计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(8):237−243 [doi: 10.15888/j.cnki.csa.009949] [CSTR: 32024.14.csa.009949]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

利用不稳定客户端增强联邦学习全局模型训练①

李晓晖, 边太成, 杨　锦, 朱习军

(青岛科技大学 信息科学技术学院, 青岛 266061)
通信作者: 朱习军, E-mail: zhuxj990@163.com

摘　要: 在联邦学习中, 不稳定客户端可能通过数据污染或恶意行为干扰全局模型的训练过程. 传统的防御方法通常侧重于排除这些客户端, 但忽视了不稳定客户端生成的数据也可以为模型提供有价值的训练信号. 为此, 本文提出了一种增强适应性的联邦对抗训练方法 (Fed-ATEA), 利用不稳定客户端生成的对抗样本来增强全局模型的鲁棒性. 该框架允许在不排除不稳定客户端的情况下, 将其生成的对抗样本融入信任组客户端的训练过程, 进而增强模型的稳健性和鲁棒性. 通过动态调整训练策略, 最大化地利用不稳定客户端提供的有益信号, 并有效抑制其负面影响. 实验结果表明, 相对其他联邦学习方法, Fed-ATEA 在应对攻击和噪声干扰时展现出更强的稳健性和鲁棒性.

关键词: 联邦学习; 不稳定客户端; 鲁棒性; 对抗性训练

引用格式: 李晓晖,边太成,杨锦,朱习军.利用不稳定客户端增强联邦学习全局模型训练.计算机系统应用,2025,34(8):237–243. http://www.c-s-a.org.cn/1003-3254/9949.html

Utilizing Unstable Clients to Enhance Global Model Training in Federated Learning

LI Xiao-Hui, BIAN Tai-Cheng, YANG Jin, ZHU Xi-Jun

(School of Information Science and Technology, Qingdao University of Science & Technology, Qingdao 266061, China)

Abstract: In federated learning, unstable clients may disrupt the training process of the global model through data pollution or malicious behavior. Traditional defense methods typically focus on excluding these clients, but overlook the fact that the data generated by unstable clients can also provide valuable training signals for the model. To address this issue, a federated adversarial training with an enhanced adaptability method, Fed-ATEA, which utilizes adversarial examples generated by unstable clients to enhance the robustness of the global model is proposed. This framework allows for the incorporation of adversarial samples into the training process of trusted clients without excluding unstable ones, thereby improving the stability and robustness of the model. By dynamically adjusting training strategies, Fed-ATEA maximizes the utilization of beneficial signals from unstable clients while effectively mitigating their negative influences. Experimental results show that, compared to other federated learning methods, Fed-ATEA demonstrates stronger stability and robustness in handling attacks and noise interference.

Key words: federated learning (FL); unstable client; robustness; adversarial training

联邦学习 (federated learning, FL) 作为一种去中心化的分布式学习方法, 在保护客户端数据隐私方面具有显著优势. 然而, 联邦学习面临的一个关键挑战是客户端的不稳定性, 包括数据污染和恶意攻击等, 这些问题严重影响全局模型的训练效果. 现有的防御方法通常侧重于排除不稳定客户端, 但这种做法往往违背了联邦学习的初衷. 为应对这一问题, 本文提出了一种增强适应性的联邦对抗训练方法 Fed-ATEA. 与传统方法

① 收稿时间: 2024-12-25; 修改时间: 2025-01-24; 采用时间: 2025-03-18; csa 在线出版时间: 2025-06-24
CNKI 网络首发时间: 2025-06-25

不同, Fed-ATEA 通过将对抗样本融入信任组客户端的训练中, 最大化地发挥不稳定客户端的对抗样本信息, 同时有效抑制其负面影响.

本文以此为研究切入点, 主要贡献总结如下.

(1) 提出 Fed-ATEA 框架, 允许不稳定客户端的对抗样本加入信任组的训练并提供支持.

(2) 不直接丢弃不稳定客户端持有的数据, 而是通过动态调整训练策略, 最大化发挥对抗样本的作用.

(3) 通过实验结果验证了 Fed-ATEA 在面对攻击和噪声干扰时的有效性, 且始终优于其他 FL 方法.

1 不稳定客户端检测与利用策略概述

在联邦学习中, 恶意客户端[1–3](不稳定客户端) 往往通过数据污染或攻击行为影响全局模型训练, 导致模型的稳健性和鲁棒性下降. 传统的防御策略通常侧重于排除这些不稳定客户端[4–8], 但未充分考虑到不稳定客户端中携带的部分有训练价值的数据. 近年来, 一些研究开始探索如何在不排除不稳定客户端的情况下, 利用其生成的数据对全局模型进行优化. 本节将概述现有的主要检测与利用策略, 并讨论其优缺点.

1.1 基于数据的检测方法

基于数据的检测方法[9]通过监控客户端上传的数据质量, 评估客户端本地模型是否偏离全局模型的预测结果. 常见方法包括统计异常值检测和标签预测差异分析. 然而, 当客户端的数据分布差异较大时, 这些方法容易误判, 导致正常数据被错误剔除. 因此, 这类方法无法充分利用不稳定客户端提供的部分有效数据, 尤其是在不稳定客户端通过逐步变化进行攻击时, 其潜在的有效数据可能会被忽视, 从而浪费宝贵的训练样本.

1.2 基于模型行为的检测方法

基于模型行为的检测方法关注客户端在本地训练对全局模型的影响. 如果某个客户端的本地模型更新过于偏离全局模型的趋势, 则可能会被判定为不稳定客户端. 这类方法通常利用梯度异常检测或聚类技术识别表现异常的客户端[10,11]. 然而, 对于一些通过逐渐变化的方式进行攻击的不稳定客户端, 早期的模型行为检测可能无法及时识别出其异常. 为了解决这一问题, 我们在本文中计划通过引入依据熵值变化的自适应检测机制, 使检测方法能够根据客户端行为的变化动态调整, 及时识别不稳定客户端的潜在威胁.

1.3 基于信誉系统的检测方法

基于信誉系统的检测方法[12,13]通过动态评估客户端的历史表现来调整其对全局模型训练的影响. 信誉较低的客户端会进行加权[14]甚至排除, 以降低其对全局模型训练的干扰. 然而, 这种方法存在局限性: 如果不稳定客户端在训练初期表现良好, 可能会被误判为可信, 从而影响系统判断的准确性. 此外, 信誉系统的设计重点通常在于排除不稳定客户端, 而未考虑如何有效利用这些客户端生成的数据. 因此, 我们提出通过动态调整信誉评估机制, 解决现有方法中未能有效利用不稳定客户端数据样本的问题, 从而将这些数据逐步融入全局模型训练, 探索其在特定情况下的积极作用.

2 对抗性增强的不稳定客户端利用框架

2.1 Fed-ATEA 算法流程概述

本节将概述 Fed-ATEA 算法的整体执行流程. Fed-ATEA 联邦学习框架旨在通过动态检测和有效利用不稳定客户端的数据, 增强信任组客户端的训练效果, 从而提升全局模型的鲁棒性. 与传统方法通过直接排除不稳定客户端的数据不同, Fed-ATEA 通过计算熵值识别不稳定客户端并生成对抗样本, 使不稳定客户端在训练过程中发挥积极作用, 帮助信任组客户端在面对噪声和攻击性数据时提高识别能力, 最终提高全局模型的性能. Fed-ATEA 算法的执行流程如算法 1.

算法 1. Fed-ATEA 对抗性训练增强算法

1) 各客户端初始化其本地模型参数θ_k^0, 并收集本地数据集D_k;

2) 在每轮训练结束后, 根据式 (1) 计算客户端C_k在未标注数据上的输出熵值H_k以衡量模型的不确定性, 并将客户端分为信任组与不稳定组;

3) 不稳定组客户端M_k基于其本地数据集D_k生成对抗样本, 并通过局部训练优化模型, 根据式 (2) 生成扰动δ, 使得生成的对抗样本能够有效辅助信任组客户端训练;

4) 信任组客户端T_k在本地数据和对抗样本的结合下进行训练并根据式 (5) 更新本地模型参数θ_k;

5) 信任组客户端的局部模型通过式 (3) 进行更新聚合, 更新全局模型θ_g.

Fed-ATEA 对抗性训练增强算法通过熵值变化检测不稳定客户端并生成对抗样本, 确保信任组能够有效利用这些样本进行训练, 而不是将不稳定客户端排除在外. 这样不仅提升了全局模型的鲁棒性, 还避免了不稳定客户端对模型性能的负面影响, 进一步增强了

训练过程的多样性和泛化能力. 信任组客户端在面对具有挑战性的对抗性训练数据时, 能够提高对不确定性和攻击性数据的识别能力. 此外, 通过强化学习和激励机制, 信任组能够有效地从对抗样本中学习, 提升全局模型在噪声数据下的适应能力和鲁棒性.

2.2 不稳定客户端检测与对抗样本生成

为了有效利用不稳定客户端的数据并避免传统方法中直接排除这些客户端的做法[15], 我们提出了一种基于对抗性训练的新方法. 该方法旨在增强信任组客户端的抗干扰能力, 而非简单地剔除不稳定客户端. 通常, 不稳定客户端的数据包含噪声或攻击性特征[16], 这些数据的分布与正常客户端有所不同. 尽管这些数据存在潜在的负面影响, 但我们认为它们也可能提供有价值的训练信号. 为了有效利用这些数据, 我们将被判别为不稳定客户端的数据用于生成对抗样本, 并通过局部训练来优化其贡献. 首先, 我们通过计算每个客户端在未标注数据上的输出熵值来识别不稳定客户端. 熵值衡量了模型输出的不确定性, 较大的熵值通常表示客户端的模型输出存在较大的不稳定性, 这可能是由数据噪声或恶意攻击引起的[17]. 假设客户端 k 对第 i 个输入样本 x_i 的预测概率为 $P(y|x_i)$, 其中 y 为标签. 客户端 k 的熵值 H_k 可以通过式 (1) 计算:

$$H_k = -\frac{1}{N}P(y|x_i)\log P(y|x_i) \tag{1}$$

其中, N 为客户端 k 上传的样本数, $P(y|x_i)$ 是客户端 k 对样本 x_i 预测的类别概率分布, $\log P(y|x_i)$ 是对类别概率的对数计算. 较大的熵值 H_k 表示客户端的模型输出不确定性较高, 表示客户端数据存在异常或受到攻击. 根据熵值的大小, 中央服务器可以评估客户端的稳定性, 并据此判断其是否为不稳定客户端.

对于被判定为不稳定客户端的数据, 我们将这些数据用于生成对抗样本. 假设不稳定客户端 k 的数据集为 $\{x_i, y_i\}$, 其中 x_i 为输入样本, y_i 为对应标签, 且样本数为 N. 我们通过对抗训练, 优化一个扰动 δ, 使得对抗样本 $\hat{x}_i = x_i + \delta$ 能有效帮助信任组客户端提升鲁棒性. 生成对抗样本的目标是最小化以下损失函数:

$$\delta = \text{argmin}_{\delta} L(\theta_k + \delta) \tag{2}$$

其中, L 为损失函数, θ_k 为不稳定客户端的局部参数模型, δ 为扰动. 通过逐步调整扰动 δ, 生成的对抗样本能够有效地帮助信任组客户端提升其对噪声数据的鲁棒性[18,19].

2.3 客户端间信息流动优化与全局模型聚合

为了避免不稳定客户端的干扰, 并最大化其对训练过程的潜在正面影响, 我们设计了一种信息流动优化机制. 该机制确保每轮训练后的全局模型能够根据信任组客户端的贡献进行调整, 从而提高模型的整体鲁棒性. 在我们的框架中, 不稳定客户端并不会直接参与全局模型的聚合. 相反, 不稳定客户端通过生成对抗样本的方式, 间接地帮助信任组客户端进行训练, 以提升信任组的鲁棒性. 这些由不稳定客户端生成的对抗样本被加入信任组客户端的训练过程中, 而不会直接影响全局模型的聚合与更新. 因此, 不稳定客户端的数据生成过程有助于增强信任组客户端对干扰性数据的识别与应对能力. 在全局模型更新过程中, 信任组客户端的贡献被加权处理, 以确保全局模型能够充分反映信任组客户端的训练成果. 假设全局模型参数为 θ, 每个信任组客户端 k 提供其本地更新 θ_k, 则全局模型的更新过程通过信任组客户端的聚合来实现. 更新过程如下:

$$\theta_g^{t+1} = \frac{1}{C_t}\sum_{k\in t} a_k \times \theta_k^{t+1} \tag{3}$$

其中, N_t 为全局模型的更新参数, C_t 为信任组客户端总数, θ_k^{t+1} 为信任组客户端 k 在当前轮训练中的本地更新, a_k 是客户端 k 的加权因子, 根据客户端的输出熵值来决定, 熵值更低的客户端将赋予更高权重.

2.4 奖励机制优化信任组训练

在我们的框架中, 不稳定客户端生成的对抗样本并非用于破坏信任组的训练, 而是作为一种数据增强策略, 旨在提升信任组在面对攻击性数据时的鲁棒性. 具体而言, 这些对抗样本被加入信任组客户端的训练数据集中, 使信任组能够更好地适应不同类型的数据扰动, 从而增强模型的稳健性与鲁棒性. 通过使用这些对抗样本, 信任组客户端在面对噪声和攻击性数据时, 能够提高对数据的识别能力, 最终提升全局模型的鲁棒性和精度.

信任组客户端并不会直接将不稳定客户端的数据加入训练中, 而是利用这些对抗样本提升其对噪声数据的识别能力. 每一轮训练中, 信任组客户端会根据对抗样本的质量调整其学习策略, 增强其对噪声数据的适应能力. 为了实现这一目标, 我们设计了以下学习过

程. 在每一轮训练中, 信任组客户端通过强化学习调整其参数, 使得损失函数最小化, 并强化其对噪声数据的识别能力. 具体的更新规则如下:

$$\theta_k^{t+1} = \theta_k^t - \eta \nabla_{\theta_k} L_k(\theta_k^t, A_k) \tag{4}$$

其中, θ_k^t是信任组客户端 k 在第 t 轮的模型参数, η是为学习率, A_k是不稳定客户端生成的对抗样本, $L_k(\theta_k^t, A_k)$是信任组客户端 k 的损失函数, 表示模型在对抗样本上的表现. 为了进一步激励信任组客户端更好地适应不稳定客户端生成的对抗样本并优化对噪声数据的识别能力, 我们引入了强化学习中的奖励机制. 奖励项帮助信任组客户端调整训练策略, 提升其处理对抗样本的能力, 我们在信任组客户端的参数更新规则加入了奖励项$R_k(A_k)$, 使得信任组客户端在训练过程中不仅要关注损失函数的最小化, 还要注重提升其在面对对抗样本时的表现. 引入奖励项后的更新规则如下:

$$\theta_k^{t+1} = \theta_k^t - \eta \nabla_{\theta_k} (L_k(\theta_k^t, A_k) + \lambda \times R_k(A_k)) \tag{5}$$

其中, $R_k(A_k)$为奖励项, 衡量信任组客户端在处理对抗样本时的表现, λ为奖励项的权重.

3 实验分析与性能评价

3.1 数据集及实验环境

本实验使用了两个广泛应用于图像分类任务的标准数据集: CIFAR-10 和 CIFAR-100, 这两个数据集用以有效验证算法在面对不同复杂性数据时的鲁棒性和适应性. 实验在 NVIDIA RTX 3090 GPU 上进行, 训练轮次设定为 100 轮, 优化器选择 Adam, 学习率设置为 0.001, 批次大小 (batch size) 为 64, 所有实验均在此配置环境下执行, 确保实验结果的一致性和可重复性.

CIFAR-10 数据集包含 10 个类别, 每个类别有 6000 张 32×32 像素的彩色图像, 总共有 60000 张图像. 类别包括飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船和卡车. 该数据集的图像内容相对简单, 且类别间差异明显, 广泛用于图像分类的基准测试. CIFAR-10 可用以评估算法在较简单场景下的表现.

CIFAR-100 数据集是 CIFAR-10 的扩展, 包含 100 个类别, 每个类别 600 张 32×32 像素的彩色图像, 总共有 60000 张图像. 与 CIFAR-10 相比, CIFAR-100 的类别更加细致, 类别间的相似度更高, 从而增加了分类任务的难度. CIFAR-100 特别适用于验证算法在面对更高维度和更复杂数据时的鲁棒性与泛化能力, 为算法提供了一个更具挑战性的测试环境. 尤其在处理不同程度的不稳定客户端数据时的表现.

3.2 基于熵值的客户端不稳定性检测方法有效性验证

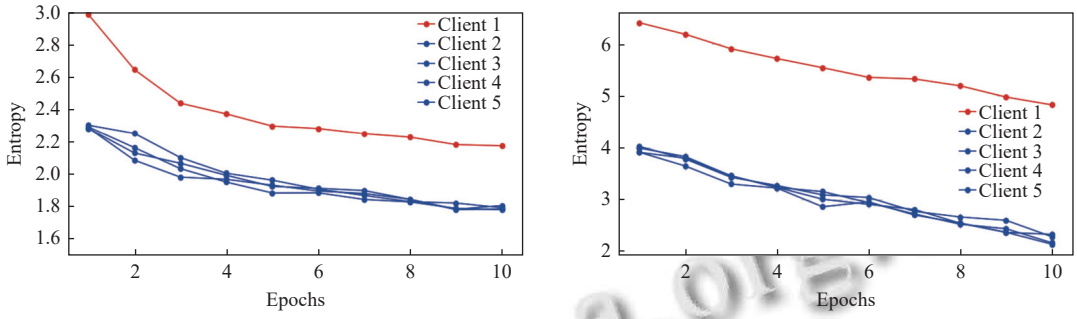
为了验证基于熵值的不稳定客户端检测方法的有效性, 我们设计了模拟不稳定客户端的实验. 在此实验中, 我们引入了人工标签噪声和数据扰动, 旨在研究其对训练过程中的模型稳定性与输出不确定性的影响. 数据集采用了广泛用于图像分类任务的 CIFAR-10 数据集. 每个客户端的数据来自 CIFAR-10 的子集, 客户端的数据分布是随机且独立的, 模拟了不同客户端在实际环境中可能遇到的异质数据情况.

在实验中, 数据集的划分是按比例随机分配的, 每个客户端的训练数据和测试数据都来自 CIFAR-10, 训练数据包含了标注信息, 而测试集则用于评估模型的泛化能力. 为了模拟不稳定客户端的行为, 我们人为地向部分训练数据中引入了标签噪声和数据扰动, 标签噪声的比例设定为 30%, 以模仿真实环境中客户端可能遭遇的噪声干扰情形. 通过这种方式, 我们增强了数据的复杂性, 从而增加了客户端模型输出的不确定性.

图 1(a) 展示了训练过程中各客户端熵值的变化情况. 图 1 中的红色曲线表示客户端 1 的熵值变化. 客户端 1 在训练初期表现出较高的熵值, 表明其模型输出的不确定性较大, 可能与标签噪声或数据扰动有关. 随着训练的进行, 熵值逐渐降低, 但始终高于其他客户端, 表明该客户端可能是不稳定客户端, 并被归类为不稳定组. 相比之下, 其他客户端 (蓝色曲线) 在训练初期的熵值较低, 波动较小, 表现出较低的不确定性, 因此被归类为信任组. 即使是在训练初期, 模型尚未收敛或数据复杂性较高时, 熵值可能自然较大. 但随着训练的进行, 模型的稳定性逐渐增强, 熵值明显下降, 进一步证实了熵值波动并非仅因训练阶段造成. 此外, 在同一数据集 (CIFAR-10) 上对不同客户端的数据分布进行对比实验, 尽管客户端的数据分布有所不同, 但干净客户端的熵值波动相对稳定, 不会因数据分布差异产生显著变化. 而引入标签噪声或数据扰动的客户端则表现出明显的熵值波动, 证明熵值变化能够有效地区分干净客户端和不稳定客户端, 并且在不同数据分布下保持较好的鲁棒性和准确性.

此外, 为了验证不同数据分布对熵值的影响, 我们分别使用了简单数据和复杂数据的子集, 并对比了

不同分布下的熵值. 根据我们在 CIFAR-100 数据集上使用相同实验设置的结果表明, 如图 1(b) 所示, 虽然不同客户端的熵值存在一定差异, 但熵值较大的客户端通常与标签噪声和数据扰动的引入密切相关, 而不仅是数据分布的复杂性所致. 基于这些实验结果, 我们能够有效地区分信任组和不稳定组, 进一步验证了基于熵值的客户端不稳定性检测方法在实际应用中的有效性.



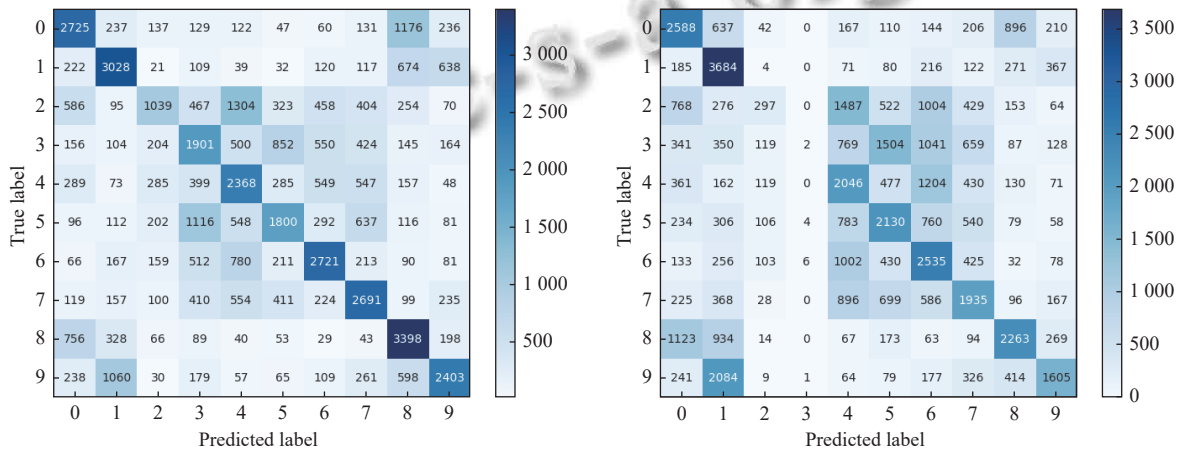
(a) 熵值变化 (CIFAR-10) (b) 熵值变化 (CIFAR-100)

图 1 基于熵值对不稳定客户端进行检测的结果

3.3 对抗样本生成与信任组训练实验

为了验证由不稳定客户端生成的对抗样本对信任组客户端训练效果的提升作用, 我们设计了两种不同的训练方式进行对比实验: 第 1 种方式: 仅使用信任组客户端的本地数据进行训练, 作为基准, 评估在没有对抗样本情况下模型的表现. 第 2 种方式: 在信任组客户端的本地数据基础上, 加入不稳定客户端生成的对抗样本, 用以评估信任组对数据扰动的鲁棒性提升. 通过这两种不同的训练策略对比, 我们可以评估对抗性训练是否能有效提升信任组客户端的训练效果, 特别是在面对数据扰动或噪声数据时的应对能力.

图 2 展示了信任组客户端在两种训练策略下的混淆矩阵对比: 一种是仅使用信任组本地数据进行训练, 另一种是在信任组本地数据的基础上加入不稳定客户端生成并处理后的对抗样本进行训练. 从混淆矩阵中可以看出, 在仅依赖本地数据进行训练时, 模型的类别区分存在较大的误差, 错误分类较为显著. 而在加入对抗样本的训练中, 混淆矩阵显示模型在类别预测的准确性上有了明显提高, 误分类样本显著减少, 表明模型的判别能力更加稳定. 通过对比两种训练策略的混淆矩阵结果, 我们可以得出结论: 引入对抗样本可以显著提升信任组客户端的分类精度和鲁棒性.



(a) 加入不稳定客户 (b) 本地数据

图 2 两种训练策略下信任组混淆矩阵对比

为了进一步验证对抗性训练对全局模型精度的影响, 我们进行了全局模型精度对比实验. 如图 3 所示, 设计了两种训练方案进行对比: 第 1 种方案仅使用信任组客户端的本地数据进行训练, 以评估不引入对抗

样本时全局模型的精度; 第 2 种方案则在信任组本地数据的基础上, 加入由不稳定客户端生成并处理后的对抗样本进行训练, 探讨不稳定客户端生成的对抗样本对全局模型精度的贡献. 通过对比这两种训练方式下的全局模型精度, 可以发现, 经过对抗性训练后的全局模型表现出了显著的提升, 进一步证明了对抗样本在增强全局模型鲁棒性方面的积极作用.

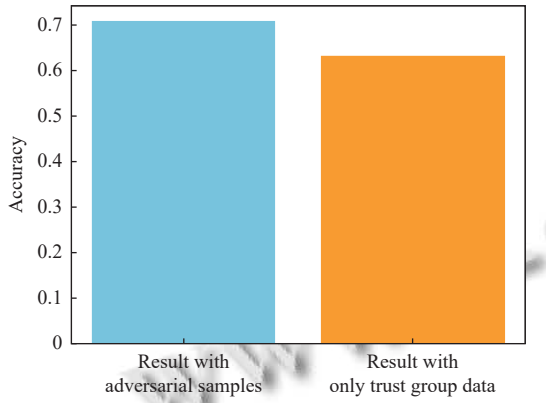


图 3 信任组客户端训练精度对比 (加入对抗样本 vs. 本地数据)

3.4 不稳定客户端比例对训练效果的影响

为了全面评估 Fed-ATEA 算法在不同不稳定客户端比例下的表现, 我们设置了不稳定客户端占比为 10%、20% 和 30% 的不同实验场景. 每个场景包含 10 个客户端. 与此同时, 还与其他几种主流联邦学习方法 FedAvg[20]、FedProx[21]、SCAFFOLD[22]和 FedDyn[23] 进行对比分析, 旨在深入探讨不同不稳定客户端比例对全局模型精度和算法鲁棒性的影响, 进而验证 Fed-ATEA 算法在处理不稳定客户端时的优势.

表 1 和表 2 显示了 Fed-ATEA 算法在不同不稳定客户端比例下 (10%、20%、30%) 的全局模型精度对比. 表 1 所展示的结果清晰地表明, 与传统方法相比, 在面对不稳定客户端的环境下 Fed-ATEA 表现出更强的稳健性和鲁棒性. 即使在较高比例的不稳定客户端场景下保持较高的训练精度.

表 1 随着不稳定客户端数量增加各 FL 方法的性能比较 (CIFAR-10) (%)

客户端比例	Fed-ATEA	FedAvg	FedProx	SCAFFOLD	FedDyn
10%	**82.97**	75.01	76.52	74.91	75.41
20%	**75.12**	68.96	72.69	68.02	69.03
30%	**70.83**	60.65	67.88	63.97	66.00

表 2 的结果进一步验证了 Fed-ATEA 在应对更具挑战性任务, 尤其是在高噪声环境下的抗干扰能力. 实验结果表明, Fed-ATEA 能够在复杂任务和高噪声干扰条件下, 依然保持优异的稳健性和鲁棒性.

表 2 随着不稳定客户端数量增加各 FL 方法的性能比较 (CIFAR-100) (%)

客户端比例	Fed-ATEA	FedAvg	FedProx	SCAFFOLD	FedDyn
10%	**48.37**	40.01	41.94	39.72	41.54
20%	**43.01**	35.28	35.57	33.89	36.00
30%	**39.64**	29.81	29.41	26.99	30.04

4 结论与展望

本研究提出了 Fed-ATEA 对抗性训练增强算法, 通过熵值动态检测和利用不稳定客户端生成的对抗样本, 提升了信任组客户端在干扰数据下的训练效果, 从而增强了全局模型的精度与鲁棒性. 在该框架中, 不稳定客户端生成对抗扰动来弥补损失, 并将这些对抗样本提供给信任组客户端训练, 有效减小了干扰数据的负面影响. 信任组客户端利用对抗样本提高了对干扰数据的识别能力, 确保了全局模型在存在不稳定客户端的影响下依然保持较好的稳健性和鲁棒性. 实验结果表明, Fed-ATEA 能有效提升信任组客户端的训练效果, 避免了不稳定客户端对全局模型的显著负面影响, 验证了算法在复杂联邦学习环境中的有效性.

未来的研究可集中在优化对抗样本生成机制, 提升不稳定数据的利用效率, 并探索更精准的不稳定客户端检测与数据利用方法. 同时, 考虑到大规模分布式环境中的挑战, 如何高效实现针对不稳定客户端的算法将成为一个重要研究方向.

参考文献

1 Kairouz P, McMahan HB, Avent B, *et al*. Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 2021, 14(1-2): 1–210.

2 Zhang ZX, Cao XY, Jia JY, *et al*. FLDetector: Defending federated learning against model poisoning attacks via detecting malicious clients. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington: ACM, 2022. 2545–2555.

3 Le JQ, Zhang D, Lei XY, *et al*. Privacy-preserving federated learning with malicious clients and honest-but-curious servers. IEEE Transactions on Information Forensics and Security, 2023, 18: 4329–4344. [doi: 10.1109/TIFS.2023.3295949]

4 Li SY, Cheng Y, Wang W, *et al*. Learning to detect

malicious clients for robust federated learning. arXiv: 2002.00211, 2020.
5 Shibahara T, Takata Y, Akiyama M, *et al*. Detecting malicious websites by integrating malicious, benign, and compromised redirection subgraph similarities. Proceedings of the 41st IEEE Annual Computer Software and Applications Conference (COMPSAC). Turin: IEEE, 2017. 655–664.
6 Gu ZP, Yang YX. Detecting malicious model updates from federated learning on conditional variational autoencoder. Proceedings of the 2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS). Portland: IEEE, 2021. 671–680.
7 Gupta A, Luo T, Ngo MV, *et al*. Long-short history of gradients is all you need: Detecting malicious and unreliable clients in federated learning. Proceedings of the 27th European Symposium on Research in Computer Security. Copenhagen: Springer, 2022. 445–465.
8 王珊, 荆桃, 肖淦文, 等. 联邦学习下高效的隐私保护安全聚合方案. 计算机系统应用, 2023, 32(11): 175–181. [doi: 10.15888/j.cnki.csa.009302]
9 连宇瀚, 廖声扬, 张坤三, 等. 基于联邦学习的异常日志检测. 计算机系统应用, 2024, 33(12): 78–88. [doi: 10.15888/j.cnki.csa.009714]
10 Geiping J, Bauermeister H, Dröge H, *et al*. Inverting gradients-how easy is it to break privacy in federated learning? Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 16937–16947.
11 Yao X, Huang TC, Zhang RX, *et al*. Federated learning with unbiased gradient aggregation and controllable meta updating. arXiv:1910.08234, 2019.
12 Kang JW, Xiong ZH, Niyato D, *et al*. Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. IEEE Internet of Things Journal, 2019, 6(6): 10700–10714. [doi: 10.1109/JIOT.2019.2940820]
13 ur Rehman MH, Salah K, Damiani E, *et al*. Towards blockchain-based reputation-aware federated learning. Proceedings of the 2020 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). Toronto: IEEE, 2020. 183–188.
14 许亚倩, 崔文泉, 程浩洋. 基于联邦学习和重要性加权的疾病得分预测. 计算机系统应用, 2022, 31(12): 375–382. [doi: 10.15888/j.cnki.csa.008871]
15 Nguyen DC, Ding M, Pathirana PN, *et al*. Federated learning for internet of things: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2021, 23(3): 1622–1658.
16 Li L, Fan YX, Tse M, *et al*. A review of applications in federated learning. Computers & Industrial Engineering, 2020, 149: 106854.
17 Karim N, Rizve MN, Rahnavard N, *et al*. UNICON: Combating label noise through uniform selection and contrastive learning. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 9666–9676.
18 Wu NN, Yu L, Jiang XF, *et al*. FedNoRo: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity. Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao: ACM, 2023. 4424–4432.
19 Song H, Kim M, Park D, *et al*. Learning from noisy labels with deep neural networks: A survey. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(11): 8135–8153. [doi: 10.1109/TNNLS.2022.3152527]
20 McMahan B, Moore E, Ramage D, *et al*. Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. PMLR, 2017. 1273–1282.
21 Li T, Sahu AK, Zaheer M, *et al*. Federated optimization in heterogeneous networks. Proceedings of Machine Learning and Systems, 2020, 2: 429–450.
22 Karimireddy SP, Kale S, Mohri M, *et al*. SCAFFOLD: Stochastic controlled averaging for federated learning. Proceedings of the 37th International Conference on Machine Learning. JMLR.org, 2020. 5132–5143.
23 Jin C, Chen XD, Gu Y, *et al*. FedDyn: A dynamic and efficient federated distillation approach on Recommender System. Proceedings of the 28th IEEE International Conference on Parallel and Distributed Systems (ICPADS). Nanjing: IEEE, 2023. 786–793.

(校对责编: 张重毅)