Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/108/2025.LD21292

Application of Principal Component Analysis and BP Neural Network Algorithm in Stock Price Prediction

Xianjun Chen[1,a,*]

[1]*Dalian University of Technology, No. 2 Linggong Road, Ganjingzi District, Dalian City, Liaoning Province, China*
a. 924206487@qq.com
**corresponding author*

Abstract: Stock price fluctuations are influenced by numerous interrelated factors. Traditional stock price prediction models, including neural networks, often fail to account for these correlations effectively, leading to lower prediction accuracy. To improve prediction performance, this paper integrates Principal Component Analysis (PCA) with a Backpropagation (BP) neural network, proposing a dynamic PCA-BP model for stock price forecasting. Simulation experiments conducted on the stock prices of selected listed enterprises demonstrate that the PCA-BP model exhibits varying performance across different experimental groups. The findings indicate that while the combined model enhances prediction accuracy, its generalization capability requires further optimization for practical applications.

Keywords: Principal Component Analysis, BP Neural Network, Stock Price Prediction

1. Introduction

With the rapid development of China's economy, the stock market has become increasingly significant in resource allocation. Its issuance and trading have greatly contributed to the growth of the real economy. As a high-risk, high-return investment method, stocks have attracted a large number of investors. However, their inherent uncertainty can also lead to substantial financial losses. Therefore, accurately predicting stock prices can help mitigate investment risks, enhance investors' decision-making efficiency, and guide them in making informed and rational investment choices.

This paper proposes a combined model integrating Principal Component Analysis (PCA) and a Backpropagation (BP) neural network. By applying PCA, redundant information among the influencing factors of stock price prediction can be reduced, the dimensionality of the BP neural network's input data can be decreased, and both the training speed and prediction accuracy of the combined model can be improved.

Neural networks exhibit strong characteristics such as self-learning, self-adaptation, and nonlinear approximation, making them highly suitable for predicting stock price trends. However, stock price fluctuations are influenced by numerous nonlinear factors that are highly correlated. Directly applying neural networks to stock price prediction can lead to an excessive number of training iterations and reduced prediction accuracy.

To address this issue, PCA is first applied to the influencing factors of stock prices to reduce dimensionality and eliminate correlations among variables. The extracted principal components are

DOI: 10.54254/2755-2721/108/2025.LD21292

then used to train the BP neural network model, which is subsequently employed to predict stock prices.

2. Background

Domestic scholars have conducted extensive research on quantitative analysis using Principal Component Analysis (PCA) and Backpropagation (BP) neural network algorithms.

In 2008, Yang Jinhui et al. [1] applied PCA and a self-organizing network to analyze financial data from 47 listed real estate companies in 2005. They selected 13 financial indicators reflecting company performance and reduced them to five principal components based on profitability, solvency, asset management capability, growth potential, and equity expansion ability. The total variance contribution rate reached 80%, effectively capturing the actual data patterns.

In the same year, Zhang Jigang and Liang Na [2] proposed a prediction model combining the Self-Organizing Map (SOM) network, PCA, and BP neural network to study real-time stock market closing price predictions. They used 150 datasets from Sinopec (600028) spanning November 2006 to April 2007 as sample data. The prediction factors included price changes, opening prices, trading volumes, transaction amounts, and highest and lowest prices. Based on the classification results of SOM neural network training, the sample data were divided into six groups. One group was selected for empirical analysis, and PCA was applied. The cumulative contribution rate of the first three eigenvalues was found to be 99.41%, which were then used as input variables for the BP neural network model. After just 15 training iterations, the model achieved a minimum error of 0.0001, effectively eliminating redundant information and improving both prediction accuracy and efficiency.

In 2011, Zhao Shian [3] utilized PCA and a support vector machine regression model to forecast stock market trends using 355 trading days of Shanghai Composite Index data (March 12, 2008 – August 19, 2009). The next 30 trading days (August 20 – September 30, 2009) were used for model validation. The Mean Absolute Percentage Error (MAPE) fitting value was 2.10, while the prediction value was 5.34. Additionally, the Prediction Mean Squared Error (PMSE) fitting value was 65.4293, and the prediction value was 16.7899. The results demonstrated that the PCA-SVM regression model outperformed the simple weighted average integration model in both fitting and prediction accuracy.

In 2016, Hu Zhaoyue and Bai Yanping [4] applied the PCA-SVM neural network model to stock price prediction. They analyzed 102 trading days of OTex (002227) stock data from August 25, 2015, to January 25, 2016, using 12 technical indicators as input variables, including: Today's highest price, lowest price, opening price, and closing price; Today's trading volume, and moving averages over 5, 10, 30, and 60 days; KDJ.K, KDJ.D, and KDJ.J indicators.

After PCA dimensionality reduction, the variables were reduced to five principal components, and the next day's closing price was set as the output variable. The Mean Squared Error (MSE) for the PCA-SVM algorithm was 0.00082, compared to 0.00224 for the SVM algorithm alone. The PCA-SVM model achieved superior results in terms of convergence speed, error reduction, and prediction accuracy. The study concluded that this model effectively shortens training time, ensures prediction accuracy, and performs well in short-term stock price forecasting, making it valuable for practical applications.

In 2018, Liu Jiaqi et al. [5] introduced the PCA-GA-BP model, integrating PCA, Genetic Algorithm (GA), and the BP neural network to enhance stock price predictions. This approach addressed the slow computation speed and susceptibility to local minima inherent in traditional BP neural networks while overcoming the limitations of conventional stock price prediction models.

In 2019, Lu Tianyu et al. [6] conducted an empirical study on weekly stock data from listed enterprises in the scientific research and technical service sector using the RESSET financial database (March 10, 2017 – March 30, 2018). They applied PCA and BP neural networks to reduce 36 indicators to eight principal components. The model was trained using PCA-processed scores of these

DOI: 10.54254/2755-2721/108/2025.LD21292

eight factors and stock prices. The error accuracy was close to 0.001, demonstrating an ideal prediction performance. The study concluded that the PCA-BP model achieves fast training speed and high prediction accuracy, providing valuable insights for stock price forecasting.

3. Methodology

Principal Component Analysis (PCA), first introduced by Pearson in 1901 and further developed by Hotelling in 1933, is a statistical technique designed to reduce the dimensionality of high-dimensional data while preserving as much information as possible. It achieves this by identifying a small number of orthogonal vectors that effectively represent the key characteristics of a dataset containing multiple variables. By minimizing information loss, PCA eliminates redundant information, improving computational efficiency and model performance. The PCA process follows these steps:

(1) Standardize the sample values: Given a dataset $\mathrm{X} = (\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_\mathrm{n})$ compute its covariance matrix: $\mathrm{D(x)} = \mathrm{B} = (r_{ij})_{n\times n}$;

(2) Compute the eigenvalues $r_1, r_2, \ldots, r_n$ of $\mathrm{D(x)}$ and their corresponding eigenvectors $\beta_1, \beta_2, \ldots, \beta_n$;

(3) Select principal components: Calculate the contribution rate of the i-th principal component to the total variance: $t_i = \delta_i \Big/ \sum_{j=1}^{n} \delta_i$. Rank them in descending order to determine the first principal component, second principal component, and so on. Choose k principal components such that: $\sum_{i=1}^{k} t_i > C$ (where C is typically within the range (0.80, 0.95)). (4) Compute the principal component transformation matrix P_c and establish the principal component equation: The eigenvectors corresponding to the eigenvalues of the k selected principal components form the principal component transformation matrix: $P_c = (\beta_1, \beta_2, \cdots, \beta_k)^T$. Transform the n-dimensional vector X into a k-dimensional vector X' using the equation: $\mathrm{X}' = \mathrm{P_c} \times \mathrm{X}$.

The Backpropagation (BP) neural network is a multi-layer feedforward neural network trained using the error backpropagation algorithm. It can learn and store complex input-output mapping relationships without requiring prior knowledge of their underlying mathematical equations. The BP network topology consists of: An input layer; One or more hidden layers; An output layer. A typical three-layer BP neural network structure is illustrated in Figure 1.

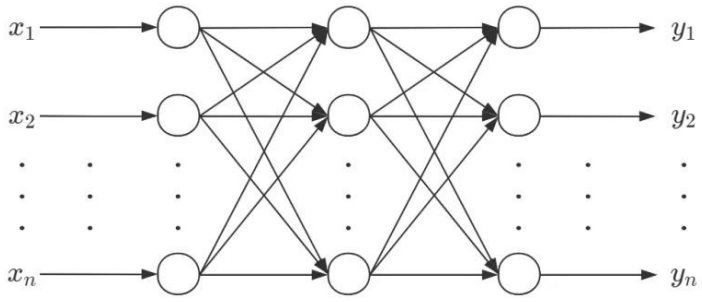

Figure 1: Three-Layer BP Neural Network Topology

The PCA-BP model integrates Principal Component Analysis with the BP neural network, forming a hybrid predictive model. By applying PCA to reduce the dimensionality of the original dataset, the number of input variables in the BP neural network is reduced, which significantly: Enhances the training speed; Reduces the correlation between attributes; Improves prediction accuracy.

Steps for Stock Price Prediction Using the PCA-BP Model:

(1) Standardize the data.

(2) Perform PCA on the dataset to extract factor scores for each principal component.

DOI: 10.54254/2755-2721/108/2025.LD21292

(3) Select appropriate principal components and use their scores as input variables for the BP neural network.

(4) Construct a three-layer BP neural network, set relevant parameters, and train the model.

(5) Adjust the parameters and retrain the model multiple times until optimal results are achieved.

(6) Use the trained PCA-BP model to predict stock prices.

4. Results

In this study, empirical research was conducted using Shanghai Composite Index data for three stocks—SHA600028, SHA600031, and SHA600585—from the Shanghai Stock Exchange (SSE) database covering the period January 10, 2022, to December 23, 2022. A total of 698 data points were selected for analysis.

During the experiment, four different network configurations were constructed to compare prediction results:

(1) BP-12 Network: Select 12 financial indicators as input variables, including: Return on common shareholders' equity; Free cash flow amount; Earnings per share; Year-on-year growth rate of net profit attributable to shareholders of the parent company in a single quarter; Price-earnings ratio over the past 12 months; Market capitalization; Average household shareholding ratio; Returns in the last five trading days; Return rank score in the last five trading days; Amount rank score in the last five trading days; Trading amount; Opening price. The closing price was used as the output variable. The BP neural network (BPNN) was directly trained and tested without preprocessing.

(2) PCA-BP-12 Network: Based on the 12 indicators in (1), Principal Component Analysis (PCA) was applied. The top four principal components were extracted based on their contribution rates. These four comprehensive variables were used as input variables for the BP neural network, forming a PCA-enhanced model.

(3) BP-8 Network: Select 8 financial indicators as input variables: Price-earnings ratio over the past 12 months; Market capitalization; Average household shareholding ratio; Returns in the last five trading days; Return rank score in the last five trading days; Amount rank score in the last five trading days; Trading amount; Opening price. The closing price was used as the output variable. The BP neural network was directly trained and tested without preprocessing.

(4) PCA-BP-8 Network: Based on the eight indicators in (3), PCA was performed. The top four principal components were extracted based on their contribution rates. These four variables were used as input variables for the BP neural network, forming another PCA-enhanced model. These four network configurations are referred to as BP-12, PCA-BP-12, BP-8, and PCA-BP-8, respectively.

Comparison of Network Training Efficiency:

BP-12 Network: For the BP-12 network, an optimal neural network structure was selected based on experimental comparison. The best-performing configuration included: Two hidden layers; 16 nodes per hidden layer; Step size of 0.2; 1,000 training iterations. Since the 12 indicators were directly used without preprocessing, the training convergence results are illustrated in the following figures:

Figure 2: Training convergence of BP-12 network (SHA600028)

DOI: 10.54254/2755-2721/108/2025.LD21292

Figure 3: Training convergence of BP-12 network (SHA600031)

Figure 4: Training convergence of BP-12 network (SHA600585)

PCA-BP-12 Network: For the PCA-BP-12 network, PCA was applied to the 12 indicators, transforming them into four principal components that retained the majority of the information. These were used as input variables for the BP neural network. The training convergence results are presented in:

Figure 5: Training convergence of PCA-BP-12 network (SHA600028)

Figure 6: Training convergence of PCA-BP-12 network (SHA600031)

DOI: 10.54254/2755-2721/108/2025.LD21292

Figure 7: Training convergence of PCA-BP-12 network (SHA6000585)

BP-8 Network: For the BP-8 network, the same neural network structure as in the BP-12 configuration was used. The eight selected indicators were directly fed into the BP neural network without preprocessing. The training convergence results are shown in:

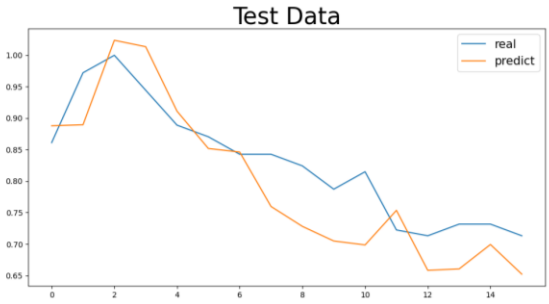

Figure 8: Training convergence of BP-8 network (SHA600028)

Figure 9: Training convergence of BP-8 network (SHA600031)

Figure 10: Training convergence of BP-8 network (SHA600585)

DOI: 10.54254/2755-2721/108/2025.LD21292

PCA-BP-8 Network: For the PCA-BP-8 network, PCA was applied to the eight selected indicators, transforming them into four principal components that served as input variables for the BP neural network. The training convergence results are displayed in:

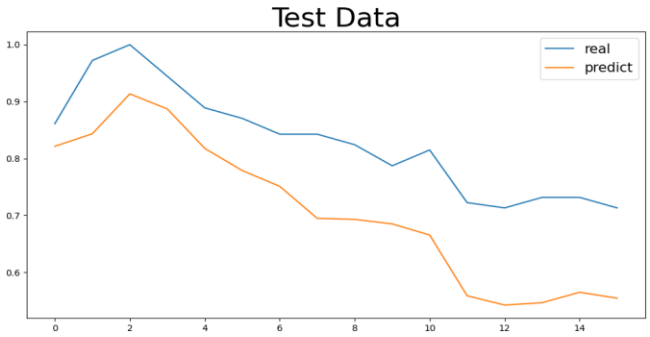

Figure 11: Training convergence of PCA-BP-8 network (SHA600028)

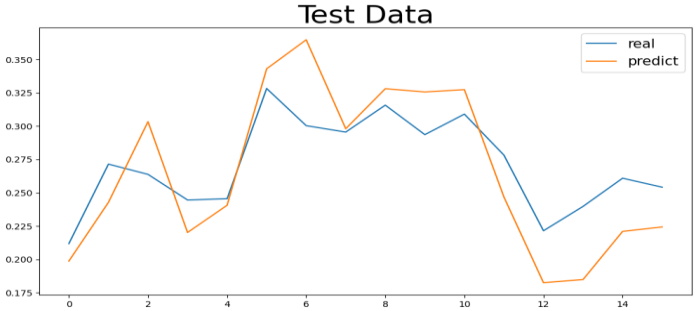

Figure 12: Training convergence of PCA-BP-8 network (SHA600031)

Figure 13: Training convergence of PCA-BP-8 network (SHA600585)

Table 1: Comparison of Prediction Accuracy Across the Four Networks

	BP-12			PCA-BP-12		
	SHA600028	SHA600031	SHA600585	SHA600028	SHA600031	SHA600585
MAE	0.0410	0.0813	0.0511	0.0711	0.0900	0.0531
MSE	0.0025	0.0076	0.0036	0.0065	0.0085	0.0033

DOI: 10.54254/2755-2721/108/2025.LD21292

Table 1: (continued).

	BP-8			PCA-BP-8		
	SHA600028	SHA600031	SHA600585	SHA600028	SHA600031	SHA600585
MAE	0.0547	0.0280	0.0617	0.1214	0.0282	0.0499
MSE	0.0040	0.0010	0.0042	0.0166	0.0011	0.0032

A comparative analysis of the prediction accuracy across the four network models is summarized in Table 1. This table provides a quantitative evaluation of the performance differences among BP-12, PCA-BP-12, BP-8, and PCA-BP-8 networks, highlighting the impact of dimensionality reduction via PCA on stock price prediction accuracy.

5. Discussion

Table 1 presents a comparison of the prediction accuracy across the four network models. The results indicate significant differences in performance across the three selected stocks.

For SHA600028, the BP neural network without data preprocessing outperformed the PCA-BP network, suggesting that the dimensionality reduction process via PCA did not enhance prediction accuracy. Additionally, the prediction accuracy using 12 indicators as input was generally higher than that using 8 indicators.

For SHA600031, there was no substantial difference in prediction accuracy between the BP network without preprocessing and the PCA-BP network. However, the accuracy of the model trained on 12 indicators was significantly lower than that of the model trained on 8 indicators.

For SHA600585, the prediction accuracy of the BP network without preprocessing was comparable to that of the PCA-BP network. Similarly, there was no significant difference between using 12 indicators and 8 indicators as input variables.

Overall, when PCA was applied before training the neural network, the prediction accuracy tended to be slightly lower than that of the BP network trained directly on raw data. This suggests that dimensionality reduction via PCA may lead to a loss of critical information for certain stocks, potentially affecting prediction performance.

The difference between using 12 indicators and 8 indicators is primarily attributable to four additional indicators— Return on common shareholders' equity; Free cash flow amount; Earnings per share; Year-on-year growth rate of net profit attributable to shareholders of the parent company in a single quarter.

These four indicators remained unchanged over the same time period, potentially influencing the PCA transformation process and the final prediction results. Future research should further investigate the impact of these indicators on PCA's effectiveness in stock price prediction.

6. Conclusion

In summary, this study explores stock price prediction using a PCA-BP neural network model. Stock prices are influenced by numerous factors that exhibit a certain degree of correlation. Direct application of neural networks to stock price prediction can lead to increased training iterations and reduced prediction accuracy. To address this issue, Principal Component Analysis (PCA) is applied to reduce dimensionality and restructure the input data for the BP neural network model. By retaining the maximum amount of relevant information, PCA effectively reduces input dimensionality, simplifies the network structure, accelerates training speed, and enhances prediction accuracy.

DOI: 10.54254/2755-2721/108/2025.LD21292

However, this study also has certain limitations, highlighting areas for further research. The analysis was conducted on a limited subset of stock data, and broader investigations are required to derive more comprehensive insights. Additionally, the observed variations in results across different stock types suggest that further experiments and model optimizations are necessary to validate and refine our approach. Future research could explore alternative feature selection techniques, hybrid modeling approaches, and expanded datasets to enhance the robustness and generalizability of the PCA-BP neural network model.

References

[1] *Yang, J. H., Zhao, J., Ma, T. Y., & Sun, Y. F. (2008). Comprehensive performance evaluation of listed companies based on PCA-SOM. Journal of Jilin University: Information Science Edition, 26(2), 7.*

[2] *Zhang, J. G., & Liang, N. (2008). Stock price prediction based on SOM network-principal component analysis-BP network. Statistics and Decision, (6), 3.*

[3] *Zhao, S. A. (2011). Application of PCA support vector machine regression ensemble in stock market prediction. Journal of Baise University, 24(3), 5.*

[4] *Hu, Z. Y., & Bai, Y. P. (2016). Stock price prediction based on PCA-SVM combined model. Commerce, (2), 1.*

[5] *Liu, J. Q., Liu, D. H., & Lin, T. T. (2018). A brief study on stock prices based on BP neural network model. China Business Review, (8), 2.*

[6] *Lu, T. Y., Du, L. N., Wang, H. Y., Wang, Y. Q., Tao, M. W., & Zhang, X. W. (2019). Prediction of stock price trend based on principal component analysis and neural network combination. Computer Knowledge and Technology: Academic Edition, (2X), 4.*