面向非独立同分布数据的联邦学习数据增强方案

汤凌韬[1]，王迪[1]，刘盛云[2]

（1. 数学工程与先进计算国家重点实验室，江苏 无锡 214125；2. 上海交通大学网络空间安全学院，上海 200240）

摘 要：为了解决联邦学习节点间数据非独立同分布（non-IID）导致的模型精度不理想的问题，提出一种隐私保护的数据增强方案。首先，提出了面向联邦学习的数据增强框架，参与节点在本地生成虚拟样本并在节点间共享，有效缓解了训练过程中数据分布差异导致的模型偏移问题。其次，基于生成式对抗网络和差分隐私技术，设计了隐私保护的样本生成算法，在保证原数据隐私的前提下生成可用的虚拟样本。最后，提出了隐私保护的标签选取算法，保证虚拟样本的标签同样满足差分隐私。仿真结果表明，在多种 non-IID 数据划分策略下，所提方案均能有效提高模型精度并加快模型收敛，与基准方法相比，所提方案在极端 non-IID 场景下能取得 25%以上的精度提升。

关键词：联邦学习；非独立同分布；生成式对抗网络；差分隐私；数据增强

中图分类号：TP301

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2023007

Data augmentation scheme for federated learning with non-IID data

TANG Lingtao[1], WANG Di[1], LIU Shengyun[2]

1. State Key Laboratory of Mathematical Engineering and Advanced Computing, Wuxi 214125, China
2. School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract: To solve the problem that the model accuracy remains low when the data are not independent and identically distributed (non-IID) across different clients in federated learning, a privacy-preserving data augmentation scheme was proposed. Firstly, a data augmentation framework for federated learning scenarios was designed. All clients generated synthetic samples locally and shared them with each other, which eased the problem of client drift caused by the difference of clients' data distributions. Secondly, based on generative adversarial network and differential privacy, a private sample generation algorithm was proposed. It helped clients to generate informative samples while preserving the privacy of clients' local data. Finally, a differentially private label selection algorithm was proposed to ensure the labels of synthetic samples will not leak information. Simulation results demonstrate that under multiple non-IID data partition strategies, the proposed scheme can consistently improve the model accuracy and make the model converge faster. Compared with the benchmark approaches, the proposed scheme can achieve at least 25% accuracy improvement when each client has only one class of samples.

Keywords: federated learning, non-IID, generative adversarial network, differential privacy, data augmentation

0 引言

联邦学习[1-2]以深度神经网络为载体，通过本地训练和中央聚合的模式，使各节点在数据不出本地的情况下共同训练一个全局模型，有效打破了不同团体和组织间的信息壁垒。然而，联邦学习实用化面临的一个关键问题是：节点间的数据往往是非独立同分布（non-IID, non-independent and identically distributed）的。由于面向的采样对象不同或采样设备存在规格差异，各节点的本地数据往往不服从同

收稿日期：2022−08−24；**修回日期**：2022−11−16

基金项目：国家重点研发计划基金资助项目（No.2016YFB1000500）；国家科技重大专项基金资助项目（No.2018ZX01028102）

Foundation Items: The National Key Research and Development Program of China (No.2016YFB1000500), The National Science and Technology Major Project (No.2018ZX01028102)

一分布，表现出较大的差异性。non-IID 数据会影响全局模型的预测准确率，甚至导致模型不收敛，从而使联邦学习任务不能取得预期的效果。例如，2 个节点希望共同建立一个判断就诊人员是否患病的二分类模型，节点 A 只拥有患者样本，节点 B 只拥有健康人员样本，则 A 训练得到的模型倾向于将所有样本判定为“患病”，而 B 则相反，此时 2 个本地模型都不具备基本的可用性，直接对模型进行聚合容易偏离全局最优的优化方向，因此全局模型不会有较高的准确率。

一些文献就 non-IID 数据对模型精度的影响进行了分析。文献[3]证明了数据分布的差异会导致各节点训练得到的本地模型逐渐收敛到局部最优，而偏离了全局最优的方向，严重影响聚合后的全局模型精度，学者将这种现象称为“本地模型偏移”或“节点偏移”。文献[4]则认为节点在模型训练的过程中发生了“知识遗忘”，虽然所有参与节点会在本地训练一定轮次后进行参数聚合，但数据分布的固有差异仍会导致节点在下一轮本地训练中不断巩固自身样本的知识，而逐渐忘记源于其他节点的样本知识。文献[5]将实际场景下的 non-IID 数据分为标签分布偏斜、特征分布偏斜以及样本数目偏斜三类，并通过实验验证标签分布偏斜对模型精度造成的影响最大。

针对 non-IID 数据，提高模型精度的工作主要存在以下困难：1) 联邦学习对隐私保护有较高的要求，节点间无法简单地通过共享原始数据来平衡数据分布；2) 联邦学习涉及多方节点的计算和通信，任何额外的工作量都可能导致任务时长成倍增加；3) 方案应该具备普适性，不能只适用于某种特定的 non-IID 数据分布情形。

为此，本文提出了一种面向联邦学习的数据增强方案，可以在保护用户数据隐私的前提下，解决 non-IID 数据引起的模型精度下降问题，同时不影响联邦学习主任务的效率。本文的主要贡献如下。

1) 提出了一种联邦学习数据增强（DA-FL, data augmentation in federated learning）框架，通过生成虚拟样本及标签并在节点间共享，平衡节点间的数据分布差异，从而减轻训练过程中各节点的模型偏移现象。

2) 提出一种隐私样本生成（PSG, private sample generation）算法，基于生成式对抗网络（GAN, generative adversarial network）生成虚拟样本，并利用差分隐私机制保护 GAN 的训练过程，防止敌手利用虚拟样本进行逆向攻击。

3) 提出一种隐私标签选取（PLS, private label selection）算法，利用差分隐私机制防止虚拟样本的对应标签泄露用户隐私。

4) 基于 MNIST、SVHN、Cifar10 等数据集，在多种 non-IID 数据划分方式下验证了方案的有效性。实验证明，所提方案能有效提高模型准确率，加速模型收敛，并取得了比基准方法更好的效果。

1　相关工作

为解决联邦学习中 non-IID 数据引起的模型精度下降问题，相关工作主要分为 3 个方向。

1) 为本地训练的损失函数添加正则项，从而控制和减轻本地模型偏移现象[6-8]。

2) 改进中央服务器的聚合算法，使聚合后的模型更新方向更贴近全局最优[9-11]。

3) 通过节点间共享数据来实现数据的补充和增强，缓解数据的 non-IID 程度[12-14]。

事实上，除上述 3 个方向外，个性化联邦学习[15-17]根据每个节点自身的数据特点和任务目标，学习个性化的模型，也有助于缓解数据非独立同分布带来的问题，然而本文主要关注建立统一、可用的模型，因此对该方向不作展开。

添加正则项和改进聚合算法两类方法具备模块化、效率高的优势，对原有联邦学习算法只需进行少量改动，且不会明显增加系统开销。然而其缺点为：1) 效果有限，无法带来明显的模型精度提升；2) 普适性不强，只适用于某些特定 non-IID 数据分布情形，而当节点间数据分布情况发生改变时，方法效果减弱甚至降低模型精度[5]。

数据共享方法从本质上缓解了节点间数据非独立同分布的状况，并且扩充了节点的本地数据集，因此对模型精度提升更明显。然而该方法往往面临新的问题，一是增加了隐私泄露的风险，二是增大了计算和通信开销。例如，文献[4]中提出各客户端在本地随机选取部分数据进行共享，但未考虑数据隐私问题，贡献的数据是明文。文献[18]提出了 COVID-GAN，整合多种来源的数据训练一个生成式对抗网络，来估计现实世界的人口流动，以便帮助相关部门制定决策，该方法虽然避免了明文传输，但一些研究表明敌手仍能通过访问生成器实现逆向攻击[19-20]。文献[14]提出一种基于样本平均的

数据增强方法，将多个样本进行平均，客户端之间通过共享这些平均样本来辅助校正本地训练，该方法通过平均计算来隐藏个体样本信息，但未能给出严格的隐私性证明。文献[13]提出了一种零次数据增强方法，客户端可根据上一轮的全局模型参数生成虚拟数据，无须接触其他客户端的真实数据。然而该方法只支持有限的模型架构，并且为了生成虚拟数据，客户端每轮训练需要求解额外的优化问题，影响了主任务的效率。

针对这些问题，本文提出一种隐私保护的联邦学习数据增强方案，与上述工作不同，所提方案中数据增强阶段不依赖于主任务的执行流程和中间结果，因此可在主任务前任意时间进行，而不影响主任务的效率，增强了方案的实用性。另外，所提方案利用差分隐私技术保护用户样本的隐私，防止敌手进行逆向攻击，提高了方案的隐私性。

2 理论基础

2.1 生成式对抗网络

生成式对抗网络是由 Goodfellow 等[21]于 2014 年提出的一种机器学习架构，包含生成器 $\mathcal{G}$ 和判别器 $\mathcal{D}$ 这 2 个模型。训练过程可看作 2 个模型的零和博弈，生成器输入低维随机噪声，输出虚拟样本，其优化目标是尽可能让判别器将虚拟样本误判为真实样本；而判别器输入真实样本和虚拟样本，输出每条样本是真实样本的概率，其优化目标是尽可能正确区分两类样本。该过程可看作如下优化问题

$$\min_{\mathcal{G}}\max_{\mathcal{D}}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log\mathcal{D}(\boldsymbol{x})]+\mathbb{E}_{z\sim p_z}[\log(1-\mathcal{D}(\mathcal{G}(\boldsymbol{z}))] \quad (1)$$

学者后续对 GAN 进行了许多优化和改进，例如，CGAN（conditional generative adversarial network）[22]允许生成器生成指定类别的数据，DCGAN（deep convolutional generative adversarial network）[23]改变生成器和判别器的模型架构，将全连接层替换为卷积层和卷积转置层，使生成器能更好地生成复杂图像。WGAN（Wasserstein generative adversarial network）[24]用 Wasserstein 距离代替 Jensen-Shannon 散度，来解决真实样本和虚拟样本分布不重叠时生成器的梯度消失问题，从而将优化问题(1)转化为

$$\min_{\mathcal{G}}\max_{w\in W}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[f_w(\boldsymbol{x})]-\mathbb{E}_{z\sim p_z}[f_w(\mathcal{G}(\boldsymbol{z}))] \quad (2)$$

其中，f_w 是判别器尝试拟合的函数，且满足 K-Lipschitz 连续。

2.2 差分隐私

差分隐私是由 Dwork 等[25]提出的隐私保护框架，最早用于保护数据库被查询时的样本隐私。差分隐私的概念可被扩展至任意算法。

若随机算法 $\mathcal{M}$ 对任意只相差一个元素的相邻集合 D 和 D'，以及 $\mathcal{M}$ 所有可能输出组成的集合 S，满足

$$\Pr[\mathcal{M}(D)\in S]\leqslant\exp(\varepsilon)\Pr[\mathcal{M}(D')\in S]+\delta \quad (3)$$

其中，概率取自对 $\mathcal{M}$ 的随机掷币，称 $\mathcal{M}$ 满足 (ε,δ)-差分隐私。

满足差分隐私的算法简称为 DP 算法，其输出对任意数据都不敏感，因此杜绝了敌手通过输出分布的差异推断一条数据的敏感信息。差分隐私一般通过对算法输出添加噪声来实现，以高斯机制为例，假设 f 是对数据集 D 的一个查询函数，查询返回结果为 $f(D)$，此时对结果添加噪声 $\mathcal{N}(0,\sigma^2)$，当满足 $\sigma\geqslant\dfrac{c\Delta f}{\varepsilon}, c^2>2\ln\left(\dfrac{1.25}{\delta}\right)$ 时，算法 $\mathcal{M}(D)=f(D)+\mathcal{N}(0,\sigma^2)$ 满足 (ε,δ)-DP[26]，其中，$\varepsilon\in(0,1)$，$\Delta f=\max\limits_{D,D'}\|f(D)-f(D')\|$。可见噪声方差由隐私预算 (ε,δ) 和查询函数敏感度 Δf 共同决定。

2.3 满足差分隐私的机器学习

文献[27]基于差分隐私技术提出了一种典型的隐私保护机器学习框架——差分隐私随机梯度下降（DP-SGD），在模型训练过程中，对一批样本中每个样本得到的梯度进行剪裁，平均梯度后再添加噪声，最后更新模型。该方法提供了模型单步更新的隐私保证，而模型训练需要经过多轮迭代，为统计全局的隐私保护程度，文献[27]进一步提出了隐私计量方法 Moments Accountant，用于计量训练全流程的隐私损失，根据该损失可以计算满足差分隐私定义的参数 (ε,δ)。

基于 DP-SGD 框架，学者们对满足差分隐私的生成式对抗网络（DP-GAN）进行了探索[28-29]，由于只有判别器接触真实数据，故在训练中对判别器的梯度添加噪声，使其满足差分隐私，由后处理定理[26]可知，在不接触原数据的情况下，对差分隐私算法的输出做任意计算都不会增加隐私损失，因此生成器及其生成数据也满足差分隐私。

3　方案设计

3.1　整体架构

本文的核心思路是通过数据增强的方式，平衡不同节点间数据分布的差异，从而提高最终模型的表现。每个客户端基于本地数据训练一个满足差分隐私的生成式对抗网络，然后用生成器输出一定数目的虚拟样本，并上传至中央服务器，形成一个共享数据集。服务器将共享数据集下发至各客户端，客户端合并本地数据集与共享数据集从而完成数据增强，至此预处理阶段结束。方案的整体架构如图 1 所示，以客户端 1 为例描绘了本地 GAN 训练和生成虚拟样本的过程，实际上所有客户端都同样执行上述流程。

本文的数据增强方案在预处理阶段进行，而联邦学习的模型训练过程则称为主任务阶段，当主任务开始时，各客户端基于增强后的数据集进行模型训练，与正常联邦学习的流程相同，此处不再赘述。

在方案高效性方面，虚拟样本的生成和客户端本地的数据增强不依赖于联邦学习主任务的执行逻辑和中间输出，除了因客户端本身数据集规模扩大而增加的训练开销，不在主任务阶段引入额外的计算和通信开销，提高了方案的实用性。

在方案可用性方面，注意到 GAN 生成的样本是不带类别标签的，可直接适用于主任务为半监督学习的情况。而当主任务是监督学习时，本文利用 CGAN 技术，先选取一批虚拟标签，再生成对应标签的虚拟样本，后续将主要介绍主任务为监督学习的情况。

在方案的隐私性方面，本文关注个体样本的隐私，分别在虚拟标签选取过程和虚拟样本生成过程引入差分隐私，从而保证敌手无法根据客户端的虚拟样本及标签推断出特定真实样本的信息。

表 1 给出了系统参数及含义。

表 1　　系统参数及含义

参数	含义
S	中央服务器
C_i, D_i	第 i 个客户端节点及其本地数据集
γ	客户端虚拟样本共享比例
m_i	C_i 共享的虚拟样本数目
U_i	D_i 中所有样本的对应标签集合
L	全局样本类别总数
$(\boldsymbol{x}, y)$	真实样本特征及标签
$(\hat{\boldsymbol{x}}, \hat{y})$	虚拟样本特征及标签
(ε, δ)	差分隐私机制的隐私预算
σ	差分隐私机制的噪声乘子
c	训练梯度的剪裁上界
$\mathcal{A}$	隐私损失计算函数
$\mathcal{D}, \mathcal{G}$	判别器, 生成器
B	一批训练样本的数目

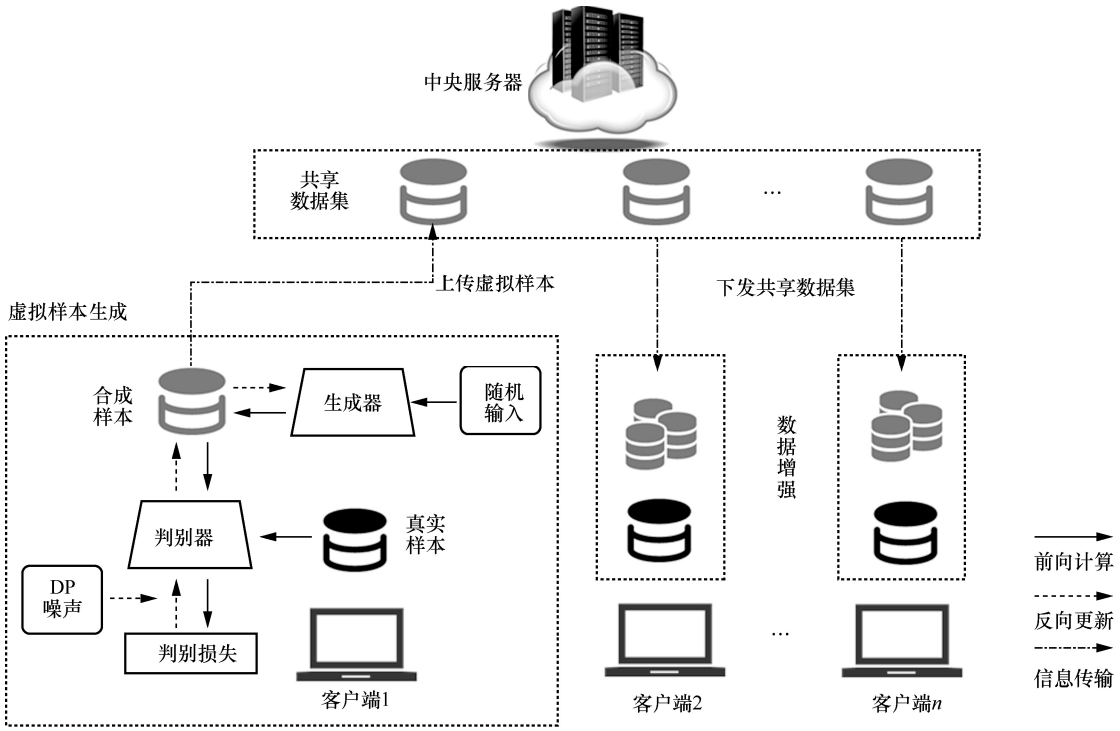


图 1　方案整体架构

3.2 联邦学习数据增强框架

本文提出的联邦学习数据增强框架 DA-FL 如算法 1 所示。

算法 1 联邦学习数据增强框架 DA-FL

输入 中央服务器 S，客户端 $C_1,\cdots,C_N$ 及其本地数据集 $D_1,\cdots,D_N$，虚拟样本共享比例 γ， PSG 算法的辅助输入 aux

输出 增强数据集 $\hat{D}_1,\cdots,\hat{D}_N$

1) for C_i do
2) $m_i \leftarrow |D_i|\gamma$
3) 从 D_i 的标签集合 U_i 随机采样 m_i 个标签 $\hat{y}_1,\cdots,\hat{y}_{m_i}$
4) $\hat{\boldsymbol{x}}_1,\cdots,\hat{\boldsymbol{x}}_{m_i} \leftarrow \mathrm{PSG}(m_i,\hat{y}_1,\cdots,\hat{y}_{m_i},D_i;\mathrm{aux})$
5) $\hat{D}_i \triangleq \{(\hat{\boldsymbol{x}}_k,\hat{y}_k)\}_{k\in[m_i]}$，上传 $\hat{D}_i$ 至 S
6) end for
7) for S do
8) 等待直到收到所有消息 $\{\hat{D}_i\}_{i\in[N]}$
9) 向 $C_i(i\in[N])$ 发送虚拟数据 $\{\hat{D}_j\}_{j\in[N],j\neq i}$
10) end for
11) for C_i do
12) 接收 $\{\hat{D}_j\}_{j\in[N],j\neq i}$
13) $\hat{D}_i \leftarrow D_i \bigcup \{\hat{D}_j\}_{j\neq i}$
14) end for
15) return $\hat{D}_1,\cdots,\hat{D}_N$

首先，每个客户端 C_i 计算所需生成的虚拟样本数目 m_i，由本地数据集 D_i 的规模乘以一个共享比例 γ 得到，即 $m_i=|D_i|\gamma$，设置参数 γ 是便于仿真时评估虚拟样本数目对联邦学习的提升效果，实际应用中各客户端的共享比例可以不同。

然后，记 U_i 为 D_i 中所有样本的对应标签集合，例如，客户端 C_i 本地共 5 个样本，其中一个样本属于类别 1，其余 4 个属于类别 2，则 $U_i=\{1,2,2,2,2\}$，易知 U_i 是一个无序的多重集，且 $|U_i|=|D_i|$。客户端 C_i 从 U_i 中随机选取 m_i 个标签 $\hat{y}_1,\cdots,\hat{y}_{m_i}$，称为虚拟标签。3.4 节将改进上述虚拟标签选取方法，使其满足差分隐私。

接着，C_i 执行 PSG 算法，生成与虚拟标签 $\hat{y}_1,\cdots,\hat{y}_{m_i}$ 对应的虚拟样本特征 $\hat{\boldsymbol{x}}_1,\cdots,\hat{\boldsymbol{x}}_{m_i}$，之后将虚拟样本和标签一并上传至中央服务器，中央服务器整合后下发至所有客户端。

最后，客户端收到源于其他节点的虚拟数据，将其加入本地数据集从而完成数据增强。

算法 1 中 PSG 算法的描述见 3.3 节。注意到，上述框架是模块化的，只涉及预处理阶段的数据增强，而不对后续的联邦学习流程做出改动。因此，现有的联邦学习主任务流程的优化算法理论上都可与本文方案相结合，从而进一步提高 non-IID 数据场景中的模型准确率。在第 4 节仿真实验中，为客观地对比不同方法的效果，采用基础的 FedAvg 算法作为本文方案的主任务算法。

3.3 满足差分隐私的样本生成

虽然 GAN 生成的样本与真实训练样本不同，但有研究表明通过模型或虚拟样本，仍能发起对训练样本的成员推断攻击[19-20]。因此，本文采用差分隐私保护真实样本的隐私性。

本文基于 DP-SGD 框架，在 GAN 训练过程中对判别器的每个梯度进行剪裁以控制其敏感度，然后将同一批次的梯度进行平均并添加噪声，同时利用 Moments Accountant 统计每轮训练产生的隐私损失。为了使生成器能生成指定类别的样本，对判别器和生成器的模型结构进行修改，用嵌入层对样本标签进行表示，并将其作为判别器和生成器的额外输入。另外，GAN 模型中常使用批归一化技术，而该方法需获取一批样本的整体统计数据，破坏了差分隐私性质[30-31]，因此将其替换为实例归一化，并禁止追踪滑动均值与方差，模型架构详见 4.1 节。

隐私样本生成算法如算法 2 所示。步骤 1)～步骤 21)是生成式对抗网络的训练主循环，其中，步骤 5)～步骤 12)为判别器的训练和更新过程，步骤 13)～步骤 18)为生成器的训练和更新过程；步骤 19)～步骤 21)利用 Moments Accountant 统计当前的累计隐私损失，并计算已消耗的隐私预算，一旦超出预先设定的隐私预算，则停止训练并撤销当前轮次的训练结果；步骤 22)～步骤 26)利用训练得到的生成器进行样本生成。

算法 2 PSG 算法

输入 生成虚拟样本数目 m，虚拟标签 $\hat{y}_1,\cdots,\hat{y}_m$，本地数据集 D，预定训练轮数 T，学习率 η，批样本数 B，隐私预算 (ε_0,δ_0)，训练梯度剪裁上界 c，噪声乘子 σ，隐私损失计算函数 $\mathcal{A}$

输出 虚拟样本特征 $\hat{\boldsymbol{x}}_1,\cdots,\hat{\boldsymbol{x}}_m$

1) 初始化判别器 $\mathcal{D}$ 和生成器 $\mathcal{G}$ 的参数 $\boldsymbol{\theta}_{\mathcal{D}},\boldsymbol{\theta}_{\mathcal{G}}$

2) for $t = 1,\cdots,T$ do　　//训练主循环
3)　从正态分布随机采样 $\boldsymbol{z}_1,\cdots,\boldsymbol{z}_B$
4)　从 D 中随机采样 $(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_B,y_B)$
5)　for $i = 1,\cdots,B$ do　　//训练判别器
6)　　$\tilde{\boldsymbol{x}}_i \leftarrow \mathcal{G}(\boldsymbol{z}_i,y_i)$
7)　　$\ell_i^{(\mathcal{D})} \leftarrow \log \mathcal{D}(\boldsymbol{x}_i,y_i) + \log(1-\mathcal{D}(\tilde{\boldsymbol{x}}_i,y_i))$
8)　　$\boldsymbol{g}_i^{(\mathcal{D})} \leftarrow \nabla_{\boldsymbol{\theta}_{\mathcal{D}}} \ell_i^{(\mathcal{D})}$　　//计算梯度
9)　　$\boldsymbol{g}_i^{(\mathcal{D})} \leftarrow \dfrac{\boldsymbol{g}_i}{\max(1,\| \boldsymbol{g}_i^{(\mathcal{D})} \| /c)}$ //梯度剪裁
10)　end for
11)　$\overline{\boldsymbol{g}}^{(\mathcal{D})} \leftarrow \dfrac{1}{B}\left(\sum_{i=1}^{B} \boldsymbol{g}_i^{(\mathcal{D})} + \mathcal{N}(0,\sigma^2 c^2 \boldsymbol{I})\right)$
　　//添加噪声
12)　$\boldsymbol{\theta}_{\mathcal{D}} \leftarrow \boldsymbol{\theta}_{\mathcal{D}} + \eta \overline{\boldsymbol{g}}^{(\mathcal{D})}$　　//更新判别器
13)　for $i = 1,\cdots,B$ do　　//训练生成器
14)　　$\ell_i^{(\mathcal{G})} \leftarrow \log(1-\mathcal{D}(\tilde{\boldsymbol{x}}_i,y_i))$
15)　　$\boldsymbol{g}_i^{(\mathcal{G})} \leftarrow \nabla_{\boldsymbol{\theta}_{\mathcal{G}}} \ell_i^{(\mathcal{G})}$
16)　end for
17)　$\overline{\boldsymbol{g}}^{(\mathcal{G})} \leftarrow \dfrac{1}{B}\sum_{i=1}^{B} \boldsymbol{g}_i^{(\mathcal{G})}$
18)　$\boldsymbol{\theta}_{\mathcal{G}} \leftarrow \boldsymbol{\theta}_{\mathcal{G}} - \eta \overline{\boldsymbol{g}}^{(\mathcal{G})}$ //更新生成器
19)　$\varepsilon_t \leftarrow \mathcal{A}(\delta_0,B,t,\sigma,|D|)$//计算已消耗的隐私预算
20)　if $\varepsilon_t \geqslant \varepsilon_0$ then break//超出隐私预算则停止训练
21) end for
22) for $k = 1,\cdots,m$ do
23)　从正态分布随机采样 $\boldsymbol{z}$
24)　$\tilde{\boldsymbol{x}}_k \leftarrow \mathcal{G}(\boldsymbol{z},\hat{y}_k)$ //生成虚拟样本
25) end for
26) return $\hat{\boldsymbol{x}}_1,\cdots,\hat{\boldsymbol{x}}_m$

定理 1 给出了算法 2 的隐私性质。

定理 1　算法 2 满足 (ε_0,δ_0)-差分隐私。

证明　如附录 1 所示。

3.4　满足差分隐私的标签选取

算法 1 中客户端除了向服务器提交虚拟样本的特征 $\{\hat{\boldsymbol{x}}_k\}_{k\in[m]}$ 外，还要提交虚拟标签 $\{\hat{y}_k\}_{k\in[m]}$，所以需要保证选取的虚拟标签也满足差分隐私。

设计标签选取方法需要兼顾隐私性和可用性。一种简单的方法是客户端为每个类别生成相同数目的虚拟样本，且虚拟样本数目为事先约定，则该标签选取过程与本地数据集无关，也不会泄露任何信息。这种方法适用于 IID 数据场景，然而 non-IID 数据场景中客户端可能只拥有某几类的样本数据，对于缺失的类别，生成器无法生成有效的虚拟样本，影响了样本的可用性。

考虑到上述类别缺失问题，以及共享数据集中样本的多样性和全面性，一个合理的方式是使共享数据集的分布逼近全局数据的分布，从而使模型在共享数据集上的优化方向趋近全局优化方向。此时，客户端选取的虚拟标签应该与本地真实标签的分布相同，即不同类别间的虚拟标签数目占比应与本地真实标签保持一致。设全局数据分为 L 个类别，客户端每个类别的真实样本数目分别为 $n_1,\cdots,n_L$，每类选取虚拟标签的数目分别为 $\hat{n}_1,\cdots,\hat{n}_L$，则应有 $\hat{n}_k = \lfloor \gamma n_k \rfloor$，$k \in [L]$。

但是，该标签采样方法是确定性的，无法抵抗敌手的逆向差分攻击，故在此基础上，引入指数机制（EM, exponential mechanism）对每种类别采样的标签数目进行扰动，具体步骤如下。

1) 对类别 k，定义效用函数为

$$u_k(D,r) = -\left|\frac{r}{\hat{n}} - \frac{n_k}{n}\right| \tag{4}$$

其中，$n = n_1 + \cdots + n_L, \hat{n} = \lfloor \gamma n \rfloor$，对 $\hat{n}_k$ 的每个候选值 r 赋予一个效用值，效用值越高则被选概率越大。式(4)的定义表明，每个类别虚拟标签的数目比例与本地真实数据越接近越好。

2) 对类别 k，令 $\hat{n}_k$ 取值为 r 的概率为

$$\Pr[\hat{n}_k = r] = \frac{\exp\left(\dfrac{\varepsilon u_k(D,r)}{2\Delta u_k}\right)}{\sum\limits_{r'\in[\hat{n}]} \exp\left(\dfrac{\varepsilon u_k(D,r')}{2\Delta u_k}\right)} \tag{5}$$

3) 客户端按式(5)中概率输出每个类别的虚拟标签数目 $\hat{n}_1,\cdots,\hat{n}_L$，已知所有类别和每个类别的标签数目，易确定所有虚拟标签 $\hat{y}_1,\cdots,\hat{y}_m$，其中 $m = \hat{n}_1 + \cdots + \hat{n}_L$ 为实际选取的虚拟标签总数。

依据上述思路，给出虚拟标签选取算法如下。

算法 3　PLS 算法

输入　虚拟样本共享比例 γ，全局样本类别总数 L，客户端样本总数 n，其中每个类别样本数 $n_1,\cdots,n_L$

输出　虚拟标签 $\hat{y}_1,\cdots,\hat{y}_m$

1) $\hat{n} \leftarrow \lfloor \gamma n \rfloor$
2) for $k \in [L]$ do
3)　　for $r \in [\hat{n}]$ do

4) 根据式(5)计算取值概率 $\Pr[\hat{n}_k = r]$

5) end for

6) 根据步骤 4)的概率输出 $\hat{n}_k$

7) end for

8) 将 $\hat{n}_1,\cdots,\hat{n}_L$ 转化为 $\hat{y}_1,\cdots,\hat{y}_m$

9) return $\hat{y}_1,\cdots,\hat{y}_m$

利用算法 3 代替算法 1 的步骤 3)，即可保证虚拟标签满足差分隐私。

定理 2 算法 3 满足 $(\varepsilon,0)$-差分隐私。

证明 如附录 2 所示。

至此，根据定理 1 和定理 2，可以得到算法 1 的隐私性质。

定理 3 算法 1 满足 (ε,δ)-差分隐私。

证明 算法 1 中每个客户端需按顺序执行算法 3 和算法 2，根据差分隐私的组合性质，假设算法 2 满足 (ε_0,δ_0)-差分隐私，算法 3 满足 $(\varepsilon_1,0)$-差分隐私，则算法 1 满足 (ε,δ)-差分隐私，其中，$\varepsilon=\varepsilon_0+\varepsilon_1,\delta=\delta_0$。证毕。

4 仿真实验

4.1 实验设置

1) 实验环境

本文的实验环境为 Amazon EC2 p3.2xlarge，硬件配置为 8vCPU、61 GB 内存、Tesla V100 GPU。本文方案基于 Pytorch 和 Opacus[31]库实现，参与对比的基准方法部分采用了 NIID-Bench[5] 和 FedLab[32]中的实现代码。

2) 数据集与数据划分

实验数据集为 MNIST[33]、FashionMNIST[34]、Cifar10[35]、SVHN[36]。文献[5]详细研究了不同的 non-IID 数据划分方式对模型精度的影响，本文从中选择了 3 种对模型精度影响最大的数据划分方式进行实验，分别如下：1-Label，每个客户端只有一种类别的样本；2-Label，每个客户端只有 2 种不同类别的样本；Dir(0.05)，客户端的样本服从 Dirichlet 分布[10] Dir(β)，其中，参数β越小表示非独立同分布程度越高，此处将β设置为一个较小的值，即β=0.05。

本文设置了 10 个客户端的联邦学习场景，针对上面 3 种数据划分方式，随机生成一组样本分布并固定，以便公平地比较不同方法的效果。图 2 展示了 non-IID 数据划分情况，每个子图展示了各客户端的样本分布，不同类别样本用不同深浅的灰色标识。

3) 模型架构

本文使用的 GAN 和 CNN 分类模型的结构如图 3 所示。其中，判别器和生成器的主体分别为 4 个卷积层（conv）和 4 个卷积转置层（upconv），均采用实例归一化。跨步（stride）、填充（padding）等参数设置如图 3 所示。判别器和生成器中间层的激活函数

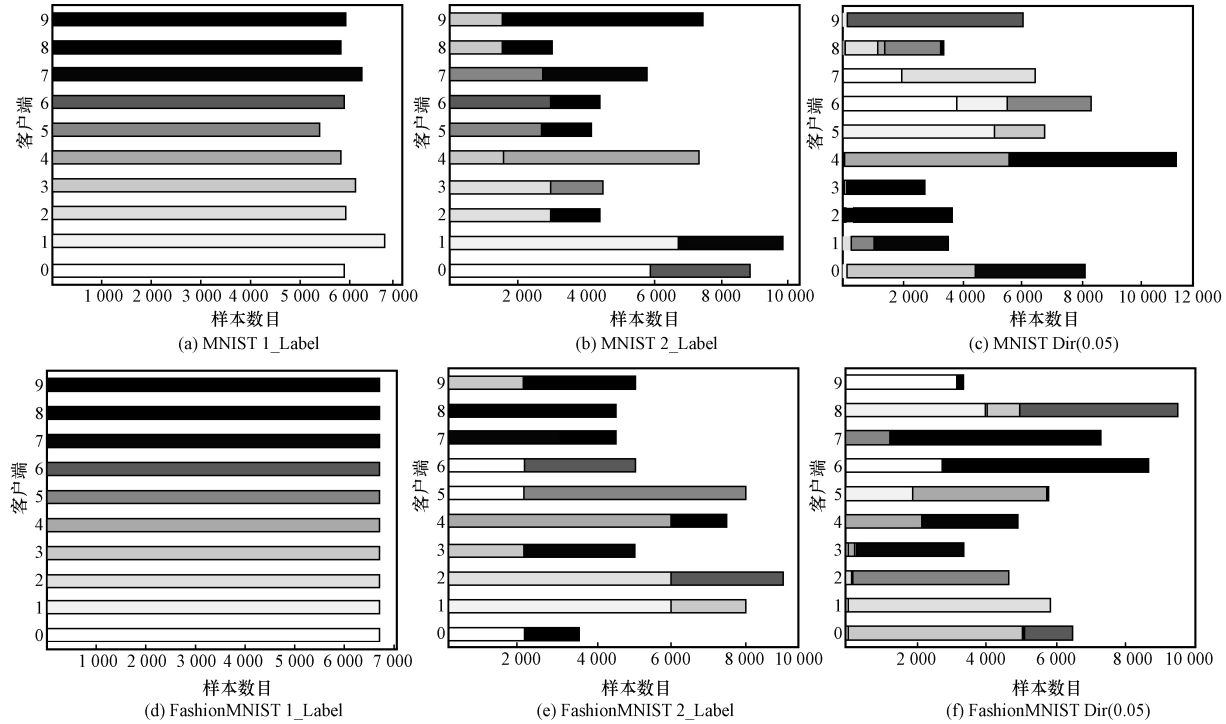
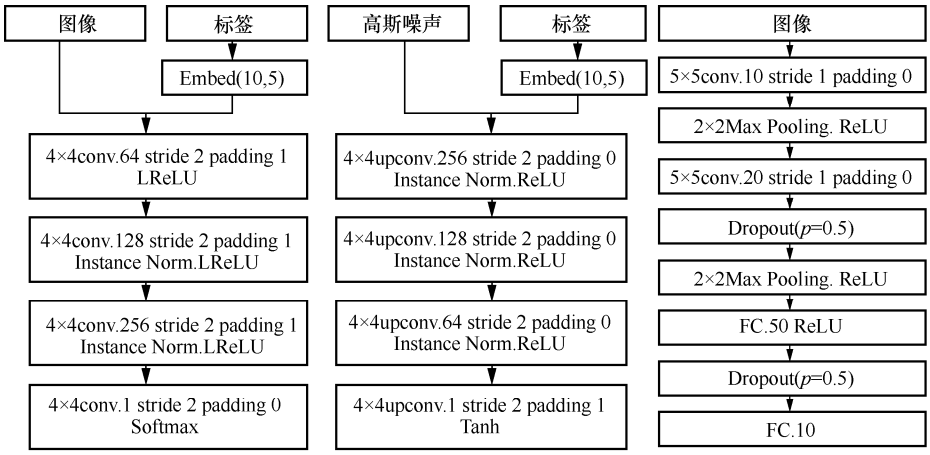


图 2 Non-IID 数据划分情况



(a) 判别器　　(b) 生成器　　(c) 分类模型

图 3　GAN 和 CNN 分类模型的结构

分别为 LReLU 和 ReLU。判别器接收 32 像素×32 像素图像和标签作为输入，输出一个判别评分；生成器接收维度为 10 的高斯噪声和标签作为输入，32 像素×32 像素图像作为输出。本文所用数据集图像规格为 28 像素×28 像素，故对输入判别器和生成器输出的图像进行 resize 处理。联邦学习主任务的分类模型主要包含 2 个卷积层和 2 个全连接层 FC，每层卷积后设置最大池化层 Max Pooling 和 ReLU 激活函数。

4) 相关参数

表 2 给出了实验参数设置。其中，每轮参与训练的客户端比例设置为 1，即所有客户端都参与训练。对于数据集 SVHN 和 Cifar10，隐私预算ε分别设置为 100 和 200。

表 2　　实验参数设置

阶段	参数名称	参数值
数据增强阶段	虚拟样本共享比例γ	0.01
	GAN 预定训练轮数/次	50
	Batch Size	256
	生成器输入噪声维度	10
	噪声乘子σ	0.5
	剪裁上界 c	2
	隐私预算(ε,δ)	50, 0.000 01
	优化算法	Adam
	优化器参数(学习率,β_1, β_2)	0.0002, 0.5, 0.999
联邦学习阶段	客户端数/个	10
	节点间总通信轮数/次	50
	客户端本地训练轮数/次	5
	每轮参与训练的客户端比例	1
	Batch Size	32
	优化算法	SGD
	优化器参数(学习率,动量)	0.01, 0.5

4.2　方案有效性验证

本节验证方案的有效性。基于图 2 所示的数据划分方式，测试了联邦学习经过 50 轮通信后的全局模型准确率。同时，在相同的参数设置下，将本文方案与 FedAvg[1]、FedProx[6]、SCAFFOLD[7]、FedNova[9]、FedMix[14]进行了对比。其中，对于本文方案，测试了虚拟样本共享比例为 0.01 和 0.05 这 2 种情况；对于 FedProx，超参数μ测试了{0.001, 0.01, 0.1, 1}4 种取值；对于 FedMix，超参数λ测试了{0.05, 0.1, 0.2}3 种取值，分别报告最好的一组结果。另外，对每个数据集测试了集中训练（centralized training）的模型精度，该结果用来估计给定模型架构、训练算法和超参数后，所能达到的模型精度上界。

由表 3 可知，本文方案在 3 种数据划分方式下，都取得了相对较高的模型准确率，特别是 1-Label 的极端 non-IID 场景下，本文方案在各数据集上都取得了比基准方法更高的模型准确率。由 2-Label 和 Dir(0.05)的实验结果可见，样本数目的不均衡对模型精度的影响相对较小，而客户端本地数据的类别多样性对模型精度的影响较大。在本文方案中，每个客户端的增强数据集包含了所有类别的样本，因此能取得较好的模型表现。

图 4 给出了不同方法训练中的模型准确率变化情况，其中，本文方案设置γ=0.05。从图 4 可知，本文方案在 non-IID 数据场景中可以使模型快速收敛，在 1-Label 下，基准方法训练过程中的模型准确率振荡幅度较大甚至不收敛，而本文方案中模型在前 5 轮通信即可收敛至极值点附近。相比于上述

表 3 不同方法的模型测试准确率对比

方法	MNIST			FMNIST			Cifar10		SVHN	
	1-Label	2-Label	Dir(0.05)	1-Label	2-Label	Dir(0.05)	1-Label	2-Label	1-Label	2-Label
FedAvg[1]	50.44%	90.99%	96.63%	39.13%	69.92%	77.58%	13.99%	**45.62%**	9.69%	48.48%
FedProx[6]	48.56%	87.56%	96.32%	46.48%	62.39%	77.46%	12.28%	31.46%	15.53%	**58.86%**
SCAFFOLD[7]	9.52%	91.53%	97.82%	10.00%	68.87%	75.38%	10.00%	42.71%	9.33%	34.53%
FedNova[9]	42.40%	88.86%	96.96%	36.47%	70.13%	77.38%	10.73%	44.97%	11.16%	47.76%
FedMix[14]	23.32%	65.64%	91.24%	36.18%	51.03%	64.28%	9.98%	43.61%	19.59%	44.89%
本文方案(γ=0.01)	75.13%	93.21%	97.11%	75.11%	**79.12%**	**82.84%**	**25.95%**	44.27%	21.46%	51.48%
本文方案(γ=0.05)	**85.99%**	**94.26%**	**97.85%**	**75.54%**	78.56%	82.41%	23.25%	40.02%	**42.94%**	52.41%
centralized training		99.16%			90.15%			64.87%		87.90%

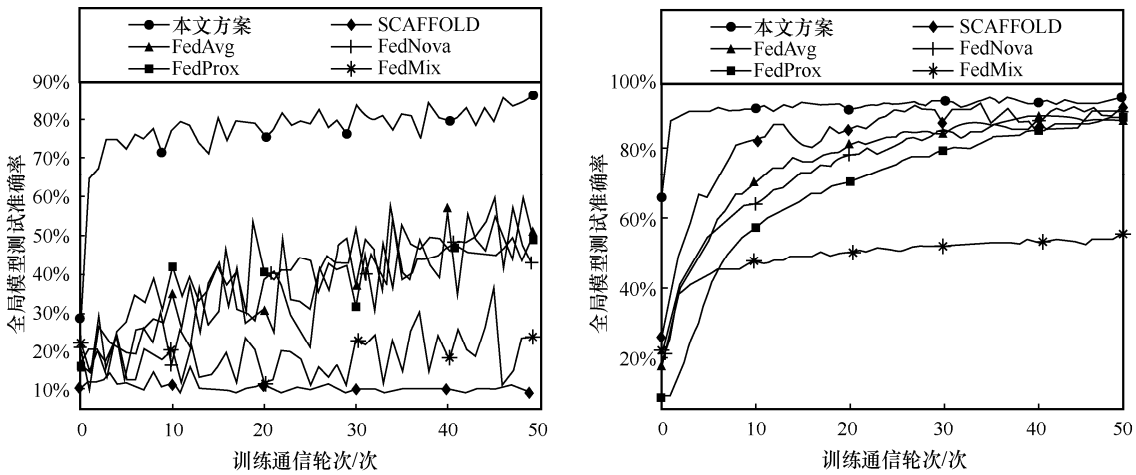


图 4 不同方法训练中的模型准确率变化情况

情形，在 2-Label 下，基准方法与本文方案的最终模型准确率差距缩小，但基准方法的收敛速度较慢，所需通信轮次较多。联邦学习主任务阶段往往涉及多个节点参与，节约此阶段的训练轮次具有重要的实际意义。

4.3 隐私预算对方案效果的影响

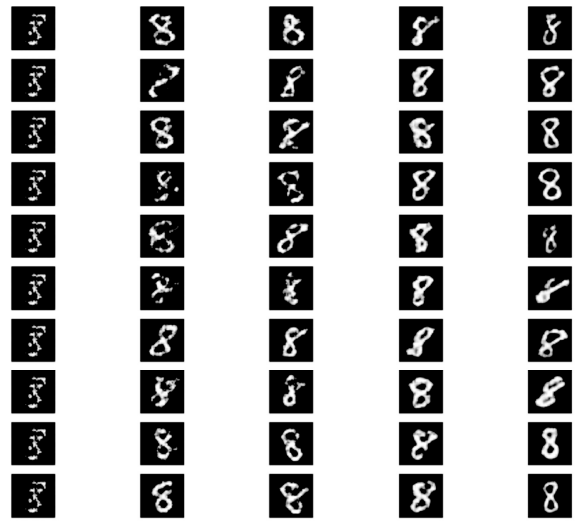
本节研究差分隐私的隐私预算对方案效果的影响。基于 MNIST 数据集在 1-Label 下进行实验，令 $\gamma=0.01,\delta=10^{-5}$，分别测试 $\varepsilon=1,5,20,50,\infty$ 这 5 种情况下，主任务模型经过 50 轮通信后的准确率，其中 $\varepsilon=\infty$ 表示不对 GAN 训练添加噪声。

由表 4 可知，当不添加噪声时，GAN 生成的样本能帮助主任务模型达到最高的准确率；当隐私预算为 5～50 时，模型准确率相对接近；当隐私预算为 1 时，模型准确率明显降低。上述情况体现了 DP-GAN 可用性和隐私性之间的矛盾，隐私保护程度越强，生成的样本质量越低。

表 4 不同隐私预算时的模型准确率

ε	模型准确率
1	58.81%
5	74.55%
20	75.87%
50	75.13%
∞	88.33%

图 5 展示了不同隐私预算时的虚拟样本，此处选取只有样本类别“8”的客户端，对不同的隐私预算 $\varepsilon=1,5,20,50,\infty$ 分别训练一个生成器，然后固定一组输入噪声，观察每个生成器输出的虚拟样本。由图 5 可知，随着隐私预算的减少，虚拟样本质量略有降低，当 $\varepsilon=1$ 时发生了模式崩塌，对于不同的输入噪声，生成器只输出相同的图像，说明对梯度添加的噪声过大，影响了判别器的正常更新，从而无法正确指导生成器优化。



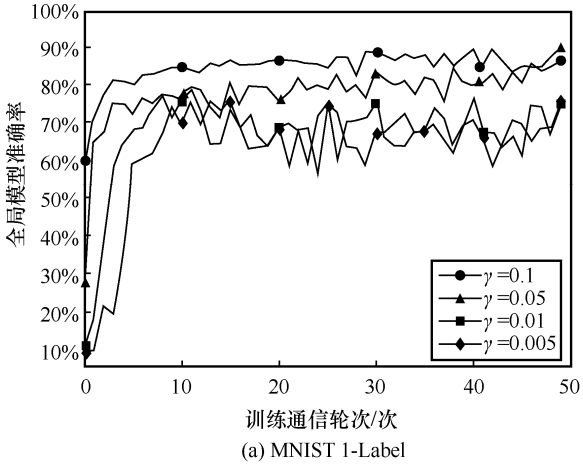
(a) ε=1　(b) ε=5　(c) ε=20　(d) ε=50　(e) ε=∞

图 5　不同隐私预算时的虚拟样本

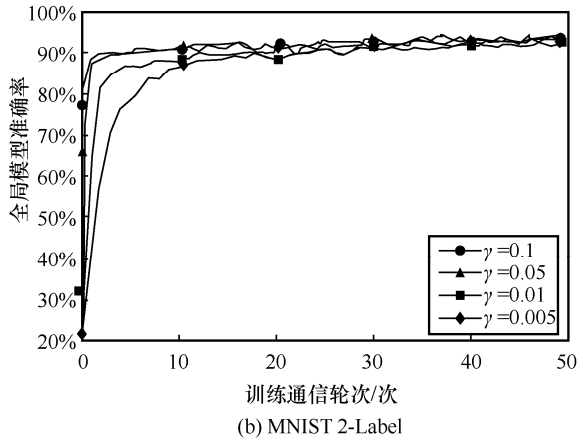
4.4　样本共享数目对方案效果的影响

本节研究客户端贡献的虚拟样本共享数目对方案效果的影响。基于 MNIST 数据集进行实验，样本共享比例分别设置为 $\gamma=0.1,0.05,0.01,0.005$，观察训练过程中模型准确率的变化情况。

由图 6 可知，在 1-Label 中，节点间数据分布差异较大，增大虚拟样本的共享数目有助于平衡全局数据的分布，从而增强训练稳定性，提高最终模型的精度。在 2-Label 中，节点间数据分布差异变小，γ 值对最终模型准确率的影响也变小，4 种取值都能获得相近的模型表现，但增大γ 仍有助于提高模型收敛速度。由表 3 可知，Cifar10 数据集训练过程中γ=0.05 时的模型准确率反而低于γ=0.01，这是因为 GAN 训练过程中的噪声导致生成样本质量较低，造成了数据分布与样本质量间的矛盾，加入更多的虚拟样本更好地平衡了数据分布，但降低了总体样本质量。



(a) MNIST 1-Label



(b) MNIST 2-Label

图 6　不同样本共享比例的模型准确率变化曲线

4.5　方案效率测试

本节测试方案的执行效率，主要验证以下两点。1) 主任务效率：方案的主任务阶段耗时是否与基准方法相近；2) 总体效率：考虑预处理阶段耗时，方案的总体耗时是否仍处于可接受范围。

基于表 2 的默认参数设置，在 6 个场景下对不同方案进行效率对比，结果如图 7 所示，其中 Ours-Main 和 Ours-Pre 分别代表本文方案的主任务阶段和预处理阶段。因为联邦学习是同步系统，每个通信轮的耗时取决于执行最慢的节点，而在 2-Label 和 Dir(0.05)中存在明显的样本数目偏斜，所以主任务阶段耗时比 1-Label 更长。

本文方案主任务阶段采用的是 FedAvg 算法，主要区别是由于数据增强，客户端的本地数据集规模增大，故由图 7 可知，本文方案主任务阶段的耗时与 FedAvg 等基准方法相近。其次，本文方案数据增强阶段的耗时约为主任务阶段的 10.2%～16.7%，2 个阶段的总体耗时相比于基准方法处于可接受范围。

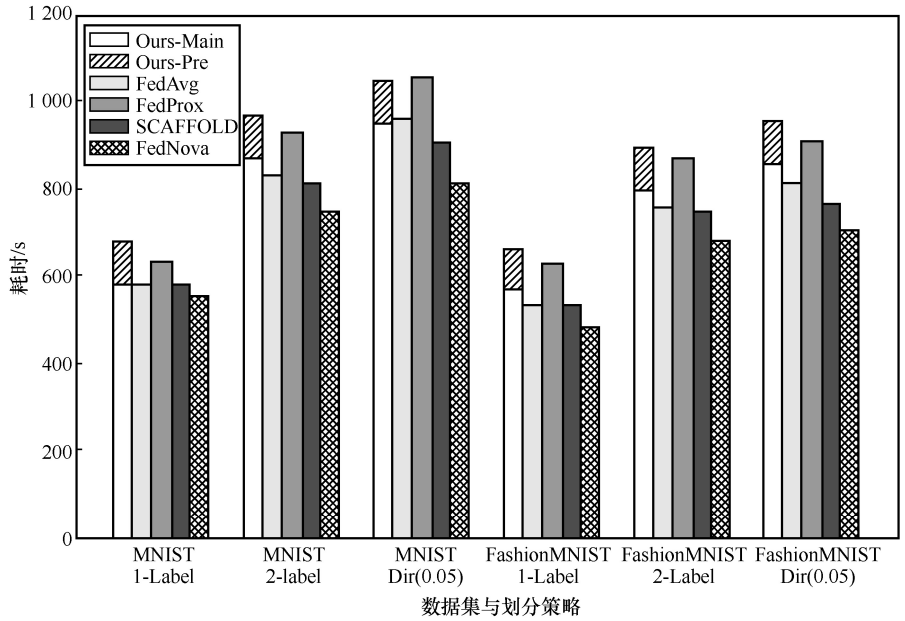


图 7　不同方案效率对比

5　结束语

本文提出一种面向非独立同分布数据的联邦学习数据增强方案，所有客户端在本地训练一个生成式对抗网络，然后生成一定数目的虚拟样本，客户端间通过共享虚拟样本来增强本地数据。在生成式对抗网络训练过程中，对判别器添加合适的噪声，使虚拟样本满足差分隐私，从而保证原始数据的隐私。同时，设计了满足差分隐私的标签选取算法，避免在数据共享过程中虚拟标签泄露隐私。与已有工作相比，所提方案在多种数据划分下都取得了更高的模型精度和更快的模型收敛速度。在未来的工作中，将进一步研究 DP-GAN 可用性与隐私性之间的矛盾，在合理的隐私预算下，生成更复杂的、高可用的虚拟样本，提高方案在面向复杂数据集时的有效性。

附录 1　定理 1 的证明

基于 Moments Accountant 技术[27]证明定理 1。首先，定义调用一次算法 $\mathcal{M}$ 所产生的隐私损失为随机变量 Z 为

$$Z(o;\mathcal{M},D,D')=\log\frac{\Pr[\mathcal{M}(D)=o]}{\Pr[\mathcal{M}(D')=o]} \tag{6}$$

其中，D,D' 是相邻数据集，o 属于 $\mathcal{M}$ 的输出域。可以通过计算 Z 矩母函数的值来估计隐私损失的范围，定义

$$\alpha_{\mathcal{M}}(\lambda;D,D')=\log\mathbb{E}_{o\sim\mathcal{M}(D)}[\exp(\lambda Z(o;\mathcal{M},D,D'))] \tag{7}$$

$$\alpha_{\mathcal{M}}(\lambda)=\max_{D,D'}\alpha_{\mathcal{M}}(\lambda;D,D') \tag{8}$$

引理 1[27]　对任意 $\varepsilon>0$，算法 $\mathcal{M}$ 满足 (ε,δ)-差分隐私，其中，$\delta=\min_{\lambda}\exp(\alpha_{\mathcal{M}}(\lambda)-\lambda\varepsilon)$。

记算法 2 为 $\mathcal{M}$，则由引理 1 可知，为保证算法 $\mathcal{M}$ 满足差分隐私，只需约束 $\alpha_{\mathcal{M}}(\lambda)$ 的上界，并且，由 $\alpha_{\mathcal{M}}(\lambda)$ 可以进一步计算得到隐私预算 (ε,δ)。$\mathcal{M}$ 共包含 T 轮训练，记第 t 轮训练为子算法 $\mathcal{M}_t$，$\mathcal{M}_t$ 又包含 2 个子算法：判别器 $\mathcal{D}$ 的训练过程 $\mathcal{M}_t^{(\mathcal{D})}$，生成器 $\mathcal{G}$ 的训练过程 $\mathcal{M}_t^{(\mathcal{G})}$。

下面证明对每个 t，$\mathcal{M}_t^{(\mathcal{D})}$ 的隐私损失存在上界。算法 2 中步骤 7)和步骤 8)可合并写为

$$\boldsymbol{g}_i^{(\mathcal{D})}\leftarrow\nabla_{\boldsymbol{\theta}_{\mathcal{D}}}[\log\mathcal{D}(\boldsymbol{x}_i,y_i)+\log(1-\mathcal{D}(\tilde{\boldsymbol{x}}_i,y_i))] \tag{9}$$

式(9)表示由真实样本和虚拟样本共同计算得到的判别器梯度，将该计算过程抽象为

$$f(\boldsymbol{x}_i)=\boldsymbol{g}_i^{(\mathcal{D})} \tag{10}$$

设 batch size 为 B，则式(9)需执行 B 次，然后对每个梯度进行剪裁并添加噪声，最后计算平均梯度。为方便分析，令剪裁上界 c=1，于是 $\mathcal{M}_t^{(\mathcal{D})}$ 可表示为

$$\mathcal{M}_t^{(\mathcal{D})}=\frac{1}{B}\left(\sum_{i=1}^{B}\boldsymbol{g}_i^{(\mathcal{D})}+\mathcal{N}(0,\sigma^2\boldsymbol{I})\right),\|\boldsymbol{g}_i^{(\mathcal{D})}\|\leqslant 1 \tag{11}$$

引理 2[27]　设 $f:D\to\mathbb{R}^p$ 且 $\|f(\cdot)\|_2\leqslant 1$。令 $\sigma\geqslant 1$，$\mathcal{I}$ 是从[n]中随机采样的一个子集，$\mathcal{I}$ 中每个元素被选中概率为 q 且相互独立。记 μ_0,μ_1 分别为 $\mathcal{N}(0,\sigma^2),\mathcal{N}(1,\sigma^2)$ 的概率密度函数，且 $\mu\triangleq(1-q)\mu_0+q\mu_1$，则对算法 $\mathcal{M}(d)=\sum_{i\in\mathcal{I}}f(d_i)+\mathcal{N}(0,\sigma^2\boldsymbol{I})$ 有

$$\alpha_{\mathcal{M}}(\lambda)=\log \mathrm{E}_{z\sim\mu}\left[\frac{\mu(z)^{\lambda}}{\mu_0(z)}\right] \tag{12}$$

若 $q<\dfrac{1}{16\sigma}$，则对满足 $0<\lambda\leqslant\sigma^2\ln\left(\dfrac{1}{q}\sigma\right)$ 的任意整数 λ，$\alpha_{\mathcal{M}}(\lambda)$ 存在上界，即

$$\alpha_{\mathcal{M}}(\lambda)\leqslant\frac{q^2\lambda(\lambda+1)}{(1-q)\sigma^2}+O\left(\frac{q^3\lambda^3}{\sigma^3}\right) \tag{13}$$

结合式(10)、式(11)、式(13)，引理 2 给出了 $\alpha_{\mathcal{M}_t^{(\mathcal{D})}}(\lambda)$ 的一个近似上界，再结合引理 1 可知 $\mathcal{M}_t^{(\mathcal{D})}$ 满足差分隐私。在实际应用中，一般通过数值积分的方法计算式(12)从而确定 $\alpha_{\mathcal{M}}(\lambda)$ [27,37]，算法 2 中将该过程抽象为函数 $\mathcal{A}$，输入当前的环境参数，输出已消耗的隐私预算。

接着，证明第 t 轮中 $\mathcal{M}_t^{(\mathcal{G})}$ 不会带来额外的隐私损失。由于生成器的更新只需与判别器交互，而不接触真实样本，可将 $\mathcal{M}_t^{(\mathcal{G})}$ 视为一个作用于 $\mathcal{M}_t^{(\mathcal{D})}$ 输出结果的随机映射。因此，由引理 3 可知训练 $\mathcal{G}$ 不会导致额外的隐私损失，即 $\alpha_{\mathcal{M}_t}(\lambda)=\alpha_{\mathcal{M}_t^{(\mathcal{G})}\circ\mathcal{M}_t^{(\mathcal{D})}}(\lambda)=\alpha_{\mathcal{M}_t^{(\mathcal{D})}}(\lambda)$。

引理 3　后处理定理[26]令 $\mathcal{M}:D^{|\mathcal{X}|}\to R$ 是满足 (ε,δ)-差分隐私的随机算法，且 $f:R\to R'$ 是任意随机映射，则 $f\circ\mathcal{M}:D^{|\mathcal{X}|}\to R'$ 同样满足 (ε,δ)-差分隐私。

至此，证明了算法 2 每一轮训练 $\mathcal{M}_t$ 满足差分隐私。回顾算法 2 包含 T 轮迭代，故需计算 $\mathcal{M}$ 的总体隐私损失。

引理 4[27]　假设算法 $\mathcal{M}$ 由一系列子算法 $\mathcal{M}_1,\cdots,\mathcal{M}_T$ 组成，其中 $\mathcal{M}_t:\prod_{i=1}^{t-1}\mathcal{R}_i\times D\to\mathcal{R}_t$，子算法间的输入和输出可能存在顺序依赖。则对于任意λ满足

$$\alpha_{\mathcal{M}}(\lambda)\leqslant\sum_{t=1}^{T}\alpha_{\mathcal{M}_t}(\lambda) \tag{14}$$

引理 4 说明统计变量 $\alpha_{\mathcal{M}_t}(\lambda)$ 满足线性的组合性质，因此结合式(13)，可以对算法 2 总体的隐私损失进行估计，即 $\alpha_{\mathcal{M}}(\lambda)\leqslant\dfrac{Tq^2\lambda^2}{\sigma^2}$，根据文献[27]可知存在常数 c_1,c_2 使对任意 $\varepsilon<c_1q^2T$，令 $\sigma\geqslant c_2q\dfrac{\sqrt{T\log\left(\frac{1}{\delta}\right)}}{\varepsilon}$ 时，算法 2 满足 (ε,δ)-差分隐私。

至此，证明了给定参数 ε,δ,q,T 时，通过选取合适的噪声乘子 σ 可使整个训练过程满足 (ε,δ)-差分隐私。实际执行过程中，算法 2 的噪声乘子是预先确定的，通过隐私计量函数 $\mathcal{A}$ 计算当前已消耗的隐私预算 (ε_t,δ_t)，当其超过既定隐私预算 (ε_0,δ_0) 时，则停止训练。

最后，算法 2 利用生成器 $\mathcal{G}$ 生成虚拟样本特征 $\hat{\boldsymbol{x}}_1,\cdots,\hat{\boldsymbol{x}}_m$，由于无须接触真实样本，且 GAN 的训练过程满足 (ε_0,δ_0)-差分隐私，根据引理 3 可知，算法 2 满足 (ε_0,δ_0)-差分隐私。

证毕。

附录 2　定理 2 的证明

算法 3 中对每个样本类别 $k\in[L]$，客户端生成的虚拟标签数目为 $\hat{n}_k$，$\hat{n}_k$ 的取值概率 $\Pr[\hat{n}_k=r]$ 正比于 $\exp\dfrac{\varepsilon u_k(D,r)}{(2\Delta u_k)}$，其中效用函数 u_k 定义为 $u_k(D,r)=-\left|\dfrac{r}{\hat{n}}-\dfrac{n_k}{n}\right|$。

将数据集 D 中任一样本替换为另一条样本，得到相邻数据集 D'，记基于 D' 生成类别 k 的标签数目为 $\hat{n}'_k$，易见在 2 个相邻数据集上，u_k 最多产生的变化为 $\dfrac{1}{n}$，即敏感度 $\Delta u_k=\dfrac{1}{n}$。约束 $\hat{n}_k$，$\hat{n}'_k$ 有相同取值时概率差异为

$$\frac{\Pr[\hat{n}_k=r]}{\Pr[\hat{n}'_k=r]}=\frac{\exp\left(\dfrac{\varepsilon u_k(D,r)}{2\Delta u_k}\right)\sum\limits_{r'\in[\hat{n}]}\exp\left(\dfrac{\varepsilon u_k(D',r')}{2\Delta u_k}\right)}{\exp\left(\dfrac{\varepsilon u_k(D',r)}{2\Delta u_k}\right)\sum\limits_{r'\in[\hat{n}]}\exp\left(\dfrac{\varepsilon u_k(D,r')}{2\Delta u_k}\right)}\leqslant$$

$$\exp\left(\frac{\varepsilon}{2}\right)\frac{\sum\limits_{r'\in[\hat{n}]}\exp\left(\dfrac{\varepsilon\left(u_k(D,r')+\Delta u_k\right)}{2\Delta u_k}\right)}{\sum\limits_{r'\in[\hat{n}]}\exp\left(\dfrac{\varepsilon u_k(D,r')}{2\Delta u_k}\right)}=$$

$$\exp\left(\frac{\varepsilon}{2}\right)\exp\left(\frac{\varepsilon}{2}\right)\cdot 1=\exp(\varepsilon)$$

由此可知输出一个类别的标签数目满足 $(\varepsilon,0)$-差分隐私，由差分隐私组合性质可知，输出所有 L 个类别的标签数目满足 $(L\varepsilon,0)$-差分隐私。因为生成的虚拟标签是相互独立的，不存在先后次序关系，所以 L 个类别的标签数目唯一决定了所有的虚拟标签 $\hat{y}_1,\cdots,\hat{y}_m$，令 $\varepsilon'=\dfrac{\varepsilon}{L}$，至此证明了算法 3 满足 $(\varepsilon',0)$-差分隐私。

证毕。

参考文献：

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

[2] KONEČNÝ J, MCMAHAN H B, RAMAGE D, et al. Federated optimization: distributed machine learning for on-device intelligence[J]. arXiv Preprint, arXiv: 1610.02527, 2016.

[3] ZHAO Y, LI M, LAI L Z, et al. Federated learning with non-IID data[J]. arXiv Preprint, arXiv: 1806.00582, 2018.

[4] XU C C, HONG Z W, HUANG M L, et al. Acceleration of federated learning with alleviated forgetting in local training[J]. arXiv Preprint, arXiv: 2203.02645, 2022.

[5] LI Q B, DIAO Y Q, CHEN Q, et al. Federated learning on non-IID

data silos: an experimental study[J]. arXiv Preprint, arXiv: 2102.02079, 2021

[6] LI T, SAHU A K, ZAHEER M, et al. Federated optimization in heterogeneous networks[J]. arXiv Preprint, arXiv: 1812.06127, 2018.

[7] KARIMIREDDY S P, KALE S, MOHRI M, et al. Scaffold: stochastic controlled averaging for federated learning[C]//International Conference on Machine Learning. New York: PMLR, 2020: 5132-5143.

[8] LUO M, CHEN F, HU D P, et al. No fear of heterogeneity: classifier calibration for federated learning with non-IID data[C]//Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Massachusetts: MIT Press, 2021: 1-13.

[9] WANG J Y, LIU Q H, LIANG H, et al. Tackling the objective in-consistency problem in heterogeneous federated optimization[C]// Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Massachusetts: MIT Press, 2020: 1-13.

[10] HSU T M H, QI H, BROWN M. Measuring the effects of non-identical data distribution for federated visual classifica-tion[J]. arXiv Preprint, arXiv: 1909.06335, 2019.

[11] LIN T, KONG L J, STICH S U, et al. Ensemble distillation for robust model fusion in federated learning[J]. arXiv Preprint, arXiv: 2006.07242, 2020.

[12] GOETZ J, TEWARI A. Federated learning via synthetic data[J]. arXiv Preprint, arXiv: 2008.04489, 2020.

[13] HAO W T, EL-KHAMY M, LEE J, et al. Towards fair federated learning with zero-shot data augmentation[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2021: 3305-3314.

[14] YOON T, SHIN S, HWANG S J, et al. FedMix: approximation of mixup under mean augmented federated learning[J]. arXiv Preprint, arXiv: 2107.00233, 2021.

[15] FALLAH A, MOKHTARI A, OZDAGLAR A. Personalized federated learning: a meta-learning approach[J]. arXiv Preprint, arXiv: 2002.07948, 2020.

[16] SMITH V, CHIANG C K, SANJABI M, et al. Federated multi-task learning[C]//Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). Massachusetts: MIT Press, 2017: 1-11.

[17] GHOSH A, CHUNG J, YIN D, et al. An efficient framework for clustered federated learning[J]. IEEE Transactions on Information Theory, 2022, 68(12): 8076-8091.

[18] WAHEED A, GOYAL M, GUPTA D, et al. CovidGAN: data augmentation using auxiliary classifier GAN for improved COVID-19 detection[J]. IEEE Access, 2020, 8: 91916-91923.

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

[20] CHEN D F, YU N, ZHANG Y, et al. GAN-leaks: a taxonomy of membership inference attacks against generative models[C]//Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2020: 343-362.

[21] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th Conference on Neural Information Processing Systems (NeurIPS 2014). Massachusetts: MIT Press, 2014: 2672-2680.

[22] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. arXiv Preprint, arXiv: 1411.1784, 2014.

[23] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv Preprint, arXiv: 1511.06434, 2015.

[24] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]//International Conference on Machine Learning. New York: PMLR, 2017: 214-223.

[25] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of Cryptography Conference. Berlin: Springer, 2006: 265-284.

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

[27] 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. New York: ACM Press, 2016: 308-318.

[28] XIE L Y, LIN K X, WANG S, et al. Differentially private generative adversarial network[J]. arXiv Preprint, arXiv: 1802.06739, 2018.

[29] ZHANG X Y, JI S L, WANG T. Differentially private releasing via deep generative model[J]. arXiv Preprint, arXiv: 1801.01594, 2018.

[30] DAVODY A, ADELANI D I, KLEINBAUER T, et al. Robust differentially private training of deep neural networks[J]. arXiv Preprint, arXiv: 2006.10919, 2020.

[31] YOUSEFPOUR A, SHILOV I, SABLAYROLLES A, et al. Opacus: user-friendly differential privacy library in PyTorch[J]. arXiv Preprint, arXiv: 2109.12298, 2021.

[32] ZENG D, LIANG S Q, HU X J, et al. FedLab: a flexible federated learning framework[J]. arXiv Preprint, arXiv: 2107.11621, 2021.

[33] 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.

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

[35] KRIZHEVSKY A. Learning multiple layers of features from tiny images[D]. Toronto: University of Toronto, 2009.

[36] NETZER Y, WANG T, COATES A, et al. Reading digits in natural images with unsupervised feature learning[C]//Proceedings of the 25th Conference on Neural Information Processing Systems (NIPS 2011). Massachusetts: MIT Press, 2011: 1-9.

[37] MIRONOV I, TALWAR K, ZHANG L. Renyi differential privacy of the sampled gaussian mechanism[J]. arXiv Preprint, arXiv: 1908.10530, 2019.

[作者简介]

汤凌韬（1994− ），男，江苏启东人，数学工程与先进计算国家重点实验室博士生，主要研究方向为信息安全、机器学习和隐私保护等。

王迪（1993− ），女，江苏徐州人，数学工程与先进计算国家重点实验室硕士生，主要研究方向为人工智能芯片设计。

刘盛云（1985− ），男，云南昆明人，博士，上海交通大学助理教授，主要研究方向为区块链、联邦学习、分布式存储等。