from training set;

Step2: Use tree building algorithm to form a decision tree from present window;

Step3: Samples from training set (except window) are applied to perform classification judgment by acquired decision tree to find out samples of misjudgment.

Step4: If there have misjudging samples, insert them into windows to turn to Step2 or go to end [10].

2.3. C4.5

C4.5 algorithm is an extension of ID3 (interactive dichotomizer) algorithm and the divide-and-conquer approach (Quinlan, 1993; Winston, 1992) which main improvements included handling of continuous attributes, dealing training data with missing attribute values and a process for pruning a built tree [11-12].

The splitting node strategy is based on the computation of the information gain ratio. The basic idea is that each node should hold a question concerning the attribute which is the most informative amongst the set of attributes not yet considered in the path from the root to that node. Information value also called entropy measures how informative is the association of an attribute with a node (Gray, 1990). The notion of gain ratio (Quinlan, 1993) is useful to rank attributes.

The decision tree induction purpose is the classical over-fitting problem can be addressed via pruning strategies [13].

3. Modeling Process

3.1. Data Understanding

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 *etc*.

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. The field "online time" is added in the data, with year as the unit, and the calculation is "current year-account opening date", and all records with empty transaction information are deleted except the basic customer information.

Finally, mark out quarter time through adding the field “transaction time” , the field "age" and "online time" is added in the data, with year as the unit, and the calculation is "current year-account opening date", and all records with empty transaction information are deleted except the basic customer information. The final data handling result is 5730 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 field "age" and "online time" is added in the data, with year as the unit, and the calculation is "current year-account opening date", and all records with empty transaction information are deleted except the basic customer information. The final data handling result is 6485 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 field "age" and "online time" is added in the data, with year as the unit, and the calculation is "current year-account opening date", and all records with empty transaction information are deleted except the basic customer information. The final data handling result is 6887 pieces.

3.3. Modeling

3.3.1. Data Source: In the process of modeling, the data in the first quarter are taken as the training set of modeling, and those in the second and third quarters are taken as the test set.

3.3.2. The Characteristic Selection Indicators: The characteristics influencing the loss of customers in securities industry include the following: (1) Basic customer property, such as gender, age, bank opening time, occupation, hobby, native place and title *etc*., these data are obtained when the customers open account and are permanently stored in the customer database. Customers of different backgrounds have different social behavioral characteristics and hobbies, for example, occupation influences income and age influences the type of securities purchased. (2) Customer transaction condition: such as monthly (weekly) average transaction volume, customer fund yield rate, market price of stock, fund balance, handling rate and customer value type, the customers' detailed transaction information can be obtained from their historical transaction library.

After the securities customers are canceled, all their ending balance, ending funds, ending market value and ending total assets are 0, so these properties are not taken as the characteristic properties of customer loss. The characteristic selection indicators finally determined include: customer status, initial guarantee, initial fund, transaction frequency, total commission, total transaction volume, average transaction amount per time, market value transfer-in, time increment market value transfer-in, market value transfer-out, time increment balance transfer-out market value, Balance transferred-in value, Accumulative balance transferred-in value, Balance transferred-out value, Accumulative balance transferred-out value, bank transfer-in, bank transfer-out, time increment bank transfer-in, time increment bank transfer-out, initial total assets, accumulated assets, time increment asset profit and loss amount, profit and loss rate, times of subscription for new shares, amount of subscription for new shares, entrusted amount of online transaction, telephone entrusted amount, onsite entrusted amount, time of entrusting, time of dealing, time of order withdrawal, entrusting days, share A commission, share A transaction volume, fees of share A exchange, net commission of share A, warrant commission, warranty transaction volume, warrant exchange fee, warrant net commission, fund acquired amount, fund subscribed amount, redeemed amount, turnover rate, age and online time.

3.3.3. Parameters Configuration: In the process of modeling the decision-making tree, the specific parameters are set as follows. Customer status is selected as the dependent variable, which has only property values, i.e. normal customer and cancelation. The characteristic selection indicator is selected as the independent variable, including total commission, initial fund, total commission, average transaction amount per time, market value transfer-in, time increment market value transfer-in, market value transfer-out etc. Parameters configuration are presented in Figure 1.

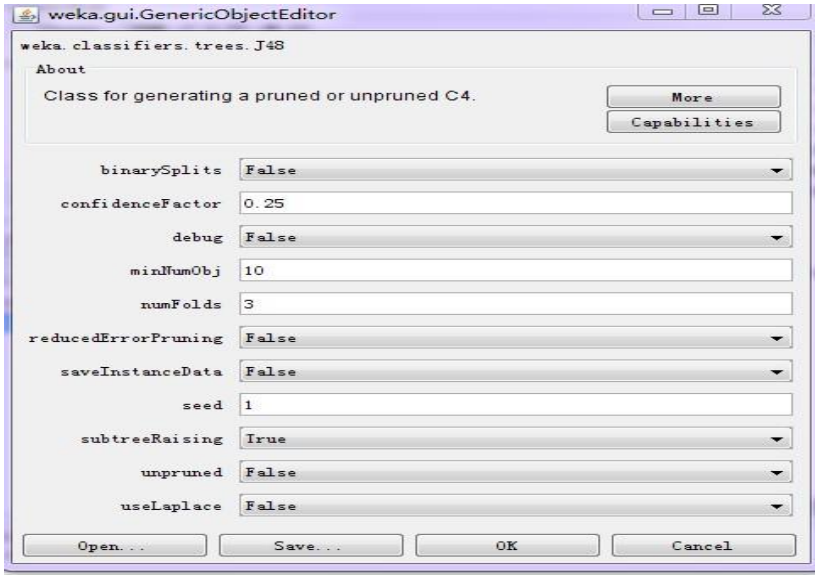

Figure 1. Parameters Configuration

3.3.4. The Results of Modeling: The results of C4.5 modeling is presented in Table1 and Figure 2.

Table 1. Results Table of Decision Tree

```
Scheme:weka.classifiers.trees.J48 -C 0.75 -M 10
Relation:QueryResult-weka.filters.unsupervised.attribute.Remove-R1-3,5-6,10-12,30,58
Instances:   5730
Attributes:   50
J48 pruned tree
Balance transferred-out value <= -1316: cancelation (40.0/8.0)
Balance transferred-out value > -1316
|   accumulated assets <= 0
|   |   profit and loss amount <= 0.09: cancelation (19.0/6.0)
|   |   profit and loss amount > 0.09: normal (74.0/4.0)
|   accumulated assets > 0: normal (5597.0/5.0)
Number of Leaves  : 4
Size of the tree :      7
```

Through the result of decision-making tree model of customer loss, the following business rules can be found.

(1)If the balance entry transfer-out market value <= -1316, then the cancelation confidence is 80%.

(2) If the balance entry transfer-out market value >-1316 and accumulated assets>0, then normal customer confidence is 99.9%.

(3) If the balance entry transfer-out market value >-1316 and accumulated assets<0 and profit and loss amount>0.09, then normal customer confidence is 94.6%.

(4) If the balance entry transfer-out market value >-1316 and accumulated assets<0 and profit and loss amount<0.09, then normal customer confidence is 68.4%.

These classification rules for the generation of decision-making tree guiding significance to the daily customer management and maintenance work of the securities companies, especially Rules (1) and Rule(3), which remind the customer maintenance personnel of the securities companies to closely concern the customers' yield and balance transferred-out value. If balance transferred-out value is approximate to -1316 or accumulated assets are approximate to 0, appropriate measures should be taken to detain customers.

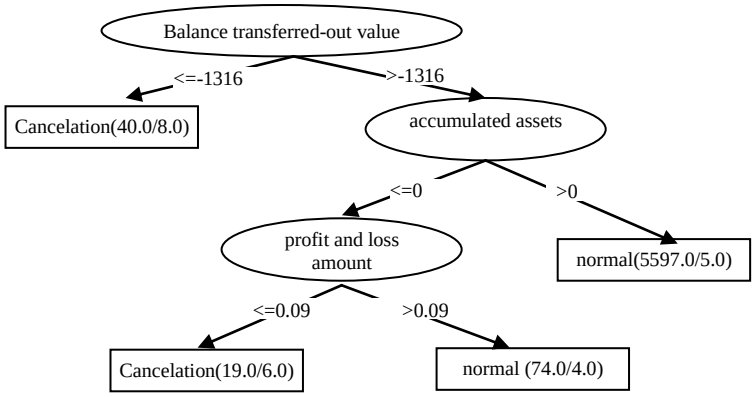

Figure 2. Results Figure of Decision Tree

4. Model Assessment

K-Fold Cross-Validation means to divide the original data into k sets with even quantity, and then carry out k-time iteration. The process of iteration is that a different set is selected from the k sets as the test set, and the remaining k-1 sets are used to train the classifier. In this way, k-time training and test are carried out, and then the error rates of k-time tests are averaged to get an overall estimation of comprehensive error. This is equivalent that each set participates in k-1 trainings and 1 test. If the category ratio of each divided set is consistent with the original data, it is called layered cross validation. Layering means to require the consistent category ration in the divided sets. Generally, layering technique can be used to improve the result. Result of 10-fold cross-validation is presented in Table 2.

Table 2. Result of 10-Fold Cross-Validation

=== Stratified cross-validation ===						
=== Summary ===						
Correctly Classified Instances 5700 99.4764 %						
Incorrectly Classified Instances 30 0.5236 %						
Kappa statistic 0.7295						
Mean absolute error 0.0071						
Root mean squared error 0.0649						
Relative absolute error 37.8376 %						
Root relative squared error 67.1545 %						
Total Number of Instances 5730						
=== Detailed Accuracy By Class ===						
TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
0.997	0.241	0.998	0.997	0.997	0.948	normal
0.759	0.003	0.707	0.759	0.732	0.948	cancelation
=== Confusion Matrix ===						
A		b		classified as		
5659		17		a = normal		
13		41		b = cancelation		

The result shows that there are 5679 classifications in the matrix, including 5659 normal classifications and 17 cancelation classifications; there are 54 cancelations, in which 13 are classified as normal wrongly and 41 are normal.

Because the model is created with the data of the first quarter, it is still insufficient to use the data of the first data for 10-fold cross-validation, and the data of the second and third quarters should also be used for model evaluation. Results of the second and third quarters evaluation are presented in Table 3 and Table 4.

Table 3. Assessment Result of Second Quarter

Correctly Classified Instances	6471	99.7841 %
Incorrectly Classified Instances	14	0.2159 %
Kappa statistic		0.7575
Mean absolute error		0.0451
Root mean squared error		0.0038
Relative absolute error		40.0364 %
Root relative squared error		66.4694 %
Total Number of Instances	6485	
=== Confusion Matrix ===		
A	b	classified as
6449	6	a = normal
8	22	b = cancelation

Table 4. Assessment Result of Third Quarter

Correctly Classified Instances	6876	99.8403 %
Incorrectly Classified Instances	11	0.1597 %
Kappa statistic		0.7835
Mean absolute error		0.0028
Root mean squared error		0.0389
Relative absolute error		32.7943%
Root relative squared error		60.0358%
Total Number of Instances	6887	
=== Confusion Matrix ===		
A	b	classified as
6856	2	a = normal
9	20	b = cancelation

Three the above three evaluation validations, the accurate rate of overall instance of the model is up to 99%, in which the accurate rate of judgment to the normal customers is up to 99% and the minimum rate to the judgment of canceled customers is up to 68.9%, so the forecast result of this model is relatively accurate. The customer maintenance personnel can forecast the customer loss and detain the customer to be lost accordingly with this model, so as to reserve the customers.

5. Modeling Application

After creation and evaluation, the customer loss model is submitted to the customer maintenance personnel, who will judge whether the customers are to be lost according to the specific customer condition and count the customers to be lost, so as to detain them before loss. Meanwhile, this model can be applied in different data sets. The model can be used to mark the category of an example and score one application. It can also be used to select the record conforming to the specific requirements in the database for further analysis with OLAP tool etc. In addition, it is also possible to investigate the customer satisfaction, so as to find out the real reason for customer loss and propose targeted strategy to prevent customer loss [14].

6. System Design and Implementation

6.1. Development Environment

The operating system is Windows 7, the development platform is eclipse, the JDK version is J2SDK1.6.0.03, test database is Microsoft SQL Server 2000 +SP4, data warehouse is Microsoft SQL Server 2000 Analysis, database driver is Jdbc4. OLAP display is JPivot.

6.2. Development Process

The system design is based on the open-source data mining system Weka and open-source OLAP display tool, and JSP language is used to develop the securities

data mining system. The distributed structure mode(Browser/Server) is adopted in system design. Functional module is shown in Figure 3.

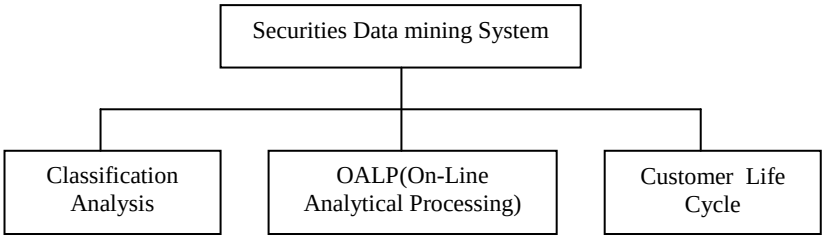

Figure 3. System Functional Module

The classification model is to use J48 in the classification algorithm in Weka software as the core algorithm, use servlet technology of JSP to call the j48 algorithm in Weka and publish the classification result to the internet through servlet.

6.3. Runtime Environment

This system adopts the B/S distribution structure model developed based on JSP, which is characterized by platform crossing, free installation, use convenience and simple operation. As long as an application is deployed in the server, users can use the system through the browser without installing software in their computer.

6.4. System Function Testing

The data of the first quarter in 2011 are used s the system test data to test the system function and use condition.

6.4.1. Classification Analysis: The classification is set by selecting the database table name and property and classification property (must be character type) to be mined. The historical modeling records can also be reserved in the system, and the specific information of model can be displayed through the model view record in the table. It is also needed to verify the correctness of the model. It is possible to select the model and data to be evaluated. Here, the data in the second quarter are selected to evaluate the result. The results are shown in the Figure 4.

```
Correctly Classified Instances 6467 99.7224 %
Incorrectly Classified Instances 18 0.2776 %
Kappa statistic 0.6772
Mean absolute error 0.0034
Root mean squared error 0.0468
Relative absolute error 36.1142 %
Root relative squared error 69.0336 %
Total Number of Instances 6485

=== Confusion Matrix ===

a b <-- classified as
6448 7 | a = normal
11 19 | b = cancelation
```

Figure 4. Evaluation Results of the Data in the Second Quarter

6.4.2. OALP Analysis: OALP (On-Line Analytical Processing) is to carry out online data access, processing and analysis trough a specific topic and display the system condition comprehensively to the users by intuitive means from more dimensions and data. A data warehouse is established and a multi-dimension database is created according to the analysis topic, for the purpose of using OLAP to meet the decision-making support. The application of OLAP can be started from

customer asset analysis, customer transaction analysis and customer profit and loss analysis *etc*.

6.4.3. Customer Life Cycle: By selecting a property and modifying the reduction proportion value of the property, it is possible to calculate the customers belonging to this range and display the change charge of the customer property. For example, through the customer number, it is possible to view the bar chart of the customer in each quarter which is shown in Figure 5.

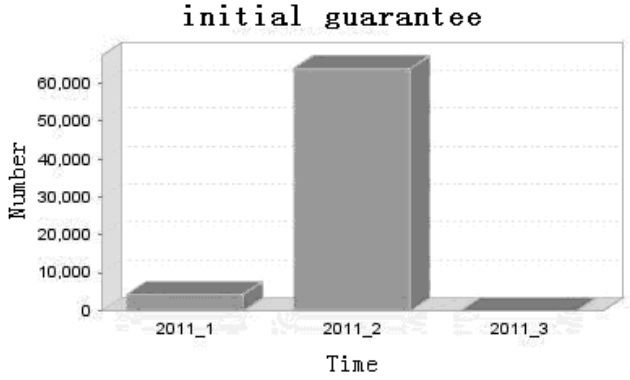

Figure 5. The Bar Chart of the Customer in Each Quarter

7. Conclusions

The customer transaction data are mined and analyzed through business intelligence technology and a customer loss warning model is established to research the API of Weka, analyze and design the securities data mining system. According to the process of creating customer loss warning model, the core technology (classification analysis) of business intelligence and OLAP analysis are introduced into this system. This system is a B/S mode structure compiled in JSP language. Weka software is open-source software compiled in JAVA language, JSP language can be used to call the API of Weka and design the system into a system of B/S mode. This system is characterized by platform crossing, free installation, use convenience and simple operation.

There are still many shortcomings in this research, for example, it has always been very difficult to select the warning characteristics of customer loss model. The data preprocessing in the system has not been improved and there has been no forecast to the single customer loss. The basic customer data and basic transaction are confidential, so there still has some difficulty in OLAP. In addition, it is also required to investigate and research the customer satisfaction, so as to find out the real reason for customer loss, and meanwhile, it is also needed to subdivide the customers, so as to take more specific measures. The customer loss waning model plays a certain basic role in the customer loss management theory, and there is still a lot of work to be done, whether in theoretical basis or practical application.

Acknowledgements

The project was supported by the Doctoral Scientific Research Foundation of Jimei University (Q201405) and Huang Huizhen Foundation of Jimei University (SD 201406)

References

[1] W. Hong-Bo and Z. Lei, "Analysis of Customer Behavior Loyalty in Insurance", Journal of Harbin university of science and technology, vol. 15, no. 1, **(2010)**, pp. 129-132.

[2] X. Fang, "Application of Data Mining in Customer Churn Management of Securities Company", Science and Technology Management Research, no. 10, **(2011)**, pp.180-183.
[3] R. A. Soeini and K. Vahidy, "Rodpysh. Evaluations of Data Mining Methods in Order to Provide theOptimum Method for Customer Churn Prediction", Case Study Insurance Industry, vol. 24, **(2012)**, pp. 290-297.
[4] 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.
[5] 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.
[6] Q. Kong and X. Liang, "The research on structural path between relationship benefits and relationship outcomes in customer loyalty programs: Using data mining", Advanced Materials and Engineering Materials, vol. 457-458, **(2012)**, pp. 1118-1121.
[7] D. Wei and J. Wei, "A mapreduce implementation of C4.5 decision tree algorithm", International Journal of Database Theory and Application, vol. 7, no. 1, **(2013)**, pp. 49-60.
[8] N. Kerdprasop and K. Kerdprasop, "A Robust Tree Induction Method Based on Heuristics and Cluster Analysis", International Journal of Database Theory and Application, vol. 5, no. 2, **(2012)**, pp. 17-34.
[9] ZY. hang, Y. Wang, H. Chunfang and Y. T. Ting, "Research on forecast model and application of customer loyalty under the background of big data", International Journal of Multimedia and Ubiquitous Engineering, vol. 9, no. 10, **(2014)**, pp. 209-222.
[10] M. Yong, "Advanced Mathematics Teaching Evaluation Model base on Decision Tree Algorithm", JCIT, vol. 8, no. 9, **(2013)**, pp. 602 ~ 608.
[11] C. J. Mantas and J. A. Credal, "Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data", Expert Systems with Applications, vol. 41, no. 10, **(2014)**, pp. 4625–4637.
[12] S. Guang-Ling and H. Zhong-xiao, "An Improved Algorithm Based on CART Decision", Journal of Harbin University of science and technology, vol. 14, no. 2, **(2009)**, pp. 17-20.
[13] J.-H. Cheng a,*, H.-P. Chen a and Y.-M. Lin b, "A hybrid forecast marketing timing model based on probabilistic neural network, rough set and C4.5", Expert Systems with Applications, vol. 37, no. 3, **(2010)**, pp. 1814–1820.
[14] Z. Yihua, W. Yuan, H. Chunfang and Y. Tingting, "Modeling and Application Research on Customer Churn Warning System Based in Big Data Era", International Journal of Multimedia and Ubiquitous Engineering, vol. 9, no. 9, **(2014)**, pp. 281-298.

