面向联邦学习标签翻转攻击的客户端选择防御方法

李建鑫，陈思光
（南京邮电大学物联网学院，江苏 南京 210003）

摘　要：联邦学习允许多个客户端仅共享模型更新而不上传本地数据以协作训练一个全局模型，但正是由于它这种基于分布式的全局聚合模式，导致联邦学习易受到标签翻转攻击的恶意影响。为此，提出了一种面向联邦学习标签翻转攻击的客户端选择防御算法。具体地，该算法基于客户端与辅助客户端模型的余弦相似度以及客户端模型的准确率获得每个客户端的可靠因子，并依据可靠因子进行加权聚合，以此获得全局模型。通过赋予良性客户端更高的权重，可显著降低恶性客户端对全局模型的影响，提高模型的准确率。结合客户端的历史良性情况，融合汤普森采样方法，计算每个客户端被选择进行聚合的概率，确定下一轮参与聚合的客户端。通过筛选更加良性的客户端进行聚合可有效防御标签翻转攻击，提升模型鲁棒性。仿真结果表明，与现有的联邦平均（FedAvg, federated averaging）算法和通过信任引导的拜占庭鲁棒联邦学习（FLTrust, Byzantine-robust federated learning via trust bootstrapping）算法相比，该算法能够更有效地防御标签翻转攻击并获得更高的准确率。

关键词：联邦学习；标签翻转攻击；汤普森采样；客户端选择
中图分类号：TP393
文献标志码：A
doi：10.11959/j.issn.2096−3750.2025.00403

Client selection for federated learning against label flipping attacks

LI Jianxin, CHEN Siguang
School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract: Federated learning (FL) allows multiple clients to train a global model collaboratively by sharing only model updates without uploading local data. But due to its distributed global aggregation mode, FL is vulnerable to the malicious impact of label flipping attacks. Therefore, a client selection algorithm was proposed for FL against label flipping attacks. Specifically, the algorithm obtains the reliability score of each client based on the cosine similarity of client model and auxiliary client model and the accuracy of client model, and carries out weighted aggregation according to the reliability score to obtain the global model. By assigning higher weights to benign clients, the influence of malicious clients on the global model can be significantly reduced and the accuracy of the model can be improved. Then, Thompson sampling method was integrated to calculate the probability of each client being selected for aggregation and determine the clients participating in the aggregation in the next round based on the historical benign data of the clients. By screening more benign clients for aggregation, label flipping attacks can be effectively prevented and the robustness of the model was improved. Simulation results show that compared with the existing FedAvg and FLTrust algorithms, the proposed algorithm can defend against label flipping attacks more effectively and achieve higher accuracy.

Key words: federated learning, label flipping attack, Thompson sampling, client selection

收稿日期：2023−11−01；**修回日期：**2024−06−30
通信作者：陈思光，sgchen@njupt.edu.cn
基金项目：国家自然科学基金项目（No. 61971235）；江苏省“333 高层次人才培养工程”；南京邮电大学“1311”人才计划资助
Foundation Items: The National Natural Science Foundation of China (No. 61971235), 333 High-Level Talents Training Project of Jiangsu Province, 1311 Talents Plan of NJUPT

0　引言

近年来，随着电子信息技术的不断普及，移动设备和感知设备被广泛部署，它们每天产生大量数据，给机器学习带来了新前景[1]。传统的机器学习大多采用集中式的方法来训练机器学习模型，用户将本地数据上传到中心服务器，然后通过数据预处理、数据分割、模型选择、模型训练等步骤实现机器学习，使用所有用户的私有数据对模型进行训练固然会给模型性能带来较大的提升[2]，然而，这种集中式的机器学习方法存在泄露用户数据，侵犯用户隐私的风险，这时候，联邦学习就应运而生了，它最初是在2017年由谷歌提出的[3]，并作为一种新兴的分布式机器学习范式[4]，被广泛应用于医疗、金融、工业、销售[5]等领域，有效打破数据壁垒[6]，成为移动网络、机器学习等交叉领域的热门研究课题[7]。它的核心思想是在保护用户数据隐私的前提下，实现多方共同参与的训练，解决数据孤岛问题[8]。

然而，由于联邦学习分布式的特性[9]，客户端能够完全控制自己的本地数据[10]，服务器无法掌控客户端的行为，因此联邦学习很容易受到各种攻击。其中，较为常见的是投毒攻击。它主要分为数据投毒攻击[11]和模型投毒攻击，数据投毒攻击是指恶性客户端向数据集中注入噪声或翻转数据标签，使全局模型做出错误的预测；模型投毒攻击是指恶性客户端篡改上传至中心服务器的本地模型参数或梯度[12]，使其包含恶意信息，用于全局模型的聚合，损害全局模型。本文探讨的是标签翻转攻击，它通常是由恶性客户端发起的，通过翻转某些类别样本的标签，让模型学习到错误的知识，最终诱导全局模型将源类别的样本错误分类为目标类别[13]，以达到损害全局模型，降低模型准确率的目的。恶性客户端在不需要掌握任何先验知识的前提下，就能够轻松地实行标签翻转攻击，并给整个模型带来巨大的危害[14-15]。因此，针对标签翻转攻击的防御工作已经刻不容缓。

为了缓解标签翻转攻击给联邦学习系统带来的危害，研究者已经提出了多种防御机制。当前，防御标签翻转攻击的方法主要是从基于客户端的本地数据集和客户端上传至中心服务器的模型更新两个维度考虑的。文献[16]通过考虑客户端的本地数据集来防御标签翻转攻击，它使用支持向量机模型来对数据进行审查，通过分类模型在有毒数据集和干净数据集上产生的总误差作为判断数据集是否有毒的标准。文献[17]基于数据预处理的方法通过学习去噪函数来消除标签翻转攻击带来的影响。文献[18]通过学习有毒数据标签和干净数据标签之间的转换关系来为标签翻转攻击建模，模拟对抗性噪声。

更多的方法是从第二个维度来防御标签翻转攻击，如文献[19]提出一种自适应的联邦聚合算法，它基于模型更新在各个维度的中位数来评估每个客户端的可靠度，然后自适应地调整相应用户模型更新的权重，并且该文献还提出了一种去除恶意模型更新的对数函数。文献[20]提出用一些鲁棒性的聚合方法来取代传统的联邦平均（FedAvg, federated averaging）算法，这些鲁棒性的聚合方法包括中位数聚合法、修剪平均聚合法等。鲁棒性聚合方法是通过从本地客户端上传至中心服务器的模型更新中剔除恶意的模型更新来达到防御标签翻转攻击的目的，但是如文献[19-20]只选择模型更新中每个维度中位数或平均数来进行全局聚合不具有科学依据。同时，这种事后防御的方法在客户端选择和模型训练过程中都会有恶性客户端的参与，防御的效果并不是很理想。

文献[21]通过比较本地客户端模型更新和参考更新的相似度，实现对恶意客户端的识别，同时引入客户端信誉度以降低非独立同分布数据对恶意客户端识别的影响。文献[22]通过向全局模型更新中添加对抗样本噪声，能够在保证联邦学习准确率的同时，有效防御攻击。文献[23]采用TOP-K梯度选择方法对模型更新进行筛选，同时对选择的模型更新进行裁剪量化以保证数据隐私安全。文献[24]中的Krum算法通过计算每一个更新与其余所有更新之间的欧氏距离，选择距离最小的更新用于全局模型的聚合。文献[25]通过Multi-Krum算法过滤掉恶意的模型更新来防御标签翻转攻击。虽然Krum和Multi-Krum算法鲁棒性较高，能够剔除掉距离其他模型较远的模型，减少恶意模型的影响，但是欧氏距离的计算会浪费较多的内存和计算资源，对于大规模的网络来说这种防御方法是不可行的，同时这两种方法需要提前知道恶性客户端的数量，显然是不符合现实场景的[26]。

文献[27]提出将模型更新存储在区块链上，然后由区块链中的矿工负责剔除不可信的模型更新。文献[28]基于双陷门同态加密设计了隐私保护防御

策略，通过评估客户端模型的质量来决定不同客户端模型更新在全局聚合时的权重，实现鲁棒性的聚合。文献[29]提出的自适应联邦平均算法通过隐马尔科夫模型在每轮迭代时检测并丢弃恶意的更新，但是该方法有较大的局限性，它假设客户端的本地数据集是独立同分布的，而现实生活中的场景大多是非独立同分布的，因而不具备现实意义。

因此，针对目前已有的防御方法所存在的计算开销大，不适用于非独立同分布场景以及防御效果不佳等问题，本文从客户端选择的角度出发，提出了一种面向联邦学习标签翻转攻击的客户端选择防御方法（FedCS, client selection for federated learning against label flipping attacks），主要贡献总结如下。

(1) 结合辅助客户端与中心服务器设计了可靠因子计算及加权聚合的方法。该方法基于辅助客户端测量的客户端模型准确率及客户端与辅助客户端模型的余弦相似度可以较小的计算开销获得客户端的可靠因子，用以评价客户端在某一训练轮次的良性程度。同时，基于该可靠因子实现对模型的加权聚合，以此获得全局模型，通过赋予良性客户端更高的权重，可显著降低恶性客户端对全局模型的影响，提高模型的准确率。

(2) 设计了动态客户端选择方法。结合客户端历史良性情况，融合多臂老虎机中的汤普森采样方法，通过对客户端被判定为良性客户端的次数 reliability 和被判定为恶性客户端的次数 unreliability 的实时更新，将它们的 Beta 分布映射为对应客户端被选择的概率，实现对下一轮聚合客户端的选择。通过该方法可以筛选出更加良性的客户端进行下一轮聚合，来提升模型的鲁棒性，防御标签翻转攻击。

(3) 在独立同分布和非独立同分布的两个场景以及不同的恶性客户端比例下开展了大量的实验，实验结果表明本算法能够有效地防御标签翻转攻击，较大地提升模型准确率。

1　系统模型

本文构建了 FedCS 模型，该模型由用户层和服务器层组成，结构如图 1 所示，具体每层的功能定义如下。

1.1　用户层

用户层由 N 个客户端组成，但并不是所有的客户端都是可信的，本模型假设有比例为 a 的客户端是恶性的（如图 1 中橙色图标所示），即它们会发动标签翻转攻击，危害全局模型，降低模型的准确率。同时，本文假设客户端的数据集既可能是独立同分布的，也可能是非独立同分布的。用户层在接收到服务器层下发的全局模型和下一轮参与聚合的客户端集合后，集合中的 K 个客户端利用自己的私有数据集对全局模型进行本地训练并将训练后的模型上传到服务器层。

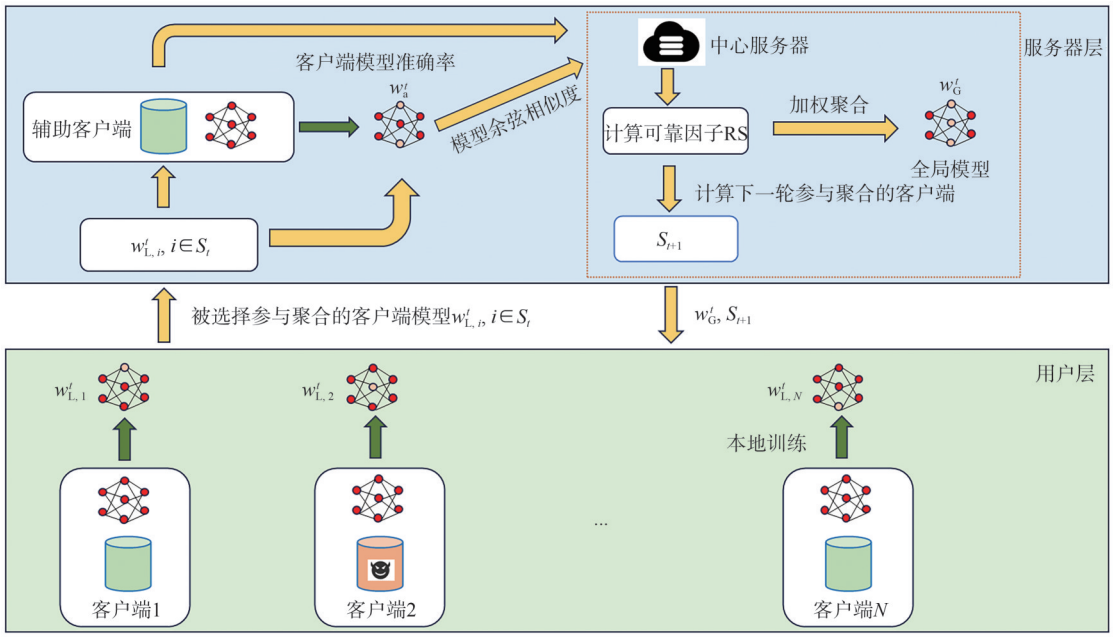

图 1　系统结构

1.2　服务器层

服务器层由辅助客户端和具有强大计算能力的中心服务器构成。假设服务器是可信的，即它不会主动发动攻击危害模型。服务器层主要包括如下4个功能。

- 辅助客户端接收中心服务器发来的全局模型，利用自己的辅助数据集对模型进行训练。
- 辅助客户端接收用户层上传的K个训练好的客户端模型，利用自己的辅助数据集测试客户端模型的准确率。
- 中心服务器计算辅助客户端与客户端模型的余弦相似度并基于准确率和余弦相似度两个指标计算K个客户端的可靠因子，然后根据可靠因子对客户端模型进行加权聚合，以此获得全局模型。
- 中心服务器基于N个客户端的历史可靠情况，融合汤普森采样方法，选择下一轮参与聚合的K个客户端，并将更新后的全局模型分发给这些客户端。

2　FedCS算法

在联邦学习中，不同的分布式客户端共同训练一个全局模型，但是，联邦学习分布式的特性使得它容易受到标签翻转攻击，恶性客户端可通过翻转本地数据集的数据标签，使得全局模型对某些样本分类错误，危害了全局模型，也降低了模型的准确率。因此本文设计了FedCS算法，该算法基于辅助客户端测量的客户端模型准确率及客户端与辅助客户端模型的余弦相似度获得客户端的可靠因子。同时，基于该可靠因子实现对模型的加权聚合，以此获得全局模型，通过赋予良性客户端更高的权重，可降低恶性客户端对全局模型的影响。该算法同时结合客户端历史良性情况，融合汤普森采样方法，实现对下一轮聚合客户端的选择。通过上述操作，可有效防御标签翻转攻击，提高模型准确率。本算法主要包括客户端可靠因子计算与加权聚合和动态客户端选择两大部分，详细设计如下。

2.1　可靠因子计算与加权聚合

本模型在服务器层设置了一个辅助客户端，假设该客户端的训练数据集是干净的，并且数据分布与整体训练数据分布保持一致，因此辅助客户端可以看作良性的，所以可以通过比较本地客户端和辅助客户端训练后模型的相似度来度量本地客户端的良性程度。本算法选择使用余弦相似度来度量两个模型之间的相似度。余弦相似度用向量空间中两个向量夹角的余弦值来衡量两个个体间差异的大小，余弦值越接近1，就表明夹角越接近0度，两个向量越相似，即本地客户端更良性。

假设两个向量分别是$\boldsymbol{v}_1$和$\boldsymbol{v}_2$，它们之间的余弦相似度$\text{proximity}\left(\boldsymbol{v}_1,\boldsymbol{v}_2\right)$可以表示为

$$\text{proximity}\left(\boldsymbol{v}_1,\boldsymbol{v}_2\right)=\cos\Theta=\frac{<\boldsymbol{v}_1,\boldsymbol{v}_2>}{\left\|\boldsymbol{v}_1\right\|\cdot\left\|\boldsymbol{v}_2\right\|}=\frac{\sum_{i=1}^{n}\boldsymbol{v}_{1i}*\boldsymbol{v}_{2i}}{\sqrt{\sum_{i=1}^{n}\left(\boldsymbol{v}_{1i}\right)^2}*\sqrt{\sum_{i=1}^{n}\left(\boldsymbol{v}_{2i}\right)^2}} \tag{1}$$

其中，Θ是向量$\boldsymbol{v}_1$和$\boldsymbol{v}_2$之间的夹角，n是向量的维数。

具体地，假设S_t为在第t轮中被选择进行聚合的本地客户端集合，将第t轮中第i个客户端训练后的本地模型定义为$w_{\mathrm{L},i}^{t}$，第t–1轮经过服务器聚合后的全局模型是w_{G}^{t-1}，通过式(2)可得到第t轮中第i个客户端的本地模型更新g_i^t，计算式为

$$g_i^t=w_{\mathrm{L},i}^{t}-w_{\mathrm{G}}^{t-1} \tag{2}$$

同时，服务器层的辅助客户端在第t轮经过辅助数据集训练后的模型为w_{a}^{t}，在第t轮的模型更新为g_{a}^{t}。计算式为

$$g_{\mathrm{a}}^{t}=w_{\mathrm{a}}^{t}-w_{\mathrm{G}}^{t-1} \tag{3}$$

因此，本地客户端与辅助客户端的余弦相似度为

$$\text{proximity}\left(g_i^t,g_{\mathrm{a}}^{t}\right)=\frac{<g_i^t,g_{\mathrm{a}}^{t}>}{\left\|g_i^t\right\|\cdot\left\|g_{\mathrm{a}}^{t}\right\|} \tag{4}$$

当客户端的数据集是非独立同分布的时候，良性客户端之间的余弦相似度并不一定很高，同时恶性客户端与良性客户端的余弦相似度可能会比较高，因此，模型的余弦相似度不能成为评判客户端良性程度的唯一指标，所以本算法融合了模型的准确率acc作为另一个指标，同时为了减少计算开销与内存占用，基于辅助客户端的少量样本数据集来测试客户端模型的准确率，计算式为

$$\text{acc}=M\left(w_{\mathrm{L},i}^{t},D_{\mathrm{a}}\right) \tag{5}$$

其中，$M(\cdot)$是用于评估客户端模型的函数，输出模型的准确率；D_{a}是辅助客户端的辅助数据集。

基于求得的proximity和acc两个指标，进行权重和运算就可得到可靠因子RS，计算式为

$$\text{RS}=\gamma\times\text{proximity}+\delta\times\text{acc} \tag{6}$$

其中，γ和δ分别是余弦相似度和准确率两个指标

在可靠因子中占的比重。

为了提升良性客户端在全局聚合中的比重，减小恶性客户端对全局模型的影响，本算法基于可靠因子对客户端模型进行加权聚合，每个客户端的模型权重是自身可靠因子与所有参与训练客户端可靠因子之和的比值，通过赋予良性客户端更高的权重，提升模型鲁棒性，有效防御标签翻转攻击，更新全局模型，最后将全局模型分发给下一轮被选中的客户端。更新后的全局模型为

$$w_{\mathrm{G}}^{t}=\sum_{i=1}^{K}\frac{\mathrm{RS}_i}{\sum_{j=1}^{K}\mathrm{RS}_j}w_{\mathrm{L},i}^{t} \tag{7}$$

其中，w_{G}^{t}指第t轮的全局模型，RS_i代表第i个客户端的可靠因子。

2.2　动态客户端选择

本文基于可靠因子，进一步融合本地客户端的历史可靠情况选择良性的客户端进行下一轮的训练。具体地，基于多臂老虎机中的汤普森采样方法筛选出更加良性的客户端进行下一轮聚合，以提升模型的鲁棒性，防御标签翻转攻击。

多臂老虎机算法是在连续决策问题中研究探索与开发权衡问题的经典框架。具体地，该算法假设有J台机器，赌徒每次选择其中一台机器拉动其杠杆，随后该机器会提供一个随机的奖励，每一台机器的奖励服从特定的概率分布。该赌徒有Q次拉动杠杆的机会，他的目标是使得总奖励最大，所以他要确定拉动杠杆的顺序，试图找到奖励最大的机器，并尽可能多地拉动这台机器的杠杆。

在联邦学习中，客户端就相当于多臂老虎机中的机器，选择某个客户端进行训练就相当于选择某个机器拉动它的杠杆，客户端训练后的模型对全局模型的贡献就相当于拉动杠杆得到的奖励值。因此，多臂老虎机算法非常适合用于联邦学习的场景中。

汤普森采样算法是多臂老虎机问题的一个流行的解决方案，首先对每个杠杆的奖励概率分布假设一个先验分布，通常选择Beta分布作为先验；然后，每次采样时，根据先验分布随机生成每个杠杆的奖励概率，并选择最高概率的杠杆进行采样；采样结束后，根据实际的奖励结果，更新每个杠杆的奖励概率分布，并进行下一次采样。Beta分布主要有α和β两个参数，α代表某事件成功的次数，β代表某事件失败的次数，α与β的相对比值越大，得到的Beta分布的概率也越大。Beta分布的概率密度函数为

$$\begin{aligned}f\left(x,\alpha,\beta\right)&=\text{constant}\cdot x^{\alpha-1}(1-x)^{\beta-1}=\\&\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\mathrm{d}u}=\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}x^{\alpha-1}(1-x)^{\beta-1}=\\&\frac{1}{\mathrm{Beta}\left(\alpha,\beta\right)}x^{\alpha-1}(1-x)^{\beta-1}\end{aligned} \tag{8}$$

其中，$\Gamma(\,)$表示gamma函数，实现数值的阶乘。

由于Beta分布的概率密度函数值域为[0,1]，Beta分布可以用于描述各种0–1区间内的事件。本文基于汤普森采样方法，根据Beta分布概率密度函数值域在[0,1]的特点将其映射为每个客户端被选择进行下一轮聚合的概率p，计算式为

$$p=\mathrm{Beta}\left(\mathrm{reliability},\mathrm{unreliability}\right) \tag{9}$$

在第一轮训练中，本算法将reliability和unreliability都设定为1，此时的Beta分布在(0,1)均匀分布，表示在第一轮训练前，每一个客户端都有均等概率被选择进行第一轮的聚合。然后，在后续轮次训练过程中，当中心服务器计算完客户端的可靠因子后，会对这些可靠因子进行降序排序，排在前80%的可靠因子对应客户端的reliability属性值会加1，其余客户端则保持reliability值不变，unreliability属性值加1，这样就可以实时地对每个客户端的reliability值和unreliability值进行更新，而reliability与unreliability的相对比值就是该客户端历史良性情况，然后就可以通过Beta分布计算每一个客户端下一轮被选中的概率。为了选择更加良性的客户端进行训练，本算法将每个客户端被选中的概率进行排序，选择概率最大的前K个客户端进行下一轮聚合，计算式为

$$S_t=\max\nolimits_K\left\{p_1,p_2,\cdots,p_N\right\}, \tag{10}$$

其中，函数$\max_K$表示从集合中选择最大的前K个元素，并返回选中元素下标组成的集合；S_t是第t轮被选中进行聚合的客户端序号的集合；p_i指在第t轮第i个客户端被选中进行聚合的概率。为便于理解本文所设计方法，将上述过程凝练成伪代码的形式，FedCS算法如算法1所示。

算法1　FedCS算法

输入：通信轮数T，本地迭代次数E，学习率η，参与训练的客户端集合S_t，辅助客户端s_{a}，辅助数据集D_{a}

输出：全局模型w_{G}^{T}，下一参与训练的客户端

集合轮 S_{t+1}

初始化：初始化全局模型 w_{G}^{0}

for 通信轮次 $t = 0, 1, 2, \cdots, T-1$ do

S_t 中的客户端和辅助客户端进行本地训练；

基于式(1)获得客户端的余弦相似度 proximity；

基于式(6)获得客户端的可靠因子 RS_i；

基于式(7)获得聚合后的全局模型 w_{G}^{t}；

将获得的可靠因子根据大小进行排序；

如果可靠因子排在前80%：

对应客户端的可靠度 reliability 值加1；

否则：

对应客户端的不可靠度 unreliability 加1；

for 每个客户端 $i = 1, 2, \cdots, N$ do

基于式(9)计算每个客户端被选择进行下一轮聚合的概率 p_i；

将概率 p_i 根据大小进行排序；

如果 p_i 排在前 K 个：

将对应的客户端下标加入集合 S_{t+1} 中；

获得下一轮参与训练的客户端集合 S_{t+1}；

end

end

3　仿真与分析

3.1　样本集与仿真设置

仿真使用的样本集为 MNIST 数据集[30]，它是一个被广泛用于计算机视觉和机器学习领域的经典数据集。它包含了一组手写数字的灰度图像，每个图像都是28像素×28像素，每个像素的值表示灰度级别，通常在0到255之间。MNIST 数据集包含10个类别，分别为数字0到9。它分为训练集和测试集两个部分。训练集共有60 000个图像，用于模型的训练；测试集共有10 000个图像，用于模型的性能评估。MNIST 数据集中的标签是用独热编码（one-hot-vectors）表示的，一个 one-hot 向量除了某一位数字是1之外，其余维度的数字都是0，比如标签1用独热编码表示为([0, 1, 0, 0, 0, 0, 0, 0, 0, 0])，标签5用独热编码表示为([0, 0, 0, 0, 0, 1, 0, 0, 0, 0])。

实验采用卷积神经网络（CNN, convolutional neural network）模型[31]，它能够自动提取输入数据的特征，实现对输入数据的高效分类和识别。其次，本实验设置了1个中心服务器和30个本地客户端，不同于传统的联邦学习架构，实验在服务器层加入了一个辅助客户端，帮助评判本地客户端的良性程度。实验设置了两个不同的场景，分别是独立同分布和非独立同分布场景，在独立同分布的场景中，首先从训练数据集中的60 000张图片中，按照每个类别选取20张图片总共200张图片分发给辅助客户端用于训练，接下来将剩余的图片均匀分发给所有的本地客户端，保证每个类别的图片在所有客户端中的占比是一样的；在非独立同分布的场景中，按照同样的方法分发数据给辅助客户端，但是需要将剩余的图片随机分发给本地客户端，保证非独立同分布的要求。此外，客户端采用随机梯度下降方法[32]进行本地训练，采取交叉熵损失函数以更好地适应分类问题，实验设置本地迭代轮次为3，全局迭代轮次为200，每一轮选取的本地客户端数量为 $K=5$，恶性客户端的比例分别是 $a = 0.1, a = 0.2, a = 0.3$，并且恶性客户端本地数据集中的数据全部被污染，即每一个数据的标签 l 会被翻转成 $10 - l$，学习率设置为0.001，余弦相似度和准确率的比重 β 和 γ 分别设置为0.3和0.7。为了评估所提面向标签翻转攻击的客户端选择防御方法，本文将所提出的方法与 FedAvg[3]算法和 FLTrust[33]算法进行对比，以突出本文方法的防御优势。

3.2　仿真结果与分析

FedAvg 算法、通过信任引导的拜占庭鲁棒联邦学习（FLTrust, Byzantine-robust federated learning via trust bootstrapping）算法和本文所提 FedCS 算法在独立同分布场景下、非独立同分布场景下模型准确率随迭代次数变化分别如图2和图3所示。可以看出，无论是在独立同分布还是非独立同分布的场景下，随着恶性客户端比例的上升，3种算法的准确率都有不同程度的下降。以独立同分布的场景为例，FedAvg 算法的准确率从92.5%下降到86.4%，下降幅度最大；FLTrust 算法的准确率从93.1%下降到91.1%；FedCS 算法的准确率从95.3%下降到92.7%，准确率均处于最高。当恶性客户端比例分别处于0.1、0.2、0.3时，FedAvg 算法和 FLTrust 算法相对 FedCS 算法的表现都较差，当处在独立同分布的场景时，FLTrust 算法虽然收敛速度不如 FedAvg 算法，但是能够取得更高的准确率，实现相对较好的防御性能；当处在非独立同分布的场景时，FedAvg 算法取得的防御效果明显优于 FLTrust 算法，这是

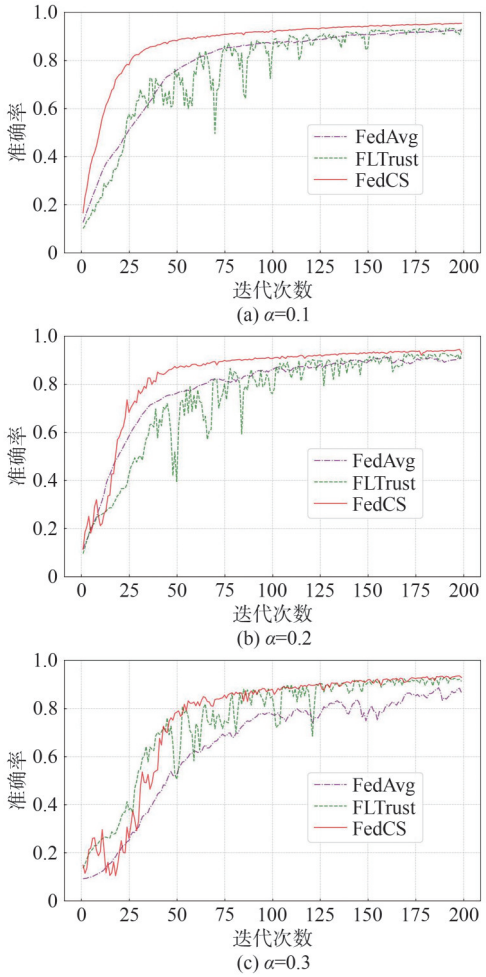

图2　独立同分布场景下模型准确率随迭代次数变化

由于FLTrust算法在服务器端设置了一个服务器模型，仅考虑本地模型更新与服务器模型更新的余弦相似度用于计算本地客户端的信任分数，最终基于客户端的信任分数对模型进行加权聚合，但是在非独立同分布的场景中，恶性客户端与良性客户端模型的余弦相似度可能会比较高，良性客户端之间模型的余弦相似度反而会比较低，所以该方法并不能很好地适用于非独立同分布场景。相比之下，本文所提出的FedCS算法融合了本地客户端和辅助客户端模型的余弦相似度和本地客户端模型在辅助数据集上的准确率两个指标获得客户端可靠因子，并基于该可靠因子实行加权聚合和客户端选择策略，通过上述操作，能够赋予良性客户端更高的权重，可显著降低恶性客户端影响，并能够较快地筛选出良性客户端进行聚合，同时又能够适用于非独立同分布场景，实现较快的收敛，有效防御标签翻转攻击。

从图2和图3也可以看出，FedCS算法较其他两种算法能够较快地实现模型收敛，在独立同分布的场景中，FedCS算法能够在第100轮左右实现收敛，在非独立同分布的场景中，FedCS算法能够在第150轮左右实现收敛；同时，相比其他两种算法，FedCS算法在整个训练过程中的准确率随迭代次数的变化波动较小，尤其在恶性客户端比例较低的时候，准确率基本随着迭代次数增加而单调增大，这是因为FedAvg算法和FLTrust算法每一轮都是随机选择客户端，这样模型的准确率就会较大程度地受到每一轮所选择客户端的影响，如果选择的都是良性客户端，那么模型准确率就能够实现较大的提升，如果选择的客户端中有较多恶性客户端，那么模型准确率就会受到较大的影响。FedCS算法在独立同分布和非独立同分布两个场景中的表现都要优于其他两种算法，并不会受到数据分布的较大影响。

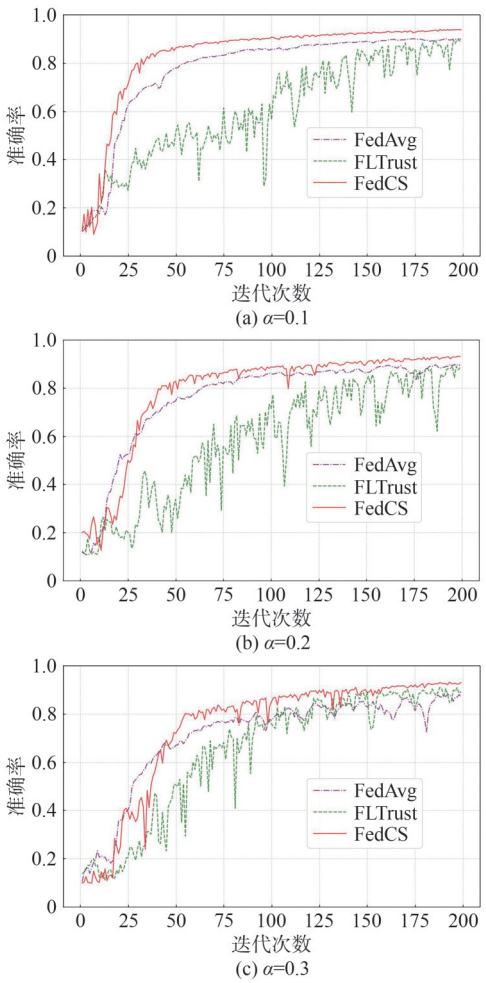

图3　非独立同分布场景下模型准确率随迭代次数变化

进一步观察图2和图3可以发现，FedCS算法在迭代的初期准确率会出现波动，这是因为第一轮给每一个客户端设置的reliability值和unreliability值都是1，所以每一个客户端有均等的机会被选择进行第一轮训练。迭代初期，恶性客户端和良性客户端训练后的本地模型在辅助数据集上进行准确率测试，由于全局迭代轮数较少，模型学习到的知识不够充分，客户端模型测得的准确率都是偏低的，他们的可靠因子也都很接近，恶性客户端也就有机会在后续的轮次中再次被选中。不过，经过20轮左右的训练，良性客户端的优势就能够体现出来，它们已经学习到了大量知识，测得的模型准确率远高于恶性客户端，所以，良性客户端的reliability值不断加大，而恶性客户端的unreliability值不断加大，后续就能够较准确地把良性客户端筛选出来用于训练，从根本上有效防御标签翻转攻击。观察图3可以发现，FLTrust方法和FedAvg方法在非独立同分布的场景中性能表现接近，这主要是由于FLTrust算法在服务器端设置了一个服务器模型，然后仅考虑本地模型更新与服务器模型更新的余弦相似度用于计算本地客户端的信任分数。最终，基于客户端的信任分数对模型进行加权聚合，但是在非独立同分布的场景中，恶意客户端与良性客户端模型的余弦相似度可能会比较高，良性客户端之间模型的余弦相似度反而会比较低；也就是说在非独立同分布的场景下，余弦相似度和客户端的良性程度并不成正比，加权聚合操作也不能够有效提高良性客户端模型的比重，所以该方法并不能很好地适用于非独立同分布场景。FedAvg算法并没有客户端选择的过程，每一轮都是随机选择客户端，并且依据本地训练样本比例决定加权聚合的权重，而这个权重同样不能反映客户端的良性程度。因此，这两种方法对模型的加权聚合都不能有效地提升良性客户端模型的比重，随机性都较强，两种方法在非独立同分布场景下性能表现较为类似。

为了进一步验证可靠因子加权计算方式的作用，对其两个指标模型准确率和模型的余弦相似度在独立同分布场景和非独立同分布场景下分别进行了消融实验。图4展示了针对可靠因子的消融实验。由图4(a)、图4(b)发现，本文所提出的对两个指标进行加权得到可靠因子的方法FedCS能够带来更高的模型准确率，分别达到了98.57%、98.3%，能够更好地防御标签翻转攻击，而只采用模型准确率或者模型的余弦相似度计算客户端的可靠因子不能够完全筛选出良性客户端，同时对客户端良性程度的评价也不够准确，尤其是在非独立同分布的场景下，加权计算可靠因子带来的性能收益更加明显，因为在非独立同分布的场景下，良性客户端的准确率以及与辅助客户端模型的余弦相似度都不一定很高，因此，只有综合考量两个因素才能够更加精准地筛选出良性客户端进行后续训练。

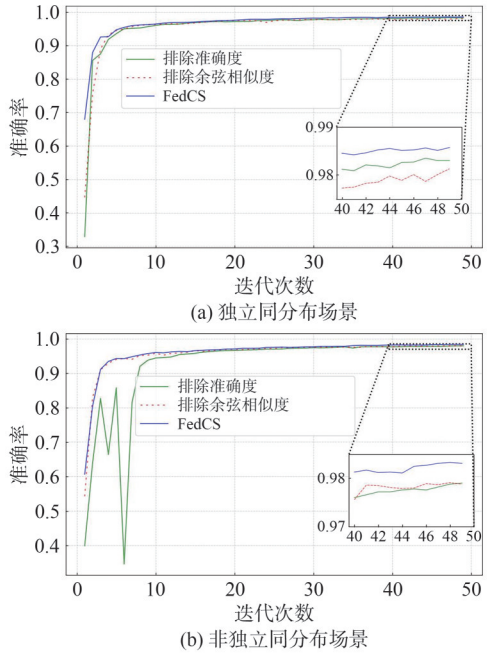

图4　针对可靠因子的消融实验

4　结束语

为了防御标签翻转攻击，提升模型的鲁棒性，本文提出了一种面向联邦学习标签翻转攻击的客户端选择防御算法。具体地，本文所提算法基于辅助客户端测量的客户端模型准确率及客户端与辅助客户端模型的余弦相似度获得客户端的可靠因子。同时，基于该可靠因子实现对模型的加权聚合，以此获得全局模型。本文所提算法同时结合客户端历史良性情况，融合汤普森采样方法，实现了对下一轮聚合客户端的选择。通过上述操作，本文所提算法能够快速有效地把恶性客户端剔除出来，选择良性客户端进行聚合。最后，与现有算法的对比表明本文所提算法能够较快地收敛，并能有效防御标签翻转攻击，提高模型准确率。

参考文献：

[1] KHAN L U, SAAD W, HAN Z, et al. Federated learning for Internet of Things: recent advances, taxonomy, and open challenges[J]. IEEE Communications Surveys & Tutorials, 2021, 23(3): 1759-1799.

[2] FANG X W, YE M. Robust federated learning with noisy and heterogeneous clients[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 10062-10071.

[3] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Proceedings of the International Conference on Artificial Intelligence and Statistics, 2022

[4] SUN J W, LI A, WANG B H, et al. Soteria: provable defense against privacy leakage in federated learning from representation perspective[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 9307-9315.

[5] 郭英芸, 高博, 张志飞, 等. 一种基于带宽分配的联邦学习激励机制[J]. 物联网学报, 2022, 6(4): 82-92.
GUO Y Y, GAO B, ZHANG Z F, et al. An incentive mechanism with bandwidth allocation for federated learning[J]. Chinese Journal on Internet of Things, 2022, 6(4): 82-92.

[6] 郭佳慧, 陈卓越, 高玮, 等. 基于背包模型的联邦学习客户端选择方法[J]. 物联网学报, 2022, 6(4): 158-168.
GUO J H, CHEN Z Y, GAO W, et al. Clients selection method based on knapsack model in federated learning[J]. Chinese Journal on Internet of Things, 2022, 6(4): 158-168.

[7] 耿光磊, 高博, 熊轲, 等. 联邦学习赋能6G网络综述[J]. 物联网学报, 2023, 7(2): 50-66.
GENG G L, GAO B, XIONG K, et al. A survey of federated learning for 6G networks[J]. Chinese Journal on Internet of Things, 2023, 7(2): 50-66.

[8] WANG Y J, LIN L, CHEN J H. Communication-efficient adaptive federated learning[C]//Proceedings of the 39th International Conference on Machine Learning (ICML). Piscataway: IEEE Press, 2022, 162: 22802-22838.

[9] SUN Y, SHEN L, SUN H, et al. Efficient federated learning via local adaptive amended optimizer with linear speedup[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 14453-14464.

[10] SHEN X C, LIU Y, LI F, et al. Privacy-preserving federated learning against label-flipping attacks on non-IID data[J]. IEEE Internet of Things Journal, 2024, 11(1): 1241-1255.

[11] SUN G, CONG Y, DONG J H, et al. Data poisoning attacks on federated machine learning[J]. IEEE Internet of Things Journal, 2022, 9(13): 11365-11375.

[12] SHI S P, HU C, WANG D, et al. Federated anomaly analytics for local model poisoning attack[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(2): 596-610.

[13] XU Q Q, YANG Z Y, ZHAO Y R, et al. Rethinking label flipping attack: from sample masking to sample thresholding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7668-7685.

[14] SUCIU O, MĂRGINEAN R, KAYA Y, et al. Technical report: when does machine learning FAIL? generalized transferability for evasion and poisoning attacks[J]. arXiv preprint, 2018, arXiv: 1803.06975v2.

[15] ROSENFELD E, WINSTON E, RAVIKUMAR P, et al. Certified robustness to label-flipping attacks via randomized smoothing[C]//Proceedings of the 37th International Conference on Machine Learning (ICML). Piscataway: IEEE Press, 2020: 8230-8241.

[16] DOKU R, RAWAT D B. Mitigating data poisoning attacks on a federated learning-edge computing network[C]//Proceedings of the 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC). Piscataway: IEEE Press, 2021: 1-6.

[17] NASEER M, KHAN S, HAYAT M, et al. A self-supervised approach for adversarial robustness[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 259-268.

[18] ZHOU D W, WANG N N, HAN B, et al. Modeling adversarial noise for adversarial training[C]//Proceedings of the 39th International Conference on Machine Learning (ICML). Piscataway: IEEE Press, 2022: 27353-27366.

[19] 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.

[20] YIN D, CHEN Y D, RAMCHANDRAN K, et al. Byzantine-robust distributed learning: towards optimal statistical rates[J]. arxiv preprint, 2018, arxiv: 1803.01498.

[21] 马鑫迪, 李清华, 姜奇, 等. 面向Non-IID数据的拜占庭鲁棒联邦学习[J]. 通信学报, 2023, 44(6): 138-153.
MA X D, LI Q H, JIANG Q, et al. Byzantine-robust federated learning over Non-IID data[J]. Journal on Communications, 2023, 44(6): 138-153.

[22] 张佳乐, 朱诚诚, 孙小兵, 等. 基于GAN的联邦学习成员推理攻击与防御方法[J]. 通信学报, 2023, 44(5): 193-205.
ZHANG J L, ZHU C C, SUN X B, et al. Membership inference attack and defense method in federated learning based on GAN[J]. Journal on Communications, 2023, 44(5): 193-205.

[23] 余晟兴, 陈钟. 基于同态加密的高效安全联邦学习聚合框架[J]. 通信学报, 2023, 44(1): 14-28.
YU S X, CHEN Z. Efficient secure federated learning aggregation framework based on homomorphic encryption[J]. Journal on Communications, 2023, 44(1): 14-28.

[24] BLANCHARD P, MHAMDI E, GUERRAOUI R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]//Proceedings of the Neural Information Processing Systems (NeurIPS). Piscataway: IEEE Press, 2017, 30: 119-129.

[25] SHAYAN M, FUNG C, YOON C J M, et al. Biscotti: a blockchain system for private and secure federated learning[J]. IEEE Transactions on Parallel and Distributed Systems, 2021, 32(7): 1513-1525.

[26] JIN S, LI Y, CHEN X, et al. Blockchain-based fairness-enhanced federated learning scheme against label flipping attack[J]. Journal of Information Security and Applications, 2023, 77: 103580.

[27] QI Y H, HOSSAIN M S, NIE J T, et al. Privacy-preserving blockchain-based federated learning for traffic flow prediction[J]. Future Generation Computer Systems, 2021, 117: 328-337.

[28] MA Z R, MA J F, MIAO Y B, et al. ShieldFL: mitigating model poisoning attacks in privacy-preserving federated learning[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 1639-1654.

[29] MUNOZ-GONZALEZ, CO K, LUPU E. Byzantine-robust federated machine learning through adaptive model averaging[J]. arXiv preprint, 2019, arXiv: 1909.05125.

[30] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[31] LEE S, HOFFMAN J, WANG Z J, et al. VIsCUIT: visual auditor for bias in CNN image classifier[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 21443-21451.

[32] FAN X, WANG Y, HUO Y, et al. BEV-SGD: best effort voting SGD against Byzantine attacks for analog-aggregation-based federated learning over the air[J]. IEEE Internet of Things Journal, 2022, 9(19): 18946-18959.

[33] CAO X Y, FANG M H, LIU J, et al. FLTrust: Byzantine-robust federated learning *via* trust bootstrapping[C]//Proceedings 2021 Network and Distributed System Security Symposium. Internet Society, 2021.

[作者简介]

李建鑫（2000– ），男，南京邮电大学物联网学院硕士生，主要研究方向为联邦学习与边缘智能。

陈思光（1984– ），男，博士，南京邮电大学物联网学院教授，主要研究方向为边缘智能与安全。