Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 8, No. 2, 2026

Financial Fraud Data Analysis Based on Genetic Neural Networks

Ganzhou Wu

School of science, Guangdong University of Petrochemical Technology, Maoming 525000, China

Abstract: It is used the Guotai An violation database to select 75 listed companies that were punished for financial fraud from 2011 to 2020 as fraud samples, and uses Beasley's matching principle to select 75 non fraud companies that match the fraudulent companies as matching samples. Among the 21 preliminary indicators, SPSS was used for significance testing, and factor analysis was used to reduce the significance indicators. Finally, 5 variables were selected as input variables for the data mining model, while fraud (marked as 2 for fraud and 1 for non fraud) was used as the output variable for identifying financial fraud. Using a GA-BP neural network algorithm to train and model the training group sample data, the accuracy rate reached 87.16%.

Keywords: Significance test, GA-BP neural network, Financial data.

1. Introduction

Financial fraud refers to companies or individuals using improper means to mislead investors and other stakeholders in financial statements in order to achieve certain benefits. Financial fraud not only causes losses to investors and stakeholders, but also has a negative impact on financial markets and economic stability [1]. Therefore, the prevention and monitoring of financial fraud have become important tasks in ensuring market fairness and stability. There are various types of financial fraud, including false sales revenue, fictitious assets, concealment of liabilities, and inflated profits. In order to better prevent and monitor financial fraud, researchers and practitioners have proposed various methods and techniques, among which data mining technology is one of the commonly used methods [2]. Data mining technology assists financial analysts or regulatory agencies in detecting potential financial fraud by mining potential patterns and information in large amounts of data [3]. The application of data mining technology in financial fraud detection has become a popular research field. Data mining techniques can extract useful information from large amounts of financial data, helping analysts discover patterns and trends hidden behind the data and identify potential financial fraud behaviors [4]. At present, data mining technology is mainly applied in anomaly detection, pattern recognition, and classification prediction in financial fraud detection [5]. From the perspective of economic and scientific development, the occurrence of financial data fraud in listed companies is inevitable due to factors such as market competition pressure, inaction of regulatory and auditing agencies, or inadequate supervision [6]. With the increasingly serious phenomenon of financial fraud, research on financial fraud identification models is receiving more and more attention, and its development is also becoming faster and faster. Early research was mainly based on traditional statistical methods such as Bayesian networks, logistic regression, discriminant analysis, etc [7]. With the development of artificial intelligence and machine learning technologies, more and more studies have begun to use machine learning based methods such as random forests, decision trees, SVM support vector machines, neural networks, etc. In addition, financial fraud identification models have been applied in various fields, including finance, accounting, auditing, etc. In the future, financial fraud identification models will be more widely applied under the trend of continuous technological development and increasing data. Zhang Jiajia used data mining techniques to construct a financial reporting fraud identification model for listed companies [8]. Research has shown that this model can effectively detect financial reporting fraud in listed companies. There are still some issues in domestic research. Firstly, the definition and classification of financial fraud need to be further clarified. Secondly, the quality and quantity of data are also important factors affecting the accuracy of financial fraud detection. In addition, domestic research needs to pay more attention to the testing and evaluation of practical application effects.

2. Data Preprocessing

The empirical research object of this article is financial fraud companies, and the sample is selected from the Guotai An violation database, mainly including six types of fraud: insider trading, delayed disclosure, major omissions, fictitious assets, fictitious profits, and false records. The research time span is 10 years from 2011 to 2020. In order to ensure the quality of the sample and avoid duplication, this study selected non-financial companies that had experienced fraud for the first time or had a gap of more than 3 years between two consecutive frauds as the research sample. After preliminary selection and screening, this article finally obtained a sample of 75 financially fraudulent listed companies, excluding those with missing data.

The sample was selected from the Guotai An violation database, which includes 21 indicators. However, applying all these indicators to data mining models may lead to overfitting and reduce modeling efficiency. In addition, there may be interrelationships between these indicators, which may also affect the interpretation of data mining results. Therefore, simplifying the number of indicators is more conducive to the interpretation and accuracy of the model structure. Therefore, before establishing a financial fraud identification model, it is necessary to screen the initially selected financial fraud identification indicators to select indicators that have a significant impact on the model structure and remove redundant information.

We choose the Mann Whitney test in non parametric testing

to remove redundant information and uses SPSS26 for analysis. The results are shown in Table 4.3. The 10 indicators of X3 board size, X5 audit opinion, X10 total asset turnover rate, X11 current asset turnover rate, X13 total asset net profit margin (ROA), X14 return on equity (ROE), X15 return on invested capital (ROIC), X17 total asset growth rate, X18 net profit growth rate, and X19 operating income growth rate show statistical significance in the significance test, with a significance level of less than 0.05. This includes 2 non-financial indicators and 8 financial indicators, which show significant differences between fraudulent and non fraudulent samples, with high discrimination.

Table 1. Mann Whitney U test for non parametric testing

	Mann-Whitney
X1	0.891
X2	0.397
X3	0.013
X4	0.581
X5	0.007
X6	0.581
X7	0.669
X8	0.252
X9	0.776
X10	0.008

These 10 indicators reflect different characteristics of financial fraud in companies, but if all indicators are applied to the data mining model, it may lead to overfitting and bias in the model. Therefore, this article adopts factor analysis to reduce the dimensionality of these indicators, in order to reduce information redundancy and improve the accuracy of the model. However, factor analysis also needs to meet some prerequisites. Usually, factor analysis can be used to reduce the dimensionality of a dataset, but before application, KMO test and sphericity Bartlett test need to be performed to evaluate whether the dataset is suitable for applying this method.

Firstly, KMO and Bartlett tests were conducted on these financial indicators to determine their correlation and whether factor analysis is suitable for dimensionality reduction. The KMO value is 0.652, which is greater than the critical value of 0.5, and the Bartlett sphericity test results in a P value of 0, which is less than the critical value of 0.05, indicating a significant correlation between these financial indicators. This meets the prerequisite for conducting factor analysis. Therefore, factor analysis can be used to reduce the dimensionality of these financial indicators. After inputting the data into SPSS26 for factor analysis dimensionality reduction, the results are shown in Table 4-5 for variance explanation. It can be seen from the table that the cumulative variance contribution rate of the first three factors reached 84.26%. This means that these three common factors can explain nearly 84% of the overall information content, meeting the standard of 80% -90% of the required information content. Meanwhile, the amount of information loss is relatively small and within an acceptable range. Therefore, extracting these three common factors can effectively summarize the information contained in the original financial indicators.

In order to better explain the significance of factors, this study used the variance maximization method of orthogonal rotation to rotate the factor loading matrix. The rotated factor loading matrix is easier to interpret and can clearly display the contribution of each variable to each factor. According to this matrix, it can be seen that the first common factor mainly includes financial indicators such as X13 net asset profit margin (ROA), X14 return on equity (ROE), X15 return on invested capital (ROIC), X17 total asset growth rate, X18 net profit growth rate, etc., indicating that this factor mainly reflects the company's profitability. The second common factor is mainly composed of X10 total asset turnover rate and X11 current asset turnover rate, indicating that this factor mainly reflects the company's operating ability.

The third factor mainly includes the X19 revenue growth rate, indicating that this factor mainly reflects the company's development capability.

After factor analysis, the 8 financial indicators were compressed into 3 factors, with F1, F2, and F3 representing profitability, operational capability, and development capability, respectively. By observing the score coefficient matrix of components in Table 4-7, the final factor score formula can be derived, which means that the score of each factor is weighted and summed by a certain weight coefficient and the value of the original indicator.

$$F1= 0.005X10-0.014X11+0.259X13+0.262X14+0.265X15+0.132X17+0.202X18+0.061X19 \quad (1)$$

$$F2 = 0.546X10+0.532X11-0.001X13+0.017X14+0.008X15+0.013X17-0.045X18+0.054X19 \quad (2)$$

$$F3 = 0.094X10+0.020X11+0.043X13+0.081X14+0.069X15+0.229X17-0.333X18+0.837X19 \quad (3)$$

After factor analysis of financial and non-financial indicators, this study ultimately selected five variables as inputs for the model, including two categorical variables X3 board size and X5 audit opinion, as well as three principal component variables F1, F2, and F3 obtained through factor analysis of financial indicators. These three variables represent the company's profitability, operating ability, and development ability, respectively. The final selected variables are detailed in Table 2, and these 5 variables are also the final input variables for the neural network.

Table 2. The variables ultimately selected for the model

variable type	Specific variables	Number of variables
categorical variable	X3, X5	2
Principal component variable	F1, F2, F3	3
total	X3, X5, F1, F2, F3	5

3. GA-BP Neural Network

Genetic algorithm is a population optimization algorithm that utilizes the evolutionary rules of the biological world and iteratively evolves through genetic operators (crossover, mutation) to find approximate or optimal solutions to problems. The steps of GA algorithm include transforming the problem into gene encoding, initializing the population, calculating the fitness value of each individual through fitness evaluation function, selecting the optimal individual, generating new individuals through crossover and mutation, and repeating this process until the maximum number of iterations is reached or the optimal solution is found. Among them, the encoding scheme is to transform the potential solutions of the problem into mapping genes, while fitness

evaluation is to calculate the fitness value of each individual based on the objective function of the problem, in order to carry out the selection of survival of the fittest. In crossover and mutation operations, individuals from the population are randomly selected for combination, and mutation operations are performed on individual genes to generate new offspring individuals, thereby continuously updating the entire population and evolving towards a better direction. Ultimately, by decoding the individuals with the highest fitness in the last generation population, an approximate or optimal solution to the problem can be obtained.

Genetic algorithm (GA) can avoid the defect of BP neural network algorithm falling into local optima, thereby improving the stability of the network. The basic idea of GA-BP neural network is to combine genetic algorithm and BP neural network, using the initial weights and thresholds of BP neural network as individuals of genetic algorithm, and evaluating the adaptability of individuals through fitness function values. Then, genetic algorithms such as selection, crossover, and mutation are used to generate new individuals, continuously optimizing the weights and thresholds of the neural network. Ultimately, the optimal initial weights and thresholds, as well as the final neural network model, can be used for tasks such as prediction and classification.

This article uses GA-BP neural network to overcome the shortcomings of BP neural network's "randomness of initial weights and thresholds". By using the "trial and error method" to determine the number of hidden layer neurons, and optimizing the initial weights and thresholds based on genetic algorithm, traditional BP neural network is used to fit the collected sample data. In this study, the sample data was divided into a training set and a testing set, and a GA-BP neural network model was constructed. The model was trained using a training set and simulated with a test set for backtesting. The research results show that there is a potential correlation between input signals and financial fraud. By using a test set for prediction, the researchers successfully evaluated the recognition ability of the model and achieved accurate identification of financial fraud in listed companies.

In this section, 150 samples will be used for the experiment, including 75 fraudulent samples and 75 non fraudulent samples. These samples will be randomly divided into two groups. One group will be used as the learning sample for the financial fraud recognition model, consisting of 55 fraudulent samples and 55 non fraudulent samples. The other group will be used as the testing sample, consisting of 20 fraudulent samples and 20 non fraudulent samples, to test the recognition ability of the model. In order to ensure the learning ability of the model, the proportion of training samples selected in this article should be greater than that of test samples, accounting for 70% of the total. This chapter uses MatlabR2025b to build a data mining model.

In this paper, the input variables have a total of 5 dimensions, namely X3, X5, F1, F2, and F3, representing two categorical variables and three principal component variables. The output variable is one-dimensional, indicating whether the company is suspected of financial fraud. If the output value is less than 1.5, it is judged as 1 (non fraudulent company), and if the output value is greater than 1.5, it is judged as 2 (fraudulent sample). The process of establishing a GA-BP neural network model can be divided into two parts: designing the BP neural network structure and optimizing the weights and thresholds of the BP neural network using genetic algorithms.

When determining the number of hidden layer nodes, this article uses empirical formulas, After repeated testing, by continuously adjusting the number of hidden layer nodes and observing the performance of the model, the final number of hidden layer nodes was determined to be 10. When determining the structure of the BP neural network, the number of input layer nodes is 5, the number of output layer nodes is 1, and the value range of the constant a is between 1 and 10. By trying different numbers of hidden layer nodes, the range of 2n hidden layer nodes was ultimately determined to be 4 to 13. Through continuous experimentation, it was found that when the number of hidden layer nodes is set to 10, the root mean square error is minimized. Therefore, the structure of the BP neural network model constructed in this article is 5-5-1, as shown in Figure 1.

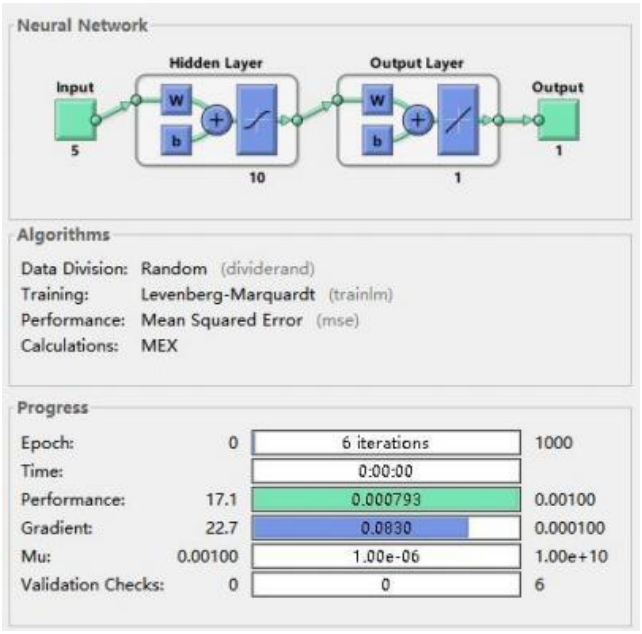

Figure 1. BP neural network structure diagram and training

In the modeling process of neural networks, it is necessary to determine the transfer function and learning function. This article chooses the tan sig function as the transfer function of the hidden layer, purelin function as the transfer function of the output layer, learngdm function as the learning function of the model, and train lm function as the training function of the model. Among them, the tan sig function is a commonly used hyperbolic tangent function that can map the input value to the interval between -1 and 1; The purelin function directly takes the input value as the output value; The learngdm function is a gradient descent learning algorithm that can minimize errors by continuously adjusting the weights and thresholds of the network; The train lm function is a training function based on the Levenberg Marquardt algorithm, which can converge to the global optimal solution faster

To train a BP neural network model, it is necessary to determine some training parameters. This article chooses a learning rate of 0.01, a minimum system error of 10-3, and a maximum training frequency of 1000 times. During the training process, 110 samples from the training sample set were introduced into the model for training. After 6 iterations of training, the model achieved the preset error and exhibited a fast training speed.

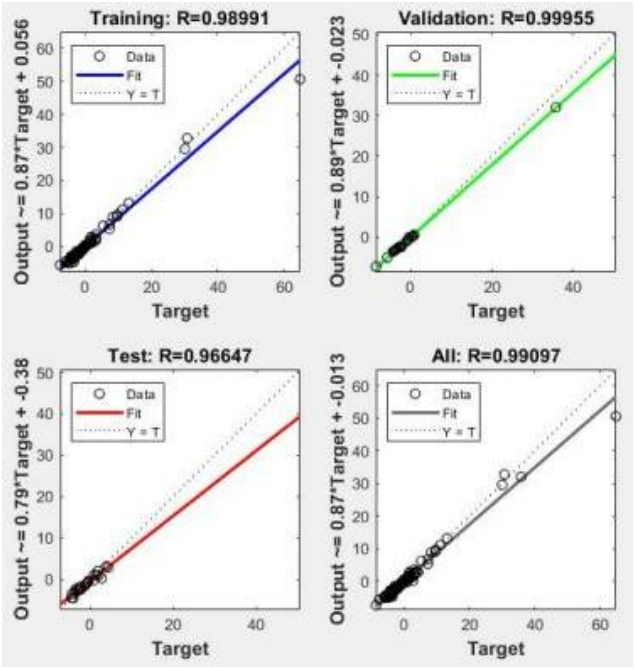

Figure 2. BP neural network training goodness of fit

According to the goodness of fit of the BP neural network training shown in Figure 2, it can be seen that the model performs well on both the training and testing sets: (1) the goodness of fit between the output values of the model on the training set and the actual values is 0.98991;

(2) The goodness of fit between the output values of the model on the test set and the actual values is 0.96647; (3) The goodness of fit of both the training and testing sets is good and close, indicating that the model has good stability.

Optimizing the weights and thresholds of BP neural network using genetic algorithm.

First, encode and generate the initial population, which involves determining the structural parameters of the neural network. The neural network structure includes the input layer, output layer, and hidden layer, with the corresponding number of neurons represented by i, j, and k respectively. The weight matrices W1 and W2 denote the connection weights between the hidden layer and output layer, while the threshold matrices B1 and B2 represent the thresholds of the hidden and output layers. These weights and thresholds are connected in sequence to form a chromosome. The encoding length is $i \times k + k \times j + k + j$. Among them, the first $i \times k$ genes correspond to W1, the subsequent $k \times j$ genes correspond to W2, and the final k and j genes correspond to B1 and B2. Using this encoding method, the initial population M is generated.Secondly, this study employs the reciprocal of MSE (Mean Squared Error) as the fitness function for the genetic algorithm to evaluate the quality of each individual. A higher fitness value indicates better fitting performance and a greater likelihood of being selected for inheritance to the next generation. Specifically, the fitness function is defined as the reciprocal of the MSE value, where MSE represents the mean squared error between the actual value A and the expected value T. Since a smaller MSE value signifies a closer match between the actual and expected values, the corresponding fitness function value will be larger.

Combining genetic algorithms with BP neural networks can effectively enhance the recognition performance of neural networks. Therefore, this paper adopts a BP neural network model based on genetic algorithms to predict the probability of corporate financial fraud. We apply genetic algorithms to the training of neural networks to obtain initial weights and thresholds.

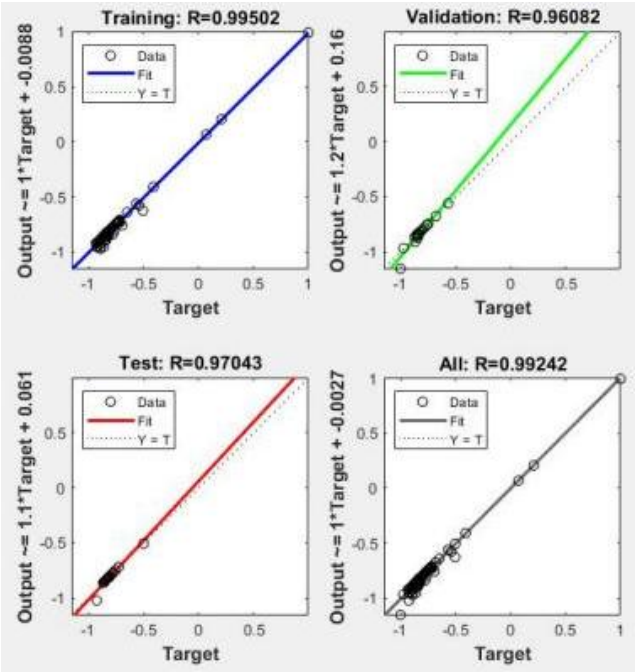

Figure 3. GA-BP neural network training goodness of fit

According to the results of the GA-BP neural network training goodness of fit in Figure 3, we can see that the model has improved the goodness of fit between the predicted results and the true values on the test set compared to using only the BP neural network model, reaching 0.97043. In addition, by observing and comparing Figure 2 and Figure 3, we can find that the GA-BP model optimized by genetic algorithm is more effective than using only the BP neural network model.

4. Experimental Results

Table 3. GA-BP neural network training set recognition rate

Category	Normal	Fraud	Recognition rate	Average recognition rate
Normal	50	5	90.20%	89.43%
Fraud	49	6	88.65%	

In Table 3, we can see the regression results for the training set of the neural network. Among the 55 samples of financial fraud in enterprises, 50 were correctly identified, while the remaining 5 samples were misjudged as non fraudulent. Therefore, the recognition accuracy of the model for fraudulent samples is 90.20%; In addition, among the 55 non fraudulent samples, 49 were correctly identified and 6 were misjudged as fraudulent samples. Therefore, the recognition accuracy of the model's non fraudulent samples is 88.65%. Overall, the recognition rate of the model is 89.43%. This indicates that the GA-BP neural network model proposed in this article demonstrates high accuracy in identifying corporate financial fraud.

Table 4. GA-BP Neural Network Test Set Recognition Rate

Category	Normal	Fraud	Recognition rate	Average recognition rate
Normal	18	2	88.56%	87.16%
Fraud	17	3	85.76%	

According to the test results shown in Table 4 GA-BP neural network test set recognition rate table, for the 40 samples in the test group, the GA-BP neural network model can recognize a large number of fraudulent and non fraudulent samples with high recognition accuracy. Among the 20 fraudulent samples, 18 were correctly identified, with a recognition accuracy rate of 88.56%. Out of 20 non fraudulent samples, 17 were correctly identified, with a recognition accuracy rate of 85.76%. Overall, the GA-BP neural network model achieved a discrimination accuracy of 87.16% in the test group.

5. Summary

The selected indicator variables were transformed into categorical variables and principal component variables to further improve the accuracy and stability of the model. In terms of data mining, this article uses the GA-BP neural network model for identifying and classifying financial fraud. Through training and testing on the training and testing sets, a discrimination accuracy of up to 87% was obtained, indicating that the model has good recognition ability and reliability.

References

[1] Cheng, X. J. (2018). Research on Financial Fraud Identification of Chinese Listed Companies Based on Data Mining [Master's thesis]. Chongqing University of Technology. [In Chinese]

[2] Zhong, S. Q. (2021). Research on Data Preprocessing Framework Based on Machine Learning [Master's thesis]. Xi'an University of Technology. [In Chinese]

[3] Huang, C. (2021). Research and Implementation of Anomaly Detection System Based on Time Series Data Mining [Master's thesis]. Zhejiang University. [In Chinese]

[4] Li, Q., & Ren, C. Y. (2016). Research on the construction and early warning of accounting fraud risk index for listed companies. Journal of Xi'an Jiaotong University (Social Sciences Edition), (01), 36–43. [In Chinese]

[5] Chen, G. X., Lv, Z. J., & He, F. (2007). Empirical study on the identification of financial reporting fraud: Based on empirical data of Chinese listed companies. Audit Research, 03, 88–93. [In Chinese]

[6] Dechow, P. M., Ge, W., Larson, C. R., et al. (2011). Predicting material accounting misstatements. Contemporary Accounting Research, 28(1), 17–82.

[7] Yuan, C. S., & Han, H. L. (2008). The mechanism and empirical test of the influence of board size on financial fraud. Business Economics and Management, (03), 44–49. [In Chinese]

[8] Zhang, J. J. (2021). Research on the Identification Model of Financial Reporting Fraud in Listed Companies Based on Data Mining [Master's thesis]. Zhejiang University. [In Chinese]