CSI 300 Index Forecast

-- Based on ARIMA Model, BP Neural Network Model and Combined Model

Yan Hu

College of Economics, Jinan University, Guangzhou 510632, China.

Abstract

CSI 300 index plays an important role in the financial market because of comprehensively reflect changes in the stock market. The ARIMA model and BP neural network and BP-ARIMA combination model are used to predict the CSI 300 index from January to July 2019, and the average absolute error and average relative error of the three models are compared in this paper. The model predicts the results and the results show that the BP-ARIMA combined model has the best prediction effect.

Keywords

CSI 300 Index; ARIMA model; BP neural network; Combined model; Entropy.

1. INTRODUCTION

In recent years, more and more investors favor the coexistence of risk and profit in the stock market. Investors hope to find out the internal law of stock market changes, and to forecast the stock trend accurately and build corresponding strategies, to avoid risk and achieve higher profits. As an important part of stock forecast, stock index can reflect the general trend of stock market to some extent. There are many kinds of time series analysis methods, one of the most widely used time series models is ARIMA model, which is due to its simplicity, feasibility and flexibility. In addition, the neural network model analysis is also a common model to analyse time series nowadays [1].

The ARIMA model and BP neural network model are established to forecast the CSI 300 Index, and the precision of the two models is compared to verify the validity of the models. By studying the CSI 300 Index which can reflect the stock market synthetically, the ARIMA model and BP neural network model are established respectively to forecast the future index. Because each model has its own characteristics and limitations, so only using a single model for forecasting is not comprehensive enough. Entropy method is used to give different weights to the two models in order to overcome the shortcomings of the single model and improve the prediction accuracy.

2. THEORETICAL MODELS

2.1. ARIMA Model

At present, ARIMA model has been widely used in linear time series forecasting. In ARIMA (p,d,q) , AR is autoregressive, p is autoregressive, MA is moving average and q is moving average, d is the number of difference times that the time series is changed to a stationary series. ARIMA model is the combination of difference operation and ARIMA Model. Getting the stable time series by difference operation firstly, then use ARIMA model to forecast. Financial time series are complex and unstable, so it is necessary to deal with the time series and get the difference series to build the model. If the time series is stationary after the difference, it is necessary to judge whether the time series after the difference is a White noise. If the time series after the difference is not a White noise, it shows that the time series contains certain deterministic information and can be modeled by ARIMA model to make predictions. The mathematical expression of the ARIMA model is:

$$\Delta^{\mathrm{d}} y_t = \theta_0 + \sum_{i=1}^{p} \phi_i \Delta^{\mathrm{d}} y_{t-1} + \varepsilon_{\mathrm{t}} + \sum_{\mathrm{j}=1}^{q} \theta_i \Delta^{\mathrm{d}} \varepsilon_{t-1} \tag{1}$$

The time series y_t, which represents the original time series and $\Delta^d y_t$ the stationary time series after d order difference, ε_t is the random error term at t time, and is an independent White noise, which obeys the normal distribution with mean zero and varianceσ^2constant. P and q represent the order of the model,Φ_i and θ_j represent the estimated parameters of the model.

2.2. BP Network

BP Network consists of input layer, one or more hidden layer and output layer. There is no coupling in the same layer node, and the output of each layer node only affects the output of the next layer node. The learning process of network consists of two parts: forward propagation and backward propagation. In the network training phase, the sample data is passed through the input layer, the hidden layer and the output layer. Compare the output to the expected value. If it does not reach the required error level or training times, that is, through the output layer, hidden layer and input layer to adjust the weights for making the network become an adaptability model. The neural network structure is shown in figure 1 below.

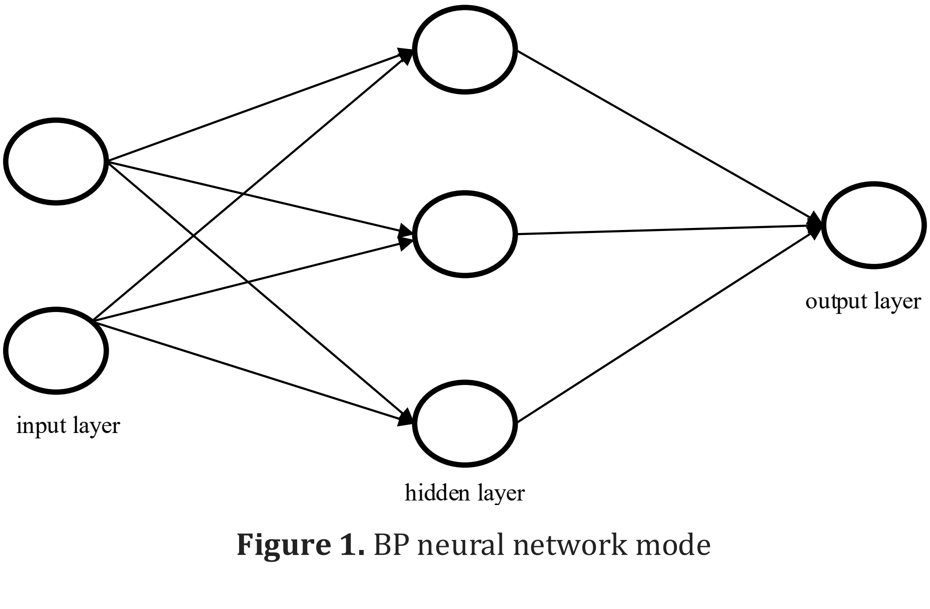

Figure 1. BP neural network mode

BP neural network model can also predict time series of single variable. For example, taking a certain time series X_1, X_2 ,... X_n , it is assumed that there is a functional relationship between the predicted value and the first m values, and constructing the functional relationship $X_t = F(X_{t-1}, X_{t-2}, \ldots X_{t-m+1})$. The predicted value X_t, as an output variable Y , the first m values $(X_{t-1}, X_{t-2}, \ldots X_{t-m+1})$ is taken as the input vector X , and the future value is predicted by BP neural network. With the development of artificial intelligence in recent years, BP neural network has become one of the main models for predicting nonlinear time series [2].

2.3. Entropy

The key of combination forecasting is how to determine the weight coefficient of each forecasting method, because different weight coefficient can get different combination forecasting results. The traditional linear programming method is often used to seek the optimal solution of the model, and it is easy to fall into the local optimum, and there is no reasonable explanation for the negative weight. The entropy method is used to get the weight coefficient of the model. The procedure for determining the weighting coefficient of a combination forecast by the entropy method is as follows:

(1) The relative error series of each single prediction method are normalized

$$p_{it} = \frac{e_{it}}{\sum_{i=1}^{N} e_{it}}, \quad t = 1,2,\ldots, \quad N \tag{2}$$

(2) Calculate the entropy of the relative error of the prediction of the single prediction method i:

$$k = 1/\ln N \tag{3}$$

$$h_i = -k\sum_{i=1}^{N} p_{it} \ln p_{it} \quad i = 1,2,\ldots, \quad m \tag{4}$$

(3) Calculate the Coefficient of variation of the prediction relative error series for the first single-term prediction method:

$$d_i = 1 - h_i \quad i = 1,2,\ldots, \quad m \tag{5}$$

(4) Calculate the weights of the various forecasting methods:

$$l_i = \frac{1}{m-1}\left(\frac{d_i}{\sum_{i=1}^{m} d_i}\right) \qquad i = 1,2,\ldots, \quad m \tag{6}$$

m is the number of models, N is the number of samples forecast [3].

3. EMPIRICAL ANALYSIS

3.1. Sample Data Selection

The model selects the CSI 300 Index from 2 Jan 2019 to 25 Jul 2019 as the raw data (the data comes from the WIND information), and uses the last ten sets of data as the prediction set to evaluate the prediction accuracy of each model.

3.2. ARIMA Modeling

3.2.1 Stability test

When modeling time series, the first thing is to judge its stability. The original data is transformed into a graph, and the time series is observed to have certain tendency, which does not fit the characteristics of zero mean and variance.

The AFC autocorrelation of the raw data is shown in figure 2. It is found that the trend to zero is very slow and the Autocorrelation is greater than zero for a long time. It indicates that the sequence is non-stationary. The unit root test is performed on the sequence, and the results are shown in Table 1. Because the P value of the original time series is greater than 0.05, the original hypothesis can not be rejected, which indicates that the original time series has units and is unstable. The P value of the unit root test is less than 0.05 for the time series after the first difference, which indicates that the time series after the difference does not have the unit root, that is, the new time series is stationary, the precondition of ARIMA modeling is satisfied and d = 1.

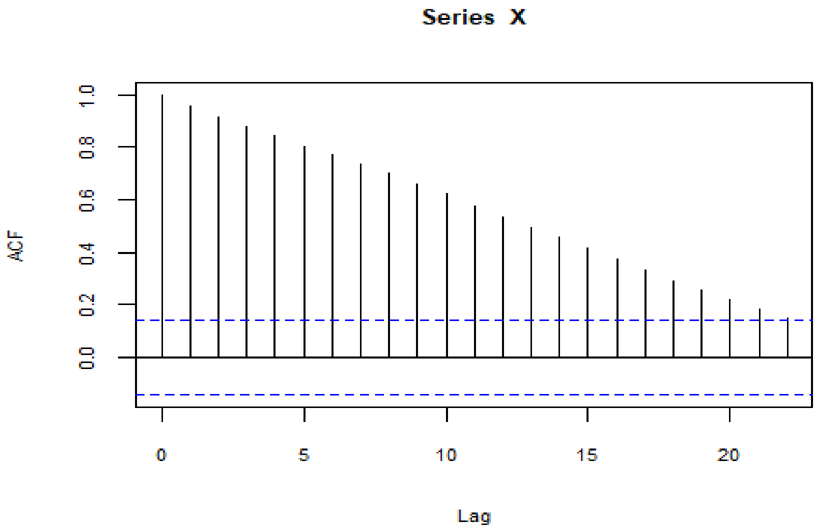

Figure 2. ACF diagram of raw data

Table 1. The results of ADF

ADF	P value
Unit root test of original sequence	0.9204
Unit root test of first-order difference sequence	<2.2e-16

3.2.2 ARIMA model fitting and parameter estimation

In order to determine the value of p and q in ARIMA (p, 1, q), the traditional method is to judge and select by observing ACF and PACF. But this approach is inaccurate, subjective and can only be used as a reference. In recent years, R language has many kinds of order-fixing methods for ARIMA. This paper uses the method of determining the order of BIC value minimum and EACF theory to judge. The results of both methods show that the fitting result of the model is ARIMA (1,1,0), so the model is selected to forecast the closing price of the CSI 300 Index.

3.2.3 model diagnosis and testing

Ljung-Box is used to test the white noise of the model ARIMA (1,1,0) residual sequence. The results show that the p value is 0.9357 and greater than 0.05, so the sequence can not reject the original hypothesis of pure randomness. We can think that the fluctuation of the series has no statistical law, that is, the series is White noise, and the model fitting effect is better. The Arima (1,1,0) model is used to forecast the following 10 sets of data, and the results are compared with the actual values as shown in Table 2.

3.3. BP Neural Network Model Establishment

3.3.1 Data pre-processing

According to the closing price of the time t, the closing price of the time t-1,t-2,t-3 can be regarded as input layer neuron, and the closing price of the time t can be regarded as output layer neuron. Because the neural network works best when the input data is scaled to a narrow range near zero, the data needs to be readjusted using normalized or normalized functions. In order to avoid the phenomenon of large error caused by the difference of the magnitude of dependent variables, the method of standard variation of range is used to process the model.

3.3.2 determination of the number of neurons in the hidden layer

The number of neurons in each hidden layer should be close to the number of neurons in the input layer and the output layer, and should not be more than twice the number of neurons in the input layer. Therefore, the number of hidden layer neurons can be reduced to 3 ~ 6. By comparing the average error between the training data and the prediction data, the number of hidden nodes is 3, so the BP neural network with 3-3-1 structure is established for prediction. The results are shown in Table 2.

3.4. Construction of BP-ARIMA Combined Model

How to determine the weight coefficient of each model method is the key to the construction of composite model. Different combination forecasting results are obtained with different weight coefficients. At present, many methods such as artificial definition, negative variation and least square are used to determine the weight. Although these methods solve the problem of weight value to some extent, the implicit information of weight value is not well considered. So this paper uses the entropy method of information theory to gets the model weight coefficient [4]. According to the relative error of the two models, the weight of ARIMA model is 0.422, and the weight of BP neural network model is 0.578. According to the weight, the predicted value of the combined model is calculated. The predicted values of the three models are shown in Table 2.

Table 2. Prediction results of three models

Date	Actual value	predicted value (ARIMA)	predicted value (BP)	predicted value (BP-ARIMA)
2019/7/12	3808.731	3801	3792.276	3795.957
2019/7/15	3824.188	3793	3813.575	3804.892
2019/7/16	3806.845	3787	3826.032	3809.561
2019/7/17	3804.638	3785	3810.334	3799.643
2019/7/18	3768.402	3809	3810.667	3809.963
2019/7/19	3807.955	3825	3776.031	3796.696
2019/7/22	3781.683	3806	3814.562	3810.949
2019/7/23	3789.914	3805	3785.108	3793.502
2019/7/24	3819.833	3768	3797.823	3785.238
2019/7/25	3851.067	3809	3823.114	3817.158

Table 3. Prediction errors of three models

Date	ARIMA absolute error	BP absolute error	BP-ARIMA absolute error	ARIMA relative error	BP relative error	BP-ARIMA relative error
2019/7/12	7.7311	16.45538	12.77373	0.00203	0.00432	0.003354
2019/7/15	31.1878	10.61303	19.29558	0.008155	0.002775	0.005046
2019/7/16	19.8449	19.1874	2.715769	0.005213	0.00504	0.000713
2019/7/17	19.6384	5.69591	4.995169	0.005162	0.001497	0.001313
2019/7/18	40.5981	42.26504	41.56159	0.010773	0.011216	0.011029
2019/7/19	17.0449	31.9246	11.25947	0.004476	0.008384	0.002957
2019/7/22	24.3168	32.87873	29.2656	0.00643	0.008694	0.007739
2019/7/23	15.0865	4.80542	3.58897	0.003981	0.001268	0.000947
2019/7/24	51.8325	22.00937	34.59473	0.013569	0.005762	0.009057
2019/7/25	42.0665	27.95276	33.90876	0.010923	0.007258	0.008805

3.5. Comparison of Three Models

In order to compare the three models better, the average absolute error and average relative error of each model are used to evaluate the advantages and disadvantages of the three models. According to table 3, the ARIMA model has a maximum absolute error of 51.8325, a minimum absolute error of 7.7311, a maximum relative error of 0.013569 and a minimum relative error of 0.00203. The maximum absolute error of BP model is 42.26504, the minimum absolute error is 4.80542, the maximum relative error is 0.011216, and the minimum relative error is 0.001268. The maximum absolute error of BP model is 41.56159, the minimum absolute error is 3.58897, the maximum relative error is 0.011029, and the minimum relative error is 0.000947. It can be seen that when the ARIMA model and BP neural network forecast separately, the range of error between forecast value and actual value is not very big, but the range of error is obviously reduced after the combination of the model. It can be seen from table 4 that the average absolute error and the average relative error of the combined model are the smallest, the BP neural network model is the second, the average absolute error and the average relative error of the ARIMA model are the largest. it is concluded that the prediction accuracy of the combined model is higher than that of the single model.

Table 4. Prediction errors of three models

	ARIMA	BP	BP-ARIMA
MAE	26.93475	21.37876	19.39593
MRE	0.007071	0.005621	0.005095

4. CONCLUSION

The prediction of CSI 300 Index is still a concern for many investors when making investment decisions. Firstly, the ARIMA model and BP neural network are established to forecast the time series of CSI 300 Index. The results obtained by R language show that both models can predict the future CSI 300 Index. The results show that both models are feasible in predicting the financial time series. However, there are many shortcomings in the single model. This paper uses the entropy method to assign different weights to the two models. The forecasting results of the three models are compared, and the BP-ARIMA combined model has the highest forecasting precision, minimum margin of error. This has provided the certain help for the investor when carrying on the investment decision.

REFERENCES

[1] X. Chen (2017). Prediction of stock-price Based on ARIMA Model and Neural network model. Journal of Quantitative Economics, vol.34, no.4, p.30-34.

[2] A. Ma, J. Xie and W. Tang (2018). A comparison of ARIMA Model, BP Neural Network Model and combined model in health policy evaluation. Chinese Journal of Health Policy, vol.11, no.1, p.76-83.

[3] Y. Cheng, Y. Cui (2011). Study on modeling and simulation of software reliability prediction based on combination model. Computer Simulation, vol.28, no.6, p.371-374.

[4] L. Liang. W. Jia and H. Wang (2009). Studies on pest population dynamics based on combined forecasting model. Journal of Yangtze University (Nat Sci Edit), vol.6, no.1, p.5-10.