2021 IEEE 3rd International Conference on Communications, Information System and Computer Engineering (CISCE 2021)

Prediction Model of Dow Jones Index Based on LSTM-Adaboost

RunKai Ying
School of Software, Faculty of Information Technology
Beijing University of Technology
Beijing China
Email: duoerxs@163.com

Yuntao Shou
School of computer and Information Engineering
Central South University of Forestry and Technology
Hunan China

Chang Liu
School of Computer Science and Technology
Huaqiao university
Xiamen China

***Abstract*—How to accurately predict the Dow Jones index is an important issue in quantitative finance. This paper designs an LSTM-AdaBoost(Long Short-Term Memory and Adaptive Boosting) prediction model based on the Dow Jones index's characteristics. The Adaboost algorithm was used to integrate multiple LSTM weak classifiers. By iteratively updating the weights of each version and combining the strategy, the improved robust classifier was formed, and the prediction for the Dow Jones index was finally obtained. Experimental results have shown that the prediction results obtained by the LSTM-AdaBoost model have been significantly higher than those of the traditional classification model, with an average increase of 43.76 percent in R_square. The establishment of this model provides a theoretical basis for predicting the Dow Jones index and provides ideas for the intersection of finance and computer industries.**

Keywords-Dow Jones Index Forecast, Quantified financial sector, Long and short memory neural network, Adaptation boosting

I. INTRODUCTION

The prediction of stock market price was always the focus of people's research, but its inherent noise environment and the enormous volatility relative to the market trend had many influencing factors. Therefore, the research process of the stock price was very complicated. In the past, stock market forecasting technology could be divided into prediction-based technology and clustering-based technology[1]. At present, the research on stock price prediction includes short and short-term memory neural networks (LSTM), time-series model, multiple linear regression model, and so on.

In the current literature, By establishing an ARIMA model, Yan Yu[2] analyzed the historical data of the closing price of the Nasdaq index stock and predicted the stock's closing price in the next seven days. However, as the stock price (index) had nonlinear fluctuation, the stock price (index) was unstable and often fluctuated because of the influence of policy and news, so that it was difficult for the time series model to handle. Li Xiaoning[3] predicted daily stock data by establishing multiple linear regression models. Although the change of multiple independent variables is considered, the calculation process is very troublesome due to the large number of independent variables.The statistical software was generally used in practical applications, but there was still a significant error between the forecast and the stock price's actual value. Huang Chaobin[4] established an LSTM model to predict the Shanghai Composite Index. Although the LSTM neural network model was easier to fit the complex nonlinear relationship, there was a lag problem in predicting stock prices. There was still a lot for optimization in the face of a more complex stock market price forecasting problem.

Therefore, because of the highly nonlinear stock price prediction problem, this paper has proposed a machine learning method based on the Dow Jones index stock data of the United States to predict the future price and then improved the LSTM model. The Adaboost model has been used to improve the parameter adjustment model, which has improved the prediction effect and has made up for the shortage of single application LSTM networks.

II. DATA PREPROCESSING

A. Data collection

This article selects the Dow Jones Index data (downloaded from the YingWei Financial Information) from January 2, 2019, to February 26, 2021 (544 data in total), including Real-time data such as closing price (close), opening price (open), highest price (high), lowest price (low), trading volume (vol), fluctuation (pct_chg), etc.

B. Data processing

Data preprocessing: because there was a lack of value in the original data set, the downloaded data and interpolation were inserted with the mean value of two adjacent days to obtain a complete data set.

Data standardization[1]: Since the magnitude of the data in the data set is not the same, for example, there are huge differences between the opening price, closing price, and transaction volume, transaction volume. To eliminate the influence of different magnitudes between data and unify the data of different magnitudes into the same magnitude, this model standardized these data, which is conducive to improving the training speed and prediction accuracy of the model. The expression is shown in formula (1):

$$x' = \frac{x - \mu}{\sigma} \quad (1)$$

978-0-7381-1215-2/21/$31.00 ©2021 IEEE

2021 International Conference on Communications, Information System and Computer Engineering (CISCE) | 978-1-6654-0352-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/CISCE52179.2021.9445928

The processed data are divided into the training set and test set in order, in which the first 70% data is used as the training set to train the model, and the last 30% data is used as the test set to evaluate the prediction effect of the model.

III. ESTABLISHMENT OF MODELS

A. Introduction to basic concepts/selection of eigenvalues

In order to make the model more concise and easier to understand, introduction of the models and selections of the eigenvalue is necessary.

1) LSTM Model Introduction and Eigenvalue Selection

LSTM is an improved RNN (Circular Neural Network), a particular form of RNN that provides feedback for each neuron[5]. The output of the RNN depends not only on the current neuron input and weight but also on the previous neuron input. Therefore, theoretically, RNN structures are suitable for processing time-series data. However, when processing a long string of data samples, the problem of exploding and vanishing gradients arises will emerge [6] so that this is a point of entry for LSTM models[7].

To deal with the problem of vanishing gradient of RNN model, the method of penetrating element is put forward, that is, adding jump connection on the time axis and then popularizing it into LSTM. The gate structure in LSTM provides a function for gradient selection. The gradient under this path can be kept constant from the last moment through the control of the gate so that it can improvise the vanishing gradient problem.

LSTM can delete or add information to the state of the cell, which is conferred by gate structures. LSTM cells consist of input gate, forgetting gate, output gate, and unit state. Input gate, forget gate, output gate is used to update, maintain and delete control information.

2) AdaBoost Method Brief Introduction and Eigenvalue Selection

AdaBoost is an integrated learning method proposed by Freund and others to solve the two-class problem[6]. The core idea is to train different weak classifiers for the same training set and then set these weak classifiers together to form a strong classifier. Adaboost can deal with classification and regression problems. In this paper, the original data is divided into two categories, a threshold is set, the classification beyond this threshold is 1, and the one not exceeding is set to 0, so that the weak classification is carried out, and update the weights of each version by iteration.

B. Dow Jones Index Forecast Method Based on Improved LSTM-AdaBoost

By setting a threshold, the training sample is set to 0 and 1 weak classification set, and then the improved LSTM-AdaBoost strong classifier is obtained by iterative updating the weight and combining the strategy. The overall framework of the improved method is shown in figure 1.

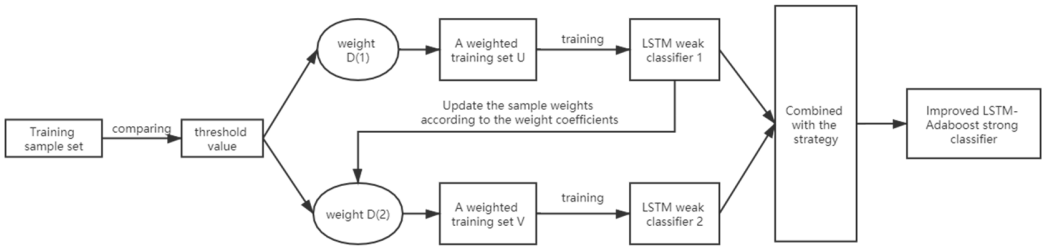

Figure 1. Improved LSTM-AdaBoost Framework

Suppose the training sample is

$$\mathrm{T} = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\} \tag{2}$$

The output weight of the k weak learner's training set is, k is the number of learners, k=2.

$$D(k) = (w_{k1}, w_{k2}, w_{k3}, \dots, w_{km}); w_{1i} = \frac{1}{m}; \quad i = 1,2,3, \dots, m \tag{3}$$

The steps are as follows:

(1) Setting Initial Conditions

Set the initial condition: set the number of iterations as T, T means the frequency of iterations as LSTM, the initialization distribution weight is 0, the input gate nodes are 6, the hidden gate nodes are 12, and the output gate node is 1.

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

The upper formula represents the calculation of the input gate, in which W_i is the weight matrix of the input gate, and b_i is the bias term of the input gate.

$$\mathrm{f}_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) d_n \tag{5}$$

$$[W_f]\begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} = [W_{fh} \quad W_{fx}][\begin{matrix} h_{t-1} \\ x_t \end{matrix}] = W_{fh} h_{t-1} + W_{fx}. \tag{6}$$

The upper formula represents the calculation of the forgetting gate, in which W_f is the weight matrix of the forgetting gate, $[h_{t-1}, x_t]$ is the connection of the two vectors into a longer vector, and b_f is the bias term of the forgetting

gate. The weight matrix W_f is composed of two matrices W_{fh} and W_{fx} splicing.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (7)$$

The upper formula represents the calculation of the output door, the same as the forgotten door.

(2) Set the improved LSTM structure

Define the threshold, compare the training data set with the threshold to carry on the weak classification, then get the LSTM-AdaBoost strong classifier through the weight iteration and the combination strategy.

(3) Calculation of Learning Error Rate of Weak Classifier

Because this model is essentially a binary classification problem, the output category is {-1,1}. So the learning error rate of the k weak classifier Gk (x) on the training set is

$$e_k = P(G_k(x_i) = y_i) = \sum_{i=1}^{m} w_{ki} I(G_k(x_i) /= y_i) \quad (8)$$

(4) Weight update

The weight coefficient Gk (x) of the k weak classifier is

$$\alpha_k = \frac{1}{2}\log\frac{1-e_k}{e_k} \quad (9)$$

(5) Cycle LSTM training until the error rate is less than the set error value and the iteration ends to obtain a strong classifier.

IV. SOLUTION OF MODELS

To compare the effect of different model prediction algorithms, the prediction accuracy of LSTM-AdaBoost, LSTM, ARIMA 3 improved methods is calculated. The single LSTM method refers to the research of Huang Chaobin [4]. The paper standardized data processing and used the "multi-layer network method"[10] to determine the number of hidden nodes and regularization coefficients of neural networks. The ARIMA method refers to the research of Yan Yu[2]. By AIC, the minimum criterion[9], the first-order difference of the initial sequence is made, and finally, determine the appropriate order and finally determine the specific model.

A. Single LSTM forecast

The data of different orders of magnitude are unified into the same order of magnitude by data standardization. The processed data are divided into the training set and test set in order, in which the first 70% of the data is used as the training set to train the model. The last 30% of the data is used as a test set to evaluate the prediction effect of the model. The projections are shown in Figure 2.

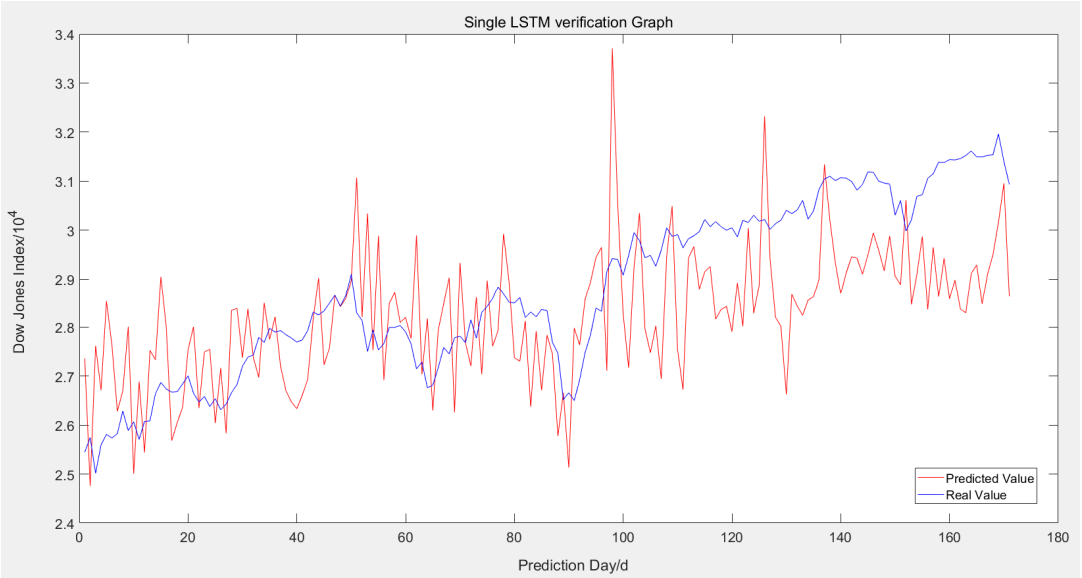

Figure 2. Single LSTM forecast

B. Single ARIMA projections

The exponential time series are smoothed by splitting the data, and the ARIMA(2,1,2), ARIMA(3,1,2), ARIMA(4,1,2) model is established by AIC as the minimum criterion[5]Determine the most appropriate order. Finally, the model ARIMA(2,1,2) is determined as the final model. The final projections are shown in Figure 3.

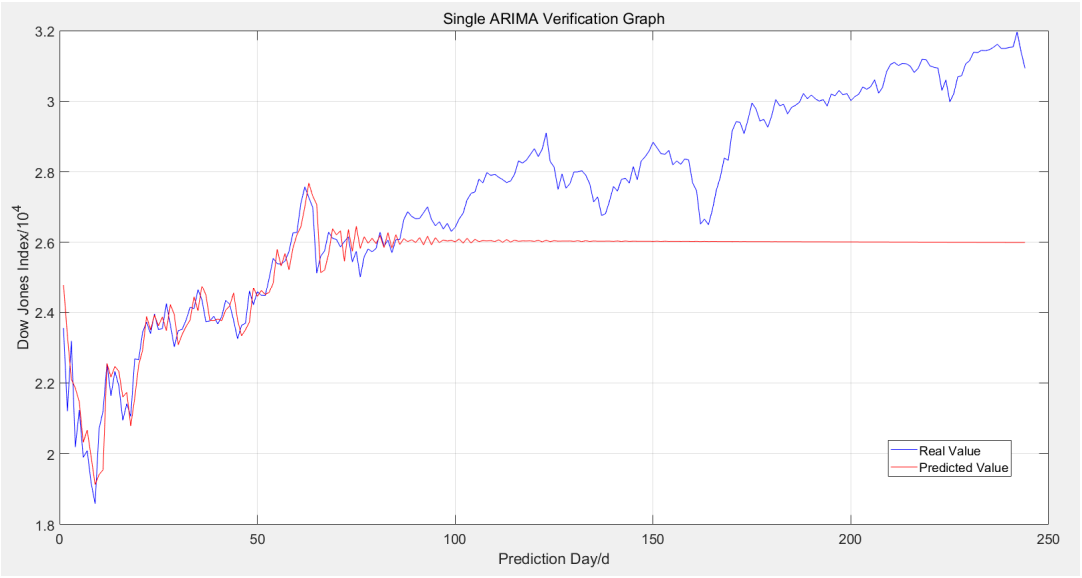

Figure 3. Single ARIMA projections

C. *Dow Jones Index Forecast Model Based on Improved LSTM-AdaBoost*

The AdaBoost algorithm is used to integrate multiple LSTM and weak classifiers, and a threshold is set. The classification beyond this threshold is 1, and the one not exceeding is 0, so that the weak classification is carried out. Through iterative updating of the weights of each version, combined with the strategy to form an improved LSTM-AdaBoost strong classifier, cycle back and forth for LSTM training until reaching the setting error threshold this way to improve the accuracy of exponential prediction. Finally, the prediction results are shown in Figure 4 the red line represents the predicted value, and the blue line represents the actual value. According to the data trend in Figure IV, the Dow Jones index prediction model's accuracy based on the improved LSTM-AdaBoost is better.

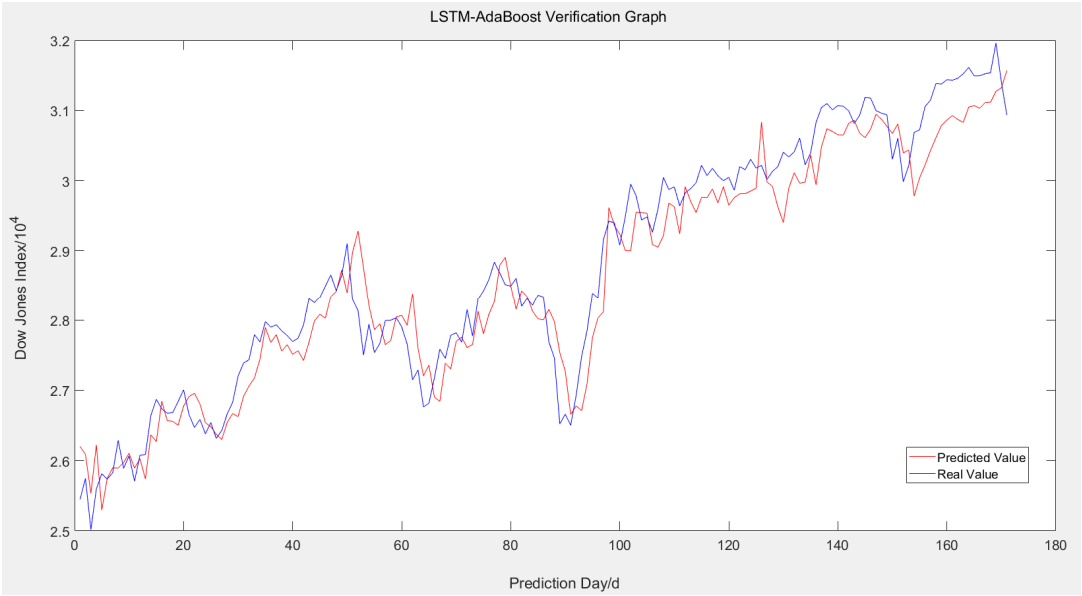

Figure 4. Improved LSTM-AdaBoost forecasting

D. *Comparison of experimental results*

The accuracy of the three methods is shown in Table 1.

Table 1. RESULTS OF LSTM-ADABOOST ALGORITHM COMPARED WITH LSTM AND ARIMA

	MSE	RMSE	R^2	MAE
LSTM	0.0118	0.1058	0.6142	0.0948
ARIMA	0.1047	0.3236	0	0.2753
Adaboost-LSTM	0.0036	0.0598	0.8830	0.0464

From the table1, it can be seen that both method one and method three have better prediction results for the model, while method two has poor prediction results, and the prediction results of a single LSTM method fluctuate significantly, which is quite different from the data at some times. A single ARIMA method is more accurate in the early prediction, but with time, the variable fluctuates less, and the prediction index tends to smooth. LSTM, ARIMA comparison LSTM-AdaBoost the improved methods, it can be found that the improved LSTM-AdaBoost methods are promoted in the accuracy of model prediction, and the effect of the improved methods on the data set is better than that of the other two methods.

V. CONCLUSION

Because of the Dow Jones index's prediction problem, this paper designs the LSTM-AdaBoost prediction model fully combines the advantages of the two algorithms and uses the Adaboost algorithm to integrate multiple LSTM weak classifiers. Iteratively updating the weights of each version, combined with strategies to form an improved LSTM-AdaBoost strong classifier, and finally, get the prediction for the Dow Jones Index. The accuracy of the model proposed in this paper is high through experimental verification. The mean square error and the determination coefficient are improved significantly, and the R_square is already 0.8830. In future work, through the vast Dow Jones index data, the existing prediction model will be further optimized to make the error between the prediction data and the actual data smaller.

REFERENCES

[1] G.Randhmaldp and K.Umark,"Systematic analysis and review of stock market prediction techniques," Computer Science Review ,Mar.2019,pp:100-190.

[2] Yan Yu and Wu Haitao, "Short-term forecast of NASDAQ index based on ARIMA model Information and Computer ,"Theoretical Edition,Mar.2020,pp:155-158.

[3] Li Xiaoning.,"Application of Multivariate Linear Regression and Time Series Model in Stock Prediction," Science and Technology Entrepreneurship monthly ,Mar.2019,pp:153-155.

[4] Huang Chaobin and Cheng Ximing,"Study on Stock Price Prediction Based on LSTM Neural Network of China Journal of Beijing University of Information Technology," Natural Science Edition,Mar.2021,pp:79-83.

[5] Liao Jinping ,Mo Yuchang and Yan Ke, "Short-term electricity forecasting model and application based on C-LSTM Journal of Shandong University," Mar.2021,Engineering Edition,pp:1-8.

[6] Jozefowicz R and Zaremba W ,Sutskever I ."An empirical exploration of recurrent network architectures,"Mar.2015.

[7] Hochreiter S and Schmidhuber J."Long short-term momory ,"Mar.1997, Neural Computer

[8] Freund Y and Schapire R E."Game theory,on-line prediction and boosting,"Mar.1996 Proceedings of the ninth annual conference on computational learning theory,Desenzano del Garda,pp:325-332.

[9] Sun Xianqiang,"Application of Time Series Model in Stock Price Prediction," unpubilshed.

[10] Li Da, Zhang Zhaosheng, Liu Peng, Wang Zhenpo, and Dong Haotian. "Classification of multi-weather vehicles based on improved long-term memory neural network-adaptive enhancement algorithm,"Apr.2020,Automotive Engineering ,pp:1248-1255.