

# 1. Restatement of the Problem

## 1.1 Background of the Problem

In recent years, with the accelerating process of industrialization and urbanization, the problem of air pollution has become increasingly serious. Air pollution can have a negative impact on the daily life of the public and even cause a series of health problems [1-3]. Carrying out environmental air quality forecasting work is an important technical means to ensure timely and proper response to heavy pollution weather, and it also has guiding significance for the joint emission reduction of regional air pollution.

Existing air quality forecasting methods mainly include numerical analysis methods and statistical analysis methods. However, numerical forecasting methods usually require accurate input data and expensive computing resources for air quality forecasting; while statistical forecasting methods have low accuracy in forecasting pollutant concentrations with nonlinear changes [4]. At present, the realization of environmental air quality forecasting through methods such as artificial intelligence and machine learning has become a research hotspot and development trend in the field of environmental protection in various countries [5-6].

## 1.2 Restatement of the Problem

Question 1: Using the basic air quality forecast data of monitoring point A, calculate the daily measured AQI and primary pollutants of monitoring point A from August 25 to August 28, 2020 in accordance with the method in the appendix.

Question 2: Under the condition that pollutant emissions remain unchanged, if the meteorological conditions in a certain area are conducive to the diffusion or settlement of pollutants, the AQI of that area will decrease; otherwise, it will increase. Classify meteorological conditions reasonably according to their degree of impact on pollutant concentrations, and expound on the characteristics of each type of meteorological conditions.

Question 3: Using the basic air quality forecast data of monitoring point A and the basic air quality forecast data of monitoring points B and C, establish a quadratic forecasting mathematical model applicable to all three monitoring points A, B, and C (the straight-line distance between each pair of monitoring points is more than 100km, and mutual influence can be ignored) to predict the daily concentration values of 6 conventional pollutants in the next three days. It is required that the maximum relative error of the AQI forecast value in the prediction results of the quadratic forecasting model should be as small as possible, and the prediction accuracy of the primary pollutant should be as high as possible. Then, use this model to predict the daily concentration values of 6 conventional pollutants at monitoring points A, B, and C from July 13 to July 15, 2021, and calculate the corresponding AQI and primary pollutants.

Question 4: The pollutant concentrations in adjacent areas are often correlated to a certain extent, and regional collaborative forecasting may improve the accuracy of air quality forecasting. There are monitoring points A1, A2, and A3 in the adjacent area of monitoring point A. Using the data in Annexes 1 and 3, establish a collaborative forecasting model including the four monitoring points A, A1, A2, and A3. It is required that the maximum relative error of the AQI forecast value in the quadratic model prediction results should be as

small as possible, and the primary pollutant forecast...

Try to achieve high accuracy. Use this model to predict the daily concentration values of 6 conventional pollutants at monitoring points A, A1, A2, and A3 from July 13 to July 15, 2021, calculate the corresponding AQI and primary pollutants. And discuss: compared with the model in question 3, can the collaborative forecasting model improve the accuracy of pollutant concentration forecasting for monitoring point A?

## 2. Problem Analysis

Question 1: Calculate by substituting data in accordance with the calculation rules in the appendix. Only select some representative data to be placed in the main text, and the rest have been included in the appendix.

Question 2: Perform unsupervised clustering based on the AQI data and related pollutant data calculated in Question 1. There are many unsupervised clustering models, such as hierarchical clustering, Gaussian mixture clustering, etc. Using the SOM self-organizing neural network clustering algorithm, input the original data into the network, and then it can automatically generate different results at different steps according to the characteristics of various types of data.

Question 3: First, analyze the known data in the question, which includes hourly pollutant concentrations at each monitoring point, primary meteorological forecast data, as well as measured pollutant concentrations and meteorological data, etc. Use the measured data to correct the errors of the forecast data, and find the relevant error correction rules through the difference between the early predicted data and the measured data. Adopt a neural network model and set up a three-layer network structure. The input layer data is the meteorological conditions of the primary forecast, and the standard output data is the difference between the actual pollutant concentration and the predicted pollutant concentration, thus establishing the relationship between the predicted meteorological conditions and the error of the actual pollutant concentration. Using a neural network model optimized based on genetic algorithm, its accuracy will be higher compared with the traditional BP neural network. After obtaining the above network relationship, if a new set of primary forecast meteorological data is obtained, secondary correction can be carried out in combination with relevant error variables.

Question 4: Regional collaborative forecasting is mainly to prevent errors in the prediction data of a certain point. In such cases, corrections can be made using the prediction data of other points (the premise for correction is that when there is a large error between the relevant data obtained from other points and the real data, correction will be carried out). Therefore, a fitting model related to location and pollutant concentration data among the four stations A, A1, A2, and A3 can first be established based on real data. Then, the relationship between these fitting models is incorporated into the primary forecast data. If there is a large gap between the primary forecast data and the fitting results, the primary forecast data will be replaced (a certain error limit can be set); if they are consistent, the primary forecast data will be retained. At this point, the model constructed in Question 3 is used for calculation to check whether the error with the real value is reduced. If it is reduced,

it indicates that regional collaborative forecasting can improve accuracy; if it is not reduced, it indicates that the effect is not good.

### 3. References

- [1] CHEN F L, CHEN Z F. Cost of economic growth: air pollution and health expenditure[J]. Science of the Total Environment, 2021, 755:142543.
- [2] LIU W L, XU Z P, YANG T. Health effects of air pollution in China[J]. International Journal of Environmental Research and Public Health, 2018, 15(7):1471.
- [3] 张雨梦, 钱鹏, 查书平. 南京市一次大气污染事件时空演化特征及影响因素[J]. 南通大学学报(自然科学版), 2018, 17(4):48-55.
- [5] 许治国. 利用 Keras 构建神经网络在空气质量预测中的应用[J]. 环境监控与预警, 2018, 10(5):18-21.
- [6] RYBACZYK Y, ZALAKVICIUTE R. Machine learning approaches for outdoor air quality modelling:a systematic review[J]. Applied Sciences, 2018, 8(1 2):2570.

### 4. Appendix

```
%二维自组织特征映射网络设计
%输入数据为各类实测污染物数据
clc
clear
close all %-----
%随机生成 100 个二维向量, 作为样本, 并绘制出其分布
P=[此处填写污染物数据]
%
%建立网络, 得到初始权值
net=newsom([0 1;0 1],[5 6]);
w1_init=net.iw{1,1};
%-----
%绘制出初始权值分布图
figure(2);
plotsom(w1_init,net.layers{1}.distances)
%-----
%分别对不同的步长, 训练网络, 绘制出相应的权值分布图
for i=10:30:100
net.trainParam.epochs=i;
net=train(net,P);
figure(3);
plotsom(net.iw{1,1},net.layers{1}.distances)
end
%程序一: GA 训练 BP 权值的主函数
```

```

function net=GABPNET(XX,YY)
% GABPNET.m
% 使用遗传算法对 BP 网络权值阈值进行优化，再用 BP 算法训练网络
%数据归一化预处理
nntwarn off
XX=[1:19;2:20;3:21;4:22];
YY=[1:4];
XX=premnmx(XX);
YY=premnmx(YY);
YY
%创建网络
net=newff(minmax(XX),[19,25,1],{'tansig','tansig','purelin'},'tra
inlm');
%下面使用遗传算法对网络进行优化
P=XX;
T=YY;
R=size(P,1);
S2=size(T,1);
S1=25;%隐含层节点数
S=R*S1+S1*S2+S1+S2;%遗传算法编码长度
aa=ones(S,1)*[-1,1];
popu=50;%种群规模
save data2 XX YY % 是 将 xx,yy 二个变数的数值存入 data2 这个
MAT-file,
initPpp=initializega(popu,aa,'gabpEval');%初始化种群
gen=100;%遗传代数
%下面调用 gaot 工具箱，其中目标函数定义为 gabpEval
[x,endPop,bPop,trace]=ga(aa,'gabpEval',[],initPpp,[1e-6
1
1],'maxGenTerm',gen, ...
'normGeomSelect',[0.09],[ 'arithXover'],[2],'nonUnifMutation',[2
gen 3]);
%绘收敛曲线图
figure(1)
plot(trace(:,1),1./trace(:,3),'r-');
hold on
plot(trace(:,1),1./trace(:,2),'b-');
xlabel('Generation');
ylabel('Sum-Squared Error');
figure(2)
plot(trace(:,1),trace(:,3),'r-');
hold on
plot(trace(:,1),trace(:,2),'b-');
xlabel('Generation');

```

```

ylabel('Fitness');

%下面将初步得到的权值矩阵赋给尚未开始训练的 BP 网络
[W1,B1,W2,B2,P,T,A1,A2,SE,val]=gadecod(x);
net.LW{2,1}=W1;
net.LW{3,2}=W2;
net.b{2,1}=B1;
net.b{3,1}=B2;
XX=P;
YY=T;
%设置训练参数
net.trainParam.show=1;
net.trainParam.lr=1;
net.trainParam.epochs=50;
net.trainParam.goal=0.001;
%训练网络
net=train(net,XX,YY);

%程序二：适应值函数
function [sol, val] = gabpEval(sol,options)
% val - the fitness of this individual
% sol - the individual, returned to allow for Lamarckian evolution
% options - [current_generation]
load data2
nntwarn off
XX=premnmx(XX);
YY=premnmx(YY);
P=XX;
T=YY;
R=size(P,1);
S2=size(T,1);
S1=25;%隐含层节点数
S=R*S1+S1*S2+S1+S2;%遗传算法编码长度
for i=1:S, x(i)=sol(i);
end;
[W1, B1, W2, B2, P, T, A1, A2, SE, val]=gadecod(x);

%程序三：编解码函数
function [W1, B1, W2, B2, P, T, A1, A2, SE, val]=gadecod(x)
load data2
nntwarn off
XX=premnmx(XX);
YY=premnmx(YY);
P=XX;
T=YY;
R=size(P,1);
S2=size(T,1);

```

```

S1=25;%隐含层节点数
S=R*S1+S1*S2+S1+S2;%遗传算法编码长度
% 前 R*S1 个编码为 W1
for i=1:S1, for k=1:R, W1(i,k)=x(R*(i-1)+k);
end
end
% 接着的 S1*S2 个编码 (即第 R*S1 个后的编码) 为 W
for i=1:S2, for k=1:S1, W2(i,k)=x(S1*(i-1)+k+R*S1);
end
end
% 接着的 S1 个编码 (即第 R*S1+S1*S2 个后的编码) 为 B1
for i=1:S1, B1(i,1)=x((R*S1+S1*S2)+i);
end
% 接着的 S2 个编码 (即第 R*S1+S1*S2+S1 个后的编码) 为 B2
for i=1:S2, B2(i,1)=x((R*S1+S1*S2+S1)+i);
end
% 计算 S1 与 S2 层的输出
A1=tansig(W1*P,B1);
A2=purelin(W2*A1,B2);
% 计算误差平方和
SE=sumsqr(T-A2);
val=1/SE; % 遗传算法的适应

```