Research on air quality prediction method based on GA-BP model

Ziqing Zhang[a], Ning Ma[b,*]
[a]Department of Information Engineering, Shandong Vocational Institute of Fashion Technology, Tai'an, 271000, China; [b]College of Economics and Management, Shandong Agricultural University, Tai'an, 271000, China

ABSTRACT

With the development of China's economy, the environmental quality is deteriorating, and the problem of air pollution has become particularly prominent. People's high quality of life is closely related to air pollution. Air quality information is information that people will inevitably pay attention to every day. Therefore, research on air quality prediction methods is of very practical significance for revealing the changing laws of urban air quality, grasping air quality, and guiding people's travel and lifestyle. This paper takes Beijing's $PM_{2.5}$ pollution as an example to study air quality prediction methods. Firstly, analyzing the correlation between air pollutant concentration and meteorological factors, establishing a GA-BP pollutant concentration prediction model with meteorological factors and historical pollutant concentration as input factors, and verifying GA-BP through a comparison experiment with the standard BP prediction model. Subsequently, based on the GA-BP pollutant concentration prediction model, a progressive prediction method was proposed, and the concentration prediction process of $PM_{2.5}$ was used to predict the concentration of other five air pollutants. Based on the prediction of pollutant concentration, it refers to the calculation method of the air quality index to predict the AQI and AQI level. Comparing the predicted level with the actual level, verifying the feasibility and accuracy of the prediction method, establishing an air quality prediction system with GA-BP hybrid algorithm as the core.

Keywords: Air quality prediction, correlation of influencing factors, GA-BP hybrid algorithm

1. INTRODUCTION

In China, economically developed areas such as the Beijing-Tianjin-Hebei region, often suffered large-scale, long-term continuous air pollution, which has caused serious social impacts. As a typical northern city, Beijing has a serious problem of urban air pollution. At the same time, as a capital city with a large population, when air pollution breaks out, it will not only cause adverse effects on the health of urban residents, but also cause widespread concern and cause huge problems.

As a key content of air quality management, air quality forecasting is closely related to people's lives. At present, China uses the air quality index (AQI) as the air quality standard. According to the calculation method of the AQI, the AQI is determined by the maximum value of the IAQI, and the air quality level is greatly affected by a single pollutant. Therefore, the key to predicting the AQI or even the level is to predict the concentration of air pollutants.

2. BASIC THEORY AND DATA SOURCES

2.1. The AQI system and related standards

The air quality level reflects the level of air pollution. Generally speaking, the lower air quality level reflected the lower concentration of air pollutants. Currently, China uses the Air Quality Index (AQI) to evaluate air quality. The AQI, reference the concentration of 6 major pollutants when calculating. They are fine particulate matter ($PM_{2.5}$), inhalable particulate matter (PM_{10}), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), and carbon monoxide (CO). The AQI is the maximum of the six pollutant sub-indexes, and is divided into six levels according to the size of the index. The AQI index only represents the level of pollution, not the specific concentration value of a certain pollutant. Because the six pollutants involved in AQI have different effects on human health, the six pollutants have different concentration limits in the calculation. Each pollutant has a corresponding Individual Air Quality Index (IAQI). The AQI ranges from 0

*maning0538@foxmail.com

International Conference on Computer Application and Information Security (ICCAIS 2021),
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12260, 122601C · © The Authors.
Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2637417

to 500, and pollutants greater than 100 are excessive pollutants. For example, today’s average $PM_{2.5}$ concentration is 75ug/m^3, then the IAQI of $PM_{2.5}$ is 100. And the IAQI corresponding to the concentration value of 500ug/m^3 is 500. AQI is the maximum value of the IQAI of various pollutants. If the IAQI's largest pollutants are two or more, they are listed as the primary pollutants. The following describes the calculation process of AQI:

Firstly, comparing with the standard (GB3095-2012), obtain the corresponding concentration limit of each pollutant. The six pollutants include $PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, and CO (among which $PM_{2.5}$ and PM_{10} are the 24-hour average concentration). And the IAQI is calculated;

$$\boldsymbol{IAQI_P = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{HI} - BP_{Lo}}\left(C_p - BP_{Lo}\right) + IAQI_{Lo}} \qquad (1)$$

In the formula:

$IAQI_P$-$IAQI$ representing pollutant P;

Cp—Actual measured concentration value of pollutant P;

BP_{Hi}—The high value of the concentration limit close to the actual concentration of pollutant P in the standard;

BP_{Lo}—The low value of the concentration limit close to the actual concentration of pollutant P in the standard;

$IAQI_H$—The $IAQI$ corresponding to BP_{Hi} in the standard;

$IAQI_{Lo}$—The $IAQI$ corresponding to BP_{Loi} in the standard;

Secondly, the one with the largest value selected from the six *IAQIs* is determined as the AQI. When the AQI is greater than 50, the pollutant with the largest IAQI is determined as the primary pollutant;

$$AQI=max\{IAQI_1, IAQI_2,IAQI_3, ...IAQI_n\} \qquad (2)$$

In the formula:

IAQI—the air quality sub-index corresponding to the six pollutants;

n—the number of pollutant items.

Thirdly, refering to the air quality level table to determine the current air quality level and related information.

In summary, the AQI of the day is the maximum value of the IAQI of the day. The daily air quality report cycle is 24 hours, and the time period is 24 hours before zero o'clock of the day. The data uses in this standard includes the 24-hour average concentration of $PM_{2.5}$, the 24-hour average concentration of PM_{10}, the 24-hour average concentration of SO_2, the 24-hour average concentration of NO_2, the 24-hour average concentration of CO, and daily maximum 8-hour average concentration of O_3. The following Table 1 is the relevant situation of the different levels of the AQI.

Table 1. Air quality rating table.

AQI	AQI level	Air quality index type	Color
0-50	1	Excellent	Green
51-100	2	Good	Yellow
101-150	3	Light pollution	Orange
151-200	4	Moderately polluted	Red
201-300	5	Heavy pollution	Purple
＞300	6	Serious pollution	Maroon

2.2. Related theories of genetic algorithm

The genetic algorithm simulates the biological evolution process in nature. In the evolution of natural organisms, genes carry the genetic information of organisms, and the chromosomes composed of genes are the basic units of heredity. In a genetic algorithm, a set of data is set to simulate genes or chromosomes. This set of data represents all possible solutions to a certain problem.

The main links of GA are: encoding and decoding, population setting, adaptability function, genetic operator. The steps to solve the genetic algorithm are as follows:

(1) According to the actual research problem, generate an initial population, which contains all the solutions of the problem.

(2) Determine the adaptability function of the population.

(3) Perform three operations in the population: mutation, selection, and crossover to continuously produce solutions

(4) Compare the adaptability of newly emerged individuals. If it meets the qualification, the solution with the highest adaptability in genetics is the approximate solution; if the conditions are not met, the iteration is continued.

2.3. Related theories of BP neural network

BP neural network, the full name is Error Back Propagation Neural Network (BPNN). The network simulates the neural network of the creatures in nature to conduct information transmission. By continuously modifying the attribute values of the network, the output value can meet the conditions, and finally the data can be accurately fitted[1, 2]. BP neural network has many characteristics, this research uses BP neural network mainly based on its self-learning and adaptability.

The working process of BP neural network is mainly divided into two parts. The first part is the forward input of the sample data, and the second part is the reverse propagation of the error.

Firstly, the forward conduction of the sample data. The input samples enter the network from the input layer, and then processed by the hidden layer, finally output by the output layer. The conduction direction of the input sample is positive, so this stage is called the forward conduction stage. Subsequently, compared the data of the output layer with the expected input, calculated the mean square error of the them , and transmitted the mean square error backwards along the direction of the input layer, so as to obtain the reference error of each layer of neurons, and this reference error as evidence for adjusting the weights or thresholds of each hidden layer unit. This is a complete learning process. The neural network continues to perform this process until the error meets the requirements of the problem[3].

Although BP neural network has many advantages, there are still some problems, for example: the initial value of the network is difficult to determine. The network initial value has clear impact on the efficiency and accuracy of the learning process. However, current determination of the initial value is mainly based on past experience and there is no scientific method. In response to this problem, this article tries to optimize the initial value through GA, by the characteristics of genetic algorithm, reduces the number of neural network training and improves efficiency.

2.4. Data sources

In this paper, the air pollution-related data used mainly come from the following sources: (1) the monitoring data from monitoring stations, (2) network acquisition, including related websites such as “China Weather Network”, “Zhenqi Network”, etc. (3) acquisition of relevant literature data, including “China Statistics Yearbook”, “China Environmental Status Bulletin”, “Beijing Environmental Status Bulletin” and other relevant materials. And the meteorological data come from the webside “China Meteorological Data Network” and “Reliable Prognosis”. The data format is the hourly monitoring value from 0-23 o’clock every day. There are many monitoring items. In this study, only the relevant data of common meteorological factors are taken, including: temperature, humidity, wind speed, air pressure, precipitation and other items.

Most of the air quality data comes from historical data recorded by the urban air quality inspection system, in the form of hourly monitoring data for each monitoring point. To conduct an overall analysis of Beijing’s air quality, the daily average data and annual average data of air quality need to be used, which requires preprocessing of the data. The monitoring station monitors the data every hour, and averages the 24 sets of daily data to obtain the daily average data. At the same time, for meteorological data, similar daily average value processing operations are also carried out, which is convenient for data application.

Because some monitoring points cannot be monitored due to regular maintenance, equipment abnormalities, network abnormalities, etc., some data will be missing. The missing and abnormal data will have a certain impact on the analysis results during statistical calculations, so the missing data must be dealt with. Since the total amount of abnormal points in the data is small, and all the data are numerical data of monitoring pollutant concentration values or numerical data of meteorological factors, the average value filling method can be used to supplement the data of abnormal points. This ensures the quality of the data in the research and reduces the impact of data quality issues on the research results.

3. CORRELATION ANALYSIS OF POLLUTANT CONCENTRATION AND METEOROLOGICAL FACTORS

3.1. Theoretical basis of the correlation between pollutant concentration and meteorological factors

There is a certain correlation between pollution sources, geographical environment, seasons, weather and changes in the concentration of air pollutants[4]. After consulting relevant resource, it is found that the correlation between air pollutants and meteorological elements is the strongest, and there is obvious a non-linear relationship[5]. Therefore, in this study, the correlation of meteorological factors is considered when air quality prediction is made. General meteorological data includes the following categories: temperature, air pressure, evaporation, precipitation, relative humidity, sunshine hours, wind direction and speed and 0cm ground temperature. Based on existing research, the five factors of relative humidity, temperature, air pressure, wind speed, and precipitation are mainly considered for the climate and pollution in Beijing. The following points are mainly explained here:

(1) Air pollution of moderate pollution level in Beijing mostly occurs in autumn, winter and spring. In these seasons, according to the climatic conditions in Beijing, winter is cold and dry, hot and rainy in summer. The concentration of $PM_{2.5}$ is obviously affected by precipitation[6]. Therefore, when considering meteorological factors, precipitation factors are indispensable.

(2) The reason why only wind speed is considered is that the size of the wind has a significant impact on the diffusion of particulate pollutants. For Beijing, no matter which direction the wind originates from, it will not affect the particle pollutants. Therefore, the influence of wind direction on the diffusion is relatively small. In this study, only the influence of wind speed on pollutants is considered.

(3) The number of sunshine hours, evaporation and 0cm ground temperature, with reference to related studies, it has not been found that these meteorological factors have direct or indirect effects on changes in pollutant concentration[7].

The meteorological data used in this correlation study comes from historical monitoring data. Among them, the measured meteorological data has 28 categories. In this paper, we use only five factors, including: wind speed, relative humidity, air pressure, precipitation and temperature, which are selected.

3.2. The influence of meteorological factors on $PM_{2.5}$

Summarizing the meteorological data and the pollutant data of the same period, we can get a scatter plot. The following Figure 1 is a scatter plot of the $PM_{2.5}$ concentration and temperature in 2016.

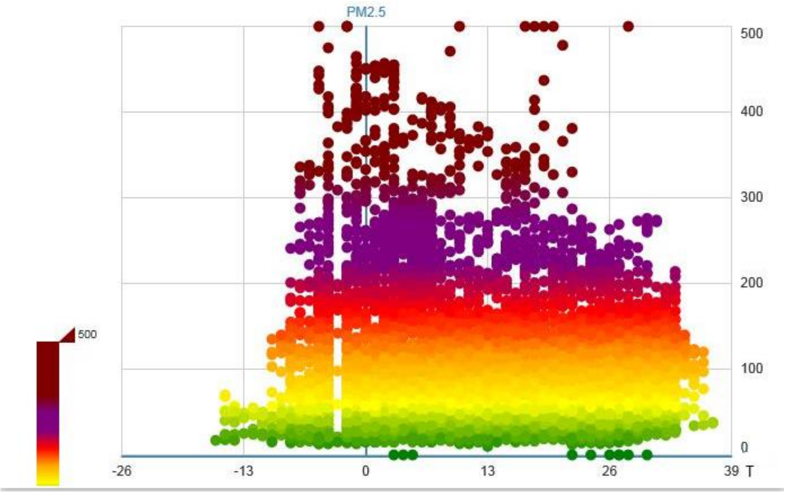

Figure 1. $PM_{2.5}$ and temperature scatter diagram.

The horizontal axis is the temperature, and the vertical axis is the $PM_{2.5}$ concentration. Figure 1 shows that the $PM_{2.5}$ are mainly concentrated between -13°C and 29°C, and most of the high-concentration points are concentrated between -7°C and -7°C. Between 26°C. At the same time, refer to the fitted images of the line graph of the daily mean value of $PM_{2.5}$ concentration in 2016 and the line graph of the daily mean value of temperature in 2016.

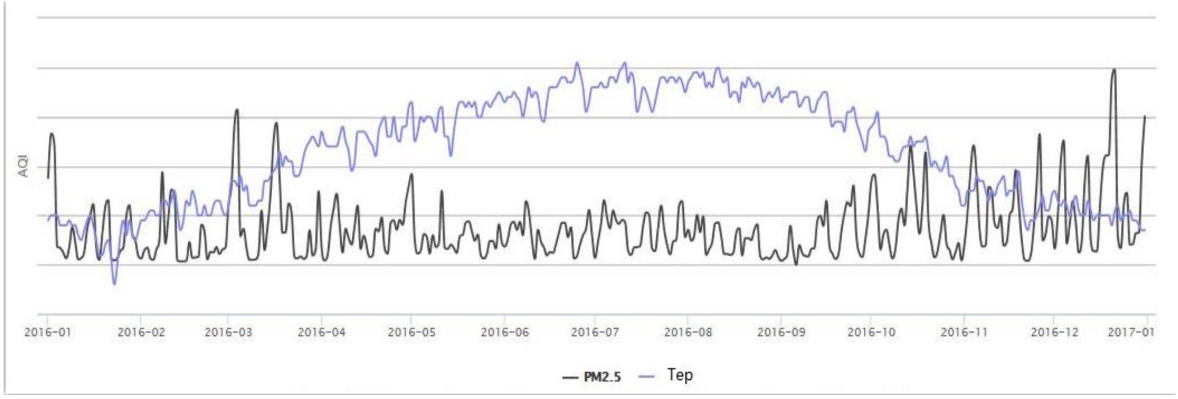

Figure 2. $PM_{2.5}$ and temperature time change curve.

In Figure 2, the horizontal axis is time, and the vertical axis is the average daily concentration of $PM_{2.5}$ and the average daily temperature. It can be seen from Figure 2 that there is a strong similarity with the change trend of the curve of $PM_{2.5}$ daily average concentration and daily average temperature[8, 9]. Basically, it can be judged that the two are related. The SPSS 21 was used to analyze the correlation, and the Pearson correlation coefficient was used as a reference at this time.

The first thing that is obtained is the normality test result between the $PM_{2.5}$ concentration and the average temperature. According to the shapiro-wilk test, it can be seen that the *P* is greater than 0.05. From this, the two variables conform to the normal distribution.

Subsequently, through analysis, the Pearson coefficient, that is, the *P* value, is obtained, *P*=-0.8467. From the above, it can be seen that the *P* value is between (-1, 1). The closer it is to -1, the stronger the negative correlation between the two. Judging by the *P* value, there is a negative correlation between the temperature and the $PM_{2.5}$ concentration. The correlation is strong.

In the same way, the $PM_{2.5}$ concentration is tested against relative wind speed, precipitation, air pressure and humidity. The Pearson coefficient of the $PM_{2.5}$ concentration and the four meteorological factors of relative humidity, wind speed, air pressure and precipitation can be obtained as shown in Table 2 below.

Table 2. Correlation coefficients between PM2.5 and various meteorological factors.

Name	Average temperature	Relative humidity	Wind speed	Air pressure	Precipitation
$PM_{2.5}$	-0.847^{**}	-0.594^{*}	-0.654^{**}	0.334^{*}	-0.494^{*}

Note: ** means passing the 0.01 confidence level test (two-sided test); * means passing the 0.05 confidence level test (two-sided test).

In Table 2, the negative correlation between $PM_{2.5}$ concentration and average temperature is the strongest. And the wind speed, precipitation, relative humidity all negatively correlated with $PM_{2.5}$ concentration, while air pressure is positively correlated with $PM_{2.5}$ concentration.

When the temperature increases, the vertical convection activity in the troposphere is strengthened, which is conducive to the diffusion of pollutants. When the temperature is low, the slowing down of convection will lead to long-term retention of pollutants. At the same time, for particulate matter, the increase in temperature helps promote its Brownian motion[10]. The influence of relative humidity on particulate matter is mainly manifested in that when the relative humidity increases, the particulate matter is easily surrounded by moisture to form large particles, which are easy to settle toward the surface[11]. The density of $PM_{2.5}$ particulate matter decreases due to the increase in moisture. The wind can dilute the pollutants, wind can also help the diffusion and transportation of the pollutants. Air pressure is positively correlated with fine particulate matter $PM_{2.5}$, mainly because when air pressure rises, it means that air convection activity is reduced, and a downdraft is formed, which leads to the accumulation of pollutants and is not easy to diffuse. Precipitation has a very significant effect on the concentration, mainly due to the erosion of precipitation, which can remove pollutants in the air.

In summary, the concentration of $PM_{2.5}$ is closely related to meteorological factors. So these five meteorological factors can be included in the parameter variables of the prediction model.

The $PM_{2.5}$ concentration of the day has a more exact correlation with the $PM_{2.5}$ concentration of the fine particulate matter of the previous day. Therefore, this study established an air quality prediction model that combines historical air pollutant concentration data with meteorological factors.

3.3. Correlation between other pollutants and meteorological factors

Through the same steps, the correlation between other pollutants and meteorological factors is analyzed. Firstly, normality test is performed, and the test result is in accordance with the normal distribution. After that, calculate the Pearson correlation coefficient between pollutants and meteorological factors. The following Table 3 shows the correlation coefficient.

Table 3. Correlation coefficients of other pollutants and meteorological factors.

Name	Average temperature	Relative humidity	Wind speed	Air pressure	Precipitation
PM_{10}	-0.8262^{**}	-0.5235^{*}	-0.634^{**}	0.4026^{*}	-0.5128^{*}
SO_2	-0.8071^{**}	-0.4124	-0.2133^{*}	0.3462^{*}	-0.5201^{*}
NO_2	-0.8242^{**}	-0.164	-0.254^{*}	0.5301^{**}	-0.323^{*}
CO	-0.602^{**}	0.1903	-0.313^{*}	0.406^{*}	-0.217^{*}
O_3	0.692^{**}	-0.5513^{*}	0.4524^{*}	-0.612^{**}	0.1343

Note: ** means passing the 0.01 confidence level test (two-sided test); * means passing the 0.05 confidence level test (two-sided test).

Since PM_{10} and $PM_{2.5}$ are homologous and have the same properties, they show the same characteristics in terms of the correlation between meteorological factors[12]. The negative correlation between PM_{10} and average temperature is the strongest, followed by relative humidity, wind speed and precipitation all negatively correlated with PM_{10} concentration, while air pressure is positively correlated with PM_{10} concentration.

Among gaseous pollutants, SO_2 has a negative correlation with wind speed, precipitation, relative humidity and temperature. And temperature has the strongest correlation with it; it has a positive correlation with air pressure, but the relationship is weak. NO_2 has a significant negative correlation with relative humidity and temperature, a positive correlation with air pressure and wind speed[13]. CO also has a significant negative correlation with wind speed, temperature, precipitation, a significant positive correlation with air pressure and air pressure[14].

The O_3 has a significant positive correlation with temperature, mainly due to the high temperature weather that promotes the strengthening of the photochemical reaction, which promotes the production of O_3[15]. A significant negative correlation with the relative humidity, because when the relative humidity is high, it means that there is less solar radiation reaching the ground, so the photochemical reaction process is weak, which is not conducive to the formation of O_3. A significant negative correlation with pressure, the higher the pressure, the lower the concentration. There is no significant correlation with precipitation, mainly because there are fewer days of precipitation during the statistical period, and the influencing factors are not as obvious as other meteorological factors. At the same time, it is positively correlated with wind speed. Studies have shown that when the wind speed is lower than 2m/s, it is more conducive to the accumulation of O_3.

In summary, temperature, relative humidity, wind speed, and precipitation are mostly negatively correlated with air pollutants, while atmospheric pressure is positively correlated; the correlation of O_3 is different from other pollutants, which is different from O_3 itself.

4. CONSTRUCTION AND DEMONSTRATION OF AIR QUALITY PREDICTION MODEL

4.1. GA-BP hybrid algorithm design

For the classic BP neural network, it is difficult to set the initial value. GA has the characteristics of global optimization. Combine BP neural network and GA to form a hybrid GA-BP model for training neural network.

In this paper, to simulate the ability of survival of the fittest through GA, set the initial value to individual, perform global optimization in the solution space of population, optimize the initial value. And select the best initial value, which can improve the training efficiency and save training time.

The GA optimizes the BP neural network, mainly for the initial value. The initial value includes: the number of hidden layers, the number of neurons contained in hidden layer, the weights between neurons, and the threshold of neurons. In this study, the case where the number of hidden layers is 1 is considered, so the problem is simplified. Through the following steps, introduce how to optimize BP neural network through GA.

(1) Optimization of coding scheme. Since genetic algorithms cannot directly use the solution space as genes, it's necessary to convert the number of thresholds, connection weights and hidden neurons into the form of chromosomes used by genetic algorithms. The optimization object this time is the initial value. Considering that the thresholds and connection weights are real numbers with high precision, it is difficult to ensure the accuracy using binary coding and the binary coding needs to be decoded. So in this optimization problem, we choose the real number coding scheme.

For the problem of selecting the initial parameters, we set the neurons in the input layer and output layer according to the research problem, which are m and n respectively. And the neurons in the hidden layer are optimized by genetic algorithm, denoted as t. When neurons in the hidden layer changes, the connection weights and the thresholds will also change inversely. According to the principle of neural network universal approximation, for a network structure containing one hidden layer, the more hidden layer neurons, the closer the network will be to any function on the bounded area. Therefore, in this study, in order to facilitate processing, we also adopt a network structure with a hidden layer structure. Then, combined the threshold of a hidden layer neuron and its associated connection weight to form a unified operation coding block. In this way, the code of the problem chromosome can be composed of three parts: the first part has only one real number code, which is used to represent the hidden layer neurons. The second part has a total of x codes, which represent the threshold. The third part is a number of coding blocks composed of the thresholds and connection weights. The number of coding blocks is determined by the hidden layer neurons. In the implementation of the above coding scheme, since the neurons in the hidden layer are not determined, the length of individual coding string is variable. The crossover operation of two parent individuals of unequal length will lead to inconsistencies within the offspring individuals. And the mutation operation may also destroy the integrity of the individual. In this paper, the maximum T allowed by the hidden layer can be set, so the maximum length L of the individual code can be calculated. We set the code length of all individuals to L. Since the hidden layer neurons may not reach the maximum value T, the data must be supplemented for such chromosomes, and 0 supplementation is selected this time.

(2) Setting the population. The population is an initial search space of GA. The population will have a direct impact on the efficiency of the genetic algorithm. Therefore, the general population size will be selected within the range of 20-100 practical experience values.

(3) Setting of adaptability function. In the evolutionary search, GA is based on adaptability function only, and no other auxiliary information is needed. The adaptability can reflect the degree to which the individual reaches the optimal solution. The adaptability function directly affects the convergence speed of the GA. In the process of optimizing the neural network by GA, the error function is used to construct the adaptability function of the GA. Combining the convergence of neural network with the evolution of GA.

(4) Genetic operator settings. The genetic operator is the basic means for the population to generate new individuals. In the process of optimizing the BP by GA, the genetic operator directly affects the generation of the initialization parameters of the neural network. Therefore, setting up appropriate selection, crossover, and mutation operators is conducive to generating optimal individuals as soon as possible and reaching the standard of adaptability function.

The above is the optimization method of the GA to the BP. After the genetic algorithm is optimized, the solution with the maximum adaptability is obtained, which is the optimal solution. According to the coding rules, decoding is performed to obtain the initial thresholds and weights. From sample set, select one and input it into the neural network for training, and compare the target samples according to the output.

4.2. Network structure settings

For air pollutant concentration prediction model, GA is directly applied to optimize the network. The network optimized by GA is applied to the air pollutant concentration prediction, which increases the model's ability to deal with complex problems. And, the GA is used to optimize the network to obtain the initial value, which can effectively improve the efficiency of the prediction algorithm.

According to requirements, the GA-BP pollutant concentration prediction algorithm has 6 neurons in the input layer, corresponding to the concentration of fine particulate matter $PM_{2.5}$, wind speed, average temperature, air pressure, relative humidity and precipitation the previous day. And output vector is the $PM_{2.5}$ concentration, so 1 output neuron is used. The network structure is a single hidden layer structure.

When setting network related parameters, the neurons in the hidden layer, weights and thresholds are calculated by GA. The momentum factor is set to 0.5, the maximum number of learning times is 100,000, learning rate is 0.1, and the target learning error is 0.001. The relevant parameters of GA are: population size is 40, the number of evolutions is limited to 200, and the maximum number of hidden layer neurons is limited to 50. Table 4 shows the relevant information.

Table 4. GA-BP related parameters.

Name	Value	Name	Value
Input neurons	6	Learning times	10000
Output neurons	4	Weight	GA optimized
Number of hidden layers	1	Learning rate	0.1
Hidden layer neurons	Max 50	Expectation error	0.001
Threshold	GA optimized	Momentum factor	0.5
Population	40	Evolutional generation	200
Cross rate	0.5	Mutation rate	0.08

4.3. GA-BP network training process and results

The model construction process is as follows:

(1) Encode the chromosomes of GA in real-number, the information includes hidden neurons, thresholds and weights.

(2) Set the relevant parameters: population size, evolutionary algebra.

(3) Normalize the manipulated data.

(4) Execute GA to select the possible solution.

(5) Assign the possible solution of GA to BP.

(7) Input training samples into the air pollutant concentration prediction model.

(8) Train the model until the error range is reached.

(9) Input the test sample into the model and get the output, then perform the denormalization operation on the data to get the corresponding prediction data.

After completing the modeling through the above process, use the sample data for network training, and finally get the corresponding output results. The following Table 5 is part of the predicted data and monitoring data obtained by the GA-BP predict model.

Mean square error (MSE) = 26.77621%.

Average absolute percentage error (MAPE) = 24.821%.

The accuracy rate is 75.179%. The error curve between the predicted concentration and the actual monitored concentration is obtained, as shown in Figure 3.

Table 5. Some test sample data of GA-BP prediction model.

Test sample number	Actual concentration (ug/m^3)	Predicted concentration (ug/m^3)
1	27	31
2	92	113
3	191	210
4	246	268
5	49	97
6	75	115
7	120	162
8	91	108
9	21	40
10	46	80

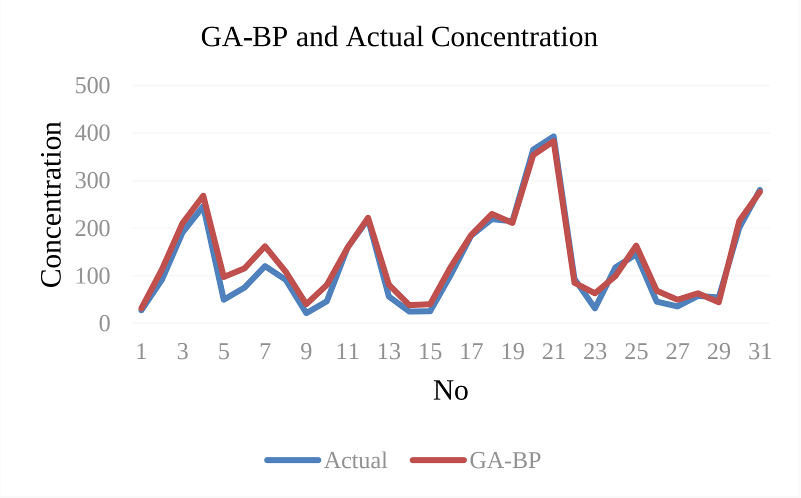

Figure 3. Curve of GA-BP predicted data and actual data.

Comparing the change trend of the monitoring point and the test, it is found that in the overall trend, the change of the predicted and the monitored is the same, which is in line with the change trend of PM$_{2.5}$. Therefore, it can be judged that the model is feasible to make prediction. Comparing the predicted results, the air pollutant concentration prediction reached a higher level.

4.4. Standard BP algorithm comparison experiment

In this study, the standard BP algorithm was selected as the comparison model to conduct a prediction comparison experiment, and the results obtained by the standard BP algorithm for air pollutant concentration prediction were compared. The accuracy and efficiency of the experimental results were compared. The three-layer structure was used in the standard BP algorithm to predict data, that is, the structure of input layer-single hidden layer-output layer. Among them, there are 6 neurons in the input layer, which same as the GA-BP model. The hidden layer still chooses one layer of structure. Based on experience, the number of hidden layer neurons is selected as 5. The basic modeling process is as follows:

(1) Set up the BP model, set the error range, learning rate, and momentum factor.

(2) Normalize the sample data.

(3) Input the training samples in the sample data into the network, and train by yourself until the error range is reached.

(4) After the network training, input the sample and denormalize the output.

Compare the actual measured data with the model predicted data. After completing the modeling through the above process, use the sample data for network training, and finally get the corresponding output results. The following Table 6 is the results of the prediction experiment conducted on the test samples of the BP neural network air pollutant prediction model.

Table 6. Some test sample data of BP prediction model.

Test sample number	Actual concentration (ug/m^3)	Predicted concentration (ug/m^3)
1	27	28
2	92	109
3	191	206
4	246	272
5	49	91
6	75	104
7	120	152
8	91	37
9	21	78
10	46	78

After calculation, the following information can be obtained: Mean square error (MSE)=28.76%; Average absolute percentage error (MAPE)=28.6379%; The accuracy rate is 71.3621%.

4.5. Analysis of experimental results

By comparing the line graphs of the predicted data of the two prediction models and comparing the concentration change trends, the results show that the two prediction models can predict the concentration, but the GA-BP model is closer to the actual monitored data. Comparing the accuracy, MSE and MAPE of the two models, it is found that the GA-BP model has advantages in these aspects[16]. Therefore, we can judge that the GA-BP prediction model is more suitable for predicting the concentration. Figures 4 and 5 below show the comparison between the results.

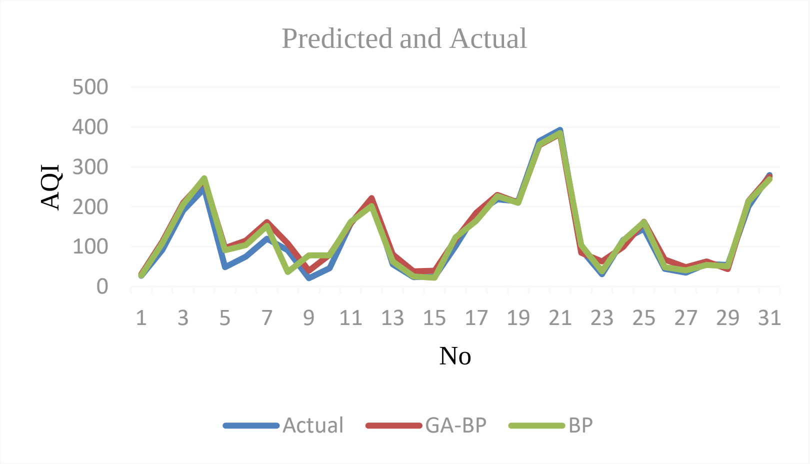

Figure 4. Two types of predicted data and actual data.

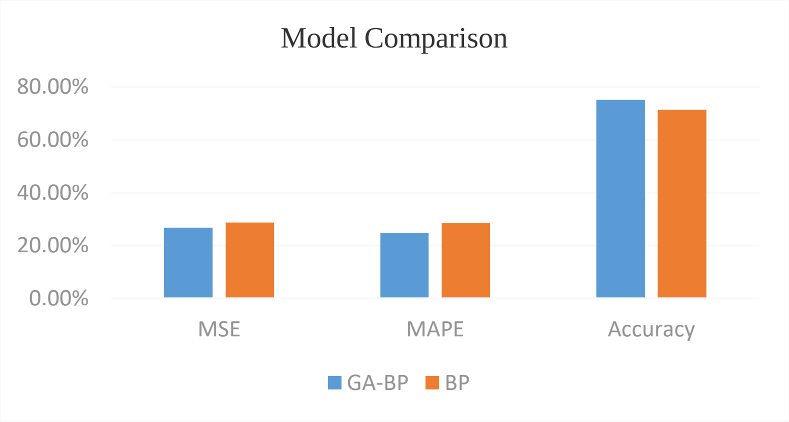

Figure 5. GA-BP model and BP model performance comparison.

Subsequently, a performance comparison was made. For the BP model, the training experienced 15.14s to reach the error range with an error of 0.0153, while the GA-BP network was optimized by genetic algorithm and reached the convergence range with 9.63s, with an error of 0.0104. The error is smaller than the BP model, and the time is also faster than the BP network prediction model. Therefore, both in terms of time and error range, the GA-BP prediction model is much better. Figure 6 shows the training error and time variation curves of the two prediction models.

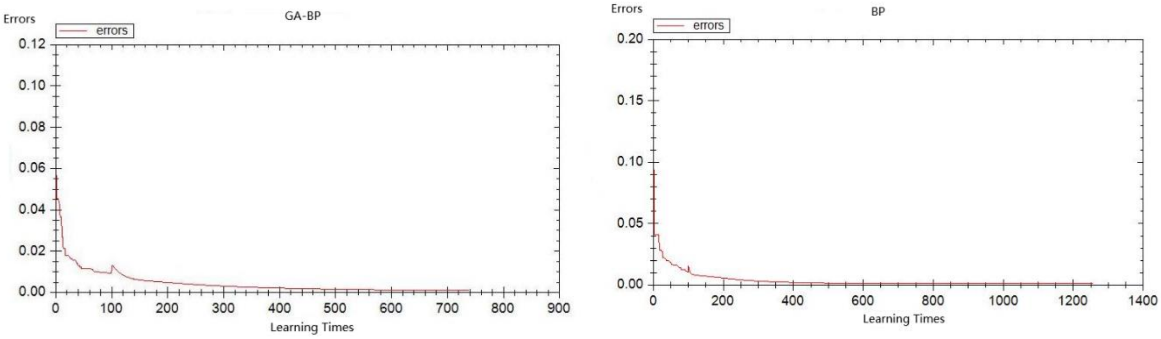

Figure 6. GA-BP and standard BP learning curve.

4.6. Research on air quality level prediction

4.6.1. Air quality index predict. After it is clear that each pollutant has the same significant correlation with meteorological factors, the GA-BP air pollutant concentration prediction model constructed through a similar process can be used. The above introduced the GA-BP hybrid algorithm pollutant concentration prediction model, and by setting the BP model comparison experiment, it showed that the GA-BP model is more suitable for air pollutant concentration prediction.

Six input factors including the previous day's pollutant concentration data, predicted daily temperature, predicted daily relative humidity, predicted daily precipitation, predicted daily wind speed, and predicted daily air pressure are used to output the predicted concentration of pollutants. The data samples are meteorological data and air quality monitoring data in the same period as the $PM_{2.5}$ model, and the training samples and test samples are divided at the same time node. After successfully constructing the GA-BP prediction model, train it until the error range is met[17]. Subsequently, input the simulation data into the prediction model to get the result. The following Table 7 shows the corresponding data obtained by the GA-BP concentration prediction model for various pollutants.

Table 7. Concentration data predicted by GA-BP.

No	$PM_{2.5}$ ug/m^3	PM_{10} ug/m^3	SO_2 ug/m^3	NO_2 ug/m^3	CO ug/m^3	O_3 ug/m^3
1	104	160	6.4	42	1.85	37.5
2	38	64	7	39.7	1.16	25.9
3	115	111	24	71.4	1.89	9.95
4	162	176	29	90.2	2.85	4.86
5	49	58	16.5	39.7	0.94	30.5
6	41	51	9.2	32	0.70	31.8
7	54	77	8.7	41.6	1.16	26.7
8	51	75.4	11.2	47	0.98	19.0
9	214	260	18.5	101	3.18	3.58
10	269	316	21.1	122	4.39	4.07

4.6.2. Air quality level predict. As mentioned, the AQI is determined by the IAQI of each pollutant. Therefore, after predicting the predicted concentration of all pollutants through the GA-BP prediction model, calculating the corresponding IAQI. The following Table 8 is the IAQI of the corresponding pollutants.

Table 8. IAQI index table.

No	$PM_{2.5}$ IAQI	PM_{10} IAQI	SO_2 IAQI	NO_2 IAQI	CO IAQI	O_3 IAQI
1	186.25	105	6.4	52.5	46.25	52.08
2	53.75	57	7	49.625	29	12.95
3	150	80.5	24	89.25	47.25	4.975
4	217	113	29	112.75	71.25	2.43
5	67.5	54	16.5	49.625	23.5	15.25
6	57.5	50.5	9.2	40	17.5	15.9
7	73.75	63.5	8.7	52	29	13.35
8	70	62.7	11.2	58.75	24.5	9.5
9	264	155	18.5	51.25	79.5	1.79
10	319	183	21.1	77.5	51.95	2.035

The AQI corresponding to each pollutant is obtained through the IAQI calculation formula, which can further determine the AQI of the predicted day. Therefore, AQI can be determined on the predicted day through a simple comparison. Compare the predicted AQI with the actual AQI. The following Table 9 is the comparison between the AQI calculated by IAQI and the actual data.

Table 9 showed that the AQI obtained through prediction is still different from the actual AQI, but the error is not large. Therefore, it can be judged that the method of predicting the concentration and AQI through the GA-BP air pollutant concentration prediction model is feasible.

Among all 10 predicted AQI data, 9 of them are determined by the IAQI of $PM_{2.5}$, and the other (number 2) is determined by the IAQI of PM_{10}. Therefore, it can be judged that the primary pollutant on 9 of these 10 days is $PM_{2.5}$, and the primary pollutant on the other day is PM_{10}. This conclusion is consistent with the actual situation. $PM_{2.5}$ has a huge impact on Beijing's air quality.

Table 9. Comparison of predicted results and actual AQI levels.

No	Predicted AQI	Actual AQI	Predicted air quality level	Actual air quality level
1	186	151	Moderately polluted	Moderately polluted
2	57	47	Good	Excellent
3	150	155	Light pollution	Moderately polluted
4	212	191	Heavy pollution	Moderately polluted
5	67.5	64	Good	Good
6	57.5	54	Good	Good
7	73.75	85	Good	Good
8	70	75	Good	Good
9	264	249	Heavy pollution	Heavy pollution
10	319	329	Serious pollution	Serious pollution

Figure 7 below shows the predicted AQI and the actual AQI. It showed that there's still a difference in the accuracy of the data, and the trend of changes in the data is consistent.

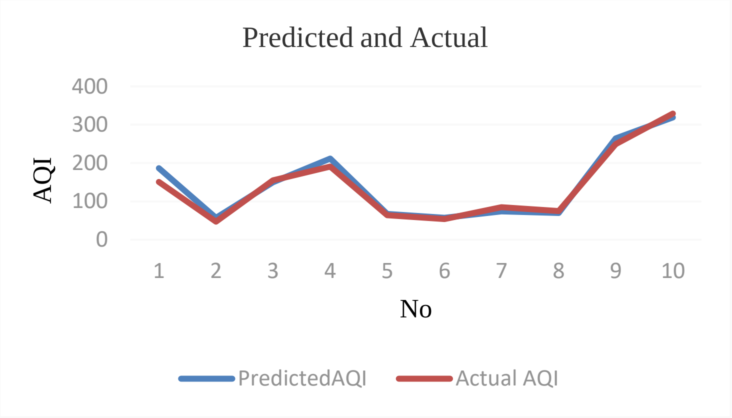

Figure 7. Comparison of predicted AQI and actual AQI.

The air quality level is assessed based on the predicted AQI, and the comparison chart between the predicted level and the actual level is drawn. The horizontal axis is the number, and the vertical axis is the air quality level. Among the ten days, seven days have the same predicted level as the actual air quality level. In the three days, the predicted level is greater than the actual air quality level on two days, and the predicted level is lower than the actual level on 1 day. The preliminary estimation accuracy rate is 70%. In the 3 days when the predict does not match, the predicted level differs from the actual level by one level, so the predicted result also has a certain reference value. Figure 8 shows the predicted level and actual level.

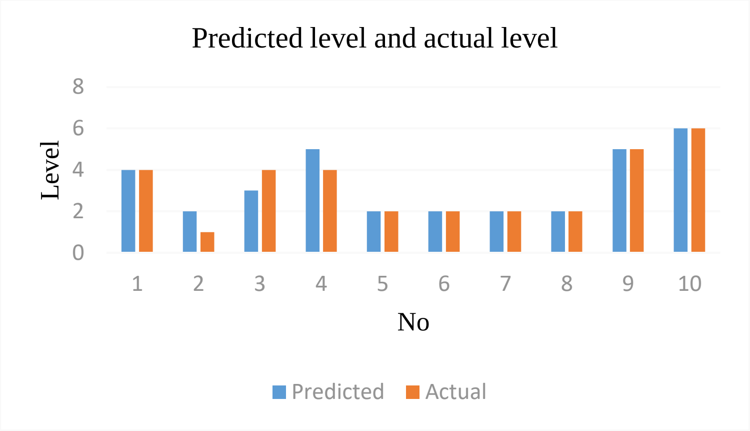

Figure 8. Comparison of predicted level and actual level.

5. CONCLUSION

In this research, by analyzing the correlation between air pollutant concentration and meteorological factors, taking $PM_{2.5}$ as an example, a GA-BP hybrid algorithm air pollutant concentration prediction model was constructed, and comparative experiments were set to prove the accuracy and efficiency of the GA-BP model much better than the BP model. Proved that the proposed method of predicting the concentration of air pollutants-AQI-air quality level progressive prediction method is feasible. The result can not only provide air quality level information, but also pollutant concentration change information. Through the combination of knowledge of different disciplines, the research ideas have been expanded and good prediction effect has been achieved.

Research on the applicability of prediction models, including climate applicability and regional applicability. Whether the GA-BP model proposed in this article is available in areas with heavy rainfall and other areas with different climatic characteristics can be further explored and explored in future work. The accuracy of the network is based on data training, and it should be continuously trained with new data to strengthen the generalization ability, so that it can gradually update the training samples over time to ensure the accuracy of the prediction.

REFERENCES

[1] Arhami, M., Kamali, N. and Rajabi, M., "Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by monte Carlo simulations," Environmental Science & Pollution Research, 20(7), 4777-4789 (2013).

[2] Pirovano, G., Colombi, C., Balzarini, A., Riva, G. M., Gianelle, V. and Lonati, G., "$PM_{2.5}$ source apportionment in lombardy (Italy): Comparison of receptor and chemistry-transport modelling results," Atmospheric Environment, 106, 56-70 (2015).

[3] Gardner, M. W. and Dorling, S. R., "Neural network modelling and prediction of hourly no_x and no_2 concentrations in urban air in London," Atmospheric Environment, 33(5), 709-719 (1999).

[4] Kwok, R., Fung, J., Lau, A. and Wang, Z. S., "Tracking emission sources of sulfur and elemental carbon in Hong Kong/pearl river delta region," Journal of Atmospheric Chemistry, 69(1), 1-22 (2012).

[5] Yang, Z. and Jian, W., "A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction," Environmental Research, 158(11), 105-117 (2017).

[6] Kaminski, W., Skrzypski, J. and Jach-Szakiel, E., "Application of artificial neural networks (ANNs) to predict air quality classes in big cities," 19th Inter. Conf. on Systems Engineering, 135-140 (2008).

[7] Fang X., Jiang W., Jian W., Zhang N., Liu H. and Xu, T., "Study on the development of numerical model system to predict urban air quality," Acta Scientiae Circumstantiae, 1(24), 111-115 (2004).

[8] Kolehmainen, M., Martikainen, H. and Ruuskanen, J., "Neural networks and periodic components used in air quality forecasting," Atmospheric Environment, 35(5), 815-825 (2001).

[9] Ning, M., Guan, J. H. and Liu, P. Z., "GA-BP Air quality evaluation method based on fuzzy theory,"

CMC-Computers Materials & Continua, 58(1), 215-227 (2019).

[10] Guo, Z., “Forecasting stock indices with back propagation neural network,” Expert Systems with Applications, 38(11), 14346-14355 (2011).

[11] Johnson, M., Isakov, V., Touma, J. S., Mukerjee, S. and Oezkaynak, H., “Evaluation of land-use regression models used to predict air quality concentrations in an urban area,” Atmospheric Environment, 44(30), 3660-3668 (2010).

[12] Ramponi, L., Benedusi, L., Toschi, A. and Pagotto, P., “Criteria for the assessment of air quality levels in homogeneous areas,” International Journal of Environment and Pollution, 40(1/3), 3-9 (2010).

[13] Chattopadhyay, S. and Bandyopadhyay, G., “Artificial neural network with backpropagation learning to predict mean monthly total ozone in Arosa, Switzerland,” International Journal of Remote Sensing, 28(19-20), 4471-4482 (2007).

[14] Varshney, K. and Poddar, K., “Prediction of wind properties in urban environments using artificial neural network,” Theoretical and Applied Climatology, 107(3-4), 579-590 (2012).

[15] Karaca, F., Nikov, A. and Alagha, O., “NN-airpol: A neural-networks-based method for air pollution evaluation and control,” International Journal of Environment & Pollution, 28(3/4), 310 (2006).

[16] Pan, L., Sun, B. and Wei, W. “City air quality forecasting and impact factors analysis based on grey model,” Procedia Engineering, 12, 74-79 (2011).

[17] Pepe, N., Pirovano, G., Lonati, G., Balzarini, A., Toppetti, A. and Riva, G. M., “Development and application of a high resolution hybrid modelling system for the evaluation of urban air quality,” Atmospheric Environment, 141, 297-311 (2016).