Financial Risk Forecasting based on Random Subspace Ensembles of Support Vector Machines: A study of Chinese commercial banks

Dechen Ding*

School of Insurance, Shandong University of Finance and Economics, Jinan, Shandong Province, 250000, China

*Corresponding author’s e-mail: dechend@163.com

Abstract. A model based on random subspace ensembles of support vector machines is proposed to forecast risk of Chinese commercial banks. For one thing the results show that the proposed model performed dramatically better when compared with the remaining ensemble learning algorithms, such as bagging, dagging, multiboosting and adaboosting. For another the results show that support vector machines could be availably applied as base learners in the ensemble learning model when compared with multilayer neural networks (MLPs), which have been perceived as benchmark methods during previous studies. Another vital finding of the paper is that macroeconomic conditions indexes play an important role in financial risk forecasting.

1. Introduction

A variety of ensemble-learning methods have applied in prediction of financial risk, including bagging and dagging, adaboosting and multiboosting\cite{1}. However, banks and their financial risk has obtained less attention considerably. Even more specifically, over the past decade most studies have been restricted to bankruptcy prediction\cite{2}. Additionally, almost all of studies in predicting financial failures of banks only pay attention to bank-level indicators, namely unsystematic risk index. This study tries to eliminate such problems by ensembling random subspace and support vector machine (SVMs as base learners) in predicting financial risk of Chinese commercial banks. SVMs are considered the most advanced method for financial risk forecasting primarily just that their good generalization performance\cite{3}. Shunning over-fitting is an vital topic to handle when training base learners in pattern recognition. The random subspace (RSS) method\cite{4} represents a parallel learning algorithm generating each base learner independently. In particular, it is favourable for parallel computing and fast learning, which makes random subspace one of the most frequently used ensemble learning method. The financial risk forecasting of banks can be considered as a two-class problem. Furthermore, bank-level data and systematic risk indexes are both important determinants in the financial risk forecasting.

The rest of this paper has been organized into five additional sections. Section 2 describes dataset and their pre-processing. Section 3 proposes the design of ensemble models (RSS\textsubscript{SVM}). Section 4 analyses the experimental results obtained by the ensemble model. SVMs serves as the base learners of the ensemble model. The results are compared with MLPs which have been perceived as benchmark methods during previous studies\cite{5}. In this section a comparison with other ensemble learning algorithms, such as bagging,
dagg, multiboosting and adaboosting is also provided. Section 5 concludes the paper and final assessments.

2. Indexes and sample

2.1 Indexes

The paper uses four categories (with reference to bank-level, banking sector, domestic macroeconomic environment and international financial risk impact) of indicators in order to capture various aspects of an Chinese commercial bank’s vulnerability to risk[6]. The indexes adds up to 34.

2.2 Sample

The sample used in this study was obtained from 12 listed commercial banks operating in China between 2004 and 2018. The total samples add up to 174 after eliminating missing value. According to the commercial bank of China banking regulatory commission (CBRC) rating report, data set is divided into two groups: 36 unhealthy (be in financial risk) and 138 healthy (be not in financial risk) sample banks. Of all samples, data of systematic risk index comes from Celn et statistics database, China statistical yearbook and China financial statistical yearbook, while data of unsystematic risk index comes from the listed bank’s annual report. To ensure the proper estimation and validation of the models, two thirds samples were randomly from each group selected as training set (i.e. total of 116 banks) and the remaining banks were kept for out-of-sample evaluation (i.e. total of 58 banks).

The normality test shows that none of the indexes are normally distributed at the conventional level, banking sector, domestic macroeconomic environment and international financial risk impact. Therefore, the non-parametric test (Mann–Whitney U test) is more suitable than the parametric one (Student t test). This test underlines that all indexes except $x_3,x_{17},x_{23},x_{26}$ show significant differences between the two groups.

Value of Sig. of Bartlett’s inspection significantly equals to 0. In consequence, the hypothesis that the correlation matrix is a unit matrix is rejected, namely indexes have strong correlation; That KMO statistics is 0.525 indicates a certain degree of information overlap exist between these indexes, which is suitable for the use of principal component analysis method to eliminate relevant attributes in bank risk. As per the variance explained criterion, the first 13 attributes whose sum of eigenvalues exhibit an explanatory power of greater than 85% in terms of the cumulative percentage of explained variation are chosen. See Table 1.

Table 1: Eigenvalues of the attributes

| Attribute | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-----------|---|---|---|---|---|---|---|---|---|----|
| Index     | $x_{11}$ | $x_{20}$ | $x_{44}$ | $x_{12}$ | $x_{10}$ | $x_{4}$ | $x_{1}$ | $x_{14}$ | $x_{30}$ | $x_{1}$ |
| Eigenvalue| 11.5 | 10.6 | 9.1 | 8.5 | 8.0 | 7.6 | 6.7 | 6.5 | 5.8 | 5.2 |
| Cumulative % of variation explained | 11.5 | 22.1 | 31.2 | 39.7 | 47.7 | 55.3 | 62.0 | 68.5 | 74.3 | 79.5 |
| Attribute | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Index     | $x_{2}$ | $x_{11}$ | $x_{15}$ | $x_{29}$ | $x_{10}$ | $x_{11}$ | $x_{24}$ | $x_{27}$ | $x_{34}$ | $x_{32}$ |
| Eigenvalue| 4.3 | 3.5 | 2.6 | 1.8 | 1.3 | 1.1 | 1.0 | 0.9 | 0.8 | 0.7 |
| Cumulative % of variation explained | 83.8 | 87.3 | 89.9 | 91.7 | 93.0 | 94.1 | 95.1 | 96.0 | 96.8 | 97.5 |
| Attribute | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| Index     | $x_{33}$ | $x_{6}$ | $x_{5}$ | $x_{12}$ | $x_{31}$ | $x_{14}$ | $x_{13}$ | $x_{4}$ | $x_{1}$ |
| Eigenvalue| 0.6 | 0.5 | 0.3 | 0.3 | 0.2 | 0.2 | 0.1 | 0.1 | 0.1 | 0.1 |
| Cumulative % of variation explained | 98.1 | 98.6 | 98.9 | 99.2 | 99.4 | 99.6 | 99.7 | 99.8 | 99.9 | 100.0 |

**Note:** Please refer to 2.1 for index explanations.

2.3 Data Pre-processing

For all data, we replaced missing values by median values, and all data were standardized using the Z-score to prevent problems with different scales.
3. Modelling
To forecast bank risk, we employed several ensemble learning algorithms using SVMs as base learners. Besides, the results were compared with another base learner-MLP. The results indicate that $\text{RSS}_{\text{SVM}}$ method performed significantly better when compared with the remaining ensemble learning algorithms, namely bagging and dagging\textsuperscript{[7]}, adaboosting and multiboosting\textsuperscript{[8]}.

Choosing the correct settings for the kernel parameters as well as the metaparameters $C$ and $\varepsilon$ are crucial for the estimation accuracy of a SVM. Existing research shows that performance of RBF kernel function is better than that of the other functions in most cases, which brings RBF kernel function more commonality. RBF function was chosen for SVM classification. Therefore, the best classification result of the SVM was tested for the following user-defined parameters: kernel functions = RBF, round-off error = 1.0E-12, search range = $[2^{-10}, 2^{10}]$, step size = 0.001. Grid search algorithm and 5-fold cross-validation was used to find the optimum settings.

The MLP was trained employing the backpropagation algorithm with momentum. The following training parameters were tested to achieve the best classification performance: the number of neurons in the hidden layer = $\{2^0, 2^1, 2^2, 2^3, 2^4, 2^5\}$, learning rate = $\{0.1, 0.2, 0.3, 0.4\}$, momentum = $\{0.2, 0.3, 0.4\}$, and the number of epochs = $\{100, 200, 400, 600, 1000\}$. Also, grid search algorithm was used to find the optimum settings of these parameters.

Subspace size is the vital parameter in learning RSS. Hence, many settingsof the subspacesize was tested to obtain the best classification performance. The number of iterations of the RSS was set to 12. Due to limited space, parameters set of the other ensemble learning algorithms used in our experiment is omitted.

To avoid overfitting, 10-fold cross-validation was used in the experiments. The classification performance was measured by the averages of standard statistics applied in classification tasks: average [%], Type I error [%], Type II error [%].

4. Results and analyses

4.1 Experimental results

The detailed classification performance on the dataset of Chinese commercial banks is shown in Table 2.

| Model | Training set | Testing set |
|-------|--------------|-------------|
|       | Type I | Type II | average | Type I | Type II | average |
| **all input attributes** | | | | | | |
| 1. MLP | 14.33 | 22.34 | 75.01 | 13.18 | 22.45 | 75.27 |
| 2. SVM | 10.22 | 17.15 | 83.11 | 10.34 | 16.77 | 83.76 |
| 3. RSS\textsubscript{MLP} | 12.23 | 17.44 | 81.64 | 12.53 | 16.65 | 82.10 |
| 4. RSS\textsubscript{SVM} | 8.11 | 11.02 | **89.87** | 8.89 | 11.89 | **88.22** |
| 5. BA\textsubscript{SVM} | 10.64 | 16.27 | 86.35 | 11.21 | 17.39 | 86.19 |
| 6. DA\textsubscript{SVM} | 11.55 | 16.32 | 86.21 | 10.36 | 15.91 | 86.11 |
| 7. MU\textsubscript{SVM} | 10.65 | 16.24 | 85.06 | 10.25 | 17.43 | 85.54 |
| 8. AD\textsubscript{SVM} | 11.66 | 15.73 | 85.92 | 11.99 | 15.33 | 86.01 |
| **Bank-level indicators only $x_{10}$ to $x_{14}$** | | | | | | |
| 1. MLP | 15.88 | 22.79 | 73.47 | 16.07 | 24.12 | 74.23 |
| 2. SVM | 13.01 | 19.43 | 81.56 | 12.54 | 20.89 | 81.04 |
| 3. RSS\textsubscript{MLP} | 14.21 | 21.24 | 77.35 | 15.77 | 22.43 | 76.07 |
| 4. RSS\textsubscript{SVM} | 10.55 | 13.44 | **86.34** | 11.45 | 13.78 | **84.85** |
| 5. BA\textsubscript{SVM} | 13.11 | 17.56 | 82.32 | 13.08 | 13.43 | 81.22 |
| 6. DA\textsubscript{SVM} | 13.38 | 19.99 | 83.43 | 14.32 | 19.47 | 82.01 |
4.2 Analyses

As can be seen from table 2, firstly, classification accuracy was memorably higher when incorporating all indexes than bank-lever indexes only. Secondly, SVM model performed observably better when compared with MLP model. Thirdly, all the ensemble models, especially RSS\textsubscript{SVM} model performed better than single classifiers (prominently better in the case of MLP). Fourthly, all classifiers performed better on Type I error, mainly due to the imbalance of the classes. Unquestionably, it is seen that RSS\textsubscript{SVM} model stands out as the best model in relation to total prediction power.

5. Conclusion and Discussion

Using a sample of 12 Chinese listed bank, we develop a random subspace ensembles of support vector machines model and examine its accuracy in classifying banks. Compared with other single classifier (especially for MLPs) and ensemble models, the fundamental conclusions draw from the experiment in relation to the prediction of bank risk can be described as follows: (i) The results indicated that variables for banking sector, macroeconomic conditions and international financial risk impact should be considered an important factor in bank risk prediction models. This finding has crucial implications for banks, regulators and other authorities. More specifically, the average accuracy of prediction was enhanced by about 3% with the help of systematic risk information. (ii) The study also showed RSS learning algorithm dramatically increases the performance of base learners MLP and SVM in case of commercial banks. (iii) Combining the contribution of systematic risk information of banks and ensemble learning approach the average accuracy was increased by about 8%. This may bring banks substantial savings because of the loss associated with potential risk.

In future, all indexes especially for systematic risk index should be adjusted with the continuous change of international financial risk impact, domestic macroeconomic environment and banking sector. In addition, other classifiers such as Naïve Bayes and ensemble learning algorithms such as rotation forest approach could be introduced to compare dictionary approach used in the study. Lastly, we tend to resolve data imbalance problem for bank risk prediction by using means such as geometric mean based boosting algorithm with over-sampling.

Acknowledgement

This work was supported by the scientific research project of National Natural Science Foundation of China Grant No: 71671037.

References

[1] Yu-Pei Huang, Meng-Feng Yen.(2019) A new perspective of performance comparison among machine learning algorithms for financial distress prediction. Applied Soft Computing, 8:8-20.

[2] Hyunchul Choi, Hyojoo Son, Changwan Kim.(2018) Predicting financial distress of contractors in the construction industry using ensemble learning. Expert Systems with Applications, 110:1-10.

[3] Francisco Climent, Alexandre Momparler, Pedro Carmona.(2018) Anticipating bank distress in the Eurozone: An Extreme Gradient Boosting approach. Journal of Business Research, In press, corrected proof, Available online 15 November.
[4] Ho, T.K. (1998) The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8):832-844.

[5] Hajek, P., Olej, V. (2010) Credit Rating Modelling by Kernel-Based Approaches with Supervised and Semi-Supervised Learning. Neural Computing and Applications, 20:761-773.

[6] Dechen Ding. (2016) Research on risk warning system of commercial Banks under the new normal of economy. Macroeconomics, 4:124-134.

[7] Ting, K.M., Witten, I.H. (1997) Stacking Bagged and Dagged Models. In: 14th Int. Conf. on Machine Learning, San Francisco, CA, pp. 367-375.

[8] Webb, G.I. (2000) MultiBoosting: A Technique for Combining Boosting and Wagging. Machine Learning, 40:66-79.