

Stock Price Prediction Model Analysis Based on Neural Network

Mingzhe Yin*

Department of Statistics, University of Minnesota, Twin City, the United States

*Corresponding author: yin00378@umn.edu

Abstract. As a matter of fact, stock price prediction is always crucial for investors and scholars to gain extra return from the stock market. Thanks to the rapid development of the machine learning techniques as well as computation ability (e.g., GPU), various deep leaning models based on the concepts of neural networks are proposed and applied in stock price prediction. On this basis, this paper will investigate the feasibility of stock price prediction based on various neural network deep learning scenarios. With this in mind, the BP neural network, principal component-based BP neural network (PCA-BP neural network) as well as GA-BP neural network are implemented to realize stock price prediction. In order to compare the effectiveness, the ARIMA model is also adopted as a benchmark. According to the analysis, neural network is suitable for stock price prediction. Overall, these results shed light on guiding further exploration of stock price prediction based on the state-of-art machine learning approaches.

Keywords: Stock price prediction; neural network; ARIMA.

1. Introduction

Having a stable stock market is significant for the development of a country's financial market [1-3]. Considering the variability of the stock market, some research on stock price prediction has attracted the attention of many scholars [4]. A large number of traditional data analysis methods have already produced research results, such as the widely used ARMA model, ARIMA model, and GARCH model. Machine learning and artificial intelligence have given us new ideas for solving such nonlinear problems [5]. Therefore, more and more scholars are introducing machine learning methods into stock price prediction [6]. For example: support vector machine regression (SVR), stock prediction method based on wavelet denoising back propagation (WDBP) neural network, gray system method, decision tree method, etc. [7].

Stock prices are a multi-independent, highly unstable, complex, and unpredictable time series data. Using the classic linear time series model to predict stock prices may not be able to extract the nonlinear part of the information, leading to problems such as large prediction errors and low prediction accuracy. The emergence of neural networks has solved this problem. On the basis of considering the noise of stock data, the GA-BP neural network model selected in this study adds global search capabilities obtained through genetic algorithms. A framework is shown in Fig. 1.

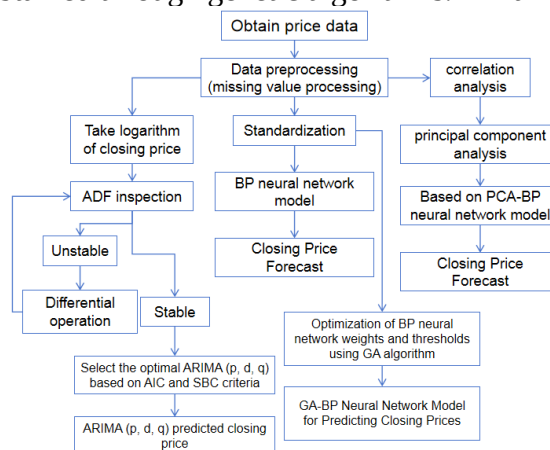


Fig. 1 The research flowchart.

2. Theoretical Basis

The training process of automatic feedback is the essence of the BP neural network model. The model compares the training results with the ideal results to perform error analysis [8-10]. This process means constant adjustments and repairs of weights and thresholds until set standards are reached. The relationship between the output and input of neuron h is expressed as:

$$net_h = \sum_{i=1}^n w_{ih} x_i \quad (1)$$

$$y_h = f(net_h) \quad (2)$$

where X is the threshold of each neuron in the hidden layer, w_{hi} is the weight between the hidden layer and the output layer, w_{ih} is the weight between the input layer and the hidden 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, Y is the output result and f is the activation function. Fig. 2 shows the structure of network.

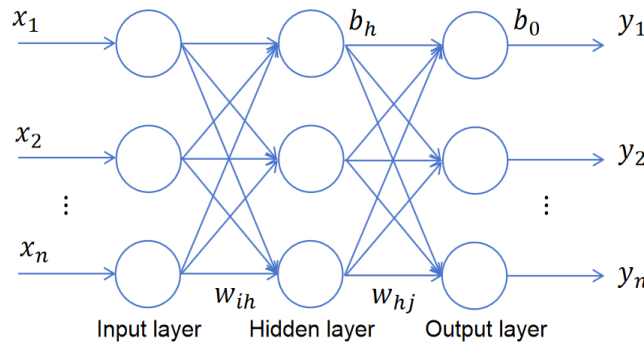


Fig. 2 The Structural Model of BP Neural Network

This analysis uses dimensionality reduction to change several correlated indicators into several unrelated comprehensive indicators without losing a small amount of information. Specific steps are as follows:

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, i = 1, 2, \dots, n; j = 1, 2, \dots, p \quad (3)$$

Here, $\bar{x}_j$ is the mean of the j -th variable, σ_j is the standard deviation of the j -th variable and Z_j is the standardized data matrix. One can calculate the standardized correlation coefficient matrix and its corresponding eigenvalues and eigenvectors:

$$R = [r_{ij}], |\lambda I_p - R| = 0, i = 1, 2, \dots, n; j = 1, 2, \dots, p \quad (4)$$

A new type of stochastic search algorithm which is called Genetic algorithm is based on global perspective. This method can generate new individuals through crossover and mutation [8]. The fitness function F needs to be determined based on actual problems, and genetic operations include three processes: selection, crossover, and mutation. There are many methods of selection operations in the GA algorithm. This article chooses the most commonly used proportional selection method, that is, the probability P_i of being selected for each individual genetic operations in the population is

$$P_i = \frac{F_i}{\sum_{i=1}^n F_i} \quad (5)$$

3. Results and Discussion

3.1. Sample data selection and preprocessing

This article selects the Shanghai Composite Index stock for analysis and prediction. The selected data includes seven variables: the date (X0), the closing price (X1), the highest price (X2), the lowest price (X3), the opening price (X4), the previous closing price (X5) and the change amount (X6). This article uses date as the index of the time series, the remaining variables are input into the model as independent variables, and the closing price is output as the only indicator. In data set, there are some

missing values selected, in this article, so the missing values need to be processed. For missing values in stock data, the following methods can be used to deal with them: directly delete the missing values and fill them with average values, regression values or normal values from the previous moment [9]. This article uses the normal value at the previous moment to fill in the missing values. One can solve the correlation coefficient between the opening price, the closing price, the previous closing price, the highest price, the lowest price, and the increase and decrease amount, and obtain the correlation coefficient Table 1. Conduct principal component analysis on the six variables, and then one gets the total variance explanation shown in Table 2.

Table 1. Correlation coefficient of Shanghai Composite Index

	X1	X2	X3	X4	X5	X5
X1	1	0.999	0.999	0.999	0.999	0.025
X2	0.999	1	0.999	1	0.999	0
X3	0.999	0.999	1	0.999	0.999	0.006
X4	0.999	1	0.999	1	1	-0.016
X5	0.999	0.999	0.999	1	1	-0.026
X6	0.025	0	0.006	-0.016	-0.026	1

Table 2. Explanation table of total variance

component	Initial eigenvalue			Extract the sum of squares of the load		
	total	Variance percentage	Accumulate%	total	Variance percentage	Accumulate%
1	4.997	83.288	83.288	4.997	83.288	83.288
2	1.002	16.693	99.981	1.002	16.693	99.981
3	0.0001	0.013	99.994			
4	0	0.005	99.998			
5	0	0.002	100			

Two principal components are extracted. The expression of the principal component is:

$$F_1 = 0.44734782X_1^* + 0.44734782X_2^* + 0.44734782X_3^* + 0.44734782X_4^* + 0.44734782X_5^* - 0.001342043X_6^* \quad (6)$$

$$F_2 = 0.027972042X_1^* + 0.001998003X_2^* + 0.008991013X_3^* - 0.012987019X_4^* + 0.022977034X_5^* - 0.9990001498X_6^* \quad (7)$$

3.2. Selection of evaluation indicators

The two methods of model evaluation, mean absolute error (MAE) and root mean square error (RMSE) are shown in this paper, and these two methods can make the model's prediction results comparable. The formulae are as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n}} \quad (8)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y} - y_i| \quad (9)$$

Among them, $\hat{y}$ is the predicted value, while y_i is the true value. Normally, the smaller the RMSE and MAE, the better the prediction effect of the model is.

3.3. Prediction Results and Analysis

In this training, this study selects the first 80% yes of data input as input data and the last 20% of data as test set. In addition, learning rate was set for 0.01, choose the activation function relu, and use the 1mean error function to define the loss in training. Of course, what is more important is the setting of the number of nodes in the input layer, hidden layer, and output layer. In the study of this paper, the number of nodes is chosen to be set to 1 to better predict the closing price.

$$l = \sqrt{m + n} + a \quad (10)$$

where n respect for the number of input layer, l respect for the number of hidden layer, m respect for the number of output layer nodes, a is always a constant, $a \in [1,10]$. To predict the Shanghai Composite Index (000001) with BP neural network model, first standardize the data. Here, maximum and minimum standardization is used, that is,

$$y = \frac{y - \min(y)}{\max(y) - \min(y)} \quad (11)$$

In this step, one inputs standardized data into the neural network model for training. Specifically, one set a neural network model with the number of nodes in the input layer as 5 and the number of nodes in the output layer as 1. Furthermore, one can speculate that the number of nodes in the hidden layer is a value between 3 and 13. Use MATLAB to test and train the possible range values of the hidden layer one by one, and get the following table. It is not difficult to conclude that when the RMSE and MAE of the model can reach the minimum value, the hidden layer node of the model is 12, which means that the prediction result is also the best.

Table 3. Corresponding errors of different hidden layers in PCA-BP

Different hidden layers	RMSE	MAE
3	1.432589158	1.248222241
4	1.069121617	0.886285816
5	1.208331946	0.864541787
6	1.528452294	1.049713746
7	1.24999793	0.955760418
8	1.245776588	0.86375491
9	1.829450188	1.367403261
10	1.931526711	1.487702557
11	1.876876527	1.575242945
12	1.156804283	0.855665777
13	1.681640046	1.257793239

By performing correlation analysis, the selected variables show obvious correlation. It is essential to use principal component analysis to downscale. Specifically, one extracts the principal components $F1$ and $F2$ as the input variables of the BP neural network, while the output variables are unchanged. In this case, the output layer node of the PCA-BP neural network is 2, which makes the model simple. Additionally, the hidden layer nodes range between 3 and 12. Next, training for different hidden layer nodes, one can get Table 4. When the number of hidden layer nodes is 10, PCA-BP corresponds to the smallest RMSE and MAE.

Table 4. Corresponding errors of different hidden layers in PCA-BP

Different hidden layers	RMSE	MAE
3	6.798554712	4.683312039
4	6.531464275	4.388189916
5	6.504174152	4.335095252
6	6.4970245	4.389630978
7	6.49122071	4.376867032
8	6.503065929	4.333717612
9	6.529578981	4.385499843
10	6.436457662	4.344930627
11	6.524074218	4.334471849
12	6.685452952	4.598161015

In BP neural network, weights W and threshold B are optimized by genetic algorithm. Specifically, the chromosomes are encoded using real number coding. The value of the overall scale is 6, the number of iterations one chooses 20, the crossover probability is chosen as 0.3, while the variance probability is 0.1. By allocating optimized weights and thresholds in the BP neural network model,

one can further predict a more accurate AAE index. At the same time, considering that the number of nodes in the input layer and the number of nodes in the output layer have not changed, it is still possible that the hidden layer of juice will have a number of nodes between 3 and 13. During the test, one found that the number of nodes in the hidden layer is 10, GA- BP corresponds to the smallest RMSE and MAE and the best prediction (seen from Table 5).

Table 5. Corresponding errors of different hidden layers in GA-BP

Different hidden layers	RMSE	MAE
3	0.343422203	0.271795054
4	0.714383575	0.633265874
5	0.608644212	0.450519719
6	0.744475806	0.421791415
7	0.754758413	0.633265874
8	0.636707719	0.540481305
9	0.640333842	0.542508164
10	0.319528582	0.215858201
11	0.606491145	0.365938561
12	0.415811836	0.341795731
13	0.406515803	0.279244819

First, one logarithmizes the original closing price, and then perform ADF stationarity test on the logarithmic data. It can be found that the absolute value of the ADF statistic after logarithmization of the original data is 0.766904, and the p value is $0.8793 > 0.05$. It shows that the original data is a non-stationary series after logarithmization. The results are shown in Table. 6.

Table 6. ADF test table after logarithmization of original data

	t-Statistic	Prob.*
ADF	0.766904	0.8793
	1% -2.5654	
Test critical values:	5% -1.94088	
	10% -1.61666	

Since the logarithmic sequence is a non-stationary sequence, it is necessary to perform a first-order difference operation on it, and then perform an ADF test on the first-order difference sequence. The test results are shown in the Table 7. The sequence after the first difference passes the ADF test, that is, the sequence after the difference is stationary. The absolute value of the ADF statistic after the first difference is 29.44964, and the p-value is 0.0000, which is far less than 0.05.

Table 7. First order differential sequence ADF test table

	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-29.44964	0.0000
	1% -3.431415	
Critical values:	5% -2.861896	
	10% -2.567002	

One calculates the variance of the sample autocorrelation coefficient according to Barlett's formula, and show that when the sample size is large enough, the sample autocorrelation coefficient approximately obeys the normal distribution $\hat{\rho}_k \sim N(0, \frac{1}{n})$. At the same time, Quenouille also proved that the sample partial autocorrelation coefficient also approximately obeys the normal distribution. Therefore, based on the properties of the normal distribution, $-\frac{2}{\sqrt{n}} \leq \hat{\rho}_k \leq \frac{2}{\sqrt{n}}, -\frac{2}{\sqrt{n}} \leq \hat{\phi}_{kk} \leq \frac{2}{\sqrt{n}}$ can be used to determine the order of the model. According to the comparison of the calculation results, it is found that except for the 4th and 6th orders of the sample autocorrelation coefficient and partial autocorrelation coefficient, which are within the range of two standard deviations, the rest are within

the range of two standard deviations. Therefore, the possible values of p and q are 4, 6, that is, four models of ARIMA(4,1,4), ARIMA(6,1,4), ARIMA(4,1,6), and ARIMA(6,1,6) can be established as given in Table 8. The optimal model is selected based on the AIC criterion and SBC criterion. From Table 2-11, one can see that the AIC and SBC values corresponding to ARIMA(6,1,6) are smaller than the AIC and SBC values of the other three models, so the better ARIMA(6,1,6) model is selected for the stock To predict the closing price, the model expression is given in Table 9.

Table 8. Model Testing

Model	ARIMA(4, 1, 4)		ARIMA(4, 1, 6)		ARIMA(6, 1, 4)		ARIMA(6, 1, 6)	
Delay order	6	12	6	12	6	12	6	12
P	0.098	0.058	0.89	0.34	0.359	0.163	0.409	0.172

Table 9. Optimal Model Selection

Model	ARIMA(4, 1, 4)	ARIMA(4, 1, 6)	ARIMA(6, 1, 4)	ARIMA(6, 1, 6)
AIC	10.42991	10.42986	10.42738	10.4253
SBC	10.43992	10.44111	10.42864	10.43781

4. Conclusion

To sum up, the goal of this article is to predict the closing price of the Shanghai Composite Index. This study discussed models including BP neural network, principal component-based BP neural network (PCA-BP neural network), GA-BP neural network and classic time series ARIMA (6,1,6), and tried to compare the implementation of these models Advantages and Disadvantages of the Forecasting Process. Specifically, a good judgment can be obtained by comparing the root mean square error and mean absolute error presented by different models. The GA-BP neural network model currently has the best performance in predicting the closing price of the Shanghai Composite Index, with the smallest prediction error.

References

- [1] Lu W, Li J, Wang J, Qin L. A CNN-BiLSTM-AM method for stock price prediction. *Neural Computing and Applications*, 2021, 33: 4741-4753.
- [2] Hu Z, Zhao Y, Khushi M. A survey of forex and stock price prediction using deep learning. *Applied System Innovation*, 2021, 4(1): 9.
- [3] Schöneburg E. Stock price prediction using neural networks: A project report. *Neurocomputing* 1990, 2(1): 17-27.
- [4] Xiao C, Xia W, Jiang J. Stock price forecast based on combined model of ARI-MA-LS-SVM. *Neural Computing and Applications*, 2020, 32: 5379-5388.
- [5] Chou J S, Nguyen T K. Forward forecast of stock price using sliding-window metaheuristic-optimized machine-learning regression. *IEEE Transactions on Industrial Informatics*, 2018, 14(7): 3132-3142.
- [6] Sun A, Zhao T, Chen J, Chang J. Comparative Study: Common ANN and LS-SVM Exchange Rate Performance Prediction. *Chinese Journal of Electronics*, 2018, 27(3): 561-564.
- [7] Dinh T A, Kwon Y K. An empirical study on importance of modeling parameters and trading volume-based features in daily stock trading using neural networks. *Informatics*, 2018, 5(3): 36.
- [8] Chang P C. A novel model by evolving partially connected neural network for stock price trend forecasting. *Expert Systems with Applications*, 2012, 39(1): 611-620.
- [9] Dash R K, Nguyen T N, Cengiz K, Sharma A. Fine-tuned support vector regression model for stock predictions. *Neural Computing and Applications*, 2021: 1-15.
- [10] Kumar D, Meghwani S S, Thakur M. Proximal support vector machine based hybrid prediction models for trend forecasting in financial markets. *Journal of Computational Science*, 2016, 17: I 1-13.