工业污染源在线监测数据的智能分析与预警模型构建

洪长吉[1]，普剑峰[2*]，谭强华[3]，黄俊[4]，罗小明[2]

1.西双版纳巅峰环境检测有限公司，云南 西双版纳 666100

2.西双版纳州生态环境局景洪分局生态环境监测站，云南 西双版纳 666100

3.西双版纳州生态环境局勐腊分局生态环境监测站，云南 西双版纳 666100

4.西双版纳州生态环境局，云南 西双版纳 666100

DOI: 10.61369/EAE.2025010017

摘　　要： 本论文围绕工业污染源在线监测数据的智能分析与预警模型构建展开研究。针对传统监测数据处理效率低、预警滞后等问题，提出基于大数据与机器学习的智能分析方法，通过数据清洗、特征提取、模型训练等流程构建预警模型。研究结果表明，该模型能够有效提升数据处理精度与预警及时性，为工业污染源监管提供科学决策支持，对推动环境保护智能化发展具有重要意义。

关 键 词： 工业污染源；在线监测；智能分析；预警模型；机器学习

Intelligent Analysis of Online Monitoring Data of Industrial Pollution Sources and Construction of Early Warning Models

Hong Changji[1], Pu Jianfeng[2*], Tan Qianghua[3], Huang Jun[4], Luo Xiaoming[2]

1.Xishuangbanna Peak Environmental Testing Co., LTD., Xishuangbanna, Yunnan 666100

2. Ecological Environment Monitoring Station, Jinghong Branch of Xishuangbanna Prefecture Ecological Environment Bureau, Xishuangbanna, Yunnan 666100

3. Ecological Environment Monitoring Station, Mengla Branch of Xishuangbanna Prefecture Ecological Environment Bureau, Xishuangbanna, Yunnan 666100

4. Xishuangbanna Prefecture Ecological Environment Bureau, Xishuangbanna, Yunnan 666100

Abstract: This thesis focuses on the intelligent analysis of online monitoring data of industrial pollution sources and the construction of early warning models. Aiming at the problems of low processing efficiency and delayed early warning of traditional monitoring data, an intelligent analysis method based on big data and machine learning is proposed, and an early warning model is constructed through processes such as data cleaning, feature extraction, and model training. The research results show that this model can effectively improve the accuracy of data processing and the timeliness of early warning, provide scientific decision support for the supervision of industrial pollution sources, and is of great significance for promoting the intelligentdevelopment of environmental protection.

Keywords: industrial pollution source; online monitoring; intelligent analysis;early warning model; machine learning

伴随着工业化进程的不断加速，工业污染的问题越来越严重。工业污染源在线监测是环境监管中的一种重要方法，累积了大量的数据，然而传统的分析方式很难发掘数据中的潜在价值，不能满足预警的实时性和准确性要求。基于这一背景，建立智能分析和预警模型以实现工业污染源的有效监管已成为目前环境保护领域中的一个研究重点。本论文研究目的在于探讨工业污染源在线监测数据智能分析的方法，建立科学高效的预警模型以帮助环境监管水平的提高。

一、工业污染源在线监测数据的特点

工业污染源在线监测数据作为环境监管体系中的核心依据，具有明显的复杂性和动态性特点，深刻地影响了智能分析预警模型建设策略。在数据来源方面，工业污染源在线监测数据涵盖了废水，废气和固体废弃物等多类污染物的排放监控，涵盖了化工，冶金和电力等高污染行业生产全过程。监测设备种类繁多，其中包括水质自动分析仪、烟气连续监测系统（CEMS）、颗粒物监测仪等，这些设备都是通过传感器来实现的、物联网终端等来实现对数据的实时采集和发送，构成多源异构数据集，从数据

作者简介：洪长吉（1983.12—），男，汉族，云南丽江人，学历：本科，职称：中级职称，研究方向：环境检测、环境工程类。

类型水平上看，工业污染源在线监测数据表现出结构化和非结构化共存的特征。结构化数据，例如污染物浓度数值，监测时间戳和设备编号，逻辑关系和存储格式明确，易于直接统计分析；非结构化数据包括了设备的运行日志、图像和视频数据等，这些数据包含了丰富的信息，但需要使用特定的技术手段进行分析和处理。另外，监测数据中也含有时序数据的特点，污染物排放浓度和流量都会随着时间推移而动态地变化，从而构成了一个带有时间序列特点的数据流，它的变化规律受到生产周期，工艺调整和环境因素等诸多因素的影响[1]。

二、智能分析方法的研究

（一）数据清洗和预处理

在采集和传输工业污染源的在线监测数据的过程中，容易受到传感器误差，网络波动和设备故障的干扰而造成数据中缺失值，异常值和噪声数据的出现。数据清洗和预处理是确保分析准确性至关重要的环节。对于缺失值的处理，对于污染物浓度和流量的时序数据通常采用线性插值法和三次样条插值法对其进行填补[2]。例如，在某一特定时间段内，如果废水的pH值数据出现缺失，可以使用三次样条插值法，并根据之前和之后的监测数据来构建一个平滑的曲线，以拟合这些缺失的数据点，并进行相应的测试，这种方法填充连续性数据时，误差率可以被限制在允许范围内。对异常值采用3σ原则辨识，数据与均值偏差大于3倍标准差后判断为异常情况，并对某燃煤电厂SO_2排放浓度进行了监控，利用这一原理成功地辨识出了传感器故障引起的超标异常，校正后的数据波动与生产实际情况相符。

在处理噪声数据时，结合了中值滤波和小波去噪的方法。对颗粒物浓度监测数据进行处理时，将中值滤波窗口布置成5×5可以有效地去除突发脉冲噪声；对含有高频噪声信号利用db4小波基对其进行三层分解和重构，去噪后信噪比提高到25dB，数据平滑度明显提高。此外，针对多源异构数据，需进行格式统一与标准化处理，将不同设备输出的浓度单位（如mg/L、ppm）统一转换为法定计量单位，通过Min-Max标准化方法将数据归一化至[0,1]区间，消除量纲差异对后续分析的影响，确保数据的可用性与一致性。

（二）数据特征提取和降维

工业污染源监测的数据维度多具有信息冗余的特点，特征提取和降维可以有效地降低数据的复杂度和提高分析效率。在特征提取中，对于时序数据利用傅里叶变换和小波变换进行频域特征提取。以某个化工园区的废水流量数据为研究对象，利用离散傅里叶变换技术，将时域信号转化为频域信号，并从中提取了主要的频率成分及其对应的能量占比，从而成功地识别了生产周期中的特征频率波动。同时通过滑动窗口对均值，方差和峰值的统计特征进行统计并构造多维特征向量。

在降维技术中，选择了将主成分分析（PCA）与局部线性嵌入（LLE）结合起来的方法。对某钢铁厂废气监测数据进行处理时，原始数据含有SO_2，NOx和颗粒物浓度共12个特征维度并经过PCA进行处理，前三个主要成分的累计方差贡献率高达92%，成功地将数据维度从12维减少到了3维；对具有非线性分布特征的数据进一步利用LLE算法对其进行优化，使其嵌入到低维空间而又能保持其局部结构，从而有效地解决了PCA对非线性数据降维的限制。实验结果显示，当结合这两种技术时，数据的存储量降低了75%，模型的训练时间减少了60%，同时关键特征的保留率超过了90%，这为后续的智能分析提供了强大的数据支撑。

（三）智能分析算法的选择

工业污染源数据具有非线性，动态性和复杂性等特点，这就给智能分析算法的研究提出了更高的要求，常见的计算方法有支持向量机（SVM）、随机森林（RF）和深度学习技术。在处理小样本数据分类时，SVM展现出了卓越的性能，而在某电镀厂废水中重金属超标的预警环节，选择了采用径向基核函数（RBF）的SVM模型，通过调整C = 10的惩罚参数和 γ = 0.5的核函数参数，成功地实现了对超出标准和正常数据的精确分类，测试集的准确率高达95%。在处理多变量回归和分类任务时，随机森林算法表现出色。在预测某火电厂NOx排放浓度的过程中，构建了一个包含500棵决策树的随机森林模型，通过对特征的重要性进行排序，筛选出了关键的影响因子，预测的均方误差（MSE）达到了0.87，这明显优于传统的线性回归模型[3]。

在深度学习算法领域，长短期记忆网络（LSTM）对于时序数据展现出了卓越的处理性能。在某工业园区的PM2.5浓度预测实验中，搭建包含2层LSTM单元（每层128个神经元）、1层全连接层的网络结构，利用过去72小时的气象数据和污染物浓度作为输入，预测未来24小时的浓度变化，模型在测试集上的平均绝对误差（MAE）为4.2μg/m³，显示出了很好的动态适应性。在实践中，需要结合数据特点和分析目标进行算法综合选择，也可以采用集成学习方法将各种算法优势进行整合，促进模型泛化能力和分析精度的提高[4]。

三、预警模型的建立

（一）模型架构设计

工业污染源预警模型架构需要综合考虑数据处理的效率，预测的准确性和实时性等因素，采取分层模块化设计，底层为数据接入层，通过物联网协议（如MQTT、CoAP）实现多源监测设备的数据实时采集，支持每秒处理1000条以上的数据流，确保数据传输的低延迟（<500ms）[5]。数据经清洗预处理后进入特征工程层，运用滑动窗口技术（窗口的尺寸设定在15分钟的数据量内）提取动态特征，并结合主成分分析（PCA）将原始20维数据降至8维，减少计算复杂度，核心预测层使用混合模型架构并集成了LSTM和XGBoost算法的优点。LSTM网络构建3层结构（每层64个神经元），通过门控机制捕捉污染物浓度的长短期时序依赖，适用于预测连续变化趋势；XGBoost模型利用树模型对非线性数据的出色拟合能力，有效地处理了如气象状况、生产压力等外部因素。两者输出结果经加权融合（权重系数由网格搜索决定，LSTM为0.6，XGBoost为0.4），形成最终预测值[6]。

上层为预警决策层，设置三级预警阈值：轻度超标（浓度超过标准值10%-30%）、中度超标（30%-50%）、重度超标（>50%）。以某石化公司的VOCs排放量为研究对象，其标准值定为80mg/m³。在预测浓度达到88mg/m³的情况下，会触发轻度预警；96mg/m³的浓度会触发中度预警；而120mg/m³的浓度则会触发重度预警。预警信息由消息队列（Kafka）向监管平台实时

推送，响应时间限定为2秒钟[7]。

（二）模型训练和优化

模型训练使用离线训练和在线更新两种策略。离线训练阶段，选取某工业园区近3年的监测数据（共100万条记录）作为训练集，划分70%用于训练、20%用于验证、10%用于测试。在LSTM网络中，学习率被设定为0.001。通过使用Adam优化器并进行50轮的迭代训练，训练集的均方误差（MSE）从最初的25.6下降到了3.2；在XGBoost模型中设定树深度6，学习率0.1和子采样率0.8，经过300次迭代后验证集的AUC为0.92，为了处理数据分布动态变化问题，该模型使用了在线学习机制[8]。当新加入的数据量达到1000条时，系统会启动增量训练，并利用FTRL（Follow-the-Regularized-Leader）算法来更新模型的参数，以确保模型能够适应生产流程的调整或季节性的排放变动。针对模型过拟合问题，在LSTM中加入Dropout层（丢弃率0.2），在XGBoost中采用正则化参数 $\lambda = 0.1$，使测试集MSE稳定在4.1左右，泛化性能显著提升。另外，利用模型蒸馏技术对复杂模型进行压缩，使推理时间由原来的150ms减少到30ms以适应实时预警的需要。

（三）模型评估指标体系

模型评估使用多维度的指标体系覆盖预测的准确性，稳定性和实用性。评估准确性的指标涵盖了均方误差（MSE）、平均绝对误差（MAE）、决定系数（R^2）以及平均绝对百分比误差（MAPE）。在某燃煤电厂进行SO_2浓度的预测时，模型的MSE达到了2.3mg/m³，MAE为1.2mg/m³，R^2为0.94，而MAPE为5.8%，这些数据充分展示了模型的高预测精度，稳定性指标是通过计算不同时间窗口（1小时、6小时、1天）的预测误差波动，从而评估模型对数据波动的抵抗能力。实验表明，该模型在各时间尺度上MSE标准差小于0.5，说明具有较好的稳定性。在实用性指标方面，重点关注预警的时效性和误报率[9]。

四、实验设计和结果分析

本实验选取某大型工业园区作为研究对象，收集近2年的工业污染源在线监测数据，涵盖废气、废水等10类污染物指标，共计120万条数据记录。实验环境采用Python 3.8开发平台，硬件配置为Intel Core i7-12700H处理器、16GB内存，使用TensorFlow 2.8与Scikit-learn 1.1.3框架搭建模型。实验被分为两组：对照组和实验组。对照组使用了传统的时间序列分析模型（ARIMA）和单一的SVM模型，而实验组则采用了本研究提出的LSTM-XGBoost融合预警模型[10]。

以PM2.5的浓度预测为研究对象，实验组的模型平均绝对误差（MAE）达到了3.8μg/m³，均方误差（MSE）为18.2μg²/m⁶，而决定系数（R^2）为0.96；而对照组ARIMA模型MAE为6.5μg/m³，MSE为34.7μg²/m⁶，R^2仅为0.82，SVM模型MAE为5.2μg/m³，MSE为26.3μg²/m⁶，R^2为0.89，实验组在复杂时序数据处理上优势显著。在进行预警性能的测试时，对于突然超出标准的事件，实验组的模型平均提前了1.5小时发出预警，预警的准确率高达92%，而误报率仅为2.5%；对照组的预警时间只有0.8小时，其准确性为78%，而误报率高达8%，从模型的运行效率角度来看，经过模型的压缩和优化，实验组的单条数据推理时间缩短到了28ms，与优化前相比减少了62%，从而满足了实时监测的需求。通过Shapley值分析可知，生产负荷（贡献度32%）、气象湿度（25%）、设备运行参数（20%）为影响污染物排放的关键因素。试验结果表明：文中所构建的融合模型无论从预测精度还是预警及时性和稳定性等方面都要好于传统方法，能够为工业污染源智能监管工作提供可靠的技术支持。

五、结束语

本研究通过建立智能分析和预警模型，实现工业污染源的在线监测数据智能分析，从而为工业污染监管工作提供一种全新的技术途径。在今后的工作中，将对模型性能进行进一步优化，并根据实际应用场景进行扩展研究，以增强其泛化能力和实用性，促进工业污染源监管朝着智能化和精准化的方向迈进。

参考文献

[1]伍恒赟，陈会明，陈谊，曹炳伟，储险峰，熊长保.工业污染源铊水质在线监测方法研究[J].江西化工，2024，41(01)：41-44.

[2]吴烨超，朱丹，丁香怡，包金婷.环境监测在节能减排工作中的作用分析[J].节能，2024，44(01)：116-118.

[3]朱明龙，倪莹，张学娟.无人机遥感监测在水源地污染源监测中的应用研究[J].清洗世界，2024，41(02)：131-133.

[4]陈林，蔡锴潮，陈臻.混合所有制改革与绿色全要素生产率——基于工业污染源重点调查的经验证据[J].商业经济与管理，2024，(11)：74-90.DOI：10.14134/j.cnki.cn33-1336/f.2024.11.007.

[5]王海榕.城市大气污染治理措施研究[J].皮革制作与环保科技，2024，5(21)：122-124.DOI：10.20025/j.cnki.CN10-1679.2024-21-42.

[6]李兵.工业排气监测的常见问题及优化措施[J].皮革制作与环保科技，2022，3(04)：110-112.

[7]苟鹏，谭铃，卜兴兵.工业污染源排放总量核算分析[J].中国资源综合利用，2023，41(08)：179-181.

[8]黄洁妮.工业污染源“一企一档”管理系统建设与应用[J].黑龙江环境通报，2024，37(07)：20-22.

[9]马智斌.关于工业污染源有组织废气监测中的常见问题探讨[J].皮革制作与环保科技，2020，1(05)：30-33.

[10]程梦婷，李凌波.工业污染源二氧化硫排放监测技术进展[J].当代化工，2017，46(10)：2116-2118.DOI：10.13840/j.cnki.cn21-1457/tq.2017.10.042.

