ISSN 2616-5775 Vol. 5, Issue 10: 41-47, DOI: 10.25236/AJCIS.2022.051007

An Improved ARIMA Method Based on Hybrid Dimension Reduction and BP Neural Network

Longhui Mo

Mathematics and Physics College, Chengdu University of Technology, Chengdu, 610059, China

Abstract: *In order to solve the problem that the ARIMA model cannot well fit the prediction of time series with high dimension and high noise, this paper proposes a method based on the combination of hybrid reduction and BP neural network. Taking the stock price as an example, the proposed method takes the intraday price as auxiliary information and uses PCA and KPCA to extract linear and nonlinear features of it respectively, and the dimensionally reduced features are then used as the input variable. BP neural network model was used to fit the residual error between the real value and the predicted value of the ARIMA model. Finally, add up the closing price prediction value obtained by the ARIMA model and the residual error prediction value obtained by BP neural network model for the final closing price prediction value. The empirical results show that compared with the ARIMA model, the proposed method has better prediction performance and fitting accuracy, and has certain robustness. This method can also be extended to other practical problems such as average temperature prediction and port ship flow prediction.*

Keywords: *ARIMA; stock price prediction; dimension reduction; BP neural network*

1. Introduction

Time series records the historical behavior of the system. A single time series value is undetectable, but the overall time series value has a certain rule. In the 1970s, American statistician Jenkins and British statistician Box proposed the Autoregressive Integrated Moving Average model (ARIMA), and its classical model is as follows:

$$\begin{cases} x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \quad \phi_p \neq 0, \theta_q \neq 0, \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0 \quad s \neq t, \\ E(x_s \varepsilon_t) = 0 \quad s < t, \end{cases} \tag{1}$$

Where $\phi_k (1 \leqslant k \leqslant p)$ and $\theta_k (1 \leqslant k \leqslant q)$ are real parameters, and ε_t is pure random series that meets the conditions. This model expresses the rule of time series in mathematical form and realizes the short-term prediction of time series value through the study of mathematical form. ARIMA model is a classical model in stationary time series analysis, but the time series seen in economic life is basically non-stationary, so we need to adopt certain methods to change it into stationary series and then establish an ARIMA model to predict and analyze the relevant time series.

For example, the stock investment market is a financial place where risks and benefits coexist. For decision-makers, if they can obtain more accurate prediction, they can more effectively avoid future risks. For regulators, getting accurate stock movements is an effective way to tighten control over the stock market. Zhang Yingchao et al. (2019)[1], tried to predict the trend of stock prices by using ARIMA(4,1,4) model, and the results showed that the prediction effect was related to the time range of prediction. The prediction accuracy was high in the short term but was poor in the long term. Nowadays, the prediction of a single model can no longer meet investors' demand for stock market prediction, not only because each model has obvious advantages and disadvantages, but also the stock market will be affected by environmental, domestic and foreign current economic factors. Therefore, many scholars combine multiple models, which can not only give full play to the advantages of a single model but also make up for the shortcomings of each model. Such as Cai Hong et al. (2011)[2] proposed a PCA-BP neural network model to perform principal component analysis on stock sequences and reduce the dimension to speed up the network prediction and improve prediction accuracy. Li Song et al. (2012) [3] proposed a particle swarm optimization algorithm-BP neural network model that adaptive mutation operator was introduced to mutate the particles trapped in the local optimal, thus improving the performance of finding the global

ISSN 2616-5775 Vol. 5, Issue 10: 41-47, DOI: 10.25236/AJCIS.2022.051007

optimal predictive value. Li Yu et al. (2017) [4] proposed a LM genetic neural network model. LM algorithm is used to improve the gradient descent algorithm of the traditional neural network, and the genetic algorithm is used to optimize the parameters of the network, so as to improve the ability of searching global optimal network and the overall convergence speed. All the above methods can obtain a more accurate stock price prediction value.

In addition to improving the algorithm, we also need to consider other economic indicators that may affect the stock price fluctuations[5]. The information of the original closing price time series may not be enough for prediction. How to extract the linear and nonlinear features between the prediction variables and the response variables and how to establish the association between the extracted auxiliary information and ARIMA are two important directions of research on machine learning algorithms. Considering the predictive advantages of the Generalized Additive Model (GAM) and BP neural network model, an improved ARIMA method based on hybrid dimension reduction and BP neural network is proposed in this paper. Firstly, we use the Auto-ARIMA model to automatically stabilize the data set composed of the stock closing price and make a linear prediction, and the residual sequence is generated. Secondly, we use PCA and KPCA to extract the linear and nonlinear features of the intraday price series (auxiliary information) and use them as input variables of the BP neural network to fit the residual between the real value and the predicted value of the ARIMA model. Finally, the results of the two parts are combined as the final predicted value. This method can make full use of the valuable information hidden in the residual, and capture the linear and nonlinear features of the auxiliary information to predict the stock closing price. The empirical results show that this method has better predictive performance than the ARIMA model and has certain robustness.

The rest of this paper is arranged as follows. The second section introduces relevant model concepts and proposes the improved ARIMA method, the third section is the prediction results and robustness analysis, and the fourth section is the conclusion.

2. Theory and method

(1) PCA and KPCA

For high-dimensional data such as intraday stock price, an important way to alleviate the curse of dimensionality is to reduce dimension. PCA aims to use the idea of spatial mapping to maintain the features of the data set with a large contribution to each other, and transform multiple indexes into a few comprehensive indexes, so as to project data into a low-dimensional subspace and achieve the effect of reducing the dimension of data space. The main process of PCA algorithm dimension reduction is shown in Figure 1.

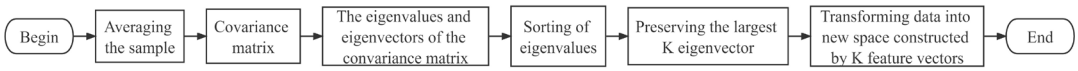

Figure 1: Main Flow Reduction Process of PCA[6]

In reality, a lot of data is nonlinear separable, so we need to introduce kernel functions to map data to high dimensional space. KPCA uses nonlinear transformation to map the input data from low dimensional space to high dimensional space so that the nonlinear problem is transformed into a linear problem. Then PCA algorithm is used to extract the main component in high dimensional space, and the purpose of reducing the dimension of the data is achieved based on maintaining the data information[7].

(2) BP neural network

BP neural network[8-9] is an "automatic feedback training process". The training process is divided into forward-propagation and backpropagation. Forward propagation is to input feature vectors into the input layer, pass through the hidden layer, and finally get the prediction results in the output layer. The process of backpropagation is to conduct the error analysis according to the results obtained by each training and the real results, and constantly revise the weights and thresholds until the results reach the set accuracy.

The relation between output and input of neuron is expressed as:

$$net_h = \sum_{h=1}^{n} w_{ih} x_h \qquad (2)$$

ISSN 2616-5775 Vol. 5, Issue 10: 41-47, DOI: 10.25236/AJCIS.2022.051007

$$y_h = f(net_h) \quad (3)$$

Where, $X = (x_1, x_2, \cdots, x_n)'$ is the input variable of the BP neural network, w_{ih} is the weight between the input layer and the hidden layer, w_{hj} is the weight between the hidden layer and the output layer, b_h is the threshold of each neuron in the hidden layer, b_0 is the threshold of each neuron in the output layer, $\widehat{Y} = (y_1, y_2, \cdots, y_n)'$ is the output result, and function f is the activation function. The BP neural network algorithm is shown in Figure 2.

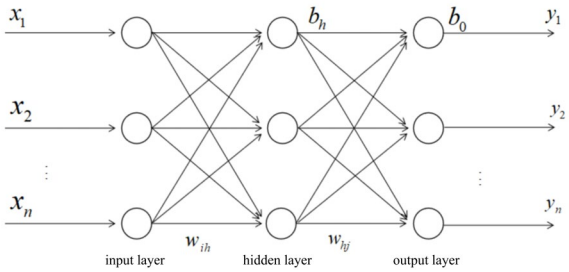

Figure 2: BP Neural Network Algorithm

According to the expected output Y and the actual output $\widehat{Y} = (y_1, y_2, \cdots, y_n)'$ of the training sample, the sum of the squared errors (SSE) of the network is calculated:

$$SSE = \sum_{k=1}^{n} \left(Y_k - \widehat{Y_k}\right)^2 \quad (4)$$

SSE is regarded as a multivariate function of X, Y, w_{ih}, w_{hj}, b_h, and b_0. According to the gradient descent method, the adjustment formula of the W parameter each time is:

$$\triangle W = -\eta \frac{\partial E}{\partial W} \quad (5)$$

$$W = \alpha W + \triangle W \quad (6)$$

Where, α is called the learning rate and η is called the impulse term.

(3) An Improved ARIMA

Previous studies have shown that ARIMA is highly suitable for detecting linear associations among variables in time series, and BP neural network is extremely sensitive to nonlinear factors[10-11]. Therefore, this paper obtained the predicted value of stock closing price through the ARIMA and BP neural network model. We take the intraday price as auxiliary information, use BP neural network model to fit residual, and finally sum up the results as the final predicted value. The specific process is shown in Figure 3.

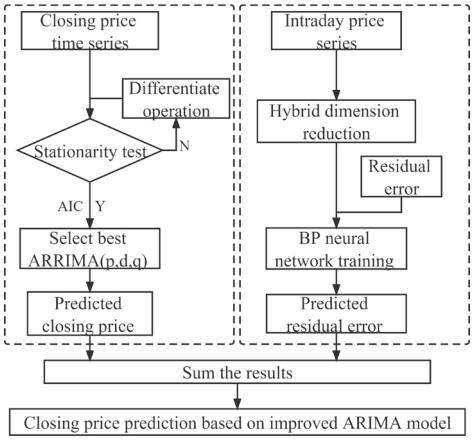

Figure 3: Improved ARIMA Model

ISSN 2616-5775 Vol. 5, Issue 10: 41-47, DOI: 10.25236/AJCIS.2022.051007

3. Data analysis

(1) Data Sources

The data in this paper comes from the Wind database. We randomly selected three stocks from Wind: Opai Home Furnishing (SH603833), Dongfang Group (SH600811) and Warbao (SZ300741), which are respectively from the material, breeding and food industries. The closing prices of these three stocks are used as the original data (242 in total), and each of them basically covers the closing prices of all trading days in 2020. The three stocks we randomly selected can all be regarded as different stochastic systems, but each stochastic system has a different complexity. We want to explore whether the ARIMA model could be improved at different complexity levels.

(2) Data Stationarity Test

Through Python visualization, we can obtain the ACF and PACF images (Figure 4-5) based on the original data of the three stocks.

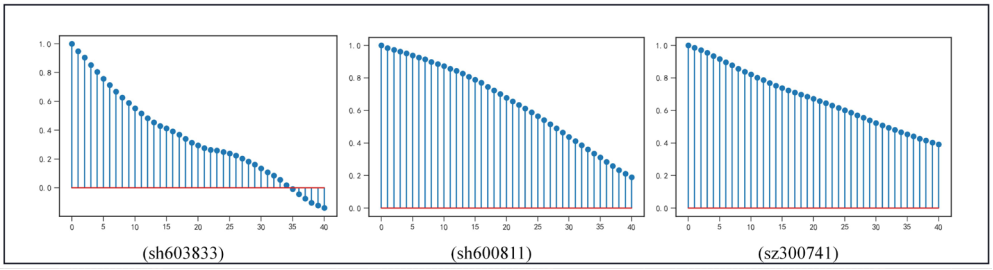

Figure 4: ACF Plot

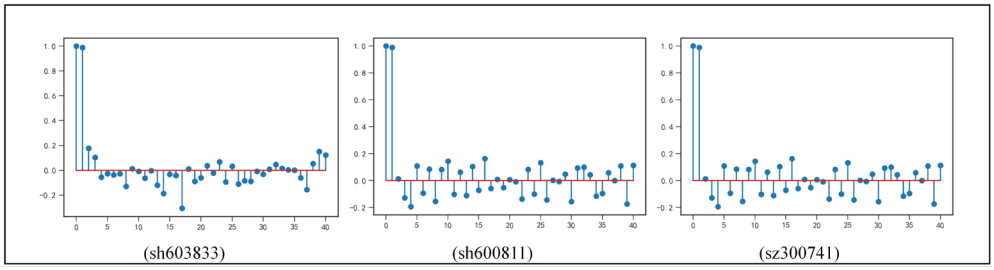

Figure 5: PACF Plot

By observing the images, we found that the closing price time series data of the three stocks showed non-stationary features, but this was far from enough for the preliminary judgment of the data. Therefore, we also conducted an ADF test on the original data. The T-statistics of the ADF test were -1.252, -0.898 and -0.898, which were all smaller than the corresponding critical values -3.4584, -2.8739 and -270.4257 when the significance level was 1%, 5% and 10%, respectively. It can be concluded that the ADF test results of the original data all fall within the acceptance of the null hypothesis interval, that is, the closing price time series data of the three stocks have unit roots and the data is not stationary.

(3) ARIMA model prediction

The Auto-Arima model in Python can automatically stabilize a data set composed of stock closing prices. We used 80% of the input data (first 193 days) as the training set and 20% (last 49 days) as the test set. AIC information criterion was used to adjust the optimal parameters of the ARIMA model through auto-regression training, and the results were shown in Table 1. Finally, we got the predicted value and residual sequence of the closing price.

Table 1: Best Model

Stock	Best model
sh603833	ARIMA (0, 0) (2, 0) [12]
sh600811	ARIMA (0,1,1) (0,1,1) [12]
sz300741	ARIMA (1,1,3) (0,1,1) [12]

(4) Hybrid dimension reduction and BP neural network fitting

We took intraday stock prices as auxiliary information, and used PCA and KPCA dimensionality reduction methods to extract 7 dimensions of linear and nonlinear features respectively, and took them as

ISSN 2616-5775 Vol. 5, Issue 10: 41-47, DOI: 10.25236/AJCIS.2022.051007

input variables of nonlinear prediction. In BP neural network, the input layer plays the role of input data, and its node number is determined by the dimension of input data. The number of nodes in the output layer is determined by the actual problems to be solved. In this paper, we wanted to predict the residual sequence, so the number of nodes in the output layer was 1. According to the actual research needs, this paper established a BP neural network model with a double hidden layer, and the number of nodes was 5 and 2 respectively to fit the residual between the real value and the predicted value of the ARIMA model. Finally, the closing price prediction value obtained by the ARIMA model and the residual prediction value obtained by BP neural network model was summed as the final closing price prediction value.

(5) The error analysis

In order to compare and analyze the effectiveness of the proposed improved ARIMA model, we selected Mean Square Error (MSE), Mean Relative Error (MRE) and Posteriori Error (C) as the accuracy evaluation indexes of the prediction model in this paper. The specific formulas are as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{y_i} - y_i\right)^2 \quad (7)$$

$$MRE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{\left|\widehat{y_i} - y_i\right|}{y_i} \quad (8)$$

$$C = \frac{S_2}{S_1} \quad (9)$$

Where, S_2 is the standard deviation of the relative value series, and S_1 is the standard deviation of the original series.

Table 2: Error Analysis Result (1: ARIMA, 2: Improved ARIMA)

	MRE		C	
	1	2	1	2
sh603833	2.9584	3.5967	0.4847	0.6170
sh600811	2.8252	3.0261	0.2757	0.3050
sz300741	3.4191	3.4336	0.1909	0.1949

By analyzing the results of the traditional ARIMA model and the improved ARIMA model (Table 2), it can be seen that the improved ARIMA model has better prediction results than the traditional ARIMA model, and the four evaluation indexes are all smaller than the traditional ARIMA model.

(6) Robustness test

Few existing studies have tested the scientificity and consistency of the prediction methods and results of the ARIMA model, so a robustness test will be added in this paper to make up for the inadequacy of existing literature. Since we cannot obtain the control data from previous studies, it is necessary to adjust the model optimization path such as the model parameter and data set division to achieve self-test. We successively took 60%, 70%, 80% and 90% of the dataset as the training set, and the rest as the training set. The results were shown in Figure 6-8 below.

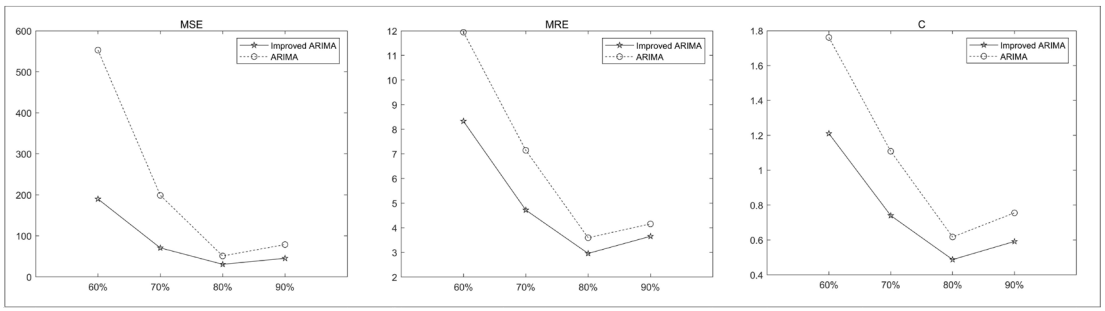

Figure 6: Error analysis of Stock sh603833

ISSN 2616-5775 Vol. 5, Issue 10: 41-47, DOI: 10.25236/AJCIS.2022.051007

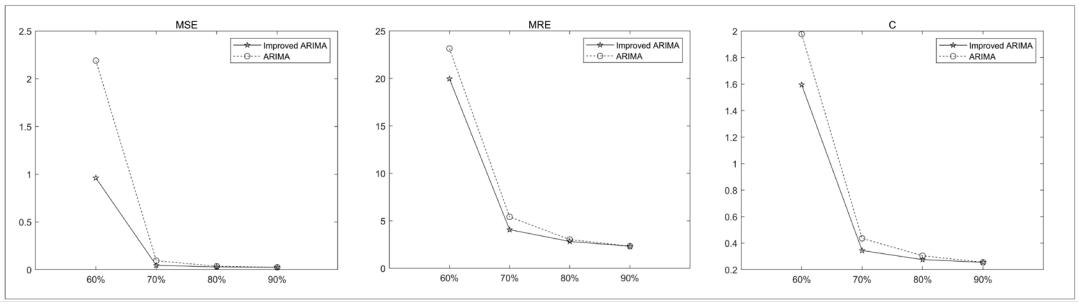

Figure 7: Error analysis of Stock sh600811

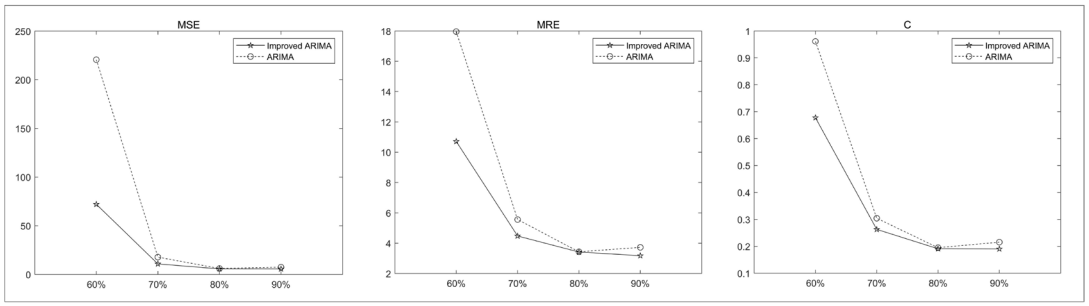

Figure 8: Error analysis of Stock sz300741

The results show that the prediction results of the improved ARIMA method based on hybrid reduction and BP neural network proposed in this paper have high reliability.

4. Conclusion

In this paper, we propose an improved ARIMA method on hybrid reduction and BP neural network. The empirical results show that the model has a good prediction effect on stock closing price and has certain robustness. At the same time, we also find that the stock closing time series does exist in both linear and nonlinear parts. The innovation of this paper is that we take the intraday price of the stock as auxiliary information, and use PCA and KPCA methods to extract its linear and nonlinear characteristics respectively. That is to say, this method can make full use of the valuable information hidden in residuals. However, people are more likely to ignore the uncertainty brought by different model combinations in the process of model selection, leading to the underestimation of the actual variance[12]. Meanwhile, because the model selection method generally bases the inference on the selected model, it may lead to the loss of information reflected by other models or information unique to other variables. In the future, we can use the model average method to solve the model combination weight problem and improve the prediction accuracy in the subsequent model optimization process.

References

[1] Yingchao Zhang, Yingjun Sun. Empirical research on Shanghai Stock Index Analysis and Prediction based on ARIMA Model [J]. Economic Research Guide, 2019(11):5.

[2] Hong Cai, Rongyao Chen. Research on Stock Price Prediction based on PCA-BP Neural Network [J]. Computer Simulation, 2011, 28(3):4.

[3] Song Li, Lijun Liu, Man Zhai. Optimization of BP Neural Network for Short-term Traffic Flow Prediction by Improved Particle Swarm Optimization [J]. Systems Engineering-Theory & Practice, 2012, 32(9):2045-2049.

[4] Jianfeng Guo, Yu Li, Dong AN. Short-term Stock price Prediction based on LM Genetic Neural Network [J].Computer Technology and Development, 2017, 027(001):152-155,159.

[5] Joseph A, Larrain M, Turner C. Daily Stock Returns Features and Forecastability [J]. Procedia Computer Science, 2017, 114:481-490.

[6] Fengjuan, Miao, Tongri, et al. Design of BP Speaker Recognition System Based on KPCA-MFCC Parameter Optimization[C]// 2018 International Conference on Mechanical, Electrical, Electronic Engineering & Science (MEEES 2018). 0.

[7] Wang H, Hu W. Optimization of Pathological Voice Feature Based on KPCA and SVM[C]// Chinese

ISSN 2616-5775 Vol. 5, Issue 10: 41-47, DOI: 10.25236/AJCIS.2022.051007

Conference on Biometric Recognition. Springer International Publishing, 2014.
[8] 2021.Chunlei Yu, Mengyue Li , Weishi Yin . Journal of Changchun University of Science and Technology: Natural Science Edition, 2021.
[9] Xiaoling Chen. Stock price prediction based on ARIMA model and neural network model [J]. Economic Mathematics, 2017, 34(4):5.
[10] Peter G, Zhang. Time series forecasting using a hybrid ARIMA and neural network model [J]. Neurocomputing, 2003.
[11] Ümit Çavuş Büyükşahin, Şeyda Ertekin. Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition [J]. Neurocomputing, 2019, 361:151-163.
[12] Xinyu Zhang, Guohua Zou. Model average method and its application in forecasting, Statistical Research, 2011, 28(6):6.