基于函数机制的差分隐私联邦学习算法

曹世翔[1)] 陈超梦[1)] 唐 朋[2),3)] 苏 森[1)]

[1)](北京邮电大学计算机学院(国家示范性软件学院) 北京 100035)

[2)](山东大学密码技术与信息安全教育部重点实验室 山东 青岛 266237)

[3)](山东大学网络空间安全学院 山东 青岛 266237)

摘 要 随着数据隐私与安全越来越多地得到社会各界的重视,联邦学习作为一种保护用户数据隐私的分布式学习技术应运而生.它可以让原本数据隔离的不同组织能够协同进行机器学习.基于差分隐私的联邦学习是一种被广泛使用的隐私保护联邦学习框架.其通过在用户上传训练梯度前对梯度引入噪声来保护本地数据的隐私.但裁剪与加噪引起的梯度精度丢失问题限制着差分隐私联邦学习的训练效果进一步提高.针对此问题,本文提出了一种基于函数机制的差分隐私联邦学习算法,用扰动目标函数替代扰动梯度,使得噪声不受梯度阈值的影响,提高梯度的真实有效性.在基础方案中,我们提出的LDP-联邦主成分分析算法让所有用户以尽量统一的方式将本地数据映射到低维,让算法更适用于联邦学习高维数据场景.另外,我们发现:由于没有对梯度进行任何限制或裁剪,基础方案在用户数据非独立同分布或存在低质量数据的场景下存在权重发散问题,导致训练效果变差.本文又针对基础方案在上述场景存在的缺陷提出增强方案:通过 ϵ-DP-最低发散度指数抽样聚合方法以优化非独立同分布与低质量数据两个应用场景下的模型训练效果.经实验验证,本文提出的算法与方案优于目前其他隐私保护联邦学习方法.其中,本算法与梯度加噪的本地差分隐私联邦学习算法相比,模型的准确度在一般场景下提升9.6%,在非独立同分布与低质量数据场景下分别提高15.6%与16.2%.

关键词 联邦学习;差分隐私;函数机制;指数机制;权重散度

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2023.02178

Differentially Private Federated Learning with Functional Mechanism

CAO Shi-Xiang[1)] CHEN Chao-Meng[1)] TANG Peng[2),3)] SU Sen[1)]

[1)](*School of Computer Science(Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing* 100035)

[2)](*Key Lab of Cryptologic Technology and Information Security, Ministry of Education, Shandong University, Qingdao, Shandong* 266237)

[3)](*School of Cyber Science and Technology, Shandong University, Qingdao, Shandong* 266237)

Abstract In recent years, data privacy and security have been getting more and more attention from all walks of life. To make distributed machine learning better popularized and applied while protecting users' privacy, the academic community has proposed a federated learning framework, whose main idea is to build a machine learning model based on data sets distributed on multiple devices. The user machine does not exchange data sets with the server, but only shares model data. In this way, federated learning not only solves the data island problem, but also improves operation efficiency. Federated learning allows multiple user devices to share models only with the server rather than data, improving the efficiency of model training without disclosing the user's data privacy, so different organizations with data isolation can cooperate with machine learning.

收稿日期:2022-09-30;在线发布日期:2023-04-21.本课题得到国家自然科学基金青年基金(No.62002203)、山东省自然科学基金青年基金(No.ZR2020QF045)资助.**曹世翔**,硕士研究生,主要研究方向为联邦学习、差分隐私.E-mail:caoshixiang@bupt.edu.cn.**陈超梦**,博士研究生,主要研究方向为联邦学习、差分隐私.**唐 朋**,博士,副教授,主要研究方向为数据安全与隐私保护.**苏 森**,博士,教授,主要研究方向为智能服务、社交网络分析、数据隐私保护.

Federated learning based on differential privacy is a widely used privacy preserving federated learning framework, which protects the privacy of local data by introducing noise into the gradient before the user uploads the results. In this method, however, gradient accuracy loss problem caused by clipping and noise limits the further improvement of the training effect of differential privacy federated learning. The magnitude of the noise added to the gradient is strongly related to the gradient clipping threshold. The functional mechanism (FM) used for local data protection at the client can avoid gradient accuracy loss. It perturbs the objective function rather than the gradient result to meet the requirements of differential privacy, so the noise is not affected by the gradient threshold thus improving the actual effectiveness of the gradient. However, FM still has some limitations, such as being not suitable for high-dimensional scenes and only protecting data privacy. Still, it cannot save the privacy of the model. This paper focuses on solving the above problems. Firstly, to solve the gradient accuracy loss problem, this paper proposes an algorithm called Differentially Private Federated Learning with Functional Mechanism and presents a basic scheme to make the algorithm more suitable for high-dimensional data federated learning scenarios. In the basic scheme, the LDP-Federated Principal Component Analysis algorithm proposed by us enables all users to map local data to low dimensions in a unified manner. Moreover, we find that because there is no restriction or clipping on the gradient, the basic scheme has weight divergence problem in the scenario where the user data is not independent and identically distributed (non-IID) or there is low-quality data, which makes the training effect worse. The non-IID user data or the existence of low-quality data are inevitable in the application of federated learning. Due to the uneven data quality of mobile devices and the strong preference for users' habits and hobbies, low-quality and non IID data are also inevitable obstacles in the federated learning application scenario. Therefore, this paper also proposes an enhanced scheme in view of the shortcomings of the basic scheme and uses the ϵ-DP minimum divergence exponential sampling aggregation method to optimize the training effect in the above scenarios. Finally, we confirmed the effectiveness and superiority of our method by experiments. The experimental results show that the algorithm and schemes proposed in this paper are better than other existing privacy-preserving federated learning methods. Among them, compared with gradient-noising based federated learning with local differential privacy, our algorithm improves by 9.6% in general scenarios, 15.6% and 16.2% in non-IID and low-quality data scenarios, respectively.

Keywords federated learning; differential privacy; functional mechanism; exponential mechanism; weight divergence

1 引 言

近年来，为了使分布式机器学习[1]能够在保护用户隐私的前提下得到更好的普及与应用，学界提出了联邦学习(Federated Learning)[2]框架，其主要思想是基于分布在多个设备上的数据集构建机器学习模型. 在联邦学习训练中，用户机器不与服务器交换数据集，只共享模型数据. 因此，联邦学习便在解决数据孤岛问题的同时也降低了对单机设备算力的要求，从而提升了运算效率. 在联邦学习的基础上，隐私保护联邦学习框架随之被提出. 它要求服务器与本地用户在一定的隐私安全协议下完成联邦学习，进一步提高安全性.

目前，基于差分隐私(Differential Privacy，DP)或本地差分隐私(Local Differential Privacy，LDP)的联邦学习[3-4]是一种广泛应用的联邦学习框架. 它通过用户在上传训练梯度前对其引入噪声来保护本地数据的隐私，但目前对该方法的研究[5]中仍然存在着梯度精度丢失导致模型准确度下降的问题：

为了限制训练中对梯度所加噪声的敏感度(sensitivity)大小,需要保证梯度有界,因此通常要对梯度进行裁剪.裁剪阈值过大会导致所需噪声量过高;裁剪阈值过小会导致裁剪梯度造成的精度损失过高.无论裁剪阈值设置过大或过小,均会影响用户所上传梯度的真实有效性,使其无法达到足够的精度.在基于本地差分隐私的联邦学习[4]中,由于每个用户在每次迭代中都要对其梯度扰动以保护本地数据的隐私,噪声积累更为严重,所以梯度精度丢失问题则表现得更为明显.

一种被用于用户端本地数据保护的函数机制[6](Functional Mechanism,FM)可在一定程度上避免上述本地差分隐私联邦学习方法中的梯度精度丢失问题.它通过扰动目标函数而非梯度结果以满足差分隐私的要求,无论梯度结果的取值范围是多少,加入的噪声量是恒定的,因此其造成性能损失更小.而FM方法仍然存在一些限制,若将该机制直接用于联邦学习,会存在以下问题:

(1)在现有的基于FM的机器学习机制[7-8]中,需要引入的噪声量级会随输入数据的维度增加而提高.对数据维度的敏感限制了FM在联邦学习高维数据场景下的应用.

(2)FM未对梯度结果做任何限制与裁剪,在低质量数据[9]或非独立同分布(Non Independent Identically Distribution,non-IID)数据[10]场景下会引发权重发散的问题,并导致了模型准确度下降.

为解决上述问题,本文提出了基于函数机制的差分隐私联邦学习(**D**ifferentially Private **Fe**derated **L**earning with Functional **M**echanism,DELM)算法.据我们所知,这是第一个将FM用于联邦学习并且优化了训练性能的算法方案.首先,我们设计了DELM基础方案,此方案利用FM解决了梯度精度丢失的问题;其次,利用Split & Shuffle[11]拆分模型并打乱参数顺序,进一步"放大"[12]了隐私,并且避免了服务器直接获取用户的完整模型梯度数据;并基于主成分分析[13-14](Principle ComponentAnalysis,PCA)提出了LDP-联邦主成分分析(Locally Differential Private Federated Principal Component Analysis,LDP-FPCA),利用LDP-FPCA预处理数据来降低输入数据维度,从而在不影响模型聚合效果的前提下避免了FM在高维数据场景下引入过量的噪声.

我们发现,由于FM未对梯度结果做任何限制与裁剪,基础方案会产生权重发散的问题,并导致了模型准确度下降.此问题在低质量数据或non-IID数据场景中更为明显.为了继续优化基础方案存在的问题,我们提出了DELM增强方案,选取具有低发散度的梯度来更新全局模型,以降低模型权重发散度[10].然而,此方案容易将用户的数据质量与数据偏向性等隐私泄露给不可信的参与者[9].针对该隐患我们又提出了ϵ-DP-最低发散度指数抽样聚合方法加以预防,使用满足DP的指数机制[9,15-16]增加了选择用户时的不确定性,进一步保护了用户隐私.最终,我们使用DELM增强方案提高了模型在低质量或non-IID数据场景下的训练效果.

本工作的主要贡献如下:

(1)我们首次提出了一种基于函数机制的联邦学习算法,并设计了DELM基础方案,提出LDP-联邦主成分分析算法使FM适用于高维数据环境,解决了本地差分隐私联邦学习中梯度精度丢失问题,并优化了服务器模型聚合的效果,同时保证了用户的数据隐私.

(2)我们针对DELM基础方案在用户持有低质量数据或non-IID数据的场景下训练效果差的缺陷,提出了ϵ-DP-最低发散度指数抽样聚合方法和DELM增强方案,进一步提高了学习效果;同时,面对用户不可信的场景,此方案避免了数据质量与数据偏向性等隐私泄露给不可信用户的风险.

(3)最后,我们通过实验评估了所提出的算法和方案,验证了我们的方法的有效性和优越性.以模型准确度为指标,相比于梯度加噪的本地差分隐私联邦学习算法,在一般场景下基础方案提升了9.6%,non-IID与低质量数据场景下增强方案分别提升了15.6%与16.2%.而在用户可信的情况下,使用非抽样最低发散度聚合方法的增强方案相比基线的提升可以达到22.9%.

2 相关工作

本章介绍相关背景与研究工作,并分析这些工作存在的局限性.我们将基于以往的研究工作中遗留的问题,提出新的方案与算法以进行改进.

2.1 隐私保护联邦学习

在谷歌最初提出的联邦学习框架[17]中,客户端直接将裁剪后的模型参数权重上传至服务器,服务器对这些参数取平均并更新模型,这也被称为FedAvg[18]算法.FedAvg是一种简单的、不保证隐私的联邦学习模型聚合算法.该算法虽然没有让客户端直接共享数据,但却令客户端对服务器上传了

模型的参数或梯度．在服务器不可信[19]的情况下，客户端的模型参数可能会被泄露，容易经受对抗性攻击[20]；同时，在客户端与服务器通讯过程中也存在中间人攻击[23]的危险．

为了进一步保护数据隐私，防止客户端的模型参数泄露，隐私保护联邦学习[5]被提出．目前，学术界与工业界已存在一些成熟的开源联邦学习框架[24]，比如PySyft[27]、Tensorflow Federated（TFF）[28]、FATE[29]．它们在隐私保护方面所用的技术包括同态加密（Homomorphic encryption）[30]、安全多方计算（Secure multi-party computing）[31]、差分隐私（Differential Privacy）[32]等．其中，差分隐私依靠对数据引入噪声来保护数据隐私，无需额外的复杂运算，效率较高，但会降低数据的准确度．能够在分布式框架下保护每个参与者的数据隐私的差分隐私机制被定义为本地差分隐私，并常用作联邦学习的隐私保护机制．由于本地差分隐私需要对数据进行足量的扰动以实现每个用户所上传数据的隐私不被泄露，隐私保护程度与模型性能往往存在权衡与折衷，因此如何在达到隐私预算的要求下尽量提高模型性能是学界较为关注的问题．

差分隐私中用来衡量数据在查询时可能被泄露隐私的可能性的变量被称为隐私预算[32]，通常用 ϵ 表示．噪声量级与 ϵ 成反比，与敏感度 Δ 成正比．因此，通常人为设置梯度裁剪阈值来使 Δ 有界，而这却影响了梯度的精度．

为了降低噪声量级，有很多研究以合理利用隐私预算为目标提出了解决方案．文献[14]提出了DP-SGD与Moments Accountant，能够精准地跟踪隐私预算．这使得隐私预算消耗随训练轮数增加的速率放缓，以得到更多单次迭代可用的隐私预算．文献[33]对DP-SGD进一步优化，严密地跟踪训练中的隐私开销，并提出了一种动态隐私预算分配方法．文献[34]则通过差分隐私模型参数生成方法来保护模型隐私，在防御成员推理攻击的场景中模型性能优于DP-SGD．文献[35]将洗牌机模型（Shuffle Model）[36-37]用于联邦学习，利用洗牌机的匿名化加强了对用户梯度的隐私保护，这被称为“隐私放大”效应[12]．文献[11]也基于洗牌机模型提出了LDP-FL（LDP Federated Learning）算法，其中，使用“先拆分模型再洗牌”的Split & Shuffle方法进一步加强了隐私放大效应．但以往的研究都没有考虑由于敏感度 Δ 受裁剪阈值影响而使梯度精度丢失的问题，这导致模型无法得到足够的准确度．

除了上述问题外，联邦学习领域另一个受关注的方向为如何在non-IID数据或低数据质量的场景下提高模型精度．这是因为在联邦学习的实际应用场景中，不同用户的习惯和取向不尽相同，导致用户设备之间数据的标签分布也相差很大；同时，不同组织的数据收集与处理的能力通常也有较大差距，最终的可用数据质量参差不齐．以上问题都会影响全局模型的收敛情况．文献[10]将此场景下模型精度的降低归因于模型权重的发散——不同训练过程的权重差异，并提出了一种为客户端分配共享数据集的策略，以降低权重的发散．该方案是有效的，但在一般情况下，并不能假设这样一个公共数据．文献[38]提出的基于Top-K稀疏化的STC（Sparse Ternary Compression）方法控制客户端只上传 k 个最大值，并且压缩上下行的通信流量以提升通信效率．文献[39]提出了一种由FedAvg[18,40]改进来的FedProx方法，通过惩罚本地模型与全局模型的差异，使得本地更新不会太过远离当前的全局模型．但以上方案未考虑优化梯度精度丢失问题，其中Top-K稀疏化也存在着用户隐私泄露给其他不可信用户的风险．

2.2 函数机制

文献[6]首次提出了FM的概念，该机制不同于以往对梯度加噪的差分隐私机器学习方案，而是对机器学习的目标函数加噪．该机制需要将目标函数转化为关于模型权重的多项式形式，然后计算出多项式系数的敏感度并对多项式系数施加满足 ϵ-DP的拉普拉斯噪声（Laplace Noise）[41]．该文献指出：FM加噪的敏感度 Δ 仅与数据维度、数据集大小相关．但由于FM对目标函数所施加的噪声的量是和数据的维度成正相关的，因此在高维数据场景下，使用函数机制会对目标函数引入过多噪声，从而影响模型的准确度．另外，FM不会对训练时的梯度结果做任何限制或裁剪，这在一般场景下是有益的，因为这会最大限度地保留了梯度的真实有效性．但在一些特殊场景下，如用户持有大量低质量或non-IID数据时，不被限制的梯度会导致权重发散的问题，并导致模型的准确度下降．因此，与梯度加噪的本地差分隐私机制相比，FM在上述问题解决之前并不适合直接用于联邦学习．

目前的研究[7]已将FM用于各类隐私保护的机器学习任务中．文献[9]将FM用于协同的深度学习中，提出SecProbe方案，首次考虑了低数据质量参与者和不可信参与者的存在．其方案要求服务器设置辅助验证数据集，每轮聚合过程中都要验证局

部模型的精度并为其量化评分，最后利用DP-指数机制来聚合局部模型. 文献[7]将FM用于公平分类，考虑决策边界的公平性问题，并进一步提出基于高斯机制的松弛函数机制. 但上述研究没有考虑高维数据场景下FM会加入过量噪声，也没有针对non-IID数据场景提出优化方案.

表1对相关工作中的一些算法与本文所提出的DELM算法进行了总结与比较.

表1 算法特性比较

算法	完整梯度精度	高维数据	隐私放大技术	non-IID数据	低质量数据	防御威胁能力
STC[38]	×	√	—	√	×	不可信服务器
LDP-FL[11]	×	×	洗牌机	×	×	不可信服务器和不可信参与者
FedProx[39]	×	√	—	√	×	不可信服务器
SecProbe[9]	√	×	—	×	√	不可信服务器和不可信参与者
DELM	√	√	洗牌机	√	√	不可信服务器和不可信参与者

3 预备知识

3.1 威胁模型

在本算法框架中共存在三个实体方：用户U、洗牌机S、分析服务器A. 这三方实体均是不可信的，会造成如下的威胁：

用户是半诚实、好奇且共谋的，他们通过额外处理从服务器上下载的最终模型或交换信息，可以获得其他参与者的隐私.

洗牌机和分析服务器均是半诚实且好奇的，它们能够正确地遵循算法和协议，但可能会在联邦学习训练过程中通过额外处理用户上传的模型更新来尝试了解用户的隐私. 因此，针对其他用户、洗牌机以及分析服务器窃取隐私的情况，需要对用户实现ϵ-DP级别的隐私保护，以预防服务器获取用户隐私. 另外，需要使用用户加密与服务器解密的方案，使洗牌机不能获取到真实数据.

同时我们假设在所有通信中都使用了安全通道，防止中间人的窥探攻击.

3.2 数据声明

用户本地数据集D包含格式如$\tau_i=(X_i, y_i)$的带标签数据样本.

X为数据样本的输入变量，包含d维数据，可表示为$(x_1, x_2, \cdots, x_d)$. y为数据样本的输出变量.

本文假设输入数据均已规范化，即：$\|x_i\|_2 \leqslant 1$, $y_i \in [-1, 1]$. 表2总结了本文所使用符号及其含义.

3.3 函数机制

FM对目标函数引入噪声，对梯度精度影响更小，并也保证为每个用户的数据实现ϵ-DP级别的隐私保护. 因为FM不要求用户裁剪梯度结果，用户通过FM训练的梯度结果可以直接上传，这会提高模型的准确性. 因此本文使用函数机制作为每个用户的本地训练时的隐私保护方式.

表2 文中符号对照表

符号	代表含义
n	用户数目
m	神经网络层数
d	输入特征的维度
$\{W_j\}_{\{j=1\}}^{m}$	神经网络模型各层的参数
ω_t	第t次迭代时的全局模型参数
M_i	用户u_i的输入特征矩阵
g_i^j	用户u_i本地模型的第j层的梯度
dvg_i^j	用户u_i本地模型的第j层的发散度

我们考虑用户在本地训练一个具有三层全连接的多层感知器(Multi-Layer Perception，MLP)神经网络[42]，第一层为输入层(Input Layer)，第二层(隐藏层，Hidden Layer)以ReLU函数作为激励函数，第三层(输出层，Output Layer)以sigmoid函数作为输出函数. 因此本模型的输出z可以表示为

$$z=\left[1+\exp\left(-\mathrm{ReLU}\left(X^T W_1\right)W_2\right)\right]^{-1} \quad (1)$$

其中，W_1为连接输入层与隐藏层的权重矩阵，W_2为连接隐藏层与输出层的权重矩阵.

数据集将以随机梯度下降(Stochastic Gradient Descent，SGD)方式训练，设抽样的批(Batch)为B，随机抽取若干τ_i，样本个数为$|B|$. 代价函数使用交叉熵函数，因此目标函数可以表示为

$$f(B, W)=\sum_{\tau_i \in B} -y_i \log(z_i)=$$

$$\sum_{\tau_i \in B} -y_i \log\left[1+\exp\left(-\mathrm{ReLU}\left(X_i^T W_1\right)W_2\right)\right]^{-1} \quad (2)$$

方便起见，定义如下的函数代换：

$$g(z)=-y_i\log[1+\exp(-z)]^{-1};\quad h(\tau_i,W)=\mathrm{ReLU}(X_i^TW_1)W_2 \tag{3}$$

目标函数可以改写为如下形式：

$$f(B,W)=\sum_{\tau_i\in B}g(h(\tau_i,W)) \tag{4}$$

根据2.2节，由于FM需要将目标函数转换为关于权重W的多项式的形式，逻辑预测问题中需要用“泰勒展开”处理激活函数（如sigmoid等），因此对于输出层为sigmoid函数的MLP神经网络，需要对其目标函数$f(B,W)$泰勒展开，得到：

$$f(B,W)=\sum_{\tau_i\in B}\sum_{k=0}^{\infty}\left[\frac{g^{(k)}(\theta)}{k!}(h(\tau_i,W)-\theta)^k\right] \tag{5}$$

当对上式取前三项（设$k\in[0,2]$时），可以得到近似的目标函数$\bar{f}(B,W)$. 为方便分析，我们设置$\theta=0$. 可得到化简结果如下式所示：

$$\bar{f}(B,W)=\sum_{\tau_i\in B}\sum_{k=0}^{2}\left[\frac{g^{(k)}(\theta)}{k!}(h(\tau_i,W)-\theta)^k\right]=\sum_{\tau_i\in B}\left[g^{(0)}(0)+g^{(1)}(0)h(\tau_i,W)+\frac{g^{(2)}(0)}{2}h^2(\tau_i,W)\right]=\sum_{\tau_i\in B}y_i\left[\log 2-\frac{1}{2}h(\tau_i,W)+\frac{1}{8}h^2(\tau_i,W)\right] \tag{6}$$

$\bar{f}(B,W)$可表示为如下式所示的关于权重W的多项式的形式，其中Φ_j表示关于W且次数为j的单项式ϕ的集合，λ_ϕ表示某个单项式ϕ对应的系数.

$$\bar{f}(B,W)=\sum_{j=1}^{J}\sum_{\phi\in\Phi_j}\sum_{\tau_i\in B}\lambda_{\phi\tau_i}\phi(W);\quad \Phi_j=\left\{W_1^{c_1}W_2^{c_2}\cdots W_d^{c_d}\Big|\sum_{l=1}^{d}c_l=j\right\} \tag{7}$$

由于$\bar{f}(B,W)$的最高次项来自$h^2(\tau_i,W)$，因此W单项式的最高次数为4，即$J=4$.

假设存在B与B'两个临近数据库，计算在这两个数据库上的目标函数$\bar{f}(B,W)$与$\bar{f}(B',W)$的敏感度Δ如下式所示：

$$\Delta=\max\left(\left(\bar{f}(B,W)-\bar{f}(B',W)\right)_1\right)=\sum_{j=1}^{J}\sum_{\phi\in\Phi_j}\left(\lambda_{\phi\tau_i}-\lambda_{\phi\tau_i'}\right)_1\leqslant 2\max_{\tau_i}\sum_{j=1}^{J}\sum_{\phi\in\Phi_j}\left(\lambda_{\phi\tau_i}\right)_1 \tag{8}$$

根据已知的$\|x_i\|_2\leqslant 1$，$y_i\in[-1,1]$，可进一步放缩，如下式所示. 并可以得到结论：敏感度Δ与数据维度d的二次方相关.

$$\Delta\leqslant 2\max_{\tau=(x,y)}\left(\frac{y}{2}\sum_{j=1}^{d}x_j+\frac{y}{8}\sum_{j,l}x_jx_l\right)\leqslant 2\left(\frac{d}{2}+\frac{d^2}{8}\right)=d+\frac{d^2}{4} \tag{9}$$

使用Laplace机制对目标函数$\bar{f}(B,W)$中权重的系数进行扰动的公式如下式所示：

$$\lambda_\phi=\sum_{t_i\in D}\lambda_{\phi t_i}+\mathrm{Lap}\left(\frac{\Delta}{\varepsilon}\right) \tag{10}$$

使用上式扰动权重系数后得到的目标函数记为$\tilde{f}(B,W)$. 并且可以得到结论：当输入数据的维度放大（或缩减）为之前的p比例后，目标函数多项式系数的所加噪声量级也会随之放大（或缩减）为之前的p^2比例.

以上关于目标函数的Δ的计算过程是由原模型神经网络的层次结构所决定的. 对于不同结构的神经网络，Δ的计算结果并不相同. 考虑模型为更深层次的卷积神经网络AlexNet[43]的情况. 该网络具有8个隐藏层，包含5个卷积层和3个全连接层，并使用ReLU作为激活函数，引入Dropout层以避免过拟合. 该网络的输出层为softmax，并搭配交叉熵损失函数. 因此AlexNet的目标函数$f(B,W)$可以如下定义，其中c为输出类别数目，$g(z)$为交叉熵损失函数，$h(\tau_i,W)$为AlexNet隐藏层的映射输出函数.

$$g(z)=-\sum_{j=1}^{c}y_j\log(z_j);\quad f(B,W)=\sum_{\tau_i\in B}g(h(\tau_i,W)) \tag{11}$$

对$f(B,W)$泰勒展开，并取前三项，可以得到近似的目标函数$\bar{f}(\mathrm{B},\mathrm{W})$，如下式所示：

$$\bar{f}(B,W)=\sum_{\tau_i\in B}\sum_{j=1}^{c}\sum_{k=0}^{2}\left[\frac{g^{(k)}(\theta)}{k!}(h(\tau_i,W)-\theta)^k\right]=\sum_{\tau_i\in B}\sum_{j=1}^{c}\left[g^{(0)}(0)+g^{(1)}(0)h+\frac{g^{(2)}(0)}{2}h^2\right]=\sum_{\tau_i\in B}\left[h(\tau_i,W)+\frac{c}{2}h^2(\tau_i,W)\right] \tag{12}$$

同样的，计算在两个临近数据库上的目标函数$\bar{f}(B,W)$与$\bar{f}(B',W)$的敏感度Δ如下式所示. 敏感度Δ同时与数据维度d和分类数c相关.

$$\Delta \leqslant 2 \max_{\tau_i} \sum_{j=1}^{J} \sum_{\phi \in \Phi_j} \left(\lambda_{\phi \tau_i}\right)_1 \leqslant$$

$$2\left(d + \frac{c \cdot d^2}{2}\right) = 2d + c \cdot d^2 \tag{13}$$

3.4 “分割混淆”洗牌方法

本算法框架采用“分割混淆”(Split & Shuffle)[11]的洗牌方法,目的是切断每个用户的梯度数据中不同的神经网络层之间的关联,起到隐私放大效应,以加强对用户数据隐私的保护. 在此方法中,服务器端与用户端不会直接通讯,而是通过其中间设置的洗牌机间接交换数据. 洗牌机执行Split & Shuffle,可以将每个用户的梯度数据拆分并打乱. 为适应后续增强方案中ϵ-DP-最低发散度指数抽样聚合方法对梯度的网络层完整性要求,本文对原始的Split & Shuffle做了一定的改动:洗牌时将本地模型梯度分割为不同的(神经网络)层,而非完全分割. 修改后Split & Shuffle的具体执行过程如下:

在每一轮迭代中,用户将本地模型的梯度数据扰动后上传至洗牌机,洗牌机收齐所有n个用户上传的梯度后,将每个用户所上传的梯度分割为m段,每一段对应一个(神经网络)层,这样就共有$n \cdot m$个梯度段. 对于每一层来说都有n段,记为一组,这样共有m组,记为$G_1, G_2, \cdots, G_m$. 然后进行洗牌操作:分别从$G_1, G_2, \cdots, G_m$每个组中不放回地随机选取一个梯度段,组成第一个具有m段的新的本地模型梯度;接着再从$G_1, G_2, \cdots, G_m$中分别不放回地随机选取一个梯度段,组成第二个新的本地模型梯度. 重复以上过程,直至得到n个重组过的本地模型梯度. 洗牌机最终将这些洗牌后的模型梯度数据上传至服务器.

因此,经过本方法洗牌后的任意一个本地模型的梯度数据最多可能来自个m 不同的用户. 分析服务器难以还原出任何一个用户的完整的本地模型梯度,更难以通过模型参数来还原本地数据. 并且,在多次迭代中,服务器也不能通过历史迭代中的梯度数据来计算得出某一个用户的真实模型参数. 因此该方法可以达到m 倍的隐私放大效果. 关于该结论的理论推导过程详见附录.

4 DELM基础方案

本文提出一种基于函数机制的差分隐私联邦学习(DELM)算法方案. 本章主要介绍DELM基础方案的流程与关键技术.

4.1 方案流程

4.1.1 系统框架

如图1所示,初始时,分析服务器A存在一个初始化的模型. U中每个独立的用户u都持有一定数量的本地数据集. 每个用户与服务器共同执行LDP-联邦主成分分析算法,得到降维后的数据. 使用该方法的目的是在不影响模型聚合性能的前提下降低用户的输入数据维度,从而降低FM引入的噪声量级,后续章节将详细说明.

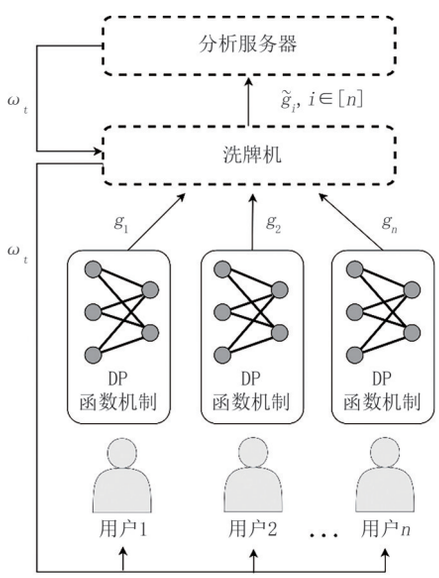

图1 DELM框架图

同步阶段,服务器A将本次迭代t的全局模型参数ω_t发送给洗牌机S,S再将其发送给用户集U,U中所有用户将本地模型同步为全局模型.

训练阶段,基于当前的本地模型,U中每个用户使用降维后的数据以及FM扰动后的目标函数进行本地模型训练.

上传阶段,用户$u_1, u_2, \cdots, u_n$分别将梯度结果$g_1, g_2, \cdots, g_n$加密上传至洗牌机S. S对用户上传的梯度数据进行Split & Shuffle[11]打乱处理,得到$\tilde{g}_1, \tilde{g}_2, \cdots, \tilde{g}_n$并将其传递给服务器$A$.

聚合阶段,服务器A使用密钥解密洗牌机S提交的用户梯度数据,对这些梯度进行加权聚合,并更新全局模型.

在当前迭代次数t未达到最大迭代次数T前,将循环依次执行以上四个阶段.

4.1.2 方案算法

本方案的算法流程如算法1所示,设原始数据

的维度为d，数据维度缩减比例为p. 在学习开始前，执行LDP-联邦主成分分析过程，各用户在此过程中不泄露隐私地计算得到维度为pd的降维数据，并将降维数据作为后续模型训练的输入数据. 之后，分析服务器A初始化模型与参数. 每一轮训练周期开始时，每个用户通过洗牌机获取分析服务器的全局模型，并更新为本地模型. 每轮迭代中，用户使用FM生成扰动后的损失函数，在本地模型的基础上运行SGD，并将梯度加密后上传至洗牌机. 当洗牌机收到所有用户上传的梯度数据后，执行Split & Shuffle，将每个用户的梯度数据拆分、混淆、打乱，再将这些结果发送给分析服务器. 分析服务器解密还原用户的梯度数据，计算出平均梯度，并以此更新全局模型.

算法1.　DELM基础方案算法.

输入：分析服务器A，洗牌机S，数据集D，数据维度缩减比例p，用户数目n，隐私预算ϵ_1, ϵ_2，学习速率η，最大迭代次数T.

输出：全局模型ω_T

1. 执行LDP-FPCA$(A, S, n, D, p, \epsilon_1)$，各用户将降维后的数据作为输入数据
2. A 初始化模型参数
3. FOR 每次迭代 $t\in[T]$DO
4. 　A 将全局模型 ω_t 发送给 S
5. 　FOR每个用户 u_i, $i\in[n]$DO
6. 　　从 S 处下载全局模型
7. 　　使用满足ϵ_2-DP的函数机制运行 SGD
8. 　　将本次训练得到的梯度 g_i 加密并上传至S
9. 　END FOR
10. 　S 对已上传的梯度执行 Split & Shuffle
11. 　S 将执行结果 $\tilde{g}_i$, $i\in[n]$ 发送给 A
12. 　A 进行解密，得到 $\tilde{g}_i$, $i\in[n]$
13. 　A 更新全局模型：$\omega_{t+1}=\omega_t-\frac{\eta}{n}\sum_{i=1}^{n}\tilde{g}_i$
14. END FOR
15. RETURN最终的全局模型 ω_T

4.2　LDP–联邦主成分分析

4.2.1　动机

3.3节得到结论：FM在联邦学习高维数据场景中会受维度增加的影响而引入更多的噪声. 因此需要设计一种数据降维的方案，来尽可能降低FM所需要引入的噪声量级.

主成分分析(PCA)[13]是常见的提取输入数据主要特征的方法，通常用于高维数据降维. PCA保留了数据的主要信息，在降低数据维度的同时也提高了特征的独立性，防止了过拟合. 一种简单的想法是让每个用户在本地进行PCA，并使用降维后的数据进行后续的模型训练. 但该方法忽略了用户间数据分布不同的问题，难以保证PCA映射方式的一致性. 这是因为PCA的输出是对数据以捕获最大方差为目标而映射出新的低维特征，不同分布的数据的映射方式并不相同. 每组用户数据经过PCA降维后，对应维度代表的信息已经不再一致. 这导致用户间本地模型参数的差异性增大，造成服务器聚合后的模型的准确度变差. 为解决上述问题，我们基于LDP-主成分分析(LDP-PCA)[44]的思想，提出了LDP-联邦主成分分析(LDP-FPCA). 此方法可以统一每个用户的数据映射方式，解决模型聚合性能差的问题. 尽管LDP-联邦主成分分析相比本地PCA对数据加了一定的噪声，但它获得了更好的数据映射一致性，提高了模型准确度. 在6.2节中的对比实验也证明：无论是训练速度还是模型准确度，LDP-联邦主成分分析比本地PCA以及LDP-PCA都要高.

4.2.2　算法流程

在此方法中，每个用户的本地数据集D_i的特征部分将被表示为矩阵$M_i\in\mathrm{R}^{|D_i|\times d}$，该矩阵的每一行即为一条数据的特征向量$X$. 首先，每个用户将计算出数据协方差矩阵$C_i=M_i{}^TM_i$，并以满足差分隐私的方式加入高斯噪声得到扰动协方差矩阵$\widetilde{C_i}$，之后将$\widetilde{C_i}$上传至服务器. 服务器聚合平均各个用户所上传的扰动协方差矩阵，得到$\widetilde{C}$. 服务器对$\widetilde{C}$进行特征值分解，得到特征向量$\tilde{V}$，再选取对应特征值最高的pd个特征向量得到$\tilde{V}$的子空间矩阵$\tilde{V}_{[pd]}=\left[\tilde{V}_1, \tilde{V}_2, \cdots, \tilde{V}_{pd}\right]$，最后将$\tilde{V}_{[pd]}$同步给客户端. 每个客户端使用$\tilde{V}_{[pd]}$计算本地降维数据为$M_{i[pd]}=M_i\cdot\tilde{V}_{pd}$. 此方法的具体算法流程如算法2所示：

算法2.　LDP-联邦主成分分析(LDP-FPCA).

输入：分析服务器A，洗牌机S，数据集D，数据维度d，数据维度缩减比例p，用户数目n，隐私预算ϵ_1.

输出：各个用户降维后的本地数据$M_{i[pd]}$, $i\in[n]$

1. FOR 每个用户 u_i, $i\in[n]$DO
2. 　将本地数据集 D_i 表示为矩阵 $M_i\in\mathrm{R}^{|D_i|\times d}$
3. 　计算得到 $C_i=M_i{}^TM_i$
4. 　生成矩阵 $Z\in\mathrm{R}^{d\times(d+1)}$，矩阵每个元素都满足：$z\sim N_d\left(0, \frac{1}{2\epsilon_1}\right)$

5. 计算得到 $\widetilde{C}_i=C_i+ZZ^T$, 并将 $\widetilde{C}_i$ 上传至 A
6. END FOR
7. A 计算得到 $\widetilde{C}=\frac{1}{n}\sum_{i=1}^{n}\widetilde{C}_i$
8. A 特征分解 $\widetilde{C}$,得到 $\widetilde{C}$ 的特征向量矩阵 $\tilde{V}$
9. A 截取 $\tilde{V}$ 的 top_{pd} 子空间矩阵 $\tilde{V}_{[pd]}$,并同步给用户
10. FOR 每个用户 u_i, $i\in[n]$ DO
11. $M_{i[pd]}=M_i\cdot\tilde{V}_{pd}$
12. END FOR
13. RETURN $M_{i[pd]}$, $i\in[n]$

5 DELM增强方案

5.1 动机

我们通过实验观察到:相比于用户数据集均匀分布的一般场景,基础方案在参与者持有较多non-IID数据或低质量数据的场景下表现不佳,模型性能表现较差. 经过分析发现:在上述场景中DELM基础方案模型准确率下降的主要原因在于:使用FM机制进行训练得到的梯度结果是无界的,这会造成某些与全局模型偏差较大的用户模型的梯度无法得到限制,影响全局模型的收敛. 训练过程中,数据质量较低或标签偏向性较强的用户模型的权重通常与其他用户的权重的差距较大,使得全局模型严重偏离最优模型.

因此我们定义梯度发散度用于衡量某次迭代中用户模型更新与全局模型更新之间的偏离程度,并提出了以达到“最低发散度”为目标的聚合方法,分析服务器通过计算每个用户上传的梯度的发散度,对模型的每一层神经网络以梯度发散度将各个用户排序,模型的每一层均得到具有最低发散度的k个用户的梯度,再求得平均值,并更新全局模型.

然而,在参与者不可信时使用最低发散度聚合机制可能会造成用户数据隐私泄露. 以一种极端情况举例:假设分析服务器筛选出发散度较低的神经网络层均来自于少数几个参与者,则其他参与者可以通过比较自己的参数和服务器发送的新参数来很容易地推断哪些参与者持有低质量的数据,或者偏向于持有某一类标签的non-IID数据. 以上恶意推断攻击会泄露用户个人习惯与偏好等隐私. 因此我们引入DP-指数机制(Differential Privacy Exponential Mechanism)[15-16]来预防该隐患,将最低发散度神经网络层的选择变成一个随机抽样过程,使每轮所选择的用户不再固定,而具有最低发散度的用户会具备更高可能性被选中.

至此,我们提出了ϵ-DP-最低发散度指数抽样聚合方法. 应用该方法的DELM增强方案可以在保护用户数据隐私的同时缓解non-IID数据问题及低质量数据问题.

5.2 方案流程

本方案的算法流程如算法2所示,增强方案算法与基础方案算法的差异在于分析服务器在模型梯度聚合时使用了ϵ-DP-最低发散度指数抽样聚合方法(从第13行至第17行). 该算法需要在开始时设置最低发散度目标个数k. 在分析服务器A完成对用户梯度数据的解密还原后,将计算每个用户的每个神经网络层梯度的发散度dvg_i^j, $i\in[n]$,$j\in[m]$. 然后执行ϵ-DP-指数机制,以$-dvg_i^j$为标准(因为我们希望更低发散度的网络层更有可能被选出),每个神经网络层抽样选出来自k个用户的梯度g_i^j, $i\in[k]$, $j\in[m]$. 然后对于每个神经网络层都计算出k个用户梯度的平均值. 最后将该梯度平均值应用到其对应的神经网络层的权重参数上,更新全局模型.

算法3. DELM增强方案算法.

输入:分析服务器 A,洗牌机 S,数据集 D,数据维度缩减比例 p,用户数目 n,隐私预算 ϵ_1, ϵ_2, ϵ_3,学习速率 η,最大迭代次数 T,最低发散度目标个数 k,神经网络层数 m

输出:全局模型 ω_T

1. 执行 LDP-FPCA$(A, S, n, D, p, \epsilon_1)$, 各用户将降维后的数据作为输入数据
2. A 初始化模型参数
3. FOR 每次迭代 $t\in[T]$ DO
4. A 将全局模型 ω_t 发送给 S
5. FOR 每个用户 u_i, $i\in[n]$ DO
6. 从 S 下载全局模型
7. 使用满足ϵ_2-DP的函数机制运行 SGD
8. 将本次训练得到的梯度 g_i 加密并上传至 S
9. END FOR
10. S 对上传的梯度执行 Split & Shuffle
11. S 将执行结果 $\tilde{g}_i$, $i\in[n]$ 发送给 A
12. A 进行解密,得到 $\tilde{g}_i$, $i\in[n]$
13. A 计算每个用户的每个神经网络层梯度的发散度 dvg_i^j, $i\in[n]$,$j\in[m]$
14. A 对$-dvg_i^j$, $\tilde{g}_i$, $i\in[n]$, $j\in[m]$ 按照 ϵ_3-DP-指数机制的概率设置,抽样得到g_i^j,$i\in[k]$, $j\in[m]$
15. FOR 模型每个神经网络层 $j\in[m]$ DO

16.　　A 更新全局模型：$\omega_{t+1}^{j}=\omega_{t}^{j}-\frac{\eta}{k}\sum_{i=1}^{k}g_{i}^{j}$

17.　END FOR

18. END FOR

19. RETURN 最终的全局模型 ω_T

我们在后续小节将详细解释ϵ-DP-最低发散度指数抽样聚合方法以及该方法在增强方案中的具体计算过程.

5.3　ϵ-DP-最低发散度指数抽样聚合方法

首先，我们需要定义分析服务器A计算神经网络中每一层梯度的发散度的方式. 设第t次迭代时当前的全局模型为ω_t，A接受到的第i组梯度为$\tilde{g}_i$（已经过洗牌机打乱，并非完全来自用户i）. size_j表示第j层网络的大小（参数个数）. $\tilde{g}_i^{(j)(l)}$表示$\tilde{g}_i$的第j层网络的第l个梯度，$\omega_t^{(j)(l)}$同理，表示第t次迭代时全局模型的第j层网络的第l个权重. 那么$\tilde{g}_i$的第j层网络的发散度可以定义为

$$dvg_i^j=\frac{1}{\text{size}_j}\sum_{l}^{\text{size}_j}\text{sign}\left(\omega_t^{(j)(l)},\omega_t^{(j)(l)}-\eta\tilde{g}_i^{(j)(l)}\right)$$
$$\text{sign}(e,e')=\begin{cases}1,\text{如果}e\text{和}e'\text{符号相反}\\0,\text{否则}\end{cases}\tag{14}$$

其含义为：统计$\tilde{g}_i^{(j)}$中可改变模型参数正负符号的梯度个数. 该计数越高，说明该层梯度偏离全局模型的程度越高，则具有更高的发散度. 为方便计算，值为0的模型参数认为符号为正.

根据ϵ-DP-指数机制，$\{\tilde{g}_i\}_{\{i=1\}}^{n}$每个神经网络层被选择的概率$P_i^j(i\in[n],j\in[m])$正比于以下指数形式的公式，这满足$\epsilon$-DP的要求：

$$P_i^j\propto\exp\left(-\frac{\epsilon}{2n\Delta dvg}dvg_i^j\right)\tag{15}$$

由于$dvg\in[0,1]$，对于在两个临近的梯度矩阵下求得的dvg与dvg'的差异Δ一定也满足$\Delta\in[0,1]$，所以$\Delta dvg=1$，可以得到：

$$P_i^j\propto\exp\left(-\frac{\epsilon}{2n}dvg_i^j\right)\tag{16}$$

将按照此概率分布，对每层神经网络筛选出来自k个用户的梯度$g_i^j,i\in[k],j\in[m]$，并求出每一层的平均梯度来更新全局模型.

对比原始的平均聚合方法，本聚合方法不可避免地造成了额外的计算开销. 但考虑到联邦学习方案本就需要服务器使用验证数据集预测评估模型的准确度，该过程会进行多次神经网络的前向运算；而发散度的计算与排序仅需对每个用户的模型参数做一次浮点减法与整数加法运算，并进行排序，其计算量相比模型评估过程要小得多，几乎是可以忽略的. 因此该处的计算开销是可以忍受的.

5.4　收敛性分析

本文通过收敛分析来证明在数据集D的数量足够大时，此方案的输出ω_T逼近于原目标函数$f(D,W)$的最优W^*[6].

为了能够使得原目标函数随$|D|$增大时可收敛，我们考虑有界函数$\frac{1}{|D|}f(D,W)$. 其可以用下式来表示（其中$\phi(W)$表示只关于权重矩阵W的任意单项多次式）：

$$\frac{1}{|D|}f(D,W)=\sum_{j=1}^{J}\sum_{\phi\in\Phi_j}\left(\frac{1}{|D|}\sum_{i=1}^{|D|}\lambda_{\phi\tau_i}\right)\phi(W)\tag{17}$$

因此模型权重的λ_ϕ系数在有界时，$\frac{1}{|D|}f(D,W)$可以逼近于一个关于权重矩阵W且以常数作为系数的多项式$h(W)$，即可以表示为$h(W)=\sum_{j=1}^{J}\sum_{\phi\in\Phi_j}c_\phi\phi(W)$. 该结论由以下推导可以证明：

$$\lim_{|D|\to\infty}\frac{1}{|D|}\sum_{i=1}^{|D|}\lambda_{\phi\tau_i}=\int_\tau\lambda_{\phi\tau_i}p(\tau)d\tau=E\left(\lambda_{\phi\tau_i}\right)=c_\phi\tag{18}$$

当按照DELM增强方案给λ_ϕ权重系数添加噪声后，经扰动的目标函数记为$\frac{1}{|D|}\tilde{f}(D,W)$. 其仍可以化简为$h(W)$. 该结论可由以下推导证明：

$$\lim_{|D|\to\infty}\frac{1}{|D|}\left(\sum_{i=1}^{|D|}\lambda_{\phi\tau_i}+\text{Lap}\left(\frac{\Delta}{\varepsilon}\right)\right)=\lim_{|D|\to\infty}\frac{1}{|D|}\sum_{i=1}^{|D|}\lambda_{\phi\tau_i}+$$
$$\lim_{|D|\to\infty}\frac{1}{|D|}\text{Lap}\left(\frac{\Delta}{\varepsilon}\right)=c_\phi+\lim_{|D|\to\infty}\text{Lap}\left(\frac{\Delta}{|D|\varepsilon}\right)\tag{19}$$

由4节7式可知Δ,ϵ均为有界实数，由此可知$\lim_{|D|\to\infty}\text{Lap}\left(\frac{\Delta}{|D|\,\varepsilon}\right)=0$. 并可得到如下结论：

$$\lim_{|D|\to\infty}\frac{1}{|D|}\tilde{f}(D,W)=h(W)\tag{20}$$

综上可知，本方案是可收敛的，当$|D|\to\infty$时，输出ω_T可逼近于原目标函数$f(D,W)$的最优解W^*.

5.5　隐私分析

由上文的分析可以得知，DELM增强方案共有三处隐私保护的步骤：用户在训练前与服务器进行

LDP-联邦主成分分析，假设其隐私预算为ϵ_1；用户在模型本地训练使用了满足差分隐私的FM，假设其隐私预算为ϵ_2；服务器在聚合局部模型时使用了满足差分隐私的DP-最低发散度指数抽样聚合方法，假设其隐私预算为ϵ_3.

由于ϵ_1与ϵ_2所对应的要保护隐私的内容都是用户本地数据的函数结果（前者为计算协方差矩阵的函数，后者为FM中计算目标函数对应多项式系数的函数结果），是对同一组数据的两次查询，因此ϵ_1与ϵ_2满足差分隐私的串行组合原理[45]. 所以这两个过程整体满足$\epsilon_1+\epsilon_2$的隐私预算.

$\epsilon_1+\epsilon_2$来自于为保护用户本地数据不泄露隐私给服务器而对目标函数系数及协方差矩阵等进行的扰动，而ϵ_3来自于为保护用户的完整的梯度参数不泄露给其他非可信用户而对最低发散度梯度抽样概率进行的扰动. 两者发生在不同的数据查询阶段，数据查询方也不一致，因此$\epsilon_1+\epsilon_2$与ϵ_3满足差分隐私的并行组合原理[45]. 该方案应满足$max(\epsilon_1+\epsilon_2,\ \epsilon_3)$-DP. 而根据3.4节的结论，洗牌机的隐私放大效应同时对ϵ_2与ϵ_3有效，因此方案最终应满足$max(\epsilon_1+\epsilon_2/m,\epsilon_3/m)$-DP.

综上，在本文中，我们设置$\epsilon_1+\epsilon_2/m=\epsilon_3/m=\epsilon$. 因此可以满足$\epsilon$—DP.

6 实验与结果分析

6.1 实验设置

6.1.1 数据集及参数设置

我们将在MNIST[46]、EMNIST[47]、CIFAR-10[48]和CIFAR-100[48]四个数据集上运行本文的DELM算法及方案，以验证其有效性.

MNIST与EMNIST：手写字符识别数据集，输入数据样本为28×28单通道图片，数据维度为$d=28\times28=784$. MNIST数据集共包含60 000项训练数据，10 000项测试数据，全部为手写数字图片；EMNIST共包含112 800项训练数据，18 800项测试数据，相比MNIST还扩充了手写大小写字母.

CIFAR-10与CIFAR-100：图像分类识别数据集，输入数据样本为32×32彩色图片，数据维度为$d=32\times32\times3=3072$. CIFAR-10与CIFAR-100均包含50 000项训练数据，10 000项测试数据. 两者不同之处在于CIFAR-10的图片标签有10个类别，而CIFAR-100有100个类别，并含有20个超类.

本文设置用户数目$n=100$，每个用户均分全部训练数据，分析服务器持有全部测试数据，用以验证模型. 以MNIST数据集为例：每个用户持有60 000/100 = 600项训练数据，分析服务器持有10 000项测试数据. 在MNIST与EMNIST的实验中，用户及服务器所使用的训练模型为3.3中的三层全连接神经网络；在CIFAR-10与CIFAR-100的实验中则使用AlexNet卷积神经网络作为模型.

6.1.2 对比方法

本文采用以下对比方法来验证DELM算法及方案的性能：

(1)FedAvg：不对数据加噪的联邦学习算法. 该算法中客户端将直接上传模型的原始梯度，服务器对客户端上传的梯度进行聚合平均，并更新全局模型. 将FedAvg作为本实验的基线，目的是将DELM模型与梯度全精度聚合的联邦学习模型进行对比.(2)LDP-FL：基于梯度加噪的本地差分隐私联邦学习算法，并使用洗牌机放大了隐私，目的是将DELM模型与梯度加噪的联邦学习模型进行对比.(3)改用本地PCA算法的DELM基础方案：将LDP-联邦主成分分析算法替换为用户本地执行PCA算法.(4)改用LDP-PCA[44]算法的DELM基础方案：将LDP-联邦主成分分析算法替换为LDP-PCA算法，该算法使用本地高斯机制对用户的数据进行降维. 使用以上两个改动方案的目的是验证LDP-联邦主成分分析算法的有效性.(5)FedProx[39]：通过限制本地更新不太远离当前的全局模型，提高在non-IID数据场景下性能表现的联邦学习算法，目的是验证DELM增强方案在non-IID数据场景下的有效性.(6)SecProbe[9]：通过服务器评估本地模型并选择性地抽样聚合，提高在低质量数据场景优化的联邦学习算法，目的是验证DELM增强方案在低质量数据场景下的有效性.(7)改用“非抽样”最低发散度聚合方法的DELM增强方案：取消ϵ-DP-最低发散度指数抽样聚合方法的抽样步骤，聚合时直接选取具有最低发散度的k个用户所上传的梯度. 改动DELM增强方案的目的是覆盖更多的应用场景，以验证方案的优越性.

6.1.3 对比场景

本文将在以下四个数据场景下进行实验，以验证DELM算法的两种方案在各自特定场景下的性能：

一般场景：将各个标签的数据集均分给用户，使

得每个用户的本地数据服从独立同分布)下. 在此场景下将进行三个对比实验：(1)在设置隐私预算 $\epsilon=1.0$、$\epsilon_1:\epsilon_2=1:2$ 且数据维度缩减比 $p=0.5$时，DELM基础方案、改用本地PCA算法的DELM基础方案、改用LDP-PCA算法的DELM基础方案、FedAvg与LDP-FL分别在训练过程中模型准确度变化，最大迭代次数 T 设置为250；(2)隐私预算比例设置为 $\epsilon_1:\epsilon_2=1:2$，依次对比在不同总预算 ϵ、不同数据维度缩减比 p 的条件下，DELM基础方案与LDP-FL算法在训练过程中模型准确度变化，其余条件与(1)保持一致.；(3)设置 $p=0.8$，依次对比在不同 $\epsilon_1:\epsilon_2$ 比例、不同 ϵ 的条件下，DELM基础方案收敛时的准确度，其余条件与(1)保持一致.

non-IID数据场景：为模拟用户持有non-IID的数据，本文采用 n-class non-IID划分方法[10]：每个用户只持有含其中 n 类标签的数据. 我们设置两个较为极端的non-IID数据场景实验——分别是1类non-IID和2类non-IID. 最低发散度目标个数 $k=50$. 实验将依次对比在以上两个non-IID场景下，DELM基础方案、增强方案、FedAvg、LDP-FL与FedProx算法在MNIST与CIFAR-10数据集下的训练效果差距.

低质量数据场景：为模拟用户持有低质量数据的场景，我们设置两个参数——低质量数据用户占比 lq_{user} 与低质量数据占比 lq_{data}，并根据该占比对用户数据进行加噪干扰，修改分类标签. 该场景下设置干扰数据占比高与低两组实验——分别是(1) $lq_{user}=lq_{data}=0.3$ 与(2) $lq_{user}=lq_{data}=0.5$. 最低发散度目标个数 $k=50$. 实验将依次对比在两个低质量数据场景下，DELM基础方案、增强方案、FedAvg、LDP-FL与SecProbe算法在EMNIST与CIFAR-100数据集下的训练效果差距.

用户可信的低质量数据场景：最后，考虑到现实应用中依然存在用户可信的应用场景，在此类场景中继续使用DELM增强方案将会浪费部分隐私预算. 即：ϵ-DP-最低发散度指数抽样聚合方法中的抽样步骤扰动了聚合结果，该步骤在用户可信场景下不必进行. 因此我们将在统一设置的低质量数据场景下重新对比DELM增强方案与改用非抽样最低发散度聚合方法的增强方案训练过程中模型准确度变化，并分别在 $k=50$，$k=20$ 两组实验设置下进行对比.

6.2　结果分析

6.2.1　一般场景实验结果分析

一般场景下，实验(1)的结果如图2所示. 在收敛时FedAvg的准确率是最高的，达到了0.973. 这是由于该算法没有隐私保护机制，未对数据进行扰动，从而在训练中保留了所有的信息，因此训练效果最好. DELM基础方案与LDP-FL在收敛时的准确度分别为0.962与0.927，这证明了在相同的隐私预算等设置下，DELM基础方案优于LDP-FL. 但是在 $t<100$ 时，DELM基础方案的模型准确度是弱于LDP-FL的，这是因为由于当前实验下的隐私预算充足，两者所引入的噪声量级均较低，相比于LDP-FL未进行数据降维并保留了原始数据信息，DELM基础方案中FM所带来的收益尚未明显体现. 因此，DELM基础方案的模型提升速度在训练前期会稍慢于LDP-FL. 而在后期DELM基础方案的准确度超过了LDP-FL，这一方面源于DELM基础方案更低的梯度精度损失，另一方面也因为数据降维后让模型具有更强的泛化性. 将DELM基础方案中数据降维方法改为LDP-PCA后，由于引入的噪声提高，数据丢失了部分信息，模型最终准确度为0.911，略低于数据未降维的LDP-FL；而将数据降维方法改为本地PCA后，模型准确率严重下降，最终仅达到0.871. 以上结果也证明了DELM基础方案中的LDP-联邦主成分分析算法的确因为统一了各用户的数据映射方式而优化了模型聚合效果.

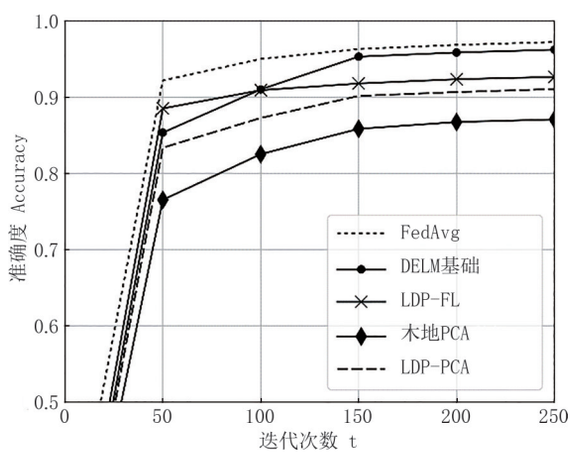

图2　一般场景下FedAvg、LDP－FL、DELM基础方案、基础方案改用本地PCA与LDP－PCA的训练收敛情况

实验(2)测试了在不同的 ϵ 与 p 下不同算法训练效果变化. 如图3(a)所示为 $\epsilon=1$ 时(MNIST数据集)的准确度变化实验结果. 在收敛时LDP-FL准

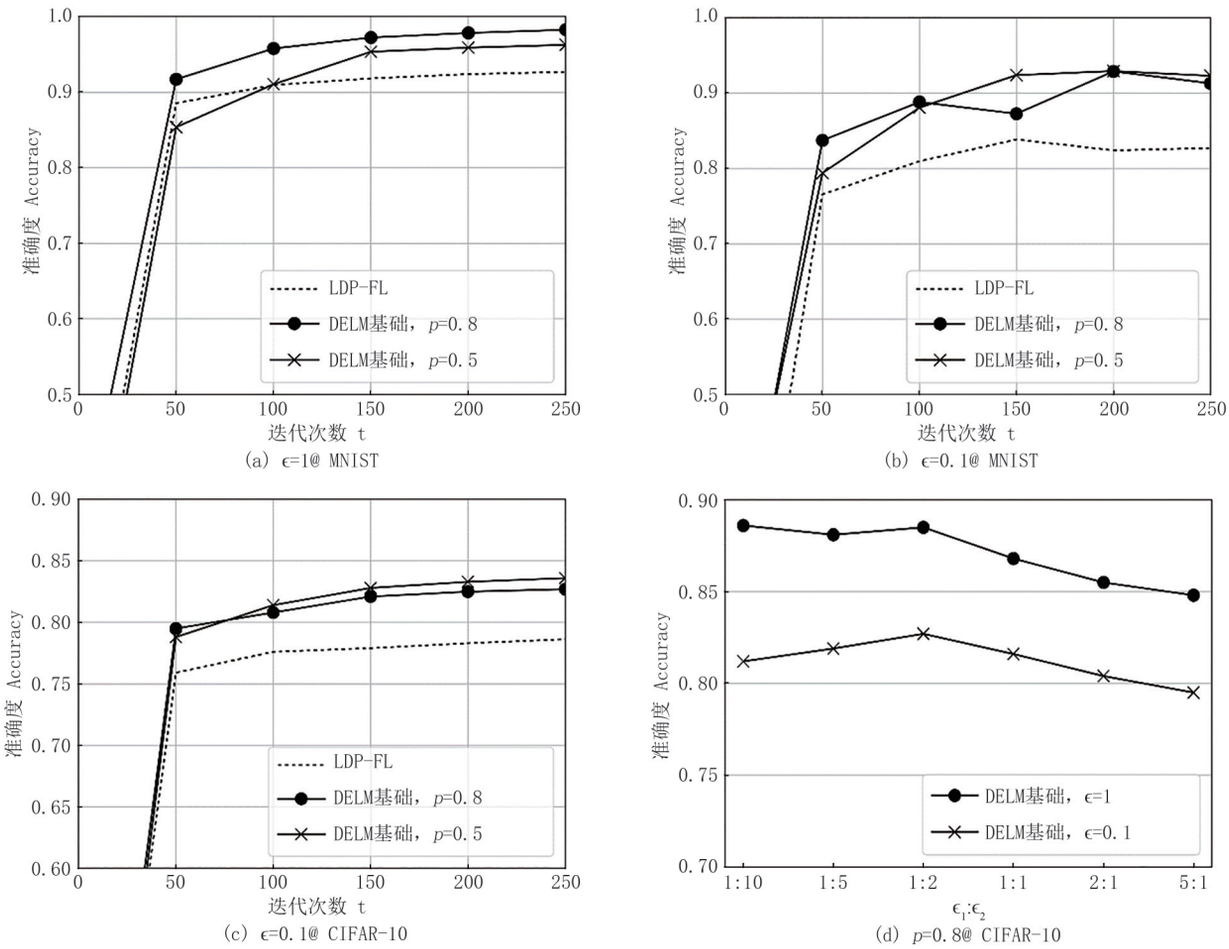

图3 一般场景下DELM基础方案与对比方法在不同参数设置下的准确度

确率达到0.927，DELM基础方案的准确率为0.981($p=0.8$)和0.962($p=0.5$). 可以看到DELM基础方案相比LDP-FL略有提升，最高提升幅度为5.4%. 由于此时隐私预算较高，提升不够明显，且p越高时效果越好. 原因可能为：较高的p可使本地数据保留较多的信息，由于此时隐私预算充足，这些信息大部分是准确的，因此模型性能较好. 如图3(b)所示为$\epsilon=0.1$时(MNIST数据集)的实验结果. 在收敛时LDP-FL准确率达到0.826，DELM基础方案准确率为0.913($p=0.8$)和0.922($p=0.5$). 可以看出当隐私预算不充足时梯度加噪过量对LDP-FL的影响较大，DELM基础方案相比LDP-FL的提升更明显，最高达到了9.6%. 且在p更小时的性能更有优势，这是因为当隐私预算较低时PCA降低数据维度后将弱化噪声的影响，同时也与输入数据维度更低而带来了更高的模型泛化性有关. 由于隐私预算较低，各算法方案受到更高的噪声的影响，准确率存在着不同程度的波动，甚至出现了训练后期准确率倒退的问题. 但比较两组数据的结果可以发现，DELM基础方案受ϵ变化的影响更小，相比其他算法更能适应低隐私预算的场景，可验证其优化了梯度精度丢失问题. 如图3(c)所示为$\epsilon=0.1$时(CIFAR-10数据集)的实验结果. 在收敛时LDP-FL准确率达到0.786，DELM基础方案准确率为0.827($p=0.8$)和0.836($p=0.5$)，同样可以体现DELM基础方案的性能更优.

实验(3)的结果如图3(d)所示. 可以看出：在保证ϵ_1充足的前提下，适当地提升ϵ_2的占比一般是对训练效果有益的. 原因在于：根据5.5总结的等式$\epsilon_1+\epsilon_2/m=\epsilon_3/m=\epsilon$，当$\epsilon$确定后，$\epsilon_3$的取值也已确定，而降低$\epsilon_1$将会换来$m$倍$\epsilon_2$的提升，让函数机制中的噪声量更小，从而为训练带来收益. 因此，在$\epsilon=1$设置下，$\epsilon_1:\epsilon_2=1:10$时模型的准确度达到最高，为0.886；而在$\epsilon=0.1$设置下，隐私预算相对不足，$\epsilon_1:\epsilon_2=1:2$时准确度达到了最高，为0.827，此时若再降低该比例则会引起LDP-联邦主成分分析算法可用的隐私预算过低，进而影响模型训练效果.

6.2.2　non-IID与低质量数据场景实验结果分析

在两个non-IID数据场景下我们测试了不同算法的训练性能. 如表3所示为不同算法在模型收敛时的结果对比. DELM增强方案在两个场景下的训练效果均强于其他算法. 在2类non-IID场景下，DELM增强方案相比FedProx分别提升3.1%@MNIST与3.7%@CIFAR-10，相比基础方案分别提升5.6%@MNIST与5.7%@CIFAR-10. 在1类non-IID场景下，更极端的数据分布使得模型准确度整体下降，而DELM增强方案的性能下降幅度相比其他算法更小，这也导致各算法间的差距更为明显. DELM增强方案相比LDP-FL分别提升15.6%@MNIST与12.3%@CIFAR-10，相比FedProx分别提升8.6%@MNIST与9.9%@CIFAR-10. 这说明最低发散度的抽样聚合方法一定程度上降低了non-IID数据引发的权重发散影响. 在大多数结果中DELM基础方案的准确度要低于FedAvg而高于LDP-FL，可能的原因是：DELM基础方案相比于LDP-FL提高了梯度的精度，但精度仍低于不加噪的FedAvg.

表3　non-IID数据场景下各算法模型收敛性能对比

算法	数据集	收敛准确度(1类/2类)	迭代次数	训练时间
FedAvg	MNIST	0.792/0.874	200	29 m
	CIFAR-10	0.723/0.855	200	1 h 36 m
LDP-FL	MNIST	0.756/0.847	100	35 m
	CIFAR-10	0.719/0.840	150	1 h 50 m
FedProx	MNIST	0.826/0.896	250	2 h 03 m
	CIFAR-10	0.743/0.852	300	4 h 12 m
DELM基础方案	MNIST	0.829/0.871	250	1 h 36 m
	CIFAR-10	0.722/0.832	250	3 h 07 m
DELM增强方案	MNIST	0.912/0.927	250	1 h 40 m
	CIFAR-10	0.842/0.889	250	3 h 18 m

如表4所示为低质量数据场景下不同算法方案在模型收敛时的结果对比. 由于数据的输出类别增多，模型拟合难度增大，该场景下各模型的准确率整体偏低. 在低干扰场景下，DELM增强方案相比LDP-FL分别提升12.5%@EMNIST与10.8%@CIFAR-100，相比DELM基础方案分别提升8.0%@EMNIST与10.3%@CIFAR-100. 当切换到高干扰场景后，DELM增强方案的模型精度的降低程度仍然要低于其他算法. DELM增强方案相比LDP-FL分别提升16.2%@EMNIST与12.7%@CIFAR-100，相比SecProbe分别提升5.1%@EMNIST与5.2%@CIFAR-100. 这说明最低发散度的抽样聚合方法在一定程度上降低了低质量数据引发的权重发散影响.

表4　低质量数据场景下各算法模型收敛性能对比

算法	数据集	收敛准确度(干扰高/低)	迭代次数	训练时间
FedAvg	EMNIST	0.670/0.716	200	30 m
	CIFAR-100	0.411/0.458	200	1 h 35 m
LDP-FL	EMNIST	0.603/0.657	100	38 m
	CIFAR-100	0.378/0.419	150	1 h 53 m
SecProbe	EMNIST	0.714/0.756	300	2 h 21 m
	CIFAR-100	0.453/0.497	300	4 h 47 m
DELM基础方案	EMNIST	0.658/0.702	250	1 h 33 m
	CIFAR-100	0.402/0.424	250	3 h 10 m
DELM增强方案	EMNIST	0.765/0.782	250	1 h 37 m
	CIFAR-100	0.505/0.527	250	3 h 16 m

6.2.3　用户可信的低质量数据场景实验结果分析

如图4所示为非抽样最低发散度聚合方法的增强方案在低质量数据条件下的训练效果与其他算法的对比. 非抽样的增强方案的准确度分别为0.954($k=50$)和0.922($k=20$)，可以看到使用非抽样的增强方案相比LDP-FL的提升最高可以达到22.9%，且相比原本的增强方案提升分别为3.3%和2.9%. 说明本方案算法在用户可信的条件下可以得到进一步的提升；另一方面，原本的增强方案在用户不可信条件下性能的减弱的程度也是可以忍受的. 并且可以看到，$k=50$时两个方案的训练效果高于$k=20$时的效果. 这是因为实验中携带正确数据的用户数为50，当k接近该数目时会使算法更能充分利用正确的用户数据.

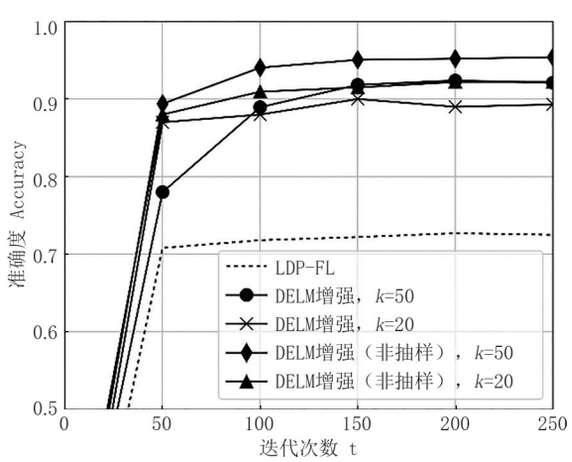

图4　非抽样增强方案的训练效果@MINST

7 结束语

本文分析了差分隐私与联邦学习的相关算法、机制等目前的研究现状以及一些瓶颈，讨论了目前本地差分隐私联邦学习框架存在的裁剪与加噪引起的梯度精度丢失问题. 同时，本文分析了函数机制对于解决联邦学习以上问题的一些潜力，也讨论了函数机制中噪声对数据维度敏感的问题. 基于此，我们提出了基于函数机制的差分隐私联邦学习算法，并设计了DELM基础方案，将函数机制用作客户端保护数据隐私的手段，缩减了噪声的量级，提高了模型准确度，并利用LDP-联邦主成分分析算法弱化了数据维度对噪声量级的影响，同时该方案使用洗牌机技术起到了隐私放大效应. 进一步，我们提出了ϵ-DP-最低发散度指数抽样聚合方法与DELM增强方案，避免了在non-IID或低质量数据场景下的模型权重发散，同时避免用户隐私泄漏给其他不可信用户. 最后，本文用实验验证了我们提出的算法，并证明了我们的方法的有效性与优越性.

本文提出的DELM算法与方案还存在一定的局限. LDP-联邦主成分分析算法要求每个用户上传其本地数据的协方差矩阵，尽管在整个训练过程只发生一次上传，但这仍然增加了通信的带宽与时延开销. 在未来的工作中，我们将考虑使用一些矩阵压缩算法来处理用户上传数据的过程，从而减少通信开销. 由于协方差矩阵以及用户数据的特殊性，可重点研究对称矩阵与稀疏矩阵的压缩存储优化方案，并最终实现一个适合联邦学习场景并且通讯高效的方案.

致 谢 感谢对本文提出宝贵建议的老师、同学以及所有评审专家！

参 考 文 献

[1] Kraska T, Talwalkar A, Duchi J C, et al. MLbase: A Distributed Machine-learning System//Sixth Biennial Conference on Innovative Data Systems Research.Asilomar, USA,2013, 1: 2-1

[2] Bonawitz K, Eichner H, Grieskamp W, et al. Towards federated learning at scale: System design//Proceedings of Machine Learning and Systems. Stanford, USA, 2019, 1: 374-388

[3] Cormode G, Jha S, Kulkarni T, et al. Privacy at scale: Local differential privacy in practice//Proceedings of the 2018 International Conference on Management of Data. Houston, USA, 2018: 1655-1658

[4] Truex S, Liu L, Chow K H, et al. LDP-Fed: Federated learning with local differential privacy//Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking. Heraklion, Greece, 2020: 61-66

[5] Kairouz P, McMahan H B, Avent B, et al. Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 2021, 14(1 - 2): 1-210

[6] Zhang J, Zhang Z, Xiao X, et al. Functional Mechanism: Regression Analysis under Differential Privacy. Proceedings of the VLDB Endowment, 2012, 5(11):1364-1375

[7] Ding J, Zhang X, Li X, et al. Differentially private and fair classification via calibrated functional mechanism//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(01): 622-629

[8] Adesuyi T A, Kim B M. A layer-wise perturbation based privacy preserving deep neural networks//International Conference on Artificial Intelligence in Information and Communication.Okinawa, Japan, 2019: 389-394

[9] Zhao L, Wang Q, Zou Q, et al. Privacy-preserving collaborative deep learning with unreliable participants. IEEE Transactions on Information Forensics and Security, 2019, 15: 1486-1500

[10] Zhao Y, Li M, Lai L, et al. Federated learning with non-iid data. arXiv preprint arXiv:1806. 00582, 2018

[11] Sun L, Qian J, Chen X. LDP-FL: Practical Private Aggregation in Federated Learning with Local Differential Privacy//The 32nd International Joint Conference on Artificial Intelligence. Montreal, Canada, 2021: 1571-1578

[12] Balle B, Bell J, Gascón A, et al. The Privacy Blanket of the Shuffle Model//39th Annual International Cryptology Conference. Santa Barbara, USA, 2019: 638-667

[13] Abdi H, Williams L J. Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2010, 2(4): 433-459

[14] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy//Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria, 2016: 308-318

[15] Awan J, Kenney A, Reimherr M, et al. Benefits and Pitfalls of the Exponential Mechanism with Applications to Hilbert Spaces and Functional PCA//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA 2019: 374-384

[16] Dong J, Durfee D, Rogers R. Optimal differential privacy composition for exponential mechanisms//Proceedings of the 37th International Conference on Machine Learning. 2020: 2597-2606

[17] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2017: 1273-1282

[18] Konečný J, McMahan H B, Yu F X, et al. Federated learning: Strategies for improving communication efficiency. arXiv

preprint arXiv:1610.05492, 2016
[19] Lyu L, Yu H, Yang Q. Threats to federated learning: A survey. arXiv preprint arXiv:2003.02133, 2020
[20] Bhagoji A N, Chakraborty S, Mittal P, et al. Analyzing federated learning through an adversarial lens///Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA 2019: 634-643
[21] Qiu S, Liu Q, Zhou S, et al. Review of artificial intelligence adversarial attack and defense technologies. Applied Sciences, 2019, 9(5): 909
[22] Phan H, Thai M T, Hu H, et al. Scalable Differential Privacy with Certified Robustness in Adversarial Learning//Proceedings of the 37th International Conference on Machine Learning. 2020:7683-7694
[23] Conti M, Dragoni N, Lesyk V. A survey of man in the middle attacks. IEEE Communications Surveys & Tutorials, 2016, 18(3): 2027-2051
[24] https://github.com/OpenMined/PySyft
[25] https://tensorflow.google.cn/federated/federated_learning?hl=zh-CN
[26] https://github.com/FederatedAI/FATE
[27] Ziller A, Trask A, Lopardo A, et al. Pysyft: A library for easy federated learning. Federated Learning Systems: Towards Next-Generation AI, 2021: 111-139
[28] Sun Z, Kairouz P, Suresh A T, et al. Can you really backdoor federated learning? .arXiv preprint arXiv:1911.07963, 2019
[29] Yang Q, Liu Y, Cheng Y, et al. Federated learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2019, 13(3): 1-207
[30] Fontaine C, Galand F. A survey of homomorphic encryption for nonspecialists. EURASIP Journal on Information Security, 2007: 1-10
[31] Du W, Atallah M J. Secure multi-party computation problems and their applications: a review and open problems// Proceedings of the 2001 workshop on New security paradigms. New Mexico, USA,2001: 13-22
[32] Dwork C. Differential privacy: A survey of results// International conference on theory and applications of models of computation. Xi'an, China, 2008: 1-19
[33] Yu L, Liu L, Pu C, et al. Differentially private model publishing for deep learning//IEEE symposium on security and privacy (SP). San Francisco, USA 2019: 332-349
[34] Mao Y, Hong W, Zhu B, et al. Secure Deep Neural Network Models Publishing Against Membership Inference Attacks Via Training Task Parallelism. IEEE Transactions on Parallel and Distributed Systems, 2021, 33(11): 3079-3091
[35] Liu R, Cao Y, Chen H, et al. FLAME: Differentially Private Federated Learning in the Shuffle Model//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(10): 8688-8696
[36] Girgis A, Data D, Diggavi S, et al. Shuffled model of differential privacy in federated learning//International Conference on Artificial Intelligence and Statistics. 2021: 2521-2529
[37] Girgis A M, Data D, Diggavi S, et al. Shuffled model of federated learning: Privacy, accuracy and communication trade-offs. IEEE Journal on Selected Areas in Information Theory, 2021, 2(1): 464-478
[38] Sattler F, Wiedemann S, Müller K R, et al. Robust and communication-efficient federated learning from non-iid data. IEEE transactions on neural networks and learning systems, 2019, 31(9): 3400-3413
[39] Li T, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks//Proceedings of Machine Learning and Systems. Austin, USA, 2020, 2: 429-450
[40] Li X, Huang K, Yang W, et al. On the convergence of fedavg on non-iid data. arXiv preprint arXiv:1907.02189, 2019
[41] Sarathy R, Muralidhar K. Evaluating Laplace noise addition to satisfy differential privacy for numeric data. Transactions on Data Privacy, 2011, 4(1): 1-17
[42] Gardner M W, Dorling S R. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment, 1998, 32(14-15): 2627-2636
[43] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84-90
[44] Wang D, Xu J. Principal component analysis in the local differential privacy model. Theoretical Computer Science, 2020, 809: 296-312
[45] Kairouz P, Oh S, Viswanath P. The composition theorem for differential privacy//International conference on machine learning. Lille, France, 2015: 1376-1385
[46] LeCunY.. (1998). The MNIST Database of Handwritten Digits. [Online].Available: http://yann.lecun.com/exdb/mnist/
[47] Cohen G, Afshar S, Tapson J, et al. EMNIST: Extending MNIST to handwritten letters//International Joint Conference on Neural Networks. Anchorage, USA, 2017: 2921-2926
[48] KrizhevskyA., NairV., and HintonG..CIFAR-10 and CIFAR-100 datasets.Available: http://www.cs.toronto.edu/~kriz/cifar.html
[49] Balle B, Barthe G, Gaboardi M. Privacy amplification by subsampling: Tight analyses via couplings and divergences// Advances in Neural Information Processing Systems. Montréal, Canada, 2018: 6280-6290

附录. 本文Split & Shuffle方法的隐私放大系数的证明

下面，将对本文Split & Shuffle方法的隐私放大系数等于m这一结论进行理论推导证明：

定义. 假设一个本地模型的梯度数据为D，D由各网络层的梯度$\{d_i, i\in[m]\}$组成. D的相邻数据集记为$D'=\{d_i', i\in[m]\}$. D和D'仅有一个梯度层不同，假设不同的梯度层为(d_k, d_k'). 存在一个不使用洗牌机方法的随机化机制M，M对D和D'满足ϵ-DP. 接着，将本文Split & Shuffle方法记为$\mathcal{A}$，将复合函数$\mathcal{A}\circ M$记为$\mathcal{A}_M$. 使用$\mathcal{A}_M$处理数据D，并将

$\mathcal{A}_M(D)$生成的序列记为$\{z_i, i\in[m]\}$. 实际上，对于每个$z_i, i\in[m]$均对应一个概率分布函数$\mathcal{A}_M^{(i)}(d_i)$. 该函数表示每一个梯度层的数据在经过Split & Shuffle的一系列处理后的最终输出的概率分布，从而分析和量化隐私泄露程度. 下面要证明的结论为：$\mathcal{A}_M(D)$满足$\frac{\epsilon}{m}-\mathrm{DP}$.

证明. 假设由d_k计算并乱序洗牌后得到的结果为z_k. 由于z_k对结果分布的影响与其他$z_i(i\neq \mathrm{k})$不同，因此要分别进行讨论. 洗牌后，对于服务器来说，每个梯度层$z_i, i\in[m]$为z_k的概率$\Pr[z_i=z_k]=1/m$. 因此有以下等式：

$$\mathcal{A}_M^{(i)}(d_i)=\frac{1}{m}\mathcal{A}(M(d_i))\Big|_{z_i=z_k}+\left(1-\frac{1}{m}\right)\mathcal{A}(M(d_i))\Big|_{z_i\neq z_k} \tag{21}$$

同理，对于数据D'，满足以下等式：

$$\mathcal{A}_M^{(i)}(d_i')=\frac{1}{m}\mathcal{A}(M(d_i'))\Big|_{z_i=z_k}+\left(1-\frac{1}{m}\right)\mathcal{A}(M(d_i'))\Big|_{z_i\neq z_k} \tag{22}$$

文献[49]中对于以上形式的概率分布输出存在如下结论：

$$\begin{cases} Y=(1-\gamma)X_0+\gamma X_1, X_0与X_1具有\epsilon-不可区分度 \\ Y'=(1-\gamma)X_0'+\gamma X_1', X_0'与X_1'具有\epsilon-不可区分度 \end{cases}$$

$$\Downarrow$$

$$Y与Y'具有\log\left(1+\gamma\left(e^{\epsilon}-1\right)\right)-不可区分度 \tag{23}$$

由于$\mathcal{A}$是对机制M的扰动结果进行洗牌混淆的操作，所以$\mathcal{A}(M(d_i))$具有不低于$M(d_i)$的隐私保护水平，$\mathcal{A}(M(d_i))$的两个条件项一定具有ϵ-不可区分度；$\mathcal{A}(M(d_i'))$同理. 因此(21)、(22)式满足(23)式的结论，可得出$\mathcal{A}_M^{(i)}(d_i)$与$\mathcal{A}_M^{(i)}(d_i')$具有$\log\left(1+\frac{1}{m}(e^{\epsilon}-1)\right)$-不可区分度，进而得出推论：机制$\mathcal{A}_M^{(i)}(d_i)$满足$\log\left(1+\frac{1}{m}(e^{\epsilon}-1)\right)$-DP.

隐私放大系数推导和证明中常常采用近似方法. 这里进行泰勒级数展开，可以得到以下结果：

$$\lim_{\epsilon\to 0}\log\left(1+\frac{1}{m}(e^{\epsilon}-1)\right)=\lim_{\epsilon\to 0}\log\left(1+\frac{1}{m}\epsilon\right)=\frac{\epsilon}{m} \tag{24}$$

说明在ϵ趋近于0时，可近似得出$\mathcal{A}_M$对梯度层d_i, d_i'满足$\frac{\epsilon}{m}-\mathrm{DP}$.

而D与D'分别为若干个不相交梯度层的组合，且只有一组梯度层有差异，因此$\mathcal{A}_M(D)$的隐私水平ϵ_D可以满足以下并行组合定理：

$$\epsilon_D=\max_i \epsilon_{d_i}=\frac{\epsilon}{m} \tag{25}$$

说明在ϵ趋近于0时，可得到本文Split & Shuffle方法的隐私放大系数为m的结论.

CAO Shi-Xiang, M. S. candidate. His research interests include federated learning and differential privacy.

CHEN Chao-Meng, Ph. D. candidate. His research interests include federated learning and differential privacy.

TANG Peng, Ph. D., associate professor. His research interests include data security and privacy protection.

SU Sen, Ph. D., professor. His research interests include intelligent services, social network analysis and data privacy protection.

Background

Federated learning allows different organizations to build machine learning models on datasets distributed across multiple devices, so that different organizations can jointly use their originally isolated data by sharing model parameters. At present, federated learning has been widely used in industry, because compared with traditional machine learning, federated learning can not only prevent data from being directly shared to third parties, but also reduce the computing power requirements of stand-alone devices to a certain extent. On the basis of federated learning, privacy preserving federal learning was proposed. Compared with traditional federated learning, privacy preserving federated learning can further prevent model parameter leakage, prevent man-in-the-middle attacks, malicious collection of client

model parameters by server and adversarial attacks, thereby protecting data and model privacy.

At present, there are popular open source federated learning frameworks including PySyft, Tensorflow Federated (TFF) and FATE. The privacy protection technologies they use mainly include Homomorphic Encryption (HE), Secure Multi-party Computing (MPC), Differential Privacy (DP). Among them, HE and MPC can guarantee the integrity of data, but the efficiency is low due to complex encryption and decryption operations. Since DP needs to introduce noise to the data, it will affect the integrity of the data compared with the former, but its operation efficiency is very high without additional complex operations. This paper focuses on federated learning based on differential privacy, and optimizes the noise introduced by DP to reduce its disturbance to the data, thereby improving the performance of DP in federated learning.

Existing research on differentially private federated learning focuses on the scheme of adding noise to client-side model parameters or gradients, and this method introducing quantitative noise to the user's gradient, and the magnitude of noise is often controlled by the gradient clipping threshold, which will affect the real validity of the gradient and the accuracy of the model. Previous studies have only started from the perspective of improving the privacy budget available for a single iteration to alleviate the over noise problem. There is no consideration of how to reduce the sensitivity to alleviate gradient accuracy loss problem. Therefore, this paper proposes a differential privacy federated learning algorithm based on Functional Mechanism (FM), which replaces the traditional method of perturbing gradient by perturbing the objective function, and solves gradient accuracy loss problem. Then, this paper proposes a basic scheme to solve the problem that FM is sensitive to the dimension of the input data. Finally, based on the basic scheme, this paper proposes the -DP-minimum divergence exponential sampling aggregation method and enhanced scheme to improve the model training effect in the two application scenarios of non-IID and low-quality data. After experimental verification, the algorithms and schemes proposed in this paper are effective and superior, and have significant performance improvement compared with the existing solutions, especially in non-IID and low-quality data scenarios.