DAGUARD：联邦学习下的分布式后门攻击防御方案

余晟兴[1]，陈泽凯[2]，陈钟[1]，刘西蒙[2]
（1. 北京大学计算机学院，北京 100871；2. 福州大学计算机与大数据学院/软件学院，福建 福州 350108）

摘 要：为了解决联邦学习下的分布式后门攻击等问题，基于服务器挑选最多不超过半数恶意客户端进行全局聚合的假设，提出了一种联邦学习下的分布式后门防御方案（DAGUARD）。设计了三元组梯度优化算法局部更新策略（TernGrad）以解决梯度局部调整的后门攻击和推理攻击、自适应密度聚类防御方案（AdaptDBSCAN）以解决角度偏较大的后门攻击、自适应裁剪方案以限制放大梯度的后门增强攻击和自适应加噪方案以削弱分布式后门攻击。实验结果表明，在联邦学习场景下，所提方案相比现有的防御策略具有更好的防御性能和防御稳定性。

关键词：联邦学习；分布式后门攻击；聚类；差分隐私
中图分类号：TN92
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2023086

DAGUARD: distributed backdoor attack defense scheme under federated learning

YU Shengxing[1], CHEN Zekai[2], CHEN Zhong[1], LIU Ximeng[2]
1. School of Computer Science, Peking University, Beijing 100871, China
2. College of Computer and Data Science/College of Software, Fuzhou University, Fuzhou 350108, China

Abstract: In order to solve the problems of distributed backdoor attack under federated learning, a distributed backdoor attack defense scheme (DAGUARD) under federated learning was proposed based on the assumption that the server selected no more than half of malicious clients for global aggregation. The partial update strategy of the triple gradient optimization algorithm (TernGrad) was designed to solve the backdoor attack and inference attack, an adaptive density clustering defense scheme was designed to solve the backdoor attacks with relatively large angle deflection, the adaptive clipping scheme was designed to limit the enhancement backdoor attack that amplify the gradients and the adaptive noise-enhancing scheme was designed to weaken distributed backdoor attacks. The experimental results show that in the federated learning scenario, the proposed scheme has better defense performance and defense stability than existing defense strategies.

Keywords: federated learning, distributed backdoor attack, cluster, differential privacy

0 引言

近年来，物联网和移动设备在移动通信领域有着广泛应用，并且在日常生活中也越来越普遍。由于其本地数据及算力极其有限，用户通常将数据和计算外包给云服务器集中处理。数据在外包计算的过程中面临隐私泄露的风险，因此联邦学习（FL, federated learning）应运而生。与传统的集中式深度学习不同，FL[1]允许客户端将数据集留在本地进行训练，本地训练后仅上传模型权重或梯度进行全局模型的训练，这种方法间接实现了不同客户端之间的协作学习，极大地降低了数据泄露的风险，节省了通信开销。随着新兴隐私保护法规的盛行，FL因其能够潜在保护用户数据而受到了广泛的认可

收稿日期：2023−01−12；**修回日期：**2023−04−12
通信作者：陈钟，zhongchen@pku.edu.cn
基金项目：国家自然科学基金资助项目（No.62072109, No.62102422）
Foundation Item: The National Natural Science Foundation of China (No.62072109, No.62102422)

和使用。例如，相关网站使用 FL 来实现信用风险预测[2]；在 Melloddy 项目中，10 多家领先的制药公司利用 FL 进行药物发现[3]；Google 在 Android Gboard[4]上部署 FL 进行键盘输入联想预测，其中 FedAvg[1]是由 Google 开发的一种流行的 FL 方案，该方案的全局模型更新为各客户端本地模型更新的加权平均值，权重为各客户端本地训练数据集的大小。

FL 由于其分布特性，很容易受到恶意客户端对抗操纵的影响，恶意客户端可能是攻击者伪造的客户端或向攻击者妥协的真实客户端。恶意客户端通过毒化本地数据[5-6]或者篡改本地模型梯度进行中毒攻击[7]，进而损坏全局模型。被损坏的全局模型会将目标测试样本预测为攻击者选择的错误标签，而其他非目标测试样本将不受影响[8-10]。目前，在 FL 中普遍使用的 FedAvg 全局模型聚合方式较脆弱，单个恶意客户端就可以采用多种攻击方式将其攻破[11-12]。

近年来，FL 攻击已经受到广泛的讨论，如 FL 容易受到后门攻击[9-10,13]以及推理攻击[14-16]的影响。后门攻击通过操纵模型更新或者梯度来影响全局模型，即攻击者选择的输入会导致全局模型预测错误。在推理攻击中，对手通过分析模型更新来学习客户端本地数据的信息。现有研究[12-13,17-18]致力于在少数恶意客户端并且服务器确保诚实的背景下增强 FL 的鲁棒性，例如，Blanchard 等[12]提出的 Krum 方案在 N 个本地更新中选择与上一轮迭代更新距离最小的一个作为每次迭代的全局更新，然而，上述机制的一个主要缺点是只适用于诚实客户端占绝大多数的情况；Median 方案[11]中，服务器选择所有上传的模型更新的中位数作为全局更新，但无法保证较高的准确率；Shen 等[19]提出的后门防御策略无法有效地抵御分布式后门攻击（DBA）[9]。FLAME 方案[20]虽然对后门攻击有较好的防御效果，但是由于其需要上传完整的模型，无法抵御推理攻击并且在某些数据集下防御稳定性较差。针对目前防御方法存在的问题，为有效保护联邦学习下的模型安全，本文提出了联邦学习下的分布式后门攻击防御方案（DAGUARD）。

本文的主要贡献如下。

1) 为了防御推理攻击和对梯度局部调整的后门攻击，本文采用 TernGrad[21]的方式对神经网络每层的梯度进行 Ternarize 转换，即使用每层绝对值最大的梯度作为当前层的梯度。

2) 根据联邦学习下服务器每轮挑选不超过半数恶意客户端进行全局聚合的假设，利用基于密度的带噪声应用空间聚类（DBSCAN）[22]设计了自适应密度聚类方法。由于大多数的中毒模型梯度相比于良性模型梯度有较大的角度偏差，一个较好的聚类策略可以在很大程度上消除恶意客户端的攻击。FLAME 方案[20]采用的是 HDBSCAN（hierarchical DBSCAN）[23]聚类方法，其设置聚类数目上限仅为客户端数量的一半，聚类结果不够准确，无法有效剔除与良性梯度相近的恶意梯度。本文采用自适应中位数作为标准，动态调整 DBSCAN[22]领域半径进行相近恶意梯度后门攻击的防御。

3) 本文基于 TernGrad 方法设计了自适应裁剪方案和自适应加噪方案。目前的 FLAME 方案[20]采用的裁剪方法是直接对梯度大小进行裁剪，恶意梯度可以通过适当缩放躲避裁剪，而本文方案是对各客户端每轮训练后的梯度经过 TernGrad 方法转换后进行裁剪，可以更好地削弱恶意客户端模型梯度的增强攻击。同时，采用差分隐私加噪的方式可以削弱联邦后门攻击，本文根据神经网络每层的最大梯度更新的第二范数计算出每层的高斯噪声 σ，为每层神经网络添加自适应高斯噪声，平滑经过 DBSCAN 聚类后的模型更新，有效减少后门攻击的影响。

4) 本文设计的 DAGUARD 方案在不同非独立同分布情况下均具有较好的防御效果，且在不同数据集和数据投毒率下均有较高的防御稳定性，实验表明 DAGUARD 的防御效果优于目前主流的 FedAvg、Median 以及 FLAME 方案。

1　相关工作

1.1　聚类

聚类[24]是一种无监督的机器学习算法，它将数据分成多个有意义的子组，这些子组使聚类后的簇内差异最小化，簇间差异最大化，目前常用的聚类算法大致可以分为四类：基于层次、基于分区、基于网格和基于密度。基于层次的聚类算法[25]是从最初的一些集群开始逐渐收敛的解决方案，其主要缺陷在于计算复杂度较高，并且如果数据存在奇异值，则会对聚类效果产生很大的影响。基于分区的聚类算法将数据集划分为初始 K 个聚类，并根据目标函数迭代提高聚类质量，如 K-means[26]就是基于分区的聚类算法，而该类算法需要明确指定聚类数目且聚类效果受其影响较大。在基于网格的聚类算

法[27]中，整个数据集被一个规则的超网格覆盖，同一个网格中的数据点被归为一簇。在基于密度的聚类算法[22]中，当区域内点的密度大于最小密度值时，该区域被称为密集区域或密度相连区域。由于基于密度的聚类算法基于密集连通性扩展集群，该类算法可以找到任意形状的集群。DBSCAN 就是基于密度的聚类算法，因此其可以对任意形状的稠密数据集进行聚类且可发现异常点。

1.2 联邦学习

假设有 n 个客户端，每个客户端都有训练数据集 $D_i, i=\{1,\cdots,n\}$，协同训练全局模型 W。集中学习的本地数据集必须在训练前由中央服务器收集，而联邦学习[6]仅要求客户端将本地模型（$\{w_i \mid i \in n\}$）上传到服务器，在服务器上进行联邦聚合得到全局模型，表示为

$$W=\frac{1}{n}\sum_{i=1}^{n} w_i \tag{1}$$

具体来说，联邦学习主要优化损失函数，表示为

$$\min F(w)=\sum_{i=1}^{n}\frac{k_i}{K}L_i(w) \tag{2}$$

其中，$L_i(w)$ 和 k_i 是损失函数和第 i 个客户端的本地数据集大小。

1.3 分布式后门攻击

分布式后门攻击[9]使用多个不同色彩或不同灰度的补丁作为触发器并将其分成几个部分，分别设置在不同的客户端上。不同于传统的集中式后门攻击，在分布式后门攻击中，每个恶意客户端会被分配后门触发器的一部分客户端进行协同攻击。如果指定触发部分被中心服务器所学习，则该触发器被触发，后门攻击成功。独立的触发器的攻击强度相比于集中式触发器弱，具有更高的隐蔽性，其中分布式后门攻击将一个集中式攻击公式分解为 M 个分布式子攻击问题[9]，表示为

$$\begin{aligned}&\operatorname{argmax}\left(\sum_{j\in S_{\text{poi}}^{i}} P\left[G^{t+1}(R(x_j^i,\Phi_i))=\tau;\Gamma;I\right]+\right.\\&\left.\sum_{j\in S_{\text{cln}}^{i}} P\left[G^{t+1}(x_j^i)=y_j^i\right]\right)\end{aligned} \tag{3}$$

其中，i 表示第 i 个攻击者，$i\in\{1,\cdots,M\}$，j 表示第 j 个数据库，t 表示第 t 轮次，P 表示预测准确率，I 表示投毒间隔，G 表示全局模型，D 表示数据库，S_{poi}^{i} 表示投毒数据库，S_{cln}^{i} 表示良性数据库且满足 $S_{\text{poi}}^{i}\cup S_{\text{cln}}^{i}=D_i$ 以及 $S_{\text{poi}}^{i}\cap S_{\text{cln}}^{i}=\Phi$，函数 R 将任何类中的良性数据库转换为具有攻击者选择触发模式的投毒数据库，参数 Φ_i 可被分解为触发位置、触发大小和触发间隙，Γ 表示中毒目标标签，y_j^i 表示未中毒的标签。

2 理论知识

2.1 DBSCAN

DBSCAN[22]是基于密度的带噪声应用空间聚类，其根据密度的方差区分高维数据库的噪声。DBSCAN 根据预先设定的超参数领域半径 Eps 和簇内最小样本数目 MinPts，将数据点分为核心点、边界点以及噪声点。当某个数据点在 Eps 半径内至少包含了 MinPts 个数据点时，该数据点为核心点；当某个数据点的 Eps 半径内包含的数据点少于 MinPts 个且该数据点在其他核心点的领域半径内，则该数据点为边界点；既不是核心点也不是边界点的则是噪声点。

数据点之间的距离关系可分为密度直达、密度可达和密度相连。当数据点 q 是核心点时，数据点 p 在其 Eps 领域半径内，则 p 和 q 是密度直达的；当数据点 p 与数据点 q 之间存在一系列节点 $l_i\in\{q,\cdots,p\}$，且对于任意的 l_i 到 l_{i+1} 是密度直达时，则 p 对于 q 是密度可达的；当数据点 p 与数据点 q 的领域半径内存在核心点 o，其对于 p 和 q 是密度可达的，则 p 与 q 是密度相连的。最终根据数据点之间的距离关系，将高密度区域聚成簇，并导出密度连接集合。

2.2 TernGrad

Wen 等[21]提出的 TernGrad 采用三元组逐层量化以及梯度裁剪的方式提高联邦学习模型训练时的通信效率。在每轮迭代训练过程中，客户端数量为 n，其中 $i\in\{1,\cdots,n\}$，梯度数量为 m，各本地量化后的局部梯度为

$$\widetilde{g}_t^{(i)}=\text{Ternarize}\left(g_t^{(i)}\right)=s_t^{(i)}g_t^{(i)}\circ b_t^{(i)} \tag{4}$$

其中，缩放因子 $s_t^{(i)}\triangleq\max\left(\text{abs}\left(g_t^{(i)}\right)\right)$，max 函数计算出所有元素的最大值，abs 计算出当前元素的绝对值，$b_t^{(i)}=\left\lfloor\frac{[g_t^{1(i)},\cdots,g_t^{m(i)}]}{s_t^{(i)}}\right\rfloor$，$\circ$ 表示 Hadamard 乘积。TernGrad[21]主要更新变化幅度大的梯度元素，其收敛性已得到证明。针对不同客户端，计算不同的 $s_t^{(i)}$ 来保持各客户端的特征。

2.3　差分隐私

差分隐私（DP）[27]是对抗敌方不同攻击的一种强标准，其具体定义如下。

定义 1　相邻数据库。如果对于 2 个数据库 x，y，有 $\|x-y\|_1 \leqslant 1$，那么数据库 x 和 y 称为相邻数据库[28]。

定义 2　$(\varepsilon,\delta)-\mathrm{DP}$ [29]。考虑 2 个相邻的数据集 D 和 D'，它们仅在一个数据样本中有所不同。对于任何确定性查询函数 $f:D\to\mathbb{R}^M$ 和随机机制 $M:\mathbb{R}^M\to O$，$M\circ f$ 实现 $(\varepsilon,\delta)-\mathrm{DP}$ 且输出的任何子集 $S\subseteq O$ 满足

$$\Pr[M(f(D))\in S]\leqslant \mathrm{e}^{\varepsilon}\Pr[M(f(D'))\in S]+\delta \quad (5)$$

其中，Pr 为期望概率；ε 为隐私预算，表示隐私保护程度；$\delta\in[0,1]$ 为松弛因子，表示可容忍违背严格差分隐私的概率，主要通过加入高斯噪声来满足差分隐私。

定义 3　高斯机制[30]。令 $\varepsilon\in(0,1)$，对于任何函数 f，定义算法 $M=f(x)+n$，其中 n 遵守高斯分布 $n\sim N(0,\sigma^2)$。当参数 $\sigma\geqslant c\Delta_2\dfrac{f}{\varepsilon}$ 时，算法 $M=f(x)+n$ 满足 $(\varepsilon,\delta)-\mathrm{DP}$，其中，常数 c 满足 $c^2>2\ln\left(\dfrac{1.25}{\delta}\right)$，$\Delta_2 f=\max_{x,y\in\mathbb{N}^{|x|}}\|f(x)-f(y)\|_2$，$x$ 和 y 表示相邻数据库。

3　问题定义

3.1　攻击模型

在典型的 FL 设置中，来自外部对手的威胁主要可以分为以下两类。

1) 试图向全局模型注入联邦学习后门的恶意客户端 A^b。

2) 诚实且好奇的聚合器 A^h，其能够遵循训练协议正确地执行计算，但会通过推理攻击[31]获取关于客户端的训练数据信息。

A^b 对 $c\left(c<\dfrac{n}{2}\right)$ 个恶意客户端及其训练数据、模型参数等具有完全控制[31]，也对 FL 中聚合器的操作有充分的了解，可以随时在训练期间采取任意适当的攻击策略，例如，同时注入一个或者多个后门攻击，A^b 也无法对其他诚实的客户端以及中心聚合器的执行过程进行操作。

A^h 为遵循训练协议的半诚实攻击者，其可以访问所有本地模型 W^i 并对每个本地模型 W^i 进行模型推理攻击，以此提取客户端数据信息。

3.2　攻击方式

在分布式后门攻击中，由于触发器被分为多个并由不同的攻击者持有，各攻击者的后门触发器攻击能力不稳定，有时单个触发器无法直接改变预测结果[12]，但攻击的总体效果可以导致恶意梯度与良性梯度出现角度偏离或者幅度偏差，因此分布式后门攻击具有很高的隐蔽性，A^b 设置的独立后门触发器具体的攻击方式如下。

1) 模型更新幅度偏差较大的恶意攻击。A^b 设置的后门触发器对良性模型影响较大，导致恶意模型更新幅度远超过良性模型更新，以此来扩大恶意攻击。

2) 模型更新角度偏离较大的恶意攻击。A^b 设置的后门触发器从角度偏转对良性模型进行攻击，通过采用较大的毒化率或者大量的局部训练周期[9]来实现良性模型更新角度偏离较大的恶意攻击。

3) 模型更新具有较小的角度偏离和幅度偏差的恶意攻击。A^b 通过限制训练和攻击规模来实现较小的角度偏离和幅度偏差的恶意攻击。

3.3　防御目标

在 FL 设置中，能有效解决分布式后门攻击的通用防御方案需要实现以下目标。

1) 主任务准确率高。必须保证全局模型的高准确率表现以维持其有效性。

2) 后门攻击成功率低。为了防止对手对目标的攻击，需要消除后门模型更新的影响，尽可能降低后门攻击成功率。

3) 防御稳定性。防御方案必须适用于通用的攻击模型，即不需要知道后门攻击方法的先验知识，并且在不同数据集和不同条件参数下的防御表现较稳定。

4　本文方案

4.1　方案描述

本节描述了轻量级联邦学习下的分布式后门攻击高效防御方案（DAGUARD）的设计过程，如图 1 所示，恶意客户端通过对训练数据插入分布式后门触发器进行后门攻击，其中，g^{attack} 表示经过后门攻击后的梯度，g 表示良性客户端训练的局部梯度，G 表示经过后门防御后的全局梯度，目标标签表示后门任务将良性标签替换成的恶意标签。DAGUARD 方案采用了 TernGrad 方法、自适应

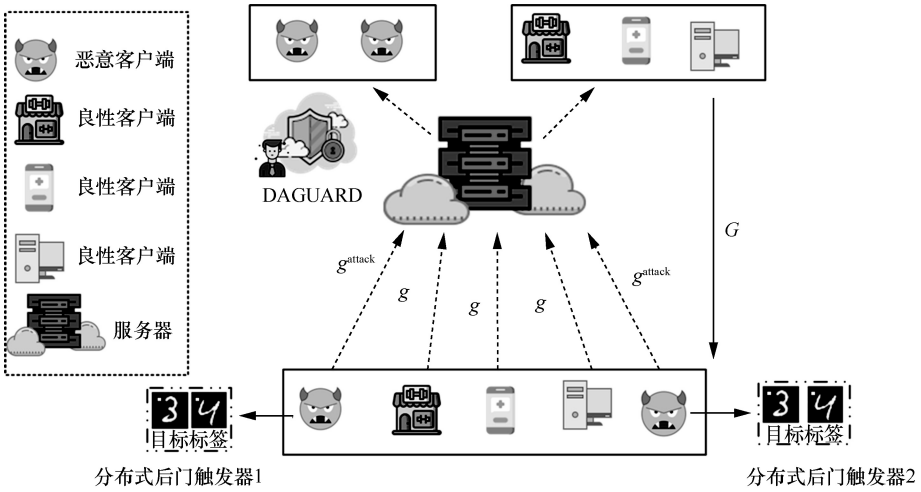


图 1　DAGUARD 的设计过程

DBSCAN 聚类、自适应裁剪、自适应加噪策略来最大限度地筛选出恶意梯度，并且根据分布式后门攻击的特点进行防御，从而更加有效地抵御分布式后门攻击。

4.2　TernGrad 局部更新

为提高防御模型的效率，避免梯度局部调整的后门攻击以及服务器的推理攻击，采用 TernGrad 方法[21]将本地更新后的梯度进行 Ternarize 转换。由于每层神经网络的模型梯度分布类似于高斯分布[32]，在每层进行 Ternarize，选择每层中的最大梯度作为该层的梯度，可以最大限度地保留梯度更新的重要信息而不泄露完整的梯度信息，从而避免服务器受到推理攻击。TernGrad 局部更新（TernGradUpdate）如算法 1 所示。

算法 1　TernGradUpdate

输入　所有客户端本地迭代轮次 T，客户端数量 n，客户端 i 的第 t 轮更新权重 g_t^i，TernGrad 转换后的梯度 g^i，梯度 g^i 的二范数 g_e^i

输出　最优局部模型 W_T^i

1)　for t in $[1,T]$ do
2)　　for i in $[1,n]$ do
3)　　$g_t^i = \text{Model}(\text{Datasets}, W_{t-1}^i)$ //更新模型
4)　　$g^i = \text{Ternarize}(g_t^i)$ //TernGrad 转换
5)　　$g_e^i = \sqrt{(g^i)^2}$ //计算梯度的二范数
6)　　发送 g^i, g_e^i 到服务器
7)　　从服务器接收全局模型 g_b
8)　　$W_t^i = W_{t-1}^i + g_b$ //更新模型
9)　　end for
10)　end for

4.3　自适应 DBSCAN

由于大多数的恶意梯度相比良性梯度有较大的角度偏离，首先采用聚类的方法剔除有较大角度偏离的恶意梯度，如图 2 所示。Blanchard 等[12]和 Ganju 等[15]提出的基于聚类的防御策略将所有模型分为良性簇和恶意簇两类，但当聚类中不存在恶意模型时，该策略可能导致一部分良性模型被错误地删除，进而降低了模型的准确率。另外，Blanchard 等[12]提出的方案允许偏转角度为 0°～90°，在非独立同分布场景下大部分良性客户端将会被剔除。Ganju 等[15]提出的方案在前 10 轮采用 K-means 聚类算法，其后的轮次通过与前 10 轮被剔除的特征进行比对来划分恶意模型与良性模型，这种做法容易将良性模型与恶意模型错误地划为同一簇中，无法将恶意模型从良性模型中分离，因此上述方案均无法抵御联邦学习下的分布式后门攻击。而 FLAME 方案[20]采用 HDBSCAN 聚类方法，仅设置了簇内最小样本数量超参数，虽然根据梯度之间的余弦距离进行聚类，但若存在与良性梯度较接近的恶意梯度，该方法不考虑其是否应该被划分到其他恶意簇中，而是仅依据距离偏转聚集一定数量的梯度，无法针对性地约束聚类，得到的聚类有一定的不可控性，当恶意梯度临界于恶意簇与良性簇时，HDBSCAN 无法准确地将恶意梯度划分到恶意簇。

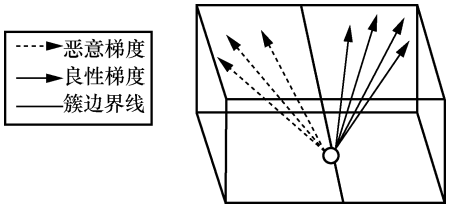


图 2　聚类分离恶意梯度与良性梯度

由于 HBDSCAN 无法有效应对恶意梯度既可以划分到良性簇又可以划分到恶意簇的情况，本文采用 DBSCAN 的密度聚类方式，根据恶意客户端数量不超过$\left\lfloor \frac{n}{2} \right\rfloor$的假设，将数据点之间余弦距离的中位数作为 Eps 领域半径的阈值并将$\left\lfloor \frac{n}{2} \right\rfloor$作为 MinPts 簇内最小样本数目。由于良性客户端具有相近的特征分布， MinPts 最大限度地将良性客户端聚成一簇。同时，由于恶意客户端与良性客户端的距离偏差较大，以余弦距离中位数作为 Eps 可以将恶意客户端聚成一簇。该方法明确设置了领域半径和最小样本数目 2 个约束条件，每次聚类一个新梯度后，聚类圆心都会发生相应调整，每个梯度会根据密度距离关系被划分到最合适的簇中，如果一个梯度既可以分到 A 簇也可以分到 B 簇，该梯度会被当作噪声点排除，这种方法可以根据设定的超参数 {Eps, MinPts} 将每个梯度划分到密度距离关系最接近的簇中，最大限度地将恶意梯度从良性梯度中剔除。

由于恶意客户端使用中毒数据导致全局模型角度较大地偏离良性客户端，因此不再采用欧几里得距离作为测量参数更新的标准，而是采用梯度之间的余弦距离作为角度距离，客户端 u 的相似度得分用 ρ_u 表示，即

$$\rho_u = \frac{1}{m}\sum_{i=1,\ i\neq u}^{m}\frac{g_i g_u}{\|g_i\|\|g_u\|} \tag{6}$$

其中， g_i 和 g_u 分别表示客户端 i 和 u 的梯度。

为了剔除恶意客户端与良性客户端之间较大的角度偏离的影响，本节提出了 AdaptDBSCAN 聚类的方式剔除恶意梯度攻击。自适应 DBSCAN 协议 （AdaptDBSCAN）如算法 2 所示。

算法 2　AdaptDBSCAN

输入　客户端上传局部梯度向量 $\boldsymbol{G}$，客户端数量 n，领域半径 Eps，簇内最小样本数目 MinPts

输出　聚类结果 I

1) Dis = COSDIS($\boldsymbol{G}$) //计算数据点之间的余弦距离，具体如式(6)所示
2) Eps = Median(Dis) //计算距离中位数
3) $\text{MinPts} = \left\lfloor \frac{n}{2} \right\rfloor$
4) I = DBSCAN(Dis, Eps, MinPts)

4.4　自适应裁剪

剔除了角度偏离较大的恶意客户端后，为进一步削弱相近角度中被放大的恶意后门梯度的影响，需要对高量级模型的梯度进行裁剪[33]，如图 3 所示。当攻击者通过缩放攻击使恶意模型与良性模型之间的欧几里得距离接近于良性模型之间的欧几里得距离时，FLAME 方案[20]无法有效地对恶意梯度进行裁剪，而本文设计的自适应裁剪方案对客户端每轮训练后的梯度进行 TernGrad[21]转化，放大了恶意梯度更新进行统一裁剪，从而避免了缩放梯度造成的裁剪逃逸问题，在每轮训练后挑选神经网络每层的最大梯度作为当前层的梯度，并计算梯度中位数裁剪各层更新的梯度来抵御恶意梯度增强攻击。由于后门放大攻击中恶意梯度大小远超过良性梯度，影响最终全局模型的方向以及大小，因此本文设计了一种自适应裁剪方案，以 g_α 为半径对梯度进行裁剪。如图 3 所示，通过限制恶意梯度大小能有效地削弱其影响。对梯度进行裁剪时需要最大限度地保留良性梯度特征，若 g_α 的值设置过小，将会删除大部分良性客户端的梯度，从而导致全局梯度被后门攻击成功；若 g_α 的值设置过大，将无法有效地削弱恶意客户端梯度的影响。

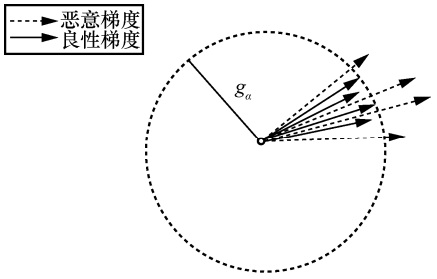


图 3　自适应裁剪削弱恶意梯度

根据恶意客户端数量不超过$\left\lfloor \frac{n}{2} \right\rfloor$的假设，通过计算各客户端模型更新的欧几里得距离，并选择中位数作为裁剪的边界。当客户端梯度欧几里得距离 g_e^i 超过 g_α 时，将被压缩成 g_α，否则梯度保持不变。g_e^i 可在第 i 个客户端本地预先计算，具体自适应裁剪计算式为

$$\begin{aligned} g_\alpha &= \text{Median}([g_e^1, \cdots, g_e^n]) \\ g_b^i &= \text{Sign}(g^i)\text{Min}(g_\alpha, g_e^i) \end{aligned} \tag{7}$$

其中，$g_e^i = \sqrt{(W_t^i - W_{t-1}^i)^2} = \sqrt{g_i^2}, i \in \{1, \cdots n\}$，$g^i$ 是第 i 个客户端本地迭代更新后的梯度，g_b^i 是第 i 个客户端裁剪后的梯度大小。

4.5 自适应加噪

Du 等[34]实验表明，通过加入差分隐私噪声可以有效抵御异常样本的影响，在训练过程中给模型添加适当的差分隐私噪声，有毒样本对模型的影响就越低。如图 4 所示，适当的噪声使加噪后的梯度偏离恶意梯度而靠近良性梯度，因此添加适当的差分隐私噪声可以增加模型抵御后门攻击的鲁棒性。

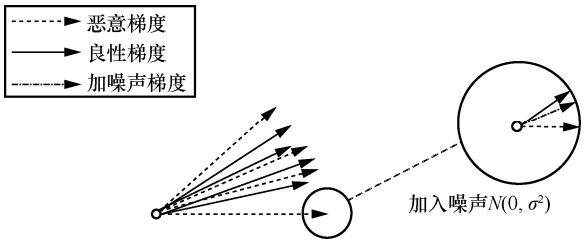


图 4 自适应加噪平滑恶意梯度与良性梯度距离

虽然引入高斯噪声可以减轻聚合后模型受后门攻击的影响，然而选择合适的噪声水平 σ 是至关重要的，因为它直接影响良性模型的梯度有效性。

如果 σ 选择过小，聚合后的模型就可能仍然被保留的后门影响导致抵御失效；如果 σ 选择过大，将会影响聚合模型的精度，导致模型鲁棒性较差，无法正确表达。因此本文基于优化方案设计了一种实用且有效的噪声界限方案，将神经网络各层的最大梯度的更新值作为 $\Delta_2 f$，从而计算出各层高斯噪声的 σ。该方法相比于 FLAME 方案中直接使用梯度更新作为加噪依据的方法突出了更新的程度，能够更有效地平滑恶意梯度的影响，σ 的计算式为

$$\sigma = \frac{\Delta_2 f}{\varepsilon}\sqrt{2\ln\frac{1.25}{\delta}} = \frac{g_\alpha}{\varepsilon}\sqrt{2\ln\frac{1.25}{\delta}} \tag{8}$$

其中，ε 是隐私预算，δ 是松弛因子，$g_\alpha = \text{Median}(g_e^1, \cdots, g_e^n)$。如图 4 所示，通过加入 $N(0,\sigma^2)$ 来平滑恶意梯度与良性梯度之间的距离，所加入的高斯噪声的 σ 取决于敏感因子 $\Delta_2 f$ 的设置并采用裁剪边界作为高斯噪声的边界，以此来削弱后门攻击。

4.6 联邦学习分布式后门防御

如 4.2～4.5 节所讨论的，联邦学习分布式后门防御方案主要由四部分构成：TernGrad、AdaptDBSCAN、自适应裁剪、自适应加噪。本节将详细介绍联邦后门攻击防御算法。为了进一步抵御联邦后门攻击，本文提出了分布式后门防御方案 DAGUARD，用于解决在联邦学习下的分布式后门攻击。该方案采用 TernGrad 来解决恶意客户端对局部梯度进行操纵的影响，通过自适应 DBSCAN 将恶意梯度与良性梯度聚类后剔除，并通过中位数定理设置自适应裁剪边界，对超过良性梯度大小的梯度进行裁剪，最后通过加入自适应高斯噪声来削弱联邦后门攻击的影响。分布式后门防御（DAGUARD）如算法 3 所示。

算法 3 DAGUARD

输入 所有客户端的更新梯度 G，所有客户端更新的二范数 G_e

输出 全局模型更新 g_b

1) $G = [g^1, \cdots, g^n]_n$， $G_e = [g_e^1, \cdots, g_e^n]_n$
2) $\text{inds} = [c_1, \cdots, c_m]_m = \text{AdaptDBSCAN}(G)$
3) $G_\alpha = \text{Median}(G_e)$ //自适应裁剪边界
4) for t in $[1, m]$ do
5) $g_b^{c_i} = \text{Sign}(g^{c_i})\text{Min}(g_\alpha, g_e^{c_i})$ //裁剪后梯度
6) end for
7) $\sigma = \frac{\Delta_2 f}{\varepsilon}\sqrt{2\ln\frac{1.25}{\delta}} = \frac{g_\alpha}{\varepsilon}\sqrt{2\ln\frac{1.25}{\delta}}$
8) $g_b = \frac{1}{m}\sum_{i=1}^{m} g_b^{c_i} + N(0,\sigma^2)$ //添加高斯噪声
9) 向各客户端发送全局模型更新 g_b

5 实验性能评估

5.1 实验设置

本节实验环境为 Intel(R) Xeon(R) Gold 64 内核，2.3 GHz，128 GB 内存，Ubuntu 16.04 服务器，PyTorch 深度学习框架。

5.1.1 数据集

本节实验基于 2 个图像识别任务来评估 DAGUARD 的性能：MNIST[35]的数字识别任务和 FASHION[36]的图像分类任务。MNIST 数据集由 10 个类组成，共包括 60 000 个图像训练样本和 10 000 个图像测试样本，且每个图像样本都是 28 像素×28 像素。FASHION 数据集由 10 个类组成，有 60 000 个训练样本和 10 000 个测试样本，每个样本为 28 像素×28 像素灰度图像。

5.1.2 基线

本文与无后门防御方案 FedAvg[1]、Yin 等[11]提出的 Median 中位数防御方案和 Nguyen 等[20]提出的 FLAME 防御方案进行了实验对比。为了证明本文提出的 DAGUARD 防御的有效性，本文分别在不同量级的数据集上进行了一系列实验，MNIST 和

FASHION 使用的是 CNN，其结构主要包括全连接层、池化层、ReLU 函数和卷积层。

5.1.3　评估指标

为了衡量 DAGUARD 的防御能力，本节使用以下指标来评估。

1) 后门任务攻击的成功率（BA）。BA 表示在后门任务中对全局模型的攻击成功率，若模型针对目标样本的预测结果为恶意客户端所选择的错误输出，则代表攻击成功。对手的目标是最大化 BA，而有效的防御可以阻止对手增加 BA。

2) 后门任务攻击平均成功率（ABA）。ABA 表示后门任务中对全局模型所有迭代轮次的攻击成功率的平均值。

3) 主任务准确率（MA）。MA 表示模型在对良性数据集进行预测时的准确性。对手的目标是最小化对 MA 的影响以减小被检测到的机会。防御系统不应该对 MA 产生负面影响。

4) 数据投毒率（PDR）。PDR 表示训练数据集的中毒数据的比例。高 PDR 可能会增加 BA，然而也可能使恶意模型与良性模型更容易被区分，从而更容易被发现。

5) 非独立同分布比例（NIR）。NIR 表示数据集中非独立同分布数据所占的比例。

5.2　实验性能

5.2.1　不同数据集的防御性能对比

为了衡量本文提出的 DAGUARD 方案的实用性，本文在不同数据集下与目前主流的防御方案 Median 和 FLAME 以及联邦聚合策略 FedAvg 进行对比，使用 MA 和 BA 来衡量各方案的防御性能表现，各方案在不同数据集中的 MA 表现分别如图 5 和图 6 所示；各方案在 MINIST 数据集和 FASHION 数据集的 BA 对比分别如表 1 和表 2 所示。

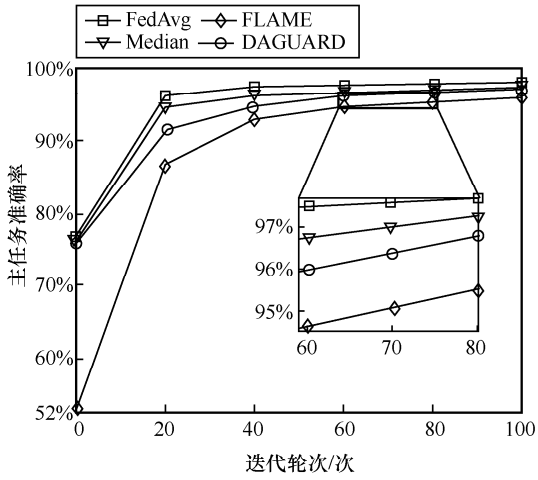
实验设置每轮服务器挑选 10 个客户端，其中恶意客户端数量 num=4，数据集分别为 MNIST 和 FASHION，客户端数量设置为 100，批处理大小设置为 64。在 MNIST 数据集下的 MA 对比实验结果如图 5(a)～图 5(c)和图 6(a)～图 6(c)所示，其中，图 5(a)～图 5(c)为不同 PDR 条件下的 MA 变化趋势，图 6(a)～图 6(c)为不同 NIR 条件下的 MA 变化趋势，可以看出，本文方案的 MA 表现要优于 FLAME 方案，但各方案在 MA 上的表现相差不大，在模型达到收敛后，MA 表现最差的方案与表现最好的 FedAvg 方案差距不超过 2%。从表 1 可以明显看出各方案在 BA 上的差别，FedAvg 方案由于没有进行后门攻击的防御，其 BA 很快就达到 100%，即被完全攻破；由于 Median 方案通过对模型更新取中位数作为更新的值，在一定程度限制了更新的幅度，在特定设置下对联邦后门攻击起到了防御作用，但是 Median 方案在 PDR=0.312 50、不同 NIR 条件下和 NIR=0.25、不同 PDR 条件下的 BA 变化幅度很大，且当 PDR=0.312 50、NIR=0.25 时，BA 达到了 64.15%；当 NIR=0.25、PDR=0.468 75 时，

表 1　　**各方案在 MINIST 数据集的 BA 对比**

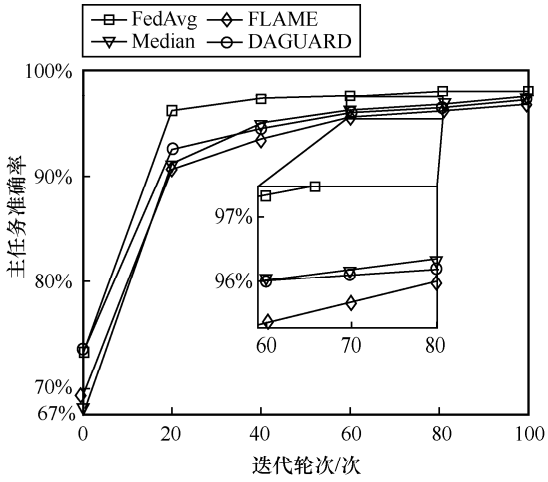
防御方案	NIR（PDR=0.312 50）			PDR（NIR=0.25）		
	0.25	0.50	0.75	0.156 25	0.312 50	0.468 75
FedAvg	100%	100%	100%	100%	100%	100%
Median	64.15%	6.18%	10.80%	6.23%	64.15%	56.72%
FLAME	1.58%	1.31%	1.98%	45.52%	1.58%	1.56%
DAGUARD	1.17%	1.08%	1.02%	0.95%	1.90%	1.79%

表 2　　**各方案在 FASHION 数据集的 BA 对比**

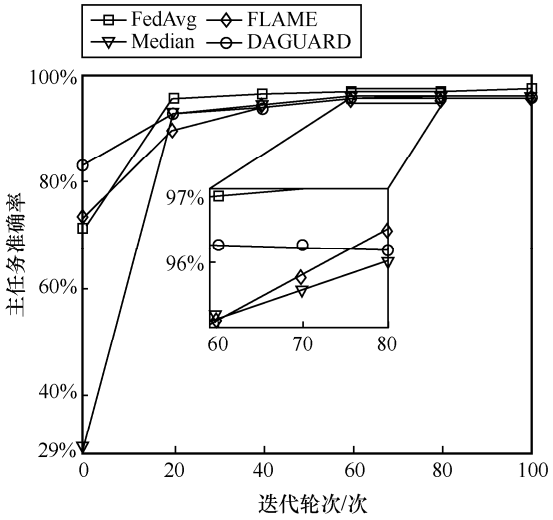
防御方案	NIR（PDR=0.312 50）			PDR（NIR=0.25）		
	0.25	0.50	0.75	0.156 25	0.312 50	0.468 75
FedAvg	100%	100%	100%	100%	100%	100%
Median	99.96%	99.92%	100%	99.94%	99.96%	99.96%
FLAME	30.15%	17.75%	54.15%	24.16%	30.15%	26.26%
DAGUARD	11.4%	10.92%	17.44%	17.61%	11.40%	13.55%



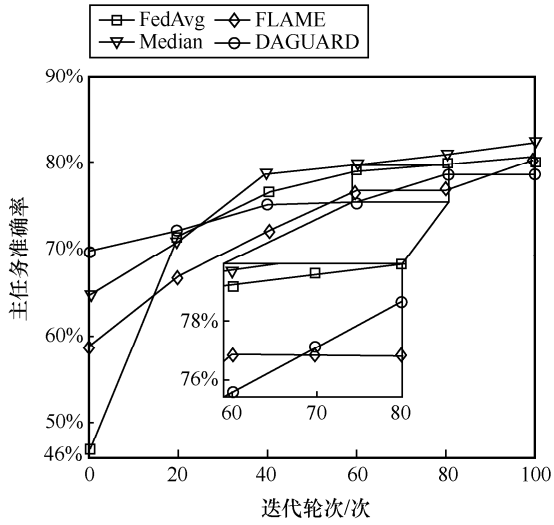
(a) MNIST (num=4, NIR=0.25, PDR=0.156 25)



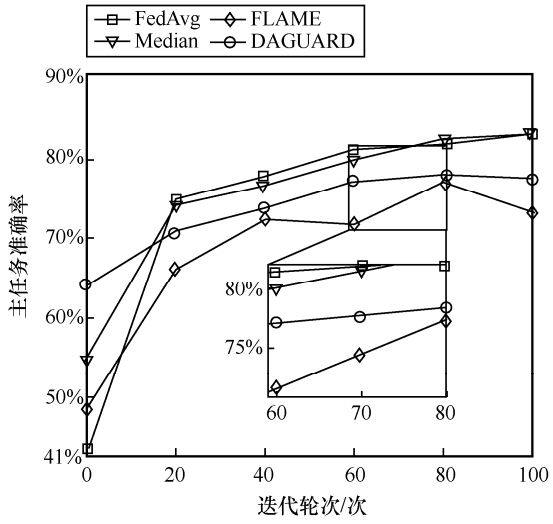
(b) MNIST (num=4, NIR=0.25, PDR=0.312 50)



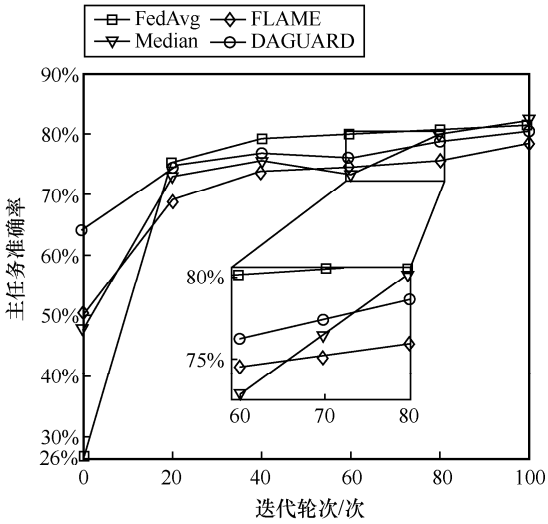
(c) MNIST (num=4, NIR=0.25, PDR=0.468 75)



(d) FASHION (num=4, NIR=0.25, PDR=0.156 25)

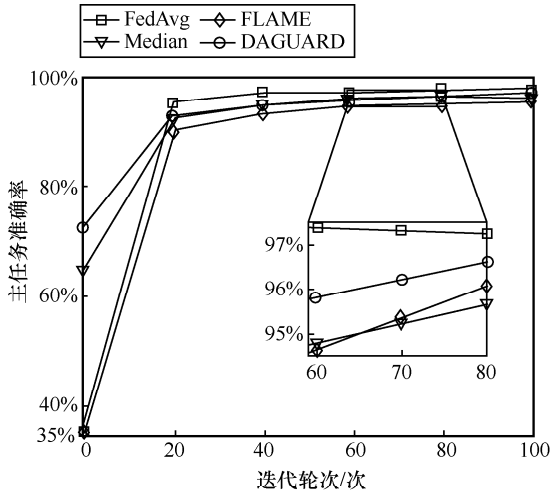


(e) FASHION (num=4, NIR=0.25, PDR=0.312 50)

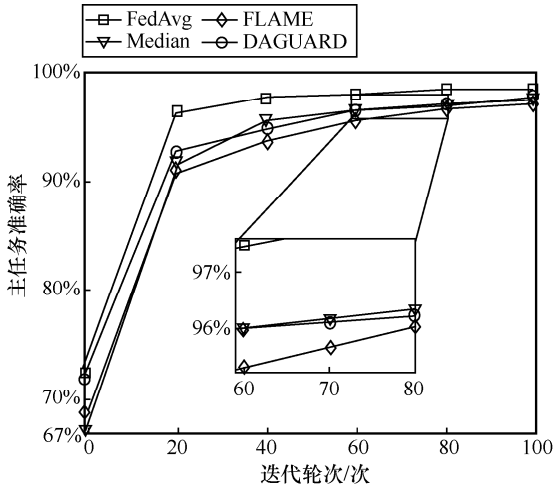


(f) FASHION (num=4, NIR=0.25, PDR=0.468 75)

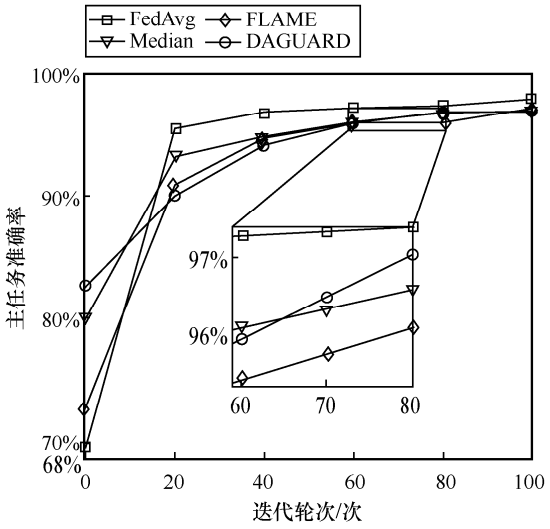
图 5　各方案在不同数据集和不同 PDR 条件下的 MA 表现



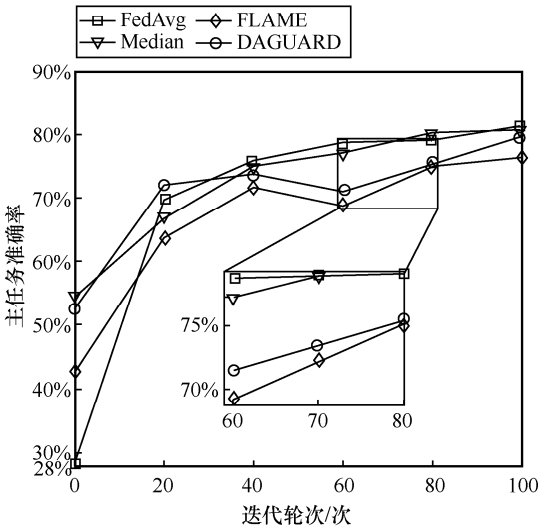
(a) MNIST (num=4, NIR=0.25, PDR=0.312 50)



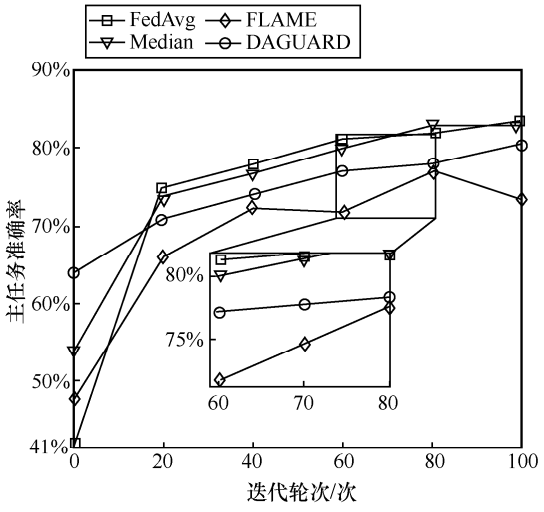
(b) MNIST (num=4, NIR=0.50, PDR=0.312 50)



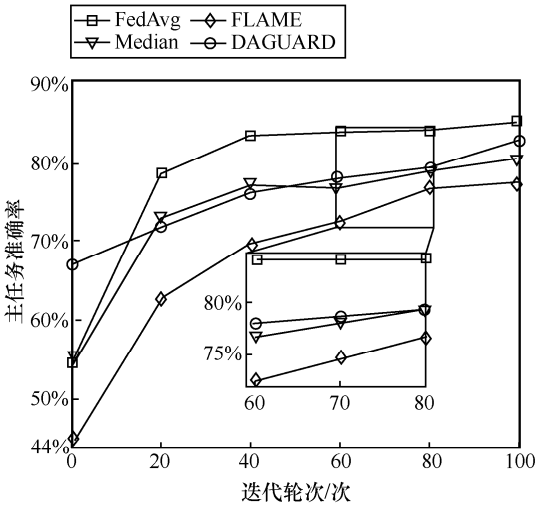
(c) MNIST (num=4, NIR=0.75, PDR=0.312 50)



(d) FASHION (num=4, NIR=0.25, PDR=0.312 50)



(e) FASHION (num=4, NIR=0.50, PDR=0.312 50)



(f) FASHION (num=4, NIR=0.75, PDR=0.312 50)

图 6　各方案在不同数据集和不同 NIR 条件下的 MA 表现

BA 达到了 56.72%；相比 Median 和 FLAME 采用的 HDBSCAN，本文方案将恶意梯度与良性梯度聚成不同簇，并采用裁剪和加噪来限制模型更新，对联邦后门攻击起到了更好的防御作用。虽然 FLAME 和本文方案的 BA 总体上较接近，但本文在 MA 胜出的情况下 BA 仍然更低，当 PDR=0.156 25、NIR=0.25 时，FLAME 方案的 BA 甚至达到了 45.52%，表现不够稳定；而本文的 DBSCAN 相比 HDBSCAN 具有更好的聚类效果，且采用 TernGrad 方式有效限制了缩放攻击，相比 FLAME 方案具有更好的防御能力。实验结果表明，在 MINIST 数据集中，本文方案具有更好的防御表现。

在 FASHION 数据集下的实验对比如图 5(d)～图 5(f)和图 6(d)～图 6(f)所示，其中，图 5(d)～图 5(f)为不同 PDR 条件下的 MA 变化趋势，图 6(d)～图 6(f)为不同 NIR 条件下的 MA 变化趋势，本文方案在 MA 的表现上仍优于 FLAME 方案，并且本文方案与 MA 中表现最好的无防御 FedAvg 方案相差不大。表 1 中，由于 FASHION 数据集的样本特征比 MNIST 数据集更多，Median 方案的 BA 均在 99.90%左右，完全无法抵御联邦后门攻击；而 FLAME 受到 FASHION 数据集的影响相比本文方案更大，当 NIR=0.75、PDR=0.312 50 时，BA 达到了 54.15%，且在不同设置下，BA 大部分超过了 20%；相比之下，本文方案在不同的设置下 BA 均未超过 20%，且大部分不超过 14%，考虑到在 MA 上本文方案表现相比 FLAME 更好，因此在 FASHION 数据集中，本文方案仍然具有更好的防御能力。

5.2.2 不同 PDR 的防御稳定性对比

实验设置客户端数量为 100，批处理大小为 64，每轮服务器挑选 10 个客户端，每轮投毒客户端数量设置为 4，NIR 分别设置为 0.25、0.50 和 0.75。不同 PDR 与 NIR 下各方案在 MINIST 数据集和 FASHION 数据集的 ABA 对比如表 3 所示。本节对比不同 PDR 下各方案的表现。

在 MINIST 数据集中，FedAvg 方案的 ABA 在各参数条件下均超过 94%，无法抵御后门攻击；Median 方案的 ABA 随着 PDR 的增加而上升，甚至一度超过 10%，防御效果欠佳；FLAME 方案的防御表现不够稳定，在 PDR=0.156 25 时其 ABA 表现较差，在其他情况下也都略差于本文方案。

在 FASHION 数据集中，FedAvg 方案仍然无法抵御后门攻击；Median 方案的 ABA 最低超过 40%，最高超过 60%，防御效果较差；FLAME 方案依然不够稳定，在 PDR=0.312 50 时其 ABA 一度超过了 20%；而本文方案的 ABA 始终控制在 10%以下，性能好且比较稳定。

5.2.3 不同 NIR 的防御稳定性对比

实验设置客户端数量为 100，批处理大小为 64，每轮服务器挑选 10 个客户端，每轮投毒客户端数量设置为 4，PDR 分别设置为 0.156 25、0.312 50 和 0.468 75，如表 3 所示。本节对比不同 NIR 下各方案的表现。

在 MINIST 数据集中，FedAvg 方案无法抵御后

表 3 不同 PDR 与 NIR 下各方案在 MNIST 和 FASHION 的 ABA 对比

防御方案	PDR	MNIST			FASHION		
		NIR=0.25	NIR=0.50	NIR=0.75	NIR=0.25	NIR=0.50	NIR=0.75
FedAvg	0.156 25	95.23%	94.71%	94.65%	96.92%	91.05%	91.06%
	0.312 50	98.18%	94.30%	94.28%	93.24%	97.02%	97.02%
	0.468 75	97.46%	98.46%	98.36%	97.88%	98.62%	98.62%
Median	0.156 25	2.58%	2.19%	2.17%	47.23%	43.51%	43.52%
	0.312 50	12.92%	2.81%	2.71%	66.17%	62.75%	62.74%
	0.468 75	15.42%	9.32%	9.29%	56.50%	61.21%	61.19%
FLAME	0.156 25	11.16%	0.98%	0.97%	9.54%	7.81%	7.80%
	0.312 50	1.59%	0.83%	0.82%	9.38%	27.24%	27.22%
	0.468 75	1.74%	0.86%	0.85%	11.89%	7.25%	7.23%
DAGUARD	0.156 25	1.27%	0.75%	0.73%	9.38%	6.37%	6.35%
	0.312 50	0.85%	0.99%	0.98%	9.04%	6.31%	6.29%
	0.468 75	0.84%	0.78%	0.80%	6.60%	6.90%	6.85%

门攻击；Median 方案在 NIR=0.25 时防御表现欠佳；FLAME 方案的防御表现不够稳定，在 NIR=0.25 时其 ABA 表现较差，在其他情况下也都略差于本文方案。

在 FASHION 数据集中，FedAvg 方案仍然无法抵御后门攻击；Median 方案的 ABA 最低超过 40%，最高超过 60%，防御效果较差；FLAME 方案依然不够稳定，在 NIR=0.50 和 0.75 时，其 ABA 甚至将近 30%；而本文方案的 ABA 始终控制在 10%以下，防御表现较好并且十分稳定。

综上所述，相比 Median、FedAvg 以及 FLAME 的防御方案，本文方案在不同实验设置下都有较高的 MA、较低的 BA 以及最稳定的表现，对联邦学习下的分布式后门攻击有较好的防御能力。实验结果表明，本文方案的综合防御性能优于目前的主流方案。

6　结束语

本文提出了一种联邦学习下的分布式后门攻击的防御方案（DAGUARD）旨在解决联邦学习中存在的分布式后门攻击问题。针对现有的防御策略，考虑在恶意客户端不超过半数的情况下，本文设计了 TernGrad 防御、自适应密度聚类 DBSCAN、基于 TernGrad 的梯度自适应裁剪以及梯度自适应加噪方案。实验结果表明，相比无防御策略的 FedAvg 方案、Median 防御方案以及 FLAME 防御方案，DAGUARD 具有更强的防御能力，且在不同的 NIR 和 PDR 下具有更高的稳定性。

参考文献：

[1] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Artificial intelligence and statistics. New York: PMLR, 2017: 1273-1282.

[2] LIU Y, FAN T, CHEN T J, et al. FATE: an industrial grade platform for collaborative learning with data protection[J]. The Journal of Machine Learning Research, 2021, 22(1): 10320-10325.

[3] KURUPATHI S R, MAASS W. Survey on federated learning towards privacy preserving AI[C]//Proceedings of Computer Science & Information Technology (CS & IT). Chennai: AIRCC Publishing Corporation, 2020: 235-253.

[4] BOGDANOVA A, NAKAI A, OKADA Y, et al. Federated learning system without model sharing through integration of dimensional reduced data representations[J]. arXiv Preprint, arXiv: 2011.06803, 2020.

[5] BIGGIO B, NELSON B, LASKOV P. Poisoning attacks against support vector machines[J]. arXiv Preprint, arXiv: 1206.6389, 2012.

[6] NELSON B, BARRENO M, CHI F J, et al. Exploiting machine learning to subvert your spam filter[C]//Proceedings of the 1st USENIX Workshop on Large-Scale Exploits and Emergent Threats. Berkeley: USENIX Association, 2008: 1-9.

[7] FANG M H, CAO X Y, JIA J Y, et al. Local model poisoning attacks to Byzantine-robust federated learning[C]//Proceedings of the 29th USENIX Conference on Security Symposium. Berkeley: USENIX Association, 2020: 1623-1640.

[8] BHAGOJI A N, CHAKRABORTY S, MITTAL P, et al. Analyzing federated learning through an adversarial lens[C]//International Conference on Machine Learning. New York: PMLR, 2019: 634-643.

[9] XIE C, HUANG K, CHEN P Y, et al. DBA: distributed backdoor attacks against federated learning[C]//Proceedings of the 8th International Conference on Learning Representations. [S.l.]: OpenReview, 2020: 1-19.

[10] BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learning[C]//International Conference on Artificial Intelligence and Statistics. New York: PMLR, 2020: 2938-2948.

[11] YIN D, CHEN Y, RAMCHANDRAN K, et al. Byzantine-robust distributed learning: towards optimal statistical rates[C]//International Conference on Machine Learning. New York: PMLR, 2018: 5650-5659.

[12] BLANCHARD P, EL-MHAMDI E M, GUERRAOUI R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 118-128.

[13] NGUYEN T D, RIEGER P, MIETTINEN M, et al. Poisoning attacks on federated learning-based IoT intrusion detection system[C]//Proceedings of 2020 Workshop on Decentralized IoT Systems and Security. Reston: Internet Society, 2020: 1-7.

[14] SHOKRI R, STRONATI M, SONG C Z, et al. Membership inference attacks against machine learning models[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 3-18.

[15] GANJU K R, WANG Q, YANG W, et al. Property inference attacks on fully connected neural networks using permutation invariant representations[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2018: 619-633.

[16] PYRGELIS A, TRONCOSO C, CRISTOFARO E D. Knock knock, who's there? membership inference on aggregate location data[J]. arXiv Preprint, arXiv: 1708.06145, 2017.

[17] CHEN Y D, SU L L, XU J M. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent[C]//Proceedings of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems. New York: ACM Press, 2018: 96.

[18] XU J, HUANG S, SONG L, et al. SignGuard: Byzantine-robust federated learning through collaborative malicious gradient filtering[J]. arXiv Preprint, arXiv: 2109.05872, 2021.

[19] SHEN S Q, TOPLE S, SAXENA P. Auror: defending against poisoning attacks in collaborative deep learning systems[C]//Proceedings of the 32nd Annual Conference on Computer Security Applications. New York: ACM Press, 2016: 508-519.

[20] NGUYEN T D, RIEGER P, CHEN H, et al. FLAME: taming backdoors in federated learning[C]//Proceedings of the 31st USENIX Security Symposium. Berkeley: USENIX Association, 2022: 1415-1432.

[21] WEN W, XU C, YAN F, et al. TernGrad: ternary gradients to reduce

communication in distributed deep learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 1508-1518.

[22] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Palo Alto: AAAI Press, 1996: 226-231.

[23] CAMPELLO R J G B, MOULAVI D, SANDER J. Density-based clustering based on hierarchical density estimates[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2013: 160-172.

[24] HAN J, PEI J, TONG H. Data mining: concepts and techniques[M]. San Francisco: Margan Kaufmann, 2022.

[25] MURTAGH F, CONTRERAS P. Algorithms for hierarchical clustering: an overview[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012, 2(1): 86-97.

[26] KRISHNA K, NARASIMHA M M. Genetic K-means algorithm[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 1999, 29(3): 433-439.

[27] AMINI A, WAH T Y, SAYBANI M R, et al. A study of density-grid based clustering algorithms on data streams[C]//Proceedings of 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Piscataway: IEEE Press, 2011: 1652-1656.

[28] DWORK C. Differential privacy: a survey of results[C]//International Conference on Theory and Applications of Models of Computation. Berlin: Springer, 2008: 1-19.

[29] HUANG Z H, HU R, GUO Y X, et al. DP-ADMM: ADMM-based distributed learning with differential privacy[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1002-1012.

[30] DWORK C, ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, 2013, 9(3/4): 211-407.

[31] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for privacy-preserving machine learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2017: 1175-1191.

[32] ANDERSON A G, BERG C P. The high-dimensional geometry of binary neural networks[J]. arXiv Preprint, arXiv: 1705.07199, 2017.

[33] SUN Z, KAIROUZ P, SURESH A T, et al. Can you really backdoor federated learning?[J]. arXiv Preprint, arXiv: 1911.07963, 2019.

[34] DU M, JIA R, SONG D. Robust anomaly detection and backdoor attack detection via differential privacy[J]. arXiv Preprint, arXiv: 1911.07116, 2019.

[35] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[36] XIAO H, RASUL K, VOLLGRAF R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms[J]. arXiv Preprint, arXiv: 1708.07747, 2017.

[作者简介]

余晟兴（1995− ），男，福建福州人，北京大学博士生，主要研究方向为机器学习、隐私保护、区块链、可验证计算等。

陈泽凯（1998− ），男，广东汕头人，福州大学硕士生，主要研究方向为安全多方计算、联邦学习等。

陈钟（1963− ），男，江苏徐州人，博士，北京大学教授、博士生导师，主要研究方向为网络与信息安全、区块链等。

刘西蒙（1988− ），男，陕西西安人，博士，福州大学教授、博士生导师，主要研究方向为云安全、应用密码学和大数据安全等。