

融合边缘计算的多源异构数据智能融合方法研究

王伟¹
WANG Wei

摘要

针对智能配电网等对实时性和异构性要求高的场景的数据处理需求,文章提出一种融合边缘计算的多源异构数据智能融合方法,把三层动态协同架构搭建完成,搭配轻量级容器化联邦学习、边缘流式处理引擎以及跨域认证协议,实现多模态数据对齐、增量式知识融合与设备动态选择优化。实验结果显示该方法能显著降低融合时延和带宽消耗,提升异常检出率与处理效率,具备良好的推广应用前景。

关键词

边缘计算;多源异构数据;智能融合;深度学习

doi: 10.3969/j.issn.1672-9528.2026.04.026

0 引言

配电网的数据量和复杂程度正在不断提升,传统云-边缘计算模式已经无法匹配数据处理与传输的实际需求,这对智能配电网的普及形成了限制^[1],本次研究把融合边缘计算的多源异构数据智能融合方法作为核心探索方向,通过搭建三层动态协同架构,搭配轻量级容器化联邦学习、边缘侧流式处理引擎以及面向安全传输的跨域认证协议,完成了数据高效预处理、动态资源调度与全局知识整合的目标,重点攻克了多模态数据对齐、联邦学习优化及增量式知识融合等关键技术难题,为边缘计算环境下的智能数据处理工作提供了新的参考方向。

1 边缘智能融合架构设计

1.1 三层动态协同架构

1.1.1 边缘节点层

系统选用 NVIDIA Jetson TX2 平台,其配置包括具备 256 CUDA 核心的 NVIDIA Pascal GPU 和 8 GB LPDDR4 内存,能够充分支撑深度学习的部署需求,边缘节点层承担数据的初步预处理与过滤职责,其核心目标在于提升原始数据质量并减轻传输负载。在该平台上,TensorFlow Lite 框架运行卷积神经网络,把数据清洗、去噪与特征提取操作逐一落实,有效保障流入后续环节的数据质量^[2],TensorFlow Lite 通过应用量化技术优化模型,在显著减少内存占用与计算负担的同时,确保 CPU 负载低于 15%。

1.1.2 近边缘层

Kubernetes 的容器编排机制能够保障资源的实时扩展与高效调度,确保推理过程保持高效运行状态,近边缘层依托

Kubernetes 构建弹性计算资源池,把 GPU 资源进行动态分配以此有效支撑各类推理任务。该层充分利用 GPU 加速能力优化计算资源利用效率,其能够依据任务的实际负载需求,动态地分配或回收 GPU 资源,从而大幅提升大规模数据处理速度。

1.1.3 云中心层

图 1 详细呈现了三层动态协同架构的设计,云中心层可把构建全局知识图谱、维护电力设备故障知识库作为核心功能,该层能完成包含 12 万实体的知识库数据整合工作并开展故障预测相关作业。云中心采用分布式计算架构,给深度学习的模型训练与推理过程提供支撑,还融合卷积神经网络(convolutional neural networks, CNN)与循环神经网络(recurrent neural network, RNN)模型共同分析设备故障模式,同步实现设备状态评估与智能决策目标,其具备实时更新知识库的能力,可保障推理能力的时效性。

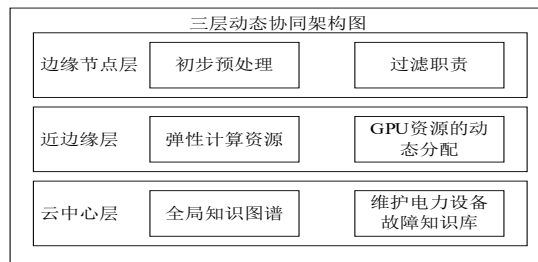


图 1 三层动态协同架构图

1.2 核心使能技术

1.2.1 轻量级容器化联邦学习

采用 Docker 实现联邦学习的容器化部署。联邦学习任务通过 TensorFlow 联邦学习框架(tensorflow federated, TFF)框架在各边缘设备上统一管理。Docker 镜像体积严格控制在 80 MB 以内,以适应边缘资源约束^[3]。边缘设备利用本地数

1. 山东省诸城市民兵装备仓库 山东潍坊 262200

据处理能力训练模型，产生的模型权重更新经由加密通道传输至中央服务器。中央服务器执行更新聚合操作，形成全局模型并将其回传至各设备节点。梯度聚合算法支撑整个协作过程，具体的聚合公式为：

$$\theta_{\text{global}}^{t+1} = \frac{1}{N} \sum_{i=1}^N \theta_i^{t+1} \quad (1)$$

式中： $\theta_{\text{global}}^{t+1}$ 为全局模型的参数； θ_i^{t+1} 为第*i*个设备的本地模型更新；*N*为设备数量。

每轮本地迭代结束后，更新结果均会经历加密处理后传送至中央服务器，以完成聚合操作和全局模型更新。

1.2.2 边缘侧流式处理引擎

边缘节点上部署 Apache Flink 作为核心流数据处理平台。数据接入模块作为输入管道将原始数据导入 Flink 引擎。引擎内部利用自定义算子执行聚合与过滤等关键处理环节。为应对连续数据流，采用时间窗口机制将数据划分为连续片段进行处理，并通过并行计算策略提升系统吞吐量，最终达成 2.7 GB/s 的数据处理能力^[4]。Flink 优化策略包含使用内存映射文件存储处理中间结果，并依托分布式框架实现负载均衡和细粒度的资源调度。

1.2.3 跨域认证协议

面向跨域交互的安全需求，协议选用国密 SM9 算法执行密钥交换与数据加密。边缘设备与云端服务节点利用 SM9 算法协商建立共享密钥，该密钥继而用于加密通信链路。基于传输层安全性（transport layer security, TLS）协议传输的数据包，由云端持有合法共享密钥予以解密。SM9 算法具备硬件加速支持的潜力，能有效减轻边缘设备计算负担。系统依赖 SM9 的身份认证机制完成合法设备身份验证。

2 智能融合关键技术

2.1 多模态数据对齐技术

2.1.1 时空配准算法

把尺度不变特征变换算法用于从图像和点云中检测关键特征点的工作，激光雷达点云数据与可见光图像的配准操作以基于几何变换的算法为实现基础，通过 RANSAC 算法对初步匹配点进行粗匹配并把噪声与误匹配点过滤掉，为最终求解点云到图像坐标的变换矩阵，引入最小二乘法对几何变换模型进行优化求解，其核心原理在于最小化特征点之间的几何距离，具体求解目标用公式表示为：

$$\hat{T} = \arg \min_T \sum_{i=1}^n \|x_i - T(x'_i)\|^2 \quad (2)$$

式中： $\hat{T}$ 为变换矩阵； x_i 和 x'_i 分别为图像和点云中的特征点；*T*为变换函数。该优化过程是实现点云与图像数据空间对齐的核心步骤。

2.1.2 频域特征融合

加权系数 α 可对两种信息源在最终融合结果中的影响强度做动态校准，面向工业振动信号与红外热像图的多模态融合需求，借助快速傅里叶变换把原始振动信号从时域转换至频域，红外热像图的特征同样在频域空间表示，使用加权融合算法对两模态的频域特征 $F_{\text{vibration}}(\omega)$ 与 $F_{\text{thermal}}(\omega)$ 进行组合集成^[5]，融合操作用公式表示为：

$$F_{\text{fused}}(\omega) = \alpha F_{\text{vibration}}(\omega) + (1 - \alpha) F_{\text{thermal}}(\omega) \quad (3)$$

式中： $F_{\text{fused}}(\omega)$ 为融合后的频域特征； $F_{\text{vibration}}(\omega)$ 和 $F_{\text{thermal}}(\omega)$ 分别为振动信号与红外热像的频域表示； α 为加权系数。

通过此机制实施频域特征的有效融合，能够在综合不同模式信息的基础上保障关键判别特征的完整性。

2.2 边缘联邦学习优化

2.2.1 梯度稀疏化传输

首先设定一个绝对值稀疏化阈值 ϵ ，把原始梯度 *g* 中绝对值低于 ϵ 的分量强制置零，得到稀疏化梯度 *S(g)*，和传输完整梯度数据的传统模式不同，该方法仅选择保留并传输超出重要阈值的关键梯度元素，只有满足 $g > \epsilon$ 条件的梯度分量才会被纳入传输更新范围。引入梯度稀疏化策略的目的是缓解联邦学习中模型梯度传输引起的带宽压力，本地模型参数的更新过程用公式表示为：

$$\theta_{t+1} = \theta_t - \eta \cdot S(g) \quad (4)$$

式中： θ_{t+1} 为更新后的模型参数； θ_t 为当前模型参数； η 为学习率；*S(g)*为稀疏化后的梯度；*g*为原始梯度。稀疏化函数 *S(g)* 对梯度进行阈值处理，将小于 ϵ 的梯度置为零，减少通信中的冗余数据量。

2.2.2 基于 Q-learning 的动态设备选择策略

为解决联邦学习训练轮次中的设备参与效率问题，借助 Q-learning 强化学习开展动态设备选择机制的设计工作，把每个边缘设备的实时状态 *s* 联合表征为其当前的计算负载水平与可用网络带宽^[6]。设备是否参与即将开展的该轮次模型训练被定义为可行动作 *a*，状态 - 动作价值函数 *Q(s,a)* 的更新依照经典的 Q-learning 算法公式进行：

$$Q(s,a) = Q(s,a) + \alpha [R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)] \quad (5)$$

式中：*Q(s,a)*表示在状态 *s* 下选择动作 *a* 的价值；*R(s,a)*表示动作后获得的即时奖励； γ 表示折扣因子； α 表示学习率； $\max_{a'} Q(s',a')$ 表示下一状态下的最大价值。通过 Q-learning 算法，设备选择策略根据当前设备状态自适应地调整，优化训练过程中设备的参与度，减少不必要的计算和通信负担。

2.3 增量式知识融合

2.3.1 在线聚类算法处理概念漂移

在工业监控和传感器数据分析工作中，设备状态或环境条件发生变化会把数据分布的根本属性改变，这种数据统计性质随时间产生变动的情况就是概念漂移。为了能实时适配这类变化，本文设计了基于增量学习的在线聚类算法，通过持续更新聚类模型来处理数据流中的变化^[7]，具体实现阶段会用到 K-means 算法开展在线学习，接收到新数据后会第一时间评估该数据点对现有聚类的影响，各个聚类中心会结合新增数据完成调整，具体更新规则为：

$$c_k(t+1) = c_k(t) + \frac{1}{n} \sum_{i=1}^n (x_i - c_k(t)) \quad (6)$$

式中： $c_k(t+1)$ 为更新后的聚类中心； $c_k(t)$ 为当前聚类中心； n 为聚类内数据点的数量。

这套更新规则把当前聚类中心做逐步的增量修正，能让聚类随数据的变化灵活调整，可解决传统静态聚类方法难以处理概念漂移的问题。

2.3.2 边缘节点本地缓存机制

在实现过程中采用最近最少使用缓存替换算法（least recently used, LRU），该策略依据数据的访问频率与时间先后顺序把缓存内容做好动态管理，为提高数据访问效率并减轻云端负担，边缘节点调用本地缓存机制对实时数据开展存储和管理相关工作，通过完成缓存策略的优化操作，边缘节点能够在网络受限或中断的情况下继续进行数据处理，确保系统的稳定性^[8-9]，具体的缓存更新过程通过公式进行动态调整：

$$\text{cache_size}(t) = \text{cache_size}(t-1) + \Delta\text{cache} \quad (7)$$

式中： $\text{cache_size}(t)$ 表示时间点 t 时的缓存大小； $\text{cache_size}(t-1)$ 表示时刻 $t-1$ 的缓存占用大小； Δcache 表示缓存更新的大小。

通过此算法，系统能够根据数据访问模式自动决定哪些数据应保留在缓存中，哪些应被驱逐，确保缓存的高效利用，并避免内存溢出或性能瓶颈。

3 实验验证与性能分析

3.1 测试环境

实验采用 NVIDIA Jetson TX2 和华为 Atlas 500 智能小站分别作为边缘节点与近边缘计算平台。Jetson TX2 配备 256 核 GPU 与 8 GB 内存，承担数据预处理和 TensorFlow Lite 推理任务；Atlas 500 配置英特尔 Xeon 处理器与 64 GB 内存，支持 GPU 加速并通过 Kubernetes 实现资源动态调度^[10]。测

试数据包括 UCI 工业传感数据集与自建的智慧园区多模态数据集（含 6 类异常），涵盖温湿度、振动与图像等多源异构信息，用于模拟真实环境下的数据融合与推理过程。

3.2 性能分析

为评估本文提出的边缘计算方法与传统云端融合方法在实际应用中的性能差异，本研究对比了多项关键指标，包括融合时延、带宽消耗、异常检出率、内存占用、收敛速度以及处理数据量。表 1 详细展示了两种方法在这些指标上的对比结果，并以百分比符号表示提升幅度。可以看到，本文方法在多个关键性能指标上都明显优于传统云端融合方法，可处理的数据量的显著提升证明该方法能够应对大规模数据流的处理需求，收敛速度加快表明其在模型训练方面具备优势，内存占用的降低保障了方法在边缘设备上的高效运行，异常检出率的提升体现了数据识别与处理的准确性，把融合时延与带宽消耗大幅降低后，还显著增强了系统的实时处理能力与带宽利用率。

表 1 传统云端融合与本文方法关键指标对比

方法	融合时延 /ms	带宽消耗 /(Mbit·s ⁻¹)	异常检出率 /%	内存占用 /MB	收敛速度 /轮	处理数据量 /(GB·s ⁻¹)
传统云端融合	1 200	98.7	82.1	256	200	0.3
本文方法	76	15.3	95.7	183	62	2.7
提升幅度	-93.70%	-84.50%	16.90%	-28.50%	68.00%	800%

4 结语

实验评估证实，本研究构建的多源异构数据智能融合方法把边缘计算与三层次协同架构做了深度整合，相较于常规的云端融合策略，本方案在核心性能指标层面有全面优势，尤其在融合时延控制、网络带宽占用和异常检测精度等方面的提升十分明显。这一方法能显著增强数据处理过程的实时性表现，面对大规模、高异质性的数据流处理场景时，其具备的效能优势会更加突出，该研究成果为智能配电网等多个应用领域的边缘计算落地打下了坚实基础，后续的工作可将更复杂运行环境下该系统适应性能的持续优化作为核心方向。

参考文献：

- [1] 赵永柱. 基于边缘计算的智能配电网多源数据处理与融合技术[J]. 电气自动化, 2024, 46(5): 79-81.
- [2] 赵静, 余恒洁, 林聪, 等. 基于边缘计算的智能电网多源异构数据监测[J]. 信息技术, 2023(9): 77-82.
- [3] 辛鲁斌, 韩玲, 李良志. 基于多源数据融合的滑坡智能识别[J]. 地球科学与环境学报, 2023, 45(4): 920-928.

一种基于聚类与多距离度量的下采样算法

杨 涛¹
YANG Tao

摘 要

数据进行采样处理可以有效解决许多领域在模型训练时所面临的数据不平衡问题。为提升数据下采样算法性能,有效剔除冗余样本,文章设计了一种基于聚类和多距离度量的下采样算法。首先,使用 K-means 算法对样本进行聚类;然后,根据簇内样本的分布情况,选择多数类较多的用于下采样;接着,从马氏距离、欧氏聚类和余弦相似度多个距离维度度量簇内样本分布情况;最后,根据综合距离和多数类样本间数量关系,确定单簇采样数量,剔除冗余样本,解决样本数量不平衡问题。在四个真实数据集和五个分类器上的实验结果表明,该算法能够有效提升分类模型的性能。

关键词

不平衡数据; K-means; 多距离度量; 下采样; 数据处理

doi: 10.3969/j.issn.1672-9528.2026.04.027

0 引言

在实际的模型训练与测试工作中,常受限于数据的不平衡问题。所谓数据不平衡,就是指数据集中,某些类别的样本数量多于其他类别,从而导致模型的训练和测试结果向多数类偏移,影响模型的整体性能。

在包括信用评分^[1]、故障车辆检测^[2]、医疗保险欺诈检测^[3]、网络异常流量检测^[4]以及轴承故障诊断^[5]在内的众多领域,均存在数据的不平衡问题。为了解决数据不平衡所带来的负面影响,各种方法应运而生,包括基于数据采样的数据层面的解决方法^[6];基于特征工程^[7]的方法;以及基于模型优化改进^[8]的模型层面的方法。其中,数据层面的采样方法能够适应不同领域的多种类型数据,因而应用广泛。采样算法根据其对样本的处理方式,可以细分为上采

样(over-sampling)、下采样(under-sampling)和集成采样(ensemble-sampling)。

上采样方法通过合成数据集中的少数类样本来平衡数据集,常用的少数类样本合成方法以 SMOTE (synthetic minority over-sampling technique) 及其各种衍生算法^[9]为主,具有代表性的有 SMOTE、ADASYN (adaptive synthetic sampling)、K-means SMOTE^[10]等算法。在处理不平衡程度较高的数据集时,算法需要合成大量样本,这可能会导致合成的少数类样本数量超过原有的少数类样本数量,造成数据集失真。

下采样算法通过清除多数类样本中的冗余样本和噪声样本来降低数据集的不平衡度,从而降低数据集的不平衡特性对后续模型的影响。最直接的采样算法是随机下采样,该算法通过随机删除多数类样本来降低数据不平衡度^[6],这可能会损失重要样本,因而较少使用;而 NCR (neigh-

1. 重庆工信职业学院 重庆沙坪坝 400037

[4] 李爱华,续维佳,石勇.基于数据融合的商务智能与分析架构研究[J].计算机科学,2022,49(12):185-194.

[5] 张合沛,陈贇.面向多源异构信息融合的隧道群组装备数据提取技术研究[J].隧道建设(中英文),2022,42(1):41-47.

[6] 蒲未来,刘敦龙,桑学佳,等.融合多源异构数据的滑坡变形阶段智能判别方法[J].灾害学,2023,38(4):179-186.

[7] 吴羿龙,余珊,陈莹捷,等.融合多源异构大数据的同安内涝风险评估[J].测绘与空间地理信息,2025,48(4):70-73.

[8] 杨童博,魏政,周坤.基于近似失效点的多源异构数据融合可靠性评估模型[J].机械强度,2025,47(3):136-142.

[9] 卢加荣,廖彬,刘怡,等.融合多源异构数据的 ICO 欺诈预测与可解释分析模型[J].计算机应用研究,2025,42(2):357-364.

[10] 费英群,田林.面向多源异构的电力工程数据融合处理技术研究[J].电子设计工程,2025,33(1):104-108.

【作者简介】

王伟(1978—),男,山东诸城人,本科,高级工程师,研究方向:电子信息。

(收稿日期:2025-09-04 修回日期:2026-04-08)