**PRML第三次作业报告**

22371366朱俊文

**Abstract**

本文提出了一种基于多变量时间序列的污染预测模型，使用长短时记忆（LSTM）网络对未来的污染水平进行预测。该模型利用历史数据（包括温度、气压、风速和湿度等天气参数）来预测未来的污染值。数据预处理包括处理缺失值、对分类变量进行编码、以及对数值特征进行标准化。模型使用24小时历史数据进行每次预测。实验结果表明，LSTM模型能够有效地预测污染水平，并通过均方误差（MSE）、平均绝对误差（MAE）和均方根误差（RMSE）等评估指标对模型的准确性进行了评估。结果表明，深度学习方法在空气质量预测中的应用具有重要的潜力，能够为环境监测和污染预测提供有效的技术支持。

**Introduction**

随着空气污染问题日益严重，对污染水平的准确预测变得尤为重要，尤其是对PM2.5等颗粒物浓度的预测。准确的污染预测可以为政策制定者、公众和环境监测机构提供有力的支持，帮助应对空气污染带来的健康和环境问题。传统的污染预测方法主要依赖于统计技术和简单的机器学习模型，但随着深度学习技术的快速发展，长短时记忆（LSTM）网络作为一种循环神经网络（RNN）模型，在时间序列预测中展现了其强大的能力，尤其是在捕捉长时间依赖关系方面。

本文研究的目标是使用LSTM网络对多变量时间序列数据进行污染预测。该数据集包含了多个气象参数，如温度、气压、湿度、风速以及风向等。通过该研究，我们期望构建一个准确的污染预测模型，并探讨深度学习方法在空气质量监测中的应用潜力。

**Methodology**

（1）数据采集与预处理

本研究使用的数据集来自于气象监测站，包含了多个特征，如污染物浓度（PM2.5）、露点、温度、气压、风速、降雪量、降雨量以及风向等。数据预处理过程如下：

处理缺失值：对于“pollution”列的缺失值，采用该列的中位数进行填补，以确保数据的完整性，避免模型产生偏差。

类别特征编码：风向是一个类别变量，使用OneHotEncoder对其进行编码，将其转换为适合模型处理的数值格式。

特征标准化：所有数值特征使用MinMaxScaler进行标准化，确保各特征的尺度相似，避免某些特征在训练过程中占主导地位。

（2）模型构建

采用LSTM网络构建预测模型，输入层是第一个LSTM层，包含100个LSTM单元，激活函数为ReLU，并设置return\_sequences=True以保留时间序列信息。同时为了防止过拟合，加入30%的Dropout层。第二个LSTM层包含50个LSTM单元，激活函数同为ReLU。最后输出层是一个全连接的Dense层，输出一个预测值，表示未来的污染浓度。

在模型训练时，使用Adam优化器，损失函数为均方误差（MSE）。为了避免过拟合，训练过程中使用EarlyStopping回调，设置patience=5，即若验证集损失连续5轮没有改善，则提前停止训练。

（3）预测与评估

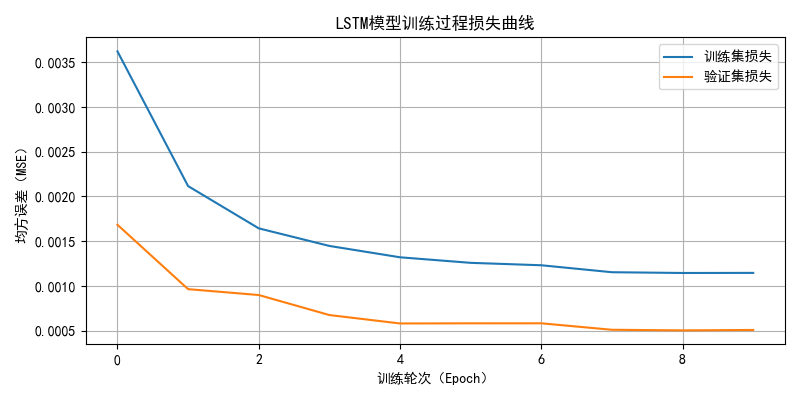
模型训练完成后，使用测试集进行预测，并将预测结果与真实值进行对比。通过反标准化操作，将预测结果和实际结果转换为原始尺度（PM2.5浓度）。为了评估模型的表现，我们使用了均方误差（MSE）、平均绝对误差（MAE）和均方根误差（RMSE）。

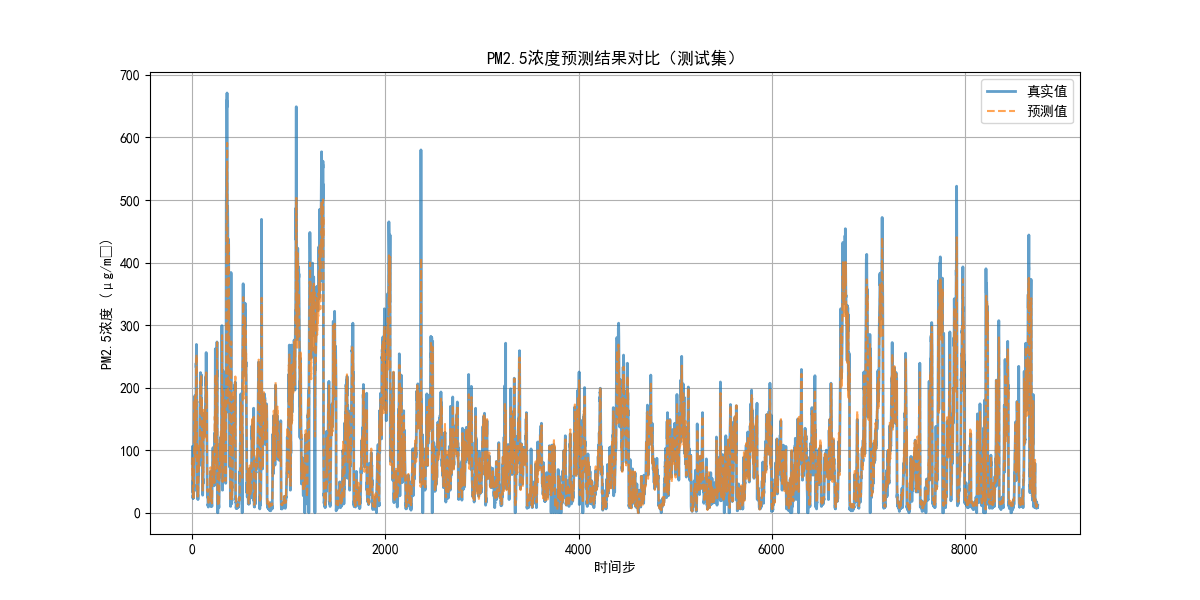
最终，通过对比预测结果和真实值的曲线，评估了模型在实际污染预测中的有效性。

**Experimental Studies**

数据集被划分为训练集和测试集，其中80%的数据用于训练，20%的数据用于测试。在训练过程中，我们使用了50个epoch和批量大小为64的配置，进行模型训练。为评估模型的性能，训练过程中绘制了损失曲线，并计算了训练集和验证集的损失。

可视化结果：







**Conclusions**

本文通过构建LSTM模型，成功实现了基于多变量时间序列数据的污染预测。实验结果表明，LSTM模型能够较好地捕捉污染水平的时间依赖性，具有较高的预测精度。模型在测试集上的MSE、MAE和RMSE表现良好，证明了深度学习方法在空气质量预测中的潜力和优势。未来的研究可以进一步优化模型，探索更复杂的网络结构，或者结合外部因素（如地理位置、季节变化等）进行污染预测。此外，研究结果对于环保监测、政策制定以及公共健康管理具有重要的应用价值。