Corporate bankruptcy prediction based on the Adaboost algorithm for optimisation of long and short-term memory networks

Siyu Li

School of business, Hunan Normal University, Changsha, Hunan, 410006, China

17752675672@163.com

Abstract. In this study, a Long Short-Term Memory (LSTM) network was used to model and predict time series data, providing an innovative approach to corporate bankruptcy prediction. Subsequently, the predictions and real labels of the LSTM models were used to train the Adaboost algorithm to improve the accuracy and robustness of the models. Eventually, multiple trained LSTM models are combined into a more robust integrated model by adjusting the weights. It is worth mentioning that the integrated model achieves 95.57% prediction accuracy on the training set and 94.39% prediction accuracy on the test set, which indicates that the model has good prediction effect and generalisation ability. The method proposed in this study is of great significance, firstly, by combining LSTM and Adaboost algorithm, we not only improve the accurate prediction ability of corporate bankruptcy, but also enhance the ability of identifying anomalies. Second, by combining multiple LSTM models and adjusting the weights to form a more powerful integrated model, we effectively improve the overall prediction performance. This approach can provide financial institutions, investors, and government regulators with a more reliable and accurate tool for assessing corporate bankruptcy risk, which can help identify potential risks and take appropriate measures in a timely manner.

Keywords: Long and short-term memory networks, Adaboost, Classification prediction.

1. Introduction

Corporate bankruptcy prediction is an important topic in the field of financial risk management, which is of great significance to investors, suppliers, banks and other stakeholders [1]. The background of research on corporate bankruptcy prediction can be traced back to the 1960s, when scholars began to explore how to use financial indicators and statistical methods to predict corporate bankruptcy risk. With the changes in the global economic environment and the increase in uncertainty in the financial market, enterprises are facing various internal and external challenges, and bankruptcy prediction has become one of the focuses of corporate management and regulatory authorities.

Over the past decades, machine learning algorithms have played an increasingly important role in corporate bankruptcy prediction. Compared with traditional statistical methods, machine learning algorithms can better handle large-scale data, discover hidden patterns, and have stronger predictive capabilities [2]. Common machine learning algorithms include decision tree [3], support vector machine [4], logistic regression [5], random forest [6], etc. These algorithms can analyse and learn from the historical data, so as to build a prediction model for the risk of corporate bankruptcy. In corporate bankruptcy prediction, machine learning algorithms can help identify companies that are potentially facing financial difficulties and issue warning signals in advance so that relevant stakeholders can take appropriate measures to reduce losses. By analysing a large amount of financial data, market data and macroeconomic data, machine learning algorithms can identify features and patterns that are closely related to corporate bankruptcy and provide timely and effective decision support for decision makers.

In summary, corporate bankruptcy prediction, as one of the important topics in the field of financial risk management, has made significant progress with the help of machine learning algorithms. In the future, with the continuous improvement of data collection technology and algorithm optimisation, it is believed that machine learning will play an increasingly important role in the field of corporate bankruptcy prediction and provide more reliable and efficient risk management tools for all parties. At present, the long short-term memory network (LSTM) shows great potential for application, in order to take advantage of the potential of LSTM, this paper uses LSTM to optimise the Adaboost model for corporate bankruptcy prediction, which provides a new way of thinking for corporate bankruptcy prediction.

2. Data sources

Data were collected from the Taiwan Economic Journal over a ten-year period. The definition of corporate bankruptcy is based on the commercial regulations of the Taiwan Stock Exchange. The data contains a total of 2,550 entries, each of which lists various business indicators of the firm, such as ROA(C), pre-tax interest and depreciation ROA(A), pre-tax interest and depreciation ROA(B), after-tax interest and depreciation, operating gross margin, realised sales gross margin, operating profit margin, pre-tax net interest margin, after-tax net interest margin, non-industrial income/expense/revenue, sequential interest rate (after-tax), and operating expense ratio, cash flow rate, interest-bearing debt interest rate, and tax rate, etc., and the predictor is corporate bankruptcy (1 indicates normal operations and 2 indicates bankruptcy), we display some of the data as shown in Table 1.

Table 1. Part of the dataset.

Gross Profit to Sales	Liability to Equity	Degree of Financial Leverage (DFL)	Net Income Flag	Equity to Liability	Bankrupt
0.60145329	0.290201893	0.026600631	1	0.016468741	1
0.610236526	0.28384598	0.26457682	1	0.020794306	1
0.60144934	0.290188533	0.02655472	1	0.016474114	1
0.583537612	0.281721193	0.026696634	1	0.023982332	1
0.59878151	0.27851379	0.024751848	1	0.035490201	1
0.590172327	0.2850871	0.026675366	1	0.019534478	1
0.619948867	0.292504124	0.026622298	1	0.015663075	2
0.60173934	0.278607306	0.027030517	1	0.034888556	2
0.603613451	0.276422514	0.02689063	1	0.065826497	2
0.599205074	0.279387519	0.027243015	1	0.030800865	2
0.614021193	0.278356432	0.026971091	1	0.036571691	2
0.623709203	0.277892082	0.027390858	1	0.04038102	2

3. Pearson correlation analysis

Pearson's correlation analysis is a statistical method used to measure the degree of linear correlation between two variables and is commonly used to understand the relationship between two variables as well as to predict trends between them. By calculating the covariance and standard deviation of the two variables, a Pearson's correlation coefficient ranging from -1 to 1 is finally obtained. Correlation analysis of some of the variables with corporate insolvency is carried out and correlation heat map is plotted and the results are shown in Figure 1.

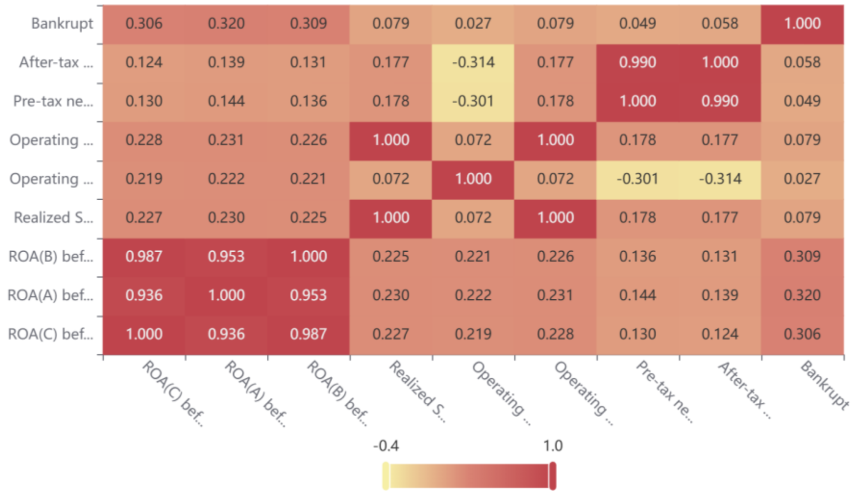

Figure 1. Correlation heat map.

From the correlation heat map, it can be seen that there is a relatively strong correlation relationship between ROA(C), pre-tax interest and depreciation ROA(A), pre-tax interest and depreciation ROA(B), after-tax interest and depreciation, operating gross margin and realised gross sales margin and corporate bankruptcy, and it is possible to use the machine learning method to predict corporate bankruptcy.

4. Pearson correlation analysis

4.1. Long Short-Term Memory

Long Short-Term Memory (LSTM) network is a deep learning model commonly used to process sequential data, LSTM network solves the long-term dependency problem in traditional Recurrent Neural Networks (RNN) by introducing a gating mechanism.

The core principle of LSTM is that its internal structure contains three key gating units: forgetting gate, input gate and output gate. These gating units help the LSTM network to learn long-term dependencies and efficiently capture important information in sequential data [7].The network structure of LSTM is shown in Fig. 2.

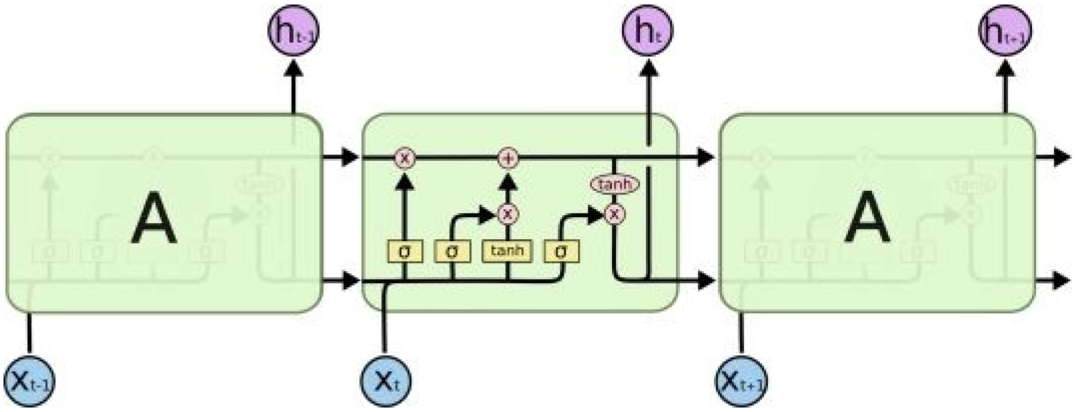

Figure 2. The network structure of LSTM.

Firstly, there is the forgetting gate, which determines which information in the past memory needs to be retained and which information needs to be forgotten. A value between 0 and 1 is output through a sigmoid activation function indicating how much information needs to be retained in the corresponding position in each memory cell.

Next comes the input gate, which is responsible for updating the contents of the memory cells. It first determines which information needs to be updated through the sigmoid activation function, and then generates a new candidate value to be used to update the memory cell through the tanh activation function.

Finally, there is the output gate, which calculates the final output based on the current input and the previously saved memory states. The sigmoid function is first used to determine which information will be output to the next layer or as the final prediction, and then the tanh function is used to map the content of the memory cells between -1 and 1 as the output [8].

4.2. Adaboost

Adaboost is an integrated learning method that aims to build a strong classifier by combining multiple weak classifiers. The principle is based on continuously adjusting the weights of the training samples so that the samples that were misclassified by the previous round of classifiers receive more attention in the next round, thus improving the accuracy of the overall model.

First, Adaboost assigns an initial weight to each training sample, which is usually equal. Then, in each round of training, Adaboost selects a weak classifier that is currently optimal and adjusts the weights of the samples based on its classification results. Samples that are misclassified will receive higher weights, while those that are correctly classified will receive lower weights. Doing so ensures that the next round of training, the samples that were misclassified in the previous round receive more attention in order to train a more accurate model [9].

Next, after each weak classifier is trained, Adaboost determines the weight that the classifier will have in the final model based on its accuracy. Classifiers with higher accuracy will be given greater weights and therefore play a greater role in the final model. By iterating this process, Adaboost is able to combine multiple weak classifiers into a powerful integrated model with good generalisation to all types of data.

4.3. LSTM Optimisation of Adaboost Algorithm

LSTM is a special kind of recurrent neural network, mainly used for processing and predicting time series data. And Adaboost is an integrated learning method that builds a more powerful classifier by combining multiple weak classifiers. Combining LSTM with Adaboost can improve the modelling and prediction of time series data by taking advantage of the long-term memory of LSTM and the integration advantage of Adaboost.

Firstly, LSTM, as an RNN, can effectively capture long-term dependencies in time series data. It controls the input, output and forgetting of information through a gating mechanism, thus retaining important information and discarding irrelevant information during training. This enables LSTM to better capture complex dependencies between data when dealing with time series data, thus improving the accuracy and generalisation of the model [10].

Second, Adaboost, as an integrated learning method, iteratively trains multiple weak classifiers and adjusts the sample weights according to their performance, eventually combining these weak classifiers into a more powerful classifier. When combining LSTM with Adaboost, LSTM can be used as the base classifier and the weights between different base classifiers can be adjusted by the Adaboost algorithm to further improve the performance of the overall model.

Firstly, LSTM is used to model and predict the time series data; then the Adaboost algorithm is trained based on the prediction results generated by the LSTM model as well as the real labels; finally, multiple LSTM models after adjusting the weights are combined into a more powerful integrated model. In this way, the respective advantages of LSTM and Adaboost can be fully exploited and better results can be achieved in time series data modelling and prediction tasks.

5. Result

The dataset is divided according to the ratio of 7:3, with 70% of the data used for model training and 30% of the data used to test the prediction effect of the trained model. This experiment is carried out using a local server and in the experimental setup, the maximum number of training sessions is set to 1000, the initial learning rate is set to 0.01, the learning rate reduction factor is set to 0.1, the number of hidden layer nodes is set to 6, and the number of weak regressors is set to 10. In addition, the machine used for this experiment has a CPU of 32G, a graphics card of 3090, and the experiment is based on Matlab R2019b.

In this paper, the model prediction effectiveness is evaluated using the confusion matrix as well as ACCURACY, which is a two-dimensional table that shows the accuracy of the prediction results of the classification model. Accuracy is the ratio of the number of samples correctly predicted by the model to the total number of samples, and it is one of the most intuitive assessment metrics. The confusion matrix for the training set and test set is shown in Figure 3.

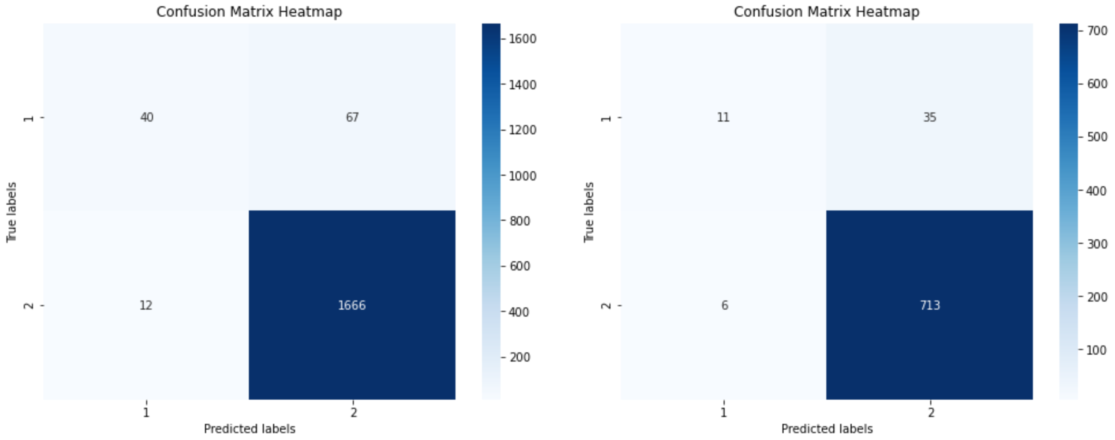

Figure 3. The confusion matrix for the training set and test set.

From the confusion matrix, it can be seen that there are 1706 correct predictions and 79 incorrect predictions in the training set, in which 12 enterprises that should have been predicted as bankrupt are predicted as normal operation and 67 enterprises that should have been predicted as normal operation are predicted as bankrupt. There were 724 correct predictions and 41 incorrect predictions in the test set, of which 6 businesses that should have been predicted to be insolvent were predicted to be operating normally and 35 businesses that should have been predicted to be operating normally were predicted to be insolvent. The prediction accuracy of the training set is 95.57% and the prediction accuracy of the test set is 94.39%, and the model achieves a good prediction effect and generalisation ability.

6. Conclusion

In this paper, we use Long Short-Term Memory Network (LSTM) to optimise the Adaboost model to predict corporate bankruptcy, and achieve satisfactory results. Firstly, this paper uses LSTM to model and predict time series data, making full use of the advantages of LSTM network in processing series data. Subsequently, the Adaboost algorithm was trained by combining the prediction results of the LSTM model with real labels to further improve the accuracy and stability of the model. Finally, multiple LSTM models with adjusted weights were combined into a more powerful integrated model to further enhance the prediction results.

The results of the confusion matrix analysis show that 1,706 samples were correctly predicted in the training set, and only 79 samples were incorrectly predicted. Among them, 12 enterprises that should have been predicted as bankrupt were misjudged as normal operation, and 67 enterprises that should have been predicted as normal operation were misjudged as bankrupt. A total of 724 samples in the test set were correctly predicted, and only 41 samples were incorrectly predicted. Specifically, in the test set 6 firms that should have been predicted as insolvent were mispredicted as operating normally and 35 firms that should have been predicted as operating normally were mispredicted as insolvent.

Overall, an accuracy of 95.57% and 94.39% was achieved on the training and test sets, respectively. This indicates that the proposed LSTM-based optimised Adaboost model for predicting corporate bankruptcy has high accuracy and generalisation ability. This method not only performs well on the training set, but also can achieve quite good results on unknown data. By combining LSTM with Adaboost and constructing an integrated model using multiple LSTM models, accurate prediction of corporate bankruptcy can be effectively performed. This method not only improves the prediction effect and generalisation ability, but also provides a new idea and method for enterprise risk management. It is hoped that this method can play a greater role in practical applications and promote more in-depth exploration and application in related fields.

References

[1] Šneidere, Ruta, and Inga Būmane. "INSOLVENCY OF A COMPANY AND THE METHODS OF FINANCIAL ANALYSIS TO FORECAST IT." Economics & Management (2007).

[2] Antonowicz, Paweł, Kamila Migdał-Najman, and Krzysztof Najman. "Financial predictors of corporate insolvency-assessment of the forecast horizon of variables in models of early warning against corporate bankruptcy." e-mentor. Czasopismo naukowe Szkoły Głównej Handlowej w Warszawie 101.4 (2023): 39-44.

[3] Kušter, Denis, et al. "Early Insolvency Prediction as a Key for Sustainable Business Growth." Sustainability 15.21 (2023): 15304.

[4] DI CARLO, A. "Forecasting and preventing bankruptcy: A conceptual review." African journal of business management 12.9 (2018): 231-242.

[5] Voda, Alina Daniela, et al. "Corporate bankruptcy and insolvency prediction model." Technological and Economic Development of Economy 27.5 (2021): 1039-1056.

[6] Correia, Cláudia, et al. "How can insolvency in tourism be predicted? The case of local accommodation." International Journal of Tourism Cities 8.4 (2022): 1127-1140.

[7] Alexandrovna-Chernyavskaya, Svetlana, et al. "Practical means to forecast potential bankruptcy and financial insolvency of companies." Revista de Investigaciones Universidad del Quindío 34.S2 (2022): 276-283.

[8] Aleksandrovna, Borisuk Anastasia. "Analysis and forecasting of financial insolvency of enterprises." (2023).

[9] CHERNYAVSKAYA, SVETLANA ALEXANDROVNA, et al. "Analytical tools for forecasting financial insolvency and potential bankruptcy of a company." The journal of contemporary issues in business and government 27.2 (2021): 3937-3943.

[10] Dunis, Christian L., and J. Alexandros Triantafyllidis. "Alternative forecasting techniques for predicting company insolvencies: The UK example (1980-2001)." Neural Network World 13.4 (2003): 326-360.