Impact Analysis of Financial Early Warning Indicators Based on Random Forest

Shi-yong XIONG¹, Chen LU¹,* , Liang CHANG¹ and Ai-rong XIE²

¹Chongqing University of Posts and Telecommunications, Chongqing, China
²Communication NCO Academy, Army Engineering University of PLA, Chongqing, China

*Corresponding author

Keywords: Financial early warning, Random Forest, Financial indicators.

Abstract. In order to improve the indicator selection method for financial early warning, this paper combines the idea of K-fold cross-validation to improve the sampling method of Random Forest (RF) and proposes the K-fold random forest algorithm (KRF). The experimental results show that the KRF algorithm has a better classification performance than the RF algorithm, and improves the accuracy of the RF algorithm on the indicator. Finally, the importance of the selected financial indicators to the financial early warning is determined. A more scientific and accurate indicator system will provide a research basis for further financial early warning research.

Introduction

In the context of the government encouraging mass entrepreneurship, following the development of the Internet, China's financial industry is already developing vigorously, and the problems that come with it are becoming more and more obvious. How to judge whether the financial situation of the entrepreneurship company is good or not and how to judge the financial situation of a certain enterprise by the financial institutions are all urgent to be resolved. As early as the middle of the 20th century, scholars from all walks of life have conducted research on enterprise financial early warning from the aspects of economics and statistics. They mainly conducted early warning analysis through relevant knowledge, and successively experienced single variable early warning model and multivariate early warning model. Up to now, due to the maturity of data mining technology and machine learning technology, more and more scholars have carried out some research on multivariate models from these two points, including multivariate discriminant model, logistics model, neural network model. The research on financial indicators mainly covers two categories. One is to verify specific financial indicators based on human experience and to select one or more indicators that best represent financial status. The other is to add non-financial indicators to analyze the financial situation of enterprises from various indicators. Most of the current research is aimed at establishing a suitable corporate financial early warning model. These studies have not thoroughly studied the impact of various indicators on the overall prediction results through data mining techniques, so it is impossible to determine how the impact of an indicator will affect the prediction result, and the result of inaccurate results for dynamic prediction. The Random Forest algorithm (RF) proposed by Leo Breiman in 2001 can better overcome subjective biases and has better generalization ability. Feature selection based on random forest algorithm can remove redundant data without affecting classification accuracy, extract key features, and have good processing ability in feature selection [1].

Random Forest Algorithm

The random forest algorithm is an integrated learning algorithm based on the decision tree, combining Bagging integrated learning theory and random subspace method. It contains multiple decision trees trained by Bagging integrated learning technology. When inputting samples to be classified, the final The classification result is determined by the vote of the output of a single decision tree.
A large number of theoretical and empirical studies have proved that the random forest algorithm has higher classification accuracy, has good scalability and parallelism for high-dimensional data classification problems, has a good tolerance for outliers and noise, and is not easy to fit. In addition, random forests are data-driven nonparametric classification method. It only needs to train classification rules by learning the given samples and does not need prior knowledge [2].

**Out of Bag (OOB)**

A plurality of training samples are randomly selected from the original samples, and each training sample quantity is equal to the original sample quantity, and a regression decision sub-tree T is constructed for each training sample, and finally, the average value of each tree is taken as the final prediction results [3]. Assuming S is the original sample and N is the number of samples in S, then the probability that each sample in S is not extracted is:

\[
\lim_{N \to \infty} p(OOB) = \lim_{N \to \infty} (1 - \frac{1}{N})^N = \lim_{N \to \infty} \frac{1}{(1 + \frac{1}{N-1})(1 + \frac{1}{N-1})^{N-1}} = \frac{1}{e} \approx 0.38
\]

Eq. 1 indicates that approximately 36.8% of the samples are not extracted each time, which is the number of OOB.

**Feature Importance Measure**

The basic idea of random forest algorithm for the measurement of individual feature importance: After adding noise to a related feature, the prediction accuracy will decrease, and the change of accuracy will measure the importance of this feature.

**Importance Measurement Algorithm**

**Step1.** For each decision tree in the random forest, calculate the out-of-bag data error \( err_{OOB1}(x_i) \) the feature \( x_i \) using the corresponding out-of-bag data;

**Step2.** Randomly add noise interference to the feature \( x_i \) of all samples of the OOB data, and calculate its out-of-bag data error again, which is recorded as \( err_{OOB2}(x_i) \);

**Step3.** If there are \( N \) trees in the random forest, Eq. 2 is the importance of the feature \( x_i \):

\[
I(x_i) = \frac{\sum_{i=1}^{N} err_{OOB2}(x_i) - err_{OOB1}(x_i)}{N}
\]

If \( I(x_i) \) is higher, the more important the variable \( x_i \) is; if \( I(x_i) < 0 \), it means that the variable \( x_i \) has obvious noise and has a negative impact on the model.

**K-fold Random Forest Algorithm**

Since the OOB data of random forests generally cannot reach 36% of the original data in actual situations, the feature importance measurement is not accurate enough due to the lack of test data when calculating the error outside the bag. Therefore, under the premise of not affecting the classification accuracy of random forests, the accuracy of the measurement of feature importance metrics is improved. This paper improves the sample sampling of random forests with the idea of K-fold cross-validation and proposes a K-fold random forest algorithm (KRF).
KRF Algorithm Flow

**Step1.** The sample set N is equally divided into K parts, a set of K samples is taken as extra-bag data, and n samples are randomly selected from the remaining K-1 groups, and the n samples are used. Train a decision tree as a sample at the root of the decision tree;

**Step2.** When each sample has M attributes when each node of the decision tree needs to be split, m attributes are randomly selected from the M attributes, satisfying m<<M. An information gain strategy is then used to select an attribute as the number attribute of the node.

**Step3.** Each node in the decision tree formation process must be split according to step b until it can no longer be split. The so-called no longer split, that is, all reach the leaf node, if the next attribute selected by the node is just the attribute used when the parent node splits, the node is the leaf node.

**Step4.** Repeat steps 1-3 to create a large number of decision trees that form a random forest.

Algorithm Performance Analysis

In order to avoid the generality, the performance evaluation of the improved algorithm is tested by using the classification performance evaluation indicator of the random forest algorithm as the standard. In this paper, three different data sets of the UCI database are used as experimental data sets (Table 1), and the random forest algorithm and KRF algorithm are used for simulation experiments.

| Data set | Total sample number | Number of features |
|----------|---------------------|-------------------|
| Blood    | 748                 | 5                 |
| Iris     | 150                 | 4                 |
| Yeast    | 1484                | 9                 |

In the simulation experiment, the experiment of each data set was repeated 500 times, in which 4/5 of the data was taken as the data set and 1/5 was the test set. The final result is the mean of 500 experimental results.

Analysis of Experimental Results

The experimental accuracy values are shown in Table 2 after testing on the three data sets.

| Data set | RF     | KRF    |
|----------|--------|--------|
| Blood    | 0.563  | 0.723  |
| Iris     | 0.892  | 0.931  |
| Yeast    | 0.795  | 0.823  |

It can be seen from Table 2 that the accuracy of the KRF algorithm in the three data sets is higher than that of the traditional random forest algorithm, indicating that the modified algorithm has better classification performance.

Financial Indicator Importance Calculation Experiment

Since the financial data of general enterprises is not easy to obtain, this paper chooses domestic listed companies as the research object, the listed companies include financial normal companies, financial losses for two consecutive years (ST) companies and financial losses for three consecutive years have delisting risks (*ST) companies that has a certain representativeness for the classification of financial crisis warnings. This paper uses the financial data of the listed company of Guotaian database as the test data set. The data set contains 539 normal listed company financial statement data, 62 ST enterprise financial statement data and 33 *ST enterprise financial statement data. Therefore, the classification attribute contains three categories, which are represented by 0, 1, and 2, respectively, 0 for normal enterprises, 1 for ST enterprises, and 2 for *ST enterprises. Secondly, there are as many as 184 financial indicators in the dataset. This paper selects financial indicators from five aspects: solvency, profitability, operational capability, growth ability, and cash.
flow indicators. After preliminary screening, the correlation is large and has not been deleted. After
the indicators of practical significance, there are 27 remaining financial indicators, as shown in
Table 3.

| Table 3. Primary selection financial indicators. |
|-----------------------------------------------|
| **Solvency**                                  |
| X1 Current ratio                               |
| X2 Quick ratio                                 |
| X3 Cash flow to debt ratio                     |
| X4 Interest coverage ratio                    |
| X5 Shareholder equity ratio                    |
| X6 Assets and liabilities                      |
| **Profitability**                             |
| X7 Gross profit margin                         |
| X8 Sales margin                                |
| X9 Roe                                        |
| X10 Earnings per share                         |
| X11 Return on total assets                     |
| X12 Net asset interest rate                    |
| **Operating capacity**                        |
| X13 Inventory turnover                         |
| X14 Fixed asset turnover                       |
| X15 Current asset turnover                     |
| X16 Accounts receivable turnover               |
| X17 Accounts payable turnover rate             |
| X18 Total asset turnover                       |
| **Growth ability**                            |
| X19 Earnings per share growth rate             |
| X20 Operating income growth rate               |
| X21 Gross profit growth rate                   |
| X22 Total asset growth rate                    |
| X23 Net asset growth rate                      |
| X24 Growth rate of shareholders’ equity        |
| **Cash flow indicator**                       |
| X25 Sales cash ratio                           |
| X26 Sales cash ratio                           |
| X27 Net cash flow from sales                   |

**Classification Accuracy Analysis**

Correlation experiments were carried out on the dataset mentioned above, and the classification
accuracy evaluation indicator of classification accuracy was used to compare the traditional random
forest algorithm and KRF algorithm.

First, a random forest model containing 60 trees, 70 trees, 80 trees, 90 trees, 100 trees and 110
trees was generated using the traditional random forest algorithm. The test data set was performed
by these traditional random forest models. After testing, the classification accuracy of the traditional
random forest model is obtained, and then the important measure of each financial indicator is
carried out from the traditional random forest model. By calculating and calculating the out-of-bag
error rate before and after adding the noise data to each indicator, each the importance value of the
indicator. Then, using the same data set, the classification accuracy of the proposed KRF algorithm
and the importance of each financial indicator were tested under the same conditions. Finally, the
classification accuracy data of the two are compared. The classification accuracy comparison results
are shown in Fig. 1. It can be concluded that, in general, the KRF algorithm has a higher
classification accuracy.

![Figure 1. Comparison of classification accuracy.](image)

From the comparison of classification accuracy, it is obvious that the importance of each
Indicator in the KRF algorithm is obviously accurate to the traditional random forest algorithm, and
the importance value of each Indicator calculated by the KRF algorithm is shown in Table 4.
Table 4. Value of each indicator.

| Indicator number | Importance Value | Indicator number | Importance Value | Indicator number | Importance Value |
|------------------|------------------|------------------|------------------|------------------|------------------|
| X1               | 0.8992           | X2               | 0.8991           | X3               | 0.8392           |
| X4               | 0.8458           | X5               | 0.7979           | X6               | 0.8301           |
| X7               | 0.7787           | X8               | 0.8832           | X9               | 0.8291           |
| X10              | 0.8953           | X11              | 0.7392           | X12              | 0.8538           |
| X13              | 0.8007           | X14              | 0.8946           | X15              | 0.8321           |
| X16              | 0.8825           | X17              | 0.7038           | X18              | 0.8927           |
| X19              | 0.7951           | X20              | 0.7872           | X21              | 0.7643           |
| X22              | 0.8332           | X23              | 0.7644           | X24              | 0.8201           |
| X25              | 0.8893           | X26              | 0.8974           | X27              | 0.8802           |

The verification of the important results of each indicator is performed by manual comparison and experimental comparison. The artificial comparison refers to the comparison with the important indicators based on the expert subjective screening method to determine whether the experimental results are reasonable. The comparison results are shown in Table 5.

Table 5. Comparison of the top ten indicators of importance and expert experience data.

| Indicator          | Importance Value | Literature               |
|--------------------|------------------|--------------------------|
| X1 Current ratio   | 0.8992           | Ohlson(1980) [4]          |
| X2 Quick ratio     | 0.8991           | Beaver(1966) [5]          |
| X26 Sales cash ratio | 0.8974       | Shi-nong WU(2001) [6]     |
| X14 Fixed asset turnover | 0.8946    | Shi-nong WU(2001)         |
| X18 Total asset turnover | 0.8927     | Shi-nong WU(2001)         |
| X25 Sales cash ratio | 0.8893       | Yu-Chen(2011) [7]         |
| X3 Sales margin    | 0.8832           | Yu-Chen(2011)             |
| X16 Accounts receivable turnover | 0.8825 | Yu-Chen(2011)             |
| X12 Net asset interest rate | 0.8538       | Altman(1994) [8]          |
| X4 Interest coverage ratio | 0.8458     | Altman(1994)              |

The experiment comparison refers to the experiment using other financial early warning models, and selects 10 indicators with a high degree of importance among all indicators, and applies them to the selected financial early warning model. According to the prediction accuracy, the importance calculation of the indicators can be reflected. Whether it is accurate or not, the selected models are SVM_KNN model [9] and PSO_SVM model [10]. The experimental results are shown in Table 6.

Table 6. Model experiment results.

| Indicator          | KNN_SVM          | PSO_SVM          |
|--------------------|------------------|------------------|
|                   | Comprehensive accuracy | Positive class accuracy | Negative class accuracy | Comprehensive accuracy | Positive class accuracy | Negative class accuracy |
| Original indicator | 85.392% | 90.452% | 80.331% | 84.545% | 87.452% | 81.638% |
| New indicator      | 88.978% | 91.518% | 86.437% | 90.093% | 90.448% | 89.737% |

Summary

In this paper, through the in-depth analysis of random forest algorithm, in order to improve the accuracy of its index error analysis, combined with the idea of K-fold cross-validation, the sampling method is improved, and a K-fold random forest algorithm is proposed. The experimental results show that the proposed algorithm can not only improve the classification accuracy of the traditional random forest algorithm but also improve its accuracy in the analysis of index error. At the same time, it provides a more scientific method for the research of financial early warning indicators, and provides the preconditions for the subsequent financial early warning research.
References

[1] Xin-wen Liu, A study on early warning of China’s financial risk based on comprehensive improved random forest algorithm, J. Computer Applications and Software. 35 (2018) 73-78+84.

[2] Ri-sheng Wang, Hong-wei Xie, Jian-cheng An, Improvement of Random Forests Algorithm Based on Classification Accuracy and Correlation, J. Science Technology and Engineering. 17 (2018) 67-72.

[3] Xiao-feng Ye, Ya-hui Lu, Credit Assessment Model Based on Random Forest and Naive Bayes, J. Mathematics in Practice and theory. 47 (2018) 68-73.

[4] James A. Ohlson, Financial Ratios and the Probabilistic Prediction of Bankruptcy, J. Journal of Accounting Research, 18(1980) 109-131.

[5] William H. Beaver, Financial Ratios as Predictors of Failure, J. Journal of Accounting Research. 4 (1966) 71-111.

[6] Shi-nong Wu, Xian-yi Lu, A Study of Models for Predicting Financial Distress in China's Listed Companies, J. Economic Research Journal. 6 (2011) 46-55+96.

[7] Yu Chen, An Empirical Study on the Prediction of ST Companies in China's Securities Market, J. Economic Science. 6 (2011) 57-67.

[8] Edward I. Altman, Marco G, Varett F. Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks, J. Journal of Banking and Finance. 18 (1994) 505-529.

[9] Guang-hua Xu, Company Financial Early-warning System Based on Machine Learning, D. Xiamen University, Xiamen, 2014.

[10] Ling-yan Wu, Hua Han, Fei Tang, Multi-classification of Financial Early-warning Based on PSO-SVM Model, J. Journal of WUT (Information & Management Engineering), 35 (2013) 265-269.