抗拜占庭攻击的梯度净化联邦自适应学习算法

杨辉[1]，邱子游[1]，李中美[2]，朱建勇[1]
（1.华东交通大学电气与自动化工程学院，江西 南昌 330013；2.华东理工大学信息科学与工程学院，上海 200237）

摘　要：在工业大数据之下，数据安全和隐私保护是关键挑战之一。传统的数据共享和模型训练方法在应对数据泄露和恶意攻击（尤其是复杂的拜占庭攻击和投毒攻击）时效果有限，因为传统联邦学习通常假定所有参与方都是可信的，这使得模型在遭遇投毒攻击时性能显著下降。为解决这个问题，本文提出一种抗拜占庭攻击的梯度净化联邦自适应学习算法，通过滑动窗口梯度过滤器和符号聚类过滤器识别恶意梯度，滑动窗口方法检测异常梯度，而符号聚类则根据梯度方向一致性筛选出偏离的对抗性梯度，经过过滤后，使用基于权重的自适应聚合规则对剩余的可信梯度进行加权聚合，动态调整参与方梯度的权重，降低恶意梯度的影响，从而增强模型的鲁棒性。实验结果显示，尽管新型投毒攻击的强度更高，但所提算法能有效防御这些攻击且减轻模型性能的损失。相比于传统防御算法，所提算法不仅提高了模型的准确性，还提升了其安全性。

关键词：联邦学习；拜占庭攻击；投毒攻击；模型鲁棒性；工业大数据
中图分类号：TN92
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024209

Gradient purification federated adaptive learning algorithm for Byzantine attack resistance

YANG Hui[1], QIU Ziyou[1], LI Zhongmei[2], ZHU Jianyong[1]
1. School of Electrical and Automation Engineering, East China Jiaotong University, Nanchang 330013, China
2. School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

Abstract: In the context of industrial big data, data security and privacy are key challenges. Traditional data-sharing and model-training methods struggle against risks like Byzantine and poisoning attacks, as federated learning typically assumes all participants are trustworthy, leading to performance drops under attacks. To address this, a Byzantine-resilient gradient purification federated adaptive learning algorithm was proposed. The malicious gradients were identified through a sliding window gradient filter and a sign-based clustering filter. The sliding window method detected anomalous gradients, while the sign-based clustering filter selected adversarial gradients based on the consistency of gradient directions. After filtering, a weight-based adaptive aggregation rule was applied to perform weighted aggregation on the remaining trustworthy gradients, dynamically adjusting the weights of participant gradients to reduce the impact of malicious gradients, thereby enhancing the model's robustness. Experimental results show that despite the increased intensity of new poisoning attacks, the proposed algorithm effectively defends against these attacks while minimizing the loss in model performance. Compared to traditional defense algorithms, it not only improves model accuracy but also enhances its security.

Keywords: federated learning, Byzantine attack, poisoning attack, model robustness, industrial big data

收稿日期：2024–08–16
通信作者：朱建勇，zhujyemail@163.com
基金项目：国家自然科学基金资助项目（No.62363010, No.61733005）；工业控制技术国家重点实验室开放课题基金资助项目（No.ICT2024B50）
Foundation Items：The National Natural Science Foundation of China (No.62363010, No.61733005), The Open Research Project of the State Key Laboratory of Industrial Control Technology of China (No.ICT2024B50)

0 引言

在工业大数据时代，私有数据的分散性带来了数据孤岛问题[1]，限制了数据的有效利用，特别是在处理如稀土金属等战略性资源时，数据的隐私和安全保护显得尤为重要。联邦学习作为一种新兴的分布式机器学习范式[1-2]，为解决这一问题提供了有效途径。通过允许各参与方在本地训练模型并将上传模型更新至中央服务器进行聚合，联邦学习有效保护了数据的隐私性[3]。然而，面对复杂的工业大数据环境，传统的联邦学习算法在应对数据分布不均和动态计算资源时，难以达到最佳性能。为此，Wang 等[4]提出联邦自适应学习的概念。它不仅继承了联邦学习保护数据隐私的优点，还引入了自适应机制，以动态调整学习过程中的各种参数和策略。这些参数可能包括学习率、批处理大小、模型架构等，旨在根据各参与方的数据分布、计算能力和网络条件等实际情况进行最优配置[5]，从而提升模型训练效率和最终性能。

然而，面对工业大数据环境的复杂性和动态性，联邦自适应学习算法往往难以达到最优性能。尤为突出的是，联邦自适应学习面临拜占庭式失败安全挑战。在联邦学习环境中，恶意参与方[6]（即拜占庭节点）可能通过发送伪造或精心设计的模型更新来破坏训练过程，导致模型性能下降甚至失效[7-8]。这些恶意更新可能包括与真实梯度相反方向的梯度参数、参数边界值或其他微小但足以影响模型性能的扰动[9]。

为了应对拜占庭攻击，学术界和工业界提出了多种防御算法。Yin 等[10]提出通过截断均值来抵御拜占庭节点的鲁棒聚合方法。该方法在面对多个恶意节点协同的情况下表现不佳，因为这些节点可能操纵更新值，使得恶意更新不被截断，从而影响聚合结果。Krum 等[11]提出贝叶斯推断方法识别并隔离恶意客户端，但其计算复杂度较高，且参与客户端数量较大时易导致计算开销显著增大和模型训练时间延长。Zhang 等[12]利用同态加密和安全多方计算保护联邦学习中的隐私和安全，尽管这些技术在隐私保护上具有优势，但它们的通信和计算成本较高，容易影响联邦学习的效率，尤其是在大规模数据和复杂模型下。Li 等[13]提出一种新的聚合规则，通过对每个客户端的贡献进行重新加权，减少恶意更新的影响。然而，该方法对恶意参与方的识别和重新加权依赖于特定的假设，如果恶意客户端能够规避这些假设，方法的有效性可能受到限制。此外，过于频繁的重新加权过程可能影响整体模型收敛速度。

综上所述，本文提出一种抗拜占庭攻击的梯度净化联邦自适应学习（MGF-WAAFL）算法，以实现拜占庭鲁棒联邦学习。MGF-WAAFL 利用一种新的信号梯度过滤技术来识别恶意梯度，并与基于权重和动量的聚合规则集成，以期算法能够抵御多种模型投毒攻击，提升模型鲁棒性。通过对恶意节点的精确识别和控制，MGF-WAAFL 有望显著提升联邦学习的安全性和模型性能。

1 基本知识

1.1 联邦学习

联邦学习是一种兼顾高效和隐私保护的机器学习方法。联邦学习框架如图 1 所示，由客户端局部模型更新和中心服务器聚合 2 个重要的部分组成。在整个学习过程中，只涉及局部模型的上传和全局模型的分发，而参与训练的各客户端的原始数据始终保留在本地数据库，因此，联邦学习为私密数据的安全提供了保护。

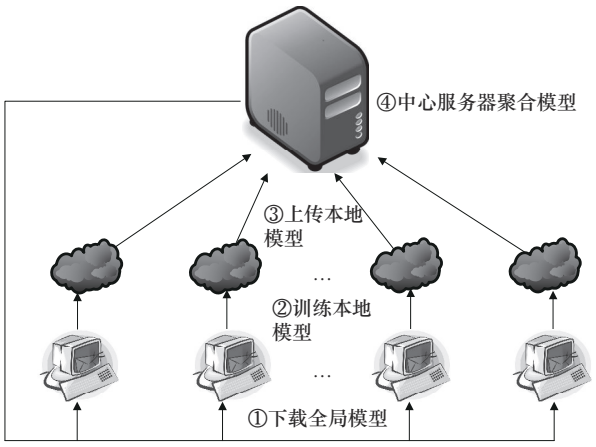


图 1 联邦学习框架

典型联邦学习需要多轮训练，每一轮不同客户端用本地数据进行局部模型训练，然后客户端和中心服务器进行双向通信，一轮迭代过程包括以下步骤（以第 t 轮模型训练和客户端 k 为例）。

1) 第 t 轮训练时，不同客户端从服务器下载全局模型 ω_t。

2) 第 k 个客户端第 t 轮通信训练本地数据得到的局部模型为 ω_{kt+1}。

3) 各方客户端将更新后的局部模型上传给中

心服务器。

4) 中心服务器接收数据后通过加权聚合操作，更新全局模型ω_{t+1}，然后将其分发给第$t+1$轮参加训练的客户端。

综上，联邦学习在工业[14]、生物识别[15]、医疗保健[16]、金融分析[17]、区块链[18-21]等领域具有重要的研究价值和广泛的应用前景。

1.2　拜占庭攻击

在联邦学习中，拜占庭敌手被定义为可完全控制多个用户的设备、数据，并可上传任意数据的攻击者[18]。拜占庭攻击是联邦学习中的一种典型攻击，旨在篡改参与者提交的模型更新参数，使模型参数的实际收敛过程偏离预期路径，从而对全局模型的精度和收敛性产生负面影响。拜占庭攻击的实现策略多种多样，通常可以分为两大类：梯度操纵攻击和数据中毒攻击。梯度操纵攻击是通过直接操纵拜占庭客户端发送的梯度更新，来影响全局模型的训练过程。以下介绍几种常见的梯度操纵攻击方法。

1) 随机攻击。在随机攻击中，拜占庭客户端生成并发送完全随机的梯度。这些梯度通常来自一个多维高斯分布，例如$g_m \leftarrow N(\mu,\sigma^2)$，其中在实验中常设定$\mu=(0,\cdots,0)$和$\sigma=0.5$。这种策略通过引入大量随机噪声，破坏了梯度的平均值，使得全局模型无法有效地收敛。

2) 噪声攻击。噪声攻击是在真实的梯度g_b上添加随机噪声$N(\mu,\sigma^2)$，生成的攻击梯度为$g_m=g_b+N(\mu,\sigma^2)$。这种方法保持了真实梯度的基本结构，但通过细微的噪声干扰，逐渐降低模型的性能。

3) 符号翻转攻击。在符号翻转攻击中，拜占庭客户端发送与真实梯度方向相反的梯度$g_m=-g_b$。这种攻击方法通过直接翻转梯度的方向，使得全局模型在优化过程中朝错误的方向更新，严重干扰模型的学习过程。

4) LIE攻击。LIE攻击[22]通过对恶意梯度的元素进行精细调整，使得它们在看似合理的同时，对模型造成显著的破坏。在这种攻击中，拜占庭客户端发送的恶意梯度看起来与正常梯度类似，但实际上已被巧妙地调整，以最大化对模型的负面影响。其基本原理是拜占庭客户端首先估算坐标均值（μ_j）和标准差（σ_j），然后通过以下方式构造恶意梯度向量的元素

$$(g_m)_j=\mu_j-z\sigma_j, j\in[1,d] \tag{1}$$

其中，d表示梯度向量的维度，且攻击因子z是一个正数，依赖于总客户端数目和拜占庭分数，可以通过累积标准正态函数$\phi(z)$来确定

$$z_{\max}=\max_z\left\{\phi(z)<\frac{n-\left[\frac{n}{2}+1\right]}{n-m}\right\} \tag{2}$$

5) ByzMean攻击。ByzMean攻击结合了拜占均值和LIE攻击的策略[23]。部分拜占庭客户端发送恶意的LIE梯度，而另一部分发送相对温和的梯度，使得整体上看起来更为合理，难以被检测，但依然有效地干扰模型的训练过程。其核心思想是使梯度的平均值成为任意恶意梯度。具体而言，恶意客户端被分为两组：一组有m_1个客户端选择任意的梯度向量g_{m1}，另一组有$m_2=m-m_1$个客户端选择梯度向量g_{m2}，使得所有梯度的平均值恰好等于g_{m1}。这可以表示为

$$g_{m2}=\frac{(n-m_1)g_{m1}-\sum_{i=m+1}^{n}g^{(i)}}{m_2} \tag{3}$$

所有现有的攻击方法都可以集成到这种ByzMean攻击中，使得这种混合攻击比任何单一攻击更加强大。例如，可以将g_{m1}设置为随机噪声或者由LIE攻击生成的梯度。

6) Min-Max/Min-Sum攻击。这些攻击方法通过优化恶意梯度的分布，使其尽量靠近良性梯度的集群[24]。Min-Max攻击确保恶意梯度在良性梯度集群的边界内，确保恶意梯度不会过于偏离正常梯度的范围，从而难以被检测和阻止，具体公式为

$$g_m=f_{\text{avg}}(g_i\in[n])+\gamma\boldsymbol{r}_p \tag{4}$$

其中，$\boldsymbol{r}_p$是扰动向量，γ是缩放系数。

Min-Sum攻击通过最小化恶意梯度与所有良性梯度的平方距离和来达到最大化破坏效果，通过微小的梯度调整，试图最大化对全局模型的破坏效果，具体公式为

$$g_m=f_{\text{avg}}(g\in[n])+\gamma\sum_{i=1}^{n}(\boldsymbol{r}_p)_i^2 \tag{5}$$

其中，$(\boldsymbol{r}_p)_i$是扰动向量$\boldsymbol{r}_p$的第i个元素。

数据中毒攻击通过在训练数据中引入错误或恶意样本，来间接干扰模型的训练过程[25]。这类攻击不直接修改梯度，而是通过操纵输入数据来影响梯度的生成，主要有标签翻转攻击。在标签翻转攻击中，拜占庭客户端将本地训练数据的标签翻转。

例如，对于一个分类任务，通常标签的取值范围是$\{0,1,\cdots,C-1\}$，其中C是类别的总数。在标签翻转攻击中，如果一个客户端的本地训练数据中的标签为l，那么该标签会被修改为$C-1-l$。

2　算法设计

基于恶意梯度过滤的联邦自适应学习（MGF-WAAFL）是一种设计用于抵御拜占庭攻击的创新性框架。现有方法通常存在静态参数设置、信息损失、未充分考虑数据异构性和设备差异性等问题。例如，Krum等[11]出的贝叶斯推断方法在计算复杂性和对先验知识的依赖上存在局限性，难以适应动态变化的攻击模式和复杂的数据分布。Li等[13]提出的重新加权聚合规则在重权调整的复杂性和数据异质性挑战方面也面临困境，可能无法完全隔离协同攻击中的恶意更新。MGF-WAAFL通过引入动态自适应的多重过滤器设计、多维度的梯度分析，以及适应非独立同分布数据和设备异构性的策略，提供了更加全面和灵活的防御机制。该框架使用滑动窗口技术动态调整过滤阈值，结合基于范数的阈值过滤器和基于符号的聚类过滤器，有效识别和过滤恶意更新，同时保留对全局模型有利的信息。此外，MGF-WAAFL还采用自适应的权重调整和动量优化技术，以适应不同客户端的数据和计算能力差异。整体而言，MGF-WAAFL能够防御多种类型的攻击，包括协同攻击和模型投毒，显著增强了联邦学习系统的鲁棒性和安全性。其工作流程如图2所示。

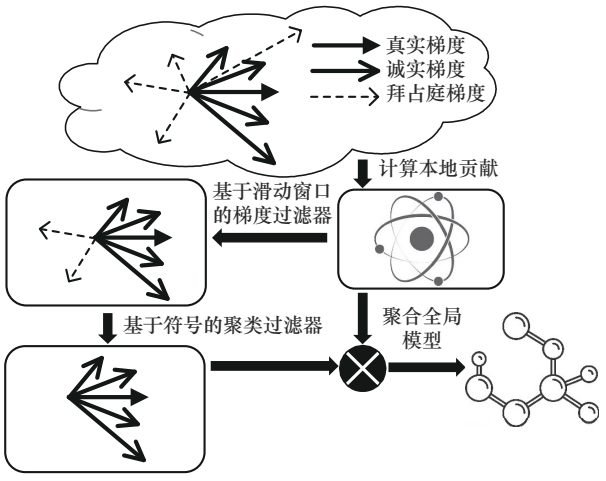


图2　MGF-WAAFL算法工作流程

2.1　基于滑动窗口的梯度过滤器

滑动窗口法是一种常用于时间序列分析的技术[26]，它在数据流动的过程中保留一段时间内的最新数据。在处理拜占庭攻击时，恶意客户端可能会发送极端的梯度更新，导致梯度范数的剧烈波动。通过滑动窗口动态调整阈值，可以有效地识别和过滤这些异常梯度，因为它们通常会导致梯度范数远离正常范围。滑动窗口的大小决定了当前计算所依赖的数据范围，这使得滑动窗口能够灵活地适应数据的变化。本文将滑动窗口应用于梯度范数的计算和过滤阈值的动态调整，以应对联邦学习中的梯度变化。该方法通过在固定大小的窗口中跟踪最近的梯度范数，并计算其中位数来确定动态的下限和上限阈值。通过设定较宽松的下限和严格的上限阈值，能够保留对模型训练影响较小的梯度，同时剔除那些异常或恶意的梯度[27]。这一机制可以有效地排除异常梯度更新，增强模型的鲁棒性。

每个客户端在训练过程中生成一个梯度向量，其范数$\|g_i\|$反映了梯度的整体大小。客户端i的梯度范数可以表示为

$$\|g_i\| = \sqrt{\sum_j g_{ij}^2} \tag{6}$$

其中，g_{ij}是客户端i在第j维度上的梯度。通过计算每个客户端的梯度范数，能够捕捉其梯度更新的整体变化情况。

窗口大小选择决定每次处理数据量。梯度过滤过程中，使用滑动窗口跟踪最近W次迭代中的梯度范数，并基于这些范数动态调整过滤阈值。每次迭代后，新的梯度范数$\|g_i\|$将被添加到滑动窗口中。如果滑动窗口已满，则移除最旧的梯度范数，使得窗口始终保持最新的W个梯度范数。滑动窗口S表示为一个长度为W的队列

$$S = \|g_t - W + 1\|,\|g_t - W + 2\|,\cdots,\|g_t\| \tag{7}$$

其中，t表示当前的迭代次数。

在每次迭代中，利用滑动窗口中的梯度范数计算中位数M。中位数M是数据集的中间值，可以有效地抵抗异常值的影响，因此适合作为动态调整过滤阈值的基准。其计算式为

$$M = \text{median}(S) \tag{8}$$

然而，当数据分布发生频繁变化时，计算出的中位数可能滞后于当前数据分布，从而影响过滤的有效性。为解决这一问题，本文引入了加权移动平均方法，通过为最近的数据点赋予更高权重，能够更灵敏地反应数据分布的变化。

具体地，首先在滑动窗口中计算加权移动平均

A_t，其根据最新的梯度范数动态调整，并用于调整中位数M的权重。加权移动平均公式为

$$A_t = \frac{\sum_{k=1}^{W} \omega_k \|g_{t-k}\|}{\sum_{k=1}^{W} \omega_k} \tag{9}$$

其中，$\|g_{t-k}\|$表示$t-k$次迭代的梯度范数，ω_k是与滑动窗口中每个梯度范数$\|g_{t-k}\|$相关的权重。通过引入加权移动平均方法，本文能够使得最近的梯度数据对阈值的调整产生更大的影响。

基于中位数和加权移动平均的动态调整，设定一个松散的下限阈值 L 和一个严格的上限阈值 R。具体公式为

$$\begin{aligned} L &= M - \alpha A_t \\ R &= M - \beta A_t \end{aligned} \tag{10}$$

其中，α是一个小于 1 的系数，用于确定下限；β是一个大于 1 的系数，用于确定上限；A_t是基于最近梯度值计算出的加权平均。通过动态调整 L 和 R，可以灵活地适应梯度变化。在每次迭代中，基于计算出的阈值 L 和 R，对每个客户端的梯度范数 $\|g_i\|$进行过滤。

对于低于下限 L 的梯度，选择保留。因为这些梯度的幅度较小，通常不会对模型训练产生显著的负面影响。对于高于上限 R 的梯度，选择剔除。因为这些梯度可能是由异常或恶意行为引起的，对模型训练的影响较大。

这一机制可以有效地排除异常梯度更新，增强模型的鲁棒性。

2.2　基于符号的聚类过滤器

在拜占庭攻击中，恶意客户端的梯度更新通常会形成与良性客户端不同的符号特征模式。为了更进一步区分良性和恶意梯度，使用了基于符号统计的特征进行聚类分析。提取一些梯度的统计量作为特征，并使用加权 MeanShift 聚类[28]算法作为无监督聚类模型，能够自适应地确定聚类的数量，同时选择规模最大的聚类作为可信集。通过分析梯度向量的符号统计特征（正符号比例、负符号比例和零符号比例），可以捕捉梯度向量的整体符号分布。基于这些特征，加权 MeanShift 聚类算法对每个客户端的符号特征进行聚类。该算法不仅能够自适应地确定聚类的数量，还通过引入权重来调整不同客户端的影响力，进而提高聚类精度。加权 MeanShift 聚类的引入，尤其在处理具有较大差异的良性客户端梯度时，能够减少误判的影响。由于权重机制能够合理调整不同客户端在聚类中的贡献，聚类过程对恶意更新的识别更加精确，进一步提高了对正常梯度的保留，从而有效降低了误判的风险。这使得该方法在动态和多变的攻击环境中能更好地维持模型的安全性和可靠性。

每个客户端在训练过程中产生的梯度向量g_i包含若干个元素，可以通过梯度向量的元素符号来提取统计特征。具体而言，计算梯度向量中正、负和零符号的比例，定义如下。正符号比例P_{pos}：梯度向量中正值元素的比例；负符号比例P_{neq}：梯度向量中负值元素的比例；零符号比例P_{zero}：梯度向量中零值元素的比例。这些符号统计特征能够有效捕捉梯度向量的整体符号分布，为进一步的聚类分析提供了基础。

在聚类前，需要首先估计合适的带宽参数。带宽决定了核密度估计的窗口大小，进而影响了聚类的密度估计和数据点的分布。带宽h估计公式为

$$h = \sigma \left(\frac{N}{d} \right)^{\frac{1}{d+4}} \tag{11}$$

其中，d是数据的维度，N是样本数量，σ是数据的标准差。

其次，将每个梯度的符号特征向量$f_i = (P_{\text{pos}}, P_{\text{neq}}, P_{\text{zero}})$作为加权 MeanShift 聚类输入，算法会自动寻找数据的密集区域并确定聚类的数量和中心。加权 MeanShift 聚类的步骤如下。

1) 选择一组初始的点作为聚类中心。通常，这些点可以是从数据集中随机选择的点或使用密度估计方法选择的点。计算密度梯度：对于每个点，计算其在特征空间中的密度梯度，并向密度最大的位置移动。设初始的聚类中心为$\{C_i\}_{i=1}^{n}$，其中C_i表示第i个聚类中心。

2) 对于每个点，计算其在特征空间中的密度梯度。计算其在特征空间中的密度梯度。不同于传统的聚类方法，考虑到每个客户端上传的梯度具有不同的权重，引入了加权核密度估计来计算密度，其中带宽 h 是关键参数。核密度估计的公式为

$$\dot{p}(x) = \frac{1}{Nh^d} \sum_{i=1}^{N} \omega_i K\left(\frac{x - x_i}{h} \right) \tag{12}$$

其中，ω_i是第i个梯度点的权重；K是核函数，通常选用高斯核函数$K(\|x\|^2) = \exp\left(-\frac{\|x\|^2}{2} \right)$；$h$是

带宽参数，决定了核的宽度。

3) 更新聚类中心：在每一次迭代中，聚类中心根据带权密度梯度进行更新。对于第t次迭代，聚类中心的更新公式为

$$c_i^{t+1}=\frac{\sum_{j=1}^{N}\omega_j x_j K\left(\frac{c_i^t-x_j}{h}\right)}{\sum_{j=1}^{N}\omega_j K\left(\frac{c_i^t-x_j}{h}\right)} \tag{13}$$

其中，c_i^t是第t次迭次时的第i个聚类中心，x_j是数据点j的位置，ω_j数据点j的权重。

4) 重复步骤 2)和步骤 3)，直到$\|c_i^{t+1}-c_i^t\|$小于给定的阈值。

5) 若当前的聚类中心与其他已经存在的聚类中心的距离小于阈值，则将两类合并；否则，将当前聚类作为新的类。

6) 重复步骤 1)~步骤 5)，直到所有样本点均被访问标记。

7) 在加权MeanShift聚类完成后，会得到若干个聚类，每个聚类包含具有相似符号特征的梯度，选择规模最大且权重较大的聚类作为可信的梯度集合。为建立选择可信的梯度集合，遵循以下原则。①规模最大原则：选择规模最大的聚类作为可信的梯度集合。假设大多数客户端是良性的，因此规模最大的聚类很可能代表了良性客户端的梯度分布[29]。②加权选择原则：根据每个聚类的总权重来评估其可信度，权重较大的聚类往往包含更可靠的客户端梯度。

这一过程有效地剔除了异常的梯度更新，并保留了可信的良性梯度。

2.3 基于权重的自适应聚合

在拜占庭攻击中，恶意客户端可能会发送不符合正常数据分布的梯度，服务器需要从众多客户端收集并聚合本地计算的梯度以更新全局模型。为了应对客户端之间可能存在的梯度差异和潜在的恶意行为，引入了一种基于权重的自适应鲁棒聚合方法，结合动量优化技术来增强模型更新的稳定性和鲁棒性[30]。在完成过滤过程后，服务器最终选择多个过滤器输出的交集作为可信集合，并通过基于权重的自适应鲁棒聚合获得全局梯度。这种方法根据每个客户端的贡献计算加权，以适应不同客户端的数据差异和计算能力，减少恶意更新的影响，提升全局模型的准确性和稳定性，从而有效防御拜占庭攻击。

基于权重的自适应聚合如算法1所示，所有与一阶矩、二阶矩、局部贡献和最终聚合比例相关的计算都是基于元素的，这意味着每个局部模型中的每个参数在每一轮中都会得到一个特定的聚合比例，而不是局部模型中所有参数的一个比例。

算法1 基于权重的自适应聚合

输入 初始本地学习率α，指数衰减率$\beta_1,\beta_2\in[0,1]$，全局训练轮次R

初始化 全局模型的初始权重ω_0、一阶矩向量m_0，并将二阶矩向量v_0初始化为零向量

1) 循环

2) for每一轮$r=1,2,\cdots,R$ do

3) for每个客户端$c\in C$ do

4) 计算梯度$g_r^{(c)}\leftarrow\nabla_\omega L(F(\omega_{r-1}))$

5) 更新一阶矩$m_r^{(c)}\leftarrow\beta_1 m_{r-1}+(1-\beta_1)g_r^{(c)}$

6) 更新一阶矩$m_r^{(c)}\leftarrow\beta_1 m_{r-1}+(1-\beta_1)g_r^{(c)}$

7) 更新二阶矩$v_r^{(c)}\leftarrow\beta_2 m_{r-1}+(1-\beta_2)\cdot(g_r^{(c)}\odot g_r^{(c)})$

8) 计算无偏估计$\hat{m}_r^{(c)}\leftarrow\frac{m_r^{(c)}}{1-\beta_1^r}$

9) 计算无偏估计$\hat{v}_r^{(c)}\leftarrow\frac{v_r^{(c)}}{1-\beta_2^r}$

10) 计算贡献$b_r^{(c)}\leftarrow\frac{\alpha}{\sqrt{\hat{v}_r^{(c)}+\varepsilon}}\hat{m}_r^{(c)}$

11) end for

12) 基于滑动窗口的梯度过滤器：使用滑动窗口方法动态调整过滤阈值，去除异常或恶意的梯度更新。

13) 基于符号的聚类过滤器：通过符号统计特征进行聚类，并筛选出可信梯度。

14) 对于过滤后的梯度以及其对应权重，计算Softmax权重$p_r^{(c)}\leftarrow\frac{\exp(\hat{b}_r^{(c)})}{\sum_{i=c}^{C}\exp(\hat{b}_r^{(i)})};\ \forall c\in C$

15) 计算全局梯度$G_r\leftarrow\sum_{i=c}^{C}p_r^{(i)}\hat{g}_r^{(i)}$

16) end for

3 实验与分析

3.1 数据集描述

以江西省某20家稀土公司的20条LaCe/PrNd分离生产线作为研究对象，其中镨（Pr）、钕

（Nd）元素合称为易萃组分（有机相），镧（La）、铈（Ce）元素称为难萃组分（水相）。经过前期数据采集以及预处理，共获取到2 000例样本作为原始数据集。为兼顾深层网络的训练效果，按8:1:1的比例将其划分为训练集、验证集和测试集。数据集中包含原料液属性、进料方式、药剂量等14个工艺输入变量以及相应萃取槽的组分含量。

3.2　实验设置

实验在配置为8 GB内存、i5处理器的Windows 10操作系统的计算机上进行，使用Pycharm 2023作为编译平台，编程语言为Python 3。所有实验基于torch框架，并通过torchvision导入数据集。在联邦学习的实验设置中，训练模型可以选择简单的多分支神经网络，主网络结构为[14-64]（即输入层有14个神经元，隐藏层有64个神经元），分支网络结构为[32-2]（即每个分支网络从主网络的隐藏层接收32个神经元的输出，并输出2个神经元）。假设有20个客户端参与训练，其中20%是拜占庭节点，这些节点可能发送恶意的梯度来干扰训练过程。为了评估不同防御算法的有效性，在训练中使用了不同的恶意客户端比例，并且在随着时间变化的攻击策略下也进行了测试。

此外，将梯度范数的下限和上限分别设置为L=0.1和R=3.0，滑动窗口W=9，并随机选择10%的坐标来计算MGF-WAAFL算法的符号统计。在数据集上的每个训练算法运行60个周期。每次局部迭代的次数设置为1，并采用动量为0.9的动量梯度下降，权重衰减设置为0.000 5。

3.3　结果分析

由于梯度的多样性，拜占庭攻击的缓解已成为一个众所周知的挑战。本节进行了广泛的实验，测试了各种攻击和防御算法的组合在数据设置下的效果。将本文提出的MGF-WAAFL算法与几种现有的防御算法（包括TrMean、Foolsgold、RSA、Multi-Krum、Bulyan、DnC和FedAvg）进行性能、开销、不同拜占庭客户端比例以及在时间变化下的比较，如表1所示。表1结果表明，本文提出的MGF-WAAFL算法在应对各种攻击时具有较高的效率和优越性。

3.3.1　性能比较

如表1所示，在不同中毒攻击下，MGF-WAAFL算法可以利用符号统计和相似性特征来过滤掉大多数恶意梯度，在无攻击情况下能够达到相当高的测试精度；在新的攻击方法中，如LIE和Min-Max/Min-Sum，能够规避基于中位数和距离的防御，从而阻碍成功的模型训练。以Multi-Krum算法的结果为例，当没有攻击时，Multi-Krum的精度下降几乎可以忽略不计（小于0.1%）。然而，在LIE攻击下，其最佳测试精度下降到56.82%，在Min-Max/Min-Sum攻击下甚至降至40%以下。在TrMean、Foolsgold、RSA和FedAvg算法下进行模型训练时，也能观察到类似的现象。

虽然Multi-Krum、DnC和Bulyan在应对简单攻击时表现良好，但这些方法在面对精心设计的攻击时仍然表现不佳。相比之下，MGF-WAAFL算法能够区分大多数这些精心设计的恶意梯度，并在各种攻击下达到令人满意的模型精度。值得注意的是，MGF-WAAFL算法具有较高的鲁棒性和保真度。此外，考虑到无攻击时拜占庭客户端的本地数据也对全局模型有贡献，因此即使是最好的拜占庭

表1　不同防御算法在不同攻击下的比较

防御算法	无攻击	Random Noise	Label-flip	ByzMean	Sign-flip	LIE	Min-Max	Min-Sum
TrMean	98.23%	98.63%	98.53%	93.31%	58.87%	48.44%	34.50%	33.89%
Foolsgold	98.00%	97.89%	98.10%	92.00%	60.00%	55.02%	50.08%	45.20%
RSA	97.50%	97.20%	97.30%	90.00%	65.00%	60.23%	55.03%	50.10%
Multi-Krum	99.20%	98.98%	99.11%	99.06%	83.26%	56.82%	39.04%	27.27%
Bulyan	99.10%	99.17%	99.12%	99.15%	98.58%	78.81%	78.86%	71.95%
DnC	99.09%	99.07%	99.08%	99.17%	82.25%	75.73%	70.12%	65.04%
FedAvg	98.50%	98.40%	98.45%	96.02%	72.12%	60.24%	52.20%	48.03%
MGF-WAAFL	99.21%	99.23%	99.34%	99.18%	99.02%	98.13%	98.15%	98.15%

攻击防御算法也会导致与基准结果之间存在小的差距。

3.3.2 不同拜占庭客户端比例下的比较

本节还评估了MGF-WAAFL算法在不同拜占庭客户端比例下的性能，对在数据集上训练的模型进行了实验。这里将通过在没有任何攻击或防御的情况下与基线相比模型预测精度的下降来衡量攻击的影响。保持总客户端数量为20，并将拜占庭客户端的比例从10%变化到40%，以研究不同防御算法在不同拜占庭客户端比例下的表现。使用默认的训练设置，并在各种最新的攻击下进行实验。特别是，将MGF-WAAFL算法的结果与TrMean、Foolsgold、RSA、Multi-Krum、Bulyan、DnC和FedAvg算法进行比较，如图3所示。从图3可以看到，本文算法能够有效地过滤掉恶意梯度，并在高比例的拜占庭客户端情况下仅导致轻微的精度下降，而其他防御算法在拜占庭客户端比例增加时受到了更大的攻击影响。特别地，在没有攻击时，FedAvg算法的性能与其他防御算法相当，但在遭受恶意攻击时，其鲁棒性明显不如MGF-WAAFL。在拜占庭客户端比例较高的情况下，FedAvg算法的精度下降更为明显，这表明其对精心设计的攻击缺乏足够的防御能力。相比之下，MGF-WAAFL算法能够在各种攻击下保持较高的模型精度，展示了其强大的鲁棒性和有效性。

3.3.3 在时间变化下的比较

在随着时间变化的拜占庭攻击策略下测试了不同的防御算法。这里仍然使用默认的系统设置，并在每个周期随机更改攻击方法（包括没有攻击的场景）。在数据集上的模型测试精度曲线如图4所示，其中基线是在没有攻击和防御的情况下进行训练的结果，只测试了最新的防御算法。另外，增加了在没有攻击和防御情况下使用FedAvg算法训练的结果作为参考。

实验结果表明，MGF-WAAFL算法能够确保成功地完成模型训练，其性能紧随基线，而其他防御算法出现了显著的精度波动和模型退化。这进一步证明了MGF-WAAFL算法的优越性。此外，在没有攻击的场景下，FedAvg算法的性能接近基线，进一步说明MGF-WAAFL在面对攻击时，尽管引入了防御机制，但对全局模型精度的影响较小，模型仍能保持接近基线的高精度表现。

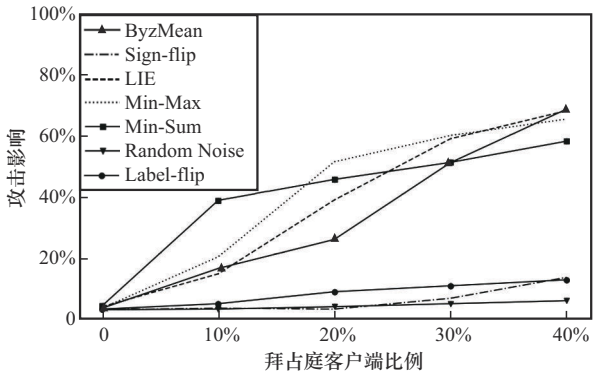
与其他防御算法相比，MGF-WAAFL算法在应对动态的攻击策略时表现出显著的稳定性，能够有效保持高精度，避免模型性能的显著下降。这说明，MGF-WAAFL在面对复杂多变的攻击场景时，依然能够提供可靠的防御效果，确保模型的鲁棒性和有效性。这些结果进一步强调了MGF-WAAFL在分布式学习中的潜在优势，尤其是在对抗复杂拜占庭攻击时的广泛应用前景。

3.3.4 通信开销比较

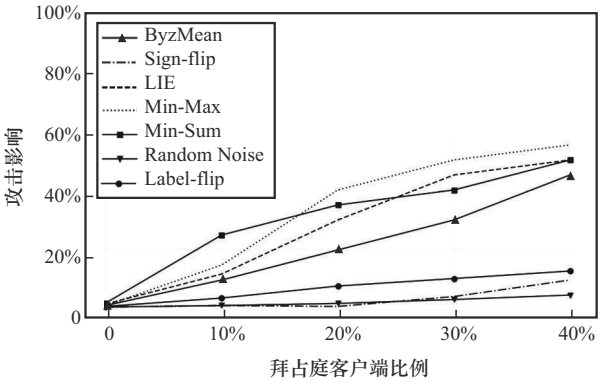
不同防御算法的计算和通信开销比较如表2所示。由表2可以看出，各种算法在计算和通信开销上存在显著差异。FedAvg和TrMean的计算和通信开销相对较低；MGF-WAAFL、Bulyan和RSA等算法由于包含复杂的恶意检测和自适应聚合机制，计算和通信开销较高。在应对拜占庭攻击时，具有防御机制的算法（如MGF-WAAFL、Bulyan、Multi-Krum、Foolsgold等）能够有效过滤恶意梯度，提高全局模型的鲁棒性。特别是MGF-

表2 不同防御算法的计算和通信开销比较

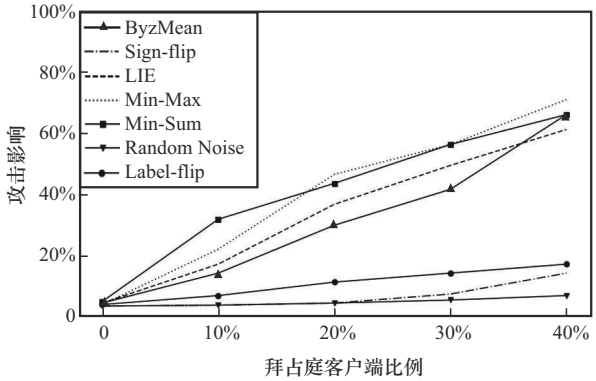
方法	恶意检测时间/s	权重计算时间/s	聚合时间/s	总运行时间/s	CPU/GPU使用率	内存消耗/MB	每轮通信数据量/MB	达到目标精度所需轮次	总通信数据量/MB
TrMean	1	0	3	4	30%	600	52	90	4 680
Foolsgold	2	1	4	7	35%	650	53	85	4 505
RSA	3	1	4	8	40%	700	55	80	4 400
Multi-Krum	3	0	5	8	40%	700	55	80	4 400
Bulyan	3	0	6	9	45%	720	56	75	4 200
DnC	2	1	5	8	38%	680	54	82	4 428
FedAvg	0	0	2	2	20%	500	50	100	5 000
MGF-WAAFL	4	2	6	12	50%	800	60	70	4 200



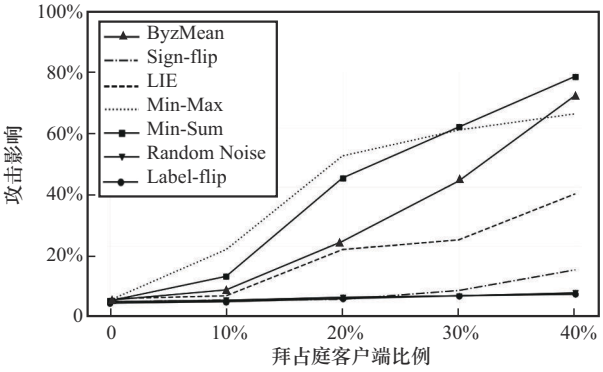
(a) TrMean



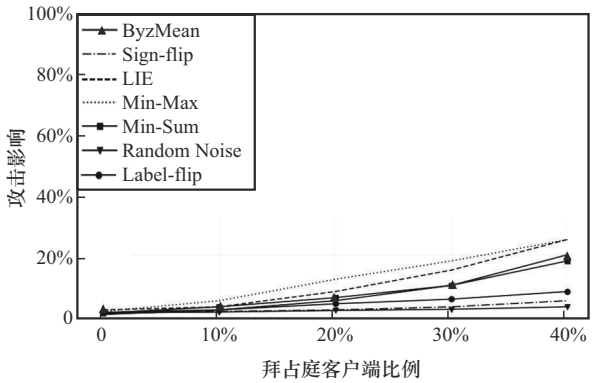
(b) Foolsgold



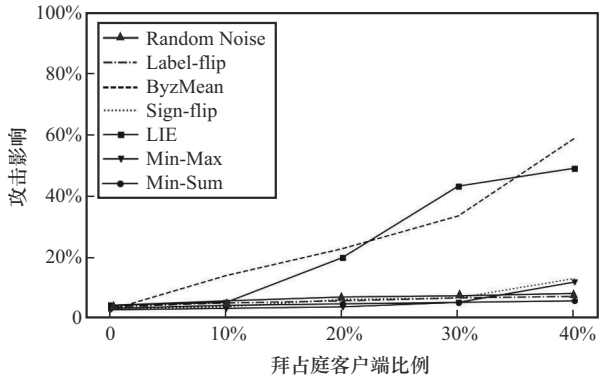
(c) RSA



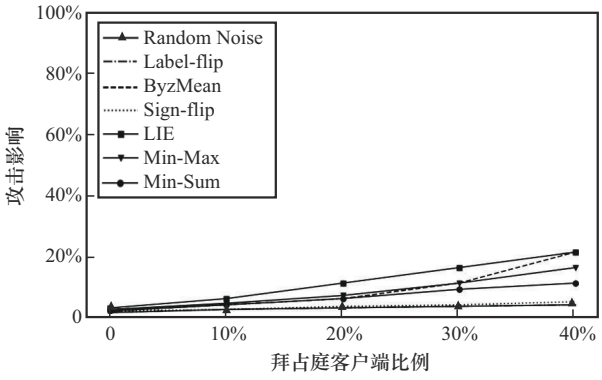
(d) Multi-Krum



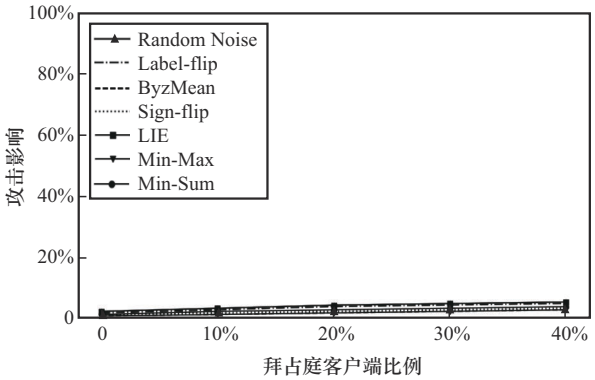
(e) Bulyan



(f) Dnc



(g) FedAvg



(h) MGF-WAAFL

图3　不同拜占庭客户端比例下的比较

WAAFL，通过基于滑动窗口的梯度过滤器和基于符号的聚类过滤器，成功筛选出可信梯度，并结合基于权重的自适应聚合，增强了模型的稳健性。尽管MGF-WAAFL和其他防御算法的开销较高，但它们提供了更强的攻击防御能力，使得全局模型在不信任环境中的表现更加稳健。

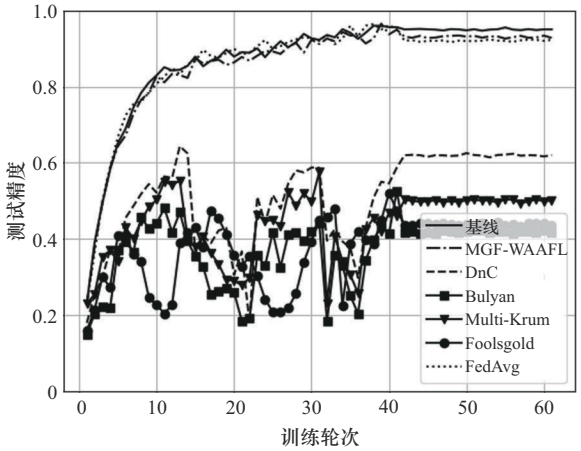


图4　在数据集上的模型测试精度曲线

4　结束语

工业大数据下的数据安全和隐私保护至关重要，尤其是在联邦学习环境下，面对复杂的拜占庭攻击和投毒攻击，传统防御算法难以应对。本文提出的MGF-WAAFL算法通过滑动窗口和符号聚类过滤器识别恶意梯度，并结合基于权重的自适应聚合，显著提升了模型的鲁棒性和安全性。尽管面临高计算与通信开销、参数敏感等问题，但是实验显示其在稀土行业有效防御攻击，保障数据合作。未来需优化计算、通信效率，自适应调参，并增强防御力，以全面提升算法效能与鲁棒性。

参考文献：

[1] 微众银行AI项目组. 联邦学习白皮书V1.0[R]. 2018.
WeBank AI　Project Team. Federated learning white paper V1.0[R]. 2018.

[2] 潘碧莹, 丘海华, 张家伦. 不同数据分布的联邦机器学习技术研究[C]//5G网络创新研讨会(2019)论文集. 广州: 移动通信, 2019: 271-276.
PAN B Y, QING H H, ZHANG J L. Research on federal machine learning technology with different data distribution[C]//5G Network Innovation Seminar (2019) Proceedings. Guangzhou: Mobile Communications, 2019: 271-276.

[3] YANG Q, LIU Y, CHEN T J, et al. Federated machine learning[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.

[4] WANG S Q, TUOR T, SALONIDIS T, et al. Adaptive federated learning in resource constrained edge computing systems[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(6): 1205-1221.

[5] WANG X, LIU X, ZHAO J. Federated adaptive learning: a new approach to optimizing federated learning systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(5): 1234-1247.

[6] YANG Z, LIU W, ZHANG W. Byzantine-robust federated learning: a review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(5): 1722-1735.

[7] XIE C, KOYEJO O, GUPTA I. Zeno: distributed stochastic gradient descent with suspicion-based fault-tolerance[J]. arXiv Preprint, arXiv: 1805.10032, 2018.

[8] WEN J, ZHANG Z X, LAN Y, et al. A survey on federated learning: challenges and applications[J]. International Journal of Machine Learning and Cybernetics, 2023, 14(2): 513-535.

[9] ENTHOVEN D, AL-ARS Z. An overview of federated deep learning privacy attacks and defensive strategies[C]//Studies in Computational Intelligence. Berlin: Springer, 2021: 173-196.

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

[11] KRUM F, ALISTARH D, ANGELOVA M. Bayesian inference for identifying and isolating malicious clients in federated learning[C]// Proceedings of the 2021 ACM Conference on Computer and Communications Security. New York: ACM Press, 2021: 1121-1135.

[12] ZHANG J, WANG T, WANG S. Privacy-preserving federated learning via homomorphic encryption and secure multi-party computation[C]// Proceedings of the 2022 Network and Distributed System Security Symposium (NDSS). Piscataway: IEEE Press, 2022: 95-110.

[13] LI X, HUANG K, LIU Y. A reweighting aggregation rule for mitigating malicious updates in federated learning[C]//Proceedings of the 29th USENIX Security Symposium. Berkeley: USENIX Association, 2020: 789-803.

[14] 刘晶, 张喆语, 董志红, 等. 基于工业物联网的区块链多目标优化[J]. 计算机集成制造系统, 2021, 27(8): 2382-2392.
LIU J, ZHANG Z Y, DONG Z H, et al. Multi-objective optimization of blockchain-based on industrial Internet of things[J]. Computer Integrated Manufacturing Systems, 2021, 27(8): 2382-2392.

[15] ZHUANG W M, WEN Y G, ZHANG X S, et al. Performance optimization of federated person re-identification via benchmark analysis[C]// Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 955-963.

[16] SHELLER M J, EDWARDS B, REINA G A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data[J]. Scientific Reports, 2020, 10(1): 12598.

[17] WeBank. Utilization of FATE in risk management of credit in small and micro enterprises[R]. 2019.

[18] PREUVENEERS D, RIMMER V, TSINGENOPOULOS I, et al. Chained anomaly detection models for federated learning: an intrusion detection case study[J]. Applied Sciences, 2018, 8(12): 2663.

[19] ZHU X D, LI H, YU Y. Blockchain-based privacy-preserving deep learning[C]//Proceedings of the International Conference on Information Security and Cryptology. Berlin: Springer, 2019: 370-383.

[20] KIM Y J, HONG C S. Blockchain-based node-aware dynamic weighting methods for improving federated learning performance[C]//Proceedings of the 2019 20th Asia-Pacific Network Operations and Man-

agement Symposium (APNOMS). Piscataway: IEEE Press, 2019: 1-4.
[21] 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.
[22] BARUCH M, BARUCH G, GOLDBERG Y. A little is enough: circumventing defenses for distributed learning[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. New York: ACM Press, 2019: 8635-8645.
[23] XIE C, KOYEJO S, GUPTA I. Zeno++: robust fully asynchronous SGD[J]. arXiv Preprint, arXiv: 1903.07020, 2019.
[24] FANG M H, CAO X Y, JIA J Y, et al. Local model poisoning attacks to Byzantine-robust federated learning[C]//29th USENIX Security Symposium (USENIX Security 20). Berkeley: USENIX Association, 2020: 1605-1622.
[25] BIGGIO B, NELSON B, LASKOV P. Poisoning attacks against support vector machines[C]//Proceedings of the 29th International Conference on Machine Learning. Piscataway: IEEE Press, 2012: 1467-1474.
[26] KEOGH E, LIN J. Clustering of time-series subsequences is meaningless: implications for previous and future research[J]. Knowledge and Information Systems, 2005, 8(2): 154-177.
[27] YIN D, CHEN Y D, RAMCHANDRAN K, et al. Byzantine-robust distributed learning: towards optimal statistical rates[C]//Proceedings of the 35th International Conference on Machine Learning. Piscataway: IEEE Press, 2018: 5650-5659.
[28] COMANICIU D, MEER P. Mean shift: a robust approach toward feature space analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(5): 603-619.
[29] BLANCHARD P, MHAMDI E M E, GUERRAOUI R, et al. Byzantine-tolerant machine learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 118-128.
[30] LI T, SAHU A K, ZAHEER M, et al. Federated learning: challenges, methods, and future directions[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(9): 3787-3807.

[作者简介]

杨辉（1965– ），男，江西高安人，华东交通大学教授、博士生导师，主要研究方向为复杂工业过程建模与优化控制、轨道交通自动化与运行优化。

邱子游（1998– ），男，广东佛山人，华东交通大学硕士生，主要研究方向为端边云协同控制平台、数字孪生。

李中美（1989– ），女，江苏苏州人，博士，华东理工大学硕士生导师，主要研究方向为人工智能（感知、认知、决策）、工业生产过程建模与优化控制。

朱建勇（1977– ），男，江西新干人，博士，华东交通大学教授，主要研究方向为复杂工业过程控制与优化、大数据分析。