DOI:10.19651/j.cnki.emt.2209707

数据异质场景下的联邦学习模型校正与聚合

邹承明[1,2] 赵 宁[1]
(1.武汉理工大学计算机与人工智能学院 武汉 430079; 2.交通物联网技术湖北省重点实验室 武汉 430079)

摘 要:作为一种分布式机器学习范式,联邦学习在用户数据隐私保护方面拥有巨大潜力,是近年来的一大研究热点。首先,针对数据统计异质场景中普遍存在的用户模型偏差问题,提出了基于生成对抗网络的虚拟特征生成与分类层校正方案。其次,针对特殊的概念偏移场景,提出了基于分类层聚类的个性化分组聚合方案。最后,整合上述两种方案,并在图像分类数据集 CIFAR-10 上进行单项实验和集成实验。实验结果显示,相较于经典的联邦平均聚合算法,本文所提出的集成方案不仅显著提升了单中心全局模型的收敛速度,也增强了多中心簇模型的个性化能力。

关键词:联邦学习;数据异质;分类层校正;分类层聚类
中图分类号:TP389.1 **文献标识码**:A **国家标准学科分类代码**:520.20

Model correction and aggregation in statistically heterogeneous federated learning

Zou Chengming[1,2] Zhao Ning[1]
(1. Department of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430079, China;
2. Hubei Key Laboratory of Transportation Internet of Things, Wuhan 430079, China)

Abstract: As a promising distributed machine learning paradigm, Federated Learning brings huge privacy-preserving potentials, and has become a hot topic of research in recent years. To tackle client drift induced by statistically heterogeneous user data, this paper first presents an intermediate feature generation method based on Generative Adversarial Networks for the aim of classifier correction. Secondly, to deal with the particular problem of concept shift, a personalized model aggregation approach is proposed on the basis of classifier clustering. Finally, the two strategies mentioned above are integrated and tested on the CIFAR-10 image classification dataset. Various empirical results show that the proposed integrated strategy, compared to the classic Federated Averaging algorithm, helps realize both better generalization of the single-center global model, and better personalization of the multi-center cluster models.

Keywords: federated learning; statistical heterogeneity; classifier correction; classifier clustering

0 引 言

在大数据的新时代背景下,分布式机器学习应运而生。基于数据并行、模型并行和图并行的大规模机器学习方法,将计算任务从中心服务器转移到了各终端设备,解决了集中式学习中训练数据过多、模型规模过大、计算量过大等问题。然而,随着近些年一些互联网企业泄露用户信息以谋取私利的不法活动被媒体曝光,人们对个人隐私数据的保护意识与日俱增,与此同时,国内外有关数据管理规范的法规条例也在不断完善健全。在这样的监管环境下,如果数据所有者不与任何其他机构直接或间接共享数据,那么这些碎片化的数据将形成一座座“孤岛”。

不同于传统的分布式学习模式,联邦学习[1]的核心思想是:用户之间通过模型聚合而非数据共享的方式,享受数据隐式整合所带来的性能增益、有效缓解数据孤岛问题。由于用户数据自始至终不离本地,附加其他隐私保护技术的支持,这就响应了用户数据隐私保护的要求。

数据统计异质性是联邦学习面临的一大挑战,即众多参与训练的用户数据非独立同分布或数据量不一致。已有研究[2]表明,数据异质性会带来用户模型偏差,导致联邦平均算法(federated averaging, FedAvg)收敛减缓甚至发散,也间接造成了通讯开销的增大。因此,设计针对数据异质场景的高效模型聚合策略,对于提升联邦学习的整体性能来说是至关重要的,具有较高的研究价值,是本文的研究对象。

收稿日期:2022-04-22

为对抗数据异质性、减小用户模型偏差,Li 等[3]提出的 FedProx 改进了 FedAvg 算法,为用户本地损失函数添加了正则项,其本质是限制用户本地模型和全局模型之间的差异。然而,FedProx 相较于 FedAvg 的提升较为有限。Wang 等[4]指出,用户局部更新次数、优化算法、或是数据分布之间的差异,都将导致优化目标偏差问题、降低分布式优化算法的收敛率。对此作者提出 FedNova 方法,在更新全局模型前,将各用户本地模型的梯度归一化,纠正了优化目标偏差。针对 CIFAR-10 数据集,Zhu 等[5]提出的 FedOVA 将十类别分类器替换为了 10 个二分类器,每个二分类器负责输出图像属于某一个类别的机率。训练期间按照各个分类标签分别进行聚合,测试期间选得分最高的二分类器的类别作为最终结果。虽然 FedOVA 有效缓解了标签分布偏移的不利影响,它不适用于用户本地数据仅包含 1 种标签的情况。上述方法均无法对抗特殊的概念偏移问题。另外,由于深度神经网络中的神经元具有置换不变性[6-7],未经神经元匹配的 FedAvg 算法有损系统的收敛性能和公平性。针对这一问题,Wang 等[6]提出的 FedMA 在模型聚合之前,首先将相似的神经元结构逐层进行对齐;Yu 等[7]设计的 Fed^2 给出了另一种对齐方法:利用分组卷积层对特征进行分离,并通过最终的分组全连接层控制梯度的反向传播和特征的归属,聚合时按照特征所属分组进行组内加权平均。然而,FedMA 算法的实现较为复杂,而 Fed^2 更改了原神经网络的结构。Xie 等[8]引入了多中心联邦学习的概念,并采用联邦随机期望最大化算法 FeSEM 来求解新目标。本文沿用了多中心联邦学习的思想,且覆盖了包括概念偏移在内的所有异质场景,提出了较为完整、简易且高效的解决方案。

本文改进了 FedAvg 算法,利用生成对抗网络(generative adversarial network, GAN)来拟合神经网络提取的特征空间的分布,进而通过虚拟特征的聚合,达到纠正模型分类层偏差、加快单中心全局模型收敛速度的目的。类似得,Luo 等[9]同样基于虚拟特征,对分类层进行了校准,但其虚拟特征是通过拟合高斯混合模型在客户端生成的,而本文的虚拟特征是通过 GAN 在服务端生成的,一定程度上避免了虚拟特征的泄漏。另外,由于 FedAvg 算法产生的单中心全局模型无法调和概念偏移所致的用户模型偏差,本文提出了基于分类层参数聚类的分组聚合方案,将拥有相似分类层参数的用户归为同一簇,采用分组聚合的方式产生多中心簇模型。最后,在 CIFAR-10 数据集上进行实验,验证了提出方案的有效性。

1 相关概念

1.1 联邦平均聚合算法

FedAvg 聚合算法的全局优化目标可表示为:

$$\boldsymbol{w}^* \triangleq \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_{n=1}^{N} f_n(\boldsymbol{w}) \tag{1}$$

其中,$f_n(\boldsymbol{w})$ 代表用户 n 的本地损失函数,$\boldsymbol{w}$ 为单中心全局模型的参数,N 为用户数。

1.2 生成对抗网络

GAN 的常见应用之一是数据增强[10-12],其目的是利用生成器与判别器之间的竞争对抗训练,学习原样本空间中隐含的分布,从而合成新的样本用于扩充原样本空间、帮助模型提升泛化性能。由原始 GAN 衍生出诸多架构变种,其中 Mirza 等[13]提出了条件 GAN(conditional GANs, cGAN),在判别器与生成器的输入端,cGAN 附加了条件信息。此时的判别器不仅能够辨别输入样本的真假,还能够判断输入样本是否与条件信息匹配。另外,尽管 GAN 已有众多成熟的应用案例且潜能巨大,其仍面临着训练梯度消失、模式崩塌、训练过程不平稳等问题。为解决上述问题,WGAN-GP[14]利用 Wasserstein 距离取代了 J-S 散度,并利用梯度惩罚技术使得判别器的优化目标函数满足 1-Lipschitz 连续性。

本文将采用结合了 WGAN-GP 的 cGAN 模型,其中生成器的输入为噪声及标签信息,生成器的输出为虚拟中间层特征。用户的原始特征将作为真实数据、生成的虚拟特征作为假数据,输入判别器。

2 方 法

2.1 基于 GAN 的虚拟特征生成与分类层校正

改进的聚合方案,每轮全局迭代分为 4 个阶段:1)用户端(Phase-1),各用户进行本地模型训练,并上传本地模型;2)服务端(Phase-2),运行 FedAvg 算法,聚合特征提取层,并下发给用户;3)用户端(Phase-3),各用户通过公共特征提取层提取真实的中间层特征,并将这些真实特征用于训练本地 GAN;加密上传本地 GAN 中的生成器,及本地的标签信息;4)服务端(Phase-4),汇总各生成器输出的虚拟特征,并逐一微调各用户的分类层;聚合更新后的分类层并下发给用户。算法流程如算法 1。

算法 1:基于 GAN 的虚拟特征聚合与分类层校正

输入:全局迭代次数 T;参与训练的用户数 N;初始全局模型 $\boldsymbol{w}^0=(\boldsymbol{g}^0,\boldsymbol{h}^0)$,其中 $\boldsymbol{g}^0$ 代表模型的特征提取层,$\boldsymbol{h}^0$ 代表模型的分类层。

1. $t=0$;
2. **while** t<T **do**
   /*(**Phase-1**)用户端 */
3. **for** n **in** {N} do:
4. 经 LE 次局部更新,求得最优解:
   $\boldsymbol{w}_n^t=\underset{\boldsymbol{w}}{\operatorname{argmin}} f_n(\boldsymbol{w})$;
5. 加密上传 $\boldsymbol{w}_n^t=(\boldsymbol{g}_n^t,\boldsymbol{h}_n^t)$ 与样本数 D_n;
6. **end for**
   /*(**Phase-2**)服务端 */

7.　运行 FedAvg 算法，聚合特征提取层：

$$\boldsymbol{g}^t = \sum_{n\in\{\mathrm{N}\}} \frac{D_n}{\sum_{n\in\{\mathrm{N}\}} D_n} \cdot \boldsymbol{g}_n^t$$

/*(**Phase-3**)用户端 */

8.　**for** n **in** {N} **do**：

9.　　通过 $\boldsymbol{g}^t$ 提取图像特征，并训练生成模型：$\delta_n = (Dis_n, Gen_n)$

10.　　加密上传本地生成器 Gen_n 及标签信息；

11.　**end for**

/*(**Phase-4**)服务端 */

12.　通过 $\{Gen_n\}$ 及标签信息生成虚拟特征，并对分类层 $\{\boldsymbol{h}_n^t\}$ 逐一微调得到 $\{\tilde{\boldsymbol{h}}_n^t\}$；

13.　运行 FedAvg 算法，聚合分类层：

$$\boldsymbol{h}^t = \sum_{n\in\{\mathrm{N}\}} \frac{D_n}{\sum_{n\in\{\mathrm{N}\}} D_n} \cdot \tilde{\boldsymbol{h}}_n^t$$

14.　将新的全局模型 $\boldsymbol{w}^t \triangleq (\boldsymbol{g}^t, \boldsymbol{h}^t)$ 发送至各用户；

15.　$t = t+1$；

2.2　基于分类层聚类的个性化分组聚合

存在概念偏移的情景中，运行 FedAvg 算法所得的单中心全局模型无法调和用户模型偏差。而基于聚类的联邦学习[15-16]已被证实能够有效对抗概念偏移。可将其定义为多中心联邦学习问题：

$$[\boldsymbol{w}_i^*] \triangleq \operatorname{argmin} \sum_{n\in\{N\}} \sum_{i=1}^{C} r_n^{(i)} f_n(\boldsymbol{w}_i)$$

$$\boldsymbol{w}_i^* = \operatorname*{argmin}_{\boldsymbol{w}_i} \sum_{n\in S_i} f_n(\boldsymbol{w}_i) \tag{2}$$

其中，$\{N\}$代表所有用户的集合，S_i 代表第 i 个用户组，C 为分组数目。$\{N\} = \bigcup_i^C S_i$ 且 $\bigcap_i^C S_i = \varnothing$。$r_n^{(i)}$ 代表用户 n 的簇归属(若 $n \in S_i$ 则 $r_n^{(i)} = 1$ 否则 $r_n^{(i)} = 0$)。w_i^* 为第 i 个用户组的最优簇模型。

类似于 Ghosh 等[17]提出的 IFCA 框架，本章的聚合方案也包括计算用户簇归属和个性化分组聚合两步骤。但与 IFCA 不同，本文的聚类依据为分类层参数的相似性，而非模型的测试精度。改进后的聚合方案工作流程如下：1)用户端(Phase-1)，各用户进行本地模型训练，并上传本地模型；2)服务端(Phase-2)：运行 FedAvg 算法，聚合特征提取层；3)服务端(Phase-3)：以本地模型的分类层为聚类对象，运行聚类算法，产生用户分组结果；对分类层执行分组聚合，产生若干簇模型，并将它们分别下发至对应的用户组。算法流程如算法 2。

算法 2：基于分类层聚类的个性化分组聚合

输入：全局迭代次数 T；用户数 N；初始全局模型 $\boldsymbol{w}^0 = (\boldsymbol{g}^0, \boldsymbol{h}^0)$。

1. $t = 0$；

2. **while** t<T **do**

/*(**Phase-1**)用户端 */

3.　**for** n **in** {N} **do**：

4.　　经 LE 次局部更新，求得最优解：$\boldsymbol{w}_n^t = \operatorname{argmin}_{\boldsymbol{w}} f_n(\boldsymbol{w})$；

5.　　加密上传 $\boldsymbol{w}_n^t \triangleq (\boldsymbol{g}_n^t, \boldsymbol{h}_n^t)$ 与样本数 D_n；

6.　**end for**

/*(**Phase-2**)服务端 */

7.　运行 FedAvg 算法，聚合特征提取层：

$$\boldsymbol{g}^t = \sum_{n\in\{\mathrm{N}\}} \frac{D_n}{\sum_{n\in\{\mathrm{N}\}} D_n} \cdot \boldsymbol{g}_n^t$$

/*(**Phase-3**)服务端 */

8.　对分类层运行聚类算法，得到各用户簇归属 $r_n^{(c)t}$，产生 C 个簇模型：$\{\boldsymbol{w}^{(c)t} = (\boldsymbol{g}^t, \boldsymbol{h}^{(c)t})\}_{c=1,2,\cdots,C}$

$$\boldsymbol{h}^{(c)t} = \sum_{n\in\{\mathrm{N}\}} r_n^{(c)t} \cdot \frac{D_n}{\sum_{n\in\{\mathrm{N}\}} r_n^{(c)t} \cdot D_n} \cdot \boldsymbol{h}_n^t$$

9.　依据 $r_n^{(c)t}$，下发簇模型 $\{\boldsymbol{w}^{(c)t}\}$ 至对应用户；

10.　$t = t+1$；

2.3　集成聚合方案

相关方法的简称与描述如表 1。针对概念偏移场景，将前文所述的 2 种方案集成后的工作流程为：首先，采用 GSCC 聚合方案，直至全局模型基本收敛、或出现连续的本地测试精度下降或损失增大(如图 4、5 中的“终止”所示)，服务器保存最优全局模型副本；其次，切换聚合方案，采用 LLCFL，直至簇模型基本收敛、或出现连续的本地测试精度下降或损失增大，服务器保存最优簇模型副本。

表 1　方法名称与对应描述

方法名	方法描述
FSCC	Feature Sharing for Classifier Correction 基于原始真实特征聚合的分类层校正
GSCC	Generative Sharing for Classifier Correction 基于生成虚拟特征聚合的分类层校正
LLCFL	Last-layer Clustered Federated Learning 基于分类层聚类与分组聚合的联邦学习

3　实验与分析

实验运行平台的硬件配置信息如表 2。

表 2　实验环境配置信息

配置项	配置信息
CPU	Intel(R) Xeon(R) @ 2.20 GHz
GPU	NVIDIA Tesla P100 PCI-E
RAM(CPU)	24 GB
RAM(GPU)	16 GB

3.1 数据集划分

数据划分方案延伸了文献[15-16]中“标签置换”的思想,通过 4 种划分方式模拟不同的数据异质场景。需要说明的是,下列场景除 Case_D 之外,每位用户将获得同等数目的样本。数据划分方案如下:

1)Case_A(特征分布偏移)

样本被划分至三组:[0,1,2]为第一组,[3,4,5]为第二组,[6,7,8,9]为第三组。此场景有 10 位用户,用户编号亦为 0~9。用户分组规则与样本分组规则相同。每位用户从所属的样本组中无放回抽取样本。

将原始样本中标签为“0”“3”“6”的样本置为标签“0”、标签为“1”“4”“7”的样本置为标签“1”,剩余标签为“2”“5”“8”“9”的样本赋予标签“2”。因此,10 位用户构成了 3 个簇,如图 1 所示。

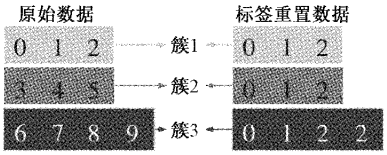

图 1 数据划分-Case_A

2)Case_B(标签分布偏移)

用户分组与样本采样规则同 Case_A,如图 2 所示。不做标签置换操作。

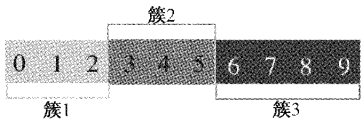

图 2 数据划分-Case_B

3)Case_C(概念偏移)

样本次序被打乱并等分为 3 组。如图 3 所示,第一组样本的前 5 个标签 i 被重置为$(i+1)\%5$,第二组样本的前 5 个标签 i 被重置为$(i+2)\%5$,剩余的第三组样本的前 5 个标签不变;全部样本的后 5 个标签不变。此场景有 9 位用户,用户分组与样本采样规则同 Case_A。

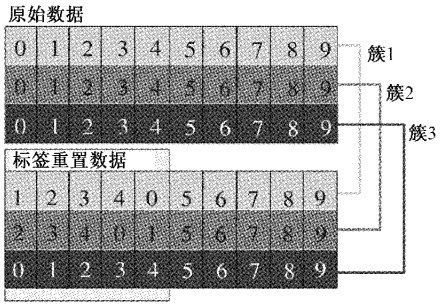

图 3 数据划分-Case_C

4)Case_D(多异质情景融合)

另外,为模拟更为复杂的异质场景,在此采用 Dirichlet 分布构造出 Case_D 数据集[18-20]。Dirichlet 分布可认为是“分布的分布”,它的概率密度函数为:

$$Dir(\boldsymbol{\theta}\mid\boldsymbol{\alpha})=\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_K)}\cdot\prod_{k=1}^{K}\theta_k^{\alpha_k-1} \tag{3}$$

其中,$\Gamma(.)$为伽玛函数(阶乘函数的延拓);$\alpha\in R^K$ 为参数,θ 为 K 维概率单纯形:

$$\sum_{k-1}^{K}\theta_k=1,\ \theta_k\geqslant 0,\alpha_k\geqslant 0 \tag{4}$$

在此划分数据集时,设置:

$$\alpha_1=\alpha_2=\cdots=\alpha_K=\alpha \tag{5}$$

可通过调整 α 的取值,控制用户数据分布的偏移程度:当 α 值越大时,各用户的样本分布越均匀;当 α 值越接近于 0 时,各用户的样本分布越不均匀。本文设置 $\alpha=0.1$。在此基础上,对用户进行与 Case_C 相同的标签置换操作,从而引入概念偏移场景。

3.2 模型与超参数配置

实验采用 3 层 CNN,模型构成如表 3 所示。卷积层之间采用 ReLU 激活函数。将此模型的卷积层作为 GSCC 的特征提取层和 LLCFL 的全局聚合层、最后的全连接层作为 GSCC 的分类层和 LLCFL 的分组聚合层。

表 3 3 层 CNN 的模型结构

结构名	类型	核尺寸	输出通道
Conv1	卷积	5 * 5	6
max_pool	最大池化	2 * 2	6
Conv2	卷积	5 * 5	12
max_pool	最大池化	2 * 2	12
fc	全连接	—	—

实验中的超参数配置为:全局迭代次数 $T=50$、局部更新次数 LE=1;batch_size 为 64、用户本地训练的优化算法为带有动量的 SGD、学习率为 0.01、学习率的指数衰减系数为 0.998;本地训练集与测试集的分配比为 1∶1。

3.3 GSCC 聚合方案实验

GSCC 的对比方案包括独立学习(Indie)、FedAvg 与 FSCC。采用基于多层感知机和 WGAN-GP 的条件生成对抗网络,其中生成器的结构如表 4 所示,判别器的结构如表 5 所示。学习率为 0.000 2、batch_size 为 8、LeakyReLU 负区间斜率为 0.2。

表 4 CGAN-生成器结构

结构名	输入维度	输出维度	激活函数
fc1	30	64	LeakyReLU
fc2	64	128	LeakyReLU
fc3	128	300	Tanh

Case_A、Case_B 的全局测试准确率(记为 G-Acc)如图 4、5。

表 5　CGAN-判别器结构

结构名	输入维度	输出维度	激活函数
fc1	310	64	LeakyReLU
fc2	64	1	—

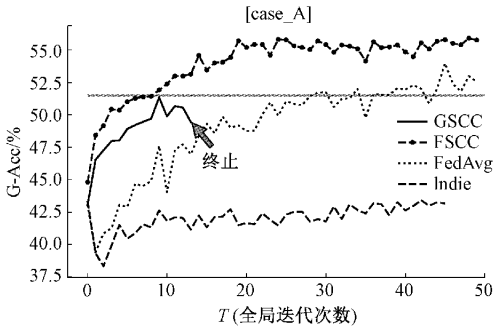

图 4　GSCC 对比实验-[Case_A]

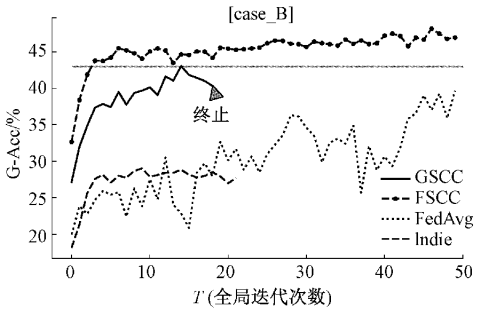

图 5　GSCC 对比实验-[Case_B]

由图 4 可知，在 Case_A 场景中：Indie 的最优值不及 44%；FedAvg 在 $t=22$ 时首次达到 50%以上，在 $t=45$ 时达到最优值 53.98%；FSCC 在 $t=3$ 时达到了 50.44%，在 $t=24$ 时达到最优值 55.82%；GSCC 在 $t=9$ 时达到 51.34%，在 $t=35$ 时达到最优值 54.16%。

由图 5 可知，在 Case_B 场景中：Indie 的最优值不及 30%；FedAvg 在 $t=12$ 时首次达到 30%以上，在 $t=49$ 时达到最优值 39.63%；FSCC 在 $t=2$ 时达到了 41.88%，在 $t=46$ 时达到最优值 48.16%；GSCC 在 $t=10$ 时达到了 40.14%，在 $t=14$ 时达到最优值 43.05%。

由上述结果可以看出，相较于 FedAvg，GSCC 不仅提升了全局模型的最高精度（提升值分别为 0.18%、3.42%），也加速了全局模型的收敛（达到最高精度的通讯效率分别为 1.3 倍、3.5 倍）。

3.4　LLCFL 聚合方案实验

本节实验选用自顶向下的层次聚类算法，簇间距离采用 Ward 测度。以 Case_C 为例，首先通过热力图观察分类层之间的相似度，距离测度除了 L1 范数（Manhattan 距离）、L2 范数（Euclidean 距离）和 Cosine 相似度之外，附加采用了 Pearson 相关系数，其兼具平移不变性和尺度不变性，适用于度量高维向量之间的线性相关性。

由图 6 可知，3 个用户簇的结构清晰。

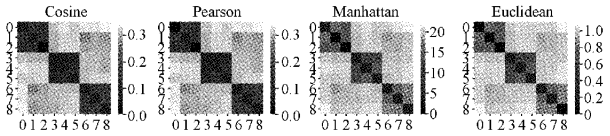

图 6　用户特征向量相似度热力图-[Case_C]

在此选取 3 个内部评价指标，用于决定最佳簇数 C 的取值（如表 6）；选取 2 个外部评价指标，用于评估聚类算法的聚类结果与真实簇归属的差距（如表 7）。可以看出，SC 和 CH 指标能够准确评估出最佳簇数 C=3，且层次聚类的结果正确（正确结果为：[0,0,0, 1,1,1, 2,2,2]）。

表 6　聚类效果内部指标评估-[Case_C]

簇数	SC	CH	DB
2	0.61	19.67	0.53
3	**0.74**	**61.49**	0.31
4	0.44	42.85	0.51
5	0.26	33.98	0.51
6	0.26	29.16	0.47
7	0.25	26.62	0.23
8	0.05	25.31	0.29
9	0.04	23.73	**0.18**

SC：Silhouette Score
CH：Calinski-Harabasz Score
DB：Davies Bouldin Score

表 7　聚类效果外部指标评估-[Case_C]

内部指标	最佳聚类结果	外部指标	
		ARI	FMI
SC	[**2,2,2,1,1,1,0,0,0**]	1.0	1.0
CH	[**2,2,2,1,1,1,0,0,0**]	1.0	1.0
DB	[0,0,7,6,5,3,4,2,1]	0.12	0.29

ARI：Adjusted Rand Score
FMI：Fowlkes Mallows Score

单中心全局模型（single-center model，SCM）在全局测试集上的测试精度不应作为聚合方案的唯一评估值，多中心簇模型（multi-center model，MCM）在本地测试集上的测试精度也应考虑在内；前者评估全局模型的泛化能力，后者评估簇模型的个性化能力。综上，本节的评估指标有 2 项：

1）簇模型在本地测试集上的最高平均准确率，记作 MCM-L-Acc；

2）全局模型在全局测试集上的最高准确率，记作 SCM-G-Acc。

由表 8 可知，$T=50$ 时，采用 Indie 独立学习的用户模

型由于不参与模型聚合,其与 FedAvg 相比 SCM-G-Acc 低了 4.99%、MCM-L-Acc 低了 0.32%;采用 FedAvg 的用户模型与 LLCFL 相比 MCM-L-Acc 低了 5.64%。类似得,由表 9 可知,采用 Indie 独立学习相较于 FedAvg 与 LLCFL,其 SCM-G-Acc 分别低了 5.87%、0.86%,MCM-L-Acc 分别低了 10.45%、17.2%。综上,采用 LLCFL 聚合方案能够有效提升多中心簇模型的个性化能力。

表 8 LLCFL 对比实验(Case_C)

方法		MCM-L-Acc		SCM-G-Acc	
		准确率/%	损失	准确率/%	损失
Indie	$T=10$	**45.33**	**1.40**	**25.65**	**2.15**
	$T=20$	47.90	1.34	同上	同上
	$T=50$	同上	同上	同上	同上
FedAvg (基准)	$T=10$	40.27	1.52	24.74	1.99
	$T=20$	44.22	1.42	**27.89**	**1.94**
	$T=50$	48.22	1.33	**30.64**	**1.92**
LLCFL	$T=10$	42.45	1.35	16.79	2.47
	$T=20$	**48.57**	**1.26**	18.35	2.57
	$T=50$	**53.86**	**1.17**	20.35	2.87

表 9 LLCFL 对比实验(Case_D)

方法		MCM-L-Acc		SCM-G-Acc	
		准确率/%	损失	准确率/%	损失
Indie	$T=10$	35.04	1.79	16.00	2.40
	$T=20$	39.05	1.82	16.29	2.33
	$T=50$	40.01	2.13	18.77	2.58
FedAvg (基准)	$T=10$	41.06	1.50	**19.14**	**2.19**
	$T=20$	43.70	1.54	**21.90**	**2.21**
	$T=50$	50.46	1.37	**24.64**	**2.20**
LLCFL	$T=10$	**45.01**	**1.40**	16.27	2.45
	$T=20$	**51.08**	**1.29**	16.75	2.63
	$T=50$	**57.21**	**1.16**	19.63	2.71

3.5 GSCC 与 LLCFL 集成方案实验

为便于显示,将方案从 GSCC 到 LLCFL 的切换点设置于 $T=13$、从 FSCC 到 LLCFL 的切换点设置于 $T=18$。由表 10 与 Case_C 结果可知:

1)在切换点之前,虽然采用 Indie 和 LLCFL 的簇模型 MCM-L-Acc 高于其他方案,但它们的 SCM-G-Acc 数值很低(17%左右)或涨幅很小;在切换点之后,GSCC/FSCC 与 LLCFL 的结合使得它们的簇模型 MCM-L-Acc 出现了跳跃式增长,且 GSCC 的最优值较 FedAvg 提升了 2.73%。

2)采用 GSCC、FSCC、FedAvg 的模型达到 SCM-G-Acc 大于 30%所需的全局迭代次数依次为 $T=11$、$T=7$、$T=43$(GSCC 的通讯效率是 FedAvg 的 3.9 倍),SCM-G-Acc 的最高取值分别为 29.75%、31.95%、30.64%。

表 10 集成方案对比实验(Case_C)

方法		MCM-L-Acc		SCM-G-Acc	
		准确率/%	损失	准确率/%	损失
Indie	$T=10$	**45.33**	**1.40**	25.65	2.15
	$T=20$	47.90	1.34	同上	同上
	$T=50$	同上	同上	同上	同上
FedAvg (基准)	$T=10$	40.27	1.52	24.74	1.99
	$T=20$	44.22	1.42	27.89	1.94
	$T=50$	48.22	1.33	30.64	1.92
LLCFL	$T=10$	42.45	1.35	16.79	2.47
	$T=20$	**48.57**	**1.26**	18.35	2.57
	$T=50$	**53.86**	**1.17**	20.35	2.87
FSCC+ LLCFL	$T=10$	31.34	1.82	**30.43**	**1.83**
	$T=20$	47.10	1.42	**31.95**	**1.82**
	$T=50$	53.66	1.22	同上	同上
GSCC+ LLCFL	$T=10$	30.61	1.89	29.27	1.93
	$T=20$	42.37	1.49	29.75	1.91
	$T=50$	50.95	1.28	同上	同上

由表 11 与 Case_D 结果可知:

1)在切换点之前,采用 LLCFL 的簇模型 MCM-L-Acc 高于其他方案,但其 SCM-G-Acc 数值很低(16%左右)且涨速较小;在切换点之后,GSCC/FSCC 与 LLCFL 的结合使得它们的簇模型 MCM-L-Acc 实现了跳跃式增长,且 GSCC 的最优值较 FedAvg 提升了 6.82%。

2)采用 GSCC、FSCC、FedAvg 的模型达到 SCM-G-Acc 大于 24%所需的全局迭代次数依次为 $T=7$、$T=9$、$T=47$(GSCC 的通讯效率是 FedAvg 的 6.7 倍),SCM-G-Acc 的

表 11 集成方案对比实验(Case_D)

方法		MCM-L-Acc		SCM-G-Acc	
		准确率/%	损失	准确率/%	损失
Indie	$T=10$	35.04	1.79	16.00	2.40
	$T=20$	39.05	1.82	16.29	2.33
	$T=50$	40.01	2.13	18.77	2.58
FedAvg (基准)	$T=10$	41.06	1.50	19.14	2.19
	$T=20$	43.70	1.54	21.90	2.21
	$T=50$	50.46	1.37	24.64	2.20
LLCFL	$T=10$	**45.01**	**1.40**	16.27	2.45
	$T=20$	**51.08**	**1.29**	16.75	2.63
	$T=50$	57.21	1.16	19.63	2.71
FSCC+ LLCFL	$T=10$	39.54	1.64	24.09	2.37
	$T=20$	44.87	1.43	**24.96**	**2.48**
	$T=50$	57.26	1.25	同上	同上
GSCC+ LLCFL	$T=10$	39.94	1.64	**24.43**	**2.35**
	$T=20$	43.01	1.53	同上	同上
	$T=50$	**57.28**	**1.26**	同上	同上

最高取值分别为 24.43%、24.96%、24.64%。

可知，基于 GSCC 与 LLCFL 的集成聚合方案有效吸取了两者的优点：不仅加速了单中心全局模型的收敛、显著提升了通讯效率，同时提升了多中心簇模型的个性化能力。

4　结　　论

本文的联邦学习模型聚合研究涉及了多种数据异质情形，提出了较为完整的解决方案：GSCC 利用生成的虚拟特征纠正用户模型偏差、增强了单中心全局模型的收敛性能；LLCFL 通过分类层参数聚类划分了用户分组，并采用分组聚合的方式形成了个性化能力更强的多中心簇模型、有效缓解了概念偏移问题；GSCC 与 LLCFL 的集成方案则兼备了两者的优点。未来将利用更多的数据集、分类模型、GAN 模型以及聚类算法（如模糊聚类）验证本文聚合方案的有效性，并开展收敛性分析。另一方面，由于现已有诸多特征反演方法（feature inversion）使用梯度反向传播等技术重建原始的图像，未来将针对 GSCC 进行隐私分析，或先将真实特征加入小方差噪声，再输入判别器。

参考文献

[1] KONEČNÝ J, MCMAHAN H B, YU F X, et al. Federated learning: Strategies for improving communication efficiency [C]. NIPS Workshop on Private Multi-Party Machine Learning, Barcelona, Spain, 2016.

[2] HSIEH K, PHANISHAYEE A, MUTLU O, et al. The non-IID data quagmire of decentralized machine learning [C]. International Conference on Machine Learning, Virtual: PMLR, 2020: 4387-4398.

[3] LI T, SAHU A K, ZAHEER M, et al. Federated optimization in heterogeneous networks [C]. Proceedings of Machine Learning and Systems (MLSys), Austin, US: 2020: 429-450.

[4] WANG J Y, LIU Q H, LIANG H, et al. Tackling the objective inconsistency problem in heterogeneous federated optimization[C]. 33rd Annual Conference on Neural Information Processing Systems, Vancouver, Canada, 2020.

[5] ZHU Y SH, MARKOS C, ZHAO R H, et al. FedOVA: One-vs-all training method for federated learning with non-IID data [C]. 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China: IEEE, 2021.

[6] WANG H Y, YUROCHKIN M, PAPAILIOPOULOS D, et al. Federated learning with matched averaging[C]. 8th International Conference on Learning Representations (ICLR), Virtual: OpenReview.net, 2020.

[7] YU F X, ZHANG W SH, QIN ZH W, et al. Fed2: Feature-aligned federated learning[C]. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore: ACM, 2021: 2066-2074.

[8] XIE M, LONG G D, SHEN T, et al. Multi-center federated learning [J]. ArXiv Preprint, 2020, ArXiv:2005.01026.

[9] LUO M, CHEN F, HU D P, et al. No fear of heterogeneity: Classifier calibration for federated learning with non-IID data[J]. Advances in Neural Information Processing Systems, 2021, 34: 5972-5984.

[10] 陈亮，吴攀，刘韵婷，等. 生成对抗网络 GAN 的发展与最新应用[J]. 电子测量与仪器学报，2020，34(6)：70-78.

[11] 王桂棠，林桢哲，符秦沈，等. 联合生成对抗网络的肺结节良恶性分类模型[J]. 仪器仪表学报，2020，41(11)：188-197.

[12] 杨鸿杰，陈丽，张君毅. 基于生成对抗网络的数字信号生成技术研究[J]. 电子测量技术，2020，43(20)：127-132.

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

[14] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein GANs [C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, US: ACM, 2017: 5769-5779.

[15] SATTLER F, MÜLLER 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.

[16] CHRISTOPHER B, FAN ZH, ANDRAS P. Federated learning with hierarchical clustering of local updates to improve training on non-IID data [C]. International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020.

[17] GHOSH A, CHUNG J, YIN D, et al. An efficient framework for clustered federated learning [C]. Advances in Neural Information Processing Systems (NIPS), Virtual: Curran Associates, 2020: 19586-19597.

[18] HSU T M H, QI H, BROWN M. Federated visual classification with real-world data distribution [C]. European Conference on Computer Vision (ECCV),

Virtual: Springer, 2020: 76-92.

[19] LI Q B, DIAO Y Q, CHEN Q, et al. Federated Learning on Non-IID Data Silos: An Experimental Study[J]. ArXiv Preprint,2021, ArXiv:2102.02079.

[20] ZHU H Y, XU J J, LIU SH Q, et al. Federated learning on non-IID data: A survey [J]. Neurocomputing, 2021, 465: 371-390.

作者简介

邹承明,工学博士,教授,主要研究方向为计算机视觉、智慧物联网、软件理论与方法(操作系统)。

E-mail:zoucm@whut.edu.cn

赵宁,硕士研究生,主要研究方向为数据挖掘、联邦学习模型聚合与通讯优化。

E-mail:mariacarter2013@163.com