

DP-FedDRC: 缓解维度坍塌的差分隐私个性化联邦学习

王民利, 张宇浩, 胡昌慧

(海南大学网络空间安全学院(密码学院), 海南 海口 570228)

摘要: 联邦学习在分布式隐私保护场景中面临数据异质性与隐私泄露双重挑战。个性化联邦学习通过约束局部模型更新方向缓解数据分布差异, 而差分隐私则提供严格的隐私保障。然而, 现有方法在同时解决这两个问题时, 未能充分考量噪声对数据非独立同分布场景下联邦学习训练过程的耦合影响, 导致隐私-效用权衡不足。实验表明, 差分隐私噪声会显著加剧数据异质性引发的维度坍塌问题。该问题导致特征空间的有效自由度衰减, 并削弱了特征的判别性表达能力, 成为模型性能下降的关键原因。为此, 提出一种融合特征去相关与特征残差约束的差分隐私个性化联邦学习框架(DP-FedDRC)。该框架通过抑制特征冗余相关性并保留关键判别信息, 缓解噪声和数据异质性共同导致的维度坍塌问题。在隐私保护方面, DP-FedDRC在本地训练中采用DP-SGD机制实现梯度裁剪与高斯噪声注入, 并基于Rényi差分隐私框架严格核算全局隐私预算。在MNIST、Fashion-MNIST和CIFAR-10数据集上的实验表明, 在同等严格的隐私约束下, DP-FedDRC相比基线算法实现了更高的模型准确率, 且计算开销仅增加8.3%~9.9%。

关键词: 联邦学习; 差分隐私; 维度坍塌; 非独立同分布数据; 特征去相关

中图分类号: TP309

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2026008

DP-FedDRC: differentially private personalized federated learning to alleviate dimension collapse

Wang Minli, Zhang Yuhao, Hu Changhui

School of Cyberspace Security (School of Cryptography), Hainan University, Haikou 570228, China

Abstract: Federated learning faces dual challenges of data heterogeneity and privacy leakage in distributed privacy-preserving scenarios. Personalized federated learning mitigates data distribution discrepancies by constraining local model update directions, while differential privacy provides rigorous privacy guarantees. However, when solving these two problems at the same time, the existing methods fail to fully consider the coupling effect of noise on the federated learning and training process in the data Non-IID scenario, resulting in insufficient privacy-utility trade-offs. Experiments show that differential privacy noise significantly exacerbates the dimension collapse problem induced by data heterogeneity. This problem leads to the decay of effective degrees of freedom in the feature space and impairs discriminative feature representation, which is a key factor in model performance degradation. To ad-

收稿日期: 2025-07-30; 修回日期: 2025-11-09

通信作者: 胡昌慧, chu@hainanu.edu.cn

基金项目: 国家自然科学基金(No.62262012); 中国国家留学基金委资助项目(No.202407560059); 海南省自然科学基金(No.124QN176); 海南省研究生创新科研课题(No.Qhyb2024-02); 海南大学科研启动基金(No.KYQD22094)

Foundation Items: The National Natural Science Foundation of China (No.62262012), China Scholarship Council Funding (No.202407560059), Hainan Provincial Natural Science Foundation of China (No.124QN176), Hainan Postgraduate Innovative Scientific Research Project (No.Qhyb2024-02), Start-up Research Foundation of Hainan University (No.KYQD22094)

dress this, a differential privacy personalized federated learning framework incorporating feature decorrelation and residual constraints (DP-FedDRC) was proposed. This framework alleviates the dimension collapse problem caused by both noise and data heterogeneity by suppressing redundant feature correlations and preserving critical discriminative information. In terms of privacy protection, DP-FedDRC implements gradient clipping and Gaussian noise injection using the DP-SGD mechanism during local training, and rigorously accounts for the global privacy budget based on the Rényi differential privacy framework. Experiments on the MNIST, Fashion-MNIST, and CIFAR-10 datasets demonstrated that under equally strict privacy constraints, DP-FedDRC achieved higher model accuracy compared to baseline algorithms, with only an 8.3%~9.9% increase in computational overhead.

Keywords: federated learning, differential privacy, dimension collapse, non-IID data, feature decorrelation

0 引言

联邦学习^[1]是一种分布式机器学习范式,旨在应对现实中普遍存在的“数据孤岛”问题,满足日益严格的数据隐私保护要求。在大量敏感数据分散存储于本地设备,难以集中利用的背景下,传统依赖中心化数据训练的机器学习方法面临显著局限。联邦学习允许多个参与方在严格保障本地数据隐私的前提下,通过协同训练共同优化一个全局模型。这一机制为医疗诊断、金融风控等对数据隐私高度敏感的领域提供了极具前景的解决方案,既能有效规避数据集中化存储与传输带来的隐私风险,又可以降低数据传输与存储的成本^[2]。

然而,联邦学习在实际部署中仍面临数据异质性与隐私泄露两大关键挑战^[3]。数据异质性主要表现为客户端本地数据的非独立同分布(non-independently identically distributed, Non-IID)特性,包括特征分布偏斜、标签分布偏斜和样本数量偏斜等形式。这种数据分布差异显著削弱了模型的泛化能力^[4],导致模型在特定客户端本地数据上的表现往往难以达到预期。与此同时,联邦学习架构下的隐私泄露风险并未被完全消除^[5-6]。具体而言,攻击者可以通过分析共享的模型参数或梯度信息发起成员推理攻击^[7-8],或利用梯度信息实施数据重建攻击以近似还原出训练集中的原始敏感数据^[9],从而对用户隐私构成严重威胁。

现有研究往往对数据异质性与隐私保护采取割裂式的优化策略,存在明显局限。个性化联邦学习(personalized federated learning, PFL)方法采用正则化技术来约束局部模型的更新,从而减少客户端之间的模型差异,提升算法收敛性,但其设计未内置隐私保护机制,模型参数或梯度的共享仍存在泄露客户端敏感信息的风险。与之相

对,差分隐私联邦学习(differentially private federated learning, DP-FL)方法通过向梯度信息或模型更新参数注入噪声实现严格的差分隐私保障,并采用自适应优化策略优化隐私-效用权衡,却未能针对性解决数据异质性引发的模型性能退化问题。

尽管部分工作尝试同时应对这两个挑战,但其设计未能深入分析差分隐私噪声在数据 Non-IID 场景下引入的额外负面效应。差分隐私与异构鲁棒的联邦学习(DP-FedCVR)算法^[10]和差分隐私联邦弹性权重整合(DP-FedEwc)算法^[11]虽然通过减少客户端之间的模型方差或引入弹性权重约束等方式提升了模型对噪声的鲁棒性,且在实验中观察到一定的性能改善,但并未深入分析噪声引起的特征表征退化问题。这一缺陷会导致现有方法在数据 Non-IID 场景下出现显著的性能下降。本文研究发现,在 Non-IID 场景下,差分隐私噪声不仅会放大客户端间的模型方差,还会扭曲特征空间的表征结构,导致特征维度高度冗余化,即加剧了维度坍塌。这一现象与 Shi 等^[12]的理论分析形成重要关联,即 Non-IID 数据会通过客户端梯度冲突引发全局模型的维度坍塌,而本文揭示了差分隐私噪声与数据异质性之间存在负面协同效应。具体而言,噪声扰动的随机性与本地数据分布的偏斜性共同作用,导致特征空间冗余化,使得现有差分隐私个性化联邦学习(differentially private personalized federated learning, DP-PFL)方法很难学到具有判别性的特征。因此,解决差分隐私噪声与数据 Non-IID 联合作用下的维度坍塌问题,成为提升 DP-PFL 性能的关键。

基于上述分析,本文提出一种融合特征去相关与特征残差约束的 DP-PFL 框架(DP-FedDRC)。

本文的主要贡献如下。

1) 实证研究发现, 差分隐私噪声与数据 Non-IID 的协同效应会加剧特征维度坍塌, 其本质是噪声扰动导致特征协方差矩阵发生奇异值谱塌缩。

2) 在差分隐私随机梯度下降 (differentially private stochastic gradient descent, DP-SGD)^[13] 的本地训练框架中, 引入特征去相关和特征残差约束的双正则化机制, 通过在本地训练中引入特征相关矩阵的最小化损失和特征残差项, 在抑制特征冗余相关性的同时, 显式保留关键判别信息。

3) 实验结果表明, 在严格的差分隐私保证下, DP-FedDRC 框架在多个基准数据集 (MNIST、Fashion-MNIST 和 CIFAR-10) 上相比基线方法能够有效缓解特征维度坍塌, 提高模型准确率, 且计算开销仅增加 8.3%~9.9%。

1 相关工作

1.1 个性化联邦学习

为应对客户端数据异质性带来的挑战, PFL 旨在通过约束客户端模型的更新方向或修正其优化偏差, 减少客户端之间的模型方差, 从而提升模型在本地数据上的适应性和收敛性。为此, 研究者提出了多种策略: FedProx (近端正则化项约束下的 PFL)^[14] 在本地损失函数中引入了邻近项, 从而限制本地模型与全局模型的偏离程度; SCAFFOLD^[15] 通过引入全局与本地梯度间的偏

差校正项来补偿客户端训练过程中出现的梯度漂移; FedDyn^[16] 利用历史全局模型信息构建动态正则化项, 引导客户端模型收敛; FedOpt^[17] 对服务器端的参数聚合策略进行了改进, 采用自适应优化算法 (如 Adam) 代替简单的参数平均。FedVRA^[18] 通过自适应客户端之间的模型方差削减和全局模型偏差校正来应对数据异质性问题。

上述方法在缓解数据 Non-IID 导致的模型性能下降方面取得了一定的成效, 但这些方法的设计范式主要聚焦于训练优化过程中的参数更新约束, 未充分考虑数据异质性对模型特征表示空间造成的结构性退化问题。FedDecorr (引入去相关正则化项的 PFL)^[12] 通过引入特征去相关正则项, 在一定程度上缓解了数据分布偏移导致的特征维度坍塌, 从而增强了全局模型的特征表征能力。然而, 在本地数据分布高度异构场景下, 该方法单纯以最小化特征间相关性为目标, 容易导致过度去相关化, 破坏数据中本应保留的语义相关结构, 反而会削弱模型的特征判别能力。更为关键的是, 包括 FedDecorr 在内的上述所有方法在设计之初均未融入隐私保护机制, 导致客户端上传的模型更新或参数仍存在泄露本地敏感信息的风险。

1.2 差分隐私联邦学习

DP-FL 系统的架构如图 1 所示。现有研究通过引入差分隐私技术来保障联邦学习的安全性, 其核心思想是在客户端的梯度或模型更新信息中注入随

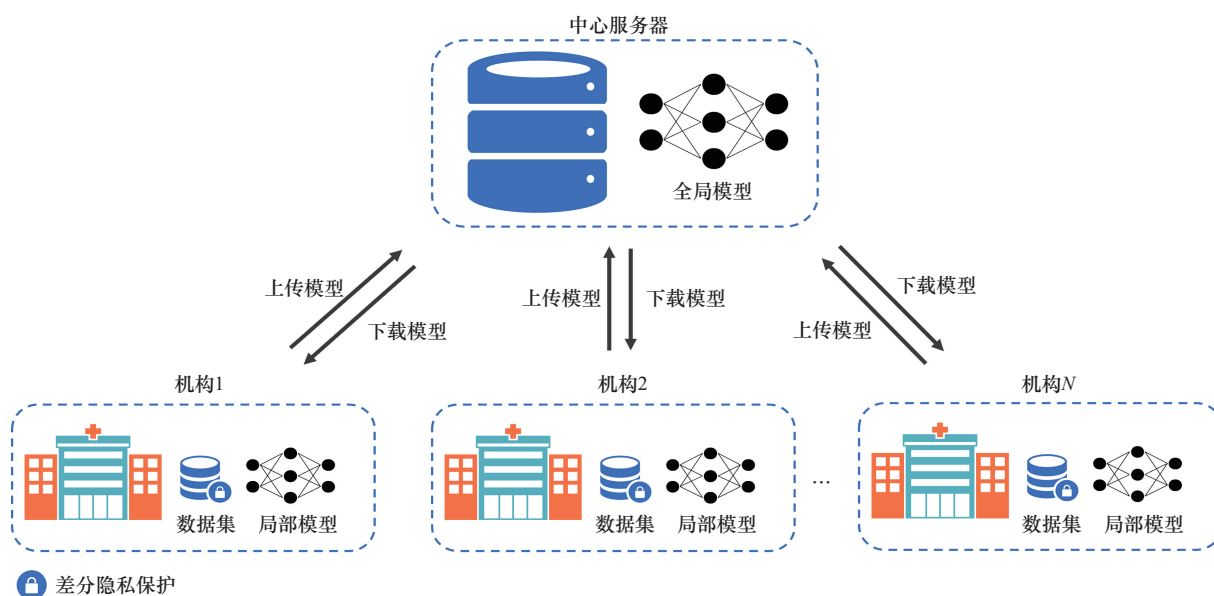


图1 DP-FL 系统的架构

机噪声,通过模糊敏感数据相关特征实现隐私保护。CLDP-SGD^[19]结合本地随机化处理和可信洗牌器机制,在降低通信成本的同时提高隐私保护强度。Fed-CDP^[20]将噪声注入与参数混淆策略相融合,通过降低梯度信息量来抵御逆向攻击。Adap DP-FL^[21]根据客户端梯度的异质性动态调整裁剪阈值和噪声注入尺度,实现隐私保护强度与模型效用的自适应权衡优化。DP-NFL^[22]首次在联邦学习中引入截断集中差分隐私技术,并采用无放回客户端采样策略来减少训练过程中的方差干扰。

部分研究方法将差分隐私与PFL算法相结合,旨在解决数据异构性问题的同时确保隐私安全。DP-FedProx^[23]和DP-SCAFFOLD^[24]将差分隐私直接嵌入现有PFL框架中。DP-FedCVR^[10]通过减少客户端之间的模型方差和在模型更新与双变量通信中同时注入噪声的设计,从优化角度增强了算法对噪声的容忍度。DP-FedEwc^[11]从模型个性化角度出发,通过引入弹性权重巩固机制约束重要参数的变化,实验证明其训练过程对差分隐私噪声引入的额外异质性具有更强的稳定性。

然而,上述方法主要从模型参数调整或训练过程优化的角度,通过提升模型对噪声的鲁棒性来改善模型性能,并未深入探究差分隐私噪声在特征表征层面引发的根本性退化问题。与这些方法不同,本文提出的DP-FedDRC框架着眼于特征表征层面的退化机制,通过特征去相关和特征残差约束显式缓解维度坍塌,从根源上提升模型在隐私保护约束下的表征学习能力与最终性能。相关方法对比见表1。

表1 相关方法对比

方法	提供 隐私保护	处理 Non-IID数据	缓解 维度坍塌
FedAvg ^[1]	×	×	×
DP-FedCVR ^[10]	√	√	×
DP-FedEwc ^[11]	√	√	×
FedDecorr ^[12]	×	√	√
FedProx ^[14]	×	√	×
FedVRA ^[18]	×	√	×
DP-FedAvg ^[25]	√	×	×
DP-FedDRC	√	√	√

注:“√”表示满足对应功能,“×”表示不满足。

1.3 差分隐私技术前沿进展

近年来,差分隐私在机器学习中的应用不断深化。Feng等^[26]将循环神经网络扩展至张量空间,实现对高阶异构时序数据的处理,并采用函数机制在目标函数中注入拉普拉斯噪声以提供差分隐私保护。Zhao等^[27]提出RobustLDP框架,通过预定义信息与反馈机制的协同设计,在频率估计与均值估计任务中实现了对操纵攻击的有效抵御。Zhang等^[28]提出MENTOR框架,基于双向有界扰动机制和高斯-拉普拉斯混合模型,在保护本地差分隐私的同时显著提升了推荐系统的预测准确性。

与上述方法不同,本文提出的DP-FedDRC框架聚焦于联邦学习场景。该方法采用基于高斯噪声的本地差分隐私机制,通过向客户端上传的梯度注入高斯随机噪声实现严格的隐私保护,并重点解决了该场景下因引入差分隐私噪声而加剧的维度坍塌问题,从而在保护用户数据隐私的同时有效维持模型的特征表达能力。

2 预备知识

2.1 联邦学习

联邦学习是一种分布式机器学习范式,旨在解决数据孤岛与隐私保护之间的矛盾。

在联邦学习体系中,假设有 N 个分布式本地参与方协同构建全局模型。第 i 个参与方持有独立的本地用户数据集 D_i ,其核心任务是在本地数据上独立训练模型,并将训练得到的模型参数上传至中心服务器。中心服务器接收各参与方的参数更新数据后,通过聚合算法生成全局更新模型参数,并将优化后的参数返回给部分或全部参与方,以启动下一轮的本地模型迭代。

中心服务器定义的全局损失函数为各参与方本地损失的加权平均,其数学表达式如式(1)所示。

$$f(w, D) = \sum_{i=1}^p \frac{n_i}{n} f_i(w, D_i) \quad (1)$$

其中, $f_i(w, D_i)$ 表示第 i 个参与方使用本地数据集 D_i 训练模型 w 时的损失函数。 n_i 为第 i 个参与方的数据样本量, $n = \sum_{i=1}^p n_i$ 为全体参与方的数据总量。

2.2 差分隐私

作为数据隐私保护的核心理论框架,差分隐

私^[29]通过严格的数学定义保证在数据分析过程中个体信息具有不可区分性。其核心机制是在查询结果中注入可控噪声，使得攻击者无法基于输出结果推断出特定个体是否存在于数据集中。

定义 1 差分隐私

一个随机化机制 $\mathcal{M}:\mathcal{D} \rightarrow \mathcal{R}$ 满足 (ϵ, δ) -DP，若对于任意相邻数据集 $D, D' \in \mathcal{D}$ 和任意输出子集 $S \subseteq \mathcal{R}$ ，则有：

$$\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \cdot \Pr[\mathcal{M}(D') \in S] + \delta \quad (2)$$

其中， ϵ 为隐私预算，用于衡量隐私保护级别； δ 为松弛参数，表示违反严格 ϵ 约束的概率上界。

本文采用传统差分隐私的一种扩展定义方法——Rényi 差分隐私 (RDP)^[30]。该方法采用 Rényi 散度来量化隐私泄露程度，能够提供更加精细和严格的隐私分析。

定义 2 Rényi 差分隐私

机制 $\mathcal{M}$ 满足 (α, ϵ) -RDP，对于任何相邻数据集 D, D' ，其输出的 Rényi 散度满足：

$$D_\alpha(\mathcal{M}(D) \parallel \mathcal{M}(D')) \leq \epsilon \quad (3)$$

其中， α 为散度阶数。

3 问题描述

Shi 等^[12]的研究表明，联邦学习中的数据异质性会导致全局模型表征空间发生维度坍塌，且维度坍塌程度与数据异质性呈正相关。本节将扩展这一发现（即在数据 Non-IID 场景下，差分隐私噪声的引入会显著加剧这一维度坍塌现象），并通过实验和理论分析加以验证。

为量化差分隐私噪声在联邦学习中对维度坍塌的影响，本节基于标准的 FedAvg^[1] 框架，使用 CIFAR-10 数据集和卷积神经网络 (CNN) 模型进行实验，并通过狄利克雷分布（浓度参数 $\alpha = 0.1$ ）划分数据集以构建典型的数据 Non-IID 场景。与 Shi 等^[12]通过调节 α 控制数据异质性不同，本文在客户端本地训练过程中注入高斯噪声，并固定 $\alpha = 0.1$ 以聚焦于噪声影响。设置噪声标准差 $\sigma \in \{0, 1, 3\}$ （ $\sigma = 0$ 为无噪声基线， σ 增大意味着隐私保护增强）。为评估全局模型表征空间退化现象，需要提取在完整测试集上的特征矩阵 $\bar{\mathbf{z}} \in \mathcal{R}^{N \times d}$ （ N 为测试样本数， d 为特征维度），计算其表征向量的协方差矩阵 Σ 。这一过程的数学表达式如式(4)所示。

$$\Sigma = \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^\top \quad (4)$$

其中， $\mathbf{z}_i$ 为第 i 个样本的特征向量， $\bar{\mathbf{z}}$ 为特征均值。对 Σ 进行奇异值分解后，按降序记录前 100 个奇异值 $\{\lambda_k\}_{k=1}^{100}$ 的对数值 ($\log \lambda_k$)，以此量化全局模型表征空间的维度坍塌程度。

如图 2 所示，在数据 Non-IID 场景下，差分隐私噪声的引入显著加剧了全局模型表征空间的维度坍塌。具体而言，在无噪声基线 ($\sigma = 0$) 下，相较于独立同分布 (IID) 场景，虽然各方法的奇异值谱已出现一定程度的衰减，但其衰减相对平缓，有效表征维度得以较好地维持，表明模型在高维特征空间中仍能保持相当的特征多样性。然而，当引入差分隐私噪声 ($\sigma > 0$) 后，奇异值谱的衰减整体显著加剧，表现为曲线整体显著下移且衰减速率明显加快，模型表征的总体强度下降。同时，这种现象随噪声标准差 σ 的增大而愈加显著。对比不同噪声水平的影响可见，从低噪声 ($\sigma = 1$) 到高噪声 ($\sigma = 3$)，奇异值谱坍塌逐步加剧，模型的有效表征维度也随之降低。噪声强度与维度坍塌之间的正相关关系，揭示了差分隐私机制在数据异质环境下对模型表征能力具有负面效应。

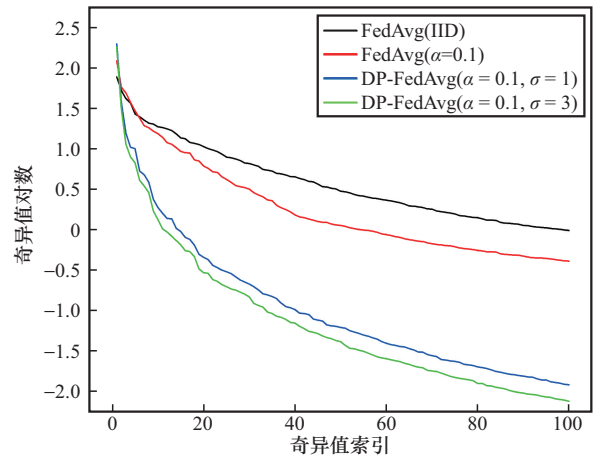


图 2 特征奇异值谱衰减

这一现象可以通过差分隐私噪声与数据异质性之间存在的负面协同效应解释，具体如下。

1) 加剧低秩偏向。数据异质性会驱动本地模型权重 $\Pi(t)$ 向低秩解演化^[12]。在模型参数聚合时注入差分隐私噪声，其扰动对权重矩阵的低方

差方向（对应较小的奇异值）的破坏性更强。噪声会掩盖这些方向上本已微弱的信号，导致优化过程难以维持其特征强度，进而加剧特征间的冗余相关性，最终加速权重矩阵的低秩化趋势。

2) 表征协方差秩降低。由式(5)可知，若 $\Pi(t)$ 的秩降低，则表征协方差矩阵 $\Sigma(t)$ 的秩（或其有效秩）必然显著减小，直接表现为奇异值谱的塌缩（即维度坍塌）。

$$\Sigma(t) = \Pi(t) \cdot \text{Cov}(X) \cdot \Pi(t)^T \quad (5)$$

上述实验清晰地表明，差分隐私噪声的引入显著加剧了数据 Non-IID 场景下的维度坍塌。这一发现揭示了一个在数据异构场景下尚未被充分研究的隐私—效用冲突：差分隐私在提供安全保障的同时，会直接引发特征表征空间的结构退化，并最终表现为模型推理精度的下降。与现有致力于在梯度层面提升噪声鲁棒性的方法不同，本文首次将特征层面的分析维度从普通的数据 Non-IID 场景拓展至隐私保护场景，从而弥补了现有 DP-PFL 方法在特征层面分析的缺失。

4 DP-FedDRC 方法设计

DP-FedDRC 旨在同时解决联邦学习中的数据异质性与隐私泄露风险问题。具体而言，该框架引入了双正则化机制。该机制在保证严格差分隐私的前提下，能够有效缓解数据异质性与差分隐私噪声共同导致的特征维度坍塌问题。表2列出了 DP-FedDRC 的相关符号及其定义。

表2 DP-FedDRC 相关符号及其定义

符号	定义
T	全局通信轮数
N	客户端总数
θ_i	全局模型参数（第 i 轮）
$\theta_{i,k}$	客户端 i 的本地模型
σ	噪声标准差
C_g	梯度裁剪阈值
λ_d	特征去相关强度
λ_r	特征残差约束强度
$\nabla_{\theta} \mathcal{L}_{\text{decorr}}$	特征去相关损失梯度
$\nabla_{\theta} \mathcal{L}_{\text{res}}$	特征残差约束损失梯度

4.1 特征去相关与特征残差约束机制

针对联邦学习中差分隐私噪声与数据 Non-IID 共同导致的特征维度坍塌问题，本文引入特征去

相关方法^[12]，通过最小化特征相关矩阵的 Frobenius 范数，显式约束特征维度间的冗余相关性。

与 DeCov^[31] 和 Structured-DeCov^[32] 等方法直接对协方差矩阵施加正则化不同，本文方法作用于特征相关矩阵。具体而言，其惩罚项仅针对相关矩阵的非对角线元素（即特征间相关系数的平方），而不影响对角线元素（各特征自身的方差），从而将优化目标聚焦于减少特征间的线性相关性，而非改变特征本身的尺度。这种设计有效消除了特征尺度与方差的干扰，使优化过程更专注于特征间的关系建模。因此，该方法在提升算法稳定性与有效性的同时，也显著降低了计算开销。

具体而言，在客户端本地训练阶段，首先对深层特征表示 $\mathbf{z} \in \mathcal{R}^d$ （分类层前输入）进行标准化处理。这一操作的数学表达式如下：

$$\hat{\mathbf{z}} = \frac{\mathbf{z} - \mu_B}{\sqrt{\sigma_B^2 + \beta}} \quad (6)$$

其中， μ_B 和 σ_B 分别为批次特征的均值与标准差， B 代表当前数据批次中所有样本的集合， β 为数值稳定性常数。基于标准化后的特征 $\hat{\mathbf{z}}$ ，计算其相关性矩阵，如式(7)所示。

$$\mathbf{K} = \frac{1}{|B|} \sum_{\mathbf{z} \in B} \hat{\mathbf{z}} \hat{\mathbf{z}}^T \quad (7)$$

最后，通过最小化该相关矩阵非对角线元素的平方和实现特征去相关。

$$\mathcal{L}_{\text{decorr}} = \frac{1}{d(d-1)} \sum_{i \neq j} \mathbf{K}_{ij}^2 \quad (8)$$

其中， d 为特征维度， i 和 j 表示 $\mathbf{K}$ 的行索引和列索引。该正则化项可以直接减少特征维度间的线性相关性，有助于抑制协方差矩阵的奇异值衰减，从而缓解维度坍塌。

尽管特征去相关具有上述优势，但在数据 Non-IID 场景下，客户端存在过度优化本地特征分布的风险。由于相关矩阵的非对角线元素（相关系数）易受本地数据分布偏斜的影响，过度最小化特征间相关性可能导致本地特征与全局语义不一致。例如，某些在全局数据中存在相关性的特征对，在本地数据中可能表现为不相关，这种局部优化将破坏模型聚合后的泛化能力。此外，差分隐私噪声会放大该问题，使本地模型更容易偏离全局语义共识。

为克服上述局限性，本文引入了特征残差约

束机制。其动机在于，浅层网络特征（如初级卷积层输出）通常包含更通用、更鲁棒的视觉基元（如边缘、纹理），这类特征对客户端本地数据分布的依赖性较小，且对差分隐私噪声干扰具有较强的鲁棒性。特征残差约束机制通过约束深层特征与投影后的浅层特征之间的差异，迫使网络在去相关的同时，保留那些具有判别性的关键信息，防止过度去相关导致特征退化。

在具体实现时，为构建特征残差，需对浅层特征进行投影变换，使其与深层特征维度对齐。当浅层特征与深层特征维度不一致时，通过可学习投影矩阵 $W \in \mathcal{R}^{d \times d_s}$ 将浅层特征 $z_s \in \mathcal{R}^{d_s}$ 映射到深层特征空间。其形式化表达如式(9)所示。

$$z_{\text{proj}} = Wz_s \quad (9)$$

其中， W 采用随机初始化，并与网络主体参数一起通过梯度下降进行端到端优化，以学习从通用浅层特征空间到特定深层特征空间的最优线性变换。这一过程在实现维度对齐的同时，能够最大限度地保留可用于约束深层特征的有用判别信息。若深层特征与浅层特征维度相同，则无须进行显式投影，此时等价于将 W 视为单位矩阵，直接计算二者间的残差。

获得维度一致的特征表示后，构建特征残差约束损失 $\mathcal{L}_{\text{res}}$ 。其形式化表达如式(10)所示。

$$\mathcal{L}_{\text{res}} = \frac{1}{|B|} \|z - z_{\text{proj}}\|_2^2 \quad (10)$$

该约束机制的核心在于，通过迫使深层特征与受差分隐私噪声影响较小、更具通用性的浅层特征保持一致，促使深层特征继承浅层特征的鲁棒性。

综上所述，本文提出的双正则化机制（特征去相关项分解冗余和有害的相关性，直接对抗维度坍塌；而特征残差约束项则有效地维持必要且通用的判别信息，防止特征退化并保障语义一致性）协同作用，使模型在数据 Non-IID 与差分隐私噪声干扰的双重影响下，仍能学习到具有高多样性和强判别性的特征表示。

4.2 本地训练与隐私保护

客户端本地训练的目标函数整合任务损失与双正则化项。其形式化表达如式(11)所示。

$$\min_{\theta} \mathcal{L}_{\text{task}} + \lambda_d \cdot \mathcal{L}_{\text{decorr}} + \lambda_r \cdot \mathcal{L}_{\text{res}} \quad (11)$$

其中， λ_d 和 λ_r 分别控制特征去相关与特征残差约束的强度， $\mathcal{L}_{\text{task}}$ 为交叉熵损失。

为符合差分隐私要求，在任务梯度计算中采用 DP-SGD 机制实施严格的隐私保护。首先对每个样本的损失梯度 g_k 进行梯度裁剪，得到裁剪后的梯度 $\bar{g}_k$ 。其形式化表达如式(12)所示。

$$\bar{g}_k = g_k / \max\left(1, \frac{\|g_k\|_2}{C_g}\right) \quad (12)$$

其中， C_g 为梯度裁剪阈值。随后，聚合该批样本裁剪后的梯度，并对其添加高斯噪声。其形式化表达如式(13)所示，其中 I 为单位矩阵。

$$\tilde{g}_{\text{task}} = \frac{1}{|B|} \left(\sum_{x \in B} \bar{g}_k(x) + \mathcal{N}(0, \sigma^2 C_g^2 I) \right) \quad (13)$$

4.3 全局聚合

在每一轮通信中，各客户端下载全局模型并执行本地训练。本地训练包含多次迭代，每次迭代采样一个批数据，依次计算受差分隐私保护的任务梯度、特征去相关正则化梯度及特征残差约束梯度，组合后更新本地模型参数。其形式化表达如式(14)所示。

$$\theta_{i,k} = \theta_{i,k-1} - \eta_l (\tilde{g}_{\text{task}} + \lambda_d \cdot \nabla_{\theta} \mathcal{L}_{\text{decorr}} + \lambda_r \cdot \nabla_{\theta} \mathcal{L}_{\text{res}}) \quad (14)$$

本地训练结束后，客户端将模型更新参数上传至中心服务器。中心服务器采用加权平均策略聚合客户端更新。其形式化表达如式(15)所示。

$$\theta_t = \theta_{t-1} + \eta_g \sum_{i=1}^N \frac{|D_i|}{\sum_j |D_j|} (\theta_{t,i} - \theta_{t-1}) \quad (15)$$

最后，基于 RDP 框架并使用子采样高斯机制的组性质计算累积隐私消耗。DP-FedDRC 的整个过程在算法 1 中描述。

算法 1 DP-FedDRC 框架

输入 学习任务数据集 $D_1, \dots, D_N$ ，客户端总数 N ，通信轮数 T ，噪声标准差 σ ，裁剪阈值 C_g ，特征去相关强度 λ_d ，特征残差约束强度 λ_r

输出 个性化模型集合 $\{\theta_{t,i}\}$ 和隐私参数 (ϵ, δ)

- 1) 中心服务器初始化全局模型 θ_0 ;
- 2) for $t = 1$ to T do
- 3) for $i = 1$ to N do
- 4) 从全局模型继承参数 $\theta_{i,0} \leftarrow \theta_{t-1}$;
- 5) 无放回均匀采样数据批次 $B_k \subseteq D_i$;

6) 计算批次中每个样本 $x \in B_k$ 的梯度 $\mathbf{g}_{\text{task}} \leftarrow \nabla_{\theta} \mathcal{L}_{\text{task}}(\theta_{i,k-1}, x)$;

7) 根据式(12)进行逐样本梯度裁剪;

8) 根据式(13)聚合梯度并添加噪声;

9) 根据式(8)计算特征去相关梯度 $\mathbf{g}_{\text{decorr}} \leftarrow \nabla_{\theta} \mathcal{L}_{\text{decorr}}(\theta_{i,k-1}, B_k)$;

10) 根据式(10)计算特征残差梯度 $\mathbf{g}_{\text{res}} \leftarrow \nabla_{\theta} \mathcal{L}_{\text{res}}(\theta_{i,k-1}, B_k)$;

11) 根据式(14)更新本地模型参数 $\theta_{i,k}$;

12) 客户端上传本轮训练参数 $\theta_{i,i} \leftarrow \theta_{i,K}$;

13) **end for**

14) 中心服务器根据式(15)加权聚合客户端更新, 得到本轮全局模型 θ_i ;

15) **end for**

16) 根据

$$\epsilon = T \cdot K \cdot \left(\frac{1}{\alpha - 1} \log \left(1 + q^2 \cdot \frac{\alpha(\alpha - 1)C_g^2}{2\sigma^2} \right) \right) + \frac{\log(1/\delta)}{\alpha - 1}$$

计算隐私参数 (ϵ, δ) 。

5 隐私分析

在 DP-FL 框架中, 隐私损失的量化与分析是确保算法符合隐私保护要求的关键。本节基于 RDP 对提出的 DP-FedDRC 算法进行严格的隐私分析, 证明其满足 (ϵ, δ) -DP 保障。

引理1 高斯机制的 RDP 保证^[30]

若机制 $\mathcal{M}$ 对函数 f 添加噪声 $\mathcal{N}(0, \sigma^2 \Delta^2 \mathbf{I})$, 其中 Δ 为 f 的 L_2 -敏感度, 则 $\mathcal{M}$ 满足 $\left(\alpha, \frac{\alpha \Delta^2}{2\sigma^2} \right)$ -RDP。

引理2 子采样的 RDP 放大^[33]

设机制 $\mathcal{M}$ 满足 (α, ϵ) -RDP, 采用无放回均匀采样 (采样率 $q = \frac{|B|}{n}$), 则子采样后的机制 $\mathcal{M}_{\text{sub}}$ 的 RDP 阶 $\epsilon_{\text{sub}}(\alpha)$ 满足:

$$\epsilon_{\text{sub}}(\alpha) \leq \frac{1}{\alpha - 1} \log \left(1 + q^2 \cdot \frac{\alpha(\alpha - 1)}{2\sigma^2} \right) \quad (16)$$

引理3 RDP 组合性^[30]

若机制 $\mathcal{M}_1, \dots, \mathcal{M}_k$ 分别满足 (α, ϵ_i) -RDP,

则其组合机制满足 $\left(\alpha, \sum_{i=1}^k \epsilon_i \right)$ -RDP。

定理1 DP-FedDRC 的隐私保证

假设每个客户端在本地训练中采用采样率为 q 的无放回均匀采样, 梯度裁剪阈值为 C_g , 噪声标准差为 σ , 本地迭代次数为 K , 总训练轮数为 T 。则 DP-FedDRC 满足 (ϵ, δ) -DP, 其中全局隐私损失如式(17)所示。

$$\epsilon = T \cdot K \cdot \left(\frac{1}{\alpha - 1} \log \left(1 + q^2 \cdot \frac{\alpha(\alpha - 1)C_g^2}{2\sigma^2} \right) \right) + \frac{\log(1/\delta)}{\alpha - 1} \quad (17)$$

定理1的证明过程如下。

1) 单次梯度更新的 RDP 分析。在每次本地迭代中, 客户端进行无放回的均匀采样, 从本地数据集 D_i 中以概率 q 随机选择批次 B_k 。对每个样本梯度 $\mathbf{g}_k$ 进行 L_2 -范数裁剪 (此操作将敏感度限制为 $\Delta = C_g$), 并在聚合梯度时添加噪声 $\mathcal{N}(0, \sigma^2 C_g^2 \mathbf{I})$, 根据引理1, 单次更新满足 $\left(\alpha, \frac{\alpha C_g^2}{2\sigma^2} \right)$ -RDP。结合无放回均匀采样的隐私放大效应 (引理2), 单次迭代的 RDP 损失如式(18)所示。

$$\epsilon_{\text{iter}}(\alpha) \leq \frac{1}{\alpha - 1} \log \left(1 + q^2 \cdot \frac{\alpha(\alpha - 1)C_g^2}{2\sigma^2} \right) \quad (18)$$

2) 多轮训练的 RDP 组合。每轮本地训练包含 K 次迭代, 总训练轮数为 T 。根据引理3, 总隐私损失如式(19)所示。

$$\epsilon_{\text{total}}(\alpha) = T \cdot K \cdot \epsilon_{\text{iter}}(\alpha) \quad (19)$$

3) RDP 到差分隐私的转换。根据标准转换公式, 对于任意 $\delta > 0$, 选择最优 α , 使得:

$$\epsilon = \epsilon_{\text{total}}(\alpha) + \frac{\log(1/\delta)}{\alpha - 1} \quad (20)$$

通过严格地应用 RDP 框架下的子采样定理、高斯机制及组合性质, 证明 DP-FedDRC 在使用无放回均匀采样的设定下满足 (ϵ, δ) -DP 保障。

6 实验

本节通过系列实验验证 DP-FedDRC 框架在

缓解 DP-FL 中维度坍塌的有效性。

6.1 实验设置

硬件与软件环境：实验在配备 13 代 Intel Core i7-13700H 处理器、NVIDIA GeForce RTX 4060 Laptop GPU 及 8 GB 内存的硬件平台上完成，操作系统为 Windows 10，基于 PyTorch 框架实现。

数据集与模型：实验选用 MNIST^[34]、Fashion-MNIST^[35]和 CIFAR-10^[36]3 个主流基准数据集进行验证，其统计信息见表 3。此外，为模拟联邦学习中的数据 Non-IID 场景，采用狄利克雷分布（浓度参数 $\alpha = 0.1$ ）对客户端数据进行差异化划分。在模型架构方面，遵循相关研究^[37]使用 CNN 进行训练。

表 3 3 个数据集的统计信息

数据集	图像类型	类别数/个	样本总数/张
MNIST	灰度图像	10	70 000
Fashion-MNIST	灰度图像	10	70 000
CIFAR-10	彩色图像	10	60 000

基准方法：在 FedAvg（标准联邦平均算法）、FedProx 和 FedDecorr 上部署 DP-SGD，构建满足样本级隐私保护的基准模型。

参数设置：系统共配置 10 个客户端，MNIST 数据集的全局通信轮数 $T = 100$ ，Fashion-MNIST 和 CIFAR-10 数据集的全局通信轮数 $T = 200$ 。本地训练采用 Adam 优化器，迭代次数 $K = 20$ ，批次大小 $B = 64$ 。学习率配置为全局 $\eta_g = 1.0$ 、本地 $\eta_l = 0.0015$ 。特征去相关强度 $\lambda_d = 0.1$ 和特征残差约束强度 $\lambda_r = 10^{-5}$ 。噪声标准差 $\sigma = 1.0$ 和梯度裁剪阈值 $C_g = 1.0$ 。所有方法保持一致以满足 (ϵ, δ) -DP 约束（ $\delta = 10^{-3}$ ）。

6.2 维度坍塌水平评估

为评估不同算法在差分隐私约束下缓解维度坍塌的有效性，本节实验除固定噪声尺度 $\sigma = 1.0$ 外，其他设置与第 3 节保持一致。

如图 3 所示，DP-FedAvg 与 DP-FedProx 均出现显著的维度坍塌现象，表现为奇异值对数值整体下移且尾部急剧衰减，这与第 3 节揭示的差分隐私噪声与数据 Non-IID 负面协同效应一致。相

比之下，引入特征去相关正则化的 DP-FedDecorr 与本文提出的 DP-FedDRC 均能够有效维持特征空间的秩结构，且尾部衰减显著缓解。此外，DP-FedDRC 与 DP-FedDecorr 在奇异值谱的整体形态和尾部维持上高度一致，但 DP-FedDRC 引入的特征残差约束项通过强制深层特征继承浅层判别基元，从而避免了过度解耦导致的特征退化。这一设计为模型提供了更具判别力的表征先验，使其在后续分类任务中取得了实质性突破。

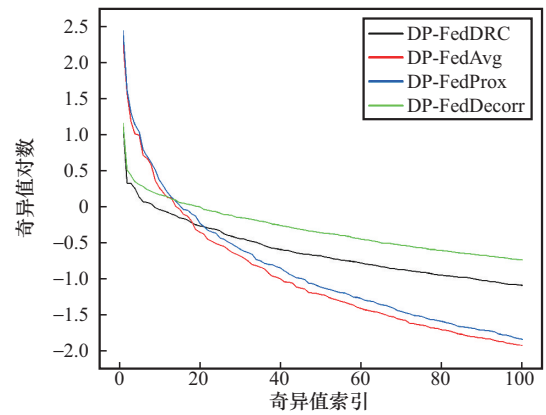


图 3 不同算法特征奇异值谱衰减对比

6.3 特征表示质量评估

为深入分析 DP-FedDRC 有效的根本原因，本节使用统一流行逼近与投影（UMAP）降维技术将算法学习到的高维特征映射至二维空间，以评估算法学习到的特征表示质量。如图 4 所示，与 DP-FedDecorr 相比，DP-FedDRC 算法提取的特征呈现出更清晰的类别聚类结构，同类样本紧密聚集，不同类别之间的边界清晰，这表明其能够学习到判别性更强的特征表示。

为量化这一差异，研究人员计算了平均类内距离、平均类间距离及类内与类间距离比，具体见表 4。结果显示，DP-FedDRC 在所有指标上均优于对比方法，尤其是获得了最低的类内与类间距离比，这说明其特征表示具有更高的类内一致性和类间区分度。

综上所述，DP-FedDRC 通过特征去相关和特征残差约束机制，有效缓解了维度坍塌，并显著增强了对判别性特征的保持能力，从而显著提升了模型在隐私保护下的分类性能。

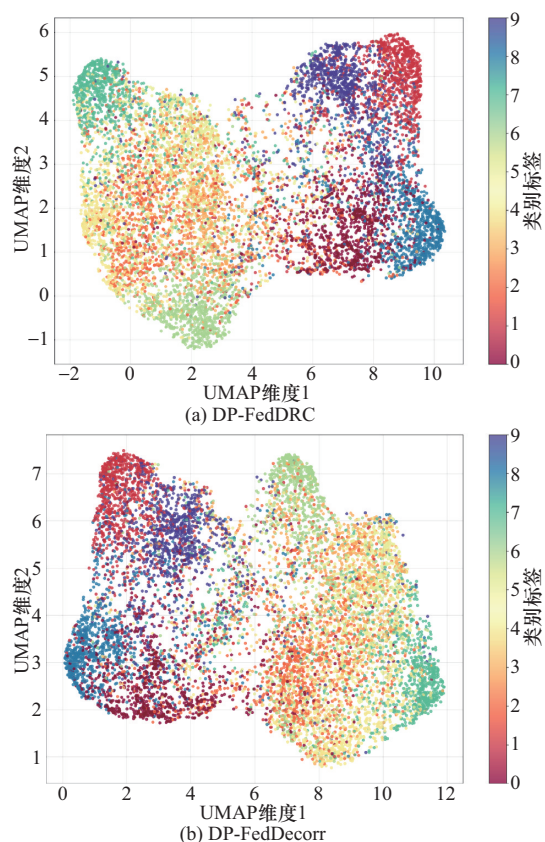


图4 UMAP特征可视化

表4 特征判别性评估结果

方法	平均类内距离	平均类间距离	类内与类间距离比
DP-FedDecorr	12.934 9	16.039 1	0.806 5
DP-FedDRC	11.681 2	15.944 5	0.732 6

6.4 差分隐私噪声下的模型性能

本节通过对比 DP-FedAvg、DP-FedDecorr、DP-FedProx 和 DP-FedDRC 4 种算法在 MNIST、Fashion-MNIST 和 CIFAR-10 数据集上的分类精度，系统验证了缓解维度坍塌对提升联邦学习模型性能的必要性，并揭示了 DP-FedDRC 算法的优越性。

如图 5(a) 所示，在 MNIST 手写数字识别任务中，经过 100 轮通信迭代后，DP-FedDRC 的最终分类精度达到 94.28%，较 DP-FedDecorr (91.89%) 提升 2.39%，较 DP-FedAvg (89.86%) 提升 4.42%，较 DP-FedProx (88.52%) 提升 5.76%。如图 5(b) 所示，在 Fashion-MNIST 衣物图像分类任务中，经过 200 轮训练后，DP-FedDRC 的分类精度达 82.9%，

较 DP-FedDecorr (81.74%) 提升 1.16%，较 DP-FedAvg (80.72%) 提升 2.18%，较 DP-FedProx (81.41%) 提升 1.49%。如图 5(c) 所示，在最具挑战性的 CIFAR-10 高维图像分类任务中，DP-FedDRC 在 200 轮训练后最终分类精度达到 64.26%，较 DP-FedDecorr (59.53%) 提升 4.73%，较 DP-FedAvg (57.65%) 提升 6.61%，较 DP-FedProx (56.71%) 提升 7.55%。

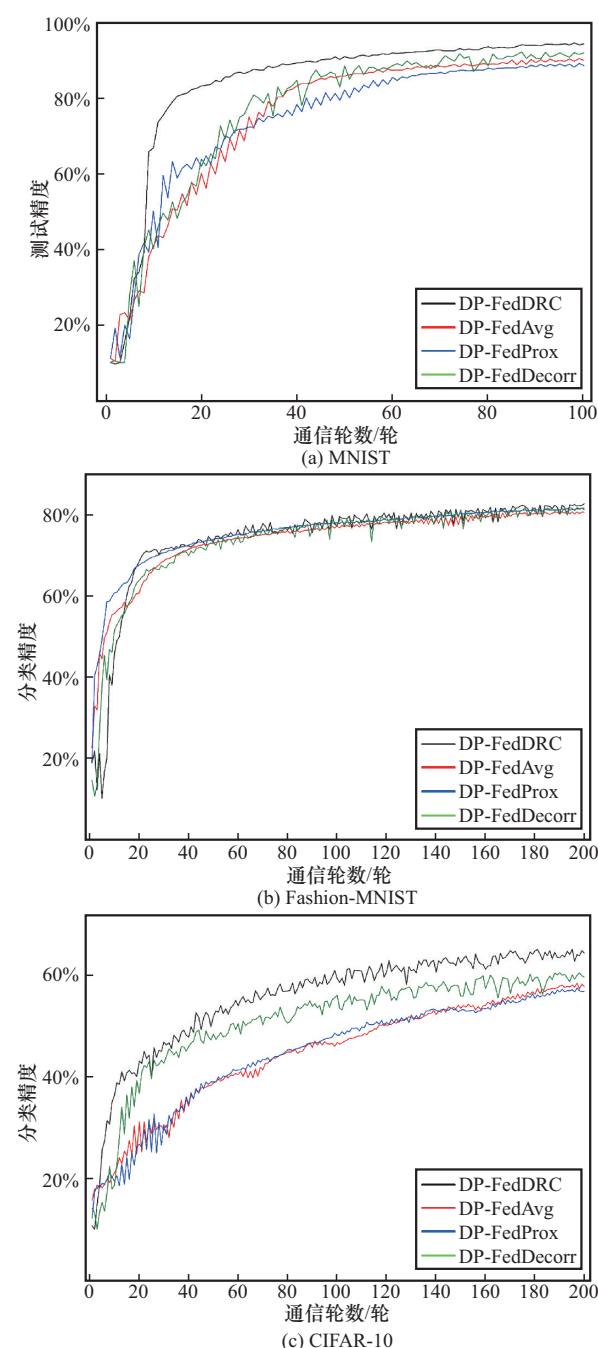


图5 各算法在3个数据集上的分类精度

实验结果表明,与基线算法相比,DP-FedDRC在3类数据集上均展现出最优性能,始终收敛于最高精度,其在高维、复杂视觉任务(如CIFAR-10)中的优势尤为突出。在MNIST等类别边界清晰、特征维度较低的简单任务中,维度坍塌效应相对较弱,因此各类算法的性能差异不大;而在CIFAR-10这类高维图像任务中,原始特征空间维度高、类别边界模糊,差分隐私机制所引入的噪声极易放大由数据Non-IID引发的局部梯度冲突,进而导致严重的特征退化。此时,DP-FedDRC通过双重正则化机制实现协同优化:一方面抑制由隐私噪声和数据异质性诱发的特征维度间冗余相关性;另一方面保留对判别性至关重要的语义信息,从而有效缓解过度去相关可能带来的表示失真。这一分析与第6.3节中基于UMAP的可视化结果、类内紧凑性与类间可区分性指标一致,进一步验证了DP-FedDRC所学习特征在结构上的优越性,并为其分类性能提升提供了直接支撑。

6.5 计算效率评估

(1) 理论分析

DP-FedDRC在提升模型性能的同时,需要额外计算特征去相关正则化项 $\mathcal{L}_{\text{decorr}}$ 和特征残差约束项 $\mathcal{L}_{\text{res}}$ 。

1) 特征去相关正则化项 $\mathcal{L}_{\text{decorr}}$: 该项的计算涉及求解批次特征相关矩阵的非对角线元素平方和,其时间复杂度为 $O(d^2)$,其中 d 为特征维度。

2) 特征残差约束项 $\mathcal{L}_{\text{res}}$: 该项需要计算浅层特征 $\mathbf{z}_s$ 与深层特征 $\mathbf{z}$ 的投影残差,时间复杂度为 $O(d \times d_s)$,其中 d_s 为浅层特征维度,通常 $d_s \ll d$ 。

因此,相较于基线算法DP-FedAvg(仅需计算主任务损失梯度),DP-FedDRC引入的额外计算复杂度为 $O(d^2 + d \cdot d_s)$ 。在联邦学习的典型设定中,特征维度 d 是固定的,且其数量级通常远小于模型的总参数量,故此项额外开销对整体计算成本的影响有限。

(2) 耗时对比

在相同硬件环境及超参数设置下,各算法单轮训练平均耗时对比如图6所示。其中,DP-FedProx在MNIST、Fashion-MNIST和CIFAR-10数据集上相比DP-FedAvg额外计算开销分别增加33.8%、56.5%和48.3%,而DP-FedDRC额外计

算开销仅分别增加8.3%、9.9%和9.1%,显著低于DP-FedProx,略微高于DP-FedDecorr。这一微小代价换来了显著的性能提升,证明算法具备优秀的效用-开销权衡能力。

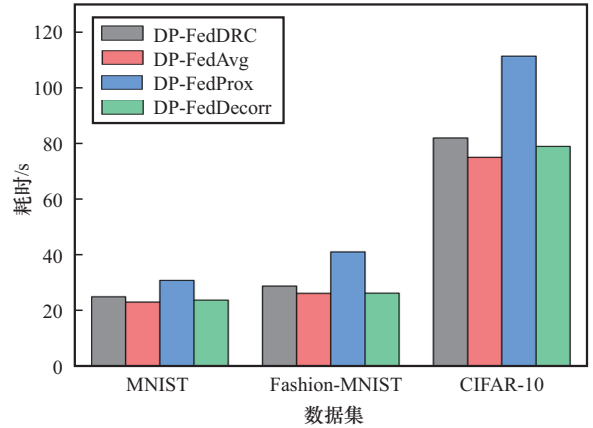


图6 各算法单轮训练平均耗时对比

6.6 特征残差约束参数的消融研究

在前文研究基础上探讨特征残差约束强度对模型性能的影响,实验中分别设置特征残差约束强度 $\lambda_r = \{0, 1 \times 10^{-6}, 1 \times 10^{-5}, 1 \times 10^{-4}\}$,其中 $\lambda_r = 0$ 对应无残差约束的基线版本(即DP-FedDecorr算法)。 λ_r 消融实验结果如图7所示,从中可以看出,DP-FedDRC算法在MNIST、Fashion-MNIST、CIFAR-10 3个数据集上的分类精度随 λ_r 的变化整体呈现先上升后下降的单峰趋势。综合各数据集的性能表现, $\lambda_r = 1 \times 10^{-5}$ 为最优约束强度,表明适当的特征残差约束能够在模型训练中保留关键判别性特征,从而有效提升模型分类精度。

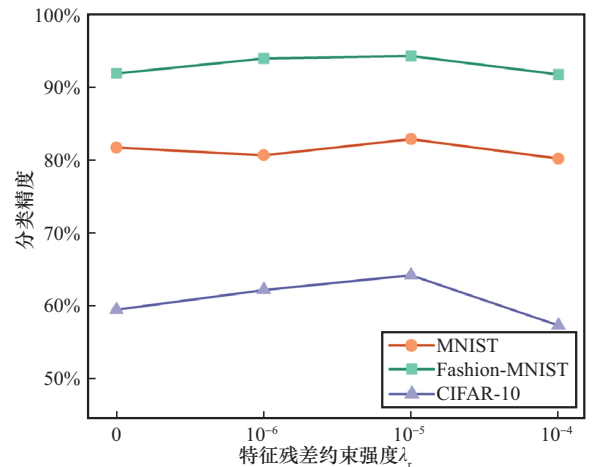


图7 λ_r 消融实验结果

7 结束语

本文针对个性化联邦学习中差分隐私噪声注入加剧维度坍塌的问题,提出了DP-FedDRC。该框架通过在本地训练中引入特征去相关与特征残差约束机制,在满足严格差分隐私保障的同时,能够有效缓解特征空间退化,从而在可控计算开销下获得优越性能。本文的研究为解决隐私保护与个性化联邦学习模型性能的协同优化提供了有效的解决方案。然而,本文研究尚存在一定的局限性:其一,隐私保障基于理论分析,未能在成员推理或梯度反演等真实攻击下评估模型的鲁棒性;其二,未考虑通信效率及边缘设备资源约束,限制了实际部署潜力。未来工作将开展多类型隐私攻击下的实证验证,优化算法轻量化以适配资源受限场景,并探索其在多模态联邦学习中的适用性。

参考文献:

- [1] McMahan H B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Proceedings of the International Conference on Artificial Intelligence and Statistics. 2016.
- [2] Yang Q, Liu Y, Chen T J, et al. Federated machine learning: concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.
- [3] Kairouz P, McMahan H B. Advances and open problems in federated learning[J]. Foundations and Trends in Machine Learning, 2021, 14(1): 1-210.
- [4] Zhao Y, Li M, Lai L Z, et al. Federated learning with non-IID data[J]. 2018, arXiv: 1806.00582.
- [5] 周传鑫, 孙奕, 汪德刚, 等. 联邦学习研究综述[J]. 网络与信息安全学报, 2021, 7(5): 77-92.
Zhou C X, Sun Y, Wang D G, et al. A summary of federal learning research[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 77-92.
- [6] 高莹, 陈晓峰, 张一余, 等. 联邦学习系统攻击与防御技术研究综述[J]. 计算机学报, 2023, 46(9): 1781-1805.
Gao Y, Chen X F, Zhang Y Y, et al. Summary of research on attack and defense technology of federated learning system[J]. Chinese Journal of Computers, 2023, 46(9): 1781-1805.
- [7] Shokri R, Stronati M, Song C Z, et al. Membership inference attacks against machine learning models[C]//Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 3-18.
- [8] Suri A, Kanani P, Marathe V J, et al. Subject membership inference attacks in federated learning[J]. 2022, arXiv: 2206.03317.
- [9] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2015: 1322-1333.
- [10] Wang X H, Wang S, Li Y W, et al. Differentially private and heterogeneity-robust federated learning with theoretical guarantee[J]. IEEE Transactions on Artificial Intelligence, 2024, 5(12): 6369-6384.
- [11] Liang J T, Su S. DP-FedEwc: Differentially private federated elastic weight consolidation for model personalization[J]. Knowledge-Based Systems, 2024, arXiv: 303. 112401.
- [12] Shi Y J, Liang J, Zhang W Q, et al. Towards understanding and mitigating dimensional collapse in heterogeneous federated learning[C]//Proceedings of the 11th International Conference on Learning Representations (ICLR). 2023: 9820-9843.
- [13] 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, 2016: 308-318.
- [14] Li T, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks[C]//Proceedings of Machine Learning and Systems. 2020, 2: 429-450.
- [15] Karimireddy S P, Kale S, Mohri M, et al. SCAFFOLD: stochastic controlled averaging for federated learning[C]//Proceedings of the 37th International Conference on Machine Learning. 2020: 5132-5143.
- [16] Durmus A E, Yue Z, Ramon M, et al. Federated learning based on dynamic regularization[C]//Proceedings of the 9th International Conference on Learning Representations (ICLR). 2021: 790-825.
- [17] Asad M, Moustafa A, Ito T. FedOpt: towards communication efficiency and privacy preservation in federated learning[C]//Applied Sciences, 2020, 10(8): 2864.
- [18] Wang S, Xu Y Q, Wang Z G, et al. Beyond ADMM: a unified client-variance-reduced adaptive federated learning framework[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 10175-10183.
- [19] Girgis A, Data D, Diggavi S, et al. Shuffled model of differential privacy in federated learning[C]//Proceedings of the 24th International Conference on Artificial Intelligence and Statistics. 2021: 2521-2529.
- [20] Wei W Q, Liu L, Wu Y Z, et al. Gradient-leakage resilient federated learning[C]//Proceedings of the 2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS). Piscataway: IEEE Press, 2021: 797-807.
- [21] Fu J, Chen Z L, Han X. Adap DP-FL: differentially private federated learning with adaptive noise[C]//Proceedings of the 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Piscataway: IEEE Press, 2023: 656-663.
- [22] Chen L, Ding X F, Bao Z F, et al. Differentially private federated learning on non-iid data: convergence analysis and adaptive optimization[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(9): 4567-4581.

- [23] Wei K, Li J, Ding M, et al. Federated learning with differential privacy: algorithms and performance analysis[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3454-3469.
- [24] Noble M, Bellet A, Dieuleveut A. Differentially private federated learning on heterogeneous data[C]//Proceedings of the 25th International Conference on Artificial Intelligence and Statistics. 2022: 10110-10145.
- [25] Li Y W, Chang T H, Chi C Y. Secure federated averaging algorithm with differential privacy[C]//Proceedings of the 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP). Piscataway: IEEE Press, 2020: 1-6.
- [26] Feng J, Yang L T, Ren B C, et al. Tensor recurrent neural network with differential privacy[J]. IEEE Transactions on Computers, 2024, 73(3): 683-693.
- [27] Zhao P N, Zhang Z K, Dong J W, et al. An attack-agnostic defense framework against manipulation attacks under local differential privacy[C]//Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2025: 3858-3876.
- [28] Zhang P F, Sun H, Zhang Z K, et al. Privacy-preserving recommendations with mixture model-based matrix factorization under local differential privacy[J]. IEEE Transactions on Industrial Informatics, 2025, 21(7): 5451-5459.
- [29] Dwork C, Roth A. The algorithmic foundations of differential privacy[M]. Hanover: Now Publishing, 2014: 211-407.
- [30] Mironov I. Rényi differential privacy[C]//Proceedings of the 2017 IEEE 30th Computer Security Foundations Symposium (CSF). Piscataway: IEEE Press, 2017: 263-275.
- [31] Cogswell M, Ahmed F, Girshick R, et al. Reducing overfitting in deep networks by decorrelating representations[J]. 2015, arXiv: 1511.06068.
- [32] Xiong W, Du B, Zhang L F, et al. Regularizing deep convolutional neural networks with a structured decorrelation constraint[C]//Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM). Piscataway: IEEE Press, 2017: 519-528.
- [33] Wang Y X, Balle B, Kasiviswanathan S P. Subsampled Rényi differential privacy and analytical moments accountant[C]//Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. 2019: 1226-1235.
- [34] 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.
- [35] Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms[J]. 2017, arXiv: 1708.07747.
- [36] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[R]. Toronto: University of Toronto, 2009.
- [37] Papernot N, Thakurta A, Song S, et al. Tempered sigmoid activations for deep learning with differential privacy[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(10): 9312-9321.

[作者简介]



王民利 (2000-), 男, 山西运城人, 海南大学硕士生, 主要研究方向为联邦学习、隐私保护。



张宇浩 (1998-), 男, 安徽宿州人, 海南大学博士生, 主要研究方向为安全多方计算、隐私保护机器学习。



胡昌慧 (1985-), 男, 山东莱西人, 海南大学教授, 博士生导师, 主要研究方向为安全多方计算、差分隐私和隐私保护机器学习。