2014

ISSN : 0974 - 7435 Volume 10 Issue 13

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(13), 2014 [7025-7031]

Research of risk assessment and prediction based on support vector machine

Dakai Li1*, Yu Li[2], Zhang Qi-Wen[3]

[1]Shandong University of Finance and Economics,Shandong Jinan, 250000, (CHINA)

[2]Qilu University Of Technology,Shandong Jinan, 250000, (CHINA)

[3]Navy Submarine Academy,Shandong Qingdao, 266000, (CHINA)

ABSTRACT

Building a suitable credit risk evaluation model is very important, that is because loan business is one of the most important assets of commercial Banks. In this paper, we introduce a kind of learning algorithm which contains small sample learning and construct a new method which is called support vector machine (SVM). SVM is developed based on a new theory that is popular used in the field of intelligent learning system in recent years. The credit risk assessment model of commercial bank is established finally. By multiple discriminate analyses and the comparison of neural network model, confirms the validity and superiority of the method when be used in risk assessment.

KEYWORDS

Neural network model; Intelligent learning system; Nonlinear technology; Non-normal distribution; Machine learning rule.

INTRODUCTION

Support vector machine (SVM) method which is widely used in the research area of artificial intelligence, is based on the theory of statistical learning which is concluded by combining VC dimension theory with principle of risk minimum. It is based on complexity and learning ability of the model to seek the best compromise, and finally get the best generalization ability.

Support vector refers to those training sample points on the edge of the area. The "machine" is actually an algorithm. In the field of machine learning, often treat some algorithms as a machine. SVM (Support vector machine) as a learning mechanism is similar to the neural network, but unlike neural network, the SVM using the mathematical method and optimization technique.

The most important assets business of commercial Banks is loan business. Therefore, how to effectively prevent such credit risks as borrowers default and bankruptcy and to build a suitable credit risk assessment model are particularly important. Statistical model and neural network model as risk assessment models are widely adopted in the world.

Forecasting model based on statistical discriminate method was proposed after illuminating study by Fisher in 1936[1]. Among them, Z score model established by Altmanis based on multiple discriminate analyses (MDA) and the improved ZETA model on the basis of this are most famous. The introduction of statistical method overcomes the poor comprehensive analysis ability and poor quantitative analysis ability of the traditional ratio analysis, but statistics is studying the asymptotic theory when the sample tends to infinity and many strict assumptions are existent, so it is difficult to achieve the ideal effect in the real applications.

The neural network (NN) technology is applied to the field of credit risk assessment in 1990’s[2], NN is a kind of nonlinear techniques which hasn’t any requirement to data distribution, can effectively solve the non-normal distribution and nonlinear problem of the credit risk assessment, its performance is superior to the statistical model in general, but the NN as a black box cannot have an explanation to its output, excessively depending on skills is the biggest bottleneck in practical application.

There are quite a few of domestic scholars research on credit risk evaluation question[3,4], but due to less accumulation of data, statistical method and the neural network method cannot fully play their effectiveness. For this situation, this paper introduces a general SVM learning algorithm based on the theory of the small sample learning[5], and uses it to establish credit risk assessment model of commercial Banks, and has achieved good results.

SVM THEORY

Statistical learning theory

Based on the research of statistics, statistical learning theory is developed which is a special theory of studying machine learning law under the condition of small sample which was put forwarded by Vapnik VN; it is an important development and complement of traditional statistics. The theory developed a new theory system for small samples statistical problems, the statistical inference rules (SIR) consider the asymptotic performances requirements. Besides, it also pursuit optimal results under the existing condition of limited information. Developing a new universal learning method whose name is SVM theory (Support Vector Machine).

Classification algorithm of SVM

SVM, which is under the condition of linear separable case, is developed from the optimal separating hyper plane. In Figure 1, we show the basic principle of SVM. On the diagram, solid and hollow points on behalf of the two kinds of samples, H is the classification hyperplane, H_1 is sample that is most close to H and H_2 is hyperplanes which is parallel to H. They have equal distance to H, the distance between them is termed classification interval. The hyperplane is the one that separates the two types with biggest distance, for such an ill-posed problem of classifying two kinds of samples, the optimal separating hyperplane has maximum stability and high generalization ability[5].

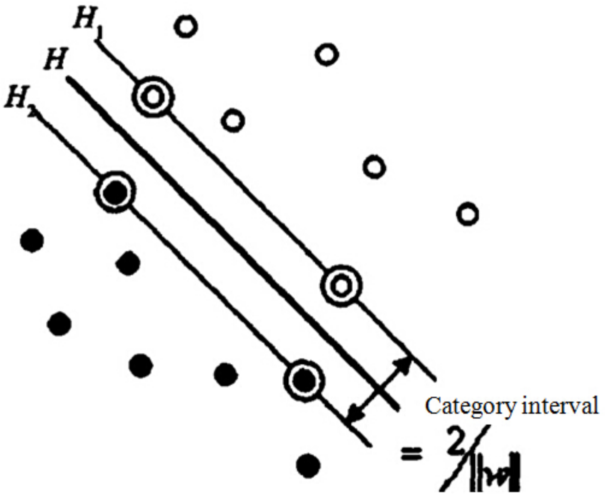

Figure 1 : Optimal hyperplane for linearly separable patterns

The general form of classified hyperplane equation can be written as $x \cdot w + b = 0$. It can be normalized as $\left(x_i, y_j\right), x_i \in R^d, y_i \in \left\{+1,-1\right\}, i = 1,\ldots,n$, which is meeting

$$y_i\left[\left(w \cdot x_i\right)+b\right]-1 \geq 0, i=1,\ldots,n \tag{1}$$

At this time, the category interval is equal to $2/\|w\|$, so let the interval biggest means let the $\|w\|^2$ smallest. The hyperplane which meets condition (1) and lets $\frac{1}{2}\|w\|^2$ smallest is called the optimal separating hyperplane. The training sample which meets the equal sign of (1) is named Support Vector.

As a conventional approach, Lagrange optimization method can be employed to transform the optimal classify surface problem into its dual problem**[6]**, which is converted to an quadratic function optimization problem with inequality constrained constraints, the unique solution is existent. The optimal classification function after solute the problem can get as the follows.

$$f\left(x\right)=\operatorname{sgn}\left\{\left(w \cdot x+b\right)\right\}= \\ \operatorname{sgn}\left\{\sum_{i=1}^{n}\alpha^* y_i\left(x_i \cdot x\right)+b^*\right\} \tag{2}$$

Among them, a$_i$* is the corresponding Lagrange multiplier for each sample, there is only part of the a$_i$* (usually a few) is not zero, the corresponding sample is support vector. b* is classification threshold and it can be obtained with any support vector, or obtained by any pair of support vector among two types of values.

In the case of linear inseparable sample set, a relaxation can be increased in the condition (1),

$$y_i\left[\left(w \cdot x_i\right)+b\right]-1+\xi \geq 0, i=1,\ldots,n \tag{3}$$

$$\left(w,\xi\right)=\frac{1}{2}\|w\|^2+C\left[\sum_{i=1}^{n}\xi_i\right] \tag{4}$$

The target is let (4) to be the smallest, namely to construct a soft interval, compromise to consider the least fault samples and the largest classification interval, the optimal classification plane can be generalized, where C>0 is a constant. It controls the degree of punishment to wrong classified of sample.

Nonlinear problem can be changed into a linear problem in high dimension space by nonlinear transformation, and solve the optimal classification plane in the high dimensional space. In general the nonlinear transform is more complex and difficult to achieve, but in fact, as long as using function K (xi, xj) to replace the inner product of the space, linear classification after a nonlinear transformation can be achieved[5], thus avoiding the concrete form of nonlinear transformation.

$$f\left(x\right)=\operatorname{sgn}\left\{\sum_{i=1}^{n}\alpha^{*}y_{i}K\left(x_{i}\cdot x\right)+b^{*}\right\} \tag{5}$$

THE CREDIT RISK ASSESSMENT MODEL BASED ON SVM

The establishment of the index system

The financial condition of enterprise can directly determines whether it can pay for its debt-service payments on time or not, we are limited to establish credit risk assessment model from a financial point of view, other non-financial indicators or qualitative analysis can be dealt with by expert system, this paper does not make discussion for this. According to the commonly used criterion and expert's advice, we use eight indexes to measure the financial situation of enterprises. They are return on sales (sales profit/sales revenue), the liquidity ratio, the assets and liabilities ratio (debt/total assets), the return on total assets (ebit/total assets), the multiple of interest safeguard, current asset turnover (net annual sales revenue/average current assets), working capital ratio/total assets (working capital), ratio of operating cash flow (operating cash flow/assets).

Dealing with sample data

The experimental data in this paper comes from a commercial bank in Fujian province, including 1,570 companies' financial data, among them 1,279 companies financial situation is good, the risk of bank loans to them is smaller, and we call these enterprises as performance enterprises. The rest of the 291 companies are in bad financial situation, if give loans to them, the probability of default is big, we call these enterprises as default enterprises. These two types of sample size are with big difference, if directly applied in SVM learning, the optimal classification surface of SVM can trend to samples with small density, can cause larger classification error in the future predict[7]. In order to get good performance of the SVM, must carry on pretreatment to the samples, make two kinds of sample with same size. In this paper, randomly selected 290 enterprises from 1,279 performance enterprises and formed a scale of sample set of 581 companies with 291 default enterprises. On the basis of this, we divided the sample set into two sets which are test sets and training sets respectively. Therefore, it is better to reflect the SVM's learning ability and model generalization ability for small sample data. In the training sample set, we randomly choose 30% (174) companies on the premise of keeping two kinds of sample size is equal. The 174 companies are employed to build the SVM model. In addition, the rest of the 70% (407) as a test sample set is used for testing the generalization ability of the model. Details of the sample set are shown in TABLE 1.

TABLE 1 :The distribution of the dataset after sampling

	Sample set size	performance enterprise number/ Percentage	default enterprise number/ Percentage
Training sample set	174	87/50	87/50
Test sample set	407	203/49.88	204/50.12
All sample sets	581	290/49.91	291/50.09

TABLE 2 : Comparison of the discriminate results of SVM, BP and MDAT

Model Type	Training sample set			Test sample set		
	Accuracy/%	Type I error rate/%	Type II error rate/%	Accuracy/%	Type I error rate/%	Type II error rate/%
MDA	78.74	7.47	13.79	78.38	6.14	15.48
BP	86.78	8.05	5.17	81.57	9.83	8.60
SVM	85.06	5.17	9.77	83.29	5.65	11.06

The svm model structure

Based on the above analysis, constructing sample set (x, y), the dimension of x is 8, y is the category of the sample properties, for performance enterprise y = 1, for default enterprises y = - 1. The different kernel functions in the SVM will form different algorithms. The most kernel function study present is the mainly polynomial function, radial basis function (RBF) and Sigmoid function. Handwritten numeral recognition experiment[5] shows that using the above three kinds of different kernel function of SVM can get similar performance results, and the distribution difference of the support vector is not big. For a specific problem, there is no general method for how to choose kernel function[6]. In this paper, kernel function is the most commonly used function when conducting SVM model.

$$K\left(x, x_i\right) = \exp\left\{-\frac{\left|x - x_i\right|^2}{\sigma^2}\right\} \tag{6}$$

THE RESULTS ANALYSIS

TABLE 2 lists the results of the SVM model, including the discriminate result in the training sample set and test sample set, at the same time compared with the model constructed by MDA and the neural network. The results of the MDA model are obtained by SPSS software analysis. In the neural network which employs BP algorithm, the number of hidden layer and the target error are obtained by cross validation method. In this paper, the target error is set as 0.1 and the number of hidden layer is set as 16. Because of the neural network method is not a stable method, so the result of neural network in the table is the average result of 10 times. First kind of mistake in the TABLE 1 is judge default enterprise as performance enterprise; the second kind of mistake is judge performance enterprise as default enterprise. Altman[8] pointed out that the loss of type 1 error is 20 to 60 times of type 2's. So it can evaluate model from two aspects, first is the overall evaluation accuracy; second is error rate of class 1.

It can be seen from TABLE 2; the overall accuracy of the SVM in the test sample concentration (i.e. prediction accuracy) reached 83.29%, significantly better than MDA model's 78. 38%, this is slightly higher than the neural network's 81.57%. In the process of the experiment, we consider the overall accuracy when constructing model, from this perspective, the SVM model has a certain advantage. From TABLE 2, we also found that the type one error rate of the SVM model is the smallest among the three kinds of models (but its performance deals with the second type errors is poorer than the neural network model), which may be caused by the distribution of sample data itself, in the results of the MDA model[9], we can see the same trend (first category error rate lower than that in second category obviously). To this problem, at present, we cannot do the interpretation of the mechanism, but we want this kind of phenomenon happens. If modeling for other new data, bigger error rate of category one of the SVM model also can appear. To effectively reduce the category one error rate, big punish coefficient can be added to category one error when we construct the SVM model, but it also may reduce the overall accuracy of the model.

From TABLE 2, we can also compare the robustness of the model. For the training sample set, the overall accuracy of neural network is the highest, reached 86.78%, the second is the SVM model of 85.06%[10], the worst effect is MDA of 78.74%. In the test sample set, accuracy (prediction accuracy) has degrees of decline. Neural network has the maximum change of 6.00%. The rate of SVM is 2.08%. The smallest rate is MDA, is only 0.46%. It can be seen that in three kinds of models, neural network model has the worst robustness, MDA's robustness is the best, the robustness of the SVM model, though not the best, but also maintain a good level, can satisfy the requirement of practical application.

CONCLUSIONS

SVM is a kind of universal learning algorithm which is based on the theory of the small sample learning. It has the strict theoretical basis[11], can well solve practical problems such as small sample, nonlinear, high dimension and local minimum point and so on. This paper proposes a commercial bank credit risk assessment model based on SVM, by comparing with the MDA model and neural network model found that the SVM model can not only improve the prediction accuracy, but also effectively reduce the mistake ratio of category one, the robustness of the model itself is strong. It has good development prospects. It is worth for further study, future work will be taken from the following aspects:

As a result of the uneven distribution of actual sample data type, this paper randomly strips out some samples, causes the waste of information[12-13]. How to construct SVM model with uneven distribution sample data under the premise of maximize use sample information is a direction of future work.

For class one and class two errors, in this paper, the SVM model adopted the same punish coefficient. Because of the two kinds of errors in the actual losses are different, and the difference is larger, if introducing different penalty coefficient for different fault when constructing model, can possible effectively prevent type one error.

In this paper's study, SVM solves only one or two types of classification problem. The SVM can be used to more complicated credit grade evaluation problems, so it can better reflect the borrower's credit situation, and provide more powerful and more detailed auxiliary tool for commercial Banks when making decision.

REFERENCES

[1] E.I.Altman; Corporate Financial Distress: a Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy [M], New York: John Wiley & Sons, **(1983)**.

[2] K.Y.Tam; Neural Network Applications Models and the Prediction of Bank Ruptcy [J], Omega the International Journal of Management Science, **19**, 429-445 **(1991)**.

[3] C.Zhitong, F.Jiazhong, C.Hongpingn, H.Guoguang, E.Ritchie; "Support Vector Machine Used to Diagnose the Fault of Rotor Broken Bars of Induction Motors", IEEE Electrical Machines and Systems (ICEMS), **2**, 891-894 **(2003)**.

[4] Nello Cristianini, John Shawe-Taylor; An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, ISBN 0-521-78019-5 **(2000)**.

[5] Zhang Xuegong; On Statistical Learning Theory and Support Vector Machine, Automation Journal, **26(1)**, 32-42 **(2000)**.

[6] Yonghua Han, Yaming Wang, Yun Zhao; Support Vector Machine-based image segmentation approach for automatic agriculture vehicle, 2012 International Conference on Image Analysis and Signal Processing (IASP), p 5 **(2012)**.

[7] J.C.Christopher Burges; "A Tutorial on Support Vector Machines for Pattern Recognition". Data Mining and Knowledge Discovery, **2**, 121-167 **(1998)**.

[8] E.I.Altman; Commercial Bank Lending: Process, Credit Scoring and Costs of Errors in Lending [J], J, Financial and Quantitative Anal., **15**, 813-832 **(1980)**.

[9] Harris Drucker, J.C.Chris, Burges, Linda Kaufman, Alex Smola, Vladimir Vapnik; "Support Vector Regression Machines". Advances in Neural Information Processing Systems 9, NIPS 1996, 155-161, MIT Press, **(1997)**.

[10] Vojislav Kecman; "Learning and Soft Computing - Support Vector Machines, Neural Networks, Fuzzy Logic Systems", The MIT Press, Cambridge, MA, **(2001)**.

[11] Chien-Yi Wang; "The fusion of support vector machine and Multi-layer Fuzzy Neural Network". Machine Learning, Jun, **(2012)**.

[12] C.Cortes, V.N.V.apnik; “Support Vector Networks”, Machine Learning, **20(3)**, 273-297 **(1995)**.

[13] N.Cristianini, J.Shawe-Tayor; “An Introduction to Support Vector Machiines and Other Kernal-based Learning Methods”, Cambridge University Press, 189 **(2000)**.