International Journal of Business and Management; Vol. 9, No. 4; 2014
ISSN 1833-3850 E-ISSN 1833-8119
Published by Canadian Center of Science and Education

Research on Imbalanced Multi-Classification of Performance Evaluation of Small and Medium-Sized Enterprises

Chen Ying[1]

[1] Sydney Institute of Language & Commerce, Shanghai University, 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: February 10, 2014 Accepted: February 28, 2014 Online Published: March 21, 2014

doi:10.5539/ijbm.v9n4p86 URL: http://dx.doi.org/10.5539/ijbm.v9n4p86

Abstract

Performance evaluation of small and medium-sized enterprises (SMEs) was the valuable problem for the researchers and the stakeholders of SMEs, which was not only for internal managers to control over the entire organization, but also for external stakeholders to familiar the SMEs. The author collected the data of 164 SMEs in east of China in 2011, and used two-step clustering method, k-means clustering method, system clustering method, neural networks method, and amended support vector machine method to analyses this problems of imbalanced multi-classification. The results of amended support vector machine were better than the results of the others.

Keywords: performance evaluation, small and medium-sized enterprises (SMEs), imbalanced multi-classification

1. Introduction

Performance evaluation was mainly reflected by the process of the operation and management of enterprises in business growth and development, achievements and contributions. Performance evaluation of SMEs referred to the operational efficiency and operating performance of the SMEs in a certain operational period, mainly including the ability of profitability, the ability of assets management, the ability of debt payment and the capability of managers or investors etc. The problems of performance evaluation of SMEs were a typically multi-classification problem, since the data in each sample set was not equal. Nowadays, there were many research findings in the problems of imbalanced multi-classification, such as image recognition, accurate face recognition, text recognition, disease diagnosis, and the evaluation of customer.

The problems of imbalanced classification included two classification problems and multi-classification problems. In most of the cases the problem of two classification problems were one of the sample data in a category far more than another, even higher than the proportion of 10:1. Besides, the two types of samples are not separable in the feature space, and especially in practice the minority class in the problems often played an important role.

For example in the problem of credit rating of the customers in commercial banks, it was always insufficient and limited the information of the default clients or the clients not receiving loans. And compared with the high quality clients already receiving the loans, the information of them were often incomplete, and did not record in the database of commercial banks, so it typically belonged to the problems of imbalanced multi-classification. Otherwise, it was also the discriminating problem to analyze the default customers in electric power companies and telecommunication companies, as well as the problem in diseases diagnosis. Currently plenty of data processing methods could not scientifically and reasonably classification solving the problems of imbalanced multi-classification. Since the prerequisite of the general algorithms required that the numbers of data in negative class and positive class or the different groups were equal or approximately equal in training set and test set, and in high dimensional space the features of the samples data in the training set could not be easily separated from each other. It would obtain the results relatively more accurate and effective. Otherwise the accuracy rate of classification was low, the speed and efficiency in training process was not high.

Performance evaluation of small and medium-sized enterprises was the typically imbalanced multi-classification

problem. From this kind of analysis the external stakeholders could know the level of performance of the certain SME, and the managers know the questions of how to evaluate performance, what measure to use, and what types of incentives to use.

2. Literature Review

The improvement of algorithm modified the inherent characteristics and original train of thought of the algorithm, which could adapt the model to analysis the problems with different characteristics of data set. Muhammad Atif Tahir proposed a novel inverse random under sampling (IRUS) method for the class imbalance problem, to severely select in sample data of majority class creating a large number of distinct training sets; and to present promising results for multi-label classification, applying on 22 UCI data sets. Zhen Jiang presented a new co-training style algorithm which employed a generative classifier (Naive Bayes) and a discriminative classifier (Support Vector Machine) as base classifiers, taking advantage of both methods. It showed that the experimental results on six datasets performed much better than the other methods, especially when the amount of labeled data was small.

At present cost sensitive learning, which was one of the algorithms the researchers most interested in, was distributed significantly different cost for different training sample data, which was usually given better learning cost to the small amount of data sample, to receive the classification algorithm similar to the balanced sample data. Zhou firstly used this method to solve the multi-classification problem, and then continuously combined with neural network to improve the algorithm. Yan combined this method with the average boosting method, receiving a better classification result. Many scholars researched the cost sensitive learning algorithm. Boosting algorithm integrated many classifiers as iterative method, usually classifying the more difficult sets of sample data, to obtain high classification efficiency after comprehensive analysis. It was a machine learning method with highly accurate classifier rate, also a way which many researchers solved the problems of imbalance classification. Many scholars combined the boosting algorithm with other method to obtain better classification results. The improved algorithm of support vector machine was an effective method to solve the imbalance classification problems.

So until now, the researchers in practice and Universities were actively exploring the solutions for the problems of imbalance multi-classification, it was necessary for the algorithms with a constant breakthrough, finding an available algorithm to solve those problems, and constantly improving the technique of data sampling, that was undersampling or oversampling, to achieve the data set balanced, and the accurate rate of classification. Undersampling was to sample less number of the data for larger number categories of data sets, reaching the results that less number of data sample was close to the other categories, to meet the results of balanced data distribution, yet it might be remove the sample data with the important information, and caused the information lost from the class with majority sample data. However, oversampling for the sample data in the minority class was copied the data, duplicating the sample data in the minority class, to achieve the sample data in different categories nearly equal, yet it would be increased the efficiency and accurate rate of calculation. Nowadays, SMOTE (Synthetic Minority Over-sampling Technique algorithm) was a mature and over-sampling of imbalanced classification algorithm, using some computing technology, to increase the new sample data in the minority class enlarging the number of sample data in the minority class, however, it easily lead to more noise interference.

3. Research Methodology

Support vector machine had transcendent advantage ability to solve classification and regression problems based on the statistical learning theory. It could be improved support vector machine algorithm by classic features, to solve classification and regression problems by different machine learning algorithms. At present, the algorithm of support vector machine deal with the classification problem with imbalanced sample data, in which the errors were still larger. Many researchers analyzed the problems of decision of adjustable boundary, and learning strategy of integrated system continuously to be added to the support vector machine learning algorithm to improve the accurate rate and efficiency of classification problems.

The basic formula of support vector machine was:

Given train sets $T=\{(x_1,y_1),\ldots\ldots(x_l,y_l)\}\in(R^n\times Y)$, in that $x_i\in R^n, y_i\in Y=\{1,-1\}, i=1,\ldots,l$. Kernel function $K(x,x')$ and penalty parameter $C>0$, construct and solve convex quadratic programming problem:

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

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

$0 \le \alpha_i \le C$, i=1,…,l.

The solution $\alpha^* = (\alpha^*_1, \ldots, \alpha^*_l)^T$.

This formula only solved the basic bi-classification problem.

According to the sample data collected by author, a small number of sample data category was negative subsets only one class of sample data, positive subsets in sample data set was the majority class. Firstly it needed determine how many categories (n classification) classification in sample data to be multi-classification analysis. Then, determine the right weight coefficient for multi-function classifiers. Author used amended support vector machine algorithm to evaluating credit rating of SMEs for imbalanced multi-classification analysis and compared the accurate rate with two step clustering, k-means clustering method, system clustering method, neural networks method.

The steps of this algorithm were: Input; Learning process; and output. Data set N were divided by M data subsets，that was，obtained the support vector C_i for separating from each positive subsets of sample data sets, the number of which i=1,2,... M was also positive category dataset.

Output process meant that to input arbitrary vector x in sample sets resulted the corresponding specific categories Y, during the learning process necessarily choosing the appropriate M. In the classifiers of support vector machine the imbalanced classification ratio was m_1: m_2 > 1:5, still had the robustness and self-adjustment ability, also it was the imbalanced classification rate higher than the original setting rate of classification, so easily causing much highly false classifiers rate for sample data in positive categories set in each kind of traditional classification methods. Thus, it was necessary to set the $M=2^{no}$, where n_0 could be equal to any positive integers, also, $\min_n \left| q - 2^n \right|$, （n=1,2…）, in this case each classification could be correctly classified to the sample data in positive categories set.

It was known that M was not only the number of categories in positive sample data, but also the number of classifiers achieving better results in system synthesis algorithm. In order to make the algorithm more reasonable and feasible, it could be calculated after n_0, M value between 2_{no-1},2_{no}, 2_{no+1}. Since only n_0 related to the proportion of accurate rate of imbalanced classification problem, in addition the scale and distribution of the sample data also affected the efficiency and results of support vector machine classification method, given $M=2_{no}$ would affect the results of classification.

The basic principle of classification method of support vector machine only used support vectors which consisted of hyper plane separating the high dimensional space; the rest in the sample data categories did not play any role. Then, for the sample data set in this imbalance multi-classification problem, classification ability of support vector machines would be affected, and there were much more sample data in positive categories, while there were less sample data in negative categories, which played a very important role in supporting hyperplane, determining the results 0 or 1of classification in two class classification problem.

Of course, in the author's case there was just one of the situations, also it was necessary to further distinguish the other sample data in positive category with intuitive inseparability. In addition, there were more noise and disturbing information in positive sample data set still existing in the process of sample collecting and function validation confirming, affecting the efficiency and accuracy of the classification method including support vector machine or the other traditional classification methods. This imbalanced classification algorithm combined the basic idea of support vector ordinal regression, using system integration algorithm, vector quantization method to accurately and efficiently solve this type of problem for multi-classification data. The steps of algorithm reflected the characteristics of the multiple classifiers of support vector machine i=1,...M, to use this series of classifiers. And the weight coefficient of the classifiers affected the distance between the hyperplane and support vectors which divided the sample data in positive categories and sample data negative categories in classification problems.

In general, the support vector machine classification problems were always the problem in highly dimensional space. General functions of classification often were not the linear classifiers, but the nonlinear classifier of power function. If increasing the dimension of sample data of support vector could also be applied, and getting

the classifiers hyperplane of classification boundary between negative and positive sample data set, also reducing the skewness.

It was a typically problem of imbalance multi-classification on performance evaluation of the SMEs, and the numbers of collected sample data in each category were not equal. The author selected the data of 164 SMEs in east of China in 2011, after getting rid of the vacant data and unqualified data and standardizing the sample data. Selecting 13 variables of sample data in SME, the capability of managers included entire period of actual operation, educational background of major managers; and the ability of asset management included the situation of primary assets, the situational of operational area and the age of corporation founded; and the ability of profitability included the profitability of shareholders, the increasing rate of sale revenue, the rate of furniture and equipment used and the increasing rate of net income; and the ability of debt payment included debt ratio and current ratio; and operational environments included industry policy and local operational environments. And the formula of decision was: The numbers of correct divided companies/the numbers of total companies; After the process of analyzing, training and testing, received the analysis results below.

Table 1. The analysis results of 13 variables

C / gamma	1000	1200	1400	1600	1800	2000	2200
0.001	0.7476	0.7500	0.7536	0.7548	0.7512	0.7571	0.7548
0.003	0.7500	0.7536	0.7631	0.7512	0.7488	0.7500	0.7524
0.005	0.7500	0.7500	0.7536	0.7560	0.7548	0.7524	0.7512
0.007	0.7536	0.7524	0.7524	0.7512	0.7560	0.7560	0.7536
0.009	0.7548	0.7500	0.7560	0.7548	0.7560	0.7560	0.7560

After ten times calculation, it was the means of the accurate rate of classification.

Then, it wished to reduce the variables of sample data in each SME, selecting the other 11 variables, the capability of managers included entire period of actual operation, educational background of major managers; and the ability of asset management included the situation of primary assets, the situational of operational area and the age of corporation founded; and the ability of profitability included the profitability of shareholders, the rate of furniture and equipment used; and the ability of debt payment included debt ratio and current ratio; and operational environments included industry policy and local operational environments. After the process of analyzing, training and testing, received the analysis results below.

Table 2. The analysis results of 11 variables

C / gamma	1800	1900	2000	2100	2200	2500	2700
0.001	0.7369	0.7381	0.7357	0.7369	0.7357	0.7345	0.7321
0.003	0.7429	0.7429	0.7440	0.7369	0.7381	0.7440	0.7488
0.005	0.7512	0.7524	0.7536	0.7440	0.7429	0.7345	0.7393
0.007	0.7357	0.7393	0.7393	0.7357	0.7357	0.7381	0.7381
0.009	0.7357	0.7369	0.7381	0.7381	0.7357	0.7369	0.7369

4. Conclusion

The author also dealt with the sample data from 164 firms with the other methods. Using the 11 variable sample data of SMEs, it was analyzed for multi classification and obtained the various results.

1) Used two step clustering methods by SPSS software to analysis; it was poor that the results of the quality test of the clustering profile measurement to separation and condensation.

2) Used K-means clustering method by SPSS software, the clustering results were shown in the following table:

Table 3. The result of clustering analysis

Group	k-means clustering method	the rate of actual sample data
1	39.6%	3.7%
2	3.7%	15.1%
3	34.7%	29.9%
4	6.1%	28%
5	5.5%	12.3%
6	4.3%	9.8%
7	6.1%	1.2%
total	100%	100%

As it can be seen, the accuracy of the experimental results was low.

3) Used system clustering method by SPSS software, it is obtained 37.57% as the average classification accuracy rate.

4) Used neural network analysis method by SPSS software, because of selected disparate factors and covariates and obtained the different results of training accuracy, the higher classification accuracy rate was 70.37% in many tests.

There were much more the sample data in positive categories than in that of the negative categories, thus, it was easily endured with interferential noise sample data and the categories of inseparability, to affect the accuracy rate. If improving the steps of the algorithm, and further detailed calculating the distance of the separating hyperplane between each class and the description feature of each subclass could improve the efficiency, the accuracy and the accurate rate of the classifiers. Firstly, it could definite distance of the sample data set of negative category and the sample data set of positive category, and the dimensional characteristics of each specific feature of support vectors, and then subdivide lots of positive sample data set.

It was certified that the improved support vector for the problem of imbalanced multi-classification was more efficient than other methods. There were many problems for further specific analysis the problem of imbalanced multi-classification, such as the accurate rate of each subclass and the decision formula of accurate rate.

References

Charles, T. H., Walter, T. H., & Suzanne, O. (2008). *Accounting* (8th ed.). Prentice Hall, Pearson Education, Inc.

Chris, S., Taghi, M. K., Jason, V. H., & Andres, F. (2011). An empirical study of the classification performance of learners on imbalanced and noisy software quality data.

David, M., Bart, B. T., & Van Gestel, J. (2007). Vanthienen Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research, 183*, 1466–1476. http://dx.doi.org/10.1016/j.ejor.2006.04.051

Der-Chiang, L., Chiao-Wen, L., & Susan, C. H. (2010). A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine, 40,* 509–518. http://dx.doi.org/10.1016/j.compbiomed.2010.03.005

Garcia, V., Sanchez, J. S., & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems, 25*, 13–21. http://dx.doi.org/10.1016/j.knosys.2011.06.013

Muhammad, A. T., Josef, K., & Fei, Y. (2012). Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recognition, 45*, 3738–3750. http://dx.doi.org/10.1016/j.patcog.2012.03.014

Nicolás, G. P., Javier, P. R., & María, G. P. (2012). Domingo Ortiz-Boyer, Colin Fyfe, Class imbalance methods for translation initiation site recognition in DNA sequences. *Knowledge-Based Systems, 25*, 22–34. http://dx.doi.org/10.1016/j.knosys.2011.05.002

Qian, L., Bing, Y., Yi, L., Naiyang, D., & Ling, J. (2012). Constructing support vector machine ensemble with segmentation for imbalanced datasets.

Qiuju, Y., Jimin, L., Zejun, H., Zanyang, X., & Heng, Z. (2012). Automatic recognition of poleward moving auroras from all-sky image sequences based on HMM and SVM. *Planetary and Space Science, 69,* 40–48. http://dx.doi.org/10.1016/j.pss.2012.04.008

Suzan, K. T., & Longin, J. L. (2011). Improving SVM classification on imbalanced time series data sets with ghost points. *Knowl. Inf. Syst.*, 281–323.

Tony, B., & Jonathan, C. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications, 36*, 3302–3308. http://dx.doi.org/10.1016/j.eswa.2008.01.005

Xiujuan, X., Chunguang, Z., & Zhe, W. (2009). Credit scoring algorithm based on link analysis ranking with support vector machine. *Expert Systems with Applications, 36*, 2625–2632. http://dx.doi.org/10.1016/j.eswa.2008.01.024

Yanmin, S., Mohamed, S. K., Andrew, K. C., & Wongb, Y. W. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition, 4*, 3358–3378.

Zhen, J., Shiyong, Z., & Jianping, Z. (2013). A hybrid generative/discriminative method for semi-supervised classification. *Knowledge-Based Systems*, *37*, 137–145. http://dx.doi.org/10.1016/j.knosys.2012.07.020