面向智慧城市的边缘计算数据清洗机制研究

张　东　陈景奕
北京泰尔凯达电信信息咨询有限责任公司　北京海淀　100080

摘　要：智慧城市建设的纵深推进催生了海量边缘数据的实时处理需求，数据质量直接决定上层决策的可靠性。本文系统梳理了边缘计算环境下数据清洗的研究进展，剖析智慧城市边缘数据的异构性、实时性与质量不确定性特征，从轻量级架构设计、缺失值处理、异常检测、隐私保护几个维度评述关键技术，并结合交通、环境、安防、能源等典型场景分析应用实践，最后探讨资源约束与清洗精度的平衡、异构适配等核心挑战，展望自适应清洗、边缘智能芯片赋能等发展趋势，为相关研究提供参考。

关键词：智慧城市；边缘计算；数据清洗；数据质量；轻量级算法；隐私保护计算

引言

智慧城市通过泛在感知网络实现对城市运行状态的全面数字化映射，边缘计算作为近数据源处理范式，有效缓解了云端集中式架构的带宽压力与响应延迟问题。然而，边缘节点部署环境的开放性与硬件资源的受限性，使得原始数据普遍存在缺失、噪声、冗余及语义冲突等质量问题。若将低质量数据直接注入分析流程，不仅会导致模型性能劣化，更可能在交通调度、应急指挥等关键场景中引发决策失误。因此，研究面向智慧城市的边缘计算数据清洗机制，兼具理论价值与现实紧迫性。

现有数据清洗研究多聚焦于云端中心化场景，假设计算资源充足、数据批次完整，难以直接迁移至边缘环境。边缘侧清洗需在毫秒级时限内完成，且须兼顾CPU占用率、能耗预算及隐私合规等硬约束，这对传统方法的计算复杂度与内存占用（memory footprint）提出了严峻挑战。本文试图弥合这一鸿沟，通过系统性文献回顾与技术解构，厘清边缘数据清洗的特殊需求与可行路径。

一、智慧城市边缘计算架构与数据特征

1. 分层架构与边缘定位

典型智慧城市边缘架构呈现“端–边–云”三级拓扑。终端层涵盖路侧单元、环境传感器、移动终端等异构设备，承担原始数据采集；边缘层由微数据中心、网关及边缘服务器构成，部署于基站、园区或道路沿线，提供首级数据处理能力；云层则汇聚跨域数据，支撑全局分析与长期存储。数据清洗的主体工作正下沉至边缘层，以削减无效数据传输、降低云端负载。边缘节点的异构性尤为突出：工业级网关配备国产化处理器与有限内存，而路侧计算单元可能集成轻量级GPU。这种硬件碎片化要求清洗机制具备弹性伸缩能力，能够依据实时资源状态动态调整算法强度。

2. 边缘数据质量困境

多源异构性：同一交叉路口可能同时输出视频流、雷达点云、地磁信号，数据格式、采样频率、时空基准各不相同，融合前的标准化清洗不可或缺。高时效压力：自动驾驶辅助决策要求端到端延迟低于50毫秒，数据清洗必须嵌入流式处理管道，而非事后批处理。质量波动剧烈：车载传感器在隧道内失锁、气象设备在极端天气下漂移等场景频发，清洗机制需具备环境自适应能力。隐私敏感：摄像头捕获的人脸、车牌信息需在边缘完成脱敏，原始数据严禁外传。

二、边缘数据清洗关键技术

1. 资源感知的轻量级架构

传统清洗框架如OpenRefine、Trifacta面向桌面级计算环境设计，内存占用常以GB计，无法直接部署于边缘。近期研究转向流式处理框架的清洗能力扩展，Apache Flink的CEP模块已支持复杂事件模式匹配，但其RocksDB状态后端启动时间超过30秒，内存占用逾

作者简介：
张东（1989.01--），男，汉族，广东汕头，本科，工程师，研究方向：信息通信。
陈景奕（1990.04--），男，汉族，四川广安，硕士，工程师，研究方向：信息通信。

500MB，难以满足边缘秒级启动需求。

更务实的方案采用分层清洗策略：边缘节点执行粗粒度过滤（如范围校验、格式检查），仅将疑似异常数据上传至区域边缘进行二次精洗，形成“边缘粗筛–近源精修”的协作链条。资源调度层面，有学者提出清洗任务的动态优先级算法，依据当前电池电量与计算队列长度，在漏检率与能耗间寻求帕累托最优。此类方法将数据清洗从静态工具转化为自适应服务，更贴合野外部署场景。

2.缺失值处理的边缘适配

边缘数据缺失成因包括通信丢包、传感器休眠及存储溢出。简单删除策略在流式场景下易引发时间序列断裂，而均值填充对非平稳过程偏差显著。当前主流方案包括：基于轻量级时间序列模型的预测填充，如将Prophet算法裁减至单变量版本，在树莓派级设备上实现分钟级趋势预测；以及利用空间相关性进行邻域插补，通过构建传感器网络的图结构，以图卷积网络推断缺失节点读数，模型参数量可压缩至数百KB量级。

值得注意的是，缺失本身携带信息——通信中断可能预示设备故障，而非单纯的数据漏洞。部分研究开始区分“可修复缺失”与“结构性缺失”，对后者直接触发告警而非盲目填充，避免将系统性误差引入下游分析。

3.异常检测的算法轻量化

统计方法因其可解释性与低计算开销，在边缘异常检测中仍占主流。改进的Z–score方法引入滑动窗口与指数衰减权重，适配概念漂移场景；基于谱残差分析的视觉异常检测，无需训练即可在嵌入式视觉芯片上实时运行。机器学习方法中，孤立森林通过子采样构建多棵随机树，时间复杂度接近线性，可在EdgeX Foundry的数据分析服务层集成实现。

4.隐私保护计算

边缘数据的隐私敏感性要求清洗过程嵌入隐私保护机制，主要技术路径包括：本地差分隐私、安全多方计算（MPC）的轻量实现、可信执行环境（TEE），本地差分隐私在终端注入噪声保护个体数据，适用于环境监测等数值型数据；安全多方计算利用秘密共享等技术实现跨节点协同清洗，结合硬件优化降低开销；可信执行环境通过硬件隔离创建安全飞地执行敏感清洗。三者结合可构建可信联邦清洗架构。

三、智慧城市典型场景应用实践

1.智慧交通：多源异构数据的实时融合

智慧交通系统是边缘计算数据清洗技术集成度最高的场景之一，完整覆盖了第3点所述的几个技术维度。

轻量级架构设计，路口边缘节点采用“端侧预筛–边缘精洗”的分层架构：路侧单元（RSU）内置国产化处理器执行格式校验与范围过滤（如车速0–120km/h阈值），剔除明显异常数据；边缘服务器部署Intel Xeon–D或NVIDIA Jetson平台，承载复杂清洗任务。资源调度模块实时监测CPU占用率与网络队列长度，当检测到信号控制优先级任务时，动态降低视频增强算法的迭代次数，确保端到端延迟<100ms。

缺失值处理，浮动车GPS在隧道、高架桥下频繁失锁，导致轨迹断点。边缘节点采用轻量级LSTM模型（参数量<500KB）预测缺失位置，结合路网拓扑约束进行地图匹配修正。对于通信中断导致的系统性缺失（如5G切换过程中的数据间隙），标记为“结构性缺失”并触发告警，避免盲目填充引入误差。

异常检测，雷达与视频的多源冲突检测是核心难点。边缘节点维护双通道校验：雷达测速与视频测速的差值超过15%时，启动第三源地磁线圈数据仲裁；若三源冲突，基于历史置信度加权决策（视频夜间置信度降级，雷达雨雾天气置信度降级）。统计方法与轻量规则引擎结合，在树莓派级设备上实现毫秒级响应。

隐私保护，车牌识别数据在边缘完成脱敏处理：检测框定位后，车牌区域采用高斯模糊（核大小自适应车牌尺寸），模糊强度满足人眼不可辨识但机器可识别（用于执法取证）的双重要求。人脸数据严格执行删除策略，特征提取后立即销毁原始图像，仅上传匿名化特征向量。

2.智慧环保：资源受限传感器网络的协同质控

环境监测场景凸显边缘节点资源受限性与数据质量波动性的矛盾，重点应用轻量级架构、缺失值处理、异常检测与一致性保障技术，冗余消除与隐私保护需求相对较低。

轻量级架构的极端适配，野外传感器节点采用TinyML架构：MCU级设备（如STM32L4，主频80MHz，内存256KB）运行裁剪后的决策树模型，仅执行本地阈值过滤与异常标记；复杂清洗任务卸载至太阳能供电的边缘网关（Raspberry Pi 4级），形成“极端边缘–弱边缘”两级架构。能耗管理模块在电池电量<20%时，自动切换至最低频采样模式（从1分钟间隔延长至15分钟），牺牲数据密度换取持续运行。

缺失值处理的时空协同，监测网络覆盖盲区（建筑物遮挡通信）采用图卷积网络（GCN）插补：构建传

感器空间邻接图，节点特征为历史监测序列，边权重为地理距离倒数。GCN模型经知识蒸馏压缩至800KB，在边缘网关实现秒级推断。区别于简单空间平均，该方法考虑风向、地形等物理约束，插补精度较反距离加权（IDW）提升23%。

异常检测的物理–数据双驱动，传感器漂移与真实污染事件的区分是核心挑战。边缘节点并行运行物理扩散模型（高斯烟羽模型）与统计异常检测（改进Z–score）：当实测值与物理预测值偏离>30%且符合下风向扩散规律，判定为真实污染事件并触发告警；若偏离但不符合物理规律，或单点异常而邻域正常，判定为传感器漂移，启动软校准程序。某工业区VOCs监测中，该方案将误报率从35%降至12%。

3.智慧安防：高实时性视频流的结构化与脱敏

针对延迟敏感（<50ms）与隐私合规场景，边缘节点多层次实现异常检测：视频质量诊断实时检测画面冻结、花屏等故障并修复或告警；YOLOv5s轻量模型实现30fps目标检测。隐私保护方面，编码前完成人脸车牌脱敏（高斯模糊/马赛克），处理延迟<20ms，同时以丢帧补偿处理缺失值。

4.智慧能源：高频负荷数据的隐私感知治理

智能电网边缘节点处理高频（1kHz–10kHz）用电波形，保障NILM精度与隐私。轻量级架构：电表端阈值过滤，集中器执行复杂清洗。异常检测采用VMD分解重构基荷，HHT定位电能质量事件，延迟<100ms。隐私保护：本地差分隐私加噪（ε=1.0）后清洗，同态加密（CKKS）密态聚合，开销增3–5倍。缺失值通过用户聚类协同插补同类曲线，精度提升40%。实现异常检测、一致性保障与隐私计算。

结论

边缘计算数据清洗是智慧城市数据价值链的关键环节，其技术演进需同时回应质量保障、实时响应、资源受限与隐私合规等多重约束。本文通过系统性技术梳理表明，轻量级算法设计、分层协同架构与硬件–算法协同优化构成当前研究主线，而自适应能力、隐私保护标准化与跨场景评测体系的缺失仍是制约产业落地的瓶颈。随着边缘智能芯片与隐私计算技术的成熟，数据清洗将从“成本中心”转化为“价值支点”，为城市治理的精细化与实时化奠定数据基础。

参考文献

[1]施巍松，刘芳，孙辉，等.边缘计算：万物互联时代新型计算模型[J].计算机研究与发展，2017，54（5）：907–924.

[2]刘晨，王珊，周烜，等.数据清洗研究综述[J].软件学报，2019，30（3）：551–569.

[3]张迪，李建中，高宏.物联网数据质量管理：问题与挑战[J].计算机学报，2021，44（6）：1061–1086.

[4]陈海明，崔莉，王继良.面向物联网的边缘智能：架构、关键技术与挑战[J].计算机研究与发展，2020，57（9）：1824–1844.

[5]杨天，李凤华，李晖.差分隐私保护技术综述[J].通信学报，2020，41（8）：158–177.

[6]吴黎兵，何德彪，陈晶，等.智慧城市数据治理：架构、技术与实践[J].电子学报，2022，50（4）：712–725.

[7]刘云浩，张兰，孙利民.边缘计算环境下的轻量级机器学习研究进展[J].自动化学报，2023，49（2）：213–230.

[8]Xu, M., Zhu, T., Liu, Y., et al. Real–time data quality assessment and cleaning for intelligent transportation systems[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(5): 4567–4579.

[9]Wang, H., Zhang, L., Chen, Y., et al. Edge–based video analytics: Balancing bandwidth efficiency and privacy protection in smart surveillance[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 13(2): 456–468.