Study on Customer Loyalty Prediction Based on RF Algorithm

Jiong Mu
College of Information Engineering and Technology Sichuan Agricultural University, Ya'an, China
Email: scmjmj@yahoo.com.cn

Lijia Xu*
College of Information Engineering and Technology Sichuan Agricultural University, Ya'an, China
Email: lijiaxu13@163.com,

Xuliang Duan and Haibo Pu
College of Information Engineering and Technology Sichuan Agricultural University, Ya'an, China
Email: xuliang Duan@163.com, HaiBo Pu@163.com

***Abstract*—In the fierce electronic commerce competition, how to utilize the existing resources to improve customer loyalty has became an important factor for enterprises to guarantee their sustained growth. Through investigation into the customer statics, dynamic and enterprise service attributes, this paper, in order to guarantee higher customer loyalty of enterprises, has used random forest algorithm to improve the accuracy of the prediction to customer loyalty, and take the appropriate measures to avoid loss of customers.**

***Index Terms*—Customer loyalty, random forests (RF), Agriculture-related enterprise**

I. Introduction

With the soaring development of e-business, the market shows openness, transparency and other characteristics in network environment. The distance between users and rivals is sharply narrowed, and users can achieve a comprehensive comparison among various products and services at low cost, therefore, the establishment and maintenance of customer loyalty is getting harder. However, cultivation and maintenance of customer loyalty is the key for an enterprise to gain profits. Domestic and overseas scholars have conducted a great many of studies on the factors influencing customer loyalty so far. According to the research results of Frederick, et al [1] , an enterprise won't achieve dominant advantage unless it focuses on and retains customer loyalty, in despite of what the enterprise adopts, traditional marketing, mixed mode or online operation. Hethil, et al [2] put forward in the online bank customer loyalty study that brand image and customer satisfaction are the main factors that influence customer loyalty; Hispen, et al [3] studied the driving model of online customer loyalty, and pointed out in the model that brand image, service quality and perceived value of customers are the key factors influencing customer loyalty; Shun, et al [4][5][6][7] proved by research that brand image, service quality, customer satisfaction and other factors have significant influence on customer loyalty. It's indicated in the survey of Bain Capital of the USA: the cost of developing a new customer is much higher than maintaining an old customer, therefore, maintaining the long-term customer loyalty can largely lower the management cost of an enterprise and thus strengthen the competitiveness on the market. Consequently, this paper provides a comprehensive analysis on customer loyalty from the aspects of customer satisfaction and perceived quality and as well as the cost of transition of constraint factors, etc, so as to achieve a reasonable model that influences customer loyalty.

II. Helpful Hints

Customer loyalty refers to the extent that a user repetitively purchases the products of an enterprise or brand to which the user prefers. For a website, customer loyalty is the behavior that a user browses it due to the user's preference to the functions of the website. It can be seen that the products or brands provided by an enterprise decide the customer loyalty, in addition, service quality of the website is a significant factor that influences the satisfaction of the customer, and the satisfaction degree will influence the customer loyalty to the enterprise. Proper transfer of costs is a regulating means for the website to lock in customers, and thus a customer loyalty evaluation index is constituted.

III. Introduction to RF Algorithm

In terms of prediction of customer loyalty, classification algorithm of machine leaning is universally adopted to build up a prediction model. There are many classification algorithms of machine learning available, and support vector machine (SVM), decision tree, neural

Manuscript received November 1, 2012; revised January 7, 2013; accepted January 7, 2013;

*Corresponding author.

doi:10.4304/jcp.8.8.2134-2138

network and among others are widely used [8] . Random forest (RF) algorithm is a combined classifier algorithm [9] put forward by Breiman in 2001. It adopts classification and regression tree (CART) as an element classifier and produces different discrepant training sample sets by Bagging method, and randomly selects characteristics for attribute split of internal node when constructing a single tree. The combination of Bagging method and CART algorithm plus the attribute splitting make RF more noise tolerant and has higher classification performance. Breiman proved that there is an upper error limit in RF algorithm, therefore, RF algorithm is employed in this paper to predict customer loyalty composed of high-dimensional data.

The basic unit of RF algorithm is a decision tree, and the structure of the decision tree is decided by a random vector. RF is a classifier composed of multiple decision trees {h(x, θ_k)}, therein, {θ_k} are mutually independent, and the vectors are in identical distribution. The final tag of the final input vector X depends on the comprehensive decision of all decision trees. When constructing k trees, k random vectors need to be generated, and these random vectors θ_1, θ_2 $\cdots\theta_k$ are independent and in identical distribution. K classifiers ($h_1(X)$, $h_2(X)$, $\cdots$, h_k (X)) and random vectors X and Y are given to define the edge function:

$$mg(X,Y) = av_k I(h_k(X)=Y) - \max_{j\neq Y} av_k I(h_k(X)=j) \quad (1)$$

Therein, I(·) is an indicator function. The edge function presents the degree that the vote of the correct classification Y of vector X exceeds the average vote of any other class. It can be seen that the greater the margin the higher the confidence level of classification.

Generalization error of classifier:

$$PE^{*} = P_{X,Y}(mg(X,Y)<0) \quad (2)$$

With the increase of trees, for any random vectors θ_i, PE^* shows the following trend:

$$\mathrm{P}_{x,y}(p_\theta(h(x,\theta)=Y) - \max_{j\neq y} p_\theta(h(x,\theta)=j<0) \quad (3)$$

It indicates RF will not over-fit. It is a key feature of this algorithm, and the generalization error PE^* approaches to an upper limit with the increase of trees, therefore, RF has very good expansibility in case of unknown prediction.

The growth process of decision tree is shown below:

Steps of RF algorithm: (1) A training sample sub-set is built up by bootstrap method, and a binary recursive survival tree is established for each sample set by top-down recursion. (RF decision tree constructed by this method is a binary tree in a single structure). The split test on each decision node in the decision tree is generated in one random set, and the split is decided by the quantitative standard splitting threshold Δ. (3) Make the survival tree grow as much as possible, and each decision tree needn't pruning. Until the sample quantity of each end point isn't smaller than d_0 (>0), the corresponding conclusion is achieved at each leaf node (tag). (4) Multiple paths from the root node to each leaf node of the decision tree compose multiple rules of classification.

IV. Establishment of Prediction Model

A. Data Attributes

Each entry of the data set includes three procedures, 24 attributes in total, and the details are shown in Table 1.

TABLE I.
DATA STRUCTURE OF CREDIT EVALUATION AND PREDICTION MODEL

Category	Test Variables	Remarks
Static Attribute	The client's age	
	Diploma	Indicated with number1-6 from low to high
	Income	
	Occupation	
	Marital status	Yes or no is respectively represented by 1,0
	Duties	From cadre to departmental level respectively represented by 1-5
Dynamic attribute of customers	Frequency of access	
	The last access time	
	Average length of stay	
	Number of pages accessed	
Service attribute	Brand image	Service attribute is expressed by 7-level scoring method
	The range of services	
	The reliability of services	
	The effectiveness of services	
	The responsiveness of services	
	The stability of services	
	The safety of services	
	The rationality of services	
	Operational services	
	Advisory service capabilities	
	Order speed	
	After service ability	
	Degree of value compared to similar goods	
	Delivery cost	

The three major properties of customer treatment are provided in Table 1: The static property mainly reflects the social relations of customers; the dynamic property mainly reflects the website application property of customers; service property mainly embodies the quality of services provided to customers. The three aspects together constitute the model customer loyalty evaluation model [10][11] . In this model, some non-numeric properties need to be numeralized, so as to ensure each entry can be embodied by a certain numerical value. If it's found that some property items are missing after some records are obtained, the median is taken as the default of the item. For example, Positions 1-5 respectively indicate "below section-level", "section-level", "division-level", "department-level" and "above department-level". In case of missing of the entry, please fill in its median 2 (section-level).

B. Normalization

After the numeralization of each attribute shown above, the difference between the values of attributes is relatively remarkable due to the difference of value ranges. To avoid deviation of analysis result caused by big difference in numeric values, each attribute value is normalized in advance. The maximum value in this item is divided by the actual score of each item provided in the table, thereby, the value of each attribute drops within [0,1].

V. Simulation experiment

A. Data Source

Based on the data required in the above established prediction model, relevant questionnaire is designed for agriculture-related enterprises in Sichuan Province, 6 points where agriculture-related enterprises are concentrated are selected, i.e. Chongzhou, Pengzhou, Dujiangyan, Shifang, Mianzhu and Hanyuan, the customer data of 36 agriculture-related enterprises from each point are selected for investment and collection, 63,880 customers in total are selected for survey, the questionnaire contents are generated in accordance with the above prediction model, and a specific survey is conducted to achieve more knowledge about the customers of agriculture-related enterprises . Survey entries with too much missing information are screened out, and 2,576 relatively complete entries are randomly extracted from the remaining ones, these sample data are designed with each entry containing 25 properties and one label indicating positive or negative sample, in this model, positive sample indicates lost customers, and negative sample indicates normal customers. Subsequently, the 2,576 samples are classified into two parts, one of which is taken as the training set of classifier, the training set contains 195 positive samples and 1,093 negative samples; the remaining samples are used as checking set which contains 187 positive samples and 1,011 negative samples, and it's used to verify the classification effect of the algorithm.

B. Selection of Standard Splitting Threshold and Parameter

In the RF algorithm, the splitting threshold Δ of leaf node has influence on both the classification effect and complexity of the algorithm, but it's also directly related to mixing ratio α [12] . In this algorithm, after comprehensive consideration of the characteristics of samples and the accuracy of algorithm classification and through multiple experiment comparisons, it's found: when $\Delta=0.31$ (corresponding to $\alpha=0.2$), the accuracy of the algorithm is high at this time, and the cost is low compared with the established forest. Therefore, during the algorithm splitting process, the judgment condition selected for leaf node splitting is: $\Delta=0.31$ is taken as the basis for leaf node splitting.

C. Measures for Sample Balance Treatment

The ratio of positive and negative samples for this survey is about 1: 8, and it's unbalanced. As each tree in the RF depends on the random vectors of independent samples with the same distribution, the final classification effect of RF algorithm may be affected if the ratio of positive and negative samples in the sample set is too big. Therefore, weighting method is adopted for pre-treatment of the imbalance of samples to achieve balanced samples [13][14] , and the positive samples and negative samples are multiplied by different weighting values respectively. If the ratio of positive and negative samples is m:n, then $\zeta=m/(m+n)$, and the weight of positive samples is selected as: $(1-\zeta)$; and that of negative samples is as: ζ.

D. Experiment Result

Comparative method is adopted in this experiment, the above data are used to compare with the experiment results by different algorithms, and typical classification algorithms are adopted in this paper: support vector machine (SVM)[15-16] and K nearest neighbor (KNN)[17-18] are compared with the improved RF algorithm.

Evaluation index adopts ROC curve which is a comprehensive index reflecting the continuous variables of sensitivity (SE) and specificity (SP). On the ROC curve, the true positive rate (SE) is the ordinate and the false positive rate (1- SP) is abscissa. Theoretically speaking, the curve is a diagonal (opportunity line) drawn from the origin to the top right corner; ROC curve is usually located above the opportunity line, the further from the opportunity line the higher the accuracy of prediction; the area under the curve (AUC) can reflect the accuracy of the diagnostic experiment, and the value of this index drops within 0.5-1, and the value 0.5-0.7 indicates lower accuracy of prediction; 0.7-0.9 indicates intermediate accuracy; and value above 0.9 indicates higher accuracy [19] .

Firstly, training and test are conducted with the above data by RF method, and it's found: the prediction accuracy is relatively high if 76% positive samples are taken when 20% of the total number of students is achieved, and 93% positive samples are taken when 40% users are read. And the improved RF algorithm is superior to the original algorithm in respect of time complexity and spatial complexity.

The comparison results achieved by the three algorithms are provided in Table 2, the support vector machine (SVM) is the lowest, the accuracy of positive samples taking is 92.64% and the AUC is 0.89. The improved RF algorithm shows the best performance in this test, and the accuracy of positive sample taking is as high as 95.35% and AUC (Area Under Curve) reached 0.96, this shows that the improved RF algorithm greatly improved the classification of positive samples, and indicates classifier constructed thereby has higher performance.

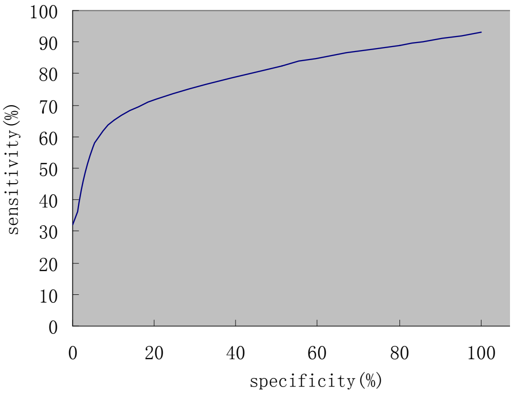


Figure 1. ROC graph using K nearest neighbor

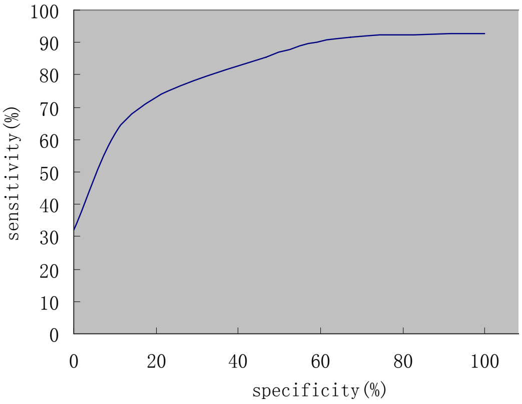


Figure 2. ROC graph using support vector machine

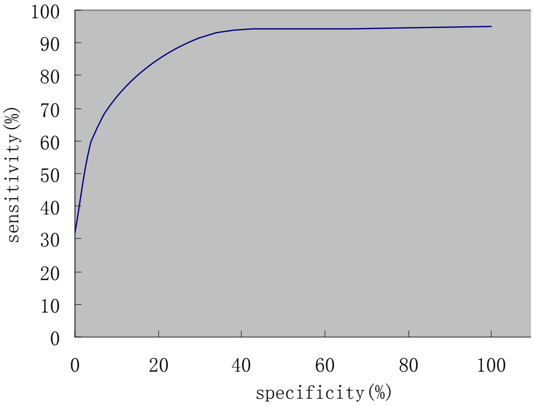


Figure 3. ROC graph using improved random forest

TABLE II.
TESTING RESULT OF EACH ALGORITHM

Classification algorithm	Accuracy (%)	AUC
KNN	93.81	0.91
SVM	92.64	0.89
IRF	95.35	0.96

VI. Conclusion

The study result indicates a customer's expectation of profits is directly proportional to the time of his/her stay in most occasions of the operation of an enterprise. It's entirely different between losing an old customer and obtaining a new customer in respect of economic benefits. The profits from repetitive purchasing of loyal customers are the precondition for the survival and development of each enterprise, and a considerable proportion of turnover is usually from a relatively small proportion of loyal customers which seems inconspicuous. Thus it can be seen that loyal customers is critically to an enterprise. If the dissatisfaction of customers to the enterprise can be predicted, the possible loss of customers discovered, and the causes of dissatisfaction of customers found out from the static data of customers, the behavioral data of exchanges between customers and the enterprise and the service data of the enterprise, active measures may be adopted and specific remedies timely prepared to retain customers according to the causes of their dissatisfaction. Whereas, the improved RF algorithm is based on such idea, it can be seen from the operation testing result achieved by improved RF algorithm, the accuracy of sample prediction may reach 95.35%, this indicates that the variation of customers' psychology and satisfaction to the services of the enterprise can be accurately figured out in accordance with the subtle behavioral features, so that customers with relatively big fluctuation in satisfaction can be screened rapidly and effectively, the cause of fluctuation worked out, and remedies made to these customers accordingly, and the customers that may be lost are retained. By this way, the customer management capacity of the enterprise can be effectively improved, and more accurate decisions on customer management can be provided to the decision makers of the enterprise.

Acknowledgment

This paper was funded by provincial science and technology department, project name: On the Elements of Successful Agricultural Enterprise E-commerce Implementation in Earthquake-stricken Area of Sichuan Provinc（2010ZR0134）.

REFERENCES

[1] F.R.Frederick,and P.Schefter, “E-loyalty: Your Secret Weapon on the Web,” *Harbard Business Review*,vol. 78, pp. 105-113, July 2000.

[2] E. L. B. Methil, and H. Nysveen, “Loyalty of On-line Bank Customers,” *Journal of Information Technology*, vol. 14,pp. 375-386, April 1999.

[3] G. L. Hispen, and C. C. L. Judy, “Towards an Understanding of the Behavioral Intention to User s Website,” *International Journal of Information Management,* vol. 20, pp. 197-208, March 2000.

[4] Y. L. Shun, S. M. Venkatesh, and E. Krishna, “Customer Value, Satisfaction,Loyalty and Switehing Costs:An Illustration from a Business-to-Business Server Context,” *Academy of Marketing Science*,vol. 32, pp. 293-311, March 2004.

[5] Z.Yang,and R.T.Peterson, “Coustomer Preceived Value,Satisfaction,and Loyalty:The Role of Switching Costs,” *Psychology and Marketing*, vol. 24, pp. 799-822, October 2004.

[6] J. Lee, and L.Feick, “The Impact of Switching Costs on the Coustomer Satisfaction-Loyalty Link:Mobile Phone Service in France,” *Journal of Service Marking*,vol. 15, pp. 35-48, January 2001.

[7] Olsen, “Comparative Evaluation and the Relationship between Quality,Satisfaction,and Repurchase Loyalty,”*Journal of the Academy of Marketing Science*,vol. 20, pp. 240-249, March 2002.

[8] S. Lessmann, and S.V O B, “A Reference Model for Customer-Centric Data Mining with Support Vector Machines,” *European Journal of Opertional Research*, Vol. 199, pp. 520-530, December 2009.

[9] L. Breiman, “ Random Forests,” *Machine Learning*, vol. 45, pp. 5-32, January 2001.

[10] S. Figini , and D. Fantazzini , “Radiom Survival Forests Models for SME Credit Risk Measurement,” *Methodology and Computing in Applied Probability*, vol. 11,pp. 29-45, January 2009.

[11] K. Coussement , and D. V. D. Poel, “Improving Customer Attrition prediction by Integrating Emotions form Client/Company Interaction Emails and Evaluation Multiple Classifiers,” *Expert Systems with Applications*, vol. 36, pp. 4626-4636, March 2009.

[12] S.L.Han, Y.J.Yang, W.H. Liu, “Incremental Learning for Dynamic Collaborative Filtering,” *Journal of Software.* . Vol.6, No 6, pp. 969-976, Jun 2011.

[13] M. Kubat, and S. Matwin, “Addressing the Curve of Imbalanced Training Sets:one-sided Selection,” *San Francisco, CA:Morgan Kaufmann Publishers*, 1997, pp. 179-186.

[14] J. Burez , and D.V.D. Poel , “Handing Class Imbalance in Customer Churn Prediction,” *Expert Systems with Applications*, vol. 36, pp. 4626-4636. March 2009.

[15] B. Yu, H. F. Li, C. Y. Fang, “Speech Emotion Recognition based on Optimized Support Vector Machine,” *Journal of Software.* Vol.7, No 12, pp. 2726-2733, Dec 2012.

[16] Z.Y.He, “Accelerometer Based Gesture Recognition Using Fusion Features and SVM,” *Journal of Software.* Vol.6, No 6, pp. 1042-1049, Jun 2011.

[17] J.P. Gou, T.S.Xiong, Y.Kuang, “A Novel Weighted Voting for K-Nearest Neighbor Rule,” *Journal of Computers.* Vol.6, No 5, pp. 833-840, May 2011.

[18] H.H.Hsu, A. C. Yang, M.D.Lu, “KNN-DTW Based Missing Value Imputation for Microarray Time Series Data,” *Journal of Computers.* Vol.6, No 3, pp. 418-425, Mar 2011.

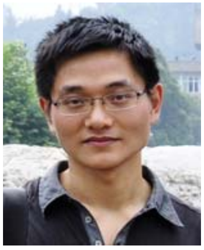
[19] A. P. Bradley, “The Use of the Area Under the ROC Curve In The Evaluation of Machine Learning Algorithms,” *Pattern Recogn*, vol. 30, pp. 1145-1159, 1997.

Jiong Mu, Born in April 1971, Sichuan Province, China. She graduated from Sichuan Normal University in 1993, and obtained a master's degree in computer applications from Sichuan University in 2007. Now she is an associate professor in Sichuan Agricultural University.

Her main research interests include computer technology in e-commerce, intelligent algorithm and optimization design， etc. She has published more than 8 papers, wherein 3 papers have been cited by EI.

Correspondence address: School of Information and Engineering Technology, Sichuan Agricultural University.

Lijia Xu, Born in December of 1973, Sichuan province, China. She graduated from Sichuan Engineering Institute in 1996, and then obtained master degree from Beijing University of Technology in 2004. She got doctor degree in automation engineering in university of Electronic Science and Technology of China in 2009. Now she is an associate professor in Sichuan Agriculture University. Her main research interests include intelligent signal processing, intelligent algorithm and fault diagnosis, etc.



Xuliang Duan, born in 1982, received the BS degree in 2005 and the MS degree in 2008 from School of Information Science & Technology of Beijing Forestry University, China.

He joined Sichuan Agricultural University in 2008 and currently is a lecturer at the Department of Computer Science. His main research interests include web engineering, natural language processing and learning technologies.

Haibo Pu, born in June of 1973, Sichuan province, China. He Graduated from Chengdu University of Technology in 1996. Now he is a lecturer in Sichuan Agriculture University. His main research interests include wireless sensor network, embedded system and intelligent algorithm, etc.