Evaluation of Stock Market Risk Model Based on Random Forest + Two-Way LSTM

Yunlan Xue[1(✉)] and Jian Yao[1,2]

[1] School of Artificial Intelligence, The Open University of Guangdong, Guangzhou 510091, China
ylxue@gdrtvu.edu.cn, abjian.yao@whu.edu.cn

[2] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430070, China

Abstract. In view of the fact that the traditional risk evaluation model has the problem of repeated indexes when evaluating the risk of stock exchange market, this paper uses random forest and two-way LSTM model to evaluate and study the risk model of stock exchange market. Firstly, random forest method is used to screen the primary indexes to remove the repetitive indexes in the index system; Secondly, the two-way LSTM model is used to improve the accuracy of various indicators and obtain the evaluation results; Finally, by comparing the evaluation results of random forest with those of two-way LSTM model and other model experimental methods, it is found that random forest plus two-way LSTM model can improve the investment income of investors more accurately.

Keywords: Stock trading · Random forest · Bidirectional LSTM · Securities trading · Risk identification and evaluation

1 Introduction

In the early stage, the prediction and evaluation of stock price and stock trading risk adopted traditional mathematical methods, such as linear regression and Markowitz's mean variance model. These basic mathematical methods are highly explanatory and understandable. However, the linear models can only capture the linear features in the financial market, and have difficulty in capturing those complex nonlinear features [1, 2]. In order to capture and fit the complex relationship between the stock price and various influencing factors, machine learning algorithms such as support vector machine, random forest, artificial neural network and integrated learning algorithms such as xgboost and adaboost are applied to stock price prediction and stock trading risk assessment to mine the nonlinear characteristics of the stock market. However, machine learning algorithm has shortcomings in the research of stock price prediction with time series characteristics [3]. The further deepening and complexity of the model structure of deep neural network strengthens the model's ability to realize nonlinear complex functions and pattern recognition of input data, and it has certain advantages in solving complex and

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 912–922, 2023.
https://doi.org/10.2991/978-94-6463-198-2_95

highly nonlinear problems [4, 5]. At present, in terms of stock price prediction, Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) are frequently used to model the time series of stock prices [6, 7], which have the following shortcomings. (1) There is a certain cross-repetition and importance difference in the expression of stock trading risk among various factors, which is not conducive to the interpretation of stock trading risk assessment results. (2) In the selection of influencing factors, it mainly focuses on hard indicators such as the macro market, the operation situation of stock issuing enterprises and the previous stock price. The investors' investment feedback to the stock market (such as the reinvestment behavior and investors' evaluation) was not included in it. (3) The different timeliness and granularity of relevant data collected in stock price prediction lead to the low overall quality of the data set of factors affecting stock price [8–10].

In view of the above, based on stock industry data, this study introduces flexible evaluation index, and integrates macro market economy condition, the condition of the micro stock issuance of the company's assets and investors' investment feedback into an organic whole to build stock trading risk evaluation index system, Then, optimize the primary indicators to reduce the interference of redundant indicators or repeated indicators on the risk evaluation results. Finally, the deep learning algorithm is used to mine and quantitatively evaluate the risk of stock trading and improve the accuracy of stock price prediction. Through the monitoring, evaluation and analysis of stock trading risks, this study can provide reliable data support and auxiliary analysis results support for investors in stock investment planning and decision-making, and help investors reasonably choose the most suitable stocks for their own investment needs from many stocks. On the other hand, this study can help industry practitioners grasp the development trends of the industry accurately, and formulate appropriate industry development strategies.

2 Stock Trading Risk Monitoring and Evaluation Model

2.1 Design Ideas

The data collected in the risk assessment of stock trading are large in volume, wide in source and diverse in form. The data after the optimization of the risk assessment index in the early stage still have a very complex implicit relationship. In order to mine the risk characteristics of stock trading and help provide stock investment reference for investors, the model should be studied to mine the complex relationship of multiple characteristics in stock trading risk. This study introduced LSTM to train and learn the dependency of the sequence data, then the complex and deep relationship between the index and risk of stock trading was mined. Through the sequence processing and pattern analysis of the monitoring indexes, the potential periodicity or semi-periodicity in stock trading can be mined.

2.2 Long and Short Term Memory Networks

Recurrent Neural Network (RNN) is a kind of neural network for processing sequence data. Compared with the general neural network, it can process the data of sequence

change. For example, the meaning of a word may have different meanings depending on the above content mentioned, and RNN can solve such problems very well. The function of ordinary RNN can be expressed as follows.

$$h', y = f(h, x) \tag{1}$$

where x is the input data in the current state; h is the input received from the previous node; y is the output in the current node state; h' is the output transferred to the next node. Y often uses h' to input into a linear layer (mainly for dimension mapping), and finally obtains the result through the Softmax layer. The calculation process is shown in Fig. 1.

LSTM proposed by Hochreiter & Schmidhuber (1997) is a kind of cyclic neural network. It recently improved and extended by Alex Graves. LSTM can learn long-term dependencies between time-step sequences of data. Different from CNN, LSTM can remember the state of the network between predictions. In order to solve the gradient vanishing and gradient explosion problems in long sequence training, all RNNs have a chain form of repeated neural network modules. In standard RNN, this repeating module has a very simple structure, such as a tanh layer. Compared with normal RNN, LSTM can perform better in longer sequences. Compared with ordinary RNN, the main input-output differences of LSTM are shown in Fig. 1. RNN has only one transfer state, h^t. However, LSTM has two transfer states, c^t (cell state) and h^t (hidden state).

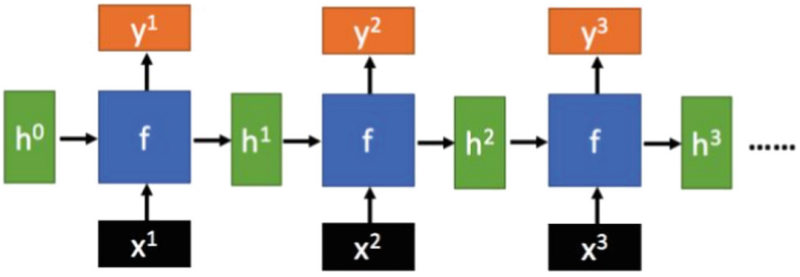

Fig. 1. Calculation process of RNN

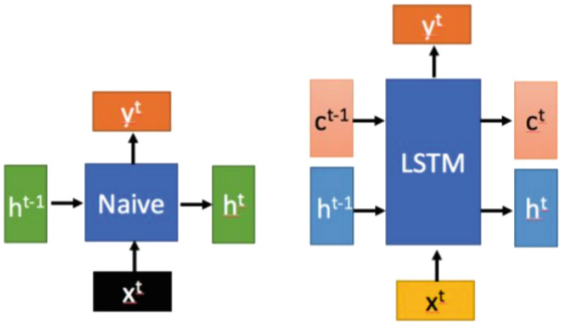

Fig. 2. The main differences of input and output between ordinary RNN and LSTM

Figure 2 shows the core components of LSTM network are sequential input layer and LSTM layer. The sequential input layer adds the sequential data to the network. The LSTM layer learns long-term dependencies between the time steps of the sequence data over a period of time. In this study, two-way LSTM model is adopted to model stock trading risks. During the training of LSTM model, only some low-level generalized spatial features can be extracted from the fixed networks in the first few layers. More advanced features can be extracted from the networks in the later layers.

In order to further enhance the sensitivity of LSTM to stock price, reduce the calculation cost and accelerate the convergence rate of the model, the neural network is improved as follows.

(1) Continuous stack of 3*3 convolution kernels is used to realize the same receptive field as 5*5 or 7*7 convolution kernels, which increases the depth of the network while reducing network parameters. By using the nonlinear activation function several times, the learning ability of the network to the feature is improved. The convolutional neural network model adopted in this study is as follows. The first convolution block contains two convolution layers. Each convolution layer is composed of 64 convolution kernels with the size of 3*3. Step size is set to 1 pixel. The output of the first convolution block is obtained by convolutional pooling operation on the input data. The second convolution block also contains two convolution layers. Each convolution layer consists of 128 convolution kernels of size 3*3. The output of the second convolution block is obtained by convolving the output of the first convolution block. The third convolution block contains three convolution layers. Each convolution layer is composed of 256 convolution kernels with the size of 3*3, which is equivalent to a convolution layer with the size of 7*7. After each convolution block is calculated, 2*2 maximum pool operations are carried out. Finally, it is composed of full connection layer and Softmax layer. The full connection layer has 200 neurons. The full connection layer summarizes the local features of the convolution layer as the global features. If too few nodes are selected, the ability of the neural network to express its understanding of the characteristic data is deprived. The ReLu activation function is used in each convolution layer.

(2) In order to prevent excessive dependence of parameters on data, further increase the ability of model generalization learning and reduce the amount of parameter calculation, access Dropout layer after fc1 and fc2 layer, and set the loss rate as 50%. The Dropout mechanism randomly inactivates half of the neurons during training.

(3) In order to accelerate the training speed of the model to ensure gradient return and avoid gradient explosion or gradient disappearance, Batch Normalization (BN) mechanism is added. Batch of regularization layer is a kind of regularization for use in the deep learning mechanism. It guarantees that each layer of the distribution are the same, which also means that the parameters will have been heavily involved in the keen range of loss function. It can avoid the gradient disappear or gradient explosion problem, can greatly speed up the training of the network model. Instead of adding the batch regularization layer, this paper integrates the batch regularization layer into the convolution layer. This modification can ensure the function of batch regularization, and can improve the training speed and the accuracy without the increase of network depth.

3 The Experimental Process

3.1 The Collection of Stock Trading Feedback Indicators

Indicators of investors' response to stock trading performance are available from Chinazqnew website. Taking investor's evaluation as an example, this study elaborates on the process of analyzing the sentiment attitude of review data crawled from Chinazqnew website.Sentiment analysis is to analyze the emotions implied in investor comment data and judge the investor's emotional attitude (positive and negative) towards the corresponding stock transaction.

snownLP, Python's Chinese NLP library, can generate values between 0 and 1 based on input comment participles. The generated value over 0.5 indicates that the review is positive, and vice versa. The closer the numbers get to 0 and 1, the more extreme the sentiment hided in the comments.

The process of snowNLP invocation is as follows. The comment participles are read in the linked list and selected one by one by s = snowNLP (). Then, the emotional attitude analysis results were obtained by s.sentiments. Finally, the bar chart is used to intuitively show the result of emotional attitude. Bayesian algorithm is used to judge emotional attitude in snowNLP library.

Figure 3 shows the chart of sentiment analysis result drawn by sentiment analysis method after investors comment on the stock of Qingdao Tgood Electric Co., Ltd. (code 300001) on the website of Chinazqnew. It is clear from the figure that there is a line next to 1.0 which is much larger than the others. It shows that for stock 300001 (GEM), investors maintain more positive and optimistic attitude, but a few are not optimistic. Investors' attitudes towards stock trading can be intuitively observed through such sentiment analysis results and visualization, which reflects investors' trust and optimism in stock trading.

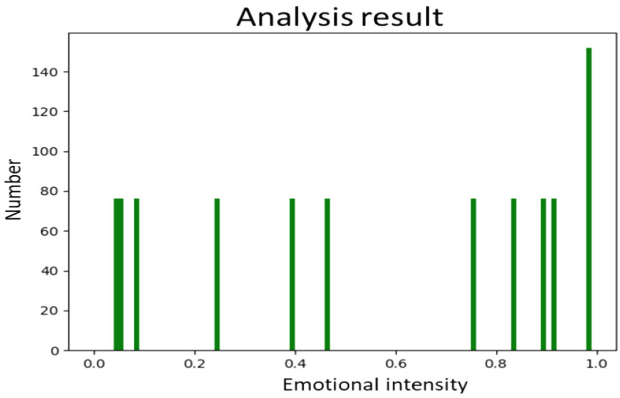

Fig. 3. The chart of sentiment analysis result

3.2 Training Set and Test Set

We used to collect annual (2008–2018) macro-market economic situation indicators, more than 40,000 stocks issued and nearly 1 million comments posted by more than 6,000 investors. The data volume involved is about 1.84 GB.

Taking the issued stocks as the basic object, the macro indexes, micro indexes, as well as investors' repurchase rate and emotional attitude (positive and negative) are stored in a series in chronological order. The minimum time unit is calculated on a daily basis to reflect the internal and external environment of the issued stocks as well as the investment trust degree to the maximum extent. Through sorting out the collected data of more than 40,000 stocks and their related 27 risk evaluation indicators, a total of 884739620 valid data were obtained after excluding delisting and suspension data. The valid data were divided into training sample set and verification sample set. In consideration of the lagging effect of macro environment, micro environment and issued stocks on stock investment risk and investment behavior, a rolling model is adopted to split the effective data set to form training sample set and verification sample set.

The effective data from 2008 to 2013 were selected for the construction and training of the stock trading risk evaluation model, and the corresponding test data set was based on the effective data from 2013 to 2014. In other words, the first five years are training sample sets, and the last one year are verification sample sets, thus forming the rolling window of dividing training sample sets and verification sample sets. A total of five sets of training sample sets and verification sample sets are generated. The sample sets are as follows: the training sample sets of 2008–2013 (2013–2014 validation sample set), 2009–2014 (2014–2015 validation sample set), 2010–2015 (2015–2016 validation sample set), 2011–2016 (2016–2017 validation sample set), and 2012–2017 training sample set (2017–2018 validation sample set).

3.3 Parameter Selection of LSTM

This study defines a two-way LSTM network for the construction of stock trading risk evaluation model, and outputs the risk of each stock trading. The network consists of a layer of 200 hidden units and a dropout layer with a drop probability of 0.5. The discard probability is a tool used to prevent over fitting. The network uses this value to randomly skip some data and avoid memorizing the training set. Hidden units often number in the hundreds. Determining the number of hidden units to include is a tradeoff. If the number is too small, the model will not have enough memory to do the learning. If it is too large, the network may over fit. In the experiment of this paper, the core parameter of attention times is set to 3. The learning rate is still set to 0.05 in order to prevent falling into the local minimum value, and the parameter initialization of the neural network obeys the uniform distribution of U $(-0.01, 0.01)$. The neuron loss rate of this neural network is set to 0.5.

4 Analysis of Risk Monitoring Results

4.1 Evaluation Method

This paper evaluates the effectiveness of the stock trading risk evaluation model based on random forest and two-way LSTM mainly from two aspects. The first is to evaluate the degree of conformity with the stock set recommended by stock research experts, which is essentially the explicit evaluation. The second is the evaluation of deviation between the stock and the investor's behavior (forming investment behavior and obtaining higher economic returns), which is essentially the implicit evaluation.

(1) Explicit evaluation methods
Explicit evaluation method is a more direct and objective evaluation method of stock trading risk evaluation model. It directly carries out validity research on stock sets recommended by stock research experts, and mainly adopts evaluation indexes including Precision, Recall and Coverage.

1) Precision
Accurate rate refers to the proportion of stocks that exist in the recommendation set of stock trading risk assessment model and are favored and recommended by experts in all investable stocks. The higher the value is, the stronger the effectiveness of the stock trading risk evaluation model is. Its calculation formula is as follows:

$$Precision = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{2}$$

where N_{TP} is the number of stocks that exist in the recommended set of the stock trading risk assessment algorithm and are favored and recommended by the expert; N_{FP} is the number of stocks that exist in the recommended set of the stock trading risk assessment algorithm but are not recommended by the expert; Precision is the accuracy index of stock trading risk evaluation model.

2) Recall
The recall rate refers to the proportion of stocks in the set of stock trading risk assessment model that are favored and recommended by experts in all stocks recommended by experts. The higher its value is, the stronger the effectiveness of the stock trading risk assessment model is. Its calculation formula is as follows:

$$Recall = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{3}$$

where N_{TP} is the number of stocks that exist in the recommended set of the stock trading risk assessment algorithm and are favored and recommended by the expert; is N_{TP} the number of stocks recommended by the expert that do not appear in the stock trading risk assessment algorithm's recommendation set; Recall is the recall rate index recommended by a stock trading risk assessment model.

3) Coverage
Coverage refers to the proportion of stocks recommended by all experts in all stocks by the risk assessment model of stock trading. The higher the value is, the stronger

the recommendation ability of the risk assessment model system of stock trading to unpopular stocks is. Its calculation formula is as follows:

$$Coverage_i = \frac{|U_{u \in U} R(u)|}{|I|} \tag{4}$$

where U is Experts set; u is a member of the expert set U; I is the total number of stocks; $R(u)$ is the number of stocks recommended by the stock trading risk assessment algorithm; Coverage is the coverage index recommended by a stock trading risk assessment algorithm.

(2) Implicit evaluation methods

In this paper, MAE (Average Absolute Error) is used as the performance evaluation index of stock trading risk evaluation model. MAE is used to calculate the error between the low risk stocks and the investor behavior (the stocks that generate investment behavior and obtain higher economic returns) in the stock trading risk evaluation model, so as to quantitatively evaluate the performance of the stock trading risk evaluation model. The smaller the value of MAE is, the smaller the gap between the successful stock trading recommended by the algorithm and the actual stock trading, and the higher the risk evaluation accuracy of the risk evaluation model.

The MAE calculation formula of stock trading risk evaluation algorithm is as follows:

$$MAE_i = \frac{\sum_{j=1}^{m_i} \left| \mathrm{ps}_{i,j} - ts_{i,j} \right|}{m_i} \tag{5}$$

where i is the investor in verification dataset; m_i is the total number of stocks invested; $\mathrm{ps}_{i,j}$ is the investment behavior of investor i to stock j under the risk assessment algorithm of stock trading; MAE is the average value of n selected investors. The calculation formula is as follows:

$$MAE = \frac{\sum_{i=1}^{n} MAE_i}{n} \tag{6}$$

4.2 Stock Trading Risk Evaluation Results

LSTM network model training and verification based on the validation sample set were carried out for the 5 sets of training sample sets and their corresponding validation sample sets obtained in Sect. 4.5.1, and the accuracy, recall rate, coverage rate and MAE of stock trading risk assessment model construction and evaluation were calculated respectively. Table 1 shows the evaluation results of stock trading risk model under the five sample sets. According to the average calculation, the accuracy rate of the stock trading risk evaluation model proposed in this paper for the stock trading risk evaluation and the recommendation of low-risk stock investment is 0.8388, the recall rate is 0.7994, the coverage rate is 0.8702, and the MAE is 0.7655. The results show that the stock trading risk evaluation model based on random forest and two-way LSTM has high risk evaluation accuracy, and can accurately recommend investable stocks with high economic return to investors, effectively reduce the probability of risk occurrence in stock investment, and effectively improve investors' investment returns.

Table 1. Stock trading risk evaluation results under five sample sets

Sample sets	Precision	Recall	Coverage	MAE
Training Sample Set 2008–2013 (Validation Sample Set 2013–2014)	83.22%	79.58%	88.14%	74.82%
Training Sample Set 2009–2014 (Validation Sample Set 2014–2015)	81.47%	77.43%	85.36%	76.19%
Training Sample Set 2010–2015 (Validation Sample Set 2015–2016)	84.92%	80.35%	86.62%	77.38%
Training Sample Set 2011–2016 (Validation Sample Set 2016–2017)	87.30%	82.74%	88.59%	78.42%
Training Sample Set 2012–2017 (Verification Sample Set 2017–2018)	82.51%	79.59%	86.37%	75.93%
Mean	83.88%	79.94%	87.02%	76.55%

4.3 Experimental Comparative Study

In order to verify the accuracy of the stock trading risk evaluation model based on random forest and two-way LSTM, this study combined the optimized random forest stock index selector with several mainstream forecasters (CNN and RNN) to quantitatively evaluate the stock trading risk respectively. The accuracy, recall rate, coverage rate and MAE are used to analyze the accuracy of stock trading risk evaluation. In this comparative experiment, CNN is taken as the reference model, and the learning rate of CNN adopts the way of momentum decay. The initial learning rate is 0.05, and the attenuation coefficient is set to 0.8. The number of convolution kernels was set to 128, and the maximum pooling was used to reduce the characteristic dimension. The initial weight value of the convolutional neural network follows the uniform distribution of U (−0.15, 0.15), and the loss rate is set to 0.5.

The above four models were used to model and evaluate the stock trading risk under the five samples. The results are shown in Table 2. As can be seen from the table, the accuracy of random forest and two-way LSTM in evaluating stock trading risk is higher than that of random forest and RNN and random forest and CNN, among which random forest and CNN has the lowest evaluation accuracy. The average precision is 0.6042, the average recall rate is 0.5572, the average coverage rate is 0.6079, and the average MAE is 0.5838. There is a big difference between the risk situation of low-risk stocks obtained by this model and the actual stock, so it is difficult to provide investors with relatively reliable guidance on stock investment.

Table 2. Stock trading risk evaluation results under four models

Model	average precision	average recall	average coverage	averageMAE
Random forest and bidirectional LSTM	83.88%	79.94%	87.02%	76.55%
Random forest and CNN	60.42%	55.72%	60.79%	58.38%
Random forest and RNN	71.36%	63.44%	66.17%	70.63%

5 Conclusion

In terms of risk monitoring results analysis and experimental comparison, the accuracy, recall rate, coverage rate and MAE are used to analyze the accuracy of stock trading risk evaluation quantitatively. The risk assessment accuracy rate is 0.8388, recall rate is 0.7994, coverage rate is 0.8702, and MAE is 0.7655, which indicates that the stock trading risk assessment model based on random forest and two-way LSTM has high risk assessment accuracy. By using CNN and common RNN combined with random forest stock index selector as the training model of stock trading risk evaluation, the accuracy of stock trading risk evaluation was compared. It is found that the accuracy of stock trading risk evaluation of random forest and two-way LSTM is higher than that of random forest and RNN and random forest and CNN.

This paper uses daily stock volume and price data. In the further study, the daily data can be extended to the level of hour and minute. Higher frequency data means more samples can be generated and deep learning can be used more effectively. On the other hand, high frequency data contains more information than low frequency data, which is more conducive to the further mining of deep learning models.

Acknowledgments. Thanks to the support of the scientific research project "Recommended learning algorithm driven by knowledge map and deep learning and its application in lifelong education (No. 2020KTSCX401)" by Guangdong Education Department and Guangdong Polytechnic vocational college. And thanks to the support of the scientific research project "Research on deep learning model based on temporal events and semantic background in event extraction and prediction tasks (No. ZD2001)" by The Open University of Guangdong (Guangdong Polytechnic Institute). In this paper, the research was sponsored by Artificial intelligence application innovation team of Guangdong Open University scientific research project (No. 20282205).

References

1. Reza Bradrania, Davood Pirayesh Neghab, Mojtaba Shafizadeh. State-dependent stock selection in index tracking: a machine learning approach[J]. Financial Markets and Portfolio Management,2021(prepublish).

2. Sandeep Patalay, Madhusudhan Rao Bandlamudi. Decision Support System for Stock Portfolio Selection Using Artificial Intelligence and Machine Learning[J]. Ingénierie des Systèmes d'Information,2021,26(1).
3. Yihua Zhong, Lan Luo, Xinyi Wang, Jinlian Yang. Multi-factor Stock Selection Model Based on Machine Learning[J]. Engineering Letters,2021,29(1).
4. Yugan Geng, Jiaming Zhu, Xia Li. Analysis of Multi-factor Quantification Stock Selection Strategy based on GEM[J]. Journal of Global Economy, Business and Finance,2021,3(1).
5. Saha A. The use of return on equity as a criterion for stock selection in the Indian equity markets[J]. Journal of Physics: Conference Series,2021,1784(1).
6. Vuković Marija, Pivac Snježana, Babić Zoran. Comparative analysis of stock selection using a hybrid MCDM approach and modern portfolio theory[J]. Croatian Review of Economic, Business and Social Statistics,2020,6(2).
7. Jingti Han, Zhipeng Ge. Effect of dimensionality reduction on stock selection with cluster analysis in different market situations[J]. Expert Systems With Applications,2020,147.
8. Yi Fu, Shuai Cao, Tao Pang. A Sustainable Quantitative Stock Selection Strategy Based on Dynamic Factor Adjustment[J]. Sustainability,2020,12(10).
9. Meiyi Zhou, Lianqian Yin. Quantitative Stock Selection Strategies Based on Kernel Principal Component Analysis[J]. Journal of Financial Risk Management,2020,09(01).
10. Kim, Kim. Variable Selection for Artificial Neural Networks with Applications for Stock Price Prediction[J]. Applied Artificial Intelligence,2019,33(1).