Optimization of Stock Price Time Series Prediction Model based on Karhunen-Loève Expansion and Information Gain Weighted Integrated Regression

Mingjun Kong[1,#], Xunhao Zhang[2,#], Faqiang Xu[2,#,*]

[1]School of Mathematical Sciences, University of Jinan, Jinan City, Shandong Province, 250002, China

[2]College of Economic and Management, Tongji University, Shanghai City, 200092, China

[*]Corresponding author: xvfaqiang@163.com

[#]These authors contributed equally.

Abstract. Time series prediction model plays an important role in stock price prediction, such as ARIMA, LSTM neural network. However, due to the need for stationary assumption of time series itself and the problems of high dimension and high noise, the common time series prediction methods have limitations. Based on this, this paper propose a framework for the optimization of the stock price time series prediction model. The proposed method uses the intra-day price as the auxiliary variable and obtains the function feature information based on Karhunen-Loève expansion. Considering that the feature variables after dimension reduction still have problems such as information loss and irrelevant noise. This paper use data enhancement method to improve the effective information of feature variables and reduce the influence of irrelevant noise. Then, since the potential model structure between the characteristic variable and the residual sequence is unknown, this paper develop a weighted ensemble regression method based on information gain to balance the variance and deviation of the prediction model, thereby improving the prediction accuracy. The actual data analysis results show that the proposed method can greatly improve the fitting accuracy of ARIMA and LSTM neural networks for stock prices. Finally, the optimization framework can also be used for the prediction of average temperature, air quality and port cargo flow.

Keywords: stock price prediction, Karhunen-Loève expansion, data enhancement, and integrated learning.

1. Introduction

With the rapid development of economy, stock has become one of the most common ways of investment. For investors, in order to minimize investment losses and obtain higher returns, this paper need to choose a more accurate method to predict the trend of stock prices[1]. The price of stocks changes all the time. This series of data that depend on time changes is called time series data. Time series analysis can find out more rules and characteristics within the data in a large number of historical data, so as to make more accurate prediction of future trends. In the classical financial time series analysis, the most widely used econometric models are ARIMA model and GARCH model[2-3]. ARIMA (Autoregressive Integrated Moving Average Model) model requires that the time series data is stable, or is stable after differential differentiation. In essence, it can only capture linear relationship, but not nonlinear relationship. GARCH model (Generalized Autoregressive conditional heteroskedasticity model) is a regression model specifically tailored to financial data. In addition to the same as the ordinary regression model, GARCH further models the variance of the error, which solves the problem caused by the second assumption (constant variance) of time series variables in traditional econometrics. Shuya Xu and Xiaoying Liang (2019) [4] applied the optimization model based on ARIMA-GARCH model to the prediction of stock prices, fitted the closing price series, and conducted error analysis. The results showed that the prediction results of the fusion model were better than those of the single model. However, the classical time series prediction method itself needs stationary hypothesis, and the stock market is further affected by market factors, political factors, macroeconomic factors and industry factors, which further makes the stock price difficult to predict. Abstractly, the actual stock price time series data contain both linear and nonlinear information, and the irrelevant noise is large and the dimension is high. Based on this, the classical time series prediction method has limitations.

With the development of machine learning, researchers began to combine traditional econometric models with machine learning and apply them to time series analysis. Zhibin Xiong (2011) [5] constructed the RMB exchange rate prediction model by integrating ARIMA and neural network. The basic idea is to give full play to the prediction advantages of the two models in linear space and nonlinear space, that is, the data structure of the exchange rate time series is decomposed into two parts: linear autocorrelation subject and nonlinear residual. Guisheng Zhang and Xindong Zhang (2016) [6] proposed SVM-GARCH prediction model by combining support vector machine (SVM) and GARCH model. The model takes advantage of SVM to deal with high-dimensional nonlinear data, not only contains the historical data information of the stock index sequence itself, but also fuses the relevant change information of the surrounding stock market closely related to the target stock index data through the nearest neighbor mutual information. Qing Yang and Chenwei Wang (2019) [7] constructed a deep LSTM neural network combined with the most cutting-edge deep neural network technology at that time, and applied it to the prediction and analysis of 30 stock indexes in the world, and compared the prediction results with three control models (SVR model, MLP neural network and ARIMA model) in short, medium and long term. Hongbing Ouyang (2020) [8] also used the LSTM model to overcome the defects that the existing model cannot reflect the complex characteristics of financial time series data such as non-stationary, non-linear and sequence correlation, and the non-linear interaction between data. These methods take into account the linear and nonlinear information of the time series to be captured, but they do not take into account that the information of the original time series may not be sufficient to predict the future stock price. Moreover, since the time series are high-frequency and high-noise data, how to carry out feature extraction and data enhancement is also a problem to be solved. In addition, if the feature extraction technology is unsupervised, the potential model structure between the original stock price time series needs to be explored.

In order to solve the appeal problem, this paper proposes a framework for the optimization of stock price time series prediction model. The innovation of the proposed method mainly includes the following aspects: Firstly, the intraday price series is used to capture the residual sequence information obtained by the original time series prediction model. Secondly, Karhunen-Loève expansion not only reduces the dimension of the intra-day price series, but also extracts its function feature information. Moreover, data enhancement method is used to further improve the fitting ability of Karhunen-Loève expansion coefficient for the residual series, and reduce the influence of irrelevant noise. Finally, considering that the potential model structure between the feature information after data enhancement and the residual series is unknown, this paper develop a weighted integrated regression method based on information gain to balance the variance and deviation of the residual series prediction model, so as to improve the final stock price prediction accuracy.

The content of this paper is arranged as follows. The first part introduces the framework for the optimization of the stock price time series prediction model. The second part shows the actual data analysis. The third part is the review and summary of the full text.

2. Theory and Methods[9-11]

2.1 Karhunen-Loève expansion

The Karhunen-Loève expansion[9], also known as functional principal component analysis, is derived as follows：Where, is the mean function,and is the projection of the centralized function on the principal component function That is, this paper have:where is the orthogonal basis. All the information about the function is contained in the Karhunen-Loève expansion coefficients, and the random function estimated by the Karhunen-Loève expansion approximation converges to the true random function in a probabilistic manner. Therefore, this paper can use the Karhunen-Loève expansion coefficient matrix instead of the original time series information, and regarding the choice of K, this paper adopts the cumulative contribution rate truncation method.

2.2 Data augmentation

Stacking is a data augmentation method that relies on multiple base estimator integration and cross-validation mechanisms to achieve the goal of enhancing useful information and reducing noise in the data[10]. Specifically, the individual learners are called primary learners and the learners used for combination are called secondary learners or meta-learners, which this paper will describe specifically in the next subsection. For each primary learner, cross-validation is used to input data for model training, and the prediction results of each primary learner are used as new feature variables. The details are shown in Figure 1 below.

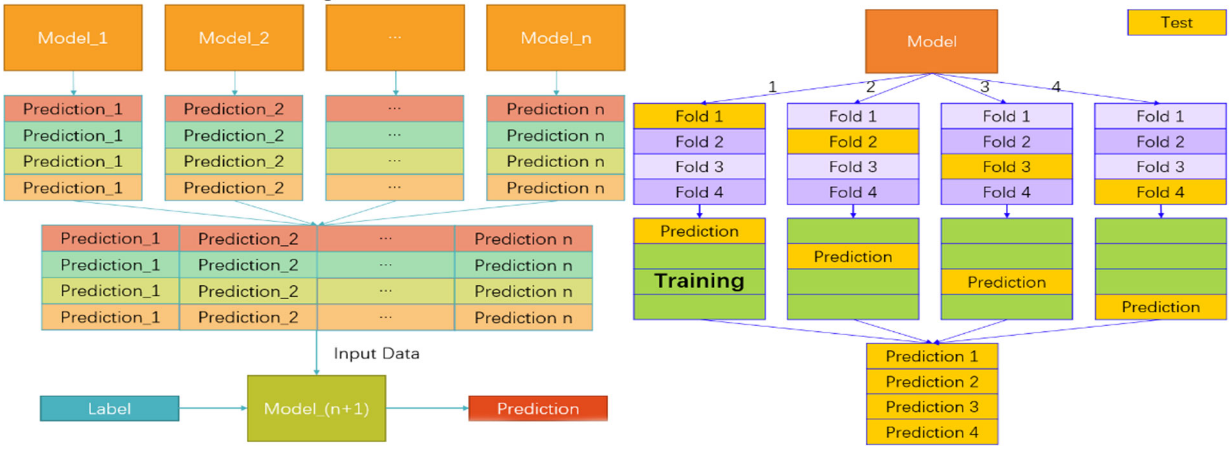


Figure 1 stacking framework

2.3 Weighted integrated regression based on mutual information

Bagging algorithm, whose full name is Bootstrap aggregating, is also known as bagging method. This algorithm was proposed and is the most famous representative of parallel integration learning methods. The Bagging algorithm can significantly improve the performance of learning methods that have low bias but large variance. In fact, by integrating the results of learning multiple Bootstrap resamples, the Bagging algorithm can reduce the variance of the initial learning method[11]. The basic algorithm is to first draw data from the original sample set with put-back. Then, for each Bootstrap resampling data set, a regression model is trained, and B regression models with known parameters are obtained. Finally, the average of these B models is calculated as the final result.

Since the prediction accuracy of each sub-model is considered to be different, this paper weight them using mutual information. The mutual information is equivalent to the Shannon entropy entropy minus the conditional entropy, i.e.

A higher mutual information represents a stronger correlation between two random variables. The specific operation applied to the weight setting is to calculate and normalize the mutual information between the predicted and true values of the response variables, and then use the normalized mutual information as the weights of each base estimator in the bagging framework.

2.4 Optimization framework

This paper optimize the prediction results of the two models to be optimized based on the predicted values of stock prices and their residual series of ARIMA and LSTM, and base expansion of intra-day prices, and predict the fitted values of the residuals through data augmentation methods and information gain weighting. This paper refer to this optimization method as FIWSK. The specific algorithm is as follows: a time series of stock prices is obtained.

On the one hand, the original model is used to predict this time series to obtain the sequence of predicted values.

Further, this paper can obtain the sequence of residuals between the predicted and true values.On the other hand, the time series of intra-day prices of the stock is obtained and subjected to Karhunen-Loève expansion.

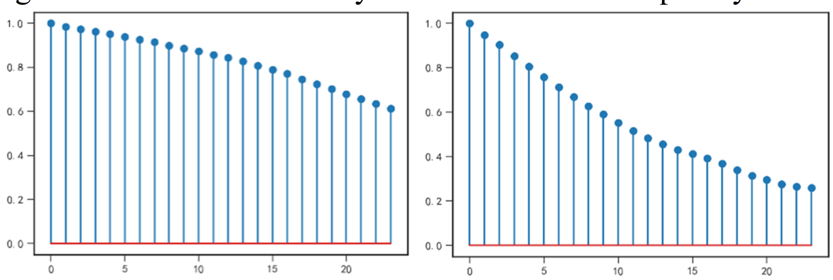
The principal component basis expansion coefficients are obtained. This paper denote the matrix of coefficients after the basis expansion as A. Next, the stacking data is augmented with the original model predictions and the stacking data augmented with the previous day's intraday prices to obtain the paired data, where the augmented data features are used as predictor variables and the residual series are used as response variables. The paired data are then brought into a Bagging framework based on mutual information weighting for training to obtain all model parameters of the framework, which includes the weight vectors.In the prediction stage, we bring the enhanced features of the intraday price data corresponding to the closing price to be predicted into the trained Bagging framework based on mutual information weighting to obtain the predicted residuals, which are weighted with the weight vectors to obtain the modified residual prediction values.

Finally, the prediction value of the original model is added to the modified residual prediction value to obtain the prediction value of the optimized model FIWSK.

3. Data analysis

3.1 Data sources and pre-analysis

In this section, three stocks are used as examples to empirically investigate the practical effects of the Karhunen-Loève expansion and information gain weighted integrated regression based stock price time series prediction optimization model. Specifically, three stocks from SSE and one stock from SZSE are selected for the empirical investigation. The basic information of the stocks is shown in the following table. SH600811 represents Dongfang Group, which belongs to the industry of grain and oil processing, SH603833 represents Oppai Home, which belongs to the industry of furniture manufacturing, and SZ300741 represents Huabao, which belongs to the industry of basic chemical. Using Wind database (https://www.wind.com.cn/), the intra-day prices and opening prices of the above three stocks in 240 groups (with a frequency of 5 minutes) were obtained, and three sets of raw financial time series were obtained. Before the pre-analysis of the functional feature information extracted after Karhunen-Loève expansion, this paper also have to pre-process the three sets of financial time series including the robustness test, and this paper mainly go to the pre-analysis of the financial time series from several perspectives such as autocorrelation plot, partial autocorrelation plot, and p-value. The details are shown in Figure 2.Where, the ordinate is autocorrelation or partial autocorrelation coefficient, and the abscissa is the lag order. From the preliminary analysis of the autocorrelation and partial autocorrelation plots of the three stocks, this paper know that the three randomly selected stocks have different degrees of volatility and are non-stationary time series. This satisfies our requirement to check whether the provided optimized model can improve the prediction accuracy of the original model in stochastic systems of different complexity.



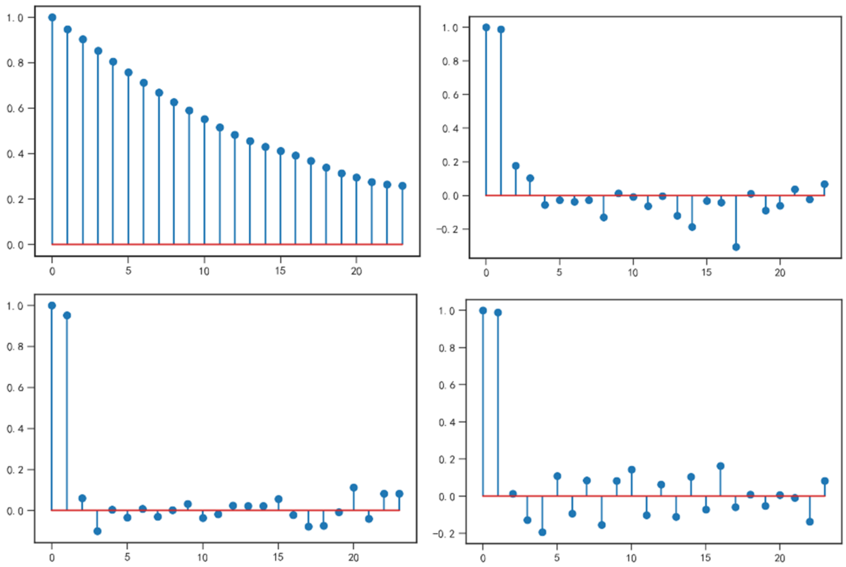


Figure 2 autocorrelogram and P-autocorrelogram of SH600811,SH600811 and SH603833

3.2 Comparative analysis of the results before and after model optimization

Since the feature vector obtained after dimensionality reduction using Karhunen-Loève expansion still suffers from information loss and irrelevant noise, this paper consider using the stacking framework for data augmentation to improve the effective information of the feature variables, reduce the influence of irrelevant noise, and obtain a new vector with less noise and more effective information. Subsequently, this paper construct multiple regression models, and put the subsets of the stacking-processed matrices into multiple regression models in turn to derive the results, and perform weighted linear combinations according to their correlations with the sample true, and finally obtain the optimized results of the model. The training set of the model contains 175,180,185,190,195,200,205,210,215,220 with 10 different training sets to test the accuracy and robustness of the model results. Using Python programming, this paper fit the three stocks through the ARIMA and LSTM neural network methods separately to derive the results and compare the obtained results with those of our stock price time series prediction optimization model based on Karhunen-Loève expansion and information gain weighted integrated regression to obtain the posterior error (BE) and mean square error (MSE) results are compared in the graph. The result are showed in Table 1-3 and Figure 3-5.

Table 1. Optimization results of ARIMA and LSTM (SH600811)

N	FIWSK1-MSE	FIWSK1-BE	ARIMA-MSE	ARIMA-BE	FIWSK2-MSE	FIWSK2-BE	LSTM-MSE	LSTM-BE
175	0.125637	0.576718	0.713995	1.365474	0.001679	0.066244	12.688349	0.671326
180	0.274476	0.787759	0.788031	1.407129	0.001442	0.061525	12.711334	0.661628
185	0.128997	0.580581	0.672285	1.334023	0.001228	0.056839	12.702678	0.664615
190	0.139730	0.607596	0.595220	1.253573	0.001606	0.065058	12.704303	0.665500
195	0.117335	0.548750	0.658714	1.316491	0.001509	0.063123	12.699451	0.670834
200	0.119794	0.561742	0.629570	1.290893	0.001512	0.063171	12.714810	0.665423
205	0.202712	0.717986	0.777209	1.425876	0.001566	0.064178	12.713702	0.664995
210	0.101212	0.516650	0.683728	1.344483	0.001572	0.064421	12.708966	0.652539
215	0.088812	0.484768	0.598012	1.257089	0.001394	0.060626	12.741882	0.658043
220	0.229353	0.765244	0.779689	1.427505	0.001396	0.060692	12.721938	0.656308

Table 2.Optimization results of ARIMA and LSTM (SH603833)

N	FIWSK1-MSE	FIWSK1-BE	ARIMA-MSE	ARIMA-BE	FIWSK2-MSE	FIWSK2-BE	LSTM-MSE	LSTM-BE
175	59.900959	0.680599	162.970908	1.085909	18.918390	0.359364	12755.4739	0.986383

180	142.394788	1.021514	288.238832	1.347816	30.398488	0.449116	12755.5706	0.986575
185	52.514974	0.632765	182.292392	1.169775	24.234669	0.399453	12756.1839	0.987151
190	173.303867	1.068827	309.197049	1.538054	17.423204	0.339443	12754.9901	0.986940
195	82.060965	0.780000	237.355249	1.355169	26.082289	0.424354	12754.7241	0.986090
200	204.222887	1.168251	350.737528	1.641493	9.818009	0.263716	12754.0441	0.986678
205	75.056005	0.742124	258.948132	1.415533	6.287704	0.215279	12753.0480	0.986402
210	54.324701	0.632572	203.870657	1.254471	6.271554	0.218490	12749.4413	0.984877
215	61.678767	0.688058	191.468238	1.201574	3.988250	0.173869	12751.3168	0.984548
220	54.915265	0.637636	241.274527	1.363730	2.391612	0.134844	12755.4739	0.986383

Table 3. Optimization results of ARIMA and LSTM (SZ300741)

N	FIWSK1-MSE	FIWSK1-BE	ARIMA-MSE	ARIMA-BE	FIWSK2-MSE	FIWSK2-BE	LSTM-MSE	LSTM-BE
175	171.548738	0.940500	600.025639	1.888782	1.075176	0.080730	2000.58756	0.978481
180	55.225301	0.567721	189.161677	1.077582	0.964192	0.077167	2000.48714	0.978389
185	59.879452	0.590510	229.621410	1.205389	0.857682	0.072810	2000.51593	0.978437
190	47.181062	0.546645	304.807856	1.393661	0.770367	0.069067	2000.56948	0.978541
195	70.994072	0.651454	257.708370	1.274919	0.654104	0.064487	2000.65334	0.978618
200	93.955585	0.746309	353.435112	1.489822	0.564087	0.059913	2000.60448	0.978645
205	47.756801	0.548941	313.664086	1.406863	0.603528	0.061951	2000.50931	0.978629
210	54.200044	0.583753	291.969510	1.364310	0.731026	0.068108	1999.65709	0.978077
215	59.246744	0.614242	302.372946	1.387235	0.733830	0.068217	2000.00189	0.978092
220	52.388869	0.577557	292.126979	1.364299	0.698986	0.066629	2000.85005	0.978527

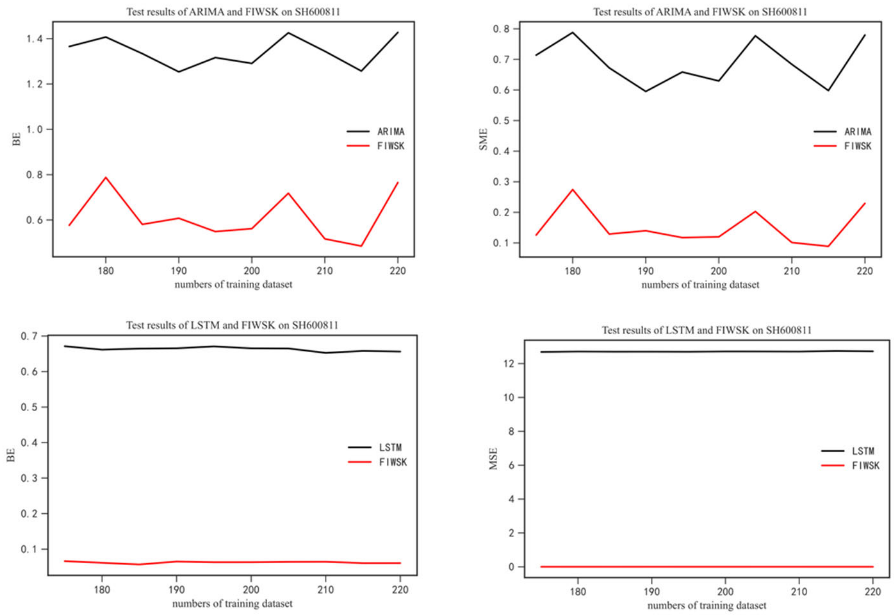


Figure 3 Comparison of optimization results of ARIMA and LSTM (SH600811)

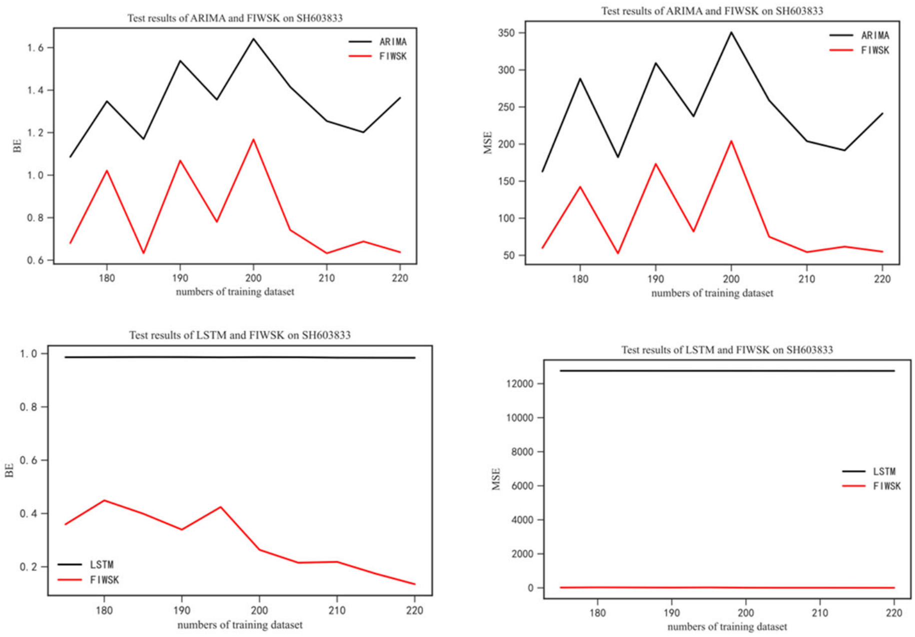


Figure 4. Comparison of the optimization results of ARIMA and LSTM(SH603833)

According to the above experimental results, this paper can find that for the prediction of financial time series of stock prices of the above three stocks, the optimization model based on Karhunen-Loève expansion and information gain weighted integrated regression for stock price time series prediction compares with the traditional ARIMA model and LSTM neural network model in terms of posterior error and mean square error. -Loove expansion and information gain weighted integrated regression stock price time series forecasting optimization model are both much lower than the ARIMA model and LSTM neural network model, which means that the optimized model is much more accurate and robust than the traditional financial time series forecasting model in terms of prediction accuracy and robustness.This is mainly because our optimization model first uses high-frequency intraday price time series to remove the residual series information obtained from the original time series forecasting model, and then this paper propose to use Karhunen-Loève expansion to reduce the dimensionality of the intraday price time series and successfully extract the chilling feature information from it, and use the stacking data enhancement framework to further improve the The Karhunen-Loève expansion further improves the fitting ability of the ground coefficients to the residual series by using the stacking data enhancement framework, minimizes the effects of extraneous noise and invalid fluctuations, and finally innovatively develops a weighted integrated regression method based on information gain to measure the variance and bias of the residual series prediction model to achieve the improvement of the final stock price prediction accuracy.

4. Conclusions and outlook

In this paper, a time series prediction model based on Karhunen-Loève expansion and information gain weighted ensemble regression is proposed and applied to the prediction of financial time series data. The proposed method uses the intra-day price as an auxiliary variable, uses Karhunen-Loève expansion to obtain its function feature information, and uses the data enhancement method to improve the effective information of Karhunen-Loève expansion coefficient variables on the residual series, reducing the influence of irrelevant noise. In addition, since the potential model structure between characteristic variables and residual series is unknown, this paper also developed a weighted ensemble regression method based on information gain to balance the variance and deviation of the prediction model, thereby improving the prediction accuracy. The empirical results show that the proposed method can greatly improve the fitting accuracy of ARIMA and LSTM neural networks for stock prices. For future research work, different base expansion methods and weighted methods can be replaced to explore the best time series prediction model under existing technical conditions. At the same time, this paper can also use clustering method to cluster the base model. Due to the low correlation between categories and categories, the base model within the category has a high degree of correlation. At this time, this paper can retain the model with the highest prediction accuracy within each category, and eliminate its model. Thus, a simplified ensemble model with weak correlation and high prediction accuracy of single model is obtained, which can further balance the variance and deviation of the prediction model.

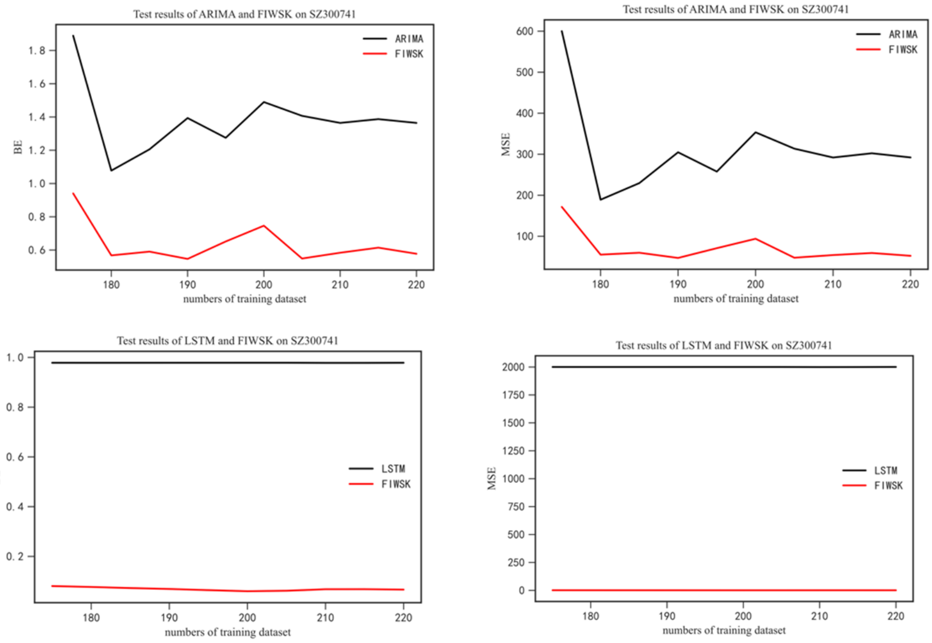


Figure 5 Comparison of optimization results of ARIMA and LSTM (SZ300741)

References

[1] Yixiao Zhang. Analysis and Forecast of Stock Price Based on Time Series[D]. North China Electric Power University, 2020.

[2] Xiaojun Li, Pan Tang. Stock price forecast based on technical analysis, fundamental analysis and deep learning[J]. Statistics & Decision, 2022, 38(02):146.

[3] Fang Wang, Xuanyi Wang, Shuo Chen. Machine learning in Economics: review and Prospect[J]. The Journal of Quantitative & Technical Economics, 2020, 37(4).

[4] Shuya Xu, Xiaoying Liang. Research on stock price prediction based on ARIMA-GARCH model[J]. Journal of Henan Institute of Education(Natural Science Edition), 2019, 28(04).

[5] Zhibin Xiong. Research on RMB exchange rate prediction model based on ARIMA fusion neural network[J]. The Journal of Quantitative & Technical Economics, 2011, 28(6).

[6] Guisheng Zhang, Xindong Zhang. Research on SVM-GARCH stock price forecasting model based on neighborhood mutual information[J]. Chinese Journal of Management Science, 2011, 24(9).

[7] Qing Yang, Chenwei Wang. Research on global stock index prediction based on deep learning LSTM neural network[J]. Statistical Research, 2019, 36(3).

[8] Hongbing Ouyang, Kang Huang, Hongju Yan. Financial time series prediction based on LSTM neural network[J]. Chinese Journal of Management Science, 2020, 28(4).

[9] Ruoqi Zhou, Junlin Li, Anqiang Dong. Research on the intensity of stock capital flow based on functional data analysis[J]. Journal of Taiyuan University of Science and Technology, 2021, 42(03).

[10] Pengfei Tang. Attribute reduction algorithm of set valued decision table based on approximate conditional entropy[J]. Intelligent Computer and Applications, 2021, 11(10).

[11] Kuying Jin, Ying Guo. Risk prediction method of unbalanced bank credit data based on stacking algorithm[C]. The 19th Annual Conference of Shenyang Science and technology, 2022.