The Application of the Integrated Machine Learning Model in the Financial Crisis of Imbalanced Samples

Zhang Lu¹, Jiapeng Liu²* and Minqi Jiang³

¹School of Economics and Management, China Jiliang University, Hangzhou, China
²*corresponding author

Abstract. In this paper, the financial crisis of the enterprise in the t-period is predicted by the financial index data of the non-financial industry A-share enterprise in the main board market of the Shanghai Stock Exchange in the t-3 period. Then we use the integrated machine learning models to select the financial crisis enterprises, and the problem of failure of classifiers in unbalanced samples is solved through bagging and sampling techniques. The highest probability of correctly selecting enterprises with financial crisis is 92.86%. With the integration of the model, the overall accuracy is improved to 88%. We help the enterprise to complete the forward-looking financial crisis forecast and provide some reference for the business activities and investment activities of the enterprise.

Keywords: Financial crisis and forecasting; Integrated machine learning; Sampling technology.

1. Introduction

The financial crisis could effect the benefit of shareholders, institutes, managers and even governments. So predicting the financial crisis has been vulnerability for the enterprises, whether in the operating or investment activities. The rise of machine learning provides a new way for the processing and application of big data. Now, many scholars combine machine learning and financial crisis warning, and have made great breakthrough. Ohlson[1] proposes to apply logistic regression to the posterior probability of the classification, and he uses the log-likelihood probability to estimate the insolvent of companies. Zou, H. and T. Hastie[2] proposed Elastic Network to overcome the disadvantages of ridge regression and lasso. Decision Tree Learning is a powerful classifier[3], and on the basis of tree classifier, Random Forest and XGBoost were proposed and has been proved that the classification effect was better in many areas[4,5].

But people used to set the sample size artificially regardless of rare amount of the enterprises with financial crisis. David and Eric[6] proposed that sampling techniques could be used to improve the performance of the imbalanced samples forecasting. And the bagging technique is proved to be available for imbalanced data[7]. Besides, Random Up-sampling Technique[8], Random Down-sampling Technique[9] and Synthetic Minority Over-sampling Technique[10] were also acquired improving performance.

Cao[11] proposed that the integrated learning mechanism can integrate the advantages of many algorithms by integrating different models. Wang and Han[12] also used integrated learning in Personal credit evaluation and got great results.

So, considered to combine the imbalanced samples and the sampling techniques, the paper attempts to apply the integrated machine learning model to the financial crisis warning of the unbalanced samples,
and by means of the optimization of the model, solves the failure of the machine learning model in an unbalanced sample and improves the result of prediction. The purpose of this paper includes three parts. The first is to test the prediction performance of the integrated machine learning model and to find the most suitable classifier for financial crisis warning. The second is to apply the idea of integrated learning to the full sample of Chinese listed companies, to reduce the coincidence prediction brought by the manual selection of samples, and to improve the probability of enterprises with financial risk in the t-period by using the sampling technique and the bagging method to improve the enterprise data in the t-3 period. Thirdly, we has integrated the model to improved the accuracy rate of the healthy enterprises’ classification with keeping the right selection of the financial warning enterprises.

2. Research Method

2.1. Sample and Indexes Selection

The paper selected A-share enterprise financial index of the main board market non-financial industry in Shanghai Stock Exchange from the Resset financial database. We set 0 variable as the healthy enterprises and 1 as the financial warning enterprises. Considering that the flag of ST or *ST is negative profit for consecutive two or three years, the t-3 period's financial indexes were selected to predict the classification of the t-period. Then we selected 107 original variables from the financial ratio data of the Resset financial database, and the index with correlation coefficient greater than 0.5 was screened out. Finally we got 48 indexes including 6 classes listed in Table 1.

Table 1. Selected financial indexes.

| Per share index | X1 per-share earnings | X2 net assets per share | X3 Total operating income per share | X4 Operating profit per share | X5 Pre-tax profit per dividend | X6 Per share provident fund | X7 Undistributed profit per share | X8 Retained earnings per share | X9 Cash flow of operating activities per share | X10 Free cash flow per enterprise |
|-----------------|-----------------------|-------------------------|------------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|----------------------------------|----------------------------------|
| **Profitability index** | X11 Rate of return on assets | X12 Net interest rate on assets | X13 Net profit rate on sales | X14 Gross margin on sales | X15 Sales cost rate | X16 Cost rate during sales | X17 Net profit/ total operating income | X18 Operating profit/ total operating income | X19 Profit before interest and tax/ gross operating income | |
| **Solvency index** | X20 ratio of expenses to sales | X21 rate of management | X22 Financial cost rate | X23 Operating profit rate | X24 Profit rate of cost and expense | |
| **Cash flow index** | X25 current ratio | X26 quick ratio | |
| **growth capacity index** | | | |
| X27 Total shareholders' equity/ liabilities | X28 EBITDA/ liabilities | X29 Total net operating cash flows/ liabilities | X30 Cash flow liability ratio | X31 Net profit growth rate of the home parent company | X32 Growth rate of cash flow per share of operating activities | X33 Growth rate of shareholders' equity relative to the beginning of the year | X34 operational capacity index | X35 fixed asset turnover | X36 Turnover rate of total assets (second) |
| **Cash flow index** | | | |
| X37 Cash from sales of goods and services/operating receipt | X38 Sales cash ratio | X39 Total assets cash recovery | X40 Balance of cash and cash equivalents per share |
| **capital structure index** | | | |
| X41 asset-liability ratio | X42 Current assets/ total assets | X43 Non-current assets/ total assets | X44 Shareholder's Equity/ All Invested Capital | X45 Total current liabilities/ liabilities | X46 Total non-current liabilities/ liabilities | X47 Shareholder's equity ratio | X48 Long-term liabilities/ shareholders' equity |

2.2 Classification Predicting Models and Sampling Technique
As is shown in Figure 1, firstly, we dealt with the unbalanced data with bagging method and sampling technique. Bagging method could enhance the accuracy of classifiers to select the financial crisis enterprises. Random Up-sampling Technique (RUT) counterweights the imbalanced samples by extracting repeated small samples randomly. Random Down-sampling Technique (RDT) screens out large samples randomly to be in equilibrium. Synthetic Minority Over-sampling Technique (SMOTE) generate new small samples through KNN to create the balanced data.

Secondly, five classifications including Logistic Regression, Elastic Net, Decision Tree, Random Forest and XGBoost would be used for prediction. The first four methods have been used in the area of financial crisis before. XGBoost (Extreme Gradient Boosting) was proposed in 2016, which had been used in other areas and proved to have better performance. The objective function of the algorithm is as follows:

$$\text{Obj}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Thirdly, the integrated learning mechanism can integrate the advantages of many algorithms by integrating different models. So we integrate the single classifier through a robust and cautious approach. The robust integration algorithm is that as long as one of the models forecasts the financial risk of the enterprise, the integration model predicts that the enterprise has financial risk, which is recorded as ME_R. The cautious integration algorithm is that only when the two classifiers predict that the enterprise will be in risk at the same time, the integration model thinks that the enterprise has financial risk, which is recorded as ME_C.

2.3 Analysis of Empirical Results
In the unbalanced full sample, we have 43 financial crisis enterprise and 1227 healthy enterprise. Table 1 describes the AUC value of the five simple classifiers.

|               | LM   | EN   | DT   | RF   | XGBoost |
|---------------|------|------|------|------|---------|
| Artificial balance | 0.5421 | 0.7582 | 0.707 | 0.7729 | 0.8095  |
| RDT           | 0.7601 | 0.8246 | 0.8376 | 0.9541 | 0.8811  |
| RUT           | 0.7508 | 0.8497 | 0.7877 | 0.9696 | 0.9529  |
| SMOTE         | 0.7859 | 0.80860 | 0.8450 | 0.9620 | 0.9446  |

From the Table 2, we can see that after the use of the bagging and sampling techniques, AUC value has been proved obviously. The optimization of classifier attached importance to information in small samples, so the accuracy of the prediction was higher than artificial balance sample, which also took large data superiority into account.
Table 3. Classification prediction probability of optimization models.

| Method  | Spec. | Sen. | Accu. | Spec. | Sen. | Accu. |
|---------|-------|------|-------|-------|------|-------|
| Only bagging |       |      |       |       |      |       |
| LR      | 0.9534 | 0.2500 | 0.9140 | 0.7097 | 0.8571 | 0.7180 |
| EN      | 0.9873 | 0.2143 | 0.9440 | 0.7903 | 0.7500 | 0.7880 |
| DT      | 0.9767 | 0.2500 | 0.9360 | 0.7246 | 0.8929 | 0.7340 |
| RF      | 0.9873 | 0.7857 | 0.9760 | 0.7839 | 0.9286 | 0.7920 |
| XGBoost | 0.9767 | 0.7857 | 0.9660 | 0.8199 | 0.9286 | 0.8260 |
| ME_R    | 0.9703 | 0.7857 | 0.9440 | 0.7288 | 0.9286 | 0.7400 |
| ME_C    | 0.9915 | 0.7857 | 0.9800 | 0.8771 | 0.9286 | 0.8800 |

| Bagging RDT |       |      |       |       |      |       |
|-------------|-------|------|-------|-------|------|-------|
| LR          | 0.9301 | 0.5714 | 0.9100 | 0.7860 | 0.7857 | 0.7860 |
| EN          | 0.8538 | 0.7143 | 0.8460 | 0.8242 | 0.7143 | 0.8180 |
| DT          | 0.8898 | 0.7143 | 0.8800 | 0.8453 | 0.7500 | 0.8400 |
| RF          | 0.9915 | 0.7857 | 0.9800 | 0.8792 | 0.8571 | 0.8780 |
| XGBoost     | 0.9725 | 0.7857 | 0.9620 | 0.9068 | 0.8571 | 0.9040 |
| ME_R        | 0.9725 | 0.7857 | 0.9620 | 0.8623 | 0.8571 | 0.8620 |
| ME_C        | 0.9915 | 0.7857 | 0.9800 | 0.9237 | 0.8571 | 0.9200 |

As is shown in Table 3, the accuracy of discriminating the financial crisis enterprises (Sen. described) has been improved after using bagging and sampling techniques. What’s more, we found an interesting phenomenon that the middle columns (Sen.) of the Random Forest and XGBoost algorithm are the same, which means that this two classifiers may perform better when being combined. So we try to integrate Random Forest and XGBoost with the integration algorithm of robust and cautious.

From the ME_R and ME_C in Table 3, we concluded that both integration algorithms could keep the accuracy of selecting the financial crisis enterprises, but cautious integration algorithm could reduce the rate of misjudgment of the healthy enterprises. On the model of bagging RUT, the total accuracy increase about 5%~9%. Therefore, we recommend the integration model under the cautious algorithm for the managers of the enterprises.

3. Conclusion

We investigated the performance of the machine learning classifiers and dealt with the problem of imbalanced samples with the application of bagging and sampling techniques. Also, we found that the cautious integration algorithm have performed better than the single classifiers and the robust one in the paper.

The empirical study shows that the use of sampling technology improves the performance of the model and improves the probability of correctly selecting financial crisis enterprises which we interested in. Among the single classifiers, the combination of XGBoost and random under-sampling is the best, which not only improves the probability of selecting financial crisis enterprises, but also outperforms the random forest in predicting the probability of healthy enterprises. In order to reduce the probability of misjudgment of the normal enterprises, we have made a simple integration of the random forest and XGBoost, so that the probability of the financial crisis enterprises in the t-period is maintained at 92.86%, the misjudgment rate of the normal enterprises is reduced to about 12%, and the overall prediction accuracy is improved.

The application of integrated machine learning can help enterprises to complete forward-looking financial crisis warning. Compared with the traditional methods, it will be more universality, which combines the background of big data era and improves the accuracy of prediction, have lower
financial and accounting professional requirements for managers which is conducive to the diversified
development of enterprises. Also it provides a new reference for enterprises to select investment
objects and daily production and management activities.

Acknowledges
This paper acknowledges the financial support from the National Social Science Foundation of China
(18BGL224), the soft science project of Zhejiang Province (2017C35027), the Natural Science
Foundation of Zhejiang Province (LQ17G030008, LQ18G010004), and the key humanities and social
science projects in Zhejiang Province university (2016GH019).

References
[1] Ohlson J A . Financial Ratios and the Probabilistic Prediction of Bankruptcy[J]. Journal of
Accounting Research, 1980, 18(1),pp.109-131.
[2] Hui Zou, Trevor Hastie. Regularization and variable selection via the elastic net[J]. Journal of
the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(2),pp.301-320
[3] Bright, Lanz. Machine learning and r language [M]. Machinery Industry Publishing House
[4] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1),pp.5-32.
[5] Chen T, Guestrin C. Xgboost: A Scalable Tree Boosting System[C].Proceedings of the 22nd
Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. ACM, 2016.
[6] David Veganzones, Eric Séverin. An investigation of bankruptcy prediction in imbalanced
datasets [J] Decision Support Systems, 2018, 112,pp.111-124
[7] Wang B, Pineau J. Online Bagging and Boosting for Imbalanced Data Streams[M]. IEEE
Educational Activities Department, 2016, 28 (12),pp.3353-3366
[8] He H, Garcia E A. Learning from Imbalanced Data[J]. IEEE Transactions on Knowledge &
Data Engineering, 2009, 21(9),pp.1263-1284.
[9] Tahir M A, Kittler J, Mikolajczyk K, et al. A Multiple Expert Approach to the Class
Imbalance Problem Using Inverse Random under Sampling[C], International Workshop on
Multiple Classifier Systems. Springer, Berlin, Heidelberg, 2009.
[10] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-sampling
Technique[J]. Journal of Artificial Intelligence Research, 2011, 16(1),pp.321-357.
[11] Fu Wei, Li can, Zhu Weidong. Research on Financial crisis early warning of small and
medium-sized Enterprises from the Perspective of Multi-source Information Fusion-data Mining
method based on Integrated Learning [J]. Financial and Accounting Newsletter, 2018(5),pp.95-99.
[12] Wang Chongren, Han Dongmei. Personal credit evaluation of Internet credit based on
hyperparameter optimization and integrated learning [J]. Statistics and decision-making, 2019,
35 (01),pp.89-93.