Bank Customer Classification Model and Application Based on SVM

[a]Hu Yi，Fang Kui, Zhu Xinghui
Hunan Agricultural University, Changsha,10537,China
[a]email: 516918002@qq.com

Abstract

Based on the data mining method, support vector machine, technology and related analysis of variable selection methods, establishing binary classification model of VIP users and Ordinary users for commercial banks, executing classification prediction and results verifying on VIP users and Ordinary users of an Agricultural Bank in Changsha, and making comparative analysis on classification prediction accuracy and the time consuming when processing by comparing with neural network.

Keywords: support vector machines, data mining, customer classification, commercial banks

1 Introduction

With the early rapid development, banks of our country have grasp a large number of user groups, and have accumulated a lot of customer data, how to dig out useful information from these complex customer data ,how to research on different group of bank customers and analyze their difference in consumer behavior, and making different marketing strategies for them. These all have immeasurable meaning for the bank business model to change from extensive model into fine model. Customer Segmentation was first proposed in 1965 by Wendell Smith[1], an American market scientist . it's a process that an enterprise classify their customer groups into several sub-process customer groups according to customers' attributes, while customers in a same type of group are similar.

Artificial neural network is a pattern recognition tool which imitate human brain processing complex information[2], while BP neural network is the most representative and widely used neural network model,it also play an important role in the classification research of bank customers. BP neural network has superior and large scaled parallel processing capabilities as well as good self-adaptive, whose predictive effect has greatly improved with respect to the traditional customer classification method. but its complex internal structure is very complex ,its operation is complicated, the classification process takes certain period of time, what most critical is that it is easy to fall into local optimum and its generalization ability is poor.

The main theory of Support vector machine how to classify bank customers is just based on the training of known data and infer functional dependence model of the dataset, then have a further judgment and prediction of the type of unknown data [3]. Support Vector Machine agency-based risk minimization principle requires the maximum classification separation distance, the most concentrated classification subset, which coincides with the needs of the customer classification.

2 1 support vector machine theory

In 1995, Vapnik proposed support vector machine theory, SVM has become the most active and the most widely used method of pattern recognition in data mining field just during the past 20 years,. The Core theory of support vector machine is mapping vector into a high-dimensional space, and establish an optimal hyper-plane H in this high-dimensional space (there is two separated hyper-plane in the two side of the optimal hyper-plane , H1 and H2) ,to enable the separation distance of the two separated hyper-planes achieve the maximum, as shown below.

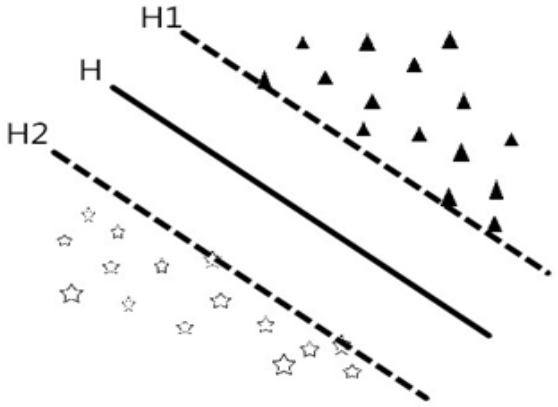

Figure1 schematic diagram of the optimal hyper-plane

SVM classification prediction process is probably like that, assuming that a known sample data is composed of n independent variables $(X_1, X_2, ...X_n)$ and a dependent variable (Y). Divide existing data sample into two portions, train samples Data1 and validation samples Data2. Unknown sample data namely forecast that sample Data3 only have n arguments, whose dependent variable is undefined. Firstly, get training model through training trained samples, and then verify the validity of the training model using the verification model, make actual prediction of the unknown predicted samples finally. Assume the testing samples Y is unknown when verify the validity of the model, and then get the predictive value of the verified sample variable by predicting independently,

then compare with the actual verified sample values Y to predict its accuracy and judge the rationality of the training model. Finally, select appropriate training model to make actual forecast on predicted sample. The process is shown below.

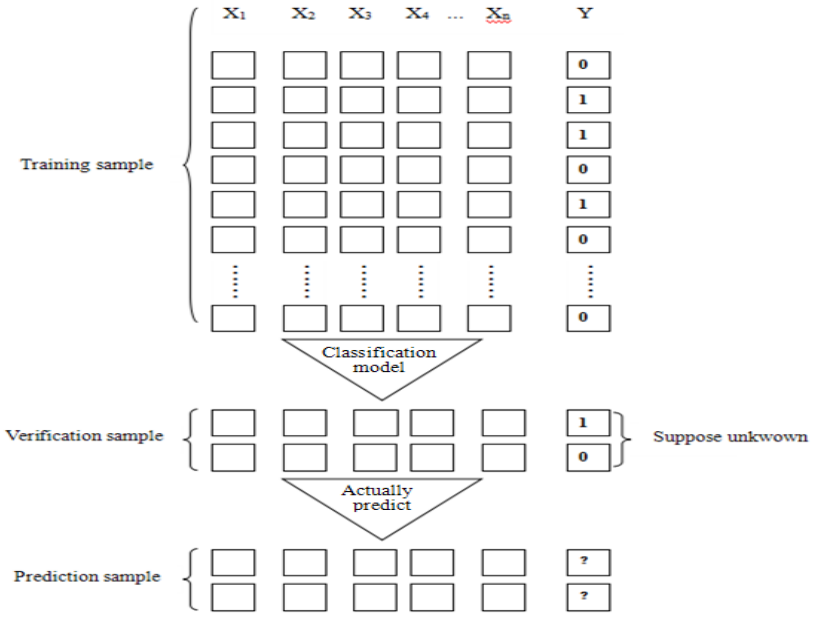

Figure 2 classification forecasting process diagram

3 Data Sources

Randomly Extract and arrange 500 data from customer information captured from the customer relationship management profile of an Agricultural Bank of Changsha, exclude their name, ID number and other properties which is unrelated with mining, leaving eight independent variable properties and one arguments property, which are gender (X_1), age (X_2), domicile (X_3), the nature of work (X_4), education level (X_5), monthly income (X_6,yuan), deposits (X_7,yuan), loans (X_8,yuan), and customer type (Y). Three attributes, monthly income, deposits and loans, have specific value, while other properties have no specific values. we classified other attributes according to the bank's statistics, as shown in the following table.

Table 1 No specific numerical attribute classification

Sex		Age		Domicile		Working Nature		Education level		Customer type	
Male	1	0-20	0	Changsha	1	government agency	0	Doctor above	0	VIP	1
Female	0	20-30	1	Not Changsha	0	institutions	1	Doctor	1	ordinary	0
		30-40	2			Soe	2	Master	2		
		40-50	3			private enterprises	3	Undergraduate	3		
		50-60	4			self-employed	4	college	4		
		60-70	5			other	5	Secondary	5		

4 Modeling and classification prediction

4.1 Independent variable standardization

Data standardization contribute to the selection of SVM kernel function parameters and improve the training rates . The entire dataset variables of this text was standardized to [-1,1]in columns according to the formula below:

$$x_i' = -1 + 2(x_i - x_{\min}) / (x_{\max} - x_{\min}) \quad (1)$$

In this formula, X_i` is the normalized data,X_i is original data X_{max} and X_{min} refer to the maximum and minimum,arguments was standardized by software DPS6.55.

4.2 Applying correlation analysis method filtering arguments

Correlation analysis is a statistical analysis method which analyze and process the relationship between two random coordinate vectors, which can get an abstract numerical that reflects the degree of interaction and the associated direction between the two vectors by mathematical calculations, that is the correlation coefficient [4]. For vectors X and Y whose length is n, their sample numerical are x_i and y_i ($i, j = 1,2,....n$), the mathematical description of the correlation coefficients of X and Y are as follows:

$$R=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}},R\in[-1,1]$$

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i \qquad \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad (2)$$

If R is positive, then it shows that the two peer vector change in the same direction that they increase and decrease at the same time. If R is negative, then it shows that the two peer vector change in the opposite direction that one increase when the other decreasing. Numerical closer to 1 indicates a higher degree of linear correlation between the two peer vectors; Numerical closer to zero, indicates the lower degree of linear correlation. We apply DPS6.55 to execute correlation analysis on the eight independent variables and the dependent variable Y, the |R| are shown in the below table.

Table2 Independent variables and the dependent variable correlation coefficient

$\lvert R\rvert$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Y	0.3156	0.8799	0.8740	0.9184	0.9326	0.9257	0.9663	0.9542

We can know from Table 2 that the impact of gender (X_1) property on the classification of bank customers is weaker, it belongs to redundant information, therefore, remove the property and leave the remaining seven independent variables as a follow-SVM training analysis.

4.3 SVM prediction

There are five common SVM kernels: linear kernel (t = 0), polynomial kernel (t = 1, d = 2), polynomial kernel (t = 1, d = 3), The radial basis kernel (t = 2) and sigmoid kernel (t = 3), and the radial basis kernel function is the most commonly used. This text also adopt radial basis to predict. SVM forecasting process complete this process by using MATLAB (R2011a) program calls Toolbox LIBSVM2.9 [5]. LIBSVM 2.9 package involves three common procedures: mygridr_mex kernel function is used to search for the optimal parameters c, g (c [10, 10], g [10, 10]) automatically, svm_train samples is used for training, svm_predict is used for prediction. Every usage of each program and parameter settings can be found in references [8].

The step of SVM training prediction are as follows:

1)Carrying out 10-fold cross verification according to the size of training sample, finding out the optimal the radial basis kernel function parameter ,c and g ,based on the mean square error minimum principle:

[c, g, mse] = mygridr_mex (train_y, train_x, 10, '- s 0 -t 2');

2) Under the combination of optimal kernel parameter, getting the training predicted structure model from training samples by using svm_train;

model = svm_train (train_y, train_x, canshu);

3)Using the training predicted structure model to execute confirmatory prediction on confirmatory samples through svmp_redict.

predict = svm_predict (test_y, test_x, model);

4) If predict highly match with the dependent variable of the actual confirmatory sample,then save it and use it to predict the predicted sample further,if not, then return to the first step 1 to find out a optimal re-kernel function parameters , till get a reasonable model.

5) Executing actual prediction on predicted samples through svmp_redict and using reasonable model to obtain the predicted outcome.

Randomly select 100 data from 500 treated customers as confirmatory sample, keep the other 400 customers as training samples to verify the predict accuracy of the confirmatory samples and measure the reasonableness of the training model.

5 Comparative Methods and Results Analysis

5.1 Comparative Method

This text introduces BP neural network as a comparative model analogy the classification ability of SVM. Input the standardized independent variables and data which is screened by related analysis variables into BPNN and SVM to implement classification forecast. BPNN prediction process using MATLAB (R2011a) program called Built-in Toolbox ,Neural Network. BPNN classification prediction process can be listed as follows:

1) First, initialize net, the network structure, of BPNN.

net = newff (sc_tr_x, sc_tr_y); net = init (net); net.trainParam.showWindow = 0;

2) Get the optimal network structure net1 through training samples optimization method.

net1 = train (net, train_x, train_y);

3)Use step 2) to get the best network institutions net1 to execute simulated prediction on confirmatory samples ,so as to get the predicted numerical of dependent variable of confirmatory sample. predict = sim (net1, test_x);

4)If predict coincide highly with the actual dependent variable of confirmatory samples, then the optimal network structure net1 is reasonable and accurate, then save it and use it to predict the predicted sample further, if not, then return to the first step 2) to executing BPNN network structure optimization method , till get a reasonable net1.

5) Executing actual prediction on predicted samples through sim () combined with reasonable net1 to obtain the predicted outcome.

4.2 Analysis

There are 32 VIP users in the 100 randomly selected samples, the other 68 are average users. BPNN and SVM applied multi-step prediction method to this

100 confirmatory samples to predict, and the SVM prediction results are shown as below.

Table 3 predicted results of SVM confirmatory sample

Actual Y	1	0	0	1	0	1	1	0	0	0	1	0	0	1	0	0	1	0	0	0
Predicted Y	1	0	0	1	0	1	1	0	1	0	1	0	0	1	0	0	1	0	0	0
Actual Y	1	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	1	0	0
Predicted Y	1	0	0	0	1	0	0	1	0	1	0	0	1	1	0	0	0	1	0	0
Actual Y	1	0	0	1	0	1	1	0	0	0	1	0	0	1	0	0	1	0	0	0
Predicted Y	1	0	0	1	0	1	1	0	1	0	1	0	0	1	0	0	1	0	0	1
Actual Y	1	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	1	0	0
Predicted Y	1	0	0	0	1	0	0	1	0	1	0	0	1	1	0	0	0	1	0	0
Actual Y	1	0	1	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0
Predicted Y	1	0	1	0	0	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0

By analyzing the results, the overall prediction accuracy that BPNN applied on bank customers classification is 87.5% ,for the VIP users, the classification accuracy is 92.3%, while the prediction accuracy of SVM model established in this paper is up to 93%, and for VIP users, classification accuracy is even as high as 96.9%, it has more significant meaning for banks to hold high quality customers, and it only takes a few seconds for the SVM to predict the entire process , either time consumed or forecast accuracy, SVM is significantly higher than the BPNN method ,which indicates that SVM model's classification recognition ability is fast, reasonable and effective.

After establishing binary classification model on Bank of the customer, we can use this model to classify and integrate unknown customers.

6 Conclusions

Based on the customer data of an Agricultural Bank of Changsha and support vector machines, this file established a SVM binary classification model for VIP users and ordinary users, which is better than BPNN binary classification model established by BP neural network, either in the aspect of time consuming or prediction accuracy,it is a relatively efficient classification method to classify bank's customers. It becomes easier to implement bank's customer relationship strategies with SVM-based binary classification model, which allows bank to attract more high-quality customer resources and reduce the loss of existing

customers. It also has a significant reference value for the design of the bank's customer relationship management system.

References

[1] Smith Wendell R.. Product differentiation and market segmentation as alternative product strategies[J]. Journal of Marketing, 1956, 11(7): 3-8.
[2] Dimitri P. B. and John T.. Neuro-Dynamic Programming[M]. Athena: Athena Scientific Press, 1996, 512.
[3] Chen Yuan. Improved support vector machine extremely applications [D]. Hunan Agricultural University, 2012.
[4] David R. H., Sandor S., John S. T.. Canonical Correlation Analysis: An Overview with Application to Learning Methods[J]. Neural Computation, 2004, 16(12): 2639-2664.
[5] Chang C. C., Lin C. J.. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27.