Research on stock prediction based on simulated annealing algorithm and ensemble neural learning

Yu Sun, Liwei Tian*
Guangdong University of science and technology, Guangdong, China

ABSTRACT

By combining the relevant theoretical knowledge of simulated annealing algorithm, the parameters of the XGBoost model are optimized, and the LSTM-SA-XGBoost stock prediction model is designed. In order to verify the prediction performance of the combined model, the combined model and the single model are respectively evaluated by the prediction model performance evaluation indicators: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Accuracy rate (Accuracy) and the f_1-score is analyzed and evaluated to verify that the two combined models proposed in this paper have good approximation ability and generalization ability in the prediction of stock fluctuations, and can improve the prediction performance of a single LSTM model or a single RNN model. The prediction performance of stock price rise and fall is verified, and the hybrid model based on LSTM neural network proposed in this paper has certain feasibility and stability in the prediction of stock price rise and fall.

Keywords: LSTM, XGBoost, simulated annealing algorithm, stock price forecast

1. INTRODUCTION

Stocks have been around for nearly 400 years. The modern economic situation has promoted the prosperity and development of the financial industry, securities industry and the stock market, and the number of shareholders has increased. Increased substantially. But the stock market is like a financial gamble. If an investor wants to win the gamble, in addition to possessing extraordinary courage and wisdom, as well as amazing perseverance and courage, he also needs a "tool" that can be strongly assisted in order to become the ultimate winner. The performance of the financial market is affected by various factors such as national policies, public opinion, and industry conditions. Under the influence of these factors, accurate stock forecasting has become the goal of people's continuous pursuit.

In 2017, David M. Q. Nelson used LSTM network in the stock prediction test to predict the future trend of stock price based on stock historical record data and stock technical analysis parameter indicators. The test results show that this method has good prediction effect[1]. Sreelekshmy Selvin and others used three different neural network models to predict the stock price of NSE listed companies, and compared the prediction performance of the three network models according to the prediction results, which further proved that LSTM model has certain advantages in time series prediction[2]. In the same year, Q Zhuge proposed a combined model of sentiment analysis model and long-term short-term memory (LSTM) time-series learning model to predict opening prices, and obtained excellent prediction performance[3]. In 2018, HK Choi used the ARIMA-LSTM hybrid model to test traditional predictive financial models, and the study found that the predictive ability of the ARIMA-LSTM model outperformed all other financial models to a considerable extent[4]. HY Kim proposed another hybrid model based on LSTM to predict the volatility of stock price. The experimental data show that the prediction error of GEW-LSTM hybrid model is the lowest[5]. In 2019, Guangyu Ding et al. proposed an associative deep recurrent neural network model with multiple inputs and multiple outputs based on long short-term memory network. The network model and the LSTM network model are analyzed, predicted and compared. When predicting multiple values at the same time, the accuracy of the correlation model is better than other models[6]. In 2020, the first mock exam was introduced into H Niu, combining variational mode decomposition (VMD) and long term short memory (LSTM) network. A new hybrid model VMD-LSTM was introduced. By comparing and analyzing the performance indices of some single models, EMD based models and other VMD based models, the VMD-LSTM prediction model's ability to predict stock price index was verified[7]. Y Liu proposed a regularized GRU-LSTM neural

[*] 656453927@qq.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125066X
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2663138

network model and applied it to the short-term forecast of two stock closing prices. The results show that the GRU-LSTM model outperforms the existing GRU and LSTM networks in stock time series forecasting model[8]. In 2021, Dilip Singh proposed a hierarchical LSTM with dense network layers that can effectively predict the closing price of certain stocks in NIFTY[9]. Although the experimental results show that combining neural network models will increase the "black box"[10]. However, according to the research of the above scholars, the combined algorithm based on LSTM and deep neural learning has a good effect in predicting stocks.

2. BASIC THEORY

2.1 LSTM

Long short term network, generally called LSTM, is a special type of RNN, which can learn long-term dependent information. LSTM was proposed by Hochreiter & schmidhuber (1997) and recently improved and popularized by Alex graves. In many problems, LSTM has achieved considerable success and has been widely used. There are input gate, forgetting gate and output gate in LSTM unit. LSTM controls the transmission state through the gating state, remembers the information that needs to be remembered for a long time and forgets the unimportant information[11].

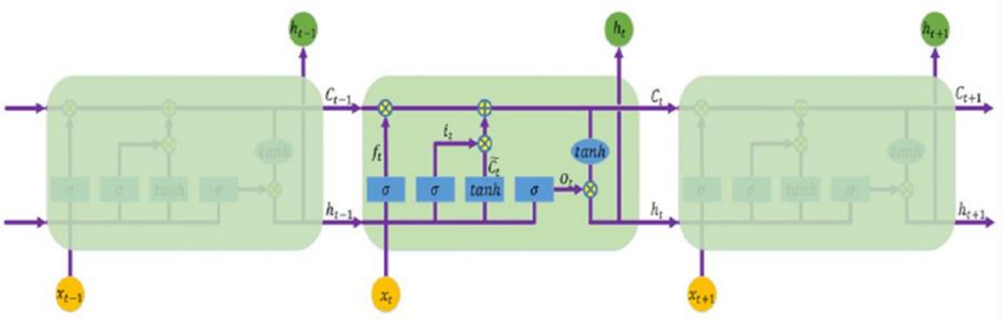

Figure 1. LSTM structure diagram.

2.2 XGBoost

The full name of XGBoost is eXtreme Gradient Boosting[12]. It is an optimized distributed gradient lifting library designed to be efficient, flexible and portable. XGBoost is a large-scale parallel boosting tree tool. It is the fastest and best open source boosting tree toolkit at present, which is more than 10 times faster than the common toolkit. In terms of data science, a large number of kaggle players choose XGBoost for data mining competition, which is a must kill weapon in major data science competitions; In terms of large-scale data in industry, the distributed version of XGBoost has wide portability and supports running in various distributed environments such as Kubernetes, Hadoop, SGE, MPI and Dask, so that it can well solve the problem of large-scale data in industry.

The mathematical expression of XGBoost prediction is shown in equation (1):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

Among them, K is the total number of trees, $f_k$is the k -th tree, and $\hat{y}_i$ is the prediction result of sample x_i .

The objective function is expressed as equation (2):

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

Where $l(y_i, \hat{y})$ is the training error of sample x_i , and $\Omega(f_k)$ is the regular term of the k -th tree.

2.3 Simulated annealing algorithm

Simulated annealing (SA) is not only a heuristic algorithm, but also a greedy algorithm, but its search process introduces random factors. When iteratively updating the feasible solution, accept a solution worse than the current solution with a certain probability, so it is possible to jump out of the local optimal solution and reach the global optimal solution[13]. The basic flow of the algorithm is shown in Figure 2.

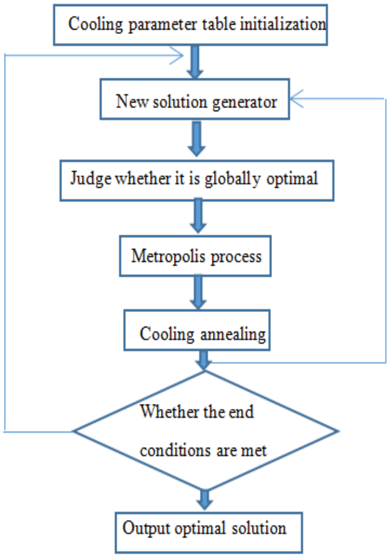
Cooling parameter table initialization
New solution generator
Judge whether it is globally optimal
Metropolis process
Cooling annealing
Whether the end conditions are met
Output optimal solution

Figure 2. Flow chart of simulated annealing algorithm.

3. LSTM-SA-XGBOOST MODEL CONSTRUCTION

Based on LSTM and XGBoost algorithm, this paper uses simulated annealing algorithm to optimize the parameters, and constructs LSTM-SA-XGBoost combined simulation algorithm to predict the rise and fall of stocks. First of all, download the data through Yahoo Finance and economics, sort out the data set, and process the missing values in the stock history data set to build the data set. Next, the "Open", "High", "Low", "Close", "Volume", "Adj Close", "Year", "Month" and "weekday" attributes in the stock historical data are analyzed by Pearson correlation coefficient, Spearman correlation coefficient and Kendall correlation coefficient respectively. According to the analysis situation, the attributes with relatively large correlation in the data set are retained, and the attribute data with small correlation or negative correlation is removed. Through calculation, it is known that the attributes that should be retained in this experiment are "Open", "High", "Low", "Close" and "Adj Close" data. Rebuild the data set again according to the retained attribute data, and divide the training set and test set. Then, the two-layer LSTM model is used to predict the five attributes of "Open", "High", "Low", "Close" and "Adj Close" in the training set. The two-layer LSTM neural network model is shown in Figure 3. Save the data results predicted by the two-layer LSTM model and construct a new test set. Since the selection of model parameters will directly affect the performance of model training, simulated annealing algorithm is used to optimize the original default parameters in the XGBoost algorithm. And use this model to complete the training and testing work. The flow of LSTM-SA-XGBoost model is shown in Figure 4.

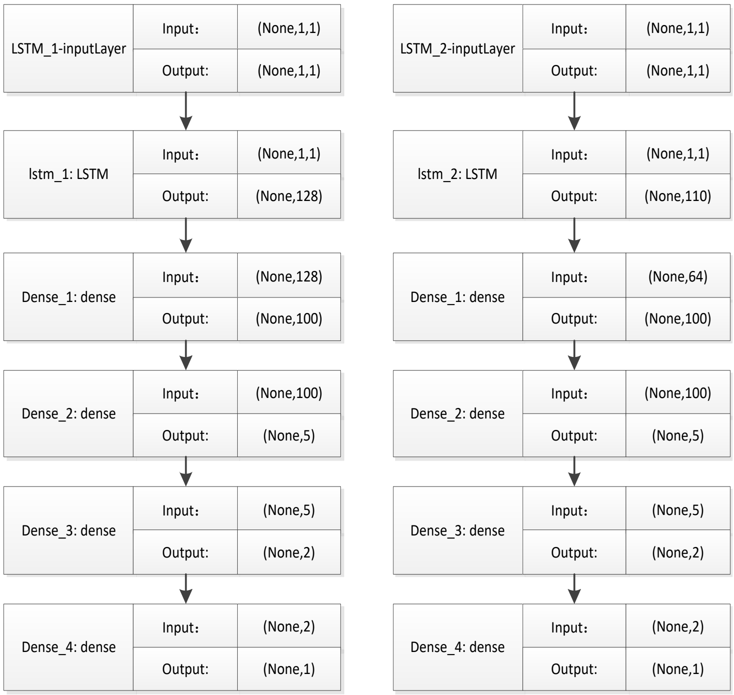

Figure 3. Construction of a two-layer LSTM model.

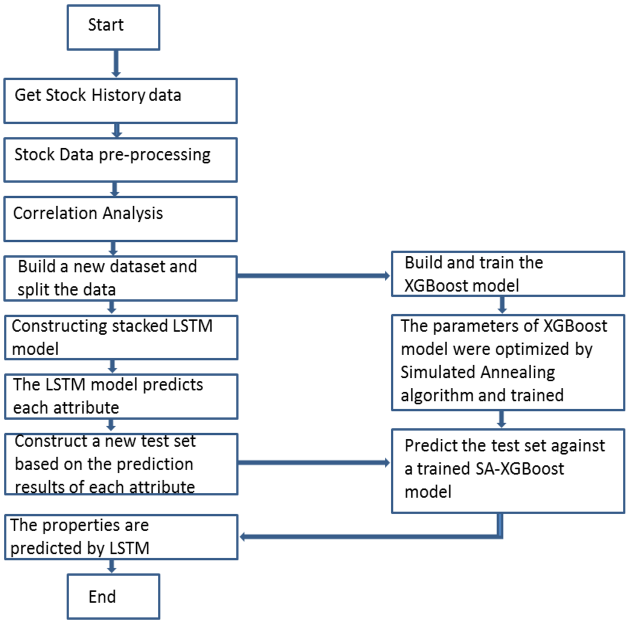
Start
Get Stock History data
Stock Data pre-processing
Correlation Analysis
Build a new dataset and split the data
Constructing stacked LSTM model
The LSTM model predicts each attribute
Construct a new test set based on the prediction results of each attribute
The properties are predicted by LSTM
End
Build and train the XGBoost model
The parameters of XGBoost model were optimized by Simulated Annealing algorithm and trained
Predict the test set against a trained SA-XGBoost model

Figure 4. LSTM-SA-XGBoost model building diagram.

The LSTM-SA-XGBoost model building process is as follows:

(1) Download the historical data of the stock index through https://finance.yahoo.com, preprocess the obtained data set, delete missing values, and construct the data set;

(2) Conduct correlation analysis on each attribute in the stock data set by means of heat map, remove the attribute data with small correlation or negative correlation, and construct a new data set.

(3) Apply the simulated annealing algorithm to the XGBoost model, complete the parameter optimization, and build the SA-XGBoost prediction model;

(4) Divide the data set. The first 80% of the data in the data set is set as the training set of the model, and the last 20% of the data is set as the test set of the model. And train the built SA-XGBoost network model, and save the training results;

(5) Use the LSTM method in the keras package to build a two-layer LSTM model, and train the attributes retained in the stock historical data set;

(6) Predict the attributes in the new stock training set, and reorganize the prediction results of each attribute to construct a new test set.

(7) Use the SA-XGBoost model saved in step (6) to complete the prediction work of the new test set;

(8) Judge the performance of the LSTM-SA-XGBoost model in the prediction of stock fluctuations.

4. EXPERIMENT AND ANALYSIS

The data history dataset deadline was updated to January 31, 2021.In this experiment, a total of 4782 stock historical attribute data of "AAPL" stock from February 1, 2002 to January 31, 2021 were randomly selected as the basis for constructing the financial time series data set of this experiment. Among them, the first 80% of the historical data of the "AAPL" stock is set as the training set of the model, and the last 20% of the data is set as the test set of the model.

4.1 Experimental evaluation index

In this experiment, the following indicators will be used to evaluate the performance of the prediction model: RMSE (Root Mean Square Error), MAE (Mean Absolute Error), Accuracy and f1-score, as shown in Formulas 3-6:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2} \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y^{(i)} - \hat{y}^{(i)}\right| \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$f1\_score = \frac{2 * pre * recall}{pre + recall} \tag{6}$$

Where y is the actual value, ŷ is the predicted value. TP is the number of correct predictions of stock price rise, FP is the number of wrong predictions of stock price rise. TN is the number of correct predictions for the stock price to fall, and FN is the number of errors to predict the decline of the stock price. Pre is the balanced precision, and recall is the recall rate. In the f1_score calculation result, 1 is the best value and 0 is the worst value.

4.2 Comparative analysis of experimental data

Pearson correlation coefficient is a kind of linear correlation coefficient used to reflect the statistics of the linear correlation degree of two variables. The value range of the Pearson correlation coefficient is [-1, +1]. If the calculation result of r is greater than 0, it indicates that the two variables are positively correlated. On the contrary, if the calculation result of r is less than 0, it indicates that the two variables are positively correlated. Variables were negatively correlated. Its mathematical expression is as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{7}$$

Among them, n represents the sample size, xi, x are the observed value and mean of the two variables, respectively.

Spearman correlation coefficient is a nonparametric index to measure the dependence of two variables. It uses monotone equation to evaluate the correlation of two statistical variables. Its mathematical expression is as follows:

$$\beta_s = 1 - \frac{6\sum {d_i}^2}{n(n^2-1)} \tag{8}$$

where n is the number of samples and d represents the rank difference between the data x and y.

Kendall correlation coefficient is a kind of rank correlation coefficient. Its calculated object is classified variables. It is used to measure the statistical value of the correlation between two random variables and to test the statistical dependence of two random variables. Its value range is [- 1, + 1]. Its mathematical expression is as follows:

$$\mathrm{R} = \frac{\mathrm{P}-(\mathrm{n}*(\mathrm{n}-1)/2-\mathrm{P})}{\mathrm{n}*(\mathrm{n}-1)/2} = \frac{4\mathrm{P}}{\mathrm{n}*(\mathrm{n}-1)} - 1 \tag{9}$$

Among them, P is the logarithm of the statistical objects with the same arrangement and magnitude of the two attribute values.

Firstly, the attributes in the stock are analyzed by Pearson correlation coefficient, Spearman correlation coefficient and Kendall correlation coefficient respectively. Through the analysis results, the attribute data with small or negative correlation with the rise and fall of the stock will be removed. According to the formula definitions of the three correlation coefficients, the data results of the correlation coefficient are calculated and analyzed. Table 1 shows the correlation coefficient values of stock fluctuations and other attributes, and the heat map of the three correlation coefficients of each attribute is shown in Figures 5~7.

Table 1. The correlation coefficient between the rise and fall of stock and each attribute.

	Open	**High**	**Low**	**Close**	**Volume**	**Adj Close**	**Year**	**Month**	**weekday**
Pearson	0.006	0.0161	0.018	0.0277	-0.03033	0.028169	0.0271	-0.0181	-0.048348
Spearman	0.045	0.0508	0.051	0.0567	-0.07210	0.056970	0.0558	-0.0085	-0.071651
Kendall	0.031	0.0355	0.036	0.0400	-0.04697	0.040146	0.0392	-0.0058	-0.052405

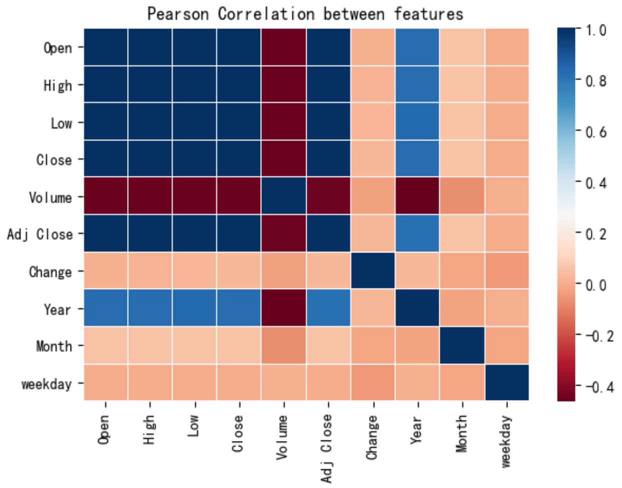

Figure 5. Pearson correlation coefficient property correlation thermograph.

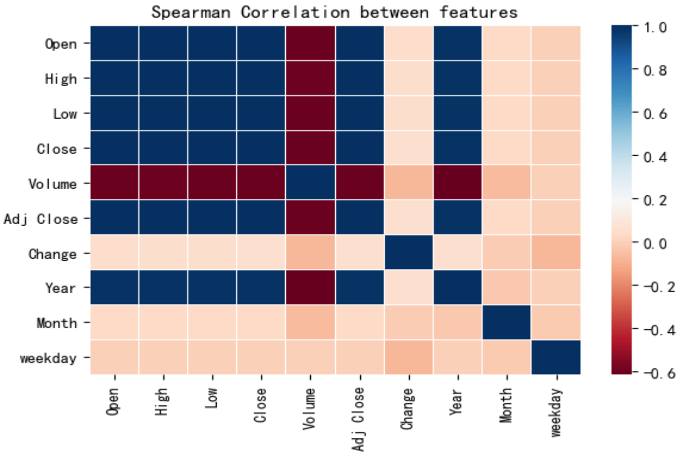

Figure 6. Spearman correlation coefficient property correlation thermograph.

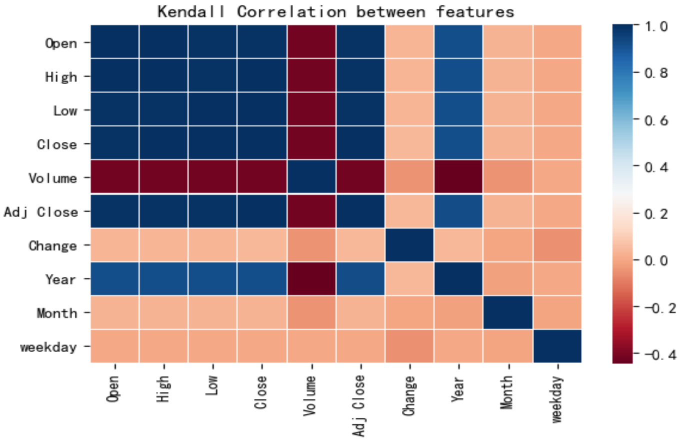

Figure 7. Kendall correlation coefficient property correlation thermograph.

According to Table 1 and Figures 5-7, it can be clearly seen that the three attributes of "Volume", "Month" and "Weekday" in "AAPL" stock are negatively correlated with the prediction results of stock rise and fall, which will affect its training effect in the process of model training. Therefore, remove the above three attributes and build a new data set for model training and prediction.

After determining the attribute data of the stock, reorganize the data set, and train and predict the "Open", "High", "Low", "Close" and "Adj Close" attributes through the two-layer LSTM. In order to better verify the advantages and disadvantages of the model, this paper sets the sliding window length of each attribute to 1, completes the model training and prediction by one-step prediction, and sets epochs to 10, batch_ Set the size to 64. The results of predicting the price change trend of "AAPL" stock by using the two-layer LSTM network model are shown in Figure 8.

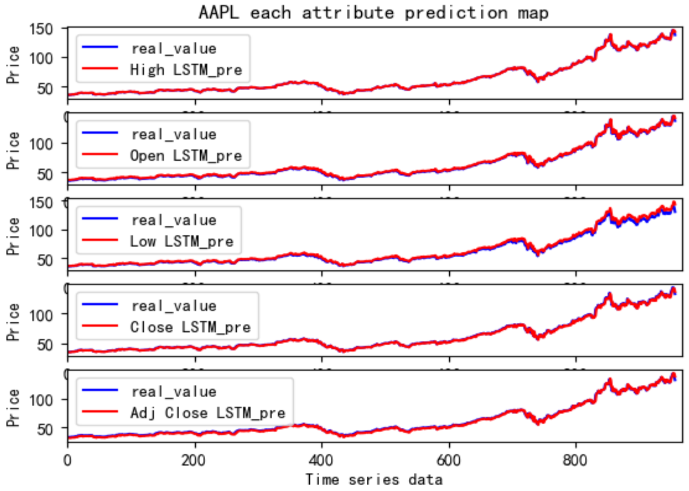

Figure 8. Forecast comparison chart of each attribute of "AAPL" stocks.

It can be clearly seen from Figure 8 that the two-layer LSTM model adopted in this paper has good performance and high fitting degree in "AAPL" stock prediction. After the prediction of LSTM model, its predicted value is constructed into a new test set. Next, the parameters of XGBoost model are optimized by simulated annealing algorithm, and SA-XGBoost model is constructed. Under the condition of 50 iterations of simulated annealing algorithm, the optimal solutions of six parameters of XGBoost model, namely "n_estimators", "max_depth", "learning_rate", "min_child_weight", "reg_alpha" and "min_child_sample", are found. The training data set of SA-XGBoost model adopts the first 80% of the attribute data set. After the training of SA-XGBoost model, the model is saved, and the new test set constructed by using two-layer LSTM prediction is predicted to obtain the final prediction results of stock time series rise and fall. In the first mock exam, we compare the prediction performance of the hybrid model with the single model, and train the four neural network prediction models of LSTM-SA-XGBoost, LSTM-XGBoost, LSTM and RNN models for the 20 time, and calculate the evaluation indexes RMSE, MAE, Accuracy and f1_score. The mean value of score is compared and analyzed. The experimental comparison data are shown in Table 2.

Table 2. Comparison results of 20-time mean values predicted by "AAPL" model.

Index	**LSTM-SA-XGB**	**LSTM-XGB**[14]	**LSTM**[15]	**RNN**[16]
RMSE	1.255	1.289	2.135	2.612
MAE	0.693	0.718	1.146	1.368
Accuracy	0.550	0.533	0.460	0.462
*f*1_score	0.698	0.693	0.002	0.0

Table 2 shows the average value of prediction performance indicators of LSTM-SA-XGBoost, LSTM-XGBoost, LSTM and RNN after training for 20 times. It can be clearly seen that LSTM-SA-XGBoost model shows the best prediction performance and stability in predicting the rise and fall of stock time series compared with other models. Among them, in the accuracy evaluation index, the model is 2% higher than LSTM-XGBoost model, 9% higher than LSTM model and 8.8% higher than single RNN prediction model. It has been significantly improved in the prediction accuracy index, such as RMSE, MAE and f1_score .The score evaluation index has also been improved to varying degrees, which further verifies that the LSTM-SA-XGBoost hybrid model proposed in this paper has good stability and prediction accuracy in stock rise and fall prediction, and the prediction performance has been improved to a certain extent.

5. CONCLUSION

Based on the LSTM-XGBoost stock price prediction model, this paper proposes a correlation analysis-based LSTM and simulated annealing optimized XGBoost hybrid model (LSTM-SA-XGBoost) to improve the performance of financial time series price prediction. First, download and organize the data set through the Yahoo Finance website, and process the missing values in the stock historical data set to construct the data set. Next, each attribute data of the stock is analyzed by three correlation coefficients. Retain the strong correlation attribute, rebuild the data set again, and divide the training set and test set. Then, the two-layer LSTM model is used to predict the training set, and the predicted data results are saved to construct a new test set. Since the selection of model parameters will directly affect the performance of model training, simulated annealing algorithm is used to optimize the default parameters in the XGBoost model. And the optimized XGBoost model is used for training, and finally a new test set is used for testing. From the experimental results, it can be known that the LSTM-SA-XGBoost hybrid model proposed in this chapter has good stability and prediction accuracy in the prediction of stock fluctuations. , the prediction performance has been improved to a certain extent.

ACKNOWLEDGEMENT

The paper is funded by Dongguan social science and technology development project (No.2020507154645) and Special project for key fields of colleges and Universities in Guangdong Province (new generation information technology): Research on cloud Storage mechanism based on blockchain technology; Cultivation Plan for Major Scientific Research Achievements of Universities: Research on the Blockchain Fragmentation Mechanism Based on Merkel Tree (GKY-2019CQYJ-13).

REFERENCES

[1] Nelson, D., Pereira, A. and Oliveira, R., “Stock market's price movement prediction with LSTM neural networks,” 2017 International Joint Conference on Neural Networks (IJCNN), IEEE, 1419-1426(2017).

[2] Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., et al., “Stock price prediction using LSTM, RNN and CNN-sliding window model,” International Conference on Advances in Computing, IEEE, 1643-1647(2017).

[3] Zhuge, Q., Xu, L. and Zhang, G., “LSTM neural network with emotional analysis for prediction of stock price,” Engineering Letters, 25(2), 167-175(2017).

[4] Choi, H. K., “Stock Price Correlation Coefficient Prediction with ARIMA-LSTM Hybrid Model,” Papers, 1808.01560(2017).

[5] Kim, H. Y. and Won, C. H., “Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models,” Expert Systems with Applications, 103, 25-37(2018).

[6] Ding, G. and Qin L., “Study on the prediction of stock price based on the associated network model of LSTM,” International Journal of Machine Learning and Cybernetics, 11(6), 1307-1317(2019).

[7] Niu, H., Xu, K. and Wang, W., “A hybrid stock price index forecasting model based on variational mode decomposition and LSTM network,” Applied Intelligence, 50(12), 4296-4309(2020).

[8] Liu, Y., Wang, Z. and Zheng, B., “Application of Regularized GRU-LSTM Model in Stock Price Prediction,” IEEE 5th International Conference on Computer and Communications (ICCC), IEEE, 1886-1890(2020).

[9] Singh, D. and Gupta, B. K., “Closing Price Prediction of Nifty Stock Using LSTM with Dense Network,” International Conference on Advances in Distributed Computing and Machine Learning, 382-392(2021).

[10] Wei, D. W. and Cui Z. W., “Research on software reliability prediction based on neural network integration,” Computer engineering and design, 4228-4231(2010).

[11] Hochreiter, S. and Schmidhuber, J., “Long Short-Term Memory,” Neural Computation, 9(8), 1735-1780(1997).

[12] Chen, T. and Guestrin, C., XGBoost: A Scalable Tree Boosting System,” Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, 785-794(2016).

[13] Ji, O., Meng, F., Zheng, H., et al., “Optimization and Integration of Logistics Facilities Resources Based on Genetic-Simulated Annealing Hybrid Algorithm,” 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), 135-142(2021).

[14] Zhang, X. and Zhang, Q., “Short-Term Traffic Flow Prediction Based on LSTM-XGBoost Combination Model,” Computer Modeling in Engineering and Sciences, 125(1), 95-109(2020).

[15] Baek, Y. and Kim, H. Y., “ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module,” Expert Systems with Applications, 113(DEC.), 457-480(2018).

[16] Ass, A. and Ss, B., “Analysis of look back period for stock price prediction with RNN variants: A case study on banking sector of NEPSE,” Procedia Computer Science, 167, 788-798(2020).