一种抗拜占庭攻击的联邦学习鲁棒聚合算法

张淑芬[1,2,3], 李　涛[1,2,3], 张镇博[1,2,3], 钟　琪[1,2,3], 景忠瑞[1,2,3]

(1.华北理工大学 理学院,河北 唐山 063210;2.河北省数据科学与应用重点实验室(华北理工大学),河北 唐山 063210;3.唐山市数据科学重点实验室(华北理工大学),河北 唐山 063210)

摘　要:针对联邦学习中现有的防御方案在模型过滤时会过度剔除良性模型的问题,提出了一种抗拜占庭攻击的联邦学习鲁棒聚合算法 FLDBA。通过 HDBSCAN 密度聚类算法对模型进行聚类,识别出良性模型集合,并求取良性模型集合中方向最具代表性的模型作为可信模型。以可信模型为基准,利用余弦相似度对聚类结果中可能被误判为异常的良性模型进行筛选,实现对误判的修正。同时设立信誉机制,对模型历史行为进行动态评估,以降低漏判对系统的影响。对于信誉较高的模型,对模型幅值进行自适应缩放,并根据其更新质量赋予不同的聚合权重,提升模型的聚合效果。实验结果表明,在抵御符号翻转攻击时,FLDBA 的准确率比 FLRAM、FLAME、RFLPA、FLTrust 和 Krum 提升了 0.18 百分点~5.13 百分点,攻击成功率降低了 40.52 百分点~61.39 百分点,具有更好的鲁棒性。

关键词:联邦学习;拜占庭攻击;鲁棒性;信誉;加权聚合

中图分类号:TP309.2;TN92　　**文献标志码**:A　　**doi**:10.13705/j.issn.1671-6833.2026.04.012

联邦学习[1]作为一种新兴的分布式机器学习[2]框架,逐渐成为解决传统集中式学习隐私泄露[3]问题的重要方案。在联邦学习中,数据并不需要集中存储,而是保留在各个参与节点上,各方通过协作训练共享模型,有效降低了信息泄露[4-5]的风险。然而联邦学习的分布式特性也引入了新的安全隐患[6]。拜占庭攻击作为一种典型的安全威胁,利用恶意客户端向全局模型注入错误模型,破坏整个系统的稳定性。如何有效抵御拜占庭攻击,保障模型训练的安全性与鲁棒性,已成为当前研究的重要课题[7]。

近年来,研究者提出了多种防御算法[8]。例如 Blanchard 等[9]提出了 Krum 算法,计算各客户端模型之间的欧氏距离,并据此为每个模型分配得分,最终选取得分最高的模型作为聚合结果。在此基础上,Blanchard 等[9]又提出了 Multi-Krum,将多个得分较高的模型进行聚合作为全局模型,进一步提升了系统的鲁棒性。当攻击者精心构造模型以伪装成“距离合理”的恶意模型时,该类算法难以准确识别出良性模型。Cao 等[10]提出的 FLTrust 算法则在服务器端引入 1 个小规模干净数据集,训练出基准模型并作为信任参考。该算法根据客户端上传模型与基准模型之间的方向一致性计算信任分数,并据此加权聚合。FLTrust 的恶意模型筛选机制相对简单,对于具备一定隐蔽性和方向伪装能力的攻击模型,识别效果仍存在局限。Mai 等[11]提出了 RFLPA 算法,在 FLTrust 的基础上构建了兼容安全聚合协议的鲁棒联邦学习框架。该方法首先对客户端上传的模型进行幅值归一化,其次计算归一化模型与基准模型的余弦相似度,进而生成信任分数并用于加权聚合。此外,RFLPA 引入了可验证的打包式 Shamir 秘密共享机制与安全点积协议,实现了在不泄露本地模型的前提下完成鲁棒聚合。该方法仍然依赖服务器持有干净数据作为信任基准,同时加密计算实现较为复杂,实际部署成本较高。Nguyen 等[12]提出了 FLAME 算法,通过模型聚类、L2 范数剪裁与差分隐私加噪 3 种机制,实现对联邦学习中后门攻击的防御。该算法在聚类阶段保留的良性模型数量有限,剪裁操作未能充分利用幅值较小但方向正确的模型,噪声注入可能干扰良性模型所携带的有用信息,从而影响全局模型的性能。Chen 等[13]提出的 FLRAM 算法通过聚类等技术对模型的方向和幅值

收稿日期:2026-01-20;**修订日期**:2026-03-01

基金项目:国家自然科学基金联合基金资助项目(U20A20179)

作者简介:张淑芬(1972—),女,河北唐山人,华北理工大学教授,主要从事云计算、数据安全、隐私保护研究,E-mail:zhsf@ncst.edu.cn。

信息进行筛选,在多项实验中展现出较强的鲁棒性。将模型的符号与幅值分开筛选导致过度剔除客户端模型,最终仅有少数良性模型参与聚合,影响全局模型精度。此外,该算法通常依赖复杂的聚类和异常检测过程,计算开销较大,难以适应大规模或资源受限的联邦学习场景。雷诚等[14]提出了 umFL 算法,通过引入更新质量评估与信誉机制,提升了联邦学习在投毒攻击下的鲁棒性。umFL 依赖全局模型作为评估基准,当全局模型被攻击污染或初始化偏差较大时,可能导致相似度评估失效,进而影响聚合效果。

在上述研究中,主要存在 3 个方面不足:其一,防御算法在模型检测阶段要么检测方式单一,难以准确识别出恶意模型;要么检测标准过于激进,导致部分良性模型被误判剔除;其二,当恶意模型被漏判并参与聚合时,系统缺乏动态纠错机制,导致恶意梯度持续干扰模型更新;其三,现有模型聚合方法普遍依赖静态或单一的权重分配机制,未能充分考虑客户端模型在不同训练轮次中的动态差异及更新质量,导致聚合权重难以真实反映模型贡献度。

针对上述问题,本文提出了一种抗拜占庭攻击的联邦学习鲁棒聚合算法 FLDBA。首先,构建动态模型过滤模块,基于客户端上传模型的方向信息,引入 HDBSCAN 密度聚类算法对模型进行自适应聚类,识别最可信的良性模型集合;其次,以该集合在方向中最具代表性的模型作为可信基准,计算其余模型与该基准的余弦相似度,以识别更多潜在的良性模型,修正模型过滤可能造成的误判;再次,设计信誉评估机制,结合动态模型过滤结果,对客户端的长期信誉与近期表现进行加权评定,以平衡长期稳定性与短期动态性,减轻恶意模型因漏判持续干扰模型更新;最后,提出自适应缩放与基于更新质量的加权聚合策略,通过动态调整高信誉模型幅值并依据更新质量分配聚合权重,使聚合过程更能体现客户端的真实贡献,提升全局模型的收敛性能。

1 背景知识

1.1 联邦学习

最初的联邦学习形式可用 FedAvg[1]算法来表示,FedAvg 通过在客户端执行本地随机梯度下降进行模型训练,并在每轮通信后由服务器对多个客户端的模型更新进行平均。本地客户端的更新公式[1]为

$$\hat{g}_i = g_i - \eta \nabla F(g_i)。 \tag{1}$$

式中:g_i 为客户端 i 在第 t 轮的本地模型;η 为学习率;$\nabla F(g_i)$ 为客户端 i 上的局部损失函数的梯度。客户端执行 e 次本地更新后,将模型 $\hat{g}_i$ 发送给服务器。服务器聚合全局模型的公式[1]为

$$\hat{g} = \sum_{i=1}^{n} \frac{|D_i|}{|D|} \hat{g}_i。 \tag{2}$$

式中:n 为参与训练的客户端总数;$|D_i|$ 为客户端 i 的数据量,$|D| = \sum_{i=1}^{n} |D_i|$ 是所有客户端数据的总量。全局模型聚合完成后,服务器将新的全局模型 $\hat{g}$ 发送回客户端,进行下一轮训练。

1.2 拜占庭攻击

拜占庭将军问题最早由 Lamport 提出,该问题描述了在一个分布式环境中,如何在部分节点不可信的情况下,保证系统的一致性和正确性[15]。Shi 等[16]将拜占庭攻击定义为一种针对分布式系统的恶意行为,例如在联邦学习中,攻击者通过参与到系统中的部分客户端,以提交恶意或篡改的模型更新的方式,干扰全局模型的训练过程。与其他攻击方式相比,拜占庭攻击具有更高的灵活性与策略性,攻击者不仅可以单独实施攻击,还可通过多个恶意客户端的协同配合,进一步放大全局模型的破坏程度。

1.3 威胁模型

在训练过程中,各个客户端在其本地数据集上进行训练,得到本地模型。良性客户端按照正常流程上传模型至服务器,而拜占庭客户端则通常通过篡改模型方向或幅值的方式,扰乱模型的聚合过程。图 1 展示了拜占庭攻击者对模型方向的干扰,可以观察到,恶意模型更新的方向明显偏离多数良性模型的更新,这种偏离可能导致最终的聚合结果远离真实的最优解,严重影响模型性能。本文假设拜占庭客户端数量少于客户端总量的 50%,攻击者可通过协同通信获取其他客户端上传的模型更新,但无法访问服务器端的防御机制。

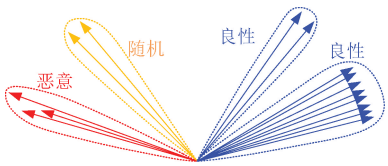

图 1 拜占庭攻击者对模型方向的干扰

Figure 1 Directional perturbation of models by Byzantine attackers

2 抗拜占庭攻击的联邦学习鲁棒聚合算法 FLDBA

2.1 动态模型过滤

FLDBA 首先在模型方向上进行恶意模型的筛

选。通过式(3)计算各个客户端模型的余弦相似度。考虑到余弦值的取值为[-1,1],值越接近1,两个模型方向越相近,而HDBSCAN密度聚类算法[12]依赖距离作为输入,距离越接近0,两个模型越相近,因此对余弦值进行转换,构造余弦距离矩阵 $\boldsymbol{O}$,其元素 o_{ij} 定义如式(4)所示[17]。

$$\cos(g_i,g_j)=\frac{\sum_{k=1}^{d} g_i^{(k)}\cdot g_j^{(k)}}{\sqrt{\sum_{k=1}^{d}(g_i^{(k)})^2}\cdot\sqrt{\sum_{k=1}^{d}(g_j^{(k)})^2}}。\tag{3}$$

$$o_{ij}=1-\cos(g_i,g_j)。\tag{4}$$

式中:o_{ij} 为矩阵 $\boldsymbol{O}$ 的元素,值越小,表示该模型方向越趋同于其他模型。

由于良性模型分布通常较为集中,而恶意模型往往表现出分布上的异常,因此可以使用HDBSCAN密度聚类算法将两者进行区分。HDBSCAN设定了一个最小簇参数 *size*,当某一簇中样本数小于 *size* 时,簇中的元素将被视为噪声。考虑到联邦学习中的恶意客户端数量一般少于 $\lfloor n/2\rfloor$,将其设置为 $\lfloor n/2+1\rfloor$。最终,选取样本数量最多的簇作为良性簇,并将簇中的模型保存至集合 C 中。

然而当部分良性客户端因数据分布差异导致模型方向偏移,或其模型方向与良性簇方向存在较大差异时,容易被误判为恶意客户端而被剔除。为降低模型检测时对良性模型的误判,求取良性簇中方向上最具代表性的模型作为可信模型 g_0,并以此为基准,将处于良性簇外,但与 g_0 方向较为接近的模型判定为良性模型。具体操作如下。

(1)设集合 C 的模型数量为 b。通过式(3)计算模型之间的余弦相似度,并以此作为元素 h_{ij} 构造余弦相似度矩阵 $\boldsymbol{H}$。

(2)对余弦相似度矩阵 $\boldsymbol{H}$ 按列进行归一化,得到归一化余弦相似度矩阵 $\overline{\boldsymbol{H}}$:

$$\overline{h}_{ij}=\frac{h_{ij}}{\sum_{k=1}^{b} h_{kj}}。\tag{5}$$

(3)计算每个客户端模型在该簇内的平均余弦相似度得分,具体计算公式为

$$\alpha_i=\frac{1}{b}\sum_{j=1}^{b}\overline{h}_{ij},\quad i=1,2,\cdots,b。\tag{6}$$

式中:α_i 可被视为"局部方向中心性"得分,数值越高代表该模型方向与簇内其他成员更加一致,可信程度越高。

(4)以 α_i 为加权系数,对簇内模型进行加权聚合,得到该簇的可信模型 g_0:

$$g_0=\sum_{i=1}^{b}\alpha_i\cdot g_i。\tag{7}$$

(5)记良性簇外的模型组成的索引集合为 C_{out},以可信模型 g_0 为基准,通过式(3)计算二者的余弦相似度 $\cos(g_i,g_0)$,若 $\cos(g_i,g_0)>\delta$,则认为该模型方向与可信模型方向较为相似,将其判定为良性模型,并保存到良性集合 B 中,即

$$B=C\cup\{g_i\mid\cos(g_i,g_0)>\delta\}。\tag{8}$$

本文在预实验阶段对 δ 的取值范围进行了敏感性分析,结果表明,当 δ 位于[0.3,0.5]时,模型在分类准确率与攻击抵御能力之间达到最优平衡。基于该实验结果,最终选取 $\delta=0.5$ 作为固定阈值。

动态模型过滤算法流程图如图2所示。

2.2 信誉机制

尽管FLDBA在筛选过程中能够保留更多的良性模型,提高全局模型的泛化性能,但其漏判风险也随之增加,部分伪装的恶意模型可能被判定为良性。并且在实际应用中,也可能面临更为复杂的情况,例如正常客户端在特殊情况下可能因受到恶意干预而上传异常模型,因此不应被长时间排除。同时,随着训练轮次的不断推进,部分高级攻击者可能会采用更隐蔽、更具有破坏性的策略,导致异常模型混入筛选结果中。针对这些问题,在服务器端设立信誉机制,根据模型的历史筛选结果对客户端进行信誉评估,选取信誉较优的客户端模型进行全局模型的聚合。定义第 t 轮客户端 i 的信誉值为 s_i^t,则第 $t+1$ 轮客户端 i 的更新公式为

$$s_i^{t+1}=s_i^t+\beta_i\cdot\Delta s_i^t。\tag{9}$$

$$\Delta s_i^t=\begin{cases}+1, & \text{若 } g_i\in B;\\ -1, & \text{若 } g_i\notin B。\end{cases}\tag{10}$$

式中:$\beta_i=1-\dfrac{|s_i^t-s_i^{t-1}|}{\max(s^t)}$,为当前轮次的信誉调整系数,通过比较客户端 i 近期的信誉波动,使信誉值较稳定的客户端获得更高的权重,而信誉波动较大的

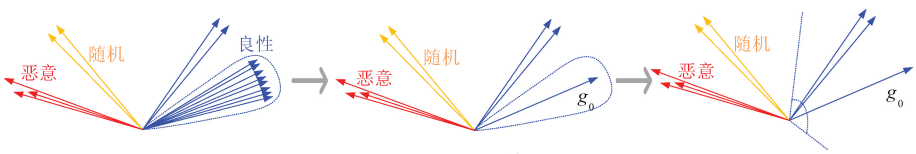

图2 动态模型过滤算法流程图

Figure 2 Workflow of the dynamic model filtering algorithm

客户端受到更强的衰减。这种机制使得系统能够稳定地保留历史信誉,避免因短期异常波动导致的误判,同时快速响应异常情况。为了防止信誉值的无限积累或过度下降,系统设计了信誉值的上限阈值 $s_{\max}=100$ 和下限阈值 $s_{\min}=-100$,信誉更新公式如下所示:

$$s_i^{t+1}=\begin{cases}s_{\max}, & 若\ s_i^t+\beta_i\cdot\Delta s_i^t>s_{\max};\\ s_{\min}, & 若\ s_i^t+\beta_i\cdot\Delta s_i^t<s_{\min};\\ s_i^{t+1}, & 其他。\end{cases}\tag{11}$$

为了使系统在聚合时能够更灵活地选取良性模型进行聚合,本文将良性集合 B 中元素的数量 $|B|$,作为最终进行模型聚合的数量。因此选取信誉评分最高的 $|B|$ 个模型,保存到集合 A 中,使得系统能够针对伪装的恶意模型及时作出应对反应。

2.3 自适应缩放

除通过操纵模型方向干扰模型聚合外,攻击者还通过改变模型幅值的方式干扰全局模型的聚合。考虑到中位数具备良好的抗极端值能力,将集合 A 中每个模型的幅值缩放至与其幅值中位数一致,降低幅值差异在聚合过程中的干扰。具体步骤如下。

(1)对集合 A 中的模型计算各个模型的幅值[18]:

$$\|g_i\|_2=\sqrt{\sum_{k=1}^{d}(g_i^{(k)})^2}。\tag{12}$$

(2)将集合 A 中良性客户端的幅值构成集合 G:

$$G=\{\|g_1\|_2,\|g_2\|_2,\cdots,\|g_{|A|}\|_2\}。\tag{13}$$

(3)取其中位数作为缩放边界 M:

$$M=\mathrm{Median}(G)。\tag{14}$$

(4)缩放各个客户端的模型幅值至模长为 M:

$$\widetilde{g}_i=\begin{cases}\dfrac{g_i}{\|g_i\|_2}\cdot M, & 若\|g_i\|_2>0;\\ g_i, & 若\|g_i\|_2=0。\end{cases}\tag{15}$$

2.4 加权聚合

考虑到不同客户端在训练过程中具有不同的更新质量得分,假设每个客户端本地都存在一个用来存放训练样本的数据集 D_i,则可用公式(16)来表示每个客户端的更新质量得分 U_i[14]:

$$U_i=|D_i|\sqrt{\frac{1}{|D_i|}\sum_{k\in D_i}L(k)^2}。\tag{16}$$

式中:$L(k)$ 为第 k 个样本在本地模型上的训练损失。

为了使得分更高的模型获得更高的聚合权重,通过以下步骤进行设置。首先,对 U_i 进行归一化处理,以消除不同得分之间的量纲差异:

$$u_i=\frac{U_i-U_{i,\min}}{U_{i,\max}-U_{i,\min}}。\tag{17}$$

其次,使用 Softmax 将 u_i 转换为聚合权重 w_i,计算公式如下:

$$w_i=\frac{\exp(u_i)}{\sum_{i=1}^{|A|}\exp(u_i)}。\tag{18}$$

最后,使用式(19)进行全局模型的加权聚合:

$$\hat{g}=\sum_{i=1}^{|A|}(\widetilde{g}_i\cdot w_i)。\tag{19}$$

3 实验

3.1 实验设置

实验使用 Python 3.9,在 Windows 10 系统中运行。CPU 为 Intel i3-12100,内存为 32 GB,显卡为 RTX 4060,机器学习框架采用 Pytorch。

采用 Fashion-MNIST[19](简称为 F-MNIST)和 CIFAR-10[20]数据集进行实验。Fashion-MNIST 数据集包含 70 000 个样本,其中 60 000 个样本用于训练,10 000 个样本用于测试。CIFAR-10 数据集由 60 000 张彩色图像组成,其中 50 000 张用于训练,10 000 张用于测试。实验训练数据采用独立同分布划分方式以验证算法的有效性,训练样本随机均匀分配至 50 个客户端,各客户端所持数据量基本一致,整体数据分布保持平衡。

采用卷积神经网络[21]对数据集进行分类。针对 F-MNIST 数据集,模型由 3 组卷积层与池化层交替组成,后接 3 层全连接层,使用 ReLU 激活函数,并通过 Dropout 防止过拟合。针对 CIFAR-10 数据集,模型由两组卷积层与池化层交替组成,后接两层全连接层,卷积层后接 BatchNorm 与 ReLU。损失函数为交叉熵损失,优化算法为随机梯度下降。客户端总数为 50,全局训练轮次为 100,本地训练轮次为 1,学习率为 0.1。

本文使用 IPM 攻击[22]、符号翻转攻击[23-24]、MIX 攻击[25]和 MINMAX 攻击[26]这 4 种攻击方式。IPM 攻击通过计算正常客户端模型的均值,并反转其方向后作为拜占庭客户端的上传模型。在符号翻转攻击中,拜占庭客户端将其本地模型的符号取反后上传至服务器。在 MIX 攻击中,对前半部分的拜占庭模型实施均值为 0,标准差为 0.5 的噪声攻击[27];对后半部分的拜占庭模型实施符号翻转攻击,并将其幅值缩放为原始值的 50%,以增强攻击

效果。在 MINMAX 攻击中,攻击者估计所有良性模型的均值与标准差,并以标准差的反方向作为扰动方向。随后通过逐步调整扰动强度,使构造出的拜占庭模型尽可能远离均值,并使得构造出的拜占庭模型与其他良性模型的最大距离不超过任意两个良性模型之间最大距离。

实验设置的对比防御基线分别为 Multi-Krum[9](简称为 Krum)、FLTrust[10]、RFLPA[11]、FLAME[12]和 FLRAM[13]。所有基线均采用原文献默认超参数进行实验。

本文从鲁棒性、聚合模型质量和效率这 3 个方面对算法进行评估。在鲁棒性方面,当恶意客户端占比为 30%时,比较各防御基线在准确率 *ACC* 和拜占庭攻击成功率 *ASR* 上的表现。聚合模型质量方面,比较最终参与聚合的模型数量,以及在参与聚合的模型中,良性和恶意模型所占的比重。在效率方面,比较 FLDBA 和其他防御基线以及 FedAvg[1] 在不同客户端数量下的运行时间。

3.2 实验结果分析

3.2.1 鲁棒性分析

本文对比了 FLDBA 与各防御基线的 *ACC* 和 *ASR* 在后 20 轮的平均值,结果如表 1 所示。

表 1 不同防御基线的 *ACC* 和 *ASR* 对比

Table 1 *ACC* and *ASR* of different defense baselines %

数据集	算法	IPM 攻击[22]		符号翻转攻击[23-24]		MIX 攻击[25]		MINMAX 攻击[26]	
		ACC	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*
F-MNIST[19]	Krum[9]	22.57	99.92	76.92	67.41	83.53	11.19	10.00	100.00
	FLTrust[10]	80.13	41.24	80.77	50.58	82.23	50.16	74.32	24.76
	RFLPA[11]	81.16	45.56	79.57	50.74	84.16	51.33	78.94	23.33
	FLAME[12]	83.31	0	81.37	63.33	80.56	37.04	78.15	0
	FLRAM[13]	83.73	0	80.56	49.63	85.33	6.81	84.16	0
	FLDBA	84.13	0	82.05	9.11	85.44	5.41	84.22	0
CIFAR-10[20]	Krum[9]	10.03	95.00	61.07	63.56	60.68	10.63	10.00	100.00
	FLTrust[10]	61.11	37.33	62.01	50.13	59.05	49.69	43.06	38.67
	RFLPA[11]	60.03	44.11	60.81	52.89	56.05	50.40	42.24	36.00
	FLAME[12]	59.80	0	58.01	68.86	10.01	53.16	61.38	0
	FLRAM[13]	62.52	0	62.32	48.89	62.54	0	61.45	0
	FLDBA	62.65	0	62.50	7.47	62.51	2.04	62.80	0

在抵御 IPM 这种简单的模型替换攻击时,FLDBA、FLAME 和 FLRAM 均展现出稳定且优异的防御性能。相较之下,Krum 的表现则明显不足。由于 Krum 仅基于模型间的距离进行筛选,忽略了模型的方向信息,导致其在面对 IPM 攻击时易受误导。IPM 攻击通过计算良性模型的均值,并翻转其方向构造恶意模型,在欧氏距离空间中,这些模型与良性模型的距离较小,从而在 Krum 中获得的评分越高。最终,Krum 将这些恶意模型误判为良性模型纳入聚合,导致其在该攻击场景下完全失效。

在抵御符号翻转攻击时,FLDBA 在两种数据集上的表现均优于其他防御基线。例如在 CIFAR-10 数据集上,其 *ACC* 分别比 FLRAM、FLAME、RFLPA、FLTrust 和 Krum 高 0.18 百分点、4.49 百分点、1.69 百分点、0.49 百分点和 1.43 百分点,*ASR* 分别低 41.42 百分点、61.39 百分点、45.42 百分点、42.66 百分点和 56.09 百分点,展现出极强的鲁棒性。

在应对 MIX 攻击时,FLDBA 在 F-MNIST 数据集上的防御效果依然显著,在 CIFAR-10 数据集上,*ACC* 比 FLRAM 低 0.03 百分点,而 *ASR* 则高出 2.04 百分点。这表明,FLDBA 整体具有较强的鲁棒性,但在面对高维复杂数据集时,由于模型特征分布较为离散,聚合效果受到一定影响,表现略逊于 FLRAM。相比之下,FLRAM 通过赋予少部分高信誉节点以较大的聚合权重,在高维数据场景下提升了防御能力,但是其鲁棒性的本质在于减少聚合数量,并将权重集中赋予少数高信誉模型。FLRAM 抑制了部分恶意模型参与聚合,但也存在误判良性模型、影响泛化能力的风险,整体防御策略仍有优化空间。

在应对 MINMAX 攻击时,Krum 完全失效,这是因为 MINMAX 攻击通过利用正常模型的均值和标准差构造出边界内的扰动模型,幅值与正常模型差异不大且高度一致,难以在距离排序中被准确识别。FLTrust 和 RFLPA 通过对模型幅值进行归一化,在一定程度上削弱了 MINMAX 攻击对聚合过程的干扰,但由于二者仅依赖与基准模型之间的余弦相似度进行评估,难以全面捕捉潜在的恶意更新。相比以上 3 种防御方法,FLRAM、FLAME 和 FLDBA 在抵

御 MINMAX 攻击时表现更优,这主要得益于三者在聚合前设置了更加全面的异常检测机制。尽管 MINMAX 攻击在距离空间中伪装得较为隐蔽,但其扰动方向的一致性与分布位置的异常特征仍能够被这些机制识别并剔除,从而显著提升了模型的鲁棒性,且 FLDBA 的防御性能最优。

3.2.2 聚合模型数量

本文统计了各基线在模型聚合时,参与聚合的模型数量在后 20 轮的平均值,实验结果如图 3 所示。

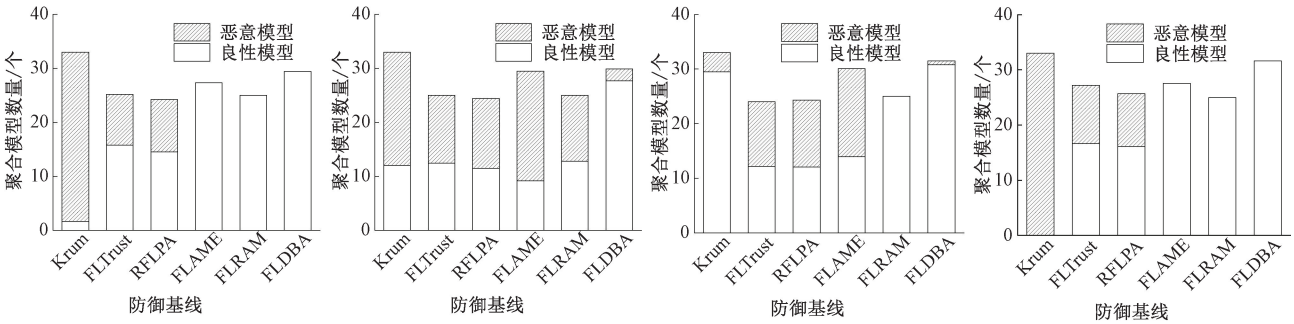

图 3 参与聚合的聚合模型数量

Figure 3 Number of models participating in aggregation

由图 3 可知,在抵御各种攻击时,FLDBA 的聚合模型数量多于多数防御基线,且恶意模型占比较低。这表明 FLDBA 在模型筛选方面表现出色,能够有效保留更多良性模型。此外,结合信誉机制,FLDBA 在降低恶意模型因漏判而被采纳方面也展现出显著优势。以抵御符号翻转攻击为例,尽管 Krum 在聚合过程中识别了较多模型,但由于其检测恶意模型的能力有限,导致大量恶意模型纳入聚合;FLTrust 和 RFLPA 通过模型归一化以及与全局模型的方向比对,在一定程度上增强了算法的鲁棒性,但由于其筛选策略不够严格,仍无法避免恶意模型的大量混入;FLAME 由于缺乏信誉机制,面对攻击者因漏判而参与聚合的情况,无法及时识别并剔除恶意模型,导致全局模型在训练过程中受到污染,进一步削弱了密度聚类方法在后续恶意模型识别中的效果,造成更多恶意模型混入聚合过程;FLRAM 虽然在拦截恶意模型方面更为高效,但由于筛选策略较为激进,导致最终参与聚合的良性模型数量少于 FLDBA,影响了模型的聚合质量。

3.2.3 防御方法的时间开销

表 2 列出了各个防御基线、FLDBA,以及 FedAvg 在不同客户端数量的情况下,执行 10 轮所耗费的时间。由表 2 可知,RFLPA 由于引入加密机制,整体时间开销较大,且随着客户端数量的增加,运行时间显著上升。Krum 和 FLRAM 由于本身算法复杂度较高,在客户端增多的情况下,运行时间呈显著的成倍增长。相比之下,FLDBA 即使在客户端规模较大时,其运行时间依然与 FLTrust、FLAME 和 FedAvg 相近,未产生明显的额外开销,在实际应用场景中具有良好的可行性。

3.2.4 消融实验

图 4 为在 CIFAR-10 数据集上,FLDBA 在是否包含信誉机制的情况下抵御符号翻转攻击能力的对比。从图 4 可以看出,当 FLDBA 中包含信誉机制时,在训练过程中能够及时识别恶意模型,并赋予其较低的信誉评分,从而有效抑制符号翻转攻击对模型聚合的干扰。相比之下,当算法中不包含信誉机制时,模型在面对符号翻转攻击时的鲁棒性明显下降,难以有效抵御攻击带来的负面影响。

表 2 不同防御基线执行 10 轮所耗费的时间对比

Table 2 Comparison of time consumption for 10 rounds of execution across different defense baselines

客户端数	t/s						
	Krum[9]	FLTrust[10]	RFLPA[11]	FLAME[12]	FLRAM[13]	FedAvg[1]	FLDBA
50	30.23	21.54	184.97	21.37	26.72	22.09	21.86
100	48.51	31.49	254.50	32.44	48.47	27.25	33.65
300	265.41	67.70	550.90	69.29	221.88	71.09	80.29

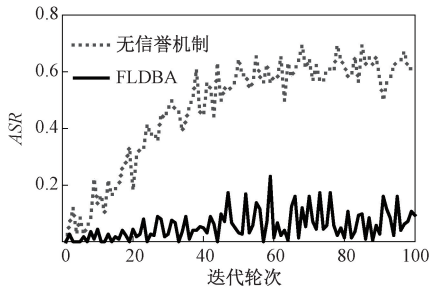

图 4　有无信誉机制时 *ASR* 对比

Figure 4　Comparison of *ASR* with and without the reputation mechanism

图 5 为在无攻击场景下,FLDBA 分别采用加权聚合与平均聚合策略时在 CIFAR-10 数据集上的准确率变化曲线。由图 5 可以看出,FLDBA 在采用加权聚合后,训练初期模型收敛速度显著加快,前 20 轮内测试准确率上升速度明显快于平均聚合策略。此外,在整个训练过程中,加权聚合始终保持更高的准确率,验证了其在提升模型性能与稳定性方面的有效性。

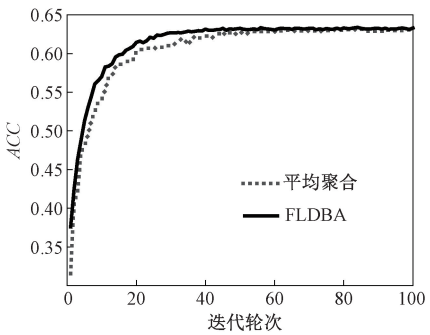

图 5　不同聚合方式时 *ACC* 对比

Figure 5　*ACC* comparison of different aggregation methods

为了评估动态模型过滤中阈值 δ 对模型性能的影响,在 CIFAR-10 数据集上进行了测试,分析了在抵御不同攻击时,不同阈值对模型性能的影响。实验结果如表 3 所示。

表 3　动态模型过滤中阈值对模型性能的影响

Table 3　Impact of thresholds on model performance in dynamic model filtering　%

阈值 δ	IPM 攻击[22]		符号翻转攻击[23-24]		MIX 攻击[25]		MINMAX 攻击[26]	
	ACC	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*	*ACC*	*ASR*
0.5	62.65	0	62.50	5.69	62.51	2.04	62.80	0
0.4	62.84	0	62.54	6.93	62.94	3.56	62.66	0
0.3	62.48	0	62.69	7.47	61.95	5.24	61.17	2.67
0.2	62.30	0	62.15	6.58	62.31	5.69	59.77	3.11
0.1	61.52	4.18	61.37	13.78	61.70	8.35	56.58	24.09
0	61.75	10.40	60.62	20.18	60.49	27.29	51.70	29.96

由表 3 可知,当阈值处于 0.2~0.5 时,模型的性能较优。当阈值低于 0.2 时,漏判的风险增加,导致在抵御某些攻击(如 MINMAX 攻击)时,模型性能明显下降。

4　结论

本文针对联邦学习中的拜占庭攻击问题,提出了一种抗拜占庭攻击的联邦学习鲁棒聚合算法 FLDBA。通过动态模型过滤、动态信誉评估、自适应缩放与加权聚合等机制,有效提升了系统的鲁棒性。实验结果表明,FLDBA 在 Fashion-MNIST 和 CIFAR-10 数据集上,相较多数防御基线在抵御拜占庭攻击方面表现更为优越,且不会带来显著的计算开销。在抵御符号翻转攻击时,FLDBA 的准确率比 FLRAM、FLAME、RFLPA、FLTrust 和 Krum 提升了 0.18 百分点~5.13 百分点,攻击成功率降低了 40.52 百分点~61.39 百分点,具有更好的鲁棒性。未来的研究可以进一步探索算法在恶意模型识别准确度上的提升,并评估其在更大规模数据集及复杂场景下的适应性表现。同时,可尝试引入降维或特征选择机制,以缓解高维空间中聚类效果下降的问题,从而进一步优化聚合性能与防御稳定性。

参考文献:

[1] Mcmahan H B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[PP/OL]. V4. arXiv (2023-01-26) [2025-11-01]. https://arxiv.org/abs/1602.05629.

[2] Shi Lei, Li Tian, Gao Yufei, et al. A review of machine learning-based methods for database tuning[J]. Journal of Zhengzhou University (Engineering Science), 2024, 45(1): 1-11. [石磊, 李天, 高宇飞, 等. 基于机器学习的数据库系统参数优化方法综述[J]. 郑州大学学报(工学版), 2024, 45(1): 1-11.]

[3] Zhang Chen, Xie Yu, Bai Hang, et al. A survey on federated learning[J]. Knowledge-Based Systems, 2021, 216: 106775.

[4] Chen Xuebin, Ren Zhiqiang, Zhang Hongyang. Review on security threats and defense measures in federated learning[J]. Journal of Computer Applications, 2024,

44(6): 1663-1672. [陈学斌, 任志强, 张宏扬. 联邦学习中的安全威胁与防御措施综述[J]. 计算机应用, 2024, 44(6): 1663-1672.]

[5] Chen Xuebin, Qu Changsheng. Overview of backdoor attacks and defense in federated learning[J]. Journal of Computer Applications, 2024, 44(11): 3459-3469. [陈学斌, 屈昌盛. 面向联邦学习的后门攻击与防御综述[J]. 计算机应用, 2024, 44(11): 3459-3469.]

[6] Sikandar H S, Waheed H, Tahir S, et al. A detailed survey on federated learning attacks and defenses[J]. Electronics, 2023, 12(2): 260.

[7] Wang Yongkang, Xia Yuanqing, Zhan Yufeng. ELITE: defending federated learning against Byzantine attacks based on information entropy[C]// Proceedings of the 2021 China Automation Congress (CAC). Piscataway: IEEE, 2021: 6049-6054.

[8] Liu Pengrui, Xu Xiangrui, Wang Wei. Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives[J]. Cybersecurity, 2022, 5: 4.

[9] Blanchard P, El Mhamdi E M, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]//Neural Information Processing Systems. Long Beach: NeurIPS, 2017: 118-128.

[10] Cao Xiaoyu, Fang Minghong, Liu Jia, et al. FLTrust: Byzantine-robust federated learning via trust bootstrapping [C]//Proceedings 2021 Network and Distributed System Security Symposium. Reston: Internet Society, 2021: 1-18.

[11] Mai Peihua, Pang Yan, Yan Ran. RFLPA: A robust federated learning framework against poisoning attacks with secure aggregation[C]// Neural Information Processing Systems. Vancouver: NeurIPS, 2024: 104329-104356.

[12] Nguyen T D, Rieger P, Chen Huili, et al. FLAME: taming backdoors in federated learning [PP/OL]. V5. arXiv (2023-08-05) [2025-11-01]. https: // arxiv.org/abs/2101. 02281v4.

[13] Chen Haitian, Chen Xuebin, Peng Lulu, et al. FLRAM: Robust aggregation technique for defense against Byzantine poisoning attacks in federated learning[J]. Electronics, 2023, 12(21): 4463.

[14] Lei Cheng, Zhang Lin. Federated learning model based on update quality detection and malicious client identification[J]. Computer Science, 2024, 51(11): 368-378. [雷诚, 张琳. 基于更新质量检测和恶意客户端识别的联邦学习模型[J]. 计算机科学, 2024, 51(11): 368-378.]

[15] Li Shenghui, Ngai E C H, Voigt T. An experimental study of Byzantine-robust aggregation schemes in federated learning[J]. IEEE Transactions on Big Data, 2024, 10(6): 975-988.

[16] Shi Junyu, Wan Wei, Hu Shengshan, et al. Challenges and approaches for mitigating Byzantine attacks in federated learning [PP/OL]. V2. arXiv (2022-10-07) [2025-11-01]. https://arxiv. org/abs/2112. 14468.

[17] Xia Peipei, Zhang Li, Li Fanzhang. Learning similarity with cosine similarity ensemble [J]. Information Sciences, 2015, 307: 39-52.

[18] Arthur D, Vassilvitskii S. k-means^{++}: the advantages of careful seeding[C]// ACM-SIAM Symposium on Discrete Algorithms. New York: ACM, 2007: 1027-1035.

[19] Xiao Han, Rasul K, Vollgraf R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms[PP/OL]. V2. arXiv (2017-09-15) [2025-11-01]. https://arxiv. org/abs/1708. 07747.

[20] Krizhevsky A. Learning multiple layers of features from tiny images[R/OL]. (2009-04-08) [2025-11-01]. https: // www. cs. utoronto. ca/~ kriz/learning-features-2009-TR. pdf.

[21] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[22] Xie Cong, Koyejo S, Gupta I. Fall of empires: breaking Byzantine-tolerant SGD by inner product manipulation [PP/OL]. V1. arXiv (2019-03-10) [2025-11-01]. https://arxiv. org/abs/1903. 03936.

[23] Li Liping, Xu Wei, Chen Tianyi, et al. RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 1544-1551.

[24] Rajput S, Wang Hongyi, Charles Z, et al. DETOX: a redundancy-based framework for faster and more robust gradient aggregation[PP/OL]. V2. arXiv (2020-03-08) [2025-11-01]. https://arxiv.org/abs/1907.12205.

[25] Alkhunaizi N, Kamzolov D, Takáč M, et al. Suppressing poisoning attacks on federated learning for medical imaging[C]//Medical Image Computing and Computer Assisted Intervention. Cham: Springer, 2022: 673-683.

[26] Shejwalkar V, Houmansadr A. Manipulating the Byzantine: optimizing model poisoning attacks and defenses for federated learning[C]//Proceedings 2021 Network and Distributed System Security Symposium. Virtual. Internet Society, 2021: 1-18.

[27] Li Shenghui, Ngai E, Voigt T. Byzantine-robust aggregation in federated learning empowered industrial IoT[J]. IEEE Transactions on Industrial Informatics, 2023, 19(2): 1165-1175.

A Robust Aggregation Algorithm Defending Against Byzantine Attacks in Federated Learning

ZHANG Shufen[1,2,3], LI Tao[1,2,3], ZHANG Zhenbo[1,2,3], ZHONG Qi[1,2,3], JING Zhongrui[1,2,3]

(1. College of Science, North China University of Science and Technology, Tangshan 063210, China; 2. Hebei Province Key Laboratory of Data Science and Application(North China University of Science and Technology), Tangshan 063210, China; 3. Tangshan Key Laboratory of Data Science(North China University of Science and Technology), Tangshan 063210, China)

Abstract: To address the issue that existing defense schemes in federated learning tend to over-prune benign models during filtering, a robust aggregation algorithm defending against Byzantine attacks in federated learning(FLDBA) was proposed. HDBSCAN density-based clustering was employed to group models, identifing the benign cluster, and the most representative model in terms of direction was selected as the trusted reference model. Using the trusted model as a benchmark, cosine similarity was utilized to screen potentially misclassified benign models within clusters, thereby correcting misjudgments. Additionally, a reputation mechanism was established to dynamically evaluate models' historical behaviors, mitigating the impact of missed detections. For models with high reputation, adaptive magnitude scaling was applied, and differential aggregation weights were assigned based on update quality to further enhance aggregation performance. Experimental results demonstrated that when defending against sign-flipping attacks, FLDBA achieved an accuracy improvement of 0.18 percentage points to 5.13 percentage points compared to FLRAM, FLAME, RFLPA, FLTrust, and Krum, while reducing the attack success rate by 40.52 percentage points to 61.39 percentage points, exhibiting superior robustness.

Keywords: federated learning; Byzantine attacks; robust; reputation; weighted aggregation