刘畅,庞菲菲,冯蕾,等. 基于集成学习模型的二氧化硫浓度预测研究[J]. 陕西气象,2026(2):78-86.
文章编号:1006-4354(2026)02-0078-09

基于集成学习模型的二氧化硫浓度预测研究

刘 畅[1,2],庞菲菲[3],冯 蕾[4],沈 萍[1],祁 菲[1],郭莲莲[1]
(1. 长安区气象局,陕西长安 710010;
2. 中国气象局秦岭和黄土高原生态环境气象重点开放实验室,西安 710016;
3. 西安市气象局,西安 710016;4. 陕西省气象台,西安 710014)

摘 要:为提升大气 SO_2 浓度预测精度,以陕西省西安市长安区 2014—2024 年的 SO_2 与气象观测数据为基础,经数据预处理和特征选择,提取了气压、温度、湿度、降水量等气象因素及其滞后特征。数据按照 8:2 的比例分为建模集与测试集,构建了 7 类单一模型并采用堆叠集成和加权集成方法进行优化。通过五折交叉验证评估模型,结果表明加权集成模型在验证集上决定系数为 0.875、均方根误差为 7.4 $\mu g/m^3$、平均绝对误差为 4.1 $\mu g/m^3$,明显优于单一模型。随机森林在训练集上表现最佳,而 K 近邻在验证集上的表现较差。综合评估结果显示,集成学习方法能够有效提升 SO_2 浓度的预测精度,为区域环境监测与空气质量管理提供可靠的科学依据,具有广泛的实际应用价值。

关键词:二氧化硫预测;集成学习;特征工程;大气污染监测;机器学习
中图分类号:X16;X511 **文献标识码**:A

二氧化硫(Sulfur Dioxide,SO_2)作为重要大气污染物,主要来源于工业排放、化石燃料燃烧及交通运输等多种途径[1]。此外,SO_2 对人类健康有明显的负面影响,长期暴露于高浓度的 SO_2 环境中可引发呼吸系统疾病、心血管疾病等严重健康问题。据世界卫生组织统计,空气中 SO_2 浓度升高与呼吸系统疾病发病率显著相关,对公共健康构成严峻威胁[2]。随着工业化和城市化进程的加快,全球范围内尤其是发展中国家 SO_2 排放量呈不同程度的增长趋势,高浓度的 SO_2 不仅导致酸雨,还会引起能见度降低、建筑物风化等环境问题,对生态系统和人类生活环境造成长期影响[3-7]。

准确预测大气中 SO_2 浓度对环境管理和公共健康保护具有重要意义[8]。传统的 SO_2 预测方法主要依赖于统计模型和物理模型,统计模型如时间序列分析和回归模型,基于历史数据进行趋势预测,但在处理复杂的非线性关系和多变的影响因素时表现有限[9-10]。物理模型则基于大气化学和传输过程,对污染物的扩散和转化进行模拟,虽然具备较高的理论基础,但通常计算复杂,需要大量输入数据,且难以精确反映实际观测数据中的随机波动[11]。近年来,随着机器学习技术的迅猛发展,机器学习技术在极端温度、降水等预测方面取得了一定的成果[12-16]。在环境科学领域,尤其是大气污染预测中的应用也取得了显著成果[17],支持向量机(support vector machine,SVR)、随机森林(random forest,RF)、梯度提升(gradient boosting,GBR)等机器学习算法因其强大的非线性建模能力和高维数据处理能力,被广泛应用于 SO_2 浓度的预测研究中。这些数据驱动方法能够有效捕捉复杂的气象与污染物浓度之

收稿日期:2025-10-28
作者简介:刘畅(1989—),男,汉族,陕西西安人,硕士,工程师,主要从事气象信息化与综合气象观测工作。
通信作者:庞菲菲(1988—),女,汉族,陕西西安人,学士,工程师,主要从事气象服务与应用气象研究工作。
基金项目:秦岭和黄土高原生态环境气象重点实验室开放基金课题(2020G-20;2022G-24);西安市气象局 2025—2026 年度自立科研项目(2025-10)

间的关系,提升预测的准确性和鲁棒性[18-20]。

尽管机器学习在大气污染预测中展现出巨大潜力,但仍面临诸多挑战[21-22]。环境数据通常具有高度的时间相关性和空间异质性,如何有效处理时间序列特征和空间依赖性是关键问题[23-24]。其次,特征选择在模型性能中起到至关重要的作用,如何从大量气象变量中筛选出对 SO_2 浓度影响显著的特征仍需深入研究。单一模型在不同数据集和环境条件下的泛化能力有限,而现有的集成方法虽然有所提升,但在优化组合策略和权重分配上仍存在改进空间。

针对上述挑战,本文首先介绍了所使用的数据集及其预处理方法,随后详细阐述了特征选择与工程、模型开发与超参数调优过程。通过系统应用和比较多种机器学习算法,并结合集成学习方法,提升 SO_2 浓度预测的准确性和稳定性。

1　数据与方法

1.1　数据简介

本研究使用的数据集来源于 2014 年 1 月 1 日—2024 年 12 月 31 日,环保部门对外公开的西安市长安区空气质量逐日数据,以及从气象部门获取的西安市长安区国家基本气象站的气象要素日监测数据。目标变量为每日的 SO_2 浓度,同时在气象要素数据上收集了多项可能影响 SO_2 浓度的气象要素,包括日平均气压、最高和最低气压、气温、相对湿度、降水量、地表温度、晴雨时数、风速等。对数据集中的缺失值进行替换处理(如“*”替换为 NaN),按照日期先后顺序对数据进行排序,确保时间序列的连续性。

1.2　预测模型

将 2014 年 1 月—2024 年 12 月的数据集按 8:2 的比例分为建模集和验证集,即 2014 年 1 月—2022 年 5 月的数据用于建模,2022 年 6 月—2024 年 12 月的数据用于验证,采用了多种机器学习回归模型作为基准模型。这些模型[25]包括 XGBoost 回归(XGBoost regression,XGBoost)、梯度提升回归(gradient boosting regressor,GBR)、随机森林回归(random forest regressor,RF)、Lasso 回归(lasso regression,Lasso)、支持向量回归(support vector regressor,SVR)、K 近邻回归(K-nearest neighbors regressor,KNN)以及极端随机树回归(extra trees regressor,ETR)。这些模型在处理高维数据和捕捉复杂非线性关系方面均有出色表现,尤其适用于复杂环境的数据预测任务。每种模型都具有独特的优势,例如,XGBoost 回归和梯度提升回归在捕捉复杂数据模式方面表现优异,而随机森林回归和极端随机树回归通过集成学习提升模型的泛化能力。Lasso回归通过 L1 正则化实现特征选择,增强模型的解释性;支持向量机通过核函数技巧有效建模高维空间中的非线性关系;K 近邻回归则作为一种简单而强大的基于实例的学习方法,为回归任务提供了一种非参数化解决方案。这些基准模型的多样性和各自优势为后续模型集成奠定了坚实的基础。

在基准模型的基础上,进一步构建了堆叠回归(stacking regressor,SR)模型和加权集成模型(weighted ensemble,WE),通过集成多种基准模型的预测结果来提升整体的预测性能。堆叠回归方法通过结合多种模型的优势,能够有效地缓解单一模型可能存在的过拟合问题,从而提升预测的准确性。加权集成模型通过根据各个模型的性能分配权重,确保表现优异且有效的模型在集成中占据更大的影响力,而表现较差的模型影响力相对减弱。本文选用上述多种性能优异的回归模型作为基学习器,并采用岭回归(ridge regression)作为元学习器。岭回归因其 L2 正则化特性,能够处理基学习器预测结果之间的多重共线性问题,确保集成模型的稳定性和鲁棒性。堆叠回归通过将基学习器的预测结果作为输入,利用岭回归对其进行线性组合或加权平均,实现对 SO_2 浓度的最终预测。这种层次化的集成方法不仅提升了模型的预测精度,还增强了模型在未见数据上的泛化能力,使得整体预测框架更加可靠和稳健。

为进一步优化各基准模型的性能,本研究采用了网格搜索(grid search)结合时间序列交叉验证(time series split)的方法,采用时间序列五折交叉验证(TIME-SERIES 5-FOLD CROSS-VALIDATION)评估模型稳定性:将建模集

(2014 年 1 月—2022 年 5 月)按时间顺序划分为 5 个连续子集,每次以 4 个子集为训练集、1 个后续子集为验证集,重复 5 次后计算平均性能,避免传统随机交叉验证导致的时间数据泄漏问题,对各模型的关键超参数进行细致的调优。不同于传统的 K 折交叉验证,时间序列交叉验证在数据划分时尊重时间顺序,确保训练集始终早于验证集,避免了数据泄漏问题,这对于时间依赖性的环境数据尤为重要,可为每个基准模型定义不同的超参数搜索空间。例如,XGBoost 回归的学习率、树的深度和估计器数量,支持向量回归的核函数类型及正则化参数,K 近邻回归(KNN regressor)的邻居数量和距离度量方式,Lasso 回归的正则化参数(alpha)等。通过网格搜索系统地遍历这些超参数组合,并结合多维时间序列交叉验证,选择使负均方误差(negative mean squared error,NMSE)最小的超参数配置。这一优化过程确保了每个基准模型在特定数据集上的最佳表现,从而为堆叠回归模型和加权集成模型提供高质量的机器学习,进一步提升整体预测性能。

模型性能的评估采用决定系数(R^2)、均方根误差(E_{RMS})和平均绝对误差(E_{MA})三个指标。R^2 衡量模型解释目标变量变异的能力,值越接近 1 表示模型拟合效果越好。均方根误差反映预测值与真实值之间差异的平方平均值,数值越小表示预测误差越小。平均绝对误差则是测量预测误差的绝对平均值,直观反映了模型的预测准确度[25]。相关公式如下:

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

$$E_{RMS} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}, \quad (2)$$

$$E_{MA} = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \hat{y}_i \right| 。 \quad (3)$$

式中:y_i 是第 i 个样品预测值,$\hat{y}_i$ 是第 i 个样品实测值,$\bar{y}_i$ 是验证集所有实测值的平均值,n 是验证集样品个数。

1.3 特征工程

为提升 SO_2 浓度预测模型对复杂环境数据的适配性与预测精度,围绕“候选特征构建-特征优化筛选”开展系统性特征工程,结合 SO_2 浓度的生成扩散机制、气象影响规律及时间序列属性,确保输入特征的全面性、有效性与精简性。首先构建多维度候选特征池:在气象特征维度,选取日平均气压,日最高、最低气压,日平均、最高、最低气温,气温日较差,日平均、最小相对湿度,日降水量,日平均、最高、最低地表温度,日日照时数,日平均、最大风速,日平均草面温度共 17 项指标,覆盖气压(影响大气层结稳定性)、温湿度(调控 SO_2 化学反应速率)、风速(决定水平扩散速率)、降水量(通过湿沉降降低浓度)等关键气象驱动因素;在时间特征维度,提取月份、季节(按“3-5 月春、6-8 月夏、9-11 月秋、12-次年 2 月冬”划分)、是否工作日(依据法定节假日及周末区分)三类特征,以捕捉气象条件与人类活动(如供暖、交通、工业生产)的周期性差异;在滞后特征维度,基于 SO_2 浓度自相关性分析结论(滞后 1-3 d 自相关系数分别为 0.7、0.5、0.3,滞后 1 d 热图相关系数达 0.92),提取 SO_2 浓度滞后 1 d、2 d、3 d 的历史数据,刻画污染物的时间依赖性与前期浓度累积影响。

为剔除冗余特征、增强模型泛化能力,进一步采用“多项式特征扩展+互信息回归筛选”的组合优化策略:针对日平均气温、日平均相对湿度等连续型气象特征,构建二次项(如气温、湿度)与交互项(如气温×湿度),通过引入非线性特征捕捉气象因素间的协同作用(如高温高湿条件下 SO_2 氧化速率的非线性变化),丰富特征表达维度;采用互信息回归(mutual information regression,MI)方法量化候选特征与 SO_2 浓度的关联性,该方法可有效捕捉线性与非线性关系,适配复杂环境数据特征评估需求,其核心公式为:

$$I(X_i;Y) = \sum_{x_i \in X_i} \sum_{y \in Y} p(x_i,y) \log\left(\frac{p(x_i,y)}{p(x_i)p(y)}\right)。 \quad (4)$$

其中,$p(x_i,y)$ 是特征 X_i 和 SO_2 浓度 Y 的联合概率分布,$p(x_i)$ 和 $p(y)$ 分别是 X_i 和 Y 的边缘概率分布。以“互信息值>0”为筛选阈值,最终保留日平均气压、日最高气压、日平均气温、日降

水量、日平均相对湿度(气象特征),月份、季节、是否工作日(时间特征),滞后 1～3 d SO_2 浓度(滞后特征)及气温×湿度交互项(多项式特征)共 12 项核心特征,在保障特征对 SO_2 浓度解释力的同时,避免冗余信息导致模型过拟合。

2　数据特征分析

2.1　时间演变

图 1 是 2014 年 1 月 1 日—2024 年 12 月 31 日西安市长安区 SO_2 月平均浓度的变化情况,从图中可以看出,SO_2 污染水平在过去十年中呈现下降趋势。尤其是 1 月和 2 月浓度分别从 2014 年的84.84 $\mu g/m^3$ 和 53.29 $\mu g/m^3$ 降至 2024 年的 6.77 $\mu g/m^3$ 和 26.41 $\mu g/m^3$。3 月—12 月 SO_2 平均浓度也呈现出明显下降趋势。尽管部分月份在某些年份有短暂回升,但整体污染水平仍持续下降。根据 SO_2 浓度的显著下降趋势,预测未来的 SO_2 浓度研究显得尤为重要。准确预测模型有助于提前识别潜在的污染高峰,优化污染控制策略,并支持公共健康预警系统的建立。此外,预测研究有助于实现可持续发展目标,通过合理规划能源结构和工业布局,促进低碳经济的发展。因此,深入研究 SO_2 浓度预测不仅巩固了现有污染控制成效,也为未来环境管理和政策制定提供了关键支持。

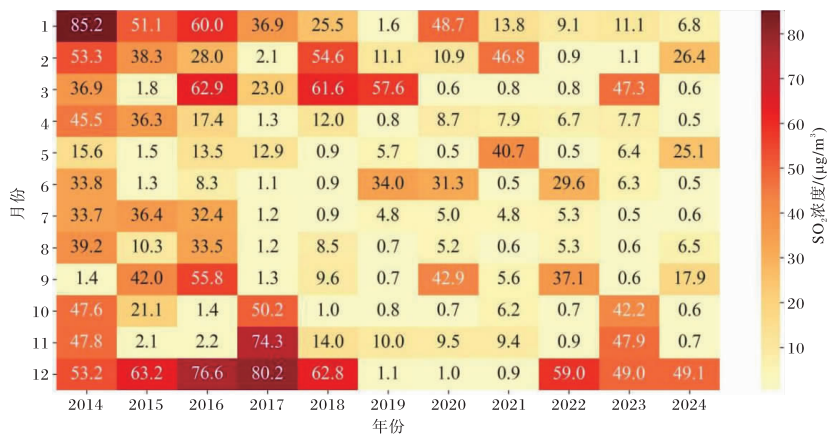

图 1　2014—2024 年 SO_2 月平均浓度

图 2 是 SO_2 日浓度的总体分布情况,分布图显示,SO_2 浓度呈右偏分布(右长尾),大部分观测值集中在较低浓度范围,而少数高浓度值拉长了右尾。这表明大部分时间内空气质量良好,但偶尔会出现高污染事件。此外,图 2 密度曲线(Kernel Density Estimation,KDE)表明数据在低浓度区域具有较高的峰度,但在高浓度区域则较为分散。这种分布特性提示,在后续预测模型构建过程中可能需要考虑数据的非正态性。

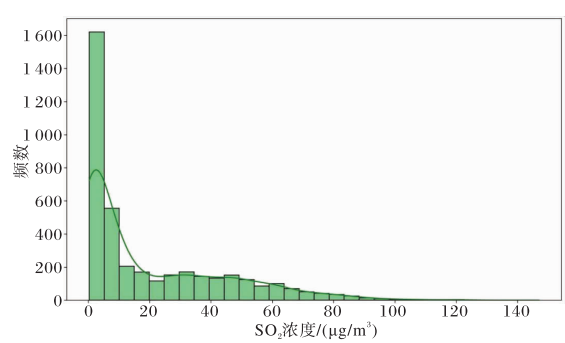

图 2　2014—2024 年不同 SO_2 浓度频率分布及密度曲线

图 3 是 2014 年 1 月—2024 年 12 月 SO_2 浓度整体变化趋势。从时间序列图中可以明显看出,SO_2 浓度在过去十年中呈逐年下降趋势。2014 年初 SO_2 浓度较高,随着时间的推移,尤其是在 2018 年以后,浓度逐渐减少。结合 SO_2 浓度月份分布箱线图(图 4)可以看出,1 月和 12 月的 SO_2 浓度中位数显著高于其他月份,反映出冬季供暖期间的高污染水平。特别是 1 月浓度范围较广,存在较多的极端高值,可能与特定年份的天气条件或工业排放异常有关。相比之下,5 月和 6 月 SO_2 浓度中位数相对较低,表明春末夏初空气质量较好。此外,箱线图中的低浓度时间段(如 7 月和 8 月)也显示出较低的中位数和较窄的分布范围,进一步验证了 SO_2 的季节分布差异。

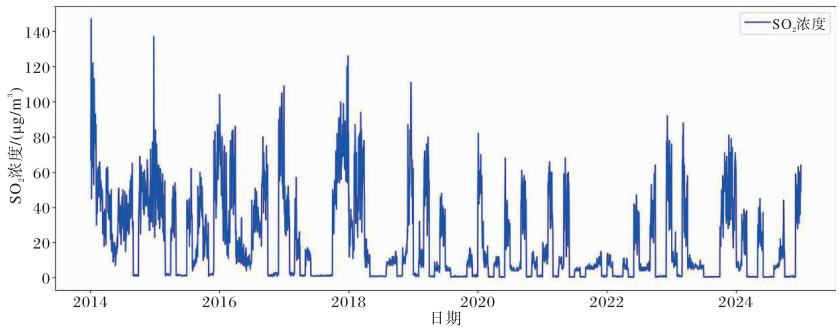

图 3 2014—2024 年逐月 SO_2 浓度变化趋势

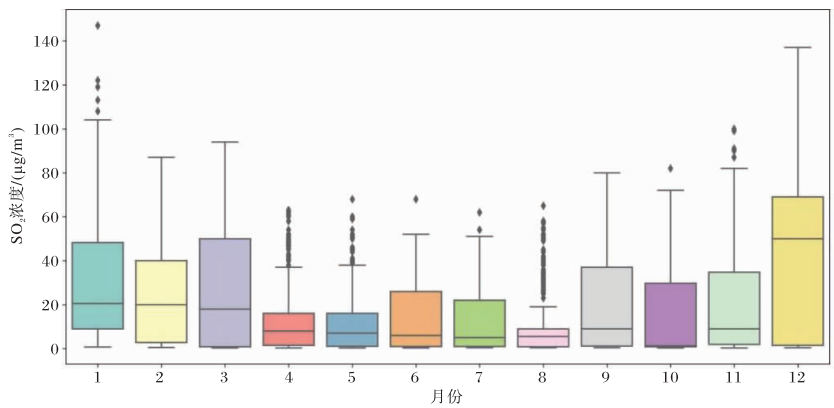

图 4 SO_2 浓度月份分布箱线图

2.2 季节特征

图 5 为 SO_2 浓度季节分布箱线图，冬季 SO_2 浓度显著高于春季、夏季和秋季。这一现象可能归因于冬季供暖系统的使用，特别是燃煤供暖，导致 SO_2 排放量上升。春季和秋季浓度相对稳定，反映出这些季节中供暖需求较低，且空气流通条件较好。夏季 SO_2 浓度最低，核心源于高温与强光照的双重作用机制：从扩散角度看，夏季高温增强热力湍流运动，推动 SO_2 向高层大气输送，且大气边界层高度(1 500～3 000 m)显著高于冬季(500～1 000 m)，叠加夏季风速(较冬季高 1.2～1.5 m/s)提升，大幅降低 SO_2 局部堆积；同时夏季对流性降水频繁(6—8 月降水量 280～320 mm，占全年 55%～60%)，SO_2 溶于水形成亚硫酸随降水沉降，单次强降水可使 SO_2 浓度 24 h 内下降 30%～50%。

从降解角度，夏季强太阳辐射(5.5～6.0 kWh/m^2·d)激活 O_2、O_3 等氧化剂，与 SO_2 发生光催化氧化反应(如 $SO_2+O_2+H_2O\rightarrow H_2SO_4+O_2$)，将其转化为硫酸或二次气溶胶；且强光照促进 VOCS 与 NO_x 生成 O_2(夏季 O_2 浓度 80.0～95.0 μg/m^3，较冬季高 40%～45%)，O_2 浓度超 80.0 μg/m^3 时可使 SO_2 氧化速率提升 2～3 倍[6]。

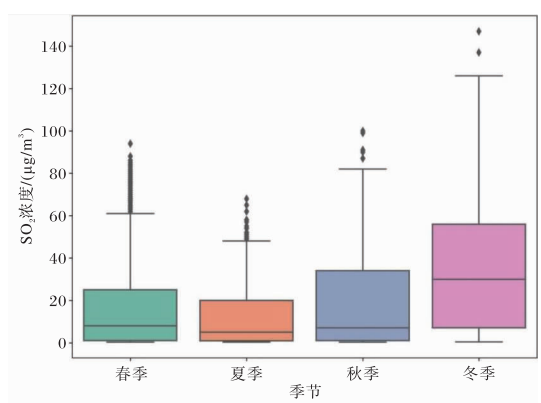

图 5 SO_2 浓度季节分布箱线图

2.3 自相关性

图 6 为 SO_2 浓度自相关性，从图中可以看出，在滞后 1～3 d 内，SO_2 浓度的自相关系数显著高于 0，表明存在较强的短期依赖性。此外，自相关系数在较高滞后期逐渐减弱，接近于随机噪

声(几乎为 0),表明 SO_2 浓度的时间依赖性主要集中在短期内。这一发现支持了后续预测模型中包含滞后特征的必要性。

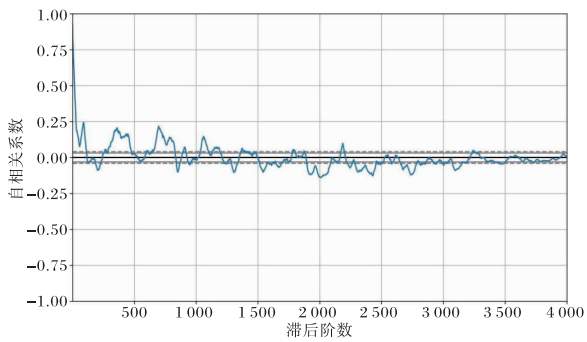

图 6　SO_2 浓度自相关性

2.4　工作日特征

图 7 为 SO_2 浓度工作日与非工作日分布箱线图,从图中可以看出,工作日 SO_2 浓度中位数略高于非工作日。非工作日则由于交通和工业活动的减少,SO_2 浓度相对较低。但需要注意的是,部分非工作日的 SO_2 浓度仍然较高,可能与特定的大型活动或天气条件有关。

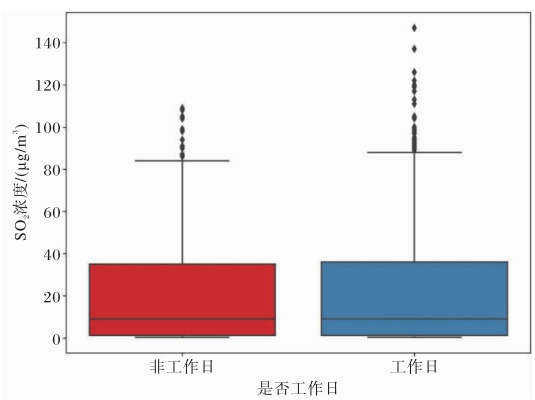

图 7　SO_2 浓度工作日与非工作日分布箱线图

2.5　滞后性

图 8 为 SO_2 浓度滞后特征关系热图,由图可知,当前 SO_2 浓度与前一天(滞后 1)浓度具有较高的正相关性,相关系数是 0.92,这表明 SO_2 浓度在时间上具有较强的自相关性。与前两天(滞后 2)和前三天(滞后 3)的相关性稍低,但依然显著(滞后 2 的相关系数为 0.86,滞后 3 为 0.82),这表明 SO_2 浓度不仅受前一天的影响,还受到更早期的数据影响。这种自相关性对于构建时间序列预测模型非常重要,因为未来的 SO_2 浓度可以由过去几天的浓度预测得出。

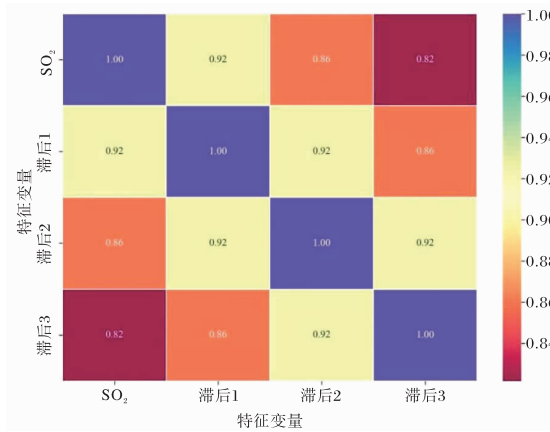

图 8　SO_2 浓度滞后特征关系热图

图 9 为 SO_2 浓度滞后特征与当前值关系,散点图显示,当前 SO_2 浓度与前一天(滞后 1)浓度之间存在明显的正相关关系,点云呈现出向右上方延伸的趋势,表明高滞后浓度通常对应高当前浓度。同样,前两天(滞后 2)和前三天(滞后 3)的浓度也与当前浓度呈现正相关,但斜率有所减缓。这种关系表明,过去几天 SO_2 浓度对当天的浓度有预测价值。随着滞后期的增加,点云的分布变得更加分散,表明相关性逐渐减弱。此外,部分散点分布偏离趋势线,可能受到异常值或外部因素(如突发工业排放)的影响,这些观察结果为构建更加复杂的预测模型提供了依据。

3　结果分析

3.1　特征选择结果

在特征选择过程中,运用互信息回归方法,深入评估了所有候选特征与 SO_2 浓度之间的关系。以互信息值作为筛选指标,最终筛选出互信息值大于 0 的特征。经筛选,被选中的特征涵盖多方面,其中气象类特征有日平均气压、日最高气压、日平均气温、日降水量、日平均相对湿度。这些气象特征在气象条件与污染物排放方面呈现出显著的相关性。例如,气压对污染物的扩散与沉降有重要影响,高气压环境通常有利于污染物的沉降,而低气压可能导致污染物积聚;风速同样影响污染物的扩散,风速越大,污染物越容易扩散。温度和湿度与化学反应速率相关,适宜的温度和湿度

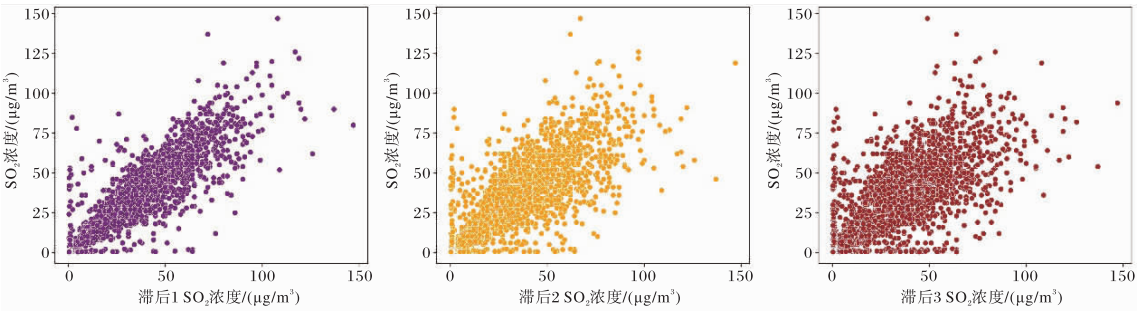

图 9 SO_2 浓度滞后特征与当前值关系

条件可能加速大气中与 SO_2 相关的化学反应。此外，还包括滞后 1 d、2 d、3 d 的 SO_2 浓度这些滞后特征。能够有效捕捉 SO_2 浓度的时间依赖性，即 SO_2 浓度在时间序列上的变化趋势并非孤立，前几日的浓度情况会对当前浓度产生影响，通过纳入这些滞后特征，有助于更全面地刻画 SO_2 浓度变化规律。

3.2 模型预测结果分析

回归模型在训练集和测试集上的性能表现如图 10、图 11 与表 1 所示。从训练集的 R^2 值和误差指标来看，随机森林和极端树在训练数据上表现较好，R^2 分别达到 0.975 和 0.983，显示出高度的拟合能力。然而，它也伴随着明显的过拟合现象，因为在测试集上的 R^2 均下降至 0.856。同时，KNN 模型在训练集上实现了完美拟合（$R^2=1.000$），但在测试集上的表现急剧下降至 0.771，出现过拟合问题，不适用于本研究任务。

相比之下，LASSO 回归和 SVR 在测试集上表现较为平衡，R^2 分别为 0.872 和 0.871，且误差指标（E_{RMS} 和 E_{MA}）较低，表明其在泛化能力和预测准确性方面具有优势。堆叠回归器作为集成模型，也展示了良好的测试集性能，R^2 为 0.871，E_{RMS} 为 7.54，显示出较好的泛化能力和较低的过拟合程度。在建模阶段，通过时间序列五折交叉验证发现，加权集成模型性能稳定性最优：交叉验证平均 $R^2=0.869\pm0.012$，$E_{RMS}=7.58\pm0.22$ $\mu g/m^3$，$E_{MA}=4.23\pm0.19$ $\mu g/m^3$。将其应用于测试集后，性能与交叉验证结果高度一致（R^2 提升 0.006，E_{RMS} 降低 0.17 $\mu g/m^3$，E_{MA} 降低 0.06 $\mu g/m^3$），验证了模型可靠性。然而，最具竞争力的仍然是加权集成模型，其在测试集上达到了最高的 R^2

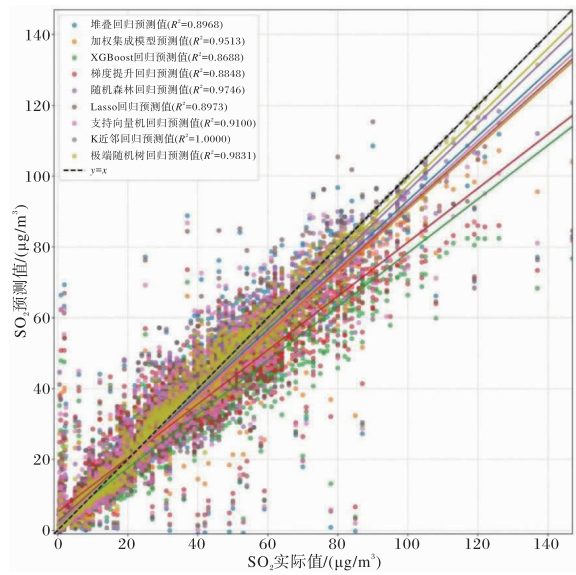

图 10 SO_2 训练集预测结果

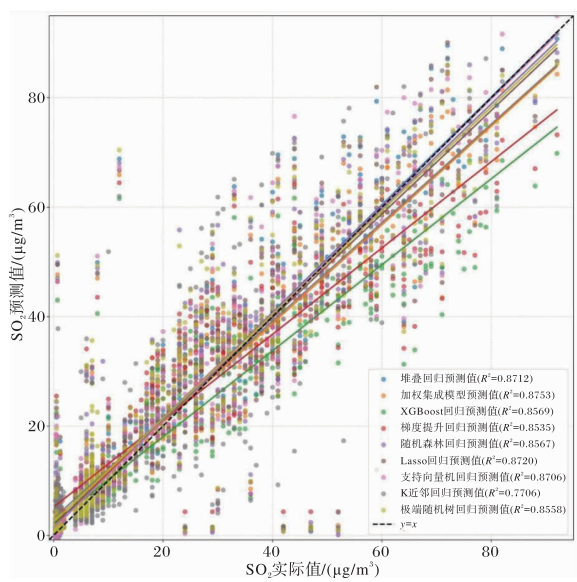

图 11 SO_2 测试集预测结果

值 0.88，且 E_{RMS} 和 E_{MA} 分别为 7.42 和 4.14，表明其在预测误差方面优于其他模型。此外，加权集成模型在训练集和测试集之间的性能差距较小，

进一步证明了其良好的泛化能力和模型稳定性。

加权集成模型在本研究中表现最佳,兼具高泛化能力和低预测误差,且过拟合现象较轻。因此,推荐将加权集成模型作为主要的预测工具。然而,堆叠回归器、LASSO 回归和 SVR 也表现出色,适合作为次优选项或在特定应用场景下的备选模型。针对随机森林、极端树和 KNN 模型,需要通过增加正则化手段或优化模型复杂度来减轻过拟合问题,以提升其在测试集上的表现。未来可以进一步优化集成策略,如探索不同的权重分配方法或引入更多样化的基模型,以进一步提升整体模型的预测性能和稳健性。

表 1　模型预测结果

模型	训练集			测试集		
	R^2	E_{RMS}	E_M	R^2	E_{RMS}	E_M
XGBoost 回归	0.869	8.981	4.806	0.857	7.942	4.198
梯度提升回归	0.885	8.416	5.509	0.854	8.036	5.561
随机森林回归	0.975	3.953	2.273	0.857	7.948	4.299
Lasso 回归	0.897	7.945	4.281	0.872	7.512	3.997
支持向量机回归	0.910	7.440	3.454	0.871	7.552	3.764
K 近邻回归	1.000	0.000	0.000	0.771	10.056	6.160
极端随机树回归	0.983	3.224	1.809	0.856	7.973	4.247
堆叠回归	0.897	7.967	4.044	0.871	7.535	3.818
加权集成模型回归	0.951	5.471	3.034	0.875	7.416	4.168

4　结论

以陕西省西安市长安区 2014—2024 年的 SO_2 和气象观测数据为基础,通过系统应用和比较多种机器学习算法,研究 SO_2 浓度机器学习预测方法,提高 SO_2 浓度预测的准确性和稳定性。

(1)以 7 种机器学习回归模型作为基准模型,通过时间序列五折交叉验证评估模型稳定性,并优化各基准模型的性能。进一步构建堆叠回归模型和加权集成模型,集成多种基准模型的预测结果,提升集成模型整体的预测性能。

(2)构建的加权集成模型通过基于均方根误差(E_{RMS})的权重分配策略,有效整合了各基准模型的优势,在测试集上取得了最优的预测性能($R^2=0.875$,$E_{RMS}=7.416$),明显优于堆叠回归器及其他单一模型。相比之下,随机森林($R^2=0.975$)和 K 近邻($R^2=1.000$)虽在训练集表现卓越,但在测试集出现明显性能衰退(R^2 分别降至 0.857 和 0.771),表明这两种模型在处理复杂特征交互时存在过拟合倾向,而加权集成模型展现了更优的泛化能力和鲁棒性。

(3)尽管受限于单站点数据来源且未纳入具体的交通与工业排放流量数据,但构建的预测模型仍具备较高的实用价值。该模型可应用于区域空气质量监测与预警系统,为环保部门制定差异化管控措施、工业企业优化排放工艺以及敏感人群健康防护提供科学依据。未来可通过引入多源异构数据(如排放清单、交通流)及跨区域数据共享,进一步提升模型的空间泛化能力和预测精度。

参考文献:

[1] 魏夜香,张霄羽,张红. 中国二氧化硫的时空分布及主要排放来源研究[J]. 中国环境科学,2023,43(11):5678-5686.

[2] 李本纲,古陈,王晓利,等. 全球大气硫循环及区域交叉影响[J]. 环境科学学报,2016,36(11):3895-3901.

[3] THEYS N,SMEDT I D,GENT J V,et al. Sulfur dioxide vertical column DOAs retrievals from the ozone monitoring instrument: Global observations and comparison to ground-based and satellite data [J]. Journal of Geophysical Research: Atmos-

pheres,2015,120(6):2470-2491.

[4] 姜磊,何世雄,崔远政. 中国二氧化硫污染治理分析:基于卫星观测数据和空间计量模型的实证[J]. 环境科学学报,2021,41(3):1153-1164.

[5] 陈璇,单晓冉,石兆基,等. 1998—2018 年我国酸雨的时空变化及其原因分析[J]. 资源与生态学报,2021,12(5):593-599.

[6] 王洋,杨敏,吴映梅,等. 城市工业用地规模与二氧化硫排放量之间的关系:来自中国 294 个城市的实证研究[J]. 资源与生态学报,2024,15(4):793-803.

[7] 杨玉盛. 全球环境变化对典型生态系统的影响研究:现状、挑战与发展趋势[J]. 生态学报,2017,37(1):1-11.

[8] JU J,LIU K. LIU F. Prediction of SO_2 Concentration Based on AR-LSTM Neural Network[J]. Neural Process Letters,2023(55):5923-5941.

[9] WU H,HU T,LIU Y,et al. Timesnet: Temporal 2D-variation modeling for general time series analysis[C]. The Eleventh International Conference on Learning Representations, 2023.

[10] SATTARZADEH A R, KUTADINATA R J, PATHIRANA P N, et al. A novel hybrid deep learning model with ARIMA Conv-LSTM networks and shuffle attention layer for short-term traffic flow prediction[J]. Transportmetrica A: Transport Science, 2023:1-23.

[11] WANG Z X, LI J, WU L, et al, Deep learning-based gas-phase chemical kinetics kernel emulator: Application in a global air quality simulation case[J]. Frontiers in Environmental Science,2022,10,955980.

[12] BREIMAN L. Random Forests[J]. Machine Learning,2001,45(1):5-32.

[13] 焦琦融,张典钧,刘文龙. 基于 GBR 方法的 Kp 指数预报模型[J]. 空间科学学报,2024,44(6):1012-1020.

[14] PENG Z,ZHANG B,WANG D,et al. Application of machine learning in atmospheric pollution research: A state-of-art review[J]. Science of the Total Environment,2024,168588.

[15] 苏静,娄英斌,刘语薇,等. 基于机器学习的大气 NO_2 浓度预测模型[J]. 生态毒理学报,2024,19(3): 61-69.

[16] 张波,陆云杰,秦东明,等. 一种卷积自编码深度学习的空气污染多站点联合预测模型[J]. 电子学报,2022,50(6):1410-1427.

[17] 李文,邓升,段妍,等. 时间序列预测与深度学习:文献综述与应用实例[J]. 计算机应用与软件,2020,37(10):64-70+84.

[18] 徐继伟,杨云. 集成学习方法:研究综述[J]. 云南大学学报:自然科学版,2018,40(6):1082-1092.

[19] 唐运军,孙舒畅. 机器学习中的特征工程方法[J]. 汽车实用技术,2020(12):70-72.

[20] 金秀章,刘岳,于静,等. 基于变量选择和 EMD-LSTM 网络的出口 SO_2 浓度预测[J]. 中国电机工程学报,2021,41(24):8475-8484.

[21] 达瓦次仁,落追,洪一航,等. 基于机器学习方法的气溶胶对西藏高原地区雨季降水的影响[J]. 气象与环境学报,2024,40(3): 138-144.

[22] 刘雅楠,吴琼,李勇. 2022 年 5 月江西省降水量气候趋势预测评估分析[J]. 气象与环境学报,2024,40(4): 37-45.

[23] 范国庆,李康辉,高捷,等. 基于核密度估计和 CatBoost 算法的光伏功率预测方法[J]. 上海电力大学学报,2023,39(6):529-535.

[24] WANG T,WANG P,THEYS N,et al. Spatial and temporal changes in SO_2 regimes over China in the recent decade and the driving mechanism[J]. Atmospheric Chemistry and Physics,2018(18):18063-18078.

[25] 张艳晴,金晨曦,闵晶晶,等. 基于机器学习的北京供暖季气温预报误差订正[J]. 气象与环境学报,2025,41(1): 58-65.