Data mining of corporate financial fraud based on neural network model

S.L. Li[1]

[1] Accounting Department, Business School, Changchun Guanghua University, Changchun, Jilin 130033, China

Abstract

Under the active market economy, more and more listed companies emerge. Because of the various interest relationships faced by listed companies, some enterprises which are not well managed or want to enhance company's value will choose to forge financial reports by improper means. In order to find out the false financial reports as accurately as possible, this paper briefly introduced the relevant indicators for judging the fraudulence of financial reports of listed companies and the recognition model of financial reports based on back propagation (BP) neural network. Then the selection of the input relevant indexes was improved. The improved BP neural network was simulated and analyzed in MATLAB software and compared with the traditional BP neural network and support vector machine (SVM). The results showed that the importance of total assets net profit, earnings per share, cash reinvestment rate, operating gross profit and pre-tax ratio of profit to debt was the top 5 among 20 judgment indexes. In the identification of testing samples of financial report, the accuracy, precision, recall rate and F value all showed that the performance of the improved BP neural network was better than that of the traditional BP network and SVM.

Keywords: back propagation neural network, financial indicators, financial report fraud, data mining.

Citation: Li SL. Data mining of corporate financial fraud based on neural network model. Computer Optics 2020; 44(4): 665-670. DOI: 10.18287/2412-6179-CO-656.

Introduction

Since the reform and opening up, China's market economy has gradually become active, the growth rate is increasing day by day, and more and more companies are listed on the stock market [1]. For listed companies, regular financial reports [2] can help managers, investors and creditors to make reasonable judgments. However, for a listed company, it is impossible to operate smoothly. The financial reports issued will always have some problems. When a company is trying to gain unfair benefits or cover up the phenomenon of the company's poor operation, it will falsify the financial reports [3]. On the one hand, false financial reports will mislead investors and creditors to make wrong judgments and cause direct economic losses; on the other hand, the market risk prediction made by the government's market supervision department based on false financial reports is wrong, which makes it difficult to avoid the risks brought by corporate violations and affects market equity [4]. Therefore, effective identification of false financial reports is conducive to maintaining market stability and reducing the economic risks of investors and creditors. Fanning et al. [5] designed a false financial report recognizer by using generalized adaptive neural network structure and adaptive logic network method. The simulation results showed that the method could identify false and real financial reports and its accuracy was higher than that of Bell's cascade logic method. Kanapickienė et al. [6] identified whether financial reports were fraudulent or not based on financial ratio analysis, made financial ratio analysis on 40 sets of fraudulent financial reports and 125 sets of real financial reports using logistic regression model, and found that the method could effectively identify the fraudulent financial reports. Lin et al. [7] classified different fraud identification factors using expert questionnaires and data mining technology and ranked their importance. Data mining technology includes logical regression, decision tree and artificial neural network, among which the classification accuracy of artificial neural network is higher than that of logical regression and decision tree. This paper briefly introduces the relevant indicators used for determining the fraudulent financial reports of listed companies and the recognition model of financial reports based on Back Propagation (BP) neural network and gives a simulation analysis on the regular financial reports and irregular financial reports in CSMAR database by using MATLAB software.

Relevant analysis of corporate financial fraud

For a listed company, the annual financial reports are intuitive data reflecting the operating conditions of the company in different periods of a year. The main content of the annual financial reports are balance sheet, cash flow statement and profit statement. In addition to the above measurable data statements, they also include non-data information that can not be directly measured, such as notes to the enterprise statements and honesty guarantee that has an impact on the operation [8]. As the financial statements can reflect a company's operating conditions, generally speaking, the financial statements need to ensure authenticity and provide accurate reference for investors. However, in the actual operation process, some companies will make false statements on financial reports for some economic purposes, including achieving financing conditions, malicious manipulation of holding prices, avoiding delisting penalties, dressing up performance, etc. The objective reasons include unreasonable company rules and regulations and imperfect securities market. Means of financial fraud usually include forging economic transaction vouchers, disguising irregular transactions between related parties, malicious use of debt restructuring, etc. The fraudulent financial reports of listed companies will cause serious damages to the operation of market economy, so the government's regulatory authorities attach great importance to the test of the authenticity of financial reports.

Table 1. Relevant indicators for identifying real and false financial reports

Number	Indicator	Number	Indicator
X_1	Council size	X_{11}	Ratio of accounts receivable to income
X_2	The first shareholder's shareholding ratio	X_{12}	Long-term asset turnover rate
X_3	Asset-liability ratio	X_{13}	Comprehensive lever
X_4	Pre-tax profit-to-debt ratio	X_{14}	Cash flow ratio
X_5	Gross operating profit margin	X_{15}	Operation index
X_6	Net profit margin of total assets	X_{16}	Cash reinvestment rate
X_7	Earnings per share	X_{17}	Free cash flow
X_8	Ratio of retained earnings to asset	X_{18}	Cash ratio
X_9	Gross asset growth rate	X_{19}	Operating capital/total assets
X_{10}	Current assets turnover rate	X_{20}	Operating capital

There are many related indicators that can be used to identify real and false financial reports. Generally speaking, the more indicators used to identify, the higher the recognition accuracy. But the premise of the above theory is that the indicators are independent of each other and the indicators have a clear relationship with the object identified. There are different degrees of correlations between the indicators in the actual financial report, and the degree of the relationship between the indicators and the object identified is different. Too many financial indicators will affect the accuracy of identification [9]. As shown in Table 1, 20 indicators for identifying the authenticity of financial reports are selected after the correlation test of the indicators. X_1~X_2 are the non-financial indicators of corporate governance structure, which have a great impact on financial reports; X_3~X_4 are the long-term solvency of a company; X_5~X_6 are the profitability indicators of a company; X_7~X_8 are the profitability indicators of shareholders; X_9~X_{12} are the development ability indicators of a company; $X_{13}$is the risk level indicator of a company; X_{14}~X_{17} are cash flow indicators of a company; X_{18}~X_{20} are a company's short-term solvency indicator.

Recognition model of financial reporting based on BP neural network

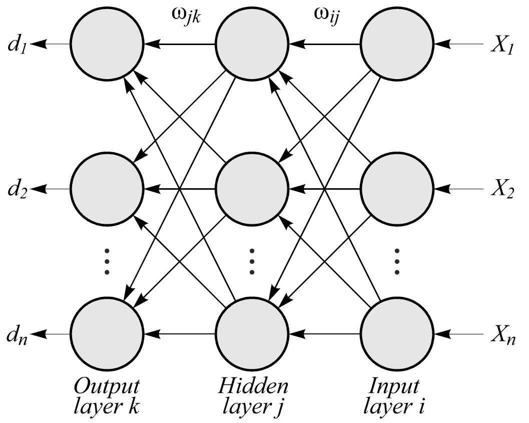

Fig. 1. BP neural network model

As shown in Fig. 1, the basic structure of BP neural network [10] is divided into input layer i, hidden layer j and output layer *k*, where *i*, *j* and *j* are the number of nodes in the corresponding layer. BP Neural Network with three-layer structure is used in this study. X_1~X_n are input vectors, indicators for identifying the authenticity of financial reports in this study; d_1~d_n are output vectors, values for determining the authenticity of financial reports in this study. The algorithm used in BP neural network is error BP algorithm. The training principle of error BP algorithm [11] is as follows. Firstly, input and predetermined output are set. Then actual output is calculated forward layer by layer and compared with the predetermined output. When there are errors, the weight is adjusted according to the opposite direction of the network to make the error between the actual output and predetermined output within the specified range.

The basic training procedures of BP neural network are as follows.

1. The neural network is initialized, including the number of nodes in the input layer, hidden layer and output layer.

2. The training sample was input and calculated using the feed forward formula:

$$\begin{cases} H_j = f(\sum_{i=1}^{n} \omega_{ij} X_i + a_j), \\ d_k = \sum_{j=1}^{l} \omega_{jk} H_j + b_k , \end{cases} \qquad (1)$$

where H_j stands for the output of the j-th node of the hidden layer, d_k is the k-th output of the output layer, ω_{ij} and ω_{ik} are the weight values transmitted from the input layer to the hidden layer and from the hidden layer to the output layer, a_j and b_k are bias terms of the hidden layer and output layer, and $f(\bullet)$ is an activation function. Sigmoid function was used as the activation function in this study.

3. Error calculation. Through the above forward calculation, the calculation result is obtained in the output layer. The result is compared to the preset expected result, and the error is calculated using the following formula [12]:

$$E = -\sum_{k=1}^{n} t_k \log(y_k) , \qquad (2)$$

where E is the error between the output vector and actual output vector obtained by calculation, n stands for the number of nodes in the output layer, y_k is the actual output vector that is set, and t_k is the label of the actual exact solution that is set.

4. Reverse adjustment. The error is determined. If the error is within the specified range, then the result will be output directly; If not, the weight and bias terms of the calculation formula in the hidden layer and the output layer will be reversely adjusted. The weight adjustment formula from the output layer to the hidden layer is:

$$\begin{cases} \omega'_{ij} = \omega_{ij} + \eta H_j (1 - H_j) X_i \sum_{k=1}^{m} \omega_{jk} e_k , \\ \omega'_{jk} = \omega_{jk} + \eta H_j e_k , \end{cases} \qquad (3)$$

where η is the learning rate, ω'_{ij} and ω'_{jk} are the weights after adjustment, and e_k is the error between the k-th output node and expected value. The adjustment formula of the bias term is:

$$\begin{cases} a'_j = a_j + \eta H_j (1 - H_j) \sum_{k=1}^{m} \omega_{jk} e_k , \\ b'_k = b_k + \eta e_k , \end{cases} \qquad (4)$$

where a'_j and b'_k are the bias terms after adjustment.

5. Whether the training stops is determined. The process from the forward calculation to the reverse adjustment of weight and bias terms according to error is considered one time of iteration. The above iteration process is repeated until the error between two adjacent iteration processes was smaller than the set threshold or the times of iterations was the maximum.

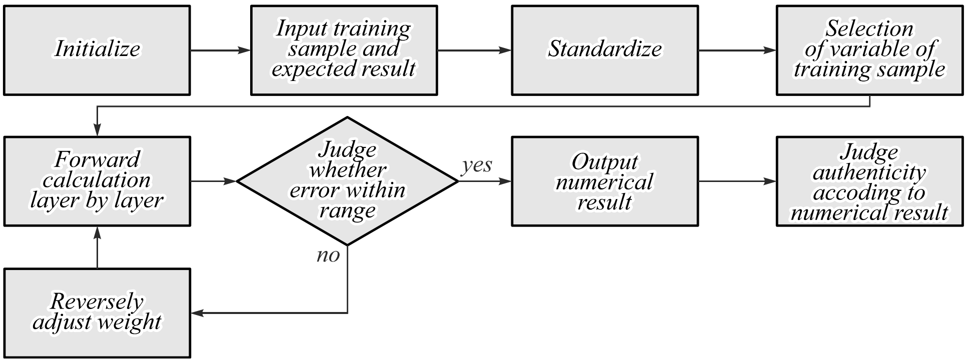
Initialize
Input training sample and expected result
Standardize
Selection of variable of training sample
Forward calculation layer by layer
Judge whether error within range
yes
no
Output numerical result
Judge authenticity accoding to numerical result
Reversely adjust weight

Fig. 2. The calculation process of the improved BP neural network

Although the traditional BP neural network described above can effectively fit the change rule to a certain extent in the training of financial report authenticity identification, there are 20 identification variables to be input in the model training as described above, which are relatively large in quantity. Although BP neural network itself can cope with many input indicators, it will still affect its performance when there are many input indicators to be processed, which is mainly reflected in the calculation efficiency. Secondly, the relationship between the 20 input identification variables and financial report is not the same. The identification variables with shallow connection play a small role in the training process of the identification model or even may interfere with other more effective identification variables, which will make the convergence curve fall into the local optimal solution. Therefore, the traditional BP neural network is improved in this study. Before training with training samples, the importance of identification variables is ranked, and the top 10 most important indicators are selected as the input variables for training. The training process of the improved BP neural network is shown in Fig. 2.

1. The weight in the model is initialized, usually as 0, but other proper values are also allowable, in order to reduce the learning time.

2. The learning samples are input, including fake and non-fake financial reports. The identification variables needed by the model are relevant indexes in financial reports, and the actual judgement is the authenticity of the report.

3. As the dimension of some indexes in the input data is different, they cannot be directly used for calculation. Therefore, it is necessary to standardize the indicators, and the related formula is:

$$X'_{ij} = \frac{X_{ij} - \bar{X}_i}{S_i} , \qquad (5)$$

where X'_{ij} is the value after the conversion of the j-th sample among the i-th class indicator, X_{ij} is the j-th sample among the i-th class indicator, $\overline{X}_i$ is the mean value of all the samples among the i-th class indicator, and Si is the mean square of the i-th class indicator.

The authenticity of the report is expressed as 0 and 1, 0 as the true report and 1 as the false report.

4. The importance of the identification indexes after normalization is calculated [13]:

$$N(i)=\left|\frac{u_1(i)-u_0(i)}{\sigma_1(i)+\sigma_0(i)}\right|, \tag{6}$$

where $N(i)$ is the distance of the i-th indicator between the true and false samples, $u_1(i)$ is the mean value of the i-th indicator in the false report sample, $u_0(i)$ is the mean value of the i-th indicator in the real report sample, $\sigma_1(i)$ is the mean square of the i-th indicator in the false report sample, and $\sigma_0(i)$ is the mean square of the i-th indicator in the real report sample. According to equation (6), the indicator is sorted in descending order, and the first ten indicators are selected as the training input variables.

5. The input training sample is processed by forward calculation according to equation (1). Then the error is calculated according to equation (2), and the weight and bias terms are reversely adjusted according to equation (3).

6. The training model is iterated repeatedly until the error converges to be stable or the times of iterations was the maximum.

Then the authenticity of the report is determined according to the value. A report is determined as real if the value is smaller than 0.5; otherwise it is determined as fake.

Example analysis

1. Experimental environment

BP neural network model algorithm was complied using MATLAB software [14]. The experiment was carried out on a laboratory server. The configuration of the server was Windows 7 system, I7 processor and 16G memory.

2. Experimental data

The illegal financial reports and normal financial reports of listed companies between 2000 and 2010 were selected from CSMAR database [15] as the training samples and testing samples. The training samples included 250 illegal financial reports which were randomly selected from the above financial reports and 250 normal financial reports with the year corresponding to the illegal financial reports. There were totally 500 training samples, and the ratio of the normal ones to the illegal ones was 1:1. Then 500 illegal financial reports and 500 normal financial reports were selected from the remaining financial reports to be used as the testing samples; and the year of two kinds of reports corresponded one by one. The selection criteria for the fraud reports included fictitious profit, fictitious asset, fraud listing and postponing disclosure; reports with any of the above violation were evaluated as the fraud report.

3. Experiment setup

The parameters of the traditional BP neural network were as follows. The number of nodes in the input layer was 20. The number of nodes in the output layer was 1. The number of nodes in the hidden layer was finally determined as 6 after test. The initial weight generated randomly in (-1, 1). The learning rate was set as 0.1.

The parameters of the improved BP neural network were the same with the traditional BP neural network except the number of nodes in the input layer; the number of nodes in the input layer was set as 10.

To better verify the performance of the improved BP neural network in recognizing fake reports, support vector machine (SVM) was used for comparison. In the SVM, the penalty parameter was 10, the kernel function was Gaussian function, and the σ^2 of the kernel function was 2.

4. Criteria for judging the recognition effect of model

The recognition results of the model and the actual results were determined using confusion matrix.

As shown in Table 2, *TP* stands for the number of reports which were real actually and were recognized as real, *FN* stands for the number of reports which were real actually and were recognized as fake, *FP* stands for the number of reports which were fake actually and were recognized as real, and *TN* stands for the number of reports which were fake actually and were recognized as fake.

Table 2. Confusion matrix

	Number of reports recognized as real	Number of reports recognized as fake
Actual number of real reports	*TP*	*FN*
Actual number of fake reports	*FP*	*TN*

Accuracy, precision, recall degree and *F* value were used to measure the recognition effect of the model. Accuracy refers to the proportion of correct classification, and its expression is:

$$P=\frac{TP+TN}{TP+FN+FP+TN}\times 100\% . \tag{7}$$

Precision refers to the proportion of reports which were fake actually among the reports which were recognized as fake, and its expression is:

$$A=\frac{TN}{FN+TN}\times 100\% . \tag{8}$$

Recall rate is the proportion of reports which were recognized as fake among the reports which were fake actually, and its expression is:

$$R=\frac{TN}{FP+TN}\times 100\% . \tag{9}$$

F value can comprehensively reflect the recognition effect of the model, and its expression is:

$$F = \frac{2 \times A \times R}{A + R}. \quad (10)$$

5. Experimental results

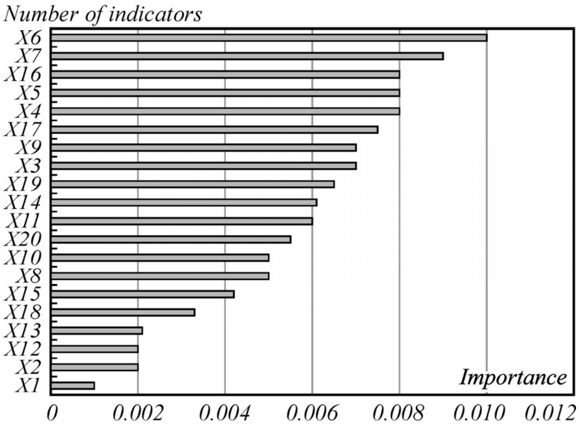

Fig. 3. Importance of indicators for judging whether a report is true or false

The importance of indicators for judging whether a report is true or false after sorting by BP neural network is shown in Fig. 3. It can be seen that the importance of net profit margin of total assets, earnings per share, cash reinvestment rate, gross operating profit and pre-tax profit-to-debt ratio were the top 5 among the 20 indicators, while the importance of council size, the first shareholder's shareholding ratio and long-term asset turnover ratio was relatively small.

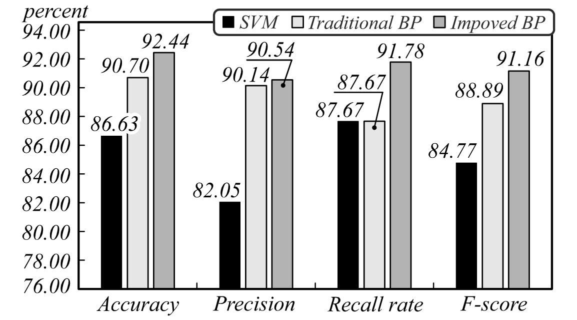

Fig. 4. Comparison of recognition effect of the model under different numbers of features

As shown in Fig. 4, the recognition accuracy of SVM was 86.63 %, the precision was 82.05 %, the recall rate was 87.67 %, and the *F* value was 84.77 %; the recognition accuracy of the traditional BP neural network was 90.70 %, the precision was 90.14 %, the recall rate was 87.67 %, and the *F* value was 88.89 %; the accuracy of the improved BP neural network was 92.44 %, the precision was 90.54 %, the recall rate was 91.78 %, and the *F* value was 91.16 %. It was seen from Fig. 4 that the improved BP neural network had the highest accuracy in judging whether a financial report was fake or not, and the SVM was the lowest; in terms of accuracy, the improved BP neural network was slightly higher than the traditional BP neural network, and both were significantly higher than the SVM; in terms of recall rate, the SVM and the traditional BP neural network were relatively close, while the improved BP neural network was significantly higher than the other two identification models; in terms of F value, the improved BP neural network was the highest and SVM was the lowest. It was concluded that the improved BP neural network had the best recognition performance and SVM had the worst recognition performance in judging whether a financial report was fake or not.

Conclusion

This paper briefly introduced the relevant indicators for identifying false financial reports of listed companies and the recognition model of financial reports based on BP neural network. Some improvements were made in the selection of relevant input indexes, and then the improved BP neural network was simulated and analyzed in MATLAB software and compared with the traditional BP neural network and SVM. The results are as follows. The importance of total assets net profit, earnings per share, cash reinvestment rate, operating gross profit and pre-tax profit debt ratio ranked top 5 among 20 judgment indexes. In the judgment of whether a financial report was fake or not, the accuracy of the improved BP network was the highest, and the accuracy of SVM was the lowest; the accuracy of the improved BP network was close to that of the traditional BP network and significantly higher than that of SVM; the recall rates of the SVM and traditional BP network were close, and the recall rate of the improved BP network was significantly higher than the other two networks; the *F* value of the improved BP network was the highest, and the *F* value of SVM was the lowest.

References

[1] Fanning K, Cogger KO. Neural network detection of management fraud using published financial data. Intell Syst Account Finance Manag 2015; 7(1): 21-41.

[2] Qi J, Yi L. Network financial fraud risk assessment system based on big data analysis. J Comput Theor Nanosci 2016; 13(12): 9335-9339.

[3] Zanin M, Romance M, Moral S, et al. Credit card fraud detection through parenclitic network analysis. Complexity 2018; 2018: 1-9.

[4] Bahnsen AC, Aouada D, Ottersten B. Example-dependent cost-sensitive decision trees. Expert Syst Appl 2015; 42(19): 6609-6619.

[5] Fanning K, Cogger KO, Srivastava R. Detection of management fraud: A neural network approach. Intell Syst Account Finance Manag 1995; 4(2): 113-126.

[6] Kanapickienė R, Grundienė Ž. The model of fraud detection in financial statements by means of financial ratios. Procedia Soc Behav Sci 2015; 213: 321-327.

[7] Lin C, Chiu A, Huang SY, et al. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. Knowl Based Syst 2015; 89(9): 459-470.

[8] Coakley JR, Brown CE. Artificial neural networks applied to ratio analysis in the analytical review process. Intell Syst Account Finance Manag 2015; 2(1): 19-39.

[9] Compin F. Tax fraud: A socially acceptable financial crime in France. J Financ Crime 2015; 22(4): 432-446.

[10] Chen T, Xu W. Post-evaluation on financial support highway traffic project based on BP neural network algorithm. J Discret Math Sci Cryptogr 2018; 21(4): 869-879.

[11] Hong Y, Sun W, Bai QL, Mu XW. SOM-BP neural network-based financial early-warning for listed companies. J Comput Theor Nanosci 2016; 13(10): 6860-6866.

[12] Wang L, Liu H, Feng C, et al. Identification of flow regimes based on adaptive learning and additional momentum BP neural network. 6th IMCCC 2016: 574-578.

[13] Liu SJ, Li SL, Jiang M, et al. Quantitative identification of pipeline crack based on BP neural network. Key Eng Mater 2017; 737: 477-480.

[14] He G, Huang C, Guo L, et al. Identification and adjustment of guide rail geometric errors based on BP neural network. Meas Sci Rev 2017; 17(3): 135-144.

[15] Yacoub HA, Sadek MA. Identification of fraud (with pig stuffs) in chicken-processed meat through information of mitochondrial cytochrome b. Mitochondrial DNA A DNA Mapp Seq Anal 2016; 28(6): 1.

Author's information

Shenglu Li (b. 1981) has gained the master's degree from Changchun University of Technology in 2006. She is now an associate professor in Changchun Guanghua University. She is interested in financial management informationization and financial intelligence. E-mail: *shenglulsl@yeah.net* .

Received October 25, 2019. The final version – December 26, 2019.