An improved LSTM stock price prediction model based on multiple basis function expansion and two-layer Bagging algorithm

Qiong Wu [1,#,*], Jiayi Lu [2,#], Heping Zhang [3,#]

[1] School of Statistics, Jilin University of Finance and Economics, Jilin, China, 130117

[2] School of Urban, Xi 'an Polytechnic University, Shaanxi, China, 710600

[3] School of Economics and Management, Jilin Jianzhu University, Jilin, China, 130000

* Corresponding Author Email: 18297747148@163.com

[#] These authors contributed equally.

Abstract. Using neural network technology to predict the time series data of stock prices is a hot issue in the field of computational science. For example, Long Short-Term memory network technology (LSTM) has been widely used in stock price prediction. However, if the original stock opening price series information is not enough to predict the future trend, the effect of LSTM model may not be ideal. In addition, improper selection of the step size will also make the prediction effect of the LSTM model poor. This paper proposes an improved LSTM model based on multi-view basis function expansion and Bagging algorithm. On the one hand, under a fixed step size, the intra-day price information of the stock is extracted by using a variety of basis function expansion forms as prior information, and then the Bagging algorithm is used to establish the relationship between the basis expansion coefficients and the prediction residual of the original LSTM model to predict the residual, thereby increasing the prediction accuracy. On the other hand, the problem of LSTM step size selection is adaptively solved by using the idea of model average instead of the model selection method, and the specific method is to combine the weighted average of LSTM models with different time lengths. The actual data analysis shows that the proposed method improves the prediction accuracy of the original LSTM model, and the t-test shows that the proposed method has significant robustness.

Keywords: stock price prediction, LSTM, basis function expansion, Bagging.

1. Introduction

The stock price trend directly affects the economic interests of investors, and also affects the country's macroeconomic policy. The establishment of accurate prediction models for stock prices can improve the transparency of the stock market, which can not only help investors avoid the risks that may occur in the stock market in advance and make more rational investments, but also assist listed companies to understand their operating conditions and future development trends. Stock price data fluctuates over time, and future stock prices are affected by historical data, so stock prices are a kind of time series data. This paper uses the time series data of stock prices of three Shanghai stock exchanges as an example, and the fluctuation trend is shown in Fig.1 as a waveform diagram.

The existing stock price time prediction methods mainly include traditional time series prediction and neural network prediction based on deep learning technology. For the traditional time series forecasting methods, for example, Vo N and Slepaczuk R [1] used the hybrid model of ARIMA and GARCH family model to test the performance of S&P 500 log return and obtained better forecasting results. Wang Y and Guo YK [2] proposed a DWT-ARIMA-GSXGB hybrid model to fit the stock index opening price, and the prediction performance was much higher than that of the single ARIMA model. Liu Q [3] et al. used the ARMA-GARCH model to simulate and predict the drought index in North China Plain more accurately. However, in the traditional time series analysis and prediction, the stock price time series data will lose some important information in the process of processing it

into a stationary series by using the difference method, resulting in the decline of the accuracy of the prediction results.

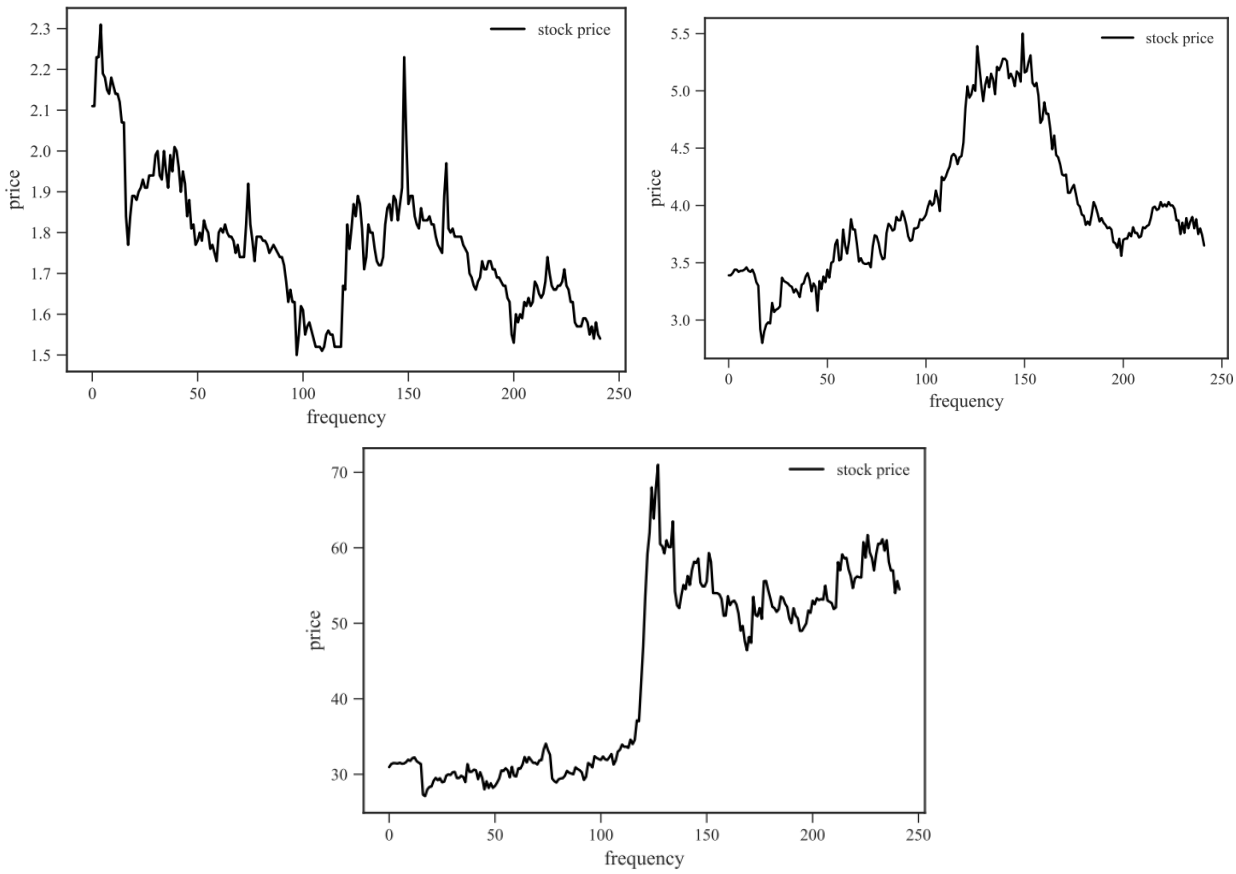


Fig 1. stock price time series

In recent years, neural network prediction methods based on deep learning have gradually replaced traditional time series prediction methods and become a common means in the field of stock price quantitative research. The main advantage lies in the fact that no additional assumptions are made on the original time series and the ability to approximate the link function of any degree of complexity. Specifically, Wang Y and Ma S [4] introduced the ARIMA-LSTM-SVR model to conduct hybrid prediction experiments on time series in different fields, and the prediction advantages of the model are more obvious compared with the traditional time series prediction model. Zhao H and Xue L et al. [5] proposed a hybrid model (LSTM-CNN-CBAM) based on the combination of LSTM and CNN algorithms to improve the feature extraction and prediction ability of the network model. Chen DW and Zhang JH [6] used a hybrid STL-LSTM model to predict the subway passenger flow. Zhang HW [7] et al. used the Bi-LSTM structure to predict image content according to context cues. Ma W [8] et al. proposed an algorithm combining LSTM and an improved residual neural network to detect abnormal traffic. The above works fully illustrate the excellent performance of LSTM models based on deep learning in time series prediction problems. However, in the problem of stock price prediction, we usually only use the opening price sequence to predict, which may have the problem that the original sequence information itself is not enough to ensure the accuracy of the prediction model. In addition, in the hyperparameters of the LSTM model, whether the step size parameter can be reasonably selected will directly affect the accuracy of the prediction model.

Based on this, this paper first uses a variety of basis function expansion methods to functionalize the intraday price information and project it into different function Spaces as prior information to form function features under different basis functions. Then, the Bagging algorithm is used twice to construct the prediction model between the basis expansion coefficient and the original LSTM prediction residual sequence, and the fusion prediction model of LSTM neural network with different

time lengths. The innovations of the proposed method are as follows: **(1)** the intraday price is integrated into the original LSTM model as prior information, which improves the prediction accuracy. **(2)** The two-layer bagging model effectively balances the variance and bias of the prediction model, and adaptively solves the problem of step size selection of the LSTM model. Finally, experiments on multiple stock price series show that the proposed method has certain competitiveness.

2. The basic theory and method

2.1. LSTM Neural Networks

LSTM neural network model is a special RNN [9]. In the neurons of the hidden layer of RNN, LSTM adds a structure called a memory cell to remember the input information data, and adds three kinds of auxiliary model structure gates, input gate, forget gate and output gate, to control the past input information. To be specific, A_{T-1} represents the state of the memory cell at the previous moment, A_{T+1} represents the state of the memory cell at the later moment. The data input of the memory cell x_t is determined by the input data xt at the current time and the output of the previous memory cell hidden layer h_{t-1}. As shown in Equation One:

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

The input gate mainly determines the input value of the current memory cell, which is determined by i_t and $\tilde{C}_t$. I_t mainly controls input values of the memory cells, and $\tilde{C}_t$ mainly controls the candidate values for generating memory cells.As shown in Equation Two:

$$i_t = \sigma(w_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(w_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

The update of the current memory cell state is mainly determined by two parts: 1. f_t acts on C_{t-1} to control the old memory of the current memory cell; 2, i_t acts on $\tilde{C}_t$ to control the new memory of the current memory cell, and the old memory and the new memory together form the new state C_t of the current memory cell. As shown in Equation Four:

$$C_t = f_t * C_{t-1} + i_{t*}\tilde{C}_t \tag{4}$$

The output gate is mainly used to decide the influence of the current memory cell C_t on the output h_t. Firstly, the current memory cell C_t is applied to the tanh activation function for normalization, and then according to x_t and h_{t-1}, it is calculated that o_t determines the influence of C_t output data on h_t. As shown in Equation Five and Equation Six:

$$o_t = \sigma(w_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

The above process is shown in Fig.2:

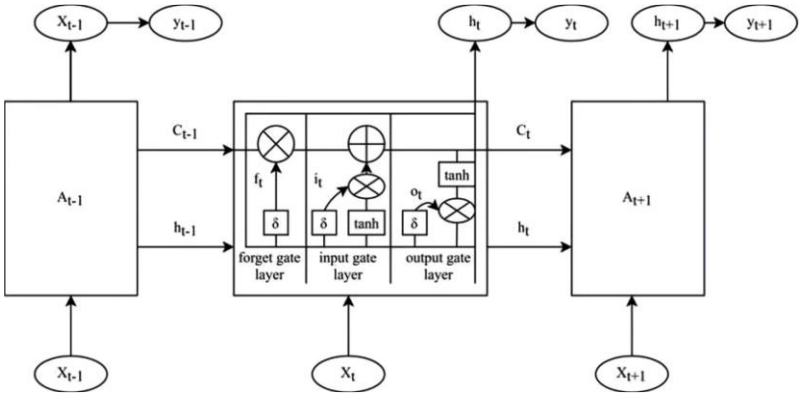


Fig 2. The structure of the LSTM network model

2.2. Basis function expansion

Basis function expansion is a common method of functionalizing discrete series. By treating time series as a discrete realization of random process, the original data information is concentrated on the basis expansion coefficients. Since different basis functions extract different information, this paper proposes to use a variety of basis function expansion methods, and then splice different basis function expansion coefficients to form a feature vector that fuses multiple basis function expansions. The main basis function expansion used in this paper is the functional principal component basis function expansion [10]:

$$x_i(s) = \sum_{k=1}^{K} C_{ik}\varphi_k(s), i = 1,2,\ldots,N \tag{7}$$

The $\varphi_k(s)$ as the functional principal component basis function; B-spline basis function expansion:

$$N_{i,0}(u) = \begin{cases} 1 & \text{if } u_i \le u \le u_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$N_{i,p}(u) = \frac{u-u_i}{u_{i+p-1}-u_i} N_{p-1}^{(i)}(u) + \frac{u_{i+p}-u}{u_{i+p}-u_{i+1}} N_{p-1}^{(i+1)}(u) \tag{9}$$

Where $U = \{u_0, u_1, \ldots, u_m\}$ is a monotone and nondecreasing sequence of real numbers, that is, $u_i \le u_{i+1}, i = 0,1,\ldots,m\text{-}1$. U_i is called a node, U is called a node vector, and Use $Ni, p(u)$ to represent the ith p(p+1 order). And Fourier basis function expansion:

$$f(t) = a_0 + \sum_{k=1}^{\infty}\left\{a_k \cos\left(\frac{2\pi k}{T_0}t\right) + b_k \sin\left(\frac{2\pi k}{T_0}t\right)\right\} \tag{10}$$

2.3. The Bagging algorithm

Bagging algorithm is a model averaging algorithm, whose idea is to continue the next training process after pulling out multiple training sets once when sampling the training set, forming several weak learners that train and predict independently, and finally combining each weak learner into a strong learner that can obtain the final prediction result [11]. The specific process is shown in Fig.3 below.

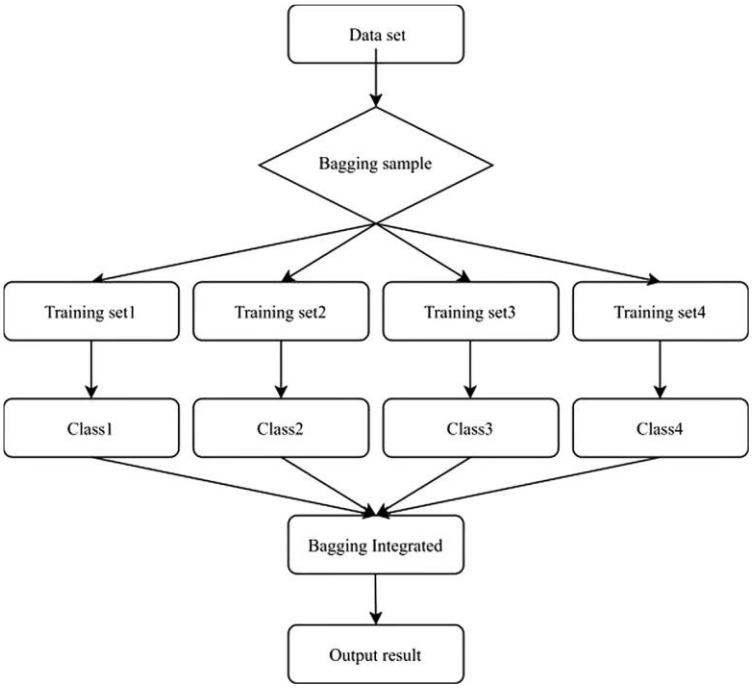


Fig 3. The structure of the Bagging algorithm

2.4. Improved LSTM model

In this paper, intra-day price information is used as prior information, and the Bagging algorithm is used to adaptively solve the problems that different synchronization length Settings have a great influence on the prediction accuracy of LSTM. The process is as follows:

Step 1: set the step size set, and perform the LSTM prediction of the original stock opening price under different step sizes (step size is fixed). The prediction value and the corresponding residual sequence of the LSTM neural network on the training set and the test set were obtained.

Step 2: FPCA basis function expansion, B-spline basis function expansion, and Fourier basis function expansion were used to functionalize the intra-day price data, and the basis expansion coefficient vectors were extracted and spliced to obtain the function feature vectors (including the training set and the test set).

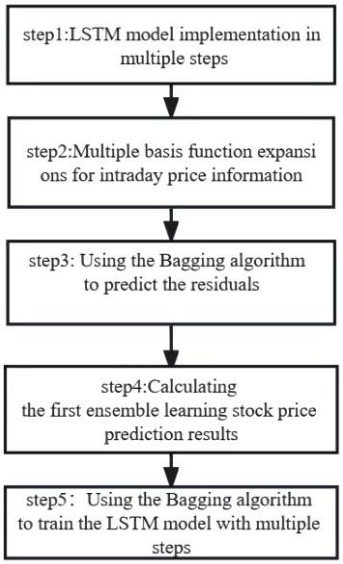


Fig 4. The flowchart of LSTM improvement model

Step 3: On the training set, the function feature vector is paired with the LSTM prediction residual, and the Bagging algorithm is used to train the model parameters of each sub-model in the Bagging algorithm, and the corresponding residual prediction value sequence (including training set and test set) is output.

Step 4: Add the corresponding residual prediction sequence with the prediction value of the original LSTM to obtain the stock price prediction value under different step sizes (fixed step sizes) (including training set and test set).

Step 5: Again, the LSTM prediction value on the training set is paired with the true value under different synchronization lengths and trained by the Bagging algorithm to obtain the model parameters in the Bagging algorithm, and output the final prediction results on the test set. The key steps of this improved method are shown in Fig.4.

3. Data analysis

3.1. Data sources and experimental Settings

The data source of the three stocks is Wind database (https://www.wind.com.cn). In this paper, the intra-day price and opening price data of the three stocks in a period of time (242 days in total) are extracted online, and the influence of holidays and other factors is removed, and a total of 174966 original time series data are obtained. The three stocks are selected based on the degree of volatility; It can be seen from Figure 1 that the degree of volatility of each stock is different. This paper wants to explore the performance effect of LSTM and the improved performance effect under different degrees of volatility, to judge the applicability and robustness of the proposed method. The specific

data sources are shown in Table I below. The experiment is implemented based on python, the algorithm used in the LSTM neural network is adam algorithm, the activation function is tanh, the learning rate is 0.005, and the loss function is the mean square loss function. The set of step sizes of LSTM is [2,5,10,15,20]. In addition, in order to check the robustness of the proposed method, we divide the training set and the test set with different numbers. The details are shown in Table II:

Table 1. Explanation of data sources

Stock code	Abbreviation	Type
600777	New wave energy	Shanghai A
600811	Oriental Group	Shanghai A
300741	Huabao Corporation	Growth Enterprise board
------	Name of company	Industry
600777	Shandong Xinchao Energy Co. LTD	Petroleum
600811	Orient Group Co. LTD	Farming, animal husbandry, feeding and fishing
300741	Warburg Essence Co. LTD	Food and beverage

Table 2. Partitioning results between training set and test set

Number of training sets (number)	135	140	145	150	155	160	165	170	175	180
Number of test sets (number)	105	100	95	90	85	80	75	70	65	60
Percentage of training set (percentage)	0.56	0.58	0.60	0.63	0.65	0.67	0.69	0.71	0.73	0.75

3.2. Results Analysis

The indicators for prediction accuracy evaluation are mean square error (MSE), Absolute error (MRE), and posterior error (BE). The smaller the value of the three, the more accurate the prediction result. The formula is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad MRE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad BE = \frac{std(y_i - \hat{y}_i)}{std(y_i)} \tag{11}$$

The experimental results are shown in Table III, Table IV, and Table V. From the preliminary results, when the number of samples in the training set is fixed, the proposed method is smaller than the original LSTM in the three evaluation indicators, which indicates that the proposed method can effectively improve the prediction accuracy of the original LSTM.

Table 3. Comparison of evaluation index results (sh600777)

sh600777	FMAL-MSE	FMAL-MRE	FMAL-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
135	0.002015	0.017283	0.381614	2.078893	0.840276	0.517507
140	0.002007	0.017900	0.372601	2.079094	0.841035	0.501266
145	0.001868	0.017179	0.365928	2.059517	0.840671	0.511098
150	0.001314	0.016369	0.368222	2.056530	0.847118	0.512291
155	0.001089	0.015159	0.351040	2.050772	0.850136	0.446642
160	0.001016	0.013911	0.361306	2.063423	0.856714	0.480113
165	0.000913	0.013617	0.352869	2.047958	0.857553	0.493853
170	0.001084	0.015122	0.432904	2.037308	0.860939	0.517329
175	0.000814	0.012818	0.448241	2.049243	0.868342	0.597493
180	0.000908	0.013981	0.481564	2.064214	0.875664	0.635522

From the empirical results, when the number of training sets and test sets is fixed, the Mean Square Error (MSE), Absolute Error (MRE) and Posterior Error (BE) of the proposed method are significantly lower than those of the original LSTM, in particular, the mean square error is reduced by about two orders of magnitude. This shows that the optimization effect of the proposed method is better and the prediction accuracy is greatly improved. On the one hand, it may be because this paper uses Bagging algorithm based on the idea of model average instead of model selection, which balances the deviation and variance of the prediction model, thus improving the prediction accuracy. On the other hand, The proposed method considers the function information of the intraday price and uses Bagging algorithm to construct a prediction model between the prediction residual sequence of the original LSTM and this function information, which is equivalent to providing prior information for the original LSTM model and improving the prediction accuracy.

Table 4. Comparison of evaluation index results(sh600811)

sh600811	FMAL-MSE	FMAL-MRE	FMAL-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
135	0.011812	0.017718	0.157294	12.87948	0.848375	0.590752
140	0.011504	0.017670	0.163862	12.73674	0.852724	0.597129
145	0.011288	0.018037	0.170334	12.54303	0.856408	0.606845
150	0.00814	0.015798	0.156769	12.4036	0.863535	0.60634
155	0.007829	0.016052	0.189937	12.06529	0.865543	0.579217
160	0.005251	0.013929	0.195244	11.95972	0.873296	0.592614
165	0.003527	0.012066	0.236444	11.71905	0.875345	0.612467
170	0.002351	0.010089	0.280642	11.70271	0.882411	0.705408
175	0.001734	0.00877	0.300612	11.71479	0.887883	0.80657
180	0.001385	0.007918	0.31613	11.76642	0.893273	0.852145

Table 5. Comparison of evaluation index results (sz300741)

sz300741	FMAL-MSE	FMAL-MRE	FMAL-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
135	24.74544	0.083982	0.704738	2858.247	0.983881	0.97831
140	5.174730	0.033000	0.496665	2883.327	0.987416	0.984269
145	3.620088	0.027307	0.439360	2876.099	0.987937	0.984777
150	3.400089	0.026026	0.411021	2867.841	0.988626	0.985235
155	3.197169	0.025923	0.389786	2854.726	0.988583	0.984671
160	1.740879	0.01933	0.299807	2863.852	0.988600	0.984575
165	1.250112	0.016107	0.271499	2873.142	0.98873	0.984713
170	0.711643	0.012283	0.230863	2911.152	0.988797	0.983250
175	0.577225	0.011184	0.217943	2942.761	0.988581	0.984439
180	0.404916	0.009469	0.178830	2952.618	0.988517	0.984252

In addition, in order to test whether the model has robustness, this paper tests the robustness of the model by setting different numbers of training sets and test sets, as shown in Table 2. The results show that when the number of samples in the training set changes, the prediction effect of the proposed method is still better than that of the original LSTM. Finally, the hypothesis test method is used to determine whether the proposed method significantly improves the prediction accuracy of the original LSTM. The specific results are shown in Table VI, respectively. From the experimental results of the three stocks, the p values of the hypothesis test of the different results of the three evaluation indicators of the original LSTM and the improved LSTM are all less than 0.0001, that is, the null hypothesis is rejected at the significance level of 0.001. The conclusion is that compared with the original LSTM, the proposed method has better performance. It has smaller mean square error (MSE), Absolute error (MRE), and posterior error (BE).

Table 6. Results of hypothesis testing

sh600777	MSE	MRE	BE
statistic	-499.404	-192.754	-17.549
P_value	<0.0001	<0.0001	<0.0001
sh600811	MSE	MRE	BE
statistic	-85.195	-142.924	-27.75
pvalue	<0.0001	<0.0001	<0.0001
sz300741	MSE	MRE	BE
statistic	-235.321	-131.409	-12.205
pvalue	<0.0001	<0.0001	<0.0001

4. Conclusion and Prospect

Considering that the information of the original stock price series is not enough to predict the future trend and the step size selection problem of LSTM, this paper first expands the intra-day price information through a variety of basis functions and forms the function feature vector. Considering that the influence relationship between the basis function coefficients and the prediction residual of the original LSTM model is unknown. Therefore, this paper uses Bagging algorithm to construct multi-model prediction to balance the bias and variance problems of a single model. Then, it considered using a set of step sizes instead of a single step size and used Bagging algorithm to adaptively solve the step size selection problem again. The actual data analysis results show that the proposed method can effectively improve the prediction accuracy of the original LSTM.

For future research directions, on the one hand, the basis expansion method is equivalent to the dimensionality reduction method, and in the process of dimensionality reduction, whether the effective information related to the response variable is fully preserved is an important issue. However, none of the basis functions used in this paper has this characteristic, and we can use the functional dimension reduction method to deal with this. On the other hand, in the Bagging algorithm, the importance of each model is obviously different, so how to assign different weights to different models is also worth considering. For this, we can solve it by heuristic model averaging or optimized model averaging methods, such as Bayes' rule and Mallos's rule[12].

References

[1] Vo, N., & Ślepaczuk, R. (2022). Applying Hybrid ARIMA-SGARCH in Algorithmic Investment Strategies on S&P500 Index. Entropy (Basel, Switzerland), 24(2),158.

[2] Wang, Y., & Guo, Y. (2020). Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. China Communications, 17, 205-221.

[3] Liu, Q., Zhang, G., Ali, S., Wang, X., Wang, G., Pan, Z., & Zhang, J. (2019). SPI-based drought simulation and prediction using ARMA-GARCH model. Appl. Math. Comput., 355, 96-107.

[4] Wang Yingwei, Ma Shucai. Time Series prediction based on Hybrid Model of ARIMA and LSTM [J]. Computer Applications and Software,2021,38(02):291-298. (in Chinese)

[5] Zhao Hongrui, Xue Lei. Research on Stock Prediction based on LSTM-CNN-CBAM Model [J] Computer Engineering and Applications,2021,57(03):203-207. (in Chinese)

[6] D. Chen, J. Zhang and S. Jiang, "Forecasting the Short-Term Metro Ridership With Seasonal and Trend Decomposition Using Loess and LSTM Neural Networks," in IEEE Access, vol. 8, pp. 91181-91187, 2020.

[7] H. Zhang, C. Ma, Z. Jiang and J. Lian, "Image Caption Generation Using Contextual Information Fusion With Bi-LSTM-s," in IEEE Access, vol. 11, pp. 134-143, 2023.

[8] Ma, W., Zhang, Y., Guo, J., & Li, K. (2021). Unbalanced abnormal traffic detection based on improved Res-BIGRU and integrated dynamic ELM optimization. Comput. Commun., 179, 112-130.

[9] Yang Guang, Chao Suya, Nie Min, Liu Yuanhua, Zhang Meiling (2023). Construction of hybrid quantum LSTM neural networks for image Classification [J/OL]. Journal of Physics, 1-29.

[10] Yan, X., Zhou, Y., Pu, X., & Zhao, P. (2021). Functional Multiple-Outcome Model in Application to Multivariate Growth Curves of Infant Data. Journal of Systems Science and Complexity, 1-23.

[11] Xu Wen, Tang Jian, Xia Heng & Qiao Junfei. (2022). Soft measurement of dioxin emission concentration based on Bagging semi-supervised deep forest regression. Journal of instruments and meters (6), 251-259.

[12] Bai Haifeng, Liu Xiaodong, Niu Shukui, He Yadong. Construction of forest fire prediction model based on Bayesian model averaging method: taking Dali Prefecture, Yunnan Province of southwestern China as an example[J]. Journal of Beijing Forestry University, 2021, 43(5): 44-52.