Prediction of PM2.5 Concentration in Chengdu Based on Optimized BP Neural Network

Yuxiang Wang

Southwest University, School of Computer Information and Science, Chongqing, China

Keywords: PM2.5, BP Neural Network, PCA, GA, Leaky-ReLU.

Abstract: Recently, air quality has always been the focus of attention in China, and the air quality has gradually improved under constant efforts. This paper aims to establish a reasonable model to predict and analyze PM2.5 concentration data in Chengdu. After an analysis of the main influencing factors of PM2.5 in Chengdu urban area of recent years, 39 sample data from 8 meteorological observation points in Chengdu in February 2019 are selected, factors including PM2.5 concentration, humidity and wind speed as input and output. The BP neural network model is used to predict the concentration of PM2.5 in Chengdu. Due to the large dimension of the input elements, principal component analysis is introduced in the input layer to achieve the goal of dimensionality reduction. The principal components of the principal component analysis method is then used to obtain the optimal initialization weights and thresholds by genetic algorithm, and the activation function uses Leaky-ReLU. By training with sample data, the network model structure is determined, and prediction accuracy is tested by using the sample data. The results show that the optimized BP neural network model has a good performance in PM2.5 concentration prediction with high precision.

1. Introduction

Traditionally, the meteorological department uses weather dynamics and modern calculation methods to build a weather forecast model based on the past collected data, and then predicts the weather conditions[1]. According to our analysis, BP neural network has strong nonlinear mapping ability, and can carry out large-scale data parallel computing. It also holds certain adaptability and self-organization ability. It is widely used in many subject areas. On this basis, a BP neural network prediction model can be established for the prediction of PM2.5 concentration in Chengdu urban area. The neural network model prediction method can solve some current problems in meteorological prediction. Multiple sets of historical data can be obtained at different observation points around Chengdu City to set up a suiTable prediction model. Applying relevant data into our model can predict the weather characteristics in a period of time[2].

2. Model Selection

According to the type of the research topic and the characteristics of the data, it is in line with the conditions of use of the BP neural network. BP neural network is a network model of feedforward transmission. Besides, the normal signal in the network propagates forward and the error signal propagates backward. The input signal propagates from the input layer to the hidden layer and then to the output layer, and the propagation of each layer only affects the next layer. If the output signal does not reach the expected value, it would propagate in the opposite direction in order to adjust each neuron. The weights and thresholds are recalculated until the final output meets expected standards. The specific structure of the BP neural network is shown in Figure 1[3].

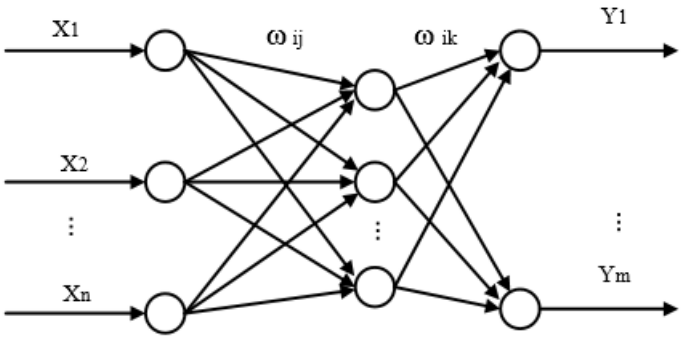

Figure 1. BP neural network topology diagram

In the figure 1, X_1, X_2 , … , X_n are input variables, Y_1 , … , Y_m are output variables, and ω_{ij}, ω_{ik} are the connection weights between the input layer and the hidden layer, and between the hidden layer and the output layer respectively. According to Fig. 1, it can be seen that BP neural network can be regarded as a nonlinear function. X_1, X_2 , … , X_n are n independent variables of function, Y_1 , … , Y_m are therefore dependent variables. Actually, it is a mapping relationship from n independent variables to m dependent variables.

BP Neural Network Prediction Model Establishment

2.1 Selection of Major Meteorological Elements

Meteorological elements refer to the description of weather conditions at a certain point in time, and may include some environmental conditions related to residents' activities, such as temperature, humidity, air pressure, altitude, precipitation, and wind speed. The current meteorological forecast is usually divided into five processes, namely meteorological observation, data collection, comprehensive analysis, forecasting consultation, and forecast product release[4]. PM2.5 refers to particulate matter containing aerodynamic diameter less than or equal to 2.5 μm in the atmospheric environment of a certain region, which is derived from two main modes, natural and artificial. In this paper, by analyzing the characteristics of Chengdu area, we select humidity, wind speed, sea level pressure and temperature as auxiliary meteorological elements for PM2.5 prediction. The data mainly comes from the air quality testing data of the Sichuan Provincial Environmental Monitoring Station. The detecting stations are distributed in 8 locations in Chengdu.

2.2 Principal Component Analysis

Principal component analysis is used to reduce the dimensionality of high-dimensional data space while trying to ensure the least data loss. Obviously, it is much easier for a system to deal with problems in a low-dimensional space than in high-dimensional ones. The dimension reduction of high-dimensional variable space is a few linear combinations of the research index system, and the comprehensive indicators formed by these linear combinations will retain as much information as possible on the original indicator variation. These comprehensive indicators are called principal components. In the BP neural network of this paper, the data dimension is large, so it is necessary to make dimensionality reduction operations. Besides, the principal components with cumulative contribution rate exceeding 80% are selected as the input index.

2.3 Genetic Algorithm for Optimizing the Network

Genetic algorithm is a random search algorithm that draws on the natural selection and natural genetic mechanism of the biological world. Compared with the traditional optimization algorithm, the genetic algorithm operates on the parameters rather than itself, starting from multiple points in parallel, rather than limiting at one point[5]. It can effectively prevent the local optimal solution from appearing in the search process. The algorithm calculates the appropriate value through the objective function, does not need other derivation and additional information, and has less dependence on the problem. The optimization rule is determined by probability.

2.4 Establish BP Neural Network Prediction Model of PM2.5

This part shows BP neural network training process. First of all, it is worth mentioning that the optimized BP neural network in this paper uses Leaky ReLU as the activation function. Compared with the common sigmoid activation function, it solves the problem of gradient disappearance and has a high convergence value therefore holding higher practical value. At the same time, this function solves the problem that the traditional ReLU is prone to have dead neurons. ReLU sets all negative values to zero. In contrast, Leaky ReLU assigns a non-zero slope to all negative values. The image of the Leaky ReLU activation function is shown as figure 2:

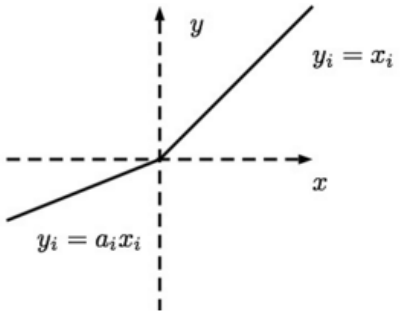

Figure 2. Leaky ReLU activation function image

Step 1. First, the original data is reduced in dimension using PCA. The GA is used with obtained principal components to optimize the initial parameters. The specific process is:

(1) Initialize the population, including cross-scale, crossover probability, and mutation probability, using real number coding. As the crossover probability increases, the randomness of the population begins to increase. As the probability of mutation increases, the diversity of the population begins to expand. As the population size increases, the risk of the algorithm falling into the local optimal solution decreases[6].

(2) Calculate the individual's fitness function, and regard the sum of the absolute errors between the predicted output value and the target output value as the value of the individual fitness function.

(3) Comparing the fitness function values, select individuals with large fitness function values, and directly inherit them to the offspring.

(4) Perform crossover and mutation operations to generate new progeny.

(5) Repeat steps (2) to (4) to continuously evolve the initial weights and thresholds until the conditions of the training objectives are met, and the optimal weights and thresholds are obtained.

Step 2. Initialize the network, including the weights ω_{ij}, ω_{ik} of the hidden layer and the output layer, the hidden layer, the output threshold a, b, the network learning rate, and the activation function between the neurons (Leaky ReLU).

Step 3. Calculate the hidden layer output value. According to the input variables X_1, X_2 , … , X_n, the weight ω_{ij} between the input layer and the hidden layer, and the hidden layer threshold a, the output variable H_j of the hidden layer can be calculated. The calculation formula is as follow:

$$H_j = f(\sum_{i=1}^{n} \omega_{ij} x_i - a_i) \quad j = 1,2, \ldots,l \tag{1}$$

Where l is the number of hidden layer nodes in the BP neural network, and f is the hidden layer activation function.

Step 4. The output layer variables are calculated. The output layer variables Y_m can be calculated according to the hidden layer variables H_j, the hidden layer and the output layer weight ω_{ik}, and the output layer threshold b.

$$Y_k = \sum_{i=1}^{l} \omega_{ij} H_i - b_i \quad ,k=1,2\ldots,m \tag{2}$$

Step 5. The error value is obtained, and the error value is calculated according to the predicted value of the neural network and the reference value, and the formula is as follow:

$$e_k = Y_k' - Y_k \quad ,k=1,2\ldots,m \tag{3}$$

Step 6. Update the neural network weights and the thresholds, and adjust the network weights and the thresholds according to the calculated error value e_k, and the calculation formula is as follow:

$$\omega_{ik} = \omega_{ij} + \eta H_j(1 - H_j)x(i) \sum_{i=1}^{l} \omega_{ij} e_k \quad ,i=1,2\ldots,n; \ ,j=1,2\ldots,l \tag{4}$$

$$\omega_{jk} = \omega_{ik} + \eta H_j e_k \quad ,j=1,2\ldots,l; \ ,k=1,2\ldots,m \tag{5}$$

$$a_j = a_j + \eta H_j(1 - H_j)x(i) \sum_{i=1}^{l} a_j e_k \quad ,j=1,2\ldots,l \tag{6}$$

$$b_k = b_k + e_k \quad ,k=1,,2\ldots,m \tag{7}$$

Where η is the learning rate of the neural network.

Step 7. Determine whether the network calculation is finished according to the judgment condition. If not, continue to repeat the calculation from step 3 until the judgment condition satisfies the value and stops the calculation [7]. In order to increase the accuracy of the training and avoid meaningless iterations, the loss function is used as the termination basis during training, and the iteration is terminated when the error of the verification data is higher than the threshold.

3. Analysis of Experimental Results

3.1 PM2.5 Training Data Set

The following Table shows the neural network training and testing data sets used in this paper. The 39 sample data from 8 meteorological observation points in Chengdu in February 2019 include PM2.5 concentration, humidity, and wind speed. In order to improve the accuracy of the model and the accuracy of the prediction, we divide the data into two broad categories. One is the sample training data set and the other is the sample verification data set, and the ratio is 77% and 23% between these two parts. In detail, includes 30 training data and 9 verifying data. The maximum value of the sample is 120.0 μm/m^3. The minimum value is 31.0 μm/m^3, and the average value is 67.9 μm/m^3. The maximum value of the verification sample is 89.0 μm/m^3, the minimum value is 49.0 μm/m^3, and the average value is 70.3 μm /m^3.

Table 1. Daily average mass concentration of PM2.5 in Chengdu urban area

sample	number	Concentration (ug/m^3)	sample	number	Concentration (ug/m^3)	sample	number	Concentration(ug/m^3)
Training samples	1	51	**Training samples**	1	48	**Training samples**	1	88
	2	36		2	59		2	74
	3	58		3	40		3	75
	4	72		4	31		4	70
	5	69		5	69		5	56
	6	120		6	65		6	64
	7	114		7	69		7	68
	8	75		8	45		8	87
	9	75		9	43		9	95
	10	74		10	64		10	83
Test samples	11	62	**Test samples**	11	69	**Test samples**	11	75
	12	79		12	89		12	67
	13	49		13	73		13	70

Among all eight observation sites, the humidity and wind speed parameter values XP, DP at various points in the Chengdu area monitored by four A, B, C, and D meteorological monitoring stations were selected as the initial values of the indicators. The coefficients are shown in Table 2:

Table 2. Initial coefficient values of humidity and wind speed

	A	B	C	D
humidity	2.01	-1.76	-0.65	1.78
wind speed	1.84	-1.93	-005	1.20

3.2 Forecast Results and Analysis

Table 3. Analysis of prediction effects

	February 1	February 2	February 3	February 4	February 5	February 6
traditional BP-NN MAE	0.1124	0.1396	0.1925	0.2479	0.2893	0.3029
optimized MAE	0.0893	0.0967	0.1321	0.1674	0.1862	0.2125
traditional accuracy	75.0%	66.3%	58.2%	42.1%	30.2%	25.0%
optimized accuracy	81.5%	73.1%	67.5%	51.3%	40.4%	33.8%

From Table 3, we can see the results of the optimized BP neural network for PM2.5 prediction in Chengdu compared with a traditional BP-NN model, and as time increases, the absolute error increases and the predicted coincidence rate decreases. However, the improved BP neural network has better accuracy and more robust maintenance.

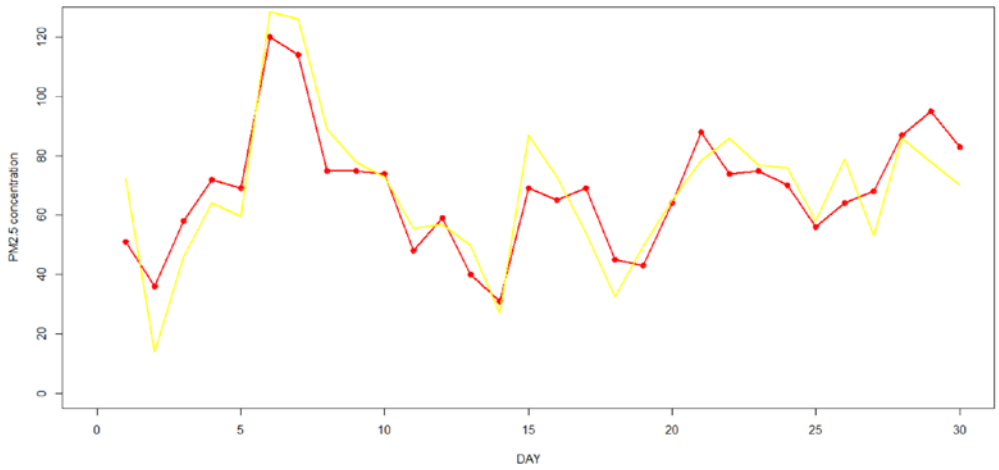

Figure 3. Comparison of 30-day predicted and actual values

In the figure 3, the red polyline is the actual fluctuation value. The yellow polyline is the predicted fluctuation value.

4. Conclusion

The experimental results prove that BP neural network can be applied in the field of meteorological prediction well, and the optimized network model can improve the accuracy of about 8.7% compared with the traditional BP neural network model. However, even the optimized model will reduce the prediction reliability as the time period increases. Future research will focus on further improving the accuracy and not reducing the accuracy significantly over time.

References

[1] Wang Zhenbo, Fang Chuanglin, Xu Guang, et al. Temporal and spatial variation of PM2.5 concentration in Chinese cities in 2014[J]. Acta Geographica Sinica, 2015, 70(11): 1720-1734.

[2] Xie Yonghua, Zhang Mingmin, Yang Le, et al. Prediction of urban PM2.5 concentration based on support vector machine regression[J]. Computer Engineering and Design, 2015(11): 3106-3111.

[3] Feng Wei, Liu Ge, Huang Yong, et al. Prediction of PM2.5 Concentration in Tianjin Based on BP Neural Network[J]. Environmental Science and Management, 2016, 41(6): 121-125.

[4] Zhang Jing. Prediction of PM2.5 concentration in Shenyang urban area based on BP neural network [A]. Chinese Meteorological Society. 35th Annual Meeting of Chinese Meteorological Society S12 Atmospheric composition and weather, climate change and environmental impact and

environmental meteorological forecast Impact Assessment [C]. Chinese Meteorological Society: Chinese Meteorological Society, 2018: 4.

[5] Wei Yunyun. Analysis and Prediction of Agricultural Industry Based on BP Neural Network Based on Genetic Algorithm[J].Journal of Science of Science and Technology,2018,38(09):15-19.

[6] Wang Deming, Wang Li, Zhang Guangming. Short-term wind speed prediction model based on genetic algorithm BP neural network [J]. Journal of Zhejiang University: Engineering Edition, 2012(6): 837-841.

[7] Liu Ling, Song Malin. Analysis and Prediction of PM2.5 Concentration in Nanjing Based on ARMA Model[J]. Journal of Zaozhuang University, 2016, 33(2): 54-62.