Eng Int Syst (2025) 6: 733–737

Engineering
Intelligent Systems

Risk Assessment of Corporate Financial Internal Control and Audit Matters: Data Mining

Shu Ren*

Anyang University, Anyang City, Henan 455000, China

Internal financial audits within enterprises are beneficial for planning the operations of enterprises and enhancing their market competitiveness. In this study, the random forest (RF) algorithm was combined with the back-propagation neural network (BPNN) algorithm to improve financial risk recognition and auditing accuracy. The RF algorithm was used to screen the main financial risk indicators, and the BPNN algorithm was employed to identify the financial risk. The RF algorithm calculated the importance of financial risk indicators in the simulation experiment. Then, the performance of the RF algorithm, traditional BPNN algorithm, and the improved BPNN algorithm was compared. The results showed that the RF algorithm effectively screened out the characteristics of important financial risk indicators. Compared with the single RF algorithm and the single BPNN algorithm, the improved BPNN algorithm had better recognition accuracy and shorter recognition time.

Keywords: financial risk, audit, data mining, random forest

1. INTRODUCTION

Given the global economic integration and the rapid development of information technology, the importance of internal control and audit of enterprises' financial status as the key link of enterprise operation and management has become increasingly prominent. Effective internal control and audit mechanisms not only help ensure the safety of enterprise assets and improve operational efficiency; as well as strengthening their market competitiveness, they also ensure that the operations of enterprises comply with laws, regulations and policies [1,2]. However, as the scale of the enterprise increases and the business becomes more complex, the traditional internal control and audit methods are gradually exposed to problems such as inefficiency, limited information processing ability, and inaccurate risk assessment. The core of an enterprise's financial internal control lies in ensuring the authenticity and integrity of financial information, preventing financial fraud, and optimizing resource allocation [3,4]. However, in the face of massive, complex, and dynamic financial data, it is difficult for traditional evaluation methods to achieve comprehensive, timely, and accurate monitoring and analysis. Especially in large enterprises operating across regions and departments, information silos and asymmetries are more common, which further increases the difficulty of financial audit risk assessment. Data mining technology, as a product of the era of big data, can extract valuable rules from massive data, thus providing a scientific basis for auditing financial risks [5]. Related studies are as follows. Xu [6] studied the multi-attribute decision problem of financial risk management in commercial banks and verified its effectiveness. Huang et al. [7] conducted a statistical analysis of the shareholder relationship network and found that the shareholder relationship can be used only as supplementary information rather than key information when predicting bond risk. Hao et al. [8] established a classification and recognition model of audit risks by using the random forest (RF) algorithm and discovered that it had high prediction accuracy and good

*Corresponding address: No. 599, Zhonghua Road, Anyang City, Henan 455000, China. Email: renshurs@outlook.com

robustness. This paper combined the RF algorithm and back-propagation neural network (BPNN) algorithm in data mining technology. The RF algorithm was employed to screen the main financial risk indicators, and the BPNN algorithm was used to identify the financial risk. In the subsequent simulation experiment, the importance of the financial risk indicators was computed using the RF algorithm. Then, the performance of the RF algorithm, traditional BPNN algorithm, and improved BPNN algorithm was compared.

2. ENTERPRISE FINANCIAL RISK AUDIT EVALUATION BASED ON DATA MINING TECHNOLOGY

The RF algorithm [9] is a data mining technology consisting of decision trees. Its function of classifying and recognizing sample data is to give classification results of samples through decision trees and then carry out statistics on the results. The RF model is constructed as follows.

① The features involved in the samples are collected from the original training sample set to form the sample feature set.

② The random sampling with replacement was performed on the original sample set to form multiple sub-training sets, as well as on the sample feature set [10] to form multiple sub-feature sets. The number of sub-feature sets is the same as that of the sub-training set, and random pairing is carried out.

③ The paired sub-training set and sub-feature set are used to train the decision tree (the number of decision trees trained is the same as the number of pairs). The decision tree is constructed as follows: the feature that can minimize the Gini index after the division of the sample set is selected as the root node, and the feature that can minimize the Gini index after the division of the subsample set [11] is selected as the branch node respectively. The operations are repeated to make the decision tree grow. Then, pruning operations are carried out on the grown decision tree [12] to avoid overfitting of the decision tree model.

④ After all the decision trees have been trained, they are combined, and the classification result of each decision tree represents one vote for that category. The final classification result is voted on. The RF algorithm can improve its classification performance by combining multiple decision trees. In addition to the audit evaluation of financial risks conducted in this paper, it can also evaluate the importance of the characteristic indicators used to evaluate financial risks [13]. The importance calculation formula is:

$$MDA(j) = \frac{\sum_t^T \left(\frac{1}{|D_t|} \left(\sum_{X_i \subset D_t} I(P(X_i) = y_i) - \sum_{X_i^j \subset D_i^j} I(P(X_i^j) = y_i) \right) \right)}{T}, \tag{1}$$

where T is the number of decision trees in the RF, D_t is the out-of-bag sample set of decision tree t, D_i^j is the sample set after D_t exchanges the j-th dimensional feature, X_i is the i-th sample, X_i^j is the sample after X_i exchanges the j-th dimensional feature, $P(\,)$ is the decision tree prediction function, and $I(\,)$ is the indicative function, which is 1 when the decision tree predicts correctly and 0 when it predicts wrongly.

In essence, the RF algorithm uses decision trees to classify sample features. This process takes only the features in order to determine the direction of bifurcation, without in-depth analysis of the relationship between features.

Deep learning technology can mine in depth the hidden rules in big data to improve its classification and recognition performance. In this study, the classic BPNN [14] is used to audit and evaluate the financial risks of enterprises. When applying the BPNN to identify enterprise financial risk, the relevant characteristic indicators are input to judge financial risk, perform forward calculation in the hidden layer, and output the classification result. If it is in the training stage, the output result is compared with the actual result, and the gap between them is used to reversely adjust the parameters in the hidden layer.

When using the BPNN to evaluate enterprise financial risk, the selection of characteristic indicators used to judge enterprise financial risk is very important. Although some characteristic indicators are related to financial risk, they may have little impact but will affect the efficiency and accuracy of calculation. Therefore, for this study, the RF algorithm was used to screen the financial risk indicators and then the screened indicators were used for training and testing. The specific steps are as follows.

① Appropriate early warning indicators are selected.

② The data samples are collected according to the selected indicators.

③ The collected samples are divided into two categories according to whether there is financial risk. Then, in order to avoid the adverse impact of size imbalance between the two types of samples on the algorithm training, the synthetic minority oversampling technique (SMOTE) algorithm [15] is used to upsample the relatively smaller set of samples and increase the number of samples.

④ According to the steps specified above, training samples are used to build a RF model, and then the model is used to evaluate the importance of financial risk characteristic indicators. Non-important indicators are screened out.

⑤ The selected feature indicators of the training samples are input into the BPNN for forward calculation. The forward calculation formula of the hidden layer of the BPNN is:

$$o_j = f\left(\sum_{i=1}^{n} \omega_i x_i - b\right), \tag{2}$$

where o_j is the output vector of the hidden layer, b is the adjustment item of the hidden layer, and $f(\,)$ is the activation function of the hidden layer.

Table 1 Relevant indicators.

Type of indicator	Level-one indicator	Level-two indicator	No.
Financial indicators	Solvency	Current assets ratio	x_1
		Cash ratio	x_2
		Asset-liability ratio	x_3
		Quick ratio	x_4
		Equity ratio	x_5
	Profitability	Net assets income rate	x_6
		Net interest rate on total assets	x_7
		Stock yield rate	x_8
		Rate of return on invested capital	x_9
		Rate of return on human input	x_{10}
	Operational capacity	Cargo turnover rate	x_{11}
		Total asset turnover rate	x_{12}
		Accounts receivable turnover rate	x_{13}
		Current assets turnover rate	x_{14}
		Fixed assets turnover rate	x_{15}
	Development capacity	Year-over-year revenue growth	x_{16}
		Net assets growth rate	x_{17}
		Total assets growth rate	x_{18}
		Net profit growth rate	x_{19}
		Monetary fund growth rate	x_{20}
Non-financial indicators	Ownership structure	The shareholding ratio of the largest shareholder	x_{21}
		The comparison of the shareholding ratio between the first and second shareholders	x_{22}
	Legal liability	Major litigation	x_{23}
		Major violation	x_{24}

⑥ The distribution probability of the result label is computed using the vector calculated forward by the hidden layer and the softmax function [16] in the output layer. The classification result is output according to the probability.

⑦ If it is in the training stage, the calculated result label is compared with the actual result label of the sample. In this paper, cross entropy is used to measure the difference between them. When the difference converges to stability, the training is over. Otherwise, the parameters in the hidden layer are reversely adjusted according to the difference, and then step ⑤ is repeated.

3. SIMULATION EXPERIMENT

3.1 Experimental Data

Given the difficulty sometimes encountered when collecting sample data from companies, the crawler program was used to crawl the financial report data of listed enterprises (excluding the insurance and financial industry) from the official website of the stock exchange. The relevant indicators for early warning of financial risk are shown in Table 1. After using SMOTE to balance the quantity, the number of risk and risk-free report samples was the same.

3.2 Experimental Setting

In order to verify the classification performance of the proposed algorithm, it was compared with a single RF algorithm and a single BPNN algorithm. The relevant parameters of the improved BPNN algorithm and two other algorithms are shown in Table 2.

3.3 Evaluation Criteria

$$\begin{cases} P = \frac{TP}{TP+FN} \\ R = \frac{TP}{TP+FP} \\ F = \frac{2 \cdot P \cdot R}{P+R} \end{cases}, \tag{3}$$

where P is the precision ratio, R is the recall rate, F is the combined value of the precision ratio and recall rate, TP represents the number of true positive cases, FP represents the number of false positive cases, FN represents the number of false negative cases, and TN represents the number of true negative cases.

3.4 Experimental Results

The RF algorithm was used to analyze the importance of indicators required for financial risk audit. From the results shown in Figure 1, it can be seen that different risk audit

Table 2 Algorithm-related parameter settings.

	Parameter	Setting	Parameter	Setting
RF model	Number of features within the sub-feature set	7	Number of decision trees	120
	Maximum tree depth	12	Minimum number of samples for leaf nodes	1
	Minimum number of samples for branch nodes	2		
BPNN	Input layer	16 nodes	Hidden layer	32 nodes
	Output layer	2 nodes	The activation function in the hidden layer	Sigmoid [17]
	Number of training	200		

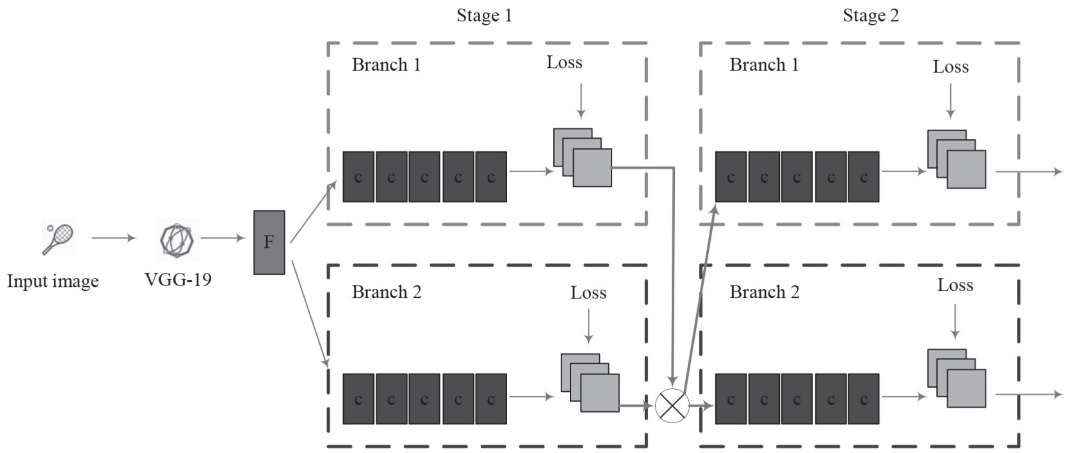

Figure 1 Relative importance of financial risk audit indicators.

Table 3 Classification performance and average time consumption of three financial risk auditing algorithms.

	RF	Traditional BPNN	Improved BPNN
Precision	0.756	0.899	0.988
Recall rate	0.758	0.897	0.986
F	0.757	0.898	0.987
Average time/s	3.67	1.98	1.37

indicators have different levels of importance, among which the current asset ratio is the most important, followed by the cash ratio, cargo turnover rate, and net interest rate on total assets. The indicators with importance lower than 0.1 include the shareholding ratio between the first and second shareholders, rate of return on human input, quick ratio, equity ratio, and rate of return on invested capital. In this paper, when training the BPNN algorithm, indicators with importance less than 0.1 were screened out to improve the efficiency of the BPNN algorithm.

The classification performance and average time consumption of the improved BPNN, RF, and traditional BPNN algorithms are shown in Table 3. It can be seen that the financial risk identification of the RF algorithm is the least accurate and took the most time; the accuracy and time consumption of the traditional BPNN algorithm for financial risk identification are in the middle; the improved BPNN algorithm has the highest accuracy in identifying financial risks and takes the least time.

The partial evaluation results of the three financial risk audit algorithms on the financial risks of enterprises and the corresponding expert audit results are shown in Table 4. Only the improved BPNN algorithm has the same audit results as those obtained manually by experts, while the audit results of the other two algorithms are different, and the audit results of the RF algorithm are the most divergent.

4. CONCLUSIONS

This paper integrated the RF algorithm with the BPNN algorithm. The RF algorithm was used to screen the main financial risk indicators, and then the BPNN algorithm was used to identify the financial risk. In the simulation experiment, the RF algorithm was used to calculate the importance of financial risk indicators. Then, the performance of the RF algorithm, the traditional BPNN algorithm, and the improved BPNN algorithm were compared. The results are as follows. The results for the RF algorithm showed that the ratio of current assets was the most important, followed by the cash ratio, the cargo turnover rate, and the net interest rate on total assets. The indicators with an importance less than 0.1 were the shareholding ratio between the first and second shareholders, the rate of return on human input, the quick

Table 4 Results of expert audit of financial risks of some enterprises and results of three financial risk audit algorithms.

Enterprise No.	Audit results of experts	Audit results of the RF algorithm	Audit results of the traditional BPNN algorithm	Audit results of the improved BPNN algorithm
11	Risk-free	Risky	Risk-free	Risk-free
19	Risk-free	Risky	Risky	Risk-free
42	Risky	Risky	Risky	Risky
51	Risky	Risk-free	Risk-free	Risky
67	Risk-free	Risk-free	Risk-free	Risk-free

ratio, the equity ratio, and the rate of return on invested capital. The improved BPNN algorithm had the highest accuracy in identifying financial risks and consumed the least time. The improved BPNN algorithm had the best alignment with the audit results obtained by experts, while the RF algorithm had the worst.

REFERENCES

1. Rahman S M K, Chowdhury M A F, Mohiuddin K M. An Investigation into the Relationship between Degree of Financial Leverage and Financial Risk of Firms: A Comparative Study between Listed MNCs and Domestic Companies of Bangladesh[J]. Asian Economic & Financial Review, 2018, 8(3):433–448.
2. Tian T. Evaluation of Financial Intelligent Systems Digitalization Maturity of Small and Medium-sized Enterprises Based on Fuzzy Logic Algorithm [J]. Engineering Intelligent Systems, 2025, 33(3): 275–284.
3. Sinnett A, Ho C K M, Malcolm B. Expanding a dairy business affects business and financial risk[J]. Animal Production Science, 2017, 57(10):2167–2174.
4. Fu W, Zhang L, Li C. Dynamic Access Control using Blockchain-based Attribute Encryption Scheme for Big Data Cloud Storage [J]. Engineering Intelligent Systems, 2025, 33(3):263–273.
5. Abeyrathna G M, Kalainathan K. Financial Risk, Financial Risk Management Practices and Performance of Sri Lankan SMEs: Special Reference to Anuradhapura District[J]. Research Journal of Finance and Accounting, 2016, 7(15):16–22.
6. Xu S. Model for evaluating the commercial banks financial risk with interval grey uncertain linguistic variables[J]. Journal of Intelligent & Fuzzy Systems, 2015, 28(2): 767–773.
7. Huang Z, Vaidyanathan S. On the Effectiveness of Graph Statistics of Shareholder Relation Network in Predicting Bond Default Risk[J]. Journal of Control Science and Engineering, 2022, 2022(Pt.1):1–11.
8. Hao Y, Qiu F. Research on the Application of DM Technology with RF in Enterprise Financial Audit[J]. Mobile Information Systems, 2022, 2022(Pt.8):4051469.1-4051469.9.
9. Lee C W, Kuo C K. Combining hazard rates with the CreditGrades model: A hybrid method to value CDS contracts[J]. International Journal of Financial Engineering, 2015, 02(04):1550037.
10. Sun J, Sun T. Research on the Effectiveness of KMV Model in China's Bond Credit Rating Market[J]. Journal of Financial Studies, 2020, 4(1):59–62.
11. Mihm B. Mispricing of Risk in Sovereign Bond Markets with Asymmetric Information[J]. German Economic Review, 2016, 17(4):491–511.
12. Robinson D. Actavis to test risk appetite with jumbo bond[J]. International Financing Review, 2015(2071):20–21.
13. Wan H, Yu Q, Ding J, Liu K. Students' behavior analysis under the Sakai LMS[C]//IEEE International Conference on Teaching Assessment, 2017:250–255.
14. Cui M. Big data medical behavior analysis based on machine learning and wireless sensors[J]. Neural Computing & Applications, 2022, 34(12):9413–9427.
15. Liu L, Zhao B, Rao Y. On the Cognitive Load of Online Learners With Multi-Level Data Mining[J]. International Journal of Information and Communication Technology Education: An Official Publication of the Information Resources Management Association, 2022, 18(Pt.1):134–148.
16. Šarić-Grgić I, Grubišić A, Šerić L, Robinson TJ. Student Clustering Based on Learning Behavior Data in the Intelligent Tutoring System[J]. International Journal of Distance Education Technologies (IJDET), 2020, 18(2):73–89.
17. Heng Shao. A New Intelligent Optimization Network Online Learning Behavior in Multimedia Big Data Environment[J]. International Journal of Mobile Computing and Multimedia Communications, 2017, 8(3):21–31.