Analysis and Forecast of Urban Air Quality Based on BP Neural Network

Wenjing Wang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 1908644938@qq.com

Shengquan Yang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: xaitysq@163.com

***Abstract*—The rapid economic development has led to the declining quality of the atmospheric environment. At present, my country is facing a very serious problem of atmospheric environmental pollution. Accurate prediction of air quality plays a vital role in the realization of air pollution control by environmental protection departments. Based on the historical air pollution concentration data, this paper establishes a BP neural network model to learn the statistical law of air pollutant values to realize the prediction of air quality in the future. Through the analysis of the target of air quality prediction, the design of an air quality prediction method based on BP neural network is designed. This method includes four stages: air pollutant concentration data collection, data processing, air quality index calculation, and prediction network construction. The experimental results show that the air quality prediction method based on BP neural network designed and implemented in this paper, combined with the developed air quality prediction system, can effectively predict the recent changes in air quality and various air pollutant concentrations. By collecting the concentration data of air pollutants and learning the changes of air pollutants to achieve air quality prediction, it provides a quantitative reference for government environmental protection departments to achieve air pollution control.**

Keywords-AQI; Air quality Prediction; BP Neural Network

I. INTRODUCTION

Air quality prediction, as the name suggests, is based on the historical emission concentration values of various pollutant items in the air to predict the concentration values of various pollutants in the air pollution in the future and the air environment quality[1]. As China's rapid economic development has led to serious atmospheric environmental pollution problems, the state and the public have paid more and more attention to the treatment and prevention of air pollution. The government environmental protection department hopes to keep abreast of the details of local air pollution and the recent changes in air pollution. The public also hopes to be able to understand the impact of air quality around them on their health in time. In recent years, the state has increased its plans for ecological environmental protection, and the plan clearly clarified that atmospheric pollution control is one of the key contents. The environmental protection departments of local governments strengthen air pollution control work, hoping to understand the changes in air quality in a timely manner by establishing an air quality prediction model.

DOI: 10.21307/ijanmc-2020-028

Xu Dahai proposed an atmospheric advection diffusion box model in 1999, in which the concept of the air pollution potential index was clearly determined, which effectively improved the accuracy of potential prediction on the basis of existing research [2]. In 2002, Liu Shi proposed a statistical model for potential prediction based on the air pollution of Changchun City. The model achieved a certain prediction effect [3]. But generally speaking, the accuracy of the potential prediction is very low, so it needs to be used together with other prediction methods, and cannot be used alone. The chemical model for high resolution of the troposphere in the atmosphere established by Lei Xiaoen is a typical numerical prediction model. Using this numerical prediction model can realize the prediction of the changing process of air pollutants in the atmosphere [4]. Due to its own characteristics, numerical prediction requires detailed geographic, meteorological, and pollution sources to realize the air quality prediction process. Collecting these data in actual situations requires huge costs and is difficult to obtain. In addition, numerical prediction models require high the amount of hardware computing resources is used to calculate the change trend of air pollution at high speed. The calculation complexity is high and it takes a long time, so the current numerical prediction model is not popular in small and medium-sized cities. Taiwan's Pai uses a gray model to achieve air quality prediction. The final actual results show that this method can achieve good results in achieving air quality prediction [5].

The time series analysis method and multiple regression model method in the statistical prediction method simplifies many change factors that affect air quality in the process of achieving air quality prediction, and makes many assumptions in the training process to achieve prediction, and finally achieves air quality the accuracy of the prediction needs to be further improved. The neural network has a good approximation effect in air quality prediction. It can continuously update the newly acquired air pollutant information to the neural network, update the prediction model in time, and improve the prediction accuracy. The neural network has a strong performance in air quality prediction. Dynamic adaptability and fault tolerance. In his research, Wang Jian pointed out that the BP neural network has advantages that other methods do not have in problems such as air quality prediction [6]. This paper uses air quality prediction based on BP neural network, and builds a neural network model to achieve air quality prediction, providing government environmental protection departments with air pollution trends.

II. AIR QUALITY RELATED FACTORS

AQI is the abbreviation of Air Quality Index. AQI does not refer to the value of a specific pollutant project, but reduces the concentration of the six air pollutant projects S02, N02, O3, CO, PM2.5 and PM10 to a single concept. Sex index form, used to represent the overall situation of air quality [7]. According to the size of the AQI value, the air pollution situation can be divided into different levels, and different air quality levels indicate the overall air quality in the local area over a period of time. The goal of this research is to make a short-term prediction of AQI in Xi'an, select the six main pollutant concentrations of AQI as features, build an air quality prediction model, improve the prediction accuracy and efficiency of the air quality prediction model, and provide environmental monitoring and governance Provide accurate air quality information.

In terms of data set acquisition, the air quality pollutant concentration data comes from the weather post website. Using web crawler technology to crawl the data of the website's air quality data module, the data from October 2013 to December 2019 can be obtained. Relevant feature data, after preprocessing the feature data to form an experimental data set. The original data does not necessarily meet the needs of the prediction model. The original data often needs to be processed before the training model is constructed, so

that the collected original data meets the needs of the model. This paper studies the air quality prediction method, and the construction of the air quality prediction model mainly needs to consider the lack of data Processing, data outlier processing, and data normalization processing.

In this paper, the mean value filling method is used to deal with missing values. The mean value filling method is to replace the missing values with the average value of historical data. This method is simple to implement and suitable for models with high accuracy requirements. Data anomaly refers to an unreasonable value in a data set. For example, taking air pollutant concentration data as an example, if the actually collected concentration data value is a negative number, the value is determined to be an abnormal value. In the research method of this paper, the outliers are regarded as missing values, and the outliers are dealt with in the way of missing values. In order to avoid the overflow of the weight of the neural network is too large or too small, to eliminate the possible impact of different variables of the input vector due to different dimensions or too large difference in value, the input vector of the neural network needs to be processed. Normalized data processing is performed on the collected original data set, so that each index of each element data of the vector is at the same order of magnitude, which is suitable for training model for learning. This article uses the Z-score standardized method, the calculation method is:

$$x^* = \frac{x - \mu}{\delta} \tag{1}$$

Among them, μ is the mean value of all sample data, and δ is the standard deviation of all sample data.

III. AIR QUALITY PREDICTION MODEL

A. BP neural network

BP neural network is an error back propagation neural network. Rumerlhar proposed an error back propagation algorithm in the study of forward neural network, referred to as BP neural network algorithm. The network of each layer of the BP neural contains many neuron nodes. There is no connection between the neurons in the layer, and all the neuron nodes between adjacent layers are fully connected [8]. The input layer is used to accept network input information. Each neuron will generate the corresponding link weight according to the input information of the obtained network. The function of the hidden layer in the BP neural network is information detection. According to Tambe's global approximation theory, even if a neural network contains only one hidden layer, as long as there are enough neuron nodes and the appropriate connection function and weight are selected, it can be arbitrary. Approximate the input and output vector of a measurable function [9]. The BP neural network can obtain information and continuously update it to the network, and constantly adjust its structure to meet the characteristics of the model, and has strong self-adaptability and fault tolerance.

The BP neural network learning process is that after receiving the initial input and the given target output, the information forward propagation learning process is performed. This process first calculates and calculates each neural unit of the input layer and each neural unit of the hidden layer. Obtain the output of each neural unit of the hidden layer, and then use the same method to calculate the output of each neural unit of the output layer to determine the error between the actual output of the output layer and the target output. If the error value is within the user's acceptable range, then Fix the weight and threshold, and end the training, otherwise it enters the second stage. The second stage is the error signal back-propagation stage. In this stage, the partial derivative of the error is first calculated

using the output of the output layer, and then the partial derivative obtained by calculation is weighted and summed with the previous hidden layer. Input layer, and finally use the partial derivative calculated by each neural unit to update the weight [10]. Repeat these two stages until the error between the actual output and the target output is reduced to an acceptable range. Figure 1 is the learning flowchart of the BP neural network:

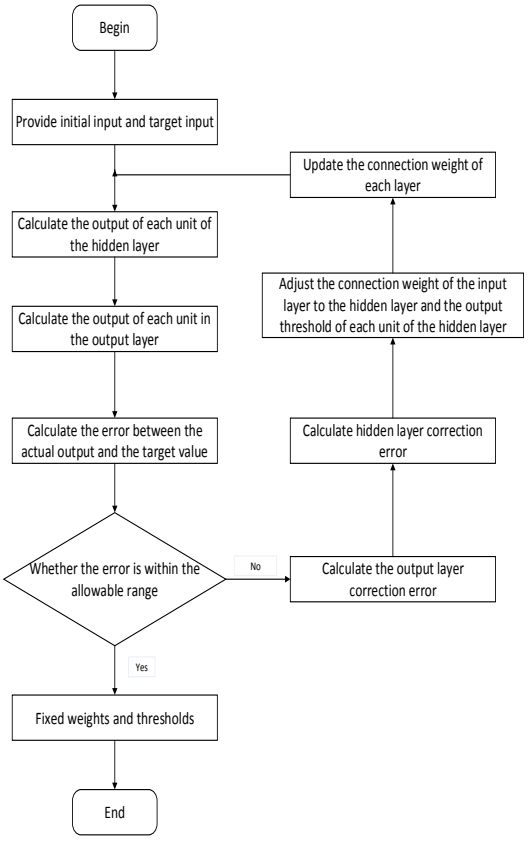

Figure 1. BP neural network learning process

From the above, the algorithm flow of BP neural network can be divided into two processes, as follows:

1) Forward propagation sub-process

It is now defined that the input value of the input layer node is X_{i}, the weight value between the input layer and the hidden layer node is W_{ih}; the threshold value of the hidden layer node is b_h and the value between the hidden layer and output layer node is W_{ho}; output The threshold of the layer node is b_{o}, the network activation function is f, the output value of the output layer node is $\mathrm{y}o_{\mathrm{o}}$, and the expected output value is y_{o}.

The forward propagation process of the BP neural network is to solve the output layer output value X_{i} from the input layer input value $\mathrm{y}o_{\mathrm{o}}$. Specific steps are as follows:

a) Calculate the input and output values of the hidden layer

Hidden layer input value:

$$\mathrm{hi}_h = \sum_{i=1}^{m} W_{ih} X_i + b_h \quad (\mathrm{h}=1,2,...,\mathrm{n}) \qquad (2)$$

Hidden layer output value:

$$\mathrm{ho}_h = f(hi_h) \quad (\mathrm{h}=1,2,...,\mathrm{n}) \qquad (3)$$

b) Calculate the input value and output value of the output layer

Input value of output layer:

$$\mathrm{yi}_o = \sum_{o=1}^{k} W_{ho} ho_h + b_o \quad (\mathrm{o}=1,2,...,\mathrm{k}) \qquad (4)$$

Output value of output layer:

$$yo_o = f(yi_o) \quad (\mathrm{o}=1,2,...,\mathrm{k}) \qquad (5)$$

2) Back propagation sub-process

The back propagation process of BP neural network is based on Widrow-Hoff learning rules. The error function is as follows:

$$E（W，\ \mathrm{b}）=\frac{1}{2}\sum_{o=1}^{k}(y_0-yo_o)^2 \qquad (6)$$

The main goal of the BP neural network algorithm is to iteratively modify the weights and thresholds between layers so as to minimize the value of the error function. According to the Widrow-Hoff learning rule, along the direction of the steepest descent of the sum of squared errors, the weights and thresholds are constantly adjusted. According to the gradient descent method, the amount of weight change is proportional to the gradient of the error function at the current position, as shown in equation (6)：

$$\Delta W=-\eta_1\frac{\partial E(W,\mathrm{b})}{\partial W} \qquad (7)$$

Also for thresholds are:

$$\Delta \mathrm{b}=-\eta_2\frac{\partial E(W,\mathrm{b})}{\partial b} \qquad (8)$$

In the formula: η_1, η_2 is the learning speed, and its value range is (0,1).

The specific steps of the BP neural network back propagation process are as follows:

a) Calculate the weight between the hidden layer and the output layer and adjust the threshold of the output layer

For W_{ho}, according to formula (6), we can get:

$$\Delta W_{\mathrm{ho}}=-\eta_1\frac{\partial E(W,b)}{\partial W_{ho}}=-\eta_1\frac{\partial E}{\partial yi_o}\frac{\partial yi_o}{\partial W_{ho}} \qquad (9)$$

From formulas (3), (4), and (5), we can get:

$$\frac{\partial \mathrm{y}\mathrm{i}_o}{\partial W_{ho}}=ho_h \qquad (10)$$

$$\frac{\partial E}{\partial \mathrm{y}\mathrm{i}_o}(y_0-yo_o)\,f'(yi_o)=\delta_{\mathrm{o}} \qquad (11)$$

From formulas (8), (9), (10), we can get:

$$\Delta W_{\mathrm{ho}}=-\eta_1\delta_{\mathrm{o}}ho_h \qquad (12)$$

Similarly, we can get:

$$\Delta \mathrm{b}_o=-\eta_2\delta_{\mathrm{o}} \qquad (13)$$

b) Calculate the weight between the input layer and the hidden layer and the adjustment amount of the hidden layer threshold

For W_{ih}, according to equation (6):

$$\Delta W_{\mathrm{ih}}=-\eta_1\frac{\partial E(W,b)}{\partial W_{ih}}=-\eta_1\frac{\partial E}{\partial hi_h}\frac{\partial hi_h}{\partial W_{ih}} \qquad (14)$$

Since hi_h affects all output layers, there are:

$$\frac{\partial E}{\partial \mathrm{hi}_h}=\sum_{i=1}^{k}\frac{\partial E}{\partial yi_o}\frac{\partial yi_o}{\partial hi_h} \qquad (15)$$

From formulas (2) and (3), we can get:

$$\frac{\partial \mathrm{yi}_o}{\partial hi_h} = W_{ho} f^{'}(hi_h) \tag{16}$$

From formula (10)、(15)、(16), we can get:

$$\frac{\partial E}{\partial \mathrm{hi}_o} = f^{'}(hi_h)\sum_{o=1}^{k}\delta_o W_{ho} = \delta_h \tag{17}$$

From equations (13), (14) and (17), we can get:

$$\Delta W_{\mathrm{ih}} = -\eta_1 \delta_{\mathrm{h}} \frac{\partial \mathrm{hi}_h}{\partial W_{ih}} \tag{18}$$

Similarly, we can get:

$$\Delta \mathrm{b}_{\mathrm{h}} = -\eta_2 \delta_{\mathrm{h}} \tag{19}$$

c) Update the weights and thresholds of the BP neural network

From (12), (13), the updated weight and output layer threshold between the hidden layer and the output layer are:

$$W_{\mathrm{ho}}^{N+1} = W_{ho}^{N} - \eta_1 \delta_o ho_h \tag{20}$$

$$\mathrm{b}_o^{N+1} = b_o^N - \eta_2 \delta_o \tag{21}$$

From equations (19) and (20), the updated weights and hidden layer thresholds between the input layer and the hidden layer can be obtained:

$$W_{\mathrm{ih}}^{N+1} = W_{ih}^{N} - \eta_1 \delta_h \frac{\partial hi_h}{\partial W_{ih}} \tag{22}$$

$$\mathrm{b}_h^{N+1} = b_h^N - \eta_2 \delta_h \tag{23}$$

B. Design of air quality prediction model

The core algorithm used in this paper is the BP neural network algorithm. According to the characteristics of the BP neural network, this topic needs to determine the number of neuron nodes in each layer of the network, and select the network activation function and initial parameters. The determination of the number of input layer nodes of the BP neural network is very important. Too many or too few selections will affect the prediction accuracy of the model. Therefore, the number of input layer nodes should be determined according to the actual application needs. This subject designs the input layer and output layer of the network based on the collected data. The number of input layer nodes is 6, which are the data of the concentration values of six pollutants such as PM2.5, PM10, SO2, NO2, C0, and O3 in a day. The number of nodes in the output layer is one, that is, the AQI value of the next day. The structure of the BP neural network in this subject is divided into three layers, with only one hidden layer. There is no theoretical guidance for determining the number of hidden layer nodes, and it is usually based on specific practical experience. The empirical formula for selecting the number of hidden layers is:

$$\mathrm{p} = \sqrt{n+q} + \alpha \tag{24}$$

In the formula, n and q represent the number of neurons in the input layer and output layer, respectively, generally take an integer between 1-10. The number of hidden layer nodes in this subject is first determined as 7.

The network activation function is an important factor that affects the performance of the BP neural network algorithm, which makes the network have nonlinear processing capabilities. There are three activation functions of BP neural network: log-sigmiod function, tanh function and ReLU function. According to the characteristics of the research data and the

characteristics of the three activation functions, in this paper, the hidden layer of the BP network selects the log-sigmiod function as the activation function, and the output layer selects the ReLU function as the activation function. Since the sample data is normalized, the value interval between the initial weight and the threshold is between [-1, 1], and they should be a set of random numbers that are not exactly equal.

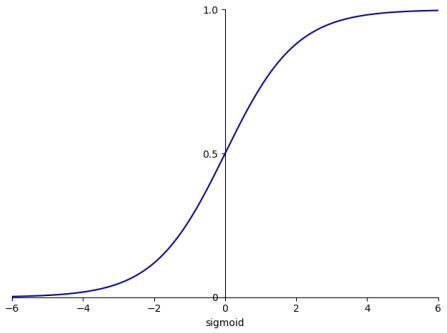

Figure 2. Log-sigmiod function

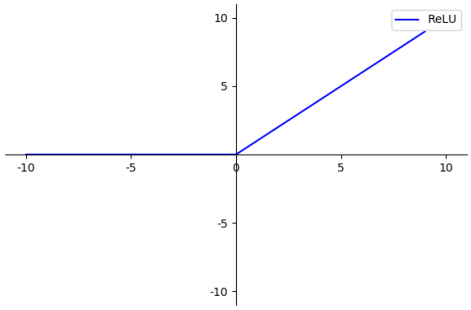

Figure 3. ReLU function

IV. EXPERIMENT

The experimental simulation platform used in this article is the Python programming language. The air data object used in the experiment is the air quality data of Xi'an from October 2013 to December 2019. All the experimental data are sorted in a continuous time series. Take the data for 30 consecutive days as the test data set, and the other as the training data set. For the evaluation of the advantages and disadvantages of the model, this paper uses the average error and the root mean square error to evaluate. The calculation formulas are shown in equations (25) and (26):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y^*_i}{y_i}\right| \quad (25)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y^*_i)^2} \quad (26)$$

Experimental results show that the prediction model established in this paper has high accuracy and high efficiency for PM2.5 concentration prediction. The simulation prediction results are shown in Figure 4 the measured values and predicted values of the first 6 groups are compared to obtain Table 1.

TABLE I. COMPARISON BETWEEN MEASURED AND PREDICTED

AQI	AQI prediction	PM2.5	PM 10	So2	No2	Co	O3
45	56.21	27	40	4	24	0.71	75
49	60.51	23	47	6	36	0.63	84
55	64.30	33	57	6	35	0.62	84
57	68.54	39	60	5	40	0.71	56
68	80.22	40	67	5	39	0.77	85
61	74.35	33	69	6	47	0.70	71

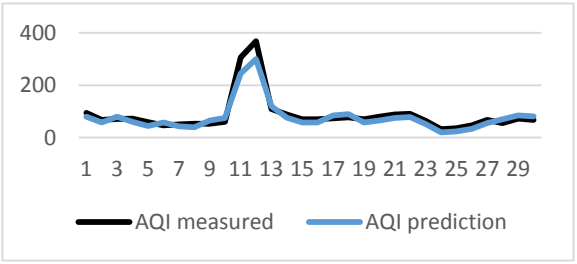

Figure 4. Comparison of sample prediction results with real values

The experimental results show that, after analyzing the prediction results, it is concluded that under the experimental conditions given in this paper, the average error of the experimental results is 0.074 and the root mean square error is 13.41. As can be seen from Figure 4-1 and Table 4-1, the BP neural network established in this paper has lower prediction error when the air quality fluctuates greatly.

This research also has some shortcomings at present. It only considers the relevant factors that can be quantitatively analyzed, and does not take into account some unexpected emergencies. For example, natural disasters, human factors, etc. Due to the unpredictable and unquantifiable characteristics of these factors, they have not been considered in the article. In the future research work, we hope to analyze these factors.

V. CONCLUSION

This article aims at the current situation of severe air pollution problems facing China. The traditional air pollutant online monitoring system cannot effectively use historical air pollutant data to provide quantitative reference for air pollution control and various control measures. The environmental protection department urgently needs to establish an air quality prediction system to realize the supervision and control of local air pollution. This paper studies a method to achieve air quality prediction based on BP neural network. By studying the change law of historical air pollutant project concentration data, it predicts the future air quality change trend for a period of time, and helps government environmental protection departments formulate air pollution control policies to provide quantification Indicators and references.

This article first explains the research background and significance of this topic, and analyzes the necessity of establishing an air quality prediction system for air pollution control. Based on the analysis of the domestic research results of air quality prediction, combined with the regional characteristics and actual conditions of prefectural and municipal government departments, a framework model for air quality prediction based on statistical prediction is proposed. Then, an air quality prediction method model based on BP neural network is established, and the realization of the method includes three stages of air pollutant project concentration data collection, data processing, and prediction algorithm network model construction. This paper uses BP neural network to predict the air quality in Xi'an. Through the analysis of experimental results, BP neural network has a significant effect in dealing with such nonlinear problems, especially in the place where the AQI fluctuation is relatively large. The research is conducive to the prediction and prevention of air pollution problems. The government can also make appropriate measures and decisions based on the prediction results, such as closing schools or reducing outdoor sports, thereby reducing the damage caused by pollution. It can also provide new ways and methods for forecasting research in other fields.

ACKNOWLEDGMENT

The Research is supported by the new network and detection control national and local joint engineering laboratory. (Financing projects No. GSYSJ2016014).

REFERENCES

[1] Ren Wanhui, Su Zongzong, Zhao Hongde. Advances in the study of urban environmental air pollution forecasting [J]. Environmental Protection Science, 2010, 36(03):9-11.

[2] Xu Dahai, Zhu Rong. Popularization and application of urban air pollution forecasting model [J]. Annual Report of CAMS, 1999(00):33.

[3] Liu Shi, Wang Ning, Zhu Qiwen, Wang Xinguo, Hu Zhongming, Chen Changsheng. Research on the Statistical Model of Air Pollution Potential Forecast in Changchun City [J]. Meteorology, 2002(01):8-12.

[4] Han Zhiwei, Du Shiyong, Lei Xiaoen, Ju Lixia, Wang Qingeng. Urban air pollution numerical prediction model system and its application [J]. Chinese Environmental Science, 2002(03): 11-15.

[5] Tzu‐Yi Pai,Keisuke Hanaki,Ren‐Jie Chiou. Forecasting Hourly Roadside Particulate Matter in Taipei County of Taiwan Based on First‐Order and One‐Variable Grey Model [J]. John Wiley & Sons, Ltd, 2013, 41(8).

[6] Wang Jian, Hu Xiaomin, Zheng Longxi, Liu Zhenshan. Research on air pollution forecasting method based on BP model [J]. Environmental Science Research, 2002(05):62-64.

[7] Wang Qingeng, Xia Sijia, Wan Yixue, Jin Longshan. Problems and new ideas in current urban air pollution forecasting methods [J]. Environmental Science and Technology, 2009, 32(03):189-192.

[8] A. Elkamel, S. Abdul-Wahab, W. Bouhamra, E. Alper. Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach [J]. Advances in Environmental Research, 2001, 5(1).

[9] Jaakko Kukkonen, Leena Partanen, Ari Karppinen, Juhani Ruuskanen, Heikki Junninen, Mikko Kolehmainen, Harri Niska, Stephen Dorling, Tim Chatterton, Rob Foxall, Gavin Cawley. Extensive evaluation of neural network models for the prediction of NO 2 and PM 10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki [J]. Atmospheric Environment, 2003, 37(32).

[10] Hunt K.J., Sbarbaro D., Żbikowski R., Gawthrop P. J. Neural networks for control systems â€ ”A survey [J]. Pergamon, 1992, 28(6).