ISBN: 978-981-18-3959-7
2022 the 12th International Workshop on Computer Science and Engineering (WCSE 2022)
doi: 10.18178/wcse.2022.06.001

Financial Risk Early Warning Model of Industrial Listed Enterprises Based on Deep Learning

Xiaohui Yu [1], Shihong Chen[1+], Shijie Chen [2] and Jiaying Tan[3]

[1] School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China

[2] Graduate School of Business, National University of Malaysia, Bangi, Malaysia

[3] School of Accounting, Guangdong University of Foreign Studies, Guangzhou, China

Abstract. In this paper, the public financial indicators of 156 listed industrial enterprises in China Stock Market in T-2 year(2017) are used to predict the company financial status in T year(2019). Based on the feature extraction of Random Forest, a double LSTM financial risk early warning model is built through Keras framework. Machine learning algorithms including LR, SVM, KNN and NBC are set as baseline models. To reduce the influence of unbalanced data on the model, the G-mean is introduced as the comprehensive measure of the model. The result shows that G-mean of RF-LSTM on the test set is much higher than that of machine learning model which verifies the practicability of the RF- LSTM model.

Keywords: LSTM, random forest, G-mean, financial risk

1. Introduction

Affected by the COVID-19, corporate financial crisis has happened frequently. How to build a financial risk early warning model with good performance has attracted more attention. This problem is usually classified as supervised learning and research objects mainly focus on listed companies due to easy access to data and label. In current studies, to improve the performance of model, the number of feature selecting has changed from single variable to multiple variables and more non-financial indicators have been added as features to train. Besides, the model algorithm has shifted from statistical algorithm to machine learning.

However, most traditional machine learning algorithms cannot fully learn the time-series characteristics of enterprise information and feature selection of the model mainly depends on human intervention. There are few industry-specific studies which lead to poor performance of the model due to differences in asset allocation among industries. Besides, the accuracy which is the main measure of model is greatly influenced by unbalanced training samples. Thus, to solve above problems, this paper establishes double LSTM deep learning model based on the feature extraction algorithm Random Forest(RF) through learning financial data of industrial listed companies in T-2 year(2017) to predict the financial status in T year(2019). Mainstream machine learning algorithms including Support Vector Machine(SVM), Logistic Regression(LR), K-Nearest Neighbor(KNN) and Naive Bayes Classifier(NBC) are set as baseline models. The G-mean value is used to comprehensively evaluate the performance of model. The result shows that RF-LSTM model has the best performance on the test set with the G-mean of 0.85 among contrast models. It verifies the practicability of the RF- LSTM model and offers reference for the investors in industrial market.

2. Related Work

The researches on financial early warning model in China generally take the listed enterprises labeled with ST (Special Treatment) in China Stock Market as the classification standard of financial risks. And the current researches can be mainly divided into feature extraction and model construction.

[+] Corresponding author. Tel.: + 8613751846364
E-mail address: 200911836@oamail.gdufs.edu.cn.

2.1. Feature Extraction

Early feature extraction methods were mainly artificial selections based on academic researches and practical experiences. The cash flow indicator and financial ratio were two single variables mainly selected in the financial risk warning model [1]. Then, the selection of features has transformed from single financial indicator to multiple financial indicators. For example, the comprehensive cash flow measurement indexes were used into model which achieved high accuracy [1]. The researches recently have added non-financial indicators as features including Stock Yield, Macro Economic Index, and Internal Governance Indicator [1].

To reduce the influence of manual feature selection on model training, various algorithms of data dimension reduction are applied. Huang Wei [2] compared model performances under different feature selections, and concluded that the performance of the model under RF was better. Besides, the mainstream methods to reduce data dimension include the Principal Component Analysis (PCA) [3], the factor analysis [4], T-test significance test [4] and Spearman correlation analysis [5], and these methods improve the performance of the model to a certain extent.

2.2. Model Construction

In terms of model construction, the financial risk early warning model was the univariate analysis model in early stage [2]. Since the model with single feature may lead to overgeneralization, the multivariate linear analysis was adapted and the Z-score model was built by Altman [2].

With the development of machine learning, machine learning algorithms are applied and optimized in this field. Common algorithms include LR [6], SVM [7], KNN [7], NBC [7] and RF [2]. To improve the performance of the LR model, the factor analysis was used for feature selection [6] and cost sensitive learning was applied to solve the problem of learning offset [3]. Chen Qicheng [7] compared NBC, LR, KNN, SVM and Quadratic Discriminant model, and concluded that NBC was only inferior to LR and SVM.

In recent years, model construction has transferred from machine learning to neural network model. Liu Hongtao [8] built a four-layer neural network model with three hidden layers and achieved the accuracy of 75%. Besides, it is recognized that LSTM model has a general prediction effect on A-share listed companies [9]. Lin Dannan et al. [10] concluded that the LSTM model had the highest early warning accuracy and stability by comparing SVM, BP neural network and convolutional neural network. Xu Wange [1] built a four-layer LSTM financial warning model with BN layer and Dropout layer to optimize the neural network.

3. Approach

3.1. Random Forest

Random Forest (RF) is a machine learning algorithm that uses multiple trees to train and predict samples, and has been widely used in data dimension reduction. It expresses the importance of a variable by calculating the cumulative mean and standard deviation of the decrease of impurity degree of a variable in each tree. The computation formulas of the importance of feature are as follows.

$$GI_m = 1 - \sum_{k=1}^{|K|} p_{mk}^2 \tag{1}$$

GI_m represents the Gini index of the node m, that is, the average change of node splitting purity in all decision trees of the RF. K represents K categories, and p_{mk} represents the proportion of category k in node m.

$$VIM_{jm} = GI_m - GI_l - GI_r \tag{2}$$

VIM_{jm} is the importance of feature X_j in node m. GI_l and GI_r represent Gini indexes of the two nodes after branching.

$$VIM_j \ = \sum_{i=1}^{n} \sum_{m \in M} VIM_{jm} \tag{3}$$

VIM_{jm} is the importance of feature X_j. Set M is the node set of feature X_j appearing in decision tree i. It is assumed that there are n trees in the random forest. The greater the importance of the feature is, the greater the contribution of the feature is.

3.2. Long and Short Term Memory

Long and Short Term Memory network(LSTM) is a special form of RNN. It is designed to solve the problem of long-term data dependencies. The structure of LSTM is shown in Fig. 1. The horizontal line running from left to right above the top of the LSTM unit represents the unit status. Information is passed from one cell to the next through this horizontal line. LSTM controls information by adding gate units.

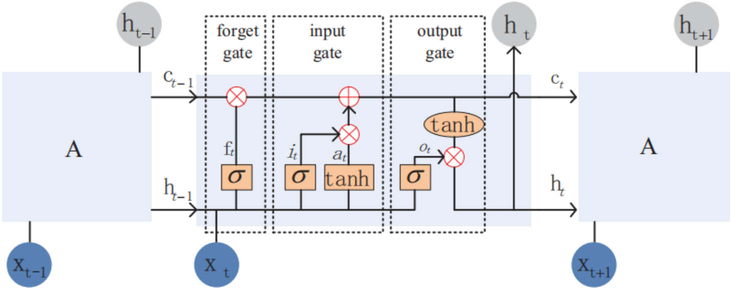

Fig. 1: The structure of one LSTM unit [11]

One LSTM unit consists of three gate units: the forget gate, the input gate and the output gate. The calculation formulas of the model are as follows. In formulas, x_t is the input for the cell at time *t*; σ is activation function; *W* is weight matric; h_t is the output of the cell at time *t;b* is bias vector.

- The input gate: a_t and i_t represent the state of the memory.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4}$$

$$a_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{5}$$

- The forget gate: f_t represents the values of the forget gate at time t and C_t is cell state at time t.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{6}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot a_t \tag{7}$$

- The output gate: o_t is the value of the output gate at time t and h_t is the output of the cell at time t.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_t = o_t \cdot tanh(C_t) \tag{9}$$

4. RF-LSTM Model Design

Based on the theories and literature mentioned before, a double LSTM deep learning model based on RF is built to predict the financial risk of the listed industrial companies.

4.1. Data Process

The raw data are all the public indicators of listed enterprises from 2014 to 2019 obtained from the Wind database. Because the financial reports in T-1 year are generally released at the beginning of T year, the financial data in T-2 year (2017) is selected to predict the financial risk of listed enterprises in T year (2019). The company with the label of ST(ST company) is recognized as negative sample, while the company without the label of ST (non-ST company) is positive sample. In fact, there are very few ST companies in China Stock Market. Thus, to obtain as many samples with practical significance as possible, the selection of listed companies has the following criteria: (1)the company has been listed for more than three years, (2)the company belongs to the industrial industry, (3)the ST company has been first tagged with ST in 2019 since 2014 and fewer than three missing indicators are allowed, (4)the non-ST company has not been tagged with ST since 2014 and missing indicators are not allowed.

After the public financial indicators of negative companies are obtained, the missing rate of each feature is calculated and the threshold value of the missing rate is set as 0.2. If the missing rate of the feature is higher than 0.2, the feature is discarded; if the missing rate of the feature is lower than 0.2, the missing value is filled with the mean of the feature.

4.2. Feature Extraction

All pre-processed samples are put into the RF model for training, and the importance of each feature is calculated. The features are retained whose importance are higher than or equal to the mean of overall feature importance.

4.3. Double LSTM Model

In LSTM, using standardized data sets can obtain more accurate training results. Thus, the data after feature extraction are normalized. The normalization formula is as follows:

$$Z = \frac{X - X_{min(axis=0)}}{X_{max(axis=0)} - X_{min(axis=0)}} \tag{10}$$

Based on a number of experiments, this paper builds double LSTM model to train input data. To alleviate the problem of gradient disappearance, the Batch Normalization (BN) layer is added. The BN is used to solve the problem of data distribution changes in the middle layers during training [1]. Its essential principle is that the input data of the layer is normalized before entering the next layer. The problem of poor generalization ability caused by overfitting can be largely alleviated by adding the Dropout layer [1]. Dropout layer sets a probability of inactivation for each layer of the network at which neurons are eliminated.

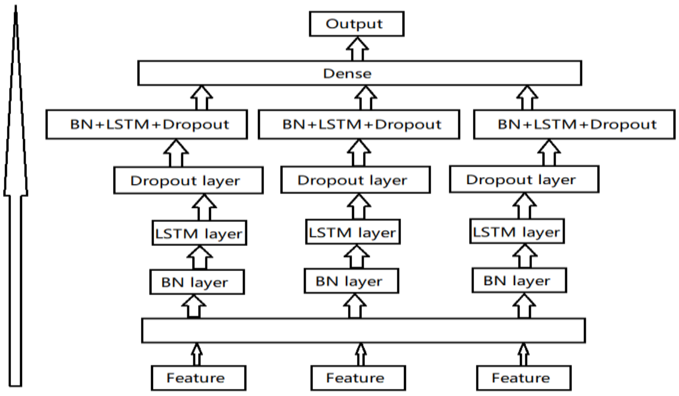

Fig. 2: The structure of double LSTM model

The structure of Double LSTM model is shown in Fig. 2. The depth direction of layer is from bottom to top. The layers in sequence are BN layer, LSTM layer, Dropout layer, BN layer, LSTM layer, Dropout layer and Dense layer.

4.4. Performance Measures

In this paper, the accuracy rate, recall rate and G-mean are used to evaluate the performance of the model. Accuracy is a universal measure of the model. However, if the samples trained are unbalanced data set, the prediction accuracy is largely affected by samples of the majority category. Therefore, in order to reduce the influence of non-ST samples on the performance measure, the G-mean is introduced for the comprehensive evaluation model [3]. The formulas are as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \tag{11}$$

$$Recall = \frac{TP}{TP+FN} \tag{12}$$

$$Specificity = \frac{TN}{TN+FP} \tag{13}$$

$$G\_mean = \sqrt{Recall * Sepcificity} \tag{14}$$

TP is the number of samples where 1 is accurately predicted to be 1, FP is the number of samples where 0 is predicted to be 1, TN is the number of samples where 0 is predicted to be 0, and FN is the number of samples where 1 is predicted to be 0. According to (14), the larger the G-mean value is, the better the model's accurate classification performance is.

5. Data Analysis

5.1. Data Set

The number of ST companies obtained by data processing is 39 while the number of non-ST companies is 1498. The label of 1 represents ST, and non-ST is 0. Since extreme unbalanced data sets lead to poor training effects of the model and LSTM model requires a large sample size, 117 samples are randomly selected from non-ST companies based on the ratio of ST: non-ST =1:3. Therefore, 156 listed industrial enterprises with 70 features are selected as experimental samples, including 39 ST samples and 117 non-ST samples. The data set is segmented according to the ratio of training set to test set =4:1. Meanwhile, the ratio of labels ST and non-ST in each data set is kept at 1:3. Thus, 124 samples are obtained as the training set and 32 samples are obtained as the test set.

5.2. Feature Extraction

156 pieces of data are put into RF model for training and 15 features are selected according to the mean of the importance of the feature. The importance of the selected features are shown in Tab.1.

Table 1: The importance of features calculated by RF

Feature	Importance
Sustainable Growth Rate	0.246
Operating Profit Margin	0.061
Return on Assets A	0.057
Net Profit Margin on Total Assets ROAA	0.054
Common Stock Return Rate	0.042
Value Preservation and Appreciation Rate of Capital	0.040
Ratio of Accounts Receivable to Revenue	0.037
Financial Liability Ratio	0.034
Ratio of Receivables	0.033
Growth Rate of Net Assets Per Share A	0.028
Z Index	0.027
Operating Profit Margin	0.019
Cash Reinvestment Ratio	0.019
Total Cash Recovery Rate	0.017
Sales Expense Ratio	0.015

5.3. Parameter Settings of Proposed Model

Based on a large number of experiments, the parameters of the Double LSTM model finally are set as follows:(1)hidden layer neurons of each LSTM layer are 256, (2)the optimizer is Adam with the initial learning rate 0.001, (3)the number of iterations epochs is 50, (4)the batch size was 64, (5)The deactivation rate of Dropout layer is 0.5, (6)the loss function is MAE.

5.4. Baseline Models

This paper adapts the traditional machine learning algorithms including LR, NBC, KNN, and SVM, and single LSTM model to compare with the Double LSTM model.

- Logistic Regression: the optimization method of LR is set as Liblinear by Solver, the regularization category penalty is L2, and the regularization force *C* is 1.
- Naive Bayes Classifier: Gaussian Bayesian algorithm is adopted.
- K-Nearest Neighbor: the value of *K* is 5.
- Support Vector Machine: Kernel Function is RBF, and the penalty coefficient *C* is 1.
- Single LSTM model: the model consists of one BN layer, one LSTM layer and one Dropout layer.

5.5. Analysis of Results

The performance of each model on the test set after training is shown in Tab.2:

Table 2: Performance of models

Model	Accuracy(%)	Recall(%)	G-mean
LR	71.88	50	0.63
NBC	71.88	62.5	0.69
KNN	87.5	62.5	0.77
SVM	87.5	62.5	0.77
Single LSTM	81.25	62.5	0.74

Double LSTM	90.62	75	0.85

The performance of LR is the worst among the models, particularly in identifying ST companies, although it achieves the same level of accuracy as NBC. KNN and SVM have same performance and are the best among traditional machine learning algorithms. However, the G-mean values in all the machine learning algorithms are relatively low. The Single LSTM model works better than LR and NBC to some degree, but it has no advantage over KNN and SVM. However, the values of three measures of LSTM model increases significantly after adding another LSTM layer. The Double LSTM model obtained the G-mean of 0.85 with the accuracy of 90.62% and the recall of 75%, which proves it is the best model among baseline models and verifies its application in financial risk warning.

6. Conclusion

This paper proposes RF-LSTM model aiming to predict the financial risks of industrial listed companies in China through training the selected indicators in the existing public annual reports. Based on the unbalanced samples in this paper, the G-mean value is introduced as performance measure of model to evaluate model more objectively. Experiments have been conducted and prove that the RF-LSTM model proposed has better performance than other machine learning models. Its value of G-mean confirms its ability in classification especially in identifying the ST companies. The result verifies the practical value of RF-LSTM model in industrial enterprises.

7. Acknowledgements

This work was supported by Provincial College Students Innovation Training Project (grant No. S202111846024).

8. References

[1] Xu Wange. Research on Financial Early Warning of Listed Companies in China Based on Deep Learning[D]. Chongqing University of Technology, 2020.DOI:10.27753/d.cnki.gcqgx.2020. 000336.

[2] Huang Wei. An Empirical Study on Financial Crisis Early Warning of Listed Companies Based on Different Feature Selection[D].Jinan University,2020.DOI:10.27167/d.cnki.gjinu.2020.001698.

[3] Ren Tingting, Lu Tongyu, Zhang Weinan. Research on Financial Early Warning Based on Feature Selecting and Cost Sensitive Learning[J]. Commercial Accounting,2021(20):11-16.

[4] Niu Fangqi. Research on Financial Risk Prediction Model of Small and Medium-sized Enterprises Based on LSTM Model[D]. Chinese Academy of Fiscal Sciences, 2021. DOI: 10.26975 /, dc nki. GCCKS. 2021.000214.

[5] Zhang Jing. Financial Early Warning of Real Estate Listed Companies Based on Logistic Model[J]. Productivity Research, 2021 (6) : 149-152. The DOI: 10.19374 / j.carol carroll nki. 14-1145 / f 2021.06.028.

[6] Zhong Ling. Financial Risk Early-Warning Analysis of Small and Medium-sized Enterprises Based on Factor Analysis and Logistic Regression[J]. Commercial Accounting,2021(21):66-69.

[7] Chen Qicheng. Research on Financial Risk Warning of Listed Enterprises Based on PCA-NBC Algorithm[J]. Management and Technology of Small and Medium-sized Enterprises (Mid-day),2021(12):85-87.

[8] Liu Hongtao. Research on Financial Risk Early Warning of Listed Companies Based on Deep Learning[J]. Accounting for Township Enterprises in China,2019(09):89-93.

[9] Song Ge, Ma Tao. Research on Financial Risk Early Warning Model of Listed Companies Based on Deep Learning[J]. Value engineering, 2019, 38 (01) : 53-56. DOI: 10.14018 / j.carol carroll nki cn13-1085 / n. 2019.01. 018.

[10] Lin Dannan, Li Shanshan, Xiao Shilong, Zhang Deyu. Early warning model of enterprise financial risk based on LSTM neural network[J]. Journal of Nanjing University of Science and Technology, 2021, (03) : 361-365 + 374. DOI: 10.14177 / j.carol carroll nki. 32-1397 n. 2021.45.03.015.

[11] Shifan Song,Xuejun Pan,Mingxiang Guo,Qi Lang,Xiaodong Liu. Futures Price Forecasting Based on the Feature Fusion LSTM Model Using Long-Term Price Patterns[C],Proceedings of the 40th Chinese Control Conference 2021:514519.DOI:10.26914/c.cnkihy.2021.025122.