Applied Research on AQI Prediction Based on BP Neural Network Modeling

Zhixuan Liu[1,*], Jiayan Lin[1], Lan Zhou[1], Yupeng Song[1], Xin Li[2]

[1]School of Computer and Communication Engineering, Nanjing Tech University Pujiang Institute, Nanjing, 211200, China
[2]School of Mechanical and Electrical Engineering, Nanjing Tech University Pujiang Institute, Nanjing, 211200, China
[*]Corresponding author: freestarrr071319@gmail.com

Abstract: *In recent years, air environment quality has become a hot issue of concern for people all over the world, and the prediction of air quality is of great significance for air pollution prevention and control. There is mainly a nonlinear relationship between air quality data and influencing factors, and BP neural network has a strong nonlinear mapping ability, which can fit the more complex nonlinear mapping relationship. Based on this, this paper utilizes BP neural networks to establish an air quality index AQI prediction model to predict the AQI in Nanjing, with an average relative error of about 1% and a prediction accuracy of 99%. The establishment of this model can provide reliable reference and decision-making basis for government departments and citizens.*

Keywords: *BP Neural Networks, Air Quality, AQI Prediction*

1. Introduction

With the rapid development of industrialization and urbanization, a large number of pollutant emissions have serious impacts on the environment and public health. In today's society, the quality of air has attracted widespread attention, and therefore, accurate assessment and prediction of air quality has become a central concern of current scholars. Air Quality Index (AQI), as a widely used air quality assessment index, can comprehensively reflect the concentration levels of major pollutants in the air and the potential impact on human health. Through the prediction of AQI, we can know the air quality condition in the future time period in advance, which provides an important reference basis for environmental protection and public health management.

The data of air quality is usually nonlinear, and traditional forecasting methods such as regression analysis and time series analysis in statistical models have limited modeling ability for nonlinear relations, while BP neural network has stronger nonlinear fitting ability, automatic feature extraction ability, adaptability and generalization ability, and parallel computing ability. This enables it to better capture nonlinear relationships and patterns when dealing with complex prediction problems, improve the accuracy and stability of prediction, especially for the processing of multi-classification and high-dimensional data problems, BP neural network has obvious advantages.

In air quality prediction, the BP neural network model can extract useful features from complex meteorological and pollutant data, establish an accurate prediction model, and be able to accurately predict the future AQI, thus helping decision makers to take appropriate environmental protection measures and reduce the risk of public exposure to polluted environments.

In this paper, a BP neural network model will be established to predict the air quality. The specific steps are shown in Figure 1. From the accuracy of the verification set data, the model can more accurately analyze the changing trend and influencing factors of air quality conditions, and provide reliable reference and decision-making basis for government departments and citizens.

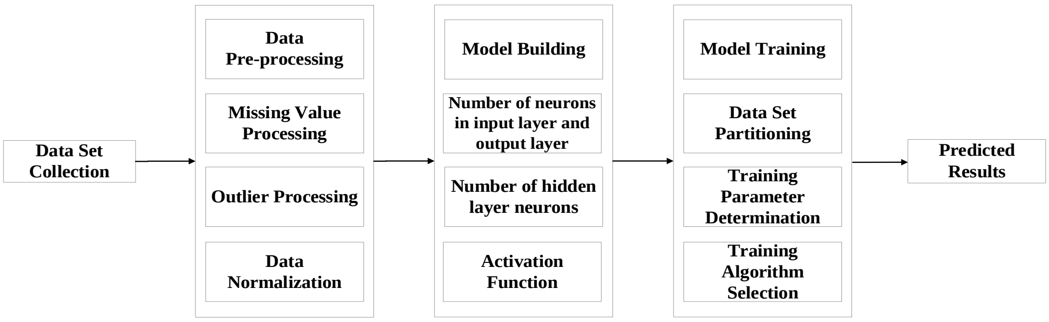

Figure 1: BP neural network prediction process

2. Related Work

In recent years, there have been many scholars who have explored the air quality prediction research in depth. For example, Tong [1] et al. (2019) established a BP neural network model to predict the concentration of $PM_{2.5}$ in Beijing, and Huijing Wu [2] et al. (2018) used a genetic algorithm to carry out a prediction study on the AQI index in Xuchang City, China. Ahmad Najim Ali [3] et al. used polynomial logistic regression to predict the air quality of New York City in the United States. The above studies used air pollutant concentration data or meteorological factors as data sets and all of them achieved some prediction results.

3. Data Sets and Preprocessing

3.1 Data Collection

The data set of this experiment is from the open data set of CSDN website, and 72,792 pieces of daily air quality data of Nanjing, Jiangsu Province are selected.

Table 1: Pollutants and AQI raw data (partial)

CO	NO_2	O_3	PM_{10}	$PM_{2.5}$	SO_2	AQI
0.77	46	18	79	59	9	80
0.69	39	25	84	63	9	85
0.6	34	31	73	57	10	74
0.54	28	40	64	54	11	74

In Table 1, the units of the data are mg/m^3 except for CO, which is in $\mu g/m^3$.

3.2 Data Pre-processing

3.2.1 Detection and Handling of Missing Values and Outliers

Missing values refer to the values of certain attributes in the data that have not been recorded or captured, which may be caused by equipment failure, data collection errors or human omissions. In order to ensure the completeness and accuracy of the data and avoid the influence on the subsequent modeling and prediction results, the mean interpolation method will be adopted here to fill in the missing values.

Outliers are observations that are significantly different from other observations, which may be caused by measurement errors, data entry errors, or real special circumstances. In order to avoid model bias and inaccuracy caused by outliers, the detected outliers are directly deleted in this paper, and since the data set used in this experiment is relatively large, direct deletion will not have too much impact on the integrity of the experimental data.

3.2.2 Data Normalization

The original data without normalization will have scale data, and directly using such data with different scales and large differences for weighting will affect the convergence speed and prediction effect of the neural network. Therefore, in order to improve the effect of model training and the accuracy of prediction, all data need to be normalized. In this paper, the maximum-minimum normalization method of [4] is used to normalize the data.

$$y = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

4. The Establishment of BP Neural Network Model

4.1 The Principle and Basic Structure of BP Neural Network

Back propagation(BP) neural network is a common type of artificial neural network used to supervise learning tasks. Based on the error backpropagation algorithm, it trains the model by the given input and output values, and constantly adjusts the weight and bias to optimize the network model so that it can better approximate the objective function.

BP neural network usually consists of an input layer, a hidden layer (there can be multiple hidden layers), and an output layer. Each layer is made up of multiple neurons, also known as nodes. The connections between neurons are represented by weights, and each neuron has a biased value. The structure diagram of BP neural network is shown in Figure 2.

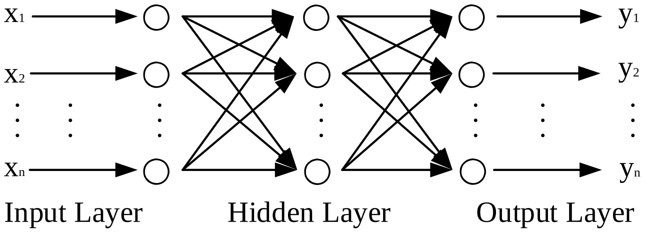

Figure 2: BP neural network structure diagram

In Figure 3, in order to realize the BP neural network algorithm, it is necessary to first set the initial weight of the neural network, select its activation function, and then determine the number of nodes of the input layer, hidden layer and output layer, and set the number of hidden layers. After selecting the predicted training algorithm mode, and then putting the input vector and the expected target value into the neural network, the model training can begin. The data will pass through the input layer, the hidden layer, and finally reach the output layer and output, and then compare the error between the expected target value and the final output value. If it appears within the set error range, the training is over. If it is outside the set error, continue training. The error is reduced by constantly updating the threshold of the weight until a predetermined stopping condition is reached (for example, the maximum number of iterations is reached or the error is less than a certain threshold), and the result of the network is finally output.

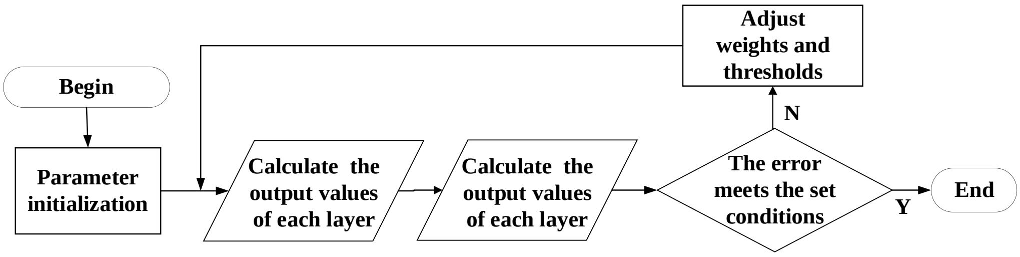
Adjust weights and thresholds
Begin
N
Parameter initialization
Calculate the output values of each layer
Calculate the output values of each layer
The error meets the set conditions
Y
End

Figure 3: BP neural network algorithm flow chart

4.2 Network Topology Design

The number of nodes in the input layer of BP neural network is closely related to the dimension of the input data. The model of this experiment mainly predicts the future air quality index through the historical pollutant data, so the number of input layers is 6 and the number of output layers is 1.

For the hidden layer, there is no mature theoretical method to determine its number, but the number of neurons in the hidden layer can be initially determined by empirical formula, and then the neural network with different numbers of neurons can be trained, and finally the optimal number of neurons can be selected according to the comparison results of error and R value [5]. The empirical formula adopted in this paper is as follows:

$$l = \sqrt{m+n} + p \tag{2}$$

Where l is the number of neurons in the hidden layer, m is the number of neurons in the input layer, n is the number of neurons in the output layer, p is a constant and $0 < p < 10$.

After conducting several experiments by enumeration method according to equation 2, the number of hidden layer neuron nodes is finally determined to be 10.

4.3 The Setting of Model Training Parameters

In this paper, the Sigmoid function in the S-type function is selected as the activation function [6], and the formula is as follows:

$$sigmoid(x) = \frac{1}{1+e^{-x}} \tag{3}$$

72792 pieces of data were divided into training sets and test sets in a ratio of 7:3. The maximum number of iterations of the neural network is set to 1 000, the learning rate is 0.05, and the target error is 0.0005.

In order to solve the bad phenomena such as slow convergence speed and overfitting of BP algorithm, we will choose Levenberg-Marquardt algorithm to train the network during model construction.

5. Result Analysis

This experiment uses matlab2023 as the experimental platform, and the processor is Intel(R) Core(TM) i3-10110U CPU @ 2.10GHz.

5.1 Result evaluation index

5.1.1 Regression R-value

R-squared in regression analysis is also known as the coefficient of determination, which is an important index to evaluate the quality of regression models. The R value ranges from 0 to 1, which reflects the interpretation degree of independent variable to dependent variable variation. The closer R value is to 1, the higher the degree of independent variable interpretation variation according to variable, and the better the model fitting effect. In general, an R-value greater than 0.5 is considered a moderate correlation in social science research, and an R-value greater than 0.8 indicates a strong linear relationship between the independent and dependent variables. In this experiment, the size of R value will be used as the evaluation standard of the degree of fitting.

5.1.2 Absolute Error and Relative Error

Absolute Error is the absolute value of the difference between the predicted value and the true value, which intuitively reflects the actual amount of deviation between the predicted value and the true value. The calculation formula is as follows:

$$A = \left| X - Y \right| \tag{4}$$

Where A represents the absolute error value, X represents the predicted value, and Y represents the true value.

Relative Error is the ratio of the absolute value of the difference between the predicted value and the true value to the true value, which reflects the proportion of the relative error to the true value, and eliminates the impact of different dimensions. The calculation formula is as follows:

$$R = \left| \frac{X - Y}{Y} \right| \tag{5}$$

Where R represents the relative error value, X represents the predicted value, and Y represents the true value.

This experiment will take the results of these two as the reference standard for the accuracy of prediction results.

5.2 Analysis of Experimental Results

Using the network and model training parameters determined above, the neural network model is obtained through training.

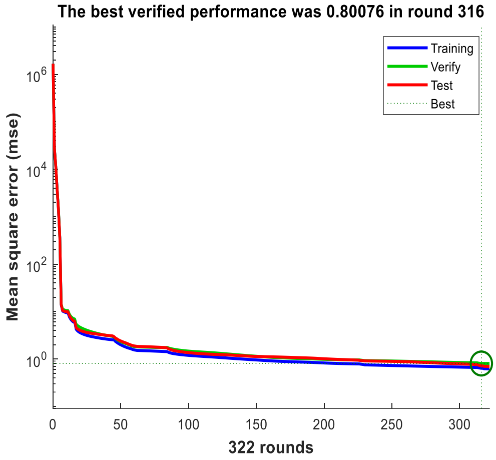

Figure 4: Error variation diagram

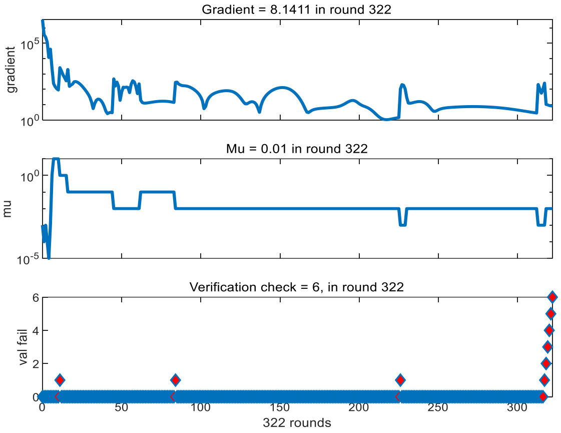

Figure 5: Gradient and step change diagram

In Figure 4, the training process stops after 322 times, and the final error is 0.80076. The process of error reduction can be clearly observed from the figure, and the error continues to decrease with the increase of the number of iterations, and the best performance is reached at the 316th round, and the training is terminated. Figure 5 can reflect the overall change of gradient and step length during training.

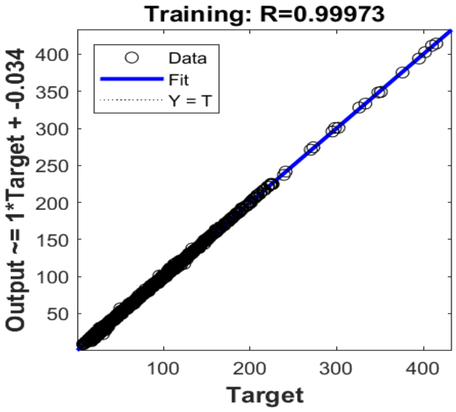

Figure 6: Regression fitting graph

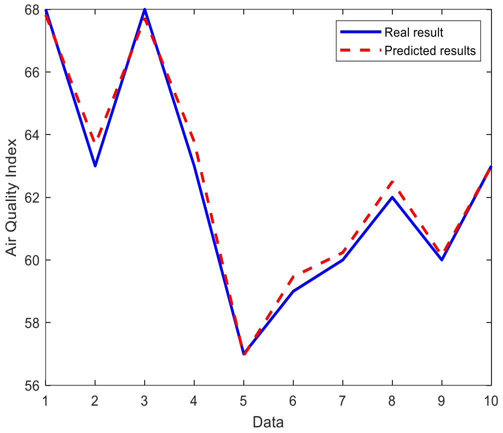

Figure 7: Contrast chart

Figure 6 is the fitting effect diagram of the training results of the BP neural network model. According to this figure, the fitting degree R value can be obtained as 0.99973, so it can be judged that the model obtained by this training has high accuracy and strong applicability.

This paper uses the trained model to forecast the AQI of Nanjing, Jiangsu Province for 10 days from May 22 to May 31, 2020, and draws a comparison curve between the predicted value and the real value, as shown in Figure 7. From the figure, we can see that the blue line and the orange line are basically consistent, indicating that the predicted value and the actual value have a high degree of fitting. A good prediction effect has been obtained.

In order to further verify the accuracy of the prediction results, this paper sorted out the table of absolute and relative errors between the prediction results and the true value, in table 2, the error between the AQI result predicted by the trained BP neural network model and the actual data is very small, and the relative error of most of them is kept within 1%, which indicates that the prediction effect is relatively ideal.

Table 2: AQI's forecast results

Data	Predicted value	True value	Absolute error	Relative error (%)
2020/05/22	67.823	68	0.171	0.259%
2020/05/23	63.687	63	0.687	1.092%
2020/05/24	67.718	68	0.282	0.415%
2020/05/25	63.776	63	0.776	1.232%
2020/05/26	56.951	57	0.048	0.085%
2020/05/27	59.464	59	0.464	0.787%
2020/05/28	60.238	60	0.238	0.397%
2020/05/29	62.502	62	0.502	0.810%
2020/05/30	60.138	60	0.138	0.231%
2020/05/31	62.997	63	0.002	0.004%

6. Conclusion

This paper takes the air quality of Nanjing as the research object, builds a BP neural network model by collecting the air pollutant data of Nanjing, trains it with the collected data, and uses the trained model to predict the AQI index of Nanjing, and takes the absolute error and relative error of the forecast result as the standard to measure the accuracy. The experimental results show that the model constructed in this paper has good fitting effect and high prediction accuracy. In the future work, we will further analyze the factors affecting the prediction results of BP neural network model, so as to improve the accuracy and stability of the prediction.

References

[1] Xue Tonglai, Zhao Donghui, Han Fei. Prediction of PM2. 5 Concentration in Beijing Based on BP Neural Network [J]. The Journal of New Industrialization, 2019, vol. 9, no. 8, pp. 87-91.

[2] Wu Huijing, He Xiaohui. Research on the Prediction of Air Quality Index Based on GA-BP Neural Network [J]. Journal of Anhui Normal University (Natural Science), 2019, vol. 42, no. 04, pp. 360-365.

[3] Ali N A, Nassreddine G, Younis J. Air Quality prediction using Multinomial Logistic Regression [J]. Journal of Computer Science and Technology Studies, 2022, 4(2).

[4] Shu He. Research on Air Quality Prediction Based on Improved BP Neural Network [D]. Nanchang University, 2020.

[5] You Jing, Zhang Linjing. Application of Bayesian Regularized BP Neural Network in Air Quality lndex Prediction [J]. Journal of Chongqing University of Science and Technology (Natural Science Edition), 2022, vol. 24, no. 01, pp. 78-82.

[6] He Fahu, Liang Jiantao. Prediction of Air Quality Index Based on LSTM [J]. Modern computer, 2021, vol. 18, pp. 64-67.