Journal of Computer and Communications, 2018, 6, 22-33
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Gold Price Prediction Based on PCA-GA-BP Neural Network

Youchan Zhu, Chaokun Zhang

School of Control and Computer Engineering, North China Electric Power University, Baoding, China
Email: 1228086656@qq.com

How to cite this paper: Zhu, Y.C. and Zhang, C.K. (2018) Gold Price Prediction Based on PCA-GA-BP Neural Network. *Journal of Computer and Communications*, **6**, 22-33.
https://doi.org/10.4236/jcc.2018.67003

Received: June 30, 2018
Accepted: July 23, 2018
Published: July 26, 2018

Abstract

Gold price is affected by a variety of factors and has highly nonlinear and random features. Some traditional forecast methods emphasize linear relations excessively and some ignore the price randomness. The predictive error is relatively large. Therefore, a BP neural network model based on principal component analysis (PCA) and genetic algorithm (GA) was proposed for the short-term prediction of gold price. BP could establish the gold price forecasting model. The weights and thresholds of BP neural network are optimized by GA, which overcome the shortcoming that BP algorithm falls into local minimum easily. PCA can effectively simplify the network input variables and speed up the convergence. The results showed that, compared with GA-BP and BP, the convergence rate of PCA-GA-BP neural network model was faster and the prediction accuracy was higher in the prediction of gold price.

Keywords

PCA, Genetic Algorithm, BP Neural Network, Gold Price

1. Introduction

For a long time, gold is a symbol of wealth and has been widely used in currency, jewelry and other industries. The current gold market has the characteristics of high-yield and high-risk coexistence. On the basis of practice, people gradually formed some gold price theories. However, a lot of research shows that due to various factors such as economic, political, human, and market factors, the price of gold has a high degree of randomness and nonlinearity [1] [2] [3] [4].

ARIMA [5] [6] usually only predicts data with a linear relationship. The grey prediction [7] method regards irregular data change as interference during forecasting. These Irregular data will be removed during the forecasting process.

This determines that the grey prediction has a strong inertia without considering the randomness of the system. It is also not sensitive about the fluctuation trends and is more suitable for predictions data with certain characteristic features. In addition, it also includes FAR model [8], GARCH model [9] and so on. But none of these methods are suitable for the prediction of gold prices.

The opening price, closing price, highest price, lowest price, change amount, change rate, trading volume and turnover of gold prices are the combined results of various macro factors and micro factors. A large number of potential laws and factor indicators determine the changes about the gold price. Through the study of these data, we can grasp the trend of gold prices to a certain extent. Therefore, using historical market data to study gold prices have certain significance. After PCA selects principal components with some rules, the original multidimensional data can be simplified, the relevance of network input data can be eliminated, redundant information can be eliminated, the input data of network can be reduced, and the main information of the original data can be retained. However, PCA cannot obtain the non-linear relationship about data. BP neural network is a feed forward neural network [10]. The BP neural network could be used as a good model for the gold price prediction due to its simple structure and easy operation, especially the ability of self-learning to realize any complex nonlinear mapping. When using BP network, it is not necessary to establish a specific mathematical model, and could find the optimal solution with iterativing the input and output data is directly. However, the BP network has the disadvantages of slow convergence rate, easily falling into local minimum, and oscillation near-optimal solution. GA can globally optimize the weights and thresholds of the neural network, obtain the approximate solution of the optimal solution. Then the BP neural network can obtain the optimal solution, achieve the goal of global optimization. GA can solve the problem of BP neural network. So, a PCA-GA-BP neural network model combining PCA, GA and BP is proposed to realize short-term prediction of gold prices.

2. Related Principles

2.1. Principal Component Analysis (PCA)

PCA is a concept in statistics. Through the analysis of original data to obtain the cumulative contribution rate, and then get the main component. That the reconstructed data retain the primary information of the original data, thus achieving the goal of reducing the correlation between the original data and reducing the data dimension [11]. The original data are represented by the matrix $X(n*p)$. The specific calculation steps are as follows:

1) Data standardization: Y_{ij} is the standardized variable.

$$Y_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j}, \quad (i = 1,2,\ldots,n; j = 1,2,\ldots,p) \tag{1}$$

where, $\bar{X}_j$ is the average of the jth variable, S_j is the standard deviation of the *j*th variable. The definitions of $\bar{X}_j$ and S_j are as follows:

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}\ ,\quad S_j = \sqrt{\frac{1}{n\text{-}1}\sum_{i=1}^{n}\left(X_{ij} - \bar{X}_j\right)^2}$$

2) *R* is the correlation coefficient matrix, as in (2)

$$R = Y^T\,Y/(n-1) \tag{2}$$

3) Calculating eigenvector matrix *A* and the eigenmatrix λ

$$RA = A\lambda \tag{3}$$

4) Determine the principal component, calculate the principal component contribution rate and cumulative contribution rate. Select the first m principal components with the cumulative contribution rate is not less than 85%. The component contribution rate of *k*th principal as in (4):

$$\lambda_k \Big/ \sum_{j=1}^{p} \lambda_j\ ,\ (k = 1, 2, \ldots, p) \tag{4}$$

where, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p)$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$

The cumulative contribution rate of the first k principal components as in (5)

$$\sum_{j=1}^{k} \lambda_j \Big/ \sum_{j=1}^{p} \lambda_j\ ,\ (k = 1, 2, \ldots, p) \tag{5}$$

5) $\alpha_1, \alpha_2, \ldots, \alpha_m$ are the feature values corresponding to feature vectors $\lambda_1, \lambda_2, \ldots, \lambda_m$ respectively. The sample data calculated by the principal component:

$$Z = \sum_{i=1}^{m} \alpha_i Y_i \tag{6}$$

2.2. BP Neural Network

The BP neural network uses a gradient descent method to change the sample I/O problem into a nonlinear optimization problem [12]. BP is a typical supervised learning algorithm [13]. Through learning the neural networks weights and thresholds repeatedly to obtain the output error minimum value. The specific process is divided into two steps:

Forward propagation: The data pass the input layer, hidden layer, and output layer. The actual output and expected output are compared at the output layer. If the actual output error is not reached the expected output error, the network enter the back propagation.

Back propagation: The error signal transmits from the output layer, then passes the hidden layer and finally reaches the input layer. During this process, every neuron's weight in each hidden layer is corrected according to the negative gradient direction of the error function, and the error signal is continuously reduced to make the actual output near the desired output.

After training, neural network grasps the relationship between input variables and output variables. Finally the output could be predicted according to the input variables base on the trained model [14] [15].

2.3. GA Optimize BP

GA simulates the evolutionary principle in the biological field [16], which can search data in parallel and randomly, regarding the problem as the biological evolution process. GA selects individuals and generates new individuals repeatedly with selection-crossover-mutation operation until the condition is satisfied [17]. The specific steps of GA to optimize the BP are as follows:

1) Encoding and initializing population: Using the floating point number encoding, each individual contains all the weights and threshold. R is the nodes number in the input layer. S_1 is the nodes number in the hidden layer. S_2 is the nodes number of the output layer. S is the individual length.

$$S = R * S_1 + S_1 * S_2 + S_1 + S_2 \tag{7}$$

The size of the population has a great influence on the global search performance of the genetic algorithm. Therefore, the size of the population must be selected according to the specific problem. Initial population size is 50.

2) Evaluation function: Inputting the training sample and calculating its error function value. Regarding the reciprocal of the error function value as the fitness. If the error is smaller, the fitness is greater, as in (8). E is the sum of squared errors between the predicted output and the expected output. δ is a positive minimum amount.

$$f = 1/(E+\delta) \tag{8}$$

3) Use roulette algorithm as the selection operator. P_i is the probability that the individual i is selected. f_i is the fitness value of individual i. n is the population size.

$$P_i = f_i \Big/ \sum_{i=1}^{n} f_i \tag{9}$$

4) Crossover: The primary search mean of GA is crossover operation. For the non-optimal individuals, let two individuals cross to produce two new individuals with a crossover probability of P_c (P_c = 0.7) and the optimal individuals can reach the next generation directly. The maximum number of iterations is 100.

5) Non-optimal individuals mutate with a probability of P_m (P_m = 0.1) to produce new individuals and optimal individuals can reach the next generation directly. The maximum number of iterations is 100.

6) Repeat steps 2) - 5) until reach the iteration goal or the number of iterations.

The PCA-processed data can be used as the input data of the GA-optimized BP network to achieve rapid convergence. At the same time, because PCA greatly reduces the complexity of input data and the complexity of neural network training, it can improve the accuracy of prediction. GA can overcome the shortcomings of the slow convergence rate of BP network and falling into local minimum easily. Combined with the three algorithms, they can exert their respective advantages, reducing input, accelerating the convergence speed, and searching for global optimal values. At the same time, they have good nonlinear modeling capabilities and overall improve the performance of the network.

3. PCA Standardized Raw Data

Selecting 110 valid data from Shanghai Gold Exchange as the raw data and the period is from 2016/5/23 to 2016/11/3. The opening price, closing price, highest price, lowest price, change amount, change rate, trading volume and turnover as eight input variables.

1) The raw data is shown in Table 1.

2) Standardized the raw data. Normalize each raw data with (1) to eliminate the magnitude and dimension difference between variables. The standardized data is shown in Table 2.

3) Calculating the correlation coefficient matrix R with (2).

$$R = \begin{Bmatrix} 1.0000 & 0.9591 & -0.1626 & -0.1659 & 0.9800 & 0.9793 & -0.2077 & -0.1160 \\ 0.9591 & 1.0000 & 0.0917 & 0.0885 & 0.9798 & 0.9820 & -0.2093 & -0.1160 \\ -0.1626 & 0.0917 & 1.0000 & 0.9993 & -0.0509 & -0.0196 & -0.0619 & -0.0612 \\ -0.1659 & 0.0885 & 0.9993 & 1.0000 & -0.0557 & -0.0209 & -0.0576 & -0.0572 \\ 0.9800 & 0.9798 & -0.0509 & -0.0557 & 1.0000 & 0.9678 & -0.2165 & -0.1250 \\ 0.9793 & 0.9820 & -0.0196 & -0.0209 & 0.9678 & 1.0000 & -0.1881 & -0.0955 \\ -0.2077 & -0.2093 & -0.0619 & -0.0576 & -0.2165 & -0.1881 & 1.0000 & 0.9942 \\ -0.1160 & -0.1160 & -0.0612 & -0.0572 & -0.1250 & -0.0955 & 0.9942 & 1.0000 \end{Bmatrix}$$

Matrix R shows that the correlation coefficient between the first variable and the second variable is 0.9591, and the correlation coefficient between the second variable and the fifth variable is 0.9798. This shows that the correlation between these data is strong and the correlation needs to be reduced.

4) Calculate eigenvalues, contribution rates and cumulative variance contribution rates with (3) (4) (5). The result is shown in Table 3.

The cumulative contribution rate of the first three principal components is 99.448%. According to the rule, the original eight variables are replaced by the first three principal components. They are I_1, I_2 and I_3 respectively. The principal component load matrix is shown in Table 4.

According to Table 4, main factor expressions can be got.

Table 1. The raw data.

Number	Opening price	Closing price	Change amount	Change rate	Lowest price	Highest price	Trading volume	Turnover
1	265.88	264.11	−0.4	−0.0015	263.6	265.88	27,147.88	7,150,839,869.60
2	263.01	262.45	−1.66	−0.0063	262.3	264.1	36,155.46	9,549,629,692.40
3	262.45	258.45	−4	−0.0152	258.4	262.5	27,182.46	7,016,190,365.60
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
109	281.99	282.57	2.46	0.0088	280.55	282.85	21,985.48	6,167,895,660.20
110	283.49	283.73	1.16	0.0041	282.85	284.78	15,124.60	4,287,224,778.60

Table 2. Standardized data.

Number	Opening price	Closing price	Change amount	Change rate	Lowest price	Highest price	Trading volume	Turnover
1	−1.4915	−1.7100	−0.2197	−0.2316	−1.5903	−1.6724	0.4715	0.3189
2	−1.7798	−1.8828	−0.7014	−0.7406	−1.7251	−1.8558	1.6494	1.4820
3	−1.8360	−2.2990	−1.5960	−1.6843	−2.1294	−2.0207	0.4760	0.2536
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
109	0.1263	0.2109	0.8737	0.8607	0.1671	0.0761	−0.2036	−0.1577
110	0.2770	0.3316	0.3767	0.3623	0.4056	0.2750	−1.1008	−1.0695

Table 3. Characteristic value and contribution rate.

Number	Eigenvalues	Contribution rate/(%)	Cumulative rate/(%)
1	4.0310	50.388	50.388
2	2.1150	26.437	76.825
3	1.8098	22.623	99.448
4	0.0317	0.396	99.844
5	0.0079	0.099	99.943
6	0.0029	0.036	99.979
7	0.0012	0.015	99.994
8	0.0005	0.006	100

Table 4. Principal component load matrix.

Variables	I_1	I_2	I_3
Opening price	−0.4891	0.1128	0.0450
Closing price	−0.4860	−0.0333	0.1511
Change amount	0.0259	−0.5956	0.3690
Change rate	0.0278	−0.5944	0.3710
Lowest price	−0.4894	0.0461	0.0861
Highest price	−0.4869	0.0375	0.1190
Trading volume	0.1734	0.3670	0.5725
Turnover	0.1287	0.3738	0.5933

$$\begin{aligned} I_1 = {} & -0.4891X_1 - 0.4860X_2 + 0.0259X_3 + 0.0278X_4 \\ & -0.4894X_5 - 0.4869X_6 + 0.1734X_7 + 0.1287X_8 \end{aligned}$$

$$\begin{aligned} I_2 = {} & 0.1128X_1 - 0.0333X_2 - 0.5956X_3 - 0.5944X_4 \\ & +0.0461X_5 + 0.0375X_6 + 0.3670X_7 + 0.3738X_8 \end{aligned}$$

$$\begin{aligned} I_3 = {} & 0.045X_1 + 0.1511X_2 + 0.369X_3 + 0.371X_4 \\ & +0.0861X_5 + 0.119X_6 + 0.5725X_7 + 0.5933X_8 \end{aligned}$$

5) The sample data after principal component analysis is shown in Table 5.

Table5. The sample data after principal component analysis.

Number	I_1	I_2	I_3
1	3.2637	0.3135	−0.3692
2	3.9712	1.7302	0.5561
3	4.0682	1.9169	−1.6446
⋮	⋮	⋮	⋮
109	−0.2921	−1.1479	0.4926
110	−0.9376	−1.1944	−0.8611

4. Experiment

4.1. Model Training

The three layer neural network structure is used. The output layer transfer function is purelin and other parameters are default. Training target is 0.001, maximum training times is 10,000, and learning rate is 0.01.

Selecting the first 100 sample as the training sample and the last 10 sample as the test sample. Taking the principal components of the first three days as input data and the output is the closing price of the 4th day. The input layer nodes number is 9, and the output layer nodes number is 1. The hidden layer nodes number is usually determined based on the empirical formula to obtain a rough range and then try to determine the nodes optimal number [18]. Empirical formula is shown by (10)

$$m = \sqrt{n+l} + a \tag{10}$$

m is the hidden layer nodes number, n is the input layer nodes number, l is the output layer nodes number, a is a constant between 1 and 10. According to the empirical formula, the hidden layer nodes numbers are between 5 and 14. For the different hidden layer nodes number, the maximum number of training steps is 2000. Basing on same sample, the trial and error method is used to train the network firstly with less hidden layer nodes. Then gradually increases the nodes number of the hidden layer. For each node which is continuously trained 70 times, then select the node that has minimum output error and the corresponding number of steps. Finally, when the nodes number of the hidden layer is 13, the output error has minimum value in the PCA-GA-BP model. Detailed results are shown in Table 6. After setting up the structure, the trained error curve of the PCA-GA-BP is shown in Figure 1.

Two neural network models, GA-BP and BP, are established to compare with the PCA-GA-BP. The GA-BP training error curve is shown in Figure 2, and the BP training error curve is shown in Figure 3.

When the training target is 0.001, Figure 1 shows that the PCA-GA-BP output error is 0.00072166 and the training steps number is 71. Figure 2 shows that the GA-BP output error is 0.00089814 and the training steps number is 889. Figure 3 shows the PCA output error is 0.00097956 and the training steps number is 3880. By comparison, the GA-BP convergence speed is faster than BP and the convergence speed of PCA-GA-BP convergence speed is faster than GA-BP.

Table 6. Results of different nodes number.

Hidden layer nodes number	Mean square error	steps
5	0.0042369	2000
6	0.0036774	2000
7	0.0029725	2000
8	0.0012786	1863
9	0.00099982	918
10	0.00097621	707
11	0.00096848	359
12	0.00096949	206
13	0.00092836	194
14	0.00097189	285

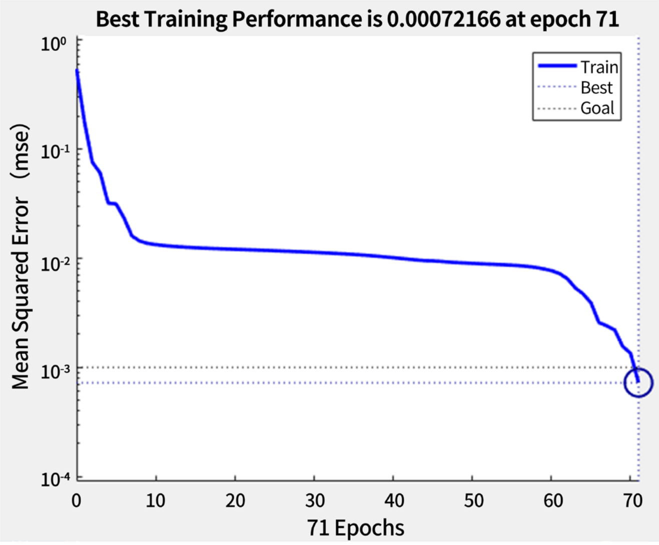

Figure 1. PCA-GA-BP model training error.

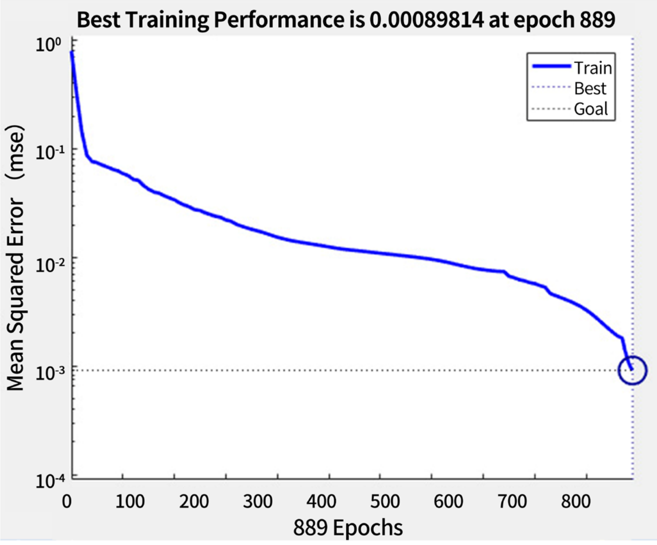

Figure 2. The GA-BP training error.

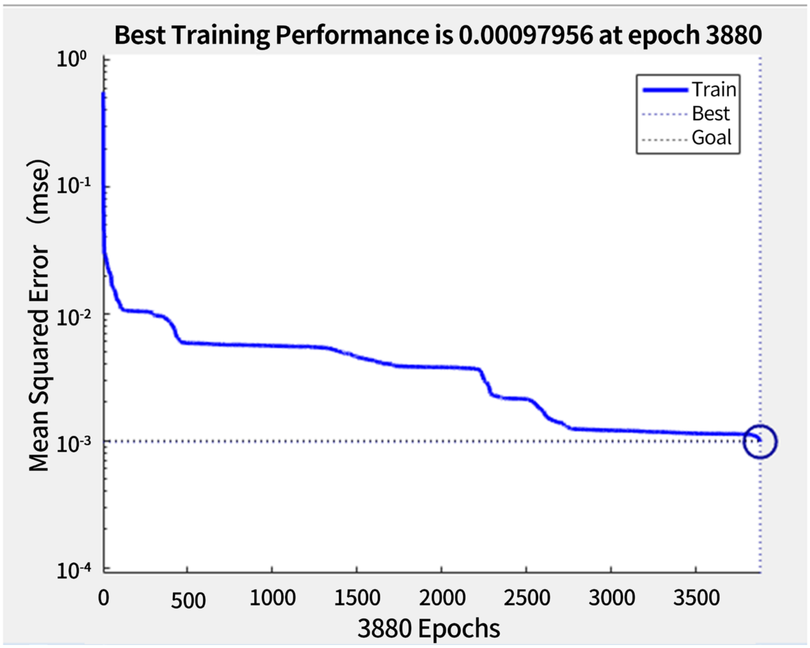

Figure 3. BP model training error.

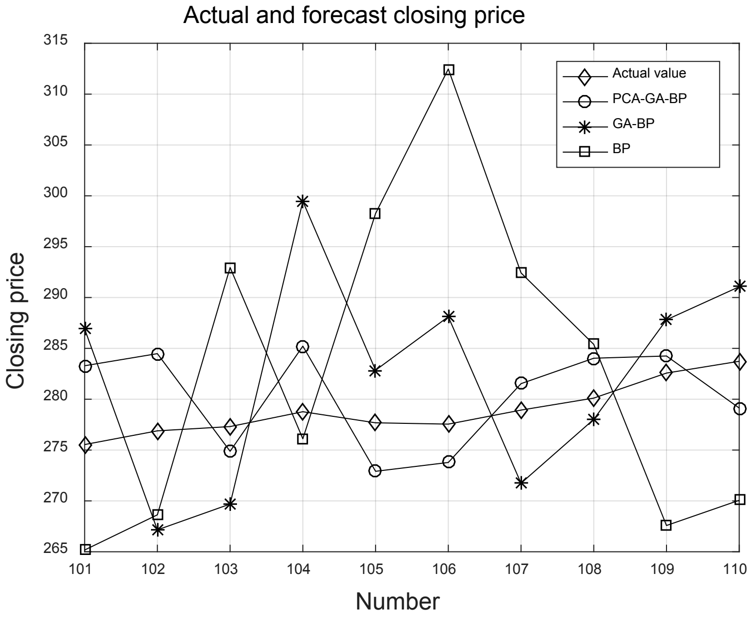

Figure 4. Actual and predictive closing price.

4.2. Model Test

Input the last 10 days data to the three neural network models respectively that have been trained. Their closing price prediction curve is shown by Figure 4 and the relative error curve is shown by Figure 5. The detailed results of the predictive closing price and the predictive error are shown in Table 7.

Table 7 can be used to calculate average relative error about the three models and the specific results are shown in Table 8.

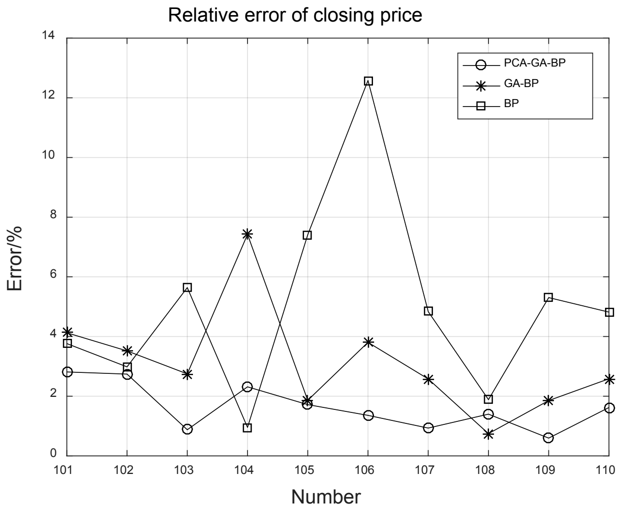

Figure 5. Relative error of closing price.

Table 7. Predictive value and predictive error of closing price.

		PCA-GA-BP		GA-BP		BP	
Number	Actual value	Predictive value	Relative error	Predictive value	Relative error	Predictive value	Relative error
101	275.54	283.29	2.813%	286.91	4.126%	265.17	3.764%
102	276.89	284.48	2.741%	267.15	3.518%	268.63	2.983%
103	277.31	274.88	0.876%	269.69	2.748%	292.95	5.640%
104	278.77	285.22	2.314%	299.46	7.422%	276.11	0.954%
105	277.68	272.91	1.718%	282.85	1.862%	298.26	7.411%
106	277.55	273.78	1.358%	288.11	3.805%	312.47	12.582%
107	278.93	281.53	0.932%	271.74	2.578%	292.42	4.836%
108	280.11	284.03	1.399%	278.02	0.746%	285.39	1.885%
109	282.57	284.25	0.595%	287.81	1.854%	267.58	5.305%
110	283.73	279.13	1.621%	291.06	2.583%	270.06	4.818%

Table 8. The average relative error.

Model	Average relative error
BP	5.018%
GA-BP	3.124%
PCA-GA-BP	1.637%

From Figure 4 and Figure 5, the PCA-GA-BP model that predicts gold price is more accurate than GA-BP and BP. Table 8 shows that the PCA-GA-BP model average relative error of prediction is only 1.637%, which is less than the GA-BP result which is 3.124% and BP result which is 5.018%. The PCA-GA-BP prediction result is closer to the actual value.

5. Conclusion

The algorithm proposed in this paper combines PCA, GA and BP neural network. PCA can simplify the network structure and reduce the dimension of input data. The genetic algorithm optimizes the weights and thresholds of the BP neural network, and overcomes the shortcoming that the BP neural network is easy to fall into the local minimum. BP neural network can predict the nonlinear relationship of gold price. The PCA-GA-BP model could predict the price of gold accurately, which has certain reference significance in the financial field. In the next step, we will continue to improve the BP network on the basis of this research, and combine other algorithms to further improve the accuracy of the forecast price.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Chen, L. (2013) Gold Price Prediction Model Based on PP-BPNN. *Computer Simulation*, **30**, 354-357.

[2] Serletis, A. and Shintani, M. (2003) No Evidence of Chaos but Some Evidence of Dependence in the US Stock Market. *Chaos Solitons* & *Fractals*, **17**, 449-454. https://doi.org/10.1016/S0960-0779(02)00387-9

[3] Liu, Y.Q., Song, J.K. and Zhou, X.J. (2003) Research on the Dynamic Trend of Gold Market. *Quantitative Economics* & *Economics Research*, **20**, 25-29.

[4] Yu, F. (2004) An Empirical Study of Recent Gold Price Fluctuations. *Industrial Economics Research*, No. 1, 30-40.

[5] Xu, L.P. and Luo, M.Z. (2011) Short-Term Analysis and Forecast of Gold Price Based on ARIMA Model. *Finance and Economics*, No. 1, 26-34.

[6] Fei, J.W. (2017) Analysis and Forecast of China's Gold Futures Price Based on ARIMA Model. *Contemporary Economics*, No. 09, 148-150.

[7] Xu, G.Y. (2014) Chinese Gold Futures Price Forecasting Model Based on Grey Forecasting Method. *Gold*, **35**, 8-11.

[8] Peng, Y.S., Zhang, D.S., Wang, R.X. and Chen, C. (2011) GARCH Prediction Model with Exogenous Variables for International Gold Prices. *Gold*, **32**, 10-14.

[9] Xia, X.Y. (2013) Research on Volatility of China's Gold Price Based on GARCH Model. *Journal of Science and Technology Pioneering*, **6**, 18-19.

[10] Zhang, K., Yu, Y. and Li, T. (2010) Wavelet Neural Network's Application in Gold Price Prediction. *Computer Engineering and Applications*, **46**, 224-226+241.

[11] Xu, X. and Lu, X.L. (2016) GA-Optimized Acceleration Feature Selection Method in Behavior Recognition. *Computer Engineering and Applications*, **5**, 139-143+166.

[12] Gao, W.H. (2010) Web Datas Mining Based on BP Neural Network. South-Central University For Nationalities, Wuhan.

[13] Liu, W., Liu, S. and Bai, R.C. (2017) Research on Mutual Learning Neural Network Training Method. *Chinese Journal of Computers*, **40**, 1291-1308.

[14] Ding, H.F. and Li, Y.H. (2016) Research on Travel Time Combination Forecast of Expressway Based on BP Neural Network and SVM. *Application Research of Computers*, **33**, 2929-2932+2936.

[15] Ren, X.L. and Lv, L.Y. (2014) A Survey about Network Important Nodes' Sorting Methods. *Chinese Science Bulletin*, **59**, 1175-1197. https://doi.org/10.1360/972013-1280

[16] Yan, X., Li, S.Y. and Zhang, Z. (2016) Application of BP Neural Network Based on Genetic Algorithm in City Water Consumption Forecast. *Computer Science*, **43**, 547-550.

[17] Ding, Y., Jiang, F. and Wu, Y.Y. (2016) GA's Application in Bus Dispatching. *Computer Science*, **43**, 547-550.

[18] Gao, Y.M. and Zhang, R.J. (2014) House Price Prediction and Anlysis Based on Genetic Algorithm and BP Neural Network. *Computer Engineering*, **40**, 187-191.