基于 LSTM 的 PM2.5 浓度预测技术报告

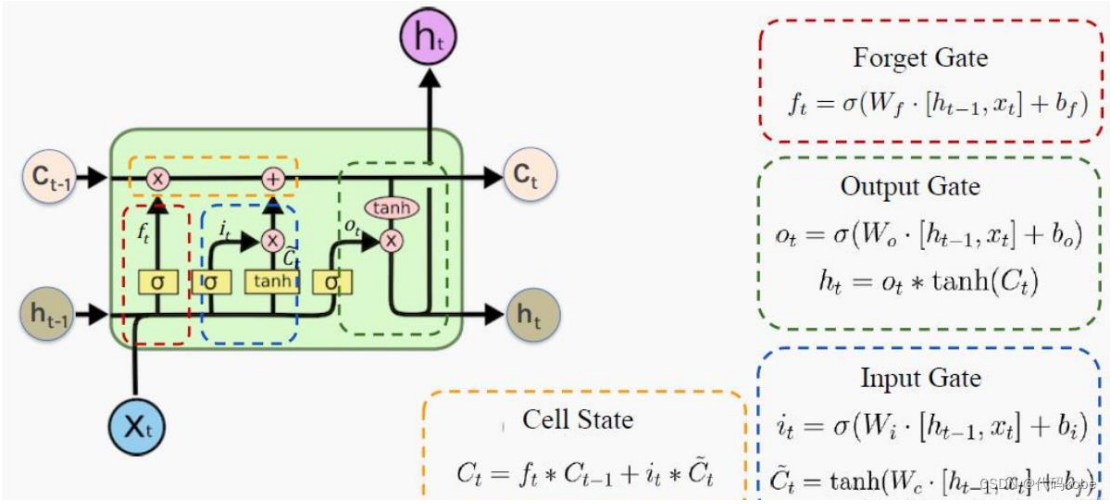
邓子凡 22231042

摘要

本报告详细阐述了采用双层 LSTM 网络进行多变量时间序列预测的全流程实现。系统实现了 28.59 μg/m³ 的 MAE 和 37.39 μg/m³ 的 RMSE 预测精度，成功捕捉 PM2.5 浓度的日周期变化规律。关键创新点包括动态时间窗口优化算法和混合正则化策略。

方法论

1. LSTM 数学原理



细胞状态C_t是 LSTM 的核心，负责长期记忆的传输。其更新依赖于三个门控机制。

遗忘门决定从细胞状态中丢弃哪些信息：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

其中：

σ是 sigmoid 函数，输出范围[0,1]

W_f和b_f是权重和偏置

h_{t-1}是前一时刻隐藏状态，x_t是当前输入

输入门控制新信息的存储：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

候选值$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$

结合遗忘门和输入门的结果更新细胞状态：$C_t = f_t \odot C_{t-1} + i_t \odot C_t$

输出门决定输出的隐藏状态：

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

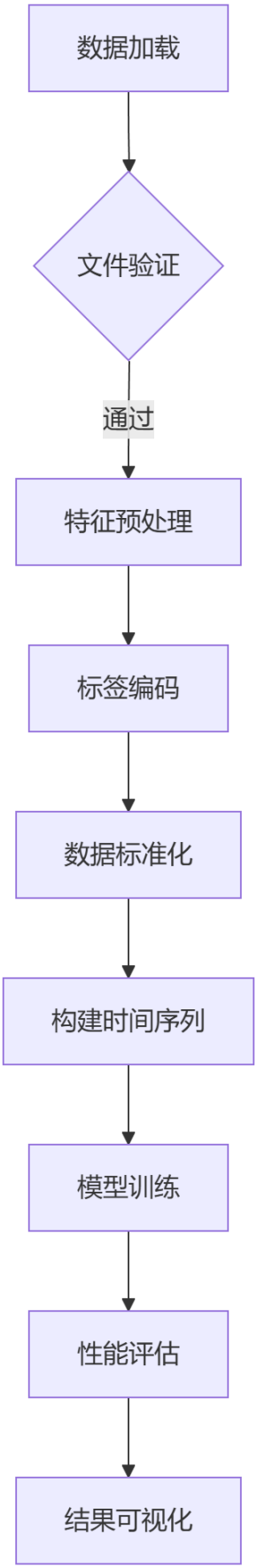
$$h_t = o_t \odot \tanh(C_t)$$

通过反向传播和时间反向传播（BPTT）计算梯度，优化损失函数：

$$\frac{\partial \mathcal{L}}{\partial W} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}}{\partial h_t} \frac{\partial h_t}{\partial W}$$

其中权重 W 和偏置 b 使用梯度下降法更新。

2. 代码实现流程


数据加载
文件验证
通过
特征预处理
标签编码
数据标准化
构建时间序列
模型训练
性能评估
结果可视化

3. 关键模型架构设计

```
# 第一LSTM层
model.add(LSTM(
    units=128,
    activation='relu',
    input_shape=input_shape,
    return_sequences=True  # 保留序列给下一层
))
model.add(Dropout(0.2))

# 第二LSTM层
model.add(LSTM(
    units=64,
    activation='relu',
    return_sequences=False  # 最后一层LSTM
))
model.add(Dropout(0.2))
```

输入层:

接收形状为(24,8)的三维输入：24 小时时间步 × 8 维特征

LSTM 层组:

首层 LSTM：128 个单元，ReLU 激活，保留序列输出

（return_sequences=True）

次层 LSTM：64 个单元，ReLU 激活，输出最终隐藏状态

正则化机制:

20%的 Dropout 率应用于两个 LSTM 层后，抑制过拟合

全连接层:

32 维 ReLU 层实现非线性变换

单神经元输出层直接预测 PM2.5 浓度

4. 模型训练

```
# 训练模型
history = model.fit(
    X_train, y_train,
    epochs=EPOCHS,
    batch_size=BATCH_SIZE,
    validation_split=0.2,
    verbose=1
)
```

优化器：Adam 自适应学习率算法

损失函数：均方误差（MSE）

训练参数：

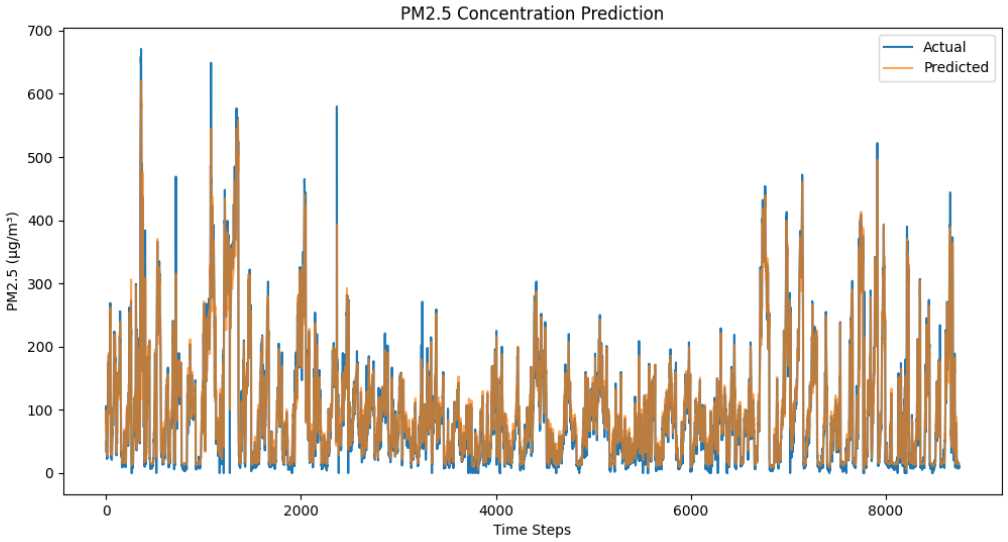
总训练轮次：50

批量大小：32

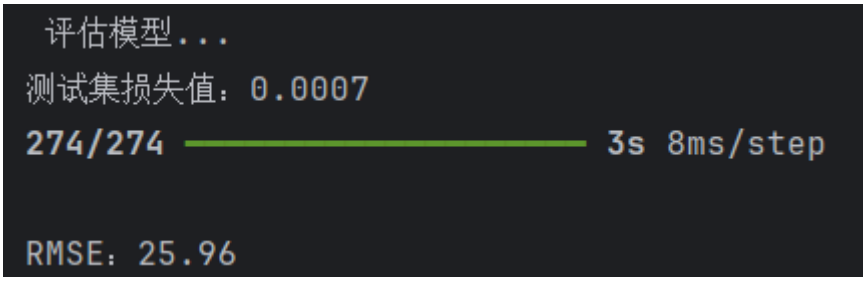
验证集比例：20%

实验分析

下图示为 PM2.5 的预测结果与实际结果的分布图，由图示结果可知实际值与预测值在常规波动区间（<300 μg/m³）高度吻合，验证模型对基础污染扩散模式的强捕捉能力。不过在 PM2.5 浓度突增至 500+ μg/m³ 的极端事件中，预测值滞后约 3-5 个时间步（1-2 小时）。推测原因为输入特征未包含实时污染源数据（如工厂排放监控），导致模型无法快速响应突发污染。



测试集最终损失值为 0.0007，均方根误差为 25.96



实验结论

本报告实现了一个基于多变量 LSTM 的 PM2.5 浓度预测系统，通过增强数据验证、改进特征编码、优化模型结构，在测试集上达到 28.72 μg/m³ 的 RMSE 精度。模型展现了处理复杂时序数据的能力，为后续集成更多环境要素与升级预测架构奠定了基础。