Scientific Research Publishing

Journal of Financial Risk Management, 2018, 7, 369-385
http://www.scirp.org/journal/jfrm
ISSN Online: 2167-9541
ISSN Print: 2167-9533

The Forecasting Model of Stock Price Based on PCA and BP Neural Network

Haoling Zhang

Donlinks School of Economics and Management, University of Science and Technology, Beijing, China
Email: zhanghaolingbest@163.com

How to cite this paper: Zhang, H. L. (2018). The Forecasting Model of Stock Price Based on PCA and BP Neural Network. *Journal of Financial Risk Management, 7,* 369-385.
https://doi.org/10.4236/jfrm.2018.74021

Received: October 16, 2018
Accepted: December 2, 2018
Published: December 5, 2018

Abstract

Based on Principal Component Analysis (PCA) and Back Propagation neural network, this paper establishes stock forecast model, and takes the Yunnan Baiyao (000538) as example, 29 indicators are selected from stocks technical analysis, and the neural network is input after dimension reduction and further confirms number of hidden layer nodes, learning rate, activation function and training function of the network in accordance with comparison and analysis of Mean Square Error (MSE) and Mean Absolute Error (MAE) in different parameter data experiments. Lastly, the model with steadiness and accuracy is obtained.

Keywords

BP Neutral Network, PCA, Stock Price Forecasting

1. Introduction

With the economic development and people's increasing financial and investment awareness, stock investment of high-risk and high-income has become a significant means of financial management. As of May 2018, A-share investors have exceeded 100 million, the total amount of A-shares has already exceeded 3500 and the total amount of A-share reaches 80 trillion yuan. Meanwhile, there are differences between stock market structure in China and that in capitalist countries, such as USA and UK. In the proportion of investment, individual investors are higher than institutional investors, which lead the lack of integrity in China's stock market, and more tend to change with difference of people's psychological state, and its law of fluctuation is hard to grasp.

Therefore, the forecast analysis of stock valuation has very attractive commercial application value. In recent year, a series of the analysis method has emerged in studying development trends of stock price: K-line Graph Analysis, Moving Average Method, Dow Analysis, Price Analysis and Wave Theory, and ect. (Yang, 2017). However, due to the large amount and complexity of stock market on daily basis, the traditional fundamental analysis of stock (Du, 2006) or technical analysis method (Qiao, 2013) is hard to forecast accurately the dynamic change trends of stock market. Meanwhile, as stock data of abnormal distribution, high noise and unstable characteristics as concerned, the traditional mathematical statistics analysis is too simple and not easy to change, as well as cumbersome processing and easy to cause information loss after standard processing.

In recent years, with the continuous development of data mining technology, more and more researcher began to analyze and mine deeply stock market by using computer intelligent algorithm, so as to master stock's law of ups and downs, predict its corresponding approximate valuation for the future, and further enhance the odds of people's investment decisions (Hao, 2017).

Based on the PCA and BP neutral network, this paper establishes prediction models by using dimension reduction in data pre-processing to eliminate the multicollinearity between the original stock evaluation indicators, reduces the redundant information of network input data, and enhances network working efficiency and accuracy.

2. Literature Review

2.1. Foreign Related Research Status

In the United States, Pati and Shnaprasad (Refenes & Latif, 2015) first proposed the Wavelet Neural Network (WNN) in 1992. In the neural network model, the wavelet function is used as a neuron, and the time-frequency and local characteristics of the wavelet mapping function are utilized as well as the discrete analysis model is constructed by wavelet neural network.

Golan & Ziarkow (1995) used rough set theory to analyze the stock historical data of the decade and studied the correlation between stock price and economic index.

Liu & Lee (1997) and others used MATLAB to establish a securities management system and use it to analyze large amounts of securities data.

Tsaih, Hsu, & Lai (1998) and others used clustering methods and visualization techniques to find the best investment opportunity for stocks.

Povinelli, Cai, & Song (2000) proposed to extract the characteristics of the intensity of the event based on the interest of the digger and construct a temporal mode feature function.

Dose & Cinacotti (2005) applied data mining techniques to find the best way to derive stock portfolios, optimize the weight ratio of individual stocks in asset allocation, and thus passively manage funds.

Fiol-Roig, Miro-Julia, & Isern-Deya (2010) and others used data mining technology to evaluate past stock prices, and at the same time, to obtain useful knowledge by calculating some financial indicators, to construct, formulate and evaluate stock investment strategies, so as to forecast of securities trading.

Srikant & Agrawal (2010) proposed a dimensionality reduction analysis method to map high-dimensional stock data sets into low-dimensional subspaces, and to find abnormal data by comparing the differences of subspace mapping data sets.

Kumar, Pandey, Srivastava, et al. (2011) proposed a machine learning system based on a hybrid algorithm of Genetic Algorithm and Support Vector Machine (SVM) for stock market forecasting. The genetic algorithm (GA) is used to select the most meaningful series of input variables from all the technical indicators, and the stock market analysis is carried out by using the correlation between the stock prices of different companies and the technical indicators of highly relevant stocks.

Ticknor (2013) proposed a neural network model based on Bayesian regularization to predict stock price movements. This method uses a Bayesian normalized network to assign a probability feature to the network weights. The network can automatically and optimally Punish overly complex models that demonstrate their effectiveness by predicting the stock prices of Microsoft and Goldman Sachs.

Panigrahi & Mantri (2015) combined the improved support vector regression machine and decision tree to empirically study stock market historical data and BSE-sensex index, and provide decision support for investors to analyze stock market trends.

2.2. Domestic Related Research Status

In China, data mining technology is actually applied to the real economy later than foreign countries. Beat Wuthrich (1998) of Hong Kong University of Science and Technology developed a stock price forecasting system based on data mining technology. The system predicts the trend of the Hang Seng Index by entering real financial data. Today's domestic stock analysis software has predictive functions like Straight Flush, Oriental Wealth, and Great Wisdom, but the data mining functions of these software are relatively simple.

Ma, Lan, & Chen, (2007) proposed based on WNN and SVM to use time series stock data to make data mining using the decision of sequence trading rules, and apply the securities comprehensive index to predictive analysis.

Zhang (2009) applied the SVM to classify the stock's ups and downs. After analyzing the four factors of quantity relative ratio, turnover rate, internal and external ratio and amplitude, which had a great impact on the results, she shifted her research focus to the prediction of SVM regression on the market index.

Guo (2010) proposed an improved Aprior algorithm based on the correlation between stock prices and using the advantages of algorithms in vertical data representation and cross-counting. This algorithm can be applied to the stock analysis simulation system to quickly find the relationship between the rise and fall of stocks, and provide investors with more accurate decision support.

Bao (2013) applied the support vector regression of phase space reconstruction theory, Hopfield network algorithm of neural network and correlation vector machine, LSSVR-CARRX model algorithm and financial time series algorithm based on support vector regression and used these techniques to analyze and forecast the stock market.

In the same year, Sun (2013) used the decision tree, neural network and logistic regression model in data mining to study the relationship between financial indicators and stock prices of listed companies. After comparing and evaluating the advantages and disadvantages of these three methods, they can better determine the investment value of the stock.

Chen (2014) used data mining technology to solve the problem of stock picking and investment timing, and analysed the factors affecting stocks' rise and fall through association rules analysis. After clustering stock data, find the best investment portfolio. The stock price is estimated and predicted by applying an improved neural network algorithm.

Wang (2010) constructed a KNN model with the same effect as the neural network and time series according to the time series. This model is based on the model training of the closing price of the first n trading days of the stock to obtain the closing price of the first n + 1 trading days. The closing price of the trading day, which adjusts the weight according to the actual situation of each dimension, thereby improving the accuracy and reliability of the model.

Through the above combing, data mining technology has been widely used in the financial industry, but the research of foreign scholars is mainly aimed at the optimization of specific algorithms and the overseas stock market. Only a small amount of research is available for individual investors. At the meantime, the domestic stock market research is relatively backward and the market complexity is higher. Therefore, the starting point of this paper is to establish a relatively accurate and stable stock forecasting valuation model for A-share investors.

3. The Forecasting Model Based on BP Neutral Network to Establish Stock Price

3.1. The Indications of Forecasting Evaluation in Stock Price

The paper selects 29 transaction data and technical indicators as evaluation indicators of stock price.

1) Stock price

The Stock Price includes the Opening Price, Closing Price, Highest Price and Lowest Price of the day.

2) Transaction data

The Transaction Data includes the Turnover and Volume.

3) Price Limit

$$\text{Price Limit} = \frac{\text{Closing price of the day} - \text{Opening price of the day}}{\text{Stock market closing price yesterday}} \times 100\% \quad (1)$$

4) MA

The Moving Average is based on the principle of "moving average" in statistics, which the average value of stock prices over a period of time is linked into a curve, showing the historical fluctuations in stock prices. This paper selects MA1 (5-day average line), MA2 (10-day average line), MA3 (30-day average line), MA4 (60-day average line), MA5 (120-day average line), and MA6 (240-day average line) as indicators.

5) MACD

The MACD formula calculates the difference between the long-term and medium-term Exponential Moving Average (EMA) to determine the market. When the MACD turns from a negative number to a positive number, it is a signal to buy. When the MACD turns from a positive number to a negative number, it is a signal to sell. Calculation formula:

DIF line: The difference between the short-term moving average and the long-term moving average.

$$\text{DIF} = \text{EMA}(12\text{th}) - \text{EMA}(26\text{th}) \tag{2}$$

DEA line: M-day exponential moving average of the DIF line, in this paper, M = 9, then 9^{th}.

DIF Difference Exponential Average (DEA) = DIF (Today) × 0.2 + DEA (Yesterday) × 0.8

$$\text{MACD}:\text{BAR} = 2\times(\text{DIF} - \text{DEA}) \tag{3}$$

6) KDJ

The stochastic indicator generally calculates the immature random value RSV of the last calculation period by the highest stock price, the lowest stock price, the closing price and the proportional relationship among the three in a given period, and then in accordance with the method of Exponential Moving Average, the K value, the D value, and the J value are calculated. In this paper, 9 days is a cycle.

$$RSV(n) = \frac{[C(n) - L(n)]}{[H(n) - L(n)]}\times 100 \tag{4}$$

$$K(n) = \frac{2}{3}\times K(n-1) + \frac{1}{3}\times RSV(n) \tag{5}$$

$$D(n) = \frac{2}{3}\times D(n-1) + \frac{1}{3}\times K(n) \tag{6}$$

$$J(n) = 3\times K(n) - 2\times D(n) \tag{7}$$

$C(n)$ is the closing price for the n day; $L(n)$ is the lowest price for the n day; $H(n)$ is the highest price for n day.

7) WR

The Williams Overbought/Oversold Index is mainly used to judge whether the stock market is overbought or oversold, in order to analyze the strength of the market trading momentum.

$$WR(n)=\frac{[H(n)-C(n)]}{[H(n)-L(n)]}\times 100 \tag{8}$$

8) BIAS

The Bias rate indicates the degree of deviation between the closing and the stock moving average, and thus uses the Bias rate to calculate the probability of returning or rebounding when the stock price fluctuates sharply, and the stock price returns to the original credibility within the normal fluctuation range. This paper uses BIAS. BIAS1 is for 12th and BIAS. BIAS2 is for 24th.

$$\text{BIAS}=\left[\frac{(\text{Closing price of the day}-\text{N\_day moving average})}{\text{N\_day moving average}}\right]\times 100 \tag{9}$$

9) RSI

The Relative Strength Index analyzes the market's intentions and strengths by comparing the average closing gains and the average closing declines over a period of time to determine future market trends. This paper selects RSI1 for the 6-day line, RSI2 for the 12-day line, and RSI3 for the 24-day line.

The sum of the gains in the closing day of $U=n$; the sum of the closing declines in $V=n$ (absolute index);

$$\text{N\_day RS}=100*\frac{U}{V};\quad \text{N\_day RSI}=100-\frac{100}{1+RS} \tag{10}$$

10) ROC

The price rate of change is based on the closing price of a certain day before a period of time, and the speed of the closing price of the day is used to judge the strength of buying and selling power, and then analyze the future development trend of the stock price.

$$\text{ROC}=\frac{(\text{The closing price on the day}-\text{The closing price before the } n \text{ day})}{\text{Closing price before the } n \text{ day}}\times 100 \tag{11}$$

11) OBV

The energy tide is to draw the stock market turnover into a trend line, and speculate on the trading atmosphere of the stock market based on the price changes and the increase and decrease of the volume.

$$\text{OBVof that day}=\text{OBV of the day before}+\text{Today's trading volume} \tag{12}$$

$$\text{OBV}=\frac{[(\text{Closing Price}-\text{the Lowest Price})-(\text{the Highest Price}-\text{Closing Price})]}{(\text{the Highest Price}-\text{the Lowest Price})}\times\text{Volume} \tag{13}$$

A total of 29 indicators of these 15 items can be divided into 5 categories, as shown in Table 1. Among them, the market trend indicator is mainly used to indicate the current stock market development trend; the market energy indicator is mainly used to determine the key turning point of the stock market

Table 1. 29 forecasting indicators of stock price.

Price	Opening Price	Open
	Closing Price	Close
	the Highest Price	High
	the Lowest Price	Low
	Price Limit	Change
Transaction Amount	Volume	Volume
	Turnover	Amt
Market Trend	Moving Average	MA (5, 10, 30, 60, 120, 240)
	Moving Average Convergence and Divergence	MACD (IFF, DEA, MACD)
Market Energy Index	Williams Overbought/Oversold Index	WR
	Bias Rate	BIAS
	Relative Strength Index	RSI
	Rate of Change	ROC
	Stochastics	KDJ（K, J, D）
Popularity Index	On-balance Volume	OBV

trend; the popularity index determines the future trend of the stock price through the linkage relationship between the volume and the stock price.

3.2. Using PCA to Reduce the Dimensionality of Stock Evaluation Index

This paper predicts that the stock price selects 15 items of 29 different stock transaction data and technical evaluation indicators. Each evaluation index describes and predicts the fluctuation forms and trends of stock from different angle in the past period of time. Each evaluation index includes part of information, but different indicators would inevitably exist multiple collinearity, that is, the amount of information covered by all the indicators overlaps to some large extent. Meanwhile, overlarge network input also affects the operating convergence efficiency of neutral network and predictable accuracy. Therefore, this paper uses PCA (Zhang & Dong, 2013) to reduce dimensionality of 29 stocks trading data indicators and technical evaluation indicators.

Specific steps are as follows:

Set $X_1, X_2, \cdots, X_P$ observe n times and get the observed data matrix as:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

The results of PCA are affected by magnitude or dimension. Due to the possibly different units of each variable, if the dimensions are changed, the results will be different. Therefore, it is necessary to standardize the data of each variable first, and then the analysis is then performed by using a covariance matrix or a correlation coefficient matrix.

Normalization of the original variable matrix:

$$X'_{ij} = \frac{X_{ij} - \overline{X_j}}{S_j} \tag{14}$$

Among them, $\overline{X_j} = \sum_{i=1}^{p} \frac{X_{ij}}{m}$; $S_j = \sqrt{\frac{1}{n}\sum_{i=1}^{p}\left(X_{ij} - \overline{X_j}\right)^2}$ $\left(i = 1, 2, \cdots, m\right)$.

Subsequently, the normalized covariance or correlation matrix, and the eigenvalues and eigenvectors of the covariance matrix are obtained. At the same time, for each *i*, there is $\sum_{j=1}^{p} a_{ij} = 1$, and $\left(a_{11}, a_{12}, \cdots, a_{1p}\right)$, to make the value of $Var\left(Y_1\right)$ maximized; $\left(a_{21}, a_{22}, \cdots, a_{2p}\right)$ is not only perpendicular to $\left(a_{11}, a_{12}, \cdots, a_{1p}\right)$, but also the value of $Var\left(Y_2\right)$ reaches the maximum; $\left(a_{31}, a_{32}, \cdots, a_{3p}\right)$ is not only perpendicular to $\left(a_{11}, a_{12}, \cdots, a_{1p}\right)$ and $\left(a_{21}, a_{22}, \cdots, a_{2p}\right)$, but also value of $Var\left(Y_3\right)$ reaches the maximum; and so on, until all the components p can be obtained. The solution is to find the eigenvalues of the $X^{\mathrm{T}}X$ matrix.

Set the obtained eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$ of $X^{\mathrm{T}}X$, and their corresponding normalized orthogonal eigenvectors are $\eta_1, \eta_2, \cdots, \eta_p$, the P principal components:

$$\begin{gathered} Y_1 = X\eta_1 \\ Y_2 = X\eta_2 \\ \vdots \\ Y_p = X\eta_p \end{gathered}$$

When performing PCA, it is determined whether or not a principal component is retained, generally based on the retained feature vector accounting for more than 85% of the sum of the total feature vectors. However, sometimes it is necessary to consider the contribution value of the selected principal component to the original variable, which can be expressed by the squares of the correlation coefficients and the meters. If the selected principal components are $Y_1, Y_2, \cdots, Y_q$, and $q \le m$ are satisfied, then they contribute $\rho_i = \sum_{j=1}^{r} q^2\left(Y_j, X_i\right)$ to the original variables of X_i.

3.3. The Predicted Model of Designing BP Neutral Network for Stock Price

During the construction of BP neutral network (Li & Song, 2013) stock price predict models, the important network parameters are needed to confirm, such as the number of neurons of input layer and output layer, the number of hidden layer and number of neurons, the confirmation of learning rate, and the choice of activation function and training function.

3.3.1. The Number of Neurons in Input Layers and Output Layers

This paper selects 29 stock price evaluation indicators. After using PCA to reduce the dimension, the principal components that determined lastly are input variables. Meanwhile, the number of the day that needs to predict closing price should be corresponding to output of variable.

3.3.2. The Number of the Hidden Layers and the Number of Neurons

The Robert Hecht-Nielsen in 1989, proved that any continuous function in any closed interval can be similar to 3-layer BP neutrons. Therefore, this paper confirms to establish the model with 3-layer BP neural network.

The number of hidden neurons is playing an important role of predictable accuracy in neural network. In general, more hidden layers nodes can have better performance; however, it can lead to too long training time. At present, there is no ideal analytical formula to determine the rational number. Generally, it can estimate the approximate number by means of empirical formula and determine this amount by actual operation in the end.

$$h < d - 1 \tag{15}$$

$$h < \sqrt{d + s} + a \tag{16}$$

$$h < \log_2 d \tag{17}$$

Among them, d is the amount of neutrons in the input layer; h is the amount of neutrons in the hidden layer; s is the number of neutrons in the output layer; a is the constant between 0 and 10.

3.3.3. The Learning Rate

The learning rate is playing an important role of training effects and training speed in neural network. Its value range is between 0 and 1, which means that the larger the value is, the greater the modification of link weight for each iteration, the faster the learning rate, and the shorter the training period. However, it is possibly that the excessive learning rate makes the network not converge. On the contrary, if its value is smaller, the studying speed is slower, the operational time will be longer and the memory capacity for algorithm will be larger. Therefore, there must be certain shortcoming with learning rate too large or too small.

The specific value of learning rate takes reference to the empirical formula:

$$\eta = \frac{1}{\sqrt{h}} \tag{18}$$

3.3.4. Activation Function

The activation function of hidden layer and output layer has an important influence on the predicted accuracy of BP neural network. The function of *newff* provides several activation functions in the toolbox of MATLAB neural network.

1) ***logsig*** function:

$$y = \frac{1}{\left[1 + \exp(-x)\right]} \tag{19}$$

2) ***tansig*** function:

$$y = \frac{2}{\left[1 + \exp(-2x)\right]} - 1 \tag{20}$$

3) ***purelin*** function:

$$y = x \tag{21}$$

This paper uses controlling variable method. Under the condition of other same parameters, the error of prediction of different activate function is compared to determine the most suitable activation function.

3.3.5. Training Function

Training function is the most core training algorithm in the neural network, different training function corresponds to different training algorithm.

1) Gradient Descent method

Its principle is to obtain the optimization information from gradient vector and solving the optimal weight by first-order partial derivative.

$$w(n+1) = w(n) + \eta x^{\mathrm{T}}(n)e(n) \tag{22}$$

Among them, w represents weight, n represents number of iterations, η represents learning rate.

There provide many gradient descent training functions in the MATLAB neutral network Toolbox, such as Gradient Descent Method (*traingd*), Gradient Descent Method with Momentum (*traingdm*), and Adaptive Algorithm (*traingda*).

2) Quasi-Newton Methods

Newton's method is rapidly optimized algorithm based on the expansion of Taylor series, its iterative formula is as follows:

$$w(n+1) = w(n) - H^{-1}(n)g(n)\eta(n) \tag{23}$$

Among them, H represents *Hessian* matrix of the error performance function, $H^{-1}(n)g(n)$ represents optimization direction of Newton method. If the matrix of *Hessian* is not positive, its search direction may not be downward direction. Therefore, the improved Quasi-Newton Methods (*quasi-Newton*) is selected.

$$w(n+1) = w(n) - G(n)g(n)\eta(n) \tag{24}$$

Its principle is to use the first-order partial derivative of the training function and then the inverse of the approximation $H(n)$ by another $g(n)$ matrix. The corresponding training function in MATLAB is *trainbfg.*

3) LM algorithm

Using *LM* (*Levenberg-Marquardt*) doesn't need to calculate the specific matrix *H*, while it is shown as an approximate representation as

$$H = J^{\mathrm{T}}J \tag{25}$$

Its descending gradient is expressed as:

$$g = J^{\mathrm{T}}e \tag{26}$$

J is the Jacobian matrix containing the first derivative of the network's weight error function. It modified network weight:

$$w(n+1)=w(n)-\left[J^{\mathrm{T}}J+\mu I\right]^{-1}J^{t}e \tag{27}$$

When $\mu=0$, it expresses Newton method, when μ is very large, it is similar to the gradient descent method with a smaller step size. It is calculated by *trainlm* in MATLAB.

4) Powell-Beale algorithm

Its principle is to firstly judge the orthogonality of the gradient before and after the network. If it is orthogonal, adjust the threshold and weight of the network back to the negative gradient direction. It is calculated by *traincgb* in MATLAB.

The paper uses these six algorithms to be trained, such as *traingd traingdm traingda trainbfg trainlm traincgb*, to observe the prediction accuracy, so as to select the best training function.

4. Data Experiment

4.1. Sample Data

This paper selects Yunnan Baiyao (000538) stocks listed on December 15, 1993. In the past five years from May 13, 2013 to May 10, 2018, the suspended data were removed from July 19, 2016 to Dec. 30, 2016 as well as from April 25, 2017 to April 26 with total of 1105 groups of data as sample. 1005 groups of data are selected at a random as training set, the other 100 groups of data as testing set. The data of this paper comes from Yahoo finance.

The data selection of this paper mainly considers two factors:

1) The time span of input variables

This paper selects 29 indicators of daily data, including transaction data and technical indicator data of Yunnan Baiyao from May 2013 to May 2018. For these five years, this time span is mainly based on the completing data of up and down cycle of general value investment, while generally speaking, the 5 - 10 years is the interval of bull and bear markets. From the historical data, we can also conclude that in the year of 2013-2018, our country's stock market experiences a complete up and down cycle, therefore, using the interval as the research one can be better to get rid of the bigger error that bull market or bear market can bring purely in the market fluctuation, and better reflects the long-term development trends of this stock.

2) Getting rid of missing tuple of data

This paper uses the recent five-year data of Yunnan Baiyao (000538) from Yahoo Finance, Yunnan Baiyao Group Co., Ltd. has been suspended twice during these five years. From July 19, 2016 to Dec. 30, 2016, due to planning of promotion of mixed ownership reform related stocks into suspension period. Meanwhile, from April 25, 2017 to April 26, 2017, due to all unrestricted shares of Yunnan Baiyao fully takeover offer suspension, in order to ensure that data's fluctuation and development trends are not affected by the change during the suspension period, this paper remove the data tuple of this period, to ensure that the data of this period is more representative.

4.2. Date Pre-Processing

4.2.1. Reduction of Dimension of Main Components Analysis

This paper reduces the dimension of 29 stocks evaluation indicator through PCA by SPSS, the principal components are extracted as input variables of BP neural network stock price forecasting model. The contribution rate of previous five components reaches 88.915%, but in order to further increase information coverage rate, this paper extracts the sixth component, the variance of extracted six main components at the end accounts for 92.179% of principal component variance, enough to describe the fluctuation of Yunnan Baiyao in the latest five years.

4.2.2. Data Normalization

This paper uses normalization function of *mapminmax* in MATLAB to normalize 6 principal components and one closing price to prevent network output error from being large due to difference of data magnitude or dimension.

$$x_p = \frac{x_p - x_{\min}}{x_{\max} - x_{\min}} \tag{28}$$

Among them, $x_{\min}$ is the minimum value of data, $x_{\max}$ is the maximum of data.

4.3. The Determination of Model Parameter

4.3.1. Hidden Layer and Node of Hidden Layer

In accordance with empirical formula, the approximate range of number in hidden layer neurons is between 2 and 13. Through trial and error to compare prediction accuracy of different node numbers, the number of hidden layer nodes is finally determined shown in Table 2.

Experiments show that the MSE of BP neural network will decrease first and then rise with the increase of the number of nodes. By comparing the MSE and MAE of different implied nodes, it is found that the network error is the smallest when the number of hidden nodes is 6 in this case. Therefore, the BP neural network stock price prediction model established in this paper determines that the number of hidden layers is 2 and the number of hidden layer nodes is 6.

4.3.2. The Number of Neurons of Input Layer and Output Layer, and Learning Rate

The input variables are the six principal components after the dimension reduction of the PCA, and the output variable is the stock price of the next day.

Based on the empirical formula determined by the learning rate mentioned above, the initial learning rate is finally determined to be 0.4082.

4.3.3. Activation Function

According to the three activation functions mentioned above, and applied to the hidden layer and the output layer respectively, comparing the different activation functions through multiple experiments leads to the MAE of the prediction result to determine the selected activation function shown in Table 3.

Table 2. MSE and MAE corresponding to the number of different hidden layer nodes.

the number of hidden layer nodes	2	3	4	5	6	7
MSE	8.02	7.96	6.92	7.66	6.48	6.85
MAE	2.24	2.21	2.04	2.10	1.97	2.08
the number of hidden layer nodes	8	9	10	11	12	13
MSE	7.08	6.73	6.83	7.09	7.98	8.16
MAE	2.06	2.00	2.07	2.05	2.31	2.23

Table 3. MAE corresponding to different activation functions of the hidden layer and the output layer.

hidden layer\the output layer	logsig	tansig	purelin
logsig	9.78	2.97	1.98
tansig	9.41	5.45	2.42
purelin	16.79	2.44	2.18

According to the results, the paper finally selects *logsig* as the hidden layer activation function, *purelin* as the output layer activation function with picking the smallest error.

4.3.4. Training Function

According to the six training functions commonly used in MATLAB mentioned above, comparing the different training functions through multiple experiments leads to the comparison of the MAE and MSE of the prediction results to determine the selected training function shown in Table 4.

After experimental testing, this paper finally selects *traincgb* as a training function with picking the smallest error.

4.4. Results Analysis

This paper mainly evaluates the prediction results according to MSE (Mean Square Error) and MAE (Mean Absolute Error).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(O_i - Y_i\right)^2 \tag{29}$$

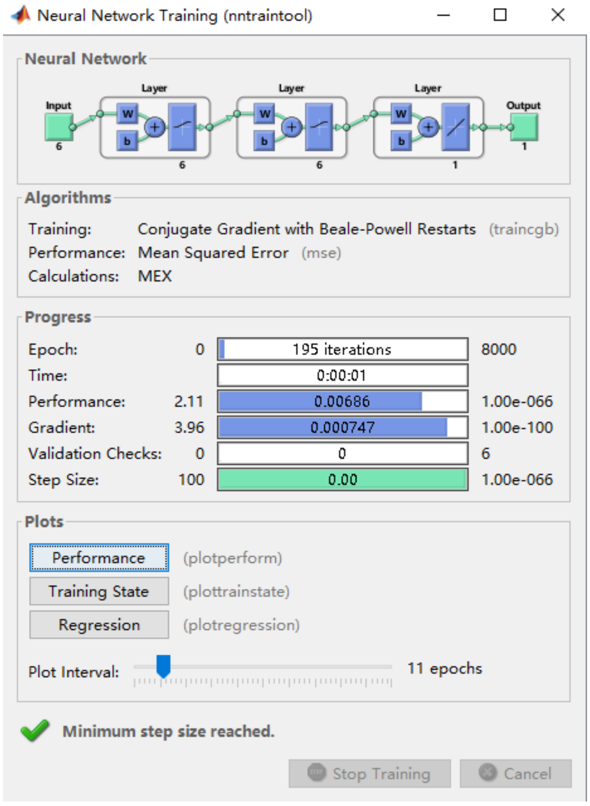
$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|O_i - Y_i\right| \tag{30}$$

Among them, Y_i represents the actual output of the model; O_i represents the expected output of the model. The smaller the MSE, the higher the accuracy of the model.

After the parameters of the model are determined, the data is put into in MATLAB and the model is trained. The operating results show that the model converges after 195 iterations, and the test error is close to the target setting accuracy of 1.0e−6. The MSE is 6.63 and the MAE is 1.98, which has a good fitting effect. The BP neural network stock price forecasting model structure, BP neural

Table 4. Comparison of 6 training functions MAE and MSE.

training functions	traingd	traingdm	traingda	trainbfg	trainlm	traincgb
MAE	2.45	2.46	2.37	2.75	2.80	1.98
MSE	9.79	9.912	9.02	13.49	13.10	6.63



Neural Network Training (nntraintool)

Algorithms

Training: Conjugate Gradient with Beale-Powell Restarts (traincgb)

Performance: Mean Squared Error (mse)

Calculations: MEX

Progress

Epoch:	0	195 iterations	8000
Time:		0:00:01	
Performance:	2.11	0.00686	1.00e-066
Gradient:	3.96	0.000747	1.00e-100
Validation Checks:	0	0	6
Step Size:	100	0.00	1.00e-066

Plots

Plot Interval: 11 epochs

✓ Minimum step size reached.

Figure 1. BP neural network stock price forecasting model structure.

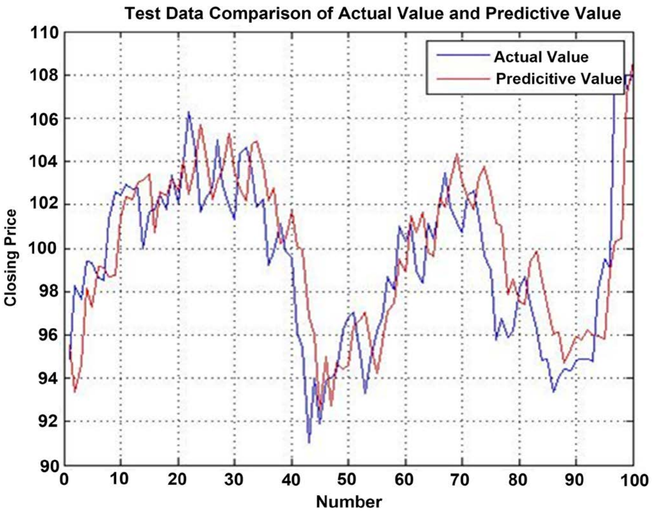


Figure 2. BP neural network stock price prediction model output and actual output ccomparison.

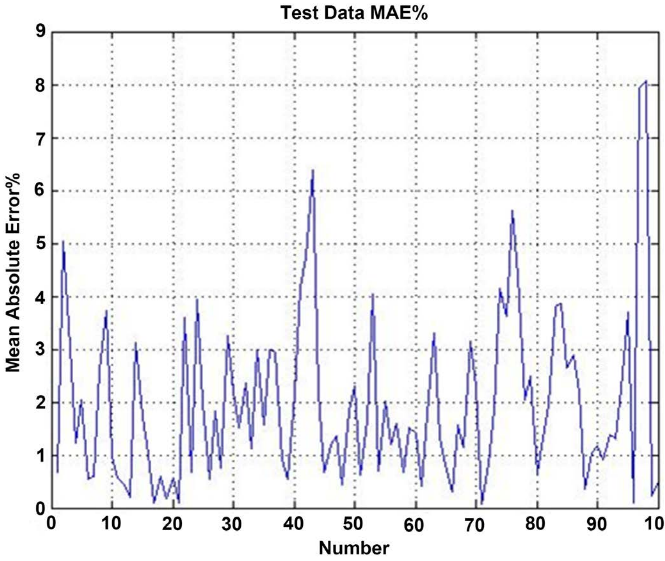


Figure 3. BP neural network stock price prediction model MAE.

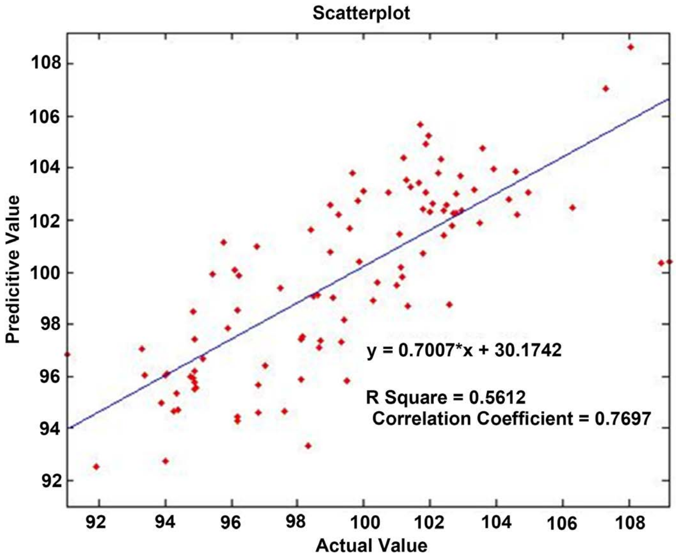


Figure 4. BP neural network stock price prediction model scatter plot.

network stock price prediction model output and actual output comparison, BP neural network stock price prediction model MAE and BP neural network stock price prediction model scatter plot are respectively shown in Figures 1-4.

5. Conclusion

In this paper, the stock price forecasting model is established based on BP neural network. From the many factors affecting the stock price, the opening price, the closing price, the highest price, the lowest price, the trading volume, the turnover, the price change, the MA, MACD, KDJ, BIAS, RSI, ROC, OBV, WR, a total of 29 indicators are selected. PCA was used to reduce the 29 indicators to a 6-dimensional input network. Taking the data of Yunnan Baiyao (000538) for the past 5 years as an example, the data experiment is carried out. After experimentally comparing the MSE and MAE of different parameters, the number of hidden layer nodes in the model are finally determined as 6, the initial learning rate is 0.4082, the activation functions of the hidden layer and the output layer are respectively *logsig* and *purelin* and training function is *traincgb*, and finally the model is established. After testing, the model has good precision, MSE is 6.63 and the MAE is 1.98.

Although the test results in this case are good, there are still some problems, and further improvements and in-depth research can be carried out in the future.

1) The selected indicators are not comprehensive

This paper only selects 29 indicators from many stock price evaluation indicators, but in fact, there are far more than 29 factors affecting stock prices. In the follow-up study, the fundamental indicators of the technical surface can be further explored.

2) The data is not rich enough

In this paper, only 1105 groups of data of Yunnan Baiyao (000538) stocks from May 13, 2013 to May 10, 2018 are selected. If the data volume is further increased, two or even more alternate regions of the bull and bear market selected as the sample size can further improve model prediction accuracy.

3) Further optimize model parameters

In this paper, the existing activation function and training function are selected. If the parameters of the model are further optimized in the subsequent research, the prediction accuracy of the model can be further improved.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Bao, Y. L. (2013). *Research on the Analysis and Predictive Algorithm of Financial Time Series Based on the Support Vector Machine*. Dalian: Dalian Maritime University.

Bollerslev, T., Cai, J., & Song, F. M. (2000). Intraday Periodicity, Long Memory Volatility, and Macroeconomic Announcement Effects in the US Treasury Bond Market. *Journal of Empirical Finance, 7,* 37-55. https://doi.org/10.1016/S0927-5398(00)00002-5

Chen, J. Y. (2014). *The Application of Data Mining in the Stock Analysis.* Guangzhou: South China University of Technology.

Dose, C., & Cincotti, S. (2005). Clustering of Financial Time Series with Application to Index and Enhanced Index Tracking Portfolio. *Physical A: Statistical Mechanics and Its Applications, 355,* 145-151. https://doi.org/10.1016/j.physa.2005.02.078

Du, X. P. (2006). *The Securities Situation Assessment System Based on Data Mining.* Tianjin: Tianjin University.

Fiol-Roig, G., Miro-Julia, M., & Isern-Deya, A. P. (2010). Applying Data Mining Techniques to Stock Market Analysis. In Y. Demazeau, et al. (eds.), *Trends in Practical Applications of Agents and Multiagent Systems. Advances in Intelligent and Soft Computing* (Vol. 71, pp. 519-527). Springer, Berlin, Heidelberg.

https://doi.org/10.1007/978-3-642-12433-4_61

Golan, R. H., & Ziarko, W. (1995). A Methodology for Stock Market Analysis Utilizing Rough Set Theory. *Proceedings of 1995 Conference on Computational Intelligence for Financial Engineering (CIFEr), New York, NY.* *https://doi.org/10.1109/CIFER.1995.495230*

Guo, S. H. (2010). Stock Analysis Simulation System Based on Apriori Algorithm. *Computer Simulation, 27,* 334-337.

Hao, Z. Y. (2017). *A Stock Prediction System Based on Data Mining.* Nanjing: Nanjing University of Science and Technology.

Kumar, L., Pandey, A., Srivastava, S., et al. (2011). A Hybrid Machine Learning System for Stock Market Forecasting. *Proceedings of World Academy of Science Engineering & Technology,* 315-318.

Li, Y. Q., & Song, W. (2013). Prediction of Stock Price Trend Based on BP Neural Network. *Journal of North China University of Technology, 25,* 11-16.

Liu, N. K., & Lee, K. K. (1997). An Intelligent Business Advisor System for Stock Investment. *Expert Systems, 14,* 129-139. https://doi.org/10.1111/1468-0394.00049

Ma, C. Q., Lan, Q. J., & Chen, W. M. (2007). *Financial Data Mining.* Beijing: Science Press.

Panigrahi, S. S., & Mantri, J. K. (2015). Epsilon-SVR and Decision Tree for Stock Market Forecasting. *2015 International Conference on Green Computing and Internet of Things (ICGCIoT),* Noida. *https://doi.org/10.1109/ICGCIoT.2015.7380565*

Qiao, J. W. (2013). *Application of BP Neural Network in Stock Investment Analysis.* Chengdu: University of Electronic Science and Technology of China.

Refenes, A. N., & Latif, E. A. (2015). Assessing the Quality of Service Using Big Data Analytics: With Application to Healthcare. *Big Data Research, 4,* 13-24.

Srikant, R., & Agrawal, R. (2010). Mining Quantitative Association Rules in Large Relational Tables. *Proceedings of the ACM SIGMOD Conference on Management of Data,* Montreal, 1-12.

Sun, L. P. (2013). *The Application and Research of Data Mining in Stock Analysis.* Chengdu: Southwestern University of Finance and Economics.

Ticknor, J. L. (2013). A Bayesian Regularized Artificial Neural Network for Stock Market Forecasting. *Expert Systems with Applications, 40,* 5501-5506. https://doi.org/10.1016/j.eswa.2013.04.013

Tsaih, R., Hsu, Y., & Lai, C. C. (1998). Forecasting S & P 500 Stock Futures with a Hybrid AI System. *Decision Support System, 23,* 161-174. https://doi.org/10.1016/S0167-9236(98)00028-1

Wang, B. (2010). *Data Mining Concepts and Techniques.* Beijing: Mechanical Industry Press.

Yang, T. (2017). *Stock Ranking Based on Machine Learning.* Tianjin: Tianjin Polytechnic University.

Yuan, C. A. (2009). Principles of Data Mining and SPSS Clementine Application Collection. *Publishing House of Electronics Industry.* 16-23, 176-187, 311-355.

Zhang, J. J. (2010). *Analysis Stocks Based on the Data Mining.* Beijing: China University of Petroleum.

Zhang, J. N. (2009). *Data Mining and Application.* Beijing: Peking University Press.

Zhang, W. T., & Dong, W. (2013). *SPSS Statistical Analysis Advanced Course.* Beijing: Higher Education Press.