Remittances Review
January, 2023
Volume: 8, No: 1, pp. 289-306
ISSN: 2059-6588 (Print) | ISSN 2059-6596 (Online)

Received: 11 November 2022 Accepted: 02 February, 2023
DOI: https://doi.org/10.33182/rr.v8i1.19

An empirical Study of the Impact of Market Trading Information on Investors' Returns in China Based on Neural Net Model

Junxiao Gui[1,2*], Nathee Naktnasukanjn[1], Wei Yuan[2], Xi Yu[3]

Abstract

This paper selects the CSI 300 index as the research object, because its constituent stocks have strict selection criteria, the index has a high market coverage and uniform industry distribution, so it can basically reflect the stock market volatility. A total of 3800 sets of data from January 2006 to December 2021 are selected, and a training group and a test group are set up. Based on the research of BP neural network, PCA analysis method and GA genetic algorithm are cited first, and the advantages and disadvantages of each research method are introduced respectively. On this basis, a multi-combination prediction model based on BP neural network model is constructed, and the fitness function of GA genetic algorithm is improved, and finally the prediction performance of different models is derived through experiments. The experimental results show that the prediction results of the multi-combination prediction model are all better than the single BP neural network prediction model; the prediction results obtained by the improved GA-BP neural network model have 4% higher accuracy compared with the unimproved GA genetic algorithm; the efficiency of optimizing the number of variables is improved by about 55%.

Keywords: *CSI 300 index, stock market, PCA, fitness function, improved GA-BP neural network*

Introduction

As the main vehicle of a country's financial activities, the stock market has a profound impact on the economic operation of a country at the macro level and on the development of enterprises and individuals at the micro level. The economic development of a country cannot be achieved without the healthy functioning of the stock market(Caldarelli, Battiston, Garlaschelli, & Catanzaro, 2004; Lee, Lee, & Hong, 2007; Shibamoto & Tachibana, 2014). The number and size of investors are gradually growing, so predicting stock trends in advance can be a good way for investors to both avoid large fluctuations in individual stocks in advance, stop losses effectively, and bring more substantial returns(Abu-Mostafa, Atiya, Magdon-Ismail, White, & Racine, 2001; Bauer, 1994; Mei, 2022).

As research continues, scholars are increasingly inclined to artificial intelligence research tools,

[1] International College of Digital Innovation, Chiang Mai University, Chiang Mai, 50000, Thailand.
[2] Overseas Education College, Chengdu University, Chengdu, Sichuan, 610000, China.
[3] University of stirling, Chengdu University, Chengdu, Sichuan, 610000, China.
Corresponding author: Junxiao Gui (g_jxqingmai@163.com)

which heralds a new level of research in the stock market. The literature (P. Zhang & Shen, 2019) uses the historical data of the stock market and the current market behavior to summarize the price fluctuation law through mathematical statistics, which is eventually reflected directly by charts or some indicator figures. The literature (Traore, Kamsu-Foguem, & Tangara, 2018) uses data from the CSI 300 index and uses the EGARCH model to show that stock market volatility and futures are correlated. The biggest feature of nonlinear forecasting method compared with linear forecasting is that it can achieve nonlinear mapping. It is able to store data in a distributed manner for the purpose of improving the accuracy of stock prediction (Di P L, 2016; Di Persio & Honchar, 2016). It is these advantages that are preferred by financial data researchers. These predictions are only qualitative predictions, exploring the interactions between variables, which are still not very effective in practical investment decisions. Along with the rapid development of data mining, a new era of stock market research has been ushered in. The literature (Adebiyi, Adewumi, & Ayo, 2014) used the SSE index as the research object and used data mining from a non-quantitative perspective to first filter the data and then classified it to make a comprehensive forecast. The literature (Hsu, 2011) used data mining techniques, combined with BP neural network as well as decision tree algorithm to mine the relationship of corporate health finance. The literature (H. Zhang, 2018) improved the BP neural network after creating an intelligent system to help investors predict the trend and then choose the buy-sell time point. The main tasks of data mining are classification, clustering, prediction, association analysis and bias detection. Based on this technique, the literature (Göçken, Özçalıcı, Boru, & Dosdoğru, 2016) combined decision tree classification algorithm to improve the prediction accuracy and predict the stock market direction well by creating multiple analysis indicators for data without identified features. This paper first introduces some traditional stock market forecasting methods and the use of data mining to predict large methods. Then some important indicators of the stock market are introduced, and a summary of the problems arising from stock market forecasting is found that the accuracy of forecasting results will be greatly reduced by direct input data due to the numerous and complicated stock market quantities. Therefore, the common PCA method of doing dimensionality reduction on the input variables is focused on, and the PCA-BP model is established; then the GA is explained in detail and the GA-BP model is proposed. For the inadequacy of the GA-BP model, an improved GA-BP model is proposed. Finally, according to the proposed model, bring in the data to do simulation and synthetically compare the predicted output of each model.

Neural Network Model

PCA-BP neural network model theory

The basic idea of principal component analysis (PCA)

PCA is commonly used in statistics as a method of data dimensionality reduction, the main principle is to transform the original possibly related variables into mutually unrelated variables through orthogonal means, and select fewer variables with overall characteristics to replace the whole according to the research needs, the transformed full-rank unrelated combination of variables is called principal components. At present, the most commonly used is to determine $F1$ and $F2$ in order of variance (at this time, it is necessary to meet $\mathrm{Cov}(F1, F2) = 0$), where $F1$ and $F2$ contain less and less information. The specific mathematical process is as follows:

Assumption X contains n samples, each consisting of p variables, denoted as:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = (X_1, X_2, ..., X_p) \quad (1)$$

$$X_J = \begin{bmatrix} X_{1j} \\ X_{1j} \\ \vdots \\ X_{1j} \end{bmatrix}, j = 1, 2, ..., p$$

Here ; PCA analysis i.e. transforming p potentially relevant original variables into m mutually irrelevant new variables as:

$$\begin{cases} F_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \\ F_2 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \\ F_m = a_{m1}X_1 + a_{m2}X_2 + ... + a_{mp}X_p \end{cases} \quad (2)$$

The matrix is of the form $F = AX$, where:

$$F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}, \mathrm{A} = \begin{bmatrix} a_{11} & a_{12} & a_{1p} \\ a_{21} & a_{22} & a_{2p} \\ & & \\ a_{m1} & a_{m2} & a_{mp} \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{bmatrix}, X = (X_1, X_2, ..., X_p) \quad (3)$$

where: A is called the principal component coefficient matrix and a_{ij} is the principal component coefficient. The model must satisfy the following conditions:

$$a_{k1}{}^2 + a_{k2}{}^2 + ... + a_{kp}{}^2 = 1, (k = 1, 2, ..., m) \quad (4)$$

$$Var(F_1) > Var(F_2) > Var(F_3) > ... > Var(F_m) \quad (5)$$

2.1.2 Steps of principal component analysis

(1) Standardization of raw data:

$$X_{ij}^{*}=\frac{X_{ij}-\overline{X_{j}}}{\sqrt{\operatorname{var}(X_{j})}},\quad i=1,2,\ldots,n;\quad j=1,2,\ldots,p \tag{6}$$

Among them:

$$\overline{X_{j}}=\frac{1}{n}\sum_{i=1}^{n}X_{ij} \tag{7}$$

$$\operatorname{var}(X_{j})=\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\overline{X_{j}})^{2} \tag{8}$$

(2) Calculate the correlation coefficient matrix:

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \tag{9}$$

Among them:

$$r_{ij}=\frac{1}{n-1}\sum_{t=1}^{n}X_{ti}X_{tj}\quad ;\quad i,j=1,2,\ldots,p \tag{10}$$

(3) Calculate the eigenvalues and eigenvectors, and transform the variables into the corresponding principal components:

$$F_{j}=\alpha_{j1}X_{1}+\alpha_{j2}X_{2}+,\ldots,\alpha_{jp}X_{p};(i,j=1,2,\ldots,m) \tag{11}$$

In general, the cumulative contribution of the k principal components is compared with 1. When the value is closer to 1, it means that the better the selection of principal components is, the better it reflects the overall information. However, in practice, it is almost impossible to reach the value of 1. The study concluded that when $kS>85\%$, the selection of principal components is considered valid and can reflect the main information of the original data.

PCA-BP neural network model

PCA uses the idea of data dimensionality reduction, which has the advantage of removing the redundant information of data under the condition of minimum loss of original data, and achieves the purpose of converting high-dimensional data into low-dimensional data for analysis, so as to speed up the prediction speed and improve the prediction accuracy.

In combination with BP neural network learning, there is a natural idea to integrate PCA method with BP neural network to build data analysis model. At this point, we need to consider the original data to be processed from a larger range of dimensionality reduction by PCA method, and then input into the BP neural network as input variables, which will greatly reduce the computational difficulty. With this idea, the cumulative contribution rate of the first two principal components reached more than 90%, which led to the conclusion that the selection of principal components was meaningful and that the variables originally input to the neural network were significantly reduced. The prediction process of PCA-BP model in this paper is shown in Figure 1.

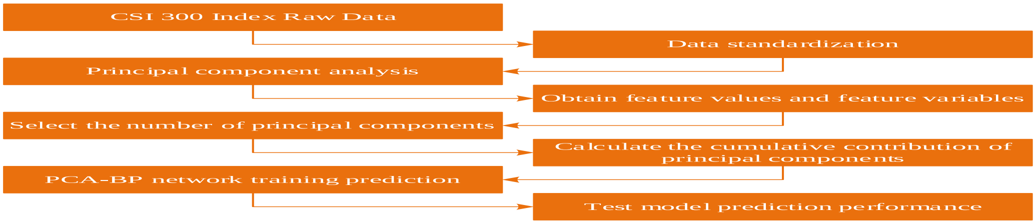
CSI 300 Index Raw Data
Data standardization
Principal component analysis
Obtain feature values and feature variables
Select the number of principal components
Calculate the cumulative contribution of principal components
PCA-BP network training prediction
Test model prediction performance

Figure 1 PCA-BP neural network stock price prediction process.

GA-BP neural network model theory

Genetic Algorithm (GA)

Genetic algorithm (GA) imitates the natural evolutionary process of living organisms, and the core mechanism is the survival of the fittest. Natural organisms evolve from lower to higher levels, and the same concepts of chromosomes, heredity, and parent-child generation are introduced in the bionic algorithm to obtain good offspring through continuous selection. In fact, genetic algorithm is a very classical global optimization algorithm, without the trouble of falling into local optimum due to gradient descent. Due to the good performance of this bionic algorithm, researchers from different disciplines have been exploring it constantly. The process of genetic algorithm is complex and includes coding representation, generation of primitive populations, chromosome selection, crossover, mutation, and then seeking optimal solutions in new subpopulations.

Basic principles of genetic algorithm

(1) Encoding representation;

Encoding is the hypothetical solution to the problem of treating chromosomes as unknown, when the genetic algorithm is applied to correspond this required solution to the chromosomes. Genetic algorithms are coded in various ways, often used for binary coding, which is convenient for pattern theorem analysis, Gray code coding, which is convenient for local search of continuous functions, and floating point coding, which is convenient for the solution of continuous asymptotic problems. Although the coding requirements are not high, but considering the data characteristics of the research object, this paper adopts the binary coding method. By encoding, the solution of the problem of things to be solved is mapped to the genetic gene string.

(2) Adaptation function;

Simply put, the higher the degree of fitness of individuals in a population in nature, the more they can overcome environmental factors to reproduce and thus survive, and vice versa. This relationship is mapped to the expression of fitness function in genetic algorithm, which is basically executed without the help of external information or auxiliary information, and the fitness function is the driving force of GA. The choice of fitness function varies for different research problems. Therefore, it is important to determine the appropriate fitness function.

(3) Parameter setting;

The parameter setting of genetic algorithm is very important for prediction, where the parameters include population size N, length of chromosomes L, mutation probability Pm, mating probability Pc, adaptation value evaluation termination condition, etc. The population size (N) affects the search ability, operation effect and convergence of the algorithm. So far, a definite N value has not been given by academia. If N is set too large, more genetic diversity of the population can be preserved, but it will greatly increase the computation and reduce the operation efficiency; if N is set too small, although it will reduce the computation, but at the same time, because the population size is too small and does not contain more individual information, it will fall into the dilemma of local optimum. In the current study, scholars generally set N between 20 and 100.

The execution process and characteristics of genetic algorithm

(1) The execution process of genetic algorithm;

The specific operation process is: Firstly, initialization is performed and binary coding is used. Then relevant parameters are set and N initial populations are randomly generated. After that, individual fitness is evaluated and population evolution is performed. Specifically, it includes selection, crossover, and mutation. Then set the termination condition to check whether the condition is satisfied and whether to continue the next round of evolution. The specific flow chart is shown in Figure 2.

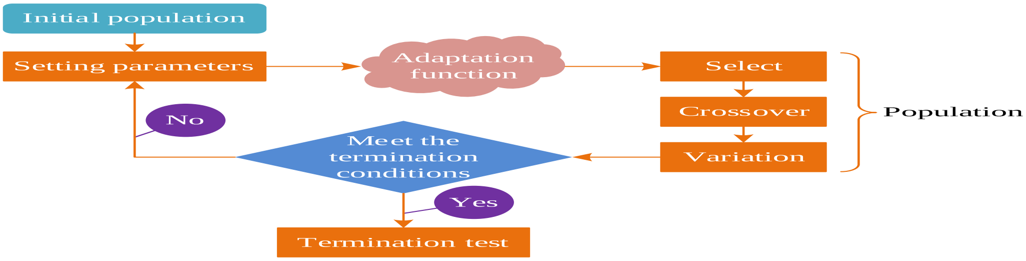

Figure 2 Flow chart of the genetic algorithm.

(2) Characteristics of genetic algorithms:

I. The solution obtained by the genetic algorithm is a set of encoding set, not the solution set of the problem itself;

II, the genetic algorithm search starts at multiple points, and not at one point, and can invisibly reach the global search for the optimal solution;

III. The fitness of the candidate solution is calculated according to certain conditions, and only the fitness function is required in the operation process;

IV. Not with definite state transition rules, using adaptive adjustment of evolutionary direction.

Improved GA-BP Neural Network Model Theory

Combined GA-BP neural network model

The principal component analysis method does a dimensionality reduction on the data, which means that the dimensionality of the features is reduced. The most obvious change is that the number of features of the data becomes less. Unlike feature filtering, the core idea of principal component analysis is to map the data through a high-dimensional space to a low-dimensional space. The result of this is that variables that are highly variable in the original data play a large role in the classification. The most important part of the principal component analysis method is to calculate the principal component, i.e., the eigenvector with the largest eigenvalue. After the processing of the principal component analysis method, the data is no longer the original input features, but a new combination of the original features.

Genetic algorithm is to make use of the idea of survival of the fittest, through the guidance of suitable fitness function within the genetic algorithm, so that the most advantageous information of the parent generation is inherited to the offspring, and through continuous evolution, so that the quality individual information is perpetuated. So using this method the most advantageous information of the original data can be filtered out. The structure of the combined GA-BP neural network model is shown in Figure 3.

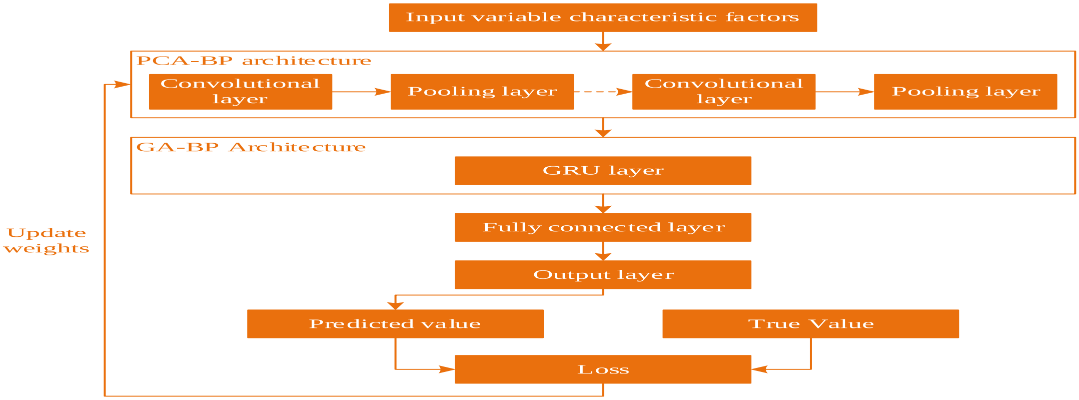

Figure 3 GA-BP neural network model

Therefore, the reciprocal of mse is used as the applicability function with the following equation:

$$f(x) = \frac{1}{mse} \qquad (12)$$

$$mse = \sum_{i=1}^{n} (P_i - t_i)^2 \qquad (13)$$

Where, x is the number of chromosomes, P_i is the actual value, t_i is the predicted value, and mse is the mean square error.

Using this fitness function the combinations with small prediction errors can be selected down and vice versa.

Analysis of GA-BP neural network model with improved fitness function

Since BP neural networks are highly stochastic, genetic algorithms are applied in order to perform a global optimization of the input variables to the neural network:

$$f(x) = \frac{1}{\omega \times mse + (1-\omega) \times n} \qquad (14)$$

where x represents a chromosome; ω is the adjustment coefficient, ω between zero and one, which adjusts the prediction error and the weight of the number of variables in the function; and n is the number of variables selected.

From a genetic point of view, different chromosomes always contain different information of genetic material and carry different numbers of genes.

In the improved genetic algorithm, when two different chromosomes are close in error, the chromosome with fewer genes will be selected down.

Therefore, the improved fitness function can achieve better experimental results.

Model evaluation indicators

In order to evaluate the prediction accuracy, we are going to introduce an evaluation metric here: the loss function. In this paper on stock market research, we take MAE, MSE, RMSE, MARE, MSRE, RMSRE, MAPE, MSPE, RMSPE and the number of variables as the loss function.

Therefore, the smaller these values are, the better the prediction is.

The formulas for these metrics are shown in Table 1, where A corresponds to the actual value at moment t, P corresponds to the predicted output at moment t, and T is the sample size.

Table 1: Evaluation criteria and its calculation formula of the prediction model

Evaluating indicator	**Computing formula**
MAE	$\text{MEA}=\text{T}^{-1}\sum_{t=1}^{\text{T}}\lvert A_t - P_t\rvert$ (15)
MSE	$\text{MSE}=\text{T}^{-1}\sum_{t=1}^{\text{T}}(A_t - P_t)$ (16)
RMSE	$\text{RMSE}=\left[\text{T}^{-1}\sum_{t=1}^{\text{T}}(A_t - P_t)^2\right]^{\frac{1}{2}}$ (17)
MARE	$\text{MARE}=\text{T}^{-1}\sum_{t=1}^{\text{T}}\left\lvert\frac{A_t - P_t}{A_t}\right\rvert$ (18)
MSRE	$\text{MSRE}=\text{T}^{-1}\sum_{t=1}^{\text{T}}\left(\frac{A_t - P_t}{A_t}\right)^2$ (19)
RMSRE	$\text{RMSRE}=\left[\text{T}^{-1}\sum_{t=1}^{\text{T}}\left(\frac{A_t - P_t}{A_t}\right)^2\right]^{\frac{1}{2}}$ (20)
MAPE	$\text{MAPE}=100{*}\text{T}^{-1}\sum_{t=1}^{\text{T}}\left\lvert(\frac{A_t - P_t}{A_t})\right\rvert$ (21)
MSPE	$\text{MSPE}=100{*}\text{T}^{-1}\sum_{t=1}^{\text{T}}\left(\frac{A_t - P_t}{A_t}\right)^2$ (22)
RMSPE	$\text{RMAPE}=\left[100{*}\text{T}^{-1}\sum_{t=1}^{\text{T}}\left(\frac{A_t - P_t}{A_t}\right)^2\right]^{\frac{1}{2}}$ (23)

Analysis of Experimental Results

Experimental data

The selection of CSI 300 index is very strict, taking the size and liquidity of stocks as the fundamental criteria, which ensures the stability of the index to a certain extent, and as a trading constituent index will constantly update the constituents due to a company's business decision, structural adjustment, etc. Therefore, the CSI 300 index basically reflects the fluctuations of the Shanghai and Shenzhen stock markets. Therefore, this paper takes CSI 300 index as the research data, and the data source is Dongfang Fortune software data download. This paper takes the historical data of CSI 300 Index from January 1, 2006 to December 31, 2021, and removes some objective factors such as holidays, there are 3966 groups of stock price, average price and volume data. According to the formula for calculating the selected variables, 3800 rows and 45 columns of sample data are finally obtained. The sample data is divided into two parts, the first 3600 as the training model and the last 200 as the test data. Figure 4 shows the daily chart of closing prices of all sample data. The time span taken for the study is large and includes some government policies and economic crisis events, which basically shows that the sample data are adequate. From Table 2, we can get that the skewness value of the closing price of CSI 300 index is greater than 0 and the kurtosis value is less than 3, that is, the data closing price does not obey normal distribution.

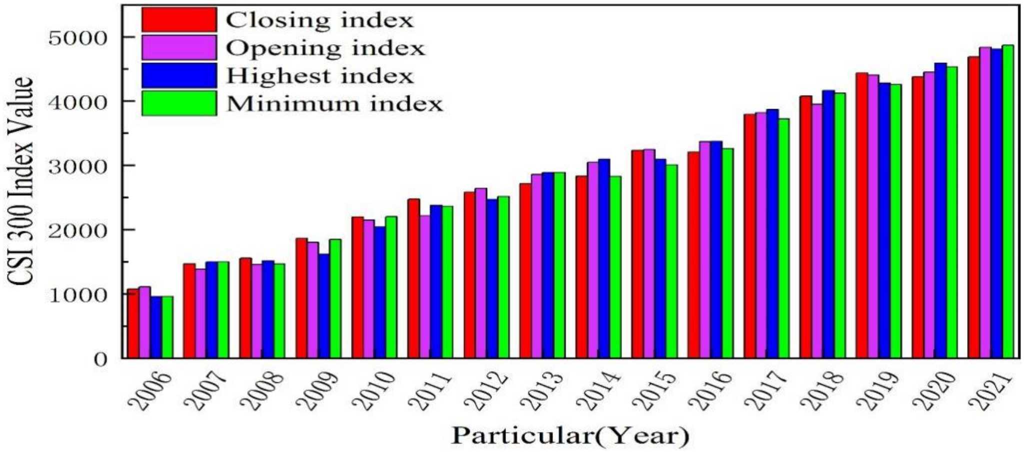

Figure 4 Daily chart of the closing price of the CSI 300 index.

Table 2: Descriptive statistical results of the closing price of the CSI 300 Index.

CSI 300 Index	All samples	Training Sample	Test Sample
Sample size	3800	3600	200
Mean value	2543.2	2496.7	3313.5
Standard deviation	1109.3	1125.8	76.2
Co-efficient	12385.5	12409.6	8908.1
Maximum value	5963.8	5748.0	3606.2

Minimum value	997.1	808.4	3159.3
Skewness	0.454	0.603	0.527
kurtosis	2.788	2.148	2.406

Investor return prediction results for a single BP neural network

The BP neural network prediction model has the following five steps: data preprocessing, structure determination, modal parameter initialization, training, and testing.

Data pre-processing

Because the sample variables are different, their corresponding unit orders of magnitude differ. This will lead to the variables with large order of magnitude seriously affect the weight of the variable in the total variables, so the data should be normalized first. In this paper, we use the normalized mapminmax function that comes with MATLAB to transform the original data with the following formula:

$$\begin{cases}[\text{inputn},\text{inputps}] = \text{mapminmax}(\text{input\_train}) \\ [\text{outputn},\text{outputps}] = \text{mapminmax}(\text{output\_train})\end{cases} \qquad (24)$$

Determining the structure of BP neural network

Scholars generally apply the following formula for the selection of the number of neurons in the hidden layer. Multiple experiments are used to select the optimal number of neurons in the hidden layer. The essence is that the number of neuron implicit layer nodes increases gradually experimentally from less to more.

Based on the previous selection to determine the number of neurons, the number of nodes in the hidden layer is calculated between, and then after several experiments, the indicators described in the previous section are used as evaluation indicators, and each indicator is determined by the average of 100 experiments, and the results are obtained after several experiments as shown in Table 3. It can be concluded that: the best results are obtained when the number of training times per training is not changed and the hidden layer variable is 18. Therefore, the 3-layer structure of the single BP neural network is determined as 50-18-1.

Table 3: Training results of the number of neurons in different hidden layers

Number of neurons in the hidden layer	MEA	MSE	RMSE	Number of training sessions
10	26.63	1342.95	37.54	100
11	25.13	1142.43	38.28	100
12	25.30	1223.23	33.17	100
13	27.13	1640.17	39.29	100
14	27.03	1081.36	32.80	100
15	27.61	1213.78	34.28	100

16	26.55	1291.48	33.51	100
17	24.60	1349.37	36.85	100
18	24.02	1048.35	32.60	100
19	28.32	1283.68	37.21	100
20	27.31	1231.01	36.28	100

Initialization of model parameters

For the selection of the transfer function, it needs to be processed by certain weighting. The input layer input variables are 40, the output layer variables are 1, as the middle implicit layer in this paper selected Sigmoid tangent as the activation function, while the output layer function selected Pure-linear transfer, training algorithm using LM function.

(The same function is used for different combinatorial neural network models below.) Also, the sample is set to be trained 1000 times, the accuracy is required to reach le^{-4}, and the learning rate is set to 0.02.

Model training and model testing

A single BP neural network program is written using MATLAB2018a. The relevant input variables are introduced and the learning training yields: Comparing the training curve with the actual curve, the BP neural network has a good training effect. A better prediction function can be generated after training. Then the test sample is simulated and compared with the actual closing price. The test sample output and error result output are shown in Figure 5.

It can be concluded that (red curve actual output curve, blue curve predicted output curve), the prediction curve of the BP neural network prediction model does not deviate from the actual value by a large margin and the fitting error is small, which can be controlled within 5%.

From the evaluation data of the divine BP via network prediction model, it can be seen that the prediction effect is good, and it is feasible to apply the model to the study of stock market direction.

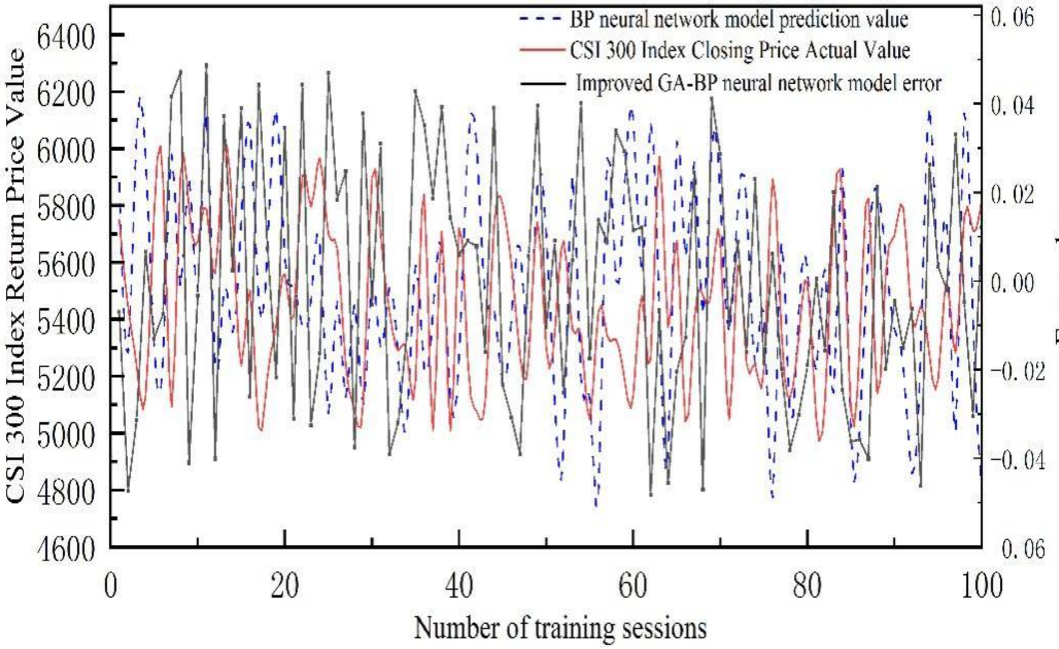

Figure 5 Simulation diagram of the test results for the BP neural network model.

Investor Return Prediction Results of PCA-BP Neural Network

Determining the PCA-BP structure

(1) The original data samples were made into a matrix of 3600 times 50, and due to the large amount of data, they were first normalized using MATLAB2018a and A1 was used to represent the processed data matrix.

(2) For the processed data, the Princomp function program in MATLAB is called to calculate the parameters corresponding to the principal component analysis. The specific values are shown in

Table 4: Some Parameters after the PCA principal component analysis

Principal components	**Eigenvalue**	**Feature contribution rate (%)**	**Cumulative contribution rate of features (%)**
1	27.38	61.81	61.81
2	15.27	30.53	79.04
3	6.343	9.982	85.92
4	2.341	6.473	92.47
5	1.027	1.905	96.39

6	0.850	0.986	97.85
......	……	……	……

Selection of the number of principal components

The variables in the input layer of the neural network are the number of principal components we need, which is currently determined by the cumulative contribution of the features counted and $Cvca \geq 85\%$. With the results in Table 4, it can be concluded that the cumulative contribution of the first three principal component variables is over 88%, which is already more than 85%. Therefore, only three principal components are needed to ensure that the majority of the characteristics of the data and the selected data are representative.

Different numbers of principal components have different effects on the BP neural network prediction, so it is important to verify whether the results above 3 principal components have optimal effects on the BP neural network prediction.

In this section, the number of principal components is determined according to 4. As can be seen from Table 4, when the number of principal components is increased to 6, the feature contribution rate has been reduced to less than 1%, and the impact is almost negligible.

Therefore, the number of principal components is selected as an integer between 1 and 5 for the experiments. The BP neural network model, evaluation indexes refer to the relevant settings. The learning rate is 0.2, the accuracy is set to le^{-4}, and the number of training is 100. The average value of 100 experiments is taken, and the standard deviation of the evaluation index is also calculated, so that the following experimental results are obtained.

From Table 5, we can see that the number of principal components should be taken as 2, and the effect is optimal at this time.

Table 5: Test results of the number of principal components of the CSI 300 Index

Principal Components	MAE	MSE	RMSE	MARE	MSRE	RMSRE	MAPE	MSPE	RMSPE
1	29.8	1527	39.09	0.083	0.012	0.887	0.020	0.013	0.116
2	27.6	913	30.32	0.006	0.009	0.683	0.047	0.008	0.122
3	25.3	842	30.67	0.001	0.009	0.592	0.058	0.009	0.026.
4	24.6	1048	32.36	0.042	0.009	0.675	0.096	0.015	0.039
5	23.5	854	29.23	0.016	0.009	0.718	0.085	0.024	0.048
6	21.3	833	28.14	0.001	0.001	0.657	0.042	0.027	0.059

After performing principal component analysis on the data by MATLAB, it can be concluded that the prediction error is minimized at a number of principal components of 2 for each training of 100. In this way, a representative analysis can be done using the variables corresponding to the principal components.

PCA-BP model training and testing results

The structure and parameters of the BP neural network model are still the same as in the previous paper (the number of implicit layers of this model is 2 through training), and the experimental results show that the two principal component variables can be selected to represent the important information of the original 40 variables, which leads to credible prediction results, and the output value is still the closing price of the second day. The optimized samples were brought into the PCA-BP neural network model for training. the performance evaluation results of the PCA-BP neural network model are shown in Table 6, which shows that the error amplitude of the PCA-BP neural network model is smaller and the error can be controlled within 3%, indicating that the prediction accuracy is higher and the prediction error is smaller compared with the single BP model. The performance index results table PCA-BP neural network model works well in solving the dimensionality reduction problem, has better self-adaptive ability, and has great advantages in dealing with nonlinear problems.

Table 6: Evaluation table of the P C A-BP neural network model

Evaluation metrics	Training samples	Test samples
MAE	12.73	15.24
MSE	253.96	844.11
RMSE	15.72	31.24
MARE	0.012	0.007
MSRE	1.6E-04	8.0E-05
RMSRE	0.134	0.009
MAPE	1.108	0.687
MSPE	0.019	0.005
RMSPE	0.135	0.089
Number of indicators	40	40

Results of Improved GA-BP Neural Network for Investor Return Prediction

Improved genetic algorithm fitness function

After the parameter setting of the GA genetic algorithm, this paper still follows the previous method regarding the parameter setting. The difference is that in the module of the fitness function, the form of the formula is changed. The number of variables is added to the formula on the basis of the prediction error. ω is the adjustment coefficient, and the value is $0<\omega<1$, which in general cannot be too small or too large, and the value of ω needs to be set according to the problem, and $0.2<\omega<0.9$ is chosen for this experiment. Table 7 shows the final number of input variables reduced to the improved GA-BP neural network when different ω is chosen. From the table, it can be concluded that the optimal number of variables obtained by the genetic algorithm is 21 at the value of ω at 0.5. After making the comparison, the experiments in this section set the value of ω to 0.5.

Table 7: Number of final input variables for the improved GA-BP neural network model.

ω **takes the value**	**MAE**	**MSE**	**RMSE**	**Number of variables**
0.3	25.796	1133.849	33.253	15
0.4	25.381	1116.826	33.471	18
0.5	21.453	780.244	28.014	21
0.6	23.597	927.535	29.635	29
0.7	25.550	997.311	31.245	34
0.8	25.341	1060.707	32.309	39

Training and testing results of the improved GA-BP neural network model

After the training of the improved GA-BP neural network model, the 40 variables are streamlined to 15, which means that the input layer variables of the BP neural network are 15. At this time, the optimized 15 input variables are brought into the BP neural network model for learning, and then 100 samples are tested later. The experimental steps of the improved GA-BP neural network model are the same as those of the GA-BP neural network model (the number of implied layers of this model is 6 through training), and the closing prices of the next 100 trading days are brought into the improved GA-BP neural network model after training. The predicted value of the improved GA-BP neural network model is closer to the actual value, and the error is within 1%, so the prediction error is smaller than the above prediction model, so it has a good prospect to be used in stock prediction.

The performance evaluation results of the improved GA-BP neural network model are shown in Table 8. This indicates that the improved GA-BP neural network model can ensure the improvement of prediction accuracy and also reduce the number of variables to a greater extent compared with the above model.

Table 8 Evaluation table of the improved GA-BP neural network model

Evaluation metrics	**Training samples**	**Test samples**
MAE	10.53	20.49
MSE	201.42	7446.13
RMSE	14.02	27.52
MARE	0.007	0.005
MSRE	1.5E-04	8.0E-05
RMSRE	0.014	0.007
MAPE	0.908	0.575
MSPE	0.011	0.005
RMSPE	0.125	0.883
Number of indicators	15	15

Evaluation of the effectiveness of trading strategy returns

Combining the above graphs, the single neural network strategy does not significantly outperform the benchmark strategy in timing during the highly volatile period of 2006-2021, i.e., it does not have significant timing ability and has a high maximum retracement level. The improved GA-BP neural network model strategy performs slightly lower in terms of annualized returns, but has a very significant advantage in terms of position returns, profit/loss ratios, and control of retracement levels. Compared with the PCA-BP neural network strategy, the position return increased by 22.69%, the win rate increased by 6.25%, the profit-loss ratio increased by 202.59%, and the maximum retracement level decreased by 48.82%; compared with the GA-BP neural network strategy, the position return increased by 156.72%, the win rate increased by 4.75%, the profit-loss ratio increased by 679.53%, and the maximum retracement level decreased by 88.64%. It can be shown that adding the improved GA-BP neural network model to the traditional neural network strategy predicts data with both certain timing ability and at the same time achieves risk control.

Conclusion

This paper first explains the algorithm principles of PCA-BP neural network and GA-BP neural network used in the study. Then this paper proposes to improve the GA-BP neural network stock price prediction model based on PCA-BP neural network and GA-BP neural network stock price prediction model for grid trading strategy to obtain backtest results. The findings of this paper are:

(1) It is practical to use the improved GA-BP neural network for stock return forecasting. Both the prediction results of the model and the strategy backtest results show that the improved GA-BP neural network model has better prediction accuracy compared with the PCA-BP neural network model and the GA-BP neural network model. Compared with the GRU model, the improved GA-BP neural network model has an average decrease of 33.8% in the MAE of the test set and 31.73% in the RMSE of the test set, with an error within 1%, indicating that the model improves the generalization ability of the model in stock return prediction to a certain extent.

(2) Constructing complex networks of upstream and downstream industry indices and incorporating complex network indicators into stock price prediction are effective features. Compared with the deep learning strategy, the position return is improved by 22.69%, the win rate is improved by 6.25%, the profit-loss ratio is improved by 202.59%, and the maximum retracement level is decreased by 48.82%; compared with the grid strategy, the position return is improved by 156.72%, the win rate is improved by 4.75%, the profit-loss ratio improves by 679.53% and the maximum retracement level decreases by 88.64%. It can be shown that the improved GA-BP neural network model forecasts data with both certain timing ability and risk control, which can be used for investors' reference and application, thus improving investors' profitability.

References

Abu-Mostafa, Y., Atiya, A., Magdon-Ismail, M., White, H., & Racine, J. (2001). Special issue on neural networks in financial engineering. *IEEE Transactions on Neural Networks, 12*(4).

Adebiyi, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Comparison of ARIMA and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics, 2014*.

Bauer, R. J. (1994). *Genetic algorithms and investment strategies* (Vol. 19): John Wiley & Sons.

Caldarelli, G., Battiston, S., Garlaschelli, D., & Catanzaro, M. (2004). Emergence of complexity in financial networks. *LECTURE NOTES IN PHYSICS-NEW YORK THEN BERLIN-, 650*, 399-424.

Di P L, H. O. (2016). Artificial Neural Networks Architectures for Stock Price Prediction: Comparisons and Applications *International Journal of Circuits, Systems and Signal Processing*(10), 403-413.

Di Persio, L., & Honchar, O. (2016). Artificial neural networks architectures for stock price prediction: Comparisons and applications. *International journal of circuits, systems and signal processing, 10*(2016), 403-413.

Göçken, M., Özçalıcı, M., Boru, A., & Dosdoğru, A. T. (2016). Integrating metaheuristics and artificial neural networks for improved stock price prediction. *Expert Systems with Applications, 44*, 320-331.

Hsu, C.-M. (2011). A hybrid procedure for stock price prediction by integrating self-organizing map and genetic programming. *Expert Systems with Applications, 38*(11), 14026-14036.

Lee, K. E., Lee, J. W., & Hong, B. H. (2007). Complex networks in a stock market. *Computer physics communications, 177*(1-2), 186-186.

Mei, D. (2022). What does students' experience of e-portfolios suggest. *Applied Mathematics and Nonlinear Sciences, 7*(2), 15-20.

Shibamoto, M., & Tachibana, M. (2014). Individual Stock Returns and Monetary Policy: Evidence from J apanese Data. *The Japanese Economic Review, 65*(3), 375-396.

Traore, B. B., Kamsu-Foguem, B., & Tangara, F. (2018). Deep convolution neural network for image recognition. *Ecological informatics, 48*, 257-268.

Zhang, H. (2018). The forecasting model of stock price based on PCA and BP neural network. *Journal of Financial Risk Management, 7*(4), 369-385.

Zhang, P., & Shen, C. (2019). *Choice of the number of hidden layers for back propagation neural network driven by stock price data and application to price prediction.* Paper presented at the Journal of Physics: Conference Series.