Stock price series prediction optimization framework based on function information stacking and model averaging

Hao Meng[1,*], Tiantian Shi[2]

[1]School of economies, Anhui University of Finance and Economics, Bengbu, 233000, China

[2]School of Economics, Lanzhou University, Lanzhou, 730106, China

[*]Corresponding author: mh010913@163.com

Abstract. Stock price series prediction has always been a hot issue in the field of quantitative finance. The commonly used models include ARIMA, GARCH, LSTM neural network and BP neural network. Aiming at these models, this paper proposes an optimization framework based on function information stacking and model averaging. The proposed method uses intra-day price information as auxiliary information and extracts functional features based on functional principal component analysis (PCA). Considering that the underlying model structure between the characteristic variables and the residual series obtained from the original time series prediction model is unknown, this paper uses Stacking method to enhance the data of the characteristic variables to reduce the impact of noise information on the prediction model. In addition, to solve the parameter optimization problem of the original model, this paper proposes a model averaging method using distance covariance weighting to deal with it. In the actual data analysis, this paper takes the LSTM neural network as an example to explore the effectiveness and robustness of the proposed method, and the results show that the proposed method has certain competitiveness. Finally, the proposed optimization method can be used to improve other time series prediction models.

Keywords: Stock Price Series Prediction, Time Series Model Optimization, Model Averaging, Distance Covariance.

1. Introduction

The problem of stock price series forecasting is essentially a time series forecasting problem. The common time series forecasting methods include ARIMA, GARCH, gray level forecasting and so on.Yang Chunjing (2018)[2]Applying ARIMA to short-term forecasting of log-return data of USD and RMB;Li Qun (2020)[3]With grey theory as the core, realized the prediction of stock series.These methods all use stationary time series data, and the conditions are harsh, ,which eventually leads to the reduction of data information, and can only make short-term prediction but not long-term prediction.To address these problems, Huo Jiangyou (2018)[4]By introducing wavelet analysis, the stock price series was regressed into two series to improve the prediction accuracy of the model.Liang Ying (2021)[5]A quadratic dimensionality reduction method combining functional dynamic principal component analysis and factor analysis was used to predict the stock price series.However, the relationship between the characteristic variables and the corresponding variables after dimensionality reduction is not taken into account.

With the development of data science, the deep learning technology is increasingly perfect. The stock price series can be predicted by LSTM, BP neural network and other methods, which can avoid the stationarity assumption, and do not need to reduce the dimension of data for medium and long-term prediction.Mei Xu, Fang Wang (2015)[6]BP neural network was combined with symbolic time series analysis method to predict stock price fluctuations using historical data.Cunhao Li (2019)[7]It is proved that LSTM time series forecasting model has more advantages than traditional stock price time series model.Cheng Wenhui, Che Wengang (2022)[8]This paper proposes a financial time series prediction algorithm based on quadratic decomposition and Long Short term memory (LSTM) network, which has good prediction accuracy in long-term prediction.Deep learning methods make up for some defects of traditional econometric models, but without considering the information of time series itself, it may not be enough to predict future changes.

In summary, an optimization framework based on function information stacking and model averaging is proposed in this paper. The framework can be used to improve any time series prediction model. The main advantages of the framework are as follows: first, the basis function expansion and stacking method are used to solve the problem of dimension disaster, and the variance and bias of the prediction model are balanced.Second, the model averaging method based on distance covariance weighting is used to solve the problem of optimizing the parameters of the original model.Experimental results show that the proposed method can significantly improve the prediction accuracy of the original model and has robustness.

2. Theory and Methods

2.1 Functional Principal component analysis

Functional Principal component Analysis (FPCA) is a generalization of traditional PCA.Similar to the situation in the traditional multivariate principal component analysis, its specific derivation is as follows:

$$X_j(t) = \mu_p(t) + \sum_{i=1}^{\infty} \xi_i^p\left(X_j(t)\right) f_i(t) \tag{1}$$

Where, is the mean function, and is the projection of the centralized function on the principal component function, namely:

$$\xi_i^p(X_j(t)) = \int_T (X_j(t) - \mu_p(t)) f_i(t) dt \tag{2}$$

That is:

$$\sum_{i=1}^{\infty} \int_T (X_j(t) - \mu_p(t)) f_i(t) dt f_i(t) X_j(t) - \mu_p(t) = \sum_{i=1}^{\infty} \int_T (X_j(t) - \mu_p(t)) f_i(t) dt f_i(t) \tag{3}$$

$X_j(t) = X_j(t) - \mu_p(t)$, there are:

$$X_j(t) = \sum_{i=1}^{\infty} \int_T X_j(t) f_i(t) dt f_i(t) \tag{4}$$

Remember,then:,where is the orthogonal basis.

$$\xi_i^X = \int_T X_j(t) f_i(t) dt \;\; X_j(t) = \sum_{i=1}^{\infty} \xi_i^X f_i(t) \;\; f_i(t), i = 1,2,...,K, K \in N$$

All the information about the function is contained in the Karhunen-Loeve expansion coefficients, and the random function approximated by the Karhunen-Loeve expansion converges to the real random function in probability.XTherefore, we can use the Karhunen-Loeve expansion coefficient matrix to replace the original time series information. For the selection of K, the cumulative contribution rate truncation method is adopted in this paper.

2.2 Stacking method

In the research process of machine learning models, Stacking can effectively integrate multiple Predictor results, which is an important means to improve the model score.

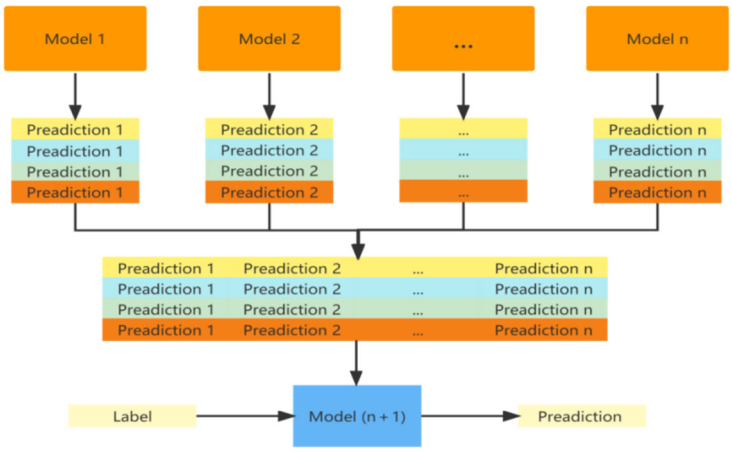


Figure 1.Stacking process

The main meaning of Stacking is that training another model on the predicted results of the model is like Stacking another model on the original model.In this way, different data information extracted by different models can be collected.Due to the noise of the data, different models tend to perform well on different features of the data, but also have poor performance.Stacking can extract the parts with good features of each model and discard the parts with poor performance at the same time, which can effectively optimize the prediction results and improve the prediction score. The specific process is shown in Figure 1:

2.3 LSTM Neural Network

As a method of deep learning, LSTM (Long short-term Memory) algorithm is a Long short-term Memory network, which is a temporal recurrent neural network.LSTM is mainly used to solve the long-term dependence problem in RNN (recurrent neural network).LSTM is also a special recurrent neural network, so it also has a chain structure, but it has a different structure than the repeating module of recurrent neural network. LSTM has four neural network layers, each of which interacts with each other in a special way, rather than a single simple neural network layer.As shown in Figure 2, the data of the previous period is input on the left side, and then D is multiplied by the forgetting gate coefficient F by a multiplier, and then multi-vector linear superposition is performed, and the current state D is finally output.Its single LSTM neurosource is shown in Figure 2 below:

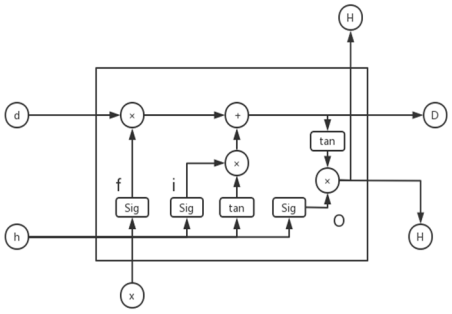


Figure 2 .LSTM model structure

2.4 Stock price series prediction optimization framework based on function information stacking and model averaging

Based on time series high dimension, big noise and the original information there might not be enough to predict task, we use the intraday prices as auxiliary information, considering the intraday price can be regarded as a random process of the implementation at a time, so we consider using functional principal component analysis as a way to extract characteristic information and using Stacking method to forecast the residual sequence.Finally, considering the parameter tuning problem

of LSTM, we use the model averaging method based on distance covariance to deal with it.Remember the Euclidean distance $d_{ij}^{X} = d(X_i, X_j)$ of the sample point X_i from X_j in R^p and Remember the Euclidean distance $d_{ij}^{Y} = d(Y_i, Y_j)$ of the sample point X_i from X_j in R^q， , and the specific formula of distance covariance is as follows:

$$v_n^2(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} d_{ij}^{X} d_{ij}^{X} + \frac{1}{n^2}\sum_{i,j=1}^{n} d_{ij}^{X} \frac{1}{n^2}\sum_{i,j=1}^{n} d_{ij}^{Y} - \frac{2}{n^3}\sum_{i=1}^{n}\sum_{i=1}^{n} d_{ik}^{X} d_{il}^{Y} \tag{5}$$

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.This improved framework is called the optimization of stock price series prediction based on function information stacking and model averaging.The specific process is shown as follows:

Step1: perform functional principal component analysis (FPCA) on intraday price data;

Step2: conduct basis expansion processing on the intraday price data to get the basis expansion coefficient;

Step3: LSTM neural network is used to predict and obtain the forecast series Y1;

Step4: Compare the real value of data y with the predicted value Y1Make the difference to get the residual series, and pair it with the intraday price basis expansion coefficient;

Step5: Stacking model regression to get the new residual series, and with the predicted value Y1Add and add to get Y2;

Step6: Calculate y vs2The distance correlation coefficient between Y and Y is normalized and taken as y2The weight of;

Step7: y2Multiply with the weight matrix to find the final ynew;

Step8: y1, ynewAnd the real value y was substituted into the evaluation function to measure the model accuracy and compared.

3. The data analysis

3.1 Data Sources

The data used in this article is from the Wind database(https://www.wind.com.cn） randomly selected from three groups of stock data, stock code SH600777 represent shandong new energy co., LTD., mainly engaged in real estate development and sales operations;The stock code SH600811 stands for Orient Group Co., LTD., which is mainly engaged in grain and oil purchase and sales business;The stock code SH603833 represents Opai Home Furnishing Group Co., LTD., which is mainly engaged in the design, research and development, production, sales and installation of customized integrated home furnishing products. The specific results are shown in Figure 3-8:

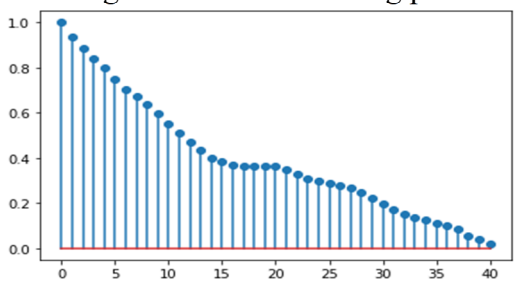


Figure 3.Autocorrelation coefficient of SH600777

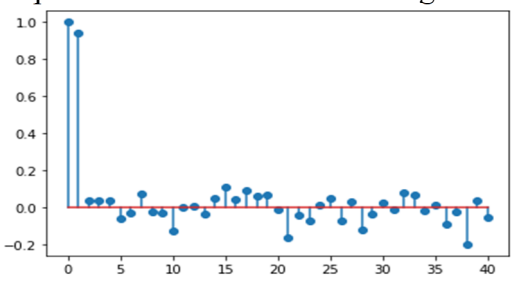


Figure 4.Partial autocorrelation coefficient diagram of SH600777

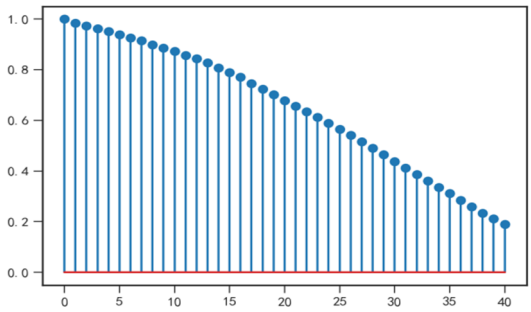


Figure 5.Autocorrelation coefficient of SH600811

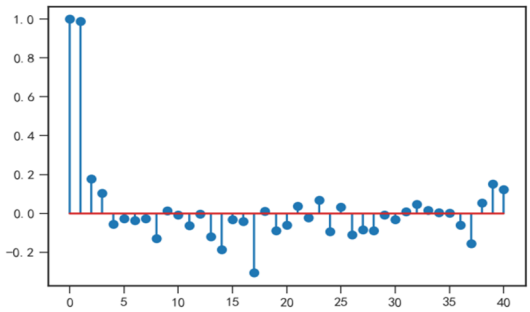


Figure 6.Partial autocorrelation coefficient of SH600811

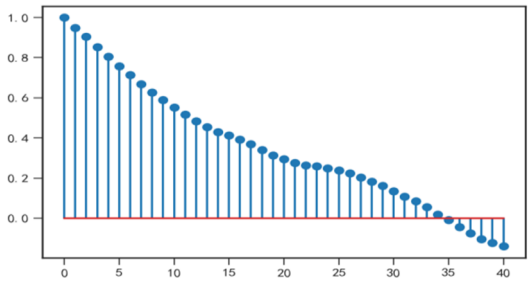


Figure 7.Autocorrelation coefficient diagram of SH603833

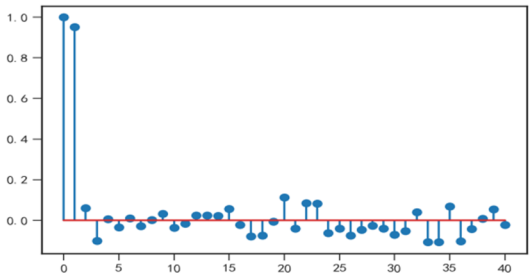


Figure 8.Partial autocorrelation coefficient diagram of SH603833

Through the analysis of the prediction results of three groups of stock data, it verifies whether the improved method in this paper is effective.Firstly, the stationarity test was conducted on the three groups of experimental data, and it was concluded that the p-value value was greater than 0.05, which was not stable.

3.2 Comparison Results

The experimental parameters of LSTM neural network were selected as 2, 3, 4, 5, and 6. The total number of samples was the opening price data of 243 days, and the frequency of intra-day prices was 240. We selected different training sets to test the robustness of the model. Specifically, the mean square error, absolute relative error and posterior error can be calculated by the following formula: The mean square error (MSE) is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{6}$$

The average relative error (MRE) is:

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{7}$$

The posterior error (BE) is:

$$BE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left(y_i - \hat{y}\right) - \mu\right)^2}} \tag{8}$$

The specific experimental results are as follows:

Table 1. Data table of SH600777

	FSLSTM-MSE	FSLSTM-MRE	FSLSTM-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
0	0.001127557	0.012707472	0.224621421	2.033713303	0.813400652	0.424432342
1	0.001041096	0.012140597	0.215993825	2.035066162	0.813619528	0.414936167
2	0.000987959	0.012016669	0.21020873	2.048371328	0.816448854	0.423420481
3	0.000975332	0.011803102	0.209306924	2.030717947	0.812681026	0.408106679

4	0.001032451	0.012036655	0.215375923	2.036944406	0.813756583	0.433688003
5	0.001016921	0.011837557	0.213383223	2.051811209	0.817145726	0.425811562
6	0.000964491	0.011344717	0.208144501	2.046823083	0.816097124	0.410154937
7	0.0009914	0.011409344	0.211265851	2.053207643	0.817637081	0.422899713
8	0.000854396	0.011001009	0.196167215	2.057274161	0.818541717	0.422961015
9	0.00082659	0.010720626	0.192915682	2.056524852	0.818442387	0.419096171

Table 2 .Data table of SH600811

	FSLSTM-MSE	FSLSTM-MRE	FSLSTM-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
0	0.001181003	0.006283366	0.055697927	1.27084569	0.089327067	0.066337453
1	0.001143685	0.006124453	0.054900999	1.27236888	0.089382853	0.06632185
2	0.001082489	0.005951348	0.053389298	1.271768764	0.089355296	0.066466778
3	0.00109087	0.005946707	0.053648737	1.27092218	0.089306989	0.066892712
4	0.001130145	0.006031923	0.05460632	1.272599614	0.089389953	0.066371525
5	0.001129965	0.005995812	0.054585453	1.272668105	0.089386675	0.066474587
6	0.001373413	0.006340277	0.060135444	1.272729653	0.089389574	0.066472041
7	0.001320271	0.006240314	0.058999572	1.27044555	0.08934817	0.065473581
8	0.001212652	0.006007142	0.056482779	1.270338676	0.089351529	0.065286245
9	0.001194501	0.005897781	0.056031093	1.278939375	0.089622877	0.066229848

Table 3 .Data table of SH603833

	FSLSTM-MSE	FSLSTM-MRE	FSLSTM-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
0	25.8680832	0.020767881	0.413221525	127.5298868	0.099572834	0.986343913
1	26.8564433	0.020838261	0.420845435	127.5261315	0.099571146	0.986433042
2	28.24938446	0.021009653	0.431453243	127.5345687	0.099573241	0.987027634
3	22.48146062	0.019507425	0.386350123	127.524227	0.099569741	0.986733664
4	25.05062001	0.019625241	0.408747204	127.519052	0.09956838	0.986402165
5	12.5294098	0.014220178	0.296811015	127.5127704	0.099566202	0.986252996
6	6.499607583	0.010644016	0.2184198	127.5030336	0.099562469	0.986163747
7	5.395646586	0.009531257	0.200801256	127.5340841	0.099576594	0.985354083
8	3.346447933	0.008575265	0.158964771	127.4936971	0.099561261	0.98496259
9	3.043229019	0.008045475	0.152005083	127.4921079	0.099561417	0.984594015

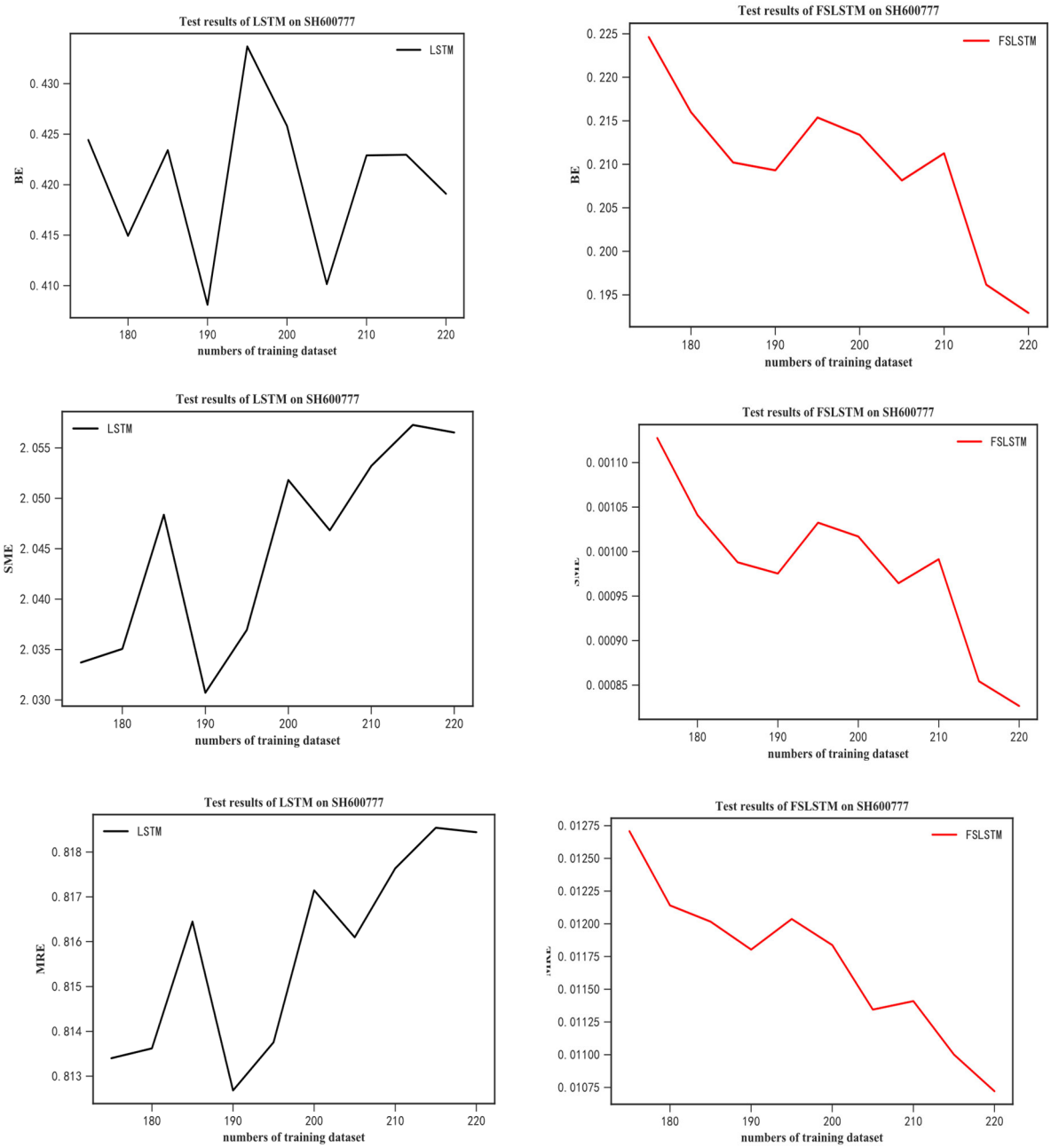
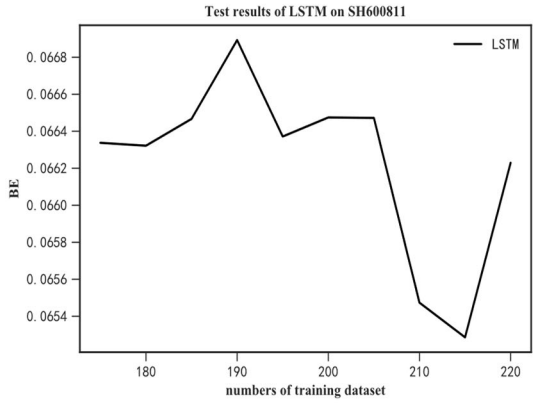
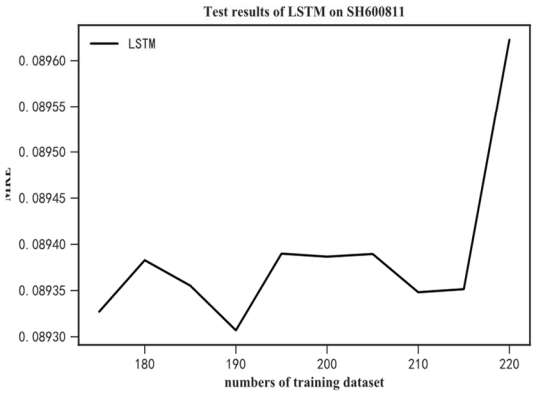
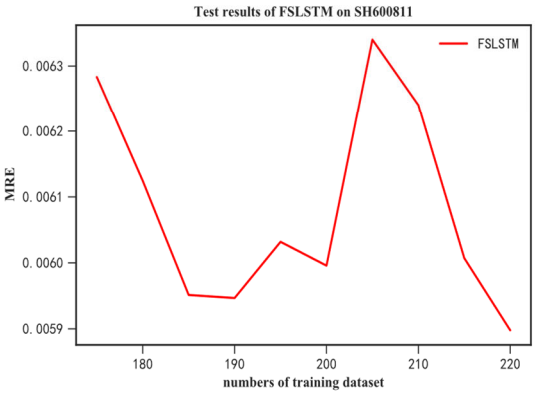


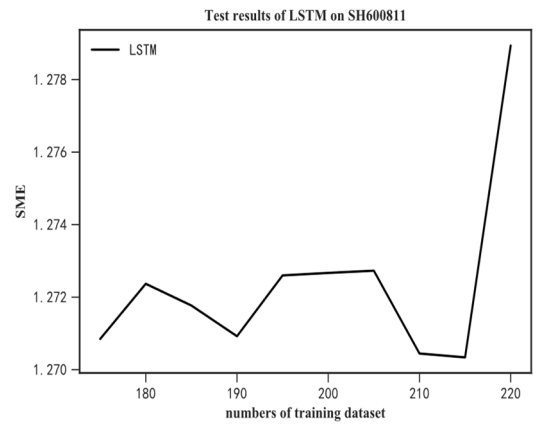
Figure 9 Robustness check of the LSTM of SH600777 against the improved method

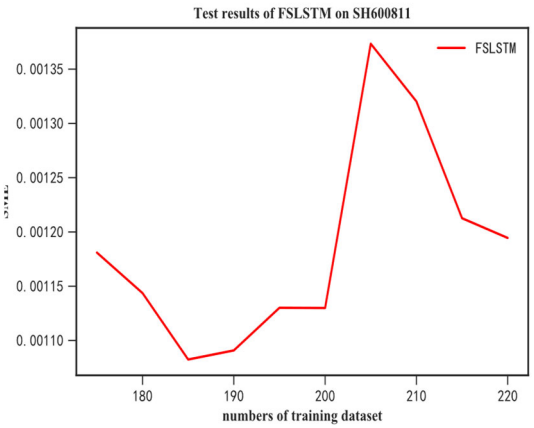

Test results of LSTM on SH600811
LSTM
BE
0. 0668
0. 0666
0. 0664
0. 0662
0. 0660
0. 0658
0. 0656
0. 0654
180
190
200
210
220
numbers of training dataset


Test results of FSLSTM on SH600811
FSLSTM
BE
0. 060
0. 059
0. 058
0. 057
0. 056
0. 055
0. 054
180
190
200
210
220
numbers of training dataset


Test results of LSTM on SH600811
LSTM
0. 08960
0. 08955
0. 08950
0. 08945
0. 08940
0. 08935
0. 08930
180
190
200
210
220
numbers of training dataset


Test results of FSLSTM on SH600811
FSLSTM
MRE
0. 0063
0. 0062
0. 0061
0. 0060
0. 0059
180
190
200
210
220
numbers of training dataset


Test results of LSTM on SH600811
LSTM
SME
1. 278
1. 276
1. 274
1. 272
1. 270
180
190
200
210
220
numbers of training dataset


Test results of FSLSTM on SH600811
FSLSTM
0. 00135
0. 00130
0. 00125
0. 00120
0. 00115
0. 00110
180
190
200
210
220
numbers of training dataset

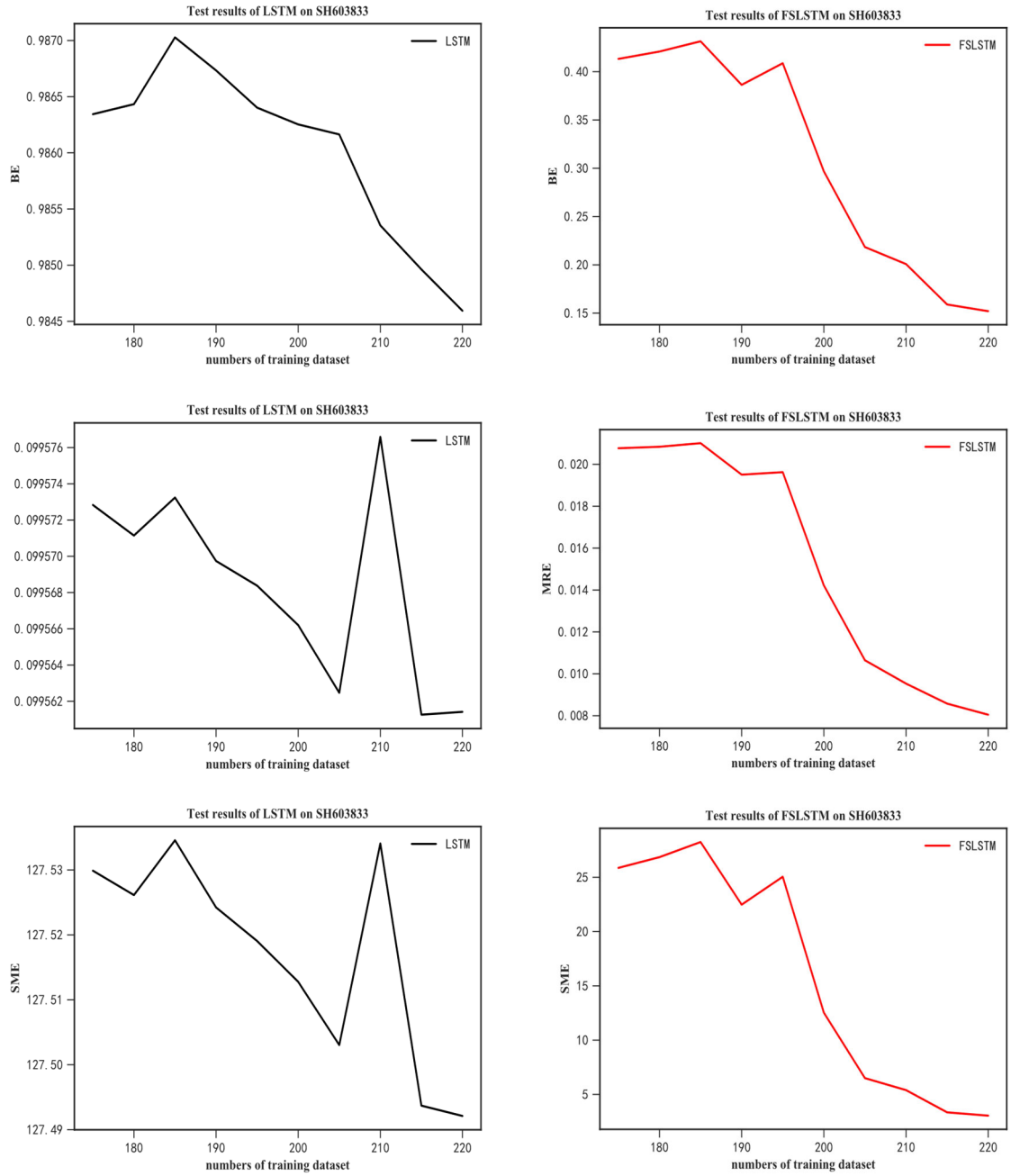


Figure 10 Robustness check of the LSTM of SH600811 against the improved method

According to the comparison of experimental results, this paper analyzes the results from two aspects. First, through the comparison of prediction accuracy, it can be seen from Table 1 to Table 3 that the prediction accuracy of the improved model is greatly improved. The reason may be that the intra-day price information is used as the auxiliary information to fit the residual prediction of the original LSTM neural network, so that the accuracy is improved. In addition, the stacking framework is used to reduce the variance and bias of the model prediction between the intraday price basis expansion coefficient and the residual series, and the influence of noise on the model is reduced. The optimization problem of the original model parameters is solved by the model averaging method weighted by distance covariance. Second, through the comparison of model robustness, the error should be consistently reduced with the increase of the number of samples in the training set, but the LSTM still shows extremely unstable fluctuations, while the improved method is relatively normal, so the proposed model is still robust.

4. Conclusion and Discussion

In this paper, the traditional LSTM neural network forecasting time series model of stock prices is optimized.Firstly, functional principal component analysis (FPCA) and basis expansion were performed on the intraday price data to obtain the basis expansion coefficient. The basis expansion coefficient was paired with the residual of the real value of the data and the predicted value of the traditional LSTM neural network. The model forecasts average as new characteristics, for the test set all base estimator weighted correction, get a new residual sequence and with the original forecast after add and get new forecast, at last, there will be a new forecast model of average method and normalized processing of distance correlation coefficient weighting matrix multiplication, it is concluded that the final forecast.There are three main research directions in the future: 1. Analyze whether the experimental results can be further optimized by using different basis function expansion methods to process intraday price data; 2.2. This paper only improves the traditional LSTM neural network model by forecasting and analyzing the unstable stock price time series, while the improvement effect of other time series prediction models remains to be studied.3. This paper takes intraday price data as auxiliary data, how to integrate other information, such as fundamental data (financial data of the company, flow data), to further improve the accuracy?

References

[1] Liang Zenghui. Research on the Relationship between volume and Price in Chinese Stock Market [D]. Southwestern University of Finance and Economics,2013.

[2] Yang chunjing. Stock price prediction based on time series model [J]. Western leather,2018,40(12):98-99.

[3] Li Qun. Research and Application of Fuzzy Similarity and Grey Model in Stock price Inflection Point Prediction [D]. Hebei university of technology, 2020. DOI: 10.27105 /, dc nki. Ghbgu. 2020.000081.

[4] Huo Jiangyou. Frequency division combination prediction of stock price volatility based on wavelet multi-resolution decomposition [D]. Jiangxi University of Finance and Economics,2018.

[5] [Liang Y. Application of functional time series analysis method in high frequency stock price prediction [D]. Xinjiang university, 2021. DOI: 10.27429 /, dc nki. Gxjdu. 2021.001340.

[6] Xu M, Wang F. Study on financial volatility based on BP neural network and symbolic time series [J]. Journal of wuhan university of technology (information and management engineering edition),2015,37(04):456-460.

[7] [Li C H. Research on event-driven stock index futures trading strategy based on LSTM model [D]. Southwestern university of finance and economics, 2019. DOI: 10.27412 /, dc nki. Gxncu. 2019.000977.

[8] Cheng Wenhui, Che Wengang. Research on financial time series forecasting algorithm based on quadratic decomposition and LSTM [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition),202,34(04):638-645.]

[9] [Li Xiaojun, Tang Pan. Stock price forecasting based on technical analysis, fundamental analysis and deep learning [J]. Statistics and decision, 2022, 38 (02) : 146-150. The DOI: 10.13546 / j.carol carroll nki tjyjc. 2022.02.029.

[10] Cheng Chaozhi. Based on the deep study of financial time series prediction research [D]. University of electronic science and technology, 2021. The DOI: 10.27005 /, dc nki. Gdzku. 2021.004845.

[11] Chen Y F. Prediction and anomaly detection of high-frequency financial time series based on LSTM and autoencoder [D]. Sichuan university, 2021. DOI: 10.27342 /, dc nki. Gscdu. 2021.000363.