

基于深度学习的环境工程检测数据异常识别模型构建与优化研究

刘娜 高洪珍^{*通讯作者}

河北华清环境科技集团股份有限公司, 河北 石家庄 054000

摘要: 环境污染越来越厉害的时候, 环境监测数据在环境工程里的运用变得特别重要, 因为采集设备, 传输系统, 外部环境这些因素的影响, 环境监测数据里经常会有些异常的数据。这些异常会影响后面的分析和判断, 传统的异常检测方法面对复杂的环境数据时有一定的局限性, 应对高维, 非线性, 动态变化的数据比较困难, 为了改善这种状况, 文章提出了一个依靠深度学习的环境工程检测数据异常识别模型, 借助创建融合 LSTM 和自编码器的很多模型, 可以完成对点异常和序列异常的有效检测。文章还采用超参数优化策略和多模型集成方法来优化模型性能, 通过实验发现该模型在多个真实环境数据集上都能达到很好的异常检测效果, 在准确率、召回率以及鲁棒性等方面都优于传统算法, 研究成果对环境监测数据质量控制与智能分析有着重要的推动作用。

关键词: 环境监测; 异常检测; 深度学习; LSTM; 自编码器; 模型优化

DOI: 10.63887/fns.2025.1.3.11

近几年, 深度学习技术因为具备很强的非线性建模能力和特征表达能力, 慢慢变成环境数据异常检测的关键方向, 文章依靠深度学习手段, 创建起一种融合 LSTM 和自编码器的多模型检测框架, 而且采用优化策略, 目的在于改善模型对于多维时序环境数据当中异常信息的识别水平, 从而达成更为智能且高效的环境数据质量控制。

1 相关研究综述

1.1 环境工程中数据异常检测的研究进展

环境监测数据被大量用于空气质量, 水质管理, 噪声监控以及工业排放监管等诸多领域, 数据采集过程会受诸多因素干扰, 异常数据问题普遍存在, 其检测与清洗成为环境数据管理的关键问题。近些年来, 研究者在环境数据异常检测方面不断探寻, 从早期依靠规则和统计分析的办法慢慢发展到依靠智能算法的自动识别体系。

1.2 常用的异常检测方法

传统的统计方法如 Z-score、IQR (四分位距)、Grubbs 检验等都是利用数据的概率分布特性来判断异常的, 这类方法适合于数据服从正态分布且噪声较少

的情况, 在环境数据中由于分布特性难以确定且分布非平稳等原因导致效果较差^[1]。

有些研究试图把规则系统同机器学习结合起来, 创建依靠专家知识的混合模型, 以此来改善异常识别的精确度和可解释性, 用物理边界条件设定初步筛选规则, 然后再加上聚类或者分类模型做二次过滤, 这种做法虽然可以提升检测效果, 不过它依赖规则设置和特定场景, 没法普遍应用到各种类型的环境数据当中。

1.3 深度学习在时间序列与多维数据中的应用研究

随着深度神经网络在序列建模、高维特征提取等方面取得的突破, 深度学习在环境数据异常检测中有着巨大的潜力。

LSTM (长短期记忆网络)、GRU (门控循环单元) 是处理时间序列数据的典型结构, 可以捕捉长时间依赖特征, 适合连续性强的环境数据 (气温变化、空气质量指数等), 已有研究证明 LSTM 可以识别周期性突变或者异常波动, 在多步预测中提升模型鲁棒性。

自编码器学习输入数据的压缩表示, 常用于无监督异常检测, 其基本原理是训练模型重构正常数据,

用重构误差判断样本是否为异常，环境数据里，自编码器能找出不符合正常模式的测量值，特别在噪声多或者标签少的时候效果好。

Transformer 结构凭借多头注意力机制在时间序列处理上有所突破，渐渐被应用于异常检测的研究当中，Transformer 能够抓住全局依赖关系并保持局部变化特性，适合做多维数据流的分析工作，像多传感器监测系统之类的东西，把时序编码同注意力加权策略结合起来使用，从而优化模型在复杂环境下的异常识别效果^[2]。

1.4 研究存在的不足和挑战

虽然深度学习方法在环境监测异常检测中有着较大的优势，但是依然存在一些问题：数据和标签少，大部分环境监测数据是未标注的时间序列数据，缺少异常样本，导致监督学习模型无法有效训练；模型难以泛化到新环境中或新监测对象上；实时性及计算问题，复杂的深度网络在部署到实际监测平台上时会出现运行时延、计算资源等问题；多源数据融合难，环境监测包含不同类型的、不同采样率的数据（如气象、水质、图像等），需要考虑数据的异质性。

2 方法论

2.1 数据来源与预处理

本研究所用到的数据是各类环境监测指标，包括空气质量、水质、噪声、工业排放等数据，数据来源包括国家环境监测平台、地方环保部门开放数据接口以及典型工业园区布设的物联网传感器网络，这些数据各有特点。有的具有较强的规律性的时间序列变化，有的则存在由人为、设备、自然因素造成的突发波动和结构性缺失，为了提高模型训练质量和检测精度，在数据预处理时，先处理缺失值和异常值，使用线性插值、K 近邻插值和基于规则的初筛方法对数据进行初步清洗。接着统一不同来源数据的时间分辨率，通过重采样和插值做到时间上的同步，保证数据结构的一致性，之后把所有的数据都实施归一化处理，用 Min-Max 办法把所有特征统一映射到[0,1]区间，去掉维度之间的差异影响，最后形成固定长度的滑动窗口输入序列，给后面的深度学习模型给予标准且结构化的训

练样本集合，从而让模型可以有效地抓住环境数据里的动态改变和潜在的异常信号^[3]。

2.2 模型构建

面对环境监测数据具备的时间依赖性、非线性特性以及多变量结构复杂性，文章设计了一种多层融合的深度学习模型，该模型融合了 LSTM、Autoencoder 以及注意力机制，目的在于优化异常检测的准确度和鲁棒性，该模型先通过 LSTM 来提取输入时间序列的长期依赖关系及动态特征，给后面的重构模块给予时序方面的信息支撑，自编码器结构会去学习并压缩正常样本的潜在空间表示，再通过解码器执行重构操作，把原始输入数据重新构建出来，用重构后的数据与原数据之间的误差当作异常评分准则，进而找出那些与训练分布差别很大的异常样本。同时为了更好地感知关键时间片段，引入注意力机制对序列中不同时间步的特征给予不同权重，使模型关注对判断最敏感的变化区间。此外为了提高模型的泛化能力和灵活性，采用模型集成方法，将多个结构不同的子模型输出加权融合，采用得分集成的方式降低单模型对特定噪声或异常类型的敏感性，从整体上提高检测的鲁棒性^[4]。

2.3 模型优化策略

为了让创建的深度学习模型在异常检测任务当中做到最佳效果，文章在训练期间采用了诸多改良办法，涉及超参数调节、改进模型训练技能以及扩充模型通用能力的方式，从超参数调节看，利用贝叶斯改善法对隐含层数目、学习速率、批次容量、滑动窗长这些核心参数实施自动化搜索，这比传统的网格搜索方法更快速准确。在改良训练策略中，设置了提早停止的办法，防止在训练集中发生过拟合现象，并采用 Dropout 技术或者 L2 正则化手段来约束网络结构，从而改善它在检测集中泛化能力的表现。同时加入了批归一化和残差连接，使得网络在深的网络中训练的稳定性变好以及更快的收敛，鲁棒性上，文章使用时间序列扰动、局部遮挡、加噪声这些扰动数据，去模拟不同的异常类型以及数据问题。增强模型对真实环境中复杂环境变化的适应能力，并且考虑多源数据有差异时，可以利用特征的标准化和多个模型同时训练的方

法，让模型在不同来源的数据上也能表现稳定。这有利于模型在更多场景下的应用

3 实验设计与结果分析

3.1 实验设置

在拥有优良的硬件性能的平台实施实验任务，以此来保证深度学习模型的训练效率以及实验过程可被多次复制，在硬件方面配置的是 Intel i7-12700H CPU 并支持多核心的并行运算，内存配置到 16G，从而可以实现大批量的数据加载和计算工作，NVIDIA RTX 3060 GPU 会大大加快神经网络前传和反传的速度，就软件平台而言，利用 Python 3.9 这个主要开发语言。结合 TensorFlow 2.12 和 PyTorch 2.0 搭建各自不同的模型结构包含 LSTM 预测结构，自编码器结构以及融合模型等等。Scikit-learn 用来实现传统机器学习的各类模

型，IsolationForest 以及 One-ClassSVM 也在内，并且这样的方式保证对比过程的完整并且有条不紊。Optuna 属于自动化超参数优化工具，它依靠贝叶斯优化策略来完成对模型超参空间的有效搜索，数据处理阶段采用 Pandas 和 NumPy 来操作，包含缺失值处理，滑动窗口创建，归一化以及时间戳同步等动作，结果展示部分利用 Matplotlib 和 Seaborn 来绘制训练曲线，性能指标图以及误差分布图，整个实验环境的配置比较稳定，工具链也较为完备，为后续的模型训练和性能评估奠定了稳固的基础。

3.2 实验结果

实验通过很多组模型在统一数据集的比较中，系统评判各方法在环境工程监测数据异常检测方面的优缺点，下表显示了主要模型在五个重要指标上——精确率，召回率，F1 分数，AUC，误报率的比较成果：

模型	Precision	Recall	F1-score	AUC	误报率
Z-score 滑动窗口	0.62	0.58	0.6	0.63	0.21
Isolation Forest	0.71	0.64	0.67	0.72	0.15
One-Class SVM	0.68	0.66	0.67	0.7	0.18
LSTM 预测模型	0.76	0.71	0.73	0.78	0.12
Autoencoder	0.81	0.77	0.79	0.83	0.1
VAE	0.84	0.79	0.81	0.85	0.09
融合模型（文章）	0.88	0.83	0.85	0.89	0.07

由表可知，文章所采用的融合模型在各个指标上都是最领先的，达到了 0.85 的 F1 分数，远远大于传统的 Z-score 方法 (0.6)、IsolationForest (0.67)，并且其控制误报率最低，只有 0.07。这种模型可以较好的平衡精确率和召回率，适合作为要求低误报高灵敏的环境监测的部署。利用 LSTM 对于时间序列的建模、自编码器对于重构误差的灵敏和注意力对于关键时间片段的敏感，这个融合模型很好的体现了结构的合理性，在真实的复杂环境中也有应用的价值。

3.3 结果讨论

文章提出的深度融合模型在多个实验数据集上具备不错的检测能力，其结构设计以及性能表现均显示

出方法可行且先进，通过与已有标注的异常样本相比较，模型可以识别绝大多数实际异常数据，这表明检测结果具有较好的可信度和一致性，尤其对于非周期性或者趋势变化不明显的异常事件而言。它具备较强的辨别能力，而且该模型还表现出良好的适应性和迁移能力，在空气，水质以及排放等不同种类的环境数据当中，都能维持稳定性能，只需微调窗口大小，层数等参数就能迅速投入新数据场景，具备很强的跨任务泛化能力。从实际应用角度来讲，此模型可以整合到智慧城市环境监控体系，工业排放在线监督平台或者地区环境警报平台等当中去，从而做到对异常事情的提早察觉并处理，不过也要留意，当下这个模型在

实际部署过程中还存在一些难题，比如高实时性需求所引发的计算费用，多种资料的同步融合困难。异常标识缺乏对监督表现的约束等等，后面的研究可着重探究模型轻量化，半监督学习机制以及多模态资料融合手段，进而改进其实用程度和推广效能^[5]。

4 研究总结

文章以环境监测数据中的异常识别为研究对象，创建起一种依托深度学习、涵盖多种模型组合检测的大框架，详细探讨了 LSTM，自动编码器以及注意机制在环境数据异常检测中是否合适及相互补充之处，而

且借助贝叶斯优化策略与集成手段改善了模型在多源，非线性且存在高维时间序列数据上的检测精度和鲁棒性，在很多真实的数据集里，此框架下的模型比传统的办法和单一模型架构有着明显更好的效果，有着更高 F1 值和更少的错误警报。研究还探讨了数据量、噪声干扰以及特征维度变化对模型性能的影响，证实了模型具备较好的适应性和扩展性，该研究给环境工程领域数据质量控制供应了一条高效又可推行的技术途径，对于改善环境监测系统的智能化水平有着一定的理论和实际意义，而且为以后的研究给予了可行的模型根基和改良方向。

参考文献

- [1]陈鸣. 网络安全态势感知与异常检测技术研究 [J]. 计算机科学, 2022, 49 (10): 150 - 156.
- [2]王强. 工业物联网设备健康监测与故障诊断研究 [J]. 仪器仪表学报, 2021, 42 (8): 1 - 10.
- [3]李华. 金融交易欺诈检测方法综述 [J]. 金融理论与实践, 2020, (12): 89 - 95.
- [4]周茂袁, 伍小双. 基于深度学习的异常检测模型综述[J]. 中国民航大学学报, 2023, 41 (4): 1-736
- [5]贾峰. 基于深度学习算法的电力调度数据网络异常检测方法[J]. 信息与电脑, 2023, 35(12): 79-81