International Journal of Business and Management; Vol. 8, No. 12; 2013
ISSN 1833-3850 E-ISSN 1833-8119
Published by Canadian Center of Science and Education

Research on Comprehensive Evaluation of Small and Micro Businesses on Improved Support Vector Machine of Imbalanced Multi-Classification

Chen Ying[1]

[1] Sydney Institute of Language & Commerce, Shanghai University, Shanghai, P. R. China

Correspondence: Chen Ying, Sydney Institute of Language & Commerce, Shanghai University, Chengzhong road 20, Jiading district, Shanghai, 201800, P. R. China. E-mail: dorothychen@staff.shu.edu.cn

Received: March 22, 2013 Accepted: May 17, 2013 Online Published: May 24, 2013

doi:10.5539/ijbm.v8n12p137 URL: http://dx.doi.org/10.5539/ijbm.v8n12p137

Abstract

Comprehensive evaluation of small and micro businesses was significant problem for the researchers and the managers in corporations for these years in China. Small and micro businesses played an important role in taxes payment and social employment, providing convenience of people's life in every country. The article collected the data of small and micro businesses in east of China in 2010, amended traditional algorithms of support vector machine, analyzed from four aspects, including financial condition, internal business operating and business growth, and used imbalanced multi-classification algorithm. The results of accuracy rate were acceptable and could be proved.

Keywords: comprehensive evaluation, small and micro businesses, support vector machine, imbalanced multi-classification

1. Introduction

From 2008, cost and expenses of small and micro businesses increased rapidly and the income of them decreased sharply in China, because the price of raw materials and labor was much higher than before. In addition to labor costs rising, many workers transferred to the central and western regions and the price of labor was increased in small and micro businesses. Small and micro businesses couldn't raise the prices of product and services to eliminate the influence of cost rising of production, since there was a highly competitive market with the decreasing of profit growth, no core technology, and the high degree of homogeneity of products. The orders and bills of small and micro businesses had gradually declining. The financing channels of small and micro businesses were narrow. One part of the reason was small and micro businesses lack of completely financial statements, high-quality collateral assets; another part of the reason was credit funds absent, and commercial banks were lack of motivation to provide financing services for small and micro businesses. The financial sources of small and micro businesses relied mainly on their own capital accumulation, borrowing from their relatives and friends, and supply chain financing. The financing gap of supply and demand was large, credit loan of small and micro businesses declining, interest rates rising, and the loan period being shorter.

During the process of Chinese economic reform in these thirty years, most of small and micro businesses belonged to the labor-intensive enterprises, offered social employment more than large enterprises, and had the lower threshold of employment, operation and management of them was flexible, and adaptable, making great contribution to national economic development, tax payment, providing services for people's life, and meeting the needs of all the different groups. Every country supported the development of small and micro businesses. Similarly, Chinese government had issued many policies to support small and micro businesses development

On the one hand, the credit environment for the enterprises, especially for small and micro businesses in our country was not good, some enterprises and the managers of enterprises were not aware of the importance of credit recordings. There were many credit fraud and breach of contract, when the incident occurred, and the risk of enterprise bankrupting was high. The creditors could not understand the cooperation business financial condition and credit situation completely, and distrust of small and micro businesses were more strongly. The whole society and the macroeconomic environment gave the less support and help to small and micro businesses, and there were lack of the laws and regulations of financial services and financial service system to small and

micro businesses. The large scale enterprises could loan from commercial bank. However, the growth of small and micro businesses obtained funds to raise capital mainly relying on privately lending and little chance receiving the loans from commercial bank, there was no more financing channels for small and micro businesses. Thus, financing difficulties of small and micro businesses was the most important problem.

On the other hand, commercial banks met the critical competitions and hoped to enlarge the customers and increase revenues and income. Meanwhile, the loans for small and micro businesses were always higher interest rates, loan amount was less, and the loan period was short, higher risk, lending activities was difficult to get legal protection and increase default risk.

Therefore, the researchers and managers had been trying to study how to solve the financing difficulties of small and micro businesses of our country, for the development of small and micro businesses financing mode, system, mechanism and system solutions, to promote the development of small and micro businesses s, the upgrading of the industrial structure, and maintain social stability, promote the steady development of the national economy.

Furthermore, it was necessary to accelerate the construction of social credit system, reduce the cost of information collection. The relevant departments needed to guide small and micro businesses to abide by the small and medium-sized enterprise accounting system and standards and to standardize financial reports, and aimed at the defects of the financial data of small and micro businesses not to be standardized. Thus, the comprehensive evaluation system should be built up and included more non-financial information into the scope of examination, reflecting comprehensive perspectives of small and micro businesses.

2. Literature Review

Support vector machines (SVM) was firstly put forward by Corinna Cortes and Vapnik in 1990's, which always was used to solve problems with nonlinear and high-dimensional pattern recognition. Support vector machine could analyze time series data and classified question data, and solve the problem of comprehensive evaluation and forecasting problems well.

Sample data classification problems in real life are imbalance classification, also is the number of each class of data samples are inconsistent The need for algorithm constantly breakthrough, find a good algorithm can solve the imbalanced data classification, improvement or data sampling technique, undersampling or oversampling, the data set to achieve balance, constantly improve the classification accuracy rate. Sampling is to sample more categories, number of less sampling, reached number and less data type of sample is close to the results, to meet the two types of equilibrium results in the data distribution, but may also remove an important problem with sample data, which leads to the majority class information loss.In contrast, over sampling is to copy the sample data of the minority class, repeated sampling for the minority class samples, the amount of sample data in order to achieve the most kind of close, will increase the amount of computation and computation task.

At present, cost sensitive learning is mostly used algorithm for imbalanced classification problem, which distributed different costs to different training samples, usually learning cost of small data sample is lower than others, in order to achieve balanced sample classification algorithm results. Zhou Z. H. etc. firstly (2006) used this method to solve multi-classification problem, then (2006) combined cost sensitive learning with neural network method to improves the measuring precision. Yan M. S. etc. (2007) integrated cost sensitive learning into average boosting method, and received better results. Many scholars studied cost sensitive learning algorithm. Improved support vector machine algorithm is the effective method to solve imbalance classification problems.

3. Research Methodologies

Support vector machine is established for separating hyperplane to classify each class. The training data set of binary classification problem is xi∈Rn,i=1, 2, …, l, corresponding classification level is yi∈{-1，1}，i=1,2,…,l, the formula is linear soft margin algorithm，the formula of optimization problem is:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=l}^{l}\xi_i$$

$$\text{s.t.} \quad yi((w\bullet xi)+b)\geq 1-\xi i，\ i=1，\ \ldots,l$$

$$\xi i\geq 0，\ i=1，\ \ldots，\ 1$$

The parameter C is balance the training accuracy and generalization ability.ξi is slack variable, w∈Rn is weight vector, to explain the position hyperplane separating each kind of space. B is the moving error of hyperplane location. Interpretation function is:

$$f(x)=sgn((w.x)+b)$$

Lagrange multiplier method solved to the dual problem is used to find the solution of this kind of optimization problem:

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j (x_i \bullet x_j)\alpha_i \alpha_j - \sum_{j=1}^{l} \alpha_j$$

$$\text{s.t.} \quad \sum_{i=l}^{l} y_i \alpha_i = 0$$

$$0 \le \alpha_i \le C \text{ , i=1,…,l}$$

Usually，the solution α *of dual problem is less, the corresponding decision hyperplane is decided by several support vector.

$$f(x)=sgn((w^* \bullet x)+b^*)=\text{sgn}(\sum_{i=l}^{l} \alpha_i^* y_i (x_i \bullet x) + b^*)$$

There are many practical problems are not solved by linear classification method, support vector machine is the method, which inputted n-dimensional sample data into nonlinear function Φ(•)，to map to the high dimensional space, used kernel function K (x_i, x_j) to replace the nonlinear function (Φ (x_i), Φ(x_j)), and obtained the decision function:

$$f(x)= \text{sgn}(\sum_{i=l}^{l} \alpha_i^* y_i K(x_i \bullet x) + b^*)$$

System comprehensive learning algorithm is the kind of machine learning method for high dimensional classification and regression problems. Different from the independent classification method, system comprehensive learning algorithm is the classification method to construct a series basic classifiers, then according to combined results from each classifiers, as the new sample data to classify continuously, that is, system comprehensive learning algorithm is divided into two major steps: the first step is to design and build up a series basic classifiers, the second step is to integrate the existing classifier, to do the weighted or unweighted data treatment processing, and to get the final classification result. Therefore, system comprehensive learning algorithm is widely used due to be improved learning stage efficiency and accuracy of the classification algorithm. Improved the accuracy of system comprehensive learning algorithm need much more classifiers offsetting the classified errors between the kinds of classifiers, and improving the accuracy and efficiency of the classification algorithm. Many scholars researched many aspects of classification algorithm, including subdividing and segmenting the data samples of training sets, adjusting control variables for attribute, controlling output results from the classifies, and using random number in machine learning algorithm, simultaneously, to introduce composite classifiers into it. Combined different compound classifier algorithm and different integrated comprehensive algorithm could obtain different system comprehensive machine learning algorithms.

Vector quantization, also called pattern matching quantization processing, usually used for compression of data loss. If the training set was composed by l vectors, T={x_1,x_2,…,x_l },vector quantization used set T, Code={c_1,c_2,…,c_M}，vector c_i, every vector （i=1，2，…M）also called support vector, all the support vectors to form a series of hyperplane, segmented higher dimensional space , forming a plurality of separating space:

$$V=\{V_1,V_2,...,V_M\}, \quad V_i=\{x \in T: \ \| x-c_i \| \leqq \| x-c_j \|, j \neq i\}$$

$$i=1, \ ..., \ M$$

Each separated space had support vectors. It is necessary to find the classifier rules, and super support vector V, minimize the error of classification and structural risk, which is:

$$D=\frac{1}{Mn}\sum_{i=1}^{l} \left\| x_i - \phi(x_i) \right\|^2$$

The n was the number of support vector dividing hyperplane.

This algorithm mainly includes three steps: segmentation, training, aggregation:

a) Because undersampling and oversampling methods were always defective, it was necessary to overcome the problem of information loss or interference noise increasing, it was required to modify sampling technique. Therefore, to reclassify conventional sample data packet achieved the effect of balanced classification. The negative categories (negative subsets) of collected data sample was less than and needn't further sub-classify,

and the data of positive categories (positive subsets) required detailed subdivision, the classification of samples classified into k subcategory, according to the collected data, after data cleaning the samples were divided into 7 to 9 subcategories.

b) Treating the sample of less data firstly, for sample data in 2010, the seventh classes of samples only had 2 data, however, it was negative category (negative subsets). If the traditional method was used only directly removing the negative categories, a detailed classification only for the positive class sample data. Divided into the learning algorithm of K classification, support vector machine was used to classify the sample data and keep negative data sample category. Negative data category had been sampling classified and this small amount of sample must be retained as important role.

c) After training, the analysis method for this kind of problem integrated for each independent basic categories, according to the distance between each feature vector for each class of sample data, and distinguished different categories, firstly isolated the sample data of negative categories (negative subsets), then separated the positive class (positive subsets) of sample data, if for the data sample of new test set were distinguished the data categories by previously formed learning classifier of support vector machine.

4. Conclusion

The variables had financial perspective, internal business perspective and growth perspective ,and included the life of business, educational level of business manager, debt ratio (total liabilities / total assets), current ratio (total current assets/total current liabilities), owner's equity in the enterprise, accounts receivable turnover, the growth of operating income, the growth rate of sales revenue, rate of return on owners equity, the length of company setting up, the credit record of the business, industry policy for the business, the local economic environment of business, the assets owned by business managers, the ownership of business area (leasing or owner) , rate of produce and sale (the volume of sale/the volume of production) , the equipment utilizing rate. After data standardized, support vector machine method classifier selection was carried on by Matlab 7.0 (2009 edition) and libsvm (software written by National Taiwan University Professor Lin Zhiren), to do 1-7 class multi-classification. Nuclear function is Gaussian radial nuclear function, which was mainly used to solve the classification problem of the lack of prior experience. Variable parameters were the error penalty parameters and Gaussian kernel parameter. C represented in the following table was the error penalty parameter, which meant error tolerance, the higher was the value, and the lower were the tolerable errors. This kind of Support vector machine classification model chose gamma radial basis function as kernel function, to implicitly determine the mapping data to the distribution of the new features space. The formula of experimental accuracy is:

Accuracy rate of accuracy = accurate sample classification number / number of sample data.

The selection of parameters setting of the method was gradually narrow range. Firstly it took a wide range of parameters, such as the 100-10000 experiment, and then divided the most accurate range, gradually narrowing the scope of the trial after repeated testing analysis, found the following range of parameters:

c = (1800, 1900, 2000, 2100, 2200, 2500, 2700)

gamma = (0.001, 0.003, 0.005, 0.007, 0.009)

The test results of this range were relatively better than the other parameters range selection of multiple parameters for unbalanced multi-classification problem, and the wide range of parameters were better than a single parameter experiments to explain the effectiveness of the method. Analyzing training sets and testing sets obtained the results shown in the following table:

Table 1. The accurate rate of multi-classification

gamma \ c	1800	1900	2000	2100	2200	2500	2700
0.001	0.7091	0.697	0.697	0.7091	0.7091	0.7152	0.7091
0.003	0.7758	0.7697	0.7636	0.7576	0.7636	0.7697	0.7818
0.005	0.7879	0.7818	0.7818	0.7818	1	0.7879	0.7697
0.007	0.7879	0.7818	0.7879	0.7879	0.7818	0.7818	0.7879
0.009	0.7818	0.7818	0.7818	0.7818	0.7818	0.7758	0.7758

The results of comprehensive evaluation of small and micro businesses on improved support vector machine of imbalanced multi-classification were verified effective reasonable and conceivable. The task of research in future should be focus on how to collected the other data of qualitative variable and improve the accuracy of imbalanced multi-classification, extending imbalanced multi-class classification to solve other social problems.

And it was provided the policy suggestion to help small and micro business: firstly, it is to reduce tax burden of small and micro businesses, and strengthen the various supporting policies of small businesses. Secondly, small and micro businesses need to need to strengthen the development of brand and the channels of sales. Small and micro businesses are going to register their own brands and patents. Thirdly, the government should encourage financial institutions servicing to the development of small and micro enterprise, such as the credit company of small business loans, and its innovation of business model, product innovation, credit technology and management idea. Fourthly, it is to accelerate the construction of social credit system, reduce the cost of information collection, standardize and build up the comprehensive evaluation system of small and micro businesses.

However, the limitations of the research methodology were obviously. Whether or not the research methodology solve the imbalanced multi-classification of great deal of number sample data, The decision rules of classification based on the view that design evaluation criteria still needed further study, setting and adjusting the classification rules would affect the analysis results. And it needed further research to determine the more variables to increase or decrease in the future analysis.

References

Atish, P. S., & Huimin, Z. (2008). Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decision Support Systems, 46*, 287-299. http://dx.doi.org/10.1016/j.dss.2008.06.013

Cheng-Lung, H., Mu-Chen, C., & Chieh-Jen, W. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications, 33*, 847-856. http://dx.doi.org/10.1016/j.eswa.2006.07.007

Chong, W., et al. (2008). Customer credit evaluation model based on support vector machine in e-business environment. *China Management Science, 10*, 368-373.

Cristianini, N., & Shawe-Taylor, J. (2004). *Introduction to support vector machines* (Chinese ed.). Chinese electronics industry publishing house.

Fen, D., et al. (2009). Application of ant colony neural network to credit evaluation of small and middle enterprises. *Computer Technology and Development, 10*, 218-221.

Naiyang, D., et al. (2009). *Support vector machines: theory, algorithm and development*. Chinese scientific publishing house.

Shu-Ting, L., Bor-Wen, C., & Chun-Hung, H. (2009). Prediction model building with clustering-launched classification and support vector machines in credit scoring. *Expert Systems with Applications, 36*, 7562-7566. http://dx.doi.org/10.1016/j.eswa.2008.09.028

Steven, F. (2010). Credit scoring for profitability objectives. *European Journal of Operational Research, 202*, 528-537. http://dx.doi.org/10.1016/j.ejor.2009.05.025

Sun, Y. M., Mohamed, S., Wongak, C., & Wang, Y. (2007). Cost-sensitive boost for classification of imbalaned data. *Journal of Pattern recognition, 2007*(40), 3358-3378. http://dx.doi.org/10.1016/j.patcog.2007.04.009

Wen-bing, X., et al. (2007). Based on support vector machine model for credit assessment and risk. *Chinese science abstract, 22,* 284.

Wenjing, C., & Yingjie, T. (2010). Lp-norm proximal support vector machine and its applications. *Procedia Computer Science, 1*, 2411-2417.

Wun-Hwa, C., & Jen-Ying, S. (2006). A study of Taiwan's issuer credit rating systems using support vector machines. *Expert Systems with Applications, 30*, 427-435. http://dx.doi.org/10.1016/j.eswa.2005.10.003

Xiufu, S., et al. (2009). *SVM prediction method for assessing the financial credit rating of the listed companies in stock market.* Doctoral thesis in East China Normal University.

Yi-feng, Z., et al. (2005). A comparison of multi-classification methods for credit risk assessment in commercial banks. Proceedings of *the 24th Chinese control conference,* 1734-1737.

Zhi-chun, X., & Zong-jun, W., et al. (2008). Introducing the non-financial factor empirical study on early

warning model of credit risks of the small and medium-sized enterprises. *Financial Theory and Practice, 6*, 3-6

Zhou, Z. H., & Liu X. Y. (2006a). On multi-class cost-sensitive learning. Proccedings of *the 21st National Conference on Artifieial Intelligence* (pp. 567-572), Boston.

Zhou, Z. H., & Liu, X. Y. (2006b). Training cost-sensitive neural networks with methods addressing the class imbalance Problem. *IEEE Transaetions on Knowledge and Data Engineering, 18*(l), 63-77.