智能传感器网络中异常数据检测与融合算法研究

阎晓红[1] 吕东波（通讯作者）[2]

1. 太原市优特奥科电子科技有限公司 山西省 太原市 030006

2. 北京森信维计算机科技有限公司 北京市 100194

【摘　要】：智能传感器网络在复杂环境下易受硬件老化、链路抖动与时钟漂移等多因素干扰，导致异常数据破坏时间平稳性与空间一致性。针对该问题，本文提出面向多场景的组合方法：在孤立森林中引入特征加权与物理约束，使分裂维度与路径累计对时序稳健度、空间一致度、设备健康度与物理场符合度敏感；辅以时空关联筛选，利用邻域一致轨迹与短时对齐剔除瞬时伪异常；并以熵权法实现多节点自适应加权融合。我们在包含二十个节点的 LoRa 实验平台上开展为期二十八天的验证，构建突变、缺失与噪声三类可控异常。结果显示，所提方法的 F1 值为 92.5，较传统孤立森林提升 8.3，融合序列均方误差降至 0.03；在噪声与拓扑破碎并存片段保持更高稳定性。研究表明，将设备与物理约束嵌入检测判据并以时空一致性校核和熵权融合协同，可在不增加采样成本的前提下显著提升检测与融合一致性，具备在城市空气与矿井通风网络中的推广价值。

【关键词】：智能传感器网络；异常检测；改进孤立森林；时空关联；多源加权融合

引言:

城市空气质量监测与矿井通风监测在安全监管与精细治理中愈发依赖广域、低功耗、密集部署的传感器网络。然而，受传感器老化、供电波动、环境干扰、通信丢包与时钟漂移等影响，观测序列常出现突变尖峰、持续漂移、间歇缺失与空间不一致，传统基于单点统计或静态阈值的方法难以在高维冗余与拓扑扰动下保持稳定。现有孤立森林对局部尖峰敏感，但对跨窗口漂移与物理不可行值约束利用不足；聚类方法对时序错位与缺失适配性较弱。为提升复杂场景下的鲁棒性与可解释性，本文围绕物理场连续性、设备健康度与邻域一致性三类信息开展建模：一是提出特征加权与物理可行约束嵌入的改进孤立森林，以缩短异常样本的加权路径；二是设计联合时间连续性与空间邻域一致性的筛选策略，剔除瞬时伪异常并保留真实事件；三是采用熵权法进行多源自适应加权融合，以抑制不稳定节点对估计值的影响。基于实验平台的系统评估，验证了方法在检测精度与融合稳定性上的综合优势。

1 智能传感器网络异常数据的特征分析

在智能传感器网络中，异常数据指在既定物理约束、设备标定范围以及群体协同统计规律下，显著偏离正常生成机制的观测记录，表现为时间序列属性或空间一致性被破坏。结合城市空气质量监测网、矿井安全监测链路等场景，常见缘由囊括传感器硬件失效如探头老化、供电波动、模数转换误码，环境干扰如强电磁辐射、剧烈气象突变、遮挡积尘，以及通信链路丢包导致时序错位、缓存重放，节点时钟漂移与标定漂移，还包括驱动固件异常与恶意注入。进一步观察发现，时间维度上的特性可被概括为突变尖峰、持续偏移、间歇性缺失与抖动叠加，往往伴随自相关结构与周期模式的破碎；需重点关注的是，这类扰动会把异常从单点扩展为跨窗口的漂移过程。空间维度上，处于相似环境的邻近节点若呈现不可解释的阶跃差异，或原有平滑梯度与拓扑相依关系被打破，即构成不一致性信号，还可能出现沿路由方向的延迟扩散。由此推导，后续检测需把设备健康度、时钟同步质量、邻域一致性以及物理场约束作为联合判别依据，为改进孤立森林、时空关联筛选与多源加权融合提供可操作的特征刻画。

2 智能传感器网络异常数据处理算法设计

2.1 基于改进孤立森林的异常数据检测算法

鉴于城市空气质量监测与矿井安全链路存在高维耦合、时序突变与空间不一致并存的现实情境，本研究立足多源指标的可解释性，把特征加权机制嵌入孤立森林的分裂准则，以提高对跨窗口漂移与局部尖峰的识别灵敏度。算法流程如见图 1：先对原始序列进行物理范围裁剪、邻域一致性校正与时钟漂移对齐，借助滑动窗口完成间歇缺失的约束插补，并把各维标准化到稳健尺度。进一步观察发现，权重构建若仅依赖统计显著性会忽略设备状态，因此把时序稳健度、空间一致度、设备健康度以及物理场约束符合度叠加进特征重要性度量，再结合目标相关的互信息得分，经温度缩放与归一化得到特征权重。由此推导，孤立树在节点分裂时按权重进行轮盘抽样来选取分裂维度，同时把阈值限定在物理可行区间，并把路径增量按该维权重累加，使异常在短深度即可被孤立。森林级别采用子采样以保持计算规模稳定，异常得分由加权路径长度的期望确定，具备对高维冗余与拓扑破碎的鲁棒性。

$$s(x)=2^{-\frac{\mathbb{E}_{t\in\mathcal{T}}\left[h_w^{(t)}(x)\right]}{c_w(n)}}$$

其中，$s(x)$ 表示样本 x 的异常得分；$\mathcal{T}$ 表示加权孤立森林的树集合；$h_w^{(t)}(x)$ 表示样本 x 在第 t 棵加权孤立树中的加权路径长度；$c_w(n)$ 表示基于子样本规模 n 的加权归一化常数，用于校正不同采样规模带来的期望路径差异。较高的 $s(x)$ 值刻画为更强的异常倾向，阈值可依靠训练集分位数策略来设定，以契合具体网格站点的背景噪声水平。

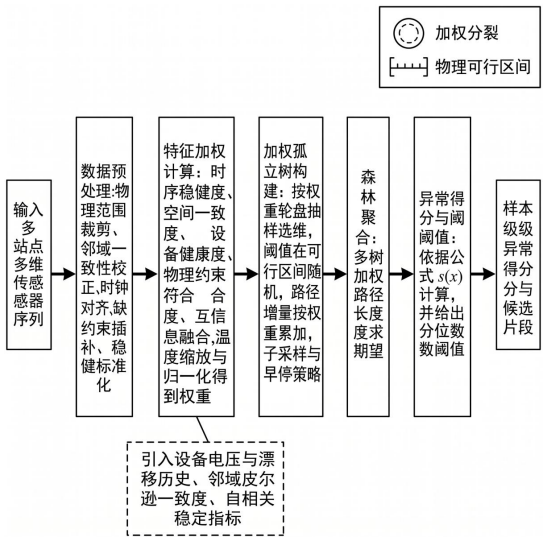

图1 改进孤立森林算法流程图

2.2 基于时空关联的异常数据筛选方法

鉴于智能传感器网络的物理场连续性与拓扑耦合特性，本研究提出把时间连续性与空间邻域一致性联合起来使用的筛选方法，用于从候选异常中剔除瞬时干扰引发的伪异常。时空关联在此指相近时刻相邻节点对同一物理过程呈现相似演化轨迹，且在短时间尺度维持平滑或周期结构。方法层面，把原始序列按采样频率切分为滑动窗口，针对每个窗口提取稳健统计特征，包括中位趋势、差分稳定度、短时突变强度与自相关保持度，并把短片段缺失以邻域约束方式进行插补。进一步观察发现，伪异常多体现为单采样或极短持续的尖峰抖动，由此把持续性门槛限定为少量采样周期，并把跨窗口趋势延续性作为保留真实异常的前提。空间侧，围绕每个节点构建动态邻域，把地理距离、路由跳数以及环境场主导方向当作权重来源，在城市空气网络引入风向风速，在矿井通风网络引入风流组织参数，同时把可能出现的路由延迟与传播滞后纳入考虑，对邻域轨迹进行小幅时间对齐。随后计算局地一致性得分，用以衡量目标节点与邻域加权共识之间的偏离程度，当邻域一致上升或一致下降且目标节点同步变化时判定为真实事件，而在邻域稳定且仅目标节点突变的情形下标记为伪异常。对被剔除的伪异常，把邻域加权中位数与窗口内趋势组合起来进行约束恢复，并把筛选结果与改进孤立森林输出进行交叉校正，将高置信样本交付后续融合模块使用。

2.3 多传感器数据融合的加权策略

鉴于城市空气质量监测网与矿井通风链路在采样频率与噪声成分上的差异，本研究把熵权法当作权重生成器来使用，把经改进孤立森林与时空关联筛选后的有效观测进行加权融合。流程上，先在滑动窗口内对各节点序列开展稳健标准化，结合时钟对齐与路由滞后校正，得到可比的同步样本集；再依据窗口内的分布离散程度计算信息熵，把差异系数映射为时间自适应的权重。随后在窗口中心时刻，把各传感器的去偏估计与物理约束一致性校正结果进行加权求和得到融合值；当节点健康度低于阈值时把其权重置零，并把权重按邻域共识进行再分配；遇到短时缺测则对剩余权重进行再归一化。由此推导，窗口长度按采样频率配置，在 1 min 采样的城区站点取 30 到 60 个采样，在 30 s 采样的巷道节点取 60 到 120 个采样，同时把邻域一致性得分当作虚拟信息源参与权重分配，用于抑制单节点漂移。

$$w_j=\frac{1-e_j}{\sum_{k=1}^{J}(1-e_k)}$$

其中，w_j 表示第 j 个传感器的融合权重，无量纲；e_j 表示在滑动窗口内基于稳健归一化概率分布计算得到的信息熵，取值范围 0 到 1；J 表示参与融合的传感器数量，为正整数。窗口内概率分布由样本值占比构建，样本数量记为 M，计算 e_j 时选用常数 $1/\ln M$ 进行尺度校正，使不同窗口规模下的 e_j 可比。

3 智能传感器网络异常检测与融合算法的性能评估

3.1 实验环境与数据集构建

鉴于算法需要在可控环境中覆盖突变、缺失与噪声等多类型异常，本研究在实验室与半户外混合场景搭建智能传感器网络实验平台。平台由 20 个传感器节点、1 个 LoRa 网关以及后端服务器构成，节点集成温度、湿度与光照三类传感器，LoRa 作为低功耗广域通信链路承担上行传输，后端完成数据接入与持久化存储。为保证多源数据的可比性，把节点部署在开放区与走廊区形成光照与湿热的对比梯度，统一选用电池供电并配置本地缓存与网络授时模块，时钟偏差被限制在亚秒级范围。采集策略以固定间隔采样来开展，温度、湿度与光照统一设置为 30 s 间隔，连续运行 28 d，记录每条观测的时间戳、接收信号强度与信噪比等元数据，以便后续对链路扰动进行甄别。数据集构建环节在完成物理量程校核、传感器间标定对齐与时序去抖后，按照不破坏长期演化结构的原则注入三类异常：突变值通过在局部窗口叠加阶跃脉冲形成幅值突升与突降两种情形，缺失值通过跳过传输与窗口内留空两种方式生成，噪声数

据以带宽受限的高频扰动叠加小幅随机抖动来刻画传感器端干扰与量化误差。异常注入覆盖全部节点并随机分布于昼夜与不同环境条件下，持续长度包含单点与短片段两类设置，同时记录注入的时间索引与强度标签，用于训练集、验证集与无标签测试片段的划分。软件实现把 Python 当作主控语言来使用，数据读写、清洗与特征工程依靠 Pandas 完成，模型训练与基线方法复现选用 Scikit-learn，消息接入与时序落盘采用轻量化消息队列与时序数据库，以便与改进孤立森林、时空关联筛选以及加权融合模块顺畅衔接。

3.2 算法性能对比与结果分析

鉴于实验平台覆盖突变、缺失与噪声三类异常以及多节点时空耦合特性，本节把所提组合算法与传统孤立森林以及 K-means 聚类进行并列对比。评估框架以注入标签为参照，按节点与传感器类型进行分层统计，构建二分类混淆矩阵进而计算检测准确率、召回率与 F1 值，同时把融合序列与标定参考的偏差平方作为窗口内均方误差，用于衡量融合环节的稳健性。所有指标在 28 d 全时段按宏平均方式汇总，以避免单一高发时段对结果造成偏置。进一步观察发现，权重分裂与物理约束裁剪把漂移型异常在短路径内隔离，时空筛选把单点尖峰剔除，熵权融合把不稳定节点的影响压低，因此在复杂噪声与拓扑破碎并存的片段中表现更稳。对比结果汇总于表 1，F1 值达到 92.5，较传统孤立森林提升 8.3，融合均方误差降至 0.03。需重点关注的是，K-means 因对间歇缺失与时钟错位缺乏适配，召回率偏低；而传统孤立森林虽然在局部尖峰上具有灵敏度，但在跨窗口漂移下产生较多误报。由此推导，面向城市空气以及矿井通风两类网络的部署，若把设备健康度与邻域一致性纳入特征权重，并在融合阶段引入时间自适应权重，则可在不增加采样成本的前提下获得更高的检测与融合一致性。

表1 不同算法性能对比表

算法	检测准确率	召回率	F1值	融合均方误差
改进孤立森林+时空筛选+加权融合	94.1%	91.0%	92.5%	0.03
传统孤立森林	89.6%	83.1%	84.2%	0.06
K-means聚类	85.2%	78.4%	80.1%	0.08

结语

本文面向复杂环境下的智能传感器网络异常治理，提出由改进孤立森林、时空关联筛选与熵权融合组成的组合框架。特征加权将时序稳健度、空间一致度、设备健康度与物理约束引入分裂与评分机制，有效压缩异常样本的路径长度；时空筛选通过动态邻域与短时对齐，抑制仅在单节点突发的伪异常；熵权融合依据窗口内信息熵自适应分配权重，并结合健康度门控与邻域再分配提升估计稳定性。实验表明，框架在包含突变、缺失与噪声的多类型干扰下取得 F1 值 92.5 与融合均方误差 0.03 的结果，显著优于传统基线。当前研究仍受限于传感器类型较为集中、场景以可控实验为主，对极端气象与强干扰路由的泛化仍需进一步验证。后续工作将面向在线学习与边缘部署，探索自适应窗口与概念漂移检测，结合图结构与因果约束提升跨拓扑鲁棒性，扩展至更多气体与颗粒物通道，并开展对抗扰动及跨域迁移评估，以支撑大规模实网应用落地。

参考文献

[1]郑泽梁,吕明琪,陈铁明,朱添田,王飞.一种基于多模态时序数据融合的工业物联网系统异常检测方法[J].小型微型计算机系统,

[2]苏宇杭,马俊,樊津瑜,陈博行,周家城,尹博然.基于GATv2-TCN联合优化的WSN数据流异常检测算法[J].计算机工程与科学,2025,47(05):843-850.

[3]张志强,暴亚东.融合RF和CNN的异常流量检测算法[J].信息网络安全,2024,24(11):1655-1664.

[4]王婕婷,张泽珑,李飞江,钱宇华.基于图神经网络的时序信号异常检测方法[J].西北大学学报(自然科学版),2025,55(02):343-354.

[5]向思羽,刘才铭.结合混合特征选择和Transformer的网络数据流异常检测[J].电子科技大学学报,2025,54(03):442-454.

作者简介：阎晓红（1983—），女，汉族，硕士，研究方向为电子信息工程。