基于机器学习的空气污染物与气象要素关系及 AQI 预测研究

黄妍玲 严皓文 张浩淼[通讯作者]

重庆电子科技职业大学，重庆市，401331；

摘要：结合污染物和气象因子，首先通过灰色关联度分析确定与空气质量指数关联度较高的污染物。然后，基于选出的污染物，利用 Spearman 等级相关系数筛选出相关气象因子作为机器学习的输入变量。通过典型相关性分析，表明在进行污染物预测时，气象变量应当被赋予更高的权重，而不仅依赖污染物的历史数据。最后，采用四种机器学习模型对 AQI 进行了预测，结果表明，四种模型的 R2 均超过 0.95，表现出较好的预测性能。但是，在梯度提升树和决策树模型上存在过拟合的现象。因此，BP 神经网络模型在泛化能力上表现最佳，其次为随机森林模型。

关键词：空气质量指数；空气污染物；气象因子；预测模型；机器学习

DOI：10.69979/3041-0673.25.09.072

1. 资料与方法

1.1 资料说明

本文的污染物资料来源于我国环境监测总站（https://www.cnemc.cn/），包括六种主要空气污染物：PM2.5、PM10、NO2、SO2、CO 和 O3。气象资料来源于美国国家海洋和大气管理局国家环境信息中心（NCEI），隶属于美国国家海洋大气管理局，所选气象元素包括：平均气温，平均露点温度，海平面平均压力，能见度平均，风速平均，降水量平均。观测站于 2024 年的重庆市沙坪坝区。

1.2 研究方法

1.2.1 筛选出空气污染物与气象因素影响大的因子

首先，对采集到的数据进行预处理，包括异常值检测、遗漏值处理、数据质量控制和数据归一化等。然后，利用特征筛选等算法实现对六种污染物与空气质量指数之间的相关程度，从中筛选出与 AQI 相关性最高的污染物，在将选出指标与气象因素进行选择。然后对所选取的污染物与气象因子的相关性进行分析。

根据相关性分析结果，选择出与空气质量变化最显著的污染物和气象因子，作为预测的因变量。最后，结合这些因变量，采用四种机器学习方法对数据进行建模训练，并对各类模型的性能进行评估与比较，以确定最佳的预测模型。具体流程如图 1 所示。

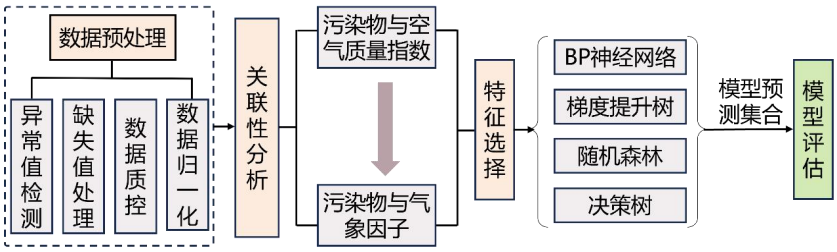

图 1 实验流程

1.2.2 基于机器学习的 AQI 预测研究方法

本文选取四种机器学习进行预测，其模型的特点和参数的选择如下表 1 所示：

表 1 模型特点和参数

模型	模型特点	参数
BP 神经网络	包含输入层、隐藏层和输出层、正向传播计算预测结果、反向传播调整权重和偏置、将误差平方和最小化等内容的误差反向传播算法[4]。适用于容易产生过拟合的非线性关系的解决。	激活函数:identity;学习率:0.1;L2 正则项:1;隐藏第 1 层神经元数量:100;迭代次数:1000
梯度提升树	通过多个弱分类器（通常是决策树）逐步优化预测结果，加法模型，减少误差。通过限制树的深度、正则化和早停技术进行优化[5]。	损失函数:friedman_mse;节点分裂评价准则:friedman_mse;基学习器数量:100;学习率:0.1;无放回采样比例:1;内部节点分裂的最小样本数:2;叶子节点的最小样本数:1;树的最大深度:10;叶子节点的最大数量:50

随机森林	通过对多棵决策树的投票机制进行预测，其表现与决策树之间存在着反比关系。袋外数据法是用来分析特征重要性的数据法，越大的误差越能说明特征的重要性[6]。	节点分裂评价准则:mse;内部节点分裂的最小样本数:2;叶子节点的最小样本数:1;树的最大深度:10;叶子节点的最大数量:50;决策树数量:100
决策树	归纳学习算法，通过递归划分数据集，直到数据集无法再划分。根据特征值分支，各节点代表 1 个特征[7]。	节点分裂评价准则:friedman_mse 特征划分点选择标准:best;划分时考虑的最大特征比例:None;内部节点分裂的最小样本数:2;叶子节点的最小样本数:1;树的最大深度:10;叶子节点的最大数量:50

1.2.3 模型评价指标

本文采用平均绝对误差 MAE、平均方根误差 RMSE、决定系数 R2 三种方法对模型进行性能评估。在 MAE 和 RMSE 数值越小的情况下，预测模型的准确率越高；R2 离 1 越近，模型的贴合度越好。

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|\hat{y}_i - y_i|$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}(\hat{y}_i - y_i)^2}{\sum_{i=1}(y_i - \bar{y})^2}$$

其中，$\hat{y}_i$为预测值；y_i为真实值；m为样本数。

2 结果与分析

2.1 污染物和气象要素筛选

为了提高预测精度，本文通过筛选空气污染物浓度和气象要素中的主要影响因素，将其作为机器学习模型的输入变量。

2.1.1 污染物筛选

将每一种污染物与 AQI 之间的关系用灰色关联度进行分析。结果显示，AQI 与 PM2.5、PM10、SO2、NO2、O3、CO 这 6 项污染物的关联度均大于 0.5，且平均关联度达到 0.84，表明 AQI 与这些污染物的关联性较强。具体关联度值如表 2 所示:

表 2 AQI 与 6 种污染物灰色关联度

污染物	PM2.5	PM10	SO2	NO2	O3	CO
AQI	0.899	0.947	0.829	0.847	0.686	0.849

灰色关联度的取值范围为[0,1]，越接近 1 说明关联度越强，反之关联度较弱[8]。根据表 2 的结果，得到污染物排序为 PM10>PM2.5>CO>NO2>SO2>O3。基于这个排序，选择机器学习模型的输入变量排在前 3 位的污染物(PM10、PM2.5、CO)。

2.1.2 气象因子的筛选

使用 Spearman 相关系数分析 PM10、PM2.5 和 CO 与六种气象因素之间的关系。最终选取与这些污染物具有较高相关性的 4 个气象要素以代替原有气象因子。相关系数热力图，如图 2 所示。

图 2 相关系数热力图

从图 2 中可知，PM2.5、PM10 和 CO 与能见度、海平面气压、露点温度和温度的相关系数最高。其中，与海平面气压呈正相关的 PM2.5、PM10 和 CO 表明，污染物在高气压环境下蓄积的可能性较大。与能见度、露点温度和温度之间呈反比例关系，意味着当污染物浓度增加时，光的传播能力、空气中水蒸气的凝结及大气的热力学特性都会随之变化，从而导致能见度降低和温度发生变化。

2.2 污染物与气象要素关系

根据筛选出的影响空气质量指数的主要污染物和气象因素作为输入变量，如表 3 所示。

表 3 AQI 预测模型输入变量

污染物	输入变量
污染物指标	PM2.5、PM10、CO
气象因子	能见度、海平面气压、露点温度和温度

将以上筛选出的输入变量气象因子 X 与污染物 Y 建立典型相关性分析(CCA)[9]，以探讨污染物与气象因素之间的关系，但需要测试两组变量之间是否有显著的相关性，若关联性不显著，则意义不大，即$Cov(X,Y)=0$；若显著则建立典型相关性分析。因此，设原假设H_0：两组变量之间无相关性；备择假设H_1：两组变数之间是有关联性。检验结果见表 4。

表 4 典型相关系数检验

	相关系数	特征值	威尔克统计	F	分子自由度	分母自由度	P 值
第 1 对	0.832	0.693	0.26	12	950.116	52.586	0.000***
第 2 对	0.391	0.153	0.846	6	720	10.454	0.000***
第 3 对	0.031	0.001	0.999	2	361	0.174	0.841

注：***、**、*分别代表 1%、5%、10%的显著性水平

从表中可以看出，前两对典型变量的 P 值都在 0.01 以下，说明“相关性不存在”这一原假设在 99%的置信水平下可以被拒绝。故前两对典型变量相关性显著。不过，仅有第 1 对的典型变量相关系数大于 0.8，说明相关性较强。基于此，本文选择第 1 对典型变量进行污染物和气象因子之间的冗余度分析，结果如表 5 所示。

表 5 典型变量的解释能力

变量关系	方差解释比例	变量关系	方差解释比例
污染物的方差比例由自身的典型变量所解释	34.018%	气象因子的方差比例由自身的典型变量所解释	52.161%
污染物的方差比例由气象因子的典型变量所解释	75.283%	气象因子的方差比例由污染物的典型变量所解释	23.57%

根据表 5 知，气象因子对污染物浓度变化的解释力较强，达到了 75.283%，表明污染物浓度的变化主要由气象因素主导。相对而言，污染物对气象因子的影响较为有限，仅为 23.57%。另外，污染物自身变化对其浓度的解释力较小，仅为 34.018%，而气象因子的自我解释力较强，为 52.161%，表明气象因子变化受多种因素共同作用。总体上，在解释污染物浓度变化时，气象因子占主导地位。因此，在进行污染物预测时，应更加重视气象变量，而非仅依赖污染物的历史数据。

利用这些标准化后的典型变量系数，建立详细的典型相关模型，如下所示。

$$Y_1=0.008\times PM10-0.053\times PM2.5+0.164\times CO$$

$$X_1=-0.187\times 温度+0.114\times 露点温度+0\times 海平面气压+0.523\times 能见度$$

通过对污染物和气象因子的典型变量系数与典型载荷因子的分析，发现气象因素(例如温度、露点温度和能见度等)与空气污染物浓度(如 PM10、PM2.5 和一氧化碳浓度)之间存在着复杂的相互关系，并且这种相互作用非常明显。例如，能见度与 PM2.5 浓度之间存在显著的负相关关系，即较高的能见度通常对应较低的 PM2.5 浓度。

3 结论与讨论

本文运用灰色关联度和 Spearman 等级相关系数，选出了 6 种污染物与气象因子，接着通过典型相关性分析研究了它们之间的关系。最终，将选出的污染物和气象因子作为输入变量，以建立重庆沙坪坝区的空气质量指数预测模型，主要结论如下：

（1）在 6 种污染物中,PM10、PM2.5 和 CO 对 AOI 的影响较为显著。另外，PM2.5、PM10 和 CO 在能见度、海平面气压、露点温度及温度方面具有较强的相关性。

（2）气象因子对污染物浓度变化具有较强的解释力，达到 75.28%，而污染物对气象因子的影响程度相对较弱，只有 23.57%。因此，在进行污染物预测时，气象因子应当被赋予更高的权重，而不仅仅依赖污染物的历史数据。

（4）四种机器学习模型对 AQI 的预测均表现出较好的性能，R^2 均超过 0.95。但是，在梯度提升树和决策树模型上存在着过度拟合的现象。综合考虑，在泛化能力方面，BP 神经网络表现最好，随机森林模型则紧随其后。

参考文献

[1]刘金培,罗瑞,陈华友,等.基于多尺度 3D-CNN-CBAM 的空气质量指数时空预测[J].控制与决策,2025,40(02):404-412.DOI:10.13195/j.kzyjc.2024.0105.

[2]马俊文,严京海,孙瑞雯,等.基于 LSTM-GCN 的 PM2.5 浓度预测模型[J].中国环境监测,2022,38 (5):153-160.

[3]张小曳,徐祥德,丁一汇,等.2013～2017 年气象条件变化对中国重点地区 PM2.5 质量浓度下降的影响[J].中国科学:地球科学,2020,50(04):483-500.

[4]赵艺伟.基于机器学习的空气质量预测模型研究[D].安徽理工大学,2024.DOI:10.26918/d.cnki.ghngc.2024.001097.

[5]陈建坤,牟凤云,张用川,等.基于多机器学习模型的逐小时 PM2.5 浓度预测对比[J].南京林业大学学报(自然科学版),2022,46(05):152-160.

课题：重庆电子科技职业大学校级课题(24XJXSCX01)：基于多源传感数据的空气环境监测系统的研究与应用