

基于动态聚类 and 信誉评估的抗攻击联邦框架

张唯一, 罗启瑞, 张琳

(南京邮电大学计算机学院, 江苏 南京 210003)

摘要: 联邦学习在数据异构场景下的实际应用面临着严峻挑战, 即恶意客户端的投毒攻击会降低全局模型性能。针对该挑战, 提出了一种基于动态聚类和双维度信誉评估的鲁棒联邦学习框架动态聚类联邦学习 (dynamic clustering federated learning, DCFL)。首先, DCFL 采用无须预设聚类数量的 HDBSCAN 算法实现客户端自适应聚类, 并基于相似性度量对噪声点进行重分配, 将具有相似数据分布特征的客户端动态分组, 避免将异构的良性客户端误判为恶意客户端。其次, 在类簇内部, 综合考察客户端模型更新相似度和损失一致性, 利用双维度的互补指标综合评估客户端更新质量。最后, 设计了动态阈值调整策略, 基于信誉加权的簇内聚合策略有效强化了类簇的个性化功能。实验验证了 DCFL 在 4 种不同数据分布和安全环境下的性能优势, 尤其是在高异构有攻击场景下, 相比传统聚类联邦学习算法测试准确率提升了约 9%, 证明其在复杂联邦环境中具有良好的鲁棒性和适应性。

关键词: 联邦学习; 动态聚类; 数据异构性; 恶意攻击; 信誉评估

中图分类号: TP309

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025069

Anti-attack federated framework based on dynamic clustering and reputation evaluation

ZHANG Weiyi, LUO Qirui, ZHANG Lin

College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract: The practical application of federated learning in scenarios with heterogeneous data faces a serious challenge, which is that poisoning attacks by malicious clients can degrade the global model. A robust federated learning framework DCFL (dynamic clustering federated learning) was proposed based on dynamic clustering and dual-dimensional reputation evaluation. Firstly, DCFL employed the HDBSCAN algorithm for adaptive client clustering without predefined cluster numbers, and redistributed noise points based on similarity measurements, dynamically grouping clients with similar data distribution characteristics, preventing misclassification of heterogeneous but benign clients as malicious. Secondly, within clusters, both model update similarity and loss consistency were examined comprehensively as complementary indicators to evaluate client update quality. Finally, a dynamic threshold adjustment strategy was designed, while reputation-weighted intra-cluster aggregation effectively enhances cluster personalization. Experiments validated DCFL's performance advantages across four different data distribution and security environments. Particularly in high-heterogeneity attack scenarios, DCFL improved testing accuracy by ap-

收稿日期: 2025-05-14; 修回日期: 2025-07-25

通信作者: 张琳, zhangl@njupt.edu.cn

基金项目: 国家自然科学基金 (62372247, 61872194); 南京邮电大学校级自然科学基金 (NY222142)

Foundation Items: The National Natural Science Foundation of China (62372247, 61872194), School-Level Natural Science Foundation of Nanjing University of Posts and Telecommunications (NY222142)

proximately 9 percentage points compared to traditional clustered federated learning algorithms, demonstrating its robustness and adaptability in complex federated environments.

Keywords: federated learning, dynamic clustering, data heterogeneity, malicious attack, reputation evaluation

0 引言

为解决传统中心化机器学习因需要收集用户原始数据而带来的隐私泄露风险, 谷歌公司提出了联邦学习这一分布式学习范式^[1]。在联邦学习中, 各客户端接收全局模型参数后, 在本地数据集上进行训练, 仅需上传模型更新参数而无须共享原始数据, 从而在保护用户隐私方面显著优于传统机器学习方法。然而, 在实际应用过程中, 恶意客户端可能通过多种手段破坏联邦学习系统。例如, 在投毒攻击中, 恶意客户端凭借上传的虚假模型参数, 能够有效降低全局模型的性能。

联邦学习可以整合分散的客户端共同训练全局模型, 但参与训练的客户端之间的个性化特征也不容忽视。现实中, 来源不同的数据往往具有统计异构特性, 这种非独立同分布 (Non-IID) 特性可能导致客户端模型更新方向产生分歧, 引发“客户端漂移”现象, 进而造成全局模型性能下降、收敛速度缓慢甚至训练发散等问题。当前, 研究者已提出若干技术方案应对这一挑战, 其中聚类联邦学习是一条重要的技术路线。

然而, 现有聚类方法在面对恶意攻击时表现不佳, 同时数据异构性也给传统的恶意客户端检测方法带来了挑战。现有的恶意客户端检测方法大多基于数据同分布的假设构建, 在高度异构场景下容易将具有特殊数据分布的良性客户端误判为恶意节点。

综上所述, 可将需要解决的核心问题概括为: 在数据高度异构的联邦学习环境中, 如何设计一个既能够应对数据分布差异又可以抵御恶意攻击的鲁棒性框架。针对该问题, 本文提出一种基于动态聚类与信誉评估的鲁棒联邦学习框架动态聚类联邦学习 (dynamic clustering federated learning, DCFL)。其核心思想是首先对异构客户端进行聚类, 然后在近似同分布的类簇内, 通过评估模型更新质量动态更新客户端信誉值, 并基于信誉值加权聚合类簇模型。

本文的主要贡献如下。

1) 设计了一种基于动态聚类和双维度信誉评估的鲁棒联邦学习框架 DCFL。

2) 设计了一种无须先验知识的动态聚类机制, 基于密度感知自动聚类和基于相似性度量的噪声点重分配策略, 实现对异构客户端的精确分组。

3) 设计了一种双维度更新质量评估体系, 引入相似分布下类簇内损失一致性作为互补指标。从模型参数的方向性偏离和数值异常两个维度综合评估客户端模型相似度和损失一致性。

4) 采用基于信誉度的加权聚合策略保护类簇的个性化特征, 并在信誉更新过程中引入自适应动态阈值调整策略。

1 相关工作

为应对联邦学习中的数据异构性挑战, 基于全局模型优化的方法是一类重要的解决方案。万伟等^[2]提出的 FedUp 框架通过最小化全局经验损失函数的上界来保证模型具有较小的泛化误差, 但该方法难以充分利用数据的个性化特征。Ma 等^[3]提出的 PFL 框架通过识别模型层级重要性来优化聚合效果。Zhang 等^[4]提出的 FedALA 方法针对客户端的本地目标自适应地聚合全局和本地模型。这些方法能够较好地保留客户端特性, 但大多采用统一的聚合策略, 没有充分考虑恶意客户端的影响。应作斌等^[5]针对联邦学习中的固定聚合权重问题, 通过动态学习聚合权重并引入差分隐私机制, 在 Non-IID 环境下实现了更好的隐私保护和模型性能。

聚类方法是应对数据异构性问题的另一有效策略。传统的聚类方法主要基于预设聚类数量的算法实现。Ghosh 等^[6]提出的 IFCA 算法通过交替估计用户聚类身份和优化模型参数实现个性化聚类。Long 等^[7]提出的 FeSEM 算法基于模型参数聚类客户端, 从而训练多个全局模型, 但该方法较依赖先验知识。这些传统方法难以适应动态变化的联邦学习环境, 因此研究者提

出了自适应聚类算法来解决聚类数量预设问题。朱素霞等^[8]提出的ACFL算法通过自适应加速聚类和正向反馈机制划分不同任务簇，提高了聚类效率，但该方法在复杂攻击环境下的鲁棒性仍有待验证。

恶意攻击是联邦学习面临的一大挑战，现有防御方法可分为隐私保护和异常客户端检测两类。Liu等^[9]提出的PEFL框架采用同态加密防御推理攻击，并通过去除恶意梯度和惩罚恶意客户端来防御投毒攻击。康海燕等^[10]提出了一种减少性能损失的本地化差分隐私联邦学习机制，探索了差分隐私在联邦学习中的应用。Yang等^[11]提出将原始输入拆分为性能敏感特征和性能鲁棒特征，并通过加噪保障隐私安全。该方法虽然能够有效地保护数据隐私，但防御主动恶意攻击的能力有限。异常客户端检测方法通过分析客户端行为特征主动识别恶意节点。刘艳等^[12]通过计算本地更新参数和全局模型参数的余弦相似度来判断本地更新的必要性。Zhang等^[13]提出的FL-Detector算法基于Cauchy中值定理和L-BFGS算法判断客户端更新的一致性。Chen等^[14]设计了基于残差检测和相似度聚合的联邦学习方法，能够准确捕获和处理对抗性模型更新。Gu等^[15]提出的基于谱分析的FedCut方法，将攻击者检测问题转化为加权图中的社区检测问题。这些方法在理想环境下表现出色，但在异构场景下，数据分布的自然差异与恶意行为的人为异常在特征空间中往往重叠，容易产生误判。徐明迪等^[16]提出结合个性化层和原型学习机制，基于3阶段防御来抵御异构环境下的投毒攻击。本文提出的DCFL框架通过动态聚类技术将异构性处理与恶意检测分离，在近似同分布的类簇环境中识别恶意行为。

2 方案实现

2.1 框架概述

DCFL框架如图1所示，其工作流程可以划分为以下4个阶段。

1) 初始训练阶段：在训练初期，所有客户端共同参与训练一个全局模型。

2) 动态聚类阶段：当客户端之间的模型更新差异超过阈值且平均更新幅度趋于稳定时，采用

HDBSCAN算法基于客户端之间的相似度进行动态聚类。该阶段无须预先设定类簇数量，其核心任务在于将识别出的噪声点依据相似度度量重新分配到最匹配类别中。

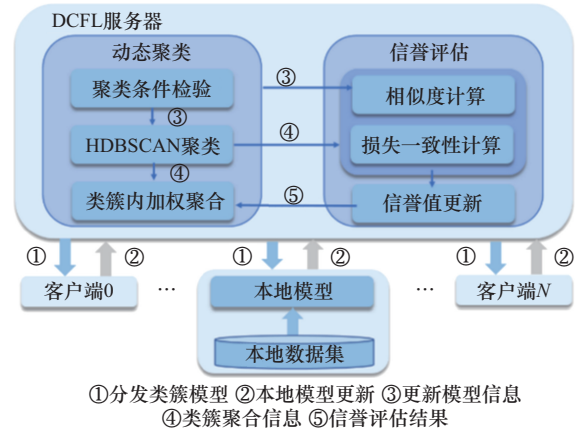


图1 DCFL框架

3) 信誉评估阶段：聚类后，类簇内部形成近似同分布的环境。DCFL从双维度评估客户端更新质量，一是计算客户端模型更新与类簇中心的相似度，二是客户端损失与类簇内平均损失的一致性。

4) 加权聚合阶段：以信誉值为权重对客户端更新进行加权聚合，结合自适应动态阈值调整策略，实现对客户端行为的精确评估。该机制既能保护类簇模型的个性化性能，又能有效降低恶意客户端对联邦学习系统的影响。

2.2 动态聚类算法

动态聚类算法主要包含3个部分，即聚类触发机制、基于HDBSCAN的动态聚类和噪声点重分配策略。

在训练初期，每个客户端模型的收敛程度不同，过早聚类可能导致错误分组，因此本文采用基于更新稳定性的聚类触发机制。在每一轮训练中，计算所有客户端模型更新的平均范数 mean_norm 和最大范数 max_norm 。当同时满足式(1)~式(3)时，触发聚类操作。

$$\text{mean_norm} = \left\| \frac{1}{|C|} \sum_{i \in C} \Delta \theta_i \right\| < \varepsilon_1 \quad (1)$$

$$\text{max_norm} = \max_{i \in C} \|\Delta \theta_i\| > \varepsilon_2 \quad (2)$$

$$t > T_{\text{warmup}} \quad (3)$$

其中， C 为当前类簇， $\Delta \theta_i$ 为客户端 i 权重更新的

向量, ε_1 和 ε_2 为聚类阈值, T_{warmup} 为预热轮次。式(1)表明 mean_norm 较小, 大多数客户端模型已经接近稳定点; 式(2)表明 max_norm 较大, 客户端之间存在显著差异, 部分模型参数梯度变化较大, 全局模型未完全收敛; 式(3)表明模型已经完成充分的初始训练, 模型更新趋于稳定。综合判断, 当客户端模型趋于收敛而全局模型尚未完全收敛时, 聚类操作才启动。

聚类触发后, DCFL 采用基于密度的层次聚类算法 HDBSCAN^[17] 作为核心聚类机制。该算法的优势在于能够自动识别不同密度的类簇, 无须提前设定类簇数量, 可以识别任意形状的聚类结构, 并对噪声数据具有鲁棒性。通过计算两个客户端之间模型参数更新的余弦相似度 (其梯度更新的余弦相似度可以表示为 α_{ij}), 判断其收敛方向是否一致, 进而可以判断两个客户端是否属于同一类簇。每轮训练结束后, 计算客户端模型更新之间的余弦相似度矩阵 $\mathbf{S} \in R^{N \times N}$ 。

在动态聚类过程中, HDBSCAN 算法会将位于低密度区域的客户端标记为噪声点, 该区域的客户端包含异构的正常客户端和行为隐蔽的恶意客户端两类对象。本文在 DCFL 中设计了基于相似性度量的噪声点重分配策略: 对于每一个噪声客户端, 计算其与所有已有类簇模型的相似度 s_{ik} , 并将其分配至相似度最高的类簇。该策略能够将异构的正常客户端归入与其更新方向最接近的类簇, 从而保留其个性化特征, 同时初步隔离潜在的恶意节点。后续由信誉机制完成二次过滤。

动态聚类算法的具体流程如算法1所示。

算法1 动态聚类算法

输入 客户端集合 C , 相似度矩阵 $\mathbf{S}$, 聚类阈值 ε_1 和 ε_2 , 预热轮次 T_{warmup}

输出 聚类后的客户端类簇集合 $\{C_1, C_2, \dots, C_k\}$

1) if $t > T_{\text{warmup}}$ and $\text{mean_norm} = \left\| \frac{1}{|C|} \sum_{i \in C} \Delta \theta_i \right\| < \varepsilon_1$ and $\text{max_norm} = \max_{i \in C} \|\Delta \theta_i\| > \varepsilon_2$: //满足聚类条件

2) $\mathbf{D} \leftarrow 1 - \mathbf{S}$ //将相似度矩阵转化为标准化的距离矩阵

3) $\{C_1, C_2, \dots, C_k\} \leftarrow \text{HDBSCAN}(\mathbf{D})$ //采用 HDBSCAN 算法进行聚类

4) for each $c_i \in C$:

5) if $\text{label}_i = -1$ //噪声点处理

6) 计算噪声点与每个 center_k 的相似度 s_{ik}

7) $\text{label}_i \leftarrow \arg \max_k s_{ik}$ //基于相似性度量的重分配策略

8) else:

9) 保持原类簇

10) return $\{C_1, C_2, \dots, C_k\}$

2.3 类簇内信誉机制

2.3.1 双维度更新质量评估

聚类后, 同一类簇内客户端 i 和 j 的数据分布 P_i 和 P_j 相似。计算客户端 θ_i 和类簇模型 θ_j 的模型相似度 α_{ic} , 以捕捉参数空间的方向性偏离。假设恶意客户端 m 主要将类别 y_A 的标签翻转为 y_B , 而其他类别保持不变。此时 $\Delta \theta_m^{\text{adv}}$ 主要影响 y_A 和 y_B 相关的特征空间, 而对高维参数空间大部分维度的影响较小, 两者的余弦相似度依然可能较高。由此可见, 仅靠余弦相似度难以全面评估客户端的更新质量。为此, 本文引入损失一致性作为补充指标。

损失一致性指标基于一个重要理论, 即在数据分布相似的类簇内, 正常客户端的损失函数值往往一致。以下为理论证明过程。

定理1 类簇内损失一致性

对于数据分布相似的客户端 i 和 j , 即 $d(P_i, P_j) \leq \varepsilon$, 经过联邦训练后, 其损失函数值之差存在上界。

根据 Wasserstein 距离的性质, 客户端 i 和 j 的数据分布 P_i 和 P_j 相似时, 对于参数为 θ 的模型 f_θ , 存在常数 $M > 0$, 使得客户端 i 和 j 之间的梯度更新之差 $\leq M \cdot \varepsilon$, 即式(4)成立。

$$\begin{aligned} & \left\| \nabla_{\theta} \mathcal{L}_{P_i}(\theta) - \nabla_{\theta} \mathcal{L}_{P_j}(\theta) \right\| = \\ & \left\| \mathbb{E}_{(x,y) \sim P_i} [\nabla_{\theta} l(f_{\theta}(x), y)] - \mathbb{E}_{(x,y) \sim P_j} [\nabla_{\theta} l(f_{\theta}(x), y)] \right\| \leq \\ & M \cdot d(P_i, P_j) \leq M \cdot \varepsilon \end{aligned} \quad (4)$$

其中, M 表示损失函数梯度 $\nabla_{\theta} \mathcal{L}_{P_i}(\theta)$ 对分布变化

的敏感系数, $\mathbb{E}_{(x,y) \sim P_i}[\cdot]$ 表示期望损失。式(4)表明, 当客户端之间的数据分布相似时, 其训练后的梯度结果也相似。

客户端 i 和 j 经过 t 步后的模型差异通过递归方式可以推导出, 具体形式如式(5)所示。

$$\|\theta_i^t - \theta_j^t\| = \left\| -\eta \sum_{k=0}^{t-1} \left(\nabla_{\theta} \mathcal{L}_{P_i}(\theta_i^k) - \nabla_{\theta} \mathcal{L}_{P_j}(\theta_j^k) \right) \right\| \leq (5) \eta \cdot M \cdot \varepsilon \cdot t \cdot C$$

其中 C 表示梯度的平滑性系数。若两个客户端的梯度相似, 则会产生相似的模型参数。

为了计算客户端 i 和 j 的损失差异, 可结合损失函数关于模型参数的 Lipschitz 连续性假设和式(5), 求解在数据分布相似情况下客户端之间的损失差异的上界, 如式(6)所示。

$$\begin{aligned} \|\mathcal{L}_{P_i}(\theta_i^t) - \mathcal{L}_{P_j}(\theta_j^t)\| &= \\ \|\mathcal{L}_{P_i}(\theta_i^t) - \mathcal{L}_{P_i}(\theta_j^t)\| + \|\mathcal{L}_{P_i}(\theta_j^t) - \mathcal{L}_{P_j}(\theta_j^t)\| &\leq (6) \\ (K \cdot \eta \cdot M \cdot t \cdot C + N) \cdot \varepsilon &= \delta(\varepsilon) \end{aligned}$$

其中, $\mathcal{L}_{P_i}(\theta_i^t)$ 表示分布 P_i 下在轮次 t 时的损失值, η 表示学习率, 常数 K 表示损失函数对参数的敏感性, 常数 N 表示损失函数值对数据分布变化的敏感程度。该上界表明, 客户端之间的数据分布差异 ε 越小, 客户端之间的损失值差异越小。

设客户端 i 的损失值为 $\mathcal{L}_i$, 类簇内的平均损失值为 μ_c , 损失值的标准差为 σ_c , 客户端 i 的损失一致性 β_{ic} 如式(7)所示。

$$\beta_{ic} = \exp\left(-\frac{|\mathcal{L}_i - \mu_c|}{\sigma_c}\right) \quad (7)$$

β_{ic} 越接近 1, 表明客户端损失值与类簇内的平均损失值越接近。

在双维度评估体系下, 当前轮次客户端的行为评分由模型相似度和损失一致性两项指标加权求得, 如式(8)所示。

$$\text{Score}_i = \lambda \cdot \alpha_{ic} + (1 - \lambda) \cdot \beta_{ic} \quad (8)$$

其中, λ 表示动态调整因子。在训练初期, 模型参数处于随机状态, 余弦相似度计算容易受到参数空间噪声的干扰, 而损失一致性则相对更加稳定和可靠, 此时值较小, 通过 Sigmoid 函数平滑增长至稳定。当类簇内平均损失异常偏高时, 表明模型可能出现发散或受到恶意攻击, 系统将自

动调低 λ 至较小值。该双重策略确保了评估体系在不同训练阶段和异常情况下的自适应性。

2.3.2 信誉值更新

DCFL 统计客户端的正面和负面行为次数, 并依据文献[18]采用 Beta 函数的期望计算其信誉值 R_i , 具体如式(9)所示。其中, a_i 表示客户端的正面行为次数, b_i 表示负面行为次数。

$$R_i = \frac{a_i + 1}{a_i + b_i + 2} \quad (9)$$

本文在 DCFL 框架中引入了基于客户端信誉值的自适应动态阈值机制, 通过阈值参数 h 动态调整各客户端更新的评判标准。具体而言, 参数 h 取所属类簇内评分的 α 分位数。若某客户端评分低于阈值, 则认为其在近似独立同分布条件下对该类簇模型的贡献较小, 进而判定该轮更新为负面行为。在训练前期, 模型参数波动较大, α 分位数设为 0.2。随着全局训练轮数的增加, 系统对客户端更新质量的要求逐步提高, α 分位数随之递增至 0.5。信誉值的更新策略如式(10)所示, 其中 Δ_a 和 Δ_b 为奖惩幅度。

$$\begin{cases} a'_i = a_i + \Delta_a, \text{Score}_i \geq h \\ b'_i = b_i + \Delta_b, \text{Score}_i < h \end{cases} \quad (10)$$

类簇内信誉评估算法实现过程如算法 2 所示。

算法 2 类簇内信誉评估算法

输入 类簇的客户端列表 C_k , 类簇模型 θ_c

输出 更新后的客户端信誉度 R

1) for each $c_i \in C_k$:

2) $\alpha_{ic} \leftarrow \text{similarity}(\theta_i, \theta_c)$ // 计算相似度

3) $\beta_{ic} \leftarrow \text{LossConsistency}(\theta_i, \theta_c)$ // 计算损失一致性

4) $\text{Score}_i = \lambda \cdot \alpha_{ic} + (1 - \lambda) \cdot \beta_{ic}$ // 计算客户端更新质量

5) $h_k \leftarrow \text{Quantile}(\{\text{Score}_i\}, \alpha)$ // 计算类簇的更新质量动态阈值

6) for each $c_i \in C_k$:

7) if $\text{Score}_i \geq h_k$:

8) $a'_i = a_i + \Delta_a$

9) else:

10) $b'_i = b_i + \Delta_b$

11) $R_i = \frac{a_i + 1}{a_i + b_i + 2}$ //更新客户端信誉值

12) return 更新后的客户端信誉度 R

2.4 算法的复杂度分析

为评估 DCFL 算法的复杂度, 本文将 DCFL 算法与相关的聚类联邦学习算法 CFL^[1]、FeSEM^[7]进行了对比分析。

在时间复杂度方面, DCFL 算法的计算开销主要集中在相似度矩阵计算 $O(N^2)$ 和 HDBSCAN 密度聚类 $O(N \log N)$, 总时间复杂度为 $O(N^2)$ 。其中, N 表示参与联邦学习的客户端数量。相比之下, CFL 算法采用递归二分聚类策略, 时间复杂度为 $O(N^2)$ 。FeSEM 算法基于期望最大化框架, 其时间复杂度为 $O(T \times N \times K)$, 其中 T 表示 EM 算法的迭代轮数, K 表示预设的聚类数量。

在空间复杂度方面, DCFL 需要存储客户端间的相似度矩阵、聚类分配信息和信誉值向量, 总空间复杂度为 $O(N^2 + N)$ 。CFL 的空间复杂度为 $O(N \times M)$, 其中 M 表示模型参数维度, 主要用于存储客户端模型参数。FeSEM 算法由于需要维护多个聚类中心和 EM 算法的中间变量, 其空间复杂度达到 $O(N \times K \times M)$ 。

尽管 DCFL 在某些情况下的理论复杂度并不具备绝对优势, 但在实际应用中具有显著的效率优势。DCFL 基于密度聚类, 能够在单次聚类操作中完成聚类分配, 并自适应地确定最优聚类结构。相较于需要递归二分的 CFL 和反复迭代的 FeSEM, DCFL 减少了聚类次数和单次聚类的时间。更重要的是, DCFL 引入的双维度信誉评估机制, 使其在确保算法有效性的同时, 即使在异构场景和存在恶意客户端的复杂环境中, 仍能保持较高的实际执行效率。

3 实验与结果分析

3.1 实验设置

本文实验采用 EMNIST 数据集评估 DCFL 性能, 模型使用卷积神经网络 (CNN) 模型, 所有实验均基于 Pytorch 框架实现。

本文实验采用的 CNN 包含输入层、卷积层 (第1个卷积层输出6个通道, 第2个卷积层输出

16个通道)、池化层和全连接层。

客户端训练参数设置如下: 参与联邦学习的客户端总数为 50 个, 客户端基于本地数据集的训练迭代轮数 (epoch) 为 5, 本地训练的批量大小 (batch_size) 为 128, 学习率为 0.1, 动量为 0.9。

Non-IID 数据模拟: 本文通过两种方式构建客户端的 Non-IID 数据环境。其一, 基于 Dirichlet 分布进行随机采样, 生成一个类别标签分布矩阵, 其行向量表示类别 K 在各客户端上的概率分布。该分布由参数向量 α 控制, α 值越小, 表明客户端之间数据的异构性越显著。其二, 将全部客户端四等分, 并分别对其本地图像数据进行 0° 、 90° 、 180° 和 270° 的旋转, 从而生成 4 种具有清晰聚类结构的客户端。

恶意客户端模拟: 研究人员在不同数据分组中指定恶意客户端, 并采用两种攻击方式。在数据层面, 实施标签翻转攻击, 同时针对 EMNIST 数据集的特点, 构建一个视觉相似字符映射表 (如 “D” $\rightarrow$ “O” 和 “1” $\rightarrow$ “7”), 使攻击更加隐蔽, 不易被察觉。在模型层面, 实施模型替换攻击, 恶意客户端在完成本地训练后, 将在其模型参数中添加标准差 $\sigma = 0.1$ 的高斯噪声 $N(0, \sigma^2)$ 。该攻击方法不直接修改数据, 而是在参数空间进行扰动, 因此更加隐蔽。

3.2 实验结果分析

本文从 4 个方面评估 DCFL 的性能, 并将 DCFL 与基准算法 FedAvg、聚类算法 CFL^[1] 和 FeSEM^[7]、恶意客户端检测算法 eFL^[12] 和 FLDetector^[13] 进行对比。

3.2.1 聚类效果

本文采用调整兰德指数 (adjusted Rand index, ARI) 和纯度 (Purity) 两项指标评估聚类效果。ARI 用于衡量聚类结果与实际分类之间的一致性, 其值越大, 表明聚类结果与实际分类越吻合。图 2 展示了在无恶意客户端 ($m = 0$) 和存在恶意客户端 ($m = 10$) 两种情况下的聚类 ARI 值。Purity 则通过计算每个聚类中优势类别的比例来反映类簇的纯净程度, 结果如图 3 所示。Purity 值越大, 表示类簇纯净度越高。Purity 不

关注类簇的完整性，而 ARI 兼顾纯净度和完整性（对将同一类别分散至多个聚类的情況会施加惩罚）。

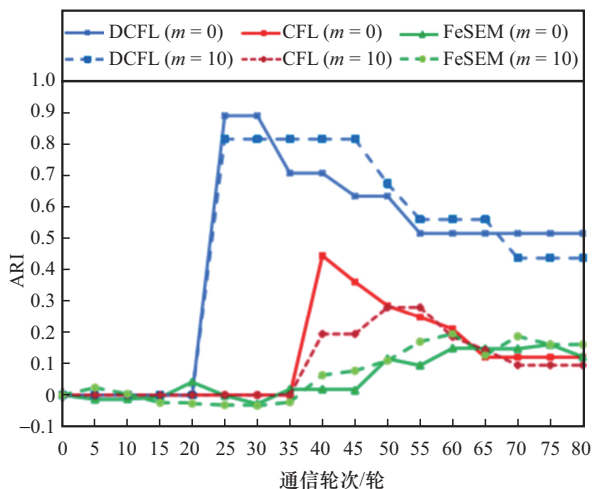


图2 聚类ARI值

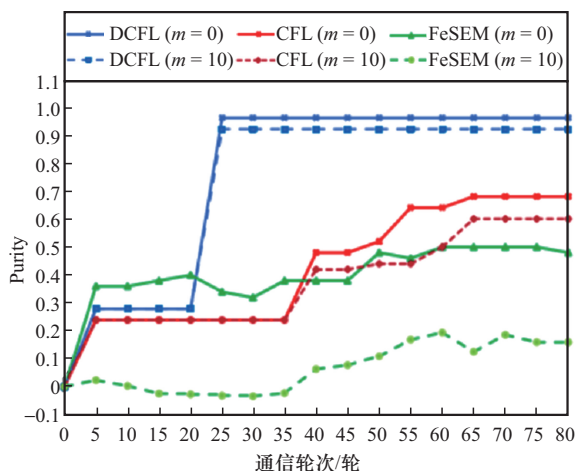


图3 聚类Purity值

DCFL约在第25轮完成第一次聚类，其ARI值达到0.9（约为CFL算法的两倍）。Purity值迅速升至0.95左右，这表明DCFL仅需极少次聚类即可获得与真实数据分布高度一致的结果。该优势主要源于DCFL采用的HDBSCAN算法能够识别不同密度的类簇。而CFL依赖迭代二分聚类过程，效率较低。随着训练的推进，DCFL的Purity值基本保持稳定，未出现明显下降；而ARI值在后期有所下降，最终稳定在0.5左右。这一现象与参数设置密切相关：实验中设定 $\text{min_cluster_size} = 5$ ，导致训练后期部分本属同一类别的客户端被进一步划分为更小的两个类

簇，从而引起ARI下降并趋于稳定。在存在恶意客户端的环境中，DCFL的ARI与Purity值与无恶意客户端时的结果相近。这是因为其所采用的层次化聚类机制善于处理复杂的客户端关系。

3.2.2 双维度更新质量评估有效性证明

DCFL在评估客户端更新质量时提出：在数据分布相似的条件，各本地客户端的损失值也应趋于一致。因此，损失一致性被纳入更新质量评估体系，作为一项重要指标。为验证该机制的有效性，本文设置了一组对比实验：在 $\alpha=0.6$ 且存在10个恶意客户端的环境中，比较了两种不同信誉值计算方法对模型测试准确率的影响，结果如图4所示。

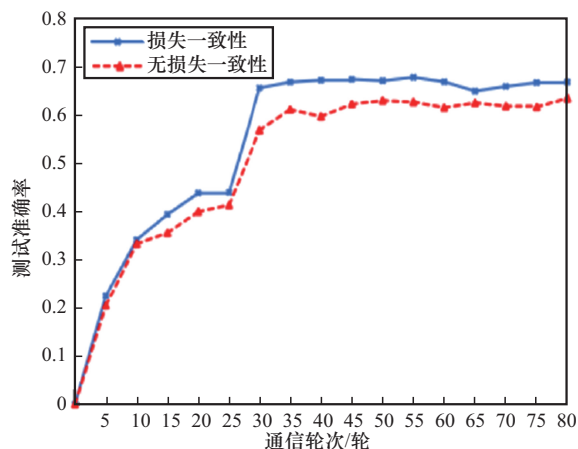


图4 损失一致性对模型测试准确率的影响

在训练初始阶段，两种方法的性能比较接近。自第25轮左右开始聚类后，两种方法的测试准确率都实现显著的跃升。其中，采用损失一致性的方法提升幅度更大，测试准确率迅速达到0.65左右。在后续训练过程中，采用损失一致性的方法展现出更优的稳定性，准确率曲线波动较小，始终保持5%~8%的优势。

3.2.3 信誉值更新有效性证明

在本文实验中共设置了10个客户端，其中客户端0与客户端1被设定为恶意客户端， $\alpha=0.6$ 。实验过程中形成了3个类簇，其组成如下。

1) 类簇1：客户端0、2、4、7，内含恶意客户端0。

2) 类簇2：客户端1、6、9，内含恶意客

户端1。

3) 类簇3: 客户端3、5、8, 全部为正常客户端。

图5与图6分别展示了包含恶意客户端的两个类簇中各客户端信誉值的变化情况。

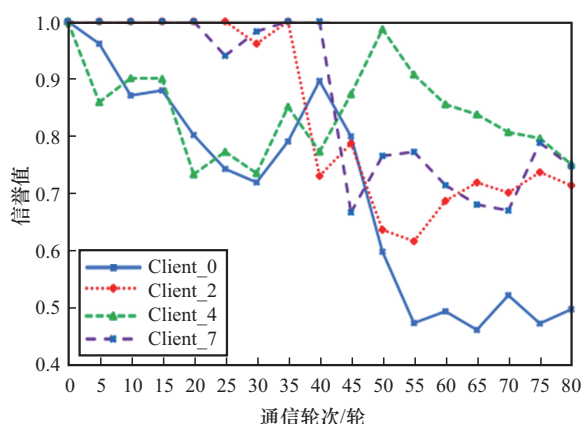


图5 客户端0、2、4、7的信誉值变化

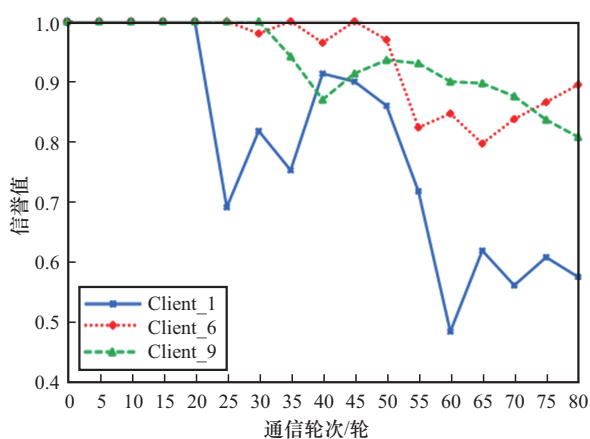


图6 客户端1、6、9的信誉值变化

前期, 恶意客户端0和正常客户端4与其他客户端的行为差异较大, 信誉值较低(约为0.7)。约在第30轮完成第一次聚类后, 恶意客户端0的信誉值出现短暂提升, 但随后迅速下降, 最终稳定在0.5左右, 表明系统成功识别出其恶意行为。恶意客户端1在聚类后信誉值急剧下降, 最终稳定在0.6左右。后期, 恶意客户端0和1均成为各自类簇中信誉值最低的客户端。正常客户端4在聚类后信誉值回升, 虽然后面又有波动, 但整体维持在类簇内较高水平。这一结果表明, DCFL能够随着训练轮次的增加, 持续积累行为证据并有效评估客户端的

信誉。

3.2.4 不同数据分布和安全环境下的性能评估

研究人员设置了5组对比实验: ① $\alpha=1$, $m=0$ (低异构无攻击场景); ② $\alpha=1$, $m=10$ (低异构有攻击场景); ③ $\alpha=0.6$, $m=0$ (高异构无攻击场景); ④ $\alpha=0.6$, $m=10$ (高异构有攻击场景); ⑤ $\alpha=0.6$, $m=10$ (高异构, 模型替换攻击场景)。

DCFL、CFL、FeSEM和FedAvg联邦学习算法在理想情况下(低异构无攻击场景)的测试准确率曲线如图7(a)所示。从图中可以看出, DCFL的表现最佳, 证明DCFL的聚类能力与信誉评估机制能够为联邦学习提供额外的性能增益。

在数据分布相对均匀但存在恶意客户端的(低异构有攻击)场景下, DCFL、CFL、FeSEM、eFL、FLDetector和FedAvg算法的测试准确率曲线如图7(b)所示。从图中可以看出, 大多数的算法在早期即达到约0.7的准确率。DCFL在第20轮左右触发了首次聚类, 恶意客户端被分到了更小的类簇中, 在经过进一步聚类和信誉机制更新后, DCFL展现出对恶意行为的持续抵抗能力。DCFL在第40轮后准确率恢复至较高值, 最终稳定在0.73左右, 与无恶意客户端场景下的性能基本持平。FLDetector和eFL表现出色, 略优于DCFL。而CFL和FedAvg的准确率在后期出现波动, 性能相对较弱。

在数据高度异构但不存在恶意客户端(高异构无攻击)的场景下, DCFL、CFL、FeSEM和FedAvg算法的测试准确率曲线如图7(c)所示, 从图中可以看出, 各算法的模型性能均大幅下降。DCFL在第30轮左右出现性能跃升, 通过建立合理的聚类结构, 最终准确率稳定在0.67左右, 优于其他算法。CFL算法次之, 在第45轮左右达到约0.6的准确率。而传统的FedAvg算法在该场景下的准确率一直较低, 仅维持在0.45左右。

在数据高度异构且存在恶意客户端(高异构有攻击)的场景下, DCFL、CFL、FeSEM、eFL、FLDetector和FedAvg算法的测试准确率曲线如图7(d)所示。其中, DCFL的表现最出色。DCFL的准确率在第25轮出现明显的跃升(从0.4迅速

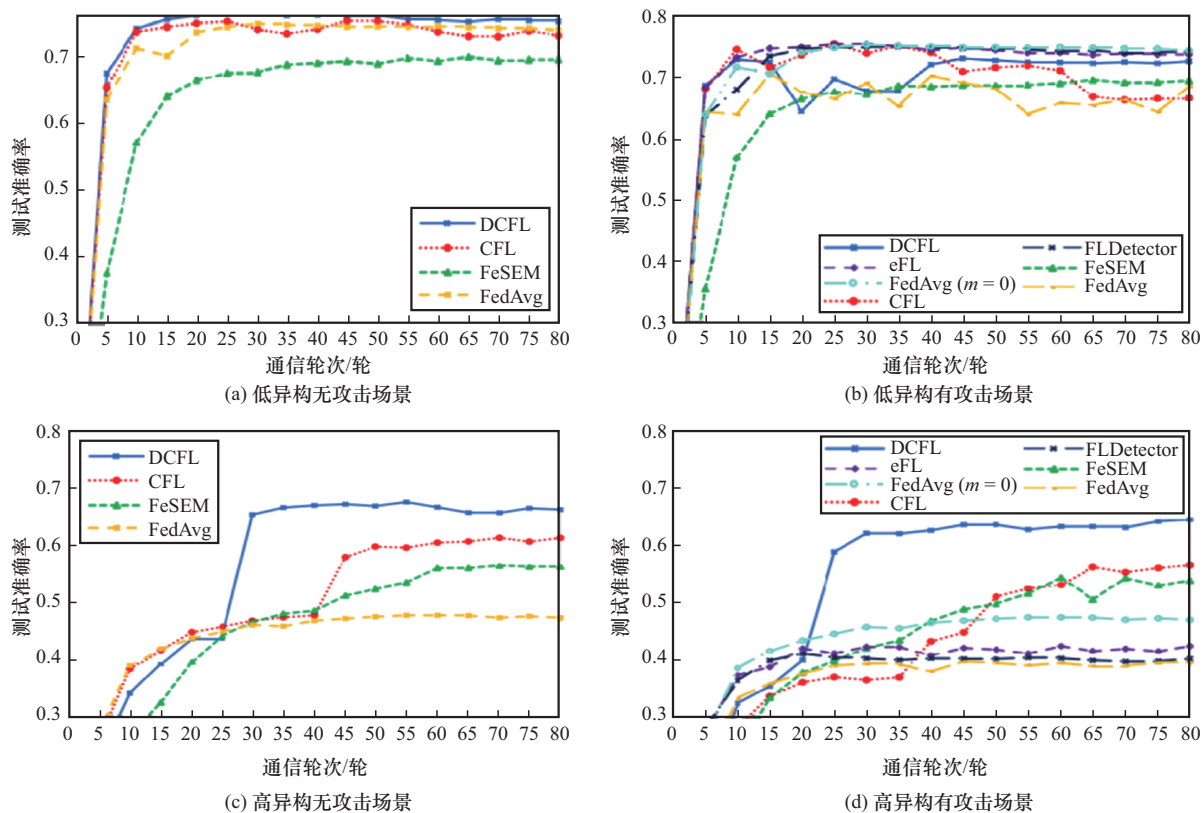


图7 各算法在不同场景下的测试准确率对比

提升至0.6以上)。此时聚类结构与信誉体系开始发挥协同作用,其准确率已接近无攻击场景下的水平。这主要得益于DCFL能够将数据分布相似的客户端聚合到同一类簇,有效隔离恶意客户端,从而抑制恶意行为的扩散。此外,在近似同分布的类簇环境中,损失一致性指标与模型相似度互为补充,能够及时捕获客户端的恶意行为。此时,DCFL与其他算法的性能差距达到最大,相比非聚类算法提升超过20个百分点,较CFL提高了约9个百分点,充分体现了其综合设计优势。

图8展示了各算法在模型替换攻击场景下的测试准确率对比情况。DCFL在此场景下依然展现出卓越的鲁棒性,约在第65轮触发聚类机制后,其测试准确率迅速提升至0.6左右并趋于稳定,相比CFL与FeSEM等聚类算法提升约10个百分点。这一结果进一步证明了DCFL所采用的双维度信誉评估机制能够有效识别参数空间中的异常扰动,而其动态聚类策略则成功将恶意客户端隔离在较小的类簇中,减少了对全局模型性能

的负面影响。

此外,DCFL显著提升了模型的收敛速度。其动态聚类机制能够将数据分布相似的客户端有效聚合,从而减少类簇内部的梯度方差。实验结果表明,在高异构无攻击场景下,DCFL仅需30轮训练即可达到收敛状态。同时,双维度信誉评估机制能够帮助模型快速识别恶意客户端,并通过降低恶意客户端的权重,减少其对训练过程稳定性的干扰。

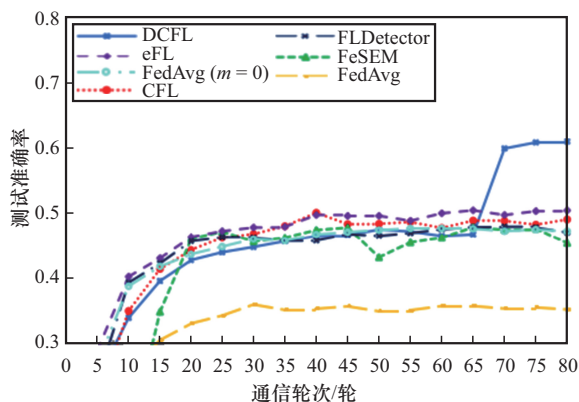


图8 各算法在模型替换攻击场景下的测试准确率对比

4 结束语

DCFL 是一种基于动态聚类与双维度信誉评估的鲁棒联邦学习框架。该框架无须预先指定类簇数量, 即可高效区分客户端的异构数据分布特征。在每个聚类内部, DCFL 综合利用客户端模型更新相似度与损失一致性评估客户端模型的更新质量, 降低恶意客户端的信誉值。基于信誉值的加权聚合机制在保护各类簇个性化特征的同时, 显著降低了恶意客户端对全局模型的负面影响。实验结果表明, DCFL 能够适应不同数据分布与安全场景, 有效缓解联邦学习中的数据异构性问题, 并抵御包括标签翻转和模型替换在内的典型恶意攻击。然而, 在实际应用中, 数据分布的动态变化和客户端的异步性等因素仍可能影响 DCFL 的性能。未来工作将进一步优化其对复杂环境的适应能力, 并探索更加灵活的聚类与信誉评估策略。

参考文献:

- [1] SATTLER F, MÜLLER K R, SAMEK W. Clustered federated learning: model-agnostic distributed multitask optimization under privacy constraints[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(8): 3710-3722.
- [2] 万伟, 胡胜山, 陆建荣, 等. 联邦学习在高度数据异构场景下的泛化鲁棒性增强[J]. 中国科学: 信息科学, 2024, 54(3): 566-581.
WAN W, HU S S, LU J R, et al. Enhancing generalization robustness of federated learning in highly heterogeneous environments[J]. Scientia Sinica (Informationis), 2024, 54(3): 566-581.
- [3] MA X S, ZHANG J, GUO S, et al. Layer-wised model aggregation for personalized federated learning[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 10082-10091.
- [4] ZHANG J Q, HUA Y, WANG H, et al. FedALA: adaptive local aggregation for personalized federated learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2023: 11237-11244.
- [5] 应作斌, 方一晨, 张怡文. 动态聚合权重的隐私保护联邦学习框架[J]. 网络与信息安全学报, 2022, 8(5): 56-65.
YING Z B, FANG Y C, ZHANG Y W. Privacy-preserving federated learning framework with dynamic weight aggregation[J]. Chinese Journal of Network and Information Security, 2022, 8(5): 56-65.
- [6] GHOSH A, CHUNG J, YIN D, et al. An efficient framework for clustered federated learning[J]. IEEE Transactions on Information Theory, 2022, 68(12): 8076-8091.
- [7] LONG G D, XIE M, SHEN T, et al. Multi-center federated learning: clients clustering for better personalization[J]. World Wide Web, 2023, 26(1): 481-500.
- [8] 朱素霞, 顾玢珂, 孙广路. 基于相似度加速的自适应聚类联邦学习[J]. 通信学报, 2024, 45(3): 197-207.
ZHU S X, GU B K, SUN G L. Adaptive clustering federated learning via similarity acceleration[J]. Journal on Communications, 2024, 45(3): 197-207.
- [9] LIU X Y, LI H W, XU G W, et al. Privacy-enhanced federated learning against poisoning adversaries[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 4574-4588.
- [10] 康海燕, 冀源蕊. 基于本地化差分隐私的联邦学习方法研究[J]. 通信学报, 2022, 43(10): 94-105.
KANG H Y, JI Y R. Research on federated learning approach based on local differential privacy[J]. Journal on Communications, 2022, 43(10): 94-105.
- [11] YANG Z Q, ZHANG Y G, ZHENG Y, et al. FedFed: feature distillation against data heterogeneity in federated learning[C]//Proceedings of Advances in Neural Information Processing Systems. Vancouver: NeurIPS, 2023: 60397-60428.
- [12] 刘艳, 王田, 彭绍亮, 等. 基于边缘的联邦学习模型清洗和设备聚类方法[J]. 计算机学报, 2021, 44(12): 2515-2528.
LIU Y, WANG T, PENG S L, et al. Edge-based model cleaning and device clustering in federated learning[J]. Chinese Journal of Computers, 2021, 44(12): 2515-2528.
- [13] ZHANG Z X, CAO X Y, JIA J Y, et al. FLDetector: defending federated learning against model poisoning attacks via detecting malicious clients[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2022: 2545-2555.
- [14] CHEN Z Y, LIANG J, YU Z X, et al. Resilient collaborative caching for multi-edge systems with robust federated deep learning[J]. IEEE Transactions on Networking, 2025, 33(2): 654-669.
- [15] GU H L, FAN L X, TANG X X, et al. FedCut: a spectral analysis framework for reliable detection of Byzantine colluders[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(9): 5905-5920.
- [16] 徐明迪, 李正潇, 王梓行, 等. Robust-PFedproto: 基于个性化层的鲁棒联邦原型学习方法[J]. 网络与信息安全学报, 2025, 11(3): 67-80.
XU M D, LI Z X, WANG Z H, et al. Robust-PFedproto: robust federated prototype learning based on personalized layers[J]. Chinese Journal of Network and Information Security, 2025, 11(3): 67-80.
- [17] TAHVILI S, HATVANI L, FELDERER M, et al. Automated functional dependency detection between test cases using Doc2Vec and clustering[C]//Proceedings of the 2019 IEEE International Conference on Artificial Intelligence Testing (AITest). Piscataway: IEEE Press, 2019: 19-26.

- [18] 雷诚, 张琳. 基于更新质量检测和恶意客户端识别的联邦学习模型[J]. 计算机科学, 2024, 51(11): 368-378.

LEI C, ZHANG L. Federated learning model based on update quality detection and malicious client identification[J]. Computer Science, 2024, 51(11): 368-378.

[作者简介]



张唯一 (2001-), 女, 安徽阜阳人, 南京邮电大学硕士生, 主要研究方向为联邦学习、隐私保护。



罗启瑞 (2004-), 男, 江苏徐州人, 南京邮电大学本科生, 主要研究方向为联邦学习、隐私保护。



张琳 (1980-), 女, 江苏丰县人, 博士, 南京邮电大学副教授、硕士生导师, 主要研究方向为网络安全、可信计算、联邦学习。