ISSN 2616-5775 Vol. 5, Issue 10: 14-20, DOI: 10.25236/AJCIS.2022.051003

An improved ARIMA Stock Price Forecasting Method Based on B-spline Expansion and Model Averaging

Minsong Gao[1], Chuyu Feng[2]

[1]Department of Mathematical Sciences, Anhui University, Hefei, 230601, China
[2]School of International Business Administration, South China Normal University, Guangzhou, Guangdong, 510000, China

Abstract: *Aiming at the limitation of ARIAM model in predicting stock prices with relatively complex fluctuation trends, this paper proposes an improved ARIMA method (BMA-ARIMA) based on B-spline basis expansion and model averaging. The proposed method uses intraday prices as auxiliary information and considers its functional characteristics. By combining the two methods of parallel model averaging and recursive model averaging, the intraday price function features extracted based on the B-spline expansion method can be used to the greatest extent to fit the residuals predicted by the ARIMA model, thereby improving the prediction accuracy of ARIMA. In addition, since the regression prediction method in the model averaging stage is model-free, when the nonlinear regression model is selected, it can capture the linear and nonlinear information of the intraday price function characteristics. Specifically, this paper uses CART as the base model, and the actual data analysis results show that BMA-ARIMA can improve the prediction accuracy of the ARIMA model and has certain robustness. Finally, the method can theoretically be extended to time series forecasting in the fields of medicine, water conservancy, electric power and environmental science.*

Keywords: *ARIMA; B-spline basis expansion; Model average; Auxiliary information*

1. Introduction

In the financial field, most of the data appear in the form of time series, how to establish an accurate prediction model based on the time series features has become a hot topic. Due to its simple form, the ARIMA model only needs endogenous variables and does not need to rely on other exogenous variables, and is widely used in financial time series forecasting: Liu Song (2021) [1] et al. established ARIMA model to predict the stock price of Southwest Securities; Wu Yuxia (2016) [2] et al. established an ARIMA model to predict the trend of stock price changes in the ChiNext market; Dong Bolun (2015) [3] et al. established an ARIMA model to predict agricultural stock prices. They used the ARIMA model to predict stock prices in different industries, and obtained better short-term forecasting effects.

However, when using the ARIMA model for prediction, the time series is required to be stationary or stationary after difference, and only the information of the series to be predicted is used in the prediction process, and other auxiliary information, such as intraday price data, cannot be used. In addition, ARIMA model also has some problems in actual prediction, such as high data requirements and weak anti-interference ability.

Based on these, many scholars have proposed many methods to improve the ARIMA prediction: or example, Bai Yujie (2011) [4] et al. used wavelet transform to decompose and reorganize the time series to reduce noise to obtain a stationary time series, and then used the ARIMA model to predict , which provided a new forecasting approach for rainfall forecasting; Pan Difu (2008) [5] et al. introduced the Kalman prediction method to improve the ARIMA model in the wind speed prediction of wind farms, and proposed the Kalman time series method, which improved the accuracy of multi-step prediction; Fu Xinzhong (2009) [6] et al. combined the characteristics of BP neural network and ARIMA model, and proposed an ARIMA-ANN combined prediction model. This model adds the prediction result of ARIMA and the error prediction value of BP neural network as the prediction value, which has the advantages of high prediction result accuracy and strong fault tolerance; These methods improve the ARIMA prediction model from different aspects, and all of them can improve the prediction accuracy to a certain extent.

ISSN 2616-5775 Vol. 5, Issue 10: 14-20, DOI: 10.25236/AJCIS.2022.051003

In summary, we consider the intraday price as auxiliary information, and propose an improved ARIMA method based on B-spline basis expansion and model averaging. The innovation of this method is that it first uses intraday price as auxiliary information and extracts its functional features, and then uses the method of nested recursive model averaging (boosting) in parallel model averaging (bagging) to weigh the deviation and variance of the base model. By improving the fitting accuracy between the function characteristics and the ARIMA model prediction residual, it indirectly reduces the final time series prediction error. Specifically, the method first uses the ARIAM model to predict the predicted value of the stock opening price and calculates the residual, then introduces the intraday price as auxiliary information and uses the B-spline basis to expand to extract the functional features of the auxiliary information, and then using the method of nested recursive model averaging in parallel model averaging, with CART as the base model, the function features of auxiliary information and residuals are modeled and the residuals are predicted. Finally, the sum of the prediction residuals and the predicted value obtained by the ARIMA model is used as the predicted value of the final time series. The experimental results show that the proposed method is more competitive than the ARIMA model.

2. Theory and method

2.1 ARIMA model

ARIMA model, also known as autoregressive moving average summation model, is a famous time series forecasting method proposed by Box and Jenkins in the early 1970s, also known as Box-jenkins model. In ARIMA(p,d,q), p is the number of autoregressive items, q is the number of moving average items, and d is the number of differences made when the time series becomes stationary. The general operator is expressed as follows:

$$(1+\varphi_1 B+\varphi_2 B^2+\cdots+\varphi_p B^p)(1-B)^d Y_t=(1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q)e_t \quad (1)$$

When using the ARIMA model for modeling, the difference order should be selected first through the unit root test, and then the optimal model parameters should be selected according to the AIC criterion. The BMA-ARIMA method uses the function auto_arima in the python third-party library pmdarima to automatically establish the optimal ARIMA model.

2.2 B-spline basis expansion

The B-spline basis expansion is to approximate the given data through the B-spline curve:

$$P(u)=\sum_{i=0}^{n} P_i N_{i,k}(u) \quad (2)$$

The basis function coefficient $P_1, P_1, \cdots, P_n$ is obtained by the least square method, and it is used as the function feature of this group of data. Among them, $N_{i,k}(u)$ is called the k-order B-spline basis function, which is defined by the De Boor-Cox recurrence formula. The recurrence formula is as follows:

$$N_{i,0}(u)=\begin{cases}1 & u_i<x<u_{i+1}\\ 0 & else\end{cases} \quad (3)$$

$$N_{i,p}(u)=\frac{u-u_i}{u_{i+p}-u_i}N_{i,p-1}(u)+\frac{u_{i+p+1}-u}{u_{i+p+1}-u_{i+1}}N_{i+1,p-1}(u) \quad (4)$$

When calculating the basis function, it is necessary to specify the node vector $U=(u_0,u_1,u_2,\cdots,u_n)$ and the degree p in advance.

2.3 Model averaging method

Model averaging is essentially a kind of ensemble learning method, which is one of the commonly used methods in statistics and has a very wide range of application value. The principle can be

ISSN 2616-5775 Vol. 5, Issue 10: 14-20, DOI: 10.25236/AJCIS.2022.051003

understood as: Assuming that there are n models, let the predicted value of the i-th model at time t be $\hat{y}_{it}(i=1,2,\cdots,n)$, then the final estimated value obtained by the model averaging method is $\hat{y}_t=\sum_{i=1}^{n}\eta_i\hat{y}_{it}$, $\sum_{i=1}^{n}\eta_i=1$, where η_i is the weight of the i-th model, and the weights of each models are generally equal . When dealing with uncertain models, the model-averaging method can usually achieve better prediction results than using a single model. The two most common model averaging methods are parallel model averaging (Bagging) and recursive model averaging (boosting). For details, please refer to Xu Jiwei (2018) [7] et al. Ensemble Learning Methods: Research Review.

2.4 BMA-ARIMA

This paper improves the prediction accuracy of the ARIMA model by using the functional characteristics of intraday prices to fit the residuals obtained by ARIMA prediction. Since the B-spline basis function is used in extracting function features and the model averaging method is used in fitting residuals, we call this optimization method the improved ARIMA method based on B-spline basis expansion and model averaging (BMA- ARIMA). This method makes up for the shortcomings of the ARIMA model that only uses the information of the time series to be predicted, and cannot deal with stochastic systems with complex fluctuation trends. The specific implementation of this method is as follows:

1). First, establish an ARIMA model according to the opening price Y of the opening stock, make predictions according to the ARIMA model, obtain the predicted value $\hat{Y}$ of the opening price, and calculate the prediction error $e=Y-\hat{Y}$;

2). Expand the intraday price Z with the B-spline basis, and obtain the basis function coefficient X as the function characteristic of the intraday price;

3). Parallel model averaging: Bootstrap the sample to obtain B sample data subsets, each with the same weight.

4). Parallel model averaging: Boosting method is used for each sample subset, where the CART model is used as the underlying model, and the weight of each underlying model is a decreasing function of the prediction residuals of the model.

5). Through steps 3, 4, and 5, the fitted value $\hat{e}$ of the ARIMA prediction residuals of the training sample and the new sample can be obtained. The final predicted value of BMA-ARIMA method is $\tilde{Y}=\hat{Y}+\hat{e}$.

2.5 Model accuracy evaluation indicators

How to objectively evaluate the accuracy of a model needs to introduce three indicators: mean relative error (MRE), mean square error (MSE), posterior error (BE), and their corresponding calculation formulas are as follows:

$$MRE=\frac{1}{n}\sum_{k=1}^{n}\frac{|e_k|}{Y_k} \tag{5}$$

$$MRE=\frac{1}{n}\sum_{k=1}^{n}\left(Y_k-\hat{Y}_k\right)^2 \tag{6}$$

$$BE=\frac{S_2}{S_1} \tag{7}$$

Where e_k is the residual sequence, Y_k is the true value, S_1 is the standard deviation of the original sequence, S_2 is the standard deviation of the relative value sequence. The smaller the three indicators of the model, the higher the prediction accuracy.

ISSN 2616-5775 Vol. 5, Issue 10: 14-20, DOI: 10.25236/AJCIS.2022.051003

3. Data Analysis

3.1 Data pre-analysis

In order to verify the effectiveness of BMA-ARIMA on random systems with different degrees of complexity, this paper randomly selects three stocks for stock price prediction, which are Shandong Xinchao Energy Co.,LTD. (sh600777), Dongfang Group Co.,LTD. (sh600811) and Huabao Flavors Co.,LTD. stock (sz300741). The three companies are engaged in the fragrance industry, investment holding, and oil and gas industry. There is no direct connection between these three industries, and the inherent volatility mechanism of stocks is different, so they have different levels of complexity. The opening prices of three stocks in 241 days, and the average transaction price per minute of each day which is called the intraday price, are selected. The dimension of intraday prices is 240, and the data comes from the wind database. In order to discuss the stability of the opening prices of the three stocks, we draw the autocorrelation function diagram and partial autocorrelation function diagram of the opening prices of the three stocks, as shown in Figure 1 to Figure 3 below:

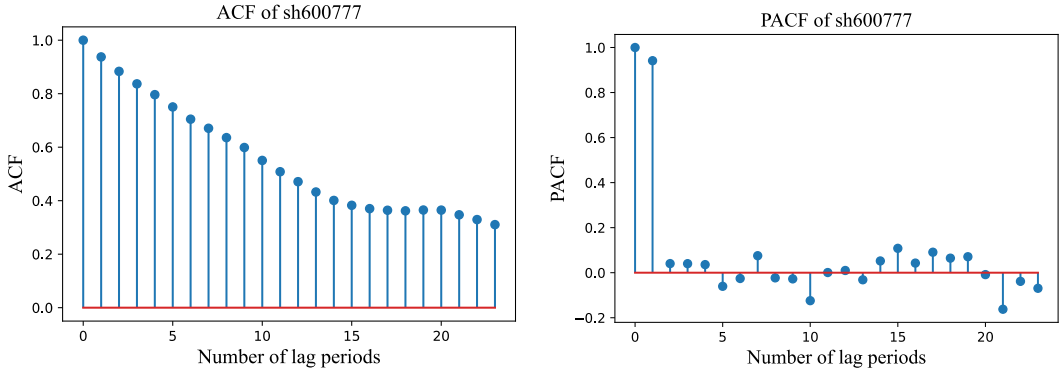

Figure 1: ACF and PCF of sh600777

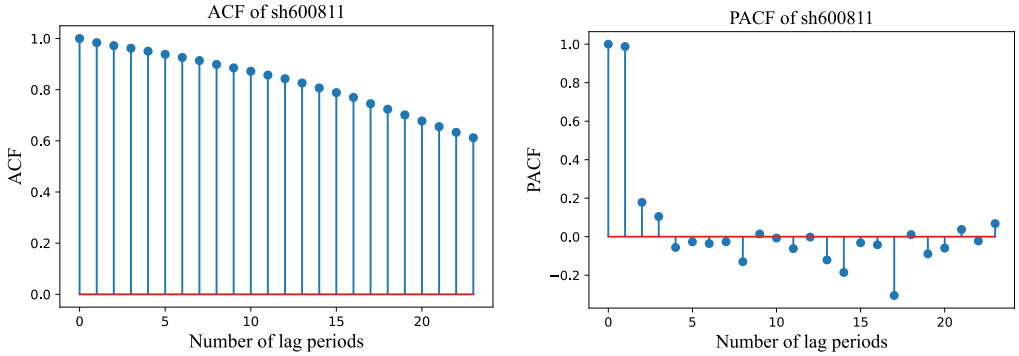

Figure 2: ACF and PCF of sh600811

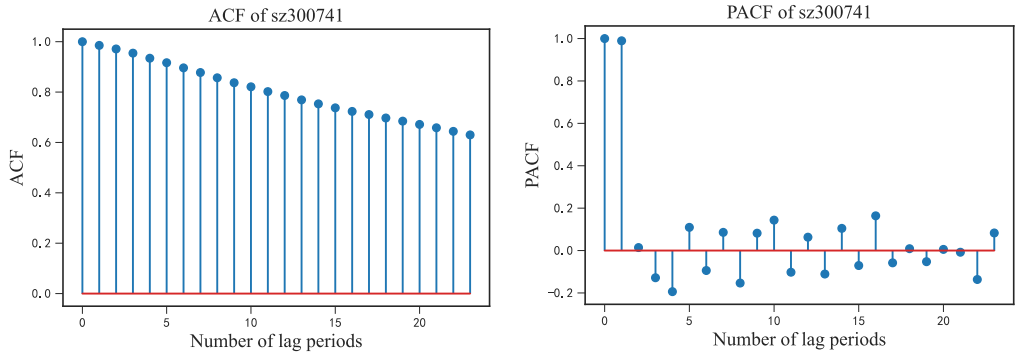

Figure 3: ACF and PCF of sz300741

From Figure 1 to Figure 3, the autocorrelation functions of the opening prices of the three stocks still take a large value when the lag period is large, and there is no exponential decreasing trend.

ISSN 2616-5775 Vol. 5, Issue 10: 14-20, DOI: 10.25236/AJCIS.2022.051003

Therefore, we can preliminarily believe that the opening prices of the three stocks are all non-stationary time series. In order to further explore the stationarity of the series, we conducted the ADF test on the opening prices of the three stocks. The ADF test P values corresponding to sh600777, sz300741, and sz300741 were 0.07, 0.65, and 0.79, all of which were greater than 0.05, so we have reason to believe that the three stocks opening prices are all non-stationary time series.

3.2 Empirical results

Next, take 80% of the samples as the training set and 20% as the test set, use the ARIMA method and the BMA-ARIMA method to predict, and calculate the predicted mean square error MSE, average relative error MRE, and posterior error BE , the hyperparameters of the BMA-ARIMA method are obtained using the cross-validation method. The specific results are shown in Table 1 below. From the results in Table 1, when the training set is 80%, the MSE, MRE, and BE predicted by the BMA-ARIMA method are smaller than those predicted by the ARIMA method. The BMA-ARIMA method obtained by improving the ARIMA method can be used for stocks of different complexity, and the prediction effect is better than that of the ARIMA method.

Table 1: Three prediction error indicators of three stocks

	sh600777			sh600811			sz300741		
	MSE	MRE	BE	MSE	MRE	BE	MSE	MRE	BE
ARIMA	0.0029	1.9623	0.3328	0.0136	2.0892	0.1875	4.5707	3.0258	0.1673
BMA-ARIMA	0.0008	1.0637	0.1802	0.0054	1.1594	0.1144	2.3858	1.9027	0.1196

3.3 Robustness check

We take the percentages of samples in the training set as 60%, 70%, 80%, and 90%, respectively use the ARIMA method and the BMA-ARIMA method for prediction, and calculate the predicted MSE, MRE, and BE, and draw the training set at different percentages. The MSE, MRE, and BE of the ARIMA prediction method and the BMA-ARIMA prediction method are listed below. The specific results are shown in Table 2 to Table 4.

Table 2: Three indicators of sh600777 in different percentage of training sets

Indicators	MSE				MRE				BE			
Percentages	60%	70%	80%	90%	60%	70%	80%	90%	60%	70%	80%	90%
ARIMA	0.0075	0.0189	0.0029	0.0032	3.6410	5.5747	1.9623	2.1063	0.4888	0.7382	0.3328	0.3496
BMA-ARIMA	0.0053	0.0155	0.0008	0.0012	2.7085	4.5500	1.0637	1.1620	0.4090	0.6580	0.1802	0.2115

Table 3: Three indicators of sh600811 in different percentage of training sets

Indicators	MSE				MRE				BE			
Percentages	60%	70%	80%	90%	60%	70%	80%	90%	60%	70%	80%	90%
ARIMA	0.0619	0.0202	0.0136	0.0135	4.3859	2.6274	2.0892	2.0360	0.3605	0.2205	0.1875	0.1875
BMA-ARIMA	0.0533	0.0113	0.0054	0.0049	3.5542	1.7063	1.1594	1.0821	0.3288	0.1600	0.1144	0.1122

Table 4: Three indicators of sz300741 in different percentage of training sets

Indicators	MSE				MRE				BE			
Percentages	60%	70%	80%	90%	60%	70%	80%	90%	60%	70%	80%	90%
ARIMA	6.3596	3.7741	4.5707	4.3569	3.4830	2.9392	3.0258	2.8392	0.1907	0.1551	0.1673	0.1646
BMA-ARIMA	5.9829	1.7805	2.3858	1.9491	2.7760	1.8289	1.9027	1.6162	0.1774	0.1065	0.1196	0.1076

In order to intuitively compare the prediction effect of the ARIMA method and the BMA-ARIMA method, we draw a line graph with the percentage of the training set as the abscissa and the MSE, MRE, and BE predicted by the two methods as the ordinate according to Table 2-Table 4, as follows Figure 4-Figure 6: From Figure 4-Figure 6, for stocks in three different industries randomly selected, The MSE, MRE, and BE of the BMA-ARIMA prediction method are smaller than those of the ARIMA prediction method under different percentages of training sets, that is, the improved BMA-ARIMA prediction method can not only improve the prediction accuracy, but also have robustness.

ISSN 2616-5775 Vol. 5, Issue 10: 14-20, DOI: 10.25236/AJCIS.2022.051003

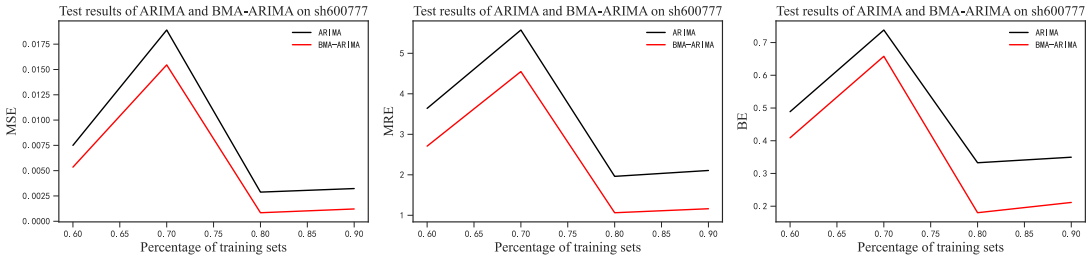

Figure 4: Three indicators of sh600777 in different percentage of training sets

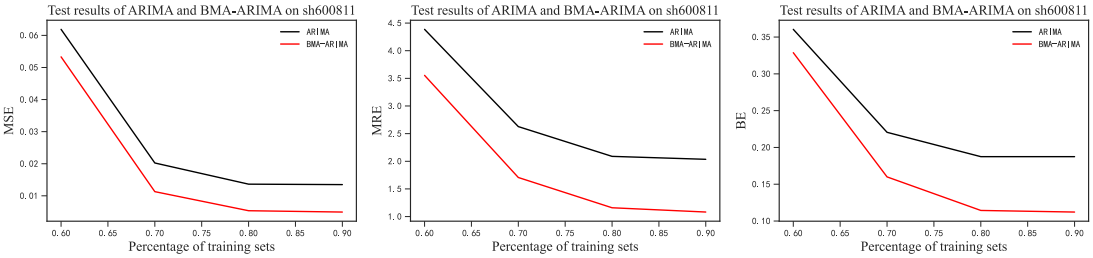

Figure 5: Three indicators of sh600811 in different percentage of training sets

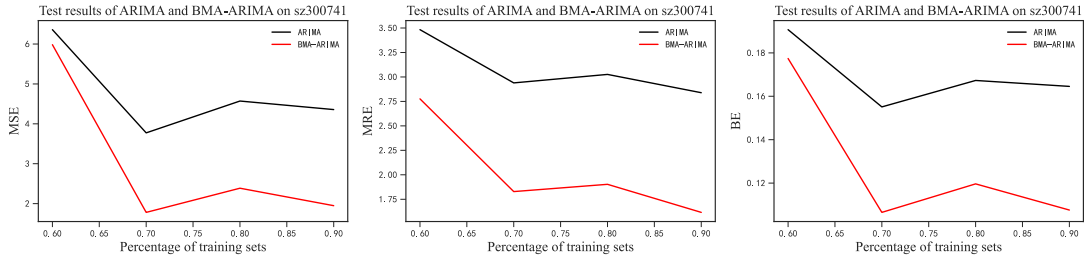

Figure 6: Three indicators of sz300741 in different percentage of training sets

4. Conclusion

The BMA-ARIMA method introduces the intraday price as auxiliary information, and uses the B-spline basis function to expand the function features of the auxiliary information, which makes up for the disadvantage that the ARIMA method cannot extract the nonlinear relationship of the time series. For the BMA-ARIMA model, in addition to introducing intraday prices as auxiliary information, we can also introduce other auxiliary information, such as introducing the intraday trading volume of stocks as auxiliary information. Besides, in addition to using B-spline basis expansion, we can also use wavelet basis expansion, Fourier basis expansion and other auxiliary information to extract function features. When using auxiliary information to model the prediction error, the statistical characteristics of auxiliary information can also be considered, so that auxiliary information can be fully used to improve the prediction accuracy. In actual forecasting, there may be missing values in auxiliary information or time series. In this case, we can use the nearest filling method or the KNN filling method in machine learning to fill in the missing values, and then use the BMA-ARIMA method for prediction. The BMA-ARIMA method obtained by improving ARIMA can improve the prediction accuracy, has generalization and robustness, and can not only be used for the prediction of stock opening prices, but also in the fields of meteorology, transportation, medicine, etc. For example, we can use BMA-ARIMA to predict the concentration of PM2.5 in the atmosphere, traffic flow at intersections, cardiovascular and cerebrovascular diseases, etc.

References

[1] Liu Song & Zhang Shuai. (2021). Empirical research on stock price forecasting using ARIMA model. Economic Research Guide (25), 76-78.

[2] Wu Yuxia & Wen Xin. (2016). Short-term stock price forecast based on ARIMA model. Statistics and Decision (23), 83-86. doi: 10.13546/j.cnki.tjyjc.2016.23.051.

ISSN 2616-5775 Vol. 5, Issue 10: 14-20, DOI: 10.25236/AJCIS.2022.051003

[3] Dong Bolun & Xu Dongyu. (2015). Prediction and Analysis of Agricultural Products Stock Prices Based on ARIMA Model. Modern Business (03), 186-188. doi: 10.14097/j.cnki.5392/2015.03.100.
[4] Bai Yujie. (2011). Application of improved time series model in rainfall forecasting. Computer Simulation (10), 141-145.
[5] Pan Difu, Liu Hui & Li Yanfei. (2008). An improved algorithm for short-term multi-step forecasting of wind speed in wind farms. Chinese Journal of Electrical Engineering (26), 87-91.
[6] Fu Xinzhong, Feng Lihua & Chen Wenchen. (2009). Application of ARIMA and ANN combined prediction model in medium and long-term runoff forecasting. Chinese Journal of Water Resources and Water Engineering (05), 105-109.
[7] Xu Jiwei & Yang Yun. (2018). Ensemble learning method: A review of research. Journal of Yunnan University (Natural Science Edition) (06), 1082-1092.