DOI:10.11918/202504085

隐私增强的安全联邦入侵检测方法

金志刚[1],丁 禹[1],武晓栋[2,3],陈旭阳[1]

(1.天津大学 电气自动化与信息工程学院,天津 300072;2.内蒙古工业大学 新能源学院,内蒙古 鄂尔多斯 017010;
3.内蒙古自治区新能源与储能技术重点实验室,呼和浩特 010051)

摘 要:入侵检测系统(intrusion detection system,IDS)面临着生成式模型逆向攻击的安全考验,而对于联邦式IDS,联邦GAN(generated adversarial network)攻击是其极为典型的数据安全威胁。为提升联邦式IDS的数据隐私安全,本研究提出通用的隐私增强的安全联邦入侵检测方法(privacy-enhanced federated intrusion detection,PEFID),并在多样化的攻防对抗仿真中验证其性能。PEFID从特征层面和模型层面共同增强数据隐私。在特征层面,提出改进的自适应隐私增强模块调整表征学习的泛化程度,权衡隐私保护与任务学习。此外,向中间层隐变量注入可控扰动,进一步弱化梯度的可追踪性。在模型层面,提出结合预测置信度的标签平滑策略以应对标签反转。各节点可根据预测置信度个性化调整软标签值,赋予受害者数据更加宽容的软标签值以阻止攻击深入。CICIDS2018和UNSW-NB15数据集上的验证实验表明:在多种网络场景中,PEFID均可有效防御联邦GAN攻击;与其他防御方案相比,PEFID能够在可控的时间复杂度下实现隐私与性能间的平衡;即使在单点防御失效时,PEFID仍能够保持优秀的防御效用。本文所提方法兼具通用性与轻量化,可适配于现有的联邦式入侵检测系统,以极小的性能代价显著增强数据隐私。

关键词:入侵检测系统;联邦学习;深度学习;模型逆向攻击;隐私保护
中图分类号:TP393.08 **文献标志码**:A **文章编号**:0367-6234(2026)05-0025-08

A privacy-enhanced secure federated intrusion detection method

JIN Zhigang[1], DING Yu[1], WU Xiaodong[2,3], CHEN Xuyang[1]

(1. School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China;
2. School of Renewable Energy, Inner Mongolia University of Technology, Ordos 017010, China;
3. Inner Mongolia Key Laboratory of New Energy and Energy Storage Technology, Hohhot 010051, China)

Abstract: Intrusion detection systems (IDS) face security challenges of generative model inversion attacks. And Federated GAN Attacks are the particularly characteristic data security threat to federated IDS. To improve data privacy in federated IDS, a universal privacy-enhanced federated intrusion detection (PEFID) method is proposed and is validated in diverse attack-defense simulation scenarios. PEFID jointly enhances data privacy at both the feature level and the model level. From the feature level, an improved adaptive privacy enhancing module is proposed to adaptively adjust the regularization degree of representation learning to balance privacy protection and task learning. Besides, controllable perturbations are injected into the hidden variables to further degrade the traceability of the gradient. From the model level, a label smoothing strategy combined with prediction confidence is proposed to deal with label inversion. Each client can individually adjust the soft label value according to the prediction confidence, assigning victim data a more lenient soft label value to mitigate the consistent attack. Experimental results on the CICIDS2018 and UNSW-NB15 datasets show that PEFID can effectively resist federated GAN attacks in various network scenarios. Compared with other methods, PEFID can better balance privacy and performance with controllable time complexity. It can still maintain superior defensive efficacy even in the case of single point penetration. The proposed method is both universal and lightweight, which can be adapted to existing federated IDS to significantly enhance data privacy with minimal performance cost.

Keywords: intrusion detection system; federated learning; deep learning; model inversion attack; privacy protection

“数字中国”战略的推进使得网络设备与网络数据爆发式增长,愈加猛烈的网络攻击也随之发生。入侵检测系统(intrusion detection system, IDS)可以实时监测和分析网络流量,识别潜在的攻击行为并

收稿日期:2025-04-30;**录用日期**:2025-06-04;**网络首发日期**:2025-07-24
网络首发地址:https://link.cnki.net/urlid/23.1235.t.20250723.1553.002
基金项目:国家自然科学基金(52471364)
作者简介:金志刚(1972—),男,教授,博士生导师
通信作者:金志刚,zgjin@tju.edu.cn

及时响应应急措施[1],是现代网络防御的重要一环。融合深度学习(deep learning,DL)的DL-IDS具有更强的特征捕捉与识别能力,极大提升了检测性能和检测速度[2]。在此基础上,联邦学习(federated learning,FL)[3]有效解决了DL-IDS的训练数据稀缺问题,进一步拓展了其应用前景。目前,以基于FL的DL-IDS为范式的协同入侵检测系统广泛应用于工业物联网[4]、车联网[5]等分布式网络,以提升其整体安全防护能力。

作为安全网关,IDS是流量必经之处,持有丰富的数据信息。因此,越来越多的网络攻击将矛头直指IDS本身,以窃取其拥有的数据隐私信息[6]。虽然FL-IDS避免了数据传输导致的直接泄露风险,但隐含在模型梯度中的数据信息仍有被敌手利用的可能。模型逆向攻击(model inversion attack,MIA)[7]便可通过窃取模型参数或联邦通信中的梯度信息逆向推理训练数据,不断优化拟造样本以重构目标数据[8]。其中,攻击性更强大的生成式模型逆向(generative model inversion,GMI)攻击[9]无需预训练就可以重构出更加准确的拟造样本。Hitaj等[10]提出的联邦GAN攻击(federated GAN attacks)是针对联邦学习的一类典型GMI攻击,其攻击者伪装成一个正常客户端加入联邦学习,在不用付出任何代价的条件下便能在本地重构目标数据。随着联邦学习的进行,其重构样本便能持续提升对于目标样本的拟合能力。

联邦GAN攻击是基于深度学习的FL-IDS面临的深度隐私推理攻击,然而目前的防御方案对IDS的安全需求与联邦GAN攻击的内在特征考量不足。首先,隐私、性能与效率间的矛盾是隐私保护的核心问题,有全时、实时、精确响应需求的IDS尤其重视其三元平衡。然而,目前的防御方案无法完全适配安全联邦IDS的设计需求。多数研究聚焦于GMI防御中的安全多方计算[11]、同态加密[12]和差分隐私(differential privacy,DP)[13]等技术。安全多方计算涉及由物理层至应用层的一系列安全协议,安全成本巨大,其在FL-IDS中的应用缺乏可行性;同态加密虽能提供优秀的隐私保护,但其复杂的加解密机制同样会带来大量的计算开销,对于时间敏感的IDS并不可行;组级DP[14]通过对梯度或模型加噪提升隐私,但以严重的性能受损为代价,需严格控制隐私预算。例如,高媛等[15]提出一种差分隐私联邦方法DP-FLAGD,结合本地视角与全局视角自适应分配隐私运算以提升系统性能。

其次,联邦GAN攻击有其独有特点,简单的泛化训练防御能力有限。虽然泛化训练控制是直接且有效的对抗MIA的手段[16],但联邦GAN攻击可通过独有的标签反转机制加剧受害者的本地过拟合,削弱其隐私保护。Scheliga等[17]设计了一个通用的隐私增强模块(privacy enhancing module,PRECODE),利用信息瓶颈的思路筛除数据的冗余特征,保留核心特征。在计算机视觉领域,Luo等[18]基于生成对抗网络(generative adversarial networks,GANs)重构训练图像,在保留分类特征的同时模糊视觉特征,有效减弱MIA的攻击性。马睿[19]提出一种基于小样本扩充的联邦学习数据保护方案,利用对抗训练的思路增强IDS的泛化能力。Struppek等[20]发现负标签样本平滑可有效防御基于样本优化的MIA,拓展了标签平滑的应用思路。

综上,为提升现有FL-IDS对于联邦GAN攻击的鲁棒性,本文提出隐私增强的安全联邦入侵检测方法(privacy-enhanced federated intrusion detection,PEFID),分别从数据特征层面与网络模型层面增强入侵检测系统隐私,避免敏感信息泄露。在数据特征层面,提出改进的自适应隐私增强模块模糊表征学习,根据历史训练信息自适应调整特征泛化程度,并通过可控扰动注入弱化梯度优化过程的可追踪性。在网络模型层面,设计结合预测置信度的标签平滑策略泛化模型训练。各节点可根据模型的预测置信度个性化调整软标签,受攻击的标签将被赋予更加宽容的软标签值以阻止攻击的进一步深入。最后,结合多场景的攻防模拟实验探究本文所提方法的有效性和鲁棒性。

1 威胁模型

为方便分析,考虑一个仅有双方参与的联邦场景(图1),其中攻击者A拥有$\{a,b\}$类数据,受害者V拥有$\{a,c\}$类数据。攻击者试图窃取其未拥有的c类数据,联邦GAN攻击流程如下。

1)联邦学习预期执行。当达到预定通信轮次或全局模型性能满足某一阈值时,转至下一步。

2)攻击者A启动本地预部署的GANs。鉴别器D拷贝收到的全局模型参数$\boldsymbol{\omega}_{\mathrm{g}}$,并依此训练本地生成器G。

3)生成器生成一批次c类数据,并将数据标签设置为b。

4)生成数据与A的原始数据共同训练本地模型,上传攻击者模型参数$\boldsymbol{\omega}_{\mathrm{A}}$至服务器。

5)受害者V正常训练并上传模型参数$\boldsymbol{\omega}_{\mathrm{V}}$至服务器。服务器聚合最新的全局模型,并广播至所有节点。

6)若未达到联邦学习最大通信轮次,跳转至2)。

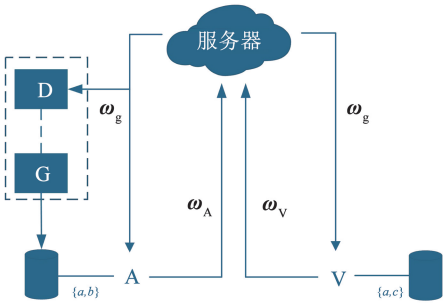

图 1　联邦 GAN 攻击流程

Fig. 1　Attack process of federated GAN attacks

上述联邦 GAN 攻击流程可推广至任意多实体联邦学习框架中。作为联邦学习的半诚实参与者，攻击者拥有网络结构、数据标签的全部信息，隶属于白盒攻击。联邦 GAN 攻击的成功依赖于全局模型中隐含的数据信息，这也是深度学习模型的普遍困扰。此外，攻击步骤 3）中的标签反转会持续诱导受害者模型过拟合，用以不断促使其提供更加精细的数据特征。对抗生成技术允许生成器在更广泛的特征空间内搜寻目标向量，而无需公开数据集的预训练或是对目标类的先验知识[21]。

分析可知，联邦 GAN 攻击的防御有两个关键。1）干扰全局模型的输出反馈，破坏生成器训练收敛。敌手生成器依靠全局模型的反馈优化训练，受干扰的输出反馈可提供错误的梯度下降方向阻止生成器收敛。2）遏制标签反转的负面影响，阻止隐私的持续泄露。标签反转诱使受害节点过拟合，导致数据特征不断精细。遏制标签反转的影响以阻止攻击的持续深入。

2　PEFID 整体设计

2.1　自适应隐私增强模块

为防止联邦学习中的梯度泄露，Scheliga 等[17]提出即插即用的隐私增强模块 PRECODE，通过对数据特征变分建模以模糊梯度中蕴含的原数据信息。PRECODE 的核心思想是通过引入信息瓶颈（information bottleneck，IB）使信息传递过程中只有最重要的信息能够通过。具体说来，对于输入向量 $\boldsymbol{x}$ 和输出向量 $\boldsymbol{y}$，信息瓶颈理论寻找中间隐变量 $\boldsymbol{z}$ 使得

$$\max[I(\boldsymbol{x};\boldsymbol{z})-\beta I(\boldsymbol{z};\boldsymbol{y})] \tag{1}$$

式中：β 为权重系数，$I(\cdot)$ 为互信息熵。在神经网络中互信息熵难以计算，可利用变分自编码器构建变分瓶颈以近似信息瓶颈，其优化目标为

$$\min L(\theta)=L_{\text{task}}(\boldsymbol{y},\hat{\boldsymbol{y}})+\beta\cdot D_{\text{KL}}(N(\mu_z,\sigma_z),N(0,1)) \tag{2}$$

式中：θ 为神经网络参数，$\hat{\boldsymbol{y}}$ 为预测向量，L_{task} 为任务损失，D_{KL} 为隐变量 $\boldsymbol{z}$ 与标准正态分布间的 KL 散度，用来衡量 $\boldsymbol{z}$ 的泛化程度。PRECODE 允许模型捕捉和保留有用的特征而抑制无关的噪声。对于网络流量，协议、时序、统计等元数据信息会被重点关注，而数据包内容将被相对模糊。

β 控制着 PRECODE 的泛化程度，权衡着性能与泛化间的关系，需要根据对任务的先验知识进行调整，无法适应动态变化的网络环境。此外，固定的 β 缺乏任务弹性，无法在学习过程中灵活调整。因此，本文提出改进后的自适应隐私增强模块（adaptive PRECODE，A-PRECODE），可面向不同的任务自适应调整权重系数，其结构如图 2 所示，其中 CE(·) 为交叉熵损失函数。首先，在原基础上添加自适应的权重调整策略，根据历史训练情况动态调整 β，以平衡数据隐私与模型性能间的关系，即

$$\beta^t=\left(1+\eta\frac{L_{\text{task}}^{t-1}-L_{\text{task}}^{t}}{L_{\text{task}}^{t}}\right)\cdot\beta^{t-1} \tag{3}$$

式中：η 为调整步长，t 为本地训练轮次。在任务性能优秀时，A-PRECODE 更加关注隐私问题，进而提升模型泛化程度；在任务表现差时，其将着力于任务学习。此外，A-PRECODE 额外向重构隐变量 $\boldsymbol{z}$ 注入可控的微小噪声，进一步弱化梯度的可追踪性，干扰敌手生成器的反馈训练以增强隐私，即

$$\boldsymbol{z}=\mu_z+\boldsymbol{\varepsilon}\cdot\sigma_z+\boldsymbol{n} \tag{4}$$

式中：$\boldsymbol{\varepsilon}\sim N(0,1)$，$\boldsymbol{n}\sim N(0,s^2)$ 为额外的高斯扰动，s^2 为控制注入噪声的强度。

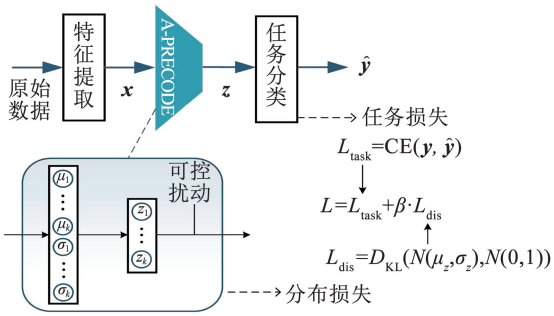

图 2　A-PRECODE 结构

Fig. 2　Structure of A-PRECODE

2.2　结合预测置信度的标签平滑策略

标签平滑是一项正则化技术，通过平滑硬标签提升模型泛化能力，避免过拟合。有研究证明[16]，采用独热标签训练会导致模型对训练数据过拟合，数据信息将隐含在模型参数中，更容易遭受推理攻击。对于类别数为 M 的硬标签 $\boldsymbol{y}_i=\mathbf{e}_i\in\mathbb{R}^M$（$M$ 维空间的标准基向量），标签平滑结果为

$$\boldsymbol{y}^{\text{LS}}=\left[1-\frac{M}{M-1}\alpha\right]\cdot\boldsymbol{y}_i+\frac{\alpha}{M-1}\cdot\mathbf{1},\mathbf{1}\in\mathbb{R}^M \tag{5}$$

式中:$\alpha \in [0,1]$为平滑系数。

联邦 GAN 攻击的攻击者通过标签反转诱导受害者不断提供样本信息,促使其对于训练数据的过拟合。本文提出结合模型预测置信度的标签平滑策略,通过全局模型在本地数据集的预测置信度向量 $\boldsymbol{C} \in \mathbb{R}^M$ 周期性调整软标签向量,具体为

$$\boldsymbol{y}_t^{\mathrm{LS}} = \left(\cdots, \frac{[1 - \Delta \boldsymbol{C}(i)] \cdot \boldsymbol{y}_{t-1}^{\mathrm{LS}}(i)}{\sum_{i=1}^{M} [1 - \Delta \boldsymbol{C}(i)] \cdot \boldsymbol{y}_{t-1}^{\mathrm{LS}}(i)}, \cdots \right)^{\mathrm{T}} \in \mathbb{R}^M \tag{6}$$

式中,$\Delta \boldsymbol{C} = \boldsymbol{C}_{t-1} - \boldsymbol{C}_t$ 为预测置信度差分。结合预测置信度的标签平滑策略允许每个客户端对其拥有的每类标签独立计算平均置信度,拥有独立的软标签向量。对于受害者,个性化标签平滑允许适当增加被执行反转标签的样本的软标签值以降低本地过拟合,防止隐私泄露的不断深入。为降低计算开销,每 5 轮全局通信,执行一次标签更新。结合预测置信度的标签平滑策略实施细节见如下伪代码。其中:节点 k 的数据总条数为 n_k,标签为 i 的数据条数为 n_k^i,其构成的数据子集为 D_k^i。

伪代码:结合预测置信度的标签平滑策略

输入:训练轮次 T,客户端数量 N,数据集 $D_{k \in N} = \{\boldsymbol{x}_r, \boldsymbol{y}_r\}^{r \in n_k}$,客户端模型 $\boldsymbol{\omega}_k$,历史标签平滑向量 $\boldsymbol{y}_{t-1,k}^{\mathrm{LS}}$,历史预测置信度向量 $\boldsymbol{C}_{t-1}^{k,i}$

输出:更新后的标签平滑向量 $\boldsymbol{y}_{t,k}^{\mathrm{LS}}$,更新后的预测置信度向量 $\boldsymbol{C}_t^{k,i}$

For t in T:
  For 所有客户端 **parallel do**:
    模型进入测试模式
    For i in M:
      $\boldsymbol{C}_t^{k,i} = \frac{1}{n_k^i} \sum_{i \in M}^{n_k^i} \mathrm{Softmax}(\boldsymbol{\omega}_k, D_k^i)$
    End for
    由式(6)计算更新后的 $\boldsymbol{y}_{t,k}^{\mathrm{LS}}$
  End for
End for
Return $\boldsymbol{y}_{t,k}^{\mathrm{LS}}, \boldsymbol{C}_t^{k,i}$

3 实验概况

3.1 数据集

本文选择适用于联邦式 IDS 的 CICIDS2018[22] 和 UNSW-NB15[23] 入侵检测数据集。由于本文仅研究数据隐私问题,考虑数据重构的可能性,因此并不关注数据本身的良恶性。CICIDS2018 数据集共包含 7 类数据,对其中数据量较多的 Benign、Infiltration、Dos、Bot 与 Brute-Force 类作欠采样以保持较为平衡的数据分布,用于模拟一个节点一类数据的简单情形。此时,联邦 GAN 攻击的受害者节点仅有 1 个。UNSW-NB15 数据集共有 10 类数据,本文使用官方发布的 UNSW-NB15-training 训练集与 UNSW-NB15-testing 测试集。同样,本文舍弃其中数据量较少的后 4 类数据:Analysis、Backdoor、Shellcode 和 Worms。UNSW-NB15 数据集用于模拟实际网络中数据量不平衡的情形。具体数据分布详见表 1。

表 1 数据集样本分布

Tab. 1 Distribution of dataset samples

数据集	标签	类别	训练集/条	测试集/条
UNSW-NB15	0	Normal	50 000	37 000
	1	Generic	40 000	18 871
	2	Exploits	33 393	11 132
	3	Fuzzers	18 184	6 062
	4	Dos	12 264	4 089
	5	Reconnaissance	10 491	3 496
CICIDS2018	0	Benign	23 852	5 123
	1	Infiltration	23 989	6 010
	2	Dos	24 127	5 872
	3	Bot	24 127	5 932
	4	Brute-Force	24 067	5 998
	5	DDos	1 641	434
	6	Web	568	164

原始数据特征需经最大最小归一化映射至[0,1]。其中,UNSW-NB15 数据集样本共 46 个特征,需剔除其中的 3 个非数值特征:proto、service 和 state。而 CICIDS2018 数据集样本的 80 个特征均为数值特征,无需额外处理。

3.2 实验环境与评价指标

实验在 Windows 11 操作系统,Python3.8 和 PyTorch1.13.0 版本下进行。实验硬件环境为 CPU:12th Gen Intel(R) Core(TM) i7-12700H 2.30 GHz,GPU:NVIDIA GeForce RTX 3060 LapTop。

本文评价指标有检测准确率 *Acc*(accuracy)、攻击准确率 *Raa*(attack accuracy rate)、性能下降率 *Rpd*(performance degradation rate)与响应时间 *Rt*(response time)。*Acc* 衡量 IDS 的检测性能,由深度学习混淆矩阵计算,即

$$\text{检测准确率} = \frac{T_{\mathrm{P}} + T_{\mathrm{N}}}{T_{\mathrm{P}} + F_{\mathrm{N}} + T_{\mathrm{N}} + F_{\mathrm{P}}} \tag{7}$$

式中:T_{P} 与 T_{N} 分别为真阳样本和真阴样本,即正确分类的恶意样本与良性样本,F_{P} 与 F_{N} 分别是假阳样本与假阴样本。*Raa* 评价联邦 GAN 攻击的成功程度,为重构样本被集中训练模式下的代理 IDS 模型正确分类的比例。代理模型的网络结构和训练条

件与实验所用一致。需要注意，Raa 仅从特征层面关注重构代表样本向量与真实样本向量间的相似性。样本层面的评估还需进一步的逆向特征工程，涉及敌手先验知识评估、域间关系约束等。但特征的差异增大也同样会增大样本复原的难度。Rpd 为隐私增强前后的模型性能下降程度，即

$$\text{性能下降率} = \frac{|\Delta \text{ 检测准确率}|}{\text{检测准确率}} \tag{8}$$

式中，分子为检测模型添加防御前后的检测准确率绝对差值。Rt 为完整训练一个全局轮次所需的平均时间，反映算法复杂度。

3.3　攻防设置

实验设置一个中央服务器和数个子节点。实验开始前，选择其中一个子节点为潜在的攻击节点，并指定攻击重构的目标类数据。在联邦学习进行至预设轮次的 50% 时，攻击节点发动联邦 GAN 攻击。每次实验，攻击节点有唯一的目标类重构数据。攻击开始后，攻击节点会生成 1 000 条重构数据并与本地数据集融合用于训练本地模型。重构数据集会在每个全局轮次中优先于本地训练更新。

每组实验设置与数据标签相匹配的节点数量，每个节点仅拥有一类流量数据，节点编号与其拥有的数据标签一致。基线入侵检测网络为多层感知机模型，其主体由 4 层全连接网络组成，神经元数量为[128，256，64，32]。自动编码器与解码器网络分别用于原始网络流量特征的转化输入与最终的决策分类。基线模型采用 SGD 优化器训练，学习率为 0.01，衰减因子为 0.998。敌手采用条件生成器 CGAN(conditional GAN)结构，通过一维卷积层拟合时序特征间的依赖关系，而后通过 3 层全连接层依次重构数据特征。生成器模型采用 Adam 优化器，学习率为 0.005。鉴别器模型接受自全局模型参数，无本地优化过程。实验的部分超参数设置为：权重系数 $\beta = 0.5$，调整步长 $\eta = 0.05$，噪声强度 $s^2 = 0.01$，标签平滑系数 $\alpha = 0.1$。

3.4　防御效用分析

为全面研究 PEFID 在多样的联邦场景、数据分布与攻防目标下针对联邦 GAN 攻击的防御效用，在 UNSW-NB15 与 CICIDS2018 数据集开展本实验，实验结果如表 2 所示。其中，基线表示未添加任何隐私保护策略，PEFID 表示添加本文所提出的隐私增强方法。

据表 2 可知，与基线相比，本文所提方法实现了显著的隐私提升与可控的性能下降。整体看出，在多种攻防场景中，PEFID 的攻击准确率 Raa 较基线模型明显下降，且性能下降率 Rpd 不高于 6.02%。这说明 PEFID 中的 A-PRECODE 模块能够根据不同的训练任务与难度，自动调整隐私程度，实现安全保护下的最好检测效果。观察细节，可以得出以下结论。1)联邦 GAN 攻击的成功多依赖于全局模型性能。以 UNSW-NB15 数据集为例，标签为 1 的流量数据量远多于 4 和 5，模型学习更加深刻，联邦 GAN 攻击的重构效果更好，即实验组 5-1(攻击节点 5，受害节点 1。若无特殊说明，为简化表达后文依此表示)的 Raa 明显高于 2-4 与 2-5。2)由于 CICIDS2018 的数据量更少，模型对于训练数据过拟合更加严重，联邦 GAN 攻击防御难度更大。3)联邦 GAN 攻击所执行的伪样本生成与标签反转会对全局模型收敛产生一定的负面影响。特别在 CICIDS2018 数据集的实验组 2-5 中，伪样本与原始样本的比例接近于 0.6，基线 Acc 降幅达 1%；而 UNSW-NB15 数据集样本量更大，联邦 GAN 攻击并未对全局模型收敛性能产生明显影响，攻击隐蔽性更强。此外，基线模型 3 个实验组的 Acc 均能够稳定收敛在51.16%，攻击难以察觉。

表 2　PEFID 防御效用结果

Tab. 2　Defensive efficacy results of PEFID

数据集	攻击节点	受害节点	Raa/%		Acc/%		Rpd/%
			基线	PEFID	基线	PEFID	
UNSW-NB15	5	1	63.44	6.62	**51.16**	48.59	5.02
	2	4	47.85	5.87		49.53	3.19
	2	5	37.87	0.36		50.13	2.01
CICIDS2018	6	1	77.21	12.17	63.38	59.82	5.62
	6	3	74.39	14.35	63.17	59.37	6.02
	2	5	39.75	6.82	62.57	60.65	3.07

为深入探究 PEFID 的隐私增强机制，本文在实验组 5-1 的基础上观察生成器训练损失与标签平滑效用。图 3 展示了敌手生成器自联邦 GAN 攻击启动后的训练损失情况，直观反映了联邦 GAN 攻击的失败。由图 3 可以看到，在基线模型中敌手生成器训练过程稳定，并于 30 轮次后(联邦学习第 80 轮

次)收敛。然而,受 PEFID 保护下的敌手生成器训练一直处于震荡状态且无法收敛至最佳。这是因为 A-PRECODE 模块改变了网络的表征学习,潜在变量的随机状态与可控扰动的叠加使得敌手生成器无法依据全局模型的反馈持续获得可信的梯度下降方向,从而破坏其训练。

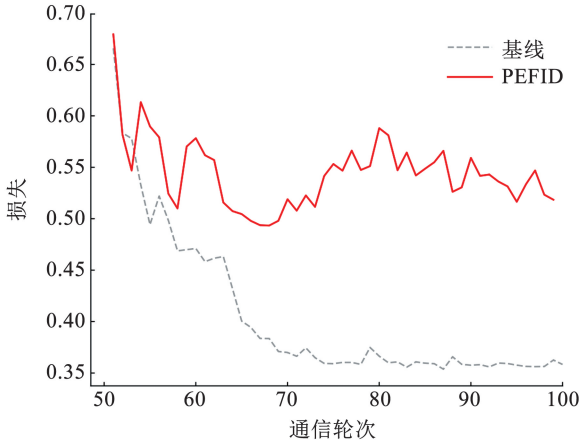

图 3 生成器损失曲线

Fig. 3 Generator loss curve

图 4 为实验组 5-1 训练结束时攻击节点与受害节点的标签平滑结果,纵轴经对数变换以突显差异。实验设置初始标签平滑系数 $\alpha = 0.1$,标签平滑值下限为 0.01。由图 4 可以看到,受害节点标签 5 的软标签值由初始的 0.02 上升至 0.16;对应地,目标类标签值由 0.9 降至 0.71。这表明为防止联邦 GAN 攻击标签反转诱使受害节点本地过拟合,结合预测置信度的标签平滑策略可动态分配各标签平滑值,为受害类与被反转类标签给予更加宽容的平滑值。

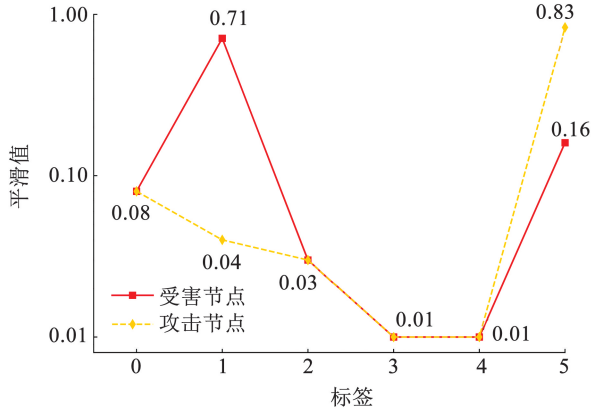

图 4 个性化标签平滑结果

Fig. 4 Result of personalized label smoothing

在实际网络中,流量多以非独立同分布形式存在,各节点数据在特征空间均有不同程度的重叠。此时任何拥有部分目标类数据的节点均为联邦 GAN 攻击的受害节点,且攻击节点可通过搭便车行为在自身无任何数据的情况下无代价发动联邦 GAN 攻击。为模拟在非独立同分布场景中敌手利用搭便车行为发动的联邦 GAN 攻击,本文开展如下攻防仿真:数据集 UNSW-NB15 经狄利克雷采样(采样参数为 1.0)[24],并分配给 6 个良性节点(编号 0 ~ 5),攻击节点 6 号无任何本地数据,通过搭便车行为加入联邦学习,在首轮全局通信中随机生成模型参数并上传。实验结果见表 3。

表 3 搭便车行为下的 PEFID 防御效用结果

Tab. 3 Defensive efficacy results of PEFID under free-riding

攻击节点	目标数据	*Raa*/%		*Acc*/%		*Rpd*/%
		基线	PEFID	基线	PEFID	
6	1	78.73	17.53	72.05	68.42	5.04
6	4	53.81	11.94	72.46	71.18	1.77

由表 3 结果可知,在非独立同分布场景中,联邦 GAN 攻击的威胁更大,防御难度也更大。这是因为单节点单数据的分布可视为最极端的非独立同分布场景,全局模型的聚合存在巨大困难。而在经狄利克雷采样后,全局模型聚合效果有所提升,联邦 GAN 攻击的成功率相应提升。且受害节点范围增加,提升了联邦 GAN 攻击的防御难度。然而,PEFID 仍能保持不错的性能表现,其 *Rpd* 仍与表 2 结果基本相当。这说明本文所提方法拥有优秀的通用性和适应性。面对多样化的联邦学习设置与攻防对抗,PEFID 可动态调整隐私保护强度,尽可能维持高效的检测性能,自主适配学习任务。

3.5 对比实验

本文对比了两类常用的防御方案,差分隐私与对抗训练,实验结果见表 4。其中,差分隐私为 OPACUS 开源库的 DPAVG 算法。作为组级差分隐私策略,DPAVG 为本地模型参数提供噪声保护。ε 为隐私预算,越小的预算拥有越好的隐私保护能力,而代价是更为严重的性能受损。对抗训练设置基于文献[19],于本地部署 GAN 执行对抗训练以扩充训练数据集,提升模型泛化能力。p 为对抗样本占原始样本的比例。对比实验使用数据集为 CICIDS2018,实验组为 6-1。

观察表 4 可知,本文所提 PEFID 在隐私增强和性能表现方面更加全能均衡,且额外引入的响应时间也在可接受范围之内。对比结果显示,DPAVG 在隐私预算 $\varepsilon = 0.1$ 时取得了最低的 *Raa*,然而代价是剧烈的性能受损,*Rpd* 跃升至 12.73%。在 $\varepsilon = 0.7$ 时 DPAVG 获得了较好的检测性能,但隐私保护能力却随之下降;对抗训练方案中,*Raa* 受 p 变化影响微小,而 *Rt* 和 *Acc* 变化更加敏感:一方面,随着样本量的增加,对抗训练所需的响应时间显著增加;另一

方面，对抗样本增多导致模型泛化程度愈高，*Acc* 反而呈下降趋势。

与之相比，PEFID 的最大优势是自适应性。DPAVG 与对抗训练均需要不断调试其参数以取得理想的隐私能力和检测性能。而凭借 A-PRECODE 模块的自适应特性，PEFID 可在学习过程中动态调整训练策略，权衡隐私和性能。通过设置结合预测置信度的标签平滑策略的更新周期，PEFID 的整体训练时间也可实现控制，以有限成本增强 FL-IDS 的隐私安全。

表 4　对比实验结果

Tab. 4　Results of comparison experiment

模型		*Raa*/%	*Acc*/%	*Rpd*/%	*Rt*/s
基线		77.21	63.38		26.3
DPAVG	$\varepsilon=0.1$	**9.17**	55.31	12.73	26.8
	$\varepsilon=0.7$	16.95	59.28	6.47	26.7
对抗训练	$p=0.2$	15.74	**61.42**	3.09	41.9
	$p=0.5$	14.65	57.81	8.79	**52.6**
PEFID		12.17	59.82	5.62	29.4

3.6　白盒场景防御实验

作为一种白盒攻击，联邦 GAN 攻击的防御方案势必需要考虑到敌手能力之内的破防手段。本文所提出的 PEFID 部署在节点本地，攻击者可对其结构与机制非法篡改以削弱防御。具体来说，对于 A-PRECODE模块，攻击者可忽略式(3)中的分布损失(即 $\beta=0$)，同时不添加额外扰动用于精细化本地学习；对于结合预测置信度的标签平滑策略，攻击者可拒绝执行软标签训练与个性化标签的动态更新，采用硬标签训练以最大化操纵受害节点的训练。在 CICIDS2018 数据集下，白盒场景中的攻防模拟结果如表 5 所示，Partial-PEFID 表示攻击者本地防御失效后的白盒攻击模型。

表 5　白盒场景防御实验结果

Tab. 5　Defense experiment results in white-box scenario

攻击节点	受害节点	模型	*Raa*/%	*Acc*/%	*Rpd*/%
		基线	77.21	63.38	
6	1	PEFID	12.17	59.82	5.62
		Partial-PEFID	13.53	59.68	5.83
		基线	39.75	62.57	
2	5	PEFID	6.82	60.65	3.07
		Partial-PEFID	7.41	60.33	3.58

由表 5 可知，与 PEFID 相比，Partial-PEFID 防御方案下的联邦 GAN 攻击成功率有一定上升，但 *Acc* 基本不受影响。在敌手破解本地防御后，其本地的 A-PRECODE 退化为简单的全连接层，失去了特征模糊效用，且硬标签训练无法阻止标签反转导致的类别混淆。但受 PEFID 保护的其他节点仍能高效防范隐私泄露。这是因为敌手依靠全局模型反馈输出优化生成器，而单一节点的防御漏洞不足以影响全局模型的隐私保护。全局模型参数依然由绝大多数节点经 A-PRECODE 模糊后的表征学习训练所得，且个性化的标签平滑策略仍能够在本地有效缓解过拟合。实验结果表明，即使遭遇敌手破除部分的本地防御，PEFID 在特征层面与模型层面的防御效用仍然出色，具备在开放联邦框架中的实用性。

4　结　论

本文通过深入研究联邦 GAN 攻击特点，提出了隐私增强的安全联邦入侵检测方法，以提升现有 FL-IDS 的数据隐私，得到如下结果。

1)联邦 GAN 攻击的威胁根源为模型的过拟合，标签反转是其关键手段。数据量、数据分布、模型聚合方式均会影响联邦 GAN 攻击的攻击强度。

2)本文所提出的隐私增强的安全联邦入侵检测方法 PEFID，可自主调整特征模糊和任务学习权重，权衡隐私保护和检测性能，具有通用性、鲁棒性和轻量化特点。

3)实验证明，PEFID 可有效提升现有联邦入侵检测系统的安全鲁棒性，增加敌手的数据复原难度。

参考文献

[1]武晓栋，金志刚，陈旭阳，等. 对抗学习辅助增强的增量式入侵检测系统[J]. 哈尔滨工业大学学报，2024，56(9)：31
WU Xiaodong，JIN Zhigang，CHEN Xuyang，et al. Adversarial learning-augmented incremental intrusion detection system[J]. Journal of Harbin Institute of Technology，2024，56(9)：31. DOI：10.11918/202403066

[2]金志刚，刘凯，武晓栋. 堆叠循环卷积和头部注意力的工业互联网入侵检测[J/OL]. (2024-06-24). https://afffgc1d129f57bb244a4hcc6ncwqfk90x6opofgfy.eds.tju.edu.cn/kcms/detail/23.1235.t.20240622.1013.002.html
JIN Zhigang，LIU Kai，WU Xiaodong. Stacked recurrent convolutions and head attention for industrial internet ntrusion detection[J/OL]. (2024-06-24). https://afffgc1d129f57bb244a4hcc6ncwqfk90x6opofgfy.eds.tju.edu.cn/kcms/detail/23.1235.t.20240622.1013.002.html

[3]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：PMLR，2017(54)：1273

[4]RUZAFA-ALCÁZAR P，FERNÁNDEZ-SAURA P，MÁRMOL-CAMPOS E，et al. Intrusion detection based on privacy-preserving federated learning for the industrial IoT[J]. IEEE Transactions on Industrial Informatics，2023，19(2)：1145. DOI：10.1109/TII.2021.3126728

[5] LI Yang, MOUBAYED A, SHAMI A. MTH-IDS: A multitiered hybrid intrusion detection system for internet of vehicles[J]. IEEE Internet of Things Journal, 2022, 9(1): 616. DOI:10.1109/JIOT.2021.3084796

[6] 肖雄, 唐卓, 肖斌, 等. 联邦学习的隐私保护与安全防御研究综述[J]. 计算机学报, 2023, 46(5): 1019
XIAO Xiong, TANG Zhuo, XIAO Bin, et al. Survey on privacy protection and security defense of federated learning[J]. Chinese Journal of Computers, 2023, 46(5): 1019. DOI:10.11897/SP.J.1016.2023.01019

[7] RIGAKI M, GARCIA S. Asurvey of privacy attacks in machine learning[J]. ACM Computing Surveys, 2024, 56(4): 1. DOI:10.1145/3624010

[8] LEWIS C, VARADHARAJAN V, NOMAN N, et al. Mitigation of gradient inversion attacks in federated learning with private adaptive optimization[C]//2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS). Jersey, NJ: IEEE, 2024: 833. DOI:10.1109/ICDCS60910.2024.00082

[9] ZHANG Yuheng, JIA Ruoxi, PEI Hengzhi, et al. The secret revealer: Generative model-inversion attacks against deep neural networks[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA: IEEE/CVF, 2020: 250. DOI:10.1109/CVPR42600.2020.00033

[10] HITAJ B, ATENIESE G, PEREZ-CRUZ F. Deep models under the GAN: Information leakage from collaborative deep learning[C]// Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: CCS, 2017: 603. DOI:10.1145/3133956.3134012

[11] XU Runhua, BARACALDO N, ZHOU Yi, et al. HybridAlpha: An efficient approach for privacy-preserving federated learning[C]// Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2019: 13. DOI: 10.1145/3338501.3357371

[12] ZHANG Li, XU Jianbo, VIJAYAKUMAR P, et al. Homomorphic encryption-based privacy-preserving federated learning in IoT-enabled healthcare system[J]. IEEE Transactions on Network Science and Engineering, 2023, 10(5): 2864. DOI:10.1109/TNSE.2022.3185327

[13] WANG Tianhao, ZHANG Yuheng, JIA Ruoxi. Improving robustness to model inversion attacks via mutual information regularization[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence. California: AAAI Press, 2021: 11666. DOI:10.1609/aaai.v35i13.17387

[14] 唐湘云, 王伟, 翁彧, 等. 联邦遗忘学习隐私安全与算法效率研究综述[J]. 计算机学报, 2025, 48(9): 2064
TANG Xiangyun, WANG Wei, WENG Yu, et al. A survey on privacy security and computation efficiency in federated unlearning[J]. Chinese Journal of Computers, 2025, 48(9): 2064. DOI:10.11897/SP.J.1016.2025.02064

[15] 高媛, 石润华, 刘长杰. 自适应差分隐私的联邦学习方案[J]. 智能系统学报, 2024, 19(6): 1395
GAO Yuan, SHI Runhua, LIU Changjie. Federated learning scheme with adaptive differential privacy[J]. CAAI Transactions on Intelligent Systems, 2024, 19(6): 1395. DOI:10.11992/tis.202306052

[16] AHMED F, SÁNCHEZ D, HADDI Z, et al. MemberShield: A framework for federated learning with membership privacy[J]. Neural Networks, 2025, 181: 106768

[17] SCHELIGA D, MÄDER P, SEELAND M. PRECODE-a generic model extension to prevent deep gradient leakage[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI: IEEE/CVF, 2022: 1849

[18] LUO Xinjian, ZHANG Xianglong. Exploiting defenses against GAN-based feature inference attacks in federated learning[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2025, 19(3): 1. DOI:10.1145/3719350

[19] 马睿. 基于联邦学习的协同入侵检测技术研究[D]. 哈尔滨: 哈尔滨工程大学, 2023
MA Rui. Research on collaborative intrusion detection technology based on federated learning[D]. Harbin: Harbin Engineering University, 2023

[20] STRUPPEK L, HINTERSDORF D, KERSTING K. Be careful what you smooth for: label smoothing can be a privacy shield but also a catalyst for model inversion attacks[PP/OL]. (2023-10-10)[2025-04-18]. https://doi.org/10.48550/arXiv.2310.06549

[21] 王冬, 秦倩倩, 郭开天, 等. 联邦学习中的模型逆向攻防研究综述[J]. 通信学报, 2023, 44(11): 94
WANG Dong, QIN Qianqian, GUO Kaitian, et al. Survey on model inversion attack and defense in federated learning[J]. Journal on communications, 2023, 44(11): 94. DOI:10.11959/j.issn.1000-436x.2023209

[22] THAKKAR A, LOHIYA R. A review of the advancement in intrusion detection datasets[J]. Procedia Computer Science, 2020, 167: 636

[23] MOUSTAFA N, SLAY J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)[C]//2015 Military Communications and Information Systems Conference (MilCIS). Canberra: IEEE, 2015: 1

[24] JIN Zhigang, DING Yu, WU Xiaodong, et al. Federated dual correction intrusion detection system: efficient aggregation for heterogeneous data[J]. Computer Networks, 2025, 259: 111116. DOI:10.1016/j.comnet.2025.111116

(编辑 吕雪梅)