# 恶意参与者多数情景下的聚合模型保护算法

张恩[a,b],高婷[a],黄昱晨[a]

(河南师范大学 a.计算机与信息工程学院;b.智慧商务与物联网技术河南省工程实验室,河南 新乡 453007)

**摘 要**:隐私保护联邦学习能够帮助多个参与者构建机器学习模型.然而,该方法很难防御恶意参与者占多数时的投毒攻击.此外,用户或服务器可能会私自出售聚合模型.针对以上问题,提出了一种抗大多数恶意参与者的安全聚合方案,同时保护聚合结果不泄露.在训练阶段,参与者使用差分隐私噪声和随机数保护局部模型;然后参与者对其余的差分隐私模型进行准确率测试,并将结果记录在一个向量中;最后参与者与服务器执行不经意传输协议,得到聚合模型.通过安全分析证明了安全性和正确性.实验结果表明算法在恶意参与者占多数时仍能保持良好的检测能力,并在一定程度上保证了参与者的公平性.

**关键词**:联邦学习;隐私保护;不经意传输;同态哈希

**中图分类号**:TP309.2 **文献标志码**:A **文章编号**:1000-2367(2025)04-0058-08

联邦学习(federated learning,FL)[1]作为一个分布式学习框架允许客户在保护数据隐私的情况下合作训练聚合模型,虽然避免了直接共享用户本地隐私数据,但仍面临诸多挑战.即使参与者只上传模型参数,敌手仍然能够从中推理出隐私数据[2].此外,恶意参与方通过操纵本地数据集或局部模型更新向聚合器上传虚假的模型参数,这种行为可能导致聚合模型的错误预测和不准确性[3-4].基于统计学方法或基于距离的拜占庭鲁棒性判别方案无法抗恶意参与者大多数的情况,服务器端利用干净小型验证数据集判别方案违反了联邦学习的隐私保护原则(即剩余的数据本地化原则).另一方面,聚合模型也具有数据价值,代表着一种重要的知识产权[5],一些工作通过验证模型水印以判断它们是否被未被授权的第三方窃取[6].安全聚合与模型产权保护研究已成为联邦学习中研究者关注的热点.

文献[7-9]均假定服务器维护一个公共干净的验证数据集,服务器使用此数据集评估局部模型更新的准确性或相似度,将性能较差的模型判定为有毒模型.GUERRAOUI 等[10]提出 Bulyan,通过将 Krum 和修剪平均值相结合,确保聚合梯度的每个维度上的多数一致.SHAYAN 等[11]提出了 Biscotti,该方案在区块链上应用 Krum 算法检测局部模型并结合秘密共享聚合全局模型.TAO 等[12]提出了一种拜占庭弹性分布式梯度下降算法,该算法可以处理重尾数据并在标准假设下收敛.LI 等[13]使用核密度估计方法测量相邻局部模型之间的相对分布以区分恶意和干净的更新.但是当拜占庭客户占多数时,这些方案无法保证模型的鲁棒性.ZHOU 等[14]结合范数检测与准确率检测生成了混合检测策略,通过调整范数检测和准确率检测的比重以适应不同比例恶意参与者的情景,但是在进行联邦学习时,通常很难确定恶意参与者的数量.MA 等[15]结合 Pailliar 同态加密和零知识证明,以保证局部模型隐私并过滤出恶意参与者的异常模型,但是对于参数通常高达数百万的机器学习模型来说,同态加密的开销较大并难以有效实现.LIM 等[16]提出了两种不同的递归

**收稿日期**:2024-04-12;**修回日期**:2024-05-07.

**基金项目**:国家自然科学基金(62072159;62002103;6207608);河南省科技攻关项目(232102211057).

**作者简介(通信作者)**:张恩(1974-),男,河南新乡人,河南师范大学教授,博士,CCF 高级会员(47887S),研究方向为隐私保护机器学习、安全多方计算,E-mail: zhangenzdrj@163.com.

**引用本文**:张恩,高婷,黄昱晨.恶意参与者多数情景下的聚合模型保护算法[J].河南师范大学学报(自然科学版),2025,53(4):58-65.(Zhang En,Gao Ting,Huang Yuchen.The aggregation model protection algorithm in scenarios with majority of malicious participant[J].Journal of Henan Normal University(Natural Science Edition),2025,53(4):58-65.DOI:10.16366/j.cnki.1000-2367.2024.04.12.0001.)

神经网络下的水印嵌入方案,以保护图像字幕模型.李璇等[17]利用深度学习后门技术在不影响主任务准确率的情况下仅对少量触发集样本造成误分类实现模型的产权保护.但这些方法是在模型版权被盗取出现争议之后,利用水印为版权归属提供有力的证据.XU 等[18]引入两个非共谋服务器并通过高度集成加性同态加密和混淆电路从而保护了所有用户相关信息的隐私性,但在现实情况下很难保证两个服务器不会合谋.

针对以上问题,本文提出了适用于恶意参与者多数情景下的聚合模型保护算法,实验结果表明本方案在恶意参与者占大多数的情况下仍能够发挥良好的检测作用,并且聚合模型精度与数据集质量成正比,从而保证贡献度越高的参与者得到的聚合模型性能也越好.

# 1　基础知识

## 1.1　联邦学习

根据各参与方数据分布的情况不同,联邦学习被分为横向联邦学习、纵向联邦学习和联邦迁移学习[19].横向联邦学习的本质是样本的联合,适用于参与者间业务相同但接触客户不同,即特征重叠多,用户重叠少的场景.纵向联邦学习的本质是特征的联合,适用于用户重叠多,特征重叠少的场景.当参与者间特征和样本重叠都很少时可以考虑使用联邦迁移学习.

本文主要关注横向联邦学习的场景.给定具有 $W$ 个样本的数据集 $D=\{(u_w,v_w)\}$,其中 $u_w$ 是第 $w$ 个样本的特征向量,$v_w$ 是标签.神经网络函数的输出可以表示为 $f(u,x)=v'$,其中 $x$ 为模型参数.数据集 $D$ 的损失函数表示为:$L_f(D,x)=\dfrac{1}{D}\sum_{\langle u_w,v_w\rangle\in D}\| v'_w-v_w\|_2$.

联邦学习的训练目标是通过改变 $x$ 来最小化损失函数,其每轮迭代的计算公式为:$x^{t+1}=x^t-\lambda\nabla L_f(D,x')$. 其中,$\lambda$ 是学习率,它代表了每次迭代中模型调整的步长.服务器使用算术平均算法或加权平均算法将 $N$ 个参与者提交的所有局部模型聚合为一个全局模型.全局模型的计算方法为:$X_{\text{global}}=\sum_{i=1}^{N}(|D_i|x_i)/(\sum_{i=1}^{N}|D_i|)$. **1.2　同态哈希**

同态哈希[20]是一种具有同态特性的抗碰撞哈希函数,可以将任意大小的数据映射为固定大小的数据而且满足同态映射.简单地说,给定一个消息 $m_i\in Z_q$,一个抗碰撞的同态哈希函数 $HH:Z_q\to G_1\times G_2$ 可表示为:$HH(m_i)=(A_i,B_i)=(g^{HH_{\xi,\psi}(m_i)},h^{HH_{\xi,\psi}(m_i)})$,其中,$\xi$ 和 $\psi$ 都是在有限域 $Z_q$ 中随机选择的密钥,$HH_{\xi,\psi}()$ 是一个单向同态哈希函数,单向哈希函数 $HH_{\xi,\psi}()$ 的安全性保证了从 $HH_{\xi,\psi}(m)$ 反转来恢复 $m$ 是不可行的.给定 $HH(m_1)=(g^{HH_{\xi,\psi}(m_1)},h^{HH_{\xi,\psi}(m_1)})$ 和 $HH(m_2)=(g^{HH_{\xi,\psi}(m_2)},h^{HH_{\xi,\psi}(m_2)})$,同态哈希函数有以下性质:

1)可加性(在指数中)可以表示为 $HH(m_1+m_2)\leftarrow(g^{HH_{\xi,\psi}(m_1)+HH_{\xi,\psi}(m_2)},h^{HH_{\xi,\psi}(m_1)+HH_{\xi,\psi}(m_2)})$.

2)乘以一个常数 $\alpha$ 可以表示为 $HH(\alpha m_1)=(g^{\alpha HH_{\xi,\psi}(m_1)},h^{\alpha HH_{\xi,\psi}(m_1)})$.

## 1.3　不经意传输

不经意传输(oblivious transfer,OT)是密码学中经常用到的一个安全的两方通信协议,被广泛应用于隐私集合交集、安全多方计算等领域[21].不经意传输协议理想函数:参数,消息的长度为 $L$;输入,接收方输入一个选择比特 $b\in\{0,1\}$,发送方输入一对消息 $m_0,m_1\leftarrow\{0,1\}^L$;输出,发送 $m_b$ 给接收方.在这个协议中,发送方有一对消息 $m_0,m_1$,接收方有一个选择比特 $b$,协议执行结束后,接收方可以获得 $m_b$,而不能获得关于 $m_{1-b}$ 的任何信息,发送方也无法知道接收方获得了哪一条消息.

# 2　系统概述

## 2.1　网络模型

本文系统模型由 3 种实体构成.

密钥生成中心(key generation center,KGC):KGC 的作用是生成公私钥对和同态哈希所用参数,并且

为每个参与者生成 $T$ 个随机数(即在联邦学习的每一轮中为每一个参与者生成一个随机数).随后 KGC 离线,不再参与学习进程.在密码学领域,KGC 是一种极为常见的存在.

服务器($S$):$S$ 的主要职责是协调参与者的信息通信,包括初始化全局模型的状态,以及有效地转发和处理参与者之间的通信消息.参与者($P_i$):假设共有 $n$ 个参与者,每个 $P_i$ 拥有本地数据集 $D_i$,在本文方案中,由于隐私保护的要求,$P_i$ 训练局部模型之后会将其用两种方式加密,并由服务器转发给各个参与者.$D_i$ 也将作为测试集验证其余参与者的模型准确率,然后 $P_i$ 与 $S$ 执行 OT 协议得到干净的局部模型并聚合为全局模型.

### 2.2 安全模型

本文方案中,KGC 作为可信实体,为系统生成必要的公私钥对与算法参数,$S$ 和小部分 $P_i$ 都是半诚实的实体,虽然他们严格遵守安全聚合协议,也希望获悉或收集其余参与者的隐私信息.同时,本文还考虑到大部分 $P_i$ 可能通过上传恶意梯度信息来破坏模型的训练.基于以上存在的安全隐患,本文引入敌手 $A^*$,其拥有的能力如下:1) $A^*$ 可以监听通信信道或攻击 $S$,获取模型训练过程中 $P_i$ 上传的本地梯度信息.通过分析这些梯度信息,$A^*$ 可能能够进行模型反推,进而推理出参与者 $P_i$ 的本地敏感训练数据. 2)$A^*$ 可以攻击一个或多个拜占庭节点来构造并上传恶意的梯度信息实现对模型训练的干扰,达到投毒的目的.

在攻击模型中,敌手 $A^*$ 不能同时攻破多个参与者和服务器(即 $P_i$ 与 $P_j$ 不能共谋,$P_i$ 与 $S$ 不能共谋),该项限制条件普遍存在于安全计算协议中,而且在现实应用中也很难实现该项限制.

## 3 方案设计

本节介绍了方案的具体流程,下面给出了本文中使用的符号及其含义.

$P_i$:第 $i$ 个联邦学习参与者;$u_i^t$:$P_i$ 在第 $t$ 轮添加了随机数的模型;$S$:服务器;$z_i^t$:在第 $t$ 轮时的投毒检测向量;$D_i$:$P_i$ 的本地数据集;$z_i^t[j]$:$z_t^i$ 的第 $j$ 位;$x_i^t$:$P_i$ 在第 $t$ 轮的局部模型;$\rho$:准确率阈值;$\delta_i^t$:$P_i$ 在第 $t$ 轮添加的差分隐私噪声;$e$:本地训练迭代次数;$y_i^t$:$P_i$ 在第 $t$ 轮添加了噪声的局部模型;$\alpha_{i,j}^t$:$y_i^t$ 在 $D_i$ 上的测试准确率;$r_i^t$:$P_i$ 在第 $t$ 轮的随机数;$\eta$:学习率.

### 3.1 初始化

KGC 将安全参数 $\kappa$ 作为输入为每个参与者生成一对公私钥对 $\{pk_i, sk_i\}$ 并产生本文算法所需要的参数,如同态哈希和 OT 协议的参数.

同时 KGC 每一轮都为每个 $P_i$ 生成一个随机数 $r_i^t$,即共生成 $nT$ 个随机数,在第 $t$ 轮训练中的随机数满足以下性质:$\sum_{i=1}^{n} r_i^t = 0$.并将这些信息以安全的方式发送给各个参与者,同时公开这些随机值的同态哈希值 $HH(r_i^t)$. 服务器 $S$ 初始化全局模型 $x_0$ 并定义联邦学习的迭代轮数 $T$.

### 3.2 局部模型训练与加密

参与者 $P_i$ 使用 $D_i$ 训练出局部模型 $x_i^t$ 后在 $x_i^t$ 分别加上两个不同的值来加密 $x_i^t$,即差分隐私噪声 $\delta_i^t$ 和随机数 $r_i^t$.$P_i$ 的局部模型训练与加密过程如算法 1 所示.

算法 1 局部模型训练与加密

输入:本地数据集 $D_i$,本地训练次数 $e$.

输出:$y_i^t, HH(y_i^t), HH(x_i^t), HH(\delta_i^t), m_{i,j}^t, HH(u_i^t)$.

1: for $i \leftarrow 1$ to $e$ do

2: $x_i^t \leftarrow x_i^{t-1} - \nabla l(D_i, x_i^{t-1})$;

3: end for

4:生成差分隐私噪声 $\delta_i^t$;

5: $y_i^t = x_i^t + \delta_i^t, u_i^t + x_i^t + r_i^t$;

6: $m_{i,j}^t = \{Enc_{pk_j}(r_i^t), Enc_{pk_j}(u_i^t)\}$;

7:公开 $y_i^t, HH(y_i^t), HH(x_i^t), HH(\delta_i^t), HH(u_i^t)$;

8:将 $m_{i,j}^t$ 发送给 $S$;

由于局部模型参数将共享给所有参与者,本文使用了差分隐私噪声来保护局部模型参数 $x_i^t$ 得到 $y_i^t = x_i^t + \delta_i^t$.

联邦学习迭代总数为 $T$,根据序列组合性,为了满足全局 $\varepsilon$ 差分隐私要求,第 $t$ 次迭代满足 $\varepsilon_t$ 差分隐私要求,需要保证 $\sum_{t=1}^{T} \varepsilon_t = \varepsilon$.本文平均分配隐私预算,所以每次迭代的隐私预算是 $\frac{\varepsilon}{T}$.如果使用 $y_i^t$ 进行模型聚合的

话,则全局模型为:$\sum_{i=1}^{n} y_i^t = \sum_{i=1}^{n} (x_i^t + \delta_i^t) = \sum_{i=1}^{n} x_i^t + \sum_{i=1}^{n} \delta_i^t$.

此时的全局模型包含了大量的噪声 $\sum_{i=1}^{n} \delta_i^t$,这将对全局模型的性能造成消极影响,因此 $y_i^t$ 的作用是使其余参与者方便检测出 $x_i^t$ 是否是干净的,而不参与模型聚合.最终参与模型聚合的是 $u_i^t$,从而得到不带噪声的聚合模型.

恶意的参与者可以上传无毒的 $y_i^t$ 通过其余参与者的投毒检测,同时上传有毒的 $u_i^t$ 参与模型聚合从而达到投毒的目的.为了避免这种行为即 $y_i^t$ 与 $u_i^t$ 是由不同的 $x_i^t$ 加密而来的,本文要求 $P_i$ 同时向服务器发送 $x_i^t$,$\delta_i^t$ 和 $u_i^t$ 的同态哈希值,服务器再将这些信息转发给其余参与者,以便 $P_j$ 能够验证 $P_i$ 是否在后续通信中更改了输入.

### 3.3　全局模型检测

得到其余参与者的信息之后,$P_i$ 首先验证 $HH(y_j^t) = HH(x_j^t)HH(\delta_j^t)$ 是否成立,如果成立,则使用自己的本地数据集作为测试集,验证 $y_j^t$ 的精确度,如果精确度 $\alpha_{i,j}^t$ 大于等于阈值 $\rho$,则认为是干净的.如果不成立或者精确度 $\alpha_{i,j}^t$ 小于 $\rho$ 则认为是有毒的.$P_i$ 准备一个 $n$ 位的二进制向量 $z_i^t$,将无毒的 $y_j^t$ 对应位置设置为 1,即如果 $\alpha_{i,j}^t \geq \rho$,则 $z_i^t[j]=1$,否则,等于 0.$z_i^t$ 的第 $j$ 位指示了第 $j$ 个参与者的局部模型是干净的还是有毒的,从这里可以看到,由于每个参与者的数据集的质量是不同的,因此他们的检测能力也不同,则向量 $z_i^t$ 也不同. 算法 2 给出了参与者进行投毒检测的步骤.

算法 2　抗大多数恶意参与者的投毒检测算法

输入:本地数据集 $D_i$,差分隐私模型 $y_i^t$,准确率阈值 $\rho$.

输出:$z_i^t$.

1:设置一个 $n$ 位的二进制向量 $z_i^t$ 并初始化为全 0;
2:for $j \leftarrow 1$ to $n$ do
3:　if $HH(y_j^t) = HH(x_j^t)HH(\delta_j^t)$ then
4:　　$\alpha_{i,j}^t \leftarrow y_j^t$ 在 $D_i$ 上的测试准确率;
5:　　if $\alpha_{i,j}^t \geq \rho$ then
6:　　$z_i^t[j] = 1$;
7:　　end if
8:　end if
9:end for
10:return $z_i^t$

### 3.4　局部模型聚合

$P_i$ 使用自己的向量 $z_i^t$ 作为不经意传输协议的输入,而服务器则将 $m_{j,i}^t = \{Enc_{pk_i}(u_i^t), Enc_{pk_i}(r_i^t)\}$ 作为输入,$P_i$ 作为接收者解密得到 $\beta_i^t$.其中,如果 $z_i^t[j]=1$ 则 $\beta_i^t[j]=u_j^t$,否则 $\beta_i^t[j]=r_j^t$. 算法 3 给出了保护全局模型的聚合算法的详细步骤.

算法 3　保护全局模型的聚合算法

输入:$z_i^t$,$\{Enc_{pk_i}(u_j^t), Enc_{pk_i}(r_j^t)\}$.

输出:$x_i^{t+1}$.

1:for $j \leftarrow 1$ to $n$ do
2:　if $HH(u_j^t) = HH(x_j^t)HH(r_j^t)$ then
3:　　$Enc_{pk_i}(\beta_i^t[j]) \leftarrow OT(z_i^t, m_{j,i}^t)$;
4:　　$\beta_i^t[j] = Dec_{sk_i}(Enc_{pk_i}(\beta_i^t[j]))$;
5:　end if
6:end for
7:$x_i^{t+1} = \sum_{j=1}^{n} \beta_i^t[j]$;
8:return $x_i^{t+1}$

为了防止 $P_j$ 发送有毒的 $u_j^t$,即 $P_j$ 篡改了自己的输入,$P_i$ 验证自己的选择向量 $z_i^t$ 对应位为 1 的加密模型是否满足 $HH(u_j^t) = HH(x_j^t)HH(r_j^t)$,如果成立则将其聚合得到全局模型 $x_i^{t+1} = \sum_{j=1}^{n} \beta_i^t[j]$.由于 OT 协议的性质,$P_i$ 只能拿到 $Enc_{pk_i}(u_j^t)$,$Enc_{pk_i}(r_j^t)$ 中的一个,因此 $P_i$ 无法重构出 $P_j$ 的本地模型.同时由于服务器不知道 $P_i$ 的输入 $z_i^t$,因此服务器无法获得协议的输出结果,即服务器不知道 $P_i$ 的聚合模型,所以该方案有效保护了每个参与者的聚合模型不泄露.

# 4 安全分析

## 4.1 正确性分析

**定理 1** 正确性.如果参与者和服务器按照上述流程执行协议,最终可以得到正确的聚合结果.

**证明** 假设对于 $P_i$ 来说,诚实的参与者集合为 $G_i$,恶意的参与者集合为 $B_i$,则:

$$\sum_{j=1}^{n}\beta_i^t[j]=\sum_{P_j\in G_i}\beta_i^t[j]+\sum_{P_j\in B_i}\beta_i^t[j]=\sum_{P_j\in G_i}u_j^t+\sum_{P_j\in B_i}r_j^t=\sum_{P_j\in G_i}(x_j^t+r_j^t)+\sum_{P_j\in B_i}r_j^t=\sum_{P_j\in G_i}x_j^t+\sum_{P_j\in B_i\cap G_i}r_j^t=\sum_{P_j\in G_i}x_j^t=x_i^{t+1}.$$

## 4.2 安全性分析

**定理 2** 局部与聚合模型隐私性.本文算法在联邦学习过程中,可以保证用户局部与聚合模型不泄露给敌手.

**证明** 服务器方拿到的关于 $P_i$ 的消息有 $y_i^t$,$Enc_{pk_j}(u_i^t)$,$Enc_{pk_i}(r_i^t)$ 以及一些同态哈希值 $HH(y_i^t)$,$HH(x_i^t)$,$HH(\delta_i^t)$,$HH(u_i^t)$.

首先,根据差分隐私的性质,满足 $(\max\varepsilon_i)$ 差分隐私,其中 $\varepsilon_i$ 是每个参与者 $P_i$ 的隐私预算.在 $y_i^t$ 发布后,攻击者就无法从 $y_i^t$ 中推断出敏感数据信息.

其次,根据同态哈希函数的单向性,服务器不能从 $HH(y_i^t)$,$HH(x_i^t)$,$HH(\delta_i^t)$,$HH(u_i^t)$ 中求逆恢复出 $y_i^t$,$x_i^t$,$\delta_i^t$,$u_i^t$,所以服务器无法从这些同态哈希值中得到 $P_i$ 的隐私信息.

最后,证明服务器无法从 $Enc_{pk_j}(u_i^t)$,$Enc_{pk_i}(r_i^t)$ 中得到 $u_i^t$ 和 $r_i^t$.构造模拟器 $S_1$ 模拟服务器的视图.模拟器 $S_1$ 选择两个均匀分布的随机值 $R_1$ 和 $R_2$,并用 $P_j$ 的公钥加密得到 $Enc_{pk_j}(R_1)$ 和 $Enc_{pk_j}(R_2)$.根据公钥加密的随机性,服务器无法区分模拟器 $S_1$ 随机生成的 $Enc_{pk_j}(R_1)$,$Enc_{pk_j}(R_2)$ 和真实执行过程中 $P_i$ 发送的 $Enc_{pk_j}(u_i^t)$,$Enc_{pk_i}(r_i^t)$,二者在计算上是不可区分的,即:$\{S_1(R_1,R_2)\}\overset{c}{\equiv}\{view(u_i^t,r_i^t)\}$.

接下来证明服务器无法得到 $P_i$ 的聚合模型.构造模拟器 $S_2$ 模拟服务器的视图,模拟器 $S_2$ 随机采样一个 $n$ 位的二进制向量 $z_2$,并与服务器交互执行 OT 协议.根据 OT 协议的安全性,服务器无法区分交互方的输入是 $z_2$ 还是 $z_i^t$,二者在计算上是不可区分的,即服务器的视图在理想世界和真实世界中是不可区分的.

综上,本文方案能够保证用户局部模型与聚合模型的隐私性.

# 5 实验分析

本节通过在真实数据集上的实验来评估所提方案的性能.在本研究中,本文利用一台配备 AMD 锐龙 74800U CPU、1.8 GHz 和 16.0 GB RAM 的笔记本微型计算机对所提出的方案进行了评估.使用 Python 语言进行了模型加密和聚合,并在模型训练中使用了广泛使用的 MNIST 数据集.该数据集包含 6 万张训练图像和10 000张测试图像,每个样本都是一个从 0 到 9 的灰度手写数字,分辨率大小为 28×28.此外,训练数据集在参与者之间平均分布.使用卷积神经网络(CNN)对模型进行训练,其体系结构包括两个卷积层,一个全连接层和一个 softmax 输出层.在整个实验过程中,设置的批处理大小为 10,学习率为 0.01.利用标签翻转攻击[22]模拟投毒攻击,修改训练数据的标签,同时保持样本特征不变.

## 5.1 准确率与损失

本文假设存在 100 个参与者其中有 70%是恶意参与者,对比了准确率阈值分别在 70%、80%和 90%时聚合模型的准确率与损失.根据图 1 和图 2 可以看到在 MINIST 数据集上,阈值大小与聚合模型的准确率成正比,与损失成反比,阈值设定越大则模型的准确率越高,模型损失值越小即模型性能越好.

为了模拟拥有不同质量的数据集,使用[0,1]范围内的随机噪声替换原本数据集的 Pa 部分.通过改变 Pa 的值来模拟不同质量的数据集.表 1 显示了当准确度阈值 $\rho$ 为 70%时,Pa 分别在 0%,10%,30%和 50%时不同迭代轮数的模型准确度.实验表明随着迭代轮数的增加数据质量越高的参与者得到的聚合模型准确

度越高,反之亦然.这说明该方案能够保证参与者的公平性,从而提升参与者参与联邦学习的意愿.

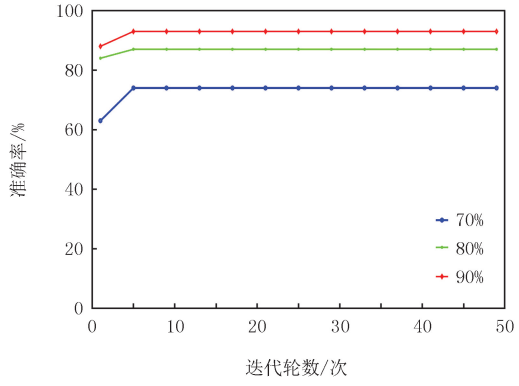


图1　不同准确率阈值下的聚合模型准确率

Fig.1　Aggregation model precision under different precision thresholds

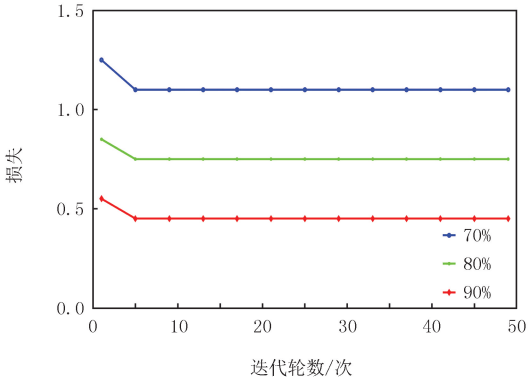


图2　不同准确率阈值下的聚合模型损失

Fig.2　Aggregation model loss under different precision thresholds

**表 1　拥有不同数据质量的聚合模型准确率对比**

**Tab.1　Comparison of accuracy of aggregation models with different data qualities**

| Pa | 迭代轮数 | | | | | Pa | 迭代轮数 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 3 | 5 | 10 | 20 | | 1 | 3 | 5 | 10 | 20 |
| 0% | 85.6 | 87.7 | 88.9 | 89.4 | 90.1 | 30% | 74.5 | 76.9 | 72.9 | 87.8 | 88.0 |
| 10% | 80.7 | 85.3 | 88.2 | 89.1 | 89.2 | 50% | 72.1 | 74.7 | 77.2 | 80.1 | 80.1 |

## 5.2　方案对比

表 2 将本文方案与文献[23－25]进行了对比.文献[23]提供了隐私保护,使用干净验证集保证恶意客户端占多数时仍能够保护模型的鲁棒性,但是没有考虑到保护模型产权;文献[24]提供了隐私保护,使用 Krum 算法只保证在恶意客户端占少数时模型的鲁棒性,此外,该文献也未考虑模型产权的保护;文献[25]使用同态加密保护局部模型的隐私,但是该框架未考虑模型鲁棒性与产权保护;相比较而言,本文保护了模型的隐私性、恶意客户端占多数时模型的鲁棒性以及聚合结果.

**表 2　不同算法功能对比**

**Tab.2　Function comparison of different algorithms**

| 方案 | 隐私性 | 鲁棒性 | | 模型产权保护 |
|---|---|---|---|---|
| | | 恶意客户端少数 | 恶意客户端多数 | |
| 本文 | 是 | 是 | 是 | 是 |
| 文献[23] | 是 | 是 | 是 | 否 |
| 文献[24] | 是 | 是 | 否 | 否 |
| 文献[25] | 是 | 否 | 否 | 否 |

由于文献[25]不能保证模型的鲁棒性,因此将本文所提算法与服务器端干净验证集[23]、Krum[24]和裁剪均值算法[3]防御投毒攻击的能力进行了对比.这里将准确率阈值设置为 90%,分别比较了恶意参与者比例为 10%、30%、50%和 70%情况下 4 种方案的聚合模型性能.

由图 3 和图 4 可以看到,当恶意参与者比例较小时,4 种检测方案的差距不大,模型准确率均保持 90%左右.当恶意参与者比例达到 50%时,Krum 和裁剪均值算法的准确率开始出现了下降,同时损失也相应地快速增长.当客户端的数据集质量良好,则所提方案与干净验证集防御投毒攻击的能力近乎相等,且恶意客户端的数量变化对聚合模型性能影响较小.

由于服务器可以与参与者并行计算,所以通过统计单个实体在单次迭代中所需的时间进行测试.如图 5 所示,将本文算法与文献[23－25]所提算法的单次迭代时间进行了对比.文献[23－24]算法单次迭代的时间开销分别约为本文算法的 7 倍和 5 倍.本文算法在保护模型隐私性、鲁棒性以及模型知识产权的情况下,单次迭代时间略高于文献[25].

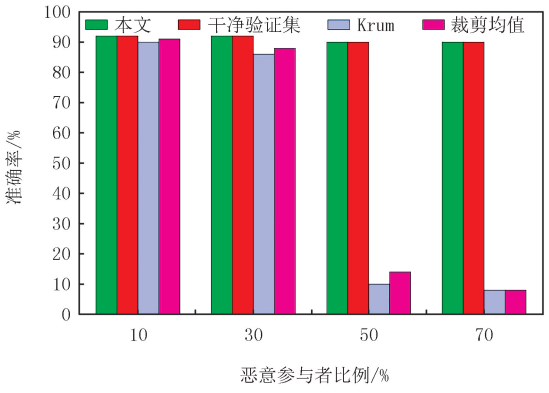


图3 4种方案聚合模型准确率

Fig.3 Aggregation model accuracy of four schemes

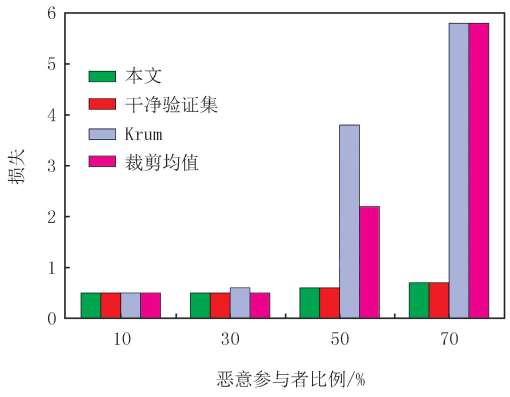


图4 4种方案聚合模型损失

Fig.4 Aggregation model loss of four schemes

## 6 结 论

本文设计了一个抗恶意参与者大多数且保护聚合模型不泄露的安全联邦学习算法.在参与者上传局部模型时分别使用差分隐私和随机数保护了局部模型不泄露;在模型检测阶段通过使用本地数据集作为验证集从而实现了抗恶意参与者大多数,同时在聚合阶段保护聚合模型不泄露给服务器.实验结果表明,本文算法准确率阈值设置越大模型性能越好.此外,本文算法即使在恶意参与者占多数时仍然能够检测出有毒模型.作为未来的发展方向,本文计划探索更高效的模型检测方案,并研究减少联邦学习计算和通信开销的方法.

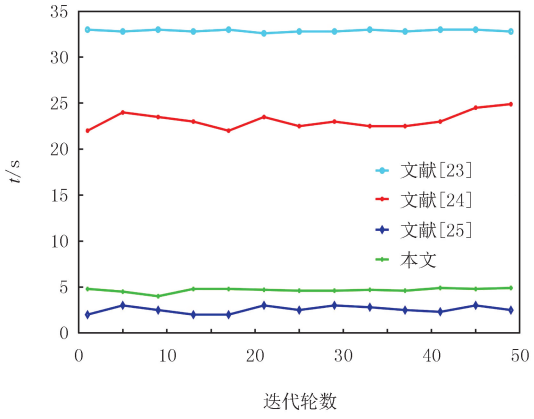


图5 不同算法运行时间对比

Fig.5 Comparison of running times of different algorithms

## 参 考 文 献

[1] MCMAHAN 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.[s.l.]:PMLR,2017.

[2] YANG H M,GE M Y,XIANG K L,et al.Using highly compressed gradients in federated learning for data reconstruction attacks[J].IEEE Transactions on Information Forensics and Security,2022,18:818-830.

[3] SHEJWALKAR V,HOUMANSADR A.Manipulating the Byzantine:optimizing model poisoning attacks and defenses for federated learning[C]//Proceedings 2021 Network and Distributed System Security Symposium.[S.l.]:Internet Society,2021.

[4] 范海菊,马锦程,李名.基于深度神经网络的遗传算法对抗攻击[J].河南师范大学学报(自然科学版),2025,53(2):82-90.
FAN H J,MA J C,LI M.Genetic algorithm based on deep neural network for countering attacks[J].Journal of Henan Normal University (Natural Science Edition),2025,53(2):82-90.

[5] 张蕴萍,翟妙如.数据要素的价值释放及反垄断治理[J].河南师范大学学报(哲学社会科学版),2022,49(6):59-65.
ZHANG Y P,ZHAI M R.Value release of data elements and antimonopoly governance[J].Journal of Henan Normal University(Philosophy and Social Sciences Edition),2022,49(6):59-65.

[6] LI Y M,ZHU M Y,YANG X,et al.Black-Box Dataset Ownership Verification via Backdoor Watermarking[J].IEEE Trans.Inf.Forensics Secur,2023,18:2318-2332.

[7] MA Z R,MA J F,MIAO Y B,et al.Pocket diagnosis:secure federated learning against poisoning attack in the cloud[J].IEEE Transactions on Services Computing,2022,15(6):3429-3442.

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

[9] XIE C,KOYEJO S,GUPTA I.Zeno++:Robust fully asynchronous SGD[C]// Proceedings of the 37th International Conference on Ma-

chine Learning.[s.l.]:PMLR,2020.
[10] GUERRAOUI R,ROUAULT S.The hidden vulnerability of distributed learning in byzantium[C]//Proceedings of the 35th International Conference on Machine Learning.[s.l.]:PMLR,2018.
[11] SHAYAN M,FUNG C,YOON C J M,et al.Biscotti:a blockchain system for private and secure federated learning[J].IEEE Transactions on Parallel and Distributed Systems,2021,32(7):1513-1525.
[12] TAO Y M,CUI S J,XU W L,et al.Byzantine-resilient federated learning at edge[J].IEEE Transactions on Computers,2023,72(9):2600-2614.
[13] LI X Y,QU Z,ZHAO S Q,et al.LoMar:a local defense against poisoning attack on federated learning[J].IEEE Transactions on Dependable and Secure Computing,2023,20(1):437-450.
[14] ZHOU J,WU N,WANG Y S,et al.A differentially private federated learning model against poisoning attacks in edge computing[J].IEEE Transactions on Dependable and Secure Computing,2023,20(3):1941-1958.
[15] MA X,ZHOU Y Q,WANG L H,et al.Privacy-preserving Byzantine-robust federated learning[J].Computer Standards & Interfaces,2022,80:103561.
[16] LIM J H,CHAN C S,NG K W,et al.Protect,show,attend and tell:empowering image captioning models with ownership protection[J].Pattern Recognition,2022,122:108285.
[17] 李璇，邓天鹏，熊金波，等.基于模型后门的联邦学习水印[J].软件学报，2024,35(7):3454-3468.
LI X,DENG T P,XIONG J B,et al.Federated learning watermark based on model backdoor[J].Journal of Software,2024,35(7):3454-3468.
[18] XU G W,LI H W,ZHANG Y,et al.Privacy-preserving federated deep learning with irregular users[J].IEEE Transactions on Dependable and Secure Computing,2022,19(2):1364-1381.
[19] 高莹，陈晓峰，张一余，等.联邦学习系统攻击与防御技术研究综述[J].计算机学报，2023,46(9):1781-1805.
GAO Y,CHEN X F,ZHANG Y Y,et al.A survey of attack and defense techniques for federated learning systems[J].Chinese Journal of Computers,2023,46(9):1781-1805.
[20] BELLARE M,GOLDREICH O,GOLDWASSER S.Incremental cryptography:the case of hashing and signing[C]// Advances in Cryptology — CRYPTO'94.Berlin,Heidelberg:Springer Berlin Heidelberg,1994:216-233.
[21] 张恩，秦磊勇，杨刃林，等.基于弹性秘密共享的多方门限隐私集合交集协议[J].软件学报，2023,34(11):5424-5441.
ZHANG E,QIN L Y,YANG R L,et al.Multi-party threshold private set intersection protocol based on robust secret sharing[J].Journal of Software,2023,34(11):5424-5441.
[22] HUANG L,JOSEPH A D,NELSON B,et al.Adversarial machine learning[C]//Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence.Chicago:ACM,2011:43-58.
[23] LIU H,ZHANG S P,ZHANG P F,et al.Blockchain and federated learning for collaborative intrusion detection in vehicular edge computing[J].IEEE Transactions on Vehicular Technology,2021,70(6):6073-6084.
[24] 方晨，郭渊博，王一丰，等.基于区块链和联邦学习的边缘计算隐私保护方法[J].通信学报，2021,42(11):28-40.
FANG C,GUO Y B,WANG Y F,et al.Edge computing privacy protection method based on blockchain and federated learning[J].Journal on Communications,2021,42(11):28-40.
[25] KU H C,SUSILO W,ZHANG Y D,et al.Privacy-Preserving federated learning in medical diagnosis with homomorphic re-Encryption[J].Computer Standards & Interfaces,2022,80:103583.

## The aggregation model protection algorithm in scenarios with majority of malicious participant

Zhang En[a,b], Gao Ting[a], Huang Yuchen[a]

(a. College of Computer and Information Engineering; b. Engineering Lab of Intelligence Business and Internet of Things of Henan Province, Xinxiang 453007, China)

**Abstract**: Privacy-preserving federated learning can help multiple participants build a machine learning model. However, this method is difficult to defend against poisoning attacks when malicious participants are in the majority. Additionally, users or servers may privately sell the aggregated model. To address these issues, a secure aggregation scheme is proposed to resist most malicious participants while protecting the privacy of the aggregated result. In the training phase, participants use differential privacy noise and random numbers to protect their local models. Then, participants test the accuracy of differential privacy models of other participants and record the results in a vector. Finally, participants and the server execute the oblivious transfer protocol to obtain the aggregated model. The security and correctness are proved through a security analysis. The experimental results show that the algorithm can maintain good detection ability even when malicious participants are in the majority and ensure the fairness of the participants to some extent.

**Keywords**: federated learning; privacy-preserving; oblivious transfer; homomorphic hash

[责任编校　陈留院　杨浦]