

A high frequency stock price prediction model based on Hermite basis expansion and LSTM neural network

Tiantian Shi^{1,*}, Hao Meng²

¹School of Economics, Lanzhou University, Lanzhou, China, 730106

²School of economics, Anhui University of Finance and Economics, Bengbu, China, 233000

*Corresponding author: tt2802256897 @163.com

Abstract. Based on Hermite basis function expansion and ensemble learning, an improved LSTM neural network method is proposed in this paper. The proposed method can be used for high frequency stock price prediction. Considering the characteristics of high frequency time series data such as high dimensionality, large noise and instability, etc. In this paper, the function information extracted from the Hermite basis function expansion is used to predict the residual sequence predicted by the LSTM neural network. Since the components of the function feature vector are unknown to the underlying model structure of the response variable, the proposed method is processed by Bagging framework. It not only captures the structure of the latent model, but also balances the variance and bias of the model. In addition, the number of prediction periods of the LSTM neural network is a hyperparameter, and the model averaging method based on distance covariance weighting is considered in this paper for optimization. The results of actual data analysis show that the proposed method can effectively optimize the prediction accuracy of the LSTM neural network, and has certain robustness. Finally, on the one hand, this optimization framework can be used to improve other time series prediction models. On the other hand, the proposed method can play an important role in forecasting problems such as daily average temperature prediction and real-time monitoring of atmospheric environmental quality.

Keywords: High frequency stock price prediction, LSTM neural network, model averaging, basis function expansion.

1. Introduction

As we all know, the stock market is complicated and volatile, and the stock price is affected by many factors such as economic cycle, financial policy and international environment, so investors need to take huge risks while obtaining benefits. High-frequency stock prices belong to high-frequency time series data, which has the characteristics of high dimensionality, loud noise and instability, etc., while high-frequency financial data contains more information and has a significant effect on capturing micro changes of the stock market. Therefore, the use of high-frequency time series to predict the trend of stock prices can more accurately avoid risks and obtain more ideal returns.

The traditional time series prediction methods mainly solve the model parameters on the basis of determining the time series parameter model, and use the solved model to complete the prediction work. Li Zehui (2019) [1] selected China Merchants Bank as the research object, established ARIMA model, and studied the volatility of return rate based on this model. Yang Yun (2021) [2] et al. proposed an improved LOBNN&AR-GARCH model to achieve short-term prediction of stock prices. However, the above methods have dimension disaster, instability and other problems, and long-term prediction is needed. Therefore, ARIM, GARCH and other models have limitations, because they cannot capture nonlinear information and function information. Zhou Hua (2018) [3] et al. used the principal component analysis method to construct the absorption rate, which was used to measure the changes in the degree of risk association between various industries in China over time. Zhao Dandan (2019) [4] et al. used support vector machine (SVM) and kernel principal component analysis (PCA) to build a systemic risk early warning model for China's banking industry to predict the level of systemic risk in China's banking industry. This kind of dimensionality reduction and then prediction method does improve the accuracy, but the disadvantage is that it does not take into account the model structure between the dimensionality reduction variable and the response variable. Shen Zejun (2019)

[5] et al. proposed a BP neural network algorithm based on granular computing thinking. Shi Jiannan (2020) [6] et al., proposed a stock price time series prediction method based on dynamic mode decomposition-Long short-term Memory Neural network (DMD-LSTM) to solve the problems of difficulty in effective feature extraction and low price prediction accuracy. LSTM and BP neural network prediction methods do not need stationarity assumption, nor dimensionality disaster, but without extracting function information, the noise of data will sometimes be enhanced incorrectly. In addition, it is difficult to choose the period number of predictor variables.

To sum up, this paper proposes an improved LSTM neural network prediction model based on Hermite basis function expansion and ensemble learning. The advantages of this method mainly include the following aspects: At the same time of reducing the dimension of time series, the function information and nonlinear information are extracted by using Hermite basis function expansion, which has the advantages of reducing the computational complexity and improving the estimation accuracy. The Bagging method is used to capture the potential model structure between the dimensionally reduced variables and the residual sequence, and to balance the bias and variance of the prediction model, so as to reduce the variance of the model, improve the generalization error and improve the accuracy. In addition, the forecasting methods in the Bagging framework are "model" free. To solve the problem of LSTM neural network parameter selection, a model averaging method based on distance covariance weighting is proposed in this paper. Experimental results show that the proposed method has a significant improvement on the original LSTM neural network.

2. Theory and Methods

2.1 Hermite basis function expansion

The basis expansion method refers to using the basis function expansion to approximate the real function, the specific form is as follows:

$$f(x) = \sum_n a_n H_n(x) \quad (1)$$

Where a_n is the expansion coefficient of the basis function, and the Hermite basis function $H_n(x)$ is expressed as follows:

$$H_n(x) = (-1)^n e^{x^2} (e^{-x^2})^{(n)} \quad (2)$$

Which is polynomial of degree n . $H_n(x)$ The N th derivative of is the product of some polynomial of degree n with, which is called Hermite polynomial.

2.2 Bagging framework

Bagging is a method used to improve the accuracy of learning algorithms. The following is the flow of Bagging algorithm: The input is the sample set $D = \{(x, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$, the weak learner algorithm, the number of iterations T of the weak learner, and the output is the final strong learner $f(x)$. Specifically, for $t=1, 2, \dots, T$: firstly, the training set is randomly sampled for T times, and a total of m times are collected to obtain a sampling set D_t containing m samples. Then, the weak learner $G_t(x)$ is trained with the sampling set D_t . Finally, the predicted value obtained by simple arithmetic average or other weighting methods is used as the final output. The specific process is shown in figure 1:

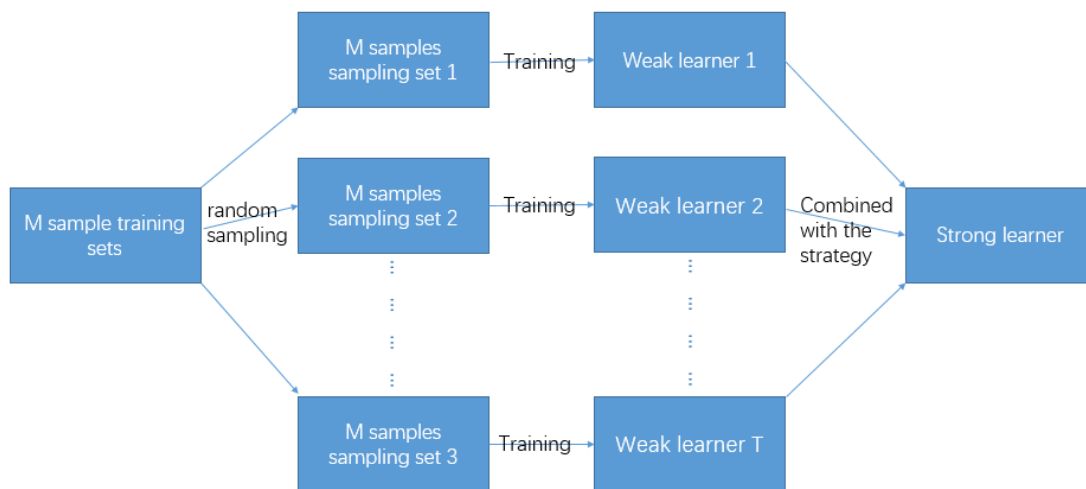


Fig. 1 Bagging framework

2.3 LSTM Neural networks for Hermite basis function expansion and ensemble learning

LSTM neural network is a short short-term memory network, which is an improved recurrent neural network. It can solve the problem that RNN cannot deal with the dependence of long distance. In this paper, we use the function information extracted based on the Hermite basis function expansion to predict the residual sequence obtained by the LSTM neural network, and use the Bagging framework to capture the latent model structure of the components of the function feature vector on the response variables, and balance the variance and bias of the model. In addition, we use the model averaging method based on distance covariance weighting to optimize the number of prediction periods of the LSTM neural network. The specific calculation process is shown as follows:

Step1: LSTM neural network is used to directly predict the stock price, and the result is $\hat{Y}_{n+1}, \hat{Y}_{n+2}, \dots, \hat{Y}_T$. The residual between the predicted value and the real Y is generated, e ;

Step2: Perform Hermite basis function expansion on the data first to get A vector;

Step3: Use Bagging framework for ensemble learning of E to predict the new residual sequence $\hat{e}$;

Step4: Pair the A vector with the $\hat{e}$ obtained in Step2 to get the data pair $(a, \hat{e})^{T-n}$;

Step5: Use vector A to predict the residual $\hat{e}$, and add back the predicted value in Step1 to get $\hat{Y}_{new}$;

Step6: Use the model averaging method based on distance covariance weighting to optimize $\hat{Y}_{new}$, and get the final model $\hat{Y}$.

As for the distance correlation coefficient, it is used to study the independence between two variables. Specifically, the Euclidean distance between sample points and Euclidean distance in space is denoted as $d_{ij}^Y = d(Y_i, Y_j)$, and the specific formula of distance covariance is as follows:

$$\nu_n^2(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^n d_{ij}^X d_{ij}^X + \frac{1}{n^2} \sum_{i,j=1}^n d_{ij}^X \frac{1}{n^2} \sum_{i,j=1}^n d_{ij}^Y - \frac{2}{n^3} \sum_{i=1}^n \sum_{i=1}^n d_{ik}^X d_{il}^Y \quad (3)$$

When variables X and Y are completely independent, the distance covariance is the smallest and the result is 0, indicating that there is no duplicated information between the two variables. On the contrary, the greater the distance covariance is, the greater the interdependence between the two variables is, that is, the more duplicated information between the two variables is.

3. The data analysis

3.1 Data sources and pre-analysis

All data analyzed in this paper were obtained from WIND database (<https://www.wind.com.cn/>). We obtain three stock data by random selection to determine whether our improved methods are effective for random systems with different levels of complexity. The following are three randomly selected stocks: the stock code SH600777 corresponds to Shandong Xinchao Energy Co., LTD., whose business scope includes oil and gas exploration, exploitation, sales, etc.; The stock code SH600811 corresponds to Dongfang Group Co., LTD. The company mainly invests in and operates seven industries: modern agriculture and health food, oil and gas and new energy, information security, finance, resources and products, port transportation, and new urbanization development; The stock code SH603833 corresponds to Opai Home Furnishing Group Co., LTD., which is a comprehensive modern integrated home furnishing service provider in China. We use the standard LSTM neural network and the improved method to predict three stock prices respectively, so as to verify whether the proposed method effectively improves the LSTM prediction method. We take the first 10,000 time points of the three stocks as the research object, and the trend charts of the three stocks are shown in Figure 2:

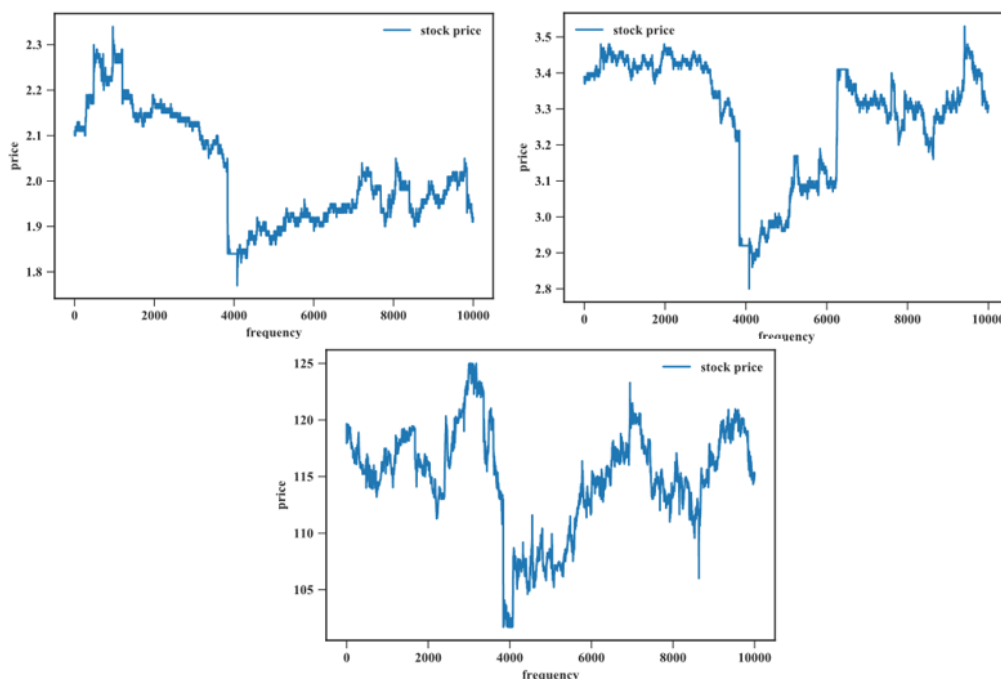


Fig. 2 Price movements of the three stocks

3.2 Comparative Analysis

In this paper, 2, 3, 4, 5, and 6 were selected as parameters of LSTM, and different training sets were selected to test the robustness of the model. The results of mean square error (MSE), mean relative error (MRE) and posterior error (BE) of the experimental model and the traditional LSTM model are shown in Table 1-Table 3 and Figure 3-Figure 5. Specifically, the mean square error, absolute relative error and posterior error.

Table 1. Comparison of results for each error value under stock symbol SH600777

	AFWLSTM-MSE	AFWLSTM-MRE	AFWLSTM-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
0	0.855639	0.253375	0.241292	2.513752	0.789939	0.773137
1	0.817311	0.24522	0.235451	2.52493	0.791701	0.750043
2	0.831803	0.24214	0.235808	2.519568	0.790863	0.757062
3	0.794595	0.231818	0.230453	2.529805	0.792468	0.750314
4	0.531104	0.195375	0.191515	2.496183	0.787189	0.77832
5	0.375754	0.171795	0.162145	2.542361	0.794428	0.716277
6	0.380699	0.170384	0.163105	2.539822	0.794043	0.734327

Table 2. Comparison of results for each error value under stock symbol SH600811

	AFWLSTM-MSE	AFWLSTM-MRE	AFWLSTM-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
0	6.731247	0.448425	0.153576	6.840357	0.800765	0.406727
1	3.204504	0.358175	0.107677	6.865003	0.802183	0.380903
2	3.02662	0.344436	0.104667	6.870349	0.802496	0.3807
3	2.723063	0.329229	0.099366	6.814898	0.799286	0.402925
4	2.271916	0.300835	0.090874	6.909323	0.804759	0.369655
5	2.03964	0.285661	0.085906	6.837844	0.800632	0.40049
6	1.816814	0.268538	0.081241	7.009808	0.810554	0.345899

Table 3. Comparison of results for each error value under stock symbol SH603833

	AFWLSTM-MSE	AFWLSTM-MRE	AFWLSTM-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
0	2458.415	0.592209	0.328194	13059.45	0.995181	0.957476
1	2413.124	0.584789	0.32692	13056.6	0.995074	0.956679
2	1742.684	0.540041	0.280426	13061.33	0.995252	0.957671
3	1649.267	0.523934	0.272688	13058.1	0.99513	0.957238
4	1634.351	0.517279	0.271073	13062.18	0.995282	0.958441
5	1525.464	0.49855	0.260994	13062.05	0.995277	0.958435
6	1408.653	0.479574	0.251845	13060.74	0.995229	0.957698

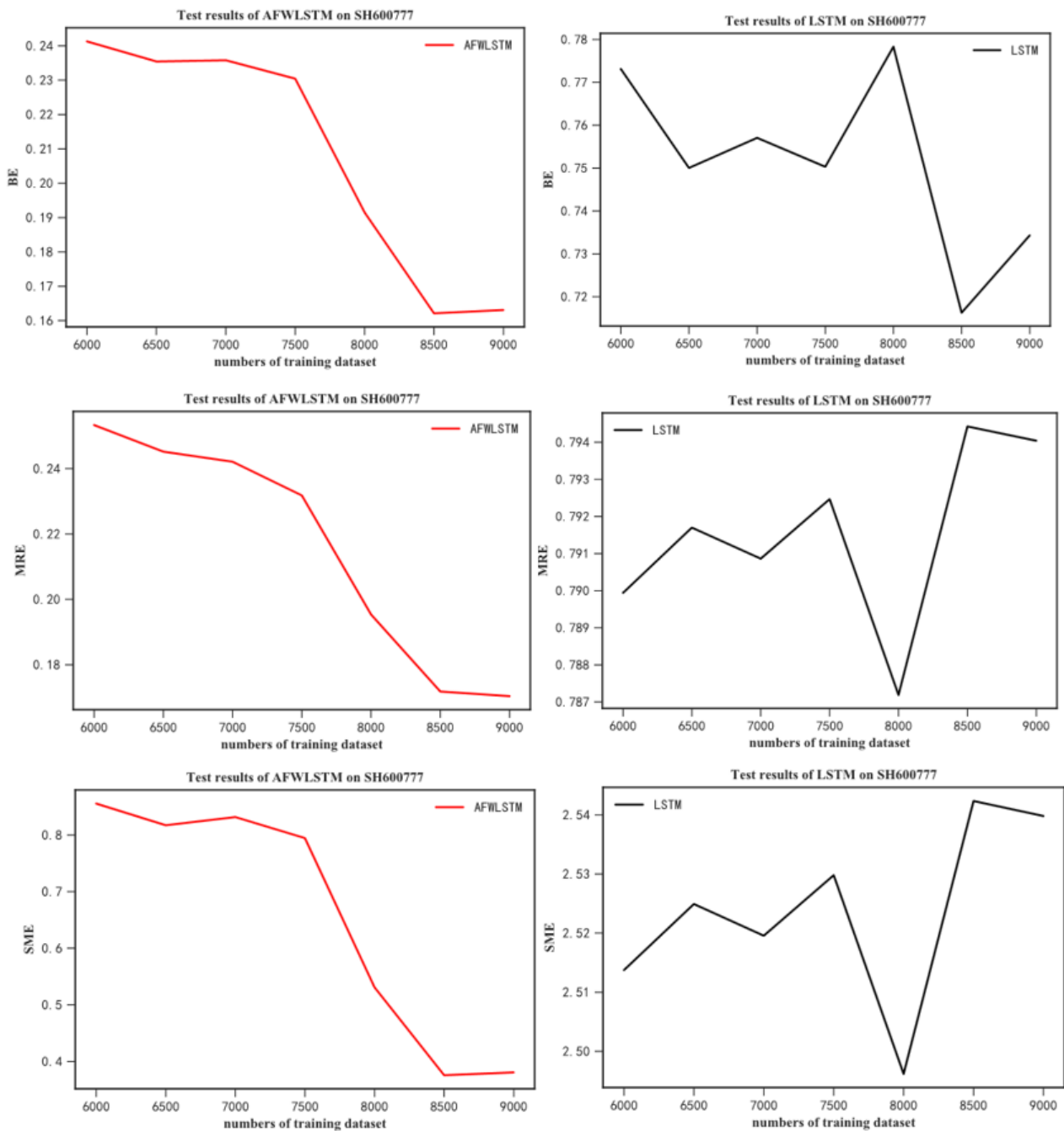


Fig. 4 Trend comparison of error values between LSTM and AFWLSTM of SH600777

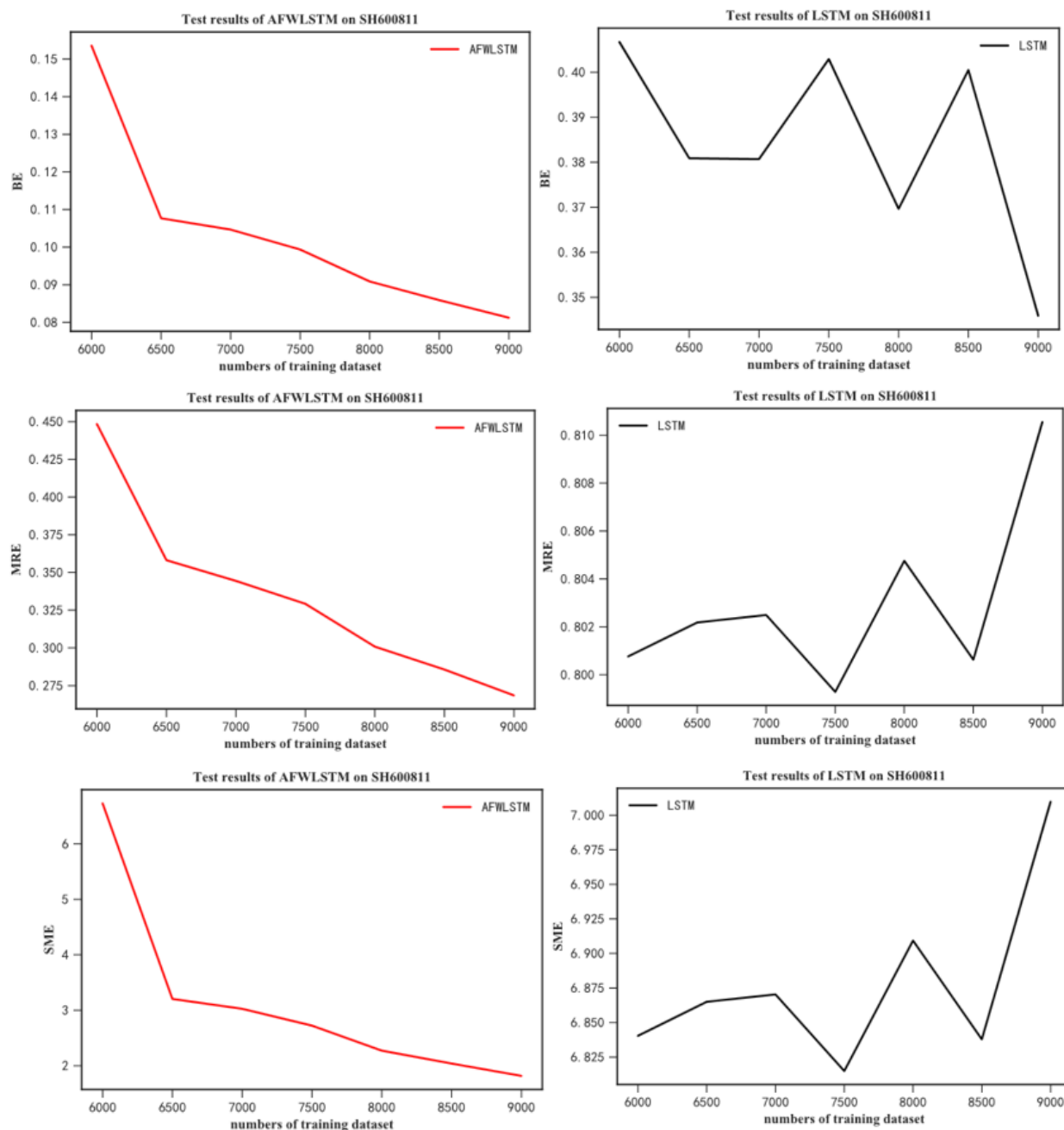


Fig. 5 Trend comparison of error values between LSTM and AFWLSTM of SH600811

According to the above trend diagram, we can clearly see that the standard LSTM neural network and the improved method have obvious differences in the results of predicting the three stock prices respectively, and the improved method has smaller error and a downward trend. On the one hand, it shows that the model has higher prediction accuracy, on the other hand, it also shows that the model has certain robustness. In summary, the experimental results prove that the proposed method has a significant improvement on the original LSTM neural network. The main reasons may be as follows: Firstly, while reducing the dimension of time series, the improved method uses Hermite basis function expansion to extract the function information and nonlinear information, which reduces the computational complexity and improves the estimation accuracy. Secondly, the model averaging method is used to balance the bias and variance of the prediction model, which reduces the variance of the model and improves the generalization error.

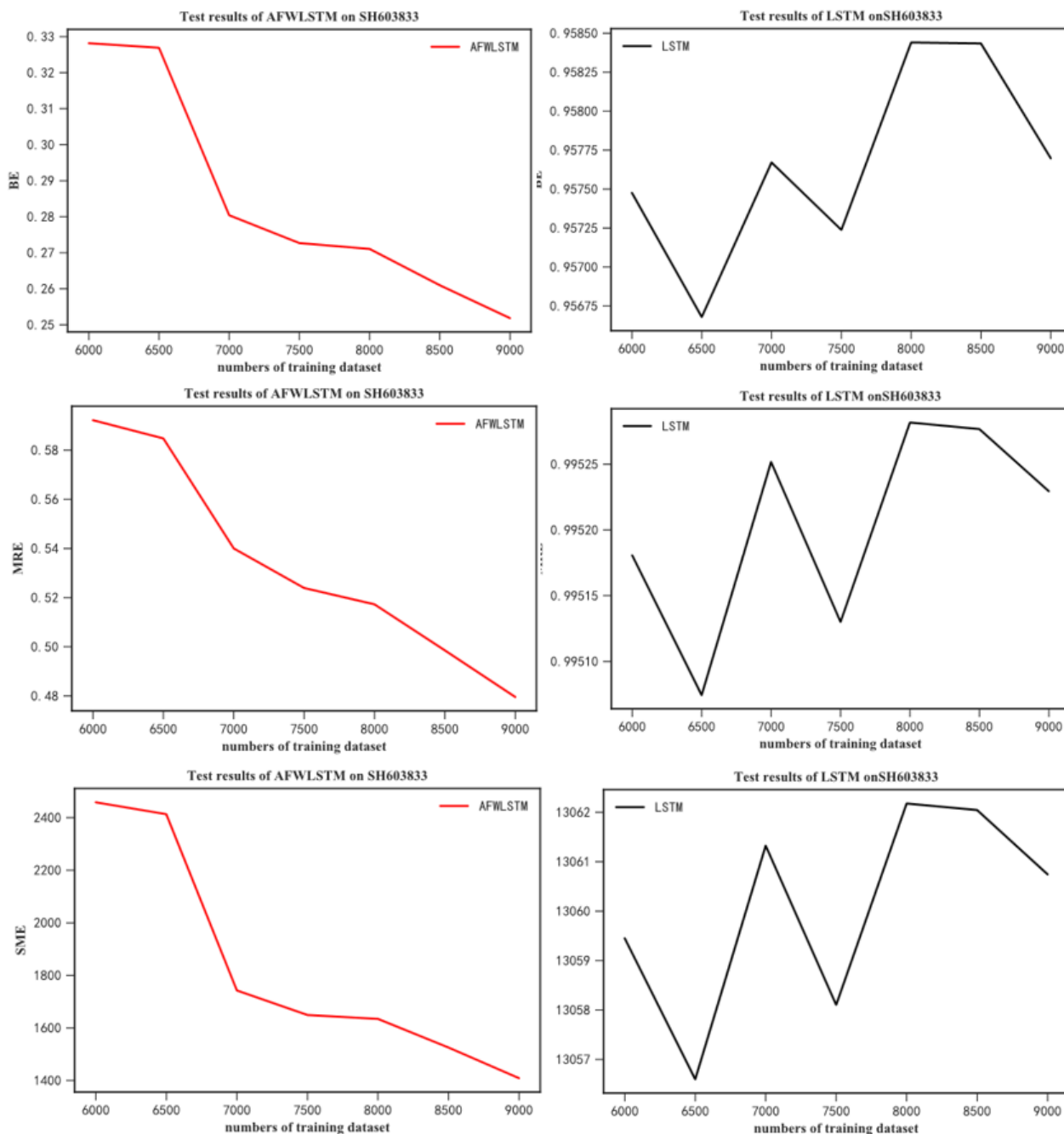


Fig. 6 Trend comparison of error values between LSTM and AFWLSTM of SH603833

4. Conclusion and Discussion

In this paper, we propose an LSTM neural network method based on Hermite basis function expansion and ensemble learning improvement to predict high-frequency stock prices. First, we use the function information extracted from the Hermite basis function expansion to predict the residual sequence obtained by the LSTM neural network, and then use the Bagging framework to process it, so as to capture the underlying model structure and balance the variance and bias of the model. In addition, we consider the model averaging method based on distance covariance weighting to optimize the number of prediction periods of LSTM neural network. The above data analysis results show that the proposed method can effectively optimize the prediction accuracy of the LSTM neural network, and has certain robustness.

For future research directions, this paper firstly uses Hermite basis function expansion method to improve the LSTM neural network, but whether other different basis function expansions have other

effects on the improved results of the LSTM neural network needs to be further verified. Secondly, in this paper, we use the model averaging method based on distance covariance weighting to optimize the prediction period number of LSTM neural network, so whether other weighting methods or model averaging methods have better optimization effect on the prediction period number of LSTM neural network needs further testing. Finally, for dense time data or sparse time series data, whether the proposed method is still applicable, and how to deal with time series data with different frequencies, the above issues need further study.

References

- [1] TAN Yue. Research on CNN-BILSTM stock price prediction model and quantitative trading strategy based on Attention mechanism [D]. Harbin industrial university, 2021. DOI: 10.27061 / , dc nki. Ghgdu. 2021.003454.
- [2] Yang Haimin, Pan Zhisong, Bai Wei. A survey of time series prediction methods [J]. Computer science, 2019, 46(01): 21-28.
- [3] Zhou Hua, Pang Jiaren, Wang Ziyue. Based on the method of principal component analysis of China's financial systemic risk measurement study [J]. Journal of insurance research, 2018 (04) : 3-17. DOI: 10.13497 / j.carol carroll nki is. 2018.04.001.
- [4] QIN X Y. Research and application of model selection and model averaging method [D]. Shandong University of Technology, 2017.
- [5] [5] Li Yuelong, Tang Dehua, Jiang Guiyuan, Xiao Zhitao, Geng Lei, Zhang Fang, Wu Jun. Short-term traffic flow prediction with residual LSTM based on dimension weighting [J]. Computer engineering, 2019, 45(06): 1-5. DOI: 10.19678 / j.issn.1000-3428.0052120.
- [6] Song Yuping, Sun Yankun. Prediction of High Frequency Financial time series: An improved ARIMA model based on Adaptive filtering method [J]. Journal of jilin province industrial and commercial college, 2021 ((02) : 82-86. The DOI: 10.19520 / j.carol carroll nki issn1674-3288.2021.02.012.
- [7] Xie Heliang, Zhang Mound. Kalman filter in the application of high frequency financial time series models to predict [J]. Journal of statistics and decision, 2017 (13) : 82-84. The DOI: 10.13546 / j.carol carroll nki tjyjc. 2017.13.019.
- [8] Yang Yun, Chen Liang, Fan Chongjun, Yang Jin. Application of Improved LOBNN & AR-GARCH Model in Stock Forecasting [J]. Operations research and management, 2021, 30(10): 153-158.
- [9] Shen Zejun, Yang Wenyuan. Application of BP neural network based on granular computing thinking in financial trend forecasting [J]. Journal of microcomputer systems, 2019, 40(03): 527-532.
- [10] Zhao Dandan, Ding Jianchen. Research on Systemic Risk Warning of China's Banking Industry -- Based on SVM Model [J]. Journal of international business (foreign economic and trade university), 2019 (4) : 100-113. The DOI: 10.13509 / j.carol carroll nki ib. 2019.04.008.
- [11] Shi Jian-nan, Zou Jun-zhong, Zhang Jian, Wang Chun-mei, Wei Zuo-chen. Based on the DMD - LSTM model stock price time series prediction research [J]. Computer application research, 2020 ((03) : 662-666. DOI: 10.19734 / j.i SSN. 1001-3695.2018.08.0657.