计算机科学
COMPUTER SCIENCE

一种基于自适应加权的鲁棒联邦学习算法

张连福, 谭作文

引用本文

张连福, 谭作文. 一种基于自适应加权的鲁棒联邦学习算法[J]. 计算机科学, 2023, 50(6A): 230200188-9.

ZHANG Lianfu, TAN Zuowen. Robust Federated Learning Algorithm Based on Adaptive Weighting[J]. Computer Science, 2023, 50(6A): 230200188-9.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

差分隐私研究进展综述

Review of Differential Privacy Research

计算机科学, 2023, 50(4): 265-276. https://doi.org/10.11896/jsjkx.220500292

RCP:本地差分隐私下的均值保护技术

RCP:Mean Value Protection Technology Under Local Differential Privacy

计算机科学, 2023, 50(2): 333-345. https://doi.org/10.11896/jsjkx.220700273

一种非独立同分布问题下的联邦数据增强算法

Federated Data Augmentation Algorithm for Non-independent and Identical Distributed Data

计算机科学, 2022, 49(12): 33-39. https://doi.org/10.11896/jsjkx.220300031

面向网络侦察欺骗的差分隐私指纹混淆机制

Differential Privacy Based Fingerprinting Obfuscation Mechanism Towards NetworkReconnaissance Deception

计算机科学, 2022, 49(11): 351-359. https://doi.org/10.11896/jsjkx.220400285

基于联盟链的能源交易数据隐私保护方案

Privacy-preserving Scheme of Energy Trading Data Based on Consortium Blockchain

计算机科学, 2022, 49(11): 335-344. https://doi.org/10.11896/jsjkx.220300138

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200188

一种基于自适应加权的鲁棒联邦学习算法

张连福[1,2]　谭作文[1]
1 江西财经大学信息管理学院计算机科学与技术系　南昌 330013
2 宜春学院数学与计算机科学学院　江西 宜春 336000
(zlf_jx@163.com)

摘　要　联邦学习(Federated Learning,FL)允许多个数据所有者联合训练机器学习模型,而无需他们共享私有训练数据。然而,研究表明,FL容易同时遭受拜占庭攻击和隐私泄露威胁,现有的研究都没有很好地解决这一问题。在联邦学习场景中,保护FL免受拜占庭攻击,同时考虑性能、效率、隐私、攻击者数量、简单可行等问题,是一个极具挑战性的问题。为解决这一问题,基于l_2范数和两次归一化方法提出了一种隐私保护鲁棒联邦学习算法DP-FedAWA。提出的算法不需要训练过程之外的任何假设,并且可以自适应地处理少量和大量的攻击者。无防御设置下选用DP-FedAvg作为比较基线,防御设置下选用Krum和Median作为比较基线。MedMNIST2D数据集上的广泛实验证实了,DP-FedAWA算法是安全的,对恶意客户端具有很好的鲁棒性,在Accuracy,Precision,Recall和F1-Score等性能指标上全面优于现有的Krum和Median算法。

关键词:自适应加权;l_2范数距离;两次归一化;拜占庭攻击;鲁棒联邦学习;差分隐私
中图法分类号　TP391

Robust Federated Learning Algorithm Based on Adaptive Weighting

ZHANG Lianfu[1,2] and TAN Zuowen[1]
1 Department of Computer Science and Technology,School of Information Management,Jiangxi University of Finance and Economics,Nanchang 330013,China
2 College of Mathematics and Computer Science,Yichun University,Yichun,Jiangxi 336000,China

Abstract　Federated learning allows multiple data owners to jointly train machine learning models without sharing private training data. However,studies have shown that FL is vulnerable to Byzantine attacks and privacy breaches,this problem has not been well addressed by existing studies. In the federated learning scenario,protecting FL from Byzantine attacks while considering performance,efficiency,privacy,number of attackers,simplicity and feasibility is a challenging problem. To solve this problem,a privacy preserving robust federal learning algorithm DP-FedAWA is proposed based on l_2-norm distance and quadratic normalization. The proposed algorithm does not require any assumptions outside the training process and can deal with a few or a lot of attackers adaptively. In no defense setting,DP-FedAvg is used as the comparison baseline,while Krum and Median are used as the comparison baseline in the defense setting. Extensive experiments on MedMNIST2D data set confirm that the proposed DP-FedAWA algorithm is safe and robust to malicious clients,and comprehensively outperforms the existing Krum and Median in Accuracy,Precision,Recall and F1-Score.

Keywords　Adaptive weighting,l_2-norm distance,Quadratic normalization,Byzantine attacks,Robust federated learning,Differential privacy

1　引言

机器学习(Machine Learning,ML)模型在医疗保健领域有很好的应用前景,例如可以用于医疗诊断工具构建[1]、患者相似性学习[2]、疾病危险因素预测[3]、肿瘤分型[4]或基因序列数据分析[5]等。一般来说,ML算法需要访问大量的训练数据,才能达到最佳的精度。然而,在医疗保健领域,集中式收集所有数据是不可行的。另一方面,EHR数据包含患者的敏感信息,发布这些数据将侵犯他们的隐私。此外,一些法律法规和医疗监管政策也限制了医疗保健数据的访问。最近出现的联邦学习(Federated Learning,FL)作为一种计算范式,允许用户协作学习一个全局机器学习模型,而不暴露他们的本地数据。这就缓解了在医疗场所之外共享原始医疗数据的必要性,并且可以显著降低带宽成本,同时很好地保护用户隐私。

然而,最近的研究[6]表明,这种深度学习模型的权重参数中仍然含有一些敏感信息,还将揭示有关训练数据的隐私信息。针对这一问题,已有不少保护隐私的联邦学习方案先后被提出。可是这些方案假设服务器是诚实但好奇的,只考虑

基金项目:国家自然科学基金(61862028);江西省教育厅青年科技项目(GJJ210529)
This work was supported by the National Natural Science Foundation of China(61862028) and Youth Projects Science and Technology of Jiangxi Provincial Department of Education(GJJ210529).
通信作者:谭作文(tanzyw@163.com)

了模型训练过程中的隐私保护，通常不能防御恶意客户端的拜占庭攻击。但是研究表明，FL 还非常容易遭受拜占庭攻击。拜占庭攻击是指，客户端在提交模型更新时，提交任意的随机数，或者选择不提交。这可能是出于机器故障或者网络故障等不可抗力的原因，也可能是恶意的用户故意提交错误的更新来破坏训练过程。Blanchard 等[7]已经证明，一个恶意用户就足以破坏训练的收敛性，并破坏最终全局模型的性能。针对 FL 中的拜占庭攻击，研究人员已经提出了很多防御方案，其大体可以分为两类。第一类是基于距离的防御方案。这类方案受到以下假设的限制：客户端的数据分布必须是独立同分布的，恶意客户端的数量小于良性客户端的数量。另一类是基于性能的防御方案。这类方案依赖于一个干净的辅助数据集来训练检测器或协助检测恶意模型。然而，这些辅助数据是独立地从部分客户端采集的，这违反了隐私保护的原则，妨碍了其实用性。为了同时防御联邦学习过程中的隐私威胁和拜占庭攻击威胁，研究者们提出了一些保护隐私鲁棒联邦学习方案。然而，这些方案都存在一定的局限性，降低了其实用性。

因此，在联邦学习的场景中，保护 FL 免受拜占庭攻击，同时考虑性能、效率、隐私、攻击者数量、简单可行等问题，是一个极具挑战性的问题。为了解决这些问题，我们提出了一种隐私保护鲁棒聚合算法 DP-FedAWA 来防御 FL 中的拜占庭攻击，其同时满足训练过程中的隐私保护。该算法由 4 个关键部分组成：对局部模型进行添加差分隐私噪声的训练；选择合适的参数来计算局部模型之间的二范数距离；根据二范距离对各个局部模型设置不同的信用评分；根据信用评分对各个模型进行自适应聚合。DP-FedAWA 不需要训练过程之外的任何假设，并且可以自适应地处理少量和大量的攻击者。在 MedMNIST2D 数据集上进行的广泛实验，证实了提出的方法在抗拜占庭攻击和模型分类精度等各种评价指标方面的有效性。

2 相关工作

(1)保护隐私联邦学习。在 FL 系统中，主流的隐私保护技术有差分隐私[8-10]、同态加密[11-12]和安全多方计算技术，如秘密共享[13-16]。然而，这些方案假设服务器是诚实但好奇的，只考虑了模型训练过程中数据和模型的隐私保护，通常不能防御恶意客户端提供错误梯度或模型参数。

(2)鲁棒联邦学习。为了规避或缓解 FL 系统中异常更新的影响，许多鲁棒聚合算法逐渐被提出来。根据文献调研分析，现有的防御方案大体可分为两类，即基于性能的防御方案和基于距离的防御方案。

基于性能的防御方案通过直接验证其性能来检测异常更新，这种方案比其他解决方案更可靠。Li 等[17]利用预训练的自动编码器来评估性能。对于良性模型更新，自动编码器将输出与输入相似的向量，但异常更新将产生较大的间隙。然而，训练一个自动编码器是耗时的，并且很难得到包含足够良性模型更新的训练集。相比之下，Xie 等[18]提出的方案 Zeno 在服务器端只需要一个小的验证集。具体来说，Zeno 使用验证集为每个候选梯度计算一个分数，更新的分数越高意味着性能越好，表明可靠性的概率越高。但 Zeno 需要了解攻击者的数量。为了解决这一问题，Cao 等[19]提出了一种拜占庭鲁棒分布式梯度算法，该算法利用一个小的干净数据集计算噪声梯度，过滤掉从恶意客户端接收到的信息。但是，Cao 等[19]提出的防御方案严重依赖于超参数的设置，在实际应用中可能很难发现合理的超参数设置。据此，Cao 等[20]提出联邦学习方法 FLTrust。在此方法中，服务提供者自己引导信任，即利用每次本地更新与服务器更新(基于中央服务器收集的干净数据集)之间的余弦相似度来设置本地更新的信任分数。由于与服务器更新一致，可靠的更新将获得较高的信任分数，而恶意更新将获得较低的信任分数。信任分数越高，聚合时对应的本地更新的权重越大。然而，基于性能的防御方案依赖于一个干净的辅助数据集来进行检查，这妨碍了其实用性。

基于距离的防御方案认为明显远离其他更新的更新是恶意的。这种方法通过中值或均值计算比较更新之间的距离来规避异常更新。Blanchard 等[21]提出了 Krum 及其变体 Multi-Krum。在 Krum 中，中央服务器基于欧氏距离在多个局部更新中选择一个最优更新，并丢弃所有其他更新。而 Multi-Krum 则选择多个更新并计算平均值来更新全局模型。与 Multi-Krum 类似，FABA[22]的目标是通过丢弃远离平均梯度的梯度来去除上传梯度中的异常值。但是，Multi-Krum 和 FABA 都需要提前知道恶意客户端的数量，因此很难应用到实际应用中。Yin 等[23]提出了两种鲁棒分布式梯度下降算法 Median 和 TrimmedMean。Bulyan[24]结合 Krum 和 TrimmedMean 选择最优模型。由于基于中值的算法比基于均值的算法更优越，鲁棒性更好，其他算法使用坐标中值[23]、几何中值[25]和近似几何中值[26]来聚合全局模型。然而，这些基于距离的防御方案只适用于恶意参与者占比不超过总节点数一半的情况，这限制了其实际使用场景。

此外，Yan 等[27]提出了一种联邦学习场景下增强模型鲁棒性的对抗训练方法。该方法通过加入对抗样本，迭代对抗样本以及正常样本等进行训练，调整各训练样本下损失函数的权重，完成本地训练以及全局模型的更新。然而，该方法要求客户端首先基于本地数据生成相应的对抗样本，额外增加了客户端的计算开销。

(3)保护隐私鲁棒联邦学习。上述鲁棒联邦学习方案在一定程度上缓解了拜占庭攻击的威胁，但它们都忽略了模型训练过程中的隐私问题，因而无法防御各种推理攻击，可能导致模型训练过程中用户训练数据的隐私泄露。为了同时防御联邦学习过程中的隐私威胁和拜占庭攻击威胁，研究者们最近提出了一些保护隐私的鲁棒联邦学习方案。Chang 等[28]设计了一个鲁棒协作机器学习框架 Cronus。作者利用黑盒局部模型之间的健壮知识迁移来防御联邦学习的中毒攻击。但是，该框架存在一定的局限性，因为它会降低精度，并具有较高的计算成本。Miao 等[29]设计了一个基于区块链的隐私保护拜占庭鲁棒联邦学习方案来减轻恶意客户端的影响。方案使用余弦相似度来识别恶意客户端，利用同态加密来提供安全聚合，并使用区块链实现去中心化存储，保证中间参数信息无法被篡改。然而，区块链等外部模块将带来较大的网络和计算开销，降低了其实用性。Tang 等[30]提出了一个鲁棒的隐私保护系统 PILE。该系统通过可验证的扰动模块，使机密局部梯度可验证，确保了模型训练的鲁棒性，实现了局部梯度和全局模型的隐私保护。但该方案中验证协议比较复杂，并且 Paillier 加密技术的使用需花费更多的时间。

3 背景知识

3.1 联邦学习

联邦学习可以看作一种特殊的分布式机器学习[31]。该种体系结构通常包含服务器和参与者集合 $\mathcal{K}:=\{1,2,\cdots,K\}$,其中每个参与者 $k\in\mathcal{K}$ 拥有一个大小为 $|\mathcal{D}_k|$ 的本地数据集。FL 的目标是训练一个具有参数 $w\in\mathbb{R}^d$ 的模型,并最小化以下经验风险函数:

$$\min_w F(w):=\sum_{k\in\mathcal{K}}\mu_k F_k(w) \tag{1}$$

其中,$\mu_k=\frac{|\mathcal{D}_k|}{\sum_{k\in\mathcal{K}}|\mathcal{D}_k|}$ 表示参与者 k 的权重因子,F_k 表示参与者 k 的损失函数。

3.2 差分隐私

差分隐私是一种数学上严格的隐私保护方法,可以有效地防御成员推理攻击和属性推理攻击。差分隐私通常用于对包含敏感信息的数据库执行各种操作。差分隐私的目的是,函数 M 的输出不完全依赖于数据库中的任何单一实例,即无论某一样本是否在数据库中,M 都极有可能产生相同的输出。

定义 1 (ε,δ)-差分隐私[32]:随机算法 $M:D\to R$ 满足(ε,δ)-差分隐私当且仅当对于任意相邻数据集 $d,d'\in D$ 和任意输出 $S\subseteq R$,满足以下条件:

$$Pr[M(d)\in S]\leqslant e^{\varepsilon}Pr[M(d')\in S]+\delta \tag{2}$$

其中,$M(d)$和 $M(d')$分别表示算法 M 在数据集 d 和d'上的输出,Pr 是算法的输出概率,ε 是隐私预算。隐私预算 δ 表示可容忍的隐私预算超过 ε 的概率。差分隐私有两个重要性质:串行组合性和并行组合性。它们用于判断分布式算法是否满足差分隐私,以及分配机器学习中的隐私预算。

3.3 l_2范数

l_2范数(l_2 norm),也称欧几里得范数(Euclidean norm)、欧几里得距离。对于 n 维向量$\vec{a}$,其l_2范数定义为该向量所有元素的平方和的开平方,如式(3)所示。

$$\|\vec{a}\|_2=\sqrt{\sum_i^n(a_i)^2} \tag{3}$$

对于两个 n 维向量$\vec{a}$,$\vec{b}$,其L_2范数可以认为是空间中该两个点间的距离,如式(4)所示。

$$\|\vec{a}-\vec{b}\|_2=\sqrt{\sum_i^n(a_i-b_i)^2} \tag{4}$$

我们利用二范数公式来计算参与者提交的局部模型与其他参与者提交的局部模型之间的距离,并将其距离之和的倒数作为该模型的得分,任何表现不佳的局部模型更新都将被分配更低的权重。

3.4 FL 聚合算法

DP-FedAvg[33]是非对抗性设置中流行的 FL 聚合方法,而 Krum[21]和 Median[23]是拜占庭鲁棒 FL 聚合方法。本文选用 DP-FedAvg 作为非对抗性设置下的比较基准,选用 Krum 和 Median 作为对抗性设置下的比较基准。

DP-FedAvg:该算法是在 FedAvg 的基础上,引入了差分隐私机制。它假定中心服务器是诚实的,通过加权平均获得全局模型后注入高斯噪声,因此可能导致本地参与方数据隐私的泄露。

Krum:该算法在每个局部模型更新中,计算与最近邻的 $K-M-2$ 个其他模型的欧几里得距离的平方之和,并将其作为该模型的得分,然后选择得分最低的模型作为聚合模型,如式(5)所示。

$$Krum:=\{\Delta_i\,|\,i=\arg\min_{i\in K}\sum_{i\to j}\|\Delta_i-\Delta_j\|_2^2\} \tag{5}$$

其中,$i\to j$ 是Δ_i的 $K-M-2$ 个最近邻的索引(按欧几里得距离的平方测量),K 是客户端的总数,M 是恶意客户端的数量。Krum 算法在拜占庭客户端不超过一半的情况下,能保证模型正常收敛。

Median:Median 定义为所有给定更新集的坐标中值,即:

$$med:=Median(\{\Delta_k:k\in K\}) \tag{6}$$

对于第 i 个全局模型参数,其坐标中值定义为:

$$med_i:=Median(\{\Delta_k^i:k\in K\}) \tag{7}$$

Median 可以容忍少于 50%的恶意客户端。

4 问题描述

4.1 设计目标

本文关注的是 Cross-Silo 场景的联邦学习,其中用于训练的数据被水平划分,即不同数据集共享相同的特征空间,但数据集中的样本不同。在 Cross-Silo FL 中,K 个医院或医疗研究机构和 1 个服务器参与训练过程。这些医院或医疗研究机构拥有大量用于训练模型的数据,它们参与交互式 FL 协议。在该协议中,它们与服务器共享参数更新,即本地模型参数,并且不会中途退出。服务器聚合各局部模型参数来训练一个全局模型。目标是设计一个 FL 方法,使得其在确保性能和效率的前提下,实现隐私保护和对恶意客户端的拜占庭鲁棒性,同时优于现有的鲁棒聚合算法。

4.2 威胁模型

考虑一个攻击场景,其中一个客户端子集是恶意的,并由对手 $\mathcal{A}$ 控制。将所有参与的客户端中恶意客户端的比例表示为恶意客户端率(Malicious Client Rate,PCR),定义为 $MCR=m/n$,其中 m 和 n 分别表示恶意客户端和所有客户端的数量。这里允许恶意客户端的数量大于 50%。现有的很多拜占庭鲁棒 FL 方法,如 Krum,Trim-mean 和 Median 等,都要求恶意参与者不超过 50%,这在 FL 系统中是不现实的。然而,我们也不考虑对手 $\mathcal{A}$ 控制了 80%以上客户端的场景,因为这样的 FL 系统毫无意义。恶意客户端可以操纵 FL 协议,并可以在 FL 的每次训练迭代过程中向服务器发送任意的本地模型参数来干扰全局模型训练,或提取其他客户端的私有训练数据。我们认为服务器是半诚实的,即诚实但好奇的。它确实遵循协议,但试图尽可能多地推断参与节点的私有训练数据。

5 DP-FedAWA 算法

5.1 概述

针对联邦学习中拜占庭攻击和隐私泄露威胁,我们提出了隐私保护鲁棒聚合算法 DP-FedAWA。该算法的系统架构图如图 1 所示。在模型聚合方面,它在技术思路上类似于 Krum 算法,都是基于欧几里得距离来计算局部模型之间的距离,然后赋予相应的信用评分。但 Krum 算法需要预先知道恶意节点数 M,然后计算与最近邻的 $K-M-2$ 个其他模型的欧几里得距离的平方之和作为该模型的得分,并且最终选择得分最低的一个模型作为聚合模型。其缺点是没有充分

利用更多的训练数据，从而影响了最终模型的性能。Median方案也是一样，它选择所有给定更新集的坐标中值，其本质也是选择一个最优模型。我们的改进之处在于，利用欧几里得距离来识别恶意更新，并为它们分配一个低权重，以减少它们对全局模型的负面影响，这样又可充分利用更多的训练数据，从而提升最终模型的性能，并加速模型的收敛。

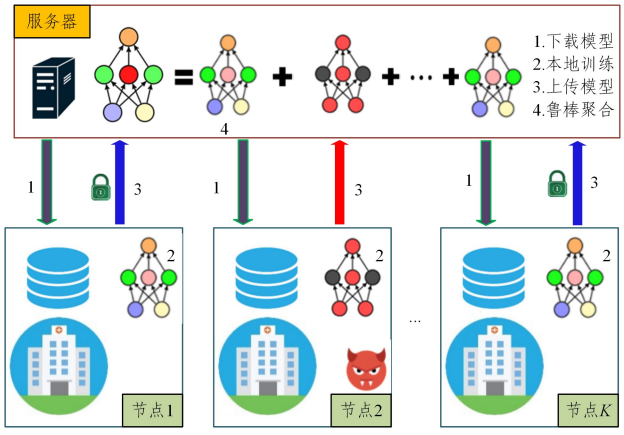


图1 具有拜占庭参与者的FL系统架构图

Fig. 1 Architecture of FL system with Byzantine participants

5.2 算法设计

DP-FedAWA算法的关键是计算局部模型的信用得分，即最终全局模型聚合时的最优加权系数。设计思路是，首先计算局部模型与其他模型的欧几里得距离的平方之和，然后将该距离之和的倒数作为该模型的初次得分。为了让良性更新获得更高的得分，异常更新获得更低的得分，对该初次得分连续进行两次归一化操作，得到最终得分，并将该最终得分作为该局部模型的信用得分。最后，服务器依据信用得分对各个模型进行自适应聚合。在我们看来，异常更新之间以及异常更新与正常更新之间的二范数距离应该会更大。基于这种观察，中心服务器可以发现异常更新，并为其分配更低权重。具体来说，主要包含以下3步：计算局部模型的初次得分，归一化初次得分，以及聚合各局部模型。

(1)计算局部模型的初次得分

假定有客户端参与者集合$\mathcal{K}:=\{1,2,\cdots,K\}$。先计算局部模型$\Delta_i$与其他模型的欧几里得距离的平方之和$d_i$，如式(8)所示。

$$d_i \leftarrow \sum_{i \to j} \| \Delta_i - \Delta_j \|_2^2 \tag{8}$$

其中，$i,j \in \mathcal{K}$，$i \to j$是Δ_i与$K-1$个邻居节点的索引。

类似地，对于模型i的第k个参数，距离计算公式如式(9)所示。

$$d_i^k \leftarrow \sum_{i \to j, k \in M} \| \Delta_i^k - \Delta_j^k \|_2^2 \tag{9}$$

其中，M为模型i的参数集合。需要说明的是，式(9)更为实用，因为在实际训练中，全局模型是按照不同的参数来聚合的，而不是直接按照整个模型来聚合的。但为了简单起见，本文以式(8)为准，即以模型为单位，来计算出它们的聚合权重系数。

接着，将该距离之和的倒数作为该模型的初次得分，如式(10)所示。

$$TS_i \leftarrow 1/d_i \tag{10}$$

为了让良性更新获得更高的得分，异常更新获得更低的得分，对该初次得分进行两次归一化。

(2)归一化初次得分

归一化是指，将数据按比例缩放，使之限定于一个较小的特定区间。由于最终全局模型聚合需要先获得各个局部模型的加权系数p_i，并且满足$p_i \in [0,1]$和$\sum_{i \in \mathcal{K}} p_i = 1$，而步骤(1)中计算得到的局部模型的初次得分是一个不小于0的任意实数，不能直接作为聚合加权系数，还必须归一化处理。第一次归一化时，拟采用ℓ_p范数归一化($p=1$)方法，求出在本轮迭代中，本模型得分在所有局部模型得分中的百分占比。假定在第r轮训练迭代中，参与者集的初次得分是一个K维向量s，其中$s_1, s_2, \cdots, s_K$是向量中的元素，分别表示第$1,2,\cdots,K$个模型的得分。经过第一次归一化操作之后，模型i得到新的得分TS_i'，如式(11)所示。

$$TS_i' \leftarrow \frac{TS_i}{\| s \|_1} = \frac{TS_i}{|s_1| + |s_2| + \cdots + |s_K|} \tag{11}$$

经过第一次归一化处理，也许还不能较好地区分出良性与异常更新。因此，还需接着进行第二次归一化操作。第二次归一化拟采用归一化指数函数softmax函数。假定在第r轮训练迭代中，第一次归一化后，模型i的新得分为TS_i'，则经过第二次归一化后，可以获得最终得分μ_i，如式(12)所示。

$$\mu_i \leftarrow \frac{e^{\alpha TS_i'}}{\sum_{j=1}^{K} e^{\alpha TS_j'}} \tag{12}$$

其中，α是一个用于算法优化的超参数，j为第j个局部模型的索引，K为客户端参与者总数。显然，$\mu_i \in [0,1]$，且$\sum_{i \in \mathcal{K}} \mu_i = 1$。经过两次归一化后，可以获得模型聚合中的最优加权系数，确保任何表现不佳的异常更新均被分配更低的权重。

(3)聚合各局部模型

假定在第r轮训练结束后，模型i共享的模型参数为w_i^r，模型i的最终得分为μ_i，则服务器按式(13)聚合各局部模型，得到一个全局模型w_{global}^r。

$$w_{\text{global}}^r = \sum_{i=1}^{K} \mu_i w_i^r \tag{13}$$

其中，K为客户端参与者总数。

5.3 训练算法流程

隐私保护联邦学习算法DP-FedAWA的本地模型训练过程如算法1所示，完整流程如算法2所示。该算法完整流程为：1)服务器初始化全局模型，随机选定部分节点参与训练，并将初始全局模型发送给各参与节点；2)参与方在本地训练模型和更新权重，并将DP噪声添加到局部模型梯度参数中，将更新后的模型上传给服务器；3)服务器接收各参与方提交的局部模型参数，选择合适的参数来计算局部模型之间的二范数距离，并根据二范距离对各个局部模型设置不同的信用评分；4)服务器根据信用评分对各个模型进行自适应聚合。重复上述步骤，直至达到设定的迭代次数或预期的模型精确度。该算法防御拜占庭攻击主要由服务器端自主完成，对客户端不会增加额外的计算和通信开销。模型训练中用到的数学符号及其含义如表1所列。

算法1 LocalUpdate(w,D,b,η,T,σ,C)//本地训练过程

输出：本地更新后的模型w^T

1. $w^0 \leftarrow w$
2. for t=1,2,…,T do
3. 从D中随机选取小批量数据L_b
4. 计算梯度 $g_t \leftarrow \nabla \text{Loss}(L_b; w)$
5. 剪裁梯度 $\bar{g}_t \leftarrow g_t / \max(1, \| g_t \|_2 / C)$

6. 添加噪音 $\tilde{g}_t \leftarrow \frac{1}{b}(\sum \bar{g}_t + \mathcal{N}(0,\sigma^2 C^2))$

7. 模型更新 $w^t \leftarrow w^{t-1} - \eta \tilde{g}_t$

8. end for

9. return w^T.

算法 2 DP-FedAWA //完整算法流程

输入:学习率 η,全局总轮数 R,本地迭代次数 T,客户端集合 $C=\{C_1, C_2, \cdots, C_K\}$,本地数据集 $D=\{D_1, D_2, \cdots, D_K\}$,差分隐私参数 σ,梯度剪裁边界 C,批大小 b,每轮训练中选中的客户端数 M

输出:全局模型 w

1. w←服务器初始化全局模型w^0

2. for r=1,2,…,R do

3. 服务器随机选择 M 个客户端$C_1, C_2, \cdots, C_M$并将 w 发送给它们

//客户端本地训练

4. for $i=C_1, C_2, \cdots, C_M$ do in parallel

5. w_i=LocalUpdate(w,D,b,η,T,σ,C)

6. 上传w_i给服务器

7. end for

//服务器计算加权系数

8. for $i=C_1, C_2, \cdots, C_M$ do

9. $TS_i \leftarrow \frac{1}{d_i}$

10. $TS_i' \leftarrow \frac{TS_i}{\| s \|_1}$

12. end for

13. $w=\sum_{i=1}^{M} \mu_i w_i$// 服务器聚合模型

14. end for

15. return w.

表 1 数学符号及其含义

Table 1 Mathematical symbols and their meanings

Symbols	Parameter meaning
w	Global model parameter
D	Local dataset $D=\{D_1, D_2, \cdots, D_K\}$
b	Batch size
η	Learning rate
T	Number of local iterations
σ	Differential privacy parameter
C	Gradient clipping boundary
R	Global rounds
M	Number of clientsselected
P	Client sets $P=\{P_1, P_2, \cdots, P_K\}$

6 安全性分析

本节证明提出的算法满足数据安全性和具有拜占庭鲁棒性。

定理 1 在 DP-FedAWA 算法中,参数服务器或恶意参与者无法从全局模型参数中推断出关于训练数据的信息。

证明:在 DP-FedAWA 算法中,多个参与节点联合训练一个全局模型。各参与节点首先根据其灵敏度 Δf 和 ε 产生高斯机制噪声,然后将其添加到训练中的梯度参数中。根据差分隐私的性质,只要算法 1 中 Step6 的高斯噪声的标准差满足 $\sigma \geqslant c\Delta f/\varepsilon$,且 $c^2 > 2\ln(1.25/\delta)$,则可保证算法中每一轮训练都满足(ε,δ)-差分隐私。假定全局训练总轮数为 R,本地迭代次数为 T,那么根据差分隐私的组合性质可知,DP-FedAWA 算法至少满足$(RT\varepsilon, RT\delta)$-差分隐私。因此,可以确保参数服务器或恶意参与者无法通过推理攻击,如 GAN 攻击[34]、模型反向攻击[6]和成员推理攻击[35]等,从全局模型参数推断出关于训练数据的信息。此外,在联邦学习过程中,训练数据不会离开数据所有者,数据隐私不会直接泄露。因此,可以保护训练数据的隐私性。

定理 2 DP-FedAWA 算法能够抵抗拜占庭攻击,实现鲁棒聚合的目标。

证明:拜占庭设备D_b通过发送恶意的模型参数w_b来攻击系统。假定诚实设备D_h发送的模型参数为w_h。为了攻击系统,w_b应尽可能偏离w_h。$d_i \leftarrow \sum_{i,j \in \mathcal{N}} \| w_i - w_j \|_2^2$能够正确地反映$w_i$偏离$w_j$的程度,故$d_b > d_h$。模型$w_i$的初次得分$TS_i \leftarrow 1/d_i$表示$w_i$在模型聚合中的重要性,拜占庭设备的$TS_b$会较小而诚实设备的$TS_h$会更大。经过一次归一化后,拜占庭设备的二次得分$TS_b' \leftarrow \frac{TS_b}{\| s \|_1}$也会较小。同理,再次归一化后,拜占庭设备的权重$\mu_b \leftarrow \frac{e^{\alpha TS_b'}}{\sum_{j=1}^{K} e^{\alpha TS_j'}}$也会更小。由此可见,经过两次归一化后,任何表现不佳的异常更新都将被分配更低的权重,从而确保可以获得模型聚合中最优的加权系数。

故 DP-FedAWA 算法能够抵抗拜占庭攻击,实现模型的鲁棒聚合。

7 实验结果及分析

7.1 数据集和模型

本文所有实验基于 MedMNIST v2 数据集[36]。这是上海交通大学研究人员近期创建的医疗版 MNIST 数据集。该数据集是二维和三维生物医学图像分类的大规模基线实验数据,包括 12 个 2D 数据集(28 * 28,224 * 224)和 6 个 3D 数据集(28 * 28 * 28),涵盖了生物医学图像中的主要数据模式,支持各种任务类型。对于 2D 数据集,支持 ResNet18 和 ResNet50 分别在 28 * 28 和 224 * 224 分辨率上进行测试;对于 3D 数据集,支持 2.5D,3D,ACS 卷积的 ResNet18 和 ResNet50 进行测试。同时,还支持 3 种自动机器学习模型 autosklearn,AutoKeras 和 Google AutoML Vision。在本文中,为了简单起见,选择了 2D 数据集中的 BloodMNIST,OrganAMNIST 和 OrganCMNIST 数据集进行实验,并假定各参与客户端的数据分布均匀。数据集描述如表 2 所列。模型选用 ResNet18(28 * 28)。为了保证比较的公平性,所有实验均取测试集上 3 次结果的平均值作为平均性能指标。

表 2 数据集描述

Table 2 Description of datasets

Dataset	Data type	Training set	Verification set	Test set
BloodMNIST2D	Multi-Class(8)	11959	1712	3421
OrganAMNIST2D	Multi-Class (11)	34581	6491	17778
OrganCMNIST2D	Multi-Class (11)	13000	2392	8268

7.2 评价指标

利用准确率(Accuracy)、精度(Precision)、召回率(Recall)和 F1 评分(F1 -Score)这 4 个评价指标来评估 DP-FedAWA 算法在不同情况下的性能表现。

$$Accuracy=(TP+TN)/(TP+TN+FP+FN) \quad (14)$$

$$Precision=TP/(TP+FP) \quad (15)$$

$$Recall = TP/(TP+FN) \tag{16}$$

$$F1\text{-}score = 2 \cdot Precision \cdot Recall/(Precision+Recall) \tag{17}$$

其中，真阳性(True Positives，TP)表示被正确地划分为正例的样本数；假阴性(False Negative，FN)表示被错误地划分为负例的样本数；真阴性(True Negative，TN)表示被正确地划分为负例的样本数；假阳性(False Positives，FP)表示被错误地划分为正例的样本数。

7.3 实验设置

联邦学习框架及攻击类型：联邦学习框架采用基于 Pytorch 的客户端/服务器架构。拜占庭攻击主要考虑高斯攻击(Gaussian attack)。这种攻击方法涉及到改变局部模型参数，从而使中央服务器聚合一个损坏的全局模型。具体方法是拜占庭节点生成一个向量，其中每个元素来自高斯分布 $\mathcal{N}(0, \sigma^2)$，高斯噪声的标准差 $\sigma=100$，并将其作为局部模型的参数进行处理。

比较基线及实验参数设置：非对抗性设置下，选用 DP-FedAvg 算法作为比较基线；对抗性设置下，选用 Krum 和 Median 算法作为比较基线。3 种数据集上的通用参数设置为：优化器为 Adam，损失函数为 Cross-entropy，batch 大小为 128，本地训练 epoch 为 1，差分隐私预算 ε 为 1，δ 为 1×10^{-5}，加噪方式为本地模型训练时往梯度参数中添加高斯噪音，初始学习率为 0.01，在训练到 50% 和 75% 时，学习率分别减半。对于 BloodMNIST2D 数据集，客户端节点总数为 20，每轮随机选择 10 个节点参与训练，全局训练轮数为 100；对于 OrganAMNIST2D 数据集，客户端节点总数为 60，每轮随机选择 10 个节点参与训练，全局训练轮数为 80；对于 OrganCMNIST2D 数据集，客户端节点总数为 30，每轮随机选择 10 个节点参与训练，全局训练轮数为 100。

7.4 性能比较和结果分析

下面分别对 BloodMNIST2D，OrganAMNIST2D 和 OrganCMNIST2D 数据集上的实验结果进行讨论分析。限于篇幅，只展示了 Accuracy 和 Loss 曲线，没有展示 Precision，Recall 和 F1 -Score 曲线。

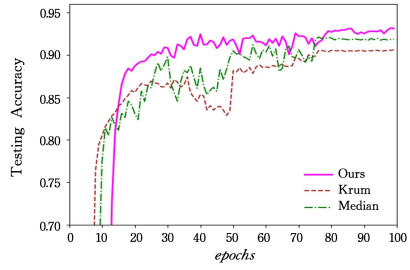
(1)BloodMNIST2D 数据集实验结果及分析

表 3 和图 2 展示了对 BloodMNIST2D 数据集进行高斯攻击的实验结果，证明了我们的算法在面对高斯攻击时，与其他算法相比，具有明显的鲁棒性。结果表明，在恶意节点占 10%和 30%时，算法 DP-FedAWA 保持了很高的准确率，准确率分别到达 92.6%和 91.8%，而 Krum 和 Median 的准确率低于我们提出的方法。实验结果还表明，Krum 支持不超过 50%的恶意节点，Median 支持的恶意节点数低于 50%，而我们方法可以支持高达 70%的恶意节点。当恶意节点达到 50%时，我们的方法的准确率也高达 90.4%，明显优于 Krum 算法。从 Loss 曲线可以看出，算法 DP-FedAWA 中的 Loss 值一直减小，并且趋于稳定，说明提出的方法能够正常收敛。与其他方法相比，DP-FedAWA 中的 Loss 值最小，说明其性能更佳。DP-FedAvg 是非对抗性设置下的聚合方法，从表 3 可见，在高斯攻击下，当恶意节点占 10%时，它的准确率已非常低。

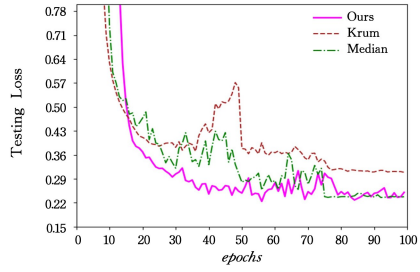
表 3 BloodMNIST2D 数据集上的性能对比

Table 3 Comparison of performance on BloodMNIST2D dataset

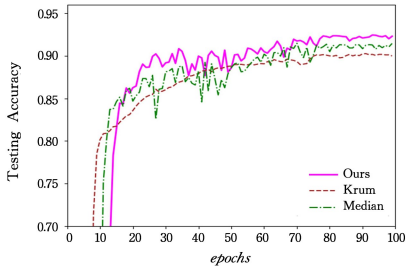
malicious node	Evaluation index	Ours	Krum	Median	DP-FedAvg
10%	*Accuracy*	**0.926**	0.906	0.919	0.204
	F1-score	**0.917**	0.891	0.906	0.121
	Precision	**0.924**	0.893	0.909	0.080
	Recall	**0.910**	0.889	0.904	0.246
30%	*Accuracy*	**0.918**	0.899	0.913	0.194
	F1-score	**0.908**	0.881	0.898	0.040
	Precision	**0.917**	0.888	0.904	0.024
	Recall	**0.900**	0.875	0.892	0.125
50%	*Accuracy*	**0.904**	0.880	N/A	N/A
	F1-score	**0.895**	0.860	N/A	N/A
	Precision	**0.899**	0.870	N/A	N/A
	Recall	**0.892**	0.851	N/A	N/A
60%	*Accuracy*	**0.852**	N/A	N/A	N/A
	F1-score	**0.827**	N/A	N/A	N/A
	Precision	**0.851**	N/A	N/A	N/A
	Recall	**0.804**	N/A	N/A	N/A



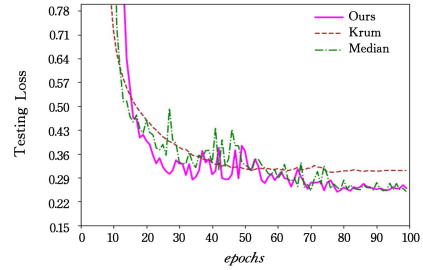
(a)恶意节点占 10%时的 *Accuracy*



(b)恶意节点占 10%时的 *Loss*



(c)恶意节点占 30%时的 *Accuracy*



(d)恶意节点占 30%时的 *Loss*

图 2 BloodMNIST2D 数据集上实验结果

Fig. 2 Experimental results on BloodMNIST2D dataset

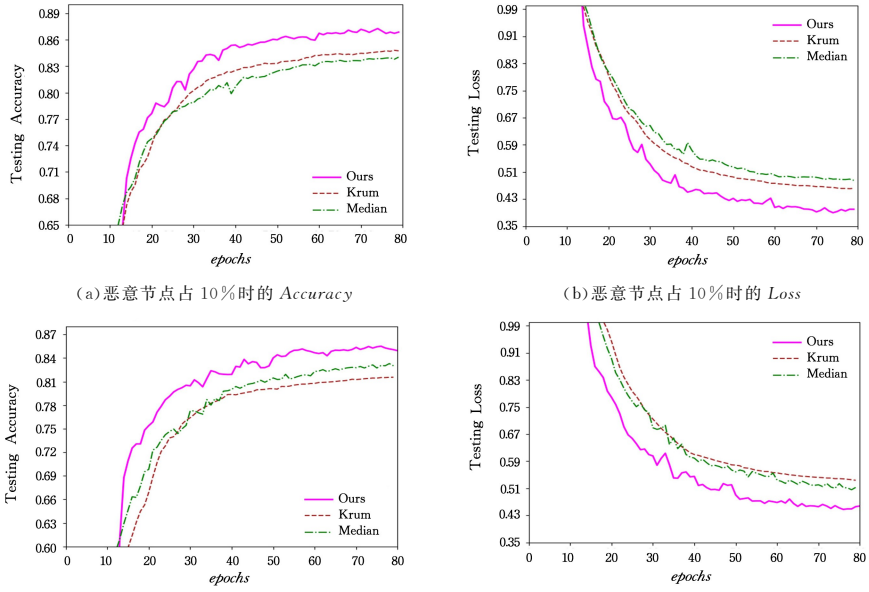
(2)OrganAMNIST2D 数据集实验结果及分析

表 4 和图 3 展示了对 OrganAMNIST2D 数据集进行高斯攻击的实验结果,证明了我们的算法在面对高斯攻击时,与其他算法相比,具有明显的鲁棒性。结果表明,在恶意节点占 10%和 30%时,DP-FedAWA 保持了较高的准确率,准确率分别到达 86.8%和 84.9%,而 Krum 和 Median 的准确率低于提出的方法。实验结果还表明,Krum 支持不超过 50%的恶意节点,Median 支持的恶意节点数低于 50%,而我们的方法可以支持高达 70%的恶意节点。当恶意节点达到 50%时,我们的方法的准确率也高达 83.1%,明显优于 Krum 算法。从 Loss 曲线可以看出,算法 DP-FedAWA 中的 Loss 值一直减小,并且趋于稳定,说明提出的方法能够正常收敛。与其他方法相比,DP-FedAWA 中的 Loss 值最小,说明提出的方法的性能更佳。DP-FedAvg 是非对抗性设置下的聚合方法,从表 4 可见,在高斯攻击下,当恶意节点占 10%时,它的准确率已非常低。

表 4 OrganAMNIST2D 数据集上的性能对比

Table 4 Comparison of performance on OrganAMNIST2D dataset

malicious node	Evaluation index	Ours	Krum	Median	DP-FedAvg
10%	*Accuracy*	**0.868**	0.848	0.838	0.174
	F1-score	**0.865**	0.848	0.841	0.069
	Precision	**0.867**	0.857	0.848	0.045
	Recall	**0.864**	0.840	0.834	0.149
30%	*Accuracy*	**0.849**	0.815	0.832	0.123
	F1-score	**0.844**	0.817	0.831	0.063
	Precision	**0.850**	0.827	0.840	0.044
	Recall	**0.839**	0.808	0.822	0.112
50%	*Accuracy*	**0.831**	0.802	N/A	N/A
	F1-score	**0.824**	0.797	N/A	N/A
	Precision	**0.833**	0.802	N/A	N/A
	Recall	**0.815**	0.793	N/A	N/A
60%	*Accuracy*	**0.717**	N/A	N/A	N/A
	F1-score	**0.697**	N/A	N/A	N/A
	Precision	**0.708**	N/A	N/A	N/A
	Recall	**0.686**	N/A	N/A	N/A



(a)恶意节点占 10%时的 *Accuracy*　(b)恶意节点占 10%时的 *Loss*

(c)恶意节点占 30%时的 *Accuracy*　(d)恶意节点占 30%时的 *Loss*

图 3 OrganAMNIST2D 数据集上的实验结果

Fig. 3 Experimental results on OrganAMNIST2D dataset

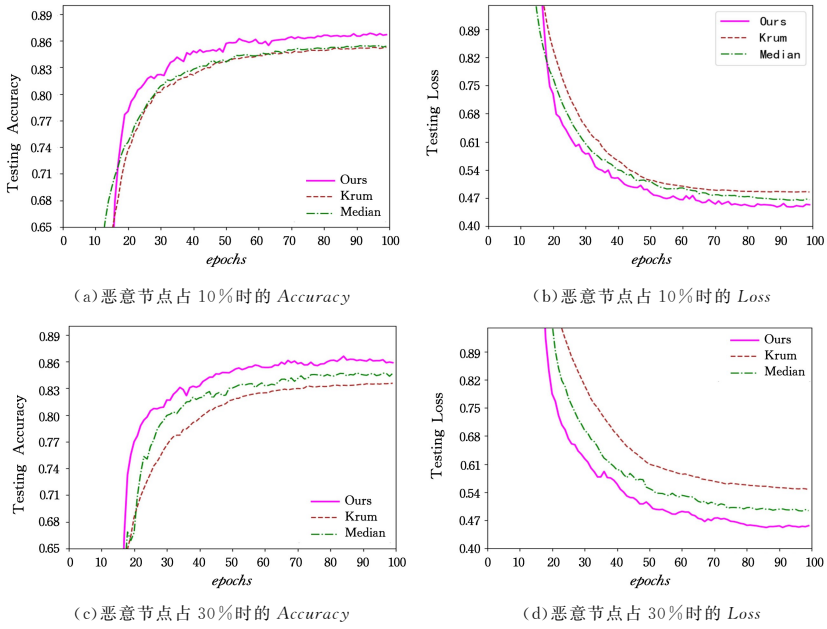
(3)OrganCMNIST2D 数据集实验结果及分析

表 5 和图 4 展示了对 OrganCMNIST2D 数据集进行高斯攻击的实验结果,证明了我们的算法在面对高斯攻击时,与其他算法相比,具有明显的鲁棒性。结果表明,在恶意节点占 10%和 30%时,DP-FedAWA 保持了较高的准确率,准确率分别到达 86.6%和 86%,而 Krum 和 Median 的准确率低于我们提出的方法。实验结果还表明,Krum 支持不超过 50%的恶意节点,Median 支持的恶意节点数低于 50%,而我们的方法可以支持高达 70%的恶意节点。当恶意节点达到 50%时,我们的方法的准确率也高达 84.4%,明显优于 Krum 算法。从 Loss 曲线可以看出,算法 DP-FedAWA 中的 Loss 值一直减小,并且趋于稳定,说明提出的方法能够正常收敛。与其他方法相比,算法 DP-FedAWA 中的 Loss 值最小,说明提出的方法的性能更佳。DP-FedAvg 是非对抗性设置下的聚合方法,从表 5 可见,在高斯攻击下,当恶意节点占 10%时,它的准确率已非常低。

表 5 OrganCMNIST2D 数据集上的性能对比

Table 5 Comparison of performance on OrganCMNIST2D dataset

malicious node	Evaluation index	Ours	Krum	Median	DP-FedAvg
10%	*Accuracy*	**0.866**	0.850	0.854	0.220
	F1-score	**0.854**	0.836	0.841	0.051
	Precision	**0.859**	0.840	0.849	0.035
	Recall	**0.850**	0.832	0.833	0.095
30%	*Accuracy*	**0.860**	0.834	0.845	0.146
	F1-score	**0.849**	0.821	0.836	0.051
	Precision	**0.855**	0.826	0.847	0.035
	Recall	**0.843**	0.816	0.826	0.094
50%	*Accuracy*	**0.844**	0.830	N/A	N/A
	F1-score	**0.826**	0.817	N/A	N/A
	Precision	**0.832**	0.826	N/A	N/A
	Recall	**0.821**	0.809	N/A	N/A
60%	*Accuracy*	**0.730**	N/A	N/A	N/A
	F1-score	**0.692**	N/A	N/A	N/A
	Precision	**0.705**	N/A	N/A	N/A
	Recall	**0.680**	N/A	N/A	N/A



(a)恶意节点占 10%时的 *Accuracy*　(b)恶意节点占 10%时的 *Loss*

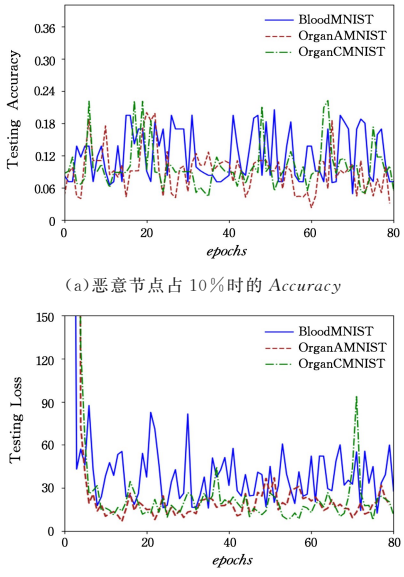
(c)恶意节点占 30%时的 *Accuracy*　(d)恶意节点占 30%时的 *Loss*

图 4　OrganCMNIST2D 数据集上的实验结果

Fig. 4　Experimental results on OrganCMNIST2D dataset

(4)无防御情况下实验结果及分析

DP-FedAvg 是非对抗性设置下的聚合方法。从图 5 可见，在相同的攻击下，当恶意节点占 10%时，它在 3 种数据集上的准确率都非常低。从 Loss 曲线也可以看出，这时 DP-FedAvg 方法的 Loss 值非常大，训练无法收敛。这从另一个角度体现出我们提出的方法的优越性。



(a)恶意节点占 10%时的 *Accuracy*

(b)恶意节点占 10%时的 *Loss*

图 5　无防御情况下恶意节点占时 10%的 *Accuracy* 和 *Loss*

Fig. 5　Accuracy curve and loss curve when malicious nodes account for 10% without defense

结束语　联邦学习允许多个数据所有者联合训练一个 ML 模型，而无需共享原始训练数据。然而，研究表明，FL 容易同时遭受拜占庭攻击和隐私泄露威胁，现有的方案都存在一定的局限性，降低了其实用性。为了解决这些问题，我们提出了一种隐私保护的鲁棒联邦学习算法 DP-FedAWA，来防御 FL 中的拜占庭攻击，同时满足训练过程中的隐私保护需求。方案不需要训练过程之外的任何假设，并且可以自适应地处理少量和大量的攻击者。MedMNIST2D 数据集上的广泛实验，证实了 DP-FedAWA 算法是安全的，对恶意客户端具有较好的鲁棒性。在未来的工作中，我们将重点研究针对异构数据分布环境中的安全攻击和隐私泄漏的有效防御措施，并进一步研究如何在更具挑战性的对手模型下保护隐私。

参 考 文 献

[1] MCKINNEY S M, SIENIEK M, GODBOLE V, et al. International evaluation of an AI system for breast cancer screening [J]. Nature, 2020, 577(7788): 89-94.

[2] LEE J, SUN J, WANG F, et al. Privacy-preserving patient similarity learning in a federated environment: development and analysis[J]. JMIR Nedical Informatics, 2018, 6(2): e7744.

[3] ELSHAFEEY N, KOTROTSOU A, HASSAN A, et al. Multicenter study demonstrates radiomic features derived from magnetic resonance perfusion images identify pseudoprogression in glioblastoma[J]. Nature Communications, 2019, 10(1): 3170.

[4] KAISSIS G, ZIEGELMAYER S, LOHÖFER F, et al. A machine learning model for the prediction of survival and tumor subtype in pancreatic ductal adenocarcinoma from preoperative diffusion-weighted imaging[J]. European Radiology Experimental, 2019, 3(1): 1-9.

[5] LU H, ARSHAD M, THORNTON A, et al. A mathematical-descriptor of tumor-mesoscopic-structure from computed-tomography images annotates prognostic-and molecular-phenotypes of epithelial ovarian cancer[J]. Nature Communications, 2019, 10(1): 764.

[6] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]// Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 2015: 1322-1333.

[7] BLANCHARD P, EL MHAMDI E M, GUERRAOUI R, et al. Machine learning with adversaries: Byzantine tolerant gradient

descent[J]. Advances in Neural Information Processing Systems,2017,30.

[8] ABADI M,CHU A,GOODFELLOW I,et al. Deep learning with differential privacy[C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016:308-318.

[9] DWORK C. Differential privacy[C]// Proceedings of the Automata,Languages and Programming:33rd International Colloquium(ICALP 2006). Venice,Italy,2006:1-12.

[10] GEYER R C,KLEIN T,NABI M. Differentially private federated learning:A client level perspective[J]. arXiv:171207557,2017.

[11] SATHYA S S,VEPAKOMMA P,RASKAR R,et al. A review of homomorphic encryption libraries for secure computation[J]. arXiv:181202428,2018.

[12] XU G,LI H,ZHANG Y,et al. Privacy-preserving federated deep learning with irregular users[J]. IEEE Transactions on Dependable and Secure Computing,2020,19(2):1364-1381.

[13] KELLER M,PASTRO V,ROTARU D. Overdrive:making SPDZ great again[C]// 37th Annual International Conference on the Theory and Applications of Cryptographic Techniques(EUROCRYPT 2018). Tel Aviv,Israel,2018:158-189.

[14] BOYLE E,GILBOA N,ISHAI Y. Function secret sharing:Improvements and extensions[C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016:1292-1303.

[15] MOHASSEL P,ZHANG Y. Secureml:A system for scalable privacy-preserving machine learning[C]// Proceedings of the 2017 IEEE Symposium on Security and Privacy(SP). IEEE,2017:19-38.

[16] TANG L T,WANG D,ZHANG L F,et al. Federated learning scheme based on secure multi-party computation and differential privacy[J]. Computer Science,2022,49(9):297-305.

[17] LI S,CHENG Y,LIU Y,et al. Abnormal client behavior detection in federated learning[J]. arXiv:191009933,2019.

[18] XIE C,KOYEJO S,GUPTA I. Zeno:Distributed stochastic gradient descent with suspicion-based fault-tolerance[C]// Proceedings of the International Conference on Machine Learning. PMLR,2019:6893-6901.

[19] CAO X,LAI L. Distributed gradient descent algorithm robust to an arbitrary number of byzantine attackers[J]. IEEE Transactions on Signal Processing,2019,67(22):5850-5864.

[20] CAO X,FANG M,LIU J,et al. Fltrust:Byzantine-robust federated learning via trust bootstrapping[J]. arXiv:201213995,2022.

[21] BLANCHARD P,MHAMDI E,GUERRAOUI R,et al. Machine learning with adversaries:byzantine tolerant gradient descent[C]// Proceedings of the Neural Information Processing Systems. 2017.

[22] XIA Q,TAO Z,HAO Z,et al. FABA:an algorithm for fast aggregation against byzantine attacks in distributed neural networks[C]// Proceedings of the IJCAI. 2019.

[23] YIN D,CHEN Y,KANNAN R,et al. Byzantine-robust distributed learning:Towards optimal statistical rates[C]// Proceedings of the International Conference on Machine Learning. PMLR,2018:5650-5659.

[24] GUERRAOUI R,ROUAULT S. The hidden vulnerability of distributed learning in byzantium[C]// Proceedings. of the International Conference on Machine Learning. PMLR,2018:3521-3530.

[25] CHEN Y,SU L,XU J. Distributed statistical machine learning in adversarial settings:Byzantine gradient descent[C]// Proceedings of the ACM on Measurement and Analysis of Computing Systems. 2017:1-25.

[26] PILLUTLA K,KAKADE S M,HARCHAOUI Z. Robust aggregation for federated learning[J]. IEEE Transactions on Signal Processing,2022,70:1142-1154.

[27] YAN M,LIN Y,NIE Z S,et al. Training Method to Improve Robustness of Federated Learning[J]. Computer Science,2022,49(S1):496-501.

[28] HONGYAN C,VIRAT S,REZA S,et al. Cronus:Robust and heterogeneous collaborative learning with black-box knowledge transfer[J]. arXiv:191211279,2019.

[29] MIAO Y,LIU Z,LI H,et al. Privacy-preserving Byzantine-robust federated learning via blockchain systems[J]. IEEE Transactions on Information Forensics and Security,2022,17:2848-2861.

[30] TANG X,SHEN M,LI Q,et al. PILE:Robust Privacy-Preserving Federated Learning via Verifiable Perturbations[J]. IEEE Transactions on Dependable and Secure Computing,2023:1-18.

[31] TAN Z,ZHANG L. Survey on privacy preserving techniques for machine learning[J]. J Softw,2020,31(7):2127-2156.

[32] DWORK C,ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends © in Theoretical Computer Science,2014,9(3):211-407.

[33] MCMAHAN H B,RAMAGE D,TALWAR K,et al. Learning differentially private recurrent language models[J]. arXiv:171006963,2017.

[34] 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. 2017:603-618.

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

[36] YANG J,SHI R,WEI D,et al. Medmnist v2:A large-scale lightweight benchmark for 2d and 3d biomedical image classification[J]. arXiv:211014795,2021.

ZHANG Lianfu, born in 1978, Ph.D candidate, is a member of China Computer Federation. His main research interests include information security and privacy-preserving machine learning.

TAN Zuowen, born in 1967, Ph.D, professor, PhD supervisor, is a member of China Computer Federation. His main research interests include cryptography, blockchain and privacy-preserving machine learning.