计算机应用研究
Application Research of Computers
ISSN 1001-3695,CN 51-1196/TP

《计算机应用研究》网络首发论文

题目：一种基于联邦学习参与方的投毒攻击防御方法
作者：刘金全，张铮，陈自东，曹晟
DOI：10.19734/j.issn.1001-3695.2023.07.0340
收稿日期：2023-07-12
网络首发日期：2023-11-02
引用格式：刘金全，张铮，陈自东，曹晟．一种基于联邦学习参与方的投毒攻击防御方法[J/OL]．计算机应用研究．
https://doi.org/10.19734/j.issn.1001-3695.2023.07.0340

网络首发时间：2023-11-02 15:09:40
网络首发地址：https://link.cnki.net/urlid/51.1196.TP.20231102.1032.022

一种基于联邦学习参与方的投毒攻击防御方法 *

刘金全[1]，张 铮[1]，陈自东[2]，曹 晟[2†]

(1. 国能大渡河大数据服务有限公司 数据安全组，成都 610041; 2. 电子科技大学 计算机科学与工程学院(网络空间安全学院)，成都 611731)

摘 要：联邦学习分布式的训练结构易受到投毒攻击的威胁，现有方法主要针对中央服务器设计安全聚合算法以防御投毒攻击，但要求中央服务器可信且中毒参与方数量需低于正常参与方。为了解决上述问题，提出了一种基于联邦学习参与方的投毒攻击防御方法，将防御策略的执行转移到联邦学习的参与方。首先，每个参与方独立构造差异损失函数，通过计算全局模型与本地模型的输出并进行误差分析，得到差异损失权重与差异损失量。其次，依据本地训练的损失函数与差异损失函数进行自适应训练。最终，依据本地模型与全局模型的性能分析进行模型选取，防止中毒严重的全局模型干扰正常参与方。在 MNIST 与 FashionMNIST 等数据集上的实验表明，基于提出的算法的联邦学习训练准确率优于 DnC 等投毒攻击防御方法，在中毒参与方比例超过一半时，正常参与方仍能够实现对投毒攻击的防御。

关键词：联邦学习；投毒攻击防御；训练权重；鲁棒性
中图分类号：TP309.2 doi: 10.19734/j.issn.1001-3695.2023.07.0340

Defense method on poisoning attack based on clients in federated learning

Liu Jinquan[1], Zhang Zheng[1], Chen Zidong[2], Cao Sheng[2†]

(1. *CHN Energy Dadu River Big Data Services Co, Ltd, Data Security Group, Chengdu* 610041, *China*; 2. *School of Computer Science & Engineering* (*School of Cyber Security*), *University of Electronic Science & Technology of China, Chengdu* 611731, *China*)

Abstract: The distributed training structure of federated learning is vulnerable to poisoning attacks. Existing methods mainly design secure aggregation algorithms for central servers to defend against poisoning attacks, but require the central server to be trusted and the number of poisoned participants to be lower than normal participants. To address the above issues, this article proposes a poison attack defense method based on federated learning participants, which transfers the execution of defense strategies to the participants of federated learning. Firstly, each participant independently constructs a differential loss function, calculates the output of the global and local models, and conducts error analysis to obtain the weight and amount of differential loss. Secondly, we perform adaptive training based on the local trained loss function and differential loss function. Finally, our approach selects models based on the performance analysis of local and global models to prevent severely poisoned global models from interfering with normal clients. Experiments on datasets such as MNIST and FashionMNIST have shown that the federated learning training accuracy based on our algorithm is superior to poison attack defense methods such as DnC. Even when the proportion of poisoned participants exceeds half, normal participants can still achieve defense against poison attacks.

Key words: federated learning; poisoning attack defense; training weight; robustness

0 引言

面对信息社会产生的数据孤岛，谷歌在 2017 年首次提出联邦学习技术[1~3]，期望平衡数据价值与隐私保护的矛盾。联邦学习允许相互不信任的参与方在不共享其本地数据的情况下协作训练统一的通用模型或各自的个性化模型。在聚合操作中，参与方被服务器随机选取，并使用本地数据计算模型更新梯度信息，并与其他参与方共享信息，服务器执行聚合算法，并使用聚合梯度更新全局模型。

对于联邦学习算法，如 FedAvg[1]和 FedProx[4]，都在分布式的参与方上计算模型。由于参与方的安全防御能力存在差异，使得联邦学习容易受到投毒攻击的威胁[5, 6]。攻击方控制或毒害部分联邦学习参与方，称为中毒参与方，并将其与中央服务器共享恶意更新，以降低全局模型的性能。在联邦学习中，投毒攻击方的攻击方式主要有 2 类：1)破坏训练数据集，防止模型收敛或往指定的方向收敛，这称为数据投毒攻击[7]；2)构建恶意模型或恶意梯度参与全局模型聚合，干扰全局模型的生成，这称为模型投毒攻击[8]。由于数据存储和模型训练过程在参与方本地进行，所以上述攻击都是针对参与方实施的攻击。

由于中央服务器聚合参与方提交的模型更新，是联邦学习的枢纽，目前联邦学习投毒攻击防御方法研究集中在基于中央服务器设计拜占庭鲁棒聚合算法，识别并剔除可能的中毒参与方。这些防御方法也可分为 3 类：1)利用服务器设计

收稿日期：2023-07-12；修回日期：2023-09-11 基金项目：四川省重点研发计划项目(2021YFG0113，2023YFG0118)
作者简介：刘金全(1987—)，男，重庆人，高级工程师，硕士研究生，主要研究方向为数据安全与隐私计算；张铮(1997—)，男，四川成都人，助理工程师，硕士研究生，主要研究方向为大数据与联邦学习；陈自东(2000—)，男，四川宜宾人，硕士研究生，主要研究方向为鲁棒联邦学习；曹晟(1981—)，男(通信作者)，湖北武汉人，研究员，博导，博士，CCF 高级会员(16703S)，主要研究方向为信息安全与区块链(caosheng@uestc.edu.cn).

聚类算法或权重函数[9~12]，剔除或降低可能的中毒参与方，即与多数参与方上传的更新不一致的参与方。2)分析恶意更新与正常更新的特点，通过先验知识建立假设，服务器将满足此假设的参与方都视为恶意参与方。如 FoolsGold 算法[13]，依赖于恶意更新的随机性低于正常更新的假设，实现了中毒参与方数据量超过一半的投毒攻击防御。3)服务器共享部分测试数据集，利用测试数据集识别上传的恶意更新。但此方法需要向服务器共享部分真实数据集，数据的隐私性受到影响。上述 3 类方法虽然对投毒攻击实现了一定程度的防御，但由于服务器缺少真实数据集，难以对参与方提交的模型更新进行精准的判断，以上方法都不能解决中毒参与方数量高于正常参与方情况下的投毒攻击防御问题。

针对上述问题，本文与之前中央服务器设计防御算法不同，提出了一种基于联邦学习参与方的投毒攻击防御方法。本文算法不依赖于中心服务器进行防御，同时可以保证正常参与方在任意比例中毒参与方环境中的鲁棒性。该方法在 FedAvg 算法的框架下，将参与方视为防御策略的独立执行方，通过参与方额外在本地训练中利用差异计算函数(如均方误差等)计算全局模型参数与参与方本地模型的差异损失权重，并在训练损失函数中嵌入了差异损失权重与差异损失函数，利用全局模型和参与方本地模型的差异进行自适应的个性化训练。在联邦学习结束时，对全局模型以及本地模型进行评估，获取最优的模型。由于参与方为策略的独立执行主体，所以当服务器被攻击或任意比例的参与方被攻击时，未中毒的正常参与方仍可以保证本地模型的鲁棒性。

1　相关工作

1.1　联邦学习

联邦学习参与方代表算力的提供方，通常由个人终端或不同企业与部门组成，负责保存用户或企业的私有数据。参与方进行本地训练并上传参数给聚合服务器，并由聚合服务器对所有参与的参与方进行聚合并同步，开始一轮新的训练。这种联合协作训练的方式可以在保证模型性能的前提下，避免个人数据的泄露，并有效解决数据孤岛的问题。联邦学习常用框架有两种，一种是参与方-服务器架构[14]，另一种是对等网络架构[15]。在参与方-服务器架构中，各个数据拥有方利用本地数据和算力，根据其特定条件和规则，在本地进行模型训练，然后将训练得到信息通过差分隐私或同态加密后，由聚合服务器进行计算。在对等网络架构中，不借助第三方，而是通过参与方的直接通信，降低了由于服务器受到攻击所带来的风险，但是需要更复杂的加解密操作来实现信息共享。目前研究热点更多集中在参与方-服务器框架，最为广泛使用的 FedAvg 算法架构如图 1 所示。

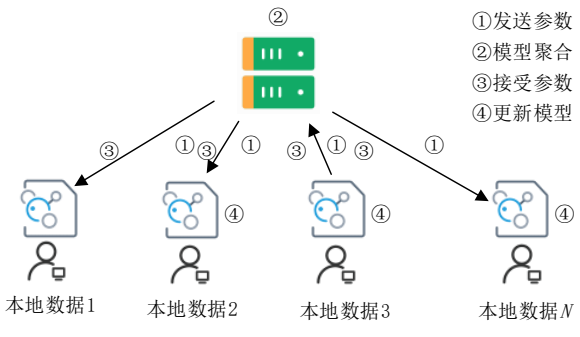

图 1　FedAvg 算法架构

Fig. 1　Fedavg algorithm architecture

在非独立同分布的本地数据集下，参与方共同优化的全局模型如下：

$$\min\{F(\mathrm{w})\} \cong \sum_{i=1}^{N} \psi_i F_i(w) \tag{1}$$

其中，N 为参与方数量，ψ_i 是参与方的权重，F 是训练损失函数。

1.2　联邦学习投毒攻击

恶意攻击方的主要目的是通过执行投毒攻击来控制本地模型的行为，以此影响全局模型。攻击方在参与方的训练或再训练过程中，通过修改训练数据集或模型参数的方式来控制机器学习模型。

数据投毒是指攻击者对训练数据集中的样本进行污染，如翻转标签或添加噪声，降低数据质量，从而影响全局模型。针对联邦学习的数据投毒攻击主要是标签翻转攻击与后门攻击。标签翻转攻击[16, 17]是指将联邦学习参与方的训练数据集标签集中式或分布式进行翻转，使得模型学习错误的对应关系。对于更为隐蔽且危害性更大的后门攻击，则有更为详细的研究。Zhang 等人[18]提出了一种单线后门攻击并将其命名为神经毒素，攻击在训练过程中变化较小的参数以植入后门。此外，Gong 等人[19]中也设计了一种攻击方式，提出了针对联邦学习使用多个局部触发器从而有效协调后门攻击。

1.3　联邦学习投毒攻击的防御方法

针对联邦学习的投毒攻击防御方法主要是设计安全聚合算法[20]，需假定服务器可信，从而利用服务器进行恶意参与方的识别与恶意更新的检测。其中，根据防御方式的不同，可以分为基于模型参数差异的防御与基于验证数据集的防御。基于模型参数差异的防御方法对参与方上传的模型参数进行验证，区分恶意更新与正常更新。Krum 算法和 Multi-Krum 算法在联邦学习的投毒攻击与防御中得到了广泛的应用。

(1)Krum 算法[10]假设参数服务器在每次迭代中掌握拜占庭参与方的数量信息，聚合时仅选择一个参与方模型的更新作为全局模型的更新。参与方更新的选择策略为

$$u_i^t = w_i^t - w^{t-1} \tag{2}$$

$$u^* = \arg\min \sum_{u_j \in \Omega_{j,\tau-f-2}} \| u_i - u_j \|_{i=1,2,\ldots,\tau}^2 \tag{3}$$

其中，$\Omega_{j,\tau-f-2}$ 是 $\tau-f-2$ 个距离模型 j 最近的模型更新的集合，其中距离使用欧几里德距离。而 Multi-Krum 是 Krum 算法的变体，它收集 $\tau-f-2$ 个距离最近的参与方的模型更新，并将它们聚合进行全局模型的更新。

(2)DnC 算法[21]利用基于奇异值分解(singular value decomposition，SVD)的谱方法来检测和去除异常值。为了减轻直接在高维梯度上执行 SVD 开销巨大的缺陷，DnC 通过对其输入梯度进行随机抽样来降低维数。

首先，随机选取一个维度构造该维度下梯度的下采样集，并使用下式计算中心下采样集。

$$\nabla^{\mathrm{c}} = \nabla_i - \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \nabla_i \tag{4}$$

其中，n 是参与方的个数，▽表示梯度的下采样集。利用中心下采样集的奇异特征向量计算离群值向量 s_i。

$$s_i = \left(\left\langle \nabla_i - \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \nabla_i, v \right\rangle \right)^2 \tag{5}$$

其中，v 表示中心下采样梯度的投影。利用 s_i 移除具有最高分数的部分梯度，得到良性梯度集并进行梯度聚合，实现投毒攻击的防御。

但上述算法都有恶意参与方必须小于正常参与方数量的

限制。除本文提出的参与方防御算法外，目前突破限制的方法分为两类[22, 23]，一类利用正常更新的特点进行分析，如通过余弦相似度计算相似度分数选取参与方进行聚合等，或依据恶意更新的多样性低于正常更新的特点[13]，取消对恶意方数量的限制，根据历史更新与参与方的最大余弦相似度调整权重并计算聚合结果。另一类则需要额外引入验证数据集，需要聚合服务器拥有部分或相似的正常样本，利用准确率进行判断。但通过聚合服务器或参与方节点评估进行交叉验证[24]，数据集隐私保护以及协调聚合服务器计算存在困难。

2　模型假设

2.1　威胁模型

对投毒攻击方的能力做了 3 点假设：①攻击方可以秘密修改参与方的训练数据集。②攻击方可以对任意数量的参与方进行投毒攻击。③攻击方可以对参与方发送服务器的模型参数信息进行修改。

2.2　防御目标

从保真性、鲁棒性 2 个方面来评估本文算法。

①保真性：由于全局模型来自于本地模型的聚合，所以不存在投毒攻击时，本文提出的算法相比于 FedAvg 算法具有接近的性能。

②鲁棒性：在存在投毒攻击的环境中，随着投毒方比例的增加，本地模型逐步降低对全局模型的学习权重，从而本文提出的算法能够降低投毒数据对参与方的影响。

2.3　防御方能力

每个参与方都为防御方，防御方需要降低投毒攻击方对本地模型的影响。这里提出 4 点假设：①服务器无法访问参与方本地训练数据集。②在每轮迭代中，服务器可以获取到每个参与方的本地模型。③每轮迭代中，防御方不可知自己是否被成功投毒攻击。④参与方之间的样本独立同分布。

3　研究方法

本文改进了参与方进行联邦训练的方法。中毒参与方与正常参与方完成联邦训练后将参数传递给服务器进行聚合。参与方通过本地模型与全局模型的相似度来决定对全局模型的相信程度，从而判断全局模型被投毒攻击的严重程度。

3.1　基于参与方的投毒攻击防御算法框架

本算法的训练方式包括 4 个部分：聚合(平均收集的模型参数以获取全局模型)，加权(参与方计算模型的差异度构建差异损失权重)、训练(通过均方误差构建本地损失与差异损失并进行协同优化，更新本地模型参数)和模型选取(测试性能，选取目标模型，并反馈给服务器)。图 2 描述了本文聚合算法的具体流程，其中 M 代表本地模型参数，GM 表示全局模型参数。

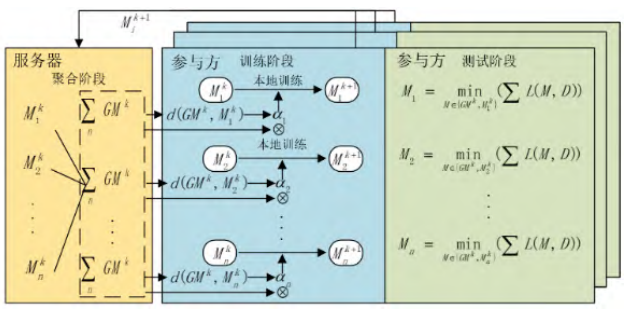

图 2　基于参与方的投毒攻击防御算法框架

Fig. 2　Participatory based algorithm framework for poisoning attack defense

服务器仅对参与方模型进行平均，参与方首先基于梯度下降算法对本地数据集训练，生成该轮迭代需要贡献的本地模型参数。服务器对各个参与方的模型参数使用联邦平均算法进行简单聚合后，下传给各参与方。定义本地模型训练的损失函数由两个部分的加权平均构成，这两个部分为模型训练正常样本产生的损失以及本地模型与全局模型的距离构成。其中，后一部分损失的权重由本地模型与全局模型的输出决定。

1) 聚合。在防御方能力①与防御方能力②的假设下，每一轮全局更新中，参与方仅上传自己的本地模型，而服务器的功能为构建全局模型，使得全局模型充分聚合多个参与方模型的信息。由于攻击方能力的假设③使得服务器不可信赖，所以无法使用服务器进行投毒模型与正常模型的判断。对于全局模型的聚合，本文参考 FedAvg 聚合算法，不存在投毒攻击时，本地模型可以从正确的全局模型中学习，满足了防御目标①保真性。

2) 加权。考虑投毒攻击的隐秘性，参与方难以对全局模型是否受到投毒攻击进行校验，因此参与方依赖本身的信息对全局模型进行评估，并得到全局模型的差异损失权重，即信赖程度。为了衡量全局模型与参与方模型差异在损失函数中所占的权重，本文构建了差异损失函数权重，差异损失权重定义为

$$\alpha_i = \frac{1}{1+\|GM^{k-1}, M_i^k\|^2} \tag{6}$$

其中，GM^{k-1} 表示第 $k-1$ 轮聚合后的全局模型，M_i^k 表示第 i 个参与方在第 k 轮聚合中的参与方模型。相较于对模型输出直接求差异，通过 L2 范式对全局模型与参与方模型的参数向量求差异更能直接衡量两个模型的差异量。由上式 4 可知，参与方模型的参数向量与全局模型的参数向量的差异量，与差异损失权重 α_i 呈负相关，当差异量减少时，权重则相应增加。通过控制差异损失权重，在本地模型与全局模型差异较大时，参与方降低对全局模型的学习率，并在训练完成后配合模型选取，实现防御目标中②鲁棒性。

3) 训练。在为差异损失分配权重后，构建损失函数开始训练。损失函数分为两个部分，分别为本地损失函数 F_1 与联邦差异损失函数 F_2。其中本地损失函数计算本地模型输出与真实标签的差异，而联邦差异损失函数则衡量本地模型输出于全局模型输出的差异。总的损失函数如下所示。

$$F_{adv}(M^k) = F_1(M^k, x_i, y_i) + \alpha_i F_2(GM^k, M^k, x_i) \tag{7}$$

F_1 函数可通过均方误差或交叉熵的方式计算参与方模型 M^k 的输出，并与真实标签 y_i 进行对比，计算得到参与方模型在 x_i 样本中产生的损失。F_2 函数同理，采用模型输出差异的对比来估计参与方模型 M^k 与全局模型 GM^k 的差异，并配合差异损失权重，实现自适应的训练。在参与方最信赖参与方模型的基础上，如果偏差值过大，则差异权重损失会相应降低，即降低对全局模型的差异损失函数的比重。

4) 模型选取。在上述多次联邦训练后，产生了由多数参与方支持的全局模型与少数参与方本地训练的个性化模型。由于防御方能力③的假设，部分被投毒的参与方可能训练了投毒模型。为了更好的让全局参与方的收益达到最优，需要利用测试集对全局模型与参与方模型进行测试，计算其在测试集上的准确率。模型选取的公式如下：

$$\min_{M\in\{M^k, GM^k\}} \sum_{x,y\in D} F_1(M, x, y) \tag{8}$$

其中，D 为测试数据集，x、y 为测试样本，M 为模型参数。求得最优的模型参数 M，使得模型在测试集上的累计损失达

到最小，从而得到最终模型。

3.2　基于参与方的投毒攻击防御算法实现

算法 1、2 分别介绍了聚合服务器与参与方的训练算法。其中，假设训练服务器不可信，所以需要对上传的梯度添加噪声处理，防止信息泄露。算法 1 通过对参与方模型参数的平均，以此聚合生成全局模型。算法 2 中参与方根据本地模型与全局模型的欧式距离来计算参与方对全局模型的置信值，并将置信值映射在[0,1]的区间内。同时优化损失函数，将损失函数考虑为本地损失与全局模型差异的结合，利用置信值与差异损失可以从全局模型中学习到一定比例的信息。

算法 1　中心服务器聚合算法

输入：参与方数量 n，目标迭代次数 m，参与方发送的参数 M_1，M_2，…，M_n

输出：聚合之后的参数 GM

a) i=1;//初始化当前聚合次数

b) M=receive();//接收参与方得到的参数

c) GM=1/n*(sum(M));//对参数进行求和并平均

d) if i>m

e)　return GM, false;//达到停止条件,停止聚合

f) end if

g) i=i+1;

h) return GM, true;

算法 2　参与方训练算法

输入：训练数据 D，测试样本 T，样本数量 N，学习率 η，中毒数据 N*，本地迭代轮次 k

输出：训练完成的本地模型 M

a) GM, flag=receive();

b) M=GM;D←N*;

c) if 第一轮迭代

d)　for i=1,2,…,N do//第一轮迭代，未添加差异损失权重

e)　(xi,yi)~D;//数据集样本采样

f)　$L_{adv}=CE(M^k,x_i,y_i)$;//CE 为交叉熵

g)　$M^{k+1}\leftarrow M^k-\eta\cdot\nabla\theta(L_{adv}(M^k))$;//梯度下降算法

h)　end for

i) else if 第二轮以上迭代

j)　for k=1,2,…,K do

k)　for i=1,2,…,N do

l)　(xi,yi)~D;//数据集样本采样

m)　$\alpha_i=\frac{1}{1+||GM^k(\mathrm{x}_i),M^k(\mathrm{x}_i)||^2}$;//权重系数

n)　$L_{adv}=CE(M^k,x_i,y_i)+\alpha_i MSE(GM^k,M^k,x_i)$;

o)　$M^{k+1}\leftarrow M^k-\eta\cdot\nabla\theta(L_{adv}(M^k))$;//参数更新

p)　end for

q)　end for

r) end if

s) return M;

4　算法分析

目前针对联邦学习的投毒攻击防御方法主要集中在设计安全的服务器聚合算法[9]，但是上述方法需假设服务器可信，参与方也无须参与投毒攻击的防御过程。本文的算法旨在将参与方加入投毒攻击的防御过程，在假设自身数据可信的情况下，推测全局模型受到毒害的程度，从而确定学习全局模型的权重。本文从以下 3 种情况对联邦训练中正常参与方与中毒参与方的不同比例的情况进行分析。

1) 不存在中毒参与方。正常参与方：本地模型与全局模型差异较小，分配权重系数较大，实现个性化联邦学习训练。

2) 中毒参与方数量小于正常参与方数量

正常参与方：由于正常参与方数量较多，通过联邦平均聚合后，全局模型参数更接近于于正常参与方的模型。全局模型与正常参与方参数差异较小，本地模型对全局模型的差异损失权重较大，从而本地模型学习全局模型的参数。

中毒参与方：当投毒模型参数与全局模型参数之间存在显著差异时，参与方主观认为全局模型受到投毒攻击，并相应地降低本地模型对全局模型的差异损失权重，更加信任本地模型。在联邦学习完成后，参与方会通过比较全局模型和本地模型在测试集上的准确率来检测是否存在本地数据被投毒，并在发现本地数据存在投毒时，选择将本地模型替换为全局模型。

3) 中毒参与方数量大于正常参与方数量

正常参与方：在中毒参与方数量大于正常参与方时，通过联邦平均聚合后，全局模型参数更接近于于中毒参与方的模型，所以性能较差。在经过参数距离计算函数(如 L2 范式等)对全局模型与本地模型的参数距离进行计算后，得到了较小的差异损失权重。正常参与方会认为全局模型被投毒攻击，从而更倾向于减少对全局模型参数的学习。

在联邦训练完成后进行模型选择，由于本地模型只对全局模型的参数学习了少量信息，所以本地模型在测试集上的准确率远远优于全局模型。最终在模型选取步骤，选取得到受投毒攻击影响微弱的本地模型。所以，即使在只有一个正常参与方的极端情况下，正常参与方也能几乎不被投毒攻击影响。

中毒参与方：由于投毒模型参数与全局模型之间存在差异较小，经过参数计算计算函数进行计算后，会得到较高的差异损失权重，本地模型学习更多的全局模型信息。导致最终全局模型与本地模型在测试集的结果都较差，中毒参与方最终只能训练得到被投毒攻击的模型。

综上所述，假设在联邦学习中存在任意数量的中毒参与方，也几乎无法对正常参与方造成影响。由参与方参与到投毒攻击防御中，弱化了服务器在投毒攻击防御中的作用，加强了联邦学习算法的安全性。

5　实验与分析

实验评估了本文算法的保真性和鲁棒性。通过与其他 3 种常见聚合算法：FedAvg、Krum、Multi-Krum 和 DnC 等对比，证明了本文算法在真实联邦计算环境中的可行性。

5.1　实验设置

该节对实验采用数据集、投毒攻击方式、评估指标以及系统设置进行介绍。

5.1.1　数据集

使用 2 个计算机视觉领域的数据集：MNIST 数据集和 FashionMNIST 数据集。对于每个数据集，以均等的概率分发给各个参与方，以模拟真实联邦学习中的各个系统。

1) MNIST 数据集。MNIST 数据集是一个经典的手写数字图像数据集，由 Yann LeCun 等创建。它包含了 60000 个训练图像和 10000 个测试图像，每个图像都是 28x28 像素大小的灰度图像，用于机器学习中的图像分类任务。该数据集已经成为了机器学习和计算机视觉领域中最广泛使用的数据集之一。在实验中，MNIST 数据样本随机分发给各个参与方。

2) FashionMNIST 数据集。不同于 MNIST 手写数据集，

Fashion-MNIST 数据集包含了 10 个类别的图像，分别是：T 恤，牛仔裤，套衫，裙子，外套，凉鞋，衬衫，运动鞋，包，短靴。与 MNIST 相同，图像是一个 28*28 的像素数组，每个像素的值为 0~255 之间的 8 位无符号整数。

5.1.2　投毒攻击方式

使用符号翻转攻击进行模型投毒，使用黑盒边缘攻击进行数据投毒，并分别对 40%的参与方以及 60%的参与方进行毒害。

1) 符号翻转攻击

在符号翻转攻击中，参与方 i 正常训练出本地模型 w_i，然后将其参数翻转后提交给服务器。即提交-w_i 给服务器。在 Krum 算法中，参与方提交模型训练梯度Δw_i 给服务器，在符号翻转攻击中，被攻击的参与方提交-Δw_i 给服务器。

2) 黑盒边缘攻击

在黑盒边缘攻击中，修改参与方的本地训练数据，而不直接窜改模型参数。本文将手写图像的标签进行标签翻转，将所有标签的值依次后移，如将标签“7”改为“8”以制作毒数据。攻击者在混有干净数据和中毒数据的数据集中训练出毒模型。对于黑盒设置，本文将 20%的干净数据和 80%的中毒数据混合在一起作为中毒参与方的数据集。

5.1.3　评估指标

对于 FedAvg、Krum、Multi-Krum 和 DnC 算法，使用全局模型的测试准确率来衡量算法性能。对于本文提出的算法，在中毒参与方少于一半时，采用参与方平均准确率来衡量，但在中毒参与方超过一半时，采用未中毒参与方的平均准确率进行衡量算法性能。除此之外，由于本文算法具有检测并抵御中毒方多于正常方的优点，能否获取有效的模型也是本文的测试指标。在投毒方少于攻击方的测试中，全局模型测试准确率是主要指标，准确率越高，说明模型效果越好。在投毒方数量多于正常参与方的测试中，正常参与方能否训练出合格模型为主要评估方式。

5.1.4　系统设置

实验设置 10 个参与方，每轮选取所有参与方进行全局模型更新。对于 Multi-Krum 算法，每轮随机选取 4 个参与方进行更新。各算法实验中使用相同的模型，并随机初始化模型参数。在符号翻转和黑盒边缘攻击中全局模型均更新 20 次。

5.1.5　对比方法

实验进行了本文与四种方法的对比：

1)FedAvg。一种简单的聚合联邦学习参与方参数的方法，不对投毒攻击行为进行主动防御，用于对比观察本文所提出的算法在不同环境中，防护策略带来的效益。

2)Krum。一种基于欧氏距离的投毒攻击防御经典算法，Kum 在若干本地梯度中选择一个与其余梯度相似度最高的梯度作为全局梯度，从而去除恶意梯度。该方法用于对比在不同投毒攻击比例下的防御效果。

3)Multi-Krum。Krum 算法的改进算法，求相似度最高的若干梯度，并将这些梯度的平均作为全局梯度。该方法用于对比在不同投毒攻击比例下的防御效果。

4)DnC。一种高鲁棒性的投毒攻击防御方法。通过选取部分维度的梯度向量，并计算其均值与奇异特征向量，获取到本地梯度的离群值，并剔除离群值较高的梯度。该方法用于对比在不同投毒攻击比例下的防御效果。

5.2　实验结果

5.2.1　保真性

如图 3 与图 4 所示，在没有攻击的情况下，本文算法的全局模型准确率和 FedAvg、Krum、Multi-Krum、DnC 等算法基本一致，均能取得较好的训练结果。

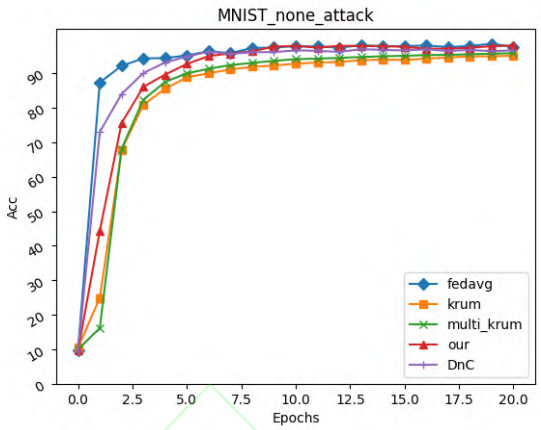

图 3　MNIST 数据集中聚合算法的保真性

Fig. 3　The fidelity of aggregation algorithms in MNIST datasets

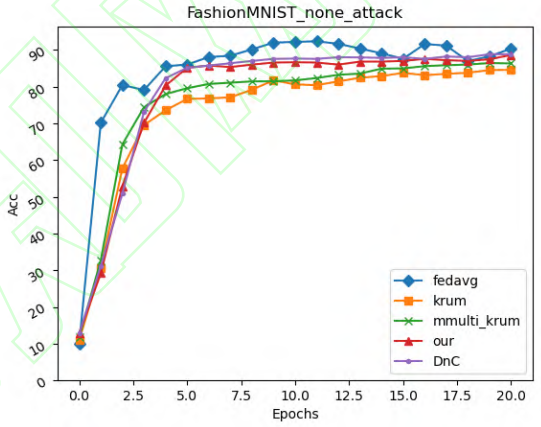

图 4　FashionMNIST 数据集中聚合算法的保真性

Fig. 4　The fidelity of aggregation algorithms in FashionMNIST datasets

由图 4 与图 5 可知，在 MNIST 与 FashionMNIST 数据集上，对于图中所有联邦学习算法，其训练的准确率上升趋势几乎一致，最终达到的模型准确率也几乎一致。对比本文方法与 FedAvg 算法的曲线，可以发现本文算法相较于其他基准算法的最终性能损失较低，在没有投毒攻击的环境更接近于 FedAvg 算法。此外，与 FedAvg 算法相比，最终的训练准确率均出现了少量降低，这是因为无论是 Multi-Krum 算法或 DnC 算法等对部分梯度进行剔除，或者本文算法是对全局模型与本地模型的对比，都不会接聚合恶意的梯度或者恶意的全局模型，由于剔除了部分正常的梯度，使得聚合的梯度多样性减少。在两个不同数据集上的测试结果表明，本文算法相较于其他基准算法具有更高的保真性。

5.2.2　鲁棒性

为了进一步验证本文算法在鲁棒性方面的优势，本文对中毒参与方的比例进行了控制，并使用符号翻转攻击、黑盒边缘攻击方式进行了实验，以验证本文算法在应对各种投毒攻击情况下的防御效果。

在无攻击环境与中毒参与方比例为 40%的投毒攻击环境中，记录每种算法最终训练完成的模型在统一的测试集上的准确率。但在中毒参与方比例超过 60%的符号翻转攻击以及黑盒边缘攻击实验中，由于本文算法会产生多个模型，所以只使用正常参与方模型的平均准确率作为对比指标，而其余对比算法因为只产生一个全局模型，所以使用全局模型在测

试集上的准确率作为对比指标。

表 1 与表 2 的数据显示，Krum、Multi-Krum 以及 DnC 算法虽然在中毒参与方比例为 40%时起到了一定的防御，但都在超过中毒参与方比例达到 60%后，失去了抵御攻击的能力。这与上述算法更相信于多数参与方有关。而 FedAvg 算法因为仅仅进行模型平均，对抵抗属于收敛性攻击的符号翻转攻击效果较差，反而对属于后门攻击的黑盒边缘攻击有一定的抵御能力，这是因为黑盒边缘攻击对模型的参数改动较少。

表 1　MNIST 数据集下联邦训练性能对比(比例 60%的攻击只统计正常参与方准确率)

Tab. 1　Comparison of federated training performance under the MNIST dataset((60% of attacks only count the accuracy of normal participants))

攻击方式	FedAvg(%)	Krum(%)	Multi-Krum(%)	DnC(%)	本文算法(%)
无攻击	97.56	94.96	95.72	96.54	**97.45**
符号翻转攻击(40%)	8.92	94.11	95.15	96.27	**91.55**
符号翻转攻击(60%)	9.10	9.80	9.56	21.89	**93.70(normal)**
黑盒边缘攻击(40%)	75.81	95.34	95.22	95.38	**94.23**
黑盒边缘攻击(60%)	53.35	10.42	10.12	14.48	**94.80(normal)**

表 2　FashionMNIST 数据集下联邦训练性能对比(比例 60%的攻击只统计正常参与方准确率)

Tab. 2　Comparison of federated training performance under the fashionmnist dataset(60% of attacks only count the accuracy of normal participants)

攻击方式	FedAvg(%)	Krum(%)	Multi-Krum(%)	DnC(%)	本文算法(%)
无攻击	90.35	84.57	86.31	88.82	**88.62**
符号翻转攻击(40%)	13.97	82.78	86.10	87.25	**84.29**
符号翻转攻击(60%)	9.02	9.30	9.21	16.83	**85.63(normal)**
黑盒边缘攻击(40%)	70.20	83.22	84.76	85.47	**84.13**
黑盒边缘攻击(60%)	62.16	12.42	8.51	13.32	**85.92(normal)**

对于符号翻转攻击与黑盒边缘攻击，不论是在 MNIST 还是 FashionMNIST 中，在超过 50%的参与方比例下，只有本文的聚合算法可以防御这两种攻击，使得模型的准确率与无攻击情况下的差异较小。在中毒参与方比例为 40%时，本文算法虽然与 Krum、Multi-Krum、DnC 算法的准确率在 MNIST 数据集相比，降低了 2.56%、3.6%与 4.72%，在 FashionMNIST 数据集上降低了 1.51%、1.81%与 2.99%。但在中毒参与方比例超过 50%时，Krum 算法以及 Multi-Krum 算法防御失效后，本文算法中的正常参与方仍可以进行防御，且在 MNIST 数据集上的精度仍超过了 90%。实验表明，在投毒参与方比例超过一半时，对比方法几乎全部失效，仅本文提出的方法实现了正常参与方对投毒攻击的防御，证明了本文算法更优的鲁棒性。

5.2.3　参与方模型性能对比

本文算法会在每个参与方产生一个本地模型，为了更好的评估本文算法产生的不同本地模型在测试集上的表现，分别选取部分参与方在测试集上的准确率进行详详细对比分析。在在边缘黑盒攻击下，选取 Krum、Multi-Krum、DnC 以及本文算法在不同参与方上的准确率进行分析。

由图 5 可知，各参与方的模型在测试集上的准确率相差较小，对投毒攻击的防御效果近似。同时可以观察到，在 client0 至 clien t4 中，本文算法由于没有直接剔除疑似的恶意参与方，相较于其他基准算法，准确率存在略微降低。此外，由于 DnC 算法基于奇异值分解的方式，在选取的 5 个客户端中，表现的防御效果最好。如图 6 所示，为了更进一步的测试攻击方比例在 60%，即超过一半比例下，各个防御算法的效果，进行了对比实验。在投毒攻击方比例为 60%时，除了本文算法在正常参与方上训练的模型，其余防御算法的模型精度出现了严重下降。这是由于 Krum、Multi-Krum、DnC 算法都利用相似性剔除少量参与方的共享。所以本文算法性能较其余对比方法具有更高的鲁棒性。

在 FashionMNIST 数据集的测试结果如下图 7 与图 8。

由图 7 发现，对于数据集 FashionMNIST，实验结果与 MNIST 数据集上的结果近似。在中毒参与方比例为 40%时，Krum、Multi-Krum、DnC 算法与本文提出的算法在每个参与方模型的准确率上相差不大，也都达到了 80%以上，实现了较好的防御。如图 8 所示，在中毒参与方比例达到 60%后，本文采用的三种基准算法的准确率也急速下降，防御方法几乎完全失效，正常参与方模型也被中毒参与方的模型影响，投毒攻击效果显著。而本文的方法虽然无法对中毒参与方的模型进行纠正，但保护正常参与方，降低正常参与方模型受中毒参与方模型影响的程度，正常参与方模型的准确率在 80%以上。

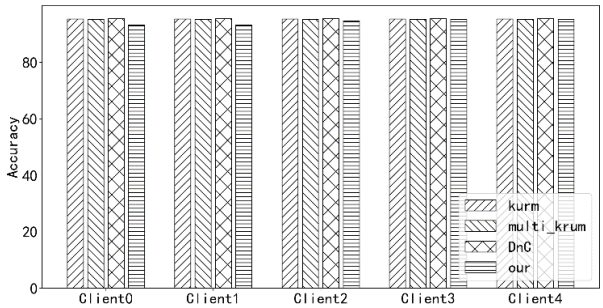

图 5　部分参与方在中毒参与方比例为 40%的准确率

Fig. 5　The accuracy of partial participant in poisoning is 40%

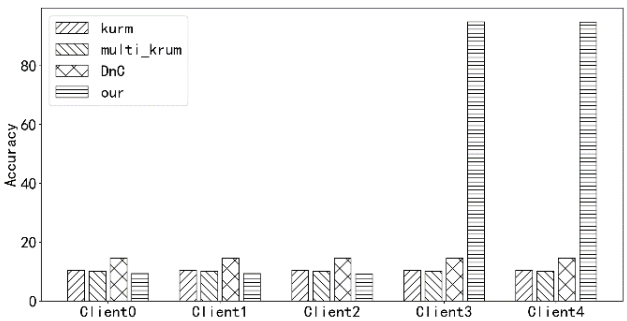

图 6　部分参与方在中毒参与方比例为 60%的准确率

Fig. 6　The accuracy of partial participant in poisoning is 60%

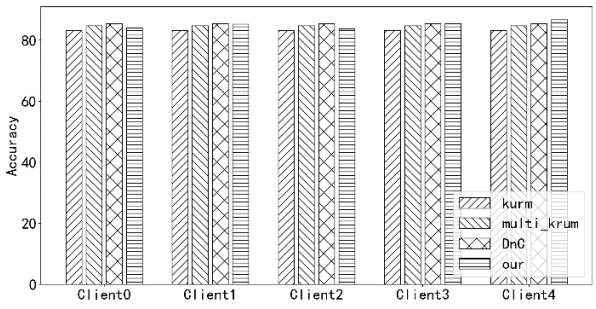

图 7　部分参与方在中毒参与方比例为 40%的准确率

Fig. 7　The accuracy rate of partial participant in poisoning is 40%

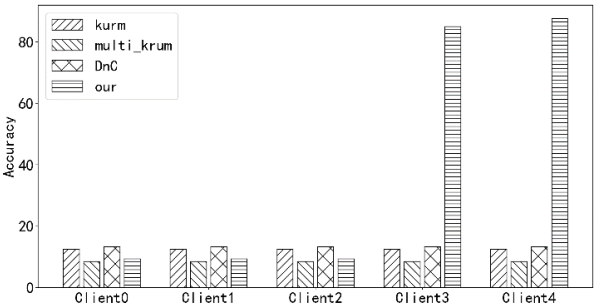

图 8　部分参与方在中毒参与方比例为 60%的准确率

Fig. 8　The accuracy rate of partial participant in poisoning is 60%

在符号翻转攻击与黑盒边缘攻击的情况下，实验的对比算法的都会因为投毒攻击而导致全局模型失效。但本文提出的算法在未中毒的参与方本地模型与全局模型差异较大的时候相信本地模型，从而降低来自全局模型的影响，最终未中毒的正常参与方仍然可以获得一个较高准确度的本地模型。因此，本文提出的算法相较于目前主流的联邦学习投毒攻击防御算法，取得了更好的效果。

6　结束语

在联邦学习迅速发展的背景下，对于分布式训练的攻击和防御方法越来越受到重视。然而，在对抗的过程中，由于联邦学习的分布式特性，仍然存在着许多攻击和挑战。本文提出的方法提高了联邦学习投毒攻击的防御能力，未来的研究工作可以将重心放在服务器与参与方进行防御配合的基础上，提高全局模型的鲁棒性。

参考文献：

[1] McMahan B, Moore E, Ramage D, *et al.* Communication-Efficient Learning of Deep Networks from Decentralized Data [C]// International Conference on Artificial Intelligence and Statistics. San Francisco: Morgan Kaufmann, 2017: 1273-1282.

[2] 孙爽, 李晓会, 刘妍, 等. 不同场景的联邦学习安全与隐私保护研究综述 [J]. 计算机应用研究, 2021, 38 (12): 3527-3534. (Sun Shuang, Li Xiaohui, Liu Yan, *et al.* A Review of Research on Federated Learning Security and Privacy Protection in Different Scenarios [J]. Computer Application Research, 2021, 38 (12): 3527-3534.)

[3] Kairouz P, McMahan H B, Avent B, *et al.* Advances and Open Problems in Federated Learning [J]. Found. Trends Mach. Learn. , 2021, 14: 1-210.

[4] Sahu A K, Li T, Sanjabi M, *et al.* Federated Optimization in Heterogeneous Networks [C]// Proceedings of Machine Learning and Systems. 2020: 429-450.

[5] Gong X, Chen Y, Wang Q, *et al.* Backdoor Attacks and Defenses in Federated Learning: State-of-the-Art, Taxonomy, and Future Directions [J]. IEEE Wireless Communications, 2022, 30 (2): 114-121.

[6] Shejwalkar V, Houmansadr A, Kairouz P, *et al.* Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Production Federated Learning [C]// 2022 IEEE Symposium on Security and Privacy (SP) . NJ: IEEE Computer Society, 2022: 1354-1371.

[7] Biggio B, Nelson B, Laskov P. Poisoning Attacks against Support Vector Machines [C]// International Conference on Machine Learning. New York: ACM, 2012: 1467-1474.

[8] 马鑫迪, 李清华, 姜奇, 等. 面向 Non-IID 数据的拜占庭鲁棒联邦学习 [J]. 通信学报, 2023, 44 (6): 138-153. (Ma Xindi, Li Qinghua, Jiang Qi, et al Byzantine Robust Federated Learning for Non IID Data [J]. Journal of Communications, 2023, 44 (6): 138-153.)

[9] 刘飚, 张方佼, 王文鑫, 等. 基于矩阵映射的拜占庭鲁棒联邦学习算法 [J]. 计算机研究与发展, 2021, 58 (11): 2416-2429. (Liu Biao, Zhang Fangjiao, Wang Wenxin, *et al.* Byzantine Robust Federated Learning Algorithm Based on Matrix Mapping [J]. Computer Research and Development, 2021, 58 (11): 2416-2429.)

[10] Blanchard P, El Mhamdi E M, Guerraoui R, *et al.* Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent [C]// Neural Information Processing Systems (NeurIPS) . San Diego: NIPS Foundation, 2017, 119-129.

[11] Lu Y, Fan L. An Efficient and Robust Aggregation Algorithm for Learning Federated CNN [C]// Proc of the 2020 3rd International Conference on Signal Processing and Machine Learning. New York: Association for Computing Machinery, 2020: 1-7.

[12] Pillutla K, Kakade S M, Harchaoui Z. Robust Aggregation for Federated Learning [J]. IEEE Trans on Signal Processing, 2019, 70: 1142-1154.

[13] Fung C, Yoon C J M, Beschastnikh I. The Limitations of Federated Learning in Sybil Settings [C]// International Symposium on Recent Advances in Intrusion Detection (RAID 2020) . Berkeley: USENIX Association, 2020: 301-316.

[14] Fang X, Ye M. Robust Federated Learning with Noisy and Heterogeneous Clients [C]// Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. NJ: IEEE Computer Society, 2022: 10072-10081.

[15] Wink T, Nochta Z. An Approach for Peer-to-Peer Federated Learning [C]// 2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W) . NJ: IEEE Computer Society, 2021: 150-157.

[16] Zhang K, Tao G, Xu Q, *et al.* FLIP: A Provable Defense Framework for Backdoor Mitigation in Federated Learning [C]// International Conference on Learning Representations. 2022: abs/2210. 12873.

[17] Cao D, Chang S, Lin Z, *et al.* Understanding Distributed Poisoning Attack in Federated Learning [C]// 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS) . NJ: IEEE Computer Society, 2019: 233-239.

[18] Zhang Z, Panda A, Song L, *et al.* Neurotoxin: Durable Backdoors in Federated Learning [C]// International Conference on Machine Learning. New York: ACM, 2022: 26429-26446.

[19] Gong X, Chen Y, Huang H, *et al.* Coordinated Backdoor Attacks against Federated Learning with Model-Dependent Triggers [J]. IEEE network, 2022, 36: 84-90.

[20] 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): 1019-1044. (Xiao Xiong, Tang Zhuo, Xiao Bin, *et al.* Review of Research on Privacy Protection and Security

Defense of Federated Learning [J]. Journal of Computer Science, 2023, 46 (5): 1019-1044.)

[21] Shejwalkar V, Houmansadr A. Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning [C]// Network and Distributed System Security Symposium. Reston: ISOC, 2021.

[22] Khazbak Y, Tan T, Cao G. MLGuard: Mitigating Poisoning Attacks in Privacy Preserving Distributed Collaborative Learning [C]// 2020 29th international conference on computer communications and networks (ICCCN) . NJ: IEEE Computer Society, 2020: 1-9.

[23] Muñoz-González L, Co K T, Lupu E C. Byzantine-Robust Federated Machine Learning through Adaptive Model Averaging [EB/OL]. (2019) [2023-08-11]. https://arxiv. org/abs/1909. 05125.

[24] Zhao L, Hu S, Wang Q, *et al.* Shielding Collaborative Learning: Mitigating Poisoning Attacks Through Client-Side Detection [J]. IEEE Trans on Dependable and Secure Computing, 2020, 18 (5): 2029-2041.