公共数据授权运营的鲁棒联邦学习入侵检测系统

卫　琳[1]，陈　超[1]，韩　丁[2]，李秋燕[2]，和孟佯[1]

（1.郑州大学 网络空间安全学院，河南 郑州 450001；2.国网河南省电力公司经济技术研究院，河南 郑州 450052）

摘　要：公共数据授权运营在实现多方数据共享复用的同时，面临针对公共数据的恶意入侵威胁，亟须可靠的入侵检测技术保障数据安全与隐私。传统基于联邦学习的入侵检测技术中存在以下问题：入侵检测数据类别不平衡，导致模型对少数类入侵样本误报率高；节点之间的数据异构性加剧了局部模型与全局模型的差异，导致收敛速度变慢；恶意节点可能通过中毒攻击破坏模型训练，影响良性节点的检测性能。为此，提出了一种面向公共数据授权运营的鲁棒联邦学习入侵检测框架（FL-RIDS）。首先，引入了一种基于有效样本数加权的入侵检测损失函数（ESLoss），通过动态调整类别权重显著提升了模型对小样本入侵类别的识别敏感度；其次，提出一种基于模型性能评估的聚类方法（SEC），通过在云服务器上评估各主体节点本地模型表现，将数据特征相似的节点进行聚类，有效缓解了异构数据导致的模型性能下降和收敛问题；最后，为确保系统对抗恶意攻击的鲁棒性，提出了一种轻量级中毒检测算法（LPD），通过监测主体节点本地模型与集群内全局模型分类器更新方向的相似度变化，动态识别训练过程中发动中毒攻击的恶意节点。实验结果表明：在 CICIDS2017 和 CAN-intrusion-dataset 两个数据集上，ESLoss 相对交叉熵损失函数 *f*1*score* 最高分别提升了 18.9 和 3.2 百分点；SEC 与 FedAvg 相对比，*f*1*score* 最高分别提升了 4.9 和 5.9 百分点，并展示出良好的抗中毒攻击能力；同时，LPD 与 FedAvg 相对比，*f*1*score* 最高分别提升了 54.2 和 68.4 百分点，能有效抵抗恶意节点在训练中途的中毒攻击。

关键词：聚类联邦学习；公共数据；入侵检测；中毒攻击防御；数据异构；类别不平衡

中图分类号：TP18；TP391；TP393　　**文献标志码**：A　　**doi**：10.13705/j.issn.1671-6833.2027.01.008

公共数据授权运营通过多主体授权流转激活数据价值，但采集、流转过程中面临非法访问、数据篡改等安全风险。与通用联邦入侵检测相比，该场景合规约束更强、跨域数据异构性更显著、攻击风险更高。例如车联网等多主体协同场景同时面临 3 类核心挑战：高危攻击样本稀缺导致类别不平衡、跨主体业务差异引发数据异构、开放接入存在恶意节点投毒风险。

入侵检测是有效的网络攻击防护手段[1]，但该场景下数据分布失衡问题尤为突出：正常流量占绝大多数，入侵样本稀少且类别分布不均，政务、医疗等关键节点的低频高危攻击样本更是匮乏，导致模型训练偏向多数类，对关键威胁漏检率高，易引发敏感数据泄露等安全事故。现有入侵检测少数类识别研究多采用采样技术[2-4]和损失函数优化[5-6]来处理不平衡问题，但过采样易引入噪声且违反合规要求，欠采样丢失关键信息，极端不平衡场景下效果有限，难以适配多主体协同部署需求。

现有集中式入侵检测无法满足隐私合规要求，联邦学习 FedAvg[7]提供了可行路径，但公共数据运营场景主体多样、数据分布差异大，严重削弱 FedAvg 有效性[8-10]。政务、医疗、金融等多领域节点的数据格式、敏感等级、业务场景差异显著，面临的入侵威胁也不同（如政务节点易受授权篡改，医疗节点易遭病历窃取），这种内生数据分布差异使局部与全局模型目标冲突，导致训练收敛慢、不稳定，全局模型性能平庸，无法应对复杂安全威胁。聚类联邦学习可缓解异构问题，但现有方法存在依赖先

收稿日期：2026-09-09；**修订日期**：2026-10-10

基金项目：河南省重点研发计划项目（251111212100）

作者简介：卫琳（1968—），女，河南济源人，郑州大学副教授，主要从事网络与系统安全、数据科学与智能计算、网络与分布式计算等方向的研究，E-mail：weilin@zzu.edu.cn。

通信作者：和孟佯（1994—），女，河南南阳人，郑州大学助理研究员，博士，主要从事新型网络技术、信息安全与人工智能等方向的研究，E-mail：myhe@zzu.edu.cn。

验知识、客户端开销大、隐私泄露风险等局限[8-9,11-12]:缺乏先验知识确定最优集群划分策略;要求参与者承担额外模型评估与聚类计算开销,与节点优先保障核心业务的资源分配原则相悖;且参与者交换额外信息会增加通信开销,还可能泄露敏感信息,违反隐私合规要求。此外,联邦学习服务器无法完全控制参与者的行为,这使得联邦学习过程容易遭受恶意主体节点的中毒攻击[13-14]。针对此类攻击,现有鲁棒聚合方案[15-16]在异构数据场景下效果不佳,构建适配该场景的理想入侵检测模型仍具挑战。

针对上述挑战,本文提出面向公共数据授权运营的鲁棒联邦学习入侵检测系统(federated learning-based robust intrusion detection system,FL-RIDS),针对公共数据授权运营场景做了专属适配。通过3项关键技术解决核心问题:针对类别不平衡问题,引入有效样本数加权损失函数(effective sample number weighted loss,ESLoss),通过量化有效样本数动态调整类别权重,在保障多数类检测性能的同时提升小样本入侵识别敏感度,无须改动节点原始数据,符合公共数据不出域的合规约束;针对数据异构问题,设计基于服务端评估结果的聚类方法(server evaluation cluster,SEC),通过云服务器评估节点本地模型表现,实现无先验集群数量、无客户端额外负担的高效聚类,同时初步排除训练初期恶意节点;针对SEC无法应对中途中毒攻击的局限,提出轻量级中毒检测算法(lightweight poisoning detection,LPD),通过监测本地与集群全局模型分类器更新方向相似度变化,动态识别并隔离恶意节点,轻量高效的检测机制可适配公共数据运营场景下大规模异构节点的部署需求。本文主要贡献如下。

(1)提出面向公共数据授权运营的FL-RIDS整体框架,引入基于有效样本数加权的ESLoss损失函数,将有效样本数思想迁移至联邦入侵检测场景,无须改动原始数据即可缓解类别不平衡导致的少数类识别精度低问题。

(2)提出基于服务端评估指标的SEC自适应聚类方法。该方法利用公共评估数据构造节点模型行为特征评价指标序列SEM(server evaluation metrics),并据此聚合同类授权主体,缓解公共数据授权运营中多主体数据异构导致的性能退化。

(3)设计轻量级中毒检测方法LPD,通过监测节点本地模型更新方向与集群全局模型更新方向的相似度下降,动态识别训练过程中被攻击或恶意的授权主体,从而增强公共数据授权运营场景下联邦入侵检测的鲁棒性。

1 面向公共数据授权运营的鲁棒联邦学习入侵检测框架FL-RIDS

1.1 FL-RIDS的工作过程

本文3层架构与核心方法均针对上述场景特性设计:ESLoss不需要改动原始数据,契合数据不出域的合规约束;SEC由服务端完成聚类,避免节点额外隐私泄露;LPD采用轻量检测机制,适配大规模主体接入需求。系统整体分为以下3层。

节点层。对应车企、交管部门、第三方服务商等授权运营主体,各主体仅基于本地数据训练入侵检测模型,不上传原始数据;边缘服务器层对应区域数据流转节点,负责模型转发与协同调度;云服务器层对应公共数据授权运营平台的模型管理与安全协同中心,负责模型评估、聚类、聚合和异常主体识别。

边缘服务器层。由计算通信能力较强的边缘服务器组成,负责节点和云服务器之间的模型转发。

云服务器层。由具有强大计算、通信和存储能力的云服务器组成,负责收集各节点模型。初始化阶段用公共数据评估节点并聚类,迭代阶段通过FedAvg算法聚合集群内节点模型,生成个性化模型并下发至对应节点。

FL-RIDS工作流程分为初始化与迭代两个阶段,如图1所示,步骤如下。

(1)系统初始化。云服务器定义联邦学习超参数,向边缘服务器下发初始模型w^0,初始化阶段无须逐轮下发全局模型。

(2)边缘转发。边缘服务器向各节点转发初始模型与超参数。

(3)本地训练。前$a-1$轮中,节点i基于本地数据D_i充分训练,第a至s轮,将训练好的模型w_i^r发送到边缘服务器。其中,a为开始上传模型的轮次,s为初始化阶段总轮次。

(4)模型转发。边缘服务器将全部节点模型转发至云服务器。

(5)云服务器评估。出于隐私保护,节点本地数据分布无法直接获取,而现有方法[8,11]要求客户端上传其他额外参数。鉴于本文目标是最大程度保护隐私,因此,利用云服务器公共数据提取节点模型中的数据分布特征。第a至s轮,云服务器接收模型后,用公共数据评估各节点模型w_i^r,并将各个指标按类别拼接为服务器评价指标序列:

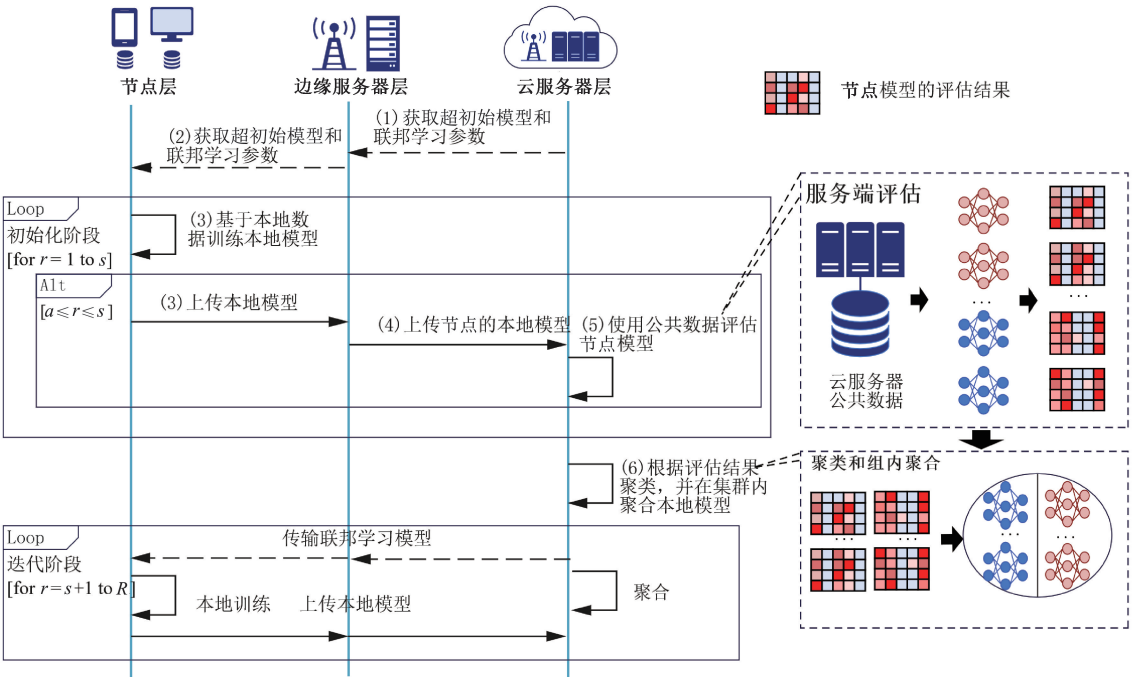


图 1　FL-RIDS 的工作流程

Figure 1　The workflow of FL-RIDS

$$\boldsymbol{SEM}_{\mathrm{i}}^{r}=\begin{bmatrix} recall_{i,1}^{r} & recall_{i,2}^{r} & \cdots & recall_{i,C}^{r} \\ precision_{i,1}^{r} & precision_{i,2}^{r} & \cdots & precision_{i,C}^{r} \\ f1score_{i,1}^{r} & f1score_{i,2}^{r} & \cdots & f1score_{i,C}^{r} \\ fnr_{i,1}^{r} & fnr_{i,2}^{r} & \cdots & fnr_{i,C}^{r} \end{bmatrix}。\quad (1)$$

式中：$\boldsymbol{SEM}_{i}^{r}$ 每一行对应各类别的某一评价指标；C 为数据类别总数。

云服务器作为平台核心枢纽，受攻击类型全、强度高，可依托自身历史安全日志、捕获攻击流量结合公开威胁情报构建公共评估数据集，天然覆盖多类型攻击与正常流量模式，可作为统一评估标尺。若评估集与节点本地数据存在分布偏差，会降低评估区分度与聚类精度，导致当前静态评估集存在一定失效风险。对此，云服务器建立常态化增量更新机制：定期整合新增安全日志与威胁情报，补充新型攻击样本，逐步校准评估集分布。更新初期，评估维度扩展会使 SEM 序列出现短期波动，可能引发聚类结果微调；随着评估集覆盖度提升，SEM 的分布表征能力持续增强，聚类稳定性与全局模型精度将逐步收敛并保持稳定。运营阶段可结合评估集更新周期同步执行聚类重校准，进一步保障框架的长期鲁棒性。

(6)云服务器聚类。基于 SEM 序列通过 SEC 算法对节点聚类，整合数据分布相似的节点。

完成初始化聚类后进入迭代阶段，云服务器在各集群内通过 FedAvg 聚合完成多轮协同训练，模型收敛后将最终模型 $\{w_k^R\}_{k\in K}$ 下发到各个节点。

1.2　基于有效样本数加权的入侵检测损失函数

数据分布不平衡是普遍存在的挑战。本文 ES-Loss 基于文献[17]有效样本数思想迁移适配：随着样本量增加，新增样本的边际效益会递减，这是因为样本间存在信息重叠，新样本多为已有样本的近似重复或微小变体。该理论适配公共数据授权运营的入侵检测场景。

多源数据汇聚、接口开放使攻击呈现两极分化特征：针对数据服务接口的高频 DDoS 攻击流量特征高度相似，样本规模增长对模型训练的边际增益快速递减；而以窃取授权数据为目标的高级持续性威胁等罕见攻击的样本量虽稀少，却携带突破防护的关键特征，直接关乎公共数据安全底线。采用基于有效样本数加权的方式，可平衡稀有与常见类别权重，显著提升模型性能。某类别的有效样本数[17]可定义为

$$E_n=(1-\beta^n)/(1-\beta)。\quad (2)$$

式中：β 为有效样本数衰减系数，取值(0,1)，β 越趋近于 1，样本重叠越低，所有样本的贡献趋于相等；β 越趋近于 0，样本之间越高度重叠，只有第 1 个样本的贡献显著；n 为对应类别的总样本数。本文将该思想迁移至入侵检测场景，使用有效样本数 E_n 加权损失，通过动态调整不同类别的损失权重，缓解数据不平衡带来的模型偏差问题。核心思想是有效样本数稀少的类别应获得更高的权重，而样本充足的类别权重应适当降低。有效样本数加权损失函数定义为

$$\mathrm{ESLoss}(p,y) = (1-\beta)/(1-\beta^{n_y}) \cdot \mathrm{CE}(p,y)。 \quad (3)$$

式中：n_y 为标签 y 所对应类别的样本总数；CE(·) 为交叉熵损失函数；p 为模型预测概率输出；y 为样本真实类别标签。

1.3 基于服务端评估指标的自适应聚类方法 SEC

鉴于轮廓系数[18]在评价聚类结果时能够同时考虑簇内样本的紧密性和簇间样本的分离性，本文提出了一种基于轮廓系数的自适应 K-means++聚类方法。SEC 的重点在于利用公共评估数据构造 SEM 序列，以模型行为相似性间接刻画节点本地数据分布相似性，从而在不访问原始数据的前提下完成授权主体聚类。该算法包括两个步骤：基于 K-means++的聚类；确定最优聚类数量 K。与传统的 K-means 相比，K-means++算法的改进主要体现在初始化过程中，通过使聚类中心之间的距离尽可能远，以加快聚类收敛速度。具体而言，K-means++首先随机选取一个 SEM 作为初始化聚类中心 $\boldsymbol{c}_1$。对于每一个 $\boldsymbol{SEM}_i$，计算与已经选择的聚类中心的最小距离：

$$D(\boldsymbol{SEM}_i) = \min_{1 \leqslant j \leqslant t} \|\boldsymbol{SEM}_i - \boldsymbol{c}_j\|^2。 \quad (4)$$

然后，使用式(5)计算各节点被选择的概率 P_i，从而选出下一个聚类中心 $\boldsymbol{c}_{j+1}$。

$$P_i = \frac{D(\boldsymbol{SEM}_i)}{\sum_{i=1}^{N} D(\boldsymbol{SEM}_i)}。 \quad (5)$$

重复上述步骤，选择 K 个质心。随后，使用 K-means 方法完成聚类。接着对于一组指定候选聚类数量 $K \in \{2,3,\cdots,K_{\max}\}$，首先使用每一个 K 值进行聚类得到聚类结果，然后在每个 K 下计算各 SEM 的轮廓系数[18]：

$$s(i) = \frac{b(i) - a(i)}{\max(a(i),b(i))}。 \quad (6)$$

式中：$a(i)$ 为 $\boldsymbol{SEM}_i$ 与所在集群中其他样本的平均距离；$b(i)$ 为 $\boldsymbol{SEM}_i$ 与离其最近的其他集群中样本的平均距离。所有轮廓系数的均值作为该 K 值的轮廓系数 S_K：

$$S_K = \frac{1}{n}\sum_{i=1}^{n} s(i)。 \quad (7)$$

最终选择 S_K 值最大的 K 值作为最优 K 值。SEC 整体时间复杂度为 $O(K_{\max} \cdot (NKD + N^2D))$，其中，$N$ 为节点数；K 为聚类数；D 为 SEM 维度，节点规模较大时轮廓系数计算为主要开销。典型场景下攻击类别有限，SEM 维度可控；类别扩充时可通过指标筛选、PCA 降维保障效率，大规模节点场景下可通过并行计算轮廓系数降低运行耗时。

1.4 基于模型分类器更新方向相似度的轻量级中毒检测算法 LPD

SEC 展现出较强的抗中毒攻击能力。对于一组数据相似的节点，若其中的恶意节点破坏系统，其云服务器端评估结果会显著偏离良性节点，可被排除在集群外。但 SEC 存在恶意节点初始化聚类成功后，训练中途发动攻击的逃逸问题，因此本文提出基于模型分类器更新方向相似度的轻量级中毒检测算法 LPD，识别恶意节点并限制其参与全局模型聚合，提升系统对隐蔽攻击的应对能力。

分类器相似度可有效度量节点数据分布相似性，监测分类器变化能反映节点数据标签变动。仅计算分类器相似度，既能有效识别恶意节点，又能显著降低云服务器计算负担。假设恶意节点初始化聚类正常，迭代阶段某一轮开始发动中毒攻击：未攻击时，其分类器更新方向会与集群全局模型趋于一致；攻击后，更新方向显著偏离，相似度下降。$h=2$ 的异构环境下的实验(2.3 节)显示，恶意节点 20 轮后发动攻击，每次持续 5 轮，随后 5 轮恢复正常，如图 2 所示，恶意节点与全局模型的相似度在攻击与正常轮次间变化明显。因此，监测节点与全局模型分类器更新方向的相似度变化可有效检测恶意节点。

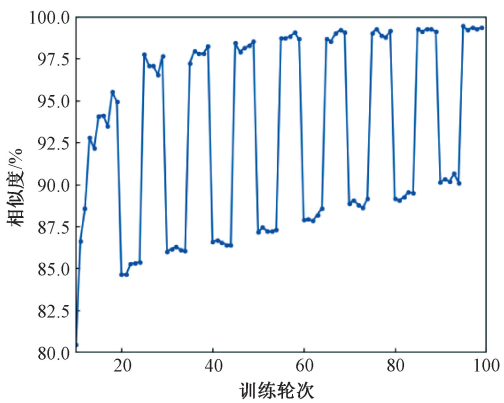


图 2 恶意节点的相似度随轮次的变化

Figure 2 Similarity variation of a malicious node with training rounds

具体而言，首先用每轮训练后节点本地分类器 φ_i^r 与迭代阶段开始时集群内全局模型分类器 φ^s 的差值来表示节点模型分类器的整体更新 $\Delta\varphi_i$：

$$\Delta\varphi_i = \varphi_i^r - \varphi^s。 \quad (8)$$

接着计算上一轮的集群内全局模型分类器 φ^{r-1} 与 φ^s 的差值 $\Delta\varphi$ 来表示集群内全局模型分类器的整体更新。为降低计算开销，采用一种直接的方法计算两者更新方向相似度，即对应参数具有相同符号的百分比：

$$sim_i = \frac{1}{M}\sum_{j=1}^{M} I(\mathrm{sgn}(\Delta m_{ji}) = \mathrm{sgn}(\Delta m_j))。 \quad (9)$$

式中：M 为模型分类器 φ 的参数数量；Δm_{ji} 为节点 i 的分类器 φ_i 的第 j 个参数的更新；Δm_j 为集群内全局模型的分类器的第 j 个参数的更新；$I(\cdot)$ 为指示函数，一致为 1，否则为 0。

得到更新方向相似度后，本文发现正常节点相似度波动较小，恶意节点攻击后相似度变化显著，因此设定相似度变化阈值 t 检测恶意节点。对于每个节点 i，本文比较本轮相似度 sim_i 和正常轮次相似度 sim_{normal} 的变化。对于恶意节点，sim_{normal} 是来自没有发起攻击的轮次的相似性，而对于正常节点，它通常来自前一轮。如果 $(sim_i - sim_{\mathrm{normal}}) < t$，则将当前节点标记为恶意，否则，更新 sim_{normal} 为本轮相似度 sim_i。中毒检测需要正常轮次的节点相似度，因此在迭代开始时需要使用第 s、$s+1$ 轮的对应全局模型分类器 φ^s、φ^{s+1} 以及第 $s+2$ 轮的各节点模型分类器 φ_i^{s+2} 来初始化各节点的正常轮次相似度。若恶意节点在基准初始化前发起早期攻击，LPD 本身无法有效识别；当前依靠 SEC 聚类的性能筛选机制可筛除大部分早期恶意节点，一定程度保障基准轮次的干净性。在 CAN-intrusion-dataset[19] 的实验表明，高强度攻击时，恶意节点相似度下降超 7%，后续维持 3%以上；良性节点波动<1%，可设-3%阈值；低强度攻击时相似度下降约 3%，后期在 1%~2%，可设-1%阈值。阈值通常略高于正常波动范围。该检测方法仅基于模型分类器的参数更新方向开展判定，与底层特征提取网络的结构无关，因此可适配 CNN、MLP、Transformer 等主流网络结构。面对更复杂的模型时，本文方法仍仅需提取分类器层参数计算更新相似度，计算开销不会随模型整体参数量同步增长，轻量特性得以保持，具备良好的可扩展性。

2 实验评估

2.1 数据集

本文使用了车联网入侵检测数据集 CAN-intrusion-dataset[19] 和公共网络入侵数据集 CICIDS2017[20] 进行方案评估。CAN-intrusion-dataset 对应车联网公共数据运营场景：授权主体含车企、交管部门、第三方服务商，数据遵循“终端-边缘-平台”三级流转，攻击类型对应越权读取、指令篡改等场景风险；CICIDS2017 对应通用公共数据平台场景：授权主体含政务部门、服务机构、运营企业，遵循分级授权管控，攻击类型对应未授权接入、数据窃取等场景风险，二者共同覆盖公共数据运营的核心特征。

2.2 ESLoss 性能评估

本文参考文献[17]中的方法构建长尾分布数据集。以最少样本类别的样本数 n_0 为基准，其余各类的样本数按指数函数 $n_i = n_0 \times \mu^i$ 进行设置，$\mu>1$ 为不平衡系数。通过调节 μ 值来构建不同不平衡程度的数据分布。数据划分均采用 60%训练集和 40%测试集比例。本文利用网格搜索法来确定 β 的最佳值，搜索范围设置为[0.990 0，-0.999 9]，步长在到达 0.999 0 之前设置为 0.001 0，随后设置为 0.000 1。最终，在 β 为 0.999 0 时达到最佳效果。

基线方法包括经典的采样方法：Borderline-SMOTE[3]、ADASYN[2]、TomekLinks[21]。也包括一些加权或修改损失函数方法：WCE[6]、DFL[8]、Focal-Loss[5]、SEQL[22]、交叉熵损失 CE。表 1 展示了不同不平衡程度的实验结果。

表 1 各方法在两个数据集上不同 μ 值的实验结果

Table 1 Experiment results of various methods under different μ values on two datasets

方法	CICIDS2017						CAN-intrusion-dataset					
	$\mu=3$		$\mu=3.5$		$\mu=4$		$\mu=3$		$\mu=3.5$		$\mu=4$	
	acc	*f1score*	*acc*	*f1score*	*acc*	*f1score*	*acc*	*f1score*	*acc*	*f1score*	*acc*	*f1score*
CE	0.943	0.508	0.961	0.513	0.971	0.516	0.940	0.808	0.946	0.841	0.970	0.829
ADASYN[2]	0.722	0.526	0.669	0.484	0.676	0.481	0.940	0.830	0.948	0.858	0.967	0.842
BorderlineSMOTE[3]	0.770	0.527	0.704	0.480	0.674	0.471	0.934	0.819	0.946	0.862	0.970	**0.858**
WCE[6]	0.876	0.619	0.903	0.588	0.928	0.655	0.939	0.810	0.936	0.813	0.920	0.827
DFL[8]	0.936	0.629	0.959	0.512	0.966	0.517	0.943	0.812	0.943	0.789	0.966	0.794
FocalLoss[5]	0.941	0.629	0.958	0.580	0.966	0.514	0.935	0.802	0.945	0.799	0.966	0.823
SEQL[22]	0.946	0.511	0.959	0.512	0.968	0.515	0.944	0.806	0.943	0.788	0.966	0.812
TomekLinks[21]	0.942	0.509	0.960	0.513	0.968	0.515	0.938	0.759	0.948	0.809	0.970	0.807
ESLoss	0.942	**0.697**	0.953	**0.630**	0.966	**0.669**	0.946	**0.840**	0.943	**0.864**	0.969	**0.858**

本文方法在绝大部分情况下显著提升,在两个数据集上,与 CE 相比 $f1score$ 最高分别提升了 18.9 和 3.2 百分点。CE 由于没有对少数类数据施加任何的关注,因此对少数类识别的效果较差。采样方法由于生成的样本与真实样本存在一定差距,加大了模型的误报率。样本数加权方法的模型过度关注少数类样本,导致了高误报率。FocalLoss 在不平衡程度低的情况下,取得较好效果,但在高度不平衡的场景下效果较差。

本文在 CICIDS2017 上探究 β 对模型性能的影响。实验选取了 4 个有代表性的 β 值:0.900 0,0.990 0,0.999 0,0.999 9,μ 设置为 3。如表 2 所示,随着 β 值从 0.900 0 增大至 0.999 9,有效样本数加权机制使模型对少数类的关注度持续增加,这导致了准确率的下降,表明对多数类的判别精度有所牺牲;而 $f1$ 则先升后降,在 β 为 0.999 0 时达到峰值 0.679。这是因为 β 为 0.900 0 时,对少数类的加权不足,模型仍偏向多数类;β 过大时,对少数类的过度关注可能引入误判,准确率显著降低。因此,0.999 0 是一个较优的平衡点,在轻微损失准确率的前提下,最大幅度提升了少数类的综合识别性能。这验证了 ESLoss 方法通过校准 β 能够有效增强对关键少数类的检测灵敏度。

表 2　不同 β 取值下的模型性能

Table 2　Model performance at different β values

β	acc	$f1$
0.900 0	0.945	0.578
0.990 0	0.941	0.619
0.999 0	0.937	0.679
0.999 9	0.911	0.673

2.3　FL-RIDS 异构性能评估

本文随机抽取了部分数据用于训练和测试。数据集的 80%作为训练集,剩余 20%作为节点的测试集。实验采用 3 层卷积和 2 层全连接的 CNN 模型,联邦学习超参数按数据集设置:CAN-intrusion-dataset 本地训练 2 轮、批次大小 64、初始学习率 0.001;CICIDS2017 本地训练 1 轮、批次大小 16、初始学习率 0.000 5,全局通信总轮数均为 100。本文按照文献[9]的方法人工划分异构数据集。具体操作是,每个节点随机在 h 类入侵数据中抽样,同时将良性数据均匀分配给每个节点。本文将 FL-RIDS 与 FedAvg[7]、FedProx[23]、IFCA[10] 以及 FedGH[9] 对比。

表 3 展示了各方法在两个数据集上不同异构程度上的实验结果。FL-RIDS 在各种情况下均显著优于其他基线方法,在两个数据集上的 $f1score$ 分别比 FedAvg 最多高出 4.9 和 5.9 百分点。虽然 FedProx 引入了正则化项来缓解异构的影响,但在处理极度异构的情况时仍表现不佳。FedGH 方法只训练一个共享全局头,虽然降低了异构数据的影响,但缺少了节点之间的知识共享。IFCA 的效果则依赖于预设的聚类数量。从变化规律来看,异构程度越高,节点数据分布差异越大,全局聚合后的性能衰减越明显;SEC 通过聚类提升簇内数据同质性,因此在高异构场景下的优化作用更突出。

2.4　中毒攻击防御

本文探讨了中毒攻击中的数据中毒攻击和模型中毒攻击,使用文献[8,24]中的攻击方式来攻击。CICIDS2017 数据集设置总节点 80 个,选取 30 个作为恶意节点,占比 37.5%;CAN-intrusion-dataset 数据集设置总节点 120 个,选取 40 个作为恶意节点,占比 33.3%。LPD 的相似度变化阈值 t 设置为 -1%。实际公共数据运营场景中恶意节点占比通常不超过 50%,当前实验已覆盖典型高风险场景。

首先设定在训练初期来发动攻击。本文对比了 FedAvg[7] 和 IFCA[10],比较了良性节点在有无中毒攻击两种场景下的性能表现。鉴于两个数据集的结论相似,此处仅以 CICIDS2017 为例,良性节点在攻击前后的性能差异汇总于表 4 中。FL-RIDS(无 LPD)表现均优于 FedAvg 和 IFCA。FedAvg 和 IFCA 性能下降幅度更为明显。模型攻击强度更高,对系统影响更大,原因是 FedAvg 聚合机制会将恶意节点的更新扩散至所有良性节点;同时,IFCA 方法未能

表 3　各方法在两个数据集上不同 h 值的实验结果

Table 3　Experiment result of various methods under different h values on two datasets

方法	CICIDS2017						CAN-intrusion-dataset					
	$h=1$		$h=2$		$h=3$		$h=1$		$h=2$		$h=3$	
	acc	$f1score$	acc	$f1score$	acc	$f1score$	acc	$f1score$	acc	$f1score$	acc	$f1score$
FedAvg[7]	0.868	0.806	0.863	0.746	0.857	0.703	0.907	0.904	0.917	0.896	0.921	0.889
FedProx[23]	0.871	0.813	0.861	0.752	0.855	0.681	0.905	0.905	0.921	0.903	0.922	0.891
FedGH[9]	0.871	0.836	0.856	0.765	0.846	0.716	0.942	0.936	0.912	0.884	0.895	0.848
IFCA[10]	0.878	0.821	0.865	0.743	0.858	0.704	0.949	0.947	0.931	0.912	0.925	0.892
FL-RIDS	**0.893**	**0.855**	**0.879**	**0.791**	**0.866**	**0.728**	**0.967**	**0.963**	**0.948**	**0.933**	**0.938**	**0.911**

表 4　CICIDS2017 数据集中,在中毒与非中毒条件下,不同攻击类型下良性节点的性能差异

Table 4　In the CICIDS2017 dataset, the performance differences of benign nodes under different attack types, both under poisoning and non-poisoning conditions

方法	良性转为攻击		攻击转为良性		所有标签反转		更新缩放攻击		相同全局模型		参数截断攻击	
	acc	*f*1*score*	*acc*	*f*1*score*	*acc*	*f*1*score*	*acc*	*f*1*score*	*acc*	*f*1*score*	*acc*	*f*1*score*
FedAvg	−0.031	−0.020	−0.002	−0.049	−0.025	−0.031	−0.212	−0.491	−0.458	−0.492	−0.212	−0.480
IFCA	−0.031	−0.009	−0.003	−0.036	−0.027	−0.045	−0.213	−0.505	−0.456	−0.505	−0.214	−0.487
FL-RIDS(无 LPD)	**0.002**	**0.001**	**0**	**0**	**−0.001**	**−0.003**	**−0.001**	**−0.004**	**−0.001**	**−0.004**	**−0.002**	**−0.002**

有效区分良性节点与恶意节点,导致良性节点遭受到较大损害。而所提方法通过将恶意节点从其预期所在的集群中排除,显著减轻了中毒攻击的影响。

恶意节点初始化阶段行为正常,迭代 40 轮后发起攻击,对比了 FL-RIDS(无 LPD)、FL-RIDS(包含 LPD)、FedAvg[7] 和 Median[16] 性能,实验结果如表 5 所示。在各种中毒攻击情况下,FL-RIDS(包含 LPD)表现均优于其他方法,在两个数据集上的 *f*1*score* 分别比 FedAvg 最多高出 54.2 和 68.4 百分点。FL-RIDS(无 LPD)通过适当的节点聚类,减少了异构节点影响,整体表现优于 FedAvg,并在一些情况下超过 Median,但部分高强度模型攻击场景,FL-RIDS(无 LPD)无法有效识别并过滤恶意节点的污染更新,出现了明显的性能衰减,这也验证了 LPD 模块在抵御投毒攻击中的作用。不同攻击方式的效果差异源于对分类器更新方向的扰动程度:高强度攻击偏离全局趋势更显著,易被相似度机制捕获;低强度攻击扰动较小,检测增益相对平缓。FL-RIDS(包含 LPD)方法能在训练中途检测并限制恶意节点在其集群内的模型聚合,从而在所有方法中表现最为突出。

表 5　训练中途中毒条件下,不同攻击类型下良性节点的实验结果

Table 5　Experiment results of benign nodes under different attack types with poisoning during training

数据集	方法	良性转为攻击		攻击转为良性		所有标签反转		更新缩放攻击		相同全局模型		参数截断攻击	
		acc	*f*1*score*	*acc*	*f*1*score*	*acc*	*f*1*score*	*acc*	*f*1*score*	*acc*	*f*1*score*	*acc*	*f*1*score*
CICIDS2017	FedAvg[7]	0.844	0.732	0.859	0.704	0.847	0.705	0.650	0.261	0.375	0.236	0.650	0.261
	Median[16]	0.861	0.727	0.851	0.693	0.856	0.712	0.829	0.657	0.821	0.644	0.841	0.680
	FL-RIDS(无 LPD)	0.855	0.762	0.873	0.764	0.862	0.759	0.650	0.261	0.403	0.245	0.651	0.268
	FL-RIDS(包含 LPD)	**0.876**	**0.785**	**0.875**	**0.776**	**0.876**	**0.784**	**0.875**	**0.782**	**0.874**	**0.778**	**0.874**	**0.781**
CAN-intrusion-dataset	FedAvg[7]	0.878	0.870	0.894	0.860	0.881	0.854	0.642	0.260	0.392	0.248	0.642	0.260
	Median[16]	0.909	0.881	0.895	0.858	0.905	0.876	0.849	0.821	0.901	0.878	0.885	0.855
	FL-RIDS(无 LPD)	0.920	0.898	0.925	0.901	0.897	0.868	0.642	0.260	0.418	0.276	0.659	0.302
	FL-RIDS(包含 LPD)	**0.950**	**0.935**	**0.944**	**0.927**	**0.947**	**0.931**	**0.946**	**0.929**	**0.948**	**0.932**	**0.946**	**0.930**

阈值 t 的灵敏度分析。本文在 CICIDS2017 数据集上以“攻击转为良性”的攻击方式进行实验。本文主要关注恶意节点检测率(即正确识别出的恶意节点占全部恶意节点的比例)和良性节点误报率(被错误标记为恶性的良性节点占全部良性节点的比例)。如表 6 所示,当阈值过小(如−0.1%)时,算法灵敏度过高,能够捕捉到更多潜在的恶意节点特征,从而提升恶意节点检测率,但同时也会将部分正常迭代的良性节点误判为恶意,导致良性节点误报率激增(0.54),影响模型正常运行。反之,当阈值过大(如−10.0%)时,算法判定标准更为严格,可将良性节点误报率降至 0,有效避免正常节点被误判,但会忽略部分恶意节点的异常特征,大幅降低恶意节点检测率,导致大量恶意节点漏检,削弱防御效果。综合来看,阈值 t 的选取需兼顾检测精度与误报控制。通常,阈值应略高于正常波动范围。

表 6　不同阈值 t 取值下的灵敏度

Table 6　Sensitivity under different threshold t values

t	恶意节点检测率	良性节点误报率
−0.1%	1	0.54
−1.0%	1	0
−3.0%	0.833	0
−5.0%	0.533	0
−10.0%	0.200	0

3　结论

本文提出的 FL-RIDS 能显著提升公共数据授权运营场景的入侵检测性能。首先,引入有效样本数加权的入侵检测损失函数 ESLoss,在维持多数类

检测性能的同时,改善类别不平衡问题。其次,针对数据异构性,设计服务端评估聚类联邦学习方法SEC,缓解非独立同分布数据导致的性能下降。最后,针对安全威胁,设计轻量级中毒检测算法 LPD,以低计算开销识别并隔离恶意节点。实验结果表明,ESLoss 与 CE 相比,*f*1*score* 最高提升了 18.9 百分点;相对 FedAvg,SEC 方法实现了最高 5.9 百分点的 *f*1 提升;LPD 算法能够以较低计算开销实现相对于 FedAvg 最高 68.4 百分点的 *f*1*score* 提升,为车联网等公共数据授权运营平台提供一种兼顾隐私保护、协同检测和鲁棒防御的入侵检测方案。但是本文仍存在一定局限性:SEC 聚类效果依赖公共评估数据的覆盖度与质量,LPD 检测性能受阈值设置影响,需要结合场景风险偏好调整参数。未来将围绕公共评估数据优化、自适应阈值设计及通信计算开销压缩,探索更轻量通用的联邦入侵检测方案。

参考文献:

[1] Wei Mingjun, Li Feng, Liu Yazhi, et al. An intrusion detection method for Internet of Vehicles based on improved WGAN-GP and ResNet[J]. Journal of Zhengzhou University (Engineering Science), 2024, 45(4): 30-37. [魏明军, 李凤, 刘亚志, 等. 基于改进 WGAN-GP 和 ResNet 的车联网入侵检测方法[J]. 郑州大学学报(工学版), 2024, 45(4): 30-37.]

[2] Ileberi E, Sun Yanxia. Advancing model performance with ADASYN and recurrent feature elimination and cross-validation in machine learning-assisted credit card fraud detection: a comparative analysis[J]. IEEE Access, 2024, 12: 133315-133327.

[3] Han Hui, Wang Wenyuan, Mao Binghuan. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[M]// Advances in Intelligent Computing. Berlin: Springer, 2005: 878-887.

[4] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[5] Dina A S, Siddique A B, Manivannan D. A deep learning approach for intrusion detection in Internet of Things using focal loss function[J]. Internet of Things, 2023, 22: 100699.

[6] Sheikh S A, Sahidullah M, Hirsch F, et al. Advancing stuttering detection via data augmentation, class-balanced loss and multi-contextual deep learning[J]. IEEE Journal of Biomedical and Health Informatics, 2023, 27(5): 2553-2564.

[7] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]// Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale: AISTATS, 2017: 1273-1282.

[8] Shan Yao, Yao Yu, Zhou Xiaoming, et al. CFL-IDS: an effective clustered federated learning framework for industrial Internet of Things intrusion detection[J]. IEEE Internet of Things Journal, 2024, 11(6): 10007-10019.

[9] Yi Liping, Wang Gang, Liu Xiaoguang, et al. FedGH: heterogeneous federated learning with generalized global header[C]// Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 8686-8696.

[10] Ghosh A, Chung J, Yin Dong, et al. An efficient framework for clustered federated learning[J]. IEEE Transactions on Information Theory, 2022, 68(12): 8076-8091.

[11] Han Weixiang, Peng Jialiang, Yu Jiahua, et al. Heterogeneous data-aware federated learning for intrusion detection systems via meta-sampling in artificial intelligence of things[J]. IEEE Internet of Things Journal, 2024, 11(8): 13340-13354.

[12] Duan Moming, Liu Duo, Ji Xinyuan, et al. FedGroup: efficient federated learning via decomposed similarity-based clustering[C]// Proceedings of the 2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom). Piscataway: IEEE, 2021: 228-237.

[13] Li Henger, Sun Xiaolin, Zheng Zizhan. Learning to attack federated learning: a model-based reinforcement learning attack framework[C]// 36th Conference on Neural Information Processing Systems. New Orleans: NeurIPS, 2022: 35007-35020.

[14] Baruch G, Baruch M, Goldberg Y. A little is enough: circumventing defenses for distributed learning[C]// 33rd Annual Conference on Neural Information Processing Systems. Vancouve: NeurIPS, 2019, 32: 1-10.

[15] Cao Xiaoyu, Fang Minghong, Liu Jia, et al. FLTrust: Byzantine-robust federated learning via trust bootstrapping [PP/OL]. V3. arXiv(2022-04-12)[2026-09-09]. https://arxiv.org/abs/2012.13995.

[16] Yin Dong, Chen Yudong, Kannan R, et al. Byzantine-robust distributed learning: towards optimal statistical rates[PP/OL]. V2. arXiv(2021-02-25)[2026-09-09]. https://arxiv.org/abs/1803.01498.

[17] Cui Yin, Jia Menglin, Lin T Y, et al. Class-balanced loss based on effective number of samples[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:

IEEE, 2019: 9260-9269.
[18] Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics, 1987, 20: 53-65.
[19] Seo E, Song H M, Kim H K. GIDS: GAN based intrusion detection system for in-vehicle network[C]//Proceedings of the 2018 16th Annual Conference on Privacy, Security and Trust (PST). Piscataway: IEEE, 2018: 1-6.
[20] Rosay A, Carlier F, Leroux P. MLP4NIDS: an efficient MLP-based network intrusion detection for CICIDS2017 dataset[M]//Machine learning for networking. Cham: Springer, 2020: 240-254.
[21] Tuysuzoglu G, Dogan Y, Ozturk Kiyak E, et al. Joint Tomek links (JTL): an innovative approach to noise reduction for enhanced classification performance[J]. IEEE Access, 2025, 13: 123059-123082.
[22] Tan Jingru, Li Bo, Lu Xin, et al. The equalization losses: gradient-driven training for long-tailed object recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 13876-13892.
[23] Li Tian, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks[C]//Proceedings of Machine Learning and Systems. 2020, 2: 429-450.
[24] Thein T T, Shiraishi Y, Morii M. Personalized federated learning-based intrusion detection system: poisoning attack and defense[J]. Future Generation Computer Systems, 2024, 153: 182-192.

Federated Learning-based Robust Intrusion Detection System for Public Data Authorized Operations

WEI Lin[1], CHEN Chao[1], HAN Ding[2], LI Qiuyan[2], HE Mengyang[1]

(1. School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450001, China; 2. Economic and Technical Research Institute of State Grid Henan Electric Power Company, Zhengzhou 450052, China)

Abstract: While public data authorized operation enables multi-party data sharing and reuse, it confronts malicious intrusion threats targeting public data, making reliable intrusion detection technologies imperative to guarantee data security and privacy. Traditional federated learning-based intrusion detection technologies have three major limitations: severe class imbalance of intrusion detection data leading to high false positive rates for minority intrusion samples; data heterogeneity across nodes exacerbating deviations between local and global models and slowing down model convergence; and poisoning attacks from malicious nodes undermining model training and degrading the detection performance of benign nodes. A federated learning-based robust intrusion detection system (FL-RIDS) for public data authorized operation was proposed. Firstly, an effective sample number weighted loss function (ES-Loss) for intrusion detection was designed to dynamically adjust class weights and significantly improve the model's sensitivity to few-sample intrusion categories. Secondly, a model performance evaluation-based clustering method (SEC) was proposed. It evaluated the performance of local models of all participant nodes on the cloud server and clusters nodes with similar data features, effectively mitigating performance degradation and slow convergence caused by data heterogeneity. Finally, a lightweight poisoning detection algorithm (LPD) was proposed to enhance system robustness against malicious attacks. It dynamically identified malicious nodes launching poisoning attacks during training by monitoring the similarity variation of update directions between local node models and cluster global model classifiers. Experimental results showed that in CICIDS2017 and CAN-intrusion-dataset, ESLoss improved the f1score by 18.9 and 3.2 percentage points respectively at the highest level compared to the cross-entropy loss function. The SEC method improved the f1score by 4.9 and 5.9 percentage points respectively at the highest level compared to FedAvg with excellent anti-poisoning performance. In addition, LPD improved the f1score by 54.2 and 68.4 percentage points respectively at the highest level compared to FedAvg, which could effectively resist poisoning attacks launched by malicious nodes in the middle of model training.

Keywords: clustered federated learning; public data; intrusion detection; poisoning attack defense; data heterogeneity; class imbalance