Research on Crop Prediction Optimization Based on BP Neural Network and Particle Swarm Algorithm

Min Cai [1], Yifan Ding [2], Rongyu Qiao [3], Xue Kang [4, *]

[1] School of Water Conservaney, North China University of Water Resources and Electric Power, Zhengzhou, China, 450046

[2] School of Foreign Studies, North China University of Water Resources and Electric Power, Zhengzhou, China, 450046

[3] School of Art, North China University of Water Resources and Electric Power, Zhengzhou, China, 450046

[4] School of Educational Science, Xinyang Normal University, XinYang, China, 464000

* Corresponding Author Email: 15836577065@163.com

Abstract. Accurate crop yield prediction is crucial for optimizing agricultural production and ensuring food security. However, traditional methods are limited by their simplicity and inability to cope with complex environments and climate changes, leading to insufficient prediction accuracy. This study integrates crop planting data from rural areas of a website for the year 2023, employing a BP Neural Network-PSO Hybrid Model. By conducting correlation analysis on existing crop data from cultivated land and combining the maximization of profit and minimization of costs, the optimal Crop Planting Strategy is determined. In the prediction process, the sigmoid function is cautiously selected as the activation function, and key parameters required for pre-adjustment in machine learning are set to ensure the model's stability and convergence speed. The training target error is set to 0.00001 to pursue higher fitting accuracy. After validation, the MSE Mean Squared Error Value is obtained as 0.085, which sufficiently indicates the model's high rationality and prediction accuracy. When importing data for prediction, the maximum total profit over seven years is calculated to be 4,135,919,090.20 CNY. By comparing the fit degrees of the values before and after, it is clear to observe an upward trend in crops. Finally, Sensitivity Analysis is used to test the stability of the model after achieving the optimal solution.

Keywords: BP Neural Network-PSO Hybrid Model, Crop Planting Strategy, MSE Mean Squared Error Value, Sensitivity Analysis.

1. Introduction

With the advancement of global agricultural technology, the precise prediction of crop yields has become an important issue in agronomic research and agricultural production. Scientific crop yield prediction can not only help farmers make reasonable planting decisions and optimize the allocation of production resources but also provide a scientific basis for government food security policies. However, traditional prediction methods are often too simple or have single parameters, making it difficult to effectively cope with the complex agricultural production environment and the impact of climate change, leading to low accuracy and stability in prediction results.

In recent years, artificial intelligence and machine learning technologies have shown strong advantages in the field of crop yield prediction. To address the optimization of crop planting strategies and the prediction of profit in agricultural product sales and distribution, this study uses a BP Neural Network-PSO hybrid algorithm. By comparing the crop yields under different planting strategies, a scientific planting strategy optimization plan is provided for farmers. This plan can recommend the best planting varieties, planting times, and fertilization plans for farmers according to the natural conditions of different regions and market demands, thereby increasing crop yields and quality, increasing farmers' economic benefits, and integrating historical crop planting and sales data to provide predictions for the maximization of profit under the optimal planting strategy in the future. This algorithm has a wide range of applications, with studies applying the BP Neural Network-PSO hybrid improved hybrid algorithm to the classification of surface defects in cold-rolled steel strips [1], and another study predicting the short-term variation trend of ρ(Chla) in Minghu based on the BP Neural Network-PSO hybrid improved hybrid algorithm [2].

Specifically, the study first constructs a crop prediction model based on the BP Neural Network through algorithms and ideas such as gradient descent algorithm, loss function, back propagation algorithm, and decision boundary. Then, based on the projection pursuit method, the unit projection vectors of each indicator are calculated as weights, and the Particle Swarm Algorithm is used to optimize it, resulting in a BP Neural Network crop prediction model based on Particle Swarm Optimization [3]. By adjusting the appropriate optimization model parameters, the optimal prediction network under the condition of path optimization is trained. Data is imported for prediction, and the optimal planting strategy is derived, with the actual result MSE value being 0.085, indicating that the model is rational and has high prediction accuracy. By comparing the fit of the values before and after, it is concluded that the crop shows an upward trend. Finally, sensitivity analysis is used to test the stability of the model.

2. Data Normality and Correlation Testing

Data for this study comes from https://www.mcm.edu.cn/.

2.1. Sample Data Multiprocessing

In response to the issues encountered in the study, it is necessary to carefully consider each dimension of the provided data to expand the possibilities for more comprehensive analysis. However, due to the large amount of text in the existing data set and the large sample size, the original data may have missing data, data anomalies, and other issues during collection or transmission, leading to a decline in data quality. In such cases, if these poor-quality data are used directly for modeling, the results will not be convincing [4]. Therefore, it is considered necessary to briefly process the original data before subsequent modeling analysis.

To meet the requirements of data analysis, first, the data format of the original data set is unified. Among them, outliers in the collected sample data are removed as follows:

Using the 3σ criterion (Laid's criterion), the variables are processed with equal precision to obtain $x1, x2, \ldots, xn$, and calculate their arithmetic mean x and the remaining error $vi = xi - x\ (i = 1,2 \ldots n)$, and then calculate the standard error σ according to the Bessel formula. If the remaining error vb of a measurement value $xb\ (1 <= b <= n)$ satisfies $vb| = |xb - x| > 3\sigma$ then xb is considered a bad value containing gross error and should be removed. The Bessel formula is as follows:

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n} v_i^2\right]^{1/2} = \left\{\left(\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right) / n / n-1\right\} \tag{1}$$

Secondly, missing value processing is carried out. The methods for handling missing data usually include: deleting feature variables, deleting samples, and data imputation. In this study, after removing outliers, the data is re-integrated and sorted to form a data set for the research object, and the missing data is filled with mean imputation for further data analysis and research. The specific formula is as follows:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2}$$

2.2. Test of normality

Normality tests are conducted on the sample data to observe whether the results are close to a normal distribution, thereby determining which correlation calculation method to use next [5]. The specific test results are shown in Figure 1, the normal Q-Q plot of expected yield, Figure 2, the normal Q-Q plot of planting cost/(CNY/acre), and Figure 3, the normal Q-Q plot of sales unit price (CNY/jin):

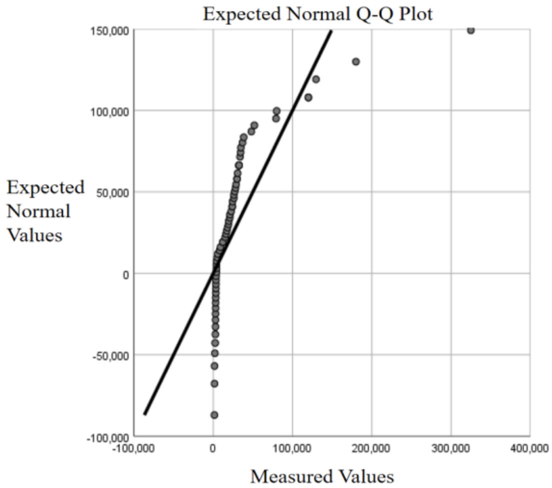

Fig. 1 Normal Q-Q plot of expected yield

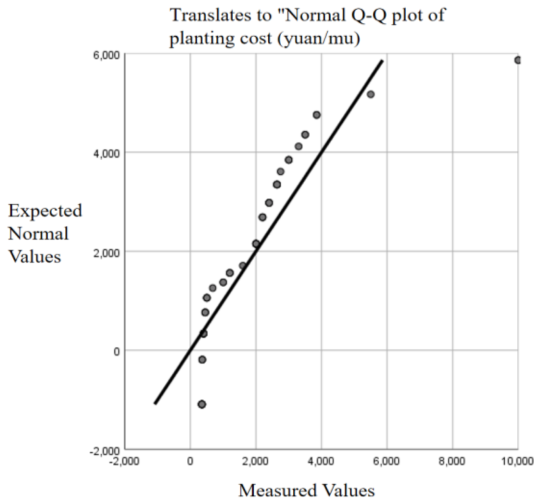

Fig. 2 Normal Q-Q plot of planting cost/(CNY/acre)

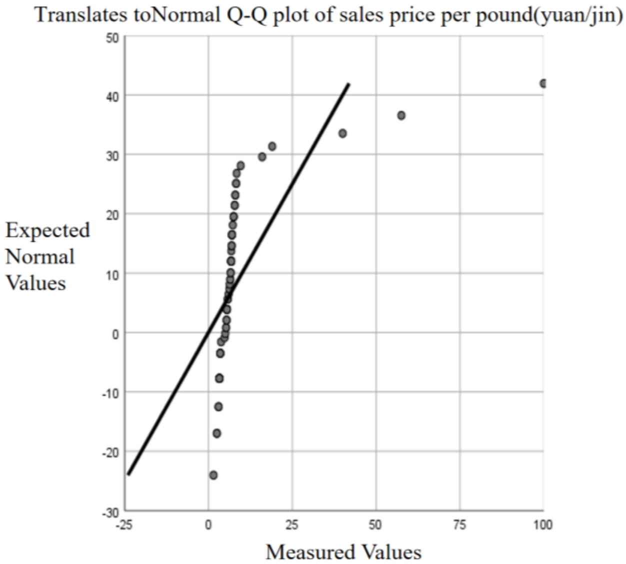

Fig. 3 Normal Q-Q chart of sales unit price (CNY/jin)

Observing the Q-Q plot, it is found that most sample points are approximately distributed near a straight line, so it is judged that the sample data is approximately normally distributed.

2.3. Spearman Correlation Test

According to the research analysis requirements, it is necessary to first conduct a correlation test between expected sales volume and sales price, planting cost. The Spearman correlation test is performed using SPSSPRO, and the specific principal formula is as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{3}$$

Where d_i is the rank difference between X_i and Y_i. (The rank of a number is the position of the number after sorting the numbers in its column from small to large) It can be proven that r_s is between -1 and 1 [6].

The following figure shows the correlation coefficient values in the form of a heat-map, mainly representing the size of the values by the depth of the color. At the same time, it can be judged whether there is a correlation between the two variables by whether the significant P-value in the image is less than 0.05. The results of the Spearman correlation test are shown in the following figures: Figure 4, the correlation heat-map between planting cost and sales unit price, Figure 5, the correlation heat-map between expected production and sales unit price, and Figure 6, the correlation heat-map between expected yield and planting cost:

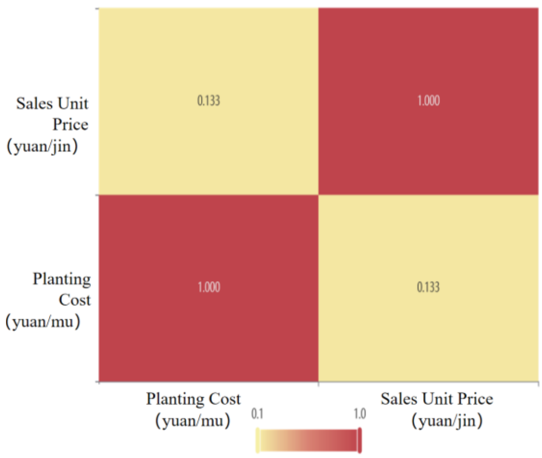

Fig. 4 Heat map of the correlation between planting cost and sales unit price

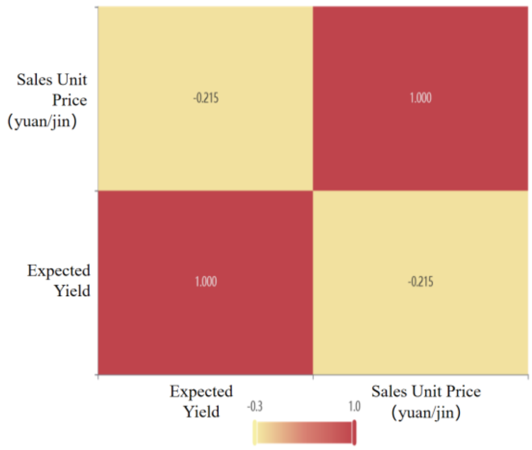

Fig. 5 Heat map of the correlation between expected production and sales unit price

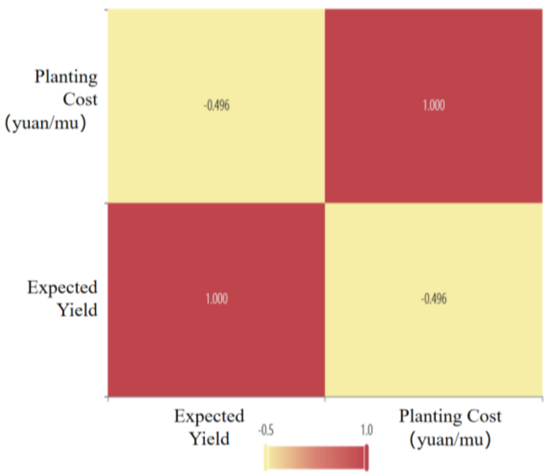

Fig. 6 Heat map of the correlation between expected yield and planting cost

Analysis of the results from the figure shows that there is no significant correlation between sales unit price and planting cost, while there is a significant correlation between sales unit price and expected yield, and between planting cost and expected yield.According to experience, the sales unit price is dominated by market supply and demand, and there is no direct connection with the planting cost; the fixed and variable parts of the planting cost and management efficiency affect the relationship between the cost and the expected output; Meanwhile, the expected production is affected by many factors and the difference between actual and expected, thus affect the sales unit price; finally, the market competition factors such as product differentiation and brand image will affect the formulation of sales unit price.

3. Coupled Optimization of BP Neural Network and Particle Swarm Algorithm

3.1. BP Neural Network Optimization Strategy Model

The BP Neural Network is a multi-layer feedforward neural network trained according to the error back propagation algorithm, which is one of the most widely used neural network models. It consists of an input layer, hidden layers (which can have multiple layers), and an output layer. Each layer contains multiple neurons, and neurons are interconnected through connections with weights. The error after output is used to estimate the error of the direct preceding layer of the output layer, and then this error is used to estimate the error of the layer before it, and so on, back propagating layer by layer to obtain the error estimates of all other layers [7].

This study uses the BP Neural Network for crop prediction on cultivated land. By establishing a neural network, the model can use self-learning and self-adaptive characteristics to establish a corresponding relationship, and then use this relationship to train, validate, simulate, and finally achieve the result through the model.

3.2. Particle Swarm Algorithm Optimization

Because BP neural network optimization model is easy to fall into the trap of local optimal solution, so increase the use of particle swarm algorithm to implement further optimization model, through the adaptation particle swarm algorithm projection tracing method, calculate the index of unit projection vector, as the weight improve model performance, to jump out of the dilemma of local optimal solution [8]. The specific implementation steps are shown as follows:

a. Establish the projection function

Let the j-th indicator of the i-th sample be $yij\ (i = 1,2 \ldots n, j = 1,2 \ldots m)$; where n is the number of samples, m is the number of indicators. If (a1…am) is an m-dimensional projection vector, then the expression for the projection characteristic value Zi of sample i in one-dimensional linear space is:

$$Z_i = \sum_{j=1}^{m} a_i y_{ij} \tag{4}$$

b. Establish the objective function

The problem that the study needs to solve is an optimization problem, so the established objective function is carried out on the basis of continuously approaching the global optimal goal.

The objective function equation of the projection pursuit method is:

$$Q(a) = S(a) * d(a) \tag{5}$$

$S(a)$ is the sample standard deviation of z, that is:

$$S(a) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2} \tag{6}$$

Where z is the mean of z, and the larger $S(a)$, the more dispersed the values. $d(a)$ is the local density of the projection value z; that is:

$$d(a) = \sum_{i=1}^{n} \sum_{i=1}^{n} (R - rij) \cdot I(R - rij) \tag{7}$$

r is the distance between projection characteristic values $rij = |2i - z|$, and its value also indicates the degree of dispersion. R is the density window width parameter, which is related to the structure of the sample data. The analysis of R shows that its value range is $rmax < R \leq 2m$, and the general value is:

$$R = r_{max} + \frac{m}{2} \tag{8}$$

Among them , $r_{max} = \max(r_{max})$.

$I(R - rij)$ is the unit step function, and its value satisfies the equation:

$$I(R - r_{ij}) = \begin{cases} 1 & R - r_{ij} > 0 \\ 0 & R - r_{ij} \le 0 \end{cases} \tag{9}$$

c. Where is the unit step function, and its value satisfies the equation

The objective function Q (a) will change according to the change of the projection vector a. The appropriate projection vectors a may show the characteristic structure of high-dimensional data, so the corresponding projection vector a [9] can be obtained by solving the maximum value of the projection target functione

d. Establish the PSO optimization solution model

From Figure 7, the Particle Swarm Optimization Algorithm Process Solution Diagram, the solution process steps are as follows:

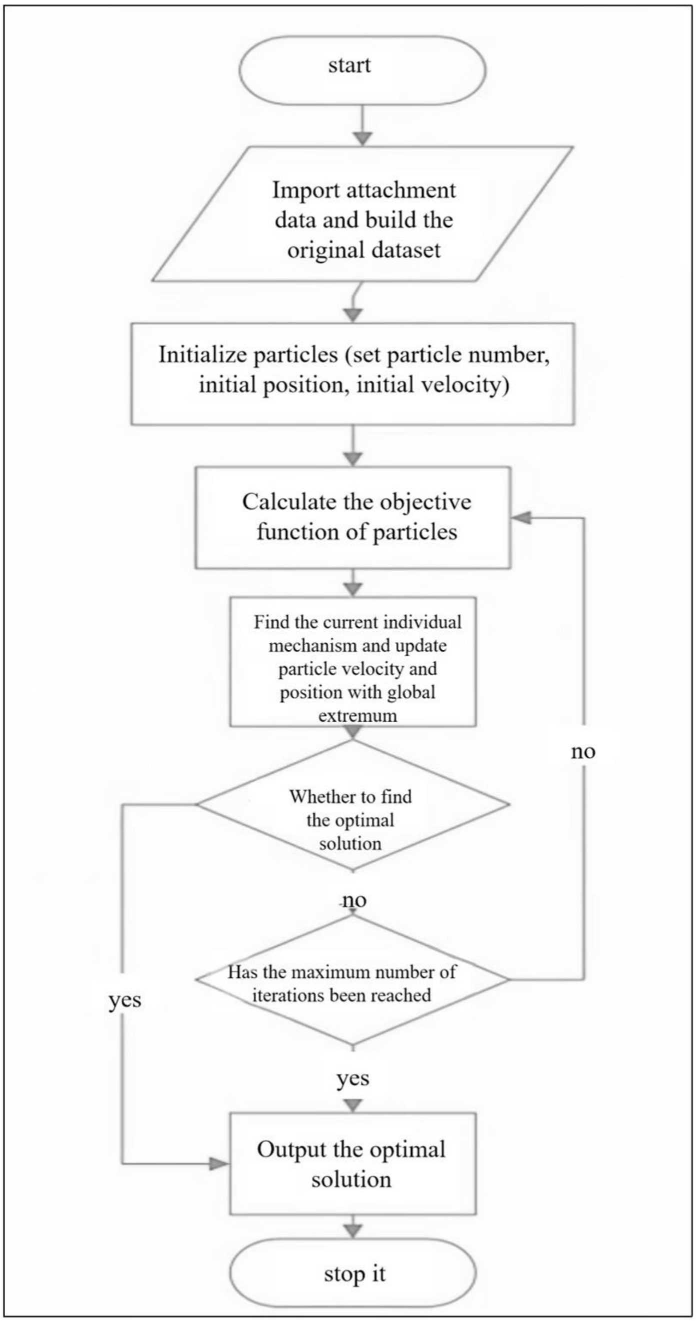

Fig. 7 Particle Swarm Optimization Algorithm Process Solution Diagram

3.3. BP Neural Network-PSO Model Prediction Solution

The improved model is applied, and the model is programmed in MATLAB. After importing the sample data for iterative calculation, the algorithm flow is shown in Figure 8, the Neural Network Simulation Iteration Curve:

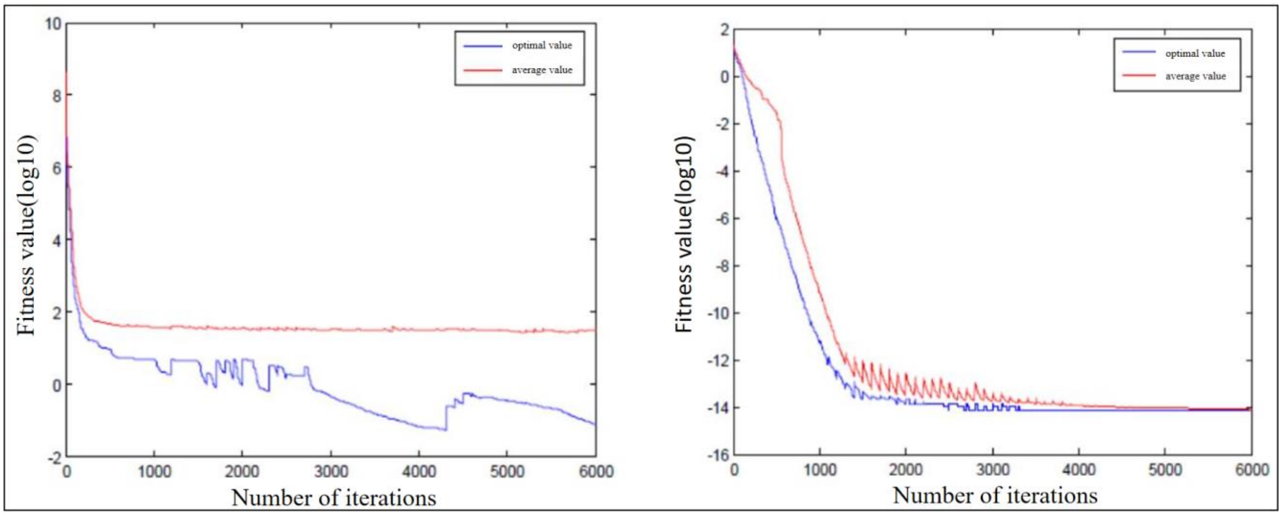

Fig. 8 Neural Network Simulation Iteration Curve

The BP Neural Network parameters set for this problem are as follows in Table 1 Specific Parameter Settings:

Table 1. Specific Parameter Settings

Maximum number of steps	Training interval	Learning rate	objective function
1000	25	0.01	0.00001

After data processing, 80% of the data is randomly divided into the training set, and 20% of the data is used as the test set. The following Figure 9 shows the Analysis of Goodness of Fit between True and Predicted Values after feature engineering:

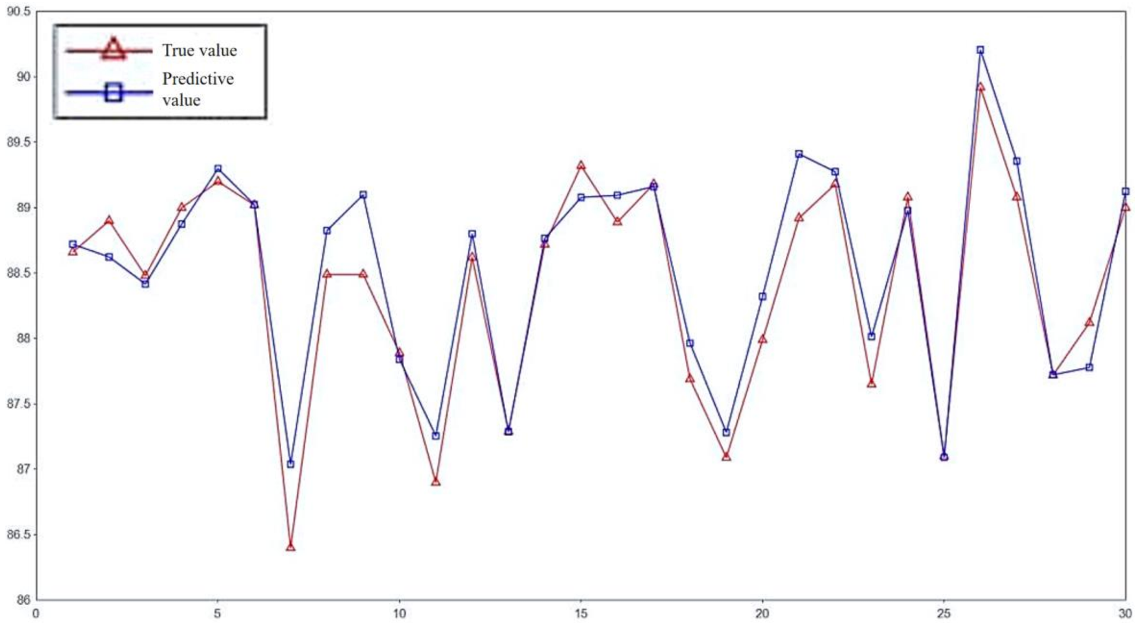

Fig. 9 Analysis of goodness of fit between true and predicted values

This study uses the mean squared error to judge the model performance evaluation index. The formula is as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 \tag{10}$$

Where n is the number of data samples, is the predicted value, and is the ideal value. The smaller the MSE value, the smaller the mean squared error of the model, and the higher the prediction accuracy [10]. The calculated MSE value is 0.085. By comparing the fit of the true and predicted values, Figure 10 shows the upward trend of crops after changes.

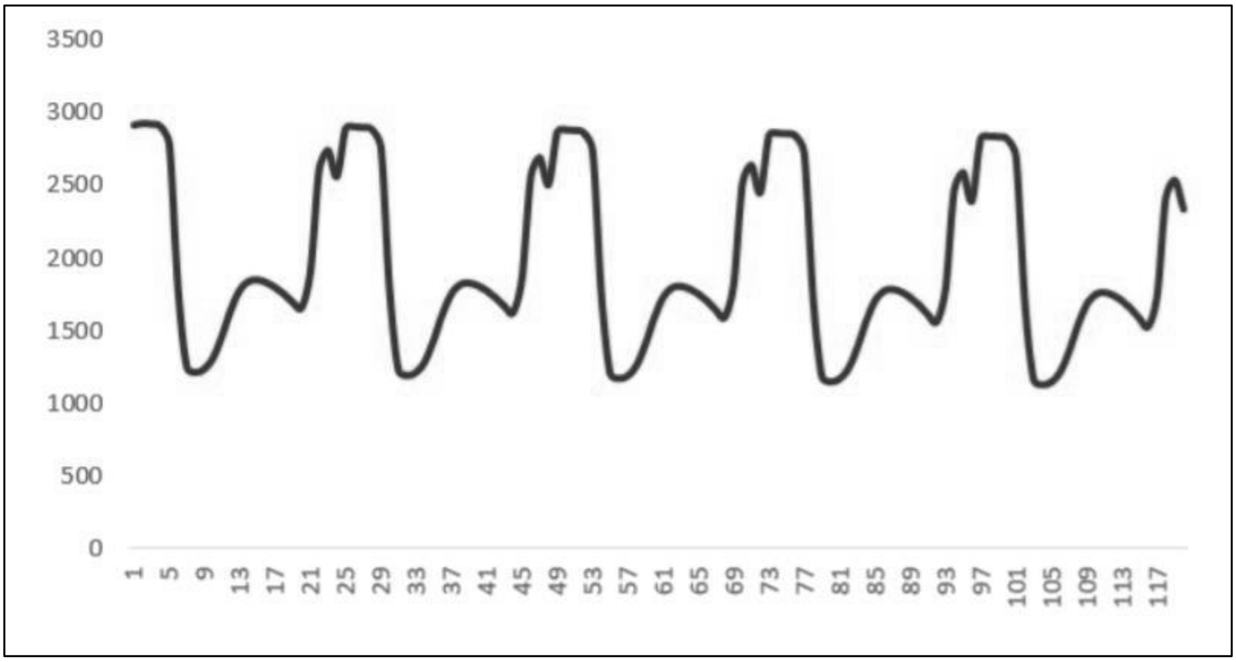

Fig. 10 The upward trend of crops after changes

3.4. Model Analysis and Testing

Finally, in the sensitivity analysis, the BP Neural Network cultivation strategy prediction model is tested by changing the learning rate parameter, and the specific analysis results are shown in Figure 11, the Schematic Diagram of Sensitivity Analysis Results:

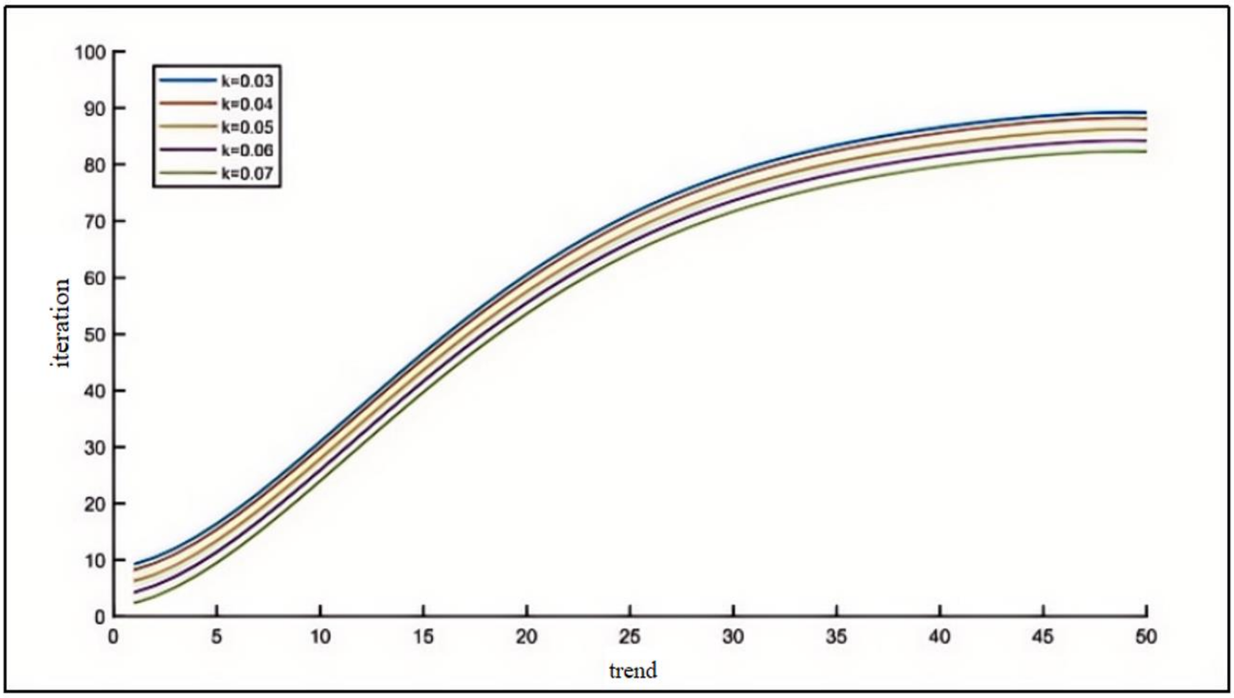

Fig. 11 Schematic diagram of sensitivity analysis results

From the above sensitivity analysis results diagram, it is known that the model trend obtained from the sensitivity test is consistent with the actual situation, and the trend after changing the learning rate parameter is consistent, proving the rationality and robustness of the BP Neural Network cultivation strategy model.

4. Conclusions

This paper has validated the application potential of the proposed research methods and frameworks in the field of agricultural informatics through a series of experiments and case studies. The experimental data clearly show the advantages of this method in improving data processing speed and enhancing the model's understanding of agricultural data. The results show that compared with traditional control methods, this method reduces the MSE mean squared error to 0.085, increases the total profit to 6.52%, and reduces the total cost loss to 6.93%, with the final solution achieving a maximum total profit of 4,135,919,090.20 CNY over seven years. When engaging in agricultural production, it is necessary to establish an appropriate optimization prediction model by analyzing historical data. Furthermore, this study's case further demonstrates the application of the framework in solving actual agricultural problems, such as crop yield prediction, pest and disease detection, and optimization of agricultural resource allocation. These application examples fully prove the practical application value and broad development space of this research method and framework in the field of agricultural informatics.

References

[1] Guo Huale. A hybrid algorithm of BP neural network and PSO and its application [D]. Liaoning: Northeastern University, 2011.

[2] Yu Yingjie, Jiang Weigang, Xu Mingfang. Prediction of chlorophyll a in water bodies based on PSO algorithm of BP neural network [J]. Environmental Science Research, 2011, 24 (5): 526-532.

[3] Shao Meng, Pan Zhengzhong, Sun Jinwei, et al. Short-term tidal current speed prediction based on EMD-PSO-BP model [J]. Journal of Ocean University of China (Natural Science Edition), 2024, 54 (11): 134-141.

[4] Zhu Xiao, Shen Xiaojing. Data preprocessing based on machine learning using MATLAB [J]. Science and Technology Information, 2024, 22 (17): 19-22.

[5] Zhou Hongwei. Several commonly used methods for normality testing [J]. Journal of Nanjing Xiaozhuang College, 2012 (3): 13-18.

[6] Li Xiaoshuang, Wang Qing, CNY Meng, et al. Correlation analysis between meteorological elements and eggplant price based on SROCC [J]. Sichuan Agricultural Science and Technology, 2024, (08): 134-138.

[7] Tian Yuhui. Research on prediction method of construction cost of large buildings based on BP neural network [J]. China Building Metal Structure, 2024, 23 (10): 16-18.

[8] Wang Xin, Zhang Fuqun, Zhang Jinping. Research on mixed gas classification and recognition method based on principal component analysis-particle swarm optimization algorithm-support vector machine [J]. Occupational Health and Emergency Rescue, 2024, 42 (05): 633-638+689.

[9] Zhao Xiaoyong, Yu Jing, Zhang Shuang, et al. Study on water resource carrying capacity based on improved projection pursuit method [J]. People's Yellow River, 2024, 46 (5): 74-79.

[10] Gao Yunxia, Shi Yisong. Collaborative design and implementation of WMS with SAP, MES, and SRM [J]. Logistics Technology and Application, 2024, 29 (7): 134-137.

[11] Hu Yu, Wang Yufei. Optimal scheduling strategy for mobile energy storage in active distribution networks based on global sensitivity analysis [J/OL]. Modern Electric Power, 1-12 [2024-10-30].