A New Customer Classification Algorithm for Electronic Commerce Enterprises

Xinwu Li
Electronic Business Department
Jiangxi University of Finance and Economics
Nanchang, 330013, China

Journal of Digital Information Management

ABSTRACT: *Correctly and effectively customer classification according to their characteristics and behaviors will be the most important resource for electronic marketing and online trading of network enterprises. A new customer classification algorithm for electronic commerce enterprises is advanced based on analyzing customer characteristics and behaviors. First, based on consumer characteristics and behavior analysis, 21 customer classification indicators, including customer characteristics type variables and customer behavior type variables, are designed, Second, aiming at the shortages of the existing BP neural network algorithm of data-mining for customer classification, the immune-genetic algorithm is used to correct BP neural network to speed up the convergence of the model. Finally the experimental results verify that the new algorithm can improve customer classification accuracy and can guarantee the effectiveness and validity of customer classification for electronic commerce enterprises in its engineering application.*

Categories and Subject Descriptors: K.4.4 [Electronic Commerce]; F.1.1 [Models of Computation]: Neural Networks

General Terms: Classification algorithm, Neural networks, E-Commerce

Keywords: Electronic Commerce, Customer Classification, BP neural network, Immune Genetics

Received: 11 May 2012, Revised 13 July 2012, Accepted 18 July 2012

1. Introduction

Customer relations management is one of the core problems of modern enterprises, whose customer oriented thought requires CRM system to be able to effectively obtain various kinds of information of customers, identify all the relations between the customers and enterprises and understand the transaction relation between customers and enterprises; meanwhile, deeply analyze customers' consuming behavior, find customers' consumption characteristics, providing personalized service for customers, supporting the decisions of enterprises. The three basic problems CRM needs to solve are how to get customers, how to keep customers and how to maximize customer value, among which maximizing customer value is the ultimate purpose, getting customers and keeping customers are both the means for realizing the purpose. The core of analyzing the three problems CRM needs to solve is to classify customers. "*Getting Customers*" and "*Retaining Customers*" need to ascertain which customers are attainable, which customers need to be kept, which customers are kept for a long term and which customers are kept for a short term, therefore, customer classification is needed. It is the same case with "*Maximizing Customer Value*". Due to different values of different customers, "*Maximum Customer Value*" of different customers should be distinguished. Thus, the core problem of enterprises to correctly implement CRM is to adopt effective method to reasonably classify customers, find customer value, focus on high-value customers with enterprises' limited resources, provide better service for them, keep "*High-value*" customers for loss prevention; also, establish the customer service system through classification, carry out differential customer service management. Hence, customer classification is becoming a more and more popular research hotspot, also a research difficulty, becoming one of the urgent problems of CRM [1].

2. Summarization of Customer Classification Methods

The widely-used methods of enterprises for customer

classification at present are mainly qualitative method and quantitative method [2-9]. As the qualitative method for customer classification is just to classify all the target customers of enterprises in the macroscopic level, customer classification is carried out according to different value emphasis of different customers. The formation of customer value is simply expressed as: *Value = Benefit — Cost*. Qualitative classification method classifies customers in a simple way, only offering guidance for customer classification of enterprise in the macroscopic level, unable to provide specific and credible basis for enterprise decisions; furthermore, as there is no strict process of argumentation, the method depends on decider's subjective inference, there may be certain deviations in the analysis process, easily resulting in faulty decisions. For this reason, to truly provide customer classification information beneficial to enterprises should depend on quantitative technology for customer classification [3, 4, 5].

Quantitative classification method is to apply quantitative analysis technology to conduct customer classification on the basis of some specific customer variables (credit level of customers, purchasing power of customers, characteristics of demand of customers, etc.). Currently, there are mainly two categories of data mining for quantitative customer classification research, which are traditional statistical method and non-statistical method. The former mainly includes cluster analysis, Bayesian Classification, factor analysis method, etc.; this statistics-based method is unable to process a great deal of sophisticated customer data, and there are some problems on the accuracy of customer classification results, so to fundamentally solve the problem of customer classification needs to rely on non-statistical customer classification method, which mainly includes neural network, fuzzy set method, association rules, genetic algorithm, etc. The classification technology based on neural network is combined with certain information technology, which is a kind of mathematical method applicable to complex variables and multi influencing factors calculation, so it is more effective in solving complex customer classification problems with better classification accuracy, however, the convergence problem of the function itself greatly limits its application value in specific project practice. Secondly, classification is mainly based on such mathematical methods as fuzzy clustering, rough set, association rules, etc., although these methods offer classification reason explanation in a relatively clear way with better classification results under the circumstances of satisfactory data conditions, the modeling process needs to provide specific mathematical equations. As a result, these methods are limited by data conditions in specific application, always having problems like insufficient classification accuracy or poor "*robustness*", limiting the application in customer classification. Due to lots of influencing factors related to customer classification, more often than not, the complicated relations are difficult to be expressed in mathematical equations [6, 7, 8, 9].

The paper use immune genetic algorithm to correct and modify BPNN model to overcome the question of slow convergence speed of BPNN. In so doing, not only the problem of convergence speed of BPNN has been solved, but also the simplicity of the model structure and the accuracy of the transformation are ensured, the and a new customer classification model is advanced to classify online trading customer.

3. Selection of Customer Classification Indicators

The selection of reasonable classification variables is the basis of correct and effective customer classification, namely establishing scientific and reasonable classification indicator system. In view of the nature of trading and own characteristics of online trading, this Paper adopts customer characteristics' type variable and customer behaviors type variable in the specific selection of customer classification variables [2].

3.1 Selection of Customer Characteristics Type Variable

Customer characteristics' type variable is mainly used for getting the information of customers' basic attributes. Such variable indicators as geographical position, age, sex, income of individual customer play a key role in determining the members of some market segment. This kind of variables mainly comes from customers' registration information and customers' basic information collected from the management system of banks, the contents of which mostly indicate the static data of customers' basic attributes, the advantage of which is that most of the contents of variables are easy to collect. But some of the basic customer-described contents of variables are lack of differences at times.

Based on analyzing and summarizing existing literatures, the customer characteristics type variables designed in this Paper include: Customer No., Post Code, Date of Birth, Sex, Educational Background, Occupation, Monthly Income, Time of First Website Browsing, and Marital Status.

3.1 Selection of Customer behavior Type Variable

Customer behavior type variable is mainly used for getting the information of customers' basic attributes. Such variable indicators as geographical position, age, sex, income of individual customer play a key role in determining the members of some market segment. This kind of variables mainly comes from customers' registration information and customers' basic information collected from the management system of banks, the contents of which mostly indicate the static data of customers' basic attributes, the advantage of which is that most of the contents of variables are easy to collect. But some of the basic customer-described contents of variables are lack of differences at times.

Based on analyzing and summarizing existing literatures, the customer behavior type variables designed in this paper

include Monthly Frequency of Website Login, Monthly Website Staying Time, Monthly Times of Purchasing, the Monthly Amount of Purchasing, Type of Consumer Products Purchased, Times of Service Feedback, Service Satisfaction, Customer Profitability, Customer Profit, Repeat Purchases, Recommended Number of Customers, Purchasing Growth Rate.

4. Derivation of Algorithm

4.1 Simultaneous Analysis and Design

De Castro indicated that there were similarities among the quality of weight value initialization of back-propagation neural network and the relationship of network output and the quality of antibody instruction system initialization in the immune system and the quality of the immune response. A simultaneous analysis and design---SAND algorithm was advanced to solve the problem regarding the weight value initialization in the back-propagation network [6]. In SAND algorithm, each antibody corresponds to a weight value vector of neuron given in one of several layers of neural networks, the length is l, and the affinity $aff(x_i, x_j)$ between antibody x_i and antibody x_j is shown by their derivative of Euclidean distance function $D(x_i, x_j)$ in Formula 1. In which, ε is a positive of value adoption 0.001. The definition of Euclidean distance function $D(x_i, x_j)$ is shown in Formula 2 [6].

$$aff(x_i, x_j) = \frac{1}{D(x_i, x_j) + \varepsilon} \quad (1)$$

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^{l} (x_{ik} - x_{jk})^2} \quad (2)$$

SAND algorithm aims to reduce the similarities between the antibodies and produce the antibody repertoire to cover the entire form space with the best, so energy function is maximized. The energy function is shown in Formula 3.

$$E = \sum_{i=1}^{N} \sum_{j=i+1}^{N} D(x_i, x_j) \quad (3)$$

In the method of Eculidean form space, the energy function is not percentage. With a view to the diversity of the vector, SAND algorithm has to define the stop condition. Given vector x_i, $i = 1, 2,...N$, its standardization is unit vector I_i, $i = 1, 2,...N$, $\bar{I}$ shows to calculate the average vector. Therefore, Formula 4 shows the diversity of unit vector, in which, $\| \bar{I} \|$ means the average vector distance from the origin of coordinate. Formula 5 shows the stop condition U of SAND algorithm.

$$\| \bar{I} \| = (I^T I)^{1/2} \quad (4)$$

$$U = 100 \times (1 - \| \bar{I} \|) \quad (5)$$

4.2 BP Neural Network Design Based on Immune Genetic Algorithm

According to the actual application, providing that both the input and output number of node and the input and output values in BPNN have been confirmed, activation function adopts S type function. The following steps show BP neural network design based on immune genetic algorithm.

(1) Every layer of BPNN carries on the weight value initialization separately by SAND algorithm.

(2) Antibody code. The initial weight value derived by SAND algorithm constructs the structures of BPNN. Each antibody corresponds to a structure of BP neural network. The number of hidden node and network weight value carry on the mixture of real code. Each antibody serials are shown in Figure 1.

(3) Fitness function design. Fitness function $f(x_i)$ is defined as the mean value function of squared error of neural network in Formula 6, in which, $E(x_i)$ is shown by Formula 7. In Formula 7, p is the total training sample, o is the number of node of output layer, T_j^n and Y_j^n are the n training sample's expected output and actual output in the j output node separately, and ξ is the constant larger than zero.

N number of hidden node	Weight value corresponding to the first hidden node	Weight value corresponding to the Second hidden node	...	Weight value corresponding to the N hidden node

Figure 3. Antibody Code

$$f(x_i) = \frac{1}{E(x_i) + \xi} \quad (6)$$

$$E(x_i) = \frac{1}{2p} \sum_{N=1}^{p} \sum_{j=1}^{o} (T_j^n - Y_j^n)^2 \quad (7)$$

(4) Genetic operation. The model here adopts the Gaussian compiling method to go on the genetic operation so as that each antibody decoding is the corresponding network structure and change the network weight value as shown in Formula 8, in which, x_i and x_i^m are the antibodies before and after the variation, $\mu(0, 1)$ shows that the mean value is zero and squared error is normal distribution random variable of l, and $\partial \in (-1, 1)$ is the individual variation rate. It is seen in Formula 8 that the variation degree varies inversely as the fitness, i.e. the lower the fitness is (the less the fitness value of objective function is), the higher the individual variation rate is, or vice versa. After the variation, all the hidden node and weight value components constitute a new antibody again.

$$x_i^m = x_i + \partial \exp(-f(x_i)) \times \mu(0, 1) \quad (8)$$

(5) Group renewal based on density. In order to guarantee the antibody diversity, improve the entire searching ability of the algorithm, the model adopts the Euclidean distance and the fitness based on the antibodies to calculate the similarity and density of the antibody. Providing that there are x_i and x_j antibodies, and $\eta \succ 0$ and $t \succ 0$, given constants, the fact that Formula 9 is satisfied indicates that x_i and x_j antibodies are similar, the number of antibody similar to the antibody x_i is the density of x_i marked by C_i. The probability of selecting antibody x_i is $p(x_i)$ as shown in Formula 10, in which, α and β is the adjustable

parameters between (0, 1), and $M(x)$ is the maximum fitness value of all the antibodies. It is seen in Formula 10 that while the antibody density is high, the probability of selecting the antibody with high fitness is low, and conversely high. Therefore, excellent individual is not only retained, but the selection of similar antibodies is reduced, and the individual diversity is guaranteed.

$$\begin{cases} D(x_i, x_j) \le \eta \\ |f(x_i) - f(x_j)| \le t \end{cases} \quad (9)$$

$$p(x_i) = \alpha C_i [1 - \frac{f(x_i)}{M(x)}] + \beta \frac{f(x_i)}{M(x)} \quad (10)$$

5 Experiment Confirmation

5.1 Object of Experimental Verification

The instance data of the experiment conduct empirical research on the customer data of the B2C transaction of certain enterprise website of the recent three years (totaling data of 41351 customers, 21 attributes in the data table are listed in the third part of the paper including customer characteristics type variables and customer behaviors type variables), making statistics on attribute values like annual transaction frequency, total amount, product cost, etc. of certain customer according to customer transaction records in information base, forming an information table (among which the decision attribute set D is null) [4].

5.2 Process of Experimental Verification

The process of the experimental verification can be listed as follows [5].First, what is to be processed during the classification is the numeric data, so the numeric coding on character data should be conducted first; Second, if the value number of certain attribute is equal to sample number, it means that it has little effect on classification, hence, remove such attribute first. Three attributes as Customer No., Post Code and Date of Birth are removed in this case. Third,establish training sample set according to domain (prior) knowledge. Times of purchasing and total amount of purchasing of each customer are two major factors of customer classification (this is the prior knowledge of domain), so select 400 pieces of typical data among all the customers to form training sample set. And divide them into five types as Gold Customers, Silver Customers, Copper Customers, General Customers and Negligible Customers according to ABC management theory. Fourth, use the customer classification algorithm above-mentioned, and the customer classification results can be expressed in Table 1. In the specific algorithm realization, this Paper simultaneously realizes ordinary K_means algorithm and customer classification algorithm based on BP neural network. The performance comparison of these three algorithms can be expressed in Table 2.

We can see from Table 1 that in the autonomous learning of algorithm of this Paper, such five factors as the educational background, income, occupation, times of purchasing, and total amount of purchasing of customers have a relatively great influence on customer classification. Through the classification result in Table 1, it can be seen that Gold Customers take up 7.16% of the total number of customers, while the profit takes up 52.24% of the total profit. These customers play a significant role in the existence and development of enterprises. However, the negligible customers account for 18.25%, who not only do not bring profit to enterprises, but also make enterprise lose money. These customers should be either further cultivated or eliminated according to the actual situation.

We can see from Table 2 that the cluster accuracy rate of algorithm in this paper is the highest, reaching 99.81 %, obviously higher than ordinary K-means algorithm and BP Neural Network algorithm; the square errors and E values on customer classification of three algorithms are 103.92, 161.74 and 120.56 respectively. The smaller the E value is, the smaller the possibility of wrong classification is. Thus it can be seen that the square error and E value of the algorithm in this paper during the classification are far more less than ordinary K-means algorithm [4] and BP

Customer Type	Number of Customers	Percentage %	Profit Contribution Proportion
Gold Customers	2876	7.16	52.24
Silver Customers	5934	14.77	30.18
Copper Customers	10207	25.41	13.16
General Customers	13821	34.41	6.01
Negligible Customers	7331	18.25	-1.59
Total	40169	100.00	100.00

Table 1. Customer Classification Result of Some Website

Algorithm	Algorithm in This Paper	Ordinary K-means Algorithm	BP Neural Network Algorithm
Accuracy Rate	99.81 %	87.98%	94%
E Value	103.92	161.74	120.56

Table 2. Classification Performance Comparison of Each Algorithm

Neural Network algorithm[6]. Therefore, it shows that the improvement on K-means clustering algorithm in this paper turns out to be a success, with reasonable classification results.

6. Conclusion

Customer relations management of online trading is still developing. But to correctly and effectively classify online trading customers is the critical issue for reforming network marketing mode, improving customer management and service level and enhancing competitiveness of network enterprises [5]. On account of the shortcomings of the typical K_means clustering algorithm in data mining, this Paper puts forward several improvement measures, and applies them into the classification of online trading customers. Simulation results indicate that the improved online trading customer classification has higher accuracy rate on customer classification and more reasonable classification results.

7. Acknowledgements

This work is supported by the National Natural Science Foundation of China under the grant No.60963012 and is supported by the education department of Jiangxi Province (2007259).

References

[1] Liu, Z. H.(2008). Study on model of customer classification based on the customer value. *A Dissertation of Huazhong University of Science and Technology*.

[2] Zhou, H. (2008). Study of Classifying Customers Method in CRM. *Computer Engineering and Design*, 29 (3) 659-661.

[3] Deng, W. B., Wang, Y. (2009). B2C Customer Classification Algorithm Based on Based on 3DM. *Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition).* 21 (4) 568-572.

[4] Guan, Y. H. (2011). Application of Improved K-means Algorithm in Telecom Customer Segmentation. *Computer Simulation*, 28 (8) 138-140.

[5] Quan, X. N. (2011). Application of K-means Based on Commercial Bank Customer Subdivision. *Computer Simulation*, 28 (6) 357-360.

[6] Yang, B. Z., Tian G. (2007). Research on Customer Value Classification based on BP Neural Network Algorithm. S*cience and Technology Management Research*. 23 (12)168-170.

[7] Bradley, P. S., Managasarian L. (2007).K-plane Clustering. *Journal of Global Optimization.* 16 (1) 23-32.

[8] Tang, Y., Rong Q. S. (2004). An Implementation of Clustering Algorithm Based on K-means. *Journal of Hubei Institute for Nationalities*. 22 (1) 69-71.

[9] Zhang, Y. F., Mao J. L. (2008). An improved K-means Algorithm. *Computer Application*. 23 (8) 31-33.