Research on Air quality based on optimized BP Neural Network genetic algorithm

Fuyou Mao, Haomin Zhao, Xiyang Liu

College of Computer Science and Technology, Shenyang Jianzhu University, Liaoning, Shenyang, 110168, China

Abstract

Air pollution is one of the primary pollution sources of ecological environment pollution, which will pose significant harm to people's physical and mental health. It is used to predict air pollution accurately. In order to improve the accuracy of the prediction model, it is necessary to explore the hidden correlation between the actual measurement data and the primary prediction data. First of all, this paper processes the first prediction data and the measured data of the three monitoring points, then establishes a neural network model with momentum factor, establishes the model and iterates through python, and obtains the most suitable weight and the threshold between the two groups of data. Finally, the machine learning genetic algorithm is used to correct the error of the forecast data to get a more accurate daily concentration of conventional pollutants.

Keywords

Air quality forecast, Neural network, Genetic algorithm.

1. Introduction

With the rapid development of China's economy, environmental problems are becoming more and more serious[1]. The pollutant gas contained in the atmosphere is getting worse. At present, a variety of pollutants exist simultaneously, and it has become a high concentration pollutant gas, causing the problem of compound air pollution. Air pollution prediction is a process of prediction based on pollution emission sources (including emission source list and real-time monitoring data) and meteorological data through the air quality model coupled with physical and chemical mechanisms. The practice of pollution prevention and control shows that by establishing the air quality prediction model, the relevant air pollution processes can be obtained in advance, and corresponding control measures can be taken in time[2].

It has become one of the ways to reduce the harm of air pollution to human health, reduce the impact on the ecological environment and effectively improve the ambient air quality. The WRF-CMAQ simulation system is usually used for air quality prediction [3] (hereinafter referred to as the WRF-CMAQ model). The so-called WRF-CMAQ model is mainly composed of WRF and CMAQ: WRF is a weather prediction model with unified mesoscale, and CMAQ can use its meteorological field data. CMAQ is a simulation system related to air quality developed by the US Environmental Protection Agency. It is based not only on the meteorological information from WRF but also on the pollution field's emission list. The simulation of different forms of change processes of pollutants is based on a series of physical and chemical reaction principles to obtain the accurate prediction results of specific time points or required time periods.

Through real-time monitoring of the concentrations of "two dust" (PM2.5, PM10) and "four gases" (O_3, Co, SO_2, NO_2), AQI, the maximum value of air quality index, is used to represent the degree of air pollution, and the air quality status is detected in time. One of the six pollutants is ozone concentration, which is difficult to predict. The reason is that ozone is a unique secondary pollutant among the six pollutants. The pollutant is not from the pollution source of direct emission but the chemical and photochemical reaction in the atmosphere, making it more difficult for the WRF-CMAQ model to predict the change of ozone concentration accurately. Therefore, if the secondary modeling can be carried out based on primary simulation results such as WRF-CMAQ and other data sources, it will help to improve the accuracy of model prediction, purify the air in time, achieve the goal of improving the environment and creating a healthy, comfortable and safe living environment.

2. BP neural network

Genetic algorithm has good applicability in solving multivariable optimization problems. Therefore, this study intends to build a genetic algorithm based on a BP neural network solution model.

Prediction methods can be divided into two analysis methods according to their types: qualitative and quantitative. In the qualitative analysis method, professional forecasters analyze the lack of data in the prediction data according to the accumulated rich experience and subjective judgment ability. The quantitative analysis method uses a mathematical model to calculate various numerical indexes for the analysis object, such as the causality analysis method and the time series analysis method. The BP neural network can be studied in the field of data prediction because its function approximation ability is outstanding. Then, the neural network belongs to the forward neural network and adopts the error back propagation learning algorithm. It comprises an input layer, multiple hidden layers, and an output layer. The main idea of the model is to train through the sample set. For example, Tao Yangwei et al. [4] improved the BP neural network by using the momentum term in the error backpropagation process.

The input layer data used many data indicators such as gross national product, energy structure, and urban population proportion to create a model that can predict China's energy demand and predict China's energy demand in the next three years. The practice of pollution prevention and control shows that to reduce the harm of air pollution to human health and the environment is reduced. The quality of environmental air is also improved, and it is necessary to establish an air quality prediction model for people's physical comfort and health, know the possible process of air pollution in advance and take corresponding control measures[5-7]. However, due to the uncertainty of simulated meteorological field and emission inventory and the incompleteness of the generation mechanism of pollutants, including ozone, the results of the WRF-CMAQ prediction model are not ideal. Based on the prediction results of the WRF-CMAQ model, we established a BP neural network model (GA-BP) based on a genetic algorithm to improve the accuracy of prediction. The structure of the BP neural network model is shown in the Figure 1.

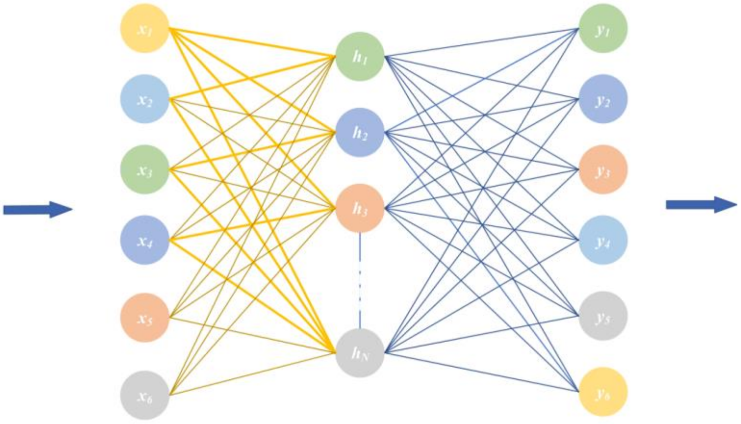

Figure.1 BP neural network model diagram

Let the input of BP neural network be $X=\{x_1,x_2,x_3,\ldots,x_6\}$,x_1-x_6 is SO_2, NO_2 Predicted concentration data of six pollutants. It corresponds to the measured concentration data of six pollutants, such as SO_2, NO_2, O_3, etc. A complete iteration of the BP neural network includes forward propagation and error backpropagation. The calculation process is as follows.

Forward propagation includes the calculation process from the input layer to the hidden layer and from the hidden layer to the output layer

$$h_{input_{in}}=\sum_{i=1}^{6}w_{i\_in}\cdot x_i \tag{1}$$

$$h_{\text{output}_j}=f\left(h_{\text{input}_j}\right) \tag{2}$$

Where " w_{i_in} " is the weight of hidden layer neurons corresponding to the i input layers, and N is the number of hidden layers. This paper sets N to 7, and it is the activation function. Since the data in this paper are greater than 0, the sigmoid function is selected as the activation function. The specific calculation process of this function is shown in the following formula:

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

The calculation formula for the output layer is:

$$O_{output}=\sum_{j=1}^{6}w_{j\_\text{out}}\cdot \text{hidden}_j \tag{4}$$

After completing a forward propagation, calculate the training error. In the initial stage of training, the forward propagation is carried out along the network to generate the error between the network output value and the expected output value. The backpropagation updates the threshold and weight and dynamically adjusts and updates them to approach the expected output value gradually. Because BP neural network has the disadvantage of slow convergence speed, aiming at this deficiency, this paper uses the momentum BP method to update the weight. The momentum factor is introduced when updating the weight:

$$\Delta w(n)=-n(1-m)\nabla e(n)+m\Delta w(n-1) \tag{5}$$

After the momentum factor is obtained, the weight is modified:

$$w(n)=w(n-1)+\Delta w(n-1) \tag{6}$$

3. Genetic algorithm

Although BP neural network model can solve the nonlinear problem between various factors affecting air quality, its accuracy and training speed are still difficult to meet the demand of minimizing the relative error of AQI in the problem. Therefore, the genetic algorithm is supplemented to form a combination model to make up for the defects of the BP neural network. When using the genetic algorithm to solve the optimal solution problem, we use binary coding to represent these variables. These binary codes are equivalent to individual gene segments, and the individual of the whole algorithm is composed of these gene segments. These individuals are part of the genetic algorithm population and follow the law of survival of the fittest. Then, when individuals continue their offspring, they use roulette to select appropriate individuals. After reaching a certain genetic algebra operation, they will finally get an optimal population that can adapt to the environment, which is also the optimal solution to this problem. Among them, the adaptive ability of genetic algorithms is mainly reflected in whether the individuals in the population can adapt to the changes in the surrounding environment to adjust their abilities. The adaptability of the genetic algorithm (GA) mainly depends on the crossover probability P_C and mutation probability P_M.

(1) Parameter initialization:

The predicted data of each monitoring point is used as the input port of the model, and the measured data of each monitoring point is used as the output part. Initialize the combined model and train it in a Python environment. The number of groups is set to 6 to ensure that the individuals in the population have sufficient population diversity. The initial population of the first generation is defined as a structure, including 6 population numbers and 6 chromosome codes. The maximum evolutionary algebra is set to 20 generations. The empirical values of crossover and mutation probability were selected as Pc=0.2 and Pm=0.1, respectively. The optimal evolution stop criterion of the GA algorithm is set as follows: the maximum evolution algebra 20 is reached, or the fitness value is maintained at 1.0e-6 for several consecutive generations.

(2) Overlapping

At this stage, two individuals are randomly selected from the parent population, and the offspring inherit the excellent chromosomes of the parent through the cross combination of two chromosomes to produce new excellent individuals. The crossing modes include single point, double point, multi-point, and arithmetic. Randomly select two chromosomes to cross and generate a matrix with one row and two columns. Check whether the matrix contains 0 elements. If so, regenerate the random number. If the final random number is less than the crossover probability, the crossover can be carried out.

$$\begin{aligned} a_{kj} &= a_{kj}(1-b) + a_{ij}b \\ a_{ij} &= a_{ij}(1-b) + a_{kj}b \end{aligned} \tag{7}$$

(3) Variation

Randomly select a chromosome. If the generated random number is greater than the mutation probability, it can be mutated. It is assuming that the jth gene of the ith individual is selected.

(4) The fitness function is used to calculate the fitness of individuals with training data

The training input data, output data, and network nodes are taken as the input, and the fitness is taken as the output. The evolutionary parameters are set as follows: the number of iterations is 20, and the learning rate is 0.5 1. The minimum target error is 0.00001.

(5) After the individual evolution of the training data, because the evolution direction is random and uncertain, we use the chrome function to control the individual in the test data to evolve towards the best individual in the training data to improve the fitness value of the test data greatly. Compare the fitness populations evolved in each generation and retain the individuals with a high degree of adaptation to the environment by eliminating those with poor fitness.

For the resulting fitness of the genetic algorithm, this paper introduces the numerical analysis method and uses the measured data and the corresponding prediction data for numerical analysis. The error of the WRF-CMAQ prediction model is corrected. In this model, the goal is to reduce the relative error to achieve the effect of error correction. Therefore, in a genetic algorithm, the overall training direction is to reduce the relative error limit of AQI. After the model training, the product of the weight is put out.

$$\begin{aligned} & s\quad w_{7\times6} = w_{\text{in}} \times w_{\text{out}} \\ & w_{in} = \begin{matrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{71} & \cdots & a_{7N} \end{matrix} \\ & w_{\text{out}} = \begin{matrix} b_{11} & \cdots & b_{16} \\ \vdots & \ddots & \vdots \\ b_{N1} & \cdots & b_{N6} \end{matrix} \end{aligned} \tag{8}$$

Maintain the most advanced optimal solution, do not participate in crossover and mutation operations, maintain excellent populations, and prevent the destruction of the optimal global solution. The worst solution does not to participate in the cross operation to prevent poor genes from passing on to the next generation.

4. Verification of GA-BP quadratic prediction model

Based on the above operations, this paper constructs a BP neural network (GA-BP) based on a genetic algorithm, which reduces the model error by dynamically adjusting the network weight so that the actual output of the network is as close as possible to the expected output

Data processing and parameter setting

The number of iterations of the neural network designed in this paper is 100, and the learning rate is 0.4. The hidden layers are 7, and the initial weight is randomly generated through uniform distribution. This paper sets an early stop mechanism for the model to prevent overfitting. Every ten iterations in the training set, the error will be calculated on the verification set. If the training result is worse than the last training result, the training will be terminated in advance, and the parameters in the last iteration result will be taken as the final parameters of the model.

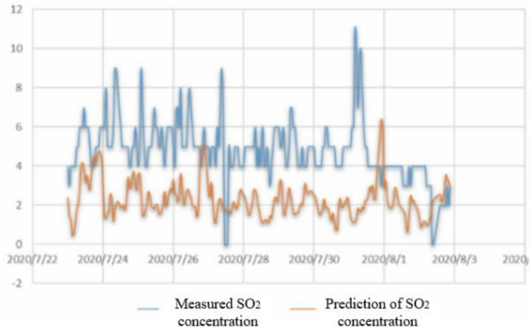

Figure.2 Comparison of SO_2 concentration measured and predicted data

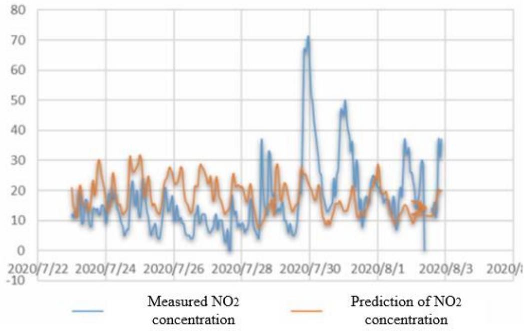

Figure.3 Comparison of NO_2 concentration measured and predicted data

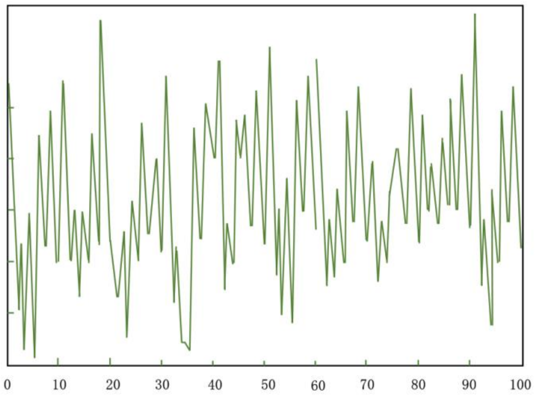

Figure.4 GA-BP model prediction output

The following Figure 5 shows the fitting accuracy of the GA-BP combined model. By calculating the prediction error, it can be concluded that the maximum error percentage of O_3

concentration data will not exceed 0.045, and the overall accuracy is 99.5%. In order to improve the relative error of O_3 concentration data, we optimize the model and add a genetic algorithm. The regression coefficient of the GA-BP combined network model is 0.99998, which proves that the measured data can be approximately regarded as a linear fitting function with time. A more accurate fitting effect is obtained.

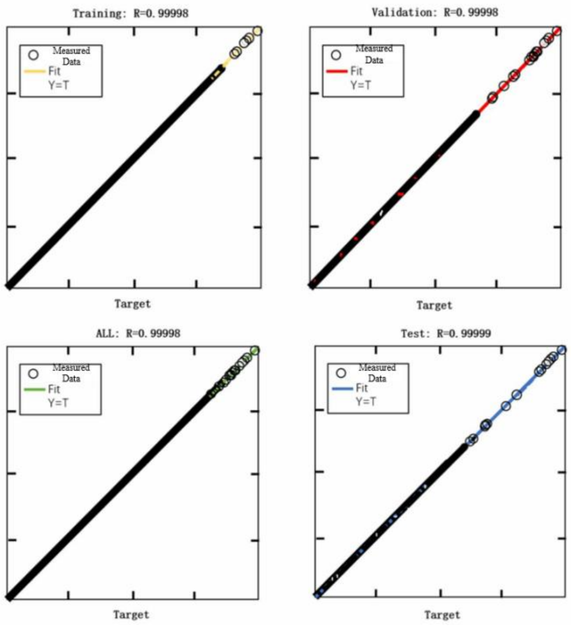

Figure.5 Regression coefficient

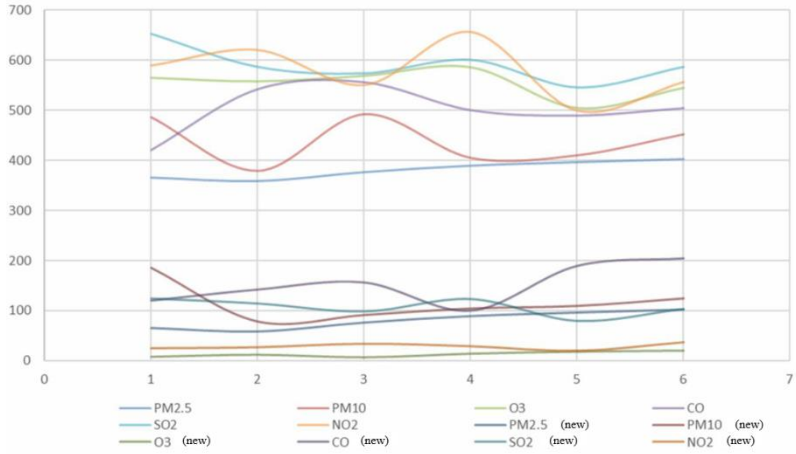

Figure.6 Error between primary prediction and secondary prediction data

Although the BP neural network model converges only at the 60th iteration

Tab.1 Daily value prediction results of the quadratic model at monitoring point A (As an example)

Forecast date	Place	SO_2 $(\mu g/m^3)$	NO_2 $(\mu g/m^3)$	PM_{10} $(\mu g/m^3)$	$PM_{2.5}$ $(\mu g/m^3)$	Maximum eight-hour moving average	CO (mg/m^3)	AQI	Primary pollutant
2021/7/13	A	4	28	10	6	94	0.2	47	NO
2021/7/14	A	5	36	9	6	118	0.2	65	O_3
2021/7/15	A	4	38	9	7	137	0.3	81	O_3

5. Conclusion

Our air pollution forecast is based on pollution emission sources (including emission source list and real-time monitoring data) and meteorological data and is predicted by an air quality model coupled with physical and chemical mechanisms. The practice of pollution prevention and control shows that through establishing an air quality forecast model, the relevant air pollution processes in advance can be obtained, and corresponding control measures can be taken in time. Therefore, this paper explores the hidden correlation between the actual measurement data and the primary prediction data because of the above process. The neural network model with momentum factor is established, and the model is established and iterated by python to find out the most suitable weight and the threshold between the two groups of data. Finally, the weight is iteratively derived using a machine learning genetic algorithm, and the Darwinian theory of evolution is used to train the initial weight in the evolutionary direction of error reduction. A more accurate relation weight is obtained. Thus, the primary prediction data error is corrected using the secondary prediction model, and a more accurate daily concentration value of conventional pollutants is obtained.

References

[1] Hao Jiming, Ma Guangda, Wang Shuxiao Air pollution control engineering [M] Beijing: Higher Education Press, 2010.

[2] PETALAS Y G, PARSOPOULOS K E, VRAHATIS M N. Improving fuzzy cognitive maps learning through memetic particle swarm optimization[J]. Soft computing, 2009,13 (1): 77-94

[3] Boxin et al Operation guide and case study of air quality model (smoke, WRF, CMAQ, etc.) [M] Beijing: China Environmental publishing group, 2019.

[4] Tao Yangwei, sun Mei, Wang Xiaofang Research on China's energy demand forecasting based on Improved BP neural network [J] Journal of Shanxi University of Finance and economics, 2010 (S2): 3-5.

[5] Dai Shugui Environmental chemistry [M] Beijing: Higher Education Press, 1997.

[6] Zhao Qiuyue, Li Li, Li Huipeng Research progress of near surface ozone pollution at home and abroad [J] Environmental Science and technology, 2018, 31 (05): 72-76.

[7] Chen Mindong Formation mechanism and research progress of atmospheric ozone pollution [J / OL] 2018, https://max. book118. com/html/2018/0201/151478594. shtm.