联邦学习中标签翻转攻击污染度感知的异常客户端模型参数重用算法

罗佳媛 谭振华 宁婧宇 贾志亮
（东北大学软件学院 沈阳 110169）

摘 要 标签翻转攻击(Label Flipping Attack)是联邦学习中的常见数据投毒攻击方法，通过置乱翻转客户端样本中的源类标签使得聚合后的服务器全局模型性能下降。现有防御技术主要通过不同客户端模型参数向量的相似度聚类来识别并舍弃异常客户端模型。此类方法在舍弃受污染参数的同时也舍弃了异常客户端模型中未被污染的有效参数。聚焦于该问题，本文提出了标签翻转攻击污染度感知的异常客户端模型参数重用算法FedCAR。首先设计了污染度指标，通过分别聚合K-means算法输出的两个簇中的客户端模型，得到异常全局模型和基础全局模型，并通过比较二者在输出层神经元参数上的差异来衡量异常模型中每类样本的污染度；进而提出了加权知识蒸馏损失函数，以参数裁剪后的异常客户端模型为教师，以基础全局模型为学生，用污染度调节学生对教师中各类知识的学习率，从而将异常客户端模型中的有效参数知识迁移至最终全局模型。实验结果表明，FedCAR在MNIST和CIFAR10数据集上的全局模型测试准确率较九种先进方法平均提升了1.81%，表明该方法能够有效挖掘并重用异常客户端模型中的效信息，从而增强联邦学习系统的泛化能力。

关键词 联邦学习；数据投毒攻击；标签翻转攻击；异常检测；知识蒸馏
中图法分类号 TP309 **DOI号** 10.11897/SP.J.1016.2026.00243

Contamination-Aware Reutilization Algorithm for Compromised Client Model Parameters against Label Flipping Attacks in Federated Learning

LUO Jia-Yuan TAN Zhen-Hua NING Jing-Yu JIA Zhi-Liang
(Software College, Northeastern University, Shenyang 110169)

Abstract Label flipping attacks represent one of the most prevalent forms of data poisoning in federated learning (FL). In such attacks, compromised clients deliberately manipulate their local training data by flipping the labels of source-class samples to a predefined target class. Such corruption misguides the optimization direction of the global model on the central server, leading to a significant decline in the test accuracy of source-class samples, while the model's performance on other classes remains largely unaffected. Existing defense methods typically rely on clustering the similarity matrix of model parameters from different clients to identify and discard compromised models during global aggregation. Although these approaches can effectively filter out contaminated parameters, they also inadvertently discard uncorrupted yet valuable parameters contained within the compromised local models, leading to the loss of useful information and consequently suboptimal global performance. To address this limitation, the paper proposes a contamination-aware reutilization algorithm for compromised client model parameters against

收稿日期：2025-07-02；在线发布日期：2025-11-02。本课题得到国家重点研发计划项目(No. 2023YFC3306201)资助。**罗佳媛**，博士研究生，主要研究领域为联邦学习与隐私保护。E-mail：jiayuanluo@stumail.neu.edu.cn。**谭振华**(通信作者)，博士，教授，中国计算机学会(CCF)高级会员，主要研究领域为网络行为分析、内容理解、隐私计算等。E-mail：tanzh@mail.neu.edu.cn。**宁婧宇**，博士，主要研究领域为安全多方计算与隐私保护。**贾志亮**，硕士，主要研究领域为联邦学习与隐私保护。FedCAR的代码已开源于：https://github.com/Vivienneljy/FedCAR。

label flipping attacks, called FedCAR. Unlike conventional defense methods that simply discard abnormal models, FedCAR aims to detect, extract, and reutilize the valuable information embedded in those models. The core idea is to estimate the level of contamination for each class in the abnormal models, and to selectively transfer reliable knowledge from these models to the final global model through a knowledge distillation process. Specifically, we first quantify the contamination level of each class to evaluate the extent to which its parameters are affected by the label flipping attack. Local models are clustered into two groups using the K-means algorithm based on the pairwise similarity of their parameters. The models in the smaller cluster are aggregated to construct a basic global model, while those in the larger cluster are aggregated to obtain a compromised global model. By comparing the neuron weights in the output layer of these two global models, FedCAR estimates the contamination level of each class. Based on this, FedCAR introduces a weighted knowledge distillation loss function. In this framework, the clipped compromised models serve as teacher models, while the basic global model acts as the student. During the distillation process, the contamination level is employed to dynamically modulate the student′s learning rate for each class-specific knowledge transfer from the teachers. Classes with higher contamination are assigned lower learning weights to suppress the influence of poisoned knowledge, whereas classes with lower contamination receive higher weights, enabling the student model to absorb valuable information. Through this contamination-aware adaptive learning strategy, FedCAR effectively transfers useful parameter knowledge from compromised local models to the final global model while mitigating the adverse effects of poisoning. Experimental evaluations conducted on the MNIST and CIFAR10 datasets demonstrate the effectiveness and robustness of the proposed method. Compared with nine state-of-the-art (SOTA) defense techniques, FedCAR achieves an average improvement of 1.81% in the global model′s test accuracy. These results clearly indicate that FedCAR not only mitigates the impact of label flipping attacks but also successfully extracts and reutilizes valuable knowledge embedded in abnormal local models, thereby enhancing the overall robustness and generalization capability of federated learning systems.

Keywords federated learning; data poisoning attack; label flipping attack; anomaly detection; knowledge distillation

1 引 言

联邦学习[1]是一种分布式机器学习技术，能够在不集中收集各方隐私数据的前提下，由服务器协调多个客户端在本地数据上协同训练机器学习模型，有效解决了数据孤岛和数据融合之间的矛盾[2]，被广泛应用于金融[3]、医疗[4]等领域。然而，由于客户端与服务器之间存在大量的参数交换过程，且客户端的本地训练过程对服务器不可见，而服务器对客户端所提交的参数又缺乏有效的验证机制，全局模型被攻击的风险大幅增加[5]。投毒攻击便是其中一种典型的安全威胁，恶意客户端向服务器发送受污染的模型更新，导致全局模型中毒[6]。

投毒攻击中最具代表性和危害性的一种形式为标签翻转攻击[7-10]。此类攻击中，恶意客户端在数据收集阶段将源类样本的标签篡改为目标类，同时保持其他数据信息不变，从而在表面上维持对非源类样本的识别准确率，同时误导本地模型将源类样本错误分类为目标类。又因为联邦学习场景中的局部污染会被全局扩散，恶意客户端的攻击将进一步破坏全局模型的决策边界，造成全局模型性能下降[11]。标签翻转攻击实施难度低、隐蔽性高、破坏性广的特点使其成为联邦学习安全领域的一大威胁，亟需高效的防御方法加以应对。

在良性客户端占多数的假设下，研究者们针对标签翻转攻击的防御方法展开了大量研究，主要分为健壮聚合和异常检测两类。基于健壮聚合的防御

方法[12-15]通常采用中位数、截尾平均等统计分析技术，在参数聚合过程中去除偏离程度较大的客户端模型参数，以此削弱异常模型对全局模型的干扰。尽管此类方法的计算开销较小，但其通常对模型参数进行逐维度处理，难以捕捉参数在高维空间中的异常，在面对复杂或精心设计的攻击时防御能力有限。为此，后续研究进一步提出了基于异常检测的防御策略[16-26]。该方法分析恶意客户端模型与良性客户端模型在参数更新上的差异，利用聚类技术识别并舍弃异常客户端模型。尽管现有研究可以有效消除异常客户端模型的影响，但全局模型性能仍有进一步提升的空间。

现有的防御方法普遍采用“识别-舍弃”的策略，即在全局聚合阶段直接剔除被判定为异常的客户端模型参数。该策略虽有助于过滤中毒信息，提升全局模型的健壮性，但在应对标签翻转攻击时存在一定局限性：尽管恶意客户端模型对源类样本的识别能力显著下降，其对非源类样本的识别能力仍保持在较高水平，如图1所示。直接舍弃这些模型不仅会浪费客户端的本地计算资源，还可能导致潜在的有效知识丢失，从而削弱全局模型的泛化能力。因此，关键在于如何有效保留异常客户端模型中未受污染的有效参数知识，最大程度地降低标签翻转攻击带来的全局性能损失。

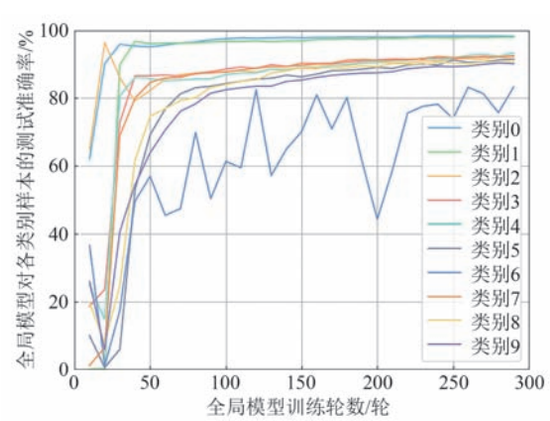


图1　FedAvg在标签翻转攻击下对各类样本的识别准确率（类别6为源类，类别2为目标类）

为缓解现有防御策略在参数利用方面的不足，本文引入知识蒸馏（Knowledge Distillation）技术，以将异常客户端模型中的有效知识迁移至全局模型。知识蒸馏作为一种经典的知识迁移方法，通过最小化教师模型与学生模型在特征表示或输出分布上的Kullback-Leibler[27]（KL）散度，结合学生输出与真实标签之间的交叉熵损失，引导学生模型学习教师模型的知识。然而，现有蒸馏方法普遍对不同类别的知识赋予统一的学习权重，缺乏区分样本污染程度的能力。当异常客户端模型作为教师时，其中包含的攻击信息可能会被一并迁移至全局模型，从而削弱其健壮性。此外，多数蒸馏方法依赖于与客户端训练数据分布高度一致的辅助数据集，而这一条件在实际部署中往往难以满足。

针对上述问题，本文提出标签翻转攻击污染度感知的异常客户端模型参数重用算法（Federated Contamination-Aware Reutilization algorithm，FedCAR）。在现有异常检测方法将客户端模型划分为正常与异常的基础上，FedCAR分别聚合两类模型以构建基础全局模型和异常全局模型，并通过比较二者在输出层神经元参数上的差异量化异常模型中各类别知识的污染度，据此动态调整基础全局模型在蒸馏过程中对异常模型各类别知识的学习权重。此外，FedCAR使用与训练数据集类别数一致但标签不同的样本作为辅助数据集，通过仅比较基础模型与异常模型在逻辑输出上的差异进行知识蒸馏。本文的主要贡献如下。

（1）针对现有防御策略参数利用不足的问题，提出标签翻转攻击污染度感知的异常客户端模型参数重用算法FedCAR。该方法量化异常客户端模型中各类样本的污染度，并将其作为权重引入加权知识蒸馏过程，从而动态调整基础全局模型对不同类别知识的学习率，有效重用了异常客户端模型的有效参数。

（2）基于MNIST和CIFAR10数据集的实验结果表明，FedCAR的全局模型测试准确率较FLTrust、FLAME、XMAM、LFighter和DPFLA等九个基线模型平均提升了1.81%，有效挖掘并重用了异常客户端模型中未受污染的有效知识。

本文第1节为引言；第2节介绍相关工作；第3节明确本方法的威胁模型；第4节介绍本文设计的标签翻转攻击污染度感知的异常客户端模型参数重用算法；第5节通过对比实验验证了本方法的有效性；第6节总结全文。

2　相关工作

目前，国内外研究者针对联邦学习已提出了多种攻击策略，主要分为模型投毒攻击和数据投毒攻击两类。同时，对应的防御策略也被提出，主要分为

健壮聚合和异常检测两类。本文针对数据投毒攻击下的标签翻转攻击的防御方法展开研究，旨在挖掘并重用异常客户端模型中的有效参数知识，主要用到了异常检测和知识蒸馏技术。在介绍所提方法前，本节对相关技术的研究现状展开分析。

2.1 联邦学习与投毒攻击

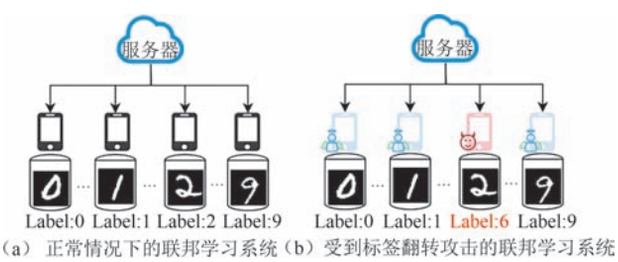
联邦学习可分为五个阶段[28]：(1)数据收集阶段：所有客户端采集用于训练的私有数据，并本地存储。(2)客户端选择阶段：服务器从所有可用客户端中选取部分参与本轮训练，并下发当前全局模型。(3)本地训练阶段：客户端基于本地数据对全局模型进行多轮梯度下降，得到本地模型。(4)模型上传与聚合阶段：服务器按预设的聚合算法整合各客户端上传的本地模型更新，生成新的全局模型并进入下一轮迭代。(5)模型推理阶段：将训练完成的全局模型部署至目标应用场景，对输入的实际样本进行预测。

尽管联邦学习避免了原始数据的集中存储，降低了直接泄漏的风险，但其各执行阶段仍存在安全漏洞，易被攻击者利用实施投毒攻击[29]。在数据收集阶段，恶意客户端可任意篡改本地数据或注入中毒样本，以植入攻击信息[30]。在客户端选择阶段，由于缺乏身份验证机制，攻击者可伪造或篡改客户端身份，多次参与训练，放大恶意更新在全局聚合中的影响[28]。在本地训练阶段，服务器无法直接监督客户端训练过程，恶意客户端可任意篡改本地超参数或损失函数，以改变优化目标[31]。在模型上传阶段，由于缺乏更新验证机制，恶意客户端可注入精心构造的恶意参数。在全局聚合阶段，传统加权平均策略的健壮性有限，易受极值更新干扰，导致全局模型偏离正常优化轨迹[32]。最后，在推理阶段，中毒的全局模型在特定条件下会输出错误的预测。根据攻击目标不同，投毒攻击可分为模型投毒攻击(Model Poisoning Attack)与数据投毒攻击(Data Poisoning Attack)两类。

模型投毒攻击[29]中，恶意客户端通过修改本地训练的超参数或损失函数，或在提交模型更新前直接篡改模型参数，从而干扰全局模型的正常聚合。模型替换攻击[33](Model Replacement Attack)通过放大本地更新幅度，间接提高其在全局聚合中的权重，同时控制更新范数以规避检测，从而实现隐蔽投毒。符号翻转攻击[34](Sign Flipping Attack)通过翻转本地更新的符号，误导全局模型的训练方向，从而阻碍其正常收敛。

数据投毒攻击[35]中，恶意客户端在数据收集阶段注入恶意数据或直接篡改原始数据，使本地模型在恶意的数据分布上学习。边缘案例后门[36](Edge-case Backdoors)通过诱导全局模型在小概率但合法的边缘样本上产生错误预测实现投毒。分布式后门攻击[37](Distributed Backdoor Attack，DBA)将全局触发器拆分为相互独立的局部触发器并分配给不同恶意客户端，以提高后门攻击的隐蔽性。Gong等人[38]通过优化触发器掩码的值分布，构建了一种能够轻易激活特定恶意神经元的触发器。

标签翻转攻击[10]是一种常见的数据投毒攻击，攻击者通过将训练数据中源类样本的标签篡改为目标类破坏模型的决策边界，如图2所示。例如，将手写数字识别中数字“2”的标签翻转为“6”，或者将交通标志识别中的“停止”标志标记为“继续”。此类攻击可在数据收集阶段离线进行，计算开销较小，技术门槛低，且对模型性能影响显著[39]。



(a) 正常情况下的联邦学习系统 (b) 受到标签翻转攻击的联邦学习系统

图2 标签翻转攻击

2.2 标签翻转攻击的防御

在恶意客户端数量少于良性客户端的假设下，现有的防御方法主要分为健壮聚合和异常检测两类。

基于健壮聚合的防御方法通过不同聚合规则剔除各维度上统计异常的客户端更新，以保障全局模型的健壮性。Krum[12]通过最小化到其他客户端模型的欧几里德距离选择参与聚合的模型。Coordinate-Wise Median[13]选取所有客户端模型在每一维度上的中值作为下一轮训练的全局模型。Trimmed Mean[13]在每个维度上去除若干个最大值和最小值，再对剩余参数求平均。Shalom等人[14]通过评估每个客户端更新与更新中位数的差异过滤异常参数。TFL-DT[15]基于客户端多维行为画像动态调整聚合权重，提升可信客户端贡献。FLTrust[40]基于服务器在辅助数据集上训练的参考模型，评估客户端模型更新的方向偏离程度，动态调整其信任分数和聚合权重。尽管此类方法计算开销较低，但

因逐维处理参数，难以捕捉高维异常模式，因而后续研究转向基于异常检测的防御策略，以应对更复杂攻击。

基于异常检测的防御方法对客户端模型更新间的余弦相似度等指标进行二聚类，并将较小簇判定为异常，在聚合时剔除异常模型。Tolpegin等人[16]对客户端模型的输出层参数进行PCA降维并聚类，从而在训练早期也能实现高效检测。Zang等人[17]利用KL散度弥补余弦相似度在捕捉高维模型参数差异上的局限性。针对车载网络的标签翻转攻击，Lfgurad[18]将客户端模型最后一层的激活值输入支持向量机，有效分离了异常客户端模型。LSM[19]提出了一种基于本地模型损失和样本量的评分机制，结合曼哈顿相似度构建可信的异常检测方法。利用正常与异常客户端模型对随机矩阵的输出差异，XMAM[20]避免了对辅助数据集的依赖。MCDFL[21]借助生成对抗网络提取客户端模型的潜在特征，并通过对比客户端模型对该特征及其本地数据的输出差异评估其数据质量。pFL-IDS[22]基于客户端模型与全局模型在输出层参数上的余弦相似度筛选正常客户端，并动态调整其聚合权重。CONTRA[23]通过客户端之间的余弦相似度评估对齐水平，并将对齐水平更高的客户端判定为异常。LFighter[24]将客户端模型输出层中梯度值最高的两个神经元作为潜在源类和目标类神经元。RSim-FL[25]利用客户端模型逻辑输出与全局模型逻辑输出之间的皮尔逊相似度区分客户端。FLAME[26]利用HDBSCAN[41]动态聚类客户端更新之间的余弦相似度。DPFLA[42]引入适用于奇异值分解(Singular Value Decomposition，SVD)的可移除噪声，并结合K-Means算法进行攻击检测，有效保障了联邦学习系统的安全性和隐私性。

现有防御方法尽管能在一定程度上防御标签翻转攻击，但通常仅聚合正常客户端模型参数，未能充分挖掘异常客户端模型中未被污染的有效知识。为此，本文提出FedCAR算法，通过挖掘并重用异常客户端模型的有效参数知识，提升全局模型的泛化能力。

2.3　知识蒸馏

知识蒸馏(Knowledge Distillation，KD)以概率分布的形式将教师模型知识迁移至结构更轻量的学生模型中，从而在降低计算开销的同时尽可能保留其性能。在联邦学习中，因其无需共享模型参数或原始数据，知识蒸馏被广泛用于应对模型异构[43]、数据异构[44]等问题。

FD[45]首次将知识蒸馏引入联邦学习以解决模型异构问题，客户端基于私有数据计算各类样本的平均逻辑输出，并在本地训练中将全局聚合后的逻辑输出作为教师输出进行知识蒸馏。DS-FL[46]提出熵减聚合策略，证明在对全局逻辑输出应用Softmax时，设置小于1的温度有助于降低熵值，提升训练效果。MHAT[47]引入用于客户端本地蒸馏和全局模型蒸馏的公开辅助数据集，以提升知识迁移效果。FedAD[48]则结合中间层特征图与输出层逻辑输出进行蒸馏。

知识蒸馏技术也可用于提高全局模型在数据异构情况下的泛化能力。Fedzkt[49]利用生成器生成的合成样本对全局模型进行零样本微调。Dafkd[50]按重要性加权聚合客户端模型的逻辑输出，并将其作为教师输出进行蒸馏。FedNTD[51]在计算蒸馏损失时忽略实际类别对应的logit，以缓解训练中的遗忘问题。FedGEN[52]利用服务器端的轻量级生成器为每个客户端生成增强样本，以全局知识引导本地训练。

近年来，研究者进一步将知识蒸馏用于防御联邦学习投毒攻击。HYDRA-FL[53]为缓解攻击放大问题，提出双层知识蒸馏方法，同时计算中间层特征图与最终层逻辑输出的蒸馏损失，以降低对逻辑输出的依赖。SPFL[54]提出注意力引导的自蒸馏方法，良性客户端利用本地可信历史特征监督全局模型的训练，以实现本地模型的自我净化。BadCleaner[46]通过蒸馏多个良性客户端的模型知识，有效削弱了全局模型对触发器的注意力。Fedredefense[55]观察到正常模型更新可通过本地知识蒸馏重构，而异常更新难以还原，据此提出基于更新重构误差的攻击检测方法。

尽管上述方法在提升全局模型健壮性方面取得了一定进展，但若将现有知识蒸馏方法直接用于异常客户端模型参数重用，仍存在两个问题：其一，现有方法通常对所有知识赋予相同的学习权重，易将被污染知识一并引入基础全局模型；其二，此类方法大多依赖于与客户端训练数据类别完全一致的辅助数据集，但在实际场景中这一前提难以满足，服务器更可能收集到类别数相同但类别标签不同的辅助数据。例如，当客户端数据集为包含10个类别的手写数字数据集时，服务器可能更容易获取同样具有10个类别的时尚产品数据集。为此，本文提出加权知识蒸馏算法，以污染度为学习权重，引导基础全局

模型吸收异常客户端模型中未被污染的有效知识。

3 威胁模型

在横向联邦学习场景下，本文对标签翻转攻击的防御技术展开研究。为明确本方法的适用范围，从攻击目标、攻击知识、攻击能力三个维度建立威胁模型，如图3所示。

图3 威胁模型

在攻击目标方面，多个恶意客户端共同实施相同的标签翻转攻击，其目的在于使全局模型错误地将源类样本预测为目标类，同时不影响全局模型对其他样本的识别准确性。

在攻击知识方面，恶意客户端掌握全局模型的聚合算法、服务器部署的防御策略等背景知识，但无法获取任何关于良性客户端的信息。

在攻击能力方面，恶意客户端仅能操控本地训练过程，而无法干预良性客户端的本地训练过程和服务器的防御策略。

4 FedCAR

本文提出的标签翻转攻击污染度感知的异常客户端模型参数重用算法FedCAR包括五个步骤，如图4所示。服务器随机初始化全局模型的参数权重w_G后，重复以下5个步骤，直到全局模型收敛。

步骤1： 服务器发送全局模型。服务器从N个客户端中随机选择n个参与通信，并将全局模型参数$w_G=\{w_G^e\}$发送给被选中的客户端集合$C=\{C_i\}$。其中，w_G^e表示全局模型在第e维上的参数权重，C_i表示被选中的第$i\in[1,n]$个客户端。

步骤2： 客户端进行本地训练。客户端C_i以η_i为学习率，L_i为损失函数，使用本地数据集D_i对全局模型参数w_G进行训练，利用随机梯度下降(Stochastic Gradient Descent，SGD)得到本地模型$w_i=\{w_i^e\}$，如公式(1)所示。训练完毕后，客户端C_i将本地模型更新$u_i=w_i-w_G$发送给服务器。

$$w_i=w_G-\eta_i\nabla L_i(w_G,D_i) \tag{1}$$

步骤3： 服务器检测异常客户端模型参数。服务器利用公式(2)计算所有客户端更新之间的余弦相似度矩阵S，并将其输入K-means算法中进行二聚类，得到两个簇$C_a,C_b=Kmeans(S)$。设簇$C_*\in\{C_a,C_b\}$中客户端数量为$|C_*|$，则服务器将较大簇$C_{ben}=\arg\max|C_*|$中的客户端标记为正常，而较小簇$C_{mal}=\arg\min|C_*|$中的客户端标记为异常。

$$S=[s_{i,j}],\ s_{i,j}=\frac{\langle u_i,u_j\rangle}{\|u_i\|_2\cdot\|u_j\|_2}\ (\forall i,j\in[1,n]) \tag{2}$$

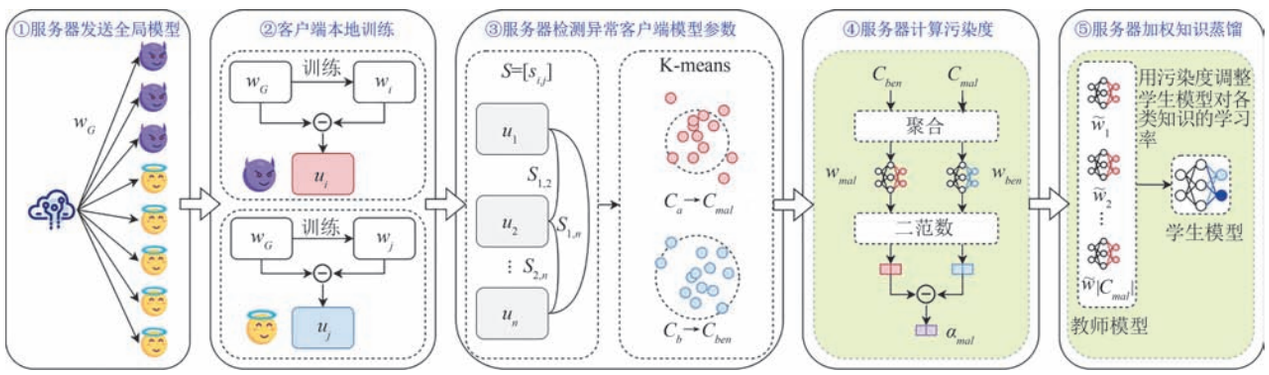


图4 标签翻转攻击污染度感知的异常客户端模型参数重用算法

步骤4： 服务器计算异常客户端模型的污染度。服务器利用公式(3)分别聚合簇C_{ben}和C_{mal}中的客户端模型参数，得到基础全局模型w_{ben}和异常全局模型w_{mal}，然后对比两个全局模型在输出层参数上的差异，以量化异常客户端模型中各类样本的污染度。

$$w_{ben}=\frac{1}{|C_{ben}|}\sum_{i\in[C_{ben}]}w_i$$
$$w_{mal}=\frac{1}{|C_{mal}|}\sum_{i\in[C_{mal}]}w_i \tag{3}$$

步骤5： 服务器重用异常客户端模型参数。首先以基础全局模型$w_{ben}=\{w_{ben}^e\}$在各维度上的参数值w_{ben}^e作为裁剪阈值，对簇C_{mal}异常客户端模型$w_i=\{w_i^e\}$对应维度的参数w_i^e进行截断，以削弱异常客户端模型的毒性，得到参数裁剪后的模型$\tilde{w}_i=\{\tilde{w}_i^e\}$，如公式(4)所示。然后，以裁剪后的异常客户端模型为教师模型，基础全局模型为学生模型，各类样本的污染度为学习权重，进行加权知识蒸馏，形成最终的全局模型。

$$\tilde{w}_i^e=\frac{w_i^e}{\max\left(1,\left\|w_i^e\right\|_2/w_{ben}^e\right)} \tag{4}$$

FedCAR算法的具体实现如算法1所示。

算法1. FedCAR算法

输入：N个客户端$\{C_i\}$的本地数据$\{D_i\}$、损失函数$\{L_i\}$和学习率$\{\eta_i\}$，通信轮数T，辅助数据集D_{aux}，初始局模型w_G

输出：最终全局模型w_G

1. PROCEDURE $FedCAR(\{C_i,D_i,L_i,\eta_i\},T,D_{aux},w_G)$
2. FOR $t=1,2,\cdots,T$ DO
3. 服务器从N个客户端中随机n个参与训练
4. // 客户端进行本地训练
5. FOR C_i IN $\{C_i\}$ DO
6. 客户端C_i接收服务器下发的全局模型w_G
7. 训练本地模型$w_i=w_G-\eta_i\nabla L_i(w_G,D_i)$
8. 上传本地更新$u_i=w_i-w_G$
9. END FOR
10. // 服务器进行攻击检测
11. 用公式(2)计算客户端更新之间的余弦相似度S
12. 对相似度矩阵进行二聚类$C_a,C_b=Kmeans(S)$
13. 将较大簇$C_{ben}=\arg\max|C_*|$中的客户端标记为正常
14. 将较小簇$C_{mal}=\arg\min|C_*|$中的客户端标记为异常
15. // 服务器进行参数重用
16. 算法2计算最终全局模型$w_G=WKD(D_{aux},C_{ben},C_{mal})$
17. END FOR
18. RETURN w_G

4.1 污染度

由于标签翻转攻击破坏了特征与标签之间原有的映射关系，异常客户端模型在各个输出层神经元参数上与正常客户端模型存在显著差异。对此，本文设计了一种用于评估异常客户端模型中各类样本受污染程度的污染度指标，如图5所示。

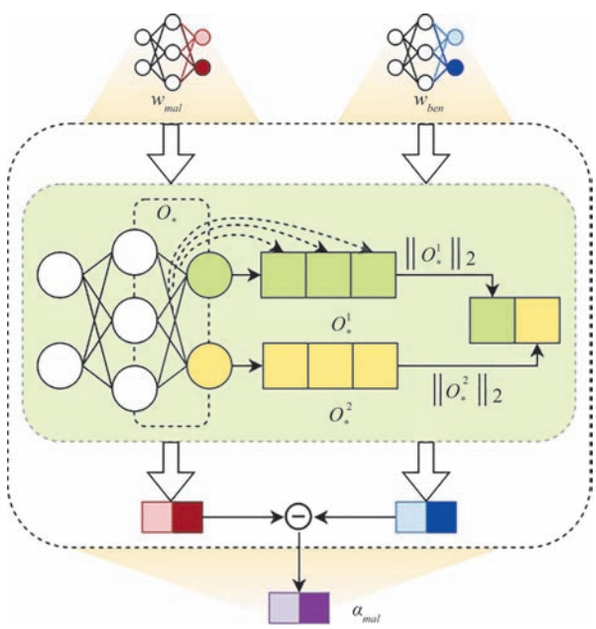


图5　污染度

服务器分别提取基础全局模型w_{ben}和异常全局模型w_{mal}的输出层参数$o_{ben}=\{o_{ben}^k\}$和$o_{mal}=\{o_{mal}^k\}$，其中o_{ben}^k和o_{mal}^k表示两个全局模型中与第$k\in[1,K]$个输出层神经元相关的参数向量。然后，分别计算两个全局模型中每个输出层神经元对应的参数范数$\|o_{ben}^k\|_2$和$\|o_{mal}^k\|_2$，并计算上述两个范数向量之间的欧氏距离，以量化异常客户端模型中各类样本的污染度$\alpha_{mal}=\{\alpha_{mal}^k\}$，如公式(5)所示。$|\alpha_{mal}^k|$越大，第$k$类样本越有可能受到标签翻转攻击的污染。

$$\alpha_{mal}^k=\|o_{ben}^k\|_2-\|o_{mal}^k\|_2 \tag{5}$$

4.2 加权知识蒸馏

为将异常客户端模型中未受污染的有效知识迁移至基础全局模型中，服务器应以裁剪后的异常客户端模型为教师模型，以基础全局模型为学生模型，进行知识蒸馏。然而，已有的知识蒸馏方法对所有

知识赋予相同的学习权重，忽视了不同类别知识受攻击影响的差异，可能导致攻击信息在蒸馏过程中被引入学生模型。此外，这些方法通常依赖于与训练数据类别完全一致的辅助数据集，但在实际场景中，这一假设难以满足，服务器往往更容易收集一个类别数量相同但类别标签不同的数据集。

针对上述问题，本文设计了加权知识蒸馏损失函数，如图6所示。为了引导基础全局对异常参数知识和正常参数知识加以区分地学习，本文引入基于污染度的权重调整机制，赋予污染度较低的样本信息更高的学习权重。同时，为解决数据集不匹配的问题，服务器在知识蒸馏过程中仅计算基础全局模型与异常客户端模型在逻辑输出上的差异，而不对比基础全局模型的逻辑输出与样本独热编码标签的差距。

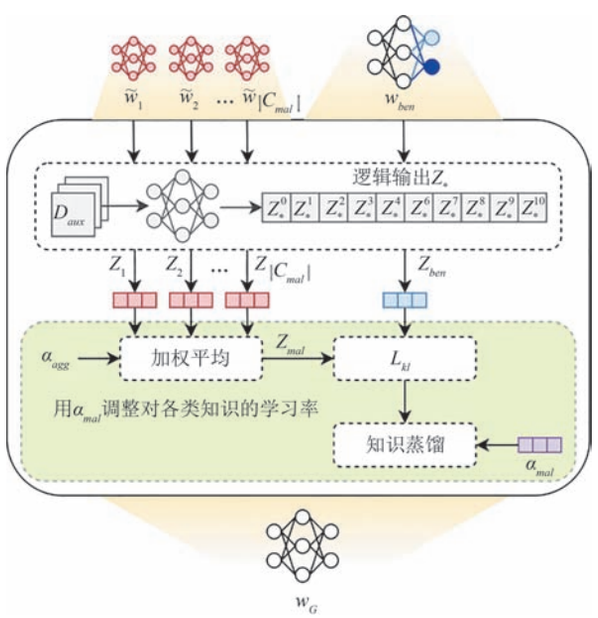


图6 加权知识蒸馏

首先，服务器收集一个与客户端训练数据集一样包含K类样本的辅助数据集D_{aux}，并以该数据集为输入，计算基础全局模型的逻辑输出z_{ben}，如公式(6)所示。

$$z_{ben}=\{z_{ben}^{k}\}=f(D_{aux},w_{ben}) \tag{6}$$

同时，计算所有异常客户端模型的逻辑输出并聚合。由于不同异常客户端模型包含不同的样本信息，而这些信息对基础全局模型的重要性各不相同，因此采用可学习参数$\alpha_{agg}=\{\alpha_{agg}^{i}\}$来调整不同逻辑输出的聚合权重。各异常客户端模型的初始聚合权重相同，均为$\alpha_{agg}^{i}=1/|C_{mal}|$，聚合后的异常客户端模型逻辑输出$z_{mal}$如公式(7)所示。

$$z_{mal}=\{z_{mal}^{k}\}=\sum_{i\in[C_{mal}]}\alpha_{agg}^{i}\cdot f(D_{aux},\tilde{w}_i) \tag{7}$$

其次，用KL散度衡量基础全局模型和异常客户端模型在逻辑输出上的分布差异L_{kl}。为了增强基础全局模型对各类别样本的区分能力，本文将原始K分类任务拆解为K个二分类任务。对于第k类样本的预测概率，计算向量$\left[z_{mal}^{k},1-z_{mal}^{k}\right]$与$\left[z_{ben}^{k},1-z_{ben}^{k}\right]$之间的KL散度$L_{kl}^{k}$，如公式(8)所示。

$$L_{kl}^{k}=z_{mal}^{k}\cdot\log\left(\frac{z_{mal}^{k}}{z_{ben}^{k}}\right)+(1-z_{mal}^{k})\cdot\log\left(\frac{1-z_{mal}^{k}}{1-z_{ben}^{k}}\right) \tag{8}$$

最后，将基础全局模型每类知识的学习率设置为翻转后的污染度$\tilde{\alpha}_{mal}^{k}=1-\left|\alpha_{mal}^{k}\right|$，并用加权知识蒸馏损失函数$L_{wkd}(\cdot)$对聚合权重$\alpha_{agg}$和基础全局模型参数$w_{ben}$进行优化，如公式(9)所示。

$$\min_{\alpha_{agg},w_{ben}}L_{wkd}\left(\{\tilde{\alpha}_{mal}^{k}\},\{L_{kl}^{k}\}\right)=\sum_{k=1}^{K}\tilde{\alpha}_{mal}^{k}\cdot L_{kl}^{k} \tag{9}$$

在训练过程中，服务器首先固定基础全局模型的参数权重w_{ben}，使用随机梯度下降算法学习聚合权重α_{agg}；其次，固定聚合权重α_{agg}，再次使用随机梯度下降算法优化基础全局模型参数w_{ben}。上述过程的具体实现如算法2所示。

算法2. 加权知识蒸馏算法

输入：辅助数据集D_{aux}，正常客户端C_{ben}，异常客户端C_{mal}

输出：下一轮全局模型w_G

1. PROCEDURE $WKD(D_{aux},C_{ben},C_{mal})$
2. // 计算各类样本的污染度
3. 用公式(3)计算基础和异常全局模型w_{ben}和w_{mal}
4. 分别提取w_{ben}和w_{mal}的输出层参数o_{ben}和o_{mal}
5. 计算异常客户端模型的污染度$\alpha_{mal}=\left\{\left\|o_{ben}^{k}\right\|_2-\left\|o_{mal}^{k}\right\|_2\right\}$
6. // 以污染度为权重进行知识蒸馏
7. 计算基础全局模型的逻辑输出$z_{ben}=f(D_{aux},w_{ben})$
8. FOR i IN C_{mal} DO
9. 用公式(4)计算参数裁剪后的异常客户端模型$\tilde{w}_i$
10. 计算异常客户端模型的逻辑输出$z_i=f(D_{aux},\tilde{w}_i)$
11. END FOR
12. 用初始权重$\alpha_{agg}^{i}=1/|C_{mal}|$聚合$z_i$得到$z_{mal}$
13. 固定基础全局模型的参数w_{ben}，用公式(9)优化α_{agg}
14. 固定聚合权重α_{agg}，用公式(9)优化w_{ben}得到w_G
15. RETURN w_G

5　实验设计

5.1　实验设置

5.1.1　实验数据

本文以MNIST[56]和CIFAR10[57]数据集为客户端的训练集。特别的，在FedCAR中，服务器分别采用Fashion MNIST[58]和CIFAR100[57]作为对应的辅助数据集进行加权知识蒸馏；而在FLTrust中，服务器使用和训练集类别一样的样本作为辅助数据。所有辅助数据集均包含300个样本，各数据集的信息如表1所示。

表1　实验数据

数据集	MNIST	Fashion MNIST	CIFAR10	CIFAR100
训练集样本量	60 000	60 000	50 000	50 000
测试集样本量	10 000	10 000	10 000	10 000
类别数	10	10	10	100
通道数	1	1	3	3
尺寸	28×28	28×28	32×32	32×32

MNIST数据集是一种经典的手写数字图像数据集，涵盖了从数字0到数字9的10个类别，共包含70 000张灰度图像，其中训练集有60 000个样本，测试集有10 000个样本，每张图像的尺寸为28×28像素。

Fashion MNIST数据集是一个经典的时尚产品数据集，包含了裤子、套头衫、裙子等10类产品，共有70 000张灰度图像，其中训练集有60 000个样本，测试集有10 000个样本，每张图像的尺寸为28×28像素。与MNIST数据集相比，Fashion MNIST的纹理特征更加复杂。

CIFAR10数据集是一个经典的自然图像分类数据集，包含了飞机、汽车、猫、狗等10类常见物体，共有60 000张彩色图像，其中训练集包含50 000个样本，测试集包含10 000个样本，每张图像的尺寸为32×32像素。与MNIST等灰度图像数据集相比，CIFAR10在背景、纹理和颜色等方面更加复杂，因而更具挑战性。

CIFAR100数据集包含了如苹果、自行车等100个类别的物体，共有60 000张彩色图像，其中训练集有50 000个样本，测试集有10 000个样本，每张图像的尺寸为32×32像素。为了与CIFAR10数据集的类别数相匹配，在知识蒸馏过程中，服务器随机选取CIFAR100中的10类样本进行训练。

5.1.2　全局模型

本文选择卷积神经网络作为全局模型。用于训练MNIST数据集的全局模型由两个卷积层、一个随机失活层和两个全连接层组成。用于训练CIFAR10数据集的全局模型由两个卷积层和三个全连接层组成，具体结构如表2所示。

表2　全局模型

数据集	类型	通道数	卷积核	激活函数
MNIST	输入数据	1	-	-
	卷积	10	5*5	-
	最大池化	10	2*2	ReLU
	卷积	20	5*5	-
	最大池化	20	2*2	ReLU
	随机失活	320	-	-
	全连接	50	-	ReLU
	全连接	10	-	Log Softmax
CIFAR10	输入数据	3	-	-
	卷积	6	5*5	ReLU
	最大池化	6	2*2	-
	卷积	16	5*5	ReLU
	最大池化	16	2*2	-
	全连接	120	-	ReLU
	全连接	84	-	ReLU
	全连接	10	-	Log Softmax

5.1.3　实验参数

本文选取以下九种方法作为基线，包括FedAvg[1]、Krum[12]、Trimmed Mean（TMean）[13]、Coordinate-Wise Median（Median）[13]、FLTrust[40]、FLAME[26]、XMAM[20]、LFighter[24]和DPFLA[42]。其中，FedAvg代表无防御策略下的基础方案。

在客户端数据集独立同分布场景下，联邦学习系统包含100个客户端，其中包括30个恶意客户端和70个良性客户端。服务器与客户端在MNIST数据集上训练时进行300轮通信，在CIFAR10数据集上训练时进行1000轮通信。每轮训练中，随机选取50个客户端参与模型更新。客户端在本地训练阶段采用批量梯度下降策略，每批包含600个样本，并基于这些数据进行3轮训练，以生成本地模型。在加权知识蒸馏阶段，服务器对基础全局模型进行1轮训练。

在所有实验中，恶意客户端实施标签翻转攻击，将源类和目标类分别设置为类别6和类别2。在数据预处理阶段，所有恶意客户端将本地所有源类样本的标签翻转为目标类，并使用污染后的数据集进行本地模型训练。恶意客户端在整个训练期间保持

固定，仅在被服务器选中参与训练的轮次中发动攻击。此外，恶意客户端与良性客户端的超参数设置一致。除非另有特别说明，以下实验均在上述设定下开展。

5.1.4 性能指标

本文将从测试准确率的角度衡量全局模型的性能，并用假阳率和假阴率评估攻击检测效果。

测试准确率ACC是指全局模型正确预测的样本数$|D_{right}|$占测试样本总数$|D_{test}|$的比例，如公式(10)所示。该指标越大，说明全局模型性能越好。

$$ACC=\frac{|D_{right}|}{|D_{test}|}\times 100\% \tag{10}$$

假阳率FPR表示被误判为异常的良性客户端占全部良性客户端的比例。若TP表示被正确判定为异常的恶意客户端数量，TN表示被正确判定为正常的良性客户端数量，FP表示被错误判定为异常的良性客户端数量，FN表示被误判为正常的恶意客户端数量，则假阳率FPR的计算公式如式(11)所示。该指标越大，说明被误判的良性客户端越多。

$$FPR=\frac{FP}{FP+TN}\times 100\% \tag{11}$$

假阴率FNR表示被误判为正常的恶意客户端占全部恶意客户端的比例，如公式(12)所示。该指标越大，说明被误判的恶意客户端越多。

$$FNR=\frac{FN}{FN+TP}\times 100\% \tag{12}$$

5.2 实验结果

5.2.1 评估全局模型的收敛性

为了验证全局模型的收敛性，本文以MNIST和CIFAR10为训练集，以Fashion MNIST和CIFAR100为辅助数据集，从测试准确率的角度与九种基线模型进行性能比对。图7描绘了训练过程中全局模型测试准确率和损失值的变化情况，表3展示了全局模型在收敛后的测试准确率对比结果。

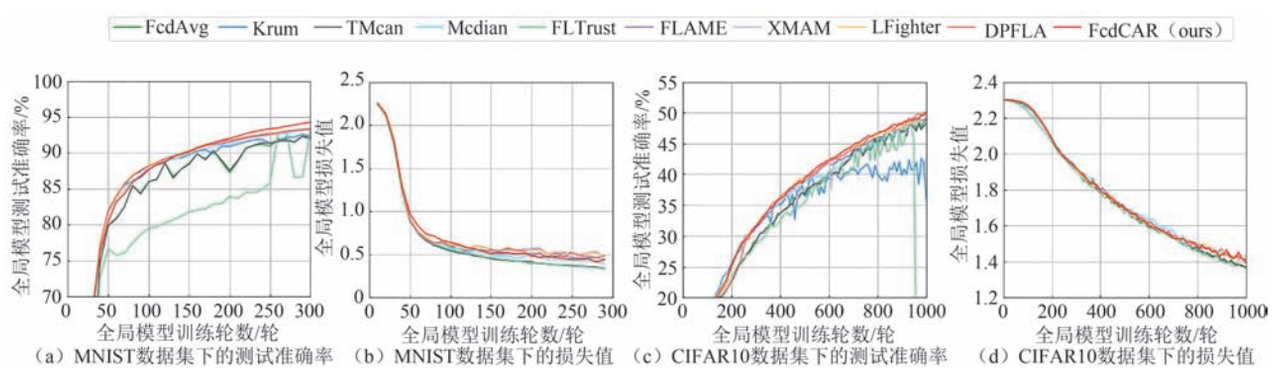


图7 全局模型的性能比对

表3 对比不同方法的性能

方法	MNIST	CIFAR10
FedAvg	91.91	48.96
Krum	92.36	35.96
TMean	92.16	48.18
Median	93.43	49.19
FLTrust	93.37	48.36
FLAME	93.26	49.08
XMAM	93.49	49.55
LFighter	93.44	49.65
DPFLA	93.38	**50.09**
FedCAR	**94.28**	49.98

在MNIST数据集上，所有方法的损失值均在50轮通信内快速下降，达到0.9左右；50轮之后，收敛速度减慢，损失值趋近于0.3。最终，Krum和TMean的测试准确率分别为92.36%和92.16%，相比FedAvg有小幅提升。FLTrust的全局模型测试准确率在前250轮内稳定上升至85%，随后在93%到86%之间波动，最终达到93.37%。相比之下，Median方法在训练过程中更加稳定，准确率较FLTrust提高了0.06%，且优于FLAME的93.26%和DPFLA的93.38%。上述结果表明，基于健壮聚合的防御方法可以在统计上有效剔除恶意客户端模型参数。XMAM和LFighter的准确率分别为93.49%和93.44%，整体表现优于基于健壮聚合的方法，展现出更强的异常检测与防御能力。然而，上述方法的准确率始终未能突破93.5%。相比之下，本文提出的FedCAR的损失值从最初2.25的下降至0.44，并以94.28%的测试准确率在所有方法中表现最佳，较XMAM提升了0.79%。

在CIFAR10数据集上，各方法的损失值在

1000轮通信内缓慢下降，最终趋近于1.35。在无防御场景下，FedAvg方法的测试准确率达到48.96%。在有防御场景下，Krum方法防御失效，准确率比FedAvg低13.00%，表明其在有限通信轮数内难以抵御标签翻转攻击，健壮性较弱。TMean和FLTrust方法的整体性能与FedAvg相近，测试准确率分别比FedAvg略低0.78%和0.6%。然而，FLTrust在训练过程中性能波动较大，稳定性相对较差。Median方法的测试准确率为49.19%，在四种基于健壮聚合的防御方法中性能最优，同时较FLAME方法提升了0.11%。XMAM和LFighter方法性能接近，最终准确率分别为49.55%和49.65%。DPFLA方法取得了50.09%的测试准确率，对此类复杂数据集表现出较强的适应性。相较之下，FedCAR方法的损失值从初始的2.30下降至1.39，最终测试准确率达到49.98%，仅比最优的DPFLA方法低0.11%。

综上所述，FedCAR在两个数据集上均表现出良好的收敛性，并实现了较高的测试准确率，验证了其在不同任务中的适应性和有效性。

5.2.2　评估异常客户端识别效果

为评估各异常检测方法在识别异常客户端模型方面的性能，本文以MNIST和CIFAR10为训练集，统计了FLAME、XMAM、LFighter、DPFLA以及本文提出的FedCAR五种方法在所有通信轮次中的平均假阳率和平均假阴率，如表4所示。

表4　对比异常客户端的识别效果

方法	MNIST		CIFAR10	
	FPR	*FNR*	*FPR*	*FNR*
FLAME	19.77	**0**	23.14	**0**
XMAM	0.10	**0**	4.43	0.77
LFighter	**0**	**0**	**0**	**0**
DPFLA	**0**	**0**	**0**	16.00
FedCAR	**0**	**0**	0.15	**0**

实验结果表明，FedCAR在两个数据集上的假阳率和假阴率均接近0%，识别性能与LFighter相当，且整体优于FLAME与XMAM等方法。在MNIST数据集上，FLAME的假阳率高达19.77%，表明存在较高比例的良性客户端被误判为异常。XMAM的假阳率虽仅为0.10%，但在CIFAR10数据集上升至4.43%，表明其在更复杂的数据分布下易受到干扰。DPFLA在CIFAR10数据集上的假阴率达到16%，对部分异常客户端模型存在漏检问题。LFighter在两个数据集上的假阳率和假阴率均为0%，表现出较强的异常检测能力，这得益于其仅选取部分模型参数进行聚类，降低了冗余参数空间对聚类的干扰。但其与其他传统方法相同，仍采用“识别-舍弃”策略，未能利用异常客户端模型中的有效知识，限制了全局性能提升。相比之下，FedCAR虽在CIFAR10数据集上出现极少量假阳，但假阴率始终为0%，未引入受污染参数。同时，参数重用的策略在一定程度上缓解了假阳导致的知识损失，从而保持了较高的全局模型测试准确率。未来将优化识别策略，在保留重用优势的同时进一步降低假阳率。

5.2.3　污染度可视化

为评估污染度指标在挖掘异常模型中受污染信息方面的能力，图8通过热力图可视化了训练过程中异常全局模型对MNIST和CIFAR10数据集10类样本的污染度绝对值$|\alpha_{mal}|$，其中颜色越深表示该类样本受到的污染越严重。

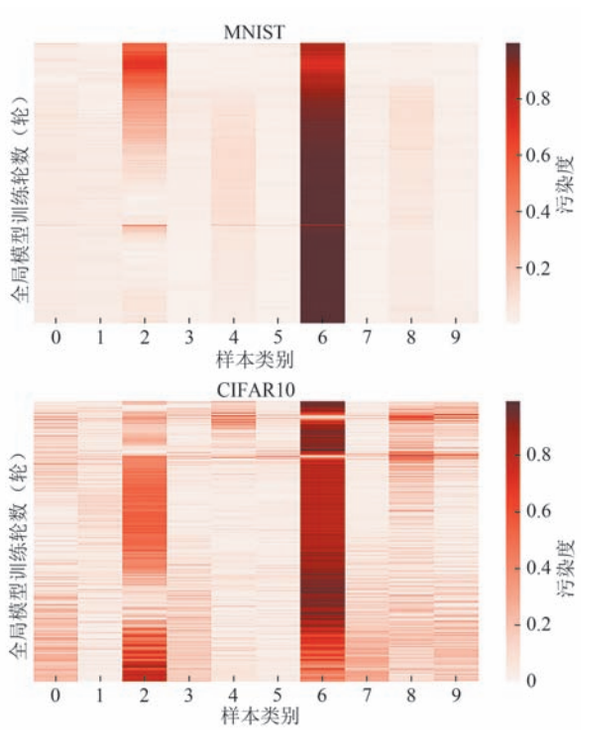


图8　污染度可视化

实验结果表明，异常全局模型中不同类别样本的污染度呈现出“源类>目标类>除了源类和目标类的其他类”的关系。在MNIST数据集上，源类（类别6）在整个训练过程中污染度始终较高，反映了恶意客户端对该类样本进行标签翻转操作的直接

影响；目标类（类别2）的污染度则随通信轮数增加逐渐下降，说明攻击的干扰在训练过程中被部分削弱。相比之下，在CIFAR10数据集上，源类（类别6）在多数训练过程中的污染度依旧显著高于其他类别，显示出攻击的持续影响。目标类（类别2）在模型收敛阶段仍保持较高污染度。这一现象可能源于CIFAR10数据的复杂性，其较低的类内相似性与较高的类间相似性，使得恶意客户端注入的中毒信息更难被识别与消除，导致污染长期残留于该类别。综上所述，污染度指标可以有效评估异常模型中各类知识的受污染程度，为后续的异常客户端模型参数重用奠定了良好的基础。

5.2.4 消融实验

为验证所提出的异常客户端模型参数重用算法中各关键步骤的有效性，本文以MNIST为训练集，Fashion MNIST为辅助数据集，设计了异常检测、参数裁剪、知识蒸馏三个步骤的消融实验。全局模型在不同设置下的测试准确率如表5所示，实心圆表示“使用”，空心圆表示“不使用”。其中，异常检测（步骤3）对客户端更新间的余弦相似度进行聚类以识别异常客户端；参数裁剪（步骤5）在基础全局模型的指导下对异常客户端模型参数进行截断；知识蒸馏（步骤5）进行加权知识蒸馏。由于参数裁剪和知识蒸馏均依赖于异常检测提供的异常客户端识别结果，因此本文不考虑在未使用异常检测的情况下单独使用参数裁剪或知识蒸馏的实验配置。

表5 消融实验

异常检测	参数裁剪	知识蒸馏	*ACC*
○	○	○	91.91
●	○	○	93.42
●	●	○	92.32
●	○	●	93.98
●	●	●	**94.28**

在未引入任何防御方法的情况下，FedAvg方法的全局模型测试准确率为91.91%。单独使用异常检测步骤后，准确率提升了1.51%，表明该步骤可以有效识别并剔除异常客户端模型。若在此基础上进一步引入参数裁剪，被判定为异常的客户端模型在经过参数裁剪后也被纳入聚合过程，导致准确率较仅使用异常检测时下降1.1%，但仍高于无防御场景。这说明参数裁剪在一定程度上具有缓解标签翻转攻击的能力，但也可能将部分异常参数引入全局模型，影响模型性能。同时使用异常检测和知识蒸馏时，全局模型的准确率提升至93.98%，优于仅使用异常检测的设置，表明知识蒸馏可以有效重用异常客户端模型中的有效参数。进一步整合异常检测、参数裁剪和知识蒸馏三个步骤后，FedCAR方法取得最优性能，准确率达94.28%。上述结果表明，若缺少参数裁剪，可能导致全局模型吸收异常客户端模型中的有害信息；而缺乏知识蒸馏，则会丢失异常模型中的有效知识。二者的结合实现了异常客户端模型参数的有效重用。

5.2.5 评估异常客户端模型参数重用算法有效性

为验证FedCAR的有效性，本文将其集成至FLAME、XMAM、LFighter和DPFLA四种异常检测方法中，即在这些方法完成异常与正常客户端模型的识别后，采用FedCAR对异常客户端模型中未受污染的有效知识进行重用。表6展示了引入FedCAR前后四种基于异常检测的基线方法对各类别（类别0到类别9）样本及所有样本的全局模型测试准确率。

实验结果表明，FedCAR能够有效提升四种异常检测方法的性能。引入FedCAR后，FLAME、XMAM、LFighter和DPFLA四种方法的全局模型测试准确率分别提升了0.84%、0.78%、0.79%和0.87%，验证了FedCAR在重用异常客户端模型参数方面的有效性。进一步分析各类别样本的测试准确率发现，FedCAR对类别2的准确率提升较小，与引入前相近，而对其他类别样本均带来了不同程度的性能提升。其中，对类别5的提升最为显著，四种方法的测试准确率均从约91%提升至约94%。其余类别的平均提升幅度约1%。该结果说明，FedCAR不仅有效避免了攻击信息的干扰，还能挖掘并重用异常客户端模型中保留的有效知识。

5.2.6 参与方数量对全局模型性能的影响

为了验证参与方数量对全局模型性能的影响，本文以MNIST为训练集，Fashion MNIST为辅助数据集，评估了十种算法在不同参与方占比$P_{in}=n/N=\{3\%, 5\%, 10\%, 30\%, 50\%, 70\%, 90\%\}$下的全局模型测试准确率，实验结果如表7所示。实验中，恶意客户端占比固定为30%，服务器每轮从包含30个恶意客户端和70个良性客户端的集合中随机选取$100\cdot P_{in}$个参与训练。实验结果表明，FedCAR在不同的参与方数量设置下均表现出色，性能稳定。

在参与方数量较少的情况下，即$P_{in}\leqslant 10\%$时，多数基线方法的防御效果接近无防御的FedAvg方

表6　引入FedCAR前后四种基于异常检测的基线方法对各类别样本及所有样本的全局模型测试准确率

样本	FLAME		XMAM		LFighter		DPFLA	
类别	前	后	前	后	前	后	前	后
0	97.45	**98.06**	97.96	**98.06**	97.45	**98.06**	97.86	**98.06**
1	98.06	**98.59**	98.33	**98.59**	98.33	**98.59**	98.24	**98.59**
2	**92.25**	91.76	**92.25**	92.15	**92.35**	92.05	**92.15**	92.05
3	92.87	**93.17**	93.07	**94.06**	92.57	**92.97**	92.48	**93.07**
4	93.58	**94.50**	93.18	**94.70**	93.28	**94.70**	93.28	**94.50**
5	91.03	**93.83**	91.48	**93.83**	91.37	**94.06**	92.15	**94.96**
6	95.20	**95.93**	94.99	**95.93**	94.99	**95.93**	94.99	**95.93**
7	91.73	**92.70**	91.93	**92.70**	92.02	**93.19**	91.73	**93.09**
8	89.84	**90.45**	90.35	**90.35**	90.04	**90.66**	90.45	**91.38**
9	89.00	**90.19**	89.49	**89.99**	89.99	**89.99**	89.89	**90.58**
所有	93.26	**94.10**	93.49	**94.27**	93.44	**94.23**	93.38	**94.25**

表7　参与方数量对全局模型性能的影响

P_{in}	3%	5%	10%	30%	50%	70%	90%
FedAvg	92.79	93.23	86.94	91.88	91.91	92.54	92.74
Krum	92.44	84.05	93.19	92.84	92.36	92.41	90.50
TMean	92.77	93.23	86.05	92.22	92.16	92.46	92.57
Median	92.41	92.41	91.52	93.36	93.43	93.36	93.37
FLTrust	91.25	91.68	92.48	93.03	93.37	93.31	93.38
FLAME	92.22	92.48	91.39	93.90	93.26	93.25	93.34
XMAM	92.79	93.23	93.33	92.70	93.49	93.33	93.41
LFighter	92.43	92.68	93.42	93.40	93.44	93.49	93.43
DPFLA	92.79	93.13	93.52	93.42	93.38	93.41	93.44
FedCAR(Ours)	**93.81**	**93.97**	**93.96**	**94.34**	**94.28**	**94.20**	**94.26**

法，而FedCAR的准确率始终维持在93%以上。TMean与FedAvg表现接近，在$P_{in}=10\%$时性能均最差，全局模型测试准确率分别为86.05%和86.94%。Krum在$P_{in}=5\%$设置下防御几乎失效，测试准确率较FedAvg下降了9.18%。Median和FLTrust表现相对稳定，测试准确率维持在91%以上。而异常检测方法的健壮性整体优于健壮聚合方法，即使在低参与率条件下，仍能维持约92%的全局模型准确率。其中，XMAM和DPFLA在$P_{in}=5\%$时的准确率均突破93%，分别为93.23%和93.13%。相较之下，FedCAR在所有设置下均取得最高性能，即使$P_{in}=3\%$的设置下也能达到93.81%的测试准确率，显著优于所有基线方法。这主要是因为在参与方数量较少的情况下，客户端数据的多样性受限，使得全局模型对整体数据分布的表征能力不足。而现有防御方法中直接舍弃异常客户端模型参数的做法，进一步加剧了有效知识的损失，导致全局模型性能下降。相比之下，本文方法能够最大限度地挖掘并利用异常客户端模型中的有效知识。

在参与方数量较多的情况下，即$P_{in}>10\%$时，FedAvg对参与方数量变化较为敏感，随着客户端数量的增加，全局模型的测试准确率由$P_{in}=30\%$时的91.88%上升至$P_{in}=90\%$时的92.74%。TMean方法的测试准确率与FedAvg相近，表明其在聚合过程中仍引入了被污染的模型参数，防御效果有限。Krum方法的表现相对稳定，在多数实验中优于FedAvg，全局模型测试准确率在$P_{in}=30\%$的情况下最优。Median和FLTrust的表现接近，准确率维持在约93.30%。FLAME方法仅在$P_{in}=30\%$的设置中略优于Median，提升幅度为0.54%，其余设置下均未超过Median。XMAM方法在大多数设置下优于FLAME，尤其在$P_{in}=50\%$时全局模型测试准确率可达93.49%。DPFLA方法$P_{in}=90\%$设置下表现最优，准确率较FedAvg提升了0.7%。LFighter方法整体性能更为优越，测试准确率始终高于93.40%，并在多数设置中优于其他基线方法。相比之下，FedCAR的全局模型的测试准确率维持在94%以上，表现出更强的健壮性，虽然在$P_{in}=70\%$时准确率较低，但仍较其他异常检测方法高出0.71%到0.95%。

综上所述，FedCAR能有效应对参与方数量波动带来的挑战，具备更强的攻击防御能力。

5.2.7　恶意客户端数量对全局模型性能的影响

为了验证恶意客户端数量对全局模型性能的影响，本文以MNIST为训练集，Fashion MNIST为辅助数据集，评估了十种算法在不同恶意客户端占比$P_{atk}=\{3\%, 5\%, 10\%, 20\%, 30\%, 40\%, 50\%\}$下的全局模型测试准确率，实验结果如表8所示。实验中，共包含100个客户端，其中$100\cdot P_{atk}$个被设置为恶意，其余$100\cdot(1-P_{atk})$个为良性客户端。每轮训练中，服务器从全部客户端中随机选取50个参与本轮模型更新。

实验结果表明，当恶意客户端比例较低时，所有基线方法的全局模型测试准确率均维持在92%以上；然而，随着恶意客户端比例升至50%，大多数基线方法性能显著下降。相比之下，FedCAR在不同攻击强度下始终保持更高且稳定的测试准确率，显著优于其他方法。

在恶意客户端数量较少的情况下，即$P_{atk}\leqslant 10\%$时，FedAvg、TMean和Median方法在不同设

表8 恶意客户端数量对全局模型性能的影响

P_{atk}	3%	5%	10%	20%	30%	40%	50%
FedAvg	93.43	93.42	93.39	93.18	91.91	92.47	89.82
Krum	92.25	92.05	92.70	92.57	92.36	92.66	91.44
TMean	93.43	93.42	93.41	93.18	92.16	92.24	89.10
Median	93.40	93.46	93.46	93.47	93.43	93.11	92.00
FLTrust	92.78	92.65	92.13	93.19	93.37	93.24	93.15
FLAME	93.27	93.28	93.31	93.31	93.26	93.41	84.57
XMAM	93.45	93.44	93.41	93.46	93.49	93.42	92.45
LFighter	93.46	93.45	93.47	93.51	93.44	93.46	84.16
DPFLA	93.37	93.32	93.37	93.32	93.38	93.34	92.54
FedCAR(Ours)	**94.28**	**94.23**	**94.27**	**94.37**	**94.28**	**94.23**	**93.63**

置下的测试准确率均维持在93.40%左右，显示出一定的抗攻击能力。Krum和FLTrust的准确率约为92.50%，整体低于FedAvg，可能引入了少量受污染参数。FLAME和DPFLA方法性能相对稳定，准确率均保持在93.3%。XMAM和LFighter的整体性能更优，准确率始终接近或高于93.4%，在多数设置中展现出更强的防御能力。相比之下，本文提出的FedCAR方法在所有恶意客户端占比下的全局模型准确率均在94.2%以上，较其他方法平均提升约0.8%。

在恶意客户端数量较多的情况下，即$P_{atk}>10\%$时，FedAvg和TMean方法在$P_{atk}=20\%$时测试准确率均为93.18%，但随着恶意客户端数量增加，二者准确率呈现出显著的下降趋势，在$P_{atk}=50\%$时分别降至89.82%和89.10%，表明这两种方法的健壮性较差，易受恶意客户端数量的影响。Krum和Median方法整体表现较为稳定，测试准确率分别维持在约92%和93%。FLTrust的性能与Median接近，其在$P_{atk}=30\%$时表现最优，测试准确率达93.37%。FLAME、XMAM和LFighter三种异常检测方法整体性能接近，其中FLAME的性能在三种方法中略低，测试准确率在93.30%左右；XMAM在$P_{atk}=50\%$时仍能维持92.45%的准确率，是三者中唯一在强攻击下仍能保持稳定性能的方法；而LFighter方法在$P_{atk}<50\%$的多数实验中优于前两者，测试准确率最高达93.51%。尽管DPFLA相较于前三种方法整体性能略低，但在$P_{atk}=50\%$时依然保持了92.54%的准确率。相比之下，FedCAR方法表现出最为稳定和健壮的防御效果，即使在$P_{atk}=50\%$的极端条件下，仍能达到93.63%的测试准确率，其余所有设置中准确率均高于94.20%。

综上所述，本方法在面对不同规模的标签翻转攻击时，显著优于现有防御方法，能够在恶意客户端占比较高的条件下仍保持全局模型性能的稳定。

5.2.8 辅助数据集规模对全局模型性能的影响

为评估辅助数据集规模对全局模型性能的影响，本文以MNIST为训练集，Fashion MNIST为辅助数据集，设计了三组实验，分别将辅助数据集的大小$|D_{aux}|$设置为100、300、600。图9对比了FedCAR和FLTrust在不同辅助数据集规模下的测试准确率。

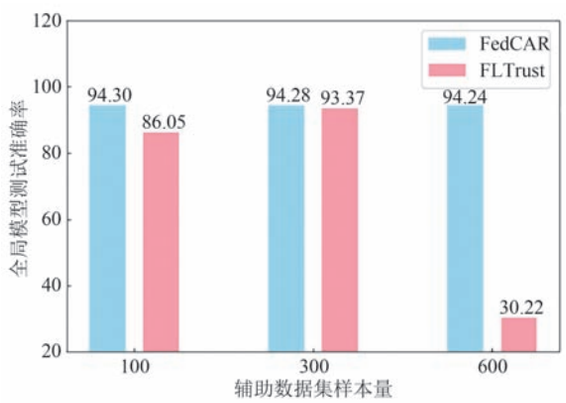


图9 辅助数据集大小对全局模型性能的影响

实验结果显示，FLTrust的性能波动较为显著，当辅助数据集包含300个样本时，全局模型测试准确率最高可达93.37%；然而，随着样本量增加至600，其准确率骤降至30.22%，最大波动幅度高达63.15%，反映出该方法对辅助数据规模极为敏感。相比之下，FedCAR在辅助数据集包含100个样本时表现最佳，全局模型的测试准确率达94.30%；当样本数量增加至600时，准确率小幅下降了0.06%。这一现象主要源于辅助数据集（Fashion MNIST）与原始训练数据（MNIST）之间的分布差异，过多使用辅助数据集进行训练可能导致全局模型分布偏移，从而干扰其对目标任务的学习。尽管如此，FedCAR在各辅助数据规模下的准确率始终保持在94%以上，表现出良好的稳定性。

5.2.9 异构数据分布下全局模型的收敛性评估

为验证所提方法在数据异构场景下的健壮性，本文采用基于狄利克雷分布（Dirichlet Distribution）的样本划分策略构造非独立同分布（non-IID）数据集，并在此基础上评估各方法的收敛性。对于每个样本类别，从参数为α的迪利克雷分布中采样一个长度为N的概率向量$p=\{p_i\}=\mathrm{Dirichlet}(\alpha)$，并按照分配比例$p_i$从原始数据集中抽取对应数量的样本

给客户端C_i。其中，超参数α控制了客户端之间数据的异构程度，α越小，类别分布越不均衡，异构性越强，本文实验中设置$\alpha=0.1$。表9展示了各方法在MNIST和CIFAR10数据集上的测试准确率。

表9　对比各方法在数据异构场景下的测试准确率

方法	MNIST	CIFAR10
FedAvg	93.11	48.13
Krum	80.07	29.85
TMean	93.11	48.09
Median	89.32	34.42
FLTrust	92.49	22.58
FLAME	92.09	46.75
XMAM	92.87	48.04
LFighter	91.85	47.66
DPFLA	93.31	44.39
FedCAR	**93.40**	**48.48**

实验结果表明，在数据异构场景中，多数方法的性能较同构场景出现不同程度的下降，而本文提出的FedCAR仍实现了最优性能。在无防御设置下，FedAvg在MNIST和CIFAR10两个任务中分别达到了93.11%和48.13%的准确率，具备一定健壮性。TMean的表现与FedAvg相近，而Krum与Median的性能下降显著，其中Krum在CIFAR10上的准确率仅为29.85%，反映其在non-IID环境中无法有效辨别异常更新。FLTrust在CIFAR10上表现最差(22.58%)，表明其基于可信样本统计的聚合机制在复杂分布下泛化能力有限。

异常检测方法的性能整体优于健壮聚合方法，主要原因在于后者依赖于异常客户端更新显著偏离多数客户端的假设。然而，在non-IID场景下，客户端更新之间的差距本身较大，易导致误判。相比之下，异常检测方法从客户端行为特征出发，实现了更精确的检测。在各类异常检测方法中，FLAME、XMAM与LFighter展现出更好的适应性。XMAM在两个任务上性能接近FedAvg，LFighter在CIFAR10上达到了47.66%，显示其较强的健壮性。DPFLA在MNIST上较FedAvg提升了0.2%，但在CIFAR10上准确率下降至44.39%，表明该方法在处理复杂数据时存在性能瓶颈。FedCAR在两个数据集上均取得最高准确率，分别为93.40%和48.48%，验证了其利用污染度指标与加权知识蒸馏机制对异常客户端中有效知识的保留与重用能力。综上，FedCAR能够在不完全舍弃异常更新的前提下实现更优的识别性能和更稳健的全局模型收敛。

6　总　结

联邦学习分布式的特性虽然有效解决了数据孤岛问题，但也带来了多种安全挑战，标签翻转攻击便是其中之一。针对现有攻击防御策略中参数利用不足的问题，提出了标签翻转攻击污染度感知的异常客户端模型参数重用算法FedCAR，该方法通过对比异常全局模型与基础全局模型在输出层参数上的差异，量化攻击对各类样本的污染度，并据此调节基础全局模型在知识蒸馏中对异常客户端模型中各类知识的学习率。实验结果表明，FedCAR在MNIST和CIFAR10数据集上的全局模型测试准确率较九个基线模型平均提升了1.81%，有效挖掘并重用了异常客户端模型中的有效知识。未来研究将聚焦于提升异常客户端识别算法的准确性。

参　考　文　献

[1] 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. Fort Lauderdale, USA, 2017: 1273-1282

[2] Zhao Z, Mao Y, Liu Y, et al. Towards efficient communications in federated learning: A contemporary survey. Journal of the Franklin Institute, 2023, 360(12): 8669-8703

[3] Long G, Tan Y, Jiang J, et al. Federated learning for open banking. Federated Learning: Privacy and Incentive, 2020: 240-254

[4] Nguyen D C, Pham Q V, Pathirana P N, et al. Federated learning for smart healthcare: A survey. ACM Computing Surveys, 2022, 55(3): 1-37

[5] Zhang K, Song X, Zhang C, et al. Challenges and future directions of secure federated learning: a survey. Frontiers of Computer Science, 2021, 16(5): 165817

[6] Tian Y, Zhang W, Simpson A, et al. Defending against data poisoning attacks: From distributed learning to federated learning. The Computer Journal, 2021, 66(3): 711-726

[7] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines//Proceedings of the 29th International Coference on International Conference on Machine Learning. Edinburgh, Scotland, 2012: 1467-1474

[8] Steinhardt J, Koh P W, Liang P. Certified defenses for data poisoning attacks//Proceedings of the 31st International Conference on Neural Information Processing Systems. California, USA, 2017: 3520-3532

[9] Shejwalkar V, Houmansadr A, Kairouz P, et al. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning//Proceedings of 2022 IEEE

Symposium on Security and Privacy. San Francisco, USA, 2022: 1354-1371

[10] Yu S, Shen J, Xu S, et al. Label-flipping attacks in GNN-based federated learning. IEEE Transactions on Network Science and Engineering, 2025, 12(2): 1357-1368

[11] Cao D, Chang S, Lin Z, et al. Understanding distributed poisoning attack in federated learning//Proceedings of 2019 IEEE 25th International Conference on Pparallel and Distributed Systems. Tianjin, China, 2019: 233-239

[12] Blanchard P, El Mhamdi E M, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent// Proceedings of the 31st International Conference on Neural Information Processing Systems. California, USA, 2017: 118-128

[13] Yin D, Chen Y, Kannan R, et al. Byzantine-robust distributed learning: Towards optimal statistical rates// Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 5650-5659

[14] Shalom O, Leshem A, Bajwa W U. Mitigating data injection attacks on federated learning// Proceedings of the 49th IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, Republic of Korea, 2024: 9116-9120

[15] Guo J, Liu Z, Tian S, et al. TFL-DT: A trust evaluation scheme for federated learning in digital twin for mobile networks. IEEE Journal on Selected Areas in Communications, 2023, 41(11): 3548-3560

[16] Tolpegin V, Truex S, Gursoy M E, et al. Data poisoning attacks against federated learning systems// Proceedings of the 25th European Symposium on Research in Computer Security. Guildford, UK, 2020: 480-501

[17] Zang L, Li Y. Detection and mitigation of label-flipping attacks in fl systems with kl divergence. IEEE Internet of Things Journal, 2024, 11(19): 32221-32233

[18] Sameera K M, Vinod P, KA R R, et al. Lfgurad: A defense against label flipping attack in federated learning for vehicular network. Computer Networks, 2024, 254: 110768

[19] Yang R, He H, Wang Y, et al. Dependable federated learning for IoT intrusion detection against poisoning attacks. Computers & Security, 2023, 132: 103381

[20] Zhang J, Zhang F, Jin Q, et al. XMAM:X-raying models with a matrix to reveal backdoor attacks for federated learning. Digital Communications and Networks, 2024, 10(4): 1154-1167

[21] Jiang Y, Zhang W, Chen Y. Data quality detection mechanism against label flipping attacks in federated learning. IEEE Transactions on Information Forensics and Security, 2023, 18: 1625-1637

[22] Thein T T, Shiraishi Y, Morii M. Personalized federated learning-based intrusion detection system: Poisoning attack and defense. Future Generation Computer Systems, 2024, 153: 182-192

[23] Awan S, Luo B, Li F. Contra: Defending against poisoning attacks in federated learning//Proceedings of the 26th European Symposium on Research in Computer Security. Darmstadt, Germany, 2021: 455-475

[24] Jebreel N M, Domingo-Ferrer J, Sánchez D, et al. LFighter: Defending against the label-flipping attack in federated learning. Neural Networks, 2024, 170: 111-126

[25] Chen G, Li K, Abdelmoniem A M, et al. Exploring representational similarity analysis to protect federated learning from data poisoning//Proceedings of the ACM on Web Conference 2024. Singapore, 2024: 525-528

[26] Nguyen T D, Rieger P, Viti R D, et al. FLAME: Taming backdoors in federated learning//Proceedings of the 31st USENIX Security Symposium. Boston, USA, 2022: 1415-1432

[27] Van Erven T, Harremos P. Rényi divergence and kullback-leibler divergence. IEEE Transactions on Information Theory, 2014, 60(7): 3797-3820

[28] Gao Ying, Chen Xiao-Feng, Zhang Yi-Yu, et al. A survey of attack and defense techniques for federated learning systems. Chinese Journal of Computers, 2023, 46(09): 1781-1805
(高莹, 陈晓峰, 张一余, 等. 联邦学习系统攻击与防御技术研究综述. 计算机学报, 2023, 46(09): 1781-1805)

[29] Zhang J, Chen J, Wu D, et al. Poisoning attack in federated learning using generative adversarial nets//Proceedings of the 13th IEEE International Conference On Big Data Science And Engineering. Rotorua, New Zealand, 2019: 374-380

[30] Liu Jia-Lang, Guo Yan-Ming, Lao Ming-Rui, et al. Survey of backdoor attack and defense algorithms based on federated learning. Journal of Computer Research and Development, 2024, 61(10): 2607-2626
(刘嘉浪, 郭延明, 老明瑞, 等. 基于联邦学习的后门攻击与防御算法综述. 计算机研究与发展, 2024, 61(10): 2607-2626)

[31] Jiang Wei-Jin, Yang Xuan, Li Bi-Xia. Federated learning poisoning attack defense method based on interpretable contribution anomaly detection and dynamic pruning. Chinese Journal of Computers, 2025, 48(12):2855-2874
(蒋伟进, 杨璇, 李碧霞. 基于可解释贡献异常检测与动态裁剪的联邦学习投毒攻击防御方法. 计算机学报: 2025, 48(12):2855-2874)

[32] Gao Qi, Sun Yi, Gai Xin-Mao, et al. PRFL: Privacy-preserving robust aggregation method for federated learning. Computer Science, 2024, 51(11): 356-367
(高琦, 孙奕, 盖新貌, 等. PRFL:一种隐私保护联邦学习鲁棒聚合方法. 计算机科学, 2024, 51(11): 356-367)

[33] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning//Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. Virtural, 2020: 2938-2948

[34] Chen W, Zhang Z, Hu X, et al. Boosting decision-based black-box adversarial attacks with random sign flip//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK , 2020: 276-293

[35] Fang M, Cao X, Jia J, et al. Local model poisoning attacks to byzantine-robust federated learning//Proceedings of the 29th USENIX Security Symposium. Virtural, 2020: 1605-1622

[36] Wang H, Sreenivasan K, Rajput S, et al. Attack of the tails: Yes, you really can backdoor federated learning//Proceedings of the 34th International Conference on Neural Information

Processing Systems. Vancouver, Canada, 2020: 16070-16084
[37] Xie C, Huang K, Chen P-Y, et al. Dba: Distributed backdoor attacks against federated learning//Proceedings of 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-19
[38] Gong X, Chen Y, Huang H, et al. Coordinated backdoor attacks against federated learning with model-dependent triggers. IEEE Network, 2022, 36(1): 84-90
[39] Li D, Wong W E, Wang W, et al. Detection and mitigation of label-flipping attacks in federated learning systems with KPCA and K-means//Proceedings of the 8th International Conference on Dependable Systems and Their Applications. Yinchuan, China, 2021: 551-559
[40] Cao X, Fang M, Liu J, et al. FLTrust: Byzantine-robust federated learning via trust bootstrapping//Proceedings of Network and Distributed System Security Symposium. Virtural, 2021: 1-15
[41] Campello R J G B, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates//Proceedings of Advances in Knowledge Discovery and Data Mining. Berlin, Germany, 2013: 160-172
[42] Feng X, Cheng W, Cao C, et al. Dpfla: Defending private federated learning against poisoning attacks. IEEE Transactions on Services Computing, 2024, 17(4): 1480-1491
[43] Deng Y, Ren J, Tang C, et al. A hierarchical knowledge transfer framework for heterogeneous federated learning//Proceedings of IEEE Conference on Computer Communications. New York, USA , 2023: 1-10
[44] Song D, Xu J, Pang J, et al. Classifier-adaptation knowledge distillation framework for relation extraction and event detection with imbalanced data. Information Sciences, 2021, 573: 222-238
[45] Jeong E, Oh S, Kim H, et al. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. arXiv preprint arXiv: 1811.11479, 2018
[46] Itahara S, Nishio T, Koda Y, et al. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-iid private data. IEEE Transactions on Mobile Computing, 2021, 22(1): 191-205
[47] Hu L, Yan H, Li L, et al. MHAT: An efficient model-heterogenous aggregation training scheme for federated learning. Information Sciences, 2021, 560: 493-503
[48] Gong X, Sharma A, Karanam S, et al. Ensemble attention distillation for privacy-preserving federated learning// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 15076-15086
[49] Zhang L, Wu D, Yuan X. Fedzkt: Zero-shot knowledge transfer towards resource-constrained federated learning with heterogeneous on-device models//Proceedings of IEEE 42nd International Conference on Distributed Computing Systems. Bologna, Italy, 2022: 928-938
[50] Wang H, Li Y, Xu W, et al. Dafkd: Domain-aware federated knowledge distillation//Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 20412-20421
[51] Lee G, Jeong M, Shin Y, et al. Preservation of the global knowledge by not-true distillation in federated learning. Advances in Neural Information Processing Systems, 2022, 35: 38461-38474
[52] Zhu Z, Hong J, Zhou J. Data-free knowledge distillation for heterogeneous federated learning//Proceedings of the 38th International Conference on Machine Learning. Virtural, 2021: 12878-12889
[53] Khan M A, Chandio Y, Anwar F. HYDRA-FL: Hybrid knowledge distillation for robust and accurate federated learning. Advances in Neural Information Processing Systems, 2024, 37: 50469-50493
[54] Liu Z, He W, Chang C-H, et al. SPFL: A Self-purified federated learning method against poisoning attacks. IEEE Transactions on Information Forensics and Security, 2024, 19: 6604-6619
[55] Yueqi X I E, Fang M, Gong N Z. Fedredefense: Defending against model poisoning attacks for federated learning using model update reconstruction error//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria, 2024: 54460-54474
[56] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition//Proceedings of the IEEE. Pasadena, USA, 1998: 2278-2324
[57] Krizhevsky A, Hinton G. Learning Multiple Layers of Features from Tiny Images. 2009
[58] Xiao H, Rasul K, Vollgraf R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017

LUO Jia-Yuan, Ph. D. candidate. Her research interests include federated learning and privacy protection.

TAN Zhen-Hua, Ph. D., professor. His research interests include cyberspace security, social network analysis, verification and identification technologies.

NING Jing-Yu, Ph. D.. Her research interests include privacy protection and multi-party computation.

JIA Zhi-Liang, M. S. . His research interests include federated learning and privacy protection.

Background

The decentralized nature of federated learning (FL) makes it particularly vulnerable to poisoning attacks, among which label flipping attacks are one of the most prevalent threats. In such attacks, compromised clients modify the labels of source-class samples to a predefined target class while keeping other data unchanged. This manipulation maintains the recognition accuracy for non-source-class samples but misleads the global model into misclassifying source-class samples as the target class. With their low implementation cost, high stealthiness, and destructive impact, label flipping attacks pose a significant threat to the security of FL.

Existing defense methods can be classified into robust aggregation and anomaly detection. Robust aggregation methods apply statistical techniques to mitigate the impact of abnormal models, while anomaly detection methods employ clustering algorithms to identify and remove them. Although these approaches are effective in defending against label flipping attacks, they generally adopt an "identify-and-discard" strategy, in which abnormal model parameters are excluded during aggregation. While this helps filter poisoned knowledge, it has limitations under label flipping attack scenarios. Specifically, malicious models exhibit poor accuracy on source-class samples but still perform well on non-source-class samples. Discarding these models not only wastes client resources but also leads to the loss of valuable knowledge, thereby hindering the global model's generalization. Therefore, there is a pressing need for a defense mechanism capable of mining and reutilizing useful knowledge from abnormal models.

To minimize global performance loss, we propose FedCAR, a federated contamination-aware reutilization algorithm for compromised client model parameters against label flipping attacks. The contamination level for each class is quantified by comparing the output-layer neuron parameters of the compromised global model with those of the basic global model, which are constructed by aggregating client models from the two clusters obtained via K-means clustering. Then, a weighted knowledge distillation loss function is proposed, where clipped compromised local models serve as teachers and the basic global model acts as the student. The contamination level is used to modulate the student's learning rate for each class-specific knowledge transfer from teachers. Experimental results demonstrate the effectiveness of our methods in extracting and reutilizing valuable parameter knowledge.

This work was supported by the National Key Research and Development Program of China under Grant No. 2023YFC3306201.