A Study of Identification of Corporate Financial Fraud Using Neural Network Algorithms in an Information-Based Environment

Zilu Liang,[1] Yunji Liang[2,*]
[1]School of International Business, Dongbei University of Finance and Economics, Dalian, Liaoning 116012, China
[2]Accounting School, Harbin University of Commerce, Harbin, Heilongjiang 150028, China
Email: jispa48@163.com

Keywords: financial fraud, neural network, recognition, principal component analysis

Received: September 22, 2023

This paper provides a brief overview of corporate financial fraud behavior and the initial feature indicators utilized for detecting financial fraud. Principal Component Analysis (PCA) was employed to refine these feature indicators. Subsequently, the Back-Propagation Neural Network (BPNN) algorithm was applied for identification. Simulation experiments were conducted to test the BPNN algorithm's parameters. Additionally, a comparative analysis was conducted to compare the BPNN algorithm with the decision tree and Support Vector Machine (SVM) algorithms. The results demonstrated that PCA effectively reduced the initial set of 30 indicators to 20, retaining 90.64% of the essential information. The optimal configuration for the BPNN algorithm was seven hidden nodes and the application of the ReLU activation function. Furthermore, the BPNN algorithm outperformed the decision tree and SVM algorithms in the context of financial fraud recognition.

Povzetek: Članek opisuje uporabo PCA za zaznavanje finančnih goljufij in primerja DNN algoritem BPNN z metodo strojnega učenja odločitvenih dveves in SVM.

1 Introduction

In the era of digital transformation, enterprises are facing an increasingly complex and constantly evolving business landscape where financial fraud is becoming more prominent [1]. Corporate financial fraud encompasses actions taken by businesses to deceive investors, creditors, regulators, and other stakeholders by misrepresenting, concealing, or tampering with financial information in their financial statements, all with the aim of illicitly gaining advantages. This misconduct not only inflicts significant harm on the economic interests and reputation of the enterprise but also exerts adverse effects on the entire economic system [2]. Consequently, the timely identification and prevention of financial fraud has become an important task for both enterprises and regulatory authorities. Traditional methods of financial fraud detection typically involve manual examination by accountants, a process that demands substantial time and labor while often yielding unsatisfactory results. Leveraging their formidable capabilities in pattern recognition and learning, neural network algorithms present a promising avenue for detecting financial fraud [3]. Notable research efforts in this domain are as follows. Geng et al. [4] constructed a financial fraud detection model based on empirical data and found that the model yielded commendable identification outcomes. Li [5] utilized a financial report identification model grounded in the BPNN. Kanapickienė et al. [6] utilized financial ratios to uncover instances of financial fraud. This paper provides a concise overview of corporate financial fraud behavior and the initial feature indicators employed in its detection. This paper employed Principal Component Analysis (PCA) to refine these feature indicators and subsequently utilized the Back-Propagation Neural Network (BPNN) algorithm for financial fraud detection. Simulation experiments were performed to test the parameters of the BPNN algorithm and compare it with the decision tree and Support Vector Machine (SVM) approaches. A summary table of related works is shown in Table 1.

Table 1: A summary of related works

Main structure	Summary content
Abstract	The abstract summarizes the research content of this article and indicates that the algorithm used in this study can better identify financial fraud.
Introduction	The introduction provides a brief statement on financial fraud and the corresponding measures, along with a summary of several related works.
Main text	The main text describes a financial fraud detection algorithm based on a neural network algorithm and utilizes PCA for dimensionality reduction of indicators.
Experiment	The financial report data from the CSMAR database was used as the object for simulation experiments. The experiments tested the effects of

	different numbers of hidden layer nodes and types of activation functions on the algorithm and then compared it with the decision tree and SVM algorithms.
Result	PCA could used to reduce 30 initial indicators to 20 indicators, which contained 90.64% of the relevant information. The BPNN algorithm performed best with seven hidden nodes and using the ReLU activation function. Compared to the decision tree and SVM algorithms, the BPNN algorithm had better performance in detecting financial fraud.

2 Financial fraud recognition based on neural network algorithm

Financial fraud refers to the deliberate manipulation, modification, or fabrication of data within financial statements by either enterprises or individuals, with the intent of perpetrating fraud or concealing the true financial status. Financial fraud has significant economic and societal repercussions [7] and poses a grave threat to a company's financial stability and market standing. The origins of corporate financial fraud can be attributed to the significant importance placed on corporate financial statements and the limitations of traditional financial auditing methods [8]. However, verifying the authenticity of financial statements often proves difficult due to their complex preparation and large amounts of data involved. Traditional financial auditing methods primarily hinge on the external auditors' reviewing and sampling of financial statements, which limits their ability to uncover corporate financial fraud [9]. When employing neural network algorithms for financial statement analysis, the financial statements cease to be traditional paper-based documents; instead, they become digitized financial data stored within the corporate database. Neural network algorithms mimic the operational principles of human brain cells, allowing them to extract meaningful features from complex data and uncover potential instances of financial fraud [10].

2.1 Indicator screening for financial fraud identification

Before employing the neural network algorithm for identification, it is imperative to selectively screen relevant indicators that serve as characteristics for identifying financial fraud, which are essentially the attributes of financial statements [11]. In this study, the indicators for financial fraud identification are categorized into two groups: financial indicators and non-financial indicators. Financial indicators are directly associated with the financial data presented in the financial report, offering an intuitive reflection of the enterprise's financial performance; however, they can also be susceptible to manipulation. On the other hand, non-financial indicators shed light on the enterprise's managerial competence, and poor management practices will increase the likelihood of financial fraud [12]. The indicators outlined in Table 2 form the initial indicator system. Recognizing variations in the scales of different enterprises, the financial indicators employed are expressed as relative ratios. Additionally, it is worth noting that these indicators invariably utilize the same financial data to varying degrees during their calculation process, resulting in a certain degree of interdependence among them. The large number of feature indicators in financial statements and the multicollinearity resulting from their correlation both increase the difficulty of analyzing patterns. To mitigate the potential interference of redundant information within high-dimensional data and extract the essential patterns, dimensionality reduction is necessary [13].

Table 2: Hierarchical indicators used for financial fraud identification

Target layer	Criterion layer 1	Criterion layer 2	Indicator layer	Indicator number
Indicator system for financial fraud identification	Financial indicators	Profitability	Return on net assets	X1
			Return on total assets	X2
			Total operating cost ratio	X3
			Cost-effectiveness ratio	X4
			Net profit margin on sales	X5
			Non-operating profit ratio	X6
		Accruals	Excess cash flow variance	X7
			Operating accruals	X8
		Solvency	Current ratio	X9
			Quick ratio	X10
			Net cash ratio	X11
			Total asset-liability ratio	X12
		Turnaround capacity	Accounts receivable	X13

			turnover ratio	
			Accounts payable turnover	X14
			Current asset turnover ratio	X15
			Total asset turnover	X16
		Asset quality	Current assets ratio	X17
			Ratio of cash assets	X18
			Working capital ratio	X19
			Fixed asset ratio	X20
			Inventory current assets ratio	X21
		Development capacity	Total asset growth rate	X22
			Roe growth rate	X23
			Net profit growth rate	X24
		Risk level	Financial leverage	X25
			Business leverage	X26
	Non-financial indicators	Governance structure	Number of board meetings	X27
			Number of shareholders' meetings	X28
			Shareholding ratio of top ten shareholders	X29
		External audit	Audit opinion	X30

2.2 The process of recognizing financial fraud

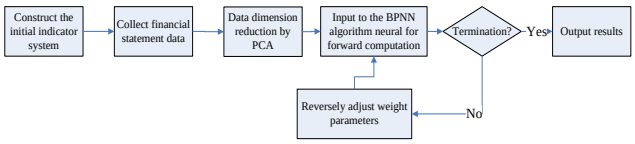

Figure 1: Recognition process of financial fraud based on multi-layer BPNN

After constructing the initial indicators for identifying financial fraud, the financial statement data is collected based on these indicators. Then, dimensionality reduction is applied to the indicators [14]. Compared with several classification algorithms mentioned earlier, the BPNN algorithm is able to approximate any continuous function in the range of real numbers by using the fully connected nodes of the hidden layer, which can effectively mine the hidden laws under the financial statement feature indicators. Its process is shown in Figure 1.

① The initial indicator system consisting of the hierarchical structure indicators shown in Table 1 constructed in the previous section is constructed.

② The data of financial statements are collected according to the constructed initial indicator system and standardized to eliminate the influence of different scales. The processing formula is:

$$x_i' = \frac{x_i - \bar{x}}{s}, \quad (1)$$

where x_i is the i-th number in the sequence, $\bar{x}$ is the mean of the sequence that x_i belongs to, s is the standard deviation of the sequence, and x_i' is the standardized data.

③ The dimensionality of the data is reduced using PCA. Firstly, sample data are transformed into a matrix in a size of $i \times j$, where i represents the number of samples, j stands for the indicator number of samples. Then, the data matrix is standardized to eliminate the dimensional difference between different indicators. Next, the correlation coefficient matrix of the standardized data matrix is constructed. The eigenvalue and corresponding eigenvector of the correlation coefficient matrix are calculated. Finally, the variance contribution rate of each indicator is calculated. Moreover, the indicator with a cumulative variance contribution rate exceeding 85% is taken as the principal component. The formula for the indicator contribution rate [15] is:

$$a_j = \frac{\lambda_j}{\sum_{k=1}^{p} \lambda_k}, \quad (2)$$

where a_j is the variance contribution ratio of the j th indicator, λ_j and λ_k are the eigenvalues of the j-th and k-th indicators, respectively, and p is the total number of eigenvalues.

④ The financial statement sample data after PCA dimensionality reduction is input into the input layer of the multi-layer BPNN, and then forward computation is carried out in multiple hidden layers with the formula:

$$o_j = f\left(\sum_{i=1}^{n} \omega_i x_i - b\right), \quad (3)$$

where o_j is the output vector of the hidden layer, b is adjustment term of the hidden layer, and $f(\bullet)$ is the activation function of the hidden layer, which can be linear or nonlinear.

⑤ If the algorithm is in the training completion stage, the judgment results are output according to the calculation results of the previous step. If the algorithm is in the training stage, whether the training is terminated or not is determined. If it is terminated, the training stops, and the results are output; if it is not terminated, the weight parameters in the algorithm are adjusted in the reverse direction, and step ④ is performed again. The termination conditions included reaching the maximum number of training or the computational error converged to the set range. Cross entropy was utilized as the measure of computational error. The calculation formula of the cross entropy is:

$$E = -\sum_i y_i' \log(y_i), \quad (4)$$

where H_1 is the cross entropy between the recognition result and the actual result during training, y_i is the recognition result during training, and y_i' represents the actual result.

3 Simulation experiment

3.1 Data sources

The required financial statements were obtained from the China Stock Market Accounting Research (CSMAR) database. 60% of these statements were utilized as a training set, while the remaining 40% served as an independent test set.

3.2 Experimental setup

(1) Parameter setting of BPNN

The input layer was configured with 20 nodes, while the output layer consisted of a single node. Two hidden layers were employed. The number of nodes in the hidden layer was set to 5, 6, 7, 8, 9 and 10. The activation functions in the hidden layer were set to ReLU, sigmoid and tahn, respectively, to test the recognition performance of the BPNN algorithm under different numbers of nodes in the hidden layer and activation functions.

(2) Performance comparison between different classification recognition algorithms

The BPNN algorithm was compared with two algorithms, SVM and decision tree. The decision tree algorithm [16] used the C5.0 decision tree model, and the depth of the model dendrogram was set to 22. In the SVM algorithm [17], the kernel function was a sigmoid function, and the penalty parameter was set to 2.

Table 3: The confusion matrix

	Judged as positive	Judged as negative
Positive actually	*TP*	*FN*
Negative actually	*FP*	*TN*

The binary confusion matrix as used to assess the algorithm performance. The confusion matrix is presented in Table 3. The corresponding calculation formulas are:

$$\begin{cases} P = \dfrac{TP}{TP + FP} \\ R = \dfrac{TP}{TP + FN} \\ F = \dfrac{2PR}{P + R} \end{cases}, \quad (5)$$

where P stands for precision, R stands for recall rate, and F represents the comprehensive consideration of the precision and recall rate.

3.3 Experimental results

Before using the classification algorithm to classify and recognize the financial statements, the initial indicators' dimensionality was firstly reduced by PCA. Table 4 shows the eigenvalues, contribution rates, and cumulative contribution rates of the initial indicators after PCA calculation. In Table 4, the indicators have been ranked according to the contribution rate from high to low. It can be seen from Table 4 that the cumulative contribution rate of the first 20 indicators after sorting was 90.64%, which meant that the first 20 indicators contained 90.64% of the information in the samples. Therefore, this paper took the first 20 indicators after sorting as the indicators after dimensionality reduction.

Table 4: Calculation results of PCA for initial indicators

Indicator number	Eigenvalue	Contribution rate/%	Cumulative contribution/%	Indicator number	Eigenvalue	Contribution rate/%	Cumulative contribution rate/%
X1	5.23	11.77	11.77	X27	0.76	1.71	84.22
X2	4.12	9.27	21.04	X19	0.74	1.67	85.89
X6	4.11	9.25	30.29	X23	0.73	1.64	87.53
X8	3.75	8.44	38.74	X3	0.71	1.60	89.13
X11	3.21	7.22	45.96	X4	0.67	1.51	90.64
X12	3.11	7.00	52.96	X5	0.61	1.37	92.01
X16	2.87	6.46	59.42	X7	0.57	1.28	93.29
X15	2.45	5.51	64.93	X9	0.52	1.17	94.46
X20	2.14	4.82	69.75	X10	0.49	1.10	95.57
X17	1.21	2.72	72.47	X13	0.43	0.97	96.53
X21	1.01	2.27	74.75	X14	0.41	0.92	97.46
X24	0.89	2.00	76.75	X18	0.37	0.83	98.29
X25	0.87	1.96	78.71	X28	0.31	0.70	98.99
X29	0.86	1.94	80.64	X22	0.24	0.54	99.53
X30	0.83	1.87	82.51	X26	0.21	0.47	100.00

Figure 2 shows the classification performance of the BPNN algorithm under different numbers of hidden layer nodes and types of activation functions. The F-value, which reflects the recognition performance of the model, was used as the measure. The classification performance of the BPNN algorithm with ReLU activation function was the best, followed by tahn and sigmoid, as shown in Figure 2 under the same number of hidden layer nodes. Additionally, when the number of hidden layer nodes was 7, the BPNN algorithm achieved the best classification performance regardless of the type of activation function.

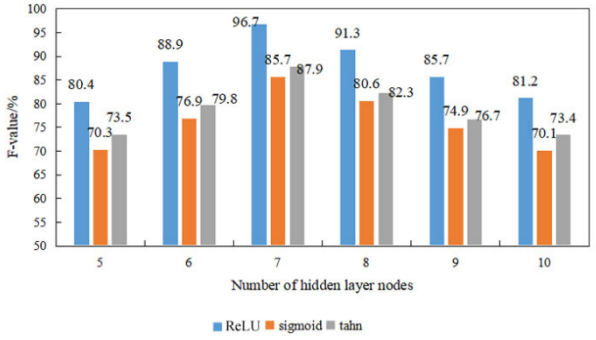

Figure 2: Classification performance of the BPNN algorithm under different numbers of hidden layer nodes and activation functions

After determining the parameter settings of the BPNN algorithm, it was compared with the other two classification algorithms, and the results are presented in Figure 3. As shown in Figure 3, both the precision, recall rate, and F-value were the highest for the BPNN algorithm, followed by the SVM algorithm, and the decision tree had the worst performance.

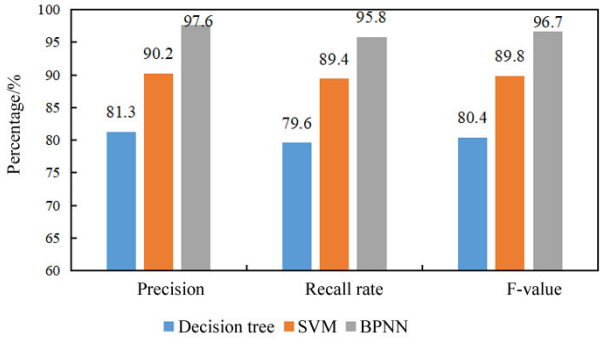

Figure 3: Recognition performance of three classification algorithms for financial frauds

4 Discussion

With the rapid development and extensive application of information technology, the forms of corporate financial fraud have become increasingly diverse and complex. Traditional methods for identifying financial fraud, such as statistical analysis and rule engines, are no longer sufficient to accurately detect new types of fraudulent behavior. Neural network algorithms, as a powerful machine learning tool, have demonstrated unique advantages in identifying corporate financial fraud. This article adopted the classic BPNN algorithm in the field of neural networks. During the process of identifying financial reports, the BPNN algorithm utilizes activation functions in the hidden layer to explore and uncover patterns and rules hidden within complex financial data. The conventional procedure involves analyzing the relevant indicators used to identify financial fraud and then inputting the data of these indicators as features into a BPNN for calculation, ultimately obtaining results. However, when selecting relevant indicators, it is

important to consider both the quantity and effectiveness of the indicators. If there are too many indicators, it will increase computational complexity, and irrelevant indicators may actually interfere with identification. Therefore, this article adopts PCA for principal component analysis of the indicators in order to eliminate those with low effective components. This approach aims to reduce algorithmic computation while maintaining recognition effectiveness as much as possible. Finally, simulation experiments were conducted, and the results are shown as mentioned above. PCA ultimately selected 20 indicators, which contain 90.64% of the relevant information. It was verified that the BPNN algorithm with relu activation function performed the best with seven hidden layer nodes. Compared to the decision tree and SVM algorithms, the BPNN algorithm exhibited superior performance. The reasons for the above results were analyzed. When using decision tree algorithm to identify financial fraud, it classifies data into two categories based on indicators sequentially. During the training process, the selection of branching nodes is determined by information entropy. However, this method assumes that the indicators are independent from each other, while in reality there exists certain correlation among different indicators. The SVM algorithm partitions data by finding support vectors that act as 'hyperplanes'. However, even with the use of kernel functions to map data into higher-dimensional spaces, it is difficult to fit nonlinear patterns. In contrast, the BPNN algorithm is capable of fitting hidden nonlinear patterns in the data using activation functions in the hidden layers, resulting in better recognition performance.

The novelty of this article lies in combining PCA with the BPNN algorithm, utilizing PCA to select effective indicators and reduce their quantity, thereby reducing the computational burden of the BPNN algorithm. However, a limitation of this study is that the range of the dataset used is not extensive enough. A future research direction is to expand the scope of the dataset.

5 Conclusion

This paper briefly described corporate financial fraud behavior and the initial feature indicators employed in its detection. It utilized PCA to refine these feature indicators and employed the BPNN algorithm for financial fraud identification. Subsequently, simulation experiments were conducted to test the parameters of the BPNN algorithm and compare the BPNN algorithm with the decision tree and SVM algorithms. The results are summarized as follows. (1) Following PCA computation, each initial indicator was ranked based on their contribution rates, from highest to lowest. The cumulative contribution rate of the top 20 indicators was found to be 90.64%, and these 20 indicators were retained as the reduced-dimension indicators. (2) The optimal configuration for the BPNN algorithm was the use of seven hidden nodes and the ReLU activation function. (3) In terms of financial fraud identification performance, the BPNN algorithm performed the best, followed by the SVM algorithm, while the decision tree algorithm exhibited the least favorable performance.

References

[1] Du M (2021). Corporate governance: five-factor theory-based financial fraud identification. *Journal of Chinese Governance*, 6, pp. 1-19. https://doi.org/10.1080/23812346.2020.1803036

[2] Zandian Z K, Keyvanpour M (2017). Systematic identification and analysis of different fraud detection approaches based on the strategy ahead. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, 21, pp. 123-134. https://doi.org/10.3233/KES-170357

[3] Han D (2017). Researches of Detection of Fraudulent Financial Statements Based on Data Mining. *Revista de la Facultad de Ingenieria*, 14, pp. 32-36. https://doi.org/10.1166/jctn.2017.6119

[4] Geng X, Yang D (2021). Intelligent Prediction Mathematical Model of Industrial Financial Fraud Based on Data Mining. *Mathematical Problems in Engineering*, 2021, pp. 1-8. https://doi.org/10.1155/2021/8520094

[5] Li S L (2020). Data mining of corporate financial fraud based on neural network model. *Computer Optics*, 44, pp. 665-670. https://doi.org/10.18287/2412-6179-CO-656

[6] Kanapickiene R, Grundienė Ž (2015). The Model of Fraud Detection in Financial Statements by Means of Financial Ratios. *Procedia - Social and Behavioral Sciences*, 213, pp. 321-327. https://doi.org/10.1016/j.sbspro.2015.11.545

[7] Lin C C, Chiu A A, Huang S Y, Yen D C (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems*, 89, pp. 459-470. https://doi.org/10.1016/j.knosys.2015.08.011

[8] Coakley J R, Brown C E (2015). Artificial neural networks applied to ratio analysis in the analytical review process. *Intelligent Systems in Accounting Finance & Management*, 2, pp. 19-39. https://doi.org/10.1002/j.1099-1174.1993.tb00032.x

[9] Compin F (2015). Tax fraud: a socially acceptable financial crime in France?. *Journal of Financial Crime*, 22, pp. 432-446.

[10] Chen T, Xu W (2018). Post-evaluation on financial support highway traffic project based on BP neural network algorithm. *Journal of Discrete Mathematical Sciences and Cryptography*, 21, pp. 869-879. https://doi.org/10.1080/09720529.2018.1480277

[11] Hong Y, Sun W, Qianling B, Mu X (2016). SOM-BP Neural Network-Based Financial Early-Warning for Listed Companies. *Journal of Computational & Theoretical Nanoscience*, 13, pp. 6860-6866. https://doi.org/10.1166/jctn.2016.5638

[12] Fanning K, Cogger K O (2015). Neural network detection of management fraud using published

financial data. *Intelligent Systems in Accounting Finance & Management*, 7, pp. 21-41.

[13] Liu S J, Li S L, Jiang M, He D (2017). Quantitative Identification of Pipeline Crack Based on BP Neural Network. *Key Engineering Materials*, 737, pp. 477-480. https://doi.org/10.4028/www.scientific.net/kem.737.477

[14] He G, Huang C, Guo L, Sun G, Zhang D (2017). Identification and Adjustment of Guide Rail Geometric Errors Based on BP Neural Network. *Measurement Science Review*, 17, pp. 135-144. https://doi.org/10.1515/msr-2017-0017

[15] Yacoub H A, Sadek M A (2016). Identification of fraud (with pig stuffs) in chicken-processed meat through information of mitochondrial cytochrome b. *Mitochondrial DNA, Part A DNA Mapping Sequencing & Analysis*, 28, pp. 855-859. https://doi.org/10.1080/24701394.2016.1197220

[16] Tang X B, Liu G C, Yang J, Wei W (2018). Knowledge-based Financial Statement Fraud Detection System: Based on an Ontology and a Decision Tree. *Knowledge Organization*, 45, pp. 205-219. https://doi.org/10.5771/0943-7444-2018-3-205

[17] Li C, Ding N, Zhai Y, Dong H (2021). Comparative study on credit card fraud detection based on different support vector machines. *Intelligent Data Analysis*, 25, pp. 105-119. https://doi.org/10.3233/IDA-195011