Universitas Ahmad Dahlan

Jurnal Ilmiah Matematika
Vol. 8, No. 1, April 2021, pp. 40-47
ISSN 2774-3241
http://journal.uad.ac.id/index.php/Konvergensi

Empirical analysis of shanghai stock exchange index based on armia model and neural network model

Xin Jian [a,1,*], Weizhang Lai [a,2]

[a] Guangxi Normal University, Guilin, China 541004
[*1] jianxinxka@163.com ; 2817320567@qq.com
*Correspondent Author

KEYWORDS

ARIMA model
NN model
short-term forecast

ABSTRACT

Stocks are an important part of the national economy. With the increase of liquidity in people's hands, more and more people choose to enter the stock market. In stock investment, accurate prediction of stock price index is of great significance to investors and promotes the development of my country's stock market. It even has an important role in accelerating my country's economic development. The paper chose the ARIMA method based on linear technology for time series forecasting and the NN model that is good at mining the implicit nonlinear relationship in the data to compare the China Sea Securities Composite Index from January 21, 2020 to December 31, 2020. Empirical analysis of closing prices and short-term forecasts are made

Introduction

Stock price index, abbreviated as a stock index, is compiled using the index method in statistics to measure and reflect the overall price level of the entire stock market or an indicator of the change and trend of a specific type of stock price. It can sensitively reflect the changes in the stock market and is known as a barometer of a country or region's economic, political, and social conditions. The Shanghai Composite Index is an important indicator of the stock volatility of the Shanghai Stock Exchange. Therefore, the accurate prediction of the Shanghai Composite Index is of great significance for both individuals and the country.

The outbreak of COVID-19 in 2020 has had a substantial impact on domestic and foreign financial markets. After the outbreak, on the first day of opening the Shanghai and Shenzhen markets, 3188 stocks in the two markets fell by their limit. The Shanghai and Shenzhen 300 Index closed down 7.88% on February 3, the largest drop since 2015 [1]. However, as the domestic epidemic is effectively controlled, the stock market situation is slowly improving. This paper uses daily data of 00001 shares of the Shanghai Stock Exchange from January 21, 2020 to December 31, 2020, and uses ARIMA and neural network(NN) methods to predict the Shanghai Stock Exchange index. Moreover, providing a specific reference for investors.

1.2 Stock forecasting methods and literature review

Stock index prediction is an important research direction of the stock market. This paper mainly uses quantitative research methods. Quantitative analysis methods are mainly divided into model analysis methods and neural network analysis methods. Model methods mainly include VAR model, ARMA model, GARCH model, Markov model, ARIMA model, etc. More and

 ISSN 2774-3241

more papers have used the ARIMA model to model and predict their data in recent years, so this paper also uses the ARIMA model for analysis. Zha Zhenghong used time series analysis based on historical data of the Shanghai Stock Exchange Index to conduct modeling analysis and research and established an ARIMA model[2]. Li Zhanjiang used ARIMA to establish a prediction model for stock index futures prices. During the 13th, the closing price of the Shanghai and Shenzhen 300 stock index futures was analyzed, and it was found that the ARIMA model had a reasonable forecast of the price trend of stock index futures[3]. Li Jiasong used the ARIMA model to predict the Shanghai and Shenzhen 300 Index, and pointed out the reasons for the difference between the actual value of the Shanghai and Shenzhen 300 Index and the predicted value of the model, and provided investors with a prediction method for the Shanghai and Shenzhen 300 Index[4]. Zhang Yingchao also used the ARIMA (4, 1, 4) model to predict future stock prices. The results show that the model can predict the future Shanghai Composite Index more accurately in the short term[5]. Huang Lixia took Ping An Bank of China as an example and selected 244 sample data of P/E ratios from January 1, 2019 to December 31, 2019 as the research objects. Based on the ARIMA model, the yield of Ping An Bank in the next 5 working days is calculated. The analysis and prediction results show that the predicted and actual values error does not exceed 6.5%. It can be seen that the prediction accuracy of the model is very high, and the predicted value is very close to the true value[6]. Liu Huihao, Jiao Wenniu, etc. (2020) collected the daily data of 26 listed banks in China in 2017, established an ARIMA model to predict stock prices, and used an intervention analysis model to study the policy effects of tightening supervision empirically. . The results show that the release of the regulatory policy in April 2017 lowered the stock price of banks in a short period of time. Dividing banks into two types: large banks and small and medium-sized banks, it can also be concluded that the regulatory policy harms the stock prices of large banks and small and medium-sized banks[7].

Another analysis method is the BP neural algorithm. Wu Wei summarized a series of methods on selecting samples, assigning initial weights, the number of hidden layers, the number of neurons in each layer, and the selection of activation functions through a large amount of experimental data analysis. And compared the results of different choices to illustrate their advantages and disadvantages[8]. Deng Kai combined genetic algorithm and BP neural network to predict the stock price of Kweichow Moutai. Empirical analysis showed that the optimized BP neural network had higher prediction accuracy and application value[9]. Yin Lu combined the genetic algorithm and the neural network model as the research object. In terms of the prediction effect, the BP neural network is feasible. GA-BP9 is adopted to improve the prediction accuracy of the model[10]. Liu Jiaqi et al. (2018) combined the principal component analysis method, genetic algorithm and BP neural network algorithm to establish a PCA-GA-BP model for predicting stock price changes. This model has improved the slowness of the BP network operation and easy to fall into disadvantages of local minima[11]. Zhang Rumeng and Zhang Huamei explored the accuracy of stock prices, the forecasting accuracy of stock prices, and the rise and fall trends by constructing a PCA-BP (Principal Component Analysis-Back Propagation) neural network comprehensive model. The research found that: Combining the principal component analysis method with the BP neural network model, the test showed that the error was the smallest when the hidden layer node is 7, and it could predict the rise and fall of stocks 100%; the PCA-BP neural network comprehensive model was constructed, and the comprehensive model predicted the rise and fall of stocks. The fall accuracy was 95%. The stock price error was relatively reduced; the comprehensive model had more advantages than a single PCA-BP neural network model, and it could better give investors practical suggestions [12]. Lin Guochao, Du Yujian took BYD (SZ002594) stock price from June 24, 2019 to April 28, 2020 as the research sample, and used two different combination models to separately train the samples and predict the changes in stock prices. It is concluded that the prediction accuracy based on the BP-GM(1,1) model is higher, and the dynamic analysis of stock prices could provide a more useful reference [13].

 ISSN 2774-3241

Method

ARIMA model

The full name of the ARIMA model is the differential autoregressive moving average model and is also denoted as ARIMA (p, d, q). It is one of the commonly used statistical models in time series forecasting. Box and Jenkins developed it in the early 1970s, so it is also called the Box-Jenkins model, Box-Jenkins method. It is essentially the ARMA model, which is an extension of the autoregressive moving average model. The ARIMA model combines three basic methods: autoregressive (AR), integral (I), and moving average (MA). In addition, it contains three parameters: p, d, q, and the meanings of which are shown in the following Table 1.

Table 1. The meaning of parameters (p, d, q)

P	The number of lags (lags) of the time series data used in the prediction model, also known as the AR hierarchy
d	Represents the number of differences made when the time series is stationary, also called the integrated term
q	Represents the number of lags (lags) of the prediction error used in the prediction model, also called the MA order

The basic idea of the ARIMA model is to treat the data sequence formed by the prediction object over time as a random sequence and use a certain mathematical model to approximate this sequence. When the model is identified, the future value can be predicted from the past and present values of the time series. Theoretically, the mathematical description of the ARIMA model is:

$$\Delta^d z_t = \theta_0 + \sum_{i=1}^{p} \phi_i \Delta^d z_{t-1} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \quad (1)$$

Among them, $\Delta^d z_t$ represents the sequence of z_t after d differences, ε_t is the random error at time t, and ε_t (t=1,2,3,...) is a mutually independent white noise sequence, which obeys zero mean, The variance is a normal distribution with a constant σ^2. ϕ_i (i=1,2,...,p) and θ_j (j=1,2,...,q) are estimated parameters. It is not difficult to see from the mathematical expression that the essence of the ARMIA model is still a linear model, which also shows that the ARMIA model is insufficient in portraying the nonlinear characteristics of the time series.

Neutral network (NN) model

The neural network was originally developed to imitate the human brain. It consists of a large number of nodes (or "neurons", or "units") and interconnections. It is a computing system that creates predictions based on existing data. The neural network includes an input layer, a hidden layer and an output layer. Among them, the input layer is a layer obtained based on existing data; the hidden layer is a layer that uses direction propagation to optimize the weight of input variables to improve the predictive ability of the model; the output layer is a data output prediction based on the input and hidden layers. So far. There are many types of prediction models used in the economic and financial fields. The multi-layer feedforward neural network based on the error back propagation algorithm is the most widely used NN model[14].

Back-ProPagation Network(BP network) is also called a back-propagation neural network. Through the training of sample data, the network weights and thresholds are constantly modified to make the error function drop in the negative gradient direction and approach the expected output. Thus, it is a kind of "signal forward propagation →The process of error backpropagation". The specific structure of the neural network is shown in Figure 1.

 ISSN 2774-3241

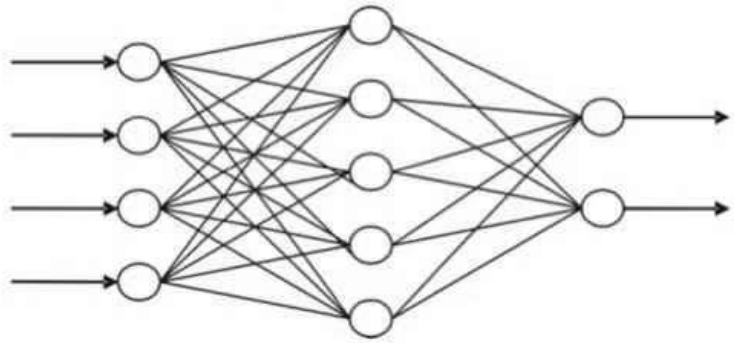

Figure 1. Three-layer BP network structure model diagram

The basic idea of the algorithm of the BP model is to describe the initial weight and threshold value of the network, calculate the output value of the network through the forward information transfer between layers, and modify the network according to the error between the actual output and the expected output of the network. Through continuous repeated training and comparison, weights and thresholds minimize the difference between the actual output expectations. The specific description is as follows [15].

Initialization of the network

Determine the input and output layer parameters n, m; and calculate the appropriate number of hidden layers by trial and error. Supposing the weight from the input layer to the hidden layer is ω_{ij}, the weight from the hidden layer to the output layer is v_{jk}, the threshold from the input layer to the hidden layer is a_j, and the threshold from the hidden layer to the output layer is b_j.The excitation function is f(x). The excitation function is the Sigmoid function. The form is:

$$f(x) = \frac{1}{1+e^{-x}} \tag{2}$$

Hidden layer output

The output of the hidden layer is:

$$H_j = f\left(\sum_{i=1}^{n} \omega_{ij} x_i + a_j\right) \tag{3}$$

Output of the output layer

$$O_j = f\left(\sum_{j=1}^{l} H_j v_{jk} + b_k\right) \tag{4}$$

Calculation of error

Taking the error formula as:

$$Q = \frac{1}{m}\sum_{k=1}^{m}(y_k - O_k)^2 \tag{5}$$

Remembering $y_k - O_k = e_k$, then Q can be expressed as $\frac{1}{m}\sum_{k=1}^{m} {e_k}^2$

Weight update

Expanding the above error definition to the hidden layer, then we have:

 ISSN 2774-3241

$$Q = \frac{1}{m}\sum_{k=1}^{m}\left[f\left(\sum_{j=1}^{l} H_j v_{jk} + b_k\right) - O_k\right]^2 \quad (6)$$

Further, expand to the input layer, then we can get:

$$Q = \frac{1}{m}\sum_{k=1}^{m}\left\{f\left[\sum_{j=1}^{l}\left(\sum_{i=1}^{n}\omega_{ij}x_i + a_j\right)v_{jk} + b_k\right] - O_k\right\}^2 \quad (7)$$

It can be seen from the above formula that the error is a function of ω_ij and v_jk of the intensity of each layer, so the intensity adjustment can be achieved by reducing the error value.

Empirical analysis of ARIMA model and NN model

This paper selects the closing price of 00001 shares of the Shanghai Stock Exchange from January 21, 2020 to December 31, 2020 as the research object. There are 230 pieces of data in total. The R software is used to construct the ARIMA and the NN models to predict the above data.

Empirical analysis of ARIMA model

We are using the R software to draw a time series chart of the Shanghai Stock Exchange Index (January 21, 2020 to December 31, 2020), as shown in Figure 2.

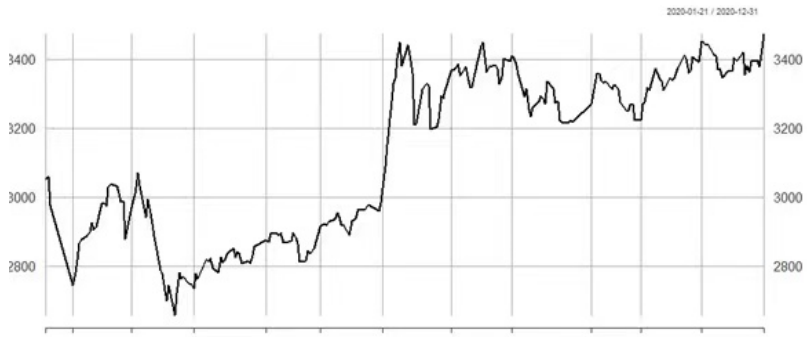

Figure 2. Shanghai Composite Index Time Series Chart

Note: The horizontal axis represents the time (from January 2020 to December 2020), and the vertical axis represents the closing price.

It can be seen from Figure2 that the general trend of the Shanghai Composite Index from January 21, 2020, to December 31, 2020, does not show a clear seasonal trend, and the fluctuation range is large, which is a non-stationary sequence. Processing the data for stationarity: Take the logarithm of the variable, record it as LY, and then perform the first-order difference and record it as DLY. It can be roughly judged by the time-series graph that the processed series has stationarity. ADF can further test the stationarity of DLY data, and the test results are shown in Table 2.

Table 2. ADF inspection result

ADF value	*Lag order*	*P-value*	*Conclusion*
-6.2253	6	0.01	sationary

The above results show that the sequence DLY is stationary, so the parameter d=1 of the model.

After the sequence is stationary, observe the autocorrelation graph and partial correlation graph, consider the minimum AIC criterion, compare and analyze different

 ISSN 2774-3241

parameters, and finally determine the model as ARIMA (4, 1, 1). The estimated results of the model are as follows.

$$\mathrm{DLY}_t = 0.849\mathrm{DLY}_{t-1} - 0.0383\mathrm{DLY}_{t-2} + 0.0204\mathrm{DLY}_{t-3} - 0.0954\mathrm{DLY}_{t-4} - 0.7949v_{t-1}$$

If the Obtained model fits well, the model's residuals should satisfy an independent normal distribution with zero means. Using the R software to draw the QQ diagram of the residuals of the sequence (Figure3), and you can find that the points on the sequence are basically and evenly distributed on the line, so the residuals satisfy the normality assumption.

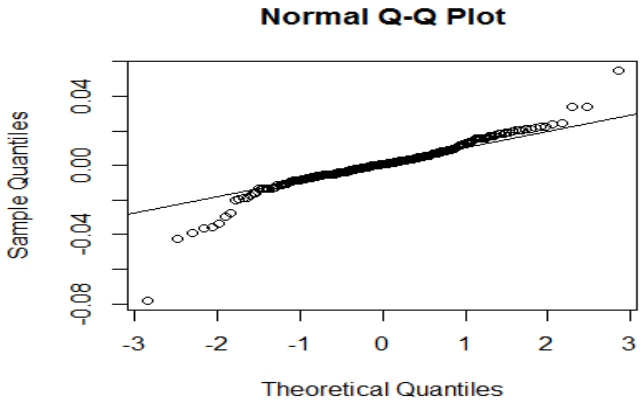

Figure 3. Normal test Q-Q plot

Finally, the white noise in the model sequence is tested. If it is white noise, all valuable information of the sequence has been fully extracted. If it is not white noise, it needs to be re-modeled. In this paper, the Box.test() function of R is used to test, and the test result is shown in Table 3.

Table 3. White noise test

X-squared	*df*	*P-value*
0.0058753	1	0.9389

The test result p=0.9389>0.05 that the model passed the white noise test, indicating that the established ARIMA model has a better fitting effect because DLYt=△LY_tthe prediction model can be expressed.

$$\mathrm{LX}_t = 1.849\mathrm{LY}_{t-1} - 0.8873\mathrm{LY}_{t-2} + 0.0587\mathrm{LY}_{t-3} - 0.1158\mathrm{LY}_{t-4} + 0.0954\mathrm{LY}_{t-5} - 0.7949v_{t-1}$$

The antilog operation of the above formula can get the prediction result of the Shanghai Stock Exchange Index. For example, using the ARIMA model to statically predict the closing price of the Shanghai Stock Exchange Index for 10 working days after December 31, 2020, the results are shown in Table 4:

Table 4. Forecast of the closing price of the Shanghai Composite Index

Serial number	*Predictive value*	*Serial number*	*Predictive value*
1	3474.940	6	3459.696
2	3476.791	7	3455.965
3	3476.041	8	3453.463
4	3469.733	9	3451.908
5	3464.274	10	3451.043

 ISSN 2774-3241

Empirical analysis of BP model

Before using the BP neural network model to predict the data, the data must be preprocessed first, and the following processing is performed:

$$\varphi(t) = \frac{\phi(\mathrm{t}) - \phi_{\min}}{\phi_{\max} - \phi_{\min}}$$

Among them, $\phi_{\min}$ and $\phi_{\max}$ The corresponding sequence's minimum and maximum values, respectively, and the data range of φ(t) after processing is between [0,1]. Using the first 80% (184) of the processed data as training data (training set) to train the BP neural network, and the remaining 20% (46) as test data (test set) to test the model Ability to predict.

In this paper, a three-layer pre-neural feedback network is used. The more complicated part of building the model is determining the network input nodes and hidden layer nodes. According to previous experience, the number of nodes is continuously changed by trial and error to improve the convergence speed and fitting ability of the network are finally determined as the input layer node is 4, the hidden layer node is 4, and the output layer node is 1, that is, a 4×4×1 network structure is adopted.

After you are ready, you can use the data to train the neural network. To facilitate comparison, draw a line graph between the predicted value and the test set (Figure 4).

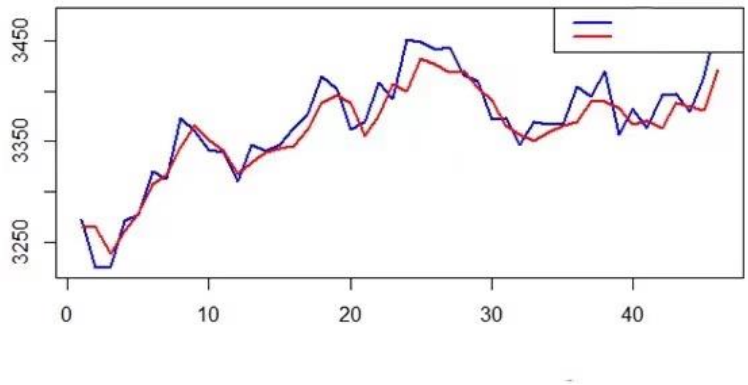

Figure 4. Trend chart of true value and predicted value

Note: The horizontal axis represents the number of days, and the vertical axis represents the closing price (the broken blue line is the test set, and the red one is the prediction set)

It can be seen from Figure 4 that the prediction points made by the BP neural network are very close to its fitting line, and it can be seen that the fitting effect of the model is very good. The above generally shows that the BP neural network was effectively forecasting the closing price of the Shanghai stock index.

Conclusion

Stock price forecasting is a significant research hotspot in the field of forecasting, with various methods used. This paper is based on the Shanghai Stock Exchange Index time series from January 21, 2020 to December 31, 2020. Domestic stock price research results selected two models commonly used in the time series forecasting-ARIMA model and BP neural network model. Through the above empirical analysis of the Shanghai Composite Index, we know that the ARIMA (4,1,1) model established by the R software and the BP model with a 4×4×1 network structure has a better fitting predictive effect. Therefore, it can provide a certain investment decision basis for investment.

 ISSN 2774-3241

References

[1] Wang Qing, Wang Zhongli, Li Shixue, Xue Fuzhong. 2020. "The short-term impact of the 'new crown pneumonia' epidemic on the price fluctuations of China's stock market". Economic and Management Review, vol.36,no.06,pp.16-27,

[2] Zha Zhenghong. 1999 "Statistical analysis and prediction of Shanghai Composite Index". Journal of Shanghai Maritime University,vol.04,

[3] Li Zhanjiang, Zhang Hao, Sun Pengzhe, Tong Guochao, Zhang Zhihao. 2013 "Research on Shanghai and Shenzhen 300 stock index futures price prediction based on ARIMA mode". Journal of Ludong University (Natural Science Edition), vol.29,no.01,pp.22-24.

[4] Li Jiasong. 2017 "Shanghai and Shenzhen 300 Index Forecast and Error Factor Analysis Based on ARIMA Model". Journal of Chifeng University, vol.33,no.1,pp. 73-75,.

[5] Zhang Yingchao, Sun Yingjun. 2019 "An Empirical Study on the Analysis and Forecast of Shanghai Stock Exchange Index Based on ARIMA Model". Economic Research Guide,vol.11,pp.131-135,

[6] Wu Wei, Chen Weiqiang, Liu Bo. 2001 "Using BP neural network to predict stock market rise and fall" . Journal of Dalian University of Technology,vol.1,pp.9-15,.

[7] Deng Kai, Zhao Zhenyong. 2009 "Research and simulation of stock market prediction model based on genetic BP network". Computer Simulation, vol.26,no.05,pp.316-319,.

[8] Yin Lu. 2010. "The theory and application of stock prediction based on GA-BP neural network". North China Electric Power University (background) ,.

[9] Liu Jiaqi, Liu Dehong, Lin Tiantian. 2018. "Research on Stock Price Based on BP Neural Network Model". China Business Journal, vol.08,pp.29-30.

[10] Zhang Rumeng, Zhang Huamei. 2020. Predicting stock prices based on the PCA-BP neural network comprehensive model. Computer Knowledge and Technology, vol.16,no.33,pp. 4-7,.

[11] Huang Lixia. 2020. "Analysis and forecast of stock price based on ARIMA model——Taking Ping An of China as an example". Science and Technology Economic Market, vol.10,pp. 62-63,.

[12] Liu Huihao, Jiao Wenniu, Liu Yue. 2020. "The impact of new regulatory regulations on the stock prices of my country's listed banks: An analysis of policy intervention based on the ARIMA model". Financial Theory Research, vol.05,pp.21-31.

[13] Lin Guochao, Du Yujian, Liu Juan. 2020. "The application of two types of combined BP neural network models in stock price forecasting". Modern Business and Trade Industry, vol.41,no.26,pp.141-143.

[14] Velldo A,Liaboa P J G,Vaughan J. 1999. "Neural network in business : A survery of applications(1992-1998)".Expert Systems with Applications, vol.17,pp.51-54.

[15] Xiong Zhibin. 2011. "Research on GDP time series forecast based on the integration of ARIMA and neural network". Mathematical Statistics and Management, vol.30,no.02,pp. 306-314.