# 1．Restatement of the Problem

Air pollution control is of great significance to the ecological environment and the health and safety of the people. In actual production and life, many air pollutants are produced. By establishing air pollution prediction models, it is possible to predict air pollution conditions in advance, allowing relevant departments to take action early to reduce the harm caused by air pollution [1]. However, the current WRF-CMAQ forecast model is limited by actual meteorological conditions, types of pollutant emissions, and unclear mechanisms of secondary pollutant generation, leading to inaccurate forecast results. Therefore, it is necessary to combine actual data from air quality monitoring stations for secondary modeling to optimize the CMAQ model's forecast results. [2].

The requirements are to establish a model based on the given monitoring station data, relevant technical standards, and actual conditions to study the following issues, while also paying attention to handling cases where data is missing or abnormal.

Problem 1**.** Based on the relevant calculation formulas and rules, combined with the actual measured data, calculate the Air Quality Index (AQI) and primary pollutant information for monitoring point A from August 25 to August 28, 2020.

Problem 2**.** Assume that the pollutant emission status in a certain area has not changed. The AQI value in the area will vary due to different meteorological conditions. For example, when meteorological conditions are favorable for pollutant dispersion or deposition, the local AQI will decrease. Use relevant data to classify the meteorological conditions based on the variation patterns of pollutant concentrations and describe their characteristics.

Problem 3**.** Establish a secondary forecast mathematical model that is applicable to three monitoring points to predict the pollutant values for each day over the next three days. The model should aim to minimize the maximum error of AQI forecast values and achieve the highest possible accuracy for predicting the primary pollutant.

Problem 4**.** In practice, the pollutant concentration values in neighboring areas often have a certain correlation. Collaborative forecasting across multiple regions can improve the accuracy of predictions. The requirement is to establish a collaborative forecast model that includes four monitoring points. The model should aim to minimize the maximum relative error of AQI forecast values and achieve the highest possible accuracy for predicting the primary pollutant.

# 2．Problem Analysis

Problem 1: Based on the calculation formulas and relevant rules provided in the appendix, write a MATLAB program. Then, substitute the corresponding data into the calculations to obtain the results.

Problem 2: Due to the presence of missing values (NA) and abnormal data in the actual measured data from the monitoring stations, it is necessary to preprocess the data matrix for smoothing and filling. For missing values (NA), use the lookup replacement method, replacing the missing values with adjacent data. The correlation between data from two consecutive days is relatively high, so the impact on the model prediction results is minimal after replacement. For abnormal data, use the box plot method for interpolation. After preprocessing the data, calculate the corresponding AQI data, and import it into the SOM neural network model for clustering, assuming four categories. Then, classify the meteorological data from the initial forecast. Combining both classifications will yield the classification of meteorological conditions.

Problem 3: The task of the secondary forecast correction model is to correct the error between the initial forecast results and the actual values. Based on the error of the initial forecast, a neural network model is used to obtain the error variation pattern, which is then applied to the initial forecast to reduce the error of the initial forecast.

Problem 4: To establish a regional collaborative model for the four monitoring stations, the data from the four stations can be fitted. The fitted data will replace the forecast data and be input into the secondary forecast correction model from Problem 3 to test whether the fitted model can reduce errors.

# 3．Model Assumptions

1. Assume that the pollutant emission status remains unchanged, meaning only weather factors affect the AQI.
2. When using MATLAB for calculations, ignore errors in significant figures..

# 4．Explanation of Symbols

Symbols

Explanation

IAQIP  Air Quality Sub-Index for pollutant P.

CP Concentration of pollutant P.

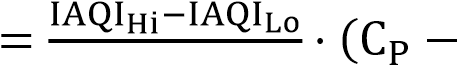
BPHi, BPLoBreakpoint concentrations that are closest to CP.

IAQIHi, IAQILo Corresponding AQI values for BPHi and BPLo.

# 5．Model Establishment and Solution

**5.1** Solution for Problem 1

To obtain the Air Quality Index (AQI), we first need to calculate the Air Quality Sub-Index (IAQI) for each pollutant. The calculation formula is as follows:

IAQIP BPHi−BPLo BPLo) + IAQILo

By referring to the tables, we can find the concentration limits for each pollutant and their corresponding IAQI values. After calculating the IAQI values for each pollutant, we select the maximum value, which is the primary pollutant for the day. The calculation rules are as follows:

AQI = max{IAQISO2, IAQINO2, IAQIPM10, IAQIPM2.5, IAQIO3, IAQICO}

The final calculation results are presented in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Monitoring date. | Location |  | AQI calculation | |
| AQI |  | Primary pollutant |
| 2020/8/25 | Point A | 65 |  | O3 |
| 2020/8/26 | Point A | 46 |  | O3 |
| 2020/8/27 | Point A | 108 |  | O3 |
| 2020/8/28 | Point A | 137 |  | O3 |

**5.2** Solution for Problem 2

5.2.1 Data preprocessing

Based on the data in Attachment 1, first handle the missing values. For missing values, use the method of value lookup replacement. Specifically, find the NA values and replace them with the preceding data in the same column.

Next, handle outliers. For outliers, use the box plot method for smoothing. A simple box plot consists of five parts: the minimum, median, maximum, and two quartiles. The first quartile, Q1, also known as the lower quartile, is the 25th percentile of the sorted data. The median, F, also known as the second quartile (Q2), is the 50th percentile. The third quartile, Q3, also known as the upper quartile, is the 75th percentile. The interquartile range (IQR) is calculated as Q3-Q1.

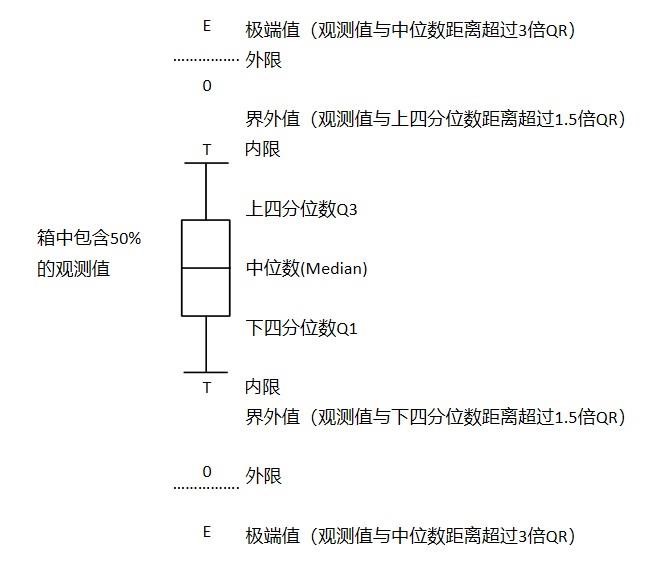


Figure 1: Simple Box Plot

5.2.2 SOM Self-Organizing Neural Network Clustering

Self-Organizing Map (SOM) neural networks can perform unsupervised clustering of data. Essentially, a SOM is a type of neural network with an input layer and a hidden layer, where each node in the hidden layer represents a class. During the training of the neural network, competitive learning is employed. Each input sample finds the most matching node in the hidden layer, referred to as its activation node. Subsequently, the parameters of the activation node are updated using the stochastic gradient descent method. Additionally, nodes close to the activation node are also updated appropriately based on their distance from the activation node.

A notable feature of the SOM model is that the nodes in the hidden layer have a topological relationship. The two-dimensional topology forms a plane.

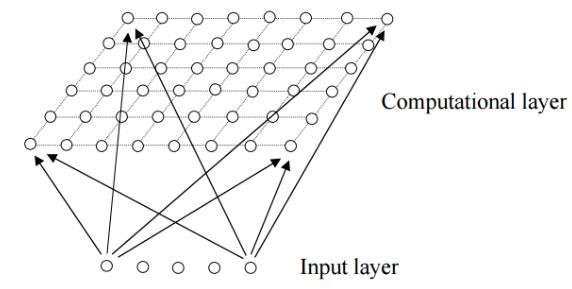


Figure 2: Two-Dimensional Topological Relationship.

SOM can discretize inputs of any dimension into a two-dimensional discrete space. The nodes in the computation layer are fully connected to the nodes in the input layer [4].

The following are the calculation steps:

1. Initialization: Randomly initialize the parameters of each node, with the number of parameters for each node matching the input dimension.
2. Finding Matching Nodes: For each input data, find the most matching node. Assuming the input is D-dimensional data, i.e.,

𝑋 = {xi, i = 1, . . . , D}, the discrimination function can be the Euclidean distance:

𝑑j(x) = ∑(xi-wji)² i=1

1. Updating Activation Nodes: After finding the activation node I(x), also update the nodes adjacent to it. Let S𝑖𝑗 represent the distance between nodes i and j. For nodes near I(x), assign an update weight:

Tj,I(x) = exp(-Sj2,I(x)/2𝜎2)

In simple terms, the closer the nodes are to each other, the smaller the update degree.

1. Updating Node Parameters: Update according to the gradient descent method:

∆wji = η(t) · Tj,I(x)(t) · (xi-wji)

Iterate until convergence.

5.2.3 Calculation Results

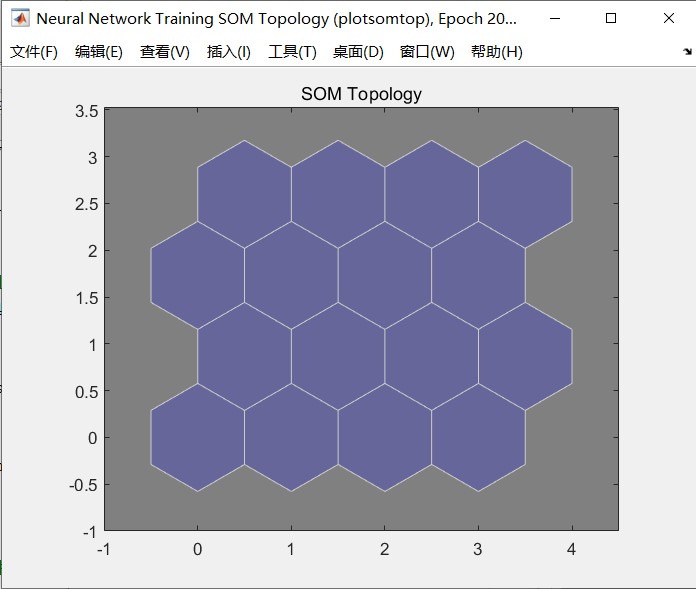


Figure 3: 16 computation layer nodes

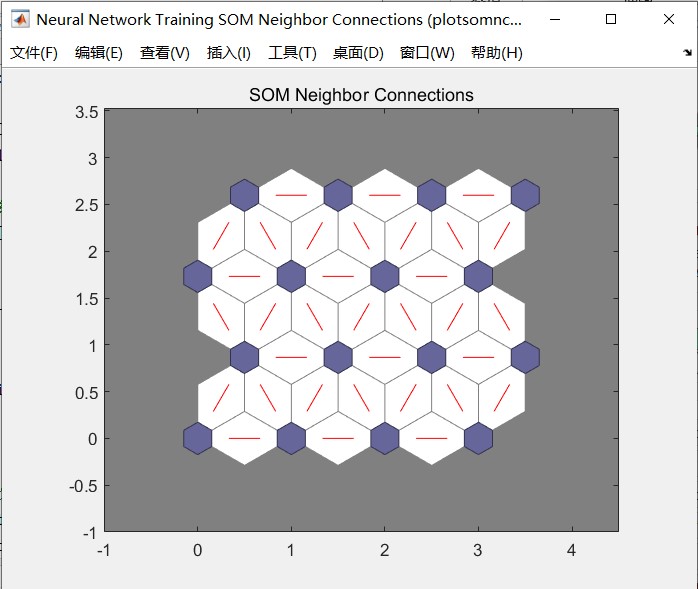


Figure 4: Relationships between nodes.

The correlation between nodes is divided by color; the lighter the color, the higher the correlation; the darker the color, the lower the correlation. As shown in the figure, the lower left is node 1, and the upper right is node 16. For example, the light color between node 1 and node 2 indicates that these two nodes are more similar and likely belong to the same category. Checking the computed classification table, node 1 is in category 2, and node 2 is also in category 2. Similarly, the dark color between node 6 and node 11 indicates they are in different categories, with node 6 in category 1 and node 11 in category 2.

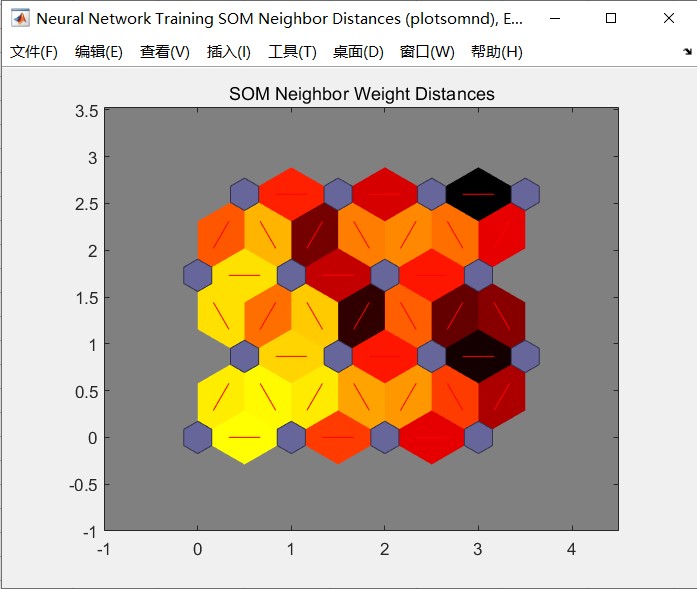


Figure 5: Correlation between nodes

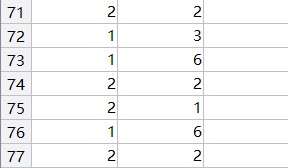


Figure 6: Partial classification result table

Comparing the classification results with the actual measured data at the monitoring points, it can be seen that higher wind speeds and humidity facilitate pollutant dispersion, resulting in lower AQI values.

**5.3** Problem Three's Solution

5.3.1 BP Neural Network Structure Optimized by Genetic Algorithm

The network structure is as follows:

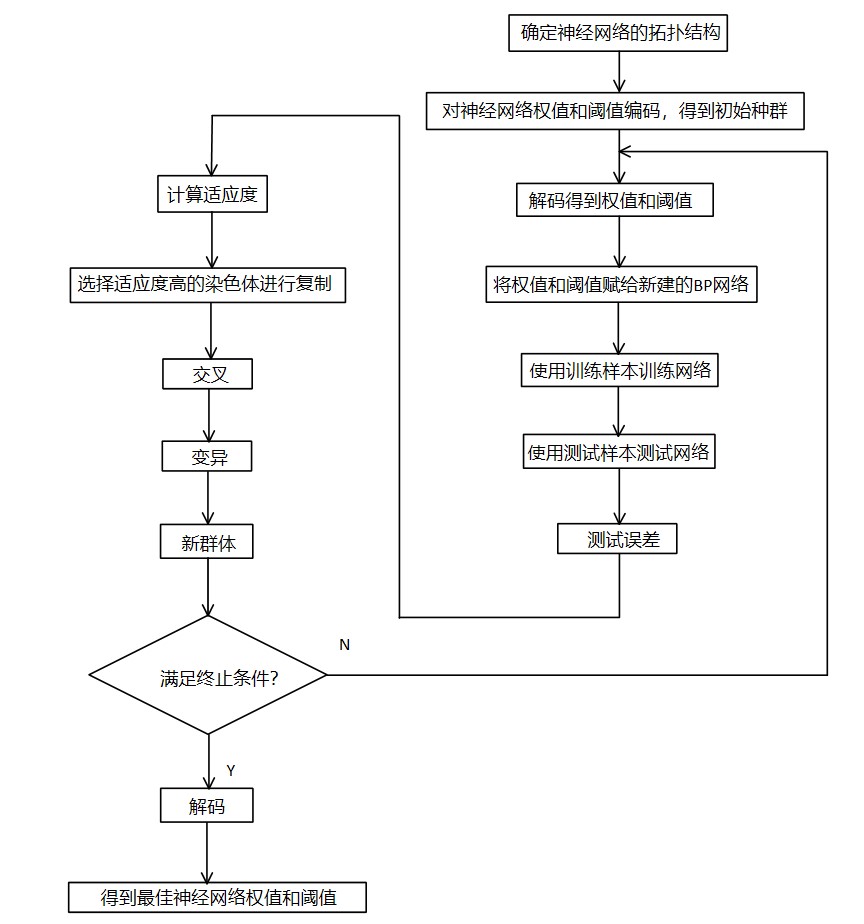


Figure 7: BP neural network structure diagram optimized by genetic algorithm

5.3.2 Implementation of the Neural Network Algorithm

The implementation of the BP neural network algorithm mainly includes the following steps [5]:

1. Network Creation

There are two important guidelines for determining the structure of a BP network:

* 1. For general pattern recognition problems, a three-layer network can solve the problem effectively.
  2. In a three-layer network, there is an approximate relationship between the number of neurons in the hidden layer (n2) and the number of neurons in the input layer (n1):

𝑛2 = 2 × 𝑛1 + 1

The transfer function for the neurons in the hidden layer uses the S-type tangent function tansig(), and the transfer function for the output layer neurons uses the S-type logarithmic function logsig(). This is because the output mode is 0-1, which perfectly meets the network's output requirements.

1. Network Training and Testing

Network training is a process of continuously adjusting weights and thresholds to minimize the output error. The training function trainlm() can use the Levenberg-Marquardt algorithm for network training. After training, the network also needs to be tested.

5.3.3 Implementation of Genetic Algorithm

Genetic algorithm optimization for the BP neural network involves using genetic algorithms to optimize the initial weights and thresholds of the BP neural network, enabling better sample prediction. The main steps include population initialization, fitness function, selection operator, crossover operator, and mutation operator.

1. Population Initialization

Individuals are encoded using binary coding. Each individual is a binary string, consisting of weights connecting the input layer and hidden layer, hidden layer thresholds, weights connecting the hidden layer and output layer, and output layer thresholds. Each weight and threshold is encoded using M bits of binary code. The encoding of all weights and thresholds together constitutes the encoding of an individual.

1. Fitness Function

The goal is to minimize the residuals between the predicted and expected values. The norm of the error matrix between the predicted and expected sample values is used as the output of the objective function.

1. Selection Operator

The selection operator uses random traversal sampling.

1. Crossover Operator

The crossover operator uses single-point crossover.

1. Mutation Operator

The mutation operator randomly selects mutant genes with a certain probability.

# 6. Model Evaluation and Application

Advantages:

1. The model clearly reflects the variations in pollutant concentrations under different meteorological conditions, and the data classification is intuitive.
2. The algorithm has significant potential for broader application.

Disadvantages:

1. The accuracy and predictive stability of the neural network model need improvement.
2. The model does not make full use of meteorological data.

# 7. References

[1]姚文强. 基于数值预报的空气质量预测模型的研究[D].浙江理工大学,2017.

[2]李昊. 基于机器学习的兰州市空气质量预报方法研究[D].兰州大学,2017.

[3]景华,陈世平.融合 SOM和多目标寻优的 Web服务组合优化方法[J].小型微型计算机系统,2021,42(09):1810-1817.

[4]Cai Ying,Wang Xu,Xiong LiRan,Ahmed Syed Hassan. Difference Analysis of Regional Economic Development Based on the SOM Neural Network with the Hybrid Genetic Algorithm[J]. Computational Intelligence and Neuroscience,2021,2021:

[5]武晨晨,苗霁,祝佳楠,张文惠.遗传算法优化 BP 神经网络的电能质量预测预警研究[J].电工电气,2021(09):18-22.

# 8. Appendix

