中国科学:信息科学 2026 年 第 56 卷 第 6 期: 1364–1389
SCIENTIA SINICA Informationis
论文

《中国科学》杂志社
SCIENCE CHINA PRESS
CrossMark
click for updates

联邦学习中基于数据选择的模型修复方法研究

支宇晗[1], 胡强[2], 谢肖飞[3], 王隆天[1], 胡铭[3], 马雷[4], 管晓宏[1], 沈超[1*]

1. 西安交通大学网络空间安全学院, 西安 710049, 中国
2. 天津大学智能与计算学部, 天津 300072, 中国
3. School of Computing and Information Systems, Singapore Management University, Singapore 188065, Singapore
4. Department of Information Science, The University of Tokyo, Tokyo 100-0001, Japan
* 通信作者. E-mail: chaoshen@mail.xjtu.edu.cn

收稿日期: 2026–01–25; 修回日期: 2026–03–15; 接受日期: 2026–04–21; 网络出版日期: 2026–05–20

国家重点研发计划 (批准号: 2023YFB3107400)、国家自然科学基金 (批准号: 62521002, U24B20185, U2441240, 62132011) 和新基石科学基金会与科学探索奖资助项目

摘要 联邦学习能够在不共享原始数据的前提下实现多客户端协同建模, 为隐私敏感场景下的模型训练提供有效解决方案. 然而, 在实际部署过程中, 联邦学习模型不可避免地会接触到分布未知的新数据, 从而导致模型性能随时间推移而下降. 如何在较低维护成本下有效修复性能退化的联邦学习模型, 仍是一个亟待解决的问题. 针对上述挑战, 本文系统研究了联邦学习场景下的基于数据选择的模型修复方法, 重点分析其在模型维护中的适用性与有效性. 具体而言, 我们首先从查询选择器与数据选择策略两个维度出发, 比较了以本地模型和全局模型作为查询选择器的差异, 并评估了多种代表性数据选择方法在联邦学习模型修复任务中的性能表现. 在 7 个数据集和 6 种不同模型架构上的实验结果表明, 当前联邦学习场景下基于数据选择的模型修复仍处于探索阶段, 现有方法并不能带来稳定的性能增益. 与以往联邦主动学习中的结论不同, 全局模型在补丁数据选择过程中表现出更优的修复效果, 表明该问题在本质上不同于联邦主动学习. 进一步的分析发现, 所选补丁数据在标签分布上保持均衡有助于提升模型修复效果. 基于上述实证发现, 我们进一步提出了一种面向联邦学习的数据选择方法 Fed-PAIR, 该方法通过显式建模成对不确定性并结合当前训练数据的分布状态, 自适应调整补丁数据的标签分配, 从而促进整体训练数据在客户端之间趋于更加均衡的分布. 实验结果表明, Fed-PAIR 在数据独立同分布与非独立同分布场景下均能够取得稳定的性能提升, 在严重标签偏斜条件下也能保持良好的鲁棒性. 上述研究不仅为联邦学习模型的高效维护提供了系统性的实证分析, 也表明针对联邦学习各端数据非独立同分布场景设计专门的数据选择策略, 是提升模型修复效果的一条有效途径.

关键词 深度学习, 数据选择, 模型修复, 联邦学习, 数据分布

1 引言

近年来, 联邦学习作为一种有效的分布式建模技术逐渐受到广泛关注, 并已在数字医疗 [1,2]、无线通信 [3]、物联网 [4] 以及移动边缘计算 [5] 等多个领域得到应用. 联邦学习的核心设计思想是在客户端利用私有数据训练

引用格式: 支宇晗, 胡强, 谢肖飞, 等. 联邦学习中基于数据选择的模型修复方法研究. 中国科学: 信息科学, 2026, 56: 1364–1389, doi: 10.1360/SSI-2026-0024
Zhi Y H, Hu Q, Xie X F, et al. Data-selection-based model repair for federated learning. Sci Sin Inform, 2026, 56: 1364–1389, doi: 10.1360/SSI-2026-0024

© 2026《中国科学》杂志社 www.scichina.com infocn.scichina.com
https://www.sciengine.com/doi/10.1360/SSI-2026-0024

多个深度神经网络 (deep neural networks, DNNs), 并由服务器端对这些模型进行聚合以生成全局模型, 从而完成特定学习任务. 联邦学习的主要目标是在充分利用各客户端私有数据的同时, 避免直接共享原始数据, 以保护客户端的隐私信息. 近年来, 随着大语言模型的发展, 联邦学习在保护高价值且获取成本极高的数据方面的重要性进一步凸显 [6~8], 使其在学术研究和工业应用中均展现出更大的潜力.

尽管联邦学习在多个领域展现出良好的应用前景, 其在实际部署过程中仍面临着与其他深度学习技术相同的核心挑战之一, 即数据分布偏移 [9]. 当测试阶段的数据分布与训练阶段的数据分布不一致时, 模型性能往往会出现显著下降, 甚至引发严重的安全隐患. 例如, 在自动驾驶场景中, 数据分布偏移极易导致严重事故 [10~12]. 数据分布偏移, 尤其是训练数据与真实环境测试数据之间的差异, 是深度学习领域长期存在的关键问题, 该问题对联邦学习模型的实际部署同样具有显著影响. 因此, 在模型部署之后对联邦学习模型进行持续维护, 以缓解分布偏移并确保模型性能持续满足用户需求, 显得尤为必要.

对深度学习模型而言, 通过引入新数据对模型进行重新训练是一种直接且有效的模型修复方式, 可用于应对数据分布变化. 然而, 该策略面临两个主要挑战: 一方面, 原始数据的采集过程通常较为容易且高度自动化, 但对这些数据进行人工标注仍然代价高昂; 另一方面, 使用全部数据对模型进行重新训练会带来显著的计算开销. 这些因素限制了基于完全重训练的模型修复方法在联邦学习场景中的直接应用. 针对上述问题, 研究者提出了基于数据选择的模型修复方法 (data selection-based model repair, DSMR) [13], 通过从大规模未标注数据池中选择一小部分具有代表性的样本作为补丁数据进行标注, 并结合原始训练数据对模型进行再训练, 从而在可控的标注预算与计算资源消耗下完成模型修复. 然而, 现有大多数 DSMR 方法主要针对单一深度学习模型进行设计. 相比之下, 联邦学习场景更加复杂, 客户端侧同时涉及本地模型与服务器端聚合得到的全局模型, 不同客户端的本地训练数据往往呈高度非独立同分布 (non-independent and identically distributed, Non-IID), 这为 DSMR 的直接应用带来了新的挑战.

在联邦学习场景下, 一个尚未得到系统回答的核心问题是: 在基于数据选择的模型修复过程中, 应由本地模型还是全局模型作为查询选择器, 才能更有效地提升模型在分布偏移场景下的性能? 此外, 训练数据的分布形式, 如独立同分布 (independent and identically distributed, IID) 与非独立同分布, 被认为是影响联邦学习模型性能的重要因素, 但其对联邦学习模型修复过程的具体影响仍缺乏系统性的实证分析. 由于联邦学习系统的复杂性, 现有研究尚未充分揭示 DSMR 方法在联邦学习场景下的实际有效性. 因此, 在将基于数据选择的模型修复方法应用于实际联邦学习系统之前, 有必要对上述问题进行系统而深入的研究.

为填补上述研究空白, 本文首先开展了一项系统的实证研究, 全面评估 DSMR 方法在联邦学习场景下的适用性与有效性. 我们在 7 个数据集、6 种模型架构以及同分布与分布偏移测试场景下进行了大规模实验, 重点围绕以下 3 个问题展开分析.

(1) 查询选择器对 DSMR 的影响. 实验结果表明, 与联邦主动学习不同, 在分布偏移场景下全局模型通常是更有效的查询选择器.

(2) 不同 DSMR 方法在联邦学习中的表现差异. 实验结果显示, 联邦学习中的 DSMR 仍处于初期阶段, 多数方法仅能带来有限且不稳定的性能增益, 难以在复杂的 Non-IID 与分布偏移条件下稳定优于随机选择. 进一步分析发现, 在多数设置下, 表现相对较为可靠的方法均属于基于不确定性度量的策略, 表明以模型预测不确定性作为核心驱动信号仍然是数据选择过程中不可或缺的关键因素.

(3) 标签分布对模型修复效果的影响. 探索性分析表明, 所选补丁数据的标签分布均衡性与模型修复效果密切相关, 而现有 DSMR 方法倾向选择偏置数据, 普遍缺乏对补丁数据与原始训练数据联合分布结构的显式约束.

基于上述实证发现, 本文进一步提出了一种面向联邦学习的数据选择方法 Fed-PAIR. 该方法同时考虑模型预测不确定性和整体数据标签均衡性, 并显式结合当前训练数据的分布状态, 自适应调整补丁数据的标签分配策略, 从而在保持不确定性驱动选择的同时, 促进补丁数据与原始训练数据在客户端间的协同再均衡. 实验结果表明, Fed-PAIR 在 IID 与 Non-IID 场景下均能够取得稳定的性能提升, 并在严重标签偏斜条件下仍能保持较为良好的鲁棒性.

综上, 本文的主要贡献包括: (1) 首次对联邦学习场景下的 DSMR 方法进行了系统性的实证评估, 揭示了查

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

询选择器、不确定性驱动策略与数据分布结构对模型修复效果的关键影响因素; (2) 基于实证结论提出了一种面向联邦 Non-IID 场景的数据选择方法 Fed-PAIR, 通过不确定性驱动与分布再均衡的联合约束, 有效改善了模型修复的稳定性与鲁棒性; (3) 构建并开源了一套可复现的实验框架, 可在不同模型架构、数据分布与分布偏移场景下评估模型修复方法, 为后续研究提供标准化基准与工具支持 (https://sites.google.com/view/smrforfl).

2　相关背景知识及研究现状

2.1　联邦学习与联邦主动学习

联邦学习 [14~16] 是一种广泛使用的分布式机器学习范式. 通常, 联邦学习由云服务器和多个客户端组成. 在每轮训练中, 云服务器首先将全局模型下发至各客户端进行本地训练, 然后每个客户端使用自身私有数据训练收到的模型并将训练后的模型上传至云服务器, 最后服务器对收集到的本地模型进行聚合以更新全局模型. 服务器端的客户端模型聚合是联邦学习的核心问题, 已有多种聚合方法被提出 [17~19]. 其中 FedAvg [14] 被认为是标准方法, 本文工作采用 FedAvg 进行联邦学习聚合. 简而言之, FedAvg 将各本地模型的参数进行平均, 以形成全局模型.

现有大多数联邦学习研究假设数据已完全标注, 但在实际中标注成本高昂, 客户端通常同时拥有标注数据和未标注数据 [20~23]. 主动学习 [24,25] 通过选择信息量大的样本进行标注, 有效缓解了数据不足问题 [26,27]. 联邦主动学习将联邦学习与主动学习结合, 以利用客户端间丰富的未标注数据 [28~31]. 早期研究 [32~34] 直接将主动学习策略应用于联邦学习, 而 LoGo [35] 通过聚类与不确定性引导采样改进了联邦主动学习方法.

2.2　基于数据选择的模型修复方法

基于数据选择的模型修复 (DSMR) 通过策略性地选取和标注有限测试数据, 对深度神经网络进行再训练或微调, 从而提升模型性能. 先前研究已探索了多种策略: 利用误分类样本 [36~40]、差分分析 (MODE) [41]、边界聚类 (MCP) [42]、基于损失的鲁棒性测试 (RobOT) [43], 以及分布感知选择 (DAT) [44]. HybridRepair [13] 将半监督学习与 DSMR 结合, SAFE [45] 利用故障特征提取, MOTS [46] 则结合不确定性和多样性进行多目标优化. 与这些针对单模型的研究不同, 本文首次将 DSMR 引入联邦学习, 并评估其在联邦学习模型上的有效性.

此外, 部分联邦学习研究也涉及数据选择, 但关注问题不同: 文献 [47] 从头选择训练数据, 文献 [48] 侧重客户端到服务器的聚合, 文献 [49] 针对存储受限的主动训练. 这些工作属于模型开发阶段, 而本文的 DSMR 方法针对模型维护阶段, 用于应对部署后联邦模型的分布偏移.

3　实验设计

为了探究基于数据选择的模型修复在联邦学习中的有效性, 本文开展了系统性实证研究, 分析不同查询选择器对 DSMR 的影响及现有 DSMR 方法的性能表现.

3.1　整体流程

图 1 展示了本文实证研究设计的概览. 本研究的框架主要由以下步骤组成.

步骤 1: 特征提取. 在每个客户端接收到新的未标注数据时, 现有 DSMR 方法首先利用 DNN 模型 (即查询选择器) 的内部层或输出层提取有用的特征, 并基于这些特征进行优先级排序. 特征提取的质量直接影响排序结果, 因此对 DSMR 至关重要. 在联邦学习环境中, 每个客户端可使用两种查询选择器进行 DSMR: 客户端本地训练的模型和服务器聚合后返回的全局模型. 因此, 首先需要探索哪种模型更适合作为查询选择器. 在后续步骤中, 分别使用上述两种模型在每个客户端提取数据特征.

步骤 2: 数据选择. 获得由不同模型提取的数据特征后 (即全局模型与本地模型), 应用 DSMR 方法对未标注数据集中的样本进行排序, 评估其修复价值, 并根据标注预算选取 Top-K 样本作为补丁数据.

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

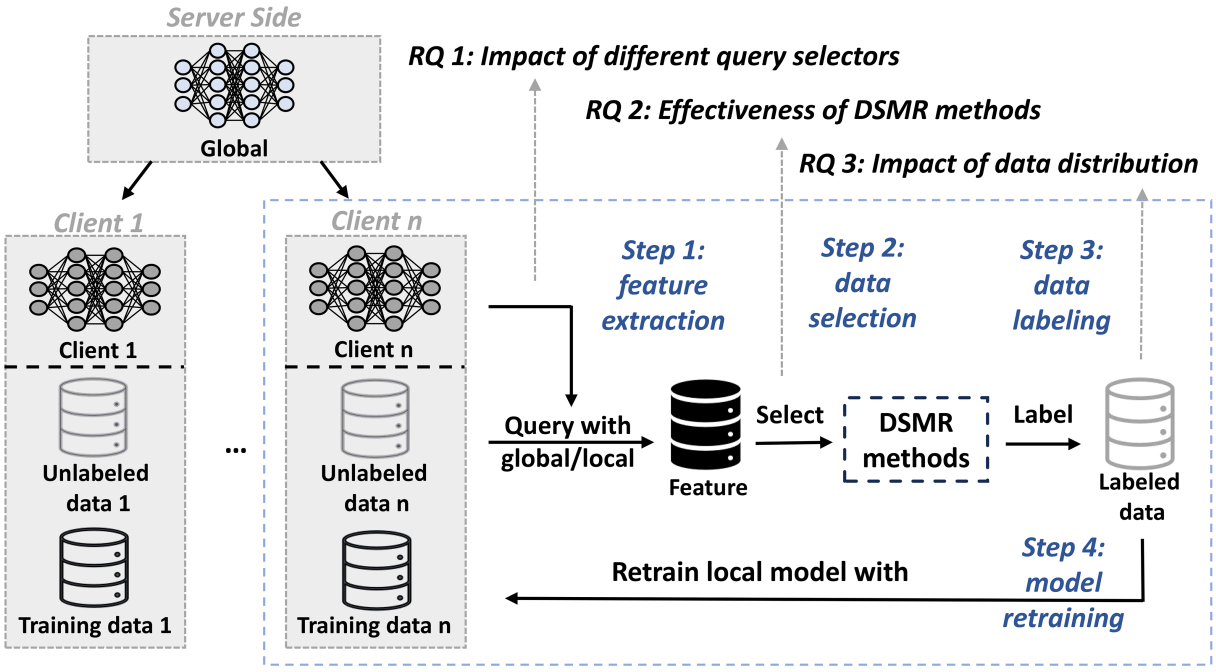


图 1　(网络版彩图) 框架设计概览.

Figure 1　(Color online) Overview of the framework design.

步骤 3: 数据标注. 对选中的数据进行标注, 并将其作为补丁数据用于模型修复.

步骤 4: 模型再训练. 最后, 每个客户端将标注后的新数据与原始训练数据合并, 对本地模型进行少量轮次的再训练. 训练完成后, 客户端更新的本地模型在服务器端进行聚合, 形成新的全局模型. 通过对比再训练前后的全局模型性能, 来评估 DSMR 的有效性.

3.2　数据集与模型

数据集. 本文选取了 7 个广泛使用的数据集, 涵盖图像分类任务和自然语言处理任务: Fashion-MNIST [50], PathMNIST [51], CIFAR-10 [52], CIFAR-100 [52], Devign [53], 20 Newsgroups [54] 和 Sentiment140 [55]. Fashion-MNIST 包含 10 类灰度服饰图像 (如大衣、衬衫). PathMNIST 来源于结直肠癌组织切片, 包含 9 类组织 (如脂肪、淋巴细胞). CIFAR-10 包含 10 类彩色图像 (如飞机、鸟类). CIFAR-100 包含 100 类彩色图像, 分为 20 个大类, 每张图像提供 “细粒度” 标签和 “粗粒度” 标签 (如 “dolphin” 为细粒度标签, “aquatic mammals” 为粗粒度标签). 上述 4 个数据集在计算机视觉领域研究广泛. 此外, 我们考虑自然语言处理领域的两个数据集: 20 Newsgroups 包含 20 个新闻组主题的文本数据 (如 space, baseball), Sentiment140 包含通过 Twitter API 收集的推文, 常用于情感分析 (正面或负面). 另外, 本文还包含与代码相关的数据集 Devign, 该数据集的构建方式为: 收集与安全相关的代码提交记录, 并从中提取所标注的易受攻击函数和非易受攻击函数. 值得一提的是, 我们是较早探索联邦学习在代码领域应用的研究之一.

模型. 为保证模型架构的多样性, 本文考虑前馈深度神经网络 (feed-forward DNNs)、预训练 Transformer 模型以及循环神经网络 (RNNs). 基于前人工作 [40, 44, 56, 57], 我们选择在这些数据集上具有较高准确率的模型: Fashion-MNIST 与 PathMNIST 采用 LeNet-5 [58], CIFAR-10 采用 ResNet-20 [59], CIFAR-100 采用 ResNet-50 [59], Devign 采用 CodeBERT [60], 20 Newsgroups 采用简单多层感知机 (MLP) [61], Sentiment140 采用长短期记忆网络 (long short-term memory, LSTM) [62]. 表 1 总结了本文使用的数据集与模型信息, 更详细的模型信息可见我们的开源网站 [63].

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

表 1　数据集与模型信息，包括训练数据集和测试数据集样本数量，在不同训练数据分布形式下训练得到的全局模型在测试数据集上的准确率. 参数 β 用于控制数据分布的不均衡程度，β 越小，数据分布越不均衡.

Table 1 Details of datasets and models, including the number of samples in the training and test datasets, and the accuracy of the global model on the test dataset under different training data distributions. The parameter β controls the degree of data distribution imbalance, where a smaller β indicates a more imbalanced distribution.

Dataset	#Training	#Test	Model	Accuracy (%)		
				IID	Non IID ($\beta = 1.0$)	Non IID ($\beta = 0.1$)
Fashion-MNIST	60000	10000	LeNet-5	90.28	88.79	78.62
PathMNIST	89996	10004	LeNet-5	87.99	80.23	61.86
CIFAR-10	50000	10000	ResNet-20	96.22	71.56	45.78
CIFAR-100	50000	10000	ResNet-50	78.62	40.30	39.64
Devign	21000	2700	CodeBERT	61.13	59.70	55.16
20 Newsgroups	11314	7532	MLP	66.70	63.08	29.29
Sentiment140	1120000	480000	LSTM	74.16	73.35	58.36

3.3　联邦学习设置

考虑到 FedAvg [14] 是最常用的联邦学习流程, 本文沿用之前工作的研究设置 [33,35], 采用 FedAvg 作为联邦学习聚合方法. 本文设定客户端数量 N 为 10, 客户端参与比例 γ 为 1.0, 即每轮训练所有客户端均参与. 本地训练轮数 E 设为 3, 与开源项目 [64] 默认值一致. 对于图像数据集, 我们将联邦学习模型训练 50 轮, 全局模型的训练日志如图 2 所示. 根据训练日志, 除 IID 设置下的 CIFAR-100 外, 50 轮足以使模型收敛. 在 IID 设置下的 CIFAR-100 数据集中, 50 轮训练会导致全局模型与本地模型同时出现过拟合现象, 因此我们对该设置下的全局模型与本地模型统一选用第 10 轮的模型作为查询选择器, 以确保两类查询选择器之间的对比公平性. 其他所有设置均使用最后一轮的模型. 需要指出的是, 上述特殊处理仅针对 IID 设置下的 CIFAR-100 数据集, 本文各实验结果均在相同数据集内部对不同方法进行横向比较, 不涉及跨数据集的性能对比, 因此该处理不会对整体实验结论的公平性产生影响. 对于代码和文本数据集, 我们将联邦学习模型训练 5 轮, 并选择最后一轮的全局/本地模型作为查询模型. 图 2 中全局模型训练日志验证了 5 轮训练足以收敛.

3.4　模型修复设置

在实验中, 我们通过选取的补丁数据对模型进行再训练以实现模型修复. 参照之前研究工作 [42,44,65] 的设置, 在再训练过程中采用与初始联邦学习训练相同的超参数设置, 比如优化器、学习率、动量、权重衰减和损失函数. 本文设定标注预算为 1%, 5% 和 10%, 并从候选数据中选取相应数量的数据, 将其加入客户端原始训练数据进行再训练. 每个本地模型再训练 5 轮后, 聚合形成再训练的全局模型. 为降低随机性影响, 每次再训练会使用相同的被选补丁数据重复三次. 本文实验结果报告的是不同案例数量和多次重复下的平均准确率.

3.5　数据准备

本文包含多种类型的数据, 以全面探索面向联邦学习模型的 DSMR 问题. 针对用于再训练的数据, 本文同时考虑 IID 与 Non-IID 设置, 并分别为原始训练数据与候选数据准备相应的数据划分. 在模型测试阶段, 除了评估联邦学习模型在分布内 (in-distribution, ID) 测试数据上的性能外, 我们还在分布偏移的测试数据上评估模型性能, 为简化表述, 将其统一称为分布外 (out-of-distribution, OOD) 测试数据, 以分析再训练模型的泛化能力.

原始数据划分. 为获得预训练的联邦学习模型, 我们将原始训练数据集中的 90% 作为原始训练数据, 剩余 10% 作为分布内测试数据, 用于联邦学习模型的训练与评估.

新数据准备. 为模拟部署后的联邦学习模型可能遇到的新环境, 我们在原始数据基础上引入分布偏移, 构造新的数据集用于模型再训练与再测试. 对于图像数据集, 我们采用图像扰动方法生成新的候选数据和测试数据. 对 CIFAR-10 数据集, 我们直接使用 CIFAR-10-C [66] 作为新数据. CIFAR-10-C 是一个基于 CIFAR-10 构建的经

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

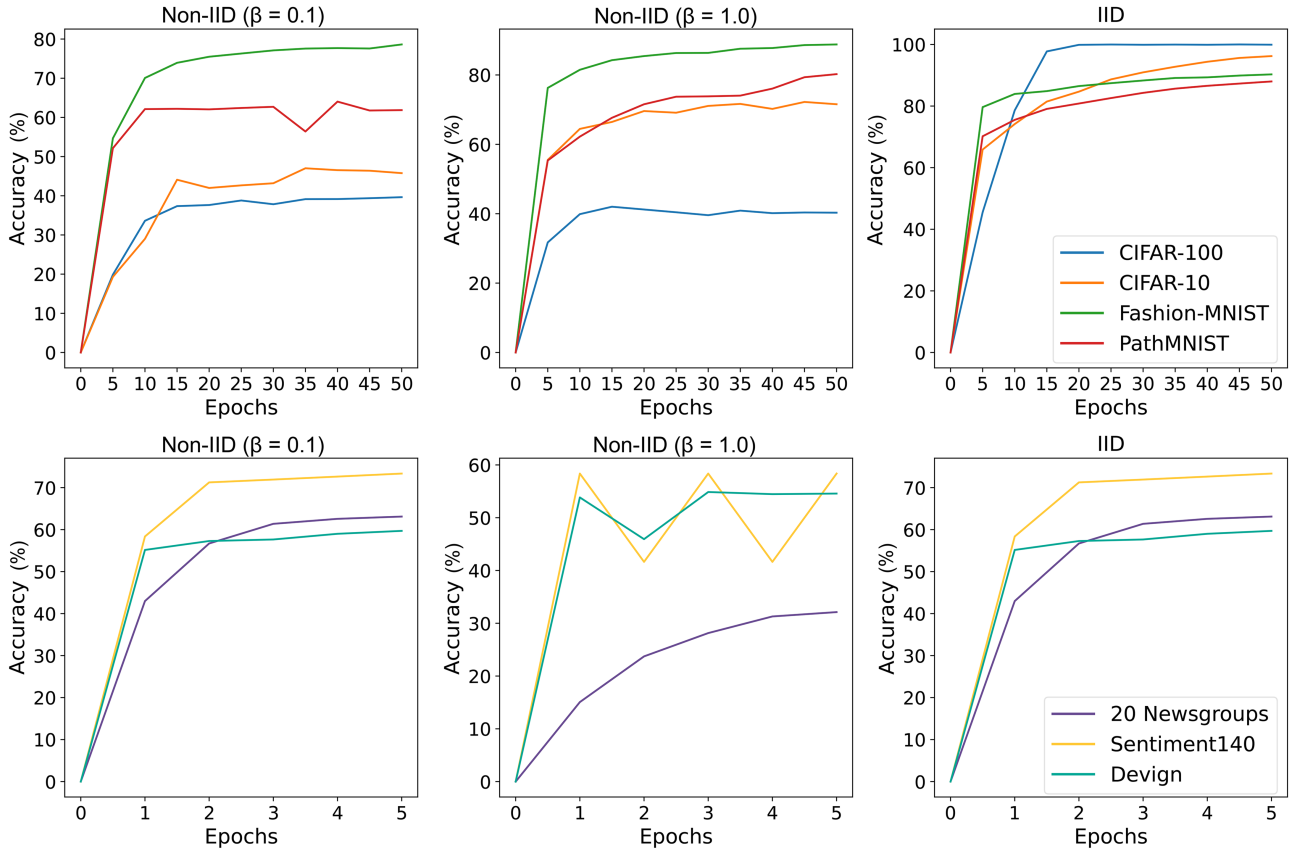


图 2　(网络版彩图) 不同训练数据分布下全局模型的训练日志. 第一行为图像任务训练日志, 第二行为自然语言处理 (NLP) 任务训练日志.

Figure 2　(Color online) The training logs of global models under different training data distributions. The first line shows the training logs on the image task, and the second line shows the training logs on the natural language processing (NLP) task.

典数据集, 包含 19 种常见视觉扰动类型 (如高斯噪声、运动模糊、雪、雾、对比度改变、JPEG 压缩等), 且每种扰动均设置 5 个严重程度等级. 参照近期工作 [67], 我们从均匀分布中随机采样 50000 张图像作为新数据, 数据规模与原始训练集保持一致, 同时覆盖不同扰动类型与严重程度. 对于其他图像数据集, 我们采用与 CIFAR-10-C 相同的扰动方法, 在其测试数据集上生成新的扰动数据. 随后, 以与 CIFAR-10-C 相同的方式随机采样新数据, 并保证其规模与原始训练数据一致. 对于每个新生成的数据集, 随机划分其中 10% 作为 OOD 测试数据, 剩余 90% 作为候选数据.

对于代码数据集, 我们采用 ALERT [68] 这一黑盒对抗代码生成方法来构造新数据. 在每个生成的数据集中, 随机划分 10% 作为 OOD 测试数据, 剩余 90% 作为候选数据.

对于文本数据集, 我们使用一个经典的自然语言处理数据增强框架 [69] 中的 4 种常用增强方法对测试数据进行处理, 以构造新数据. 具体方法包括 butter finger 扰动 (模拟键盘输入错误导致的拼写错误)、随机词删除、同义词插入以及同义词替换. 随后, 我们将每个新生成的数据集中 20% 作为 OOD 测试数据, 剩余 80% 作为候选数据.

为验证上述构造方式是否确实引入了有效的分布偏移, 我们在 7 个数据集上进行了分析: 将已在原始训练集上完成联邦训练的全局模型, 直接应用于本节构造的新测试数据集上进行评估, 见表 2. 结果表明, 与原始测试集上的性能相比, 模型在新测试集上的准确率出现普遍且显著的下滑, 平均降幅达到 16.5%. 这一结果直接证实, 本节构造的新数据与原始训练数据之间存在实质性的分布偏移, 从而为后续开展 DSMR 修复实验提供了必要的前提基础.

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

表 2　原始测试集与分布偏移测试集上的模型性能对比. Acc_{ID} 代表模型在原始测试集上的准确率 (%), Acc_{OOD} 代表模型在新测试集上的准确率 (%).

Table 2 Comparison of model performance on the original test set and the distribution-shifted test set. Acc_{ID} represents the model accuracy (%) on the original test set, and Acc_{OOD} represents the model accuracy (%) on the new test set.

Dataset	IID		Non-IID ($\beta = 1.0$)		Non-IID ($\beta = 0.1$)	
	Acc_{ID}	Acc_{OOD}	Acc_{ID}	Acc_{OOD}	Acc_{ID}	Acc_{OOD}
Fashion-MNIST	90.28	79.56	88.79	76.96	78.62	68.46
PathMNIST	87.99	51.03	80.23	47.24	61.86	45.52
CIFAR-10	96.22	59.92	71.56	56.84	45.78	36.20
CIFAR-100	78.62	68.46	40.30	34.36	39.64	30.54
Devign	55.16	55.23	59.70	59.54	61.13	59.98
20 Newsgroups	66.70	7.42	63.08	7.42	29.29	5.92
Sentiment140	74.16	72.72	73.35	72.00	58.36	59.00

Non-IID 与 IID 设定下的数据准备. 本文进一步研究数据分布对 DSMR 的影响, 因此构造了不同的数据分布设置. 原始训练数据与候选数据在联邦学习流程中将分别按照相同的数据分布方式进行划分. 具体而言, 我们采用 Dirichlet 分布 $Dir_N(\cdot)$ [70, 71] 将数据样本非均匀地分配给各客户端作为本地训练数据. 对于客户端 C_i, 其接收的数据比例记为 q_i, 其中 $q \sim Dir_N(\beta)$. 参数 β 用于控制数据分布的不均衡程度, β 越小, 数据分布越不均衡. 本文设置 3 种数据分布场景, 即 IID, Non-IID 且 $\beta = 0.1$, Non-IID 且 $\beta = 1.0$. 在 Non-IID 设置下, 不同客户端的本地数据规模和类别分布均不相同. 在 IID 设置下, 各客户端接收数量相等且类别分布一致的数据. 此外, 为贴合实际应用场景, 在 IID 与 Non-IID 两种设置中, 客户端之间均无法相互共享数据. 因此, 本文实验设置能够较好地反映真实世界中的联邦学习部署环境.

3.6　DSMR 方法

本文关注基于数据选择的联邦模型修复问题, 即在模型部署后持续接收未标注新数据的过程中, 通过选择少量高价值样本进行标注与训练, 以在有限资源条件下修复或提升模型性能. 围绕这一目标, 本文同时考虑了两类方法: 一是专门为 DSMR 设计的方法, 二是主动学习领域中的代表性策略 [72, 73]. 之所以将二者纳入同一研究框架进行对比, 是因为它们在核心目标上高度一致 —— 均旨在通过选择性标注来最大化模型性能提升. 然而, 二者在问题设定上存在本质区别: 首先, 模型的起点不同. 主动学习通常从随机初始化或弱监督模型出发, 通过迭代选样逐步构建训练集; 而模型修复则以一个已完成初始训练的模型为起点, 目标是修复其在新增数据上的性能短板. 其次, 未标注数据的产生方式不同. 主动学习通常假设一个静态的未标注池可供反复查询; 而模型修复面对的是部署后环境下动态遇见的新数据批次, 这些数据往往与原始训练数据存在分布偏移. 尽管二者存在上述差异, 但由于其核心目标均在于刻画样本对模型性能提升的潜在价值, 两者在问题形式上高度契合, 因此可以在不做本质修改的情况下将主动学习选择策略自然迁移并应用于 DSMR 场景. 遵循已有工作 [44], 我们在实验中选取了两种具有代表性的 DSMR 方法, 即 DeepGini 和 MCP. 对于主动学习方法, 我们引入了两种策略: 最常用的 Least Confidence (LC) 方法, 以及专门针对联邦学习提出的当前先进方法 LoGo. 我们将这两种主动学习方法引入 DSMR 场景中, 并将其作为 DSMR 方法进行评估. 此外, 鉴于在每个客户端上均存在两类可用于查询选择的模型, 我们进一步将 DSMR 方法划分为两组: 基于全局模型的方法与基于本地模型的方法, 并分别以 "-g" 和 "-l" 作为后缀进行区分.

Random 选择是数据选择任务中的基线方法, 即从未标注数据集中随机选取样本. 该方法作为一种简单且无偏的基线, 用于评估更复杂数据选择策略的有效性.

LC [74] 是一种简单且被广泛采用的主动学习策略, 其基本思想是选择模型最不确定的数据样本. 具体而言, LC 选择模型在所有类别上的最大预测概率最小的样本, 作为补丁数据.

LoGo [35] 是针对联邦学习提出的主动学习策略, 旨在缓解客户端之间的类别不平衡和数据异质性问题. 该

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

方法采用两阶段选择流程: 首先, 本地模型执行 K-means 聚类, 以保证本地数据在类别层面的多样性; 随后, 全局模型基于熵采样 [74] 从每个簇中选择最不确定的样本. 通过这种方式, LoGo 能够同时选取具有较高不确定性和多样性的数据样本.

Multiple-boundary clustering and prioritization (MCP) [42] 是一种典型的 DSMR 技术, 其核心思想是选择靠近决策边界的数据样本用于模型重训练. 该方法包含 3 个步骤: 首先, 使用 DNN 模型对每个数据样本进行推理, 得到对应的输出概率分布; 其次, 基于样本预测概率中排名前两位的类别, 进行边界区域聚类, 将数据划分为若干簇; 最后, 在每个簇内通过计算最大预测概率和第二大预测概率之比来确定样本优先级, 并从每个非空簇中均匀选取高优先级样本. 通过同时引入数据聚类与不确定性采样机制, MCP 在选择过程中兼顾了数据的多样性与不确定性, 这一思想与 LoGo 方法具有相似性.

DeepGini [40] 是一种较早提出的 DSMR 方法, 从深度神经网络的统计特性出发设计不确定性度量. 该方法基于 Gini 得分选择最不确定的数据样本, 其计算公式如下:

$$1-\sum_{j=1}^{K} p_j^2(x), \tag{1}$$

其中, $p_j(x)$ 表示测试输入 x 属于第 j 类的预测概率, K 表示类别总数. Gini 得分越高, 表明该数据样本的不确定性越大.

3.7　实现细节与实验环境

为缓解再训练过程中的随机性影响, 对每组实验重复三次并报告平均结果. 实验基于 Python 3.7 实现, 使用 Pytorch 1.8.0 深度学习框架. 联邦学习流程的实现基于开源项目 [64, 75] 并进行了相应修改. 所有实验均在一台服务器上完成, 服务器运行 Ubuntu 20.04.1 系统, 配备 64 核 AMD EPYC 7543P 32-Core CPU、4 张 NVIDIA RTX A5000 GPU, 以及 251 GiB 物理内存.

4　实证研究结果

本节给出本文的实证研究结果. 在全部实验中, 共对超过 45360 个 (具体包括 7 个数据集 × 10 个客户端 × 9 种数据分布设置 × 8 种数据选择方法 × 3 种标注预算设置 × 3 次重复实验) 深度神经网络模型进行了再训练.

4.1　研究问题 1: 查询选择器的影响

本研究从两个方面来分析查询选择器对 DSMR 有效性的影响. 首先, 通过分析全局模型与本地模型作为查询选择器时所选数据的重叠情况, 比较二者在数据选择结果上的差异. 其次, 进一步评估由不同查询选择器所选数据用于再训练后, 对模型性能提升所产生的影响.

表 3 汇总了在采用不同查询选择器时所选数据的重叠比例. 从结果可以观察到, 当分别使用全局模型与本地模型作为查询选择器时, DSMR 方法在数据选择上普遍呈现出明显差异. 具体而言, 在 63 种实验设置中, 仅有 5 种情况下数据重叠比例超过 60%. 此外, 在多数场景下, 随着训练数据 Non-IID 程度的加剧, 数据重叠比例呈现下降趋势. 例如, 在 Fashion-MNIST 数据集上, 当训练数据分布从 IID 转变为 $\beta = 0.1$ 的 Non-IID 设置时, 数据重叠比例由 57.93% 降低至 19.76%.

造成上述现象的主要原因在于, 当训练数据呈现 Non-IID 分布时, 联邦学习模型的整体准确率下降, 这一点已在表 1 中得到验证. 在模型准确率降低的情况下, 无论是全局模型还是本地模型, 其输出结果对 DSMR 方法而言均变得不够可靠. 这种不确定性的增加, 进一步放大了两类模型在数据选择结果上的差异. 相比之下, 候选数据的分布对重叠比例的影响较为有限. 例如, 在 CIFAR-100 数据集上, 不同候选数据分布条件下的重叠比例分别为 30.59%, 30.83% 与 30.80%, 差异极小. 该结果进一步表明, DSMR 的效果在很大程度上依赖于已训练联邦学习模型的准确率, 而候选数据的分布情况对其影响相对较弱. 总体而言, 上述结果验证了全局模型与本地模型在作为查询选择器时表现出本质差异, 并表明有必要进一步分析这种差异对联邦学习模型修复效果的具体影响.

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

表 3　全局模型与本地模型作为查询选择器时所选数据的重叠比例 (%), 其中 TrainD 表示训练数据分布, CandiD 表示候选数据分布.

Table 3 Overlap (%) of data selected by the global model and the local model, where TrainD refers to training data distribution and CandiD refers to candidate data distribution.

TrainD	CandiD	Fashion-MNIST	PathMNIST	CIFAR-10	CIFAR-100	Devign	20 Newsgroups	Sentiment140
IID	IID	57.93	66.33	36.84	30.59	62.26	56.98	33.51
	Non-IID (β=1.0)	58.37	67.02	38.65	30.83	57.26	56.24	33.02
	Non-IID (β=0.1)	60.20	72.29	40.97	30.80	59.41	56.83	38.33
Non-IID (β=1.0)	IID	39.31	46.30	26.73	42.53	54.44	45.89	13.81
	Non-IID (β=1.0)	39.34	47.59	27.43	42.36	54.33	45.88	14.67
	Non-IID (β=0.1)	41.08	52.82	31.06	42.49	52.01	44.66	15.98
Non-IID (β=0.1)	IID	19.76	18.12	15.01	30.26	52.01	38.77	39.14
	Non-IID (β=1.0)	20.44	19.44	14.83	30.23	50.77	37.63	37.95
	Non-IID (β=0.1)	21.97	22.88	16.25	29.81	54.01	38.20	38.27

表 4 与 5 分别给出了在图像数据集和自然语言任务数据集上的再训练模型在分布内测试集 $Test_{ID}$ 与分布外测试集 $Test_{OOD}$ 上的准确率提升结果, 表 6 给出了再训练模型在分布内测试集 $Test_{ID}$ 与分布外测试集 $Test_{OOD}$ 上取得 Top-1 与 Top-3 准确率的累计次数统计. 从平均准确率提升结果来看 (表 4 与表 5 中灰色背景部分), 可以观察到在分布外测试集上, 在 7 个数据集中的 5 个场景下, 采用全局模型作为查询选择器能够获得更高准确率的联邦学习模型. Top-1 与 Top-3 频次统计结果 (表 6 中灰色背景部分) 同样表明, 基于全局模型的数据选择在分布外测试数据上更容易获得高准确率模型, 即具有更强的泛化能力. 此外, 在分布内测试集上, 全局模型同样表现出具有竞争力的性能, 说明其在模型修复过程中能够较好地保留模型原有知识.

综上所述, 鉴于本文关注的是分布偏移场景下的模型修复问题, 全局模型更适合作为联邦学习中 DSMR 的查询选择器. 值得注意的是, 这一结论与已有的联邦主动学习研究 [35] 得到的结论相反. 该研究指出, 当数据分布不平衡参数 β 较小时, 本地模型更适合作为查询选择器. 这一差异进一步表明, 联邦主动学习与面向联邦学习模型的 DSMR 属于两个本质不同的问题.

在面向联邦学习的 DSMR 场景下, 我们推测全局模型准确率更高是其成为更优查询选择器的关键原因. 以 Fashion-MNIST 在 Non-IID ($\beta = 0.1$) 设置下为例, 全局模型在第 50 轮训练时可达到 79.62% 的准确率, 而本地模型的平均准确率仅为 47.28%, 其中最高为 59.53%, 最低仅为 26.05%, 与全局模型相比存在显著差距. 由于多数 DSMR 方法依赖模型预测输出来进行数据优先级排序, 因此, 预测更准确且置信度更高的联邦学习模型更有利于实现精确的数据选择. 这一点可以解释为何在 Fashion-MNIST 与 PathMNIST 任务中, 除 $FM\text{-}IID\text{-}Test_{ID}$ 场景外, 全局模型在多数情况下均优于本地模型. 进一步分析发现, Fashion-MNIST 与 PathMNIST 任务对应的联邦学习模型整体准确率高于其他任务. 相比之下, 在 Non-IID ($\beta = 0.1$) 设置下, Sentiment140 的准确率仅为 58.36%, 而 Devign 即便在 IID 场景下准确率也仅为 61.13%, 对于二分类任务而言均处于较低水平. 模型准确率偏低会自然导致其数据选择行为偏离其他数据集所呈现的整体趋势. 与 DSMR 不同, 联邦主动学习的起点是性能较差的模型 (通常由随机初始化得到). 在这一阶段, 无论是全局模型还是本地模型, 其预测置信度普遍较低, 难以作为可靠的样本筛选依据. 这一差异直接影响了数据选择策略的有效性: 在主动学习中, 选样策略需要引导模型从零开始快速收敛; 而在 DSMR 中, 选样策略则需借助已有模型的判别能力, 精准定位导致性能衰减的分布偏移样本. 因此, 模型起点的不同不仅构成了两类问题在设定层面的区别, 更塑造了它们对数据选择方法的差异化需求.

研究问题 1 小结: 与联邦主动学习中在 Non-IID 场景下更倾向采用本地模型作为查询选择器不同, 在应对分布偏移的联邦学习模型修复任务中, 全局模型通常是更优的查询选择器. 具体而言, 基于全局模型的 DSMR 方法更有可能获得泛化能力更强的联邦学习模型.

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

表 4　使用全局模型或本地模型所选数据进行再训练后, 在图像数据集上的分布内测试集 ($Test_{ID}$) 与分布外测试集 ($Test_{OOD}$) 上的平均准确率提升 (%), 其中 TrainD 表示训练数据分布, 最优结果以粗体标出.

Table 4　Average accuracy improvement (%) on image datasets using data selected by the global or local models against ID test set ($Test_{ID}$) and OOD test set ($Test_{OOD}$), where TrainD refers to training data distribution and the best results are in bold.

Dataset	TrainD	Test set	Candidate data distribution							
			IID		Non-IID (β=1.0)		Non-IID (β=0.1)		Average	
			Global	Local	Global	Local	Global	Local	Global	Local
Fashion-MNIST	IID	$Test_{ID}$	−0.06	−0.04	−0.03	−0.02	−0.01	0.01	−0.03	**−0.02**
		$Test_{OOD}$	0.74	0.67	0.69	0.65	0.56	0.50	**0.66**	0.61
	Non-IID (β=1.0)	$Test_{ID}$	0.23	0.18	0.22	0.19	0.20	0.18	**0.21**	0.18
		$Test_{OOD}$	1.41	1.19	1.21	1.13	1.02	0.87	**1.21**	1.07
	Non-IID (β=0.1)	$Test_{ID}$	0.96	0.86	1.00	0.52	0.00	−0.01	**0.65**	0.46
		$Test_{OOD}$	3.06	2.59	3.19	2.66	1.19	1.42	**2.48**	2.22
	Average	$Test_{ID}$	**0.37**	0.33	**0.39**	0.23	**0.06**	**0.06**	**0.28**	0.21
		$Test_{OOD}$	**1.74**	1.48	**1.70**	1.48	**0.93**	**0.93**	**1.45**	1.30
PathMNIST	IID	$Test_{ID}$	0.33	0.34	0.38	0.37	0.38	0.39	**0.37**	**0.37**
		$Test_{OOD}$	2.72	2.67	2.19	2.02	0.86	0.79	**1.92**	1.83
	Non-IID (β=1.0)	$Test_{ID}$	1.03	0.61	0.83	0.44	0.70	0.61	**0.86**	0.56
		$Test_{OOD}$	4.22	3.98	3.52	3.30	2.40	2.33	**3.38**	3.21
	Non-IID (β=0.1)	$Test_{ID}$	1.51	0.49	1.34	0.34	1.27	0.33	**1.37**	0.39
		$Test_{OOD}$	5.30	5.24	4.75	3.92	2.04	2.01	**4.03**	3.72
	Average	$Test_{ID}$	**0.96**	0.48	**0.85**	0.38	**0.79**	0.44	**0.86**	0.44
		$Test_{OOD}$	**4.08**	3.96	**3.49**	3.08	**1.77**	1.71	**3.11**	2.92
CIFAR-10	IID	$Test_{ID}$	0.08	0.13	0.08	0.15	0.14	0.15	0.10	**0.15**
		$Test_{OOD}$	2.39	2.33	2.15	2.07	1.48	1.46	**2.01**	1.95
	Non-IID (β=1.0)	$Test_{ID}$	1.46	1.56	1.35	1.48	0.68	0.75	1.16	**1.26**
		$Test_{OOD}$	4.30	4.01	4.14	3.97	2.84	2.82	**3.76**	3.60
	Non-IID (β=0.1)	$Test_{ID}$	6.09	6.78	5.34	5.41	6.81	6.18	6.08	**6.13**
		$Test_{OOD}$	7.62	7.15	6.63	6.21	7.29	6.77	**7.18**	6.71
	Average	$Test_{ID}$	2.54	**2.82**	2.26	**2.35**	2.54	2.36	2.45	**2.51**
		$Test_{OOD}$	**4.77**	4.49	**4.31**	4.08	**3.87**	3.68	**4.31**	4.09
CIFAR-100	IID	$Test_{ID}$	0.96	0.86	1.18	0.79	0.04	−0.04	**0.72**	0.54
		$Test_{OOD}$	3.06	2.59	2.93	2.85	1.37	1.26	**2.45**	2.23
	Non-IID (β=1.0)	$Test_{ID}$	0.12	0.21	0.17	0.15	0.15	0.14	0.15	**0.16**
		$Test_{OOD}$	1.14	0.83	0.91	0.96	0.97	0.89	**1.01**	0.89
	Non-IID (β=0.1)	$Test_{ID}$	0.17	0.34	0.24	0.21	0.12	0.16	0.18	**0.23**
		$Test_{OOD}$	1.68	1.73	1.62	1.57	1.51	1.49	**1.60**	**1.60**
	Average	$Test_{ID}$	0.42	**0.47**	**0.53**	0.38	**0.10**	0.09	**0.35**	0.31
		$Test_{OOD}$	**1.96**	1.72	**1.82**	1.80	**1.28**	1.21	**1.69**	1.57

4.2　研究问题 2: DSMR 方法的有效性

在研究问题 2 中, 我们系统性地评估了多种代表性 DSMR 方法在不同训练数据分布与候选数据分布设置下, 对联邦学习模型修复效果的实际影响. 表 7 汇总了在 7 个数据集上的平均准确率提升情况, 其中 "-g" 表示以全局模型作为查询选择器, "-l" 表示以本地模型作为查询选择器.

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

表 5　使用全局模型或本地模型所选数据进行再训练后, 在 NLP 数据集上的平均准确率提升 (%), 其中 TrainD 表示训练数据分布, 最优结果以粗体标出.

Table 5 Average accuracy improvement (%) on NLP datasets using data selected by the global or local models, where TrainD refers to training data distribution and the best results are in bold.

Dataset	TrainD	Test set	Candidate data distribution							
			IID		Non-IID (β=1.0)		Non-IID (β=0.1)		Average	
			Global	Local	Global	Local	Global	Local	Global	Local
Devign	IID	$Test_{ID}$	−0.42	−0.52	−1.57	0.02	−0.97	−0.89	−0.99	**−0.46**
		$Test_{OOD}$	0.22	−0.40	−1.45	0.42	−0.48	−0.62	−0.57	**−0.20**
	Non-IID (β=1.0)	$Test_{ID}$	0.74	0.64	1.25	0.79	−1.26	−1.31	**0.24**	0.04
		$Test_{OOD}$	0.24	0.26	0.80	0.48	−1.80	−1.79	**−0.25**	−0.35
	Non-IID (β=0.1)	$Test_{ID}$	0.33	0.14	0.13	0.06	0.20	0.16	**0.22**	0.12
		$Test_{OOD}$	0.65	0.48	0.32	0.40	0.50	0.44	**0.49**	0.44
	Average	$Test_{ID}$	**0.21**	0.09	−0.06	**0.29**	**−0.68**	**−0.68**	−0.17	**−0.10**
		$Test_{OOD}$	**0.37**	0.11	−0.11	**0.43**	**−0.59**	−0.66	−0.11	**−0.04**
20 Newsgroups	IID	$Test_{ID}$	−0.72	−0.74	−0.70	−0.70	−0.70	−0.73	**−0.71**	−0.73
		$Test_{OOD}$	10.92	11.54	9.79	9.94	6.97	7.11	9.23	**9.53**
	Non-IID (β=1.0)	$Test_{ID}$	1.05	1.16	1.10	1.14	1.13	1.14	1.10	**1.14**
		$Test_{OOD}$	10.51	9.89	9.81	8.59	6.53	5.86	**8.95**	8.12
	Non-IID (β=0.1)	$Test_{ID}$	23.19	23.24	23.32	23.35	23.65	23.59	23.38	**23.39**
		$Test_{OOD}$	13.65	12.74	13.04	12.49	10.02	9.15	**12.23**	11.46
	Average	$Test_{ID}$	7.84	**7.88**	7.91	**7.93**	8.02	**8.00**	7.92	**7.94**
		$Test_{OOD}$	**11.69**	11.39	**10.88**	10.34	**7.84**	7.37	**10.14**	9.70
Sentiment140	IID	$Test_{ID}$	0.34	0.36	0.27	0.17	0.13	0.13	**0.25**	0.22
		$Test_{OOD}$	0.46	0.39	0.35	0.22	0.12	0.17	**0.31**	0.26
	Non-IID (β=1.0)	$Test_{ID}$	0.81	0.48	0.67	0.99	0.46	0.6	0.65	**0.69**
		$Test_{OOD}$	0.93	0.51	0.77	1.00	0.52	0.62	**0.74**	0.71
	Non-IID (β=0.1)	$Test_{ID}$	10.43	12.93	8.33	11.63	5.26	7.19	8.01	**10.58**
		$Test_{OOD}$	8.74	10.95	6.81	9.72	4.63	6.2	6.73	**8.96**
	Average	$Test_{ID}$	3.86	**4.59**	3.09	**4.26**	1.95	**2.64**	2.97	**3.83**
		$Test_{OOD}$	3.38	**3.95**	2.64	**3.65**	1.76	**2.33**	2.59	**3.31**

表 6　在 7 个数据集上, 不同查询选择器取得 Top-1 与 Top-3 准确率的累计次数统计, 其中 TrainD 表示训练数据分布, 最优结果以粗体标出.

Table 6 Frequency of achieving Top-1 and Top-3 accuracy using different query selectors across seven datasets, where TrainD refers to training data distribution and the best results are in bold.

TrainD	Test set	Candidate data distribution (Top-1)								Candidate data distribution (Top-3)							
		IID		Non-IID (β=1.0)		Non-IID (β=0.1)		Sum		IID		Non-IID (β=1.0)		Non-IID (β=0.1)		Sum	
		Global	Local	Global	Local	Global	Local	Global	Local	Global	Local	Global	Local	Global	Local	Global	Local
IID	$Test_{ID}$	9	12	10	11	8	13	27	**36**	31	32	28	35	32	31	91	**98**
	$Test_{OOD}$	10	11	15	6	12	9	**37**	26	33	30	34	29	30	33	**97**	92
Non-IID (β=1.0)	$Test_{ID}$	9	12	7	14	9	12	25	**38**	29	34	30	33	31	32	90	**99**
	$Test_{OOD}$	18	3	13	8	15	6	**46**	17	42	21	38	25	35	28	**115**	74
Non-IID (β=0.1)	$Test_{ID}$	11	10	8	13	10	11	29	**34**	29	34	26	37	33	30	88	**101**
	$Test_{OOD}$	12	9	11	10	10	11	**33**	30	36	27	38	25	34	29	**108**	81
Sum	$Test_{ID}$	29	**34**	25	**38**	27	**36**	81	**108**	89	**100**	84	**105**	**96**	93	269	**298**
	$Test_{OOD}$	**40**	23	**39**	24	**37**	26	**116**	73	**111**	78	**110**	79	**99**	90	**320**	247

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

表 7　不同 DSMR 方法在 7 个数据集上的平均准确率提升 (%), 其中 TrainD 表示训练数据分布, CandiD 表示候选数据分布, 最优结果以粗体标出.

Table 7 Average accuracy improvement (%) of models using different DSMR methods across seven datasets, where TrainD refers to training dataset distribution, CandiD refers to candidate dataset distribution, and the best results are in bold.

Distribution		Test set	Random	LoGo	LC-g	MCP-g	Gini-g	LC-l	MCP-l	Gini-l
TrainD	IID	$Test_{ID}$	**0.06**	0.00	**0.06**	−0.04	−0.18	0.04	−0.02	−0.01
		$Test_{OOD}$	2.81	**3.03**	2.56	2.90	1.41	2.61	2.76	1.57
	Non-IID (β=1.0)	$Test_{ID}$	0.63	0.67	0.71	**0.75**	0.41	0.66	0.64	0.44
		$Test_{OOD}$	**3.40**	**3.40**	3.20	3.28	1.59	2.78	2.99	1.60
	Non-IID (β=0.1)	$Test_{ID}$	5.61	5.94	**6.11**	5.24	5.74	6.10	5.75	5.88
		$Test_{OOD}$	5.60	5.78	**5.91**	5.01	3.96	5.25	5.50	4.31
CandiD	IID	$Test_{ID}$	2.31	2.42	**2.52**	2.28	2.15	2.50	2.45	2.20
		$Test_{OOD}$	4.65	**4.81**	4.61	4.54	2.83	4.17	4.56	2.89
	Non-IID (β=1.0)	$Test_{ID}$	2.04	2.24	**2.44**	1.97	1.97	2.40	2.20	2.16
		$Test_{OOD}$	4.15	**4.29**	**4.29**	3.87	2.46	3.91	4.02	2.72
	Non-IID (β=0.1)	$Test_{ID}$	**1.96**	1.95	1.91	1.71	1.85	1.89	1.70	1.96
		$Test_{OOD}$	2.99	**3.11**	2.77	2.77	1.66	2.56	2.68	1.87
Average		$Test_{ID}$	2.10	2.20	**2.29**	1.99	1.99	2.26	2.12	2.11
		$Test_{OOD}$	3.93	**4.07**	3.89	3.73	2.32	3.55	3.76	2.49

从不同方法的相对表现来看, 各类 DSMR 方法在不同测试场景下呈现出一定的互补性特征. 在同分布测试数据 $Test_{ID}$ 上, LC-g 在多数设置下取得了最高或接近最高的平均准确率提升, 例如在训练数据分布为 Non-IID ($\beta = 0.1$) 设置下达到 6.11%, 在总体平均结果中也取得了最高的 $Test_{ID}$ 提升 (2.29%). 该结果表明, 在同分布测试场景中, 基于预测置信度的不确定性策略仍然是一种较为有效的选择. 在分布偏移测试数据 $Test_{OOD}$ 上, LoGo 方法整体取得了相对更高的平均提升, 在多个设置下取得最优或并列最优结果, 并在总体平均结果中取得最高的 $Test_{OOD}$ 提升 (4.07%). 该现象说明, 在分布偏移场景下, 显式刻画决策边界结构的策略在一定程度上有助于提升模型的泛化性能.

然而, 更为值得注意的是, 表 7 的结果同时表明, 在多数设置下, 各类精心设计的 DSMR 方法与随机选择在平均准确率提升方面的差距普遍较小, 且不同方法在不同数据分布条件下的相对优势并不稳定. 例如, 在若干设置下随机选择本身即可取得与最优方法相当的性能提升. 这一现象揭示了一个重要事实: 在联邦学习场景下, 基于数据选择的模型修复方法整体仍处于早期阶段. 在复杂的 Non-IID 与分布偏移场景中, 现有 DSMR 方法尚难以稳定地优于随机选择. 这不仅反映了联邦模型修复问题本身的高度挑战性, 也表明直接将单模型场景下的 DSMR 方法迁移至联邦学习环境往往难以取得预期效果.

研究问题 2 小结: 在联邦学习场景下, 现有 DSMR 方法在多数设置下仅能带来有限且不稳定的性能增益, 难以在整体上稳定优于随机选择, 表明联邦模型修复仍处于早期阶段, 亟需针对联邦 Non-IID 特性设计新的数据选择机制.

4.3　研究问题 3: 所选数据标签分布对 DSMR 的影响

如 4.2 小节所述, 联邦学习场景下基于数据选择的模型修复仍处于研究初期. 与精心设计的 DSMR 方法相比, 简单的随机选择在多数情况下也能获得相近的修复效果. 因此, 本节进一步探索潜在的研究方向, 以提升联邦学习场景下 DSMR 方法的有效性. 已有研究 [44] 表明, 在传统深度学习模型的 DSMR 场景中, 所选数据的标签分布均衡性有助于提升模型修复效果. 受此启发, 本节探讨在联邦学习场景下, 维持本地模型修复数据在标签分布上的均衡, 是否有助于提升 DSMR 对联邦学习模型的修复效果.

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

4.3.1　研究设计

本探索性研究包括 3 个部分: (1) 相关性分析, 用于探究各客户端所选数据的标签均衡性与基于此数据再训练的联邦学习模型性能之间的关系, (2) 受控实验, 用于分析在客户端层面上引入标签均衡数据对 DSMR 的促进作用, (3) 分析现有 DSMR 方法所选数据的标签分布特征.

相关性分析. 为了判断所选数据在不同类别之间的分布是否均匀, 量化所选数据的标签均衡性, 我们计算其标签分布的标准差 (std), 较小的标准差表明数据分布更加均衡. 在计算标准差时, 我们统计所有类别上的样本数量, 即使某一类别的样本数为 0 也会被纳入计算. 例如, 在一个三分类任务中, 若共有 10 个样本, 当各类别样本数为 (2, 3, 5) 时, 其标准差为 1.25. 而当类别分布为 (0, 1, 9) 时, 标准差增大至 4.03.

为验证数据均衡性与模型修复效果之间是否存在关联, 我们采用 Spearman 等级相关系数 (ρ) 计算所选数据标签分布的标准差与不同标注预算下再训练模型准确率之间的相关性. 具体而言, 在计算每个客户端所选数据的标准差 std_i 后, 我们根据该客户端所选数据量占全部所选数据量的比例为其分配权重 w_i, 计算如下: $w_i = \frac{|D_i|}{\sum_{i=1}^{N}|D_i|}$, 其中 N 表示客户端数量. 随后, 对 $std_1, \ldots, std_N$ 进行加权聚合, 得到整体的标签分布标准差 std. 在实验中, 每一组 (std, 准确率) 数据对共包含 72 个样本点 (3 种训练数据分布 $\times$ 3 种候选数据分布 $\times$ 8 种数据选择方法). 我们在每一组数据上分别计算 Spearman 等级相关系数 (ρ), 以分析标签均衡性与模型修复效果之间的关系.

受控实验. 为进一步探究在客户端层面上引入标签均衡的数据分布在多大程度上能够促进联邦学习模型的修复效果, 本文设计了一项受控实验, 并在多个中等类别规模的数据集上开展评估, 包括 CIFAR-10 (10 类)、Fashion-MNIST (10 类) 与 PathMNIST (9 类). 该实验需要对不同标签的数据进行显式分配, 并分别统计各标签上的性能变化, 因此数据集的类别数量会对实验设计与结果分析产生重要影响. 类别数过少 (如二分类任务) 难以体现标签分布调整带来的差异, 而类别规模过大的数据集 (如 100 个类别的 CIFAR-100) 会显著增加标签组合空间, 从而降低实验可控性与结果可解释性. 因此, 本实验选择标签数量相近且具有代表性的数据集, 以在保证分析可控性的同时提升结论的普适性.

在该实验中, 本文假设客户端之间可以进行通信, 从而能够人为控制并调整分配给各客户端的候选数据标签分布. 具体而言, 本文对候选数据进行重分配, 使得某一特定类别的数据在不同客户端之间尽可能均衡. 随后, 使用所有候选数据对模型进行再训练, 并评估再训练后模型的准确率. 在该实验中, 以候选数据服从 Non-IID ($\beta = 0.1$) 分布条件下再训练得到的联邦学习模型性能作为基线, 用以衡量引入标签均衡分布后对模型修复性能所带来的影响.

现有 DSMR 方法分析. 最后, 分析现有 DSMR 方法所选样本的数据分布特征. 具体而言, 与前述相关性分析一致, 计算不同 DSMR 方法所选数据的标签分布标准差 (std). 通过该分析, 我们能够揭示现有方法所选数据在标签分布上的潜在局限性, 并为设计更有效的基于数据选择的联邦学习模型修复方法提供经验性指导.

4.3.2　实验结果

表 8 总结了相关性分析的实验结果. 结果表明, 当标注预算为 5% 和 10% 时, 所选数据的标签分布标准差与基于这些数据再训练的联邦学习模型性能之间呈现出明显的负相关关系. 此外, 相较于 ID 测试数据, 该负相关关系在 OOD 测试数据上的明显更强, 尤其是在标注预算为 10% 且测试数据为 OOD 分布的情况下, 7 个任务中有 5 个任务的相关性结果具有统计显著性 (p-value < 0.03).

结论: 相关性分析实验结果表明, 本地模型所选数据的标签均衡性是影响 DSMR 效果的关键因素之一. 标签分布更加均衡的所选数据能够显著提升联邦学习模型的修复性能.

图 3 展示了各数据集在 $Test_{ID}$ 和 $Test_{OOD}$ 上进行受控实验的结果. 在该图中, 颜色越深表示准确率提升越大. 实验结果表明, 当某一特定标签在被选数据中呈现均衡分布时, 使用这些数据修复后的联邦学习模型在该标签上的准确率显著提高. 此外, 我们还对多个标签同时保持均衡分布的数据进行了实验, 结果同样表明, 修复后的联邦学习模型在这些标签上的准确率也有明显提升.

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

表 8　Spearman 秩相关系数 (ρ) 及其 p 值, 用于衡量所选数据标签分布标准差 (std) 与再训练后联邦学习模型准确率之间的相关性.

Table 8　Spearman's rank correlation coefficient (ρ) and p-value between standard deviation (std) of selected data and retrained model accuracy.

Dataset	Budget = 1%				Budget = 5%				Budget = 10%			
	$Test_{ID}$		$Test_{OOD}$		$Test_{ID}$		$Test_{OOD}$		$Test_{ID}$		$Test_{OOD}$	
	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value
Fashion-MNIST	−0.15	2.1×10^{-1}	0.04	7.5×10^{-1}	−0.21	7.0×10^{-2}	−0.15	2.0×10^{-1}	−0.26	2.9×10^{-2}	−0.22	6.1×10^{-2}
PathMNIST	0.02	8.4×10^{-1}	−0.11	3.5×10^{-1}	−0.32	6.3×10^{-3}	−0.63	2.6×10^{-9}	−0.30	1.1×10^{-2}	−0.71	3.1×10^{-12}
CIFAR-10	−0.07	5.3×10^{-1}	−0.13	2.9×10^{-1}	−0.05	6.8×10^{-1}	−0.20	8.6×10^{-2}	−0.01	9.5×10^{-1}	−0.35	2.5×10^{-3}
CIFAR-100	−0.22	6.9×10^{-2}	−0.15	2.1×10^{-1}	−0.25	3.5×10^{-2}	−0.28	1.6×10^{-2}	−0.51	5.2×10^{-6}	−0.43	1.6×10^{-4}
Devign	−0.00	9.9×10^{-1}	−0.10	4.2×10^{-1}	−0.10	4.2×10^{-1}	−0.07	5.3×10^{-1}	−0.13	2.3×10^{-1}	−0.17	1.5×10^{-1}
20 Newsgroups	0.00	9.8×10^{-1}	−0.19	1.2×10^{-1}	0.05	6.7×10^{-1}	−0.50	9.3×10^{-6}	0.00	9.7×10^{-1}	−0.54	1.3×10^{-6}
Sentiment140	−0.08	5.1×10^{-1}	−0.06	6.3×10^{-1}	−0.14	2.3×10^{-1}	−0.17	1.5×10^{-1}	−0.25	3.6×10^{-2}	−0.31	7.8×10^{-3}

表 9　在候选数据分布为 IID 且标注预算为 10% 的情况下, 不同方法所选择数据的标准差 (std), 其中 TrainD 表示训练数据分布, 最小的标准差以粗体标出.

Table 9　Standard deviation (std) values of data selected by different methods with a labeling budget of 10% when the candidate data distribution is IID, where TrainD refers to training dataset distribution and the minimum std values are in bold.

Dataset	TrainD	Random	LoGo	LC-g	MCP-g	Gini-g	LC-l	MCP-l	Gini-l
Fashion-MNIST	IID	**6.67**	23.85	10.78	20.87	25.21	11.77	23.60	25.56
	Non-IID (β=1.0)	**5.55**	23.54	13.01	23.96	29.20	19.21	33.36	27.06
	Non-IID (β=0.1)	**5.65**	24.43	17.76	19.38	28.75	26.88	46.06	40.69
PathMNIST	IID	**15.12**	16.69	17.94	43.52	24.81	19.09	42.88	24.11
	Non-IID (β=1.0)	**15.66**	20.60	20.08	50.87	22.71	19.51	53.31	23.18
	Non-IID (β=0.1)	**15.32**	21.09	18.64	56.00	23.02	25.76	49.74	31.05
CIFAR-10	IID	**6.21**	12.22	13.13	15.18	11.33	13.53	17.43	11.22
	Non-IID (β=1.0)	**6.18**	13.30	16.94	20.19	13.25	20.08	22.21	17.91
	Non-IID (β=0.1)	**5.35**	21.27	20.66	22.32	15.83	21.75	23.94	22.79
CIFAR-100	IID	**2.05**	2.54	2.27	2.88	2.90	2.19	2.78	2.69
	Non-IID (β=1.0)	**2.08**	2.47	2.26	2.65	2.64	2.21	2.62	2.58
	Non-IID (β=0.1)	**2.16**	2.57	2.32	2.51	2.59	2.31	2.76	2.57
Devign	IID	**56.63**	57.83	57.18	61.74	59.71	57.67	59.72	58.08
	Non-IID (β=1.0)	60.72	59.41	59.73	61.90	58.83	**57.67**	59.42	60.60
	Non-IID (β=0.1)	**51.60**	59.17	67.98	65.29	59.27	57.64	58.63	60.56
20 Newsgroups	IID	**3.32**	3.50	3.62	4.26	4.46	3.70	4.00	4.02
	Non-IID (β=1.0)	3.55	3.73	**3.49**	4.62	4.33	3.66	4.63	4.12
	Non-IID (β=0.1)	3.64	**3.30**	4.73	4.44	3.75	4.54	4.41	4.51
Sentiment140	IID	147.20	79.60	166.50	73.40	127.10	144.20	**73.10**	109.20
	Non-IID (β=1.0)	135.20	203.20	179.70	**83.80**	142.30	194.00	225.10	210.30
	Non-IID (β=0.1)	144.70	164.80	496.10	**14.10**	14.70	460.70	345.30	311.80

结论: 除了在每个客户端内部保证被选数据在各类别上的标签均衡之外, 在所有客户端之间实现标签分布的均衡, 同样是提升 DSMR 效果的一个很有前景的研究方向.

最后, 表 9 给出了在标注预算为 10% 时, 不同现有 DSMR 方法所选择数据的标签均衡性结果. 可以看到,

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

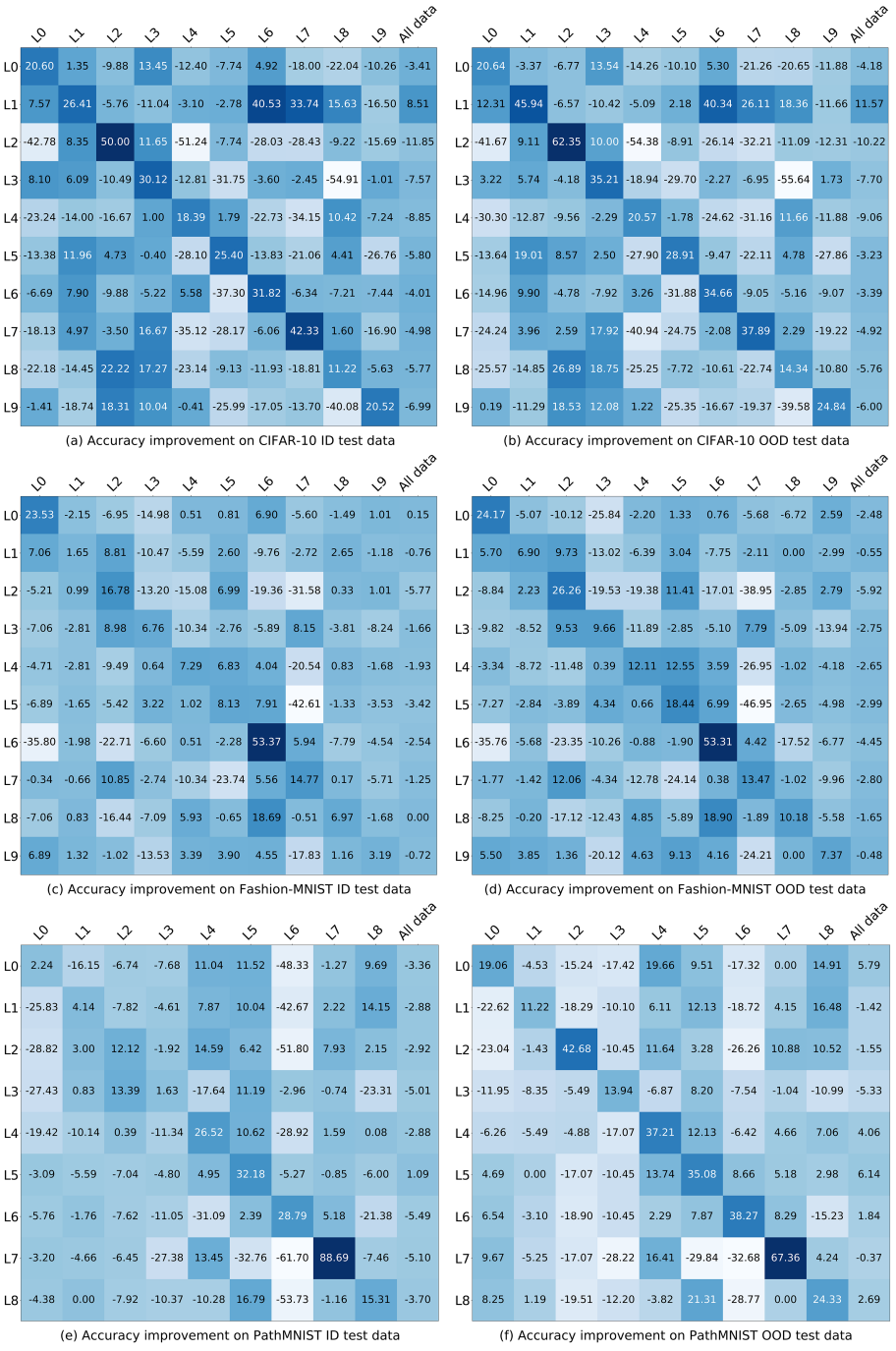


图 3　(网络版彩图) 相较于基线 Non-IID 设置, 各数据集在 $Test_{ID}$ 和 $Test_{OOD}$ 上的准确率提升 (%) 情况. 其中, 'L' 表示标签, 横轴表示某一特定标签上的准确率提升, 纵轴表示其对应数据分布被设定为均衡的标签.

Figure 3　(Color online) Accuracy improvement (%) on $Test_{ID}$ and $Test_{OOD}$ compared to baseline Non-IID setting. 'L' refers to label. The horizontal axis represents the improvement in accuracy for a specific label, while the vertical axis represents the label for which the corresponding data distribution is balanced.

与均匀随机选择相比, 所有 DSMR 方法在用于重训练联邦学习模型时, 均倾向于选择标签分布更为偏置的数据. 在所有方法中, 除随机选择外, LC 在重训练阶段选择的数据具有相对更高的标签均衡性. 这和研究问题 2 中得到的结果相符, 即 LC 是在分布内测试数据上取得最高测试准确率的 DSMR 方法之一. 该结果进一步表明: 标签均衡性是影响联邦学习模型 DSMR 效果的重要因素.

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

然而, 现有 DSMR 方法所选择数据的标签分布仍与理想情况 (即标签分布标准差 $std = 0$) 存在较大差距. 我们发现, 由于当前联邦学习模型性能较低, 尤其是本地模型, 现有主要依赖深度神经网络输出的解决方案难以有效控制标签分布. 例如, 在 Non-IID ($\beta = 0.1$) 设置下的 CIFAR-10 数据集中, 全局模型的准确率仅为 45.78%, 而本地模型的准确率介于 17.03%∼40.65%, 表现更差. 更严重的是, 即便是采用最新联邦学习方法 [16,76,77] 训练的模型, 在我们研究的任务上准确率依然偏低. 例如, 相比 FedAvg, FedMut [16] 在 CIFAR-100 数据集上使用 ResNet-18 模型, 在 Non-IID ($\beta = 0.1$) 设置下的准确率仅提高了 3.77%. 全局与本地模型之间存在较大性能差距, 加之全局模型与本地模型所选择数据的低重叠率 (见第 4.1 小节), 使得利用模型输出为数据分配标签并精确平衡标签分布变得困难.

结论: DSMR 方法在联邦学习模型修复中选择的数据存在明显的标签分布偏置.

研究问题 3 小结: 所选数据的标签均衡性是影响 DSMR 在联邦学习中有效性的关键因素之一. 更好的标签均衡性 (包括单个客户端内部以及不同客户端之间) 能够显著提升联邦学习模型的整体准确率. 然而, 现有 DSMR 方法在数据选择过程中仍然存在明显的标签偏置问题, 其在提升标签均衡性方面仍有较大的改进空间.

5　面向联邦学习的 DSMR 方法

基于前述实证研究, 尤其是研究问题 2 与研究问题 3 的结果, 我们得到两点关键认识. 一方面, 研究问题 2 的结果表明, 尽管现有 DSMR 方法在联邦学习场景下整体性能提升有限, 且在部分场景下甚至无法超过随机选择的提升效果. 但在多数场景中, 表现相对较为可靠的方法均属于基于不确定性度量的策略. 这说明, 即便在联邦学习这一复杂场景中, 以模型预测不确定性作为核心驱动信号, 仍然是数据选择过程中不可或缺的关键因素. 另一方面, 研究问题 3 的实证结果进一步揭示, 所选补丁数据在标签分布上的均衡性与模型修复效果之间存在显著关联, 能够选取标签分布更均衡数据的方法往往表现出更稳定且更优的修复性能. 然而, 现有大多数 DSMR 方法要么仅依赖单一的不确定性信号进行优先级排序, 要么仅关注补丁数据自身的分布均衡, 而缺乏对联邦学习环境中原始训练数据分布特性的系统性建模. 在联邦学习场景下, 各客户端的本地训练数据往往呈现严重的 Non-IID 分布, 即使补丁数据在全局层面保持均衡, 其与原始训练数据叠加后, 仍可能导致客户端之间以及整体训练数据的标签分布持续高度失衡. 以 MCP 方法为例, 尽管其通过多边界聚类在单模型场景下能够相对均衡地选择补丁数据, 但其隐含假设训练数据本身分布相对均衡, 并未显式建模训练数据与补丁数据之间的分布耦合关系. 在联邦学习场景中, 该缺陷将进一步放大 Non-IID 训练数据带来的标签偏置, 从而限制模型修复的实际效果.

基于上述实证结论, 我们进一步认识到, 面向联邦学习的 DSMR 方法不仅需要以不确定性作为核心选择准则, 还必须显式考虑补丁数据如何与各客户端现有训练数据协同作用, 以实现整体训练数据在客户端间的再均衡. 基于这一核心动机, 我们在 MCP 的基础上提出一种面向联邦学习的数据选择方法, 称为 Fed-PAIR (federated pairwise uncertainty-based data selection). 与现有方法不同, Fed-PAIR 并非仅追求补丁数据自身的均衡分布, 而是以当前训练数据分布为约束条件, 自适应地调整补丁数据的标签分配策略, 从而在保持不确定性驱动选择的同时, 显式促进整体训练数据以及客户端之间的标签分布趋于更加均衡的状态.

5.1　方法介绍

Fed-PAIR 是一种同时考虑样本不确定性与整体数据标签覆盖均衡性的 DSMR 数据选择方法. 其核心思想是: 通过显式建模类别对之间的决策不确定性, 并以当前训练数据的分布作为约束条件, 优先修复训练集中尚未覆盖或覆盖不足的决策边界区域, 从而实现更加有针对性且面向整体数据均衡的数据选择. 我们在算法 1 中展示了详细步骤.

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

算法 1 Fed-PAIR: 面向联邦学习的成对不确定性数据选择算法.

Require: 模型 f_θ; 训练数据 $\mathcal{D}_{train}$; 候选未标注数据 $\mathcal{D}_{cand}$; 类别数 C; 标注预算 B;
Ensure: 选中的补丁样本索引集合 $\mathcal{S}$;
根据 f_θ 计算 $\mathcal{D}_{train}$ 中样本的预测标签对 (c_1, c_2);
统计训练集中每个标签格的样本数 $M_{train}(c_1, c_2)$;
根据 f_θ 计算 $\mathcal{D}_{cand}$ 中样本 x 的 (c_1, c_2) 及不确定性比值 $r(x)$;
将候选样本按标签格 (c_1, c_2) 组织为 $C \times C$ 个集合;
初始化 $\mathcal{S} \leftarrow \emptyset$;
{第一阶段: 优先修复训练数据中最稀缺的标签格}
while $B > 0$ **do**
  选取所有候选非空且 $M_{train}(c_1, c_2)$ 最小的标签格集合 $\mathcal{G}_{min}$;
  if $\mathcal{G}_{min} = \emptyset$ **then**
    break
  end if
  for 每个 $(c_1, c_2) \in \mathcal{G}_{min}$ **do**
    if $B = 0$ **then**
      break
    end if
    从该标签格中选择 $r(x)$ 最大的样本 x;
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{x\}$;
    从候选集中移除 x, 并更新 $M_{train}(c_1, c_2)$;
    $B \leftarrow B - 1$;
  end for
end while
{第二阶段: 基于成对不确定性的全局补充选择}
while $B > 0$ **do**
  在所有剩余候选样本中, 选取 $r(x)$ 最大的样本 x;
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{x\}$;
  更新对应的 $M_{train}(c_1, c_2)$, 并从候选集中移除 x;
  $B \leftarrow B - 1$;
end while
return $\mathcal{S}$.

给定全局模型 f_θ 以及一组未标注候选样本 $\mathcal{D}_{cand}$, 对于每个样本 x, 其全局模型输出预测概率向量为 $\mathbf{p}(x) = [p_1, \ldots, p_C]$, Fed-PAIR 首先提取其预测概率最大的类别 c_1 与次大的类别 c_2, 即

$$c_1 = \arg\max_j p_j, \quad c_2 = \arg\max_{j \neq c_1} p_j, \tag{2}$$

并进一步定义成对不确定性比值为

$$r(x) = \frac{p_{c_2}(x)}{p_{c_1}(x)}. \tag{3}$$

该比值刻画了样本在类别 c_1 与 c_2 之间的判别不确定性, 数值越大表示样本越接近对应的决策边界. 与基于单类别不确定性的度量不同, 该设计能够显式刻画样本接近哪一条具体决策边界, 从而为后续有针对性的边界修复提供更精细的结构信息. 基于上述定义, Fed-PAIR 将所有样本按照 (c_1, c_2) 组织为 $C \times C$ 个标签格, 其中 C 为类别数. 每一个标签格对应于一条具体的类别决策边界区域, 形成对决策空间的细粒度划分.

与 MCP 仅依据候选数据自身进行边界聚类不同, Fed-PAIR 进一步显式引入当前训练数据的标签对分布作为选择约束. 记 $M_{train}(c_1, c_2)$ 为当前训练集中预测标签对为 (c_1, c_2) 的样本数量 (第 2 行). 当 $M_{train}(c_1, c_2) = 0$

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

表 10　Fed-PAIR 在不同训练与候选数据分布下平均准确率提升 (%) 幅度与最优基线方法的对比, 其中 TrainD 表示训练数据分布, CandiD 表示候选数据分布, 最优结果以粗体标出.

Table 10　Comparison of the average accuracy improvement (%) of Fed-PAIR over the best baseline method under different training and candidate data distributions, where TrainD refers to training dataset distribution, CandiD refers to candidate dataset distribution, and the best results are in bold.

Method	TrainD						CandiD						Average	
	IID		Non-IID (β=1.0)		Non-IID (β=0.1)		IID		Non-IID (β=1.0)		Non-IID (β=0.1)			
	$Test_{ID}$	$Test_{OOD}$	$Test_{ID}$	$Test_{OOD}$	$Test_{ID}$	$Test_{OOD}$	$Test_{ID}$	$Test_{OOD}$	$Test_{ID}$	$Test_{OOD}$	$Test_{ID}$	$Test_{OOD}$	$Test_{ID}$	$Test_{OOD}$
Best Baseline	0.06	3.03	0.75	3.40	6.11	**5.91**	2.52	4.81	2.44	4.29	1.96	3.11	2.29	4.07
Fed-PAIR-g	**0.15**	**3.29**	**0.77**	**3.65**	**6.36**	5.90	2.40	**5.29**	**2.57**	**5.17**	**2.31**	2.19	**2.43**	**4.22**
Fed-PAIR-l	0.07	3.25	0.60	3.48	**6.36**	5.76	**2.79**	4.84	2.16	4.42	2.08	**3.22**	2.34	4.16

时, 我们称该标签对对应的决策边界区域为未覆盖格.

在此基础上, Fed-PAIR 采用两阶段选择策略.

第一阶段, Fed-PAIR 按照当前训练集中标签格覆盖程度从低到高对所有候选集中非空的标签格进行排序, 即优先考虑 $M_{train}(c_1, c_2)$ 最小的标签格 (代码第 8∼10 行). 随后, Fed-PAIR 依次从这些覆盖最不足的标签格中, 为每一个标签格选择一个不确定性比值 $r(x)$ 最大的样本加入选择集合 (代码第 11∼19 行). 该阶段优先修复当前训练数据中最为稀缺的决策边界区域, 从而以最小代价逐步抬升整体训练数据在标签对层面的最低覆盖水平, 并系统性缓解由 Non-IID 训练数据引入的结构性失衡.

当第一阶段完成或剩余预算不足以覆盖所有待修复标签格时, Fed-PAIR 进入第二阶段, 在所有仍处于覆盖不足状态的标签格中, 全局选取剩余不确定性比值最大的样本, 直至达到给定标注预算 (代码第 21∼26 行).

该两阶段策略在保证高不确定性优先的同时, 显式约束所选数据在标签对层面的覆盖均衡性, 从而实现对决策空间的系统性修复, 并促进整体训练数据在客户端之间逐步趋于更加均衡的标签分布状态.

5.2　实验结果

表 10 汇报了所提出的 Fed-PAIR 方法在不同训练数据分布与候选数据分布条件下, 相对于最优基线方法的性能对比结果. "-g" 代表使用全局模型选择, "-l" 代表使用本地模型选择. "Best Baseline" 表示在各设置下表现最优的现有 DSMR 方法. 从整体平均结果可以观察到, Fed-PAIR-g 在 $Test_{ID}$ 与 $Test_{OOD}$ 上均取得了最优表现, 分别将最优基线的平均准确率从 2.29 提高至 2.43, 从 4.07 提高至 4.22, 表明该方法在分布内与分布外场景下均能够带来稳定的性能增益.

进一步观察不同训练分布条件下的结果可以发现, Fed-PAIR-g 在几乎所有的场景下 (11/12 个数据分布设置) 均能够取得比基线更好的性能表现, 并在任何设置下均优于随机选择的表现. 当训练数据分布为 Non-IID ($\beta = 1.0$) 时, Fed-PAIR-g 在 $Test_{ID}$ 与 $Test_{OOD}$ 上分别达到 0.77 与 3.65, 均为该设置下的最优结果, 表明所提出的方法在中度异质场景下能够有效改善模型修复效果. 在更为极端的 Non-IID 设置 ($\beta = 0.1$) 下, Fed-PAIR-g 在 $Test_{ID}$ 上仍保持领先, 并在 $Test_{OOD}$ 上与该设置下最优基线方法 (LoGo) 性能相当 (5.90 vs. 5.91), 表明该方法在严重标签偏斜条件下依然能够保持较为稳定的性能表现, 具有良好的鲁棒性.

此外, Fed-PAIR-g 在多数设置下均优于 Fed-PAIR-l, 再次验证了前文研究问题 1 的结论, 即在联邦 DSMR 场景中, 全局模型相比本地模型更适合作为查询选择器.

综上, 该结果表明 Fed-PAIR 能够有效地将前述关于查询选择器可靠性与数据分布均衡性的实证结论转化为实际可观的模型修复性能提升.

6　讨论

6.1　使用指导

我们总结了在联邦学习中使用 DSMR 方法的若干实践指导原则.

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

表 11　在 AccDFL 框架下, 不同 DSMR 方法在 Brain Tumor 数据集上的 $Test_{OOD}$ 准确率提升 (%). 训练数据分布为 Non-IID ($\beta = 1.0$), 数据集为 Brain Tumor. 最优结果以粗体标出.

Table 11　$Test_{OOD}$ accuracy improvement (%) of different DSMR methods on the Brain Tumor dataset under the AccDFL framework. Training data distribution is Non-IID ($\beta = 1.0$). The best result is in bold.

	Random	LoGo	MCP	LC	DeepGini	Fed-PAIR
Accuracy improvement	4.58	1.56	5.51	6.21	5.51	**7.05**

• 鉴于均衡的数据分布有助于提升 DSMR 在联邦学习中的修复效果, 在选择并标注一定预算数量的新数据后, 可进一步结合数据增强技术生成额外样本, 以改善补丁数据与原始训练数据之间的标签分布均衡性.

• 当训练数据与候选数据均呈现严重 Non-IID 分布时, 仅依赖 DSMR 来维护联邦学习模型的效果可能较为有限. 在此情形下, 将 DSMR 与其他模型维护技术相结合, 例如深度度量学习 [78] 和不确定性建模 [79], 可能是一种更为有效的替代方案.

• 在我们的实验设定中, 对于单个客户端, 服务器端不向其共享其他客户端的数据分布信息, 这在一定程度上限制了分布感知型 DSMR 方法的效果. 我们的探索性研究表明, 客户端之间保持较为均衡的数据分布有助于更有效地修复联邦学习模型. 因此, 在实际应用中, 若条件允许, 建议服务器端向各客户端共享部分分布信息 (例如平均输出概率分布), 以进一步提升 DSMR 在联邦学习场景下的有效性.

6.2　基于去中心化联邦学习架构的 DSMR 验证

前述实验均基于中心化联邦学习框架展开. 近年来, 去中心化联邦学习 (DFL) 因无需依赖中央服务器而在医疗物联网等隐私敏感场景中受到广泛关注 [80]. 为探索 DSMR 方法在更广泛联邦学习架构下的适用性, 本节在典型 DFL 框架 AccDFL [80] 上开展了补充验证.

实验采用脑肿瘤医学图像数据集[1)], 以 ResNet-18 为骨干网络, 设置 $N = 10$ 个客户端, 训练数据分布为 Non-IID ($\beta = 1.0$), 评估指标为 $Test_{OOD}$ 准确率提升. 值得注意的是, AccDFL 采用纯点对点通信拓扑, 不存在全局模型, 各客户端只能以本地模型作为查询选择器, 这与研究问题 1 中全局模型更优的结论形成直接对照.

实验结果如表 11 所示. 整体而言, 各方法在去中心化架构下仍能带来正向的准确率提升, 表明 DSMR 方法具备一定的架构普适性. 然而, LoGo 仅提升 1.56%, 低于随机选择 (4.58%), 这与其两阶段机制高度依赖全局模型进行不确定性采样密切相关, 进一步印证了研究问题 1 的结论. Fed-PAIR 在所有方法中取得最高提升 (7.05%), 表明其分布再均衡机制对查询选择器质量的依赖相对较低, 在全局模型不可用的场景下仍能保持较强的修复效果.

6.3　方法与结论的泛化性讨论

本文主要实验在 10 个客户端的设置下开展, 该规模在联邦学习研究中较为常见 [47, 81~84], 能够在控制通信与计算成本的同时稳定反映不同方法之间的差异. 然而, 在实际联邦学习系统中, 会存在客户端数量较多的情况, 这可能会带来更强的数据异质性与训练不稳定性. 为分析该因素对结果泛化性和方法有效性的影响, 本文在 Fashion-MNIST 数据集上进一步扩展了实验规模, 并分别设置 20 个与 50 个客户端进行验证, 结果在表 12~14 中展示. 实验结果显示, 在客户端数量增加的情况下, 实验结果仍呈现与主实验一致的变化趋势, 说明本文的方法以及得到的实证研究结果对客户端规模具有一定的鲁棒性.

此外, 从方法设计角度来看, 所提出框架主要在客户端内部独立运行, 不依赖客户端之间或客户端与服务器之间的数据共享, 因此其计算与通信开销不会随客户端数量线性增长. 这一特性使得该方法能够自然扩展至 cross-device 联邦学习等更大规模场景. 同时, 由于方法不涉及原始数据交换, 其适用性在差分隐私等设置受限下仍然能够保持. 然而, 需要指出的是, 当客户端规模进一步扩大时, 系统级因素 (如通信延迟、参与率波动等) 可能对整体训练动态产生额外影响, 相关复杂配置下的系统性评估仍有待未来工作进一步探索.

1) https://www.kaggle.com/datasets/jakeshbohaju/brain-tumor/data.

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

表 12　客户端数量为 20 和 50 时, 使用全局模型或本地模型所选数据进行再训练后, 在 Fashion-MNIST 数据集上的分布内测试集 ($Test_{ID}$) 与分布外测试集 ($Test_{OOD}$) 上的平均准确率提升 (%), 其中 TrainD 表示训练数据分布, 最优结果以粗体标出.

Table 12　Average accuracy improvement (%) of models on Fashion-MNIST dataset with 20 and 50 clients (using data selected by the global or local models against ID test set ($Test_{ID}$) and OOD test set ($Test_{OOD}$), where TrainD refers to training dataset distribution and the best results are in bold.

Clients	TrainD	Test set	Candidate data distribuion							
			IID		Non-IID (β=1.0)		Non-IID (β=0.1)		Average	
			Global	Local	Global	Local	Global	Local	Global	Local
20	IID	$Test_{ID}$	0.02	0.03	0.02	0.02	0.08	0.07	**0.04**	**0.04**
		$Test_{OOD}$	0.56	0.49	0.57	0.51	0.38	0.37	**0.50**	0.46
	Non-IID (β=1.0)	$Test_{ID}$	0.17	0.17	0.15	0.16	0.27	0.23	**0.20**	0.18
		$Test_{OOD}$	0.93	0.90	0.85	0.88	0.79	0.74	**0.85**	0.84
	Non-IID (β=0.1)	$Test_{ID}$	0.64	0.38	0.38	0.09	0.38	0.43	**0.47**	0.30
		$Test_{OOD}$	1.44	1.15	1.40	0.94	1.21	1.29	**1.35**	1.13
	Average	$Test_{ID}$	**0.28**	0.19	**0.18**	0.09	**0.24**	**0.24**	**0.23**	0.18
		$Test_{OOD}$	**0.98**	0.85	**0.94**	0.78	0.79	**0.80**	**0.90**	0.81
50	IID	$Test_{ID}$	−0.10	−0.10	−0.09	−0.08	−0.04	−0.04	−0.08	**−0.07**
		$Test_{OOD}$	0.50	0.49	0.47	0.44	0.38	0.37	**0.45**	0.43
	Non-IID (β=1.0)	$Test_{ID}$	−0.17	−0.13	−0.15	−0.15	−0.19	−0.20	−0.17	**−0.16**
		$Test_{OOD}$	0.45	0.43	0.45	0.41	0.45	0.36	**0.45**	0.40
	Non-IID (β=0.1)	$Test_{ID}$	0.26	0.24	0.24	0.14	0.14	0.13	**0.21**	0.17
		$Test_{OOD}$	0.54	0.48	0.69	0.51	0.58	0.34	**0.60**	0.44
	Average	$Test_{ID}$	0.00	**0.01**	**0.00**	−0.03	−0.03	−0.03	−0.01	−0.02
		$Test_{OOD}$	**0.50**	0.47	**0.54**	0.45	**0.47**	0.36	**0.50**	0.43

6.4　有效性威胁

内部有效性威胁主要来源于我们所采用的联邦学习流程以及 DSMR 方法的实现. 在联邦学习方面, 我们采用了被广泛研究的 FedAvg 方法 [14], 并基于成熟的开源项目 [64, 75] 进行实现, 该实现具有良好的可扩展性. 在 DSMR 方面, 我们复用了多种方法 [40, 42, 44] 的官方实现. 为保证实验的可复现性, 我们在项目网站上公开了完整的代码、实验配置以及全部实验结果 [63].

外部有效性威胁主要与方法、数据集和模型的选择有关. 在方法层面, 我们同时涵盖了基础方法 (如 Least Confidence) 和当前先进方法 (如 LoGo) 的联邦主动学习方法, 以及具有代表性的 DSMR 方法 (例如 DeepGini). 与以往主要聚焦于图像数据的联邦学习研究不同 [33, 35, 85], 本研究同时覆盖了图像、代码和文本 3 种模态的数据集, 共包含 7 个数据集和 6 种模型架构, 涵盖前馈神经网络、循环神经网络以及预训练 Transformer 模型. 因此, 我们的结论在一定程度上具有超出实验设置的泛化性.

本文实验采用全客户端参与 (γ=1.0) 的设置, 旨在排除客户端采样随机性的干扰, 从而更清晰地对比不同数据选择策略在模型修复任务上的本质差异. 从方法设计角度来看, Fed-PAIR 方法主要在客户端内部独立运行, 其数据选择过程不依赖于客户端的参与比例, 因此客户端参与比例的降低不会对方法本身的有效性产生本质影响. 然而, 在实际联邦学习部署中, 部分客户端参与的场景下, 客户端采样策略将与数据选择策略产生交互, 可能对整体模型修复效果引入额外的不确定性. 因此, 尽管方法本身具有鲁棒性, 相关复杂配置下的系统性评估仍有待未来工作进一步探索.

此外, 本文实验在标注预算设置上选取了 1%, 5% 和 10% 三个代表性比例. 这一选择能够初步揭示不同数据选择方法在模型修复任务上的性能趋势, 但对于实际应用中可能出现的预算极低 (如 0.5%) 或较高 (如 20%)

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

表 13　客户端数量为 20 时, 不同数据选择方法在 Fashion-MNIST 数据集上的平均准确率提升 (%), 其中 TrainD 表示训练数据分布, CandiD 表示候选数据分布, 最优结果以粗体标出.

Table 13 Average accuracy improvement (%) of different selection methods on the Fashion-MNIST dataset with 20 clients, where TrainD refers to training dataset distribution, CandiD refers to candidate dataset distribution, and the best results are in bold.

Distribution		Test set	Random	LoGo	LC-g	MCP-g	Gini-g	LC-l	MCP-l	Gini-l	Fed-Pair-g	Fed-Pair-l
TrainD	IID	$Test_{ID}$	0.06	0.07	0.04	0.03	0.05	0.04	0.05	0.04	**0.11**	0.07
		$Test_{OOD}$	0.22	0.21	0.35	0.79	0.37	0.31	0.72	0.35	**0.93**	0.82
	Non-IID (β=1.0)	$Test_{ID}$	0.24	**0.28**	0.23	0.12	0.24	0.20	0.14	0.21	0.27	0.24
		$Test_{OOD}$	0.53	0.53	0.77	1.04	0.75	0.72	1.02	0.78	**1.18**	1.12
	Non-IID (β=0.1)	$Test_{ID}$	0.41	0.44	0.60	0.20	0.60	0.34	0.14	0.42	**0.63**	0.60
		$Test_{OOD}$	1.11	1.12	1.42	1.34	1.29	1.22	1.07	1.10	**1.44**	1.39
CandiD	IID	$Test_{ID}$	0.31	0.27	**0.37**	0.10	**0.37**	0.22	0.10	0.26	0.33	0.28
		$Test_{OOD}$	0.66	0.70	0.88	1.18	0.86	0.74	0.97	0.83	**1.33**	1.20
	Non-IID (β=1.0)	$Test_{ID}$	0.21	0.20	0.22	0.08	0.25	0.09	0.03	0.15	0.38	**0.42**
		$Test_{OOD}$	0.69	0.51	0.88	1.14	0.79	0.73	0.93	0.67	**1.32**	1.28
	Non-IID (β=0.1)	$Test_{ID}$	0.19	0.32	0.28	0.18	0.27	0.27	0.19	0.27	**0.48**	0.34
		$Test_{OOD}$	0.51	0.65	0.77	0.84	0.77	0.77	0.91	0.72	**1.23**	1.09
Average		$Test_{ID}$	0.24	0.26	0.29	0.12	0.30	0.19	0.11	0.23	**0.37**	0.33
		$Test_{OOD}$	0.62	0.62	0.85	1.06	0.81	0.75	0.94	0.74	**1.24**	1.15

表 14　客户端数量为 50 时, 不同数据选择方法在 Fashion-MNIST 数据集上的平均准确率提升 (%), 其中 TrainD 表示训练数据分布, CandiD 表示候选数据分布, 最优结果以粗体标出.

Table 14 Average accuracy improvement (%) of different selection methods on the Fashion-MNIST dataset with 50 clients, where TrainD refers to training dataset distribution, CandiD refers to candidate dataset distribution, and the best results are in bold.

Distribution		Test set	Random	LoGo	LC-g	MCP-g	Gini-g	LC-l	MCP-l	Gini-l	Fed-Pair-g	Fed-Pair-l
TrainD	IID	$Test_{ID}$	−0.02	−0.02	−0.05	−0.10	−0.08	−0.04	−0.10	−0.08	**0.06**	0.01
		$Test_{OOD}$	0.18	0.31	0.36	0.59	0.39	0.33	0.56	0.41	**0.88**	0.86
	Non-IID (β=1.0)	$Test_{ID}$	−0.11	−0.13	−0.11	−0.21	−0.18	−0.14	−0.20	−0.14	−0.02	**0.00**
		$Test_{OOD}$	0.26	0.36	0.36	**0.56**	0.44	0.32	0.49	0.41	0.52	0.50
	Non-IID (β=0.1)	$Test_{ID}$	0.16	0.20	0.26	0.05	0.33	0.21	0.19	0.12	0.32	**0.35**
		$Test_{OOD}$	0.49	0.59	0.57	0.60	0.65	0.40	0.50	0.42	**0.70**	0.65
CandiD	IID	$Test_{ID}$	0.02	0.04	0.03	−0.09	0.06	0.01	−0.01	0.01	**0.09**	0.06
		$Test_{OOD}$	0.33	0.47	0.42	0.61	0.46	0.38	0.56	0.47	**0.91**	0.74
	Non-IID (β=1.0)	$Test_{ID}$	0.01	0.04	0.07	−0.11	0.05	0.01	−0.04	−0.06	**0.33**	0.11
		$Test_{OOD}$	0.33	0.47	0.47	0.61	0.52	0.38	0.54	0.44	**0.84**	0.63
	Non-IID (β=0.1)	$Test_{ID}$	0.00	−0.03	0.00	−0.05	−0.05	0.01	−0.06	−0.05	**0.19**	0.09
		$Test_{OOD}$	0.26	0.32	0.39	**0.52**	0.50	0.30	0.44	0.33	0.35	0.35
Average		$Test_{ID}$	0.01	0.02	0.03	−0.09	0.02	0.01	−0.04	−0.03	**0.16**	0.10
		$Test_{OOD}$	0.31	0.42	0.43	0.58	0.49	0.35	0.52	0.41	**0.70**	0.62

的场景未做系统探讨. 特别是在预算极低的情况下, 基于不确定性的方法是否仍能稳定优于随机选择, 尚需进一步实验验证. 未来研究可在更宽泛的预算范围内对数据选择策略的有效性进行更全面的评估.

在聚合策略方面, 本文所提出的结论与方法设计均不依赖于特定的联邦聚合算法. Fed-PAIR 方法作用于客户端本地, 其数据选择过程与服务器端的聚合方式无关, 因此本文结论在不同聚合策略下应具有一致性. 我们在 6.2 小节对 DSMR 在去中心化联邦学习架构下的有效性分析也进一步验证了这一点, 实验结果表明, 在去中心

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

化架构下 Fed-PAIR 方法仍能带来正向的准确率提升, 这为本文结论在更广泛聚合架构下的适用性提供了初步支持. 然而, 近年来涌现出多种改进的聚合策略 (如 FedProx, SCAFFOLD 等), 这些方法在处理非独立同分布数据时可能表现出不同的收敛特性. 尽管本文结论本身与聚合策略无关, 但在不同聚合策略的设定下, 数据选择方法与聚合方法的交互作用是否会对模型修复效果产生进一步影响, 仍有待未来工作系统探究.

其他有效性威胁主要来源于模型重训练和评估过程中的随机性. 为缓解该问题, 我们对所有实验结果均取三次独立运行的平均值. 此外, 不同于仅评估分布内测试准确率的既有研究 [14, 71, 85, 86], 我们进一步在分布外测试数据上评估模型的泛化性能.

7　总结

本文首次对基于数据选择的模型修复方法在联邦学习场景下的有效性进行了系统的实证研究, 旨在应对联邦学习模型在面对未知数据分布时的模型维护问题. 通过在 7 个数据集和 6 种模型架构上的全面实验, 我们揭示了 DSMR 在联邦学习场景下的若干关键行为特征. 实验结果表明, 现有 DSMR 方法在复杂 Non-IID 与分布偏移条件下整体仍处于早期阶段, 多数方法仅能带来有限且不稳定的性能增益. 进一步的分析揭示了两条关键经验规律: 一方面, 不确定性驱动仍是数据选择过程中不可或缺的核心因素; 另一方面, 补丁数据与训练数据之间的分布均衡性与模型修复效果密切相关, 而现有方法普遍缺乏对其联合分布结构的显式约束. 基于上述实证结论, 本文提出了一种面向联邦学习的数据选择方法 Fed-PAIR, 通过不确定性驱动与分布再均衡的联合约束, 显式建模补丁数据与训练数据之间的分布耦合关系, 从而在复杂 Non-IID 场景下实现更加稳定且鲁棒的模型修复. 实验结果表明, Fed-PAIR 在 IID 与 Non-IID 设置下均能够取得稳定的性能提升. 综上, 本文不仅为联邦学习场景下 DSMR 方法的实际有效性提供了系统性的实证依据, 也表明针对联邦 Non-IID 特性设计不确定性驱动且具备分布再均衡能力的数据选择策略, 是提升联邦模型长期维护效果的一条重要途径.

参考文献

1 Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. npj Digit Med, 2020, 3: 1–7
2 Shi L, Ji Q Y, Chen Q W, et al. A survey on vision Transformers for medical image analysis. J Comput Engin Appl, 2023, 59: 41–55 [石磊, 籍庆余, 陈清威, 等. 视觉 Transformer 在医学图像分析中的应用研究综述. 计算机工程与应用, 2023, 59: 41–55]
3 Niknam S, Dhillon H S, Reed J H. Federated learning for wireless communications: motivation, opportunities, and challenges. IEEE Commun Mag, 2020, 58: 46–51
4 Nguyen D C, Ding M, Pathirana P N, et al. Federated learning for Internet of Things: a comprehensive survey. IEEE Commun Surv Tutorials, 2021, 23: 1622–1658
5 Yu R, Li P. Toward resource-efficient federated learning in mobile edge computing. IEEE Netw, 2021, 35: 148–155
6 Zhuang W, Chen C, Lyu L. When foundation model meets federated learning: motivations, challenges, and future directions. 2023. ArXiv:2306.15546
7 Chen H, Zhang Y, Krompass D, et al. Feddat: an approach for foundation model finetuning in multi-modal heterogeneous federated learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 11285–11293
8 Ren C, Yu H, Peng H, et al. Advances and open challenges in federated learning with foundation models. 2024. ArXiv:2404.15381
9 Koh P W, Sagawa S, Marklund H, et al. Wilds: a benchmark of in-the-wild distribution shifts. In: Proceedings of International Conference on Machine Learning, 2021. 5637–5664
10 Guo T, Banerjee T, Liu R, et al. Building real-time awareness of out-of-distribution in trajectory prediction for autonomous vehicles. 2024. ArXiv:2409.17277
11 Filos A, Tigkas P, McAllister R, et al. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In: Proceedings of International Conference on Machine Learning, 2020. 3145–3153
12 Zhang Q, Hu S, Sun J, et al. On adversarial robustness of trajectory prediction for autonomous vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 15159–15168
13 Li Y, Chen M, Xu Q. Hybridrepair: towards annotation-efficient repair for deep learning models. In: Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis, 2022. 227–238
14 McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of Artificial Intelligence and Statistics, 2017. 1273–1282

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

15 Gao L, Fu H, Li L, et al. Feddc: federated learning with non-iid data via local drift decoupling and correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 10112–10121
16 Hu M, Cao Y, Li A, et al. Fedmut: generalized federated learning via stochastic mutation. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 12528–12537
17 Wang T, Zheng Z, Lin F. Federated learning framework based on trimmed mean aggregation rules. Expert Syst Appl, 2025, 270: 126354
18 Kim D, Lee K, Lee Y, et al. ACMFed: fair semi-supervised federated learning with additional compromise model. IEEE Access, 2025, 13: 47734–47747
19 Khan N, Nisar S, Khan M A, et al. Optimizing federated learning with aggregation strategies: a comprehensive survey. IEEE Open J Comput Soc, 2025, 6: 1227–1247
20 Uludag K, Eroglu Erdem Ç, Korçak Ö. Leveraging unlabeled data in federated learning: a review. IEEE Access, 2025, 13: 181719
21 Zhu S, Wang Y, Sun G. Federated semi-supervised learning based on truncated Gaussian aggregation. J Supercomput, 2025, 81: 269
22 Yang Q, Chen Z, Peng Z, et al. Relation-guided versatile regularization for federated semi-supervised learning. Int J Comput Vision, 2025, 133: 3312–3326
23 Nguyen V T, Beuran R. FedMSE: semi-supervised federated learning approach for IoT network intrusion detection. Comput Security, 2025, 151: 104337
24 Settles B. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison. 2009. http://digital.library.wisc.edu/1793/60660
25 Balcan M F, Beygelzimer A, Langford J. Agnostic active learning. In: Proceedings of the 23rd International Conference on Machine Learning, 2006. 65–72
26 Beluch W H, Genewein T, Nürnberger A, et al. The power of ensembles for active learning in image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 9368–9377
27 Sener O, Savarese S. Active learning for convolutional neural networks: a core-set approach. 2017. ArXiv:1708.00489
28 Jeong W, Yoon J, Yang E, et al. Federated semi-supervised learning with inter-client consistency & disjoint learning. 2020. ArXiv:2006.12097
29 Zhuang W, Gan X, Wen Y, et al. Collaborative unsupervised visual representation learning from decentralized data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 4912–4921
30 Zhuang W, Wen Y, Zhang S. Divergence-aware federated self-supervised learning. 2022. ArXiv:2204.04385
31 Li M, Li Q, Wang Y. Class balanced adaptive pseudo labeling for federated semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 16292–16301
32 Ahmed L, Ahmad K, Said N, et al. Active learning based federated learning for waste and natural disaster image classification. IEEE Access, 2020, 8: 208518
33 Ahn J H, Ma Y, Park S, et al. Federated active learning (F-AL): an efficient annotation strategy for federated learning. IEEE Access, 2024, 12: 39261–39269
34 Aussel N, Chabridon S, Petetin Y. Combining federated and active learning for communication-efficient distributed failure prediction in aeronautics. 2020. ArXiv:2001.07504
35 Kim S, Bae S, Song H, et al. Re-thinking federated active learning based on inter-class diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 3944–3953
36 Byun T, Sharma V, Vijayakumar A, et al. Input prioritization for testing neural networks. In: Proceedings of 2019 IEEE International Conference On Artificial Intelligence Testing (AITest), 2019. 63–70
37 Dang X, Li Y, Papadakis M, et al. GraphPrior: mutation-based test input prioritization for graph neural networks. ACM Trans Softw Eng Methodol, 2023, 33: 1–40
38 Gao X, Feng Y, Yin Y, et al. Adaptive test selection for deep neural networks. In: Proceedings of the 44th International Conference on Software Engineering, 2022. 73–85
39 Hu Q, Guo Y, Xie X, et al. Evaluating the robustness of test selection methods for deep neural networks. 2023. ArXiv:2308.01314
40 Feng Y, Shi Q, Gao X, et al. Deepgini: prioritizing massive tests to enhance the robustness of deep neural networks. In: Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis, 2020. 177–188
41 Ma S, Liu Y, Lee W C, et al. Mode: automated neural network model debugging via state differential analysis and input selection. In: Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2018. 175–186
42 Shen W, Li Y, Chen L, et al. Multiple-boundary clustering and prioritization to promote neural network retraining. In: Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering, 2020. 410–422
43 Wang J, Chen J, Sun Y, et al. Robot: robustness-oriented testing for deep learning systems. In: Proceedings of 2021 IEEE/ACM

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

43rd International Conference on Software Engineering (ICSE), 2021. 300–311
44 Hu Q, Guo Y, Cordy M, et al. An empirical study on data distribution-aware test selection for deep learning enhancement. ACM Trans Softw Eng Methodol, 2022, 31: 1–30
45 Attaoui M, Fahmy H, Pastore F, et al. Black-box safety analysis and retraining of DNNs based on feature extraction and clustering. ACM Trans Softw Eng Methodol, 2023, 32: 1–40
46 Hao Y, Huang Z, Guo H, et al. Test input selection for deep neural network enhancement based on multiple-objective optimization. In: Proceedings of 2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), 2023. 534–545
47 Li A, Zhang L, Tan J, et al. Sample-level data selection for federated learning. In: Proceedings of IEEE INFOCOM 2021-IEEE Conference on Computer Communications, 2021. 1–10
48 Shi H, Radu V. Data selection for efficient model update in federated learning. In: Proceedings of the 2nd European Workshop on Machine Learning and Systems, 2022. 72–78
49 Gong C, Zheng Z, Wu F, et al. To store or not? Online data selection for federated learning with limited storage. In: Proceedings of the ACM Web Conference 2023, 2023. 3044–3055
50 Xiao H, Rasul K, Vollgraf R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 2017. ArXiv:1708.07747
51 Yang J, Shi R, Wei D, et al. MedMNIST v2—a large-scale lightweight benchmark for 2D and 3D biomedical image classification. Sci Data, 2023, 10: 41
52 Krizhevsky A. Learning multiple layers of features from tiny images. Technical Report, University of Toronto, 2009. https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf
53 Zhou Y, Liu S, Siow J, et al. Devign: effective vulnerability identification by learning comprehensive program semantics via graph neural networks. In: Proceedings of Advances in Neural Information Processing Systems, 2019. 32
54 Lang K. Newsweeder: learning to filter netnews. In: Proceedings of the 12th International Conference on Machine Learning, 1995. 331–339
55 Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009, 1: 2009
56 Zhi Y, Xie X, Shen C, et al. Seed selection for testing deep neural networks. ACM Trans Softw Eng Methodol, 2023, 33: 1–33
57 Lu S, Guo D, Ren S, et al. Codexglue: a machine learning benchmark dataset for code understanding and generation. 2021. ArXiv:2102.04664
58 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86: 2278–2324
59 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
60 Feng Z, Guo D, Tang D, et al. Codebert: a pre-trained model for programming and natural languages. 2020. ArXiv:2002.08155
61 Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press, 2016
62 Graves, A. Long short-term memory. In: Supervised Sequence Labelling with Recurrent Neural Networks. Berlin: Springer, 2012
63 Zhi Y. Data selection-based model repair for federated learning. 2025. https://sites.google.com/view/smrforfl
64 Ji S. A PyTorch implementation of federated learning. Zenodo, 2018. https://doi.org/10.5281/zenodo.4321561
65 Kim J, Feldt R, Yoo S. Guiding deep learning system testing using surprise adequacy. In: Proceedings of 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE), 2019. 1039–1049
66 Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations. In: Proceedings of the International Conference on Learning Representations, 2019
67 Fang X, Ye M, Yang X. Robust heterogeneous federated learning under data corruption. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 5020–5030
68 Yang Z, Shi J, He J, et al. Natural attack for pre-trained models of code. In: Proceedings of the 44th International Conference on Software Engineering, 2022. 1482–1493
69 Dhole K D, Gangal V, Gehrmann S, et al. Nl-augmenter: a framework for task-sensitive natural language augmentation. 2021. ArXiv:2112.02721
70 Li Q, He B, Song D. Model-contrastive federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 10713–10722
71 Wang H, Yurochkin M, Sun Y, et al. Federated learning with matched averaging. 2020. ArXiv:2002.06440
72 Ducoffe M, Precioso F. Adversarial active learning for deep networks: a margin based approach. 2018. ArXiv:1802.09841
73 Gal Y, Islam R, Ghahramani Z. Deep Bayesian active learning with image data. In: Proceedings of International Conference on Machine Learning, 2017. 1183–1192
74 Wang D, Shang Y. A new active labeling method for deep learning. In: Proceedings of 2014 International Joint Conference on

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

Neural Networks (IJCNN), 2014. 112–119
75 Atul-AI08. Federated learning for NLP tasks: a Python implementation of sentiment analysis in federated settings using neural networks with PyTorch. https://github.com/Atul-AI08/Federated-Learning-for-NLP-Tasks
76 Hu M, Zhou P, Yue Z, et al. Fedcross: towards accurate federated learning via multi-model cross-aggregation. In: Proceedings of IEEE International Conference on Data Engineering (ICDE), 2024. 2137–2150
77 Xu Y, Ma Z, Xu H, et al. FedLC: accelerating asynchronous federated learning in edge computing. IEEE Trans Mobile Comput, 2024, 23: 5327–5343
78 Milbich T, Roth K, Sinha S, et al. Characterizing generalization under out-of-distribution shifts in deep metric learning. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 34: 25006–25018
79 Li X, Dai Y, Ge Y, et al. Uncertainty modeling for out-of-distribution generalization. In: Proceedings of International Conference on Learning Representations, 2022
80 Wei M, Yu W, Chen D. AccDFL: accelerated decentralized federated learning for healthcare IoT networks. IEEE Internet Things J, 2025, 12: 5329–5345
81 Xu J, Wang H. Client selection and bandwidth allocation in wireless federated learning networks: a long-term perspective. IEEE Trans Wireless Commun, 2020, 20: 1188–1200
82 Lee G H, Shin S Y. Federated learning on clinical benchmark data: performance assessment. J Med Internet Res, 2020, 22: e20891
83 Deng Y, Lyu F, Ren J, et al. AUCTION: automated and quality-aware client selection framework for efficient federated learning. IEEE Trans Parallel Distrib Syst, 2021, 33: 1996–2009
84 Roy A G, Siddiqui S, Pölsterl S, et al. Braintorrent: a peer-to-peer environment for decentralized federated learning. 2019. ArXiv:1905.06731
85 Lin T, Kong L, Stich S U, et al. Ensemble distillation for robust model fusion in federated learning. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 33: 2351–2363
86 Li T, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks. Proc Mach Learn Syst, 2020, 2: 429–450

https://www.sciengine.com/doi/10.1360/SSI-2026-0024

Data-selection-based model repair for federated learning

Yuhan ZHI[1], Qiang HU[2], Xiaofei XIE[3], Longtian WANG[1], Ming HU[3], Lei MA[4], Xiaohong GUAN[1] & Chao SHEN[1*]

1. *School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China*
2. *College of Intelligence and Computing, Tianjin University, Tianjin 300072, China*
3. *School of Computing and Information Systems, Singapore Management University, Singapore 188065, Singapore*
4. *Department of Information Science, The University of Tokyo, Tokyo 100-0001, Japan*

* Corresponding author. E-mail: chaoshen@mail.xjtu.edu.cn

Abstract Federated learning enables collaborative model training across multiple clients without sharing raw data, providing an effective solution for privacy-sensitive applications. However, in practical deployment, federated models inevitably encounter new data with unknown distributions, which leads to gradual performance degradation over time. How to effectively repair degraded federated models with low maintenance cost remains an open and important problem. To address this challenge, this paper systematically investigates data selection-based model repair methods in the federated learning setting, with a focus on their applicability and effectiveness for long-term model maintenance. Specifically, we study the problem from two key dimensions, namely the choice of query selector and the design of data selection strategies. We compare the performance of using local models and the global model as query selectors, and evaluate a variety of representative data selection methods for federated model repair. Extensive experiments on seven datasets and six different model architectures show that data selection-based model repair in federated learning is still at an early stage, and existing methods fail to provide stable performance gains across different settings. In contrast to previous findings in federated active learning, the global model consistently serves as a more reliable query selector for model repair, indicating that federated model repair is fundamentally different from federated active learning. Further analysis reveals that maintaining a balanced label distribution in the selected repair data plays an important role in improving model repair performance. Based on these empirical findings, we further propose a new data selection method for federated learning, termed Fed-PAIR. Fed-PAIR explicitly models pairwise uncertainty and incorporates the current training data distribution to adaptively adjust the label allocation of the repair data, thereby promoting a more balanced joint distribution of training and repair data across clients. Experimental results demonstrate that Fed-PAIR achieves stable performance improvements under both IID and Non-IID settings, and remains robust even under severe label skew. Overall, this work provides a systematic empirical analysis of federated model repair and shows that designing data selection strategies tailored for federated Non-IID scenarios is a promising direction for improving the effectiveness of long-term model maintenance.

Keywords deep learning, data selection, model repair, federated learning, data distribution

https://www.sciengine.com/doi/10.1360/SSI-2026-0024