计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：基于生成对抗网络的联邦学习中投毒攻击检测方案
作者：陈谦，柴政，王子龙，陈嘉伟
收稿日期：2022-12-13
网络首发日期：2023-05-23
引用格式：陈谦，柴政，王子龙，陈嘉伟．基于生成对抗网络的联邦学习中投毒攻击检测方案[J/OL]．计算机应用．
https://kns.cnki.net/kcms2/detail/51.1307.TP.20230522.1041.004.html

Journal of Computer Applications ISSN 1001-9081
计算机应用 CODEN JYIIDU http://www.joca.cn

doi:10.11772/j.issn.1001-9081.2022121831

基于生成对抗网络的联邦学习中投毒攻击检测方案

陈谦，柴政，王子龙*，陈嘉伟

（西安电子科技大学 网络与信息安全学院，西安 710071）

（*通信作者电子邮箱 zlwang@xidian.edu.cn）

摘　要： 联邦学习（FL）是一种新兴的隐私保护机器学习范式。然而，其分布式的训练结构更易受到投毒攻击的威胁：攻击者通过向中央服务器上传投毒模型以污染全局模型，减缓全局模型收敛速度并降低全局模型准确率。针对这一问题，提出了一种基于生成对抗网络（GAN）的投毒攻击抵御方案。首先，将良性本地模型输入生成对抗网络产生测试样本；然后，使用生成的测试样本检测客户端上传的本地模型；最后，根据检测指标剔除投毒模型。该方案定义了 F1 值损失和准确率损失两项检测指标用于检测投毒模型，将检测范围从单一类型的投毒攻击扩展至全部两种类型的投毒攻击；设计了阈值判定方法处理误判问题，确保了误判鲁棒性。实验结果表明，在 MNIST 和 Fashion-MNIST 数据集上，该方案能够生成高质量检测样本，并对投毒模型进行有效检测与剔除。与使用收集测试数据和使用生成测试数据但仅使用准确率作为检测指标的两种典型方案相比，全局模型准确率提升了 2.7 到 13.9 个百分点。

关键词： 联邦学习；投毒攻击；生成对抗网络；F1 值损失；准确率损失；阈值判定方法

中图分类号：TP309.2　　**文献标志码：** A

Poisoning attack detection scheme based on generative adversarial network for federated learning

CHEN Qian, CHAI Zheng, WANG Zilong*, CHEN Jiawei

(School of Cyber Engineering, Xidian University, Xi'an Shaanxi 710071, China)

Abstract: Federated Learning (FL) has emerged as a novel privacy-preserving Machine Learning (ML) paradigm. However, the distributed training structure of FL is more vulnerable to poisoning attack, where adversaries contaminate the global model through uploading poisoning models, resulting in the deceleration of convergence rate and the degradation of global model prediction accuracy. To solve this problem, a poisoning attack detection scheme based on Generative Adversarial Network (GAN) was proposed. Firstly, the benign local models were fed into the GAN to output a set of high-quality testing samples. Then, the testing samples were used to detect poisoning models. Finally, the poisoning models were eliminated according to the testing metrics. Particularly, two testing metrics named F1 score loss and accuracy loss were proposed to extend the detection scope from one single type of poisoning attack to all types of poisoning attacks. Besides, the proposed detection scheme was robust to misjudgment by introducing the threshold determination method. Extensive experimental results on MNIST and Fashion-MNIST datasets show that the proposed detection scheme can obtain a set of high-quality testing samples, and then detect and eliminate poisoning models. Compared with the global models trained with detection scheme based on directly gathering testing data from clients and the detection scheme based on testing accuracy, the global model accuracy has a significant improvement from 2.7 to 13.9 percentage points.

Keywords: Federated Learning(FL); poisoning attack; Generative Adversarial Network(GAN); F1 score loss; accuracy loss; threshold determination method

0　引言

机器学习（Machine Learning，ML）[1]被广泛应用于从数据中提取信息。随着大数据时代的来临，数据量呈现出飞速增长的趋势，收集海量的数据用于机器学习变得异常困难。随着人们隐私保护意识的增强，利用包含用户敏感信息的数据进行集中式机器学习的方式受到了极大的限制[2]，联邦学习（Federated Learning，FL）[3]应运而生。区别传统的数据集中式机器学习，联邦学习的分布式结构降低了收集训练数据带来的高昂的通信开销，节省了稀缺的网络带宽，提升了通信效率。此外，包含敏感信息的原始数据自始至终没有离开客户本地，用户隐私得以保障。因此，联邦学习以其高效

收稿日期: 2022-12-13;　**修回日期**: 2023-05-09;　**录用日期**: 2023-05-10。

基金项目：国家自然科学基金(No. 62172319，U19B200073)。

作者简介：陈谦(1993—)，男，陕西西安人，博士研究生，主要研究方向：隐私保护、机器学习、联邦学习；柴政(1999—)，男，黑龙江大庆人，硕士研究生，主要研究方向：联邦学习、投毒攻击；王子龙(1982—)，男，河南郑州人，教授，博士，主要研究方向：信息论、密码学；陈嘉伟(1998—)，男，陕西西安人，硕士研究生，主要研究方向：联邦学习、强化学习。

的通信效率和良好的隐私保护性能被广泛部署在电子医疗[4]和垃圾邮件检测[5]等场景中。

投毒攻击已经在传统的机器学习场景中得到了广泛的研究[6]，而联邦学习更易受到投毒攻击的威胁[7]。敌手挟持客户端，通过篡改训练数据间接篡改本地模型[8]，或者通过修改模型参数直接篡改上传的本地模型[9]。聚合篡改后的模型会污染全局模型，进而降低全局模型收敛速度以及全局模型准确率。依据攻击目的，投毒攻击一般可以被分为随机攻击（或无目标攻击）和有目标攻击[10]。在随机攻击中，敌手随机篡改本地训练数据或模型，使得全局模型预测准确率下降。在执行有目标攻击时，敌手通过修改训练数据的某一特定特征或特定模型参数，使模型对某一特定类型样本预测错误。

针对投毒攻击这一重大威胁，一系列防御措施相继被提出和研究，这些防御方案可以归纳为被动防御和主动防御两类。被动防御方法通常是指中央服务器依据统计方法获得本地模型分布特征，并设计相应的聚合方法，在聚合过程中剔除投毒模型，从而提升全局模型性能[11-14]。但是经过精心设计的投毒模型可以获得与正常模型相似的统计特征，进而规避聚合方法的检测以污染全局模型[8]。相较于被动防御，主动防御方法通过检测本地模型性能从而剔除投毒模型，可以有效地检测出投毒模型并完全消除投毒模型对全局模型的负面影响[15-17]。因此，主动防御方法已成为设计联邦学习中投毒攻击检测方案的新趋势。例如，Jagielski 等[15]提出一种由服务器收集部分本地训练样本并训练比对模型，通过迭代估计比对模型和本地模型残差值的检测方法，该方法可以有效抵御针对训练数据的投毒攻击。然而，当本地训练集中含有较多恶意样本时，该方法检测效果较差。同时，该方法在构建使用的训练集时需要用户上传隐私训练数据，违背了联邦学习中训练数据不出本地的初衷。Zhao 等[16]首先提出使用生成对抗网络（Generative Adversarial Network，GAN）[18]产生检测样本的方法。但由于仅采用准确率作为检测指标，该方案无法准确检测有目标攻击。此外，由于现实的本地模型预测准确率还受到用户数据非独立同分布、训练误差、传输误差等因素的影响，该方案仅仅根据单次聚合结果即判定恶意用户的操作存在误判的可能。

针对服务器无法获得优质检测数据这一问题，本文引入生成对抗网络的方法，设计并实现了基于生成对抗网络的联邦学习中投毒攻击检测方案。该方案将未受到攻击的本地模型作为生成对抗网络的输入，输出一组用于检测的测试样本，然后利用该测试样本检测本地模型的 F1 值损失和准确率损失，从而剔除全部两种类型的投毒模型。同时，针对文献[16]所提方案中的误判问题，本方案设计了阈值判定方法以提高误判鲁棒性。本文主要贡献归纳为 3 个方面：

（1）定义并应用了全面的检测指标。针对两种不同类型的投毒攻击模型，定义了 F1 值损失和准确率损失两种检测指标，并在所提方案中使用这两种指标对有目标投毒模型和无目标投毒模型进行检测和剔除，将投毒攻击的检测范围扩大至全部两种类型。

（2）提出了一种广泛适用的联邦学习中投毒攻击检测方案。针对服务器缺少检测样本的现实问题，所提方案使用生成对抗网络产生检测样本（称为生成测试集），并将其用于检测客户端上传的本地模型。该方法尤其适用于由于隐私限制、有损通信信道等因素导致的服务器无法收集高质量检测数据的现实场景。同时，引入多次检测的阈值判定方法处理误判问题，确保所提检测方案的误判鲁棒性。所提方案能够检测并剔除联邦学习系统中同时存在的两种投毒模型，提升全局模型的收敛速度和最终全局模型准确率，有效消除投毒攻击对联邦学习的威胁。

（3）设计并实现了仿真实验与有效性分析。为验证所提方案有效性，分别在 MNIST 和 Fashion-MNIST 两个图像数据集上进行仿真实验。实验结果表明，所提方案中基于生成对抗网络产生的生成测试集具有和理想测试集相近的投毒模型检测效果。同时，通过比较在生成测试集上计算得到的 F1 值损失和准确率损失，所提方案不仅获得更好的随机攻击抵御效果，而且可以更加有效地抵御有目标攻击。在两种攻击场景下，全局模型的准确率相较于有攻击场景大幅度提高到了与无攻击场景相近的水平。以 Fashion-MNIST 数据集为例，相较于收集测试数据的方法[15]与使用生成测试数据但只使用准确率作为检测指标的方法[16]，在随机攻击下，所提方案将全局模型的准确率分别提高了 13.9 个百分点和 6.7 个百分点；在有目标攻击下，将全局模型准确率分别提高了 4.5 个百分点和 2.7 个百分点。

1 联邦学习中的投毒攻击

一个典型的联邦学习系统由中央服务器和一组客户端组成，其目标是客户端在中央服务器的协调下协作训练一个机器学习模型。客户端将隐私的训练数据保存在本地，并在本地训练本地模型。随后中央服务器聚合本地模型以更新全局模型。然而这样的设置容易受到投毒攻击的威胁。攻击者可以通过上传恶意篡改的本地模型污染全局模型。

1.1 联邦学习系统

如图 1 所示，典型的联邦学习系统可以[3-6]形式化描述为，一组 n 个客户端 $C=\{c_i,i=1,2,\ldots,n\}$ 分别持有样本数量为 $|D_i|$ 的隐私数据集 D_i，其中 $i=1,2,\ldots,n$。所有客户端持有的隐私数据集的样本数量之和记为 $|D_{total}|$。客户端的目的是在中央服务器的协调下协作训练一个机器学习模型：

$$W^*=\arg\min_W F(W) \tag{1}$$

其中 $F(W)$ 表示损失函数，W 表示全局模型。在本文所考虑的联邦学习系统中，中央服务器和客户端都是诚实的，它们严格依据流程执行联邦学习步骤[3-6]。首先，每个客户端 c_i 利用自己的隐私数据集在本地计算各自的梯度：

$$g_i^t = \nabla F\left(w_i^{t-1}; D_i\right) \tag{2}$$

其中 w_i^{t-1} 表示客户端本地模型，t 表示联邦学习聚合轮次。

随后，各客户端 c_i 使用学习率 η_i 对本地模型执行梯度下降算法，其过程表示为：

$$w_i^t = w_i^{t-1} - \eta_i \bullet g_i^t \tag{3}$$

然后，各客户端 c_i 将更新后的本地模型上传到中央服务器，由中央服务器聚合本地模型并更新全局模型：

$$W^t = \sum_{i=1}^{n} q_i \bullet w_i^t \tag{4}$$

其中 q_i 为本地模型聚合权重。整个过程循环进行直到全局模型收敛到 W^*。

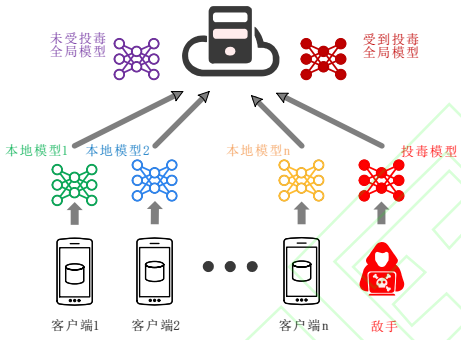


图 1 联邦学习与投毒攻击

Fig. 1 Federated learning with poisoning attacks

1.2 投毒攻击

1.2.1 威胁模型

本文考虑联邦学习中广泛应用的威胁模型[19]，其中敌手通过控制参与联邦学习的客户端完成投毒攻击。具体威胁模型定义如下：

敌手目标。敌手通过制造投毒模型并向中央服务器上传投毒模型，降低全局模型的整体预测准确率或者在特定类别样本上的预测准确率，从而达到污染全局模型目的。

敌手能力。敌手控制了 n 个客户端中的 m 个，这些被控制的客户端称为恶意客户端，其中 n 和 m 满足 $(m/n)<0.5$；同时，敌手可以访问每个聚合轮次中的全局模型参数，并可以直接操作所掌控的恶意客户端上的梯度。

敌手知识。敌手已知全局模型以及所控制客户端的本地模型和数据，因此可以使用其控制的恶意客户端训练投毒模型。但敌手无法通过任何方法获知中央服务器抵御投毒攻击的检测方法以及对应的全局模型聚合方法。

1.2.2 投毒攻击的种类

根据敌手对全局模型性能的影响程度，投毒攻击分为随机攻击和有目标攻击两种。本文设想的联邦学习系统中同时存在随机攻击和有目标攻击。

随机攻击[20-21]。根据对敌手能力与敌手知识的假设，敌手可以通过向模型的梯度添加扰动执行随机攻击[20]。敌手首先在恶意客户端上使用本地训练集计算梯度 g_j：

$$g_j = \nabla F\left(w_A; D_j\right) \tag{5}$$

其中，D_j 表示敌手控制的第 j 个客户端的数据集，$j=1,\cdots,m$；w_A 表示敌手已知的全局模型。

随后，敌手在计算得到的梯度上添加扰动向量，获得投毒梯度 $\bar{g}_j$：

$$\bar{g}_j = g_j + \nabla^p \tag{6}$$

其中 ∇^p 为一个与梯度尺寸相匹配的扰动向量。

然后，敌手使用投毒梯度，通过梯度下降算法获得随机攻击投毒模型 $\bar{w}_j$：

$$\bar{w}_j = w_A - \eta_A \bullet \bar{g}_j \tag{7}$$

其中 w_A 表示敌手已知的全局模型，η_A 表示学习率。

对应地，中央服务器聚合的全局模型如下所示：

$$W_A = \sum_{j=1}^{m} q_j \bullet \bar{w}_j + \sum_{i=m+1}^{n} q_i \bullet w_i \tag{8}$$

而受到随机攻击的全局模型 W_A 将无法对全部测试样本正确预测。例如，在医学图像分析场景中，敌手向本地模型添加符合特定分布的扰动，使本地模型与诚实客户端的模型出现较大差异，最终导致全局模型对医学图像数据集出现整体分类错误[21]。

有目标攻击[19,22-23]。同样地，敌手可以通过上传经过精心构造的模型执行有目标攻击[19]。敌手首先在其控制的每个恶意客户端本地将特定类别样本的标签为篡改为其他类别，从而构造恶性数据集[22]。然后，敌手在该数据集训练表示如下的有目标攻击投毒模型：

$$\bar{w}_j' = w_B - \eta_B \bullet g_j \tag{9}$$

其中 w_B 表示敌手已知的全局模型，η_B 表示学习率，g_j 表示使用恶性数据集计算得到的梯度。

此时中央服务器聚合得到的全局模型为：

$$W_B = \sum_{j=1}^{m} q_j \bullet \bar{w}_j' + \sum_{i=m+1}^{n} q_i \bullet w_i \tag{10}$$

与随机攻击不同，受到有目标攻击的全局模型 W_B 对大部分测试样本预测结果无影响，只在对被篡改的特定类别样本进行预测时产生偏差。例如，在入侵检测场景中，敌手将来自某一类设备的恶意流量数据标注为良性流量数据，并将篡改后的数据与其他良性数据混合以训练本地模型，最终造成全局模型虽然整体分类效果较好，但对于该特定种类恶意流量的分析准确率受到较大影响[23]。

2 投毒攻击防御方法研究现状

2.1 被动防御

被动防御中，服务器通过分析上传的本地模型的统计特征，设计抗投毒模型的聚合方法，以期在聚合本地模型过程中剔除投毒模型。Yin 等[11]提出了 Trimmed-Mean 方案和 Median 方案。这两个方案首先在模型的每个维度上对所有本地模型参数进行排序。随后，Trimmed-Mean 方案去掉规定个数的模型参数的最大值与最小值，并计算剩余模型参数的平均值；而 Median 方案则将每个维度的梯度中位数组合后作为全局模型。

投毒模型通常与正常模型有较大的差异。因此通过计算几何意义上的相似度从而检测投毒模型是一种有效的方法。Blanchard 等[12]提出的 Krum 方案通过度量欧氏距离，选择全部本地梯度中与其它本地梯度最相似的值作为全局梯度。而 Multi-Krum 方案作为 Krum 方案的变体，通过选择多个最相似的梯度并计算它们的均值，获得了更快的模型收敛速度和更高的全局模型准确率。Guerraoui 等[24]提出的 Bulyan 方案首先迭代使用 Krum 方案以选择多个梯度，之后计算选中梯度的 Trimmed-mean 值，其本质是 Krum 方案和 Trimmed-Mean 方案的结合。Muñoz-González 等[13]提出的 AFA 方案在每个联邦学习聚合轮次中首先计算收集到的梯度的加权平均值，然后计算该加权平均值与每个梯度之间的余弦相似度。最终，AFA 方案综合考虑余弦相似度的均值、中位数和标准差，检测并剔除离群梯度。陈宛桢等[14]提出了一种基于区块链的隐私保护联邦学习算法。客户端在本地训练模型参数，并以秘密共享的方式上传模型更新至附近的边缘节点；随后，边缘节点计算所有更新间的欧氏距离并将计算结果与模型更新上传至区块链；最后，区块链对模型参数之间的欧氏距离进行重构，进而去除投毒模型更新。

被动防御方法实现简单，并且在检测过程中不需要进行模型评估，因而检测效率较高。但是，本地模型特征的统计偏差将造成误判，进而影响投毒模型的检测准确率。因此，被动防御方法在实际中无法得到有效应用。

2.2 主动防御

主动防御方法是近年来联邦学习中投毒攻击检测方法的新方向。Steinhardt 等[25]提出使用数据清洗的方法，即通过对训练集进行清洗，筛选并剔除投毒数据，从而完成对投毒攻击的抵御。但是这样的方法显然无法在敌手直接篡改模型参数的本地模型投毒攻击形式中发挥作用[19]。Feng 等[17]提出了一种基于逻辑回归分类器的数据投毒防御策略，通过异常值检测的方式去除异常度超出阈值的样本。然而该方法假设服务器预先知道投毒样本在训练数据中的比例，这在联邦学习中是无法实现的。Jagielski 等[15]提出一种由服务器收集部分本地训练样本、训练比对模型，并迭代估计比对模型和本地模型残差值的检测方法。该方法可以有效抵御针对训练数据的投毒攻击。然而，当本地训练集中含有较多恶意样本时，该方法检测效果较差。且在构建该方法使用的训练集时需要用户上传隐私训练数据，违背了联邦学习中训练数据不出本地的初衷。Zhao 等[16]首先提出使用生成对抗网络产生检测样本的方法，在确保对用户隐私训练数据保护的同时，进一步缓解了投毒攻击的影响。但在联邦学习系统受到有目标攻击时，由于只采用准确率作为检测指标，该方案无法准确检测并剔除所有的投毒模型。此外，该方案仅仅根据单次聚合结果即判定恶意用户，存在误判的可能。由于现实的本地模型预测准确率还受到用户数据非独立同分布、训练误差、传输误差等因素的影响，使用该方案对此类场景下的本地模型进行检测时存在误判的可能。

在已有研究的基础上，本文定义了 F1 值损失和准确率损失两种检测指标，扩大了投毒攻击的检测范围，增强了检测方案的实用性。使用生成对抗网络产生的高质量检测样本检测客户端上传的本地模型，提高了检测准确率。引入多次检测的阈值判定方法，处理了现实场景中潜在的误判问题，提升了检测方案的误判鲁棒性。

3 生成对抗网络

GAN 自提出以来，已经在计算机视觉[26]、隐私保护[27]等领域获得了广泛的应用。GAN 可以通过机器学习模型学习原始训练数据的分布从而生成人工样本。通常来说，GAN 包括一个生成器 G 和一个判别器 H。在 GAN 的训练阶段，判别器和生成器进行对抗性博弈：生成器 G 试图用生成的人工样本欺骗判别器 H，而判别器 H 试图区分生成的人工样本和真实数据。在联邦学习中，中央服务器通常持有一个用于测试客户端上传模型准确率的测试集 D_{test} [19,26-28]。利用 GAN 和联邦学习模型对测试集进行训练，可以得到与客户端持有数据更加相似的人工样本。生成器和判别器的优化过程如下所示：

$$V_G = E_{z\sim P_z}\left[\ln\left(1-H\left(G\left(z\right)\right)\right)\right] \tag{11}$$

$$V_H = E_{x\sim P_{test}}\left[\ln H\left(x\right)\right] + E_{z\sim P_z}\left[\ln\left(1-H\left(G\left(z\right)\right)\right)\right] \tag{12}$$

其中 $x\sim P_{test}$ 为 D_{test} 中的样本，$z\sim P_z$ 表示一个随机分布的

向量。于是，GAN 的训练过程可以被描述为判别器 H 最大化 V_H 以区分真实数据和生成样本，生成器 G 最小化 V_G 使得生成的人工样本和真实数据尽可能相似。这样的对抗性训练持续进行，直到判别器无法区分生成样本和真实数据为止。

4 基于 GAN 的投毒攻击检测方案

4.1 检测指标

在本文所考虑的有目标攻击中，敌手通常将某一特定类型训练样本的标签 L_1 篡改为 L_2。如果仅考虑这两类数据，该模型可以简单地看作为一个二分类分类器模型。作为统计学中用来衡量二分类模型准确率的一种重要指标，F1 值兼顾了模型的精确率和召回率，代表了模型的全局泛化能力[29]。因此，本文设计了 F1 值损失作为有目标投毒攻击的检测指标。

定义 1（F1 值损失） F1 值损失为先前轮次全局模型与当前轮次本地模型对特定样本分类情况的 F1 值的差，即：

$$F_{W\to w_i} = F_W - F_{w_i}$$

其中 F_W 和 F_{w_i} 分别为先前全局模型和当前轮次本地模型的 F1 值。

此外，随着有目标攻击的目标样本种类增加，有目标攻击同样也会影响模型整体的准确率[19]。同时，随机攻击也会对模型准确率产生较大影响。而准确率作为机器学习领域最常见的评价指标，从直观反应了模型的性能[30]。因此，本文设计准确率损失衡量投毒攻击对模型产生的影响。

定义 2（准确率损失） 准确率损失为先前轮次全局模型与当前轮次本地模型在测试集上准确率的差，即：

$$AD_{W\to w_i} = ACC_W - ACC_{w_i}$$

其中 A_W 和 A_{w_i} 分别为先前轮次全局模型和当前轮次本地模型的准确率。

4.2 方案设计

为了抵御联邦学习中的投毒攻击，本文提出一种如图 2 所示的检测方案，评估客户端本地模型，从而检测并剔除投毒模型。其中，方案使用的检测指标为 F1 值损失和准确率损失。为了确保测试样本不依赖包含用户隐私的原始数据，本文引入生成对抗网络生成高质量的测试样本。

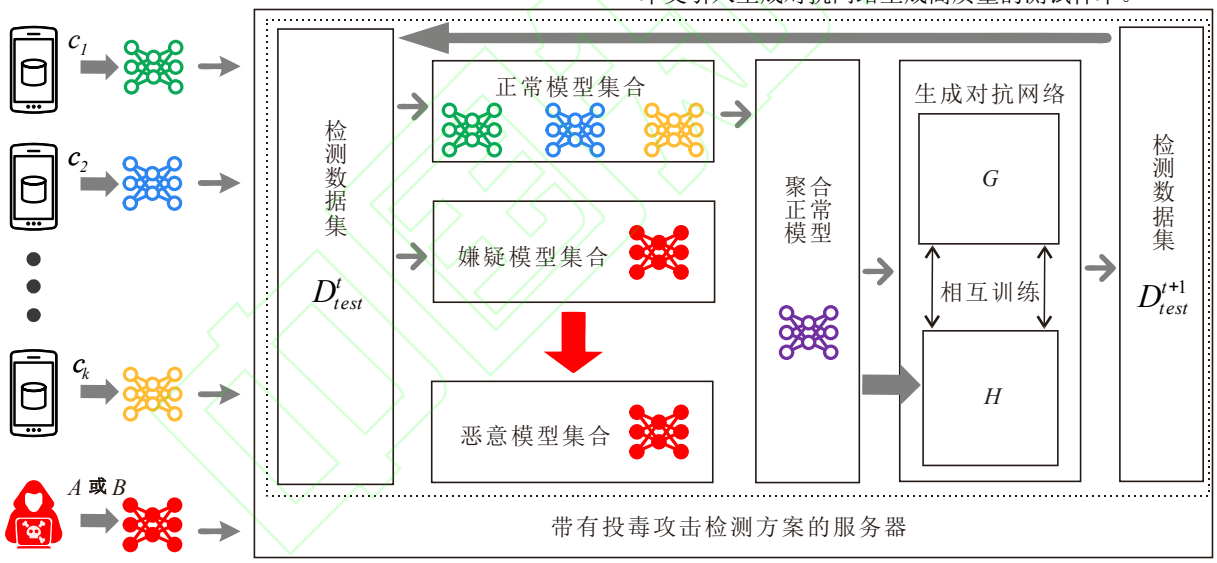


图 2 基于 GAN 的投毒攻击检测方案

Fig. 2 Poisoning attacks detecting scheme based on GAN

如算法 1 所示，在第 t 个联邦学习聚合轮次中，客户端 c_i 上传本地模型 w_i^t。中央服务器首先利用测试数据集 D_{test}^t 测试每个本地模型，并计算每个模型的 F1 值损失和准确率损失。中央服务器将 F1 值损失和准确率损失的判定阈值分别设定为 θ^t 和 γ^t，并将低于阈值的本地模型标记为正常模型，所有高于阈值的模型标记为嫌疑模型。以上判定阈值的具体值由实际应用需求决定：当联邦学习系统对投毒攻击容容忍程度较低、对模型性能时要求较高时，需要设定较小的阈值；反之，当联邦学习系统对投毒攻击容忍程度较高时，可以适当提升阈值以囊括可能存在的统计离群模型[13]。阈值设定对全局模型性能的影响见图 7 对实验结果的分析。全部正常模型对应的客户端组成的集合记为 U，全部嫌疑模型对应的客户端组成的集合记为 S。在该轮次聚合时，中央服务器仅聚合正常模型作为当前轮次的全局模型，其过程表示为：

$$W^t = \sum_{w_i\in U} \alpha_i^t \bullet w_i^t \tag{13}$$

其中 $\alpha_i^t = |D_i| / \sum_{w_i \in U} |D_i|$。随后，中央服务器将该轮次聚合的全局模型 W^t 的参数作为判别器 H 的参数，对抗训练 G 和 H，并输出检测数据集（即生成测试集）D_{test}^{t+1} 作为下一聚合轮次的测试数据集。注意到判别器 H 和机器学习模型 w 具有相同的模型结构，生成器 G 输出的测试样本和原测试集中样本尺寸相同。其训练过程如下所示：

$$V_G = E_{z\sim P_{t+1}}\left[\ln\left(1-H\left(G\left(z\right)\right)\right)\right] \tag{14}$$

$$V_H = E_{x\sim P_t}\left[\ln H\left(x\right)\right] + E_{z\sim P_{t+1}}\left[\ln\left(1-H\left(G\left(z\right)\right)\right)\right] \tag{15}$$

最后，中央服务器下发全局模型，各客户端继续使用本地数据对全局模型进行训练。上述本地模型上传、投毒模型检测、本地模型聚合、测试样本生成过程重复执行，直到全局模型收敛为止。

此外，方案还设计了阈值判定机制。规定在全局模型迭代训练过程中，超过 τ 次上传嫌疑模型的客户端被标记为恶意客户端，中央服务器在之后的聚合轮次中不再接收其上传的本地模型。安全参数 τ 在一定程度上反映了联邦学习系统对投毒攻击的容忍程度。当 τ 取较大值时，联邦学习系统对投毒攻击的容忍程度较高，例如，由于客户端的模型上传通信链路普遍存在较大噪声，本地模型性能普遍较差；对应地，当 τ 取较小值时，联邦学习系统对投毒攻击的容忍程度较低，例如，当联邦学习系统中对全局模型性能要求较高时。因此，τ 的取值视具体联邦学习系统对投毒攻击的容忍程度而定。这样的设计确保了由通信故障、本地模型差异等因素导致的误判可以被修正，从而保证了所提检测方案的误判鲁棒性。

4.3 复杂度分析

为了讨论所提方案的复杂度，规定 ϖ 为本地模型尺寸，令 Φ_1 和 Φ_2 分别表示中央服务器测试单个本地模型与训练 GAN 的计算复杂度。根据所提方案，在第 t 个联邦学习聚合轮次中，中央服务器首先使用测试数据集检测所有的本地模型，然后计算每个模型的 F1 值损失与准确率损失并分别与阈值比较，最后使用全局模型 W^t 的参数训练 GAN。因此，所提方案的计算复杂度为 $\mathcal{O}\left(n\Phi_1 + 4n + \Phi_2\right)$。此外，在每个聚合轮次中客户端与服务器只产生一次交互，因此所提方案的通信复杂度为 $\mathcal{O}\left(2n\varpi\right)$。

算法 1　投毒模型检测算法

输入:客户端集合 $N=\{c_i : i=1,2,\ldots,n\}$,经过扰动的本地模型集合 $\{\hat{w}_i^t : i=1,2,\ldots,n\}$,客户端嫌疑次数集合 $\{\lambda_i^t : i=1,2,\ldots,n\}$,安全参数 τ

输出:联邦学习全局模型 W^t

FOR 客户端 c_i IN N DO

服务器使用测试数据集 D_{test}^t 对 $\hat{w}_i^t$ 计算 θ_i^t 和 γ_i^t；

IF θ_i^t 和 γ_i^t 均小于阈值 THEN

将 c_i 添加进正常用户集合 U；

ELSE

将 c_i 添加进嫌疑用户集合 V；

c_i 被标记为嫌疑的次数 $\lambda_i = \lambda_i + 1$；

IF $\lambda_i > \tau$ THEN

从 V 中剔除 c_i；

END IF

END IF

END FOR

更新客户端集合 $N = U + V$

聚合 U 中的模型得到全局模型 W^t；

基于 W^t 训练生成器和判别器,并产生检测数据集 D_{test}^{t+1}；

RETURN W^t

5 实验结果与分析

5.1 实验数据

为了验证所提方案在部署到联邦学习系统中检测并剔除投毒模型的有效性，本文设计使用联邦学习系统训练图像分类模型，所使用的数据集为公开数据集 MNIST 和 Fashion-MNIST。其中，MNIST 为手写数字图片数据集，Fashion-MNIST 为商品图片数据集。两个数据集各包含有 6 万张训练样本和 1 万张测试样本，每个样本均为 28×28 像素的灰度图。

5.2 实验设置

实验对联邦学习系统做如下规定：中央服务器持有初始测试数据集的样本数量为 5000，每种类别的样本数量大致相同；系统中共有 30 个拥有本地模型训练能力的客户端，敌手控制其中 14 个客户端。为了模拟现实场景中联邦学习的数据非独立同分布情况，本实验将 MNIST 和 Fashion-MNIST 训练集中的样本按照期望为 1000，方差为 2500 的正态分布划分为 30 份作为各客户端的本地训练集。

MNIST 数据集和 Fashion-MNIST 数据集中的原始测试样本分别如图 3(a)和图 3(d)所示，其清晰度较高，可以表示中央服务器有能力获得高质量检测数据的理想情况。但这与现实场景中央服务器无法获得高质量测试样本的困境不符。

因此本实验对其做干扰处理，模拟在现实场景中采用收集检测数据方法时遇到的传感器故障、信道不理想、解码错误等客观问题[19]。通过在测试集中各测试样本中分别加入高斯噪声、泊松噪声和椒盐噪声中的组合，最终形成如图 3(b)和图 3(e)的低质量测试集。方便起见，将干扰处理前后的测试集分别称为理想测试集和原始测试集。通过观测可以发现，原始测试集中样本模糊不易辨认，而图 3(c)和图 3(f)所示的生成测试集中样本具有和理想测试集中样本相近的清晰度，质量较高。

图 3 MNIST 数据集和 Fashion-MNIST 数据集三种检测数据

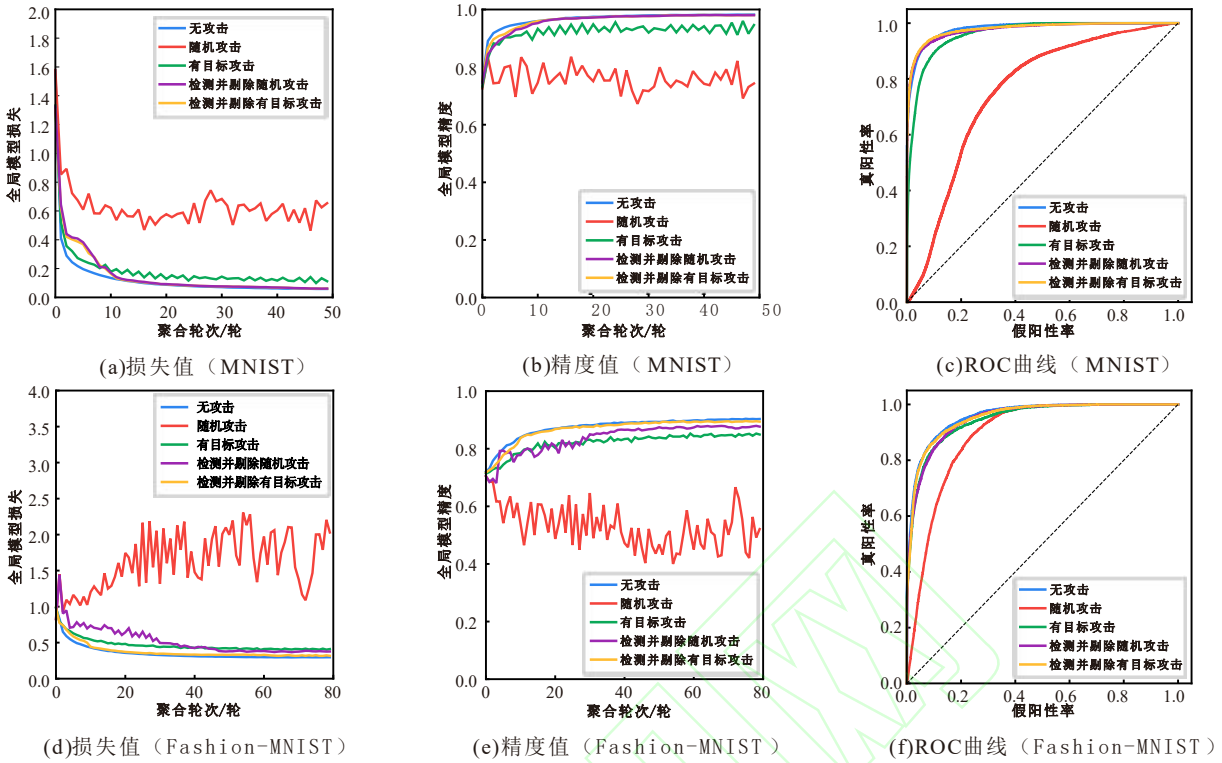
Fig. 3 Detection data of MNIST dataset and Fashion-MNIST dataset

针对不同的实验数据集，实验中图像分类模型均为卷积神经网络（Convolutional Neural Networks，CNN）。客户端和中央服务器持有的模型结构相同，均由 3 个卷积层和 2 个全连接层组成。每个卷积层后接一个池化层，全连接层后接一个柔性最大值（Softmax）激活函数。本地模型训练过程设定随机梯度下降（Stochastic Gradient Descent，SGD）为模型优化算法，并设定模型初始学习率为 0.001，衰减率为$1e^{-6}$。中央服务器用于生成测试样本的网络为 GAN。其中生成器由 1 个上采样层、2 个卷积层、2 个池化层和 1 个全连接层组成。全连接层后接一个双曲正切激活函数 tanh。此外，实验设定将当前轮次聚合后的全局模型作为判别器模型，其网络结构和图像分类模型结构相同。同时设定 GAN 使用 Adam 算法进行模型优化，其初始学习率为 0.0002，衰减率为 0.5。

将所提方案中的默认阈值分别设定 F1 值损失阈值$\theta = 0.05$、准确率损失阈值$\gamma = 5\%$以及安全参数$\tau = 5$。以上阈值为本实验设置下所能获得最优全局模型的临界值。此外，不同阈值对全局模型性能的影响在实验结果图 6 与图 7 中展现。

5.3 实验结果与分析

为了验证本文所提方案抵御投毒攻击的有效性和广泛性，分别将其部署在受到随机攻击和有目标攻击的联邦学习场景中，与未受到投毒攻击和分别受到两种投毒攻击的联邦学习进行比较。不同场景中的全局模型最终性能记录在表 1 和表 2 中。图 4(a)(b)(c)和图 4(d)(e)(f)分别展示了 MNIST 数据集和 Fashion-MNIST 数据集在不同实验设置下的全局模型训练过程。“无攻击”表示未受到任何攻击的全局模型；“随机攻击”和“有目标攻击”分别表示受到随机攻击和受到有目标攻击的全局模型；“检测并剔除随机攻击”和“检测并剔除有目标攻击”分别表示在受到两种投毒攻击时，采用所提检测方案对投毒模型进行检测并剔除投毒模型后的全局模型。以 MNIST 数据集为例分析实验结果。观察图 4(a)(b)可以发现，在受到两种投毒攻击时，全局模型收敛速度变缓，准确率下降，损失的最终收敛值上升。同时由表 1 可知，在无攻击场景下，全局模型的最终损失值为 0.0586，最终准确率为 98.2%。在随机攻击场景下，全局模型的最终损失值为 0.6541，最终准确率为 74.3%；在有目标攻击场景下，全局模型的最终损失值为 0.1121，最终准确率为 94.6%。而在部署所提检测方案后，随机攻击场景和有目标攻击场景下的全局模型最终损失值分别为 0.0625 和 0.0609，最终准确率分别为 97.4%和 98.0%，两者都接近无攻击场景下的全局模型收敛效果，即损失值仅增加了 0.0023，准确率仅减少了 0.2 个百分点。以上结果表明，在受到两种类型的投毒攻击时全局模型最终性能下降，而部署所提检测方案后的全局模型最终性能和未受到投毒攻击时接近。观察图 4(c)所示的全局模型的受试者特征（Receiver Operating Characteristic，ROC）曲线也可得到相似的结论。观察图 4(a)(b)还可以发现，在联邦学习初始阶段（0-10 联邦学习聚合轮次），部署所提检测方案的全局模型收敛速度低于未受到投毒攻击的全局模型，其主要原因是所提检测方案在联邦学习初始阶段并未生成高质量的测试样本，无法准确检测出全部投毒模型。但随着联邦学习的进行，所提检测方案可以生成高质量检测数据，从而准确高效地检测和剔除投毒模型。



(a)损失值（MNIST） (b)精度值（MNIST） (c)ROC曲线（MNIST）

(d)损失值（Fashion-MNIST） (e)精度值（Fashion-MNIST） (f)ROC曲线（Fashion-MNIST）

图 4 MNIST 数据集和 Fashion-MNIST 数据集下的全局模型性能

Fig. 4 Global model performance over the MNIST dataset and the Fashion-MNIST dataset

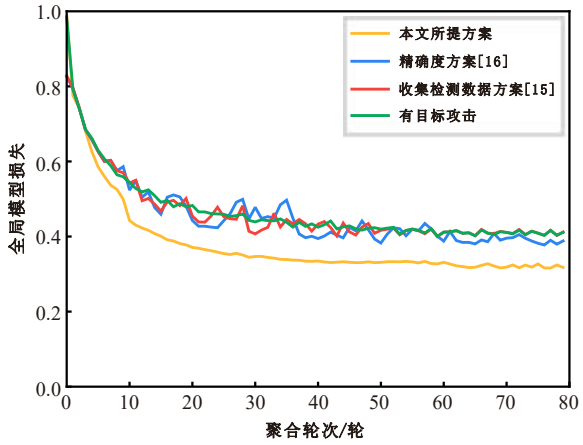
表 1 MNIST 数据集下全局模型性能

Tab. 1 Global model performance over the MNIST dataset

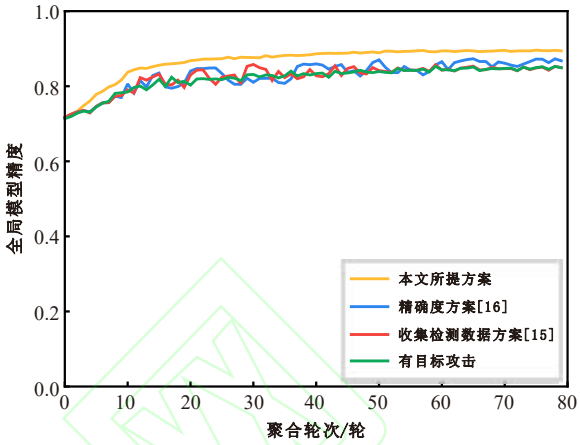
场景	全局模型损失值		全局模型准确率/%		曲线下面积	
	随机攻击	有目标攻击	随机攻击	有目标攻击	随机攻击	有目标攻击
无攻击	0.0586		98.2		0.9590	
无抵御机制	0.6541	0.1121	74.3	94.6	0.8916	0.9466
采用所提方案	**0.0625**	**0.0609**	**97.4**	**98.0**	**0.9499**	**0.9547**

为验证本文所提方案抵御投毒攻击的高效性，将其与方案[15]和[16]分别在两种攻击场景进行比较。全局模型最终性能记录在表 2 中，并在图 5 中记录了 Fashion-MNIST 数据集上采用不同方案检测并剔除两种投毒模型后的全局模型性能曲线。其中，“收集检测数据方案[15]”表示在原始测试集上比较 F1 值损失和准确率损失的方案；“准确率方案[16]”表示在生成测试集上比较准确率的方案；“本文所提方案”表示在生成测试集上比较 F1 值损失和准确率损失的方案。观察图 5(a)并结合表 2 可知，在受到有目标攻击时，全局模型损失值为 0.4110。“收集检测数据方案”通过在原始测试集上比较 F1 值损失和准确率损失，将全局模型的损失值降低为 0.4120；“准确率方案”通过在生成测试集上比较准确率，将全局模型的收敛损失值降低为 0.3881，两种方案对有目标攻击投毒的抵御效果都不明显。可见，由于仅检测在所有类别样本上的全局模型准确率，而忽略了在某一特定类别上的全局模型准确率，因此方案[15]和[16]都不能有效地检测出有目标攻击投毒模型，无法有效地抵御有目标攻击。而本文所提方案在生成测试集上比较 F1 值损失，可以有效地检测出有目标攻击投毒模型，从而将全局模型的损失值降低为 0.3176，这一数值和未受到投毒攻击时的全局模型的损失值相近。可见，所提方案可以对有目标攻击的投毒模型进行有效地检测和剔除。观察图 5(b)中的全局模型准确率变化曲线可以得到相同的结论。并由表 2 可知，在受到有目标攻击时，全局模型的最终准确率值为 84.9%。采用方案[15]和[16]时，全局模型的准确率分别为 84.9%和 86.7%。而在部署本文所提方案后，全局模型的准确率提高到 89.4%，这一数值同样和未受到投毒攻击时相近。与方案[15]和[16]相比，全局模型的准确率分别提高了 4.5 个百分点和 2.7 个百分点。图 5(c)和图 5(d)分别为随机攻击场景下的全局模型损失值和全局模型准确率变化曲线。观察图(c)并结合表 2 可知，在受到随机攻击时，全局模型损失值为 2.0298，并存在较大震荡。采用方案[15]和[16]时，全局模型的损失值分别降低为 0.7362 和 0.5492；而在本文所提方案后，全局模型的收敛损失值降低为 0.3768，这一数值和未受到投毒攻击时的全局模型的损失

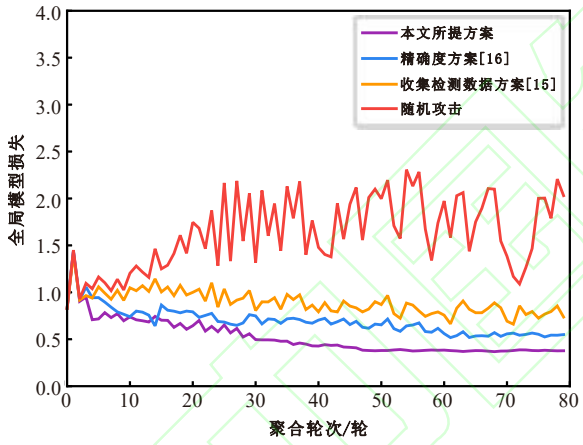
值相近。可见，所提方案可以对随机攻击的投毒模型进行有效地检测和剔除。观察图 5(d)中的全局模型准确率变化曲线并结合表 2 可以得到相同的结论。在受到随机攻击时，全局模型的准确率呈现下降趋势，最终准确率值为 52.1%，并存在较大震荡；采用方案[15]和[16]后，全局模型的准确率分别提高到 75.5%和 82.7%；而在采用本文所提方案后，全局模型的准确率提高到 87.7%，这一数值同样和未受到投毒攻击时相近。与方案[15]和[16]相比，全局模型的准确率分别提高了 12.2 个百分点和 5.0 个百分点。综合以上结果，相较于仅能抵御随机攻击的方案[15]和[16]，所提方案不仅能更好地抵御随机攻击，同时可以有效地抵御有目标攻击。



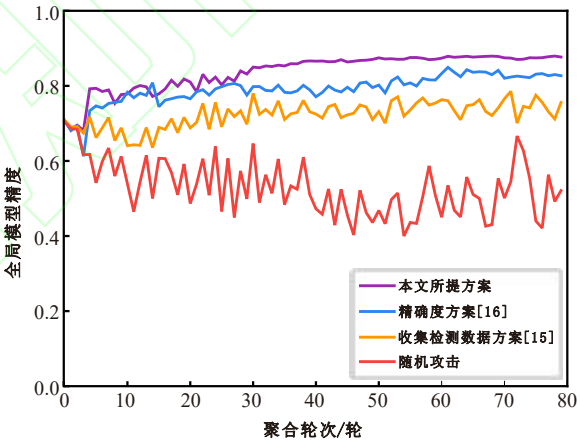
(a)全局模型损失值（有目标攻击）



(b)全局模型精度值（有目标攻击）



(c)全局模型损失值（随机攻击）



(d)全局模型精度值（随机攻击）

图 5　Fashion-MNIST 数据集下使用不同方案检测并剔除投毒模型

Fig. 5　Different schemes to detect and eliminate poisoning models over the Fashion-MNIST dataset

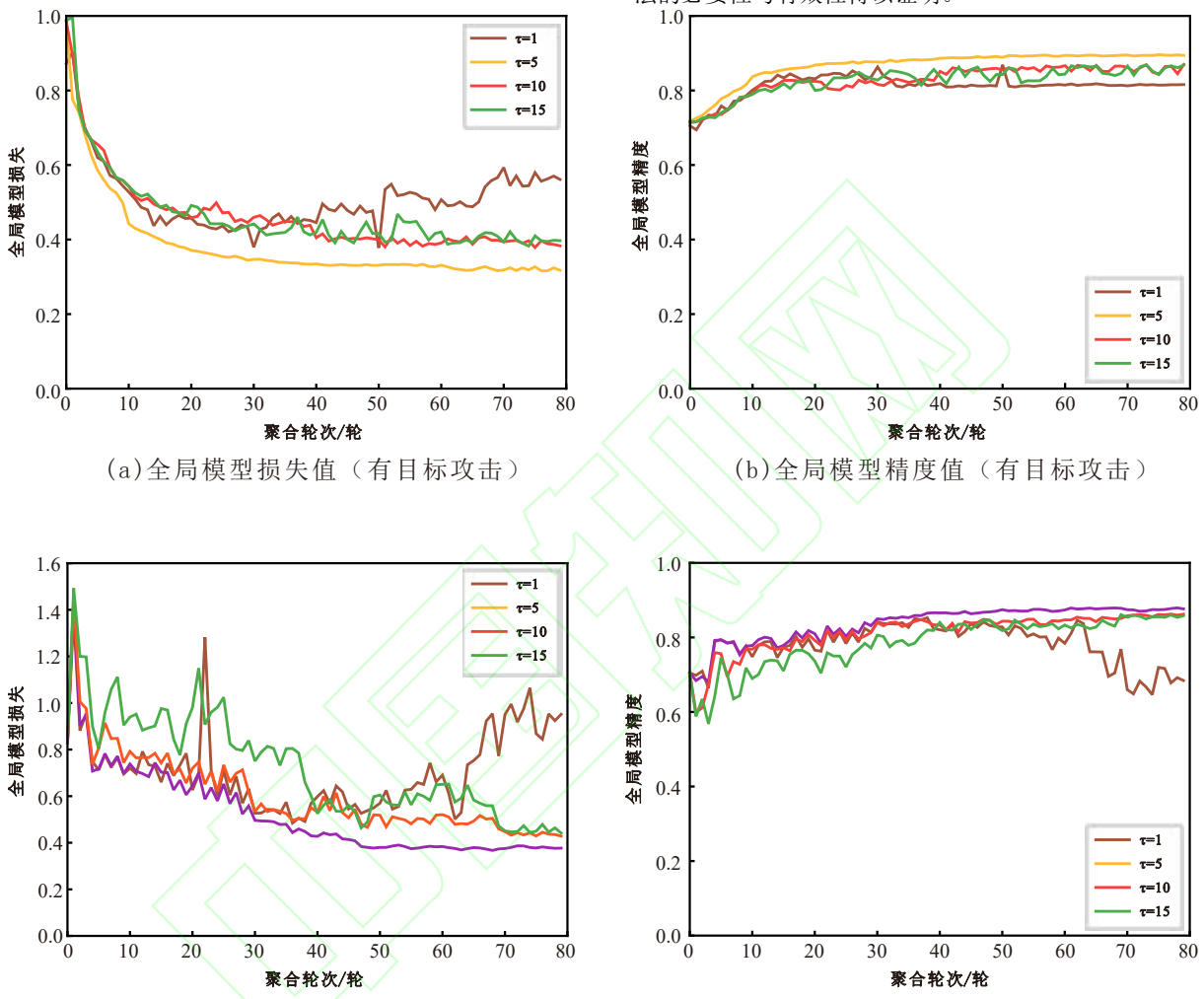
表 2 Fashion-MNIST 数据集下全局模型性能

Tab. 2　Global model performance over the Fashion-MNIST dataset

场景	全局模型损失值		全局模型准确率/%	
	随机攻击	有目标攻击	随机攻击	有目标攻击
无攻击	0.2968		90.3	
无抵御机制	2.0298	0.4110	52.1	84.9
采用所提方案	**0.3768**	**0.3176**	**87.7**	**89.4**
采用收集检测数据方案[15]	0.7362	0.4120	75.5	84.9
采用准确率方案[16]	0.5492	0.3881	82.7	86.7

为了验证所提方案中阈值判定方法对误判鲁棒性的影响，在 Fashion-MNIST 数据集上设置不同的安全参数 τ 重复实验，图 6 展示了所提方案使用不同安全参数抵御投毒攻击时的全局模型训练过程，其中，$\tau=1$ 表示不使用阈值判定方法。$\tau=5$、$\tau=10$ 以及 $\tau=15$ 分别表示某客户端在被 5 次、10 次和 15 次判定为恶意客户端后将被剔除。以图 6(a)和图 6(b)展示的有目标攻击场景为例分析实验结果。观察图 6(a)中的损失值曲线可以发现，在受到有目标攻击时，若不

使用阈值判定方法，由于客户端被判定为恶意客户端后即被剔除，因而存在有误判的可能。这导致中央服务器用于聚合的本地模型不能准确代表联邦学习系统的数据分布特征，严重降低了全局模型收敛性，全局模型损失值在 30 个聚合轮次后出现上升趋势，最终存在较大震荡。随着安全参数的逐渐增大，系统对误判的容错能力逐渐增强，全局模型损失值下降更加平缓。然而，当 τ 值持续增大时，系统的容错能力被过度放大，导致恶意客户端被确定为敌手的判定周期过长，嫌疑模型将影响更多的聚合轮次，因此全局模型损失值也存在较大震荡。以上结果表明，不使用阈值判定方法或者使用过大的安全参数都将影响所提方案的抵御投毒攻击的性能。此外，在设定的实验条件下，最优的安全参数值为 5。观察图 6(b)中的准确率曲线可以得到相似的结论：此时全局模型的最终准确率达到最高，并明显高于其他设定。阈值判定方法的必要性与有效性得以证明。



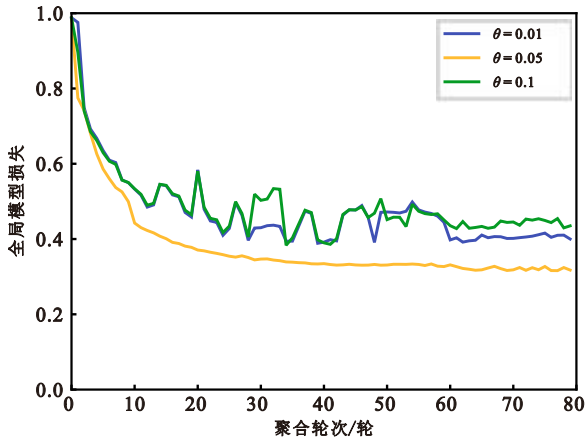
(a)全局模型损失值（有目标攻击）　(b)全局模型精度值（有目标攻击）

(c)全局模型损失值（随机攻击）　(d)全局模型精度值（随机攻击）

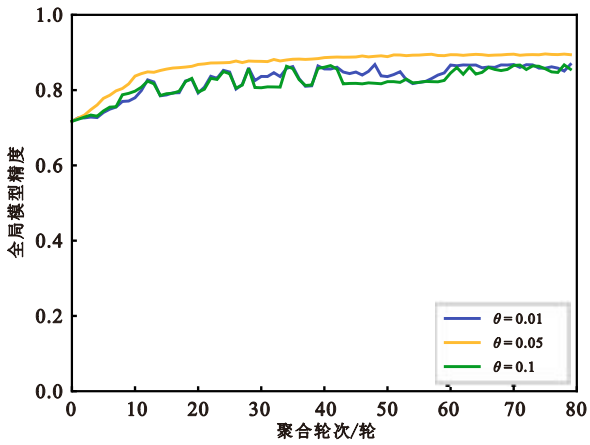
图 6　阈值判定方法对所提方案性能的影响

Fig. 6　The impact of the threshold method on the performance of proposed scheme

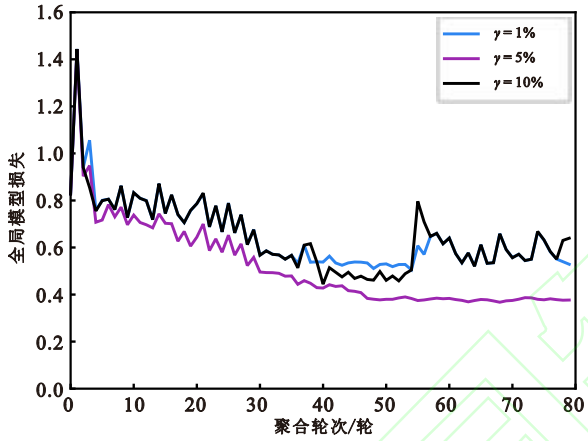
为了验证检测指标的判定阈值对所提方案性能的影响，在 Fashion-MNIST 数据集上，设置不同的 F1 值损失判定阈值 θ 和准确率损失判定阈值 γ 并重复实验。图 7 展示了所提方案使用不同判定阈值抵御投毒攻击时的全局模型训练过程。考虑到仅仅依据准确率损失无法有效抵御有目标攻击，设定在有目标攻击场景下 F1 值损失判定阈值 θ 取值为 0.01、0.05 和 0.1，图 7(a)和图 7(b)为此时的训练结果。从图中可以看到，当 θ 取 0.05 时，中央服务器能有效地检测出嫌疑模型，同时避免正常用户被误判，全局模型因此收敛最为平稳，最终准确率也最高。而当 θ 取较小的 0.01 时，中央服务器针对 F1 值的检测较为严格，存在较大的误判可能，因此全局模型收敛在过程中存在较大幅度的震荡。同样地，当 θ 取较大的 0.1 时，中央服务器的容错性能更强，但需要更多次检验才能发现投毒模型，因此全局模型的收敛速度明显变慢。以上实验结果表明，在有目标攻击场景下，使用过小或过大的 F1 值损失阈值都将降低所提方案抵御有目标攻击的有效性。此外，在随机攻击场景下，设定准确率损失判定阈值 γ 取值为 1%、5%和 10%，图 7(c)和图 7(d)为此时的训练结果。



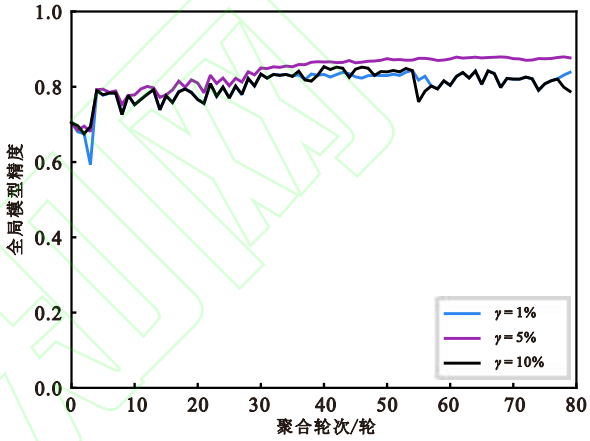
(a) 全局模型损失值（有目标攻击）



(b) 全局模型精度值（有目标攻击）



(c) 全局模型损失值（随机攻击）



(d) 全局模型精度值（随机攻击）

图 7　检测指标的判定阈值对所提方案性能的影响

Fig. 7　The impact of the threshold of the detection index on the performance of proposed scheme

6　结语

联邦学习作为一种分布式机器学习框架，以其良好的隐私保护性能和通信效率得到广泛的应用。然而其聚合本地模型的范式易于受到投毒攻击的威胁。为了检测并抵御投毒攻击，本文提出一种基于生成对抗网络的投毒攻击检测方案。所提检测方案利用生成对抗网络产生测试样本，并使用 F1 值损失和准确率损失作为检测指标对模型进行阈值检测从而剔除投毒模型。在 MNIST 和 Fashion-MNIST 数据集上进行的实验表明，通过部署所提检测方案，中央服务器可以获得高质量的测试样本，从而有效地检测并剔除全部两种类型的投毒攻击，最终获得性能良好的联邦学习全局模型。

现阶段联邦学习系统的投毒攻击与抵御措施研究都建立在无隐私保护机制联邦学习场景中。由于中央服务器可以直接获取客户端的本地模型，因此存在隐私泄漏的风险。而在使用同态加密与差分隐私等隐私保护联邦学习场景下，现有的抵御措施无法有效检测并剔除投毒模型。因此，在隐私保护联邦学习的现实场景中，针对密态模型和扰动模型的投毒攻击抵御措施有待进一步研究。

参考文献

[1] JORDAN M I, MITCHELL T M. Machine learning: trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

[2] VOIGT P, VON DEM BUSSCHE A. The eu general data protection regulation (gdpr)[S]. Cham: Springer, 2017.

[3] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[C]// Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Brookline: Microtome Publishing, 2017: 1273-1282.

[4] XU J, GLICKSBERG B S, SU C, et al. Federated learning for healthcare informatics[J]. Journal of Healthcare Informatics Research, 2021, 5(1): 1-19.

[5] CHEN Q, WANG Z L, LIN X D. PPT: a privacy-preserving global model training protocol for federated learning in P2P networks[J]. Computers & Security, 2023, 124: 102966.

[6] TIAN Z, CUI L, LIANG J, Yu, S. A comprehensive survey on poisoning attacks and countermeasures in machine learning[J]. ACM Computing Surveys, 2022, 55(8): 1-35.

[7] 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.

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

[9] FANG M H, CAO X Y, JIA J Y, et al. Local model poisoning attacks to byzantine-robust federated learning[C]// Proceedings of the 29th USENIX Security Symposium. Berkeley: USENIX Association, 2020: 1605-1622.

[10] TAHMASEBIAN F, XIONG L, SOTOODEH M, et al. Crowdsourcing under data poisoning attacks: a comparative study[C]// Proceedings of the 2020 IFIP Annual Conference on Data and Applications Security and Privacy. Cham: Springer, 2020: 310-332.

[11] YIN D, CHEN Y, KANNAN R, Byzantine-robust distributed learning: Towards optimal statistical rates[C]// Proceedings of the 35th International Conference on Machine Learning. San Diego: JMLR, 2018: 5650-5659.

[12] 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 Proceedings Systems. La Jolla: NIPS, 2017: 118-128.

[13] MUÑOZ-GONZÁLEZ L, CO K T, LUPU E C. Byzantine-robust federated machine learning through adaptive model averaging. [EB/OL]. (2019-09-11) [2022-04-25]. https://arxiv.org/pdf/1909.05125.pdf.

[14] 陈宛桢, 张恩, 秦磊勇, 等. 边缘计算下基于区块链的隐私保护联邦学习算法[J/OL]. 计算机应用. (2022-09-21) [2022-11-04] https://kns.cnki.net/kcms/detail/51.1307.TP.20220920.1049.006.html. (CHEN W Z, ZHANG E, QIN L Y, Privacy-preserving federated learning algorithm based on block chain in edge computing [J/OL]. Journal of Computer Application. (2022-09-21) [2022-11-04] https://kns.cnki.net/kcms/detail/51.1307.TP.20220920.1049.006.html.)

[15] JAGIELSKI M, OPREA A, BIGGIO B, et al. Manipulating machine learning: poisoning attacks and countermeasures for regression learning[C]// Proceedings of the 39th IEEE Symposium on Security and Privacy. Piscataway: IEEE, 2018: 19-35.

[16] ZHAO Y, CHEN J, ZHANG J, et al. Detecting and mitigating poisoning attacks in federated learning using generative adversarial networks[J]. Concurrency and Computation: Practice and Experience, 2020, 34(7): e5906.

[17] FENG J, XU H, MANNOR S, et al. Robust logistic regression and classification[C]// Proceedings of the 27th International Conference on Neural Information Proceeding Systems. La Jolla: NIPS, 2014: 253-261..

[18] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J] Communications of the ACM, 2020, 63(11): 139-144.

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

[20] SHEJWALKAR V, HOUMANSADR A. Manipulating the byzantine: optimizing model poisoning attacks and defenses for federated learning[C]// Proceedings of 28th Annual Network and Distributed System Security Symposium. Reston: INTERNET SOC, 2021:1-19.

[21] ALKHUNAIZI N, KAMZOLOV D, TAKÁČ M, et al. Suppressing poisoning attacks on federated learning for medical imaging[C]// Proceedings of the 2022 International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2022, 13438: 673-683.

[22] SUN J, LI A, DIVALENTIN L, HASSANZADEH A, CHEN Y, LI. H. Fl-wbc: enhancing robustness against model poisoning attacks in federated learning from a client perspective[C]// Proceedings of the 2021 Advances in Neural Information Proceedings Systems 34. La Jolla: NIPS, 2021:12613-12624.

[23] NGUYEN T D, RIEGER P, MIETTINEN M, SADEGHI A R. Poisoning attacks on federated learning-based IoT intrusion detection system[C]// Proceedings of the 2020 Decentralized IoT Systems and Security Workshop. Washington: Internet Society, 2020: 1-7.

[24] GUERRAOUI R, ROUAULT S. The hidden vulnerability of distributed learning in byzantium[C]// Proceedings of the 35th International Conference on Machine Learning. San Diego: JMLR, 2018: 3521-3530.

[25] STEINHARDT J, KOH P W, LIANG P. Certified defenses for data poisoning attacks[C]// Proceedings of the 31st International Conference on Neural Information Proceedings Systems. La Jolla: NIPS, 2017: 3520-3532.

[26] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 4681-4690.

[27] CHEN Z, ZHU T, XIONG P, et al. Privacy preservation for image data: a GAN-based method[J]. International Journal of Intelligent Systems, 2021, 36(4): 1668-1685.

[28] WANG Z, SONG M, ZHANG Z, SONG Y, WANG Q, QI H. Beyond inferring class representatives: user-level privacy leakage from federated learning[C]// Proceedings of the 2019 IEEE Conference on Computer Communications. Piscataway: IEEE, 2019: 2512-2520.

[29] TRUEX S, BARACALDO N, Anwar A, Steinke T, Ludwig H, Zhang R, Zhou Y. A hybrid approach to privacy-preserving federated learning[C]// Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2019: 1–11.

[30] TOLPEGIN V, TRUEX S, GURSOY ME, LIU L. Data poisoning attacks against federated learning systems[C]// Proceedings of the 2020 European Symposium on Research in Computer Security. Cham: Springer, 2022, 12308: 480-501.

This work is partially supported by the National Natural Science Foundation of China (No. 62172319，U19B200073).

CHEN Qian, born in 1993, Ph. D., candidate. His research interests include privacy preservation, machine learning, federated learning.

CHAI Zheng, born in 1999, M. S. candidate. His research interests include federated learning, poisoning attack.

WANG Zilong, born in 1982, Ph. D., professor. His research interests include information theory, cryptography.

CHEN Jiawei, born in 1998, M. S. candidate. His research interests include federated learning, reinforcement learning.