Send Orders for Reprints to reprints@benthamscience.ae

Open Access

Research on E-Commerce Customer Churning Modeling and Prediction

Xue Zhao*

International Education College, Nanyang Institute of Technology, Nanyang, 473004, China

Abstract: This paper discusses the customer churning prediction problem in electronic commerce. In electronic commerce the customer data change is non-linear and time-varying and other characteristics, using a single prediction model to accurately predict e-commerce customer loss is difficult. In order to improve the prediction accuracy rate of electronic commerce churning, the model first uses the genetic algorithm for the screening of effecting factors, and extracts the important influence factors which affect the predicting results. Then support vector machine and neural network are respectively used to carry out the forecast. Finally, using support vector machine fuses the two prediction results to acquire the prediction results of the combination model. Simulation results show that the combined model can improve the prediction accuracy rate of the electronic commerce customer churning, and provides a new prediction method for the electronic commerce customer churning.

Keywords: Customer churning prediction, e-commerce, neural network (NN), support vector machine (SVM).

1. INTRODUCTION

With the rapid development of the Internet, enterprises establish a global trading network through the Internet to provide consumers with greater choice, e-commerce has brought great changes to business process and consumer behavior. Compared with the traditional transaction methods, the biggest disadvantage of e-commerce is the churn rate is very high, therefore identifying those customers likely to churn, to take appropriate measures to reduce customer churn, maximizing the enterprise's profits, e-commerce has become a hot research [2]. Aiming to the loss forecast for e-commerce customers, many scholars have conducted in-depth and extensive research, and have achieved good results [3]. E-commerce customers currently are classified into two kinds of prediction methods: statistical analysis and artificial intelligence [4-6]. Statistical analysis includes linear regression, time series, cluster analysis, decision trees and Bayesian networks, etc. These methods are all supposed that e-commerce customer loss is presenting a linear variation, but because of the minds of consumers, buying behavior, economics, culture and other factors, the loss of e-commerce presents nonlinear and high-dimensional characteristics, statistical analysis model does not fully reflect the characteristics of e-commerce customer churning [7].

Artificial intelligence technology has self-learning ability and nonlinear processing capabilities, with respect to the statistical analysis, the correct rate of predict is improved. However, the actual customer loss data has the characteristics with noise, the sample being extremely uneven, the high dimensionality and high nonlinearity. The predict difficulty is increased, using a single prediction model is hard to achieve accurate churn prediction in e-commerce. If multiple predictive models are combined together, each of predictive models play its prediction advantage, the accuracy of customer churn prediction in E-commerce is expected to improve [8].

In order to improve the e-commerce customer churn prediction accuracy, this paper proposes an e-commerce customer churn prediction method based on a combination model. Firstly, a genetic algorithm is used to do the screen to the influencing factors of e-commerce customer churn, and then SVM (Support Vector Machine) and BPNN (BP neural network) are used to predict customer churn in e-commerce. Finally, SVM is used to combine the results of customer churn prediction.

2. THE PRINCIPLE OF E-COMMERCE CUSTOMER CHURN PREDICTION

The principle of E-commerce customer churn prediction is to collect the data of e-commerce customers over a period of time, analyze the e-commerce buying behavior of customers, and build a churn prediction model to adopt certain policy to try to prevent customer churn [9]. In e-commerce customer churn is affected by many factors, using a single predictive model is difficult to accurately predict, based on the idea of combination prediction model, this paper adopts SVM and BPNN two excellent e-commerce customer churn prediction model to do predict separately, then the two results is integrated by SVM to improve e-commerce customers prediction accuracy. Specifically, it is shown in Fig. (**1**).

3. COMBINATION FORESTING MODEL OF E-COMMERCE CUSTOMER CHURN

3.1. Impact Factor Normalization of E-Commerce Customer Churn

Under normal circumstances, the SVM and BP neural network depend on the inner product of feature vectors, with

1874-110X/14 2014 Bentham Open

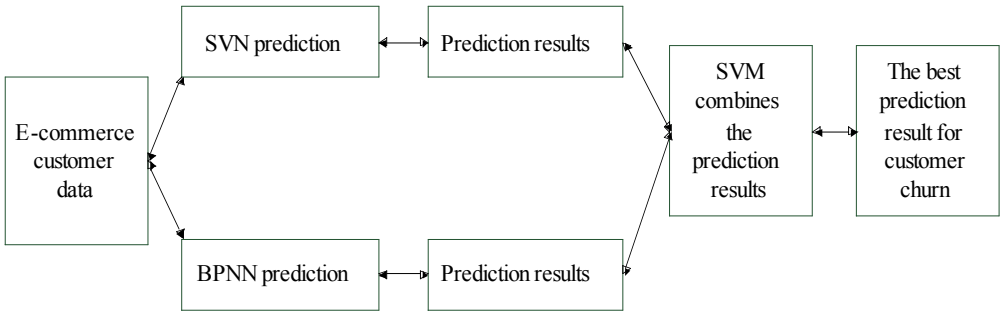

Fig. (1). The principle of e-commerce customer prediction of combined models.

the difference of e-commerce customer churn affection factors dimensionality, which may produce adversely affection to the learning speed of prediction models, in order to improve learning efficiency, the affecting factors are processed in normalization. It can be expressed as a formula in detail.

$$x_i^{'} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

In the formula, x_i is the customer churn affection factor value in e-commerce, $x_i^{'}$ is the influencing factor value after normalization, $x_{\max}$ and $x_{\min}$ represent the upper bound and lower bound of each impact factor.

3.2. The Impact Factor Screening of E-Commerce Churns

In the prediction process of e-commerce customer churn, increasing the dimension of the input data can lead to a sharp increase in the complexity of the forecasting process, the qualitative relationship between the entering dimensions and predict comprehensive effect is shown in Fig. (**2**).

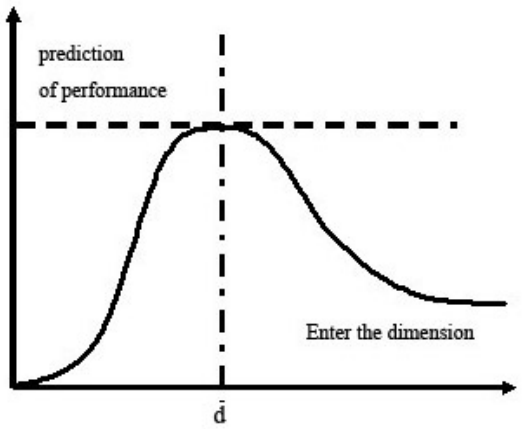

Fig. (2). The relationship between prediction performance and characteristic dimension.

It is shown from Fig. (**2**), the prediction performance and input dimension (factors) of e-commerce customer churn is not necessarily a positive correlation, SVM and BPNN have good nonlinear modeling capabilities, can better capture the nonlinear characteristics in customer churn data, but it does not reduce the input dimension function, when the input dimension is large, the training time will be longer. Genetic algorithms have good search optimization capabilities, which can be the front-end system of a predictive model. Genetic algorithm does a reasonable choice to e-commerce customer churn factors to get a new features set, and which can be applied to the follow-up e-commerce customer churn prediction.

3.3. LSSVM Algorithm

Least squares support vector machines (LSSVM) is proposed by Suykcns and others based on the improvements of the standard support vector machine. The loss function is set as squared error sum of squares, inequality constraints are changed into equality constraints, which reduces undetermined parameters and simplifies the model parameters, determining the difficulty and complexity of parameter optimization can improve the adaptability and precision of the model. For e-commerce customer training set $\{(x_i, y_i)\}, i = 1,2,...n$, x_i and y_i represents the sample input and output respectively, through the non-linear mapping function $\psi(\bullet)$ the samples are mapped into a high dimensional feature space. In the high dimensional feature space linear regression is proceeded.

Namely:

$$f(x) = \omega^T \varphi(x) + b \tag{2}$$

In the expression, ω is the weight vector, b is the offset.

According to the principle of construction risk minimization, in the expression (2), the resolved LSSSVM regressive model is:

$$\min \| \omega \|^2 + \frac{1}{2}\gamma \sum_{i=1}^{n} \xi_i^2 \tag{3}$$

$$s.t. y_i - w^T \varphi(x) + b = e_i$$

In the expression, γ is the regularization parameter.

Through leading into the Lagrangian multiplier method, the constraints optimization problem is converted into the dual space optimization problem without constraint. That is:

$$L(\omega, b, \zeta, \alpha) = \frac{1}{2}\omega^T \omega + \frac{1}{2}\gamma \sum_{i=1}^{n} \zeta_i^2 + \sum_{i=1}^{n} \alpha_i (\omega^T \varphi(x_i) - b + \zeta_i - y_i) \tag{4}$$

In the expression, α_i is the lagrangian multiplier, according to the optimization conditions,

$$\frac{\partial L}{\partial \omega}=0\,,\frac{\partial L}{\partial b}=0\,,\frac{\partial L}{\partial \zeta_i}=0\,,\frac{\partial L}{\partial \alpha_i}=0 \tag{5}$$

The follows can be concluded:

$$\omega=\sum_{i=1}^{n}\alpha_i\varphi(x_i)\,,\sum_{i=1}^{n}\alpha_i=0\,,\ \alpha_i=c\zeta_i \tag{6}$$

$$\omega\varphi(x_i)+b+\zeta_i-y_i=0$$

According to the Mercer condition: $K(x_i,x_j)=\varphi(x_i)^T\varphi(x_j)$, LSSVM prediction model is

$$f(x)=\sum_{i=1}^{n}\alpha_i K(x_i,x_j)+b \tag{7}$$

Generally, the performance of radial basic kernel function is superior to other kernel function, therefore, this paper selects LSSVM kernel function, and the LSSVM prediction model is:

$$f(x)=\sum_{i=1}^{N}\alpha_i\exp(-\frac{\|x_i-x_j\|^2}{2\sigma^2})+b \tag{8}$$

3.4. BP Neural Network Algorithm

BP neural network (BPNN) is the most widely used feed-forward neural network, which is made of the input layer, a hidden layer and the output layer, each layer constitutes of thousands of neurons. Neurons transfer function adopts the s-type transformation function. BPNN learning process is:

1) The initialization of network weights. Using a random number to assign the values to the weights and thresholds of each layer in neural network.

2) Enter the training samples of neural networks, and to do learning.

3) Calculate the output y_k of each layer in the network.

4) According to expressions (9) and (10) to calculate the back propagation of errors in each neural layer, and write down δ_{jk}^{p1}, δ_{ij}^{p1}.

$$\delta_{jk}^{p1}=(t_k^{p1}-y_k^{p1})y_k^{p1}(1-y_k^{p1}) \tag{9}$$

$$\delta_{ij}^{p1}=\sum_{k=0}^{n1}\delta_{jk}^{p1}\omega_{jk}(t)x_j^{p1}(1-x_j^{p1}) \tag{10}$$

5) If the samples number involved in the training is less than the total number of training samples, then go to step 2) to continue the learning algorithm. Otherwise, skip to step 6).

6) To correct the weight and threshold values in each layer of neural network.

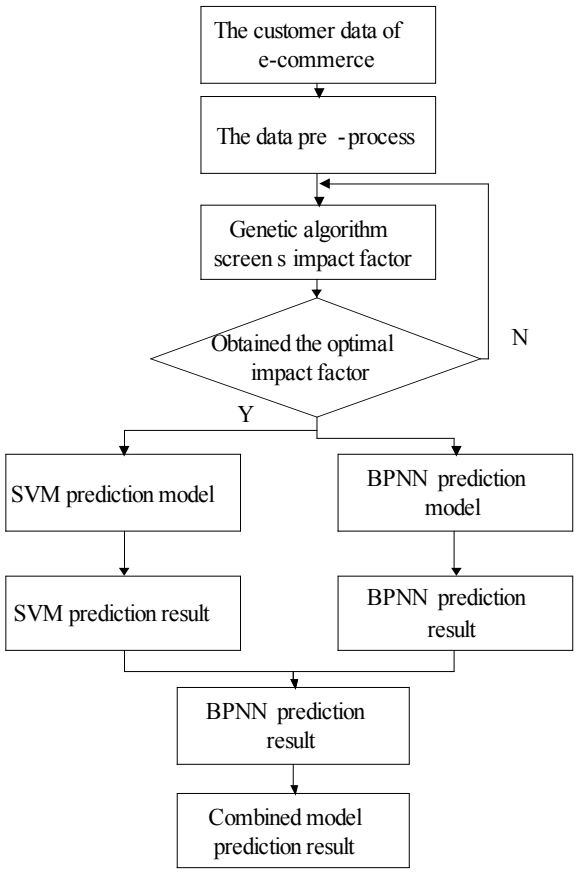

Fig. (3). The combined prediction workflow of e-commerce customer churn.

3.5. The Workflow of Combined Model

The specific workflow of customer churn in e-commerce based on a combination model is as follows:

1) Collect the e-commerce customer data from a network server.

2) Process the raw data, and eliminate some of the unwanted and abnormal data.

3) Use equation (1) to do the normalization to the impact factors of e-commerce customer churn, and improve the learning performance of the model.

4) Use genetic algorithms to screen the impact factors in e-commerce customer churn, and screen out the optimal impact factor.

5) The optimal impact factor and the corresponding customer churn data are input into SVM and BPNN to learn and build predict models.

6) The predictive model is used to predict e-commerce customer churn, respectively the prediction results of SVM and BPNN are obtained.

7) The prediction results of SVM and BPNN are input again as SVM input, the true churn value are taken as output. Re-learn and build model is to obtain the final prediction results of combined model. E-commerce customer churn workflow of combined model is shown in Fig. (**3**).

Table 1. Impact factors of customer churn in e-commerce.

Number	Name of Impact Factors	Number	Name of Impact Factors
1	Customer age	7	Repeated times of purchase
2	Customer sex	8	Marriage
3	Amount of purchase	9	Purchase times in day time
4	Purchased times in night	10	Purchase times in the evening
5	Calling times for services	11	Education degree
6	Customer credit scoring	12	Annual income

Table 2. The influence factors of e-commerce customers screened by genetic algorithm.

Number		Number	
1	Customer age	5	Repeated times of purchase
2	Customer sex	6	Purchase times in day time
3	Amount of purchase	7	Purchase times in the evening
4	Calling times for services	8	Annual income

4. SIMULATION STUDY

4.l. Data Sources

In order to test the feasibility of e-commerce customer churn prediction algorithm. An e-commerce customer data is adopted to do statistics in August to September, 2010, totally there are 2000 samples, its normal customers (non-churn) are 510, loss of customers are 1490, according to literature [10], the e-commerce customer churn impact factor are selected as shown in Table **1**. At the same time, the data are divided into two parts: training samples and testing samples, in test samples there are 400 lost customers, non-churn samples are 120 customers.

4.2. Model Implementation

First, the raw data is normalized, genetic algorithm screens the impact factor to obtain the comparative impact factor as shown in Table **2**. The data in Table **2** are input into the SVM and BPNN to learn, and mean while genetic algorithm is used to optimize both of the parameters. I n order to get the best forecasting model, SVM and BPNN optimal prediction models are used to predict the test sample. Both of the predicted results are input into the SVM learning, finally the prediction result of a combination model is obtained.

5. EXPERIMENTAL RESULTS AND ANALYSIS

The comparative models are single SVM model, BPNN model, and combined model without feature screening (SVM-BPNN), the combined model in this paper is GA-SVM-BPNN, the comparisons of prediction correct rate among several models of customer churn are shown in Fig. (**4**), and the prediction correct rates for non-loss of customers are shown in Fig. (**5**). According to Figs. (**4** and **5**), the prediction accuracy rate of the combined model is higher than

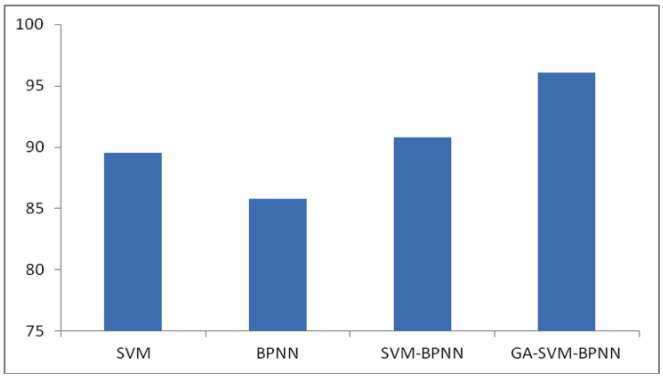

Fig. (4). The comparison among the loss customers in different models.

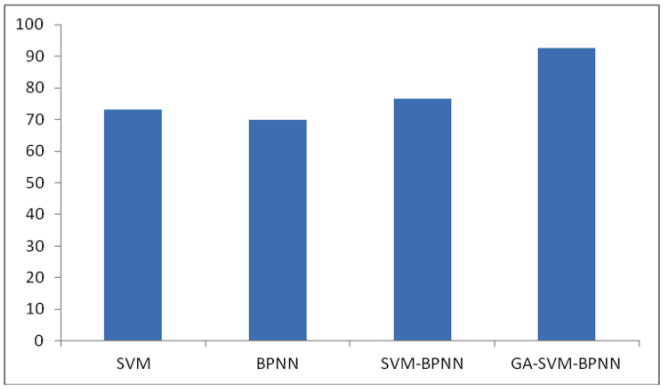

Fig. (5). The prediction accuracy comparison among non-loss customers among different models.

the rate of a single model. That's mainly due to the combined model takes full advantage of a single model to complement with each other, thereby which increases the predic-

tion accuracy rate of e-commerce customer churn. While the prediction accuracy of GA-SVM-BPNN of this paper is higher than SVM-BPNN, that's mainly because this model screens the optimal impact factor by genetic algorithm to improve the prediction accuracy, and reduce the predicted time. The prediction efficiency of the model is higher, which is a kind of e-commerce customer churn prediction method with the reliable correct rate and fast speed, and provides a new idea for e-commerce customer churn prediction.

CONCLUSION

The prediction of e-commerce customer churn is currently a hot topic. As e-commerce customer churn data is a nonlinear complex data with high-dimensional features. Single prediction model is difficult to accurately reflect the changes in characteristics. In this paper, aiming to the current shortage of e-commerce customer churn prediction model, a prediction model of e-commerce customer churn is proposed which combines the genetic algorithm, SVM and BP neural network together. Genetic algorithm is used to screen the impact factor, simplify the input of support vector machines and BP neural network. Then support vector machines and BP neural network is used to do its forecast. Finally, SVM is used to combine the prediction results of the two models. The simulation results show that, comparing to the single prediction model, in this paper the combined model improves e-commerce customer churn prediction accuracy, at the same time, which speeds up the prediction rate. The e-commerce research area has broad application prospects.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This paper is supported by the funds of the Science and Technology Research Project of Henan province (2014), China, and the project number is 142102210231.

REFERENCES

[1] X.W. Liu, "the e-commerce customer churn prediction research based on the Pareto/NBD+plain Bayesian combined model", master theses of Sichuan University, 2006.

[2] Q.J. Zhang and B.Z. Zhu, "Based on the self-organizing data mining e-commerce customer churn prediction model", Enterprise economic. pp. 95-99, January, 2011.

[3] Y. Zhao, "Customer churn analysis based on improved support vector machine (SVM)", Computer integrated manufacturing system, Vol.13 No.1, pp. 202-207, 2007.

[4] Q.J. Shi, J.M. He, and C.L. Wang, "Telecom customer churn prediction model based on improved support vector machine (SVM)", Management science, Vol. 20, No. 1, pp. 54-58, 2007.

[5] Z.H. Cheng and B.X. Liu, "The decision tree method of crisis analysis of customer churn", Journal of management science, Vol. 8, No.2, pp. 20-25.

[6] Baesens, Bart, and Verstraeten, "Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers", European Journal of Operational Research, Vol.156, No.2, pp. 508-523, 2009.

[7] Z.C. Li and M.L. Liu, "The consumer's behavior study under e-commerce environment", China management science, Vol. 10, No. 6, pp. 88-91, 2011.

[8] Y.Y. Jiang, J.H. Jiang, "The virtual enterprise risk assessment study based on BP neural network", Computer simulation, Vol.26, No.12, pp. 261-164, 2012.

[9] B.Z. Zhu, "The e-commerce customer churn prediction model based on SMC-RS-LSSVM", Journal of systems engineering theory and practice, Vo. 30, No. 11, pp. 1960-1967, 2012.

[10] Q.j. Zhang, "The three stages of prediction model of e-commerce customer churn", China soft science, Vol.6, pp.186-192, 2010.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014