Financial Risk Analysis and Early Warning Research Based on Data Mining and Statistical Analysis Technology

Chang Li*, Zuxin Meng, Laicai Chang, Dayu Pei
Guilin University of Electronic Technology, Guilin, China
*Corresponding author’s e-mail: changli@mail.guet.edu.cn

Abstract. Using the advantage of decision tree algorithm in the screening work, the traditional ID3 algorithm is improved and optimized, and a new and simplified financial index system is constructed. At the same time, combined with the unique value of artificial neural network in early warning model and data analysis, B-P model is used to build a mixed financial early warning model. In the model study, the HFPM model and Z-score model were compared and analyzed by using test samples and training samples, and the superior warning ability of the former was effectively verified.

1. Data mining and statistical analysis
In essence, data mining is a process of finding rules in the database. This idea originated from statistics at the earliest, and has been effectively integrated with various theoretical knowledge such as artificial intelligence, database and statistics so far. It is very effective to apply it to enterprise financial risk analysis. Taking artificial neural network as an example, as shown in Figure 1 below, it is the most widely used multi-layer perceptron neural network in practice, called error back propagation method, also known as BP algorithm, and it is one of the most critical models in the current application of artificial neural network[1].

![BP algorithm model](image)

In terms of enterprise financial risk and its early warning analysis, when building a model based on data mining and statistics, the most important thing is to define a scientific and reasonable
financial index system first, and then build a model according to specific early warning needs. In this paper, starting from the domestic A stock market included listed companies, using the decision tree in data mining to correctly screen out A reasonable financial index system, and combined with the BP model in the neural network for early warning analysis.

2. Establish HFPM streamlined financial indicator system

For financial risk analysis and early warning based on data mining and statistical analysis, it is necessary to use decision tree and artificial neural network to build a multivariable nonlinear financial early warning model, also known as mixed financial early warning model (HFPM). Based on the analysis of HFPM comprehensive financial indicators and relevant calculation formulas shown in Table 1 below, it can be seen that the construction steps of a concrete streamlined financial indicator system are mainly divided into the following points:

| Table 1. HFPM comprehensive financial indicators and related calculation formulas |
|---------------------------------|---------------------------------|--------------------------|
| Indicator types                  | Indicator name                  | The calculation formula   |
| An index of corporate solvency   | Debt to asset ratio             | Total liabilities/total assets ×100% |
|                                 | Liquidity ratio                 | Total current assets/total current liabilities |
|                                 | Quick ratio                     | (current assets - inventory)/total current liabilities |
|                                 | Debt to equity ratio            | Total liabilities/shareholders' equity |
|                                 | Shareholder equity ratio        | Shareholders' equity/total assets |
|                                 | Cash per share from operations  | Net cash flow from operations/total number of shares |
|                                 | Net profit cash amount          | Net cash flow/total profit |
|                                 | Cash flow liability ratio       | Net cash flow from operations/current liabilities ×100% |
|                                 | Earnings per share              | Net profit/total shares |
|                                 | Earnings per share (net)        | (Net profit - non-recurring income)/Total number of shares |
|                                 | Return on equity                | Net profit/ending net assets ×100% |
|                                 | Net profit margins              | Net profit/operating income ×100% |
| An index of corporate profitability | Operating net margin           | Recurrent net income/operating income ×100% |
|                                 | Operating gross margin          | Total current income/operating income ×100% |
| An index of a firm's ability to  | Accounts receivable turnover    | Average net sales/accounts receivable ×100% |
### Inventory turnover
Cost of Sales/Average Inventory ×100%

### Main business revenue growth rate
Increase of current year's main business income/last year's main business income ×100%

### Net profit growth rate
Increase in net profit/total net profit of last year ×100%

### Growth rate of shareholders' equity
Increase of current year's main business profit/last year's main business profit ×100%

### Year/Shareholders' Equity at the Beginning of the Year
Growth of Shareholders' Equity in Current Year/Shareholders' Equity at the Beginning of the Year ×100%

### Total asset growth rate
Growth of total assets of this year/total assets of last year ×100%

### Total net assets/total shares
(Year-end Shareholders' Equity - Accounts Receivable over 3 years - Pending Expenses - Net Loss on Pending (Current, Fixed) Assets - Organizing Expenses - Long Term Pending Expenses)/Total Capital at Year-end

### Adjusted net assets per share
Total undistributed profits/total shares

### Undistributed earnings per share
Capital reserve/total number of shares

### Net assets per share
Capital reserve/total number of shares

### Capital reserve per share

| 2.1. The ID3 algorithm |
|-------------------------|
| As a common classification algorithm in the decision tree, ID3 is assumed to have N independent results when considering A random event experiment A, which can be expressed as A1, A2... An, the probabilities of all possible outcomes are P1, P2... Pn, then they fit this formula, . At the same time, in order to calculate the uncertainty of probability experiment A, researcher Shannon referred to the function \[ H_n(A) = H(P_1, P_2, ..., P_n) = -k \sum_{i=1}^{n} P_i \ln P_i \] \, where K in the formula refers to a constant over 0, so \( H_n \geq 0 \). At this time, \( H_n \) is also called Shannon's entropy. This algorithm is used to achieve index screening. The specific pseudocode is shown as follows: |

```java
public class decision.tree {
    val [2][3]: the.data.to.build.the.decision.tree;
    tag [4]: used to record whether or not each column has been processed;
    rows: val the.number.of.rows; cols: the.number.of.columns.of;
    dataFile: data.file.name; avg[6]: the.average.value.of.each.column;
    log2=Math.log(2.0);
```
2.2. Model construction

In the study of enterprise financial early warning model, you can use the NeuroSolution 4 system of IBM to implement auxiliary operation, which is divided into the following contents:

First, the collected and sorted sample information was imported into the NeuroSolution 4 system, and the input and output columns, as well as the training samples and test samples were clearly marked. The former is for the financial indicator system. If the sample enterprise's finance is normal, then the attribute of the mark can be 1, whereas the attribute of financial failure is 1. The training samples are for the sample companies, mainly to provide an effective basis for the model construction. The test sample is used to verify the warning ability of the model composed of training samples. The information contained in the test sample is real-time and fair.[7-10]

Second, make clear the structure of artificial neural network. In practice, the concrete model with artificial neural network as the core should be defined first. Because the HFPM model is built using the B-P model, the multi-layer perceptron model provided in the software can be utilized. Then, the number of hidden layers should be defined, and it should be divided into two layers according to practical experience, and each layer should be guaranteed to contain 4 nodes. Finally, the conversion function should be defined according to the formula of

\[ f(x) = \frac{1}{1 + e^{-x}} \]

Where, Q in the formula represents the Sigmoid parameter in the form of optimized excitation function. In this paper, hyperbolic tangent S-type transfer function was selected when constructing the research model, which corresponds to "Tanhaxon" in Neurosolution 4 system. In addition, learning rate and weight correction constant should be defined. Generally speaking, the value range of the former should be controlled between 0.1 and 0.4, while the value range of the latter should be controlled between 0.7 and 0.9.

Thirdly, after the model structure and parameters are defined, the most important thing is to use training samples to design the weight matrix between layers and nodes. Because the NeuroSolution 4 system was selected in this study, the number of iterations needed to be combined for analysis, that is, 1000 times were taken from the software system, and then directly click "Trannetwork" on the menu bar.

3. Effect analysis

Since the 1940s, domestic and foreign researchers have studied and analyzed the early warning model for enterprise finance, among which the Z-score model has achieved very beneficial results in practical application, especially from the perspective of the early warning ability one year before the occurrence of enterprise financial crisis, the application effect of this model is very obvious. In this paper, when studying the financial risk early warning model, the financial statements of listed companies and their financial conditions in the fiscal year are selected, and the Z-score model is combined for a comparative study to better clarify the application of HFPM-ID3 algorithm process.

Generally speaking, when evaluating financial risk analysis and early warning models, we need to start from two aspects: on the one hand, we only need to consider the error rate of the actual forecast, and we don't need to study the cost and other influencing factors; On the other hand, it is necessary to consider the error cost from a more comprehensive perspective and compare and study the application effect of the two models.

3.1. Error rate
Before comparing the prediction effect of the actual model, the HFPM model proposed in this paper should be used for testing and analysis. By clicking "Trannetwork" on the menu bar and combining the stored data samples to test, the two options you can choose are "Testing" and "Training". In other words, you can test not only with a set sample, but also with a test sample. Although the model is designed on the basis of training samples, it is usually difficult to reach 100% in the actual test. Meanwhile, this test is still effective from the perspective of practice. Combined with the prediction results of previous models, it can be seen that the prediction accuracy of training samples is generally higher than that of test samples. As shown in Table 2 below, test sample information is selected for this study.

### Table 2. Data results of test samples

| Company number | Earnings per share | Net assets per share | Provident fund per share | Quick ratio | Asset - liability ratio | Accounts receivable turnover | Inventory turnover | Total asset growth rate | Net profit | Financially sound | Financial failure |
|----------------|---------------------|----------------------|--------------------------|-------------|------------------------|----------------------------|-------------------|------------------------|------------|------------------|------------------|
| 98             | -0.04               | 0.1                  | 0.57                     | 0.48        | 92.53                  | 8.99                       | 1.46              | -9.17                  | -6.47      | 0                | 1                |
| 99             | -1.41               | 0.2                  | 1.21                     | 0.57        | 85.82                  | 0.68                       | 2.34              | 43.4                   | 1.06       | 0                | 1                |
| 100            | -0.68               | 0.5                  | 2.37                     | 0.24        | 84.36                  | 0.86                       | 0.02              | -4.85                  | -0.11      | 0                | 1                |
| 101            | -0.06               | 0.1                  | 0.43                     | 0.11        | 31.84                  | 0.29                       | 0.78              | 22.4                   | 0.29       | 0                | 1                |
| 102            | -0.24               | 0.9                  | 0.54                     | 0.56        | 71.46                  | 4                          | 0.08              | 25.9                   | -4.92      | 0                | 1                |
| 103            | -1.79               | 1.3                  | 0.46                     | 0.25        | 1272.3                 | 0.58                       | 5.8               | 95.5                   | 0.29       | 0                | 1                |
| 104            | -0.13               | 0.9                  | 0.07                     | 0.68        | 76.42                  | 0.38                       | 0.9               | 10.9                   | 0.61       | 0                | 1                |
| 105            | -1.29               | 0.1                  | 0.65                     | 0.82        | 92.79                  | 0.35                       | 7.71              | -52.6                  | 0.47       | 0                | 1                |
| 106            | -28.3               | 3.0                  | 0.59                     | 0.22        | 358.6                  | 1.25                       | 0.28              | 71.6                   | 0.01       | 0                | 1                |
| 107            | -0.48               | 0.4                  | 2.18                     | 0.13        | 106.8                  | 0                          | 3.78              | 54.5                   | -0.42      | 0                | 1                |
| 108            | -0.88               | 0.9                  | 0.71                     | 0.24        | 203.5                  | 1.65                       | 2.53              | -62.6                  | -0.02      | 0                | 1                |
When comparing the actual warning effect, there will be two kinds of error rates, which can be called the first type and the second type from the perspective of statistics. The former model will predict the enterprises that may have problems in the future to be in good financial condition, while the latter model will predict the enterprises with good financial condition to be in trouble. In the actual comparative study, the test samples are divided into two kinds, and any kind of samples can be used for the comparative analysis of the two error rates. As shown in Table 3 below, it is the test results of training samples of HFPM model and Z-scoring model, while Table 4 is the test sample results of both. Through comparative analysis, it is found that all error rates of HFPM model are lower than those of Z-score model, especially in the second type of error rates.

|       | Financially sound | Financial failure | Financially sound | Financial failure |
|-------|-------------------|-------------------|-------------------|-------------------|
| HFPM  | 199 0.5           | 0.8               | 0.1               | 1.3               |
|       | 27.1              | 0.44              | 1.4               | 0.71              |
| Z-score| 200 0.07          | 2.0               | 1.4               | 1.4               |
|       | 24.6              | 1.95              | 1.4               | 0.7               |
|       | 201 0.8           | 3.8               | 6.5               | 2.7               |
|       | 152.57            | 11.43             | 1.4               | 0.7               |
|       | 202 0.01          | 1.9               | 9.8               | 2.5               |
|       | 18.3              | -                 | 20.13             | 1.4               |
|       | 203 0.2           | 2.0               | 3.4               | 7.7               |
|       | 18.5              | -                 | 0.21              | 1.4               |
|       | 204 0.48          | 7.7               | 5.92              | 13.27             |
|       | 5.61              | 11.93             | 18.41             | 4.27              |
|       | 205 0.3           | 3.3               | 1.7               | 57.66             |
|       | 53.11             | 3.74              | 9.73              | 4.63              |
|       | 206 0.26          | 2.1               | 0.65              | 45.07             |
|       | 0.76              | 7.31              | 16.3              | 0.95              |
|       | 207 0.3           | 2.8               | 0.61              | 46.44             |
|       | 0.57              | 143.26            | 23.5              | 0.81              |
|       | 208 0.5           | 3.0               | 0.51              | 56.43             |
|       | 0.74              | 253.9             | 5.83              | 7.28              |
|       |                   |                   | 1.23              |                   |

Table 3. Comparison of results of training samples

|       | Financially sound | Financial failure | Financially sound | Financial failure |
|-------|-------------------|-------------------|-------------------|-------------------|
| HFPM  | 199 0.5           | 0.8               | 0.1               | 1.3               |
|       | 27.1              | 0.44              | 1.4               | 0.71              |
| Z-score| 200 0.07          | 2.0               | 1.4               | 1.4               |
|       | 24.6              | 1.95              | 1.4               | 0.7               |
|       | 201 0.8           | 3.8               | 6.5               | 2.7               |
|       | 152.57            | 11.43             | 1.4               | 0.7               |
|       | 202 0.01          | 1.9               | 9.8               | 2.5               |
|       | 18.3              | -                 | 20.13             | 1.4               |
|       | 203 0.2           | 2.0               | 3.4               | 7.7               |
|       | 18.5              | -                 | 0.21              | 1.4               |
|       | 204 0.48          | 7.7               | 5.92              | 13.27             |
|       | 5.61              | 11.93             | 18.41             | 4.27              |
|       | 205 0.3           | 3.3               | 1.7               | 57.66             |
|       | 53.11             | 3.74              | 9.73              | 4.63              |
|       | 206 0.26          | 2.1               | 0.65              | 45.07             |
|       | 0.76              | 7.31              | 16.3              | 0.95              |
|       | 207 0.3           | 2.8               | 0.61              | 46.44             |
|       | 0.57              | 143.26            | 23.5              | 0.81              |
|       | 208 0.5           | 3.0               | 0.51              | 56.43             |
|       | 0.74              | 253.9             | 5.83              | 7.28              |
|       |                   |                   | 1.23              |                   |

Table 4. Comparison of results of test samples

|       | Financially sound | Financial failure | Financially sound | Financial failure |
|-------|-------------------|-------------------|-------------------|-------------------|
| HFPM  | 199 0.5           | 0.8               | 0.1               | 1.3               |
|       | 27.1              | 0.44              | 1.4               | 0.71              |
| Z-score| 200 0.07          | 2.0               | 1.4               | 1.4               |
|       | 24.6              | 1.95              | 1.4               | 0.7               |
|       | 201 0.8           | 3.8               | 6.5               | 2.7               |
|       | 152.57            | 11.43             | 1.4               | 0.7               |
|       | 202 0.01          | 1.9               | 9.8               | 2.5               |
|       | 18.3              | -                 | 20.13             | 1.4               |
|       | 203 0.2           | 2.0               | 3.4               | 7.7               |
|       | 18.5              | -                 | 0.21              | 1.4               |
|       | 204 0.48          | 7.7               | 5.92              | 13.27             |
|       | 5.61              | 11.93             | 18.41             | 4.27              |
|       | 205 0.3           | 3.3               | 1.7               | 57.66             |
|       | 53.11             | 3.74              | 9.73              | 4.63              |
|       | 206 0.26          | 2.1               | 0.65              | 45.07             |
|       | 0.76              | 7.31              | 16.3              | 0.95              |
|       | 207 0.3           | 2.8               | 0.61              | 46.44             |
|       | 0.57              | 143.26            | 23.5              | 0.81              |
|       | 208 0.5           | 3.0               | 0.51              | 56.43             |
|       | 0.74              | 253.9             | 5.83              | 7.28              |
|       |                   |                   | 1.23              |                   |
3.2. Comprehensive error cost
In order to investigate the effectiveness of the model outlined in this article, it is necessary to make a clear distinction between the cost of producing the two types of errors. Since auditors, investors and other participants all believe that the risk of such error costs has different influences on the final financial decisions, this paper uses the following relative cost ratio to verify in combination with Hopwood’s idea and the following model:
Where, the relative cost ratio is the error cost of the first type/the error cost of the second type, and the total relative cost = (PI × CI) + (PII × CII). Pi represents the error rate of the first type, CI represents the relative error cost of the first type, Pii represents the error rate of the second type, and CII represents the relative error cost of the second type.
In essence, the model with too low overall relative cost is more acceptable than the model with too high cost, as shown in Table 5 below, which is the cost comparison between the HFPM model and the Z scoring model, where A represents unit cost:

| Table 5. Comparison results of the overall relative cost of the model |
|-------------------------|----------------|----------------|
| Relative cost rate      | Hfpm model    | Z score model  |
| 1:1                     | 0.03a          | 0.34a          |
| 10:1                    | 0.03a          | 0.88a          |
| 20:1                    | 0.03a          | 1.48a          |
| 30:1                    | 0.03a          | 2.08a          |
| 40:1                    | 0.03a          | 2.68a          |
| 50:1                    | 0.03a          | 3.28a          |
| 1:1                     | 0.08a          | 0.25a          |
| 10:1                    | 0.36a          | 0.61a          |
| 20:1                    | 0.61a          | 1.01a          |
| 30:1                    | 0.89a          | 1.41a          |
| 40:1                    | 1.17a          | 1.81a          |
| 50:1                    | 1.45a          | 2.21a          |

Combined with the analysis in the above table, it can be seen that under the background that the ratio of PI/PII continues to rise, the overall relative cost difference of the two models analyzed by using test samples is getting lower and lower. Under the condition of low difference, the relative cost of Z-scoring model will exceed that of HFPM model, up to 50%.

4. Results
With the rapid development of information technology, management theory has made new achievements in practical exploration. Although in this context, the effective combination of information technology and management work has put forward more high-quality enterprise decisions, but in the trend of market development and increasingly fierce competition, the financial risks faced by enterprises will also increase. According to the experimental research and analysis results of this paper, using data mining and statistical analysis technology to analyze the financial risk of enterprises and build the corresponding early warning model is now a powerful measure to solve the financial risk of enterprises. It should be noted that when mining data to clarify research objectives, relevant contents must be collected from the internal and external
perspectives of the enterprise, and data preprocessing must be done well, so as to improve the efficiency of data application and simplify the effect of model construction. In addition, before constructing the risk warning model, sample data should be studied and collected. By clarifying the variable factors and eliminating the unnecessary influencing factors, it is helpful to reduce the subsequent modeling time and improve the efficiency and accuracy of the model application. In model research and data mining, the staff should also consider the following issues: first, whether the operational model is more effective than the model on the data set; Second, whether the accuracy of the model is higher than that of other models; Third, whether the model is built on the basis of the sample model; Fourth, whether the model operation process is lower than the effect of the part model. Because most of the current enterprise financial early warning model is based on data mining to analyze construction as the core, so researchers have to master the theoretical knowledge and application of data mining algorithm on the basis of, skillfully use the relevant operation technology, by doing a lot of uncertain and incomplete real data found in samples of valuable information. At the same time, it is necessary to pay attention to the optimization analysis of relevant models based on practical experience. For example, the HFPM-ID3 algorithm process constructed in this paper is obviously higher than the traditional Z-scoring model according to the final comparison results. This not only can grasp the enterprise financial risk as soon as possible, and put forward effective early warning, but also can solve the problems existing in the enterprise financial management on the basis of.

5. Conclusion
Above all, by integrating the enterprise financial indicators, and according to the actual demand for screening of important financial risk analysis and early warning financial indicators of ID3 algorithm to implement improvements, final design new HFPM - ID3 algorithm flow, and then use this content to the enterprise comprehensive financial indicators, screening and build HFPM streamline financial index system. At the same time, the B-P model in the artificial neural network is used to design the financial early warning model. On the basis of breaking away from the limitations of the traditional research model, the analysis is made from the attribute value and independence presented by the samples collected by enterprises, and finally the sample value that is more in line with the actual situation is obtained. Combined with the analysis of the final experimental comparison results, it can be seen that the HFPM model is more suitable for the financial decision-making needs of enterprises than the traditional Z-scoring model, and it shows a strong early warning ability in both the training samples and the test samples.

References
[1] Zhao, J., Yu, J., Wang, X. (2020) Research on Systematic Risk Early Warning System Based on Machine Learning Technology: A Case Study of Marine Economy[J]. Journal of Coastal Research,108(sp1).
[2] Shi Y., Badi, M., Rajan, D., et al. (2021) Channel Reciprocity Analysis and Feedback Mechanism Design for Mobile Beamforming Systems[J]. IEEE Transactions on Vehicular Technology, PP(99):1.
[3] Li, Q. W., Jiang, P., Li, H. (2020) Prognostics and health management of FAST cable-net structure based on digital twin technology[J]. Research in Astronomy and Astrophysics, v.20(05):49-56.
[4] Bo, M. A., Yh, B., Sc, B., et al. (2020) Risk early warning and control of food safety based on an improved analytic hierarchy process integrating quality control analysis method[J]. Food Control, 108:106824-.

[5] Wang, L., Kang, J., Xu, J., et al. (2019) Early warning of cyanobacteria blooms outbreak based on stoichiometric analysis and catastrophe theory model[J]. Journal of Mathematical Chemistry, 58(7).

[6] Tian, D., Xiong, X., Xiao, C. (2020) Early Warning and Inhibition of HVDC Subsequent Commutation Failure during Recovery Process under Grid Fault[J]. IEEE Transactions on Power Delivery, PP(99):1-1.

[7] Sharma, E., Deo, R. C., Prasad, R., et al. (2020) A hybrid air quality early-warning framework: An hourly forecasting model with online sequential extreme learning machines and empirical mode decomposition algorithms[J]. The Science of the Total Environment, 709(Mar.20):135934.1-135934.23.

[8] Zga, B., Fla, B., Dsa, B., et al. (2020) Early warning and control of food safety risk using an improved AHC-RBF neural network integrating AHP-EW - ScienceDirect[J]. Journal of Food Engineering, 292.

[9] Shi, T., Min, M., Sun, C., et al. (2020) Does insomnia predict a high risk of cancer? A systematic review and meta-analysis of cohort studies[J]. Journal of Sleep Research, 29(1).

[10] Levitin, G., Xing, L., Huang, H. Z. (2019) Security of Separated Data in Cloud Systems with Competing Attack Detection and Data Theft Processes[J]. Risk analysis, 39(4):846-858.