# 基于模型水印的联邦学习后门攻击防御方法

郭晶晶[1] 刘玖樽[1] 马 勇[2] 刘志全[3] 熊宇鹏[1]
苗 可[1] 李佳星[1] 马建峰[1]

[1](西安电子科技大学网络与信息安全学院 西安 710071)
[2](江西师范大学计算机科学技术学院 南昌 330022)
[3](暨南大学网络空间安全学院 广州 510632)

**摘 要** 联邦学习作为一种隐私保护的分布式机器学习方法,容易遭受参与方的投毒攻击,其中后门投毒攻击的高隐蔽性使得对其进行防御的难度更大.现有的多数针对后门投毒攻击的防御方案对服务器或者恶意参与方数量有着严格约束(服务器需拥有干净的根数据集,恶意参与方比例小于 50%,投毒攻击不能在学习初期发起等).在约束条件无法满足时,这些方案的效果往往会大打折扣.针对这一问题,本文提出了一种基于模型水印的联邦学习后门攻击防御方法.在该方法中,服务器预先在初始全局模型中嵌入水印,在后续学习过程中,通过验证该水印是否在参与方生成的本地模型中被破坏来实现恶意参与方的检测.在模型聚合阶段,恶意参与方的本地模型将被丢弃,从而提高全局模型的鲁棒性.为了验证该方案的有效性,本文进行了一系列的仿真实验.实验结果表明该方案可以在恶意参与方比例不受限制、参与方数据分布不受限制、参与方发动攻击时间不受限制的联邦学习场景中有效检测恶意参与方发起的后门投毒攻击.同时,该方案的恶意参与方检测效率相比于现有的投毒攻击防御方法提高了 45%以上.

**关键词** 联邦学习;投毒攻击;后门攻击;异常检测;模型水印
**中图法分类号** TP309 **DOI 号** 10.11897/SP.J.1016.2024.00662

# Backdoor Attack Defense Method for Federated Learning Based on Model Watermarking

GUO Jing-Jing[1] LIU Jiu-Zun[1] MA Yong[2] LIU Zhi-Quan[3] XIONG Yu-Peng[1]
MIAO Ke[1] LI Jia-Xing[1] MA Jian-Feng[1]

[1](*School of Cyber Engineering, Xidian University, Xi'an* 710071)
[2](*School of Computer Science and Technology, Jiangxi Normal University, Nanchang* 330022)
[3](*College of Cyber Security, Jinan University, Guangzhou* 510632)

**Abstract** As a privacy-preserving distributed machine learning paradigm, federated learning is vulnerable to poison attacks. The high crypticity of backdoor poisoning makes it difficult to defend against. Most existing defense schemes against backdoor poisoning attacks have strict constraints on the servers or malicious participants (servers need to have a clean root dataset, the proportion of malicious participants should be less than 50%, and poisoning attacks cannot be initiated at the beginning of learning, etc.). When these constraints cannot be met, the effectiveness of these schemes will be greatly compromised. To solve this problem, this paper proposes a

收稿日期:2023-06-27;在线发布日期:2023-12-28.本课题得到国家自然科学基金(62272195,61932010,62032025)、陕西省自然科学基础研究计划资助项目(2022JQ-603)、中央高校基本科研业务费专项资金(ZYTS23161,21622402)、广东省网络与信息安全漏洞研究重点实验室项目(2020B1212060081)、广州市科技计划项目(202201010421)资助.**郭晶晶**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为无线网络安全、联邦学习安全、信任管理.E-mail:jjguo@xidian.edu.cn.**刘玖樽**,硕士,主要研究方向为联邦学习安全.**马 勇**(通信作者),博士,教授,主要研究领域为云计算、边缘计算、信息安全.E-mail:may@jxnu.edu.cn.**刘志全**,博士,副研究员,中国计算机学会(CCF)会员,主要研究方向为信任管理、车载网络安全、无线网络安全.**熊宇鹏**,硕士研究生,主要研究方向为联邦学习安全.**苗 可**,本科生,主要研究方向为联邦学习安全.**李佳星**,硕士研究生,主要研究方向为联邦学习安全.**马建峰**,博士,教授,主要研究领域为密码学与网络安全.

secure aggregation method for federated learning based on model watermarking. In this method, the server embeds a watermark in the initial global model in advance. In the subsequent learning process, it detects malicious participants by verifying whether the watermark has been destroyed in the local model generated by the participants. In the model aggregation stage, the local models uploaded by malicious participants will be discarded, thereby improving the robustness of the global model. In order to verify the effectiveness of this scheme, a series of simulation experiments were conducted. Experimental results show that this scheme can effectively detect backdoor poisoning attacks launched by malicious participants in various scenarios where the proportion of malicious participants is unlimited, the distribution of participants' data is unlimited, and the attack time of participants is unlimited. Moreover, the detection efficiency of the scheme is more than 45% higher than that of the auto encoder-based poison attack defense method.

**Keywords**　federated learning; poisoning attack; backdoor attack; anomaly detection; model watermarking

## 1　引　言

各国陆续出台的数据安全与隐私保护法律法规[1-2]以及实际应用场景中普遍存在的数据孤岛问题为传统的集中式机器学习模式带来了巨大的挑战.针对这一问题,谷歌提出了联邦学习(Federated Learning,FL)的概念[3].作为一种分布式机器学习框架,联邦学习可以在各参与方的数据不出本地的情况下协同训练机器学习模型,从而保护参与方的数据隐私.联邦学习的模型训练流程为:各参与方利用本地数据进行模型训练得到本地模型,然后将本地模型信息(参数或梯度)上传至聚合服务器,聚合服务器利用给定的聚合规则对本地模型进行聚合得到全局模型供所有参与方共享.

参与方间分布式协作以及无信任关系的特性使联邦学习系统可能受到来自参与方的各种安全威胁,其中参与方发起的投毒攻击是最有威胁的攻击方式之一[4-6].攻击者可以通过向本地数据进行投毒或者直接修改本地模型信息,达到破坏全局模型完整性,降低全局模型的精度或者嵌入后门等目的.相比于其他投毒攻击来说,后门攻击更加灵活也更具有隐蔽性[7].在后门攻击中,攻击者首先在少部分训练数据中添加隐蔽的后门(如一个特殊像素点),然后将其重新标记为目标标签,当输入数据成功触发后门时,模型会输出目标标签,正常数据的预测则不受影响[8].

目前,国内外学者已经提出了很多方案来防御常见的后门攻击[9-10].这些防御方案的核心思路为:(1)根据各参与方的本地模型的某种统计(欧氏距离、中位数等)特征来识别异常本地模型;(2)基于正常参与方的模型数据来训练机器学习模型用于识别恶意模型.第一种防御思路的有效性需要建立在各参与方的本地数据满足独立同分布(IID)并且恶意参与方数量小于参与方总量的50%的基础之上,然而实际场景中这些假设条件可能无法满足.第二种防御思路的有效性需要建立在恶意参与方数量小于参与方总量的50%并且恶意参与方不能在学习初期发动攻击的基础之上.如果放松对恶意参与方的限制(恶意参与方比例不受限制,参与方数据分布不受限制,参与方发动攻击时间不受限制),那么现有方案的防御效果将会大打折扣.

针对上述问题,本文提出了一种基于模型水印的联邦学习参与方后门投毒攻击防御方法.该方法中,聚合服务器生成带有水印的初始全局模型并下发给各参与方,被嵌入的水印具有足够的鲁棒性使得良性参与方的本地模型训练过程不会对水印的完整性产生显著的影响,而发起后门投毒攻击的恶意参与方为了增强恶意本地模型对全局模型的扰动性,会对恶意本地模型参数进行一系列处理,这一过程会使模型水印的完整性遭到破坏.因此,聚合服务器可通过验证各参与方上传的模型中的水印完整性来实现对恶意参与方的检测.上传水印完整性较低的本地模型的参与方将被视为攻击者.在模型聚合阶段,恶意参与方的本地模型将被丢弃,从而提高全局模型的鲁棒性.为了验证所提方案的有效性,本文进行了大量的仿真实验,结果显示在参与方数据独立同分布与非独立同分布的场景下,所提方案均可

以有效的防御参与方发起的多种后门投毒攻击，且防御效果不受恶意参与方数量的影响. 与现有方案相比，本文所提方案具有更好的防御效果，同时具有更小的计算开销.

本文工作的主要贡献如下：

(1) 针对现有的联邦学习后门投毒攻击防御方法对联邦学习系统提出诸多约束条件的问题，本文提出了一种基于模型水印的后门攻击防御方案，该方案可以在恶意参与方比例不受限制，参与方数据分布不受限制，参与方发动攻击时间不受限制的联邦学习场景中有效检测恶意参与方发起的多种后门投毒攻击.

(2) 本文实现了在聚合服务器端部署所提方案的联邦学习系统，通过仿真实验验证了该方案防御后门投毒攻击的有效性、可靠性以及高效性.

## 2 相关工作

目前，国内外学者针对联邦学习提出了多种攻击方案，其中包括了数据投毒攻击、模型投毒攻击以及后门攻击等. 同时，也有一系列针对联邦学习攻击的防御方案被提出，然而现有方案对联邦学习敌手以及参与方有较多约束条件. 本文拟利用模型水印的特性提出一种新的联邦学习后门攻击防御方法. 在介绍所提方案前，本节对联邦学习投毒攻击、联邦学习后门攻击的防御以及神经网络模型水印的国内外研究现状进行分析.

### 2.1 联邦学习投毒攻击

分布式与隐私保护的特性使联邦学习容易遭受投毒攻击[11]. 常见的参与方投毒攻击方法包括数据投毒攻击、模型投毒攻击和后门攻击.

数据投毒攻击中，攻击者利用生成的投毒数据污染正常数据集，从而降低由被污染数据集训练得到的模型的性能. Jamie 等人[12]证实了在具有 20 个参与方的分布式机器学习系统中，攻击者仅污染了 5% 的训练数据即可实现 38% 的攻击成功率. Cao 等人[13]首次提出了联邦学习中的分布式数据投毒攻击思路. Xia 等人[14]系统地介绍了联邦学习系统中的投毒攻击，即敌手通过篡改节点训练数据来破坏联邦学习的全局模型准确性. 典型的数据投毒攻击包括了标签翻转(Label Flipping)[15]攻击等.

模型投毒攻击通过伪造模型更新或修改良性模型更新来破坏模型聚合过程. 模型投毒攻击可以跳过本地训练的流程，直接伪造模型更新[16-20]. 相较于数据投毒攻击，模型投毒的攻击效果更强. Xie 等人[16]提出了一种基于梯度方向来构造恶意梯度的投毒方法，令恶意梯度的长度和全局梯度的长度几乎一致，但其方向和全局梯度方向的内积为负，以使全局梯度指向错误的优化方向. 鉴于单个恶意节点发起的模型投毒攻击效果十分有限，Cao 等人[17]提出了基于虚节点的模型投毒攻击算法 MPAF. 在 MPAF 算法中，敌手在联邦学习系统中注册大量虚节点，在每轮训练中，虚节点接收聚合服务器分发的全局梯度，再随机生成一个与全局梯度同维度的本地模型梯度信息作为新一轮的本地模型梯度信息上传聚合服务器. MPAF 的攻击成本低，效果显著，然而攻击的隐蔽性不强. Bhagoji 等人[18]提出的模型投毒攻击方法通过在上一轮全局梯度基础上添加一个构造梯度来实现攻击，其优化目标为获取使全局模型交叉熵损失最小的构造梯度，同时在目标函数中引入恶意节点的本地训练集的训练损失，以及构造梯度和上一轮所有节点的本地模型梯度信息的欧氏距离以使攻击更加隐蔽.

后门攻击(Backdoor Attack)是一种在模型中嵌入后门的攻击方法，其攻击目标为保持全局模型在主任务(干净数据)上的良好性能的前提下提高全局模型在目标任务上的性能，确保预定义的输入(后门数据)都有很高的概率被预测为目标类别. 根据攻击者的能力，攻击者可以发动黑盒攻击或者白盒攻击. 黑盒攻击中，攻击者只能改变客户端的训练数据，而无法对其训练流程和本地模型进行修改；白盒攻击中，攻击者能够完全控制客户端的局部训练数据和训练流程，并可以直接对本地模型进行修改. 通常情况下，攻击者会同时结合数据投毒和模型投毒的方法实施后门投毒攻击，即首先对本地训练样本进行修改以嵌入后门，然后为了避免模型聚合阶段稀释后门模型的贡献，对恶意模型的参数实施放大等操作. 在后门攻击中，攻击者能够控制一个或多个参与方实施攻击. 在联邦学习环境下，分布式后门攻击比集中式后门攻击具有更强的隐蔽性和有效性. 文献[7]给出了联邦学习中分布式后门攻击的通用框架 DBA(Distributed Backdoor Attacks). Gong 等人[21]提出了合作型后门攻击(Coordinated Backdoor Attack)方法，攻击者在多个参与方中部署了不同的本地模型后门触发器，这些本地模型后门触发器将在全局模型中组成全局后门触发器，无论是全局触发器还是本地触发器均可激活全局模型中的后门. 为了提高后门触发器的有效性，攻击者还会根据全

局模型和本地模型信息动态地调整后门触发器. 这种基于模型的合作型后门触发器充分利用了联邦学习的分布式特性，使全局模型中的后门更容易被激活生效，提高了联邦学习中后门攻击的成功率. 为了提高攻击的隐蔽性，绕过常规的防御措施，文献[18]采用了求解优化问题的方法实现后门攻击，利用交替最小化(Alternating Minimization)的优化方法在模型训练过程中将扰动性目标和隐蔽性目标分开进行优化，从而在全局模型中嵌入后门的同时，保证恶意本地模型参数与良性本地模型参数的相似度较高.

### 2.2　联邦学习后门攻击的防御

目前，国内外学者提出了很多针对联邦学习后门攻击的防御方案. Hou 等人[22]提出了通过过滤客户端本地的投毒数据来消除后门攻击的方案，然而这类方案需要提前训练大量过滤器，并且无法防御白盒攻击. 针对白盒攻击，Sun 等人[23]提出了利用差分隐私技术来削弱恶意模型对全局模型的影响，这类方案在削弱恶意模型的同时也降低了模型在主任务上的性能. 还有学者提出了基于降维的后门攻击防御方案，通过训练编码器-解码器模型来输出原始模型更新的重构，通过最小化重构误差来优化编码器-解码器模型，然后在联邦学习过程中，将重构误差高于预定义阈值的更新视为恶意更新[24]. 这类方案需要额外的正常公共数据集或者要求攻击者在学习初期不能发起攻击来保证服务器能够得到正确的模型更新，从而训练变分自编码器，而这一假设在实际应用场景中通常较难实现. 此外，还有学者提出了基于距离的防御方法，该方法通过比较各本地模型之间的距离(如余弦相似度，欧氏距离)，从而去除与其他参与方的参数距离差距最大的本地模型，这种方案通常无法抵御复杂、隐蔽的攻击. 文献[25]提出了一种拜占庭鲁棒的联邦学习框架，假设服务器端拥有干净的根数据集，服务器对该根数据集进行训练得到基准模型，然后在每轮迭代中，服务器会根据各客户端的本地模型更新为其分配信任分数，最后，服务器计算归一化局部模型更新的平均值，并按信任评分进行加权作为全局模型更新. Rieger 等人[26]提出了一种用于缓解后门攻击的模型过滤方法，作者首先提出了一种描述各客户端本地数据分布的方式以及一种基于聚类的本地模型相似度度量方法，基于这两种方法来实现中毒本地模型的检测. Fung 等人[27]提出了防御联邦学习女巫攻击的方案 FoolGold，该方案根据客户端贡献的相似性调整客户端学习率来检测和防御攻击. 文献[28]提出了一种基于截尾均值的聚合算法，该算法为了减轻模型参数中的极端数值的影响，先去除了模型参数中的较大和较小值，再通过求剩余数据的均值完成聚合. 这类基于模型距离及相似性的防御方案在数据非独立同分布场景下的效果有限. 此外，还有学者提出了基于联邦遗忘的防御方案，然而这类方案的有效性需要建立在已知恶意客户端的前提之下，这在实际应用场景中也通常难以实现.

可以看出，现有的针对联邦学习的攻击防御方案大多对联邦学习系统或者恶意参与方进行了一定的约束，例如：恶意参与方比例必须小于 50%，参与方本地数据需满足独立同分布，服务器拥有干净数据集或者攻击者不能在学期初期发动攻击等. 然而这些约束条件在实际应用场景中往往无法满足. 基于这一问题，本文提出了一种基于模型水印的后门攻击防御方案，该方案可以在恶意参与方比例达到 50%，本地数据非独立同分布，攻击者随时可发动攻击的情况下，有效的检测结合白盒与黑盒攻击的复杂后门投毒攻击. 保证全局模型在主任务上的良好性能的同时确保目标任务上的精度极低. 同时，与现有的基于本地模型间的距离以及基于自编码器的防御方案相比，本文所提方法具有更好的防御效果以及更高的恶意模型检测效率.

### 2.3　神经网络模型水印

传统的数字水印技术在多个领域得到了广泛应用. 目前国内外学者也将数字水印的应用扩展到了机器学习领域，通过在神经网络模型中嵌入数字水印达到保护版权、防止篡改、模型溯源等目的. 在神经网络模型训练阶段，模型拥有者可自行设计数字水印并将其嵌入神经网络，模型拥有者可从待检测的神经网络模型中提取和恢复水印信息，将提取的水印与模型拥有者嵌入的水印进行对比，从而判断被检测模型的版权归属和完整性. 根据模型水印的嵌入和提取方法，可以将模型水印分为白盒水印、黑盒水印和无盒水印[29-32]. 在白盒模型水印中，水印的嵌入和提取均需要了解神经网络模型的内部结构和参数信息. 相比于黑盒水印与无盒水印，白盒水印不会影响神经网络模型自身的性能，正常的神经网络模型训练和模型处理不会破坏水印的完整性，然而投毒攻击中恶意参与方针对本地神经网络模型进行的特殊处理会破坏水印的完整性. 因此本文将神经网络模型水印的完整性作为检测后门投毒攻击的指标来发现恶意参与方.

## 3 基于模型水印的后门攻击防御方法

本节将详细介绍本文所提方案的系统结构、敌手假设与系统目标以及所提出的方案.

### 3.1 系统结构

本文所考虑的联邦学习系统由一个服务器 $AS$ 和 $n$ 名参与方组成. 参与方集合记为 $U=\{u_1,u_2,\cdots,u_n\}$,参与方 $u_i(i\in[1,n])$ 的本地数据集记为 $d_i$. 各参与方的本地数据可以是独立同分布或者非独立同分布. 联邦学习系统中存在 $m$ 个恶意参与方,恶意参与方集合记为 $U_q=\{q_1,q_2,\cdots,q_m\}(U_q\subseteq U)$. 服务器与参与方在联邦学习过程中的任务分别为:

**服务器 *AS*.** 负责生成初始全局模型并在其中嵌入模型水印后下发给各参与方,其中水印生成参数为服务器私有数据. 在每一轮学习过程中,服务器对其接收到的参与方本地模型进行攻击检测,基于检测结果抛弃恶意本地模型,按照预定义的聚合规则对正常本地模型进行聚合,得到更新的全局模型后下发给各参与方.

**参与方 $u_i$.** 基于全局模型及其本地数据进行模型训练得到更新的本地模型,然后将更新的本地模型上传至服务器. 若 $u_i$ 为恶意参与方,则其根据特定的攻击方法与攻击策略生成恶意本地模型,若 $u_i$ 为正常参与方,则其诚实地基于正常本地数据与全局模型进行本地模型训练.

### 3.2 敌手假设与设计目标

本文对恶意参与方有如下假设:

(1) 恶意参与方基于投毒数据集以及其拥有的正常数据集进行本地模型训练. 攻击目标是在全局模型中嵌入后门. 恶意本地模型的类型、结构与正常参与方的良性本地模型一致.

(2) 恶意参与方在攻击时可以进行共谋,合作发起投毒攻击. 恶意参与方可以获取其他正常参与方的本地模型参数信息和全局模型参数信息,进而根据这些知识对恶意本地模型参数进行隐蔽性优化. 假设攻击者可以发起以下两种攻击:

**攻击 1.** 限制与提升(constrain and scale)[8]:恶意参与方在本地模型训练过程中嵌入后门的同时限制恶意本地模型参数与良性本地模型参数间的距离. 此外,为了减轻聚合对本地模型的稀释作用,使后门能成功嵌入到全局模型中,在本地模型训练结束后,攻击者将恶意本地模型的参数更新矩阵乘以提升因子 $scale$,从而将恶意本地模型的参数扩大到防御算法允许通过的门限值附近并提高恶意本地模型在联邦聚合中的影响. 在一定的合理范围内,$scale$ 越大,恶意本地模型对全局模型的扰动越大.

服务器端的模型聚合过程可形式化地表达为式(1),其中 $w_G^{(t)}$ 为第 $t(t\in[0,P])$ 轮得到的全局模型,$w_i^{(t)}$ 为参与方 $u_i$ 在第 $t(t\in[1,P])$ 轮得到的本地模型.

$$w_G^{(t)}=\frac{1}{n}\Big(\sum_{u_i\in(U-U_q)}w_i^{(t)}+\sum_{u_i\in U_q}(w_G^{(t-1)}+scale\cdot(w_i^{(t)}-w_G^{(t-1)}))\Big) \quad (1)$$

**攻击 2.** 交替最小化(alternating minimization)[18]:攻击者将其扰动性目标和隐蔽性目标一分为二,并在生成恶意本地模型的过程中交替优化这两个目标. 在每轮本地训练结束后,恶意参与方依然对其本地模型的更新部分进行提升,之后再从提升后的本地模型参数继续进行训练,恶意参与方通过多轮迭代可以更好地分别实现扰动性目标和隐蔽性目标. 基于该策略生成的本地模型,隐蔽性和扰动性进一步提高,可以攻破经典的拜占庭鲁棒聚合算法.

在上述攻击方案中,恶意参与方可能采取三种提升策略,分别为:

(1) 每一轮本地模型训练过程中,提升因子 $scale$ 的取值为一常量;

(2) 每一轮本地模型训练过程中,提升因子 $scale$ 的取值从 1 开始,逐轮递增;

(3) 每一轮本地模型训练过程中,提升因子 $scale$ 的取值从 10 开始,逐轮递减.

基于以上敌手假设,本文所提方案的目标为:通过在服务器端对各参与方的本地模型进行检测,发现敌手发起的后门攻击,避免遭受攻击的本地模型参与全局模型的聚合,提高联邦学习的鲁棒性以及全局模型的精度.

### 3.3 所提方案

图 1 给出了本文所设计的基于模型水印的联邦学习后门攻击防御方法. 在学习开始前,服务器首先生成初始全局模型并利用私有数据生成模型水印嵌入初始全局模型中,将嵌入水印的初始全局模型下发给各参与方. 然后,各参与方进行本地模型训练并将本地模型上传至服务器,服务器基于其嵌入的水印信息对各本地模型进行检测,抛弃其中的恶意本地模型,利用良性本地模型进行聚合得到更新的全

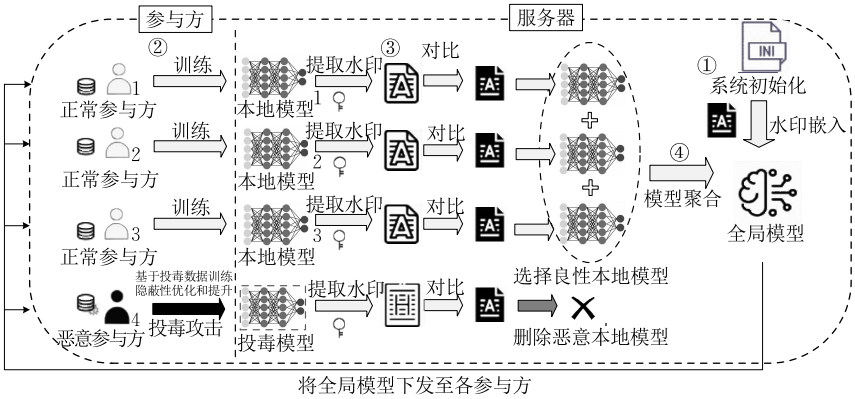



图 1　基于模型水印的联邦学习后门攻击防御方法

局模型. 下文将给出该方案的具体步骤.

(1) 系统初始化

服务器：首先随机生成 $T$ 位二进制串 $b$ 作为签名水印，以及 $T$ 行 $Z_L$ 列(假设全局模型第 $L$ 层参数的数量为 $Z_L$)的密钥矩阵 $\boldsymbol{X}$(满足标准正态分布)，正常参与方和攻击者无法获取 $b$ 和 $\boldsymbol{X}$ 的相关信息. 然后生成随机数据集 $d_{root}$ 进行初始化模型训练以嵌入水印. 这里，$T$ 表示拟嵌入到模型中的信息的大小，又称为水印容量(Capacity of Watermark).

式(2)为服务器进行水印嵌入时训练初始化全局模型的损失函数 $L(w)$. 为了将水印嵌入全局模型的第 $L$ 层，服务器需要修改原始的模型训练损失函数 $L_0(w)$，在其中增加正则项 $L_R(w)$ 以间接修改和限制模型参数 $w$，其中的 $\lambda$ 为可调整的正则超参数.

$$L(w)=L_0(w)+\lambda L_R(w) \tag{2}$$

正则项 $L_R(w)$ 的定义如式(3)所示，其中 $y_j=\sigma(\Sigma_i X_{ji}w_i)$，$\sigma(x)=\dfrac{1}{1+\exp(-x)}$ 为 sigmoid 函数.

$$L_R(w)=-\sum_{j=1}^{T}(b_j\log(y_j)+(1-b_j)\log(1-y_j)) \tag{3}$$

服务器在进行初始模型训练的过程中，每一轮训练结束均提取更新的模型中的水印信息 $b'$，并将其与 $b$ 进行对比，若水印准确率超过 90%，则说明水印已成功嵌入，停止训练. 式(4)给出了水印准确率 $acc$ 的计算方法，其中 $err^i_{init}$ 为 $b'$ 与 $b$ 相同的位的个数.

$$acc=\frac{err^i_{init}}{LEN} \tag{4}$$

最终服务器生成初始化全局模型 $M^0_G$，并分发初始化全局模型参数 $w^0_G$ 给各参与方.

接下来，假设联邦学习系统共进行 $P$ 轮模型训练，每一轮模型训练的过程如下.

(2) 本地模型训练

正常参与方 $u_i$($u_i\in U-U_q$)：利用本地数据 $d_i$ 以及上一轮的全局模型 $M^{t-1}_G$($t\in[1,P]$)进行本地模型训练，得到当前轮的良性本地模型 $M^t_{nor,i}$($t\in[1,P]$)，将 $M^t_{nor,i}$ 的参数存储于矩阵 $\boldsymbol{w}^t_i$.

恶意参与方 $q_i$($q_i\in U_q$)：利用上文介绍的某种攻击方法进行本地模型训练得到当前轮嵌入后门的恶意本地模型 $M^t_{poi,i}$($t\in[1,P]$)，将 $M^t_{poi,i}$ 的参数存储于矩阵 $\boldsymbol{w}^t_i$.

本地模型训练结束后，各参与方将本地模型参数信息 $\boldsymbol{w}^t_i$($i\in[1,n]$)上传至服务器.

(3) 本地模型异常检测

① 针对接收到的各本地模型参数信息 $\boldsymbol{w}^t_i$，服务器首先从其中提取水印，式(5)给出了提取 $\boldsymbol{w}^t_i$ 中签名水印第 $j$ 位的方法. 其中，$X_{j,r}$ 为私有水印密钥 $\boldsymbol{X}$ 的第 $j$ 行、第 $r$ 列元素，$Z_L$ 为本地模型第 $L$ 层参数的数量，$\boldsymbol{w}^t_i[L][r]$ 为本地模型第 $L$ 层中第 $r$ 个参数取值，$s(\cdot)$ 为阶跃函数.

$$b_{ij}=s\Big(\sum_{r=1}^{Z_L}X_{j,r}\cdot\boldsymbol{w}^t_i[L][r]\Big) \tag{5}$$

$$s(x)=\begin{cases}1, & x\geq 0\\ 0, & \text{else}\end{cases} \tag{6}$$

② 得到 $\boldsymbol{w}^t_i$ 中的签名水印后，利用式(4)计算水印准确率 $acc_i$，若 $acc_i<thr$(预定义水印精度门限值)，则 $\boldsymbol{w}^t_i$ 被视为恶意本地模型并被删除.

(4) 模型聚合

服务器将恶意本地模型删除后，基于其余的正常本地模型进行全局模型聚合，然后将聚合后的全局模型更新 $\boldsymbol{w}^t_G$ 下发给各个参与方.

联邦学习系统重复执行上述步骤 2～4，直到训练结束.

# 4 实验结果与分析

为了验证本文所提方案的有效性,实现了部署本文所提方案的联邦学习系统,并通过一系列仿真实验验证了本方案的有效性、安全性和高效性,同时与现有的其他防御方案进行了对比.本节将对此一一详细介绍.

## 4.1 实验环境与设置

表 1 给出了实验环境的软硬件配置.实验采用 CIFAR-10 数据集以及 Tiny-ImageNet 数据集进行模型训练.攻击者采用两种后门触发器实施攻击,分别为:使用数据集中绿色汽车图像与带条纹图案的汽车图像作为后门触发器(触发器 1),以及基于图像扭曲的 WaNet 触发器(触发器 2).实验过程中相关参数的取值如表 2 所示.

**表 1 实验环境**

| 软硬件配置 | 版 本 |
|---|---|
| CPU | Intel i7-10700F |
| GPU | GTX 1660 SUPER |
| 内存 | 16GB |
| 操作系统 | Windows 10 |
| pysyft | 0.2.9 |
| pytorch | 1.4.0 |
| torchvision | 0.5.0 |
| scipy | 1.9.2 |
| numpy | 1.23.3 |

**表 2 实验参数取值**

| 参数名称 | 取 值 |
|---|---|
| 参与方数量 $n$ | 10,20 |
| 恶意参与方数量 $m$ | 单攻击者:1<br>多攻击者:5,10 |
| 预定义水印精度门限值 $thr$ | 0.7,0.75,0.8,0.85 |
| 攻击提升因子 $scale$ | 3,5,7,9,11 |

本文将在不同场景下对所提方案的有效性进行验证,具体场景如表 3 所示.

**表 3 实验场景**

| 场景编号 | 场景配置 |
|---|---|
| 场景 1 | 数据独立同分布,恶意参与方发起攻击 1 |
| 场景 2 | 数据独立同分布,恶意参与方发起攻击 2 |
| 场景 3 | 数据非独立同分布,恶意参与方发起攻击 1 |
| 场景 4 | 数据非独立同分布,恶意参与方发起攻击 2 |

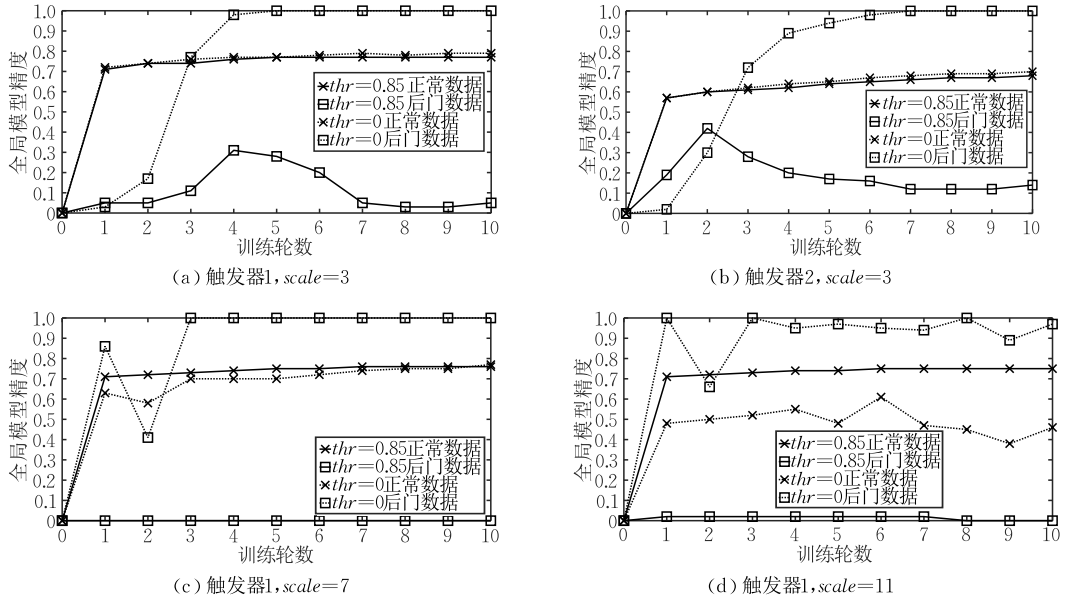
为了验证所提方案的有效性与安全性,本文将在不同场景、恶意参与方采取不同提升策略的情况下,分析部署本文所提方案后全局模型在主任务与目标任务上的精度变化情况.

## 4.2 方案有效性

### 4.2.1 单攻击者

为了验证所提方案的有效性,首先在单攻击者($n=10$,$m=1$)的情况下分析不同场景部署本文所提方案后,主任务与目标任务的预测精度变化.

图 2 给出了场景 1 下、数据集为 CIFAR-10、恶意参与方采取 $scale$ 取常数的提升策略时,部署本文所提方案后得到的主任务与目标任务的精度变化情

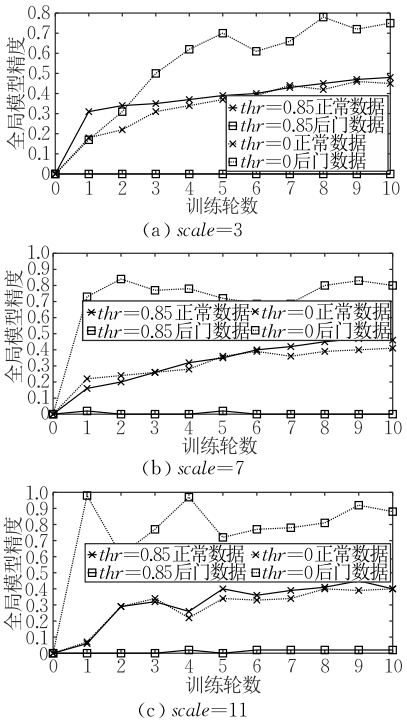



(a) 触发器1,$scale=3$ (b) 触发器2,$scale=3$

(c) 触发器1,$scale=7$ (d) 触发器1,$scale=11$

图 2 场景 1 下所提方案的有效性($n=10$,$m=1$,CIFAR-10)

况. 可以看出，当系统未部署防御机制时（$thr=0$），嵌入后门的恶意本地模型参与了模型聚合，因此，全局模型在目标任务上的预测精度随着训练轮数的增加而不断提高. *scale* 取值越大，全局模型嵌入后门越快，然而过大的 *scale* 取值（$scale=11$）会让全局模型在主任务上的精度受到明显影响. 当部署所提方案后（$thr=0.85$），无论 *scale* 取值高或低，主任务的精度都能随着训练轮数的增加而不断提高并收敛，然而目标任务的精度几乎始终保持为 0. 当数据集为 Tiny-ImageNet，恶意参与方选择触发器 1，其他实验设置与图 2 相同时，部署本文所提方案后得到的主任务与目标任务的精度变化情况如图 3 所示. 可以看出无论 *scale* 取值高或低，主任务的精度都能随着训练轮数的增加而不断提高并收敛，然而目标任务的精度几乎始终保持为 0. 可以看出敌手采用不同的触发器、不同的数据集对于本方案的有效性没有产生明显影响.



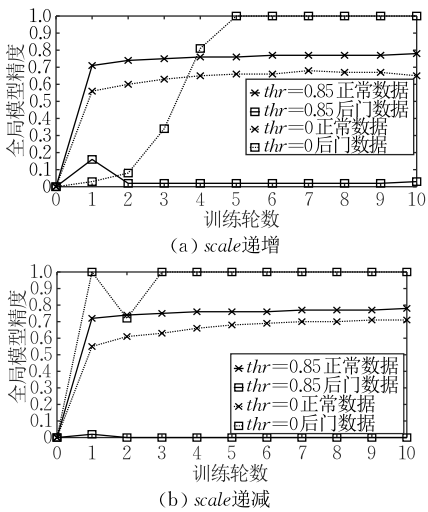
(a) *scale*=3

(b) *scale*=7

(c) *scale*=11

图 3　场景 1 下所提方案的有效性
（$n=10$，$m=1$，Tiny-ImageNet）

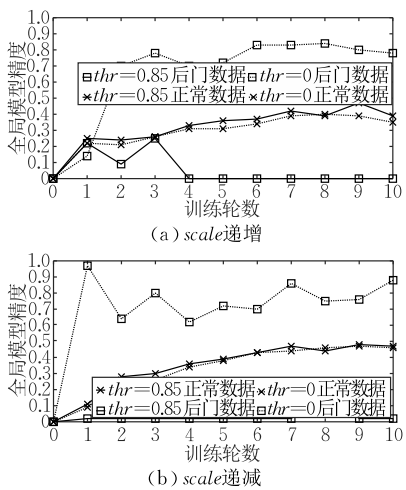
图 4 给出了场景 1 下、数据集为 CIFAR-10、恶意参与方选择触发器 1 并采用 *scale* 递增/递减的提升策略时，部署本文所提方案后得到的主任务与目标任务的精度变化情况. 可以看出，在没有任何防御机制（$thr=0$）时，在主任务的精度随着训练轮数不断提升的同时，目标任务的精度也同时提升. 当恶意参与方采取 *scale* 递减的提升策略时，由于在训练初期恶意参与方的 *scale* 取值较大，使得恶意参与方可以快速在全局模型中嵌入后门，在第一轮训练结束后全局模型的目标任务精度就达到 100%. 然而，在部署本文所提方案后（$thr=0.85$），全局模型在主任务中的精度随着训练轮数不断提升的同时，目标任务始终保持着很低的精度. 由此可以看出本方案可以在场景 1 下防御恶意参与方采取 *scale* 递增或递减的提升策略下的后门投毒攻击. 图 5 给出了数据集为 Tiny-ImageNet，其余条件与图 4 相同时，部署本文所提方案后得到的主任务与目标任务的精度变化情况. 可以看出，实验结果与图 4 相同.



(a) *scale*递增

(b) *scale*递减

图 4　场景 1 下所提方案的有效性
（$n=10$，$m=1$，*scale* 递增/递减，CIFAR-10）



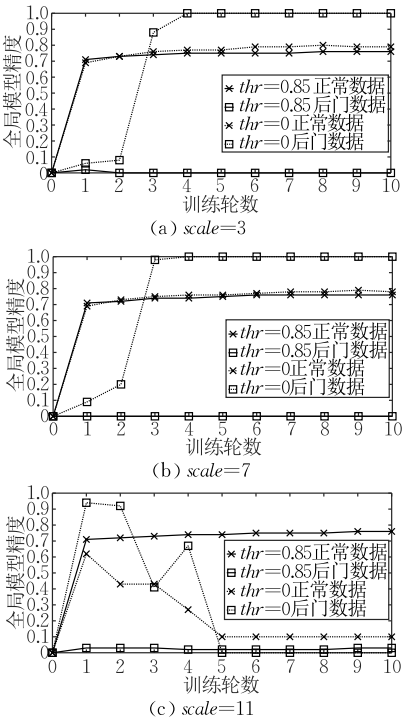
(a) *scale*递增

(b) *scale*递减

图 5　场景 1 下所提方案的有效性
（$n=10$，$m=1$，*scale* 递增/递减，Tiny-ImageNet）

图 6 给出了场景 2 下、数据集为 CIFAR-10、恶意参与方选择触发器 1 并采取 *scale* 取常数的提升策略时，部署本文所提方案后得到的主任务与目标任务的精度变化情况. 可以看出，在 *scale* 取不同值的

情况下,部署了本文所提方案的联邦学习系统($thr=0.85$)中,主任务的精度随着训练轮数的增加不断增加,同时目标任务的精度始终几乎为 0.在没有任何防御机制($thr=0$)时,主任务与目标任务的精度均随着训练轮数不断提升.当数据集为 Tiny-ImageNet,其他实验设置与图 6 实验设置相同时,本方案的有效性与图 6 所示结果一致,由于篇幅有限,不在此进行赘述.由此可见,在场景 2 下,本文所提方案可以有效的防止恶意参与方采取 *scale* 取常量的提升策略下的后门投毒攻击.



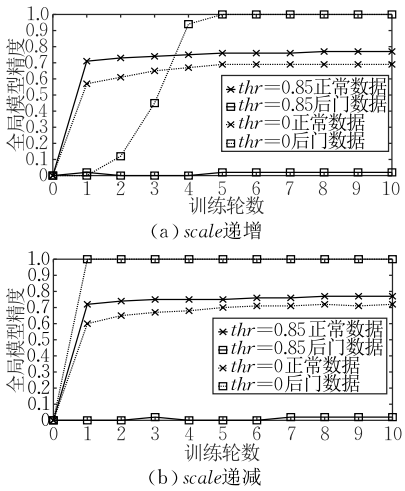
(a) *scale*=3

(b) *scale*=7

(c) *scale*=11

图 6 场景 2 下所提方案的有效性
($n=10$,$m=1$,CIFAR-10)

图 7 给出了场景 2 下、数据集为 CIFAR-10、恶意参与方选择触发器 1 并采取 *scale* 取递增或递减的提升策略时,部署本文所提方案后得到的主任务与目标任务的精度变化情况.可以看出,在没有任何防御机制($thr=0$)时,在主任务的精度随着训练轮数不断提升的同时,目标任务的精度也同时提升,恶意参与方可以迅速在全局模型中嵌入后门.当部署了本文所提方案($thr=0.85$)后,随着训练轮数的增加,全局模型的主任务预测精度不断增加,同时目标任务的精度始终几乎为 0.当数据集为 Tiny-ImageNet,其他实验设置不变时,在联邦学习系统部署本方案后主任务与目标任务的精度变化情况与图 7 一致,由于篇幅有限,不在此进行赘述.因此可看出本方案可以防御递增与递减 *scale* 提升策略下的后门投毒攻击,保证全局模型精度不受恶意参与方的影响.



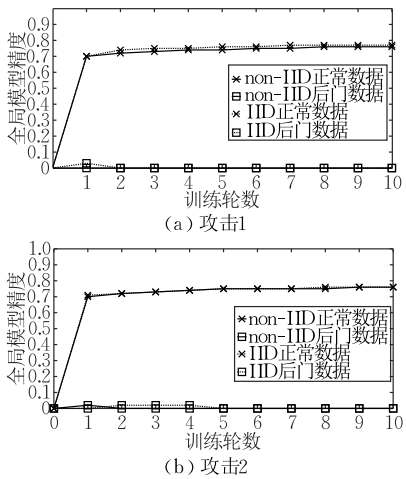
(a) *scale*递增

(b) *scale*递减

图 7 场景 2 下所提方案的有效性
($n=10$,$m=1$,*scale* 递增/递减,CIFAR10)

从图 2～图 7 可以看出,本文所提方案可以在参与方数据独立同分布的情况下,有效地防御不同数据集下恶意参与方采用不同攻击方法、不同提升策略发起的后门投毒攻击.

图 8 给出了在场景 3 与场景 4 下、数据集为 CIFAR-10、恶意参与方选择触发器 1 并采取 *scale* 取常数的提升策略时,部署本文所提方案后得到的主任务与目标任务的精度变化情况.可以看出,部署本方案后,全局模型的主任务精度随着训练轮数的增加不断提升并收敛,目标任务的精度则始终维持在极低水平.此外,主任务与目标任务的精度变化与参与方数据独立同分布下得到的主任务的精度变化趋势基本一致.当数据集为 Tiny-ImageNet,其他实验设置不变时,本方案的后门攻击防御效果与图 8 类似,在



(a) 攻击1

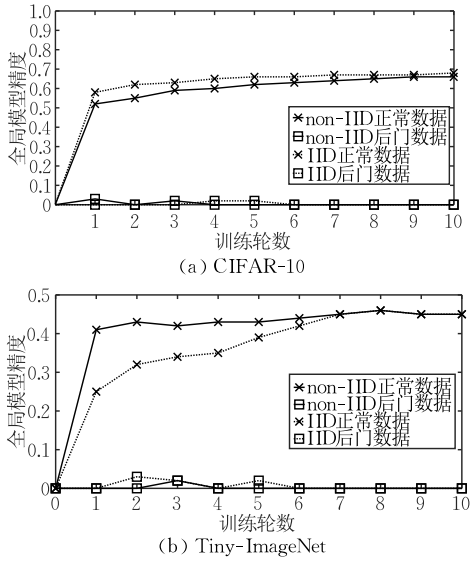
(b) 攻击2

图 8 场景 3 与场景 4 下所提方案的有效性
($n=10$,$m=1$,$scale=10$,CIFAR-10)

此不再赘述. 可以看出，本方案的防御效果不会受到参与方本地数据分布情况与数据集的影响，可以在参与方数据非独立同分布的情况下，有效的防御恶意参与方采用不同攻击方法、不同提升策略发起的后门投毒攻击.

4.2.2　多攻击者

上一节的实验说明了在单攻击者的情况下，本文所提方案能够在各种场景下有效防御恶意参与方发起的不同攻击. 本节给出在多攻击者的情况下本文所提方案的防御效果. 图 9 给出了当恶意攻击者数量为 5(占参与方总量 50%)的情况下部署本文所提方案后得到的主任务与目标任务的精度变化情况. 可以看出，在多攻击者环境中，无论参与方的数据独立同分布或非独立同分布，无论数据集为 CIFAR-10 或 Tiny-ImageNet，当恶意参与方采取 *scale* 取常数的提升策略时，部署本方案的联邦学习系统得到的全局模型在主任务上的精度依然可以随着训练轮数的增加而逐渐提高并收敛，目标任务上的精度始终维持极低水平. 因此可以看出本方案在多攻击者情况下同样可以有效防御恶意参与方发起的不同后门投毒攻击，且防御效果不受参与方数据分布的影响.



(a) CIFAR-10

(b) Tiny-ImageNet

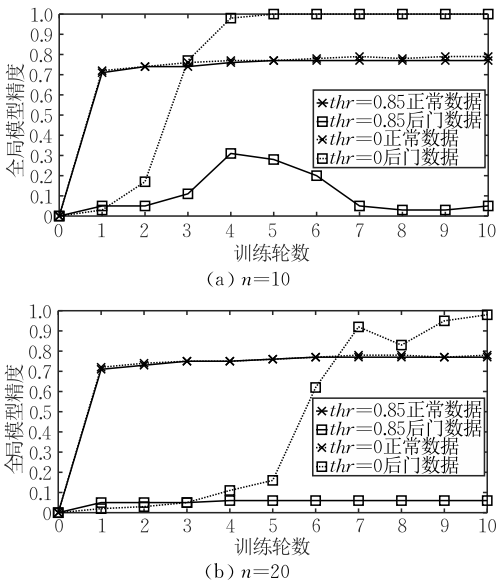
图 9　场景 2 与场景 4 下所提方案的有效性
($n$=10, $m$=5, $scale$=10，触发器 1)

在各参与方数据非独立同分布的情况下，各本地模型间的统计学相似性会遭受破坏，基于统计学的异常检测方案将不再有效. 然而从上述实验结果可以看出本方案的有效性不受各参与方数据分布情况的影响，原因在于无论各参与方的本地数据呈何种分布模式，正常的模型训练不会对模型水印的完整性产生影响，而后门投毒攻击则会破坏模式水印，那么可通过检测本地模型中的水印是否遭受破坏来判定各本地模型是否为异常模型.

### 4.3　方案可靠性

本节将分析参与方总数、恶意参与方数量、水印容量以及水印精度门限值 $thr$ 对所提方案的有效性影响，从而验证本方案的可靠性. 本节实验所采用的数据集均为CIFAR-10，敌手均采用触发器 1 发动攻击.

当恶意参与方数量为 1 时，参与方总数对所提方案的有效性影响如图 10 所示. 可以看出，当联邦学习系统中的参与方数量不同时，本方案均可有效防御恶意参与方发起的后门投毒，确保全局模型在主任务上的精度快速提高并收敛的同时目标任务的精度始终保持极低值. 由此看出本方案的有效性不受参与方总数的影响.

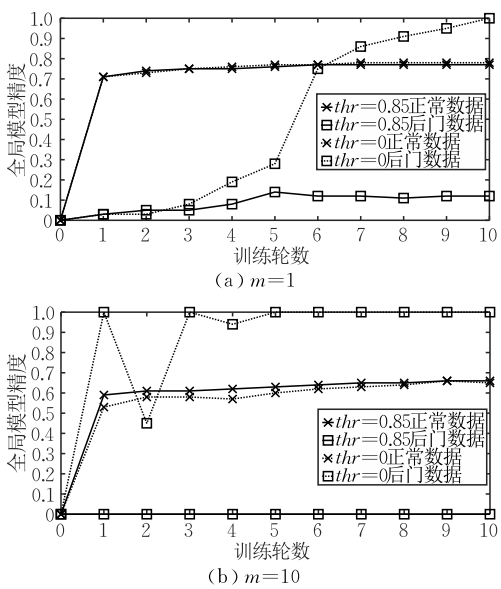


(a) $n$=10

(b) $n$=20

图 10　参与方数量对所提方案的性能影响
(场景 1, $m$=1, $scale$=3)

图 11 展示了当参与方总数为 20 时，恶意参与方数量对所提方案的有效性影响. 可以看出，在没有任何防御机制($thr$=0)时，无论恶意参与方数量为 1 或 10，全局模型在主任务上的精度随着训练轮数不断提升的同时在目标任务的精度也同时提升，恶意参与方可以迅速在全局模型中嵌入后门，恶意参与方数量越多，嵌入后门的速度也越快. 然而，在恶意参与方数量取值为 1 或 10 时，本文所提方案($thr$=0.85)均可有效防御恶意参与方发起的后门投毒，确保全局模型在主任务上的精度快速提高并收敛的同时目标任务的精度始终保持极低值. 由此看出本方案的有效性不受参与方总数的影响.

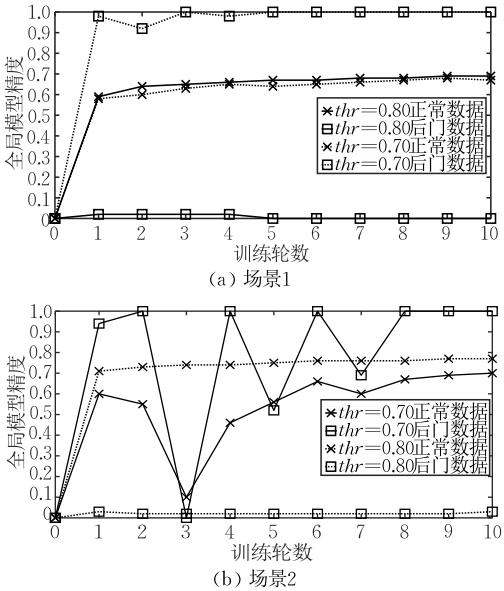



(a) $m=1$

(b) $m=10$

图 11 恶意参与方数量对所提方案的性能影响
($场景\ 2, n=20, scale=3$)

图 12 展示了在水印精度门限值 $thr$ 取值不同时单攻击者场景下所提方案的有效性. 当 $thr=0.8$ 时,随着训练轮数的增加,全局模型在主任务上的精度不断提升并收敛,其在目标任务上的精度一直几乎为 0;当 $thr=0.70$ 时,训练过程中全局模型在主任务上的精度出现波动,目标任务的精度也随着训练过程快速提升,这表明恶意本地模型参与了模型聚合过程并在全局模型中成功嵌入了后门.

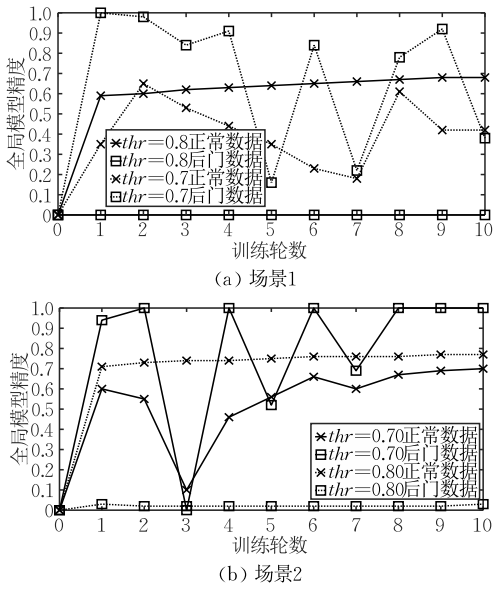



(a) 场景1

(b) 场景2

图 12 水印精度门限值 $thr$ 对所提方案的性能影响
($n=10, m=1, scale=10$)

图 13 给出了多攻击者场景下水印精度门限值 $thr$ 对所提方案的有效性产生的影响. 可以看出,与图 12 类似,当 $thr=0.8$ 时,本文方案可以有效防御恶意参与方的攻击;当 $thr=0.70$ 时,训练过程中全局模型在主任务上的精度出现波动,目标任务的精度也随着训练过程快速提升,这表明恶意本地模型参与了模型聚合过程并在全局模型中成功嵌入了后门. 与图 12 相比,恶意参与方数量的增加并未造成有效防御门限的降低,最低的有效门限应仍然在 0.8 左右.




(a) 场景1




(b) 场景2

图 13 水印精度门限值 $thr$ 对所提方案的有效性影响
($n=20, m=5, scale=10$)

可以看出 $thr$ 的取值将影响本文所提方案的有效性,无论攻击者采用何种攻击,为了保证恶意参与方发起的后门投毒攻击,最低的有效门限应设置在 0.8 左右.

图 14 给出了水印容量 $T$ 对所提方案的有效性影响. 可以看出在不同的水印容量下,全局模型在主任务上的精度不断提升并收敛,其在目标任务上的精度一直几乎为 0,由此可见水印容量 $T$ 对本方案的有效性不会产生影响. 表 4 给出了水印容量 $T$ 对所提方案的性能影响. 可以看出随着水印容量的不断增加,水印嵌入时间与水印提取时间基本处于同一数量级,未产生较大差异.

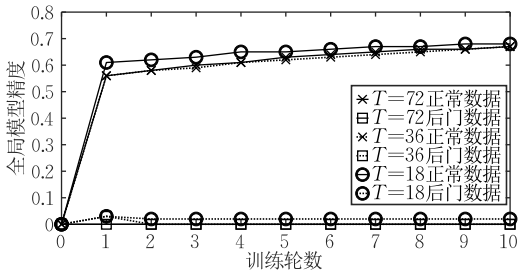



图 14 水印容量对所提方案的有效性影响
(场景 1, $n=10$, $m=1$, $scale=3$, $thr=0.85$)

**表 4　水印容量对所提方案的性能影响**

| 水印容量 | 水印嵌入时间/s | 水印提取时间/s |
|---|---|---|
| $T=18$ | 17.9200 | 0.0018 |
| $T=36$ | 13.1887 | 0.0018 |
| $T=72$ | 13.0088 | 0.0065 |

### 4.4　与现有方案的对比

为了验证本文所提方案的优势，本节将对本文所提方案与现有的其他参与方投毒攻击防御方案进行对比. 基于模型距离的防御方案[24]与基于自编码器的防御方案[33]是当前两种经典的防御思路，本节将对比本文所提方案与这两种方案的防御效果以及开销，图 15 给出了对比结果. 当 $scale=3$ 时，基于本地模型参数间距离的防御方案所得到的全局模型在主任务上的精度与部署本文所提方案时的精度接近，然而目标任务的精度随着训练也迅速提高至接近 1. 当 $scale$ 取值增加至 5 以上时，基于本地模型参数间距离的防御方案造成全局模型在主任务与目标任务上的精度均降至极低水平. 这是因为大量恶意参与方的存在导致基于本地模型参数间距离的防御门限向恶意本地模型参数的方向偏移，进而出现误检的情况，使得全局模型被嵌入后门或直接被破坏. 此外，从图中可以看出当 $scale=3$ 和 $scale=5$ 时，基于自编码器的防御方案所得到的全局模型在主任务和目标任务上的精度均出现了较大的波动，无法达到较高的水平. 当 $scale=8$ 时，由于恶意本地模型提升的幅度较大，进而与良性本地模型的差异也进一步提高，所以基于自编码器的防御方案所得到的全局模型在主任务上的精度提升至正常水平，而目标任务的精度维持在 0 附近. 而本文所提方案在恶意参与方对 $scale$ 取值不同的情况下均可以成功防御恶意参与方发起的后门投毒攻击.

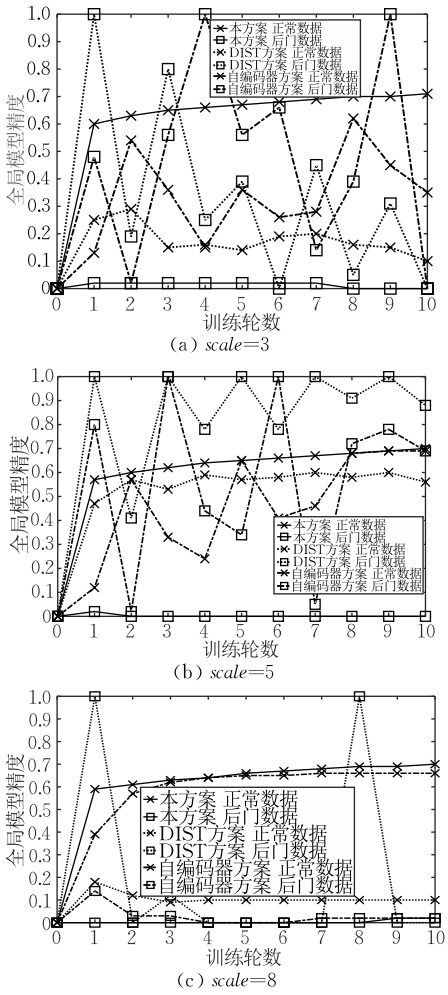



图 15　所提方案与现有方案的对比
（场景 2，$n=10$，$m=5$）

本文方案和基于自编码器的防御方案的恶意参与方检测的时间复杂度为 $O(n)$（$n$ 为参与方数量），然而自编码器的计算耗时要远大于水印提取的时间. 基于模型间距离的防御方案的时间复杂度则为 $O(n^2 \cdot num)$（$num$ 为模型的参数数量）. 图 16 给出了不同参与方数量的情况下，本方案与现有其他防御方案在服务器端产生的额外计算开销的对比结果. 可以看出本文所提方案的计算开销远小于其他方案.

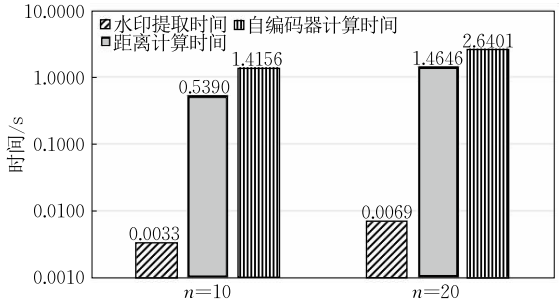



图 16　所提方案与现有方案的计算开销对比

### 4.5　分析与讨论

以上实验表明，本方案能够在数据独立同分布与非独立同分布场景下实现对多种后门投毒攻击的防御，防御效果与恶意参与方比例无关，此外，与现有其他防御方案相比计算效率有了大幅提高.

本文所提方案的有效性是建立在模型水印不会被正常训练过程所破坏的基础上的. 以上实验表明，在不同的攻击方式（静态攻击策略或动态攻击策略）下，本文所提方案均可以成功检测攻击从而避免后门被嵌入全局模型中. 但是若全局模型的部分层次未被服务器嵌入模型水印，攻击者对这些层的攻击则无法被服务器检测到，从而能够安插后门且不被发现. 因此，若攻击者已知防御算法的同时也知道全

局模型的哪些层未被嵌入模型水印，则可以在未嵌入模型水印的层上对模型进行后门投毒.因此，为了抵抗攻击者已知防御算法而进行投毒，可通过在全局模型的所有层上嵌入模型水印来实现防御目标.

文献[29]指出，假设拟在模型的第 $K$ 层(该层参数数量为 $Z_K$)嵌入 $T$ 比特的水印信息，那么当 $T>Z_K$ 时，水印的嵌入效果将会产生明显的下降(误差与模型嵌入损失均会增大)，原因在于在这种情况下会产生超定(overdetermined)问题，因此论文所提方案中，$T$ 的取值需要满足这一限制.若要嵌入更多的信息同时避免产生超定问题，可以通过多层感知机来解决.详细信息可参阅文献[29].

目前国内外学者已经针对模型水印本身提出了一系列的攻击方法，包括删除、屏蔽、篡改水印等[34].这些攻击方案均会对原有模型水印产生破坏，因此，本方案中引入模型水印也会增大敌手的攻击面.敌手可通过破坏正常本地模型中的模型水印使得该本地模型被检测为异常模型.当敌手只能控制少数节点来发动以上针对水印的攻击时，少数本地模型的缺失不会对整个模型的精度带来很大的影响.当敌手能够控制大量节点时，会造成能够通过异常检测的本地模型数量大大减少，从而会使得全局模型的收敛速度变慢，然而在这种情境下，敌手也可以采用本文所述的攻击方式进行投毒，攻击结果与针对水印的攻击所产生的结果一致.因此，可以看出，模型水印的安全性问题会令敌手有更多的攻击手段实现破坏模型完整性的目标，但是不会对服务器的检测结果以及全局模型的收敛速度与精度产生更多的影响.

## 5 结束语

联邦学习是一种新型的分布式机器学习方法，多个数据拥有方可以在不泄露本地数据的情况下协作进行机器学习模型训练.然而，分布式特性令联邦学习容易遭受来自参与方的投毒攻击，其中的后门投毒攻击具有很高的隐蔽性，对其进行高效检测是联邦学习面临的一大挑战.现有的投毒攻击防御方案大多对联邦学习系统以及恶意参与方有较多的约束条件，在许多实际场景中这些约束条件无法得到满足.因此，亟需设计一种后门投毒防御方法，能够在联邦学习系统以及恶意参与方的约束条件无法满足时有效地实现后门投毒攻击防御.

本文设计了一种基于模型水印的联邦学习后门攻击防御方法.在该方法中，聚合服务器首先在初始全局模型中嵌入私有的白盒水印，该水印在正常的机器学习模型训练过程中不会遭受破坏，因此，服务器可以通过验证其每一轮训练过程中接收到的本地模型中的水印完整性来实现对恶意参与方的检测.通过大量的仿真实验，本文验证了所提方案的有效性、可靠性以及高效性.实验结果表明，该方案在参与方数据独立与非独立同分布的情况下均可有效防御后门投毒攻击，其防御效果不受恶意参与方比例的影响.此外，该方案在防御效率方面优于现有的投毒攻击防御方法.

**致 谢** 衷心感谢编辑以及各位审稿人为本论文所花费的时间与精力！

## 参 考 文 献

[1] European Union. General Data Protection Regulation(GDPR). https://gdpr-info.eu, 2018

[2] GB/T 35273-2020. Information Security Technology-Personal Information Security Specification. National Information Security Standardization Technical Committee of China. 2020 (in Chinese)
(GB/T 35273-2020. 信息安全技术——个人信息安全规范. 全国准化技术委员会. 2020)

[3] McMahan B, Moore E, Ramage D, et al. Communication efficient learning of deep networks from decentralized data//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR. 2017, 54: 1273-1282

[4] Jere M S, Farnan T, Koushanfar F. A taxonomy of attacks on federated learning. IEEE Security & Privacy, 2020, 19(2): 20-28

[5] Bouacida N, Mohapatra P. Vulnerabilities in federated learning. IEEE Access, 2021, 9: 63229-63249

[6] Kairouz P, Mcmahan H B, Avent B, et al. Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 2021, 14(1-2): 1-210

[7] Xie C, Huang K, Chen P Y, et al. DBA: Distributed backdoor attacks against federated learning//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[8] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning//Proceedings of the International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research 108. Palermo, Italy, 2020: 2938-2948

[9] Xiao Xiong, Tang Zhuo, Xiao Bin, et al. A survey on privacy and security issues in federated learning. Chinese Journal of Computers, 2023, 46(5): 1019-1044(in Chinese)

(肖雄，唐卓，肖斌等. 联邦学习的隐私保护与安全防御研究综述. 计算机学报，2023，46(5)：1019-1044)

[10] Gu Yu-Hao，Bai Yue-Bin. Survey on security and privacy of federated learning models. Journal of Software，2023，34(6)：2833-2864(in Chinese)
(顾育豪，白跃彬. 联邦学习模型安全与隐私研究进展. 软件学报，2023，34(6)：2833-2864)

[11] Wan W，Hu S，Lu J，et al. Shielding federated learning：Robust aggregation with adaptive client selection//Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna，Austria，2022：753-760

[12] Jamie H，Olga O. Contamination attacks and mitigation in multi-party machine learning//Proceedings of the Advances in Neural Information Processing Systems 31：Annual Conference on Neural Information Processing Systems 2018. Montréal，Canada，2018：6604-6615

[13] Cao D，Chang S，Lin Z，et al. Understanding distributed poisoning attack in federated learning//Proceedings of the IEEE 25th International Conference on Parallel and Distributed Systems. Tianjin，China，2019：233-239

[14] Xia G，Chen J，Yu C，et al. Poisoning attacks in federated learning：A survey. IEEE Access，2023，11：10708-10722

[15] Tolpegin V，Truex S，Gursoy M E，et al. Data poisoning attacks against federated learning systems//Proceedings of the 25th European Symposium on Research in Computer Security. Guildford，UK，2020，Part I：480-501

[16] Xie C，Koyejo O，Gupta I. Fall of empires：Breaking Byzantine-Tolerant SGD by inner product manipulation//Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence. Tel Aviv，Israel，2019：261-270

[17] Cao X，Gong N Z. MPAF：Model poisoning attacks to federated learning based on fake clients//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans，USA，2022：3395-3403

[18] Bhagoji A N，Chakraborty S，Mittal P，et al. Analyzing federated learning through an adversarial lens//Proceedings of the 36th International Conference on Machine Learning. Long Beach，USA，2019：634-643

[19] Fang M，Cao X，Jia J，et al. Local model poisoning attacks to Byzantine-Robust federated learning//Proceedings of the 29th USENIX Conference on Security Symposium. 2020：1623-164

[20] Shejwalkar V，Houmansadr A. Manipulating the Byzantine：Optimizing model poisoning attacks and defenses for federated learning//Proceedings of the 28th Annual Network and Distributed System Security Symposium. 2021

[21] Gong X，Chen Y，Huang H，et al. Coordinated backdoor attacks against federated learning with model-dependent triggers. IEEE Network，2022，36(1)：84-90

[22] Hou B，Gao J，Guo X，et al. Mitigating the backdoor attack by federated filters for industrial IoT applications. IEEE Transactions on Industrial Informatics，2021，18(5)：3562-3571

[23] Sun Z，Kairouz P，Suresh A T，et al. Can you really backdoor federated learning？ https://arxiv. org/abs/1911. 07963v2，2019

[24] Li S，Cheng Y，Wang W，et al. Learning to detect malicious clients for robust federated learning. https://arxiv. org/abs/2002. 00211，2020，2，1

[25] Cao Xiao-Yu，Fang Ming-Hong，Liu Jia，Gong Neil Zhenqiang. FLTrust：Byzantine-robust federated learning via trust bootstrapping. Network and Distributed Systems Security (NDSS) Symposium，2021

[26] Rieger P，Nguyen T D，Miettinen M，Sadeghi A-R. DeepSight：Mitigating backdoor attacks in federated learning through deep model inspection. Network and Distributed Systems Security (NDSS) Symposium. San Diego，USA，2022

[27] Fung C，Yoon C J M，Beschastnikh I. The limitations of federated learning in sybil settings//Proceedings of the 23rd International Symposium on Research in Attacks，Intrusions and Defenses(RAID 2020). 2020

[28] Yin D，Chen Y，Kannan R，et al. Byzantine-robust distributed learning：Towards optimal statistical rates//Proceedings of the 35th International Conference on Machine Learning. Stockholm，Sweden，2018：5636-5645

[29] Yusuke U，Yuki N，Shigeyuki S，et al. Embedding watermarks into deep neural networks//Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. Bucharest，Romania，2017：269-277

[30] Bita D R，Huili C，Farinaz K. DeepSigns：An end-to-end watermarking framework for ownership protection of deep neural networks//Proceedings of the 24th International Conference on Architectural Support for Programming Languages and Operating Systems. Providence，USA，2019：485-497

[31] Yossi A，Carsten B，Moustapha C，et al. Turning your weakness into a strength：Watermarking deep neural networks by backdooring//Proceedings of the 27th USENIX Security Symposium. Baltimore，USA，2018：1615-1631

[32] Wu H，Liu G，Yao Y，et al. Watermarking neural networks with watermarked images. IEEE Transactions on Circuits and Systems for Video Technology，2021，31(7)：2591-2601

[33] Blanchard P，El Mhamdi E M，Guerraoui R，et al. Machine learning with adversaries：Byzantine tolerant gradient descent//Advances in Neural Information Processing Systems 30：Annual Conference on Neural Information Processing Systems 2017. Long Beach，USA，2017：119-129

[34] Wu Han-Zhou，Zhang Jie，Li Yue，et al. Overview of artificial intelligence model watermarking. Journal of Image and Graphics，2023，28(6)：1792-1810(in Chinese)
(吴汉舟，张杰，李越等. 人工智能模型水印研究进展. 中国图象图形学报，2023，28(6)：1792-1810)

**GUO Jing-Jing**, Ph. D., associate professor. Her research interests include wireless network security, federated learning security and trust management.

**LIU Jiu-Zun**, M. S. His research interests focus on federated learning security.

**MA Yong**, Ph. D., professor. His research interests include cloud computing, edge computing and information security.

**LIU Zhi-Quan**, Ph. D., associate researcher. His research interests include trust management, VANET security, and wireless network security.

**XIONG Yu-Peng**, M. S. candidate. His research interests focus on federated learning security.

**MIAO Ke**, undergraduated student. His research interests focus on federated learning security.

**LI Jia-Xing**, M. S. candidate. Her research interests focus on federated learning security.

**MA Jian-Feng**, Ph. D., professor. His research interests include cryptography and network security.

**Background**

As a privacy-preserving distributed machine learning paradigm, federated learning is vulnerable to poison attacks. The high crypticity of backdoor poisoning makes it difficult to defend against. Most existing defense schemes against backdoor poisoning attacks have strict constraints on the servers or malicious participants (servers need to have a clean root dataset, the proportion of malicious participants is less than 50%, and poisoning attacks cannot be initiated at the beginning of learning, etc.). When these constraints cannot be met, the effectiveness of these schemes will be greatly compromised. To solve this problem, this paper proposes a Byzantine robust Federated learning method based on model watermarking. In this scheme, the server detects malicious participants based on whether the watermark embedded in the initial global model by the server has been corrupted in local models. In order to verify the effectiveness of this scheme, a series of simulation experiments were conducted. Experimental results show that this scheme can effectively detect backdoor poisoning attacks launched by malicious participants in various scenarios where the proportion of malicious participants is unlimited, the distribution of participants' data is unlimited, and the attack time of participants is unlimited. Moreover, the detection efficiency of malicious participants in this scheme is significantly improved compared to existing poison attack defense methods.

This project is supported by the National Natural Science Foundation of China (Nos. 62272195, 61932010, 62032025), the Natural Science Basic Research Program of Shaanxi (No. 2022JQ-603), the Fundamental Research Funds for the Central Universities (Nos. ZYTS23161, 21622402), the Key Laboratory Project of Network and Information Security Vulnerability Research of Guangdong Province (No. 2020B1212060081), and the Science and Technology Plan Project of Guangzhou (No. 202201010421). The main research goal of these projects is to build a trusted and robust federated learning system. The backdoor poisoning attack defense scheme proposed in this paper can detect backdoor attacks initiated by malicious participants. The detection results can provide trust evidence for the trust evaluation of the participants in the federated learning system, and help the server to make better decisions about the trust degree of the participants, so as to build a trusted federated learning system.

Our research team has published several journal papers and has patented many inventions in security of federated learning. Several papers are listed as following.

[1] Guo Jingjing, Li Haiyang, Huang Feiran, et al. ADFL: A poisoning attack defense framework for horizontal federated learning. IEEE Transactions on Industrial Informatics, 2022, 18(10): 6526-6536

[2] Li Haiyang, Guo Jingjing, Liu Jiuzun, Liu Zhiquan. Privacy preserving byzantine robust federated learning algorithm. Journal of Xidian University, 2023, 50(4): 121-131.

[3] Guo Jingjing, Liu Zhiquan, Tian Siyi, et al. TFL-DT: A trust evaluation scheme for federated learning in digital Twin for mobile networks. IEEE Journal on Selected Areas in Communication, 2023, 41(11): 3548-3560.