可检测客户端异常更新的联邦学习分组方法

黄大广[1],雷 诚[2]

(1. 中电鸿信信息科技有限公司,江苏 南京 210000;2. 南京希音电子商务有限公司, 江苏 南京 210012)

摘 要: 鉴于联邦学习固有的分布式特性,其数据分散存储于多个参与方且各参与方独立训练模型,缺乏集中式的全局数据管控与模型统一校验机制,使其在面对中毒攻击时表现出高度的脆弱性。现有研究主要集中在单一全局模型架构下拦截恶意客户端的更新上,而一旦拦截失败,将会直接导致全局模型中毒。为此,提出一种可检测客户端异常更新的联邦学习分组方法dFL。首先,通过分组思想将客户端分成多个小组,每个小组训练一个小组模型,即使存在恶意客户端,其影响范围也仅限于小组模型;其次,在组内训练中采用异常更新检测机制,结合客户端信誉值进行权重聚合,进一步降低恶意客户端对小组模型的影响;最后,由小组模型投票表决出全局模型的最终结果。在MNIST-0.5和CIFAR10数据集上,dFL比对照算法表现出更强的稳定性。

关键词: 联邦学习;联邦分组学习;权重聚合;异常检测

DOI: 10. 11907/rjdk. 242056 扫描二维码阅读全文:

中图分类号: TP309 **文献标识码:** A **文章编号:** 1672-7800(2026)007-0040-07

Federated Learning Grouping Method Based on Abnormal Client Updates Detection

HUANG Daguang[1], LEI Cheng[2]

(1. *China Telecom Hongxin Information Technology Co. , Ltd*, *Nanjing* 210000, *China*;
2. *Nanjing Shein E-commerce Co. , Ltd*, *Nanjing* 210012, *China*)

Abstract: In view of the inherent distributed characteristics of federated learning, its data is stored in multiple participants and each participant independently trains the model. The lack of centralized global data control and unified model verification mechanism makes it highly vulnerable to poisoning attacks. Existing research mainly focuses on intercepting updates from malicious clients under the single global model architecture, and once the interception fails, it will directly lead to poisoning of the global model. Therefore, a federated learning grouping method dFL is proposed to detect abnormal updates on the client side. First of all, the client is divided into multiple groups by grouping idea, and each group trains a group model. Even if there are malicious clients, the scope of influence is limited to the group model; Secondly, in the intra group training, the abnormal update detection mechanism is used to combine the client reputation value for weight aggregation to further reduce the impact of malicious clients on the group model; Finally, the final result of the global model is voted by the group model. On MNIST-0.5 and CIFAR10 datasets, dFL shows stronger stability than the control algorithm.

Key Words: federated learning; federated group learning; weight aggregation; anomaly detection

0 引言

为了应对传统机器学习中的数据孤岛问题,McMahan等[1]于2017年提出联邦学习的概念,其允许多个客户端基于本地数据集协同训练共享全局模型,服务器与客户端之间只会进行模型参数的传递而非原始数据的传递。因此,联邦学习不仅大大节省了通信开销,而且极大程度上保护了客户端的数据隐私。然而,鉴于联邦学习的特性,服务器端并不了解客户端的训练过程,甚至会出现恶意客户端[2]。因此,研究联邦学习的中毒机制对于提高其可靠性具有重要意义。

收稿日期: 2025-06-04

基金项目: 国家自然科学基金项目(61872196)

作者简介: 黄大广(1979-),男,硕士,中电鸿信信息科技有限公司工程师,研究方向为车联网、联邦学习、位置服务等;雷诚(1999-),男,硕士,南京希音电子商务有限公司工程师,研究方向为联邦学习、信任模型、隐私保护等。

1　相关工作

中毒攻击是联邦学习面临的最直接、破坏性最强的威胁之一，该类攻击旨在破坏模型的完整性，通过篡改训练数据或模型更新来影响全局模型的性能。为此，Feng 等[3]提出一种新型的多输入功能代理重加密方案 MI-FPRE，并基于此设计了一种具有强隐私性的联邦学习框架。Chen 等[4]提出一种基于注意力机制的客户端选择算法 FEDACS，旨在解决个性化联邦学习中非独立同分布（non-IID）数据带来的挑战，通过动态选择数据分布相似的客户端参与训练，提升个性化模型的性能。Shi 等[5]提出一种有效的模型选择性聚合方法 SAM 来减轻联邦学习系统中的通信过载情况。然而，上述研究均未考虑客户端集合中存在恶意客户端的情况。恶意客户端是指那些在联邦学习过程中，故意违反协议规则、篡改数据或模型更新，以破坏系统正常运行或窃取敏感信息的客户端，对联邦学习的数据隐私保护、模型性能以及系统稳定性构成了严重威胁。Bhagoji 等[6]发现通过控制少量恶意代理可导致全局模型对一组选定的输入进行错误分类，表明有效和隐蔽的模型中毒攻击是有可能的。Cao 等[7]将虚假客户端注入联邦学习系统中，表明即使采用经典防御手段，依旧会显著降低全局模型的测试精度。Fu 等[8]研究了针对垂直联邦学习的标签推断攻击，揭示了攻击者如何利用梯度信息或已训练的局部模型来推断私有标签，进而侵犯数据隐私。Alsereidi 等[9]研究了标签翻转攻击和基于生成对抗网络生成的 EEG 数据攻击对联邦学习系统的影响，揭示了该系统在面对恶意攻击时的脆弱性。Tolpegin 等[10]研究了针对联邦学习系统的数据投毒攻击，揭示了即使少量恶意参与者也能显著影响全局模型性能。

近年来，许多专门应对联邦学习环境下中毒攻击的防御方法被提出。例如，Huang 等[11]提出一种以差分隐私为底层技术的可验证隐私保护联邦学习方案 VPPFL，既能防御中毒攻击，又能保护用户隐私，且计算量和通信成本均较小。Cao 等[12]提出 FLTrust 方法，服务器端预先收集一个干净的小型训练数据集，通过比较服务端的模型和客户端上传的本地模型来区分恶意本地模型。Zhang 等[13]提出 FLDetector 方法，通过检查模型更新的一致性来检测恶意客户端。刘艳等[14]提出 eFL 方法，通过比较全局模型和本地模型的相似度来检测恶意客户端，当相似度低于阈值时，将直接忽略客户端上传的本地模型参数。Xu[15]设计了一种动态参考梯度聚合算法来缓解中毒攻击，通过对不同本地上传的梯度进行聚类来动态划分每轮本地上传的子梯度。

然而，以上防御策略均基于单个全局模型的架构设计，当恶意客户端进行中毒攻击时，单个全局模型十分脆弱，一旦某个恶意客户端突破防御，便会直接导致全局模型中毒。此外，由于客户端本地数据对服务器是不可见的，服务器难以对客户端的更新质量进行判断。为此，本文提出可检测客户端异常更新的联邦学习分组方法（detected Federated Learning，dFL），利用分组思想来缩小恶意客户端的影响范围，并在组内训练中设计了防御性联邦学习算法，通过更新权重进一步压缩可能作恶客户端产生的影响。主要创新点为：①提出分组机制，将可能作恶的客户端的影响局限在小组内，并对分组机制的安全级别进行计算；③提出异常更新检测机制，依靠模型相似度进行计算，服务器不再对低质量更新束手无措；③提出客户端信誉值机制和新的聚合策略，为每个客户端计算信誉值，进一步降低了可能作恶的客户端在组内的影响。

2　联邦学习分组方法 dFL

dFL 模型架构如图 1 所示。

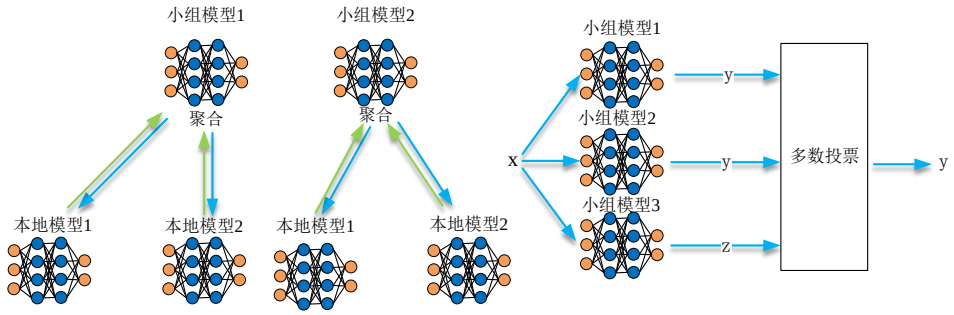


Fig. 1　dFL structure

图 1　dFL 架构

2.1　全局模型

为分散恶意客户端对全局模型的影响，本文设计了一种分组管理方法。假定存在 n 个客户端，被分成 N 个客户端小组，每个小组中包含 k 个客户端（$k < n$），将 n 个客户端确定性地分配到 N 个互不相交的小组中。当将客户端分成 N 个小组进行训练后可以得到 N 个小组模型，即 $G_1, G_2, \ldots G_n$。本文采用多数投票机制生成最终结果。假定组内训练采用传统的联邦学习算法，给定一个测试输入 x，

分别使用每个小组模型对其进行预测，结果为$f(G_u,x),u=1,2,...,N$。将$n_i(x)$表示为预测x的标签为$f(G_u,x)=i$的小组模型数量，将$p_i(x)$表示为预测x的标签为$f(G_u,x)=i$的概率，n_i为标签频率，p_i为标签概率，则$n_i(x)$、$p_i(x)$的定义分别为：

$$n_i(x)=\sum_{u\in[1,N]}D_{f(G_u,x)=i},p_i(x)=\frac{n_i(x)}{N} \tag{1}$$

式中：D为判断函数，即当$f(G_u,x)=i$时，$D_{f(G_u,x)=i}=1$，否则$D_{f(G_u,x)=i}=0$。

全局模型最终将x预测为具有最大标签频率或标签概率的类别$F(\mathrm{C},x)$，即：

$$F(\mathrm{C},x)=argmax_i(n_i(x))=argmax_i(p_i(x)) \tag{2}$$

dFL模型总是采信大多数小组模型的判断。由于每个小组模型均使用客户端的一个子集来学习，当大多数客户端为正常客户端时，大多数小组模型会使用正常客户端来学习。因此，当恶意客户端的数量有限时，多票数标签即为全局模型的结果，不会受到恶意客户端的影响。

2.2 安全级别

假设存在阈值m，当n个客户端中存在的恶意客户端数量不超过m时，dFL对测试输入x作出的预测是不变的，则称m为dFL的安全级别。在计算测试输入x的安全级别m时需要用到其最大和第二的标签频率。针对分组情况，可以确定性地将n个客户端分为N个小组进行训练，每个客户端确定性地属于唯一一个小组，可以确定性地计算最大和第二标签频率。假设C为所有正常客户端的集合，C^{all}为所有n个客户端集合，针对测试输入x的安全级别m，则有：

$$F(\mathrm{C},x)=F(\mathrm{C}^{all},x),|\mathrm{C}^{all}-\mathrm{C}|\leqslant m,m\in N \tag{3}$$

式中：N为自然数集合，即在保证全局模型不受恶意客户端影响的前提下，整个客户端集合最多存在m个恶意客户端。

给定测试输入x，假设在不存在恶意客户端的情况下，dFL预测y为x的最高频率标签，z为次高频率标签，其标签频率分别记作n_y和n_z。在存在恶意客户端的情况下，dFL依旧预测y为x的最高频率标签，z为次高频率标签，其标签频率分别记作n_y^*和n_z^*，则在满足$n_y^*(x)>n_z^*(x)$的情况下，全局模型不受来自恶意客户端的影响，即$F(\mathrm{C},x)=F(\mathrm{C}^m,x)$。假设在客户端集合$C^m$中存在$m$个恶意客户端，根据本文分组策略，它们可能会被分散到不同的客户端小组中。假定在一个小组中，只要至少存在一个恶意客户端，其小组模型便会被影响，导致预测结果由y变成z。考虑到最坏情况，即m个恶意客户端恰好被分配到m个小组中，这m个小组的预测全部由y变成z，由此可以推导出：

$$n_y^*(x)\geqslant n_y(x)-m,n_z^*(x)\leqslant n_z(x)+m \tag{4}$$

将式(4)代入不等式$n_y^*(x)>n_z^*(x)$得到：

$$m\leqslant\frac{n_y(x)-n_z(x)}{2} \tag{5}$$

综上所述，本文给出以下定理：

定理1 给定n个客户端C，将其确定性地划分成N个不相交的客户端小组。针对测试输入x，y和z分别为拥有最大和第二标签频率的预测标签，当C中存在的恶意客户端不超过m时，dFL可以确定性地将测试输入x预测为y，即：

$$F(\mathrm{C},x)=F(\mathrm{C}^m,x)=y,|\mathrm{C}^m-\mathrm{C}|\leqslant m,m\in N \tag{6}$$

式中：$m=\left\lfloor\frac{n_y(x)-n_z(x)}{2}\right\rfloor$。

2.3 针对性组内训练

分组策略可以缩小恶意客户端的影响范围，有效减少其对全局模型的影响。为准确计算出有限客户端数量的阈值，设定安全级别m，即在恶意客户端数量不超过安全级别时，可以保证全局模型不受影响。显然，更高的安全级别m意味着dFL拥有更高的防御力，可以应对更多数量恶意客户端的中毒攻击。由式(5)可知，安全级别与最高标签频率$n_y(x)$和次高标签频率$n_z(x)$的差(记作Δ)成正相关，只需增大Δ便可获得更高的安全级别。因此，dFL通过引入异常检测和客户端信誉机制来最小化恶意客户端对小组模型产生的影响，从而获得更高的安全级别m。

2.3.1 异常检测

在传统联邦学习中，当存在恶意客户端时，只需要向全局模型传递被污染过的模型参数便可以达到污染目的。因此，dFL摒弃传统聚合算法，采用异常更新检测机制来对客户端提交的模型参数进行针对性排查。常用的相似度度量算法包括马氏距离、余弦相似度、欧氏距离等，考虑到最终目的为判断两个模型之间的收敛方向是否近似，本文选用余弦相似度算法来计算小组模型与客户端模型之间的相似度。公式为：

$$sim_i^t=\cos(\boldsymbol{\omega}_{group}^{t-1},\boldsymbol{\omega}_i^t)=\frac{\boldsymbol{\omega}_{group}^{t-1}\cdot\boldsymbol{\omega}_i^t}{\left\|\boldsymbol{\omega}_{group}^{t-1}\right\|\times\left\|\boldsymbol{\omega}_i^t\right\|} \tag{7}$$

式中：$\boldsymbol{\omega}_{group}^{t-1}$表示组内训练第$t-1$轮的小组模型权重；$\boldsymbol{\omega}_i^t$表示客户端$U_i$在组内训练第$t$轮的本地模型权重；$sim_i^t$表示第$t$轮客户端$U_i$的本地模型与小组模型的相似度；$sim_i^t\in[-1,1]$，当本地模型参数更新方向与小组模型相反时，$sim_i^t$取负值。

为了更加直观地进行比较，本文对模型相似度进行归一化操作，将其映射到$[0,1]$。即：

$$sim_{i\ new}^t=\frac{sim_i^t-\cos_{\min}}{\cos_{\max}-\cos_{\min}}=\frac{sim_i^t+1}{2} \tag{8}$$

由于小组模型的训练完全来自组内客户端，其收敛方向由组内客户端决定。换言之，当组内恶意客户端不超过半数时，正常客户端与小组模型的相似度会比恶意客户端的高，并且随着训练轮次的增加，两者之间的差距会逐渐扩大。

2.3.2 客户端信誉值

为计算客户端信誉值，引入贝叶斯公式。表示为：

$$P(H|D)=\frac{P(D|H)P(H)}{P(D)} \tag{9}$$

在给定观测数据 D 的情况下，需要计算假设 H 的概率，由于假设 H 为离散型变量，根据全概率公式，则有：

$$P(D)=\sum_i P(D|H_i)P(H_i) \tag{10}$$

将式(10)代入式(9)得到：

$$P(H|D)=\frac{P(D|H)P(H)}{\sum_i P(D|H_i)P(H_i)} \tag{11}$$

假设 H 和观测数据 D 分别表示客户端 U_i 的行为假设(正常或恶意)和服务器的观测数据 D_i，则客户端信誉的计算公式为：

$$R_i=\frac{P(B_i|R_i)R_i}{\sum_i P(B_i|R_i)R_i} \tag{12}$$

为便于 R_i 的计算与更新，使用 β-分布表示客户端 U_i 的信誉，使用 gamma 函数表示为：

$$P(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1} \tag{13}$$

式中：$a>0, b>0$ 且 $x\in(0,1)$。

为准确更新客户端信誉，快速识别出恶意模型上传的参数，从而最小化有毒参数对小组模型的影响，基于相似度将每轮客户端上传的内容分为正面模型参数和负面模型参数两类。假设一个客户端 U_i 向全局模型共上传 $r+w$ 次模型参数，其中 r 次为正面模型参数，w 次为负面模型参数。所有客户端的信誉初始化服从均匀分布，即：

$$P(x)=uni(0,1)=beta(1,1) \tag{14}$$

可以得到后验分布为：

$$P(x)=\frac{Bin(r+w,r)beta(1,1)}{a+b+1}=beta(r+1,w+1) \tag{15}$$

可以看出，x 的后验分布亦服从 β-分布。因此，客户端信誉 R_i 可表示为：

$$R_i=beta(a_i+1,b_i+1) \tag{16}$$

式中：a_i 表示客户端 U_i 上传的正面模型参数次数；b_i 表示客户端 U_i 上传的负面模型参数次数。当前的客户端信誉 R_i 只是一个概率而不是一个确切的值，因此使用对 R_i 取期望值的方式来表示具体客户端信誉，即：

$$C_i^t=E[beta(a_i+1,b_i+1)]=\frac{a_i+1}{a_i+b_i+2} \tag{17}$$

根据模型之间的相似度来计算一个客户端的信誉值，依靠相似度对一个客户端的某一次更新作出判断。本文引入相似度阈值机制，每个小组的相似度阈值为本轮所有提交更新的客户端的模型相似度 $sim_1^t, sim_2^t, ..., sim_k^t$（按照升序排列）的中位数 α_t，即：

若 k 为奇数，则：

$$\alpha_t=sim_{(k+1)/2}^t \tag{18}$$

若 k 为偶数，则：

$$\alpha_t=\frac{sim_{k/2}^t+sim_{(k+1)/2}^t}{2} \tag{19}$$

将客户端模型相似度与阈值进行比较，从而判断一个客户端行为是否异常，并根据判断结果对客户端 U_i 的信誉值进行更新：

若 $sim_i^t\geqslant\alpha_t$：

$$a_i^{new}=a_i+\gamma(sim_i^t-\alpha_t) \tag{20}$$

若 $sim_i^t<\alpha_t$：

$$b_i^{new}=b_i+\delta(\alpha_t-sim_i^t) \tag{21}$$

式中：γ 和 δ 分别表示奖励系数和惩罚系数。

2.3.3　聚合

传统的联邦学习聚合算法采用平均聚合方式，公平地聚合所有客户端提交的更新，即：

$$\omega_G^t=AVG(\sum_{i\in C^{all}}\omega_i^t) \tag{22}$$

在这种情况下，恶意客户端提交的更新与正常客户端享有相同的权重，其可以更加轻松地影响小组模型。为此，本文提出一种新的聚合方法，利用客户端信誉值 C_i^t 作为权重进行聚合，即：

$$\omega_{group}^t=\sum_{i\in group}\mu_i^t\omega_i^t,\mu_i^t=\frac{C_i^t}{\sum_{i\in group}C_i^t} \tag{23}$$

聚合权重来自于客户端本身的信誉值，该值基于客户端在每轮更新中的表现计算。在恶意客户端数量有限的情况下，其更新质量会被认为比正常客户端低，导致恶意客户端的信誉值下降，聚合权重也随之降低，从而减少对小组模型的影响。

3　验证实验

3.1　实验设置

3.1.1　数据集

MNIST 为机器学习领域经典的手写数字数据集，训练集包含 60 000 个样本，测试集包含 10 000 个样本，共有 10 个类别，每幅图像均由 28 × 28 个像素点组成[16]。为模拟联邦学习环境下客户端数据的非独立同分布(Non-IID)特性，本文采用 Dirichlet 分布对 MNIST 数据集进行划分，设置 Non-IID 程度参数值为 0.5，记作 MNIST-0.5。该参数值越小，数据分布的异质性越强。

CIFAR10 为 1 个图像识别数据集，共有 10 个类别的 RGB 彩色图片，识别难度比 MNIST 更大，包含 50 000 个训练样本和 10 000 个测试样本[17]。该数据集同样设置 Non-iid 的程度为 0.5，记作 CIFAR10-0.5，以模拟非独立同分布的客户端本地数据。

3.1.2　模型架构

采用卷积神经网络模型，包括输入层、2 个卷积层、1 个池化层、2 个全连接层。采用 CIFAER10 进行训练时设置 3 个全连接层，卷积核大小为 5 × 5，激活函数为 ReLU。

3.1.3　参数设置

针对 MNIST-0.5 数据集，设置 1 000 个客户端，默认小组数量 N=200，全局训练迭代轮数为 200，客户端本地训练使用随机梯度下降法，本地迭代轮数为 5，学习率为 0.001，

batch_size 为 32。针对 CIFAR10 数据集，设置 100 个客户端，默认小组数量为 20，全局训练迭代轮数为 200，客户端本地训练使用随机梯度下降法，本地迭代轮数为 10，学习率为 0.01，batch_size 为 64。

3.1.4 中毒攻击

中毒攻击分为数据中毒攻击和模型中毒攻击，其中模型中毒攻击可以由数据中毒攻击模拟实现。dFL 采用标签翻转法生成有毒数据，首先从训练集上分离出一部分数据 $D_{poison} \in D$；其次修改训练样本 $data \in D_{poison}$ 的标签 label；最后将有毒数据分配给特定比例的客户端以模拟生成恶意客户端。

3.2 实验结果分析

以下评估 dFL 在不同数据集上对不同数量恶意客户端的表现，采用传统联邦学习[1]、eFL[14]、FedDBG[15]作为对照。

3.2.1 不同数量恶意客户端下的模型表现

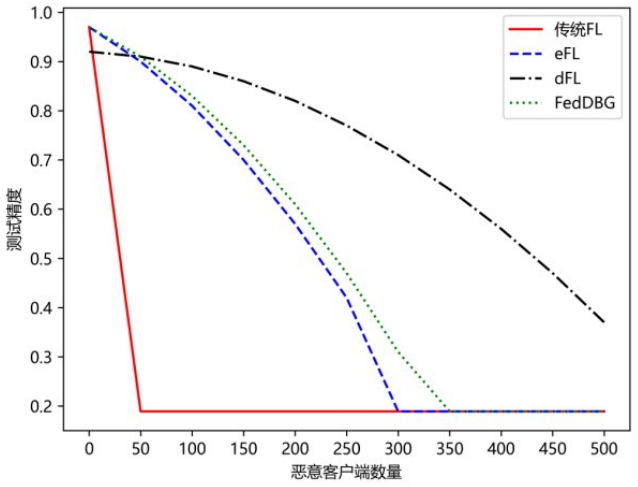
如图 2 所示，dFL 在面对不同数量的恶意客户端时表现得更加稳定，传统联邦学习的测试精度则随着恶意客户端数量的增加而快速下降。此外，当集合中不存在恶意客户端时，dFL 的测试精度低于 3 种对照方法，这是由于 dFL 对客户端进行了分组，每个客户端小组分别训练一个小组模型，然后再由小组模型投票决出最终的全局模型结果。换言之，全局模型的表现依赖于小组模型，而小组模型的表现依赖于小组中的客户端，每个小组中的客户端数量小于单一全局模型的客户端数量，这会导致每个小组模型的准确度下降。随着恶意客户端数量的增加，dFL 测试精度的下降速度比对照方法慢得多，这是由于 dFL 设置了两道防线：①分组训练多个本地模型，组内采用异常更新检测和信誉权重更新，最小化恶意客户端在组内的影响；②即使少数恶意客户端突破了第一道防线，污染了小组模型，全局模型的最终结果依旧由多数正常小组模型投票产生。

3.2.2 分组数量对模型的影响

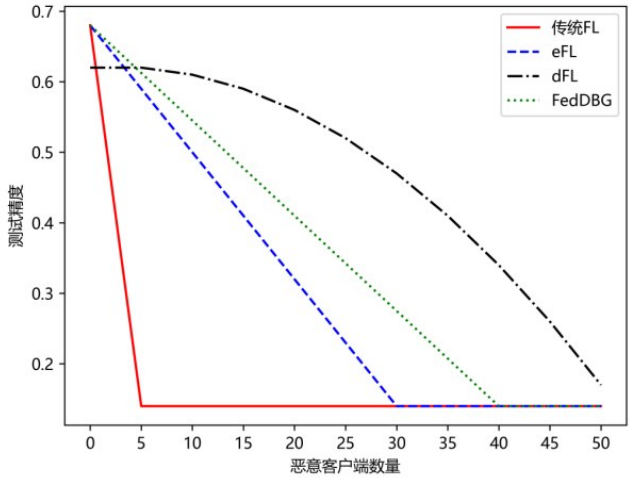
每个小组中的客户端数量小于单一全局模型的客户端数量，这会导致每个小组模型的准确度下降。因此，本文提出以下猜想：在客户端集合中不存在恶意客户端的情况下，分组数量 N 越小，全局模型的准确度越高。为验证该猜想，针对不同分组数量 N，分别在 MNIST-0.5 和 CIFAR10 数据集上进行实验，结果如图 3 所示，图中展现的趋势与猜想一致。此外，随着客户端集合中恶意客户端数量的增多，分组数量 N 越大，模型表现越稳定，测试精度越高。这是由于分组思想降低了恶意客户端对全局模型的影响，随着分组数量 N 的增大，这种影响愈加被稀释，此时 dFL 可以容忍更多恶意客户端。

3.2.3 信誉阈值对模型的影响

分别在 MNIST-0.5 和 CIFAR10 数据集上比较不同阈值（β=0.3，0.5，0.7）下 dFL 与 FedDBG 的表现，实验结果如图 4 所示。从图 4（a）可以看出，当 dFL 的信誉阈值为 0.3 时，模型的初始精度比信誉阈值为 0.5 和 0.7 时更高，因为低阈值保留了更多的客户端更新，有利于模型的初期收敛，但随着恶意客户端数量的增加，精度显著下降。当信誉阈值为 0.7 时，dFL 在高恶意客户端比例下相对稳定，但同时因为阈值过高使得正常客户端也被过滤掉，导致有效训练样本不足，影响了模型精度。当信誉阈值为 0.5 时，模型的性能与防御能力相对均衡，比信誉阈值为 0.3 和 0.7 时表现得更加稳定。从图 4（b）可以看出，所有方法在 CIFAR10 上的绝对精度均更低，这是由于该数据集更加复杂，防御难度更高。dFL 在信誉阈值为 0.5 时的表现比信誉阈值为 0.3 和 0.7 时更加优异，而 FedDBG 在恶意客户端数量>10 时精度便呈现断崖式下降。dFL 的分组机制可以稀释恶意客户端的攻击，即使信誉阈值为 0.3，其表现仍然优于 FedDBG。



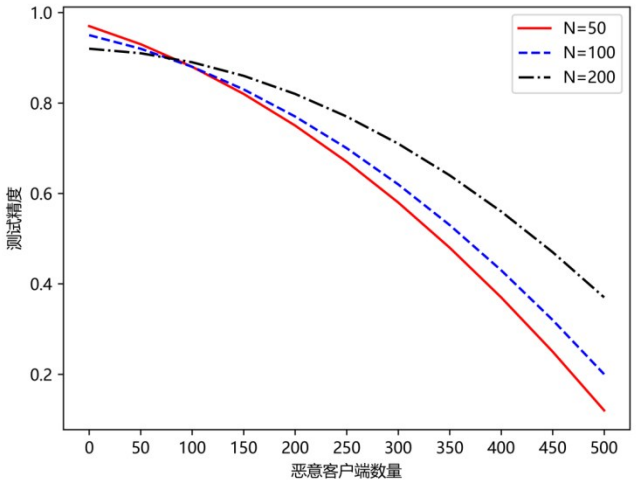
（a） MNIST-0.5



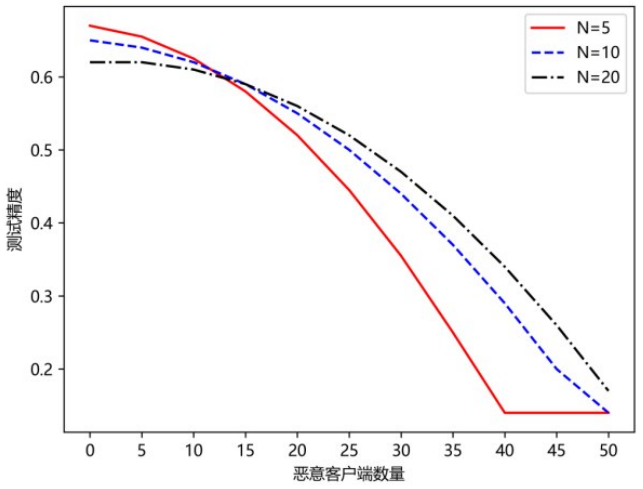
（b） CIFAR10-0.5

Fig. 2 Comparison of test accuracy of each model under different number of malicious clients

图 2 不同数量恶意客户端下各模型测试精度比较



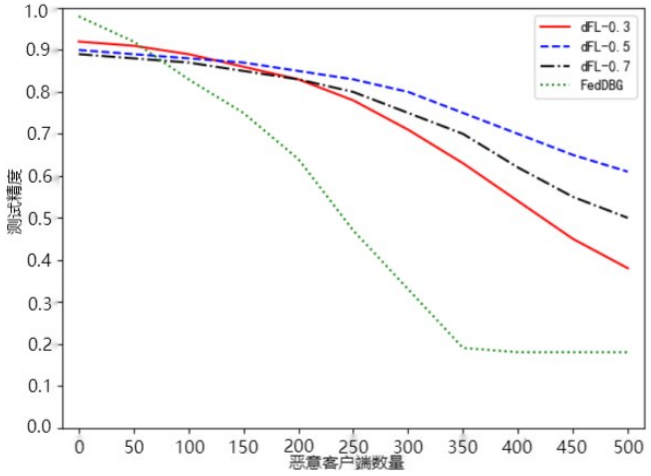
（a）　MNIST-0.5



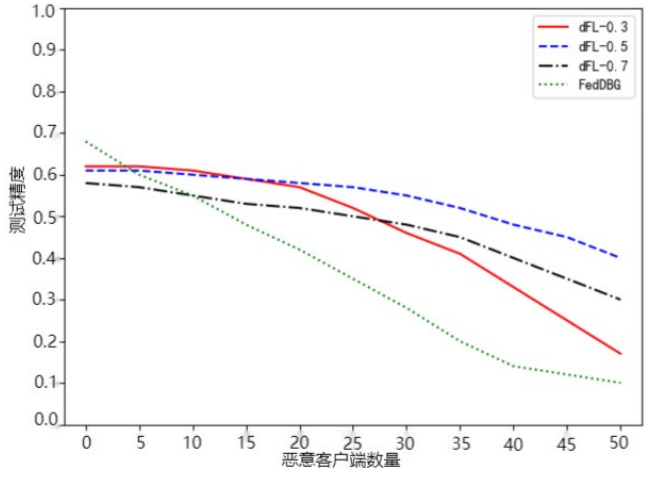
（b）　CIFAR10-0.5

Fig. 3　The impact of *N* on model

图3　分组数量*N*对模型的影响



（a）　MNIST-0.5



（b）　CIFAR10-0.5

Fig. 4　The impact of reputation value on model

图4　信誉阈值对模型的影响

4　结语

传统联邦学习凭借本地数据不共享的协作训练优势，在隐私敏感场景广泛应用，但其天然特性使其面对中毒攻击极为脆弱。为应对这一问题，本文提出一种可检测客户端异常更新的联邦学习分组方法 dFL，旨在提高联邦学习面临中毒攻击时的防御力。该方法利用分组思想稀释恶意客户端对全局模型的影响，同时结合组内异常更新检测和信誉权重聚合进一步对恶意客户端进行检测并降低其在聚合时的权重，以此减轻其对小组模型的影响，为全局模型设置了双重保险。然而，分组机制在提升防御能力的同时会降低每个小组可用的训练样本数量，导致在无恶意客户端的情况下模型精度降低，且分组数量越多，这一影响越明显。针对这一问题，后续将探索自适应分组策略和基于数据分布的智能分组方法，以在防御性能与模型精度之间实现更好的权衡[18-20]。

参考文献：

[1] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[DB/OL]. https://kim-byeong-hun. github. io/assets/pdf/9-1. pdf.

[2] ZHANG J L, CHEN B, CHENG X, et al. Poisongan: generative poisoning attacks against federated learning in edge computing systems[J]. IEEE Internet of Things Journal, 2020, 8(5): 3310-3322.

[3] FENG X Y, SHEN Q N, LI C, et al. Privacy preserving federated learning from multi-input functional proxy re-encryption[C]//2024 IEEE International Conference on Acoustics, Speech and Signal Processing, 2024: 6955-6959.

[4] CHEN Z H, LI J D, SHEN C. Personalized federated learning with attention-based client selection[C]//2024 IEEE International Conference on Acoustics, Speech and Signal Processing, 2024: 6930-6934.

[5] SHI Y C, FAN P Y, ZHU Z Q, et al. SAM: an efficient approach with selective aggregation of models in federated learning[J]. IEEE Internet of Things Journal, 2024, 11(11): 20769-20783.

[6] BHAGOJI A N, CHAKRABORTY S, MITTAL P, et al. Analyzing federated learning through an adversarial lens[C]//36th International Conference on Machine Learning, 2019: 634-643.

[7] CAO X Y, GONG N Z. Mpaf: model poisoning attacks to federated learning based on fake clients [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 3395-3403.

[8] FU C, ZHANG X, JI S, et al. Label inference attacks against vertical federated learning [C]//31st USENIX Security Symposium (USENIX Security 22), 2022: 1397-1414.

[9] ALSEREIDI M, AWADALLAH A, ALKAABI A, et al. Data poisoning against federated learning: comparative analysis under label-flipping attacks and GAN-generated EEG data [C]//2024 2nd International Conference on Cyber Resilience (ICCR), 2024: 1-5.

[10] TOLPEGIN V, TRUEX S, GURSOY M E, et al. Data poisoning attacks against federated learning systems [C]//25th European Symposium on Research in Computer Security, 2020: 480-501.

[11] HUANG Y X, YANG G, ZHOU H, et al. VPPFL: a verifiable privacy-preserving federated learning scheme against poisoning attacks [J]. Computers & Security, 2024, 136: 103562.

[12] CAO X Y, FANG M H, LIU J, et al. Fltrust: Byzantine-robust federated learning via trust bootstrapping [C]//28th Annual Network and Distributed System Security Symposium, 2021: 24434.

[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, 2022: 2545-2555.

[14] LIU Y, WANG T, PENG S L, et al. Cleaning and equipment clustering method of federated learning model based on edge [J]. Chinese Journal of Computers, 2021(12): 2515-2528.
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(12): 2515-2528.

[15] XU M. FedDBG: privacy-preserving dynamic benchmark gradient in federated learning against poisoning attacks [C]//2022 International Conference on Networking and Network Applications (NaNA), 2022: 483-488.

[16] BALDOMINOS A, SAEZ Y, ISASI P. A survey of handwritten character recognition with mnist and emnist [J]. Applied Sciences, 2019, 9(15): 3169.

[17] SUCHARITHA M, SESHA T, VIJAYAGOVINDAN S, et al. CIFAR-10 dataset image classification using CNN [C]//Proceedings of Eighth International Conference on Information System Design and Intelligent Applications, 2024: 395-407.

[18] MA Q P, JIA Q M, LIU J C, et al. Node grouping and time-sharing scheduling strategy for asynchronous federated learning in heterogeneous edge computing environment [J]. Journal of Communications, 2023, 44(11): 79-93.
马千飘, 贾庆民, 刘建春, 等. 异构边缘计算环境下异步联邦学习的节点分组与分时调度策略[J]. 通信学报, 2023, 44(11): 79-93.

[19] ZANG H R, YANG T T, LIU H B, et al. Research on distributed federal learning encryption verification for the Internet of Things [J]. Computer Science, 2024, 51(S1): 1062-1066.
臧洪睿, 杨婷婷, 刘洪波, 等. 面向物联网的分布式联邦学习加密验证研究[J]. 计算机科学, 2024, 51(S1): 1062-1066.

[20] WU X H, LI P, GU Y G, et al. Hierarchical federated learning algorithm based on EMD optimal matching [J]. Computer Engineering, 2025, 51(2): 170-178.
吴小红, 李佩, 顾永跟, 等. 基于EMD最优匹配的分层联邦学习算法[J]. 计算机工程, 2025, 51(2): 170-178.

(责任编辑: 尹晨茹)