基于联邦学习的本地化差分隐私机制研究

任一支[①] 刘容轲[①] 王 冬*[①] 袁理锋[①] 申延召[②]
吴国华[①] 王秋华[①] 杨昌天[①]

[①](杭州电子科技大学网络空间安全学院 杭州 310018)
[②](山东区块链研究院 济南 250000)

摘 要：联邦学习与群体学习作为当前热门的分布式机器学习范式，前者能够保护用户数据不被第三方获得的前提下在服务器中实现模型参数共享计算，后者在无中心服务器的前提下利用区块链技术实现所有用户同等地聚合模型参数。但是，通过分析模型训练后的参数，如深度神经网络训练的权值，仍然可能泄露用户的隐私信息。目前，在联邦学习下运用本地化差分隐私(LDP)保护模型参数的方法层出不穷，但皆难以在较小的隐私预算和用户数量下缩小模型测试精度差。针对此问题，该文提出正负分段机制(PNPM)，在聚合前对本地模型参数进行扰动。首先，证明了该机制满足严格的差分隐私定义，保证了算法的隐私性；其次分析了该机制能够在较少的用户数量下保证模型的精度，保证了机制的有效性；最后，在3种主流图像分类数据集上与其他最先进的方法在模型准确性、隐私保护方面进行了比较，表现出了较好的性能。

关键词：隐私保护；联邦学习；本地化差分隐私；区块链

中图分类号：TN918.1; TP309.2 **文献标识码**：A **文章编号**：1009-5896(2023)03-0784-09

DOI: 10.11999/JEIT221064

A Study of Local Differential Privacy Mechanisms Based on Federated Learning

REN Yizhi[①] LIU Rongke[①] WANG Dong[①] YUAN Lifeng[①] SHEN Yanzhao[②]
WU Guohua[①] WANG Qiuhua[①] YANG Changtian[①]

[①](*School of Cyberspace Security, Hangzhou Dianzi University, Hangzhou* 310018, *China*)
[②](*Shandong Blockchain Research Institute, Jinan* 250000, *China*)

Abstract: Federated Learning and swarm Learning, as currently popular distributed machine learning paradigms, the former enables shared computation of model parameters in servers while protecting user data from third parties, while the latter uses blockchain technology to aggregate model parameters equally for all users without a central server. However, by analyzing the parameters after model training, such as the weights of deep neural network training, it is still possible to leak the user's private information. At present, there are several methods for protecting model parameters utilizing Local Differential Privacy (LDP) in federated learning, however it is challenging to reduce the gap in model testing accuracy when there is a limited privacy budget and user base. To solve this problem, a Positive and Negative Piecewise Mechanism (PNPM) is proposed, which perturbs the local model parameters before aggregation. First, it is proved that the mechanism satisfies the strict definition of differential privacy and ensures the privacy of the algorithm; Secondly, it is analyzed that the mechanism can ensure the accuracy of the model under a small number of users and ensure the effectiveness of the mechanism; Finally, it is compared with other state-of-the-art methods in terms of model accuracy and privacy protection on three mainstream image classification datasets and shows a better performance.

Key words: Privacy preserving; Federated learning; Local Differential Privacy (LDP); Blockchain

收稿日期：2022-08-12；改回日期：2022-11-30；网络出版：2022-12-02

*通信作者： 王冬 wangdong@hdu.edu.cn

基金项目：浙江省“尖兵”、“领雁”研发攻关项目(2022C03174)，浙江省教育厅科研项目(Y202147115)，浙江省属高校基本科研业务费专项资金资助项目(GK229909299001-023)

Foundation Items: “Pioneer” and “Leading Goose” R&D Program of Zhejiang (2022C03174), The A Project Supported by Scientific Research Fund of Zhejiang Provincial Education Department (Y202147115), The Fundamental Research Funds for the Provincial Universities of Zhejiang (GK229909299001-023)

1　引言

随着深度学习技术的兴起，各领域都开始对其进行研究并投入实际应用。与此同时，越来越多的敏感数据被作为深度学习模型的训练对象，比如医疗研究中心为了得到辅助诊断新冠肺炎的模型，需收集新冠肺炎患者的肺部CT影像[1]进行训练。但实际应用中，用户数据通常具有隐私性极强的特点，并且数据安全和隐私保护的要求使得这些数据无法聚合起来，形成数据孤岛。而单个终端的数据样本量又不足以支撑大规模的深度学习模型训练，导致模型性能不佳以及难以实现数据价值等问题。

联邦学习[2]作为一种分布式机器学习技术可以解决数据孤岛问题，使机构间可以跨地域协作而数据不出本地，且多方合作构建的预测模型能够更准确地预测各种问题[3]。其旨在不直接上传本地数据的前提下，上传在本地数据集训练后的模型参数，服务器收到多名用户的模型参数后进行聚合，重复多次得到一个全局模型。群体学习[4]解决了联邦学习中服务器可能存在安全隐患的问题，基于区块链的对等网络，使参与者之间同等地传输与收集多方的模型参数，并在新一轮全局迭代前进行聚合而无需服务器管理。然而，在一些文献中指出，这些模型参数也会泄露用户隐私[5,6]，甚至据此还原原始敏感数据[7,8]。而且，随着研究的不断深入，深度学习模型中存在的隐私泄露问题也逐渐暴露[9–11]，由于深度神经网络有大量的隐藏层，这些隐藏层会形成较大的有效容量，足以将某些数据的细节编码为模型参数，甚至记忆整个数据集[5]。

为解决上述问题，文献[12–16]提出将差分隐私与联邦学习结合的方案，Tramèr等人[10]认为将差分隐私直接应用于模型参数上是更合适的策略。但由于隐私预算和性能之间的权衡，多数文献无法在具有复杂数据集的深度学习中提供切实可行的解决方案。Sun等人[7]提供的本地化差分隐私-联邦学习(Local Differential Privacy-Federated Learning, LDP-FL)方案解决了如上问题，但在少量用户下，仍存在较大的模型测试精度差。

本文根据文献[17,18]所提出的两种本地化差分隐私(Local Differential Privacy, LDP)机制进行改进，提出正负分段机制(Positive and Negative Piecewise Mechanism, PNPM)来解决上述问题。由于本机制能产生更小的方差，所以在用户数量与隐私预算较小时，能得到相对更小的模型测试精度差。本机制在隐私预算ε为1时，都在手写数字数据集[19](Mixed National Institute of Standards and Technology database, MNIST)、时装数据集[20](Fashion-Mixed National Institute of Standards and Technology database, Fashion-MNIST)、图像10分类数据集[21](Canadian institute for advanced research-10, Cifar-10)上得到了极小的模型测试精度差。

2　相关工作

联邦学习本质是一种分布式的机器学习，由中心服务器、参与者和通信网络3部分组成[22]。虽然联邦学习使数据不出用户本地，只需上传训练参数即可，但应用过程中存在一定的安全隐患。对于上传梯度更新参数，Zhu等人[23]通过截取上传的梯度数据就可还原模型训练数据。对于上传模型参数，Fredrikson等人[6]在计算机视觉模型的输出概率上使用爬山法来揭示训练数据中的个人面孔。但在卷积神经网络上恢复的图像不自然或无特征。Yang等人[24]通过辅助训练集训练第2个神经网络反演目标模型的训练集人脸图像。Zhang等人[25]通过定制辅助训练集来训练生成式对抗网络(Generative Adversarial Networks, GAN)，利用生成器来反演目标模型的训练集人脸图像。

由于上述联邦学习中存在的隐私保护问题，近几年学者提出了很多将差分隐私与联邦学习[12,13,15]结合的工作。但由于隐私预算和性能之间的权衡，多数文献无法在具有复杂数据集的深度学习中提供切实可行的解决方案。

Bhowmick等人[26]第1个将本地化差分隐私与深度学习模型的联邦学习相结合，提出了切实可行的大规模本地隐私模型训练方法，并在模型效用几乎没有下降的情况下适合大规模图像分类和语言模型。Truex等人[27]提出了精简的本地化差分隐私(α-Condensed Local Differential Privacy-Federated learning, α-CLDP-Fed)算法扩展到联邦学习算法中，可以在LDP保证下处理高维、连续和大规模的模型参数更新。但文献[26,27]很难在较小的隐私预算下得到良好的表现。Sun等人[7]提出了基于文献[17]改进的LDP机制，对本地模型参数进行扰动，首次考虑了模型每层权重的取值范围，降低了前人机制方差较大以及上述文献存在的问题。文献[17,18]提供了两种基于均值统计的本地化差分隐私方案，但在较小的隐私预算下直接对模型参数进行扰动，由于方差较大，若用户数量较少而不足以抵消噪声，聚合后的模型在反向传播中存在梯度爆炸的问题。本文结合以上两种方案进行改进，与文献[7]对比，可以在相同甚至更小的隐私预算下得到不错的表现。

3 基于联邦学习的本地化差分隐私机制

本文提出基于联邦学习的本地化差分隐私机制，联邦学习框架如图1所示。其中由服务器更新与本地更新两步组成，具体由算法1所述。

服务器每轮迭代随机挑选部分用户，所述用户：

(1)首先下载服务器初始化或聚合更新的模型参数，然后在本地数据上通过随机梯度下降并行地训练更新该模型参数；

(2)其次通过本文扰动机制对模型参数进行扰动；最后将扰动后的模型参数上传至服务器聚合，直至迭代结束。

3.1 正负分段扰动机制

目前已有多项研究提出了切实可行的大规模本地隐私模型训练方法，通过将本地化差分隐私技术与深度学习模型的联邦学习相结合，以在联合模型效用几乎没有下降的情况下适合大规模图像分类和语言模型。Duchi等人[17]和Wang等人[18]提供了两种基于均值统计的本地化差分隐私方案，其设计的均值统计机制各具优点：文献[17]设计的机制将输入值$t,t\in[-1,1]$输出为一个经隐私预算计算后的值$\pm\dfrac{\mathrm{e}^{\varepsilon}+1}{\mathrm{e}^{\varepsilon}-1}$，其中正负性通过输入值计算，输入值绝对值越大，翻转概率越小；文献[18]设计的机制将输入值$t,t\in[-1,1]$输出为扰动域$t^*\in[-C,C]$内的随机值，其中正负性通过隐私预算及输入值大小决定。前者输出值单一，但保证了输入值隐私性；后者输出值多变且降低了机制方差。

算法1 结合本地化差分隐私保护的联邦学习算法

输入：n是本地用户的数量；B是本地mini-batch大小；LE为本地迭代数量；GE为全局迭代数量；γ为学习率；Fr为用户比率。

(1)服务器更新

(2)初始化全局模型W_0；

(3)发送给用户 ；

(4)进行GE轮迭代

(5) 随机选取$m=\mathrm{Fr}\cdot n$个本地用户；

(6) 收集发送给服务器的m个更新后的模型参数；

(7) # 聚合并更新全局模型W_{GE}；

(8) for 每个相同位置权重$w\in\{W_{\mathrm{GE}}^m\}$ do

(9) $w\leftarrow\dfrac{1}{m}\sum w$;

(10) 发送给用户 ；

(11)本地更新

(12)接受发送给用户的全局模型W_{GE}，用户数量m；

(13)for 每个本地用户$s\in m$ do

(14) $W_{\mathrm{GE}+1}^s\leftarrow W_{\mathrm{GE}}$;

(15) 本地进行LE轮迭代

(16) for 每个批$b\in B$ do

(17) $W_{\mathrm{GE}+1}^s\leftarrow W_{\mathrm{GE}+1}^s-\gamma\nabla L\left(W_{\mathrm{GE}+1}^s;b\right)$;

(18) # 对本地模型参数进行扰动

(19) for 每个$w\in W_{\mathrm{GE}+1}^s$ do

(20) $w\leftarrow|w|\cdot\mathrm{PNPM}\left(\dfrac{w}{|w|}\right)$;

(21) 发送给服务器；

(22)返回。

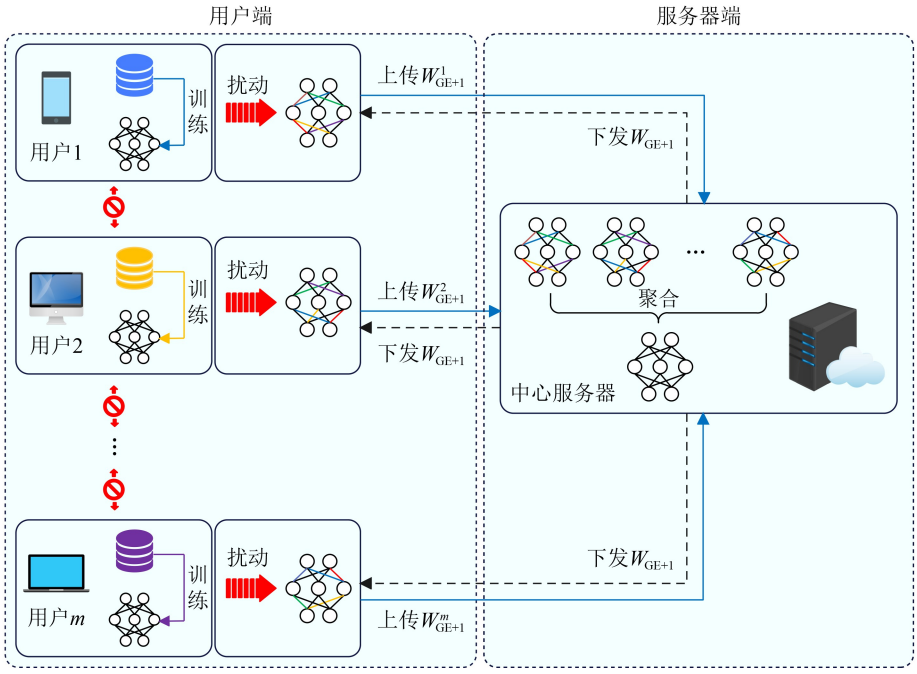

图 1 结合本地化差分隐私保护的联邦学习框架

但上述两种方案在较小的隐私预算下直接对模型参数进行扰动，由于方差较大，若用户数量较少而不足以抵消噪声，聚合后的模型在反向传播中存在梯度爆炸的问题。本文结合文献[17,18]设计的机制进行优化，提出一种正负分段机制(PNPM)(见算法2)，本机制在保证隐私性的同时，降低方差、提高聚合后模型精度。通过输入权重的正负方向$t\in\{-1,1\}$，经机制随机响应后输出一个扰动值$t^*\in[-C,-1]\cup[1,C]$，令权重的绝对值$|w|$乘以扰动值t^*，以此得到新的扰动权重$w=|w|\cdot t^*$来替代。本机制不仅直接概率性地翻转了权重的正负方向，并且通过乘以一个扰动域的数以达到权重脱敏的效果，最终在多方的参数聚合下抵消噪声，得到可用的联合模型参数。

其中扰动域的边界$C=\dfrac{\mathrm{e}^{\varepsilon}+3}{\mathrm{e}^{\varepsilon}-1}$，扰动值$t^*$的概率密度函数满足

$$\mathrm{pdf}(t^*=y|t)=\begin{cases}p, & y\in[l(t),r(t)]\\ \dfrac{p}{\mathrm{e}^{\varepsilon}}, & y\in[-r(t),-l(t)]\end{cases}\tag{1}$$

其中的概率p与范围边界满足

$$p=\frac{\mathrm{e}^{2\varepsilon}-\mathrm{e}^{\varepsilon}}{4(\mathrm{e}^{\varepsilon}+1)},l(t)=\frac{C+1}{2}\cdot t-\frac{C-1}{2},\\ r(t)=l(t)+C-1\tag{2}$$

表1为$t=-1$或$t=1$时扰动值t^*的概率密度函数输出值。

表 1　$t=-1$或$t=1$扰动值t^*的概率密度函数值

	$t=-1$		$t=1$	
	$-r(t)\le y\le -l(t)$	$l(t)\le y\le r(t)$	$-r(t)\le y\le -l(t)$	$l(t)\le y\le r(t)$
$\mathrm{pdf}(t^*=y\|t)$	$\dfrac{\mathrm{e}^{2\varepsilon}-\mathrm{e}^{\varepsilon}}{4(\mathrm{e}^{\varepsilon}+1)}$	$\dfrac{\mathrm{e}^{\varepsilon}-1}{4(\mathrm{e}^{\varepsilon}+1)}$	$\dfrac{\mathrm{e}^{\varepsilon}-1}{4(\mathrm{e}^{\varepsilon}+1)}$	$\dfrac{\mathrm{e}^{2\varepsilon}-\mathrm{e}^{\varepsilon}}{4(\mathrm{e}^{\varepsilon}+1)}$

算法2 正负分段机制PNPM

输入：$t\in\{-1,1\}$和隐私预算ε。

输出：$t^*\in[-C,-1]\cup[1,C]$。

(1)在$[0,1][0,1]$内均匀地取一个随机数x；

(2)如果$x<\dfrac{\mathrm{e}^{\epsilon}}{\mathrm{e}^{\epsilon}+1}$，那么

(3)　t^*在$[l(t),r(t)]$上均匀取一个随机数；

(4)否则

(5)　t^*在$[-r(t),-l(t)]$上均匀取一个随机数；

(6)返回t^*。

3.2　可用性与隐私性分析

经正负分段机制扰动后的用户模型将不具备原始精度，并且隐私保护程度越强，该模型精度越低，但经多方聚合后的模型参数逼近原始聚合后的参数值，从而实现个体得到隐私保护的同时，多方聚合后的模型参数可用。

以下引理确立了PNPM的可用性与隐私性理论保证。

引理1　PNPM满足ε-本地化差分隐私。并且在均值估计权重时引入了零偏差，即$\mathbb{E}\left[\overline{M(w)}\right]=\bar{w}$。并且$\mathrm{Var}[M(w)]=|w|^2\cdot\dfrac{4(\mathrm{e}^{\varepsilon}+1/3)}{(\mathrm{e}^{\varepsilon}-1)^2}$，其中$M$为PNPM，$w$为模型参数中的一个权重。

引理1直接说明了该机制的可用性，即聚合扰动模型后的联合模型逼近原始聚合后的模型效用，且较于前人机制有更小的方差。为了直观地进行对比，图2为PNPM、文献[17]、分段机制[18](Piecewise Mechanism, PM)3种机制在不同权重下的方差对比。

由于引理1证明了机制作用于权重的方差$\mathrm{Var}[M(w)]$，以及在均值估计权重时引入了零偏差，即$\mathbb{E}\left[\overline{M(w)}\right]=\bar{w}$，其中$\overline{M(w)}$为所有用户权重$w$的均值聚合估计值，由此得出以下引理确立了$\overline{M(w)}$的渐进误差边界。

引理2　对于任意的参数$w\in W$，有$1-\beta$概率使$\left|\overline{M(w)}-\bar{w}\right|<O\left(\dfrac{|w|\sqrt{\ln(1/\beta)}}{\varepsilon\sqrt{n}}\right)$。

引理2建立了$\overline{M(w)}$的精度保证，且由引理可知，PNPM使扰动后的模型参数与原模型参数的正负不可区分，且随着权重绝对值的增大，方差与渐进误差边界增大，扰动程度更大。引理的证明在附录中详细说明。

在卷积神经网络中，卷积核权重的正负直接决定了如何提取特征。为了验证卷积核权重正负对模型精度的影响，如表2所示，对不同数据集训练后的模型参数进行如下操作：在不改变卷积核参数正负方向的前提下，做如表2中两种赋值。

由结果可知，简单的赋值即可在MNIST和Fashion-MNIST数据集(灰白图数据集)训练后的模

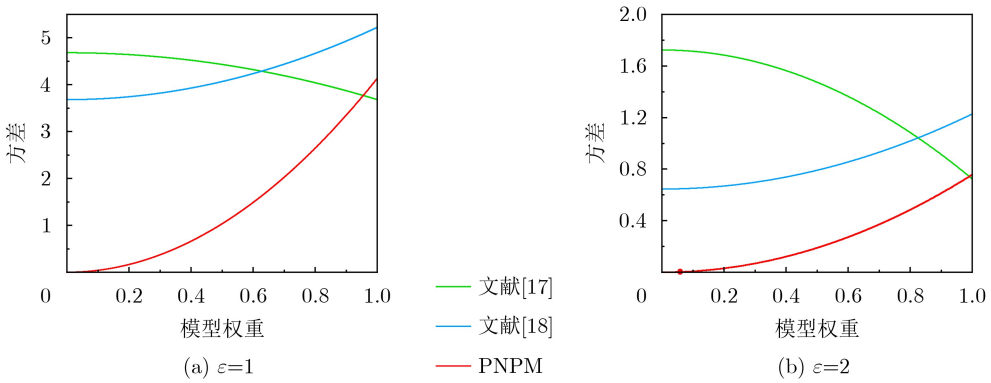

图 2 不同隐私预算下权重大小对于方差的影响

表 2 在不改变卷积核参数正负的前提下，两种赋值所能得到的精度(%)

数据集	每个权重绝对值都为0.5	每个权重在(–1,0)∪(0,1)上随机取值(20次随机中的最大精度)	未做处理的模型原精度
MNIST	78.67	89.10	98.98
Fashion-MNIST	75.26	72.83	89.89
Cifar-10	28.00	27.18	71.83

型中得到较高的精度，卷积核参数正负分布存在隐私泄露的风险。实际的病症分析中，如患者的CT影像、核磁共振影像等，仍采用灰白图进行分析。若采用如上的患者数据进行模型训练，则具备参数正负分布背景知识的攻击者可能轻松地恢复出本地训练后的模型精度，而其中的参数也将暴露用户的隐私。如果只是对权重增加一个符合拉普拉斯、高斯分布的噪声，且噪声方差不大，参数的正负方向基本没发生翻转。

本机制能让攻击者在观察到权重时，无法得出原始权重是正数或负数，其次对正负扰动后权重乘以一个扰动域内的随机数，可以破坏本地上传模型的可用性，能有效地抵御模型逆向攻击[6]与成员推断攻击[28]，保证了本地模型参数的隐私性。

4 实验结果分析

本实验在MNIST[19], Fashion-MNIST[20], Cifar-10[21]数据集上，对隐私保护机制下的模型进行隐私性和可用性的评估。对于MNIST数据集本文采用两层卷积层、两层全连接层组成的卷积神经网络，共21750个权重值；对于Fashion-MNIST和Cifar-10数据集采用3层卷积层、两层全连接层组成的卷积神经网络，分别共有41936, 56560个权重值。MNIST, Fashion-MNIST数据集实验的全局迭代轮数为10；Cifar-10数据集实验中，总用户数量为500时，全局迭代轮数为20，本地迭代次数为20；总用户数量为100时，本地迭代次数为10。在MNIST, Fashion-MNIST数据集中，用户总数为100时学习率设置为0.01，用户总数为500时学习率设置为0.05；Cifar-10数据集中，用户总数为100时学习率设置为0.06，用户总数为500时学习率设置为0.05。

4.1 隐私保证下的可用性评估

本实验分别在3个数据集上，对本地上传的模型进行扰动，统计不同隐私预算下的扰动模型精度。图3为不同隐私预算对本地上传的模型精度影响，此处的模型精度是本地训练整个数据集后的精度。

如图3所示，随着隐私预算的增大，经PNPM扰动后的模型精度也会大幅上升(图3中精度是扰动20次后取的平均值)。例如在MNIST实验中，隐私预算为2, 3时精度分别为30.16%, 72.04%，因为隐私预算的增大会导致正负翻转的概率下降，且权重值的变动更小，所以为了提供较好的隐私性，隐私预算最好设置2以下。虽然文献[17,18]的机制对模型提供了很强的隐私性，但同时也降低了可用性，即需要大量用户通过聚合抵消噪声。

4.2 多用户场景实验结果评估

本实验分别在3个数据集上，对多用户场景下

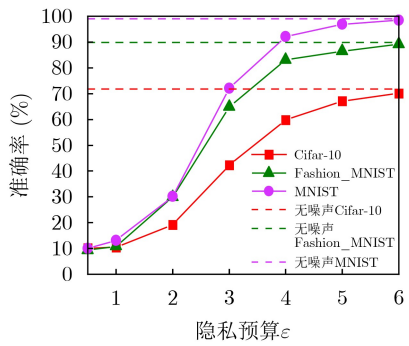

图 3 不同隐私预算对于3种数据集训练后的本地模型精度影响

未扰动与经PNPM扰动后的全局模型精度进行对比。图4(a)—图4(c)分别为3个数据集在不同用户数量中，未扰动与经PNPM扰动后的精度对比柱状图。

图4(a)隐私预算为0.5，后两者隐私预算为1，用户数量同为500。隐私预算为0.5与1时，由于采用PM和Duchi机制均存在梯度爆炸的现象，所以没有进行对比。在隐私预算ε为0.5时，若用户数量足够多，MNIST数据集训练下的精度差能控制在2%以下，如用户数量为400时，聚合经PNPM扰动后的模型精度为94.00%，原始聚合精度为95.46%；在隐私预算ε为1、用户数量为100时，Fashion-MNIST数据集训练下的精度为81.20%，原始聚合精度为83.56%，且其余用户数量下的精度差值均在5.00%以下，并且当用户达到300人以上时，精度几乎没有偏差；在Cifar-10数据集训练下的精度差均在10.00%以下，并且在用户为250人以上时，精度能达到很小的偏差。

在MNIST数据集中，Sun等人[7]设置的隐私预算为1，且当用户占比大于0.5时，与本实验得到的精度差基本相同，均在2%以下，而本实验设置的隐私预算为0.5；在Fashion-MNIST，Cifar-10数据集中，前者将隐私预算设置为5和10，而本实验在隐私预算为1和2时均能得到更小的精度差。

4.3 少用户场景实验结果评估

本实验分别在3个数据集上，对少用户场景下未扰动与分别经PNPM，文献[17,18]扰动后的全局模型精度进行对比。图5(a)—图5(c)分别为3个数据集在不同用户数量中，未扰动与分别经3种不同机制扰动后的精度对比柱状图。

图5(a)隐私预算为1，后两者隐私预算为2，用户数量同为100。部分用户数量下未展示精度，即在训练过程中出现梯度爆炸。如图5(b)所示，当隐私预算为1、用户数量为70时原始精度、经PNPM、文献[17,18]扰动后的聚合精度分别为86.07%, 85.92%, 66.98%, 54.83%。在3个数据集下，除了用户数量为10时，PNPM均能在各数据集中将精度差控制在10%以内，若用户数量只能控制在10左右，需考虑提高隐私预算。相比于其他方法，PNPM所需的隐私预算更少，并且也适用于少量用户的场景。

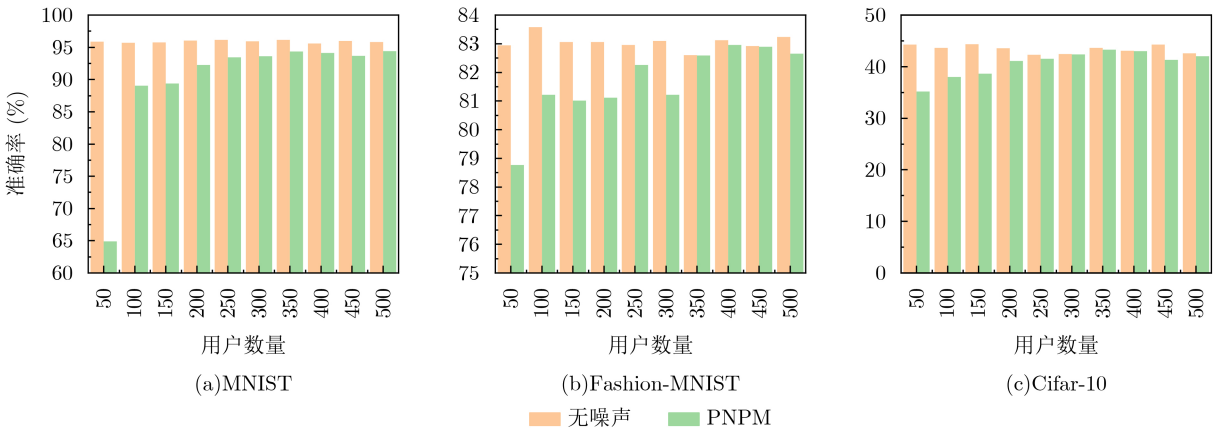

图 4 多用户场景相同隐私预算下用户数量对精度的影响

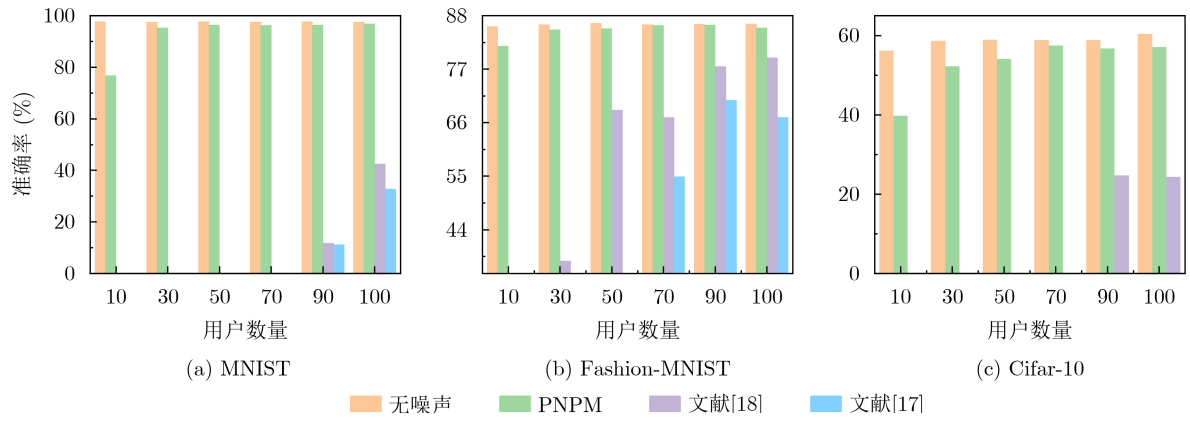

图 5 少用户场景相同隐私预算下用户数量对精度的影响

5 结束语

本文针对现有本地化差分隐私机制难以在较小隐私预算和用户数量下缩小模型测试精度差的问题，提出正负分段的差分隐私扰动机制，即正负分段机制，在聚合前对本地模型参数进行扰动。利用卷积神经网络卷积核参数正负的特性，对参数的正负方向进行翻转决定扰动域范围，其次令权重绝对值乘以该扰动域的随机数以达到脱敏效果，保证了本地用户数据的隐私性。根据机制在均值估计权重时引入零偏差以及更小方差的引理，服务器聚合扰动后的模型参数仍具有可用性。在MNIST，Fashion-MNIST，Cifar-10数据集上进行实验，验证了本文机制能在多用户场景下，采用更小的隐私预算降低模型精度差；在少用户场景下，与其他最先进的机制对比，精度差更低，可用性更高。今后的研究将考虑如何改进机制，使得提供更强隐私保护的前提下减小精度差，帮助本地化差分隐私在联邦学习中的应用。

附录

引理 1 证明 对于$\forall t^* \in [-C,-1]\cup[1,C]$以及任意的输入值$t,t'=1$ or -1，都满足$\frac{\mathrm{pdf}(t^*|t)}{\mathrm{pdf}(t^*|t')} \le \frac{p}{p/\mathrm{e}^{\varepsilon}} = \mathrm{e}^{\varepsilon}$-本地化差分隐私。并且对于任意更新后的权重$w=|w|\cdot t$

$$\begin{aligned}\mathbb{E}[t^*] &= \int_{-r(t)}^{-l(t)} \frac{p}{\mathrm{e}^{\varepsilon}} x\mathrm{d}x + \int_{l(t)}^{r(t)} px\mathrm{d}x \\ &= \frac{p}{\mathrm{e}^{\varepsilon}}\cdot\frac{1}{2}\cdot\left[l(t)^2 - r(t)^2\right] + p\cdot\frac{1}{2}\cdot\left[r(t)^2 - l(t)^2\right] = \frac{1}{2}\cdot\frac{\mathrm{e}^{\varepsilon}}{\mathrm{e}^{\varepsilon}+1}\cdot\frac{2(\mathrm{e}^{\varepsilon}+1)}{\mathrm{e}^{\varepsilon}-1}\cdot t\cdot\frac{\mathrm{e}^{\varepsilon}-1}{\mathrm{e}^{\varepsilon}} = t\end{aligned} \tag{3}$$

$$\mathbb{E}[M(w)] = \mathbb{E}[|w|\cdot M(t)] = |w|\cdot\mathbb{E}[t^*] = |w|\cdot t = w, \mathbb{E}\left[\overline{M(w)}\right] = \mathbb{E}\left[\frac{1}{n}\sum_u M(w_u)\right] = \frac{1}{n}\sum_u w = \bar{w} \tag{4}$$

而且

$$\mathrm{Var}[t^*] = \mathbb{E}\left[t^{*2}\right] - \mathbb{E}[t^*]^2 = \frac{\mathrm{e}^{2\varepsilon}+2\mathrm{e}^{\varepsilon}+7/3}{(\mathrm{e}^{\varepsilon}-1)^2} - 1 = \frac{4(\mathrm{e}^{\varepsilon}+1/3)}{(\mathrm{e}^{\varepsilon}-1)^2}, \mathrm{Var}[M(w)] = |w|^2\cdot\mathrm{Var}[t^*] = |w|^2\cdot\frac{4(\mathrm{e}^{\varepsilon}+1/3)}{(\mathrm{e}^{\varepsilon}-1)^2} \tag{5}$$

引理 2 证明 对任意的用户u，$|M(w_u)-w_u| \le |w|\cdot C + |w| = 2\cdot|w|\cdot\frac{\mathrm{e}^{\varepsilon}+1}{\mathrm{e}^{\varepsilon}-1}$，并且$\mathrm{Var}[M(w_u)] = \mathrm{Var}[M(w_u)-w_u] = \mathbb{E}\left[(M(w_u)-w_u)^2\right] - \mathbb{E}[M(w_u)-w_u]^2 = \mathbb{E}\left[(M(w_u)-w_u)^2\right]$。

通过伯恩斯坦不等式可知

$$\begin{aligned}\Pr\left[\left|\overline{M(w)} - \bar{w}\right| \ge \lambda\right] &= \Pr\left[\left|\sum_u (M(w_u)-w_u)\right| \ge n\lambda\right] \le 2 \\ &\cdot\exp\left(-\frac{1/2n^2\lambda^2}{\sum_u \mathbb{E}\left((M(w_u)-w_u)^2\right) + 2n\lambda|w|(\mathrm{e}^{\varepsilon}+1)/3(\mathrm{e}^{\varepsilon}-1)}\right) \\ &= 2\cdot\exp\left(-\frac{n\lambda^2}{2|w|^2\cdot\frac{4(\mathrm{e}^{\varepsilon}+1/3)}{n(\mathrm{e}^{\varepsilon}-1)^2} + \frac{4\lambda|w|(\mathrm{e}^{\varepsilon}+1)}{3(\mathrm{e}^{\varepsilon}-1)}}\right) = 2\cdot\exp\left(-\frac{n\lambda^2}{|w|^2\cdot O(\varepsilon^{-2}) + \lambda|w|\cdot O(\varepsilon^{-1})}\right)\end{aligned} \tag{6}$$

根据联合约束，存在$\lambda = O\left(\frac{|w|\sqrt{\ln\left(1/\beta\right)}}{\varepsilon\sqrt{n}}\right)$使$\left|\overline{M(w)} - \bar{w}\right| < \lambda$至少以$1-\beta$的概率成立。

参 考 文 献

[1] WANG Shuai, KANG Bo, MA Jinlu, *et al*. A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19)[J]. *European Radiology*, 2021, 31(8): 6096–6104. doi: 10.1007/s00330-021-07715-1.

[2] MCMAHAN H B, MOORE E, RAMAGE D, *et al*. Communication-efficient learning of deep networks from decentralized data[EB/OL]. https://doi.org/10.48550/arXiv.1602.05629, 2016.

[3] 杨强. AI与数据隐私保护: 联邦学习的破解之道[J]. 信息安全研究, 2019, 5(11): 961–965. doi: 10.3969/j.issn.2096-1057.2019.11.003.
YANG Qiang. AI and data privacy protection: The way to federated learning[J]. *Journal of Information Security Research*, 2019, 5(11): 961–965. doi: 10.3969/j.issn.2096-1057.2019.11.003.

[4] WARNAT-HERRESTHAL S, SCHULTZE H, SHASTRY K L, *et al*. Swarm Learning for decentralized and confidential clinical machine learning[J]. *Nature*, 2021, 594(7862): 265–270. doi: 10.1038/s41586-021-03583-3.

[5] SONG Congzheng, RISTENPART T, and SHMATIKOV V. Machine learning models that remember too much[C]. 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, USA, 2017: 587–601. doi: 10.1145/3133956.3134077.

[6] FREDRIKSON M, JHA S, and RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]. 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, USA, 2015: 1322–1333. doi: 10.1145/2810103.2813677.

[7] SUN Lichao, QIAN Jianwei, and CHEN Xun. LDP-FL: Practical private aggregation in federated learning with local differential privacy[C]. The Thirtieth International Joint Conference on Artificial Intelligence, Montreal, Canada, 2021: 1571–1578. doi: 10.24963/ijcai.2021/217.

[8] PAPERNOT N, ABADI M, ERLINGSSON Ú, *et al*. Semi-supervised knowledge transfer for deep learning from private training data[C]. 5th International Conference on Learning Representations, Toulon, France, 2017.

[9] PAPERNOT N, MCDANIEL P, SINHA A, *et al*. SoK: Security and privacy in machine learning[C]. 2018 IEEE European Symposium on Security and Privacy, London, UK, 2018: 399–414. doi: 10.1109/EuroSP.2018.00035.

[10] TRAMÈR F, ZHANG Fan, JUELS A, *et al*. Stealing machine learning models via prediction APIs[C]. The 25th USENIX Conference on Security Symposium, Austin, USA, 2016: 601–618.

[11] WANG Binghui and GONG N Z. Stealing hyperparameters in machine learning[C]. 2018 IEEE Symposium on Security and Privacy, San Francisco, USA, 2018: 36–52. doi: 10.1109/SP.2018.00038.

[12] LYU Lingjuan, YU Han, MA Xingjun, *et al*. Privacy and robustness in federated learning: Attacks and defenses[J]. *IEEE Transactions on Neural Networks and Learning Systems*, To be published. doi: 10.1109/TNNLS.2022.3216981.

[13] SUN Lichao and LYU Lingjuan. Federated model distillation with noise-free differential privacy[C]. The Thirtieth International Joint Conference on Artificial Intelligence, Montreal, Canada, 2021: 1563–1570. doi: 10.24963/ijcai.2021/216.

[14] MCMAHAN H B, RAMAGE D, TALWAR K, *et al*. Learning differentially private recurrent language models[C]. 6th International Conference on Learning Representations, Vancouver, Canada, 2018.

[15] GEYER R C, KLEIN T, and NABI M. Differentially private federated learning: A client level perspective[EB/OL]. https://doi.org/10.48550/arXiv.1712.07557, 2017.

[16] NGUYÊN T T, XIAO Xiaokui, YANG Yin, *et al*. Collecting and analyzing data from smart device users with local differential privacy[EB/OL]. https://doi.org/10.48550/arXiv.1606.05053, 2016.

[17] DUCHI J C, JORDAN M I, and WAINWRIGHT M J. Local privacy, data processing inequalities, and statistical minimax rates[EB/OL]. https://doi.org/10.48550/arXiv.1302.3203, 2013.

[18] WANG Ning, XIAO Xiaokui, YANG Yin, *et al*. Collecting and analyzing multidimensional data with local differential privacy[C]. 2019 IEEE 35th International Conference on Data Engineering (ICDE), Macao, China, 2019: 638–649, doi: 10.1109/ICDE.2019.00063.

[19] 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. doi: 10.1109/5.726791.

[20] XIAO Han, RASUL K, and VOLLGRAF R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms[EB/OL]. https://doi.org/10.48550/arXiv.1708.07747, 2017.

[21] KRIZHEVSKY A. Learning multiple layers of features from tiny images[R]. Technical Report TR-2009, 2009.

[22] 邱晓慧, 杨波, 赵孟晨, 等. 联邦学习安全防御与隐私保护技术研究[J]. 计算机应用研究, 2022, 39(11): 3220–3231. doi: 10.19734/j.issn.1001-3695.2022.03.0164.
QIU Xiaohui, YANG Bo, ZHAO Mengchen, *et al*. Survey on federated learning security defense and privacy protection technology[J]. *Application Research of Computers*, 2022, 39(11): 3220–3231. doi: 10.19734/j.issn.1001-3695.2022.03.

0164.

[23] ZHU Ligeng, LIU Zhijian, and HAN Song. Deep leakage from gradients[C]. The 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada, 2019: 1323.

[24] YANG Ziqi, ZHANG Jiyi, CHANG E C, *et al*. Neural network inversion in adversarial setting via background knowledge alignment[C]. The 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, 2019: 225–240. doi: 10.1145/3319535.3354261.

[25] ZHANG Yuheng, JIA Ruoxi, PEI Hengzhi, *et al*. The secret revealer: Generative model-inversion attacks against deep neural networks[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2020: 250–258, doi: 10.1109/CVPR42600.2020.00033.

[26] BHOWMICK A, DUCHI J, FREUDIGER J, *et al*. Protection against reconstruction and its applications in private federated learning[EB/OL]. https://doi.org/10.48550/arXiv.1812.00984, 2018.

[27] TRUEX S, LIU Ling, CHOW K H, *et al*. LDP-fed: Federated learning with local differential privacy[C]. The Third ACM International Workshop on Edge Systems, Analytics and Networking, Heraklion, Greece, 2020: 61–66. doi: 10.1145/3378679.3394533.

[28] SHOKRI R, STRONATI M, SONG Congzheng, *et al*. Membership inference attacks against machine learning models[C]. 2017 IEEE Symposium on Security and Privacy (SP), San Jose, USA, 2017: 3–18. doi: 10.1109/SP.2017.41.

任一支：男，教授，研究方向为大数据安全、人工智能、区块链、知识图谱.

刘容轲：男，硕士生，研究方向为深度学习、隐私保护.

王 冬：女，博士，研究方向为隐私保护.

袁理锋：男，博士，研究方向为信息安全、自然语言处理、知识图谱.

申延召：男，博士，研究方向为信息安全、隐私保护.

吴国华：男，研究员，研究方向为信息安全.

王秋华：女，副教授，研究方向为大数据安全、数据治理.

杨昌天：男，博士，研究方向为保密教育、保密管理等.

责任编辑：余 蓉