计算机科学
COMPUTER SCIENCE

基于多特征检测与自适应权重调整的鲁棒联邦学习算法

王春东, 赵立扬, 张博宇, 赵永新

引用本文

王春东, 赵立扬, 张博宇, 赵永新. 基于多特征检测与自适应权重调整的鲁棒联邦学习算法[J]. 计算机科学, 2024, 51(11A): 231100072-10.

WANG Chundong, ZHAO Liyang, ZHANG Boyu, ZHAO Yongxin. Robust Federated Learning Algorithm Based on Multi-feature Detection and Adaptive WeightAdjustment [J]. Computer Science, 2024, 51(11A): 231100072-10.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100072

基于多特征检测与自适应权重调整的鲁棒联邦学习算法

王春东　赵立扬　张博宇　赵永新
天津理工大学计算机科学与工程学院　天津 300384
天津理工大学天津市智能计算与软件新技术重点实验室　天津 300384

摘　要　联邦学习作为一种保护隐私的分布式机器学习范式,允许多个客户端在不泄露原始训练数据的情况下协同训练全局模型。然而,由于无法直接访问客户端本地训练数据和无法监控本地训练过程,联邦学习面临各种拜占庭攻击的威胁,如数据中毒和模型篡改攻击。这些攻击旨在扰乱联邦学习模型训练过程,降低模型性能。针对此问题,尽管已有许多研究提出了不同的聚合算法,但这些方法主要聚焦于单一拜占庭攻击场景,而忽略了实际环境中可能出现的混合拜占庭攻击所带来的威胁。为应对这一难题,受净水器的原理启发,提出了一种基于多特征检测与自适应权重调整的新型拜占庭鲁棒聚合算法 FL-Sieve,旨在通过多层次的筛查过滤恶意客户端。首先,算法通过角幅相似度和模型边界测度评估客户端间的特征相似性,生成相似度矩阵并计算相似性分数;接着,利用聚类算法将相似的节点归入同一簇,以确保相似的节点能够被正确分类;随后,根据预定义规则筛选潜在良性客户端;最后,根据每个客户端的信任度智能地分配权重,进一步增强防御效果和系统鲁棒性。为了验证 FL-Sieve 的性能,实验利用了 MNIST,Fashion-MNIST 和 CIFAR-10 这 3 种数据集,考虑了 Non-IID 数据分布情景和混合拜占庭攻击场景。混合拜占庭客户端的数量从 20%递增到 49%,以模拟大规模混合拜占庭客户端攻击的场景。同时也对 FL-Sieve 在 IID 和 Non-IID 数据分布以及单攻击场景下的性能进行了测试。实验结果表明,FL-Sieve 能够有效抵御不同场景下的拜占庭攻击,即使在高达 49%的混合拜占庭客户端攻击下,FL-Sieve 依然能够维持较高的主任务准确率。相比之下,几种现有的经典算法存在不同程度的失效,凸显出 FL-Sieve 的优势。

关键词:联邦学习;混合拜占庭攻击;多特征检测;动态分配权重;鲁棒聚合算法
中图分类号　TP309

Robust Federated Learning Algorithm Based on Multi-feature Detection and Adaptive Weight Adjustment

WANG Chundong,ZHAO Liyang,ZHANG Boyu and ZHAO Yongxin
School of Computer Science and Engineering,Tianjin University of Technology,Tianjin 300384,China
Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology,Tianjin University of Technology,Tianjin 300384,China

Abstract　The federated learning paradigm is designed to preserve privacy by enabling multiple clients to collaboratively train a global model without compromising the original training data. However,due to the lack of direct access to local training data and monitoring capabilities during the training process,federated learning is vulnerable to various Byzantine attacks,including data poisoning and model tampering attacks. These malicious activities aim at disrupting the federated learning model training process and degrading its performance. While several studies have proposed various aggregation algorithms to address this issue,they predominantly concentrate on single Byzantine attack scenarios,often overlooking the threats associated with hybrid Byzantine attacks that can manifest in real-world environments. To address this issue,inspired by the principle of water purifiers,we propose an innovative multi-feature detection and adaptive dynamic weighting allocation algorithm called FL-Sieve for identifying Byzantine clients,aiming to filter out malicious clients through multi-level screening. Firstly,the algorithm assesses feature similarity between clients through angular range similarity and model boundary metric,generates a similarity matrix and calculates the similarity score. Then,it performs clustering to ensure that nodes with similar features are grouped together. Subsequently,it employs predefined rules to filter potential benign clients. Finally,it intelligently allocates weights based on the trustworthiness of each client,further enhancing the defense mechanisms and system robustness. To evaluate the performance of the FL-Sieve algorithm,experiments are conducted using three datasets:MNIST,Fashion-MNIST,and CIFAR-10. The experiments consider scenarios with both non-IID data distribution and hybrid Byzantine attack situations. The number of hybrid Byzantine clients increases from 20% to 49% to simulate large-scale hybrid Byzantine client attacks. Additionally,the performance of the FL-Sieve algorithm is tested in both IID and non-IID data distribution,as well as in single attack scenarios. The experimental results demonstrate that

基金项目:国家自然科学基金(U1536122);天津市研究生科研创新项目(2022BKY158)
This work was supported by the National Natural Science Foundation of China(U1536122) and Tianjin Research Innovation Project for Postgraduate Students(2022BKY158).
通信作者:王春东(michael3769@163.com)

FL-Sieve effectively withstands Byzantine attacks in various scenarios, maintaining high main task accuracy even under the challenging condition of 49% hybrid Byzantine client attacks. In comparison, several existing classical algorithms exhibit varying degrees of failure, underscoring the significant advantages of the FL-Sieve algorithm.

Keywords Federated learning, Hybrid Byzantine attack, Multi-feature detection, Dynamic weight allocation, Robust aggregation algorithm

1 引言

联邦学习[1](FL)是一种分布式机器学习框架，旨在应对隐私安全[2]和数据孤岛[3]问题。它允许多个分布式客户端在本地协作训练一个全局机器学习模型，而无需共享各自的数据。相较于传统的集中式机器学习模式，联邦学习不仅解决了数据孤岛问题，还避免了上传可能导致隐私泄露的数据，从而减轻了客户在隐私方面的担忧。凭借在保护数据隐私和节省通信带宽方面的优势，联邦学习已广泛应用在物联网[4]、自然语言处理[5-6]以及图像处理[7]等领域。

然而，由于其分布式特性以及隐私保护措施，采用基础聚合算法 Fedavg[1]的联邦学习很容易受到拜占庭攻击[8-12]的威胁。由于服务器无法直接访问客户端数据，也无法监视客户端的训练过程，因此拜占庭客户端便可通过向服务器发送错误或误导性的信息来干扰全局模型的正常聚合。拜占庭攻击的表现如图 1 所示。在训练过程中，恶意客户端可以通过污染本地数据(例如将“猫”标签替换为“狗”)，或者篡改本地模型参数(例如恶意添加扰动)来生成恶意模型，随后将这些恶意模型上传至服务器，进而影响全局模型的性能。一些研究[13-14]表明，即使只有单个恶意客户端也可以显著影响全局模型的聚合，甚至阻止模型收敛。

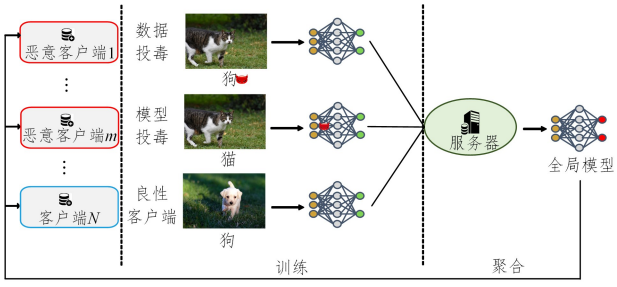


图 1 拜占庭客户端攻击联邦学习系统时的表现

Fig. 1 Performance of Byzantine client attacks federated learning system

如何抵御联邦学习中的攻击已成为当前的研究热点。目前已经提出了许多拜占庭鲁棒聚合算法，如利用统计特征估计全局模型中心或利用特定度量来区分良性与恶意客户端的 Coordinate-wise median 及 Krum 等[15-18]算法。此外，还有一些算法利用辅助数据来协助识别并排除恶意模型[10,19-21]，以避免恶意模型对全局模型产生不良影响。

尽管当前的研究已取得一定的进展，但仍存在一些不足之处。首先，基于统计均值或中值等特征的聚合算法在应用时具有一定局限性。该类方法通常只适用于客户端数据是独立同分布(Independent Identically Distribution, IID)的情况，且单一的防御指标很容易被攻击者巧妙构造的梯度所绕过。其次，基于辅助数据的检测算法需要服务器收集一定量的干净、可靠的辅助数据。然而实际操作中，这种方法面临着隐私和安全风险。出于对数据隐私的担忧，客户端往往不愿意共享其私有数据，而恶意客户端则可能会利用这一机制上传有毒数据，从而进一步危及整个防御方案的安全性和有效性。最后，现有研究都集中在单一的拜占庭攻击场景，未充分考虑到混合拜占庭攻击的潜在威胁。在现实场景中，攻击者不仅限于采用单一攻击策略，而是可能会结合多种策略来尝试破坏联邦学习系统。混合拜占庭攻击的威胁在于其综合了多种攻击手段，如数据投毒、模型篡改等，旨在通过多角度、多维度的方式对联邦学习过程进行干扰。这种灵活而复杂的攻击模式使得传统的面向单一攻击策略的防御措施可能不再适用，同时也为防御工作带来了更大的挑战。

为了应对上述挑战，受净水器过滤原理的启发，本文提出了鲁棒聚合算法 FL-Sieve(Federated Learning-Sieve, FL-Sieve)，旨在通过多层次的筛查过滤恶意客户端。在面对来自各个客户端的多样化模型更新时，本文采取一种多指标协同过滤策略，以协助识别和过滤出恶意梯度。为了全方位捕获梯度中的细微差异，本文引入了两个核心评估指标：角幅相似度(Angular Range Similarity, ARS)和模型边界测度(Model Boundary Metric, MBM)。ARS 重点关注模型梯度更新间角度和幅度差异，而 MBM 侧重于评估梯度的边界特征信息。两者相结合，提供了一个多角度、多维度捕获梯度特征的方法。为了进一步增强系统的鲁棒性，本文还开发了一种自适应动态权重分配策略，旨在实现更加高效的资源分配，并进一步加强对良性客户端的支持。

本文的主要贡献如下：

1)为了更准确地识别和防御混合拜占庭攻击这类复合攻击，本文引入了两个全新的客户端相似性指标：角幅相似度和模型边界测度。ARS 专注于模型梯度更新之间的角度与幅度差异，而 MBM 则重点关注梯度的边界特征信息，两者相结合为联邦学习算法提供了全面的梯度特征分析。

2)针对“防御混合拜占庭攻击”这一任务，提出了一种鲁棒的联邦学习聚合算法——FL-Sieve。FL-Sieve 综合应用了 ARS 和 MBM 指标，实现了对恶意梯度的高效检测和筛选。该算法不仅能有效鉴别恶意客户端提交的模型梯度，还通过自适应动态权重分配策略智能地分配学习资源，从而确保整体模型的稳定性和准确性。

3)本文在 3 个基准数据集上进行广泛实验，并与目前 5 种现有联邦学习安全策略进行对比评估。实验结果表明，即使遭遇大量混合拜占庭客户端攻击，FL-Sieve 仍然能保持出色的性能，且优于当前的同类算法。

2 相关工作

为了应对联邦学习中可能出现的攻击问题，研究者们提出了多种鲁棒的聚合算法。当前的策略主要分为两大类：基于统计特征差异的方法与基于验证评估的策略。

基于统计特征差异的方法主要关注模型更新间的特征差异，目的是识别或过滤掉潜在的恶意更新。Blanchard 等[15]提出了基于欧氏距离的 Krum 和 Multi-Krum 算法，其核心思想是选择一个或多个最近邻的更新来代替全局模型。Yin 和 Chen 等研究者[16-17]引入几何中值或坐标中值以增强系统鲁棒性，而后 Yin 等[16]进一步推出了利用截尾平均数策略滤除可能的恶意更新的 Trimmed-Mean 算法。Guerraoui 等[22]结合 Krum 和 Trimmed-Mean 的思想，提出了 Bulyan 算法。然而，Baruch 等[13]证明上述算法在处理非独立同分布（Non-Independent Identically Distribution，Non-IID）的数据时存在一定局限性。与此同时，Khazbak 等[23]提出了基于余弦相似度的评分系统，选择部分评分最高的参与方进行聚合。Fung 等[18]则通过历史聚合更新的余弦相似度进行权重调整策略，提出了 Foolsgold 算法。Lu 等[24]提出基于 L2 距离高斯分布的权重分配策略。Yu 等[25]设计了一种基于 K-means 的分组聚合策略。而 Yang 等[26]根据 Lipschitz 特征提出了基于中位数的聚合策略，但此方法主要适用于独立同分布数据集。上述基于特征差异的防御方案实施相对简单，对于引发显著参数异常的攻击具有较好的鲁棒性，然而当攻击仅引起微小变化时，其效果通常有限。

基于验证评估的策略核心在于验证每个客户端上传的模型更新的性能，从而判别潜在的恶意更新。Wang 等[27]是此类方法的早期探索者，他们提出了一个基于验证数据集分类准确率的评估策略，如果准确率低于设定的阈值则将其归为恶意方并排除，最终对筛选后的参与方求平均值。Tan 等[28]沿袭了类似的思路，并进一步引入深度强化学习，利用参与方的历史行为来优化客户端的选择。Chen 等[29]则融合了分组与验证策略，首先使用 K-means 对模型更新进行分组，然后分组分别进行验证。Kim 等[30]采用一种共识确认策略来验证模型性能，而 Rodríguez-Barroso 等[21]通过动态聚合操作符来验证模型性能。此外，为了使模型更新评估不完全依赖于准确性验证，部分学者引入了其他评估指标。例如，Xie 等[20]的 Zeno 算法结合损失函数下降和模型更新幅度，引入了“随机下降分数”评估指标。然而，Zeno 需要事先知道攻击者的数量。为了克服这个问题，Cao 等[31]结合小规模干净数据集和噪声梯度设计了一个鲁棒分布式梯度策略，以过滤受损客户端的信息。另外，Cao 等[10]提出的 Fltrust 算法把服务器自身的训练结果作为信任根，将服务器和参与方模型更新的余弦相似度作为评价指标。另外，Li 和 Gu 等[32]采用了预训练的自编码器来评估模型更新，而 Zhai 等[33]结合性能验证与自编码器验证的策略提出了 BRCA 算法。基于验证与评估的防御策略在理论上更具鲁棒性，但其实施却高度依赖于一个可信赖的验证数据集，该数据集通常假设由各客户端共享本地部分数据组成。然而，在实际应用中，构建一个全面且具有代表性的验证数据集面临着多重挑战。数据隐私保护是获取数据的首要难题。例如，在医疗领域，患者可能不愿意共享其医疗数据，未经其授权将数据用于联邦学习可能触犯相关隐私保护法规。此外，不同参与方之间存在数据管理标准的差异，增加了数据整合与标准化的复杂性。为了建立一个具有代表性的验证数据集，需要投入大量时间和资源解决数据的互操作性和一致性问题，即便成功建立了验证数据集，确保数据的真实性和可靠性同样是一项艰巨的任务。在多参与方的联邦学习环境中，恶意参与方可能故意上传错误或有偏差的数据，以破坏模型训练和性能。这种数据污染问题可能导致验证数据集的准确性和可靠性受损，从而削弱防御策略的有效性。验证数据集的质量直接关系到模型的泛化能力和对恶意行为的识别能力。在实际应用中，如果验证数据集缺乏全面性和真实性，模型可能无法有效泛化至新的数据样本，进而影响到对潜在恶意攻击的识别和防御。此外，现有防御算法通常只针对单一攻击场景，未充分考虑混合拜占庭攻击的潜在威胁。

3 相关概念及问题描述

3.1 联邦学习

在联邦学习中，各个客户端利用各自的本地数据联合训练一个全局模型。理想情况下，优化模型为：

$$\min_{w\in\mathbb{R}^d}\{F(w)\triangleq\sum_{i=1}^{N}\psi_i F_i(w)\} \tag{1}$$

其中，d 表示模型维度，N 代表客户端数量，ψ_i 是客户端权重。客户端目标函数 $F_i(\cdot)$ 定义为：

$$F_i(w)\triangleq\frac{1}{|D_i|}\sum_{j=1}^{|D_i|}\mathcal{L}(w,z_{i,j}) \tag{2}$$

其中，$\mathcal{L}(\cdot,\cdot)$ 为用户指定的损失函数。客户端 i 持有 $|D_i|$ 个数据 $z_{i,1},z_{i,2},\cdots,z_{i,|D_i|}$，$z_{i,j}$ 表示第 j 个数据样本。

在一轮迭代更新中，联邦学习可大致分为 3 个步骤：

1）服务器将聚合后的全局模型（在首轮迭代中为初始模型）下发给客户端。

2）客户端使用本地数据更新接收到的全局模型，并将本地模型更新回传服务器。

3）服务器收集客户端模型更新，并将其聚合为下一次迭代的全局模型。

3.2 拜占庭攻击

拜占庭攻击旨在通过植入有害数据或修改模型参数的方式向服务器发送错误的或对抗性的更新信息，以破坏机器学习的训练效果。在联邦学习中，与传统的集中式学习类似，最常见的攻击方式之一是数据投毒。例如，Shafahi 等[34]修改 CIFAR-10 数据集中部分青蛙图片的标签，导致模型无法正确分类这些图像。除了数据投毒外，还存在一种威胁性更大的攻击方式，即模型投毒[8,11,13,22,35-38]，相比于数据投毒，这种攻击方式对机器学习的破坏性更为显著。目前，两种广泛使用的攻击策略是符号翻转攻击和同值攻击。它们的主要策略是篡改模型参数为异常值，从而影响全局模型的最终性能。但由于这些攻击与正常模型之间存在显著差异，因此通常容易被检测到，所以大部分的防御算法已能够缓解这一问题。尽管如此，仍有一些更为复杂的攻击策略[11,13,35-36]不断涌现，如 Fang 等[11]提出的方法。该方法在上一轮全局聚合模型中加入微小的扰动向量，使每次聚合后的全局模型都轻微偏离原有的预期轨迹，随着迭代次数的增加，逐渐诱导全局模型偏离最初的期望轨迹。

3.3 攻击模型

在现实情况下，攻击者通常不会局限于单一的攻击手段，而是会倾向于采用多元化的攻击策略来规避安全机制的检测并提高攻击的成功率。相较于传统的单一拜占庭攻击，混合拜占庭攻击更能隐蔽攻击意图、使得恶意行为难以被系统迅速识别和应对。因此，除传统单拜占庭攻击外，本文同时考虑混合拜占庭攻击，与以往研究不同，本文并未对每轮攻击者的

攻击策略施加限制。换言之，本文不预设攻击者在每一轮训练中采取特定的攻击手段，而是允许恶意攻击者在训练过程中多次采用不同的拜占庭攻击手法。详细的设置将在实验部分中详述。

此外，与之前的研究[11,13-14,35]中的威胁模型类似，本文假设攻击者可以操纵一组本地客户端，以达到污染模型的目的，这些被污染的模型将会被聚合到全局模型中，从而影响全局模型的性能。攻击者的能力涵盖以下几个方面：1）攻击者具备随意操纵受控本地客户端的训练数据的能力；2）攻击者可以完全掌控本地训练过程和超参数，例如学习率等；3）在提交模型更新前，攻击者可以对其进行修改；4）攻击者可以选择是否参与全局模型的更新。

值得特别注意的是，为全面了解本文算法在面对复杂和隐蔽攻击时的鲁棒性和可靠性，本文假设存在高级攻击者。高级攻击者不仅拥有普通攻击者的能力，还额外具备了解所有客户端模型更新以及服务器聚合规则的能力。然而，高级攻击者并不具备改变聚合规则的能力，也不能篡改其他客户端训练过程和模型更新。这一威胁模型设定将有助于对算法在更复杂情景下的性能进行综合评估。

4 方法描述

本章提出了一种鲁棒联邦学习算法 FL-Sieve，旨在防御联邦学习中混合拜占庭客户端的攻击。FL-Sieve 的整体流程如图 2 所示。算法的总体执行流程详见算法 1。

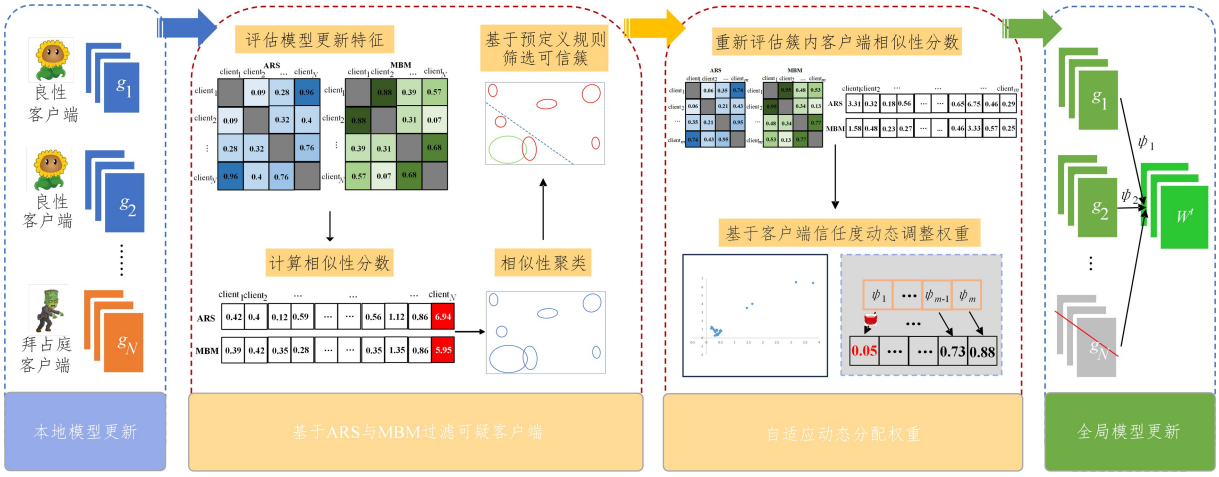


图 2 FL-Sieve 的框架图

Fig. 2 Framwork of FL-Sieve

算法 1 FL-Sieve-based Robust Federated Learning

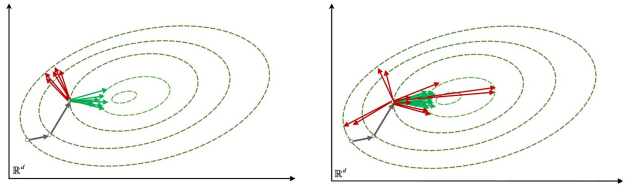
输入：迭代次数 T，数据集 D，客户端总数 N

输出：最终全局模型 w_T

1. FOR t IN 总迭代次数 TDO
2. 客户端：
3. FOR i IN 客户端总数 NDO
4. 客户端 i 接收服务器分发的全局模型 w_t，首轮为初始化模型 $w_0 \in \mathbb{R}^d$
5. 利用本地数据 D_i 对模型 w_t 进行训练，计算本地更新梯度
6. 将本地更新梯度 g_i 上传至服务器
7. ENDFOR
8. 服务器：
9. 收集所有客户端模型梯度 $G^t=[g_1,\cdots,g_N]$
10. 获取新一轮全局模型 w_{t+1}=FL-Sieve(G^t)
11. 将 w_{t+1} 分发到各个客户端
12. ENDFOR
13. 返回最终全局模型 w_T

4.1 梯度特征

在前面的讨论中提到经验丰富的攻击者可以利用精心构造的恶意梯度绕过仰仗单一度量的防御方法。例如，Min-Max/Min-Sum[36]攻击通过巧妙地缩放梯度，将其限制在良性梯度组的范围内，从而绕过基于欧氏距离的防御机制。此外，在传统单一攻击场景下，恶意更新与良性更新之间的边界明晰，类似于二元分类问题，如图 3(a)所示。这使得检测和分离这些攻击相对较为容易。但混合拜占庭攻击已经突破了这种限制，它将多种攻击技巧相结合，构建了更为复杂、更加难以侦测且更具威胁的攻击模式。攻击者在不同的环境和数据分布下执行混合拜占庭攻击，导致恶意梯度呈现多种特征。同时，由于非独立同分布数据和混合拜占庭攻击的双重挑战，不同模型之间的差异被进一步放大，使得区分良性模型和恶意模型的界限变得模糊，如图 3(b)所示。在面对如此复杂的混合拜占庭攻击时，现有的安全防御策略显然力有未逮。



(a)单拜占庭攻击 (b)混合拜占庭攻击

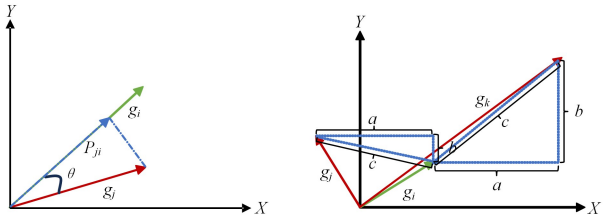
注：绿色更新是朝着真正目标前进的良性客户端所贡献的；红色更新是干扰全局模型聚合的拜占庭客户端提交的。

图 3 联邦学习中的客户端梯度更新

Fig. 3 Client gradient update in federated learning

为了有效地应对这一威胁，本文提出了两个度量指标：角幅相似度和模型边界测度。首先，本文设计了角幅相似度，它衡量了将客户端 i 的梯度更新转化成客户端 j 的梯度更新所需要的代价，旨在全面考量梯度更新之间的角度和幅度相似性。具体如图 4(a)所示，其中 g_i 与 g_j 分别表示客户端所提交的梯度更新，θ 表示两个更新间的余弦相似度，具体定义如式(3)所示，而 P_{ji} 表示 g_j 在 g_i 上的投影，如式(4)所示。式(7)描述了计算两个梯度之间角幅相似度的方法。ARS 不仅着重关注梯度方向的偏差，还综合考虑了大小的差异。这意味着即便攻击者尝试微调梯度的大小而不显著改变其方

向，ARS 依然可以有效地捕捉这种变化。



(a)角幅相似度　　(b)模型边界测度

图 4　二维空间上梯度特征的演示

Fig. 4　Demonstration of gradient features in two-dimensional space

然而，单一的指标容易被精心构建的恶意梯度规避。因此，为了更全面地评估梯度相似性和潜在的恶意变化，本文进一步设计了模型边界测度，用于评估模型梯度的边界特征信息，如图 4(b)所示，其中 c 表示客户端梯度更新间的欧氏距离，其计算方式如式(5)所示，而 a 与 b 为它们之间的曼哈顿距离，如式(6)所示。式(8)描述了计算每个客户端与其余客户端之间模型边界测度的方法，其中 α 为权重系数，用于调整两种距离度量的相对重要性。为了全面地获取模型边界特征的信息，本文采用了一种平衡的权重分配策略，对欧氏距离与曼哈顿距离进行 Min-Max 归一化后，设置 $\alpha=0.5$ 并将其相加。这样，两种距离得到相同的重视，反映出它们在构建新的距离度量时的同等关键性。MBM 综合了欧氏距离和曼哈顿距离，提供了更综合、全面的模型相似度评估指标。欧氏距离主要关注更新间的直线距离，但对于某些维度上的显著差异可能不够敏感。而曼哈顿距离则考虑了各坐标轴上的绝对轴差总和，因此更专注于梯度更新在每个维度上的差异，这有助于捕捉到那些可能被欧氏距离忽视的细微差异。这种双重视角的设计有助于更准确地捕获那些在某些特定维度上进行微小修改的攻击行为，进一步提高了对复杂数据模式的识别和描述的能力。此外，当模型边界测度相等时，角幅相似度指标也能够反映其偏离程度，两个指标之间相互补充，这种互补性有助于更全面地理解和评估客户端之间的模型梯度相似性。

$$\theta=\frac{g_i \cdot g_j}{\| g_i \| \cdot \| g_j \|} \tag{3}$$

$$P_{ji}=\frac{g_j \cdot g_i}{g_i \cdot g_i} \cdot g_i \tag{4}$$

$$c=\sqrt{\sum_{m=1}^{n}(g_{i_m}-g_{j_m})^2} \tag{5}$$

$$a+b=\sum_{m=1}^{n}|g_{i_m}-g_{j_m}| \tag{6}$$

$$ARS_{i,j}=\left(1-\frac{g_i \cdot g_j}{\| g_i \| \cdot \| g_j \|}\right) \cdot |\| g_i \|-\| P_{ji} \||\tag{7}$$

$$MBM_{i,j}=\alpha\sqrt{\sum_{m=1}^{n}(g_{i_m}-g_{j_m})^2}+(1-\alpha)\sum_{m=1}^{n}|g_{i_m}-g_{j_m}| \tag{8}$$

ARS 与 MBM 的联合应用带来了多重优势。在应对复杂的攻击策略时，单一的度量标准容易受到攻击者精心构造的攻击策略的规避。然而，通过从两个不同的视角观察相同梯度的特征，可以大幅减少潜在的盲点，显著降低攻击者绕过的风险。ARS 和 MBM 综合考量了梯度更新的方向、幅度以及各维度上的差异等信息，有助于更全面地了解梯度的特征，提高对异常梯度的检测能力。此外，ARS 与 MBM 相互验证对异常值的识别，如果攻击者试图规避 MBM 的检测，ARS 可能会揭露其行为，反之亦然。这种多维度、多角度的检测策略不仅提高了检测的准确性和可靠性，也大大降低了误报和漏报的风险。在定义了这些梯度的特征后，将其用于识别恶意梯度，目标是检测梯度中的异常值并将其滤除。

4.2　算法详细步骤

在客户端的训练过程中，FL-Sieve 与标准的联邦学习方法相似，如算法 1 中的步骤 3－7 所示。各客户端会首先获取服务器分发的全局模型，接着从本地数据集中随机选取一个数据批次进行训练，随后将得到的梯度更新发送给服务器。

在服务器端，FL-Sieve 的执行步骤如下：

1)服务器接收客户端上传的模型梯度 $G=[g_1,\cdots,g_N]$。

2)服务器评估客户端上传的模型梯度特征。

传统的方法通常将模型梯度与中值或平均值进行比较，但这种方式容易被攻击者利用，因为他们可能会刻意调整模型梯度，使其特征值更接近均值或中心值以规避检测。此外，仅将模型梯度与部分客户端或中心值进行比较可能导致某些客户端主导整个决策，而其他客户端的贡献被忽略。为了避免这些问题，本文采用了一种更全面的方法，即取每个客户端与其余客户端的特征相似性之和作为每个客户端模型的梯度特征，以反映其与整体之间的关系。算法 2 第 2－4 行描述了每轮计算客户端之间的角幅相似度和模型边界测度，生成相似度矩阵的过程。其中，$ARS_{i,j}^t$ 与 $MBM_{i,j}^t$ 描述了第 t 轮第 i 个客户端与第 j 个客户端间的模型梯度特征相似性。算法 2 第 6－9 行描述了计算每个客户端相似性分数的步骤。

3)依据梯度特征的相似性进行聚类。

在进行相似性聚类分析时，采用哪种聚类算法是一个挑战。针对这一问题，本文在 MNIST 数据集上进行了比较研究，考察了 K 均值聚类(K-means)、基于密度的聚类(Density-Based Spatial Clustering of Applications with Noise，DBSCAN)以及高斯混合模型聚类(Gaussian Mixture Model，GMM)，测试结果显示 GMM 表现最为优异。此外，确定聚类中的簇数量同样是一大挑战。错误的簇数量选择可能导致过度分割或过度汇总，对最终结果造成不利影响。为此，本文首先预设一个簇数量的范围，然后利用贝叶斯信息准则(Bayesian Information Criterion，BIC)对聚类结果进行评估，最终选择 BIC 值最小的情况对应的簇数量。BIC 的计算方式如式(9)所示：

$$BIC=-2\ln(L)+d\ln(N) \tag{9}$$

其中，L 为 GMM 模型中最大似然函数的最大值，N 和 d 分别表示样本数量和维度。经过测试，将簇的数量范围设定为[1，6]，这不仅有助于寻找最优簇数量，还避免了过大的搜索范围引发的计算耗时问题。

4)将模型梯度划分到各簇。

根据步骤 3 中的聚类结果，将所有模型梯度划分到各个簇中，得到模型梯度划分结果 $G'=[G_i,i\in C]$。

5)计算各个簇的评价指标。

对于聚类结果中的各个簇，计算各个簇内客户端特征的平均值，从而得到簇特征集合 $S=[S_i,i\in C]$。S 中的每个值代表各个簇的衡量指标。

6)筛选出潜在良性客户端集群。

在混合拜占庭攻击背景下，拜占庭客户端针对全局模型展开协同攻击，这种干扰会导致恶意客户端的梯度数据分布与良性客户端存在偏差。这种差异可以被 ARS 与 MBM 这两个指标所捕获，表现为恶意客户端簇的模型梯度特征分数往往显著超过正常客户端簇。因此，首先需要排除一些偏差较大的客户端集群，防止它们对系统的更新方向、幅度和决策边界产生重大影响。具体而言，剔除相似性分数高于平均水

平的客户端集群，同时保留与全局模型更新趋势保持一致的客户端簇。而在筛选后的客户端簇中，最大的簇往往包含着最多的良性客户端，故本文选择低于簇均值且包含元素最多的簇作为潜在的良性客户端集群集群，以聚合其模型更新。这一策略能够在一定程度上削弱恶意攻击对全局模型的影响，确保联邦学习过程中的鲁棒性和安全性。通过这种方式，得到了本轮中更新的全局梯度 G_{update}。

7）重新分配簇内客户端权重。

在上述过程中，尽管已经选择了适合的簇用于聚合，但考虑到本文涉及混合拜占庭攻击的场景，所选簇内仍可能存在极少数的逃逸拜占庭客户端。因此如何处理簇内可能存在的少量恶意客户端，以减轻它们对全局模型的影响成为了另一个需要解决的问题。本文设计了一个自适应动态分配权重的策略，旨在根据每个客户端的信任度智能地分配学习资源，实现智能资源分配，从而进一步提升防御效果和系统鲁棒性。

具体而言，由于只有在模型梯度特征相似的情况下，客户端才会被归入同一簇，因此，恶意客户端为了破坏全局模型的性能，通常会在决策边界附近引入扰动。更具体地说，即使同一簇内存在少量拜占庭客户端，簇中心附近的客户端往往会是良性客户端，而靠近簇内决策边界附近的可能是偏差较大的良性客户端甚至是拜占庭客户端。为了最大化地抑制这些潜在的恶意梯度，本文采用了一种基于客户端到簇中心距离的动态权重分配策略，通过敏感系数来调节权重随距离的衰减速度，确保那些对模型更新贡献较大的客户端能获得更高的权重。这意味着客户端距离簇中心越远，其信任度越低，获得的权重就越小，从而有助于把握模型的整体更新安全，强化模型的稳健性。如算法 2 伪代码中的 23－25 行所示，其中 $center=\min(X')$ 代表簇的中心，β 是敏感系数，而 distances 表示客户端与簇中心之间的距离组。基于这一动态权重分配策略，最终获得了本轮模型更新的权重 ψ。

8）利用 SGD 算法更新下一轮迭代的模型。

9）经过 T 轮之后，返回最终全局模型 w_T。

算法 2 FL-Sieve Function

输入：客户端模型梯度集 G^t

输出：全局模型梯度 w_{t+1}

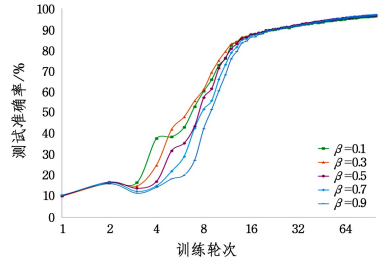
1. Step 1：计算所有模型梯度特征
2. FOR i,j IN(len(G^t)) Do
3. $ARS^t \leftarrow ARS^t_{i,j}$
4. $MBM^t \leftarrow MBM^t_{i,j}$
5. ENDFOR
6. FOR i,j IN(len(G^t)) Do：
7. $x^{(i)}_{ARS} \leftarrow \sum_{j,j\neq i}^{N} ARS^t_{i,j}$
8. $x^{(i)}_{MBM} \leftarrow \sum_{j,j\neq i}^{N} MBM^t_{i,j}$
9. $x'^{(i)} \leftarrow (x^{(i)}_{ARS}, x^{(i)}_{MBM})$
10. ENDFOR
11. $X \leftarrow [x'^{(1)}, x'^{(2)}, \cdots, x'^{(N)}]^T$
12. Step 2：筛选可信客户端
13. 依据特征相似性进行聚类
    C=Clustering(X)
14. 将各客户端的模型梯度分配到对应的簇中
    $G'=[G_i, i\in C]$
15. 计算各个簇内相似度均值 $S=[S_i, i\in C]$
16. $G_{list}=[]$.
17. FOR i,S_i IN ENUMERATE (S)
18. IF $S_i \leqslant$ mean(S)
19. G_{list}.append(G_i)
20. ENDFOR
21. 保留潜在良性客户端集群
    G_{update}=maxgroup(G_{list})
22. Step 3：自适应动态分配权重
23. 重新评估 G_{update} 簇内客户端梯度特征，得到 X'
24. 评估簇内客户端到簇中心的距离
    distances=|X'－center|
25. 动态计算权重 $\psi=\exp(-\beta * distances)$
26. $w_{t+1}=w_t-\frac{\sum(\psi * G_{update})}{\sum\psi}$
27. 返回全局模型 w_{t+1}

5 实验

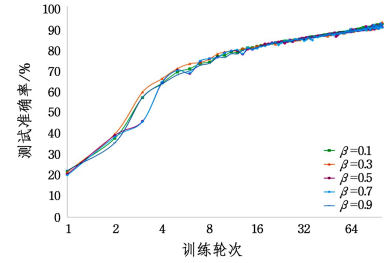
5.1 实验设置

1）系统设置

实验通过在不同的数据集上使用多种网络模型进行训练，以展示本文提出的聚合算法的普适性。具体来说，针对 MNIST[39]，Fashion-MNIST[40] 和 CIFAR-10[41] 数据集，分别采用 LeNet，CNN 和 VGG-9 模型进行训练。默认情况下，总共设置 100 个联邦学习客户端，拜占庭客户端的设置参考了后续拜占庭攻击设置下的场景 1 和场景 2。在所有实验中，每个训练过程运行 100 个 epoch，并将局部迭代始终设置为 1。此外，本文还在数据集上测试了 FL-Sieve 在面对 49% 混合拜占庭客户端攻击时，不同敏感系数设置下的全局模型性能。如图 5 所示，当敏感系数设定为 0.3 时，全局模型的收敛速度和最终性能相对较优。因此，实验将敏感系数 β 设置为 0.3。更多参数细节如表 1 所列。



(a) MNIST



(b) Fashion-MNIST

图 5 49%混合拜占庭客户端攻击下，FL-Sieve 在各个数据集上不同敏感系数设置下的性能表现

Fig. 5 Performance of FL-Sieve algorithm with different sensitivity coefficient settings on various datasets under 49% hybrid Byzantine client attack

2）数据集

本文应用了基于 PS 架构的联邦学习框架，即存在一个服务器和多个客户端。在任务学习方面，本文使用了 3 个计算机视觉领域的数据集。对于每个数据集，通过 Dirichlet 分

布[42]将训练数据不均等地分发给客户端，以模拟真实的联邦学习系统中各个客户端持有的非独立同分布数据集，分布 $X\sim Dir(\gamma,N)$ 中的参数 γ 越小，客户端本地数据的非独立同分布程度越高。3 种数据集描述如下：

(1)MNIST。MNIST 是一个手写数字图像数据集，包含 60000 张训练图片和 10000 张测试图片，每个图片是一张 28×28 的灰度图像，一共 10 个类别。在 Non-IID 设置中，通过 Dirichlet 分布对数据进行分发。具体而言，在 MNIST 中，每个类有 6000 个训练数据，通过分布 $X\sim Dir(0.9,N)$ 得到 M 个 N 维的 Dirichlet 分布。根据这 M 个分布，按数据类依次向 N 个客户分发训练数据，数据量达到平均值的客户端将停止接收数据。因此，除了数据类别比例不同之外，客户端的数据量也可能不同。

(2)Fashion-MNIST。Fashion-MNIST 是一个服饰图像分类数据集，涵盖了 10 种类别共 70000 个不同商品的正面图片。训练集和测试集划分与 MNIST 一致，其中 60000 张图片用于训练，10000 张图片用于测试，且图片是 28×28 的灰度图。与 MNIST 的处理一致，Fashion-MNIST 根据分布 $X\sim Dir(0.9,N)$ 分发数据给客户端。

(3)CIFAR-10。CIFAR-10 是一个自然图像分类数据集，包含 10 个类别共 60000 张图片，其中 50000 张用于训练，10000 张用于测试，图片是 3×32×32 的彩色图。与 MNIST 的处理方式相同，CIFAR-10 根据分布 $X\sim Dir(0.9,N)$ 分发数据给客户端。

3)拜占庭攻击

实验进行两类拜占庭攻击，包括数据投毒攻击和模型投毒攻击。

(1)符号翻转攻击[8]。符号翻转攻击客户端 m 首先训练出真实的本地模型梯度 g_m，然后将其翻转为 $g_m=-1\times g_m$ 后提交给参数服务器。

表 1　实验数据集和模型参数设置

Table 1　Experimental datasets and model parameter settings

数据集	类别	特征	模型	学习率	批大小	动量	训练轮次
MNIST	10	784	LeNet(2 个卷积层，3 个全连接层)	0.01	64	0.5	100
Fashion-MNIST	10	784	CNN(6 个卷积层，2 个全连接层)	0.01	64	0.5	100
CIFAR-10	10	3072	VGG-9(6 个卷积层，3 个全连接层)	0.001	32	0.9	100

(2)标签翻转攻击。标签翻转攻击客户端在训练过程中翻转本地样本标签，将每个训练样本的标签从 i 翻转到 $C-1-i$，其中 C 是标签的总类别，$i\in\{0,1,\cdots,C-1\}$。

(3)随机参数攻击[37]。随机参数攻击客户端发送随机化的模型梯度，由多维高斯分布 $N(\mu,\sigma^2\mathbf{I})$ 生成。在实验中取 $\mu=(0,\cdots,0)\in\mathbb{R}^d$ 和 $\sigma=0.5$ 进行攻击。

(4)噪声干扰攻击。噪声干扰攻击客户端通过在诚实的梯度中加入高斯噪声 $g_m=g_m+\mathcal{N}(\mu,\sigma^2\mathbf{I})$，以发送扰动梯度，破坏模型准确性。实验采用与随机参数攻击相同的高斯分布参数。

(5)同值攻击[8]。同值攻击客户端将本地模型更新都修改为 $g_m=1$ 并提交给参数服务器，以干扰全局模型聚合结果。

(6)Zero-gradient。Zero-gradient 客户端 $m(m\in B)$ 通过发送根据式(10)所构造的更新使得所有客户端更新总和为零。其中 g_m^t 表示第 t 轮 Zero-gradient 客户端 m 提交的更新，B 是 Zero-gradient 客户端的数量，式(11)表示本轮其余客户端的更新聚合。

$$g_m^t=-\frac{1}{B}\sum_{m'\in N-\mathfrak{B}}g_{m'}^t \tag{10}$$

$$\sum_{m'\in N-\mathfrak{B}}g_{m'}^t \tag{11}$$

(7)Little is Enough[13]。Little is Enough 攻击旨在通过隐藏在良性梯度方差中的微小扰动来阻止模型收敛。客户端首先估计坐标平均值(μ_j)和标准差(σ_j)，随后发送恶意梯度至服务器。其中恶意梯度构造如式(12)所示：

$$g_m=\mu_j-z\cdot\sigma_j,j\in[d] \tag{12}$$

其中，正攻击因子 z 取决于客户端的总数和拜占庭分数，d 是模型维度。其目的是绕过基于坐标中值和修剪均值的防御。在实验中设置默认值 $z=0.3$。

(8)Min-Max/Min-Sum[36]。恶意梯度是其余良性客户端集合的扰动版本，如式(13)所示，其中 ∇^p 是扰动向量，γ 是缩放系数，这两种攻击如式(14)和式(15)所示。第一个 Min-Max 攻击使恶意梯度靠近良性梯度团，而 Min-Sum 攻击确保恶意梯度到所有良性梯度的距离平方和的上界为任何良性梯度到其他良性梯度的距离平方和。为了使攻击影响最大化，所有恶意梯度保持相同。默认情况下，选择 $-std(g^{\{i\in[n]\}})$，即逆标准差作为扰动向量 ∇^p 的值。

$$g_m=f_{\text{avg}}(g^{\{i\in[n]\}})+\gamma\nabla^p \tag{13}$$

$$\arg\max_{\gamma}\max_{i\in[n]}\|g_m-g^{(i)}\|\leqslant\max_{i,j\in[n]}\|g^{(i)}-g^{(j)}\| \tag{14}$$

$$\arg\max_{\gamma}\sum_{i\in[n]}\|g_m-g^{(i)}\|^2\leqslant\max_{i\in[n]}\sum_{j\in[n]}\|g^{(i)}-g^{(j)}\|^2 \tag{15}$$

为了全面模拟联邦学习中真实情况下的拜占庭攻击，根据第 3 节中的攻击模型设定，本文进行了两种不同攻击场景的实验。

场景 1(混合拜占庭攻击)　存在 n 个恶意攻击者，他们共同控制着 m 个不同类型的恶意客户端，包括 m_1 个符号翻转攻击客户端、m_2 个标签翻转攻击客户端、m_3 个随机参数攻击客户端、m_4 个噪声干扰攻击客户端、m_5 个同值攻击客户端、m_6 个 Zero-gradient 攻击客户端、m_7 个 Little is Enough 攻击客户端、m_8 个 Min-Max 攻击客户端和 m_9 个 Min-Sum 攻击客户端。这些恶意客户端共同对联邦学习系统进行攻击。其中 $m_i\in\{1,41\}$ 随机选取并且 $m=\sum_{i=1}^{9}m_i<50$，良性客户端的数量为 100 减去拜占庭客户端的数量。

场景 2(单攻击)　存在单恶意攻击者控制一组客户端，采取任意同种攻击手段对联邦学习系统发起攻击。

虽然一些文献中假设攻击者能够控制更大比例的恶意客户端，但本文认为在真实的联邦学习场景中单个恶意攻击者控制超过 20%的客户端是不太可能的。例如，Gboard[5,40,43]应用拥有数百万的用户，即使攻击者只控制其中一小部分用户的设备，也需要耗费巨大的资源进行入侵，这在实际情况下是不太现实的。在恶意客户端选择方面，为了更全面地考察算法在复杂环境下的抵御能力，并模拟实际可能面临的严峻挑战，本文在场景 1(混合拜占庭攻击)中假设所有攻击手段都存在，包括符号翻转、标签翻转、随机参数攻击等不同类型

的恶意行为；在恶意客户端数量方面，为了全面评估系统在不同攻击强度下的防御效能，本文实验涵盖了恶意客户端在系统中所占比例从20%～49%的多个级别，旨在模拟在现实环境中可能遭遇的从轻微到严重的不同规模攻击者的威胁。此举意在全面评估系统在不同攻击密度条件下的抗压能力。这样不仅能够评估算法对于少量恶意客户端的鲁棒性，还能有效检验在恶意客户端数量显著上升的情况下，系统稳定性和安全性的保持程度。而在单攻击场景下，本文假设恶意客户端占客户端总数的20%。

4）评估指标

对于拜占庭攻击，攻击者的目标是降低模型在测试集上的准确率。因此，本文使用全局模型的主任务准确率（Main Task Accuracy，MTA）来评估不同聚合算法优劣。MTA的定义如式(16)所示，其中，$|S_{\text{clean}}|$和$|D_{\text{clean}}|$分别代表预测准确的良性样本数和总的良性样本数。为了评估模型性能，本文在一系列固定次数的迭代训练中使用测试精度来进行评估。考虑到训练的不稳定性以及在面对强攻击时可能出现的准确性波动，本文在每个训练周期结束时进行测试，每个实验重复3次，并取平均结果。更高的主任务准确率表示相应的聚合算法在抵御攻击时表现出更强的鲁棒性。

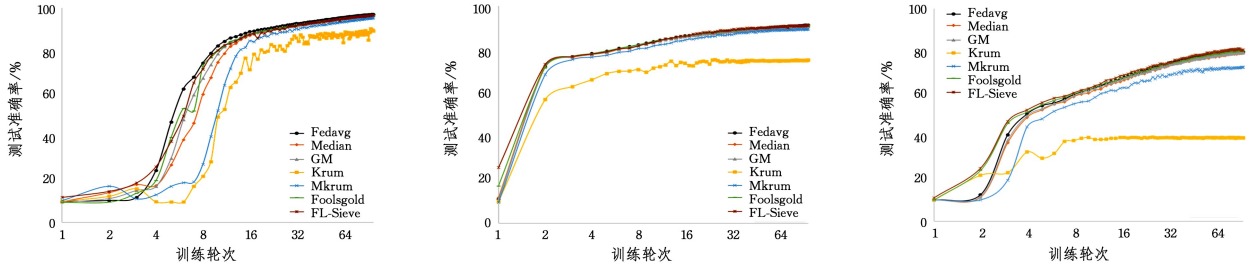
$$MTA=\frac{|S_{\text{clean}}|}{|D_{\text{clean}}|}\times 100\% \tag{16}$$

5）基准算法

在实验中，除了联邦学习中经典的Fedavg[1]算法外，还对比了5种经典的抵抗拜占庭攻击的聚合算法：Coordinate-wise median[16]（简称Median）、Geometric median[17]（简称GM）、Krum、Multi-Krum[15]（简称Mkrum），以及Foolsgold[18]。

5.2 实验结果

1）保真性。如图6所示，在3个数据集上，本文提出的聚合算法FL-Sieve与Median，Foolsgold和GM在未受到拜占庭攻击影响的情况下，均展现出与基准算法Fedavg相近的性能。尽管Mkrum具有较快的收敛速度，但其只能收敛到模型的次优解。此外不难看出，Krum算法在训练过程中为确保全局模型的鲁棒性会牺牲一定精度。总而言之，本文提出的聚合算法在各实验场景下都具有较强的保真性。



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10

图6 无攻击情况下各聚合算法的性能表现

Fig. 6 Performance of each aggregation algorithm without attack

2）鲁棒性。如表2所列。在MNIST数据集上，FL-Sieve在不同比例的混合拜占庭客户端攻击下均保持了约96%的高准确率，防御表现出色。相比之下，Fedavg和Foolsgold均出现了防守不足的问题。Median和GM在混合拜占庭客户端比例较低（20%和30%）时的准确率也很高，某些情况下甚至超越了FL-Sieve。但当恶意客户端比例逐渐增加到49%时，准确率急剧下降。Krum和Mkrum中也可观察到类似的现象，在有少数混合拜占庭客户端的情况下，它们表现出一定程度的鲁棒性，但这是以牺牲模型性能为代价，且它们在各比例下的准确率均低于FL-Sieve。

在Fashion-MNIST数据集中，FL-Sieve的准确率在各个混合拜占庭客户端比例下均接近91%。与此类似，Median和GM的准确率在混合拜占庭客户端比例为20%和30%时都在90%左右，但当比例逐渐增加到49%时，准确率大幅下降。其他算法，如Fedavg，Krum，Mkrum和Foolsgold的准确率均显著低于FL-Sieve。

在CIFAR-10数据集中，FL-Sieve的准确率持续稳定在79%～80%之间。与此相比，其他算法在此数据集上的表现始终都不及FL-Sieve。特别是Krum算法，其准确率在混合拜占庭客户端数量持续增加的情况下始终维持在40%以下水平。值得注意的是，Foolsgold算法一直陷入了防御失败的状态，未能取得令人满意的效果。

总体来看，FL-Sieve在三大数据集上均表现出强大的稳健性，特别是在混合拜占庭客户端比例较高时。虽然其他算法在某些情境下可能有出色的表现，但它们的鲁棒性和稳定性并不如FL-Sieve，尤其是在高比例恶意客户端的情况下。

表2 混合拜占庭攻击场景下各算法的实验结果

Table 2 Experimental results of various algorithms in hybrid Byzantine attack scenarios

数据集	算法	20%拜占庭客户端	30%拜占庭客户端	40%拜占庭客户端	49%拜占庭客户端
MNIST	Fedavg	9.79	10.95	11.34	9.82
	Median	96.44	**96.33**	45.65	41.23
	GM	**96.72**	96.22	43.19	44.49
	Krum	88.36	86.48	10.09	9.82
	Mkrum	92.99	92.82	11.94	14.32
	Foolsgold	14.21	15.36	12.7	14.03
	FL-Sieve	96.64	96.3	**95.58**	**96.33**
Fashion-MNIST	Fedavg	9.8	10.09	10.01	9.94
	Median	**90.91**	90.93	44.15	41.04
	GM	90.85	90.99	55.37	46.26
	Krum	75.04	74.95	15.39	14.15
	Mkrum	87.08	87.31	15.87	16.10
	Foolsgold	14.55	16.04	10.02	9.83
	FL-Sieve	90.87	**91.07**	**91.11**	**90.39**
CIFAR-10	Fedavg	10.93	9.15	9.45	10.18
	Median	**79.76**	78.58	37.05	33.76
	GM	78.86	77.67	45.23	38.36
	Krum	39.44	39.35	12.01	10.06
	Mkrum	72.4	72.22	14.37	16.49
	Foolsgold	16.49	18.2	12.84	13.78
	FL-Sieve	79.33	**79.94**	**80.38**	**80.16**

3)泛化性。先前实验主要关注 Non-IID 数据分布和混合拜占庭攻击场景。为了全面评估 FL-Sieve 应对各种拜占庭攻击的泛化能力，本文进一步扩展了评估的范围，考察了在 IID 和 Non-IID 两种数据分布场景下，FL-Sieve 在应对单一攻击时的性能表现。评估结果如表 3 所列，从表中可以清楚地看出，在 3 个不同的数据集中，无论攻击者采用何种攻击策略，FL-Sieve 均表现出卓越的能力，能够准确地识别和过滤掉拜占庭客户端，并持续获得较高的主任务准确率。以上结果验证了该算法的通用性，以及在涉及单一攻击情境下的适用性。

表 3　FL-Sieve 在 IID 和 Non-IID 情景下应对单攻击的实验结果

Table 3　Experimental results of FL-Sieve for single attacks in IID and Non-IID scenarios

数据集	数据类型	简单攻击				高级攻击				
		标签翻转	随机梯度	噪声干扰	符号翻转	同值攻击	零梯度攻击	LIE	Min-Max	Min-Sum
MNIST	IID	97.21	97.13	97.12	97.07	96.97	97.04	96.71	96.82	96.77
	Non-IID	96.48	96.84	96.37	96.22	96.04	95.77	96.23	96.3	96.48
Fashion-MNIST	IID	92.77	92.81	91.98	91.69	92.04	91.38	91.29	91.83	92.2
	Non-IID	91.11	91.19	90.97	90.84	90.87	90.95	90.88	91.02	91.11
CIFAR-10	IID	81.13	81.4	82.19	82.26	82.83	80.16	80.88	81.21	81.05
	Non-IID	80.55	80.6	81.03	81.09	81.16	80.73	80.97	80.85	80.55

结束语　本文提出了一种基于多指标检测与自适应动态权重分配的新型拜占庭鲁棒聚合算法 FL-Sieve。FL-Sieve 首先评估客户端提交的梯度之间的特征相似性，然后进行聚类，确保相似的节点被归入同一簇；随后 FL-Sieve 基于筛选规则优选潜在良性客户端；最后 FL-Sieve 依据客户端的信任度动态分配权重，以应对潜在的逃逸拜占庭客户端带来的不利影响，从而进一步增强防御效果和系统的鲁棒性。本文进行了全面的实验来评估 FL-Sieve 在各种数据集和攻击场景下的有效性。结果表明，FL-Sieve 在各种复杂攻击场景下均表现出良好的性能。

本文提出的算法虽显著增强了联邦学习模型的鲁棒性，但当前的研究重点主要集中在混合拜占庭攻击的防御上。因此，下一步的研究将关注如何抵御混合后门攻击，以进一步提高联邦学习模型的安全性。此外，未来的工作还将探索更多样化的特征，并进一步改进特征融合策略，以提高对恶意客户端的精准识别能力。

参 考 文 献

[1] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]// Artificial Intelligence and Statistics. PMLR, 2017: 1273-1282.

[2] LU Z, KUO-HUI Y, GERHARD H, et al. Security and Privacy for the Industrial Internet of Things: An Overview of Approaches to Safeguarding Endpoints [J]. IEEE Signal Processing Magazine, 2018, 35(5): 76-87.

[3] ZHOU C X, SUN Y, WANG D G, et al. Survey of federated learning research[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 77-92.

[4] 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.

[5] HARD A, RAO K, MATHEWS R, et al. Federated Learning for Mobile Keyboard Prediction[J]. arXiv: 1181.03604, 2018.

[6] LEROY D, COUCKE A, LAVRIL T, et al. Federated learning for keyword spotting[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019). IEEE, 2019: 6341-6345.

[7] LIU Y, HUANG A, LUO Y, et al. Fedvision: An online visual object detection platform powered by federated learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 13172-13179.

[8] LI L, XU W, CHEN T, et al. RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 1544-1551.

[9] WU Z, LING Q, CHEN T, et al. Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks [J]. IEEE Transactions on Signal Processing, 2020, 68: 4583-4596.

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

[11] FANG M, CAO X, JIA J, et al, Local model poisoning attacks to byzantine-robust federated learning[C]// 29th USENIX Security Symposium (USENIX Security 20). 2020: 1605-1622.

[12] KAIROUZ P, MCMAHAN H B, AVENT B, et al. Advances and open problems in federated learning[J]. Foundations and Trends © in Machine Learning, 2021, 14(1/2): 1-210.

[13] BARUCH G, BARUCH M, GOLDBERG Y. A little is enough: Circumventing defenses for distributed learning [C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 8635-8645.

[14] BAGDASARYAN E, VEIT A, HUA Y, et al, How to backdoor federated learning[C]// International Conference on Artificial Intelligence and Statistics. PMLR, 2020: 2938-2948.

[15] BLANCHARD P, EL MHAMDI E M, GUERRAOUI R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 118-128.

[16] YIN D, CHEN Y, KANNAN R, et al. Byzantine-robust distributed learning: Towards optimal statistical rates[C]// International Conference on Machine Learning. PMLR, 2018: 5650-5659.

[17] CHEN Y, SU L, XU J. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent [C]// Proceedings of the ACM on Measurement and Analysis of Computing Systems. 2017: 1-25.

[18] FUNG C, YOON C J M, BESCHASTNIKH I. The limitations of

federated learning in sybil settings[C]// 23rd International Symposium on Research in Attacks, Intrusions and Defenses ({RAID} 2020). 2020: 301-316.

[19] LI S, CHENG Y, WANG W, et al. Learning to detect malicious clients for robust federated learning[J]. arXiv: 2002. 00211, 2020.

[20] XIE C, KOYEJO S, GUPTA I. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance[C]// International Conference on Machine Learning. PMLR, 2019: 6893-6901.

[21] RODRÍGUEZ-BARROSO N, MARTÍNEZ-CÁMARA E, LUZÓN M V, et al. Dynamic defense against byzantine poisoning attacks in federated learning[J]. Future Generation Computer Systems, 2022, 133: 1-9.

[22] GUERRAOUI R, ROUAULT S. The hidden vulnerability of distributed learning in Byzantium[C]// International Conference on Machine Learning. PMLR, 2018: 3521-3530.

[23] KHAZBAK Y, TAN T, CAO G. MLGuard: Mitigating poisoning attacks in privacy preserving distributed collaborative learning[C]// 2020 29th International Conference on Computer Communications and Networks (ICCCN). IEEE, 2020: 1-9.

[24] LU Y, FAN L. An efficient and robust aggregation algorithm for learning federated cnn[C]// Proceedings of the 2020 3rd International Conference on Signal Processing and Machine Learning. 2020: 1-7.

[25] YU L, WU L. Towards byzantine-resilient federated learning via group-wise robust aggregation[J]. Federated Learning: Privacy and Incentive, 2020, 12500: 81-92.

[26] YANG H, ZHANG X, FANG M, et al. Byzantine-resilient stochastic gradient descent for distributed learning: A lipschitz-inspired coordinate-wise median approach[C]// IEEE 58th Conference on Decision and Control (CDC 2019). IEEE, 2019: 5832-5837.

[27] WANG Y, ZHU T, CHANG W, et al. Model poisoning defense on federated learning: A validation based approach[C]// International Conference on Network and System Security. Cham: Springer International Publishing, 2020: 207-223.

[28] TAN J, LIANG Y C, LUONG N C, et al. Toward smart security enhancement of federated learning networks[J]. IEEE Network, 2021, 35(1): 340-347.

[29] CHEN Z, TIAN P, LIAO W, et al. Zero knowledge clustering based adversarial mitigation in heterogeneous federated learning [J]. IEEE Transactions on Network Science and Engineering, 2020, 8(2): 1070-1083.

[30] KIM W, LIM H. FedCC: Federated Learning with Consensus Confirmation for Byzantine Attack Resistance (Student Abstract)[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 12981-12982.

[31] CAO X, LAI L. Distributed gradient descent algorithm robust to an arbitrary number of byzantine attackers[J]. IEEE Transactions on Signal Processing, 2019, 67(22): 5850-5864.

[32] GU Z, HE L, LI P, et al. FREPD: A Robust Federated Learning Framework on Variational Autoencoder[J]. Comput. Syst. Sci. Eng., 2021, 39(3): 307-320.

[33] ZHAI K, REN Q, WANG J, et al. Byzantine-robust federated learning via credibility assessment on Non-IID data[J]. Mathematical Biosciences and Engineering, 2022, 19(2): 1659-1676.

[34] SHAFAHI A, HUANG W R, NAJIBI M, et al. Poison frogs! Targeted clean-label poisoning attacks on neural networks[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 6106-6116.

[35] SHEJWALKAR V, HOUMANSADR A. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning[C]// NDSS. 2021.

[36] XIE C, KOYEJO O, GUPTA I. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation[C]// Uncertainty in Artificial Intelligence. PMLR, 2020: 261-270.

[37] LIN J, DU M, LIU J. Free-riders in federated learning: Attacks and defenses[J]. arXiv: 1911. 12560, 2019.

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

[39] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[C]// Proceedings of the IEEE. 1998: 2278-2324.

[40] XIAO H, RASUL K, VOLLGRAF R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms [J]. arXiv: 1708. 07747, 2017.

[41] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images [DB/OL]. https://learning2hash.github.io/publications/cifar2009learning/.

[42] HSU T M H, QI H, BROWN M. Measuring the effects of non-identical data distribution for federated visual classification[J]. arXiv: 1909. 06335, 2019.

[43] DAVENPORT C. Gboard passes one billion installs on the play store[J/OL]. https://www.androidpolice.com/2018/08/22/gboard-passes-one-billion-installs-play-store, accessed: 2023-12-2.

WANG Chundong, born in 1969, Ph.D, professor, Ph.D supervisor, is a senior member of CCF (No. 16230M). His main research interests include network and information security, artificial intelligence technology and edge computing.