

## 基于 LSTM-XGBoost 组合模型的南昌市空气质量预测研究

张宇恒<sup>1</sup> 钟豪杰<sup>1</sup> 徐 亮<sup>1,2</sup>

ZHANG Yuheng ZHONG Haojie XU Liang

## 摘 要

为解决传统单一 RNN 模型在空气质量时序预测中存在的长期依赖捕捉不足、易出现梯度消失与爆炸等问题,文章提出了一种基于 LSTM-XGBoost 的组合空气质量预测方法。该方法通过 LSTM(长短期记忆网络)的门控机制精准捕捉空气质量数据的长期依赖关系,利用 XGBoost 基于梯度提升决策树框架的特征重要性分析与集成学习优势强化模型非线性拟合能力,并采用 Stacking 集成策略实现两种模型的优势互补。以江西省南昌市 2024 年上半年 AQI 数据为研究对象开展实证分析,结果表明:组合模型的  $\rho$  指标较单一 LSTM 和 XGBoost 模型分别提升 0.01% 和 0.13%, nRMSE 指标分别降低 4.5% 和 1.5%, MAPE 指标分别降低 22.5% 和 9.0%; 2024 年上半年南昌市空气质量整体处于“良好”等级。提出的组合模型为城市空气质量精准预测提供了更可靠的方法支撑,具有一定的理论参考与实际应用价值。

## 关键词

空气质量指数; LSTM; XGBoost; 组合模型

doi: 10.3969/j.issn.1672-9528.2025.11.039

## 0 引言

随着工业化进程的加快,空气污染成为严重环境问题之一<sup>[1]</sup>,国家越来越重视空气质量的重要性,因此实现空气质量的准确预测也日益重要。文琴<sup>[2]</sup>基于 LSTM 模型对四川省内江市空气质量指数进行预测,但预测值与实际值存在一定差距。刘旭<sup>[3]</sup>采用 XGBoost 模型与非随机样本增容相结合的方法进行空气质量分类研究,但只考虑了满足正态分布的数据,而实际研究中并不是所有的大气组分数据均满足正态分布。刘敏等人<sup>[4]</sup>通过计算每个权重的误差来调整 BP 神经网络的连接权重,从而预测分析呼和浩特市空气质量,但忽略了 BP 神经网络的预测结果较难收敛这一问题。周闯等人<sup>[5]</sup>采用主成分分析应用于合肥市的空气质量评估,最终拟合结果与真实值同样存在差距。以上各模型虽都能实现空气质量的预测分析,但预测效果都不太理想。

Semmelmann 等人<sup>[6]</sup>结合 LSTM 和 XGBoost 模型,显著提高了智能能源社区提前一天负载预测的准确性。李子熠等人<sup>[7]</sup>也采用了 LSTM-XGBoost 组合模型对美国驻中国北

京大使馆地区的空气中  $PM_{2.5}$  含量进行了准确预测。在其他预测空气质量的模型中,很多学者运用了其他组合模型。吉丹俊<sup>[8]</sup>使用 ARIMA、Prophet 以及 LSTM 模型对扬州市 AQI 进行时间序列分析预测,最终发现 LSTM 模型的预测结果比较准确。栗治杰等人<sup>[9]</sup>对比了单一 LSTM、RNN、ARIMA(1, 1, 1)模型和 PCA-LSTM-Attention 组合模型,验证了组合模型的高效与精确。路凯丽等人<sup>[10]</sup>对比了 CNN-GRU 集成深度学习模型、广义回归、优化 BP 神经网络模型来预测 AQI,最后实验得出了组合 CNN-GRU 模型误差值均小于 0.2,同样体现了组合模型的优势。胡青等人<sup>[11]</sup>采用的是 IPSO-BP 组合模型,与未优化 BP 神经网络相比,显著提升了其预测精度及拟合优度。因此,为进一步提升空气质量预测模型的准确度,本文采用 LSTM-XGBoost 组合模型对南昌市空气质量进行预测分析,选取了江西省南昌市 2024 年上半年六种空气组分( $PM_{2.5}$ 、 $SO_2$ 、 $PM_{10}$ 、 $NO_2$ 、 $O_3$  及  $CO$ )的空气质量指数 AQI 作为参数样本,然后 LSTM 引入记忆单元以及一系列门控机制来提升技术来提高模型准确性和泛化能力,以此更进一步来提高对 AQI 的预测精确性。

## 1 理论介绍

## 1.1 AQI: 空气质量指数

AQI 是一个根据空气成分占比来定义的指标, $PM_{10}$ 、 $PM_{2.5}$ 、 $SO_2$ 、 $NO_2$ 、 $O_3$ 、 $CO$  六种大气污染物是主要的参考指标,形式简化为概念性指数值。从 6 个主要的空气污染物 AQI 值中挑选出最大值,将其定义为当天空气的 AQI 值。通过 AQI

1. 东华理工大学人工智能与信息工程学院 江西南昌 330013  
2. 东华理工大学江西省核地学数据科学与系统工程技术研究中心 江西南昌 330013  
[基金项目] 江西省生态文明建设制度研究中心项目 (JXST2204); 江西省教育厅基金项目 (GJJ2200751); 东华理工大学博士科研启动基金项目 (DHBK2019291)

值对空气污染程度和空气质量状况进行了分级表示，详细的划分标准如表 1 所示。计算方法为：

$$I_{AQI}=\frac{I_h-I_l}{C_h-C_l}+I_l \quad (1)$$

式中：AQI 的浓度值为  $I_{AQI}$ ， $I$  是对应于  $C_h$  的指数限值； $I_l$  是对应于  $C_l$  的指数限值， $C_l \leq C$  的浓度限值， $C_h \geq C$  的浓度限值。

表 1 空气质量指数分级参考表

| AQI | 0~50 | 51~100 | 101~150 | 151~200 | 201~300 | >300 |
|-----|------|--------|---------|---------|---------|------|
| 级别  | 一级   | 二级     | 三级      | 四级      | 五级      | 六级   |
| 类别  | 优秀   | 良好     | 轻污染     | 中污染     | 重污染     | 严重污染 |

## 1.2 数据平稳性检验

为了使训练结果更加精确，在训练前通过时序图来检验数据的平稳性，研究平稳性很差的数据并没有实际意义。时序图检验，是一种基于时间序列数据可视化呈现的时间序列分析方法，常用来初步判断数据的平稳性。对于平稳性的判定，通常认为一系列数据围绕一个常数均值进行上下波动，并且要求波动范围保持相对稳定。首先将收集的 2024 年上半年南昌市的空气质量数据进行数据清洗，筛选出包含 AQI 数值的项，同时确保数据按照时间顺序排序，每个时间段都有对应数值。

### 1.2.1 时间序列图检验

对数据进行时间序列图绘制，如图 1 所示，大部分数据在均值为 54.09 附近，基本趋于稳定波动。

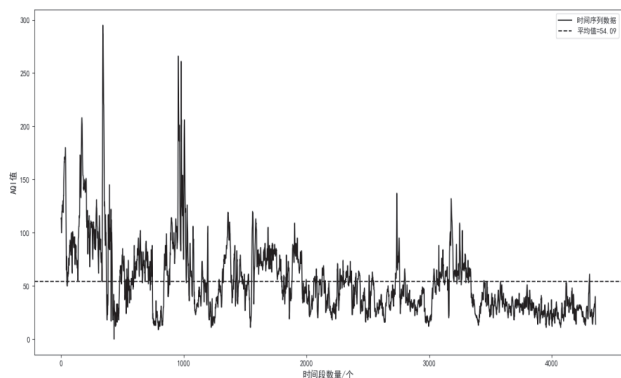


图 1 时间序列图检验

### 1.2.2 自相关检验

自相关检验用于分析时间序列数据中现有值与预测值之间的相关性，自相关的计算过程为：根据滞后期  $n$  将序列拆为等长的两个序列，计算这两个序列的相关性，得到整体滞后期为  $n$  时的自相关性，根据公式计算出结果为列表：[1, 0.978 237 35]，第一个元素是滞后期为 0 时的自相关性，第二个元素是滞后期为 1 时的自相关系数。

$$S^2(X)=\frac{1}{n}\sum_{i=1}^n(X_i-\bar{X})^2 \quad (2)$$

$$r(n)=\frac{1}{n}\sum_{i=1}^n(A_i-\bar{X})(B_i-\bar{X}) \quad (3)$$

$$ACF(n)=\frac{r(n)}{S^2(X)} \quad (4)$$

式中： $S^2(X)$  为样本方差； $\bar{X}$  为样本均值； $r(n)$  为自协方差； $n$  为滞后期； $A_i$  和  $B_i$  为两个不同观测值。

如图 2 所示，ACF 图中数据逐渐趋于 0，且主要集中在正值区域，说明时间序列具有正相关性，可判断数据基本处于平稳状态。

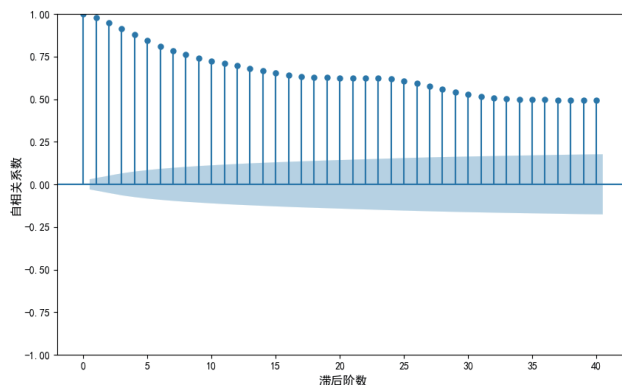


图 2 数据自相关图

### 1.2.3 Hurst 指数检验

仅根据时序检验图与数据自相关图来得出结论是不足以说明的，本文引入数据的平稳性采用 Hurst 指数检验法。黎明辉<sup>[12]</sup>基于 Hurst 指数，对量化投资策略进行了研究，Hurst 指数检验是分析时间序列长期依赖性和分形特征的重要方法，适用于空气质量数据等具有潜在长记忆性的序列，基于本文研究背景，选用该方法进行检验。以下是 Hurst 指数检验步骤：

(1) 数据准备及明确计算原理。去除原始数据中的 NaN 值，确认有效数据点个数 4 358。根据 Hurst 指数原理，量化时间序列的长期记忆性 ( $0 < H < 1$ )，其中：

- ①  $H=0.5$ ：随机游走，即无记忆性；
- ②  $0.5 < H < 1$ ：持续性，代表趋势增强；
- ③  $0 < H < 0.5$ ：反持续性，均值回归。

(2) 改进  $R/S$  分析（重标极差分析）计算 Hurst 指数<sup>[13]</sup>。通过计算不同时间尺度子序列极差和标准差比值，对子序列进行最小划分尺度为 10，最大划分长度为 1 089 ( $4\ 358/4$ )，采用对数间隔生成 20 个不同尺度。计算  $R/S$  公式为：

$$R/S=\frac{\text{极差}}{\text{标准差}} \quad (5)$$

(3) 线性回归，拟合 Hurst 指数。对子序列长度

$\log(n)$  和对应的  $\log(R/S)$  进行线性回归, 通过  $\log(R/S) = H \times \log(n) + C$ , 计算得到 Hurst 指数  $H=0.929$ , 接近 1 表明时间序列具有强持续性。这里的  $C$  是回归方程的截距, 提供序列基础波动水平的信息, 不影响整体趋势, 取  $C=0$ 。

(4) 使用 Hurst 库计算对比结果。通过库方法运算,  $H=0.721$ , 具体和 R/S 方法的差异在于库方法未去均值以及未优化子序列划分, 导致估计值偏低, 但结果仍介于 0.5~1 之间。

(5) 可视化验证。通过图 3 的线性关系, 验证了时间序列的强持续性。

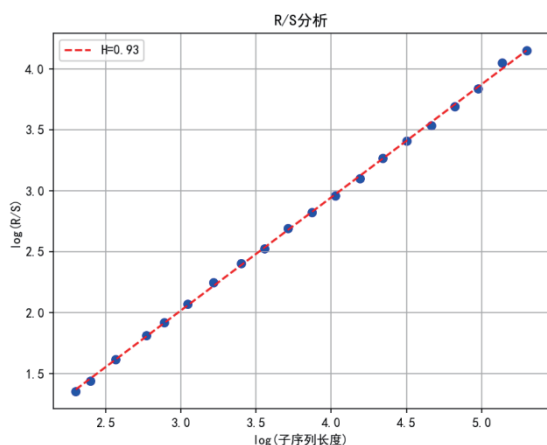


图 3 R/S 分析图

### 1.3 传统 RNN、LSTM 模型基本介绍

LSTM 是一种机器学习的特殊循环神经网络, 不仅解决了传统 RNN<sup>[14]</sup> 在处理长时间序列数据时所遇到的梯度消失或梯度爆炸的相关问题, 而且在预测上也表现出色, 基本 RNN 结构图如图 4 所示。

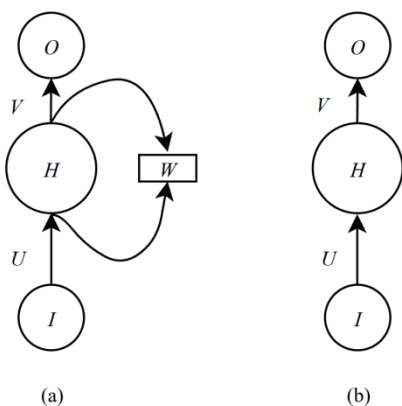


图 4 基本 RNN 结构图

具体所遇到的梯度消失和梯度爆炸的相关问题体现:

(1) 参数更新停滞: 损失函数 LOSS 下降缓慢, 50 轮训练后损失仅从 120.5 降到 119.8, 经过十余轮迭代几乎无改进。

(2) 激活函数饱和: 统计隐藏层输出的数值范围时,

使用 Tanh 激活函数输出集中在  $[-0.1, 0.1]$ , 证明梯度接近于 0。

(3) 训练数据不稳定时: 对于训练数据, 前几项为 146→159→210→253→285 陡峭式上升, RNN 在反向传播时梯度连乘导致爆炸。

因此, 由于 RNN 在长时间序列数据上处理缺陷, 本文引入的 LSTM-XGBoost 模型中的 LSTM 结构与工作原理如图 5 所示。

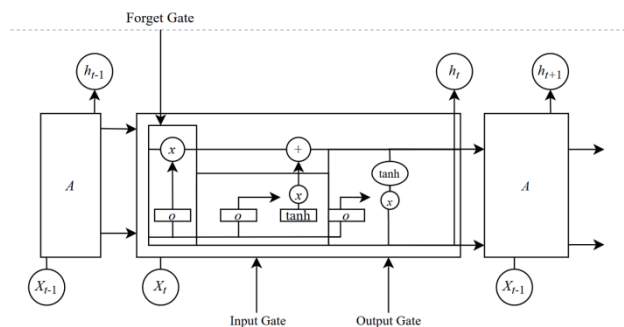


图 5 LSTM 结构与工作原理

其中各部分的数学公式为:

$$(1) \text{输入门 } i_t: i_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_i) \quad (6)$$

将之前需要留下的信息和现在需要记住的信息相加, 得到了新记忆状态。其中  $W_f$  为遗忘门的权重矩阵;  $b_i$  为对应的偏置项;  $h_{t-1}$  为上个时间步的隐藏状态;  $x_t$  为当前时间步的输入。

$$(2) \text{遗忘门 } f_t: f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (7)$$

通过  $x_t$  和  $h_t$  的操作, 并经过 sigmoid 激活函数, 得到 0, 1 的向量, 0 代表之前的记忆某一部分要忘记, 1 代表之前记忆需要留下的部分。其中 sigmoid 激活函数定义为  $\sigma$ ; 遗忘门的偏置项为  $b_f$ ; 上一个时间步的隐藏状态为  $h_{t-1}$ ; 当前时间步的输入为  $x_t$ 。

$$(3) \text{细胞状态 } C_t: C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (8)$$

式中:  $\tilde{C}_t$  是候选细胞状态。

在每个时间步中, 当前输入  $x_t$  和前一隐藏状态  $h_{t-1}$  的信息经过处理生成了一个候选细胞状态, 该候选细胞状态包含可能加入长期状态的信息。

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (9)$$

候选细胞状态的权重矩阵是  $W_c$ , 其包含当前时间步的输出信息, 而且是基于细胞状态的过滤输出。

$$(4) \text{输出门 } O_t:$$

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (10)$$

$$h_t = O_t \cdot \tanh(C_t)$$

式中: 通过整合  $C_t$ , 得到一个输出。其中,  $W_o$  是输出门的权重矩阵;  $b_o$  是输出门的偏置项;  $h_t$  是隐藏状态。

LSTM 的训练过程主要包括以下步骤: 进行数据准备,

包括收集和预处理训练模型所需的数据,接着构建 LSTM 网络结构,并初始化模型参数。然后进行前向传播,计算遗忘门、输入门、候选细胞状态和更新后的细胞状态,以及输出门和损失函数,之后进行反向传播,计算梯度并更新模型参数。在多次迭代前向传播和反向传播的过程中,不断验证模型性能,最后对模型进行评估和调优,以达到最佳的训练效果。

沿时间的反向传播和向上一层的反向传播是 LSTM 误差项的反向传播的两个层面,共同构成了 LSTM 网络的训练过程,通过不断更新网络参数来提高网络的性能和准确性,然后跳转第一步重复完成整个流程,直至网络的相关误差小于给定值才结束。

#### 1.4 XGBoost 模型基本理论介绍

XGBoost 是 Gradient Boosting 的一个优化版本并且基于决策树的集成方法。其原理是把基分类器层层叠加,训练每一层时,给予前一层基分类器分错的样本更高的权重,目标函数为:

$$\text{Obj}^{(l)} = \sum_{i=1}^n l(\hat{y}_i^{(l)}, (\hat{y}_i^{(l-1)} + f_i(x_i))) + \Omega(f_i) + \text{constant} \quad (11)$$

式中:  $\Omega(f_i)$  是控制树的复杂度的正则项;  $\text{constant}$  是常数项;  $\hat{y}_i^{(l-1)} + f_i(x_i)$  是新树的预测值;  $\hat{y}_i^{(l)}$  是将树的个数的结果进行求和。

模型步骤如下:首先初始化模型,然后计算残差并据此构建新的决策树;接着更新模型并迭代这一过程,在迭代过程中,对学习速率、最大深度等关键参数进行调整以优化模型性能;同时,关注目标函数和正则化的设置,以确保模型的准确性和泛化能力;最后,输出经过优化训练的模型结果。模型原理如图 6 所示。

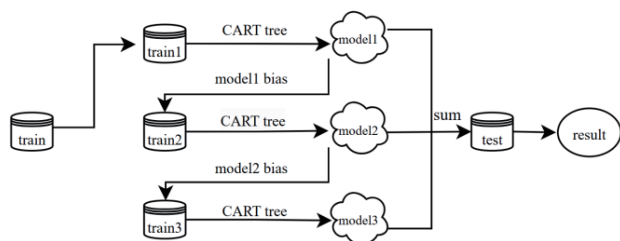


图 6 XGBoost 模型原理图

## 2 数据处理

### 2.1 数据来源

本文所选取的数据集来自 <https://quotsoft.net/air/#archive>, 把 2024 年上半年数据作为数据集, 通过数据采集与预处理技术检测到列为 type 的参数包含 AQI、PM<sub>2.5</sub>、PM<sub>2.5\_24h</sub>、PM<sub>10</sub>、PM<sub>10\_24h</sub>、SO<sub>2</sub>、SO<sub>2\_24h</sub>、NO<sub>2</sub>、NO<sub>2\_24h</sub>、O<sub>3</sub>、O<sub>3\_8h</sub>、O<sub>3\_24h</sub>、O<sub>3\_8h\_24h</sub>、CO、CO<sub>24h</sub>

共 15 类数据, 对其中参数为 AQI 的所有行和列进行输出, 再选取列为南昌市的数据, 遍历从 2024 年 1 月 1 日 0 时至 6 月 30 日 24 时的小时数据。以 70% 训练集、30% 测试集比例进行集中输出, 统一格式, 输出包含所有需要数据的 list 文件, 共有 4 358 条数据。对于数据的平稳性检验, 本文采用时序图与自相关性检验方法, ACF 图中数据逐渐趋于 0, 并且主要集中在正值区域, 说明时间序列具有正相关性, 可判断数据基本处于平稳状态。

### 2.2 数据归一化处理

杨超文等人<sup>[15]</sup>对数据归一化方法进行了详细介绍。数据读取时为 int 列表, 在准备划分数据集之前, 把数据转为二维 NumPy 数组进行归一化, 归一化后的数据都在 [0,1] 范围内, 计算公式为:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (12)$$

式中:  $X_{\text{norm}}$  为归一化后的数据;  $X$  为原始数据;  $X_{\min}$  和  $X_{\max}$  为对应数据集中的最小值和最大值。

### 2.3 对训练后的数据反归一化

组合模型 LSTM-XGBoost 训练结束以后, 得到的数据仍然在 [0,1] 之间, 最后我们通过相反操作对数据进行反归一化来完成对 AQI 值的还原。具体反归一化的计算公式为:

$$X = (X_{\max} - X_{\min}) \cdot X_{\text{norm}} + X_{\min} \quad (13)$$

### 2.4 组合模型构建特点

通过堆叠集成 XGBoost 和 LSTM 两个模型, 利用 XGBoost 强大的特征学习能力增强 LSTM 模型的输入特征, 再结合 LSTM 模型优秀的时序建模能力, 捕捉数据中长期和短期依赖关系, 提高对空气质量预测模型的性能和泛化能力。这种方法充分利用了数据中的空间信息和时间信息, 增强了两个模型之间的互补性和协同效应, 从而提高了预测模型的鲁棒性。

### 2.5 模型评价指标及效果分析

本文采用标准化均方根误差 (nRMSE)、平均绝对百分比误差 (MAPE) 及斯皮尔曼等级相关 ( $\rho$ ) 三个指标评价模型的预测性能。nRMSE 有利于比较不同单位或范围数据集的预测准确性, MAPE 更关注相对误差且对于不同量级的预测问题更具可比性<sup>[16]</sup>,  $\rho$  对于数据中的异常值或缺失值而言相对稳健。三个指标的判断标准如下: nRMSE 的值越低, 表示模型预测性能越好。MAPE 越小, 表示模型预测性能越好。当  $\rho$  接近 1 时, 表示两个变量的等级或排序完全一致, 正相关性强, 反之越接近于 -1, 负相关性越强。



以下是三个评估指标的计算公式：

$$nRMSE=\frac{1}{y_{\max}-y_{\min}}\sqrt{\frac{1}{n}\sum_{i=1}^n(y_i'-y_i)^2}$$
 (14)

$$MAPE=\frac{1}{N}\sum_{i=1}^n\left|\frac{y_i'-y_i}{y_i}\right|$$
 (15)

$$\rho=1-\frac{6\sum d_i^2}{n^3-n}$$
 (16)

由表 2 可知，文本提出 LSTM-XGBoost 组合模型的性能比其他两个单一模型好。因为  $\rho$  值越接近 1 正相关性越强且 nRMSE、MAPE 的值越小，模型精度越高 [17]，可见以上 3 个模型在接近 1 的同时，组合模型比其他两个单一模型更加接近 1。

表 2 各模型评估指标数据统计表

| 指标     | LSTM    | XGBoost | 组合模型    | 改进幅度          |
|--------|---------|---------|---------|---------------|
| $\rho$ | 0.961 3 | 0.960 1 | 0.961 4 | +0.01%~+0.13% |
| nRMSE  | 0.078 4 | 0.048 6 | 0.033 3 | -57.5%~-31.5% |
| MAPE   | 0.312 3 | 0.176 8 | 0.086 7 | -72.2%~-50.9% |

组合模型的 nRMSE 和 MAPE 两个误差指标数值都比其他两个单一模型小，足以证明组合模型更优，具体表现在：

（1）误差最小化：组合模型的 nRMSE 比最佳单一模型（XGBoost）再降低 31.5%。

（2）稳定性提升：MAPE 从 17.68% 降至 8.67%，证明对异常值的鲁棒性增强。

3 结果分析

3.1 模型参数设置

为了简要进行模型间的对比，本文设置的参数都比较小。其中，设置的 LSTM 是一个简单的顺序神经网络，包含一个具有 4 个单元的 LSTM 层、一个 ReLU 激活层以及一个具有单个输出的全连接层。该模型采用了 MSE 作为损失函数，并使用 Adam 优化器进行训练。对于 XGBoost 模型，是一个由 50 棵最大深度为 1 的树组成的梯度提升回归器，采用 0.1 的学习率，无 gamma 值限制，无 L1 正则化，旨在通过集成学习提升预测性能。

组合模型则是在数据通过 LSTM 模型训练后再进入 XGBoost 模型进行训练，可见所得的结果都比单一模型拟合度好，预测结果更加准确。该模型训练过程如下：通过堆叠集成 XGBoost 和 LSTM 两个模型来增强预测性能。首先，利用 XGBoost 模型强大的特征学习能力，挖掘数据中的非线性关系，并将 XGBoost 的预测结果作为新的特征添加到原始特征中。然后，利用 LSTM 模型优秀的时序建模能力，对增强

后的特征数据进行训练，捕捉数据中的长期和短期依赖关系。这种集成方法充分利用了数据中的空间信息和时间信息，提高了两个模型之间的互补性和预测模型的泛化能力、鲁棒性。

3.2 模型预测图

图 7~9 为三种模型预测图。

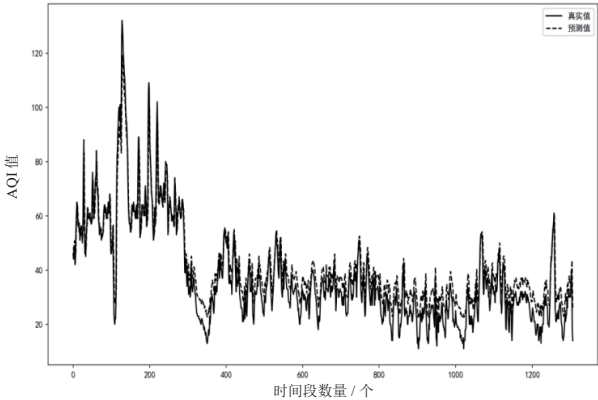


图 7 LSTM 预测图

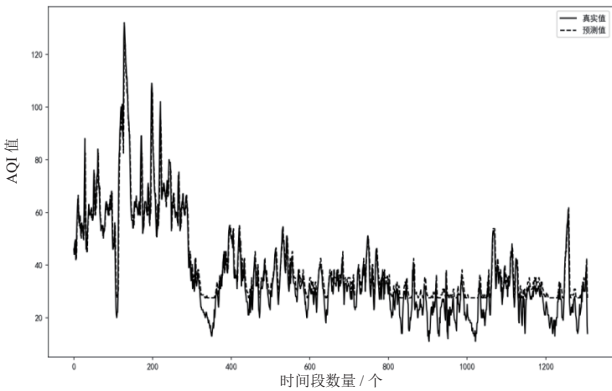


图 8 XGBoost 预测图

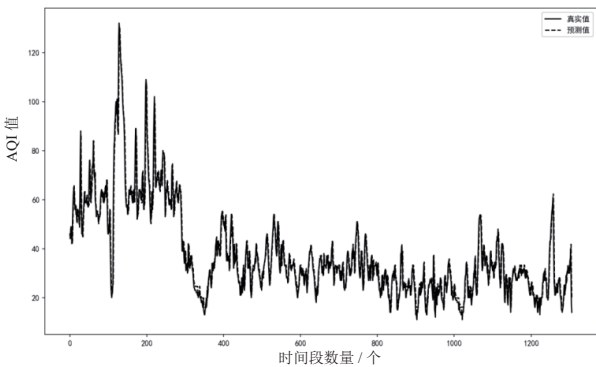


图 9 LSTM-XGBoost 预测图

通过以上模型训练结果可以看出，LSTM 和 XGBoost 预测图中真实值和预测值还存在一定差距，而组合模型的真实值和预测值基本重合，具有较高的预测准确性以及较好的泛化能力，由此体现了 LSTM-XGBoost 组合模型在空气质量指数预测分析方面的优越性，表 3 是三种模型整体拟合效果参考。

表 3 三种模型拟合效果参考

| 模型           | 趋势匹配度 | 局部波动捕捉 | 极端值预测 |
|--------------|-------|--------|-------|
| LSTM         | 中     | 弱      | 差     |
| XGBoost      | 良     | 中      | 中     |
| LSTM-XGBoost | 优     | 良      | 强     |

4 结语

本文采用 LSTM-XGBoost 组合模型旨在提高空气质量指数预测的准确性。LSTM 模型可以很好地处理长序列长期依赖性并且并行计算性能好，XGBoost 模型有益于处理缺失值，提高预测性能。综合两个模型的优点，本文以南昌市 2024 上半年数据为例进行组合模型测试，发现组合模型预测效果优于单一的 LSTM 模型和 XGBoost 模型，同时误差也比两个单一模型小。根据空气质量指数分级参考表和各模型评估指标数据统计表可见，本文预测分析的南昌市空气质量在 2024 年上半年处于“良好”类别，组合模型的标准化均方根误差、平均绝对百分比误差和斯皮尔曼等级相关系数等指标均优于单一模型，LSTM-XGBoost 组合模型的预测效果更好。

参考文献:

[1] 李华湘. 空气质量是否影响人口规模—基于十三个城市面板数据的实证研究 [J]. 经济研究导刊, 2024 (14): 64-67.

[2] 文琴. 基于 LSTM 的空气质量指数预测系统研究 [J]. 信息技术与信息化, 2022,(7): 19-22.

[3] 刘旭. 非随机样本增容 XGBoost 模型的空气质量分类研究 [D]. 昆明: 云南大学, 2022.

[4] 刘敏, 扈华. 基于 BP 神经网络的空气质量等级预测模型构建 [J]. 内蒙古农业大学学报 (自然科学版), 2023, 44 (4): 40-46.

[5] 周闯, 张琦锦, 郭映映, 等. 主成分分析法在合肥市空气质量评估中的应用 [J]. 大气与环境光学学报, 2024, 19 (4): 479-488.

[6] SEMMELMANN L, HENNI S, WEINHARDT C. Load forecasting for energy communities: a novel LSTM-XGBoost hybrid model based on smart meter data [J]. Energy informatics, 2022, 5 :102-106.

[7] 李子熠, 张天宇, 李鸿强. 基于 XGBoost 和 LSTM 组合模型的 PM2.5 浓度预测 [J]. 河北建筑工程学院学报, 2023, 41 (4): 219-223.

[8] 吉丹俊. 城市空气质量指数的时间序列分析和预测 - 以扬州市为例 [J]. 无锡商业职业技术学院学报, 2024, 24 (1):

59-65.

[9] 栗治杰, 贾东水. 基于时空规律的 PCA-LSTM-Attention 空气质量预测研究 [J]. 环境科学与管理, 2024, 49 (11): 172-177.

[10] 路凯丽, 杨露, 李涛. 基于集成深度学习模型的空气质量指数预测 [J]. 南京信息工程大学学报, 2024, 16 (1): 56-65.

[11] 胡青, 胡珍, 曲润, 等. 基于 IPSO-BP 的空气质量指数预测 [J]. 西安文理学院学报 (自然科学版), 2023, 26 (1): 92-98.

[12] 黎明辉. 基于 Hurst 指数的量化投资策略研究 [J]. 现代信息科技, 2023, 7 (22): 83-87.

[13] 宾雄志, 尹磊建, 袁帅, 等. 基于 R/S 分析法的井筒偏斜量监测数据分析研究 [J]. 建井技术, 2025, 46 (1): 85-90.

[14] 范竣翔, 李琦, 朱亚杰, 等. 基于 RNN 的空气污染时空预报模型研究 [J]. 测绘科学, 2017, 42 (7): 76-83.

[15] 杨超文, 汪宇玲, 舒志敏, 等. 基于 TensorFlow 的 LSTM 模型在空气质量指数预测的应用 [J]. 数字技术与应用, 2021, 39 (3): 203-206.

[16] 史学良, 李梁, 赵清华. 基于改进 LSTM 网络的空气质量指数预测 [J]. 统计与决策, 2021, 37 (16): 57-60.

[17] 诸鑫, 林孝先, 刘庆红, 等. 基于神经网络模型的绵阳市空气质量指数预测 [J]. 绵阳师范学院学报, 2023, 42 (8): 112-118.

【作者简介】

张宇恒 (2004—), 男, 贵州黔东南人, 本科, 研究方向: 数据科学与大数据技术。

钟豪杰 (2003—), 男, 江西新余人, 本科, 研究方向: 数据科学与大数据技术。

徐亮 (1990—), 男, 江苏泰州人, 博士研究生, 讲师, 研究方向: 地学大数据与人工智能。

(收稿日期: 2025-05-24 修回日期: 2025-11-17)