

基于 RF-1DCNN-BiGRU 的伊宁市 PM2.5 浓度预测

魏 铭^{1,2} 王可欣^{1,2} 邓秋菊^{1,2}

WEI Ming WANG Kexin DENG Qiuju

摘 要

PM2.5 浓度预测是空气质量预警、污染溯源及精细化管理的核心前提之一,对保障公众健康、提升生态环境治理水平具有重要现实意义,针对传统 PM2.5 浓度预测方法(如 ARIMA、传统机器学习模型等)难以有效捕捉数据中复杂非线性特征与长时序依赖关系、预测精度和稳定性不足的问题,文章提出一种融合随机森林(RF)、一维卷积神经网络(1DCNN)与双向门控循环单元(BiGRU)的组合预测模型。首先利用 RF 进行特征筛选,保留对 PM2.5 预测重要的特征,后利用 1DCNN 提取时序数据中的局部特征,并结合 BiGRU 捕捉 PM2.5 浓度变化的双向时序依赖关系,最终实现 PM2.5 浓度的精准预测。采用均方根误差(RMSE)、平均绝对误差(MAE)及决定系数(R^2)作为模型评价指标,对模型预测性能进行验证。以某地某时间段空气污染与气象观测数据为例,实验结果表明,所提出的 RF-1DCNN-BiGRU 模型能够有效挖掘 PM2.5 浓度与多源影响因素的时序关联特征,具备优异的预测精度与稳定性。该研究为 PM2.5 浓度的时序预测提供了可靠的模型方法与技术参考。

关键词

PM2.5 浓度预测;组合模型;特征选择;BiGRU;1DCNN

doi: 10.3969/j.issn.1672-9528.2026.03.036

0 引言

PM2.5 作为大气中主要的污染物组分,因其粒径小、比表面积大,可携带重金属、多环芳烃等有毒有害物质进入人体呼吸系统,引发哮喘、心血管疾病等多种健康问题,同时还会加剧雾霾、降低大气能见度,对交通出行与生态环境造成显著负面影响。精准预测 PM2.5 浓度变化趋势,对大气污染预警防控、环境管理决策制定及公众健康防护具有重要的现实意义。

PM2.5 浓度预测方法主要是传统统计模型、机器学习模型和深度学习模型。刘拥民等^[1]提出了一种融合分解集成框架与 Pearson 相关系数去噪的 XGBoost-ARIMA 混合预测模型预测石家庄市的 PM2.5 浓度值;张顺顺等^[2]提出基于主成分分析与支持向量机(PCA-SVM)的二次预报模型,用于优化 WRF-CMAQ 模式的空气质量预报效果,但 PCA 降低特征维度可能会丢失关键特征信息;王馨陆等^[3]以成都为对象,筛选污染物时空分布关键影响因子,用 2016—2018 年数据,通过多元线性回归、BP 神经网络和随机森林算法,构建夏

季臭氧、冬季 PM2.5 污染潜势模型,其中多元线性回归、随机森林模型分别对臭氧、PM2.5 预报性能最优;周恒左等^[4]提出基于误差变权倒数组合方法的 LSTM-XGBoost 模型,对 CMAQ 模拟的兰州近地面臭氧浓度进行订正;徐翠玲等^[5]提出考虑数据时空异质性的 STXGBoost 模型,结合多源数据估算关中地区 2022 年小时级 PM2.5 浓度;李佳成等^[6]采用 DT、RF 等四类树模型预测 AQI,结合稳定性选择法分析变量贡献。结果显示 XGBoost 与 RF 预测效果最优,PM2.5、PM10、风速、气压及部分经济变量对 AQI 影响较大。

上述传统统计模型和机器学习模型难以捕捉 PM2.5 浓度时序数据中复杂的非线性特征及多源影响因素的耦合关系,因此预测精度受到限制。深度学习模型凭借强大的特征提取与时序建模能力,在 PM2.5 预测领域得到广泛应用。Liu 等^[7]提出用麻雀搜索算法(SSA)优化 LSTM 模型的超参数,5 种模型预测 6 项空气质量指标及 AQI 的时/日/周浓度。结果显示,神经网络模型精度优于树模型,其中 LSTM-SSA 最优;刘倩等^[8]采集 2017—2020 年重庆市气象与空气质量数据,经皮尔逊相关性分析筛选变量,构建多维多步长 LSTM 预测模型。研究发现,重庆 AQI 与 PM2.5、PM10 强正相关,与能见度、湿度等强负相关;Chu 等^[9]提出 RF-CEEMDAN-Attention-LSTM 混合框架:先用随机森林筛选高相关影响因子,再用 CEEMDAN 将污染物序列分解为多维子序列,最后引入注意力与 dropout 的双隐藏层 LSTM 捕捉非线性动态;彭豪杰等^[10]提出深度学习与随机森林结

1. 重庆移通学院 重庆 401420

2. 公共大数据安全重庆市重点实验室 重庆 401420

[基金项目] 重庆市教委科学技术研究项目(KJZD-M202502401);重庆市教委科学技术研究项目(KJQN202502402)

合的 PMCOM 组合模型。该模型融合 AOD 遥感、气象再分析及 PM2.5 地面观测数据，以深度学习提取深层隐含特征，再输入随机森林训练预测；李明明等^[11]利用 2016—2020 年太原市气象与空气质量资料，结合 EOF 分解、随机森林及 CNN-LSTM 模型，探究 PM2.5 时空变化与预测。

针对 PM2.5 预测中特征冗余、时序特征捕捉不足的问题，本文提出了 RF-1DCNN-BiGRU 组合模型。其中，RF 能进行特征筛选，剔除无关变量，之后采用一维卷积神经网络（1DCNN）擅长提取时序数据中的局部关键特征，可有效挖掘污染物与气象因子间的短期关联；双向门控循环单元（BiGRU）则通过前向与后向两个独立的隐藏层结构，同时捕捉时序序列的正向与反向依赖关系，全面挖掘 PM2.5 浓度的长短期动态变化规律。

1 相关理论介绍

1.1 随机森林

随机森林（RF）是 Leo 于 2001 年提出的集成监督学习算法，通过构建并集成多棵决策树的预测结果，降低单一决策树的过拟合风险，提升模型泛化能力与预测精度。该算法通过 Bootstrap 有放回采样的样本随机化（袋外样本可用于无偏评估）和节点分裂时的特征随机化（从随机子集遴选最优特征）保障基决策树的多样性；预测阶段，分类任务采用多数投票，回归任务则取各树输出的均值。

1.2 一维卷积神经网络

1DCNN 的核心特征在于其卷积核仅沿单一维度进行滑动，通过对序列数据的一维方向执行卷积运算，实现对序列内部局部特征的有效挖掘与提取。在时间序列数据处理场景中，局部时序关联的精准捕捉是提升模型性能的关键前提。1DCNN 的上述结构设计使其能够充分适配时序数据的内在特性，在处理含有时序依赖性的任务中展现出优异性能，其卷积核沿时间维度滑动的特性，1DCNN 能精准捕捉 PM2.5 浓度与前序气象及污染物变量间的局部时序依赖，无需人工特征工程即可自动挖掘潜在动态规律。其权重共享与局部连通结构参数量少、计算高效，适用于大规模多维 PM2.5 时序数据。池化层则可增强特征鲁棒性、抑制噪声。

图 1 为 1DCNN 的基本结构，包含输入层、卷积层、池化层、全连接层及输出层。数据经输入层进入卷积层，从局部输入中进行特征提取，具体运算为：

$$x_k^l = f\left(\sum_{i=1}^j x_i^{l-1} * \omega_i^l + b_i^l\right) \quad (1)$$

式中： x_k^l 是第 l 层第 k 个输出特征映射； f 是激活函数用于引入非线性； x_i^{l-1} 是第 $l-1$ 层第 i 个输入特征映射； $*$ 为卷积运算符； ω_i^l 是第 l 层对应第 i 个输入的卷积核权重； b_i^l 是第 l 层对应第 i 个输入的偏置项。

随后通过池化层来降低特征维度以减少计算量，最大池

化计算公式为：

$$p_i^{l+1}(j) = \max_{(j-w+1) \leq t \leq jw} \{q_i^l(t)\} \quad (2)$$

式中： $p_i^{l+1}(j)$ 为池化操作的输出值； $q_i^l(t)$ 为第 l 层的第 i 个特征矢量的第 t 个神经元的值； w 为池化区域的宽度。

最终由全连接层和输出层共同完成分类或回归预测任务。

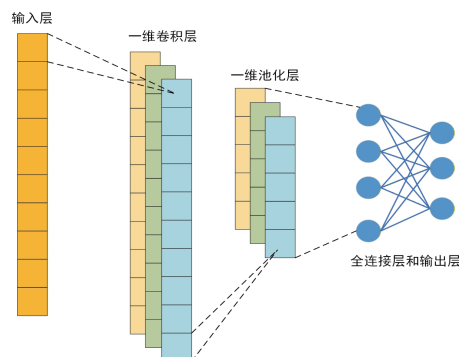


图 1 一维 CNN 结构图

1.3 双向门控循环单元

循环神经网络（RNN）虽能处理时序数据，却因梯度消失问题难以捕捉长期依赖关系。门控循环单元（GRU）作为 LSTM 的轻量化改进模型，通过重置门与更新门两个核心门控机制，实现了对信息的选择性传递与遗忘：重置门控制前一时刻状态信息的保留比例，更新门则决定当前时刻状态的更新权重，从而在简化网络结构的同时，保证对长期时序依赖的有效建模。双向门控循环单元（bidirectional GRU, BiGRU）在 GRU 基础上引入双向结构设计，由前向 GRU 与后向 GRU 两个子网络组成，如图 2 所示。

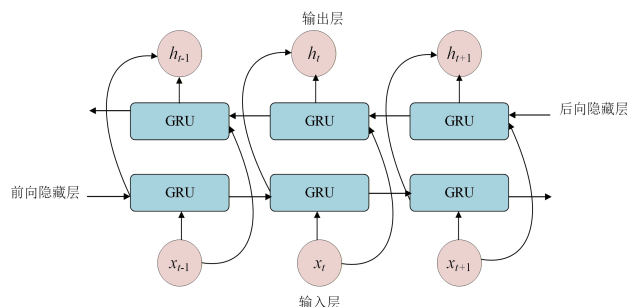


图 2 BiGRU 结构图

前向 GRU 从时序数据的起始端向末端传递信息，捕捉正向时序依赖；后向 GRU 则从末端向起始端传递信息，挖掘反向时序关联。最终，BiGRU 将两个方向的输出特征进行拼接融合，能够更全面地提取时序数据的上下文信息，BiGRU 的内部计算过程为：

$$\begin{cases} \vec{h}_t = \text{GRU}(x_t, \vec{h}_{t-1}) \\ \overleftarrow{h}_t = \text{GRU}(x_t, \overleftarrow{h}_{t+1}) \\ h_t = W \vec{h}_t + V \overleftarrow{h}_t + b_t \end{cases} \quad (3)$$

式中： $\vec{h}_t$ 和 $\vec{h}_t^{\leftarrow}$ 分别是 t 时刻的前向隐藏状态和后向隐藏状态；权重矩阵 W 、 V 分别对前向、后向隐藏状态进行线性变换；再叠加偏置项 b_t ，最终得到融合了双向时序信息的BiGRU的隐藏状态 h_t 。

2 数据预处理

2.1 数据来源

本文选取的数据集包含新疆伊宁市2020年1月1日至2021年1月1日期间每小时的6种主要空气污染物变量和6种气象变量的数据，其空气污染数据来自国家空气质量发布平台（<https://air.cnemc.cn:18007/>），气象数据来自中央气象台（<http://www.nmc.cn>），数据集的特征包括PM2.5、PM10、SO₂、CO、NO₂和O₃共6种空气污染物变量，以及温度、大气压、湿度、风速、风向、降雨量6种气象变量。

2.2 缺失值处理

针对数据集中存在的少量缺失值，本文采用线性插值法进行填补。该方法在保持时间序列连续性的同时，具有计算简便、物理意义明确等优势，适用于此类平滑且变化缓慢的环境监测数据。线性插值法计算公式为：

$$y = y_1 + \frac{x - x_1}{x_2 - x_1} \times (y_2 - y_1) \quad (4)$$

式中： y 表示在缺失时间点 x 处通过线性插值估算得到的某特征补全值； x_1 和 x_2 分别表示 x 前后最近的两个有效采样时间点，对应的特征观测值分别为 y_1 和 y_2 。

2.3 归一化与反归一化处理

本文采用Min-Max归一化方法处理，将所有特征值映射至[0,1]区间，构建维度统一、分布标准化的时序输入序列，归一化公式为：

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (5)$$

式中： x 表示原数据； x^* 表示归一化后的数据； $x_{\max}$ 表示输入数据中的最大值； $x_{\min}$ 表示输入数据中的最小值，该方法不改变原始数据的相对趋势，仅通过线性比例调整数据尺度，通过线性变换将原始数据映射至[0,1]区间内。

训练好的模型在测试集上评估性能，预测后的数据需要反归一化，用公式表示为：

$$x = (x_{\max} - x_{\min})x_i + x_{\min} \quad (6)$$

2.4 模型性能评价指标

本文采用均方根误差（RMSE）、平均绝对误差（MAE）和决定系数（ R^2 ）指标对模型性能进行评估。RMSE反映了预测值与真实值之间偏差的总体大小，MAE衡量预测误差的平均绝对幅度， R^2 则表征模型对目标变量变异的解释能力，取值越接近1，拟合效果越好，各指标的计算公式为：

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (7)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (8)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (9)$$

式中： n 表示测试集的样本规模； y_i 表示第 i 个样本的PM2.5浓度的实际值； $\hat{y}_i$ 表示第 i 个样本的PM2.5浓度的预测值； $\bar{y}$ 表示实际PM2.5浓度的平均值。

3 实验结果

3.1 随机森林特征筛选

RF模型因其良好的易用性、预测精度与鲁棒性，被广泛视为一种有效的特征重要性评估方法。本文基于RF模型量化各特征对PM2.5浓度的贡献度，据此筛选关键输入变量，以提升模型构建效率与计算性能。图3展示了各特征对PM2.5浓度预测的特征重要性排序，本文选取排名前8的特征以及PM2.5本身作为最终模型的输入特征。

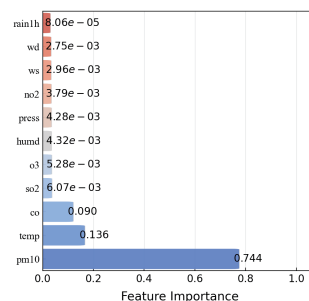


图3 各特征的特征重要性排序

3.2 模型参数设置

本文所用RF-1DCNN-BiGRU模型的参数设置为：第一个一维卷积层配置了64个卷积核，每个卷积核的尺寸为3，步长为1，填充方式采用“causal”；第二个一维卷积层则包含32个卷积核，每个卷积核的尺寸为1，步长同样为1，一维最大池化层的池化窗口尺寸为2。学习率为2e-4，epoch为200，batch_size为128，优化器为Adam，损失函数为MSE，输出层的激活函数为ReLU，时间步长为15，采用早停法，容忍轮次设置为25。

3.3 模型预测结果对比

表1呈现了4种模型在PM2.5浓度预测任务中的性能表现。图4为PM2.5浓度预测值与真实值的散点拟合图。从误差指标来看，从指标MAE来看，1DCNN-BiGRU模型的预测精度较单一1DCNN、BiGRU模型分别提升约12.8%、10.7%；从指标RMSE计算，其预测精度较单一1DCNN、BiGRU模型分别提升约12.5%、9.2%。其散点集中性略弱于RF-1DCNN-BiGRU，存在少量轻微偏离的点；拟合线斜率为0.980。

表 1 不同模型的预测精度

模型	RMSE/($\mu\text{g}\cdot\text{m}^3$)	MAE/($\mu\text{g}\cdot\text{m}^3$)	R^2
1DCNN	15.836	11.726	0.934
BiGRU	15.256	11.448	0.939
1DCNN-BiGRU	13.859	10.221	0.949
RF-1DCNN-BiGRU	13.263	9.388	0.954

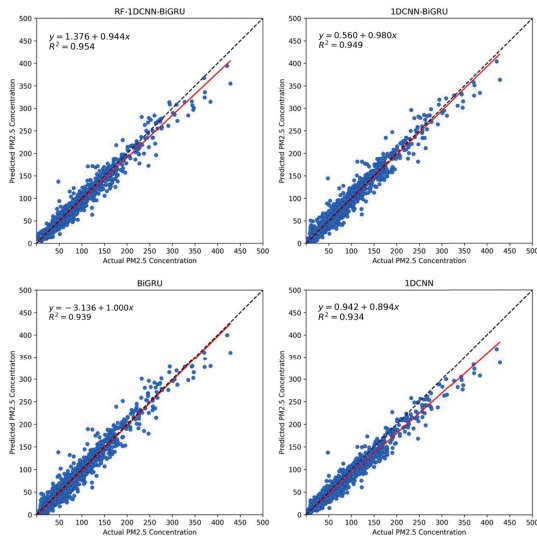


图 4 预测值与真实值的散点拟合图

进一步引入随机森林特征选择模块后, RF-1DCNN-BiGRU 模型在各项指标中均表现最优: 其 RMSE 降至 $13.263\text{ }\mu\text{g}/\text{m}^3$, MAE 仅为 $9.388\text{ }\mu\text{g}/\text{m}^3$, 决定系数 R^2 则提升至 0.954, 其散点高度集中在拟合直线附近, 几乎无明显偏离的离群点。这一结果表明, 随机森林的特征筛选机制有效过滤了冗余特征的干扰, 既降低了模型的学习负担, 也强化了对 PM2.5 浓度核心影响特征的聚焦能力, 最终实现了预测精度与泛化能力的双重提升。

4 结论

本文针对 PM2.5 浓度预测任务, 提出了融合 RF、1DCNN 与 BiGRU 的 RF-1DCNN-BiGRU 模型, 并通过对比实验验证了其性能优势。实验结果表明: (1) 单一模型 (1DCNN、BiGRU) 受限于特征提取或时序建模的单一能力, 预测精度与泛化性不足; (2) 1DCNN-BiGRU 融合模型通过局部特征与长时序依赖的协同建模, 预测精度要优于单一模型; (3) 引入 RF 特征筛选后, RF-1DCNN-BiGRU 模型的 RMSE 降至 $13.263\text{ }\mu\text{g}/\text{m}^3$ 、MAE 达 $9.388\text{ }\mu\text{g}/\text{m}^3$ 、 R^2 提升至 0.954, 模型性能进一步得到提升, 在对比模型中实现了最好的预测结果。

研究证实, RF-1DCNN-BiGRU 模型通过“特征筛选-局部特征提取-时序建模”的分层框架, 有效平衡了 PM2.5 浓度预测的精度与泛化性, 可为空气质量预警、污染管控等

场景提供可靠的技术支撑。后续研究可进一步优化特征筛选策略, 并结合多源数据(如气象、污染源)拓展模型的适用范围。

参考文献:

- [1] 刘拥民, 罗皓懿, 谢铁强. 基于 XGBoost-ARIMA 方法的 PM2.5 质量浓度预测模型的研究及应用 [J]. 安全与环境学报, 2023,23(1):211-221.
- [2] 张顺顺, 卢彦希, 罗崴, 等. 基于 AQI 的空气污染物预报研究 [J]. 能源与环境, 2024(1):97-99.
- [3] 王馨陆, 黄冉, 张雯娟, 等. 基于机器学习方法的臭氧和 PM2.5 污染潜势预报模型: 以成都市为例 [J]. 北京大学学报(自然科学版), 2021,57(5):938-950.
- [4] 周恒左, 陈恒蕊, 廖鹏, 等. 兰州市 CMAQ 近地面臭氧模拟结果的订正方法: 基于机器学习方法 [J]. 中国环境科学, 2022, 42(12):5472-5483.
- [5] 徐翠玲, 袁兵, 胡雪, 等. 基于时空 XGBoost 模型的关中地区小时级 PM2.5 质量浓度估算 [J/OL]. 环境科学, 1-17[2026-02-24]. <https://doi.org/10.13227/j.hjkk.202412262>.
- [6] 李佳成, 梁龙跃. 基于机器学习方法的空气质量预测与影响因素识别 [J]. 计算机技术与发展, 2024,34(1):164-170.
- [7] LIU X P, GUO H Y. Air quality indicators and AQI prediction coupling long-short term memory (LSTM) and sparrow search algorithm (SSA): a case study of Shanghai[J]. Atmospheric pollution research, 2022, 13(10):101551.
- [8] 刘倩, 王兆茹, 胥寒莉, 等. 基于粒子群优化 XGBoost 模型的 PM2.5 质量浓度反演 [J]. 环境科学, 2025,46(8):5023-5034.
- [9] CHU Y Y, YAO J, QIAO D W, et al. Three-hourly PM2. 5 and O₃ concentrations prediction based on time series decomposition and LSTM model with attention mechanism[J]. Atmospheric pollution research, 2023, 14(11): 101879.
- [10] 彭豪杰, 周杨, 胡校飞, 等. 基于深度学习与随机森林的 PM2.5 浓度预测模型 [J]. 遥感学报, 2023,27(2):430-440.
- [11] 李明明, 王小兰, 岳江, 等. 融合 EOF 分解及 CNN-LSTM 神经网络的 PM2.5 预测 [J]. 环境科学, 2025,46(2):715-726.

【作者简介】

魏铭(1998—), 男, 重庆永川人, 硕士, 助教, 研究方向: 深度学习。

王可欣(1992—), 女, 湖南岳阳人, 硕士, 讲师, 研究方向: 区块链技术及应用。

邓秋菊(1987—), 女, 重庆铜梁人, 硕士研究生, 教授, 研究方向: 数字图像处理。

(收稿日期: 2026-01-06 修回日期: 2026-03-11)