A High frequency stock price prediction model based on Boosting and information entropy weighted LSTM neural network

Jiaqi Liu[1,#], Zihan Zhang[1,#], Zhirong Xue[2,#,*]

[1]School of Finance, Shandong University of Finance and Economics, Jinan, Shandong, 250220, China

[2]School of Business Administration, China University of Petroleum-Beijing at Karamay, Karamay, Xinjiang, 834000, China

*Corresponding author:791293316@qq.com

#These authors contributed equally.

Abstract. With the continuous development of data acquisition technology, high-frequency time series forecasting, such as stock prices, has become a hot issue. This paper proposes a LSTM neural network based on Boosting and information entropy weighting. The proposed method extracts the functional features of high-frequency time series by using orthogonal polynomial expansion, and uses the Boosting frame to recursively fit the residual predicted by LSTM neural network. Considering that the dimension of the prediction variable of LSTM neural network is a super parameter, we propose a model average method based on information entropy weighting, which theoretically balances the variance and bias of the prediction model. In addition, the basic model in the Boosting framework can be arbitrary, which greatly expands the applicability of the proposed method. The results of real data analyses show that the proposed method can effectively improve the prediction accuracy of the original LSTM neural network and has robustness. Finally, the proposed method can be further applied to the real-time monitoring of trace elements, the monitoring of road traffic flow and the prediction of daily average temperature curve in environmental science.

Keywords: High frequency stock price forecast, Boosting, LSTM neural network, model average.

1. Introduction

High frequency stock price data refers to the real-time price sampling buying and selling data of stocks between the opening time and the closing time. It is mainly a time series displayed in chronological order with sampling frequency of hours, minutes, or even seconds. From a micro perspective, in an effective market, stock prices can truly reflect the supply and demand relationship in the stock market. Forecasting high-frequency stock prices can not only help listed companies adjust their stock liquidity, but also affect their capital structure [1]. It also helps investors to choose the right investment portfolio; From a macro perspective, the stock price index, which is composed of the stock prices of representative listed companies, is a barometer of a country's economic development and will, to a certain extent, affect the decision makers' formulation of economic policies. High frequency time series have the characteristics of high dimension and large noise, so it is difficult to fully capture the linear information, nonlinear information and function information, and the prediction accuracy based on probability is low. The analysis of financial time series has gradually become a hot spot in the field of quantitative research, which is also the focus of this paper.

Traditional time series prediction methods mainly include autoregressive moving average model (ARMA), autoregressive conditional heteroscedasticity model (ARCH) and its extensions ARIMA, GARCH, etc. ARMA was first proposed by American economist G. P. Box and British economist G. M. Jenkins (1978) [2]. ARMA model is a combination of autoregressive model (AR) and moving average model (MA), which is an important method for predicting stationary time series; Robert F. Engle (1982) [3], an American economist, first proposed the ARCH model, which provides a new method to solve the inflation volatility in Britain. However, the traditional time series prediction method is prone to the problem of dimension disaster when predicting high-frequency data, which reduces the prediction accuracy. For the dimensionality reduction of high-frequency data, the current popular methods are principal component analysis (PCA) and functional principal component analysis (KPCA). Wang Dongmei et al. (2022) [4] explored the problem of force prediction during human running by building a PCA and wavelet neural network model; Cui Jing (2021) [5] used KPCA to reduce the dimensions of high-frequency stock price data, and then used traditional time series forecasting methods for research. PCA and KPCA methods have alleviated the dimension disaster to some extent, but the potential model structure of the reduced dimension variable is unknown to the response variable, and the function information is not fully extracted, which will have a negative impact on the subsequent prediction. In recent years, machine learning algorithm has become a popular tool in high-frequency data prediction. Xu Mei et al. (2015) [6] used the error back propagation neural network (BP) to predict the high-frequency data of the Shanghai Composite Index; Li Liping et al. (2022) [7] used the long and short term memory artificial neural network (LSTM) and BP neural network to predict the closing price of Yunnan tourism stock, the conclusion showed that LSTM neural network is more effective than BP neural network is drawn. The time series prediction using LSTM, BP and other deep learning technologies does not need to meet the stationarity assumption, and does not need to consider the dimension problem, but its super parameters need to be estimated manually, and the results of different parameters vary greatly.

In conclusion, this paper proposed a LSTM neural network based on Boosting and information entropy weighting. The innovation of the proposed method mainly includes the following aspects: First, based on Chebyshev polynomial expansion, the nonlinear information and function information were extracted at the same time of dimension reduction. The other was to use Boosting method to continuously capture the potential model results between the reduced dimension feature vector and the residual sequence. Thirdly, the model average method was used to solve the problem of dimension selection of LSTM neural network prediction variables. The experimental results showed that the proposed method is competitive with the original LSTM neural network.

The rest of this paper is as follows: The second section will introduce LSTM neural network based on Boosting and information entropy weighting, the third section will show the actual data analysis results, and the fourth section is a summary of the full text.

2. Theory and Method

2.1 Chebyshev orthogonal polynomial expansion

The basis expansion method is to discretize the differential operator directly in the unstructured grid. After generating the grid, the basis function expansion is used to approximate the real function on the grid elements. The overall basis function approximation can be regarded as the combination of the element basis functions. If you take a linear function:

$$f(x) = \sum_{m=1} \beta_m T_m(x) \tag{1}$$

this form is called a linear basis expansion on x, $T_m(x)$ can be called a basis function.Chebyshev polynomials are recursively defined sequences of polynomials related to de Meyver's theorem. Chebyshev polynomials are algebraic polynomials of degree n, defined as $T_m(x) = cos(nx)$, where $-1 \leqslant x \leqslant 1$. The Chebyshev polynomial expansion method is used in our method.

2.2 Boosting method

Any other first algorithm based on Boosting framework, by Boosting framework to the operation of the training sample set, get a different training sample subset of the base classifier is generated by the training sample subset to, each get a sample set is generated by the algorithm in the sample set a base classifier, so after a given training round number n, Then the Boosting framework algorithm takes the weighted average of these n base classifiers to produce a final prediction result, which improves the accuracy of the algorithm. These basic algorithms are generally unstable weak classification algorithms, such as BP neural network , as shown in Figure 1.

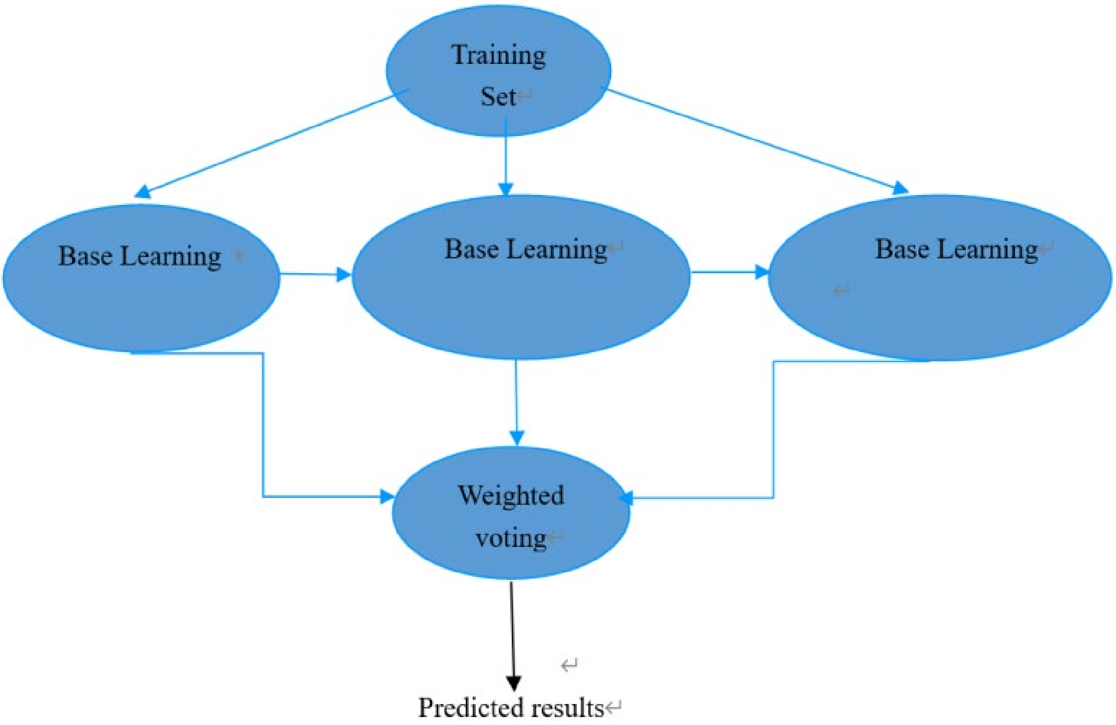

Figure 1 Boosting framework

2.3 LSTM neural network

LSTM mainly introduces gating mechanism to control the information accumulation speed, including selectively adding new information and selectively forgetting the previously accumulated information, so as to improve the long-term dependence problem of recurrent neural network (RNN) and alleviate the gradient disappearance problem in the process of long sequence training. After data X_t and H_{t-1} enter the LSTM neural network, they first go through the forgetting gate to determine what information can pass through. The second step is to generate the new information we need. This step consists of two parts. The computation of cells in the LSTM layer can be expressed as the following grouping:

$$g^{(t)} = \varphi(W_{gx}x^{(t)} + W_{gx}h^{(t-1)} + b_g) \tag{2}$$

$$i^{(t)} = \sigma(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + b_i) \tag{3}$$

$$f^{(t)} = \sigma(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + b_f) \tag{4}$$

$$o^{(t)} = \sigma(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + b_o) \tag{5}$$

$$s^{(t)} = g^{(t)} \odot i^{(t)} + s^{(t-1)} \odot f^{(t)} \tag{6}$$

$$h^{(t)} = s^{(t)} \odot o^{(t)} \tag{7}$$

The input gate uses sigmoid to determine which values to update and tanh (yellow rectangle) to generate new candidate values. Then the output value of sigmoid is multiplied by the candidate information output by tanh, that is, the information left by the final decision. The sum of the results obtained in step 1 and step 2 is the process of discarding unwanted information and adding new information, which is then transmitted to the next neuron. Figure 2 is just one neuron.

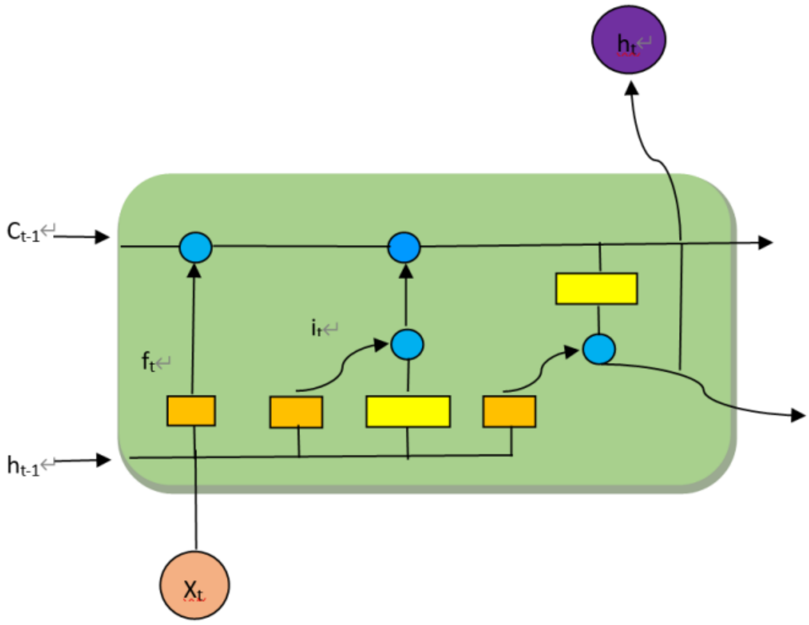

Figure 2 LSTM cell

2.4 Boosting and information entropy weighted LSTM neural network

In order to continuously improve the accuracy of errors, this paper combines the conventional Boosting framework[8] with the LSTM neural network[9] prediction method, and puts forward an innovative fitting method of residual series, as shown in Figure 3.

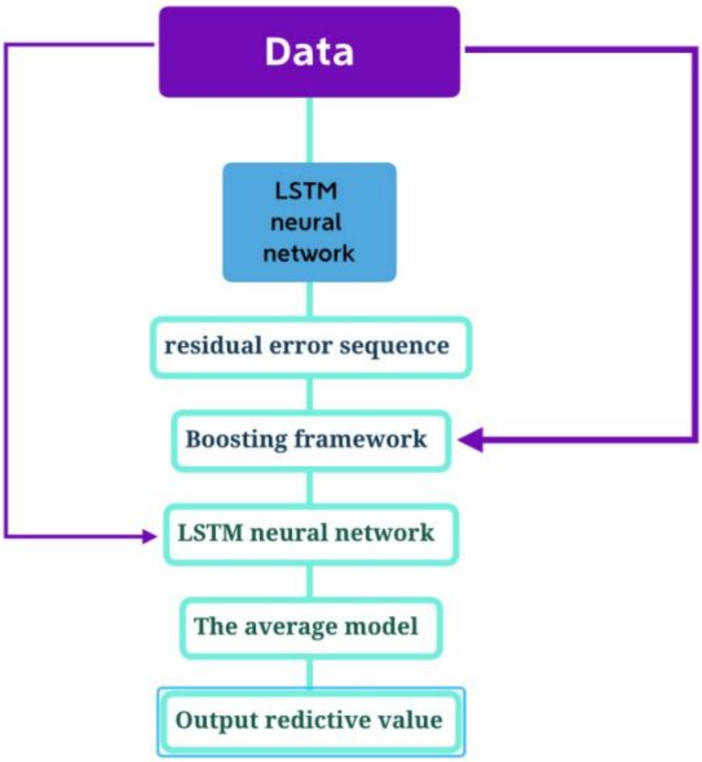

Figure 3 The flow chart of model

Step 1: Firstly, the data series after basis expansion and kernel principal component analysis were imported into the LSTM neural network for calculation. The preliminary predicted values were obtained through the LSTM neural network, and the first residual sequence was obtained by making the difference between these predicted values and the real values.

Step 2: Merge the original data and residual sequence that have been dimensionally reduced, and select any model as the basic algorithm of boosting framework based on the base expansion coefficient, and obtain more accurate residual through boosting framework. In the third step, the new prediction residuals with higher accuracy are added back to the original data sequence, and then input into the LSTM neural network for operation again to obtain a new prediction value.

Step 3: Because the number of original data sets is different, different models are generated. Therefore, in the last step, the predicted value with smaller error is given higher weight through model averaging method[10], and the final predicted value is obtained after weighted average. This predicted value is the predicted value with the smallest error and the closest to the real value. Then, using the predicted values obtained from the test set, the same method is used to calculate the training set.

3. The data analysis

3.1 The data source

The data used in this paper is from the Wind Information Financial database (www.wind.com.cn). In this paper, the intraday price data of three stocks listed on the Shanghai Stock Exchange, namely, Xinchao Energy (stock code 600777), Orient Group (stock code 600811), and Oppein Home (stock code 603833), are selected as experimental objects. The experimental data contains 242 days of price data of three stocks, a total of 174,240 hours points.

Among them, to illustrate the applicability of the experiment in the prediction of unstable time series, the experimental objects selected in this paper are from different industries. The three stocks used in the experiment samples belong to the oil exploitation, comprehensive, and furniture industries respectively, and the corresponding representative companies are Shandong Xinchao Energy Corporation Limited, Orient Group Incorporation, and Oppein Home Group Incorporation. The time series data with different degrees of instability are fully considered, which makes the experimental results more representative and stationery for the future trend prediction of the unstable time series. In this paper, mean square error (MSE), mean relative error (MRE), and bayes error (BE) are selected as the evaluation indexes of model performance.

The mean square error (MSE) is

$$MSE = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{8}$$

The mean relative error (MRE) is

$$MRE = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{9}$$

The bayes error (BE) is

$$BE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left(y_i - \hat{y}\right) - \mu\right)^2}} \tag{10}$$

The following is the intraday price chart of the three stocks:

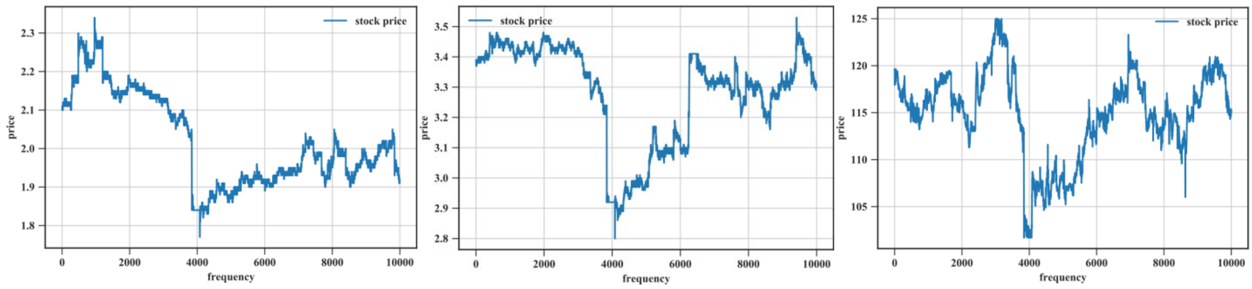

Figure 4 Stock price trend chart of three stocks

3.2 Comparing the results

In this paper, Python is used for experiments, and 6000, 6500, 7000, 7500, 8000, 8500, and 9000 data sets are used for experiments. In this paper, the data set is divided into the training set and test set, and 150, 200, 250, and 300 pieces of data are respectively selected as the training set to train the experimental model, and the results obtained are robust. The following table compares the mean square error (MSE) value, mean relative error (MRE) value, and bayes error (BE) value of the experimental model in this paper with the traditional LSTM model, as shown in Tables 1 to 3.

Table 1 Comparison table of Xinchao Energy

	BDLSTM-MSE/10-6	BDLSTM-MRE/10-3	BDLSTM-BE/10-3	LSTM-MSE	LSTM-MRE/10-1	LSTM-BE/10-1
0	4.80046	7.2411	1.56818	2.5545	7.9633	7.3088
1	3.45605	6.0717	1.37458	2.5179	7.9058	7.6358
2	0.98918	3.4545	0.80144	2.5127	7.8978	7.5857
3	0.55792	2.7034	0.62432	2.5091	7.8920	7.5723
4	0.47059	2.5373	0.57399	2.5340	7.9312	7.3040
5	0.49091	2.5681	0.58556	2.5497	7.9557	7.1563
6	0.46114	2.4913	0.56815	2.5454	7.9491	7.2286

Table 2 Comparison table of Orient Group

	BDLSTM-MSE/10-5	BDLSTM-MRE/10-3	BDLSTM-BE/10-3	LSTM-MSE/10-1	LSTM-MRE/10-1	LSTM-BE/10-1
0	24.6392	2.7370	0.91882	68.884	8.0353	3.9193
1	16.8190	2.4354	0.76645	68.036	7.9864	4.0981
2	17.4508	2.4360	0.77896	68.690	8.0243	3.8966
3	14.5637	2.2999	0.71991	68.419	8.0085	3.8543
4	9.4865	1.9888	0.58691	68.936	8.0384	3.6217
5	6.9761	1.7867	0.50465	68.958	8.0398	3.7125
6	6.7465	1.7469	0.49617	69.362	8.0631	3.5808

Table 3 Comparison table of Oppein Home

	BDLSTM-MSE/10-1	BDLSTM-MRE/10-3	BDLSTM-BE/10-3	LSTM-MSE/102	LSTM-MRE/10-1	LSTM-BE/10-1
0	4.2600	3.3384	1.38284	13.061	9.9524	9.5856
1	4.1585	3.3074	1.36604	13.060	9.9519	9.5714
2	3.9248	3.0879	1.32570	13.060	9.9521	9.5783
3	3.6942	2.9222	1.27658	13.058	9.9514	9.5742
4	3.2810	2.7727	1.20295	13.059	9.9516	9.5725
5	1.4102	2.2130	0.79522	13.058	9.9514	9.5743
6	1.2361	2.0924	0.74656	13.062	9.9528	9.5851

The results of mean square error (MSE), mean relative error (MRE) and bayes error (BE) of the experimental model in this paper and the traditional LSTM model are compared as follows, as shown in Figure 5 to 7.

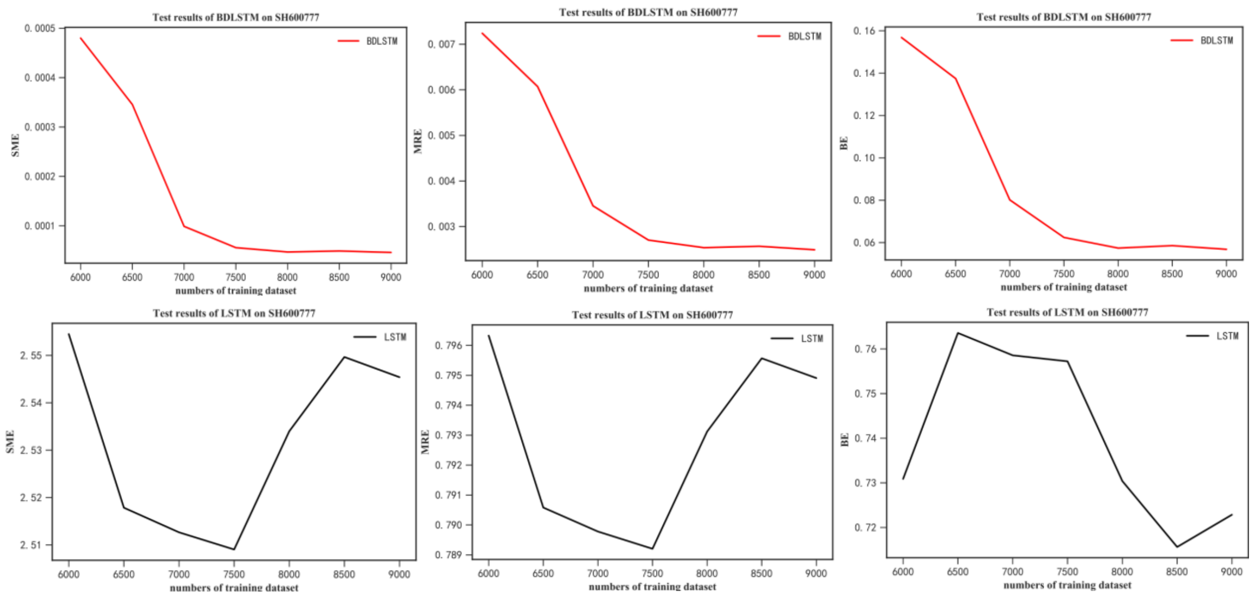

Figure 5 Comparison of Xinchao Energy

By comparing the mean square error (MSE) value, mean relative error (MRE) value, and bayes error (BE) value of the proposed model and the traditional LSTM model, it can be concluded that the fitting prediction effect of the proposed model is better, which is reflected in the improvement of accuracy and robustness. Firstly, the model in this paper improves the accuracy of prediction. The model is based on orthogonal polynomial expansion, and the nonlinear information and function information are extracted while the dimension is reduced.

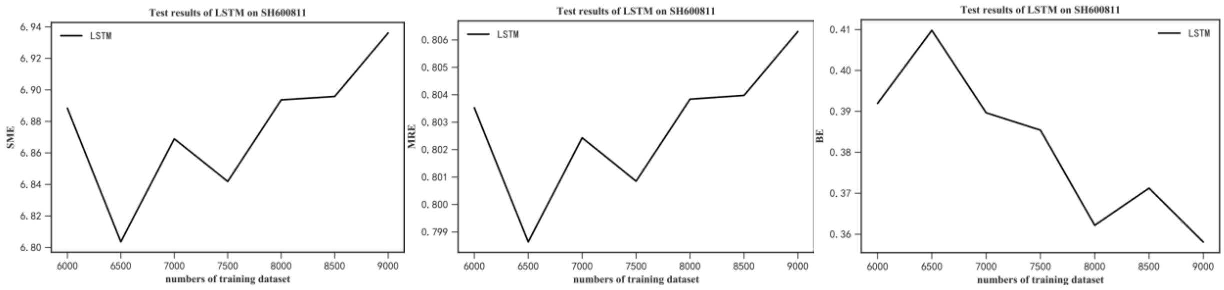

Figure 6 Comparison of Orient Group

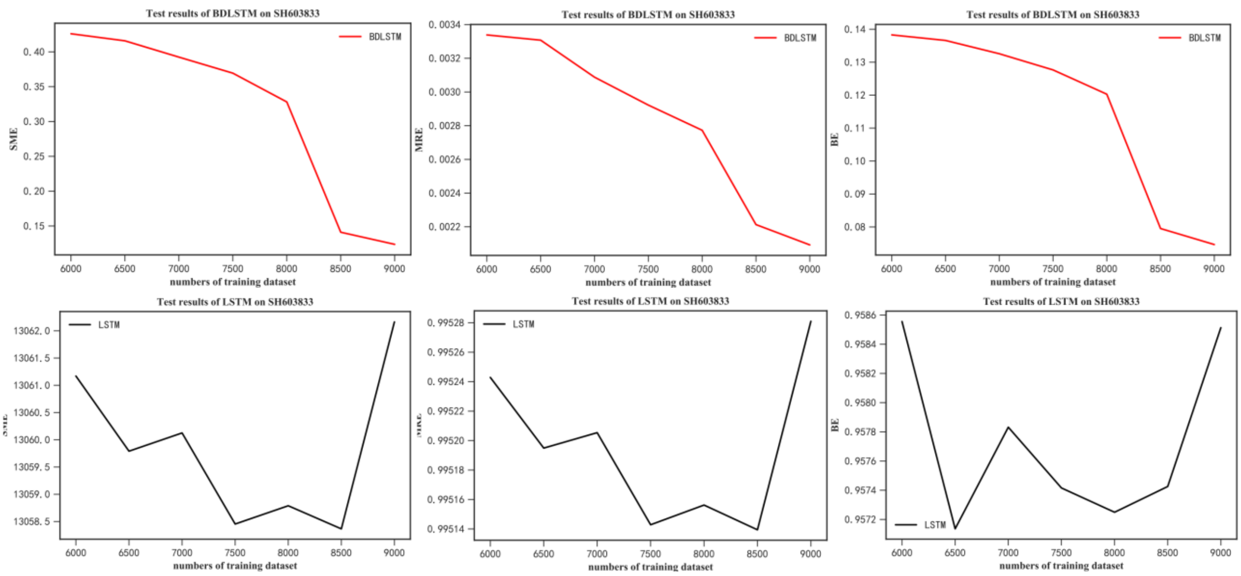

Figure 7 Comparison of Oppein Home

In addition, Boosting method is used to continuously capture the potential model results between the feature vectors and the residual sequence after dimensionality reduction, and the residual is continuously reduced through multiple residual sequence fitting. In addition, the model averaging method is used to solve the problem of selecting the dimension of the LSTM neural network predictor variables, which greatly reduces the variance and bias of the model prediction. It is worth mentioning that although most of the errors are reduced because the number of samples in the training set is increased and the deviation is greatly reduced, it can still be considered that the model in this paper effectively reduces the variance and error, and can better restore the trend of the original time series. Secondly, the proposed model is more robust than the traditional LSTM model. With the increase in the number of samples in the training set, the prediction error of the model should be consistently reduced. However, the prediction results of the traditional LSTM model show extremely unstable fluctuations, while the results of the model in this paper are relatively normal, so they can be considered to be robust.

4. Conclusions

This paper proposed a LSTM neural network prediction method based on Boosting and entropy weighted for high-frequency stock price data prediction of three representative listed companies. The innovation of this method lies in: (1) the nonlinear information and function information of the original data were effectively extracted when the dimension is reduced; (2) An integrated learning method was used to effectively capture the potential model results between the reduced dimension feature vector and the residual sequence; (3) For the dimension selection of LSTM neural network, the model average method was adopted to weight it, and the prediction effect was optimized. The final experimental results showed that the proposed method is competitive with the original LSTM neural network. To sum up, this paper improved some classic stock price prediction models through integrated learning, model average, neural network and other methods, which improves the prediction accuracy to a certain extent. We hope this model can help to predict high-frequency stock price data.

The future research directions of this paper will be divided into the following four parts: (1) Explore the impact of different basis function expansions and different Boosting basis models on the prediction results; (2) Integrated learning methods include Boosting, Bagging and Stacking. Each method has its advantages and disadvantages. Whether different integrated learning methods will cause significant differences in prediction results will become the focus of subsequent research; (3) The proposed model is used for real-time monitoring of trace elements in environmental science, monitoring of road traffic flow and prediction of daily average temperature curve, and compared with previous research results; (4) High frequency time series can be divided into dense high frequency time series and sparse high frequency time series. Whether this model is applicable to sparse high frequency series is also a positive research direction.

References

[1] Jin Shuying, Tan Xin. Research on the relationship between stock liquidity and capital cost based on the dynamic adjustment of capital structure [J]. School of Economics and Management, Shenyang Aerospace University, 2020, 37(05): 80-89.

[2] Box G E P , Jenkins G M , Reinsel C . Time series analysis: forecasting and control (third ed[J]. Oakland, California, Holden-Day, 1976, 2013.

[3] Engle R F . Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation[J]. Econometrica, 1981, 50(4):987-1008.

[4] Wang Dongmei, Guo Wenxia, Yuan Shufang, Pan Jiahui, Hao Weiya. Prediction of vertical ground reaction force in running based on principal component analysis and wavelet neural network [J]. Journal of Medical Biomechanics , 2022,37(04):706-712.

[5] Cui Jing, Functional principal component analysis and prediction of stock price [D]. Guangxi Normal University,2021.

[6] Xu Mei, Wang Fang. Forecasting financial volatility based on bp neural network and symbolic time series[J]. Journal of WUT(Information＆Management Engineering), 2015,37(04):456-460.

[7] Li Liping, Zeng Lifang, Jiang Shaoping, He Wenqian. Stock price forecast based on lstm neural network[J/OL]. Journal of Yunnan Minzu University (Natural Sciences Edition): 1-10[2022-09-21].

[8] Zhang Wensheng, Yu Yanzhao. Boosting Algorithm theory and application research [J]. Journal of university of science and technology of China: 222-230. 2016.05.

[9] Huang Yucheng, Fang Weiwei. Research on Stock Price Forecast Based on LSTM Network [J]. Modern Computer :51-60. 2021.12.05.

[10] Jiang Rongjie. Model averaging method for nonlinear models [D]. Shanghai University of Finance and Economics.2020.06.