计算机科学
COMPUTER SCIENCE

一种非独立同分布问题下的联邦数据增强算法

瞿祥谋，吴映波，蒋晓玲

引用本文

瞿祥谋，吴映波，蒋晓玲.一种非独立同分布问题下的联邦数据增强算法[J]．计算机科学, 2022, 49(12): 33-39.

QU Xiang-mou, WU Ying-bo, JIANG Xiao-ling. Federated Data Augmentation Algorithm for Non-independent and Identical Distributed Data [J]. Computer Science, 2022, 49(12): 33-39.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

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

基于TPH-YOLOv5和小样本学习的害虫识别方法

Pest Identification Method Based on TPH-YOLOv5 Algorithm and Small Sample Learning

计算机科学, 2022, 49(12): 257-263. https://doi.org/10.11896/jsjkx.221000203

基于联邦学习的暖通空调系统故障检测与诊断

Fault Detection and Diagnosis of HVAC System Based on Federated Learning

计算机科学, 2022, 49(12): 74-80. https://doi.org/10.11896/jsjkx.220700280

基于联邦学习的Gamma回归算法

FL-GRM:Gamma Regression Algorithm Based on Federated Learning

计算机科学, 2022, 49(12): 66-73. https://doi.org/10.11896/jsjkx.220600034

基于联邦学习的车联网多维资源动态分配算法

Multi-dimensional Resource Dynamic Allocation Algorithm for Internet of Vehicles Based on Federated Learning

计算机科学, 2022, 49(12): 59-65. https://doi.org/10.11896/jsjkx.211000123

边缘场景下动态权重的联邦学习优化方法

Federated Learning Optimization Method for Dynamic Weights in Edge Scenarios

计算机科学, 2022, 49(12): 53-58. https://doi.org/10.11896/jsjkx.220700136

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220300031

一种非独立同分布问题下的联邦数据增强算法

瞿祥谋　吴映波　蒋晓玲
重庆大学大数据与软件学院　重庆 401331
(201924021004@cqu.edu.cn)

摘　要　在联邦学习中,由于用户的本地数据分布会随着用户所在地以及用户偏好而变动,数据的非独立同分布下的用户数据可能缺少某些标签类别的数据,在模型聚合中显著影响了迭代更新速率和最终的模型性能。为了解决这一问题,提出了一种基于条件生成对抗网络进行联邦数据增强的算法,能够在不涉及泄露用户隐私的前提下,通过生成对抗网络模型对数据偏斜的参与者扩增少量数据,大幅提升非独立同分布数据划分下联邦学习算法的性能。实验结果表明,与当前主流的联邦算法相比,该算法在非独立同分布设置下的 MNIST,CIFAR-10 数据集上的预测精度分别提升了 1.18%和 14.6%,显示出了该算法对非独立同分布问题的有效性和实用性。

关键词:联邦学习;隐私保护;生成对抗网络;差分隐私;数据增强
中图法分类号　TP391

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

QU Xiang-mou,WU Ying-bo and JIANG Xiao-ling
School of Big Data & Software Engineering,Chongqing University,Chongqing 401331,China

Abstract　In federated learning,the local data distribution of users changes with the location and preferences of users,the data under the non-independent and identical distributed(Non-IID) data may lack data of some label categories,which significantly affects the update rate and the performance of the global model in federated aggregation. To solve this problem,a federated data augmentation based on conditional generative adversarial network(FDA-cGAN) algorithm is proposed,which can amplify data from participants with skewed data without compromising user privacy,and greatly improve the performance of the algorithm with Non-IID data. Experimental results show that,compared with the current mainstream federated average algorithm,under the Non-IID data setting,the prediction accuracy of MNIST and CIFAR-10 data sets improves by 1.18% and 14.6%,respectively,which demonstrates the effectiveness and practicability of the proposed algorithm for Non-IID data problems in federated learning.

Keywords　Federated learning,Privacy-preserving,Generative adversarial network,Differential Privacy,Data augmentation

1　引言

随着各行各业数据平台的建设,大数据驱动的机器学习技术推动现实社会逐步走向信息化和智能化,例如计算机视觉、自然语言处理以及推荐系统等。但由于用户隐私数据泄露以及行业竞争私自使用隐私数据等问题的出现,政府相继出台法律法规限制隐私数据的流通与共享,企业间形成了“数据孤岛”现象,协同隐私数据联合建模成为挑战。

谷歌的 McMahan 等提出了一种新型分布式机器学习范式——联邦学习[1]。在联邦学习中,两个或以上的联邦学习参与方协作构建一个共享的机器学习模型,任意参与方都拥有若干能够用来训练模型的训练数据,参与方拥有的数据都不会离开该参与方,参与方间通过加密传输模型训练参数,共同协作构建一个共享的机器学习模型。谷歌将联邦学习用于智能手机上的语言预测模型 Gboard 更新[2-5],能够在保护用户隐私的同时使得大量设备协同训练机器学习模型,并使参与者从中获益。

由于联邦学习中参与训练的设备归属于某个用户或企业,因此其数据分布及数据集规模往往存在极大差异,由此导致了联邦学习中的非独立同分布(Non-independent and Identical Distributed,Non-IID) 数据问题[6-9]。非独立同分布的数据问题在真实场景中广泛存在,而现有的联邦学习算法主要基于数据服从独立同分布的假设。为了验证非独立同分布对联邦学习性能的影响,Hsu 等构建了符合真实数据分布的联

到稿日期:2022-03-02　返修日期:2022-06-13
基金项目:国家重点研发计划(2019YFB1706101);重庆市技术创新与应用发展专项重点项目(cstc2019jscx-mbdxX0047);中央高校基本业务费项目(2020CDCGRJ50)
This work was supported by the National Key R&D Program of China(2019YFB1706101),Science-Technology Foundation of Chongqing (cstc2019jscx-mbdxX0047) and Fundamental Research Funds for the Central Universities(2020CDCGRJ50).
通信作者:吴映波(wyb@cqu.edu.cn)

邦学习数据集并进行了实验[10]，实验结果显示模型的性能随着用户数据分布差异的增大而下降；Li 等提出了 FedProx 算法[11]，用于对联邦平均算法 FedAvg 进行重参数化，并给出了理论收敛性的证明；Wang 等提出了 FedNova 算法[12]，消除了优化目标不一致带来的精度损失，并通过归一化平均保持了误差的快速收敛。当前主流的方法是根据用户数据量和数据分布的差异，在模型聚合中动态调节参数权重，削弱非独立同分布数据问题带来的性能影响。然而，这些方法需要依赖每轮参与训练的用户的数据分布情况，针对不同任务需重新调节权重参数，导致后续训练得到的模型泛化能力弱，模型参数优化难度大。

针对这一问题，本文在现有联邦学习算法的基础上进行了一系列的研究，提出了基于生成对抗网络的联邦学习数据增强方法，在不涉及泄露用户隐私的前提下，通过生成对抗网络模型对数据偏斜的参与者扩增数据，使得不同用户数据分布对齐。在 MNIST 和 CIFAR-10 数据集上进行了实验，结果显示该算法提升了非独立同分布数据划分下联邦学习算法的性能。

本文第 2 节介绍了相关工作；第 3 节给出了问题的定义及 FDA-cGAN 算法的细节；第 4 节进行了实验验证；最后总结全文并展望未来。

2 相关工作

联邦学习中，用户间的隐私数据具有分散和非独立同分布的特性，每个用户的本地数据分布会随着用户所在地以及用户偏好的变化而变动，其本地模型差异显著，在模型聚合中极大地影响了迭代更新速率和最终模型的性能。因此，许多研究针对非独立同分布数据问题开展了广泛的实验和优化，并提出了一系列可行的联邦学习算法。

2.1 非独立同分布问题的定义

为了阐明联邦学习非独立同分布的问题，本文给出了非独立同分布场景的定义：用户 i 的本地数据分布 $(x,y)\sim P_i(x,y)$ 与用户 j 的本地数据分布 $(x,y)\sim P_j(x,y)$ 之间存在差异，在实际场景中按照具体差异的不同，将其分为特征分布的差异与标签分布的差异。具体地，将每个用户的差异表述为特征分布 $P(x)$ 与标签分布 $P(y|x)$ 的乘积：

$$P_i(y_i|x_i)\cdot P(x_i)\neq P_j(y_j|x_j)\cdot P(x_j) \tag{1}$$

假设在联邦学习建模中存在 K 个可参与用户，使用 n^K 表示用户 k 的数据量，D_k 表示其本地数据分布，将用户 i 的第 k 条数据表述为：

$$\begin{aligned} &z^{i,k}=(x^{i,k},y^{i,k})\\ &\text{s.t.}\quad \forall i\in n^k, y^{i,k}\in D_k \end{aligned} \tag{2}$$

本文将针对数据标签非独立同分布[13]情况下的联邦学习算法进行探讨，数据标签非独立同分布情况指不同用户间数据分布的条件概率 $P_i(y|x)$ 相同，但其边缘概率 $P(y)$ 不同。例如，由于个人喜好的差异，来自不同用户的相同特征向量却有不同的标签类别。例如，大熊猫多分布在中国的动物园内，而袋鼠和树袋熊等动物分布在澳大利亚境内，标签的分布随着环境或文化的影响导致了差异的产生。后文中提到的数据非独立同分布，若非特别说明，皆指数据标签分布不平衡情况下的非独立同分布数据。

2.2 针对非独立同分布问题的联邦学习算法

大量研究表明，联邦学习对于非独立同分布数据导致的精度下降问题几乎不可避免[14-16]。由于用户局部数据分布的差异，让具有相同初始参数的局部模型收敛到不同的模型权重，通过对上传的本地模型进行平均得到的共享全局模型与理想模型之间的偏差不断增大，从而减缓了收敛速度，降低了学习的性能。

将不同用户的建模目标看作多个子任务，将联邦学习问题转化为多任务学习，由于用户数据和任务之间具有相似性，模型利用任务之间的共性和差异来共同学习。通过共享相关联任务的表示，能够提升原始任务的泛化性能[17-19]。Collins 等通过判别任务之间的相关性来决定应该共享的部分[20]，从而对模型的泛化能力进行改善。

通过少量数据以及弱监督的联邦迁移学习构建的联邦学习模型也能作为非独立同分布问题的解决方式[21-23]。在联邦学习中，每个设备节点上的数据都是以非独立同分布的方式生成的。在这些情况下，将从标记的源域数据中学习到的知识迁移到未标记的目标域中，通过优化相似任务中的模型，来建立针对目标域的高性能机器学习模型[24-26]。但在联邦学习中，在对隐私数据的建模任务样本进行选择的过程中容易出现数据泄露问题[27-28]，因此保证所有参与方共享数据表征信息时不泄露本地数据信息，捕捉参与方之间的特征空间不变性，成为了联邦迁移学习面临的重大挑战。

上述方法旨在通过对用户本地模型进行训练及对联邦全局模型聚合阶段进行优化，通过有效地提取用户间子任务的关系特征，来求解全局优化目标。因此，本文将注意力转向数据分布本身，通过优化用户的数据分布特征，来控制模型权重的一致性，提高联邦学习训练产生的最终模型结果。

3 研究方法

本文尝试从数据分布的角度去缓解联邦学习中非独立同分布数据造成的过拟合现象。本文尝试利用条件生成对抗网络去调整该类别在用户本地的数据分布，通过增强用户的本地数据，来扩增用户缺少的数据，可以使得用户的本地数据类别变得均衡。如果该分布估计足够准确，则可以减小联邦学习和传统机器学习训练的差距。

在联邦学习任务中，非独立同分布下的用户数据可能缺少某些标签类别的数据，如用户由于喜好或使用习惯，总是偏向于产生某一类标签的数据，而基本没有产生其他标签类别的数据，在偏斜的数据分布上训练模型会导致严重的过拟合现象，同时破坏模型的泛化能力，使得用户的本地训练模型更加偏向于判别该类别的数据，而对其他类别的标签数据的识别率较低。

基于此问题定义，本文提出了一种基于条件生成对抗网络的联邦数据增强算法(Federated Data Augmentation Algorithm with Conditional Generative Adversarial Network，FDA-cGAN)。FDA-cGAN 的具体训练过程如图 1 所示，该算法分为两阶段：数据增强阶段及联邦模型训练阶段。在数据增强阶段，每个客户端协作地将部分原始数据发送到中心用于训

练生成模型，将该网络分发给参与用户，生成与真实数据相似的数据来修正本地数据分布，从而实现用户本地数据增强以及缓解非独立同分布问题造成的影响。在联邦模型训练阶段，采用了一种通信效率高的分层选择策略来减少参与者的数量，所有参与者通过聚类算法被分配到不同集群中心的集合中，通过对联邦学习任务选择有代表性的参与者来减小通信成本，最终将每个客户端训练的本地模型上传到服务器，并将本地模型集聚合得到全局模型。模型训练阶段算法的过程如算法 1 所示。

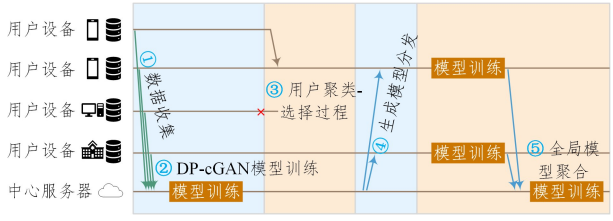


图 1 基于 FDA-cGAN 的非独立同分布数据增强技术

Fig. 1 Data augmentation technology for Non IID Data based on FDA-cGAN

算法 1 模型训练阶段算法

输入：用户的本地数据 $D=[D_1,\cdots,D_K]$；第 i 位用户的数据 $D_i=[(x_1^i,y_1^i),\cdots,(x_n^i,y_n^i)]$

输出：模型参数文件 θ

/* 数据增强步骤 */

1. 收集用户提供的可用于训练的数据 D_{train}；
2. 根据 D_{train} 训练条件生成对抗网络 θ_g；
3. 计算用户数据分布 D 之间的相似度并聚类；
4. 生成对抗网络 θ_g，在用户设备端生成数据集，使得用户间的数据分布趋于一致；

/* 联邦训练步骤 */

5. 初始化全局模型参数 θ_0
6. For 通信轮次 in (1,N) do：
7. 计算每轮参与的用户数量 $m\leftarrow\max(C\cdot K,1)$；
8. 从用户集合挑选 m 位参与训练的用户；
9. 根据用户的本地数据计算梯度；
10. 更新模型参数；
11. 保存最终训练的模型参数并下发给用户

3.1 基于差分隐私的数据增强阶段

生成对抗网络[29]（Generative Adversarial Network，GAN）是一种利用零和博弈思想的对抗学习方式，其具体结构为两个网络，即生成网络 G 和判别网络 D，通过判别函数 $D(x)$：$\mathbb{R}^n\rightarrow[0,1]$ 和生成函数 $G(x)$：$\mathbb{R}^d\rightarrow\mathbb{R}^n$ 之间的目标函数的极大值和极小值来实现。生成器随机生成噪声数据 $z\in\mathbb{R}^d$ 并转化为生成样本 $G(z)$，判别器试图将生成器产生的生成样本与来自分布的训练真实样本区分开，而生成器试图使生成的样本在分布上与训练样本相似，通过生成器与判别器的对抗来协同训练。对抗的目标损失函数如下：

$$V(D,G):=\mathbb{E}_{x\sim\mu}[\log D(x)]+\mathbb{E}_{z\sim\gamma}[\log(1-D(G(z)))] \tag{3}$$

其中，$\mathbb{E}$ 表示关于下标中指定分布的期望值，D 表示判别器网络，G 表示生成器网络。GAN 解决的极小值和极大值的描述如下：

$$\min_G\max_D \mathbb{E}_{x\sim\mu}[\log D(x)]+\mathbb{E}_{z\sim\gamma}[\log(1-D(G(z)))] \tag{4}$$

该模型通过最小化真实数据分布与生成模型分布之间的 KL 散度（Kullback-Leibler Divergence），使得生成模型尽可能接近真实数据的分布。在实际实践中，难以直接获取真实数据分布的 $P_{data}(x)$，我们使用训练数据形成的经验分布来逼近 $P_{data}(x)$，从而实现生成样本与训练样本相似的特征分布。

生成对抗网络生成的虚假数据，能够有效规避用户的隐私信息，而且生成对抗网络可以生成大量多样的样本数据，可以使用户的本地增强数据量大于所维护的公共数据集。

为了保证在生成对抗网络训练过程中不泄露用户数据，本文采用基于差分隐私方式的条件生成对抗性网络模型（Differential Privacy Conditional Generative Adversarial Network，DP-cGAN）。差分隐私最先由微软研究院的 DWork[30] 提出，用于解决数据库的隐私泄漏问题。相比原本被广泛用于解决数据隐私保护问题的语义匿名模型（k-匿名[31] 等）。差分隐私弥补了以前语义匿名模型由于理想化的假设而带来的隐私保护问题，通过控制单条数据对整个数据分析结果的影响，在模型层面实现了对差分攻击的不可分辨性。

深度学习训练机制将训练数据集作为输入，采用梯度下降法进行训练，输出训练后的参数，记作 $\mathcal{M}:\mathcal{D}\rightarrow\mathcal{T}$。如果对于任意两个毗邻的训练集 $d,d'\in\mathcal{D}$，以及任何参数范围 $S\subset\mathcal{T}$，其输出的参数分布满足：

$$\Pr[\mathcal{M}(d)\in S]\leqslant e^{\varepsilon}\Pr[\mathcal{M}(d')\in S]+\delta \tag{5}$$

在该定义下，数据集中的单个数据对模型的影响被控制在一定的范围内，实现了在模型层面对差分攻击的“不可分辨性”。

在深度学习训练中，模型参数 w 采用梯度下降法进行更新。

$$w_{t+1}\leftarrow w_t-\frac{\eta_t}{N}\cdot\nabla_{\theta_t}f(x;w_t) \tag{6}$$

本文采用由 Facebook 团队研发的基于 PyTorch 的深度学习差分隐私框架 Opacus，通过在模型优化器中引入差分隐私预算，来为模型提供可靠的差分隐私保障。该框架通过在梯度上使用差分隐私来控制个体数据对整体模型的影响，其差分隐私噪声使用高斯机制进行。其具体定义如下：假设存在一个函数 $f:\mathcal{D}\rightarrow\mathcal{T}$，敏感度为 $\Delta f=\max_{D,D'}\|f(D)-f(D')\|_2$，那么对于任意的 $\delta\in(0,1)$，给定随机噪声服从正态分布 $\mathcal{N}(0,\sigma^2)$，那么随机算法 $\mathcal{M}(d)=f(d)+\mathcal{N}(0,\sigma^2)$ 服从 (ε,δ)-差分隐私，即：

$$\mathcal{M}(d)=f(d)+\mathcal{N}(0,\sigma^2) \tag{7}$$

利用高斯机制对梯度分 3 个步骤增加差分噪声。首先，将每一个样本对应的梯度裁剪到一个固定范围 $[-C,C]$，以控制个体数据的影响，此时梯度的敏感度为：

$$\Delta_2(f)=\max_{x_i\in D}\|\nabla_\theta f(x;w)\|_2\leqslant C \tag{8}$$

然后，对裁剪后的梯度增加高斯噪声 $\mathcal{N}(0,\sigma^2)$，以得到满足差分隐私的梯度数据。最后，用这些梯度更新模型，并计算模型的隐私损失。记噪声乘子（Noise Multiplier）为 $z=\frac{\sigma}{C}$，那么该训练系统服从 (ε,δ)-差分隐私的条件为：

$$z=\frac{\sqrt{2\ln(1.25/\delta)}}{\varepsilon} \tag{9}$$

通过阈值设定的方式保证模型从任何给定样本中无法得到更多的信息，在使用高斯机制在梯度中增加噪声后，使得深度学习过程服从(ε,δ)-差分隐私，从而保证了数据的隐私性。DP-cGAN 模型训练的过程如算法 2 所示。

算法 2 DP-cGAN 模型训练

输入：用户的可用本地数据 $D=[D_1,\cdots,D_K]$

输出：模型参数文件θ_g

1. 初始化 DP-cGAN 模型参数θ_g^0；
2. For 通信轮次 t in(1,N) do：
3. For 训练用户 i in(1,m) do：
4. 根据用户的本地数据计算梯度；
5. 根据差分隐私添加本地噪声；
6. 聚合所有模型梯度；
7. 保存最终训练的模型参数θ_g并下发给用户。

基于此，该训练过程能够在不暴露任意用户信息的情况下实现对 DP-cGAN 模型的训练，并将其分发给参与训练的用户进行数据增强。服务器端训练基于差分隐私技术的 DP-cGAN，具体流程如图 2 所示。

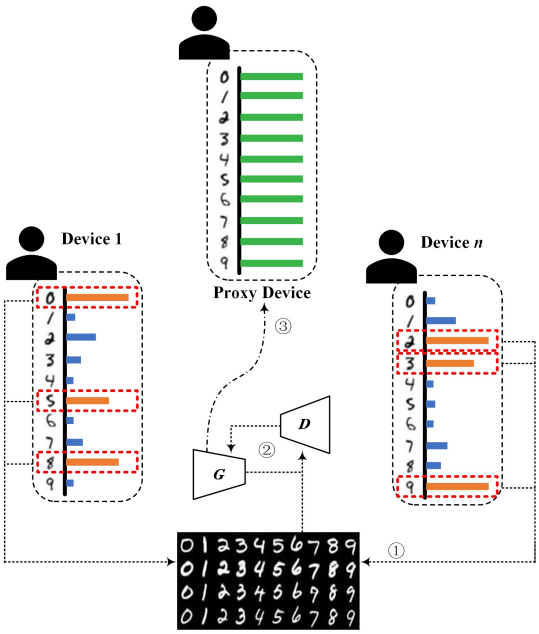


图 2 基于 DP-cGAN 的数据增强过程

Fig. 2 Data augmentation process based on DP-cGAN

在提出的 DP-cGAN 网络的训练过程中，将随机梯度下降算法与差分隐私机制相结合，为了保护这些原始训练数据不被暴露，在训练迭代中加入了高斯噪声的梯度扰动，即：

$$w^{t+1}\leftarrow w^t-\alpha\cdot g_t+\beta\cdot\mathcal{N}(0,\sigma^2) \tag{10}$$

通过聚合不同用户训练的本地模型后，即得到了用于本地数据增强的 DP-cGAN 模型。

3.2 联邦学习模型训练阶段

在联邦学习阶段采用随机梯度下降(Stochastic Gradient Descent，SGD)算法，对数据按批大小(Batch-Size)采样后对模型进行更新，具体应用流程包括如下 4 个步骤。

(1)用户选择和模型聚类。在一轮模型更新过程中，并非所有设备都会参与本轮联邦学习更新，服务器会从一组符合训练要求的客户端设备集合 K 中按照随机概率 C 进行用户采样，参数 C 用于控制全局用户数量，当 $C=1$ 时即代表全部用户参与训练。

$$\{S_t\in K\mid t=\max\{C\cdot K,1\}\} \tag{11}$$

参与者在每轮通信中更新训练模型的参数并下载全局模型，这会导致大量的通信开销[32-34]，为此采用一种分层聚类方法，根据所有客户端的本地标签分布的相似性对其进行聚类后再进行用户筛选。

$$\begin{aligned}&D_i=[d_1,d_2,\cdots,d_n]\\&S_i\leftarrow\mathrm{softmax}(D_i)\end{aligned} \tag{12}$$

其中，D_i代表用户 i 的本地数据分布情况，d_k代表其中第 k 类数据所占的百分比，通过计算其最大类别所占的比重，将其分到聚类中心S_i中，从每个聚类中心按照比例η_s抽取具有代表性的用户参与联邦训练，在不影响样本代表性的情况下减小模型的通信成本。

(2)模型广播。服务器初始化训练模型参数，配置全局聚合机制，参与本轮训练的客户端从服务器下载当前模型的结构和权重。

$$w_i^k=w^k,\forall\, i\in S_t,k\in[0,E] \tag{13}$$

(3)本地计算更新。下载到本地的训练模型根据本地数据进行本地训练，每轮训练获取本轮用户，具体方式如下：

$$f_i(w)=l(x_i,y_i;w) \tag{14}$$

通过 t 时刻的模型参数 w^t，给定该时刻输入x_i的预测值y_i，通过真实标签 y 计算出 loss 值与梯度信息。

$$g_t=\nabla f(x_i,y,w^t) \tag{15}$$

本地模型根据给定的学习率 α 及梯度信息更新第 i 个客户端的本地参数。

$$w^{t+1}\leftarrow w^t-\alpha\cdot\nabla f(x_i,y;w^t) \tag{16}$$

(4)全局聚合更新。服务器收集客户端的更新信息后，将其聚合成全局信息。

$$w_{t+1}^k\leftarrow w_t-\eta\cdot g_k,\forall\, k\in S_t \tag{17}$$

$$w_{t+1}\leftarrow\sum_{k=1}^{K}\frac{n_k}{n}w_{t+1}^k \tag{18}$$

经由此方式可得到全局联邦训练后的模型参数。基于数据增强的联邦学习，在每个客户端训练本地模型并将其上传到服务器，并将本地模型聚合到联邦全局模型中。

4 实验验证

4.1 非独立同分布数据集划分策略

现实世界的数据在不同的场景中更具随机性。第 3 节验证了非独立同分布数据对联邦学习训练的影响。为了探索基于条件生成对抗网络的联邦数据增强算法对缓解非独立同分布数据的效果，真实地模拟真实世界的标签偏态分布，本文在 MNIST 和 CIFAR-10 数据集上给出了独立同分布和非独立同分布的数据划分策略。

独立同分布数据集划分设置如下：在独立同分布数据集中，每个参与者的标签分布应该是相同的，并且拥有相同的数据量。模拟由 100 个客户端组成的独立同分布数据集，将所有样本随机打乱，对其均匀抽样出相同数据量的数据集合，并

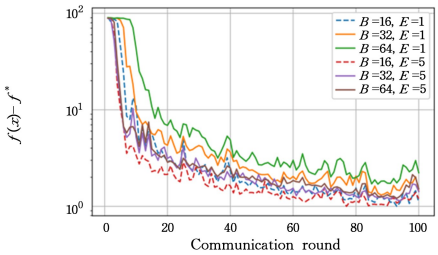
将其作为本地数据划分到参与者。

非独立同分布数据集划分设置如下:首先介绍非独立同分布数据集的划分标准,将非独立同分布数据的差异程度定义为 ϕ。例如,$\phi=0.3$ 表示 30%的数据集样本被平均划分到每个参与者中,而其余 70%的数据以最小概率阈值随机划分;当 $\phi=1$ 时即表示其为均等划分;$\phi=0$ 时表示其为随机划分。

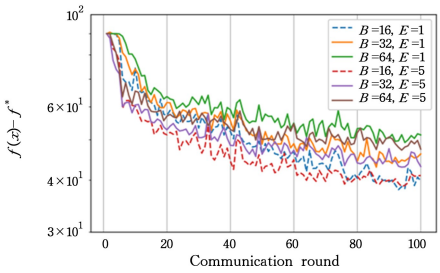
4.2 超参数搜索实验

考虑到非独立同分布数据问题对联邦学习中超参数的敏感性,通过实验搜索合适的超参数设置,并在后文中使用统一的超参数设置进行对比验证。设置局部迭代轮数 $E=[1,5]$,局部批处理数据大小 $B=[16,32,64]$,并在 MNIST 和 CIFAR-10 数据集上使用相同的学习率 $\eta=0.01$ 进行实验。

图 3 给出了局部迭代轮数和局部批数据大小对模型预测精度的影响。



(a)MNIST



(b)CIFAR-10

图 3 针对不同超参数的模型性能下降情况

Fig. 3 Performance decrease for different hyper-parameters

图 3 中,横坐标为全局通信轮数,纵坐标为模型性能下降程度。当 $B=16$ 和 $E=5$ 时,在 MNIST 和 CIFAR-10 数据集上都取得了最小的精度下降。该实验结果表明本地迭代轮数过少会导致模型训练不完全,本地迭代轮数过多会增大不同用户间的模型差异性,且随着局部批处理数据大小的增加,非独立同分布特性导致的模型性能下降更显著。因此需要在实验过程中设定固定的参数进行实验。。

实验中采用的参数设置如表 1 所列。

表 1 联邦学习参数设置

Table 1 Federatedlearning parameters setting

参数符号	描述	设定值
K	每轮参与用户数	100
C	参与用户百分数/%	10
E	本地迭代轮数	5
B	本地数据批大小	16
η	学习率	0.01
η_s	聚类时对用户的抽样比例	0.05
η_g	本地用户数据增强比例	0.2

为了验证非独立同分布数据对 MNIST,CIFAR-10 数据集的影响,本文在该实验参数设置下对传统的两层卷积神经网络(Convolutional Neural Network,CNN)以及 18 层的残差神经网络(Residual Network,ResNet-18)在 $\phi=0.2$ 的非独立同分布数据集以及不同迭代轮数设置下进行实验,测得其在不同设置的数据集下的模型预测精度,具体结果如表 2 和表 3 所列。

表 2 模型在 MNIST 数据集上不同迭代次数的精度

Table 2 Prediction accuracy under different communication rounds on MNIST dataset

模型	迭代轮数	MNIST/%	
		IID	Non-IID
ResNet	10	98.60	82.05
	50	99.16	98.81
	100	99.20	98.55
CNN	10	98.76	82.05
	50	97.89	94.77
	100	98.41	96.86

如表 2 所列,在 MNIST 数据集中,ResNet 与 CNN 模型的准确率在 Non-IID 数据设置条件下均有所下降。MNIST 数据集的训练任务中特征空间相对简单,在 IID 与 Non-IID 数据设置下的 ResNet 模型的预测精确度均能达到 98%,因此后文针对 MNIST 数据集的算法性能测试仅采用 2-layer CNN 网络模型进行验证。

表 3 模型在 CIFAR-10 数据集上不同迭代次数的精度

Table 3 Prediction accuracy under different communication rounds on CIFAR-10 dataset

模型	迭代轮数	CIFAR-10/%	
		IID	Non-IID
ResNet	10	61.62	27.90
	50	72.48	50.73
	100	74.07	62.38
CNN	10	45.33	28.19
	50	53.98	46.60
	100	54.69	47.94

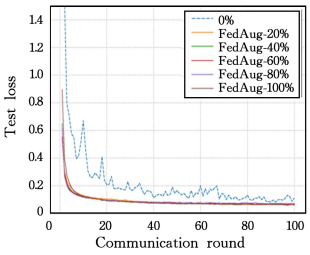
由表 3 可知,随着迭代轮数的增加,Non-IID 数据设置条件下的模型性能达到收敛后,模型性能差于同等条件的 IID 数据设置条件下的模型性能。在图像识别任务从黑白手写数字图片数据集 MNIST 转为彩色图片 CIFAR-10 数据集后,非独立同分布数据问题会造成更严重的模型性能下降。CIFAR-10 数据集的训练任务中特征空间相对复杂,在 IID 与 Non-IID 数据设置下的 2-Layer CNN 模型的预测精确度均难以达到 60%及以上的准确率,因此后文针对 CIFAR-10 数据集的算法性能测试仅采用 ResNet 网络模型进行验证。

4.3 实验过程

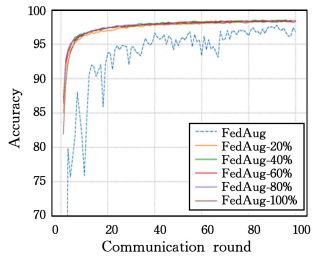
为了验证 FDA-cGAN 算法的数据增强效果,在相同的联邦设置下,将 FDA-cGAN 算法与联邦平均算法(FedAvg)进行比较。在 DP-cGAN 方法的训练过程中,设置 $\sigma=0.5$,$\delta=1.0\times10^{-5}$,训练迭代次数为 500。在 MNIST 数据集中,使用 10000 个样本的训练数据,每个类别有 1000 个样本,并应用 FDA-cGAN 模型生成 60000 个样本,并将这些样本分配给每个参与者,将其转换为独立同分布数据。

图 4 的实验结果显示,FDA-cGAN 算法在 MNIST 和 CI-

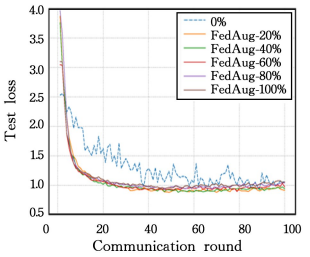
FAR-10 数据集上与传统的联邦平均算法相比有明显的改



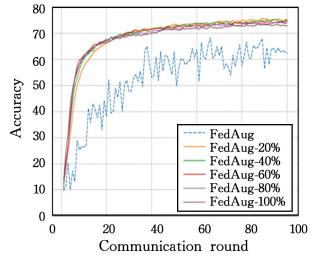
(a) 2-layer CNN 在 MNIST 数据集上的损失曲线



(b) 2-layer CNN 在 MNIST 数据集上的准确率曲线



(c) ResNet-18 在 CIFAR-10 数据集上的损失曲线



(d) ResNet-18 在 CIFAR-10 数据集上的准确率曲线

图 4 不同数据增强参数 n_g 下的模型性能

Fig. 4 Model performance with different data augmentation parameters n_g

为了验证 FDA-cGAN 算法提高图像分类任务中的预测精度不仅与样本量的增加有关，还与经数据增强后调整的数据分布有关，通过分离该算法的不同步骤(用户聚类、数据增强)进行消融实验，与经典的联邦平均模型相比，验证了在联邦学习中非独立同分布数据问题中使用生成对抗网络进行数据增强方法的有效性，其实验结果如表 4 和表 5 所列。

表 4 FDA-cGAN 算法在 MNIST 数据集上的准确率

Table 4 FDA-cGAN prediction accuracy on MNIST dataset

算法	MNIST/%	
	CNN	ResNet
FedAvg	97.05	98.80
FedAvg+cluster	97.74	99.32
FedAvg+cGAN	98.26	99.26
FDA-cGAN	98.46	99.98

表 5 FDA-cGAN 算法在 CIFAR-10 数据集上的准确率

Table 5 FDA-cGAN prediction accuracy on CIFAR-10 dataset

算法	CIFAR-10/%	
	CNN	ResNet
FedAvg	47.94	60.82
FedAvg+cluster	47.84	60.97
FedAvg+cGAN	54.02	74.84
FDA-cGAN	54.69	75.42

相比传统联邦学习算法，本文算法在数据分布及数据量两方面加强模型的训练效果，通过数据增强方法处理后，模型算法在 Non-IID 数据设置下取得了显著的性能提升，聚类算法可在降低训练过程中的通信成本的情况下不影响模型性能，该结论可从表 4 与表 5 的消融实验中得到验证。

此外，本文应用 ε，δ-差分隐私训练神经网络的方法是将高斯分布采样的噪声添加到梯度更新中，其中差分隐私的有效性取决于参数 ε，该参数被称为隐私预算。为了评估参数 ε 对生成模型的影响，使用开放资源库 Opacus 对具有不同隐私的模型进行训练，从而验证模型在不同隐私预算条件下的性能。通过设定剪裁阈值参数 $c=1$ 来控制最大梯度，设置隐私预算 $\varepsilon\in[0.5,1.0]$。

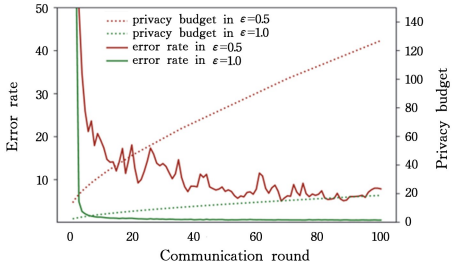


图 5 不同隐私预算下的模型错误率

Fig. 5 Mode lerror rate with different privacy budgets

如图 5 所示，当模型给定 $\varepsilon=1.0$ 时模型错误率能够快速降低，但其隐私预算随之增加，使得数据暴露的风险上升。因此，针对不同任务选择合适的隐私预算，在模型效用和隐私保障之间取得平衡。

结束语 本文提出的基于条件生成模型的联邦数据增强(FDA-cGAN)方法为缓解联邦学习中的非独立同分布问题提供了一种通用的解决方案，通过生成对抗网络进行数据增强，调整用户本地数据分布，来缓解非独立同分布数据带来的性能下降。但由于联邦学习需要高额的通信代价以及计算能力，因此在下一步工作中将考虑降低其通信成本，使用更高效的模型量化方式来降低模型大小，减轻设备通信和计算的压力。

参 考 文 献

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

[2] MCMAHAN H B, MOORE E, RAMAGE D, et al. Federated learning of deep networks using model averaging [J]. arXiv: 1602.05629, 2016.

[3] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: Concept and applications [J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.

[4] JAKUB K, MCMAHAN H B, YU F X, et al. Federated learning: Strategies for improving communication efficiency [J]. arXiv: 1610.05492, 2016.

[5] JAKUB K, MCMAHAN H B, DANIEL R, et al. Federated Optimization: Distributed Machine Learning for On-Device Intelligence [J]. arXiv: 1610.02527, 2016.

[6] ZHAO Y, LI M, SUDA N, et al. Federated learning with non-iid data [J]. arXiv: 1806.00582, 2018.

[7] BONAWITZ K, EICHNER H, GRIESKAMP W, et al. Towards federated learning at scale: System design [C]// Proceedings of Machine Learning and Systems. 2019, 1: 374-388.

[8] LARIMIREDDY S P, KALE S, MOHRI M, et al. SCAFFOLD: Stochastic Controlled Averaging for On-Device Federated Lear-

ning[C]// Proceedings of the International Conference on Machine Learning. PMLR,2020,119:5132-5143.
[9] LIX,HUANGK,YANGW,etal. Ontheconvergenceoffedavgonnon-iid data[J] arXiv:1907. 02189,2019.
[10] HSU T M H,QI H,BROWN M. Measuring the effects of non-identical data distribution for federated visual classification[J]. arXiv:1909. 06335,2019.
[11] LI T,SAHU A K,ZAHEER M,et al. Federated optimization in heterogeneous networks[C]// Proceedings of Machine Learning and Systems. 2020:429-450.
[12] WANG J,LIU Q,LIANG H,et al. Tackling the objective inconsistency problem in heterogeneous federated optimization[C]// Advances in Neural Information Processing Systems. 2020:7611-7623.
[13] KAIROUZ P,MCMAHAN H B,AVENT B,et al. Advances and openproblems in federated learning[J]. Foundations and Trends in Machine Learning,2021,14(1/2):1-210.
[14] SATTLER F,WIREDEMANN S,MULLER KR,et al. Robust and communication-efficient federated learning from non-iid data [J]. IEEE Transactions on Neural Networks and Learning Systems,2019,31(9):3400-3413.
[15] NISHIO T,YONETANI R. Client selection for federated learning with heterogeneous resources in mobile edge[C]// International Conference on Communications(ICC). IEEE,2019:1-7.
[16] WANG L,WANG W,LI B. CMFL:Mitigating communication overhead for federated learning[C]// International Conference on Distributed Computing Systems(ICDCS). IEEE,2019:954-964.
[17] SMITH V,CHIANG C K,SANJABI M,et al. Federated multi-task learning[C]// Advances in Neural Information Processing Systems. 2017.
[18] SATTLER F,MULLER K R,SAMEK W. Clustered federated learning:Model-agnostic distributed multitask optimization under privacy constraints[J]. IEEE Transactions on Neural Networks and Learning Systems,2020,32(8):3710-3722.
[19] LI R,MA F,JIANG W,et al. Online federated multitask learning[C]// International Conference on Big Data(Big Data). IEEE,2019:215-220.
[20] COLLINS L,HASSANI H,MOKHTARI A,et al. Exploiting shared representations for personalized federated learning[C]// International Conference on Machine Learning. PMLR,2021:2089-2099.
[21] PAN S J,YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering,2009,22(10):1345-1359.
[22] YANG H,HE H,ZHANG W,et al. FedSteg:A federated transfer learning framework for secure image steganalysis[J]. IEEE Transactions on Network Science and Engineering,2020,8(2):1084-1094.
[23] LIU Y,KANG Y,XING C,et al. A secure federated transfer learning framework[J]. IEEE Intelligent Systems,2020,35(4):70-82.
[24] XU M,LI X,WANG Y,et al. Privacy-preserving multisource transfer learning in intrusion detection system[J]. Transactions on Emerging Telecommunications Technologies,2021,32(5):e3957.
[25] JING Q,WANG W,ZHANG J,et al. Quantifying the performance of federated transfer learning[J]. arXiv:1912. 12795,2019.
[26] SHARMA S,XING C,LIU Y. Secure and efficient federated transfer learning[C]// International Conference on Big Data(Big Data). IEEE,2019:2569-2576.
[27] WANG Z,SONG M,ZHANG Z,et al. Beyond inferring class representatives:User-level privacy leakage from federated learning[C]// IEEE Conference on Computer Communications. IEEE,2019:2512-2520.
[28] SUN J,LI A,WANG B,et al. Soteria:Provable defense against privacy leakage in federated learning from representation perspective[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2021:9311-9319.
[29] GOODFELLOW I,POUGET-ABADIE J,MIRZA MEHDI,et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems. 2014.
[30] DWORK C. Differential privacy:A survey of results[C]// International Conference on Theory and Applications of Models of Computation. Berlin:Springer,2008:1-19.
[31] LIU J,YIN S,LI H,et al. A Density-based Clustering Method for K-anonymity Privacy Protection[J]. Journal of Information Hiding and Multimedia Signal Processing,2017,8(1):12-18.
[32] YANG Z,CHEN M,SAAD W,et al. Energy efficient federated learning over wireless communication networks [J]. IEEE Transactions on Wireless Communications,2020,20(3):1935-1949.
[33] HAMER J,MOHRI M,SURESH A T. Fedboost:A communication-efficient algorithm for federated learning[C]// International Conference on Machine Learning. PMLR,2020:3973-3983.
[34] WAHAB O A,MOURAD A,OTROK H,et al. Federated machine learning:Survey,multi-level classification,desirable criteria and future directions in communication and networking systems[J]. IEEE Communications Surveys & Tutorials,2021,23(2):1342-1397.

QU Xiang-mou, born in 1998, postgraduate, is a member of China Computer Federation. His main research interests include federated learning and data security.

WU Ying-bo, born in 1983, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include machine learning, intelligent optimization and decision.

(责任编辑:喻藜)