A Two-Stage ARIMA Model via Machine Learning and its Application in Stock Price Prediction

Fenglin Zhang[1,*], Long Chen[2], Jiwen Yu[3]

[1]School of Maritime Economics and Management, Dalian Maritime University, Dalian, Liaoning/China, 116026

[2]School of Business, Jiangnan University, Wuxi, Jiangsu/China, 214122

[3]School of Mathematics, Northwest University, Xi'an, Shaanxi/China, 710127

*Corresponding author: zfl_dmu@163.com

Abstract. Stock price prediction has always been a hot issue in the financial sector and quantitative investment. Since stock price time series data tends to have linear and nonlinear features, traditional ARIMA models exhibit certain limitations in modeling such data. Based on this, this paper innovatively uses intraday transaction data of the stock market as auxiliary information, and proposes an improved ARIMA stock price prediction model based on machine learning methods. The specific principle is to use the ARIMA model to predict the linear information of the data, and machine learning-related algorithms (RF, XGBoost, LSTM) are used to predict the nonlinear residual information. The empirical results show that compared with the traditional ARIMA model, the model can effectively improve the prediction accuracy and is robust in stock price prediction. Finally, because this framework is very flexible in content, it can be equipped with machine learning methods with the best prediction accuracy for different practical application scenarios. In addition, we can use the model averaging method in the two-stage framework to improve the accuracy, and the mixed or high-frequency data can be further mined.

Keywords: ARIMA; Machine Learning; Combined prediction; Stock Price Prediction.

1. Introduction

As a barometer of the economy, the stock market functions value discovery and resource allocation optimization. In addition, predicting stock prices is of great significance for national macro-control and risk avoidance for investors. However, accurately predicting stock prices remains challenging due to the noisy, nonlinear and non-stationary nature of stock price data.

Based on the research results of scholars in related fields, it can be found that the main methods used for stock price prediction include time series analysis methods and machine learning methods.

The methods of time series analysis mainly include the differential autoregressive moving average model (ARIMA model), and the autoregressive condition heteroscedasticity model (GARCH model). Yuxia Wu and Xin Wen[1] established the ARIMA model to predict the movement law and trend of the closing price of Huatai securities in the 250th issue. Yue Chang and Yuxu Feng[2] established an ARIMA-GARCH fitted model for stock price prediction for the CSI 300-day yield. Franses P H et al. [3] improved the model's predictive performance by correcting the residual outliers of the GARCH model. However, the above models assume that stock price data obey a specific linear relationship, but stock price data tends to be non-stationary and has a potential nonlinear relationship.

Major machine learning methods include artificial neural networks (ANNs), support vector machines (SVMs), and long-term, short-term memory neural networks (LSTM). Lei L et al. [4] used wavelet neural networks (WNNs) to predict the stock price data based on the feature dimensionality reduction of rough sets. Xinbin Yang and Xiaojuan Huang[5] used SVM nonlinear extension samples to rank time series models for stock price prediction. Kai Chen et al. [6] used LSTM models to model and forecast Chinese equity returns. However, machine learning methods rely heavily on selecting feature engineering for stock price prediction, and well-constructed feature engineering depends on personal experience. Different feature engineering has a greater impact on the model's prediction accuracy, which predicts stock prices limited by machine learning methods.

It is worth noting that the research mentioned above on stock price prediction methods is based on the data itself to fit and forecast. However, due to the non-linearity and non-stationarity of the stock price data, the residuals often contain more valid information than the original data itself. The extraction of valid information from residuals is also essential to improve stock price predictions' accuracy. In order to more effectively extract the useful information in the prediction error and eliminate the influence of systematic error on the prediction accuracy, this paper proposes a method to divide the stock price time series data into linear main part and nonlinear residual part, and predict them separately. First, we use the ARIMA model and rolling window method to predict the stock price data, so as to obtain the predicted value of the linear main part, and extract the residuals to form a new series. Subsequently, the extracted nonlinear residuals are modeled and predicted by common machine learning algorithms such as random forest, XGBoost, LSTM, etc. Finally, the prediction sequence of the nonlinear residual part is added to the prediction sequence of the linear main part of the ARIMA stock price to form the final stock price forecast value.

The main contributions of this paper are: 1) This paper proposes a method for dividing the data series into linear subject parts and nonlinear residual parts, and modeling and predicting the above two parts, respectively. 2) Combine traditional time series prediction methods with machine learning algorithms to extract as much valid information as possible in the residual sequence, which can improve the model's predictive performance. 3) The model proposed in this paper is developed in a common framework. Its applicability is very flexible, which can be equipped with machine learning algorithms with the best prediction accuracy for the data series of different application scenarios.

2. Theoretical Model

2.1 ARIMA Model

The full name of the ARIMA model is the differential autoregressive moving average model, which was jointly proposed by Box and Jenkins[7]. It is a linear model for analyzing and studying time series problems. The model can effectively measure the linearity of time series data, and has good performance in short-term prediction. In the ARIMA model, three parameters need to be set, namely the autoregressive order p , the difference order d and the moving average order q . The mathematical expressions of $ARIMA(p,d,q)$ are as follows.

$$\begin{cases} y_t' = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}' + \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} \\ y_t' = \Delta^d y_t = (1-L)^d y_t \\ (1-\sum_{i=1}^{p} \alpha_i L^i)(1-L) y_t = \alpha_0 + (1+\sum_{i=1}^{q} \beta_i L^i)\varepsilon_t \end{cases} \tag{1}$$

where, y_t' is the sequence data of phase t after differential transformation; $\alpha_i (i = 1, 2, \cdots, p)$ and $\beta_i (i = 1, 2, \cdots q)$ denote autoregressive parameters and moving average parameters respectively; L is the lag operator; $\Delta^d = (1-L)^d$ is the d-order difference; ε_t is a random error term that follows a normal distribution $N(0, \sigma^2)$.

2.2 Machine Learning related algorithms

The random forest (RF) model was proposed by Breiman[8]. As one of the bagging methods of ensemble learning, RF is extracted from the original training sample set by the bootstrap resampling technique to generate a training sample subset, and then multiple training samples are generated from the training sample set. The decision tree is used to form a random forest, and its classification or regression results are determined by the voting score of the decision tree.

XGBoost is a gradient boosting method whose objective function consists of training loss function and regularization. Through the second-order Taylor approximation of the objective function, the greedy algorithm is used to search for the segmentation point with the highest score, and the next step is to segment and expand the leaf nodes. This has the advantage of ensuring that the tree structure will not be too complicated and over-fitted in the process of minimizing the loss function, on the other hand, improving the computational efficiency[9-10].

Recurrent Neural Network (RNN) is mainly used to process time-series data, which is characterized by the fact that the output of neurons at a certain moment can be used as input to enter the neurons, so that the neural network has memory ability. Hochreiter et al.[11] proposed Long Short-Term Memory (LSTM) neural network to overcome the problem that RNN can not solve the dependence on long-term time series.

2.3 Two-stage prediction model

Suppose that the stock price data series y_t consists of two parts, the linear body part L_t , and the nonlinear residual part L_t , namely $y_t = L_t + N_t$. Firstly, Arima is used to predict the data sequence to obtain the predicted value L_t of the linear main part. Then the predicted value L_t of the linear main part is subtracted from the real value y_t of the data sequence to obtain the nonlinear residual part N_t, that is $N_t = y_t - L_t$. Then, the nonlinear residual part can be modeled and fitted by machine learning algorithms such as random forest and XGBoost, the result is N_t . Finally, the final prediction result can be obtained by adding the prediction value L_t of the linear main part and the prediction value N_t of the nonlinear residual part, that is $y_t = L_t + N_t$ .This process can be represented in Figure 1.

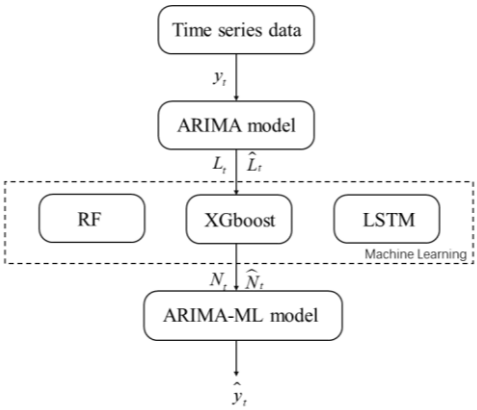

Fig. 1 Two-stage prediction model

3. Case Analysis

3.1 Experimental Data Sets and Preprocessing

1)Data Sets

Through the Baostock stock data interface, we obtained the data sets of three real stocks from January 1st, 2017, to April 18th, 2022, namely Sinopec (SH.600028), China Merchants Bank (SH.600036), China PingAn (SH.601318).

The dataset for each stock is divided into two parts, daily data and intraday trading data. The daily data contains 1,285 records, and the time interval for obtaining intra-day transaction data is 30 minutes, so each daily data corresponds to 8 intra-day data, and there are 10,280 records of intra-day

transaction data. Both parts include six factors: the opening price, closing price, highest price, lowest price, volume, and amount[12].

The candlestick charts for the three stocks over the period studied are shown in Figure 2.

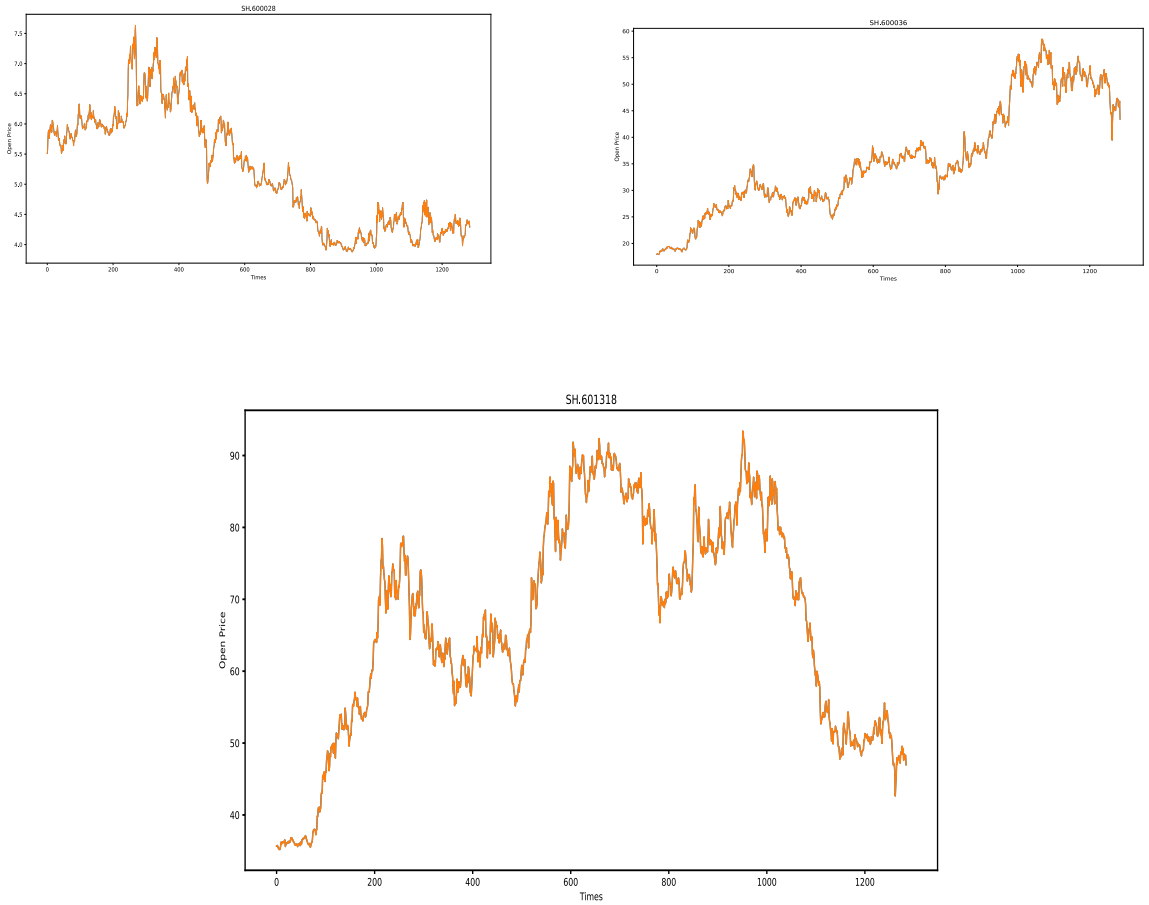

Fig. 2 SH.600028 SH.600036 and SH.601318's Candlestick chart in the selected time period

The fundamentals of the opening prices of the three sample stocks are analyzed, and the results are shown in Table 1.

Table 1. Fundamental Analysis of Stock Opening Prices

	SH.600028	SH.600036	SH.601318
count	1285	1285	1285
mean	5.22	35.82	66.78
std	0.95	10.33	15.18
min	3.86	17.60	35.23
25%	4.32	28.40	53.99
50%	5.12	34.45	68.22
75%	6.00	44.25	79.37
max	7.46	58.21	93.38

2)Standardization

In order to improve the accuracy of prediction, we normalize the residuals input to the two-stage prediction model, that is, let the variance of this dataset become 1, and the mean becomes 0. We adopt a z-score normalization strategy.

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

3)Rolling Forecast

prediction of stock data is essentially time series prediction, which may not perform well over the long term. Therefore, the method of rolling prediction is adopted, that is, the data in a time window of a certain length is used to predict the data of the next time point. When using the ARIMA model to predict daily data, 100, 150, and 250 periods of data are usually taken as rolling time windows. When using machine learning algorithms to fit residuals, the data set is generally divided into a training set and test set according to the ratio of 7:3, and uses the length of the training set as a rolling time window.

3.2 Model Accuracy Check

Let set $N_t = \{N_1, N_2, ..., N_n\}$ be the prediction residuals of the one-stage ARIMA model, and set $\hat{N}_t = \{\hat{N}_1, \hat{N}_2, ..., \hat{N}_n\}$ be the residuals after fitting by the machine learning algorithm. Denote the prediction residual as $\varepsilon_i = N_i - \hat{N}_i$.

The indicators for calculating the prediction error of the stock price prediction residual value are as follows.

1)Mean Squared Error

Mean squared error measures the difference between the estimator and the estimator as the sum of squared errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^{\,2} \quad i = 1, 2, ..., n \tag{3}$$

2)Mean Relative Error

The mean relative error measures the difference in the residual value of the predicted value relative to the original data.

$$MRE = \frac{1}{n}\sum_{i=1}^{n} \frac{|\varepsilon_i|}{N_i} \times 100\% \quad i = 1, 2, ..., n \tag{4}$$

3)Posterior Difference Test

The posterior difference test is based on the residuals of each period, and examines the probability of the occurrence of points with small residuals. The posterior error C and the small error probability P are calculated as follows.

$$C = \frac{S_2}{S_1} \tag{5}$$

$$P = P\{|\varepsilon_i - \bar{\varepsilon}_i| < 0.674 S_1\} \quad i = 1, 2, ..., n \tag{6}$$

Where, S_2 is the standard deviation of the residual sequence, and S_1 is the standard deviation of the original sequence[13]. In the subsequent analysis, we only use the posterior error C as the result of the posterior difference test.

3.3 Experimental Results

In the first stage, we take the 100-period data as the rolling prediction time window of ARIMA. In stage two, three machine learning methods, RF, XGBoost, and LSTM, are used to fit the residuals of the first stage, and the rolling method is used to forecast the residual of the opening price of the next trading day. The residual forecast value is added to the first-stage forecast result. Compared with the original forecast result of the first stage, three indicators of MSE, MRE and C are obtained to measure the error. This process is programmed in Python, and the forecast results of the three stocks are shown in Tables 2-4.

Table 2. The accuracy test of SH.600028 stock 100-period forecast after revision

	sh.600028		
	MSE	MRE	C
ARIMA	1.926	1.140	0.391
ARIMA+LSTM	1.855	1.128	0.384
ARIMA+RF	1.575	0.803	0.066
ARIMA+XGBoost	1.377	0.870	0.064

Table 3. The accuracy test of SH.600036 stock 100-period forecast after revision

0	sh.600036		
	MSE	MRE	C
ARIMA	354.277	1.506	0.260
ARIMA+LSTM	341.308	1.476	0.255
ARIMA+RF	77.157	0.982	0.047
ARIMA+XGBoost	113.354	1.159	0.059

Table 4. The accuracy test of SH.601318 stock 100-period forecast after revision

	sh.601318		
	MSE	MRE	C
ARIMA	467.438	1.297	0.083
ARIMA+LSTM	457.198	1.287	0.076
ARIMA+RF	398.941	1.044	0.076
ARIMA+XGBoost	342.285	0.990	0.075

The residuals of the final predicted values of three kinds of stocks are shown in Figure 3. Analysis of the results shows that no matter which model is used to fit the residual error in the second stage, its error is smaller than that of ARIMA prediction of the first stage. This is also determined by the nature of the model itself. From the fitting results, RF and XGBoost contribute significantly to reducing the total error of the model. It can be seen that using intra-day stock data as supplementary information to forecast the opening price of the next trading day can improve the accuracy of the forecast and make up for the limitations of the traditional ARIMA model.

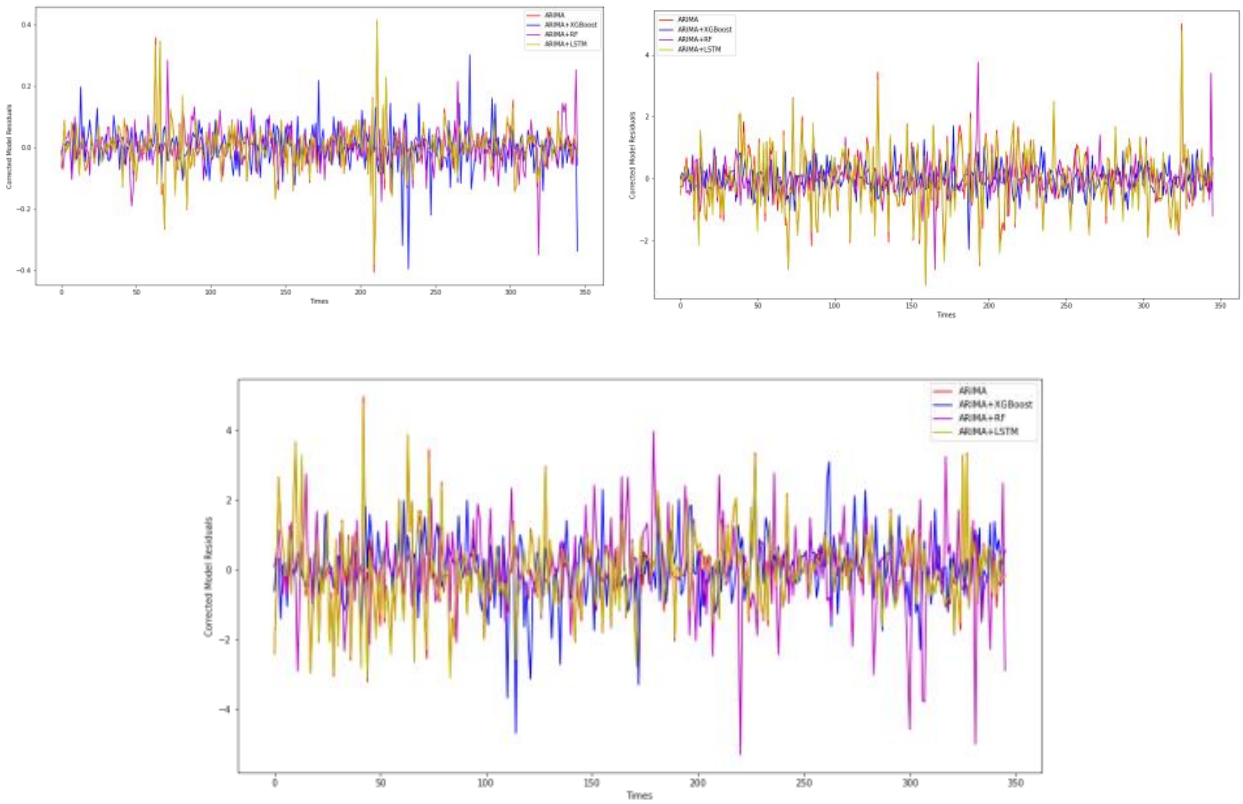

Fig. 3 SH.600028, SH.600036 and SH.601318 ‘s stock 100-period forecast revised residual comparison chart

3.4 Robustness Check

Change the number of rolling forecast periods and set the rolling time windows to 150 periods and 250 periods, respectively. For these three stocks, the revised forecast data is shown in Table 5-7.

Re-fitting under the condition of changing the number of periods, the fitting effect of the three machine learning models is still the same as that of the 100-period forecast, and the errors all have a clear downward trend. Therefore, it is considered that the model has good robustness and high stability, which can not only effectively realize the forecast of stock opening prices under rolling time windows of various lengths, but also contribute to the further promotion and generalization of the model.

Table 5. SH.600028 stock forecast accuracy

	150-period			250-period		
	MSE	MRE	C	MSE	MRE	C
ARIMA	1.730	1.125	0.386	1.633	1.153	0.395
ARIMA+LSTM	1.653	1.103	0.378	1.536	1.139	0.383
ARIMA+RF	1.198	0.710	0.060	1.089	0.756	0.059
ARIMA+XGBoost	1.411	0.845	0.068	1.360	0.843	0.064

Table 6. SH.600028 stock forecast accuracy

	150-period			250-period		
	MSE	MRE	C	MSE	MRE	C
ARIMA	331.910	1.511	0.276	311.966	1.518	0.322
ARIMA+LSTM	315.703	1.485	0.269	293.073	1.471	0.312
ARIMA+RF	81.849	0.911	0.053	98.387	1.028	0.061
ARIMA+XGBoost	98.294	1.117	0.059	116.485	1.219	0.066

Table 7. SH.600028 stock forecast accuracy

z	MSE	MRE	C	MSE	MRE	C
ARIMA	418.928	1.293	0.082	356.613	1.272	0.080
ARIMA+LSTM	398.558	1.283	0.075	304.600	1.252	0.078
ARIMA+RF	374.887	1.054	0.074	303.301	0.998	0.076
ARIMA+XGBoost	287.754	0.971	0.071	238.971	0.940	0.067

4. Conclusion

This paper proposes an improved ARIMA stock price prediction model based on machine learning methods. ARIMA models and machine learning-related algorithms (RF, XGboost, LSTM) can extract the linear body and nonlinear residuals of a stock price data series, respectively. Experimental results show that the improved ARIMA stock price prediction model based on machine learning methods constructed in this paper has higher prediction accuracy than the traditional ARIMA model. Robustness testing by using multiple stock data and changing the rolling time window of the same stock data shows that the model proposed in this paper is robust. Since many time series sequences in real life exhibit instability and nonlinearity, the model framework presented in this paper has a certain application value for studying this type of data.

On the basis of this paper, we can further study the selection of a prediction model for nonlinear residual. Different models fit different data series with different information, so they often obtain different forecasts. So which model is the best? Further discussion can be carried out based on model selection and model averaging theory. Scholars have proposed various methods and criteria for model selection, such as AIC, BIC, cross-validation, Lasso, etc. However, model selection methods often lead to uncertainty in the model selection process, which in turn underestimates the actual variance. The model averaging method takes the form of $a_1 f_1(x) + a_2 f_2(x) + \cdots + a_n f_n(x)$, where $\sum_{i=1}^{n} a_i = 1$, which in most cases can circumvent the drawbacks of the model selection method. For example, the asymptotic optimal model averaging method can directly aim to reduce the estimated or predicted risk, making the goal more explicit [14]. In addition, with the development of computer technology and the advent of the era of big data, the price information of stocks can be recorded at a higher frequency. High-frequency intraday trading price data contains much information, so the research and analysis of high-frequency data also have certain application value. At present, high-frequency data can be studied using the data smoothing method or functional data analysis method.

References

[1] Wu, Y., & Wen, X. (2016). Short-term stock price forecast based on ARIMA model. Statistics and decision-making (23),83-86.

[2] Chang, Y., Feng. Y., & Cao, X. (2018). Stock Analysis and Forecast Based on Nonlinear Time Series Model. Mathematical Practice and Understanding (22), 21-26.

[3] Franses, P. H., & Ghijsels, H. (1999). Additive outliers, GARCH and prediction volatility. International Journal of prediction, 15(1), 1-9.

[4] Lei, L. (2018). Wavelet neural network prediction method of stock price trend based on rough set attribute reduction. Applied Soft Computing, 62, 923-932.

[5] Yang, X., & Huang, X. (2010). Research on stock price prediction based on support vector machine. Computer Simulation (09), 302-305.

[6] Chen, K., Zhou, Y., & Dai, F. (2015, October). A LSTM-based method for stock returns prediction: A case study of China stock market. In 2015 IEEE international conference on big data (big data) (pp. 2823-2824). IEEE.

[7] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: prediction and control. John Wiley & Sons.

[8] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[9] Andersson, J. O. (2011). The new foundations of evolution: on the tree of life.

[10] Hendrikx, J., Murphy, M., & Onslow, T. (2014). Classification trees as a tool for operational avalanche prediction on the Seward Highway, Alaska. Cold Regions Science and Technology, 97, 113-120.

[11] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[12] Li, C., Zhang, X., Qaosar, M., Ahmed, S., Alam, K. M. R., & Morimoto, Y. (2019, August). Multi-factor based stock price prediction using hybrid neural networks with attention mechanism. In 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech) (pp. 961-966). IEEE.

[13] Lu Sunyun. (2013). Research on multi-objective allocation of water resources in the Hanjiang River Basin based on section control (PhD dissertation, Wuhan University).

[14] Zhang, X. & Zou, G. (2011). Model averaging method and its application in prediction. Statistical Research (06), 97-102.