基于自适应拜占庭防御的安全联邦学习方案

周由胜[1,2]，高璟琨[1,2]，左祥建[1]，刘媛妮[1]
（1. 重庆邮电大学网络空间安全与信息法学院，重庆 400065；2. 重庆邮电大学计算机科学与技术学院，重庆 400065）

摘　要：针对现有联邦学习方案无法自适应防御拜占庭攻击，且模型准确度低的问题，提出了一种基于自适应拜占庭防御的安全联邦学习方案。通过激励关联的自适应初步聚合和基于指数加权平均的全局聚合，在为局部模型和全局模型提供差分隐私扰动实现隐私保护的前提下最低程度地扰动全局模型，对拜占庭客户端局部模型给予不同的惩罚以自适应防御拜占庭攻击，调动参与者的积极性，并达到较高的模型准确度。实验结果表明，对于不同拜占庭客户端占比，所提方案与其他对比方案相比模型准确度分别平均提升3.51%、3.55%和5.12%，在自适应防御拜占庭攻击时达到了较高的模型准确度。

关键词：联邦学习；边缘计算；安全隐私保护；拜占庭攻击
中图分类号：TP309.2
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024138

Secure federated learning scheme based on adaptive Byzantine defense

ZHOU Yousheng[1,2], GAO Jingkun[1,2], ZUO Xiangjian[1], LIU Yuanni[1]
1. School of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
2. College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract: Aiming at the problem that the existing federated learning schemes cannot adaptively defend Byzantine attacks and low model accuracy, a secure federated learning scheme based on adaptive Byzantine defense was proposed. Through adaptive preliminary aggregation associated with incentives and global aggregation based on exponential weighted average, the global model was minimally perturbed on the premise of providing differential privacy perturbations for both the local model and the global model to achieve privacy protection. Different penalties were given to Byzantine client local models to adaptively defend Byzantine attacks, mobilized the enthusiasm of participants, and achieved higher model accuracy. Experimental results show that for different proportions of Byzantine clients, comparing the proposed scheme with other comparative schemes, the model accuracy is increased by 3.51%, 3.55% and 5.12% on average respectively, achieving higher model accuracy when adaptively defending Byzantine attacks.

Keywords: federated learning, edge computing, security and privacy protection, Byzantine attack

0　引言

与传统机器学习不同，联邦学习（FL, federated learning）采用分布式的架构，各数据源不需要传输其本地数据，只需要通过交换模型参数，即可

收稿日期：2024–04–08；**修回日期：**2024–07–05
通信作者：左祥建，zuoxj@cqupt.edu.cn
基金项目：国家自然科学基金资助项目（No.62272076）；重庆市教委科学技术研究基金资助项目（No.KJQN202200625）；重庆市自然科学基金资助项目（No.CSTB2022NSCQ-MSX0038）
Foundation Items: The National Natural Science Foundation of China (No.62272076), The Science and Technology Research Program of Chongqing Municipal Education Commission (No. KJQN202200625), The Natural Science Foundation of Chongqing (No.CSTB2022NSCQ-MSX0038)

协同训练全局模型，从而在保护数据隐私的前提下实现数据共享计算。边缘计算作为一种分布式的数据处理与计算架构，通过将计算任务分散到下属的边缘节点来解决集中式计算中服务器负载过高的问题。因此，边缘计算能够很好地与联邦学习技术结合应用，如在自动驾驶、虚拟现实、智慧城市等方面。联邦学习最早的聚合算法是2016年由Google提出的联邦平均聚合算法FedAvg[1]。在联邦平均聚合算法中，全局模型由客户端的局部模型加权平均得到。Liu等[2]和Kang等[3]也采用联邦平均聚合算法将各客户端的局部模型进行加权聚合，得到全局模型。

联邦学习中间参数传输过程中的隐私保护方式，主要包括数据加密、安全多方计算、差分隐私（DP, differential privacy）等方式。1) 在数据加密方面，金歌等[4]将众包与同态加密相结合，并将训练任务众包给各个边缘参与者，利用同态加密来防止数据隐私泄露，实现隐私保护计算。李瑞琪等[5]提出了基于NTRU（number theory research unit）的多密钥同态代理重加密方案，利用密文扩张设计了一种新的同态密文形式和运算过程，并添加代理重加密功能来保护联邦学习中的模型参数。2) 在安全多方计算方面，Xu等[6]设计了基于格的多用途秘密共享方案，通过允许参与者在本地更新各自的秘密份额，防止参与者的梯度隐私受到量子攻击，保护参与者的数据隐私。Chu等[7]将秘密共享与联邦学习技术相结合，通过基于秘密共享和Diffie-Hellman密钥交换协议的安全聚合增强联邦学习系统的隐私性，实现隐私保护。Zhang等[8]提出了基于同态加密和安全多方计算的隐私保护方案，利用Diffie-Hellman密钥交换协议和Shamir秘密共享技术实现对联邦学习的隐私保护。3) 在差分隐私方面，Abadi等[9]和高胜等[10]在联邦学习中采用差分隐私，通过添加不同类型的噪声（如高斯噪声或拉普拉斯噪声）对局部模型进行扰动从而保护模型参数，实现隐私保护联邦学习。Wu等[11]提出了一种新颖的基于DP的随机梯度下降（SGD, stochastic gradient descent）算法，通过添加差分隐私噪声降低模型的自适应成本，并证明了基于DP的SGD算法的性能限制与隐私级别和训练数据集的大小有关。Wei等[12]提出了一种模型聚合联邦学习噪声，通过适当地适应人工噪声的不同方差来满足不同保护级别下的差分隐私。Wang等[13]提出了k-子集机制，从分类数据扩展到离散型数据，实现对客户端上传数据的估计并使用局部差分隐私（LDP, local differential privacy）方式来保护客户端数据。Lang等[14]提出了一种联合隐私增强和量化的方法，将量化失真转换为具有可控方差的加性噪声项，从而利用失真增强隐私保护，提高联邦学习的隐私性。Gauthier等[15]提出了一种个性化图联邦学习框架，利用动态零集中差分隐私，通过噪声序列扰乱模型交换，实现对客户端模型的隐私保护。Yin等[16]提出了一种联邦学习混合隐私保护方法，采用函数加密算法，结合局部贝叶斯差分隐私实现隐私保护联邦学习。Zhu等[17]提出了一种分布式标签上的隐私保护垂直联邦学习系统，采用同态加密和差分隐私将高斯噪声添加到叶子权重，防止数据隐私泄露。

虽然上述方案在一定程度上可以保护联邦学习模型的数据隐私，但当系统中存在拜占庭客户端时，上述方案将会面临拜占庭攻击的威胁。Blanchard等[18]提出了Krum，通过捕获参数服务器聚合规则的充分条件，实现联邦学习的拜占庭容错。Fan等[19]提出了一种尽力而为投票功率控制策略，通过本地参与者以最大功率传输本地梯度，增强了对拜占庭攻击的鲁棒性。Data等[20]提出了一种拜占庭鲁棒SGD算法，采用高维鲁棒均值估计算法来过滤损坏的向量，对拜占庭攻击进行防御。Huang等[21]通过平滑的几何中值聚合来实现拜占庭鲁棒，对拜占庭攻击进行防御。穆旭彤等[22]将拜占庭防御与安全多方计算相结合，利用主成分分析与K-均值聚类进行拜占庭防御，并通过加法秘密共享实现隐私保护，但缺乏对参与客户端的激励，无法实现自适应拜占庭防御，不利于调动参与客户端的积极性。Lyu等[23]提出了一种隐私保护且拜占庭稳健的压缩器，通过LDP方式实现隐私保护，利用多数投票聚合符号梯度来防御拜占庭攻击，但该方案缺乏对参与客户端的激励。Zhu等[24]提出了一种具有近乎最佳统计率的拜占庭鲁棒联邦学习协议，实现了对拜占庭攻击的防御，但缺乏隐私保护和激励关联的自适应拜占庭防御。

针对上述问题，本文提出了一种基于自适应拜占庭防御的安全联邦学习方案，在拜占庭客户端引入惩罚机制，并对局部模型更新、初步聚合过程和全局聚合过程进行优化，实现联邦学习模型隐私保护与模型准确度平衡，具体贡献如下。

1) 针对现有联邦学习方案无法自适应防御拜占庭攻击的问题，提出了激励关联的自适应初步聚合算法，对不同可信程度的拜占庭客户端局部模型实施不同的惩罚处理力度，以削弱其对全局模型训练的影响，实现自适应防御拜占庭攻击，并调动联邦学习参与者的积极性。

2) 针对现有联邦学习方案模型准确度低的问题，提出了基于指数加权平均的全局聚合算法，为局部模型和全局模型提供差分隐私扰动实现隐私保护，实现对全局模型的最低程度扰动，从而达到较高的模型准确度。

1 预备知识

1.1 联邦学习

假设一个联邦学习系统中有k个客户端和一个服务器，客户端i（$i\in(1,k)$）拥有本地数据集DB_i，则客户端的本地损失函数可表示为

$$J(\boldsymbol{\theta})=\frac{1}{m}\sum_{j=1}^{m}L\left(f\left(\boldsymbol{\theta},x_j\right),y_j\right) \tag{1}$$

其中，m为样本个数，(x_j,y_j)为客户端i在本地数据集中的数据点j（$j\in(1,m)$），$\boldsymbol{\theta}$为模型参数向量，$J(\boldsymbol{\theta})$为损失函数。客户端每一轮训练的目标是最小化损失函数$J(\boldsymbol{\theta})$。客户端进行局部模型更新，如式(2)所示。

$$\boldsymbol{w}_i=\boldsymbol{w}_G^0-\alpha\frac{\partial J(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \tag{2}$$

其中，α为学习率。

训练完成后，客户端将局部模型$\boldsymbol{w}_i$上传到服务器，由服务器进行聚合，从而产生下一轮的全局模型$\boldsymbol{w}_G^1$。

$$\boldsymbol{w}_G^1=\sum_{i=1}^{k}\frac{n_i}{n}\boldsymbol{w}_i \tag{3}$$

其中，n_i为客户端i数据集DB_i的长度，n为所有客户端数据集的长度之和，即$n=\sum_{i=1}^{k}n_i$。

1.2 差分隐私

定义1 差分隐私。令$A{:}D\to R$为随机算法，D和D'是最多有一条记录不同的2个数据集，$O\in R$为算法A的输出，若算法A满足式(4)，则算法A满足ϵ-DP。

$$\Pr\left[A(D)=O\right]\leqslant \mathrm{e}^{\epsilon}\Pr\left[A(D')=O\right] \tag{4}$$

其中，$\Pr\left[A(D)=O\right]$为$A(D)=O$的概率，ϵ为隐私预算，它决定了隐私保护的效果。通常，ϵ越小，隐私保护效果越好，但模型准确度会下降。

差分隐私的一个常用性质是顺序组合性，即对于一系列随机算法$f_1,f_2,f_3,\cdots,f_n$，若$\forall f_i(1\leqslant i\leqslant n)$满足$\epsilon_i$-DP，则整个过程满足$\sum_{i=1}^{n}\epsilon_i$-DP。

定义2 灵敏度。对于任意函数$f{:}x\to\boldsymbol{R}^d$，其中$\boldsymbol{R}^d$为函数f所输出的d维向量，则灵敏度Δf可表示为

$$\Delta f=\max_{x,x'}\|f(x)-f(x')\|_p \tag{5}$$

其中，x和x'表示一对相邻的输入数据集合，$\|\cdot\|_p$表示Lp范数。

定义3 拉普拉斯机制。给定一个函数$f{:}D\to\boldsymbol{R}^d$，其灵敏度为Δf，若随机算法M的输出结果满足$M(D)=f(D)+\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)^d$，则随机算法$M$满足$\epsilon$-DP。其中，$\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)^d$为$d$维随机向量，其服从参数为$\frac{\Delta f}{\epsilon}$的拉普拉斯分布。

1.3 Multi-Krum算法

Multi-Krum算法通过尽可能筛选掉异常的梯度来保护全局模型安全。服务器在收到来自各个客户端的梯度后，对于每一个梯度$\boldsymbol{g}_i$，取与$\boldsymbol{g}_i$距离最近的$n-f-2$个梯度的距离之和作为梯度$\boldsymbol{g}_i$的得分。

$$\mathrm{score}\left(\boldsymbol{g}_i\right)=\sum_{i\to j}\|\boldsymbol{g}_i-\boldsymbol{g}_j\|_2 \tag{6}$$

其中，$i\to j$表示$\boldsymbol{g}_j$是由与梯度$\boldsymbol{g}_i$距离最近的$n-f-2$个梯度所组成集合中的其中一个成员，n表示收到的梯度总数，$\|\cdot\|_2$表示L2范数，f表示拜占庭梯度的数量。选取得分最低的$n-f$个梯度作为最终参与聚合的梯度。

1.4 层次分析法

假设γ_i^t对应的因素共有n个，分别为$W_1,W_2,W_3,\cdots,W_n$，相应的权重因子分别为$\vartheta_1,\vartheta_2,\vartheta_3,\cdots,\vartheta_n$（$\vartheta_i\in(0,1),i=1,2,\cdots,n$且$\sum_{i=1}^{n}\vartheta_i=1$），即$\gamma_i^t=\vartheta_1W_1+\vartheta_2W_2+\vartheta_3W_3+\cdots+\vartheta_nW_n$。现采用层次分析法（AHP, analytic hierarchy process）来确定γ_i^t。首先根据$W_1,W_2,W_3,\cdots,W_n$这n个因素，构造判定矩阵

$$\boldsymbol{A}=\begin{bmatrix}a_{11}&a_{12}&\cdots&a_{1n}\\a_{21}&a_{22}&\cdots&a_{2n}\\\vdots&\vdots&\ddots&\vdots\\a_{n1}&a_{n2}&\cdots&a_{nn}\end{bmatrix} \tag{7}$$

判定矩阵$\boldsymbol{A}$中的每一项$a_{ij}(i,j = 1,2,\cdots,n)$都表示因素$i$相对于因素$j$的重要程度，$a_{ij}$的取值及对应含义如表1所示。

表1　　a_{ij}的取值及对应含义

取值(标度)	含义
1	因素i和因素j同等重要
3	因素i比因素j稍微重要
5	因素i比因素j明显重要
7	因素i比因素j强烈重要
9	因素i比因素j极端重要
2,4,6,8	上述两相邻判断的中值情况
上述取值的倒数	如果因素i相对于因素j的标度为a,则因素j相对于因素i的标度为$\frac{1}{a}$

因此，判定矩阵的对角线元素总为1（因为一个因素与其自身是同等重要的），且关于对角线对称的2个位置取值互为倒数（$a_{ij}a_{ji} = 1$）。

接下来可通过算术平均、几何平均和特征值3种方法求得权重向量$\left[\vartheta_1,\vartheta_2,\cdots,\vartheta_n\right]$。为了减小误差，保证计算结果的稳健性，本文将通过这3种方法求得的3个权重向量取平均值作为$\vartheta_1,\vartheta_2,\vartheta_3,\cdots,\vartheta_n$的最终取值。因此，可由表达式$\gamma_i^t = \vartheta_1 W_1 + \vartheta_2 W_2 + \vartheta_3 W_3 + \cdots + \vartheta_n W_n$，确定$\gamma_i^t$的值。

2　系统模型

2.1　系统架构

整体系统架构如图1所示，包括一个应用服务器（主服务器）、M个边缘服务器和$N = MK$个客户端（其中下方括号里特别注明的为拜占庭客户

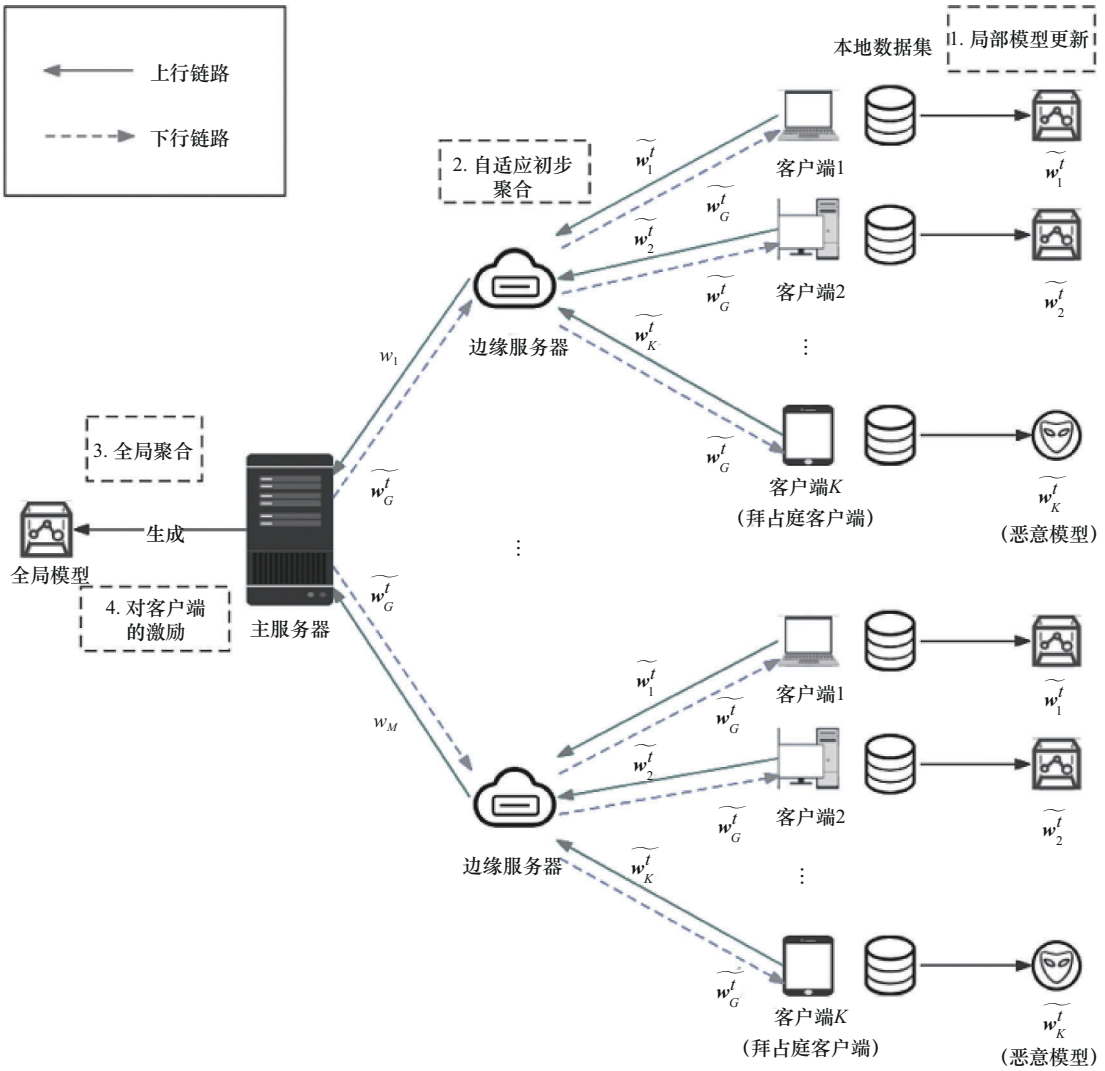


图1　系统架构

端）。客户端一共被分为M组，每个客户端P_i都拥有本地数据集D_i（$1 \leqslant i \leqslant N$），全局聚合轮数为$T$。系统包含的相关实体如下。

1) 应用服务器（主服务器）。负责完成最终的全局聚合和全局模型的发布。

2) 边缘服务器。负责对客户端产生的局部模型进行检测、处理并进行初步聚合，确定各个非拜占庭客户端的表现，从而决定奖励分配情况。

3) 客户端。首先从服务器处获取初始化全局模型，然后根据本地数据集进行局部模型训练，并将训练好的局部模型经加噪扰动后上传到边缘服务器参与自适应初步聚合。在本轮全局聚合结束后，再次从服务器处获取全局模型，开启下一轮协同训练并重复如上所述的训练过程，直至全局模型收敛或达到最大训练轮数。

2.2 威胁模型

假定本文中的边缘服务器为半诚实的，其能够诚实地执行初步聚合任务，但对客户端的隐私信息感兴趣。

外部敌手能同时窃听每个客户端上传的局部模型和主服务器广播的全局模型。尽管每个客户端的数据集都不出本地数据集，但训练出的局部模型需要与服务器共享。攻击者企图通过局部模型推导出对应客户端的相关数据信息。在客户端中，存在着一些拜占庭客户端，即内部敌手，其所上传的恶意模型会破坏全局模型聚合，从而影响全局模型的质量和收敛性。

本文用到的符号定义如表2所示。

3 方案设计

本文方案由局部模型更新、激励关联的自适应初步聚合以及基于指数加权平均的全局聚合三部分组成。首先由主服务器初始化全局模型，客户端在获取初始化全局模型之后进行局部模型更新，并将加噪扰动后的局部模型上传到边缘服务器。边缘服务器在收到来自客户端的局部模型之后，首先进行拜占庭客户端局部模型检测，检测后非拜占庭客户端会获得相应激励。然后根据检测结果进行自适应初步聚合，对非拜占庭客户端局部模型，按照其对应的初步聚合权重进行聚合；对拜占庭客户端局部模型实施不同的惩罚力度，降低其对全局模型训练的影响以自适应地防御拜占庭攻击。各边缘服务器自适应初步聚合完毕后，由主服务器进行基于指数加权平均的全局聚合，并对全局模型加以最低程度的扰动，在实现全面隐私保护的同时达到较高的模型准确度。所有过程完成后，开始下一轮全局模型训练，直至全局模型收敛或达到最大训练轮数。

表2 符号定义

符号	定义
N	客户端总数
M	边缘服务器总数
K	每组的客户端总数($K=\frac{N}{M}$)
f	全局迭代中每组所含拜占庭客户端数
T	全局训练轮数
P_i,D_i	客户端i和客户端i的本地数据集
α	学习率
β	动量参数
$\boldsymbol{w}_i^t$,$\widetilde{\boldsymbol{w}_i^t}$	局部模型和加噪后的局部模型
$\boldsymbol{\eta}_i$	对局部模型$\boldsymbol{w}_i^t$所添加的拉普拉斯噪声
γ_i^t	非拜占庭客户端局部模型参与初步聚合的权重
$\boldsymbol{w}_m$	边缘服务器的初步聚合结果
μ	全局聚合因子
$\boldsymbol{w}_G^t$,$\widetilde{\boldsymbol{w}_G^t}$	第t轮全局聚合产生的原始全局模型和第t轮全局聚合产生的加噪全局模型
$Q(\cdot)$	非拜占庭客户端P_i的模型质量
$U_i(t)$	非拜占庭客户端P_i的表现分
$\xi(i)$	客户端i的奖励权重

3.1 局部模型更新

1) 局部模型训练

首先，由主服务器初始化全局模型$\boldsymbol{w}_G^0$，并下发到客户端。在本轮及之后的训练中，客户端在获取全局模型后，利用本地数据集，通过momentum梯度下降法训练局部模型。

$$\nabla F\left(\widetilde{\boldsymbol{w}_G^{t-1}}\right)=\frac{1}{\left|b_i\right|}\sum_{j=1}^{\left|b_i\right|}\frac{\partial L\left(\widetilde{\boldsymbol{w}_G^{t-1}},d_j\right)}{\partial\widetilde{\boldsymbol{w}_G^{t-1}}} \tag{8}$$

$$\boldsymbol{v}_t=\beta\boldsymbol{v}_{t-1}+\left(1-\beta\right)\nabla F\left(\widetilde{\boldsymbol{w}_G^{t-1}}\right) \tag{9}$$

$$\boldsymbol{w}_i^t=\widetilde{\boldsymbol{w}_G^{t-1}}-\alpha\boldsymbol{v}_t \tag{10}$$

其中，$d_j=\left(x_j,y_j\right)$，$\left|b_i\right|$为批量的大小，β为动量参数($0<\beta<1$)，α为学习率($0<\alpha<1$)。

2) 局部模型扰动

训练完毕后，客户端在训练好的局部模型$\boldsymbol{w}_i^t$中添加拉普拉斯噪声$\boldsymbol{\eta}_i$后上传到附近的边缘服务器。

$$\widetilde{\boldsymbol{w}_i^t} = \boldsymbol{w}_i^t + \boldsymbol{\eta}_i = \boldsymbol{w}_i^t + \mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)^d \tag{11}$$

其中，ϵ为隐私预算，$\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)^d$为$d$维拉普拉斯噪声。

局部模型更新过程如算法1所示（以第一组为例，其他组与此同理）。

算法1　局部模型更新

输入　全局模型$\widetilde{\boldsymbol{w}_G^{t-1}}$

输出　客户端的局部模型$\left\{\widetilde{\boldsymbol{w}_i^t}\right\}_{i=1}^K$

1)//局部模型训练

2)for i=1 to K do

3)　$\nabla F\left(\widetilde{\boldsymbol{w}_G^{t-1}}\right) = \frac{1}{|b_i|}\sum_{j=1}^{|b_i|}\frac{\partial L\left(\widetilde{\boldsymbol{w}_G^{t-1}}, d_j\right)}{\partial \widetilde{\boldsymbol{w}_G^{t-1}}}$

4)　$\boldsymbol{v}_t = \beta \boldsymbol{v}_{t-1} + (1-\beta)\nabla F\left(\widetilde{\boldsymbol{w}_G^{t-1}}\right)$

5)　$\boldsymbol{w}_i^t = \widetilde{\boldsymbol{w}_G^{t-1}} - \alpha \boldsymbol{v}_t$

6)end for

7)//局部模型扰动

8)for i=1 to K do

9)　$\widetilde{\boldsymbol{w}_i^t} = \boldsymbol{w}_i^t + \mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)^d$

10)end for

11)return $\left\{\widetilde{\boldsymbol{w}_i^t}\right\}_{i=1}^K$

3.2　激励关联的自适应初步聚合

边缘服务器在收到来自客户端的局部模型后，开始进行拜占庭客户端局部模型检测。将同属于一个边缘服务器的客户端划分为一组，因此每组有$K=\frac{N}{M}$个客户端，其中包含f个拜占庭客户端（f满足$2f+2<K$）。在第t轮全局迭代中，各边缘服务器在收到其下属所有客户端的局部模型后，根据Multi-Krum算法[18]，检测出f个得分最高的局部模型（即为拜占庭客户端局部模型）。

$$\mathrm{score}\left(\widetilde{\boldsymbol{w}_i^t}\right) = \sum_{i\to k}\left(\left\|\widetilde{\boldsymbol{w}_i^t} - \widetilde{\boldsymbol{w}_k^t}\right\|_2\right)^2 \tag{12}$$

其中，$i\to k$表示$\widetilde{\boldsymbol{w}_k^t}$是由$K-f-2$个与模型$\widetilde{\boldsymbol{w}_i^t}$距离最近向量所组成的集合中的其中一个成员，$K$为各边缘服务器收到的局部模型总数，$f$为每组中拜占庭客户端的数量。

边缘服务器在检测出拜占庭客户端局部模型之后，记录其下属的各个客户端的检测结果，一方面决定可获得奖励的客户端，另一方面决定在自适应初步聚合中是否需要对局部模型进行惩罚处理。最后，各边缘服务器聚合其下属客户端的局部模型，设计自适应初步聚合过程如下（以第一组为例，其他组与此同理）。

对于非拜占庭客户端局部模型，根据其对应的初步聚合权重进行初步聚合。

$$\boldsymbol{w}_m \leftarrow \boldsymbol{w}_m + \gamma_i^t \widetilde{\boldsymbol{w}_i^t} \tag{13}$$

其中，γ_i^t为非拜占庭客户端局部模型参与初步聚合的权重($\gamma_i^t \in (0,1)$)。由于γ_i^t受多个因素的影响，为了确定γ_i^t，需要先确定这些因素各自的权重系数。层次分析法支持在多因素条件下解决因素对目标影响的问题，对问题（目标）进行分解，并求解每一个因素（准则）对目标的权重系数，并通过一致性检验提高结果的可靠性。本节将确定γ_i^t作为目标，影响γ_i^t的因素作为因素（准则）。因此，对于γ_i^t，需要采用层次分析法来确定。

对于拜占庭客户端局部模型，对其进行惩罚处理后再聚合。

$$\boldsymbol{w}_m \leftarrow \boldsymbol{w}_m + \omega^{-\mathrm{score}\left(\widetilde{\boldsymbol{w}_i^t}\right)}\widetilde{\boldsymbol{w}_i^t} \tag{14}$$

其中，$\boldsymbol{w}_m$为局部累计聚合结果（遍历所有客户端结束后即为初步聚合结果），其初始值为$\boldsymbol{0}$，$\mathrm{score}\left(\widetilde{\boldsymbol{w}_i^t}\right)$为拜占庭客户端局部模型的得分，代表模型的可信程度。

对于拜占庭客户端局部模型而言，其得分越高，代表模型可信程度越低，从而需要对其增大惩罚力度。不同得分的拜占庭客户端局部模型，在实施不同力度的惩罚处理后，在降低其对全局模型训练影响的同时，实现了自适应拜占庭防御。自适应初步聚合的整体过程如算法2所示。

算法2　自适应初步聚合

输入　客户端的局部模型$\left\{\widetilde{\boldsymbol{w}_i^t}\right\}_{i=1}^K$

输出　自适应初步聚合的结果$\boldsymbol{w}_m$

1)for i=1 to K do

2)　// 计算每个局部模型的得分

3) $\text{score}\left(\widetilde{\boldsymbol{w}}_i^t\right)=\sum_{i\to k}\left(\left\|\widetilde{\boldsymbol{w}}_i^t-\widetilde{\boldsymbol{w}}_k^t\right\|_2\right)^2$

4)end for

5)// 边缘服务器检测出f个得分最高的局部模型，即为拜占庭客户端局部模型

6)for i=1 to K do

7) //$\widetilde{\boldsymbol{w}}_i^t$未被检测为拜占庭客户端局部模型

8) if （$\widetilde{\boldsymbol{w}}_i^t$未被检测为拜占庭客户端局部模型）

9) $\boldsymbol{w}_m\leftarrow\boldsymbol{w}_m+\gamma_i^t\widetilde{\boldsymbol{w}}_i^t$

10) else

11) $\boldsymbol{w}_m\leftarrow\boldsymbol{w}_m+\omega^{-\text{score}\left(\widetilde{\boldsymbol{w}}_i^t\right)}\widetilde{\boldsymbol{w}}_i^t$

12) end if

13)end for

14)return $\boldsymbol{w}_m$

考虑到参与客户端的积极性，在激励关联的自适应初步聚合过程中，对参与客户端进行激励，以调动参与客户端的积极性。在本文的自适应初步聚合算法中，γ_i^t的值与下文对客户端激励中非拜占庭客户端的表现分和所获奖励有关。非拜占庭客户端表现分越高，所获奖励越多，说明其对应的局部模型质量就越高，因此其对应局部模型的初步聚合权重γ_i^t就越大。这能够使更高质量的局部模型取得更大的聚合权重，从而保证聚合模型的质量。

在第t轮全局迭代中（以第一组为例，其他组与此同理），边缘服务器根据自适应初步聚合阶段对其下属客户端局部模型的检测情况，计算下属的每个非拜占庭客户端的模型质量$Q\left(\boldsymbol{w}_i^t\right)$。

$$Q\left(\boldsymbol{w}_i^t\right)=\lg F\left(\boldsymbol{w}_i^t\right) \tag{15}$$

每个非拜占庭客户端P_i的表现分计算如式(16)所示。

$$U_i(t)=1-\frac{Q\left(\boldsymbol{w}_i^t\right)}{Q} \tag{16}$$

其中，$Q=\sum_{P_i\in\{P_i'|P_i'\text{为非拜占庭客户端}\}}Q\left(\boldsymbol{w}_i^t\right)$。

对于一组客户端$P_1,P_2,\cdots,P_K$，其价值$V\left(P_i\right)$定义为

$$V\left(P_i\right)=\begin{cases}U_i(t),P_i\in\left\{P_i'|P_i'\text{为非拜占庭客户端}\right\}\\0,P_i\in\left\{P_i'|P_i'\text{为拜占庭客户端}\right\}\end{cases} \tag{17}$$

最后，确定客户端的奖励权重。

对非拜占庭客户端P_i而言，奖励权重$\xi(i)$可通过式(18)计算（$P_i\in\left\{P_i'|P_i'\text{为非拜占庭客户端}\right\}$）。

$$\xi(i)=\frac{V\left(P_i\right)}{V} \tag{18}$$

其中，$V=\sum_{P_i\in\{P_i'|P_i'\text{为非拜占庭客户端}\}}V\left(P_i\right)$。

确定非拜占庭客户端的奖励权重后，非拜占庭客户端将根据奖励权重获得奖励。假设每一组的总预算为B，则有

$$\delta_i^t=\xi(i)B \tag{19}$$

其中，δ_i^t为非拜占庭客户端P_i在第t轮全局迭代中所获得的奖励。

对拜占庭客户端$P_i\in\left\{P_i'|P_i'\text{为拜占庭客户端}\right\}$而言，由于其价值$V\left(P_i\right)$为0，则对应的奖励权重也为0，因此无法参与奖励分配并获得奖励。关联自适应初步聚合的客户端激励过程如算法3所示。

算法3 客户端激励

输入 非拜占庭客户端的损失$F\left(\boldsymbol{w}_i^t\right)$，一个组的总预算$B$

输出 客户端获得的奖励δ_i^t

1)// 边缘服务器计算每个非拜占庭客户端的模型质量

2)for $P_i\in\{P_i'|P_i'\text{为非拜占庭客户端}\}$ do

3) $Q\left(\boldsymbol{w}_i^t\right)=\lg F\left(\boldsymbol{w}_i^t\right)$

4)end for

5)// 计算每个非拜占庭客户端的表现分

6)for $P_i\in\{P_i'|P_i'\text{为非拜占庭客户端}\}$ do

7) $U_i(t)=1-\frac{Q\left(\boldsymbol{w}_i^t\right)}{\sum_{P_i\in\{P_i'|P_i'\text{为非拜占庭客户端}\}}Q\left(\boldsymbol{w}_i^t\right)}$

8)end for

9)// 计算客户端的价值

10)if （P_i为非拜占庭客户端）

11) $V\left(P_i\right)=U_i(t)$

12)else

13) $V\left(P_i\right)=0$

14)end if

15)// 计算客户端的奖励权重

16)if （P_i为非拜占庭客户端）

17) $\xi(i)=\frac{V\left(P_i\right)}{\sum_{P_i\in\{P_i'|P_i'\text{为非拜占庭客户端}\}}V\left(P_i\right)}$

18)else
19)　$\xi(i)=0$
20)end if
21)// 计算客户端所获得的奖励
22)if （P_i为非拜占庭客户端）
23)　$\delta_i^t=\xi(i)B$
24)else
25)　$\delta_i^t=0$
26)end if
27)return δ_i^t

由于非拜占庭客户端表现分越高，且其局部模型在自适应初步聚合检测阶段的得分越低，说明其局部模型质量就越高，局部模型的初步聚合权重就应当越大。因此，对于γ_i^t，考虑以下3个因素，分别为非拜占庭客户端所获得的奖励δ_i^t占比W_1、非拜占庭客户端的表现分$U_i(t)$占比W_2以及非拜占庭客户端局部模型在自适应初步聚合检测阶段的表现分$W_3=1-\dfrac{\text{score}\left(\widetilde{\boldsymbol{w}_i^t}\right)}{S}$（检测得分越高，模型在自适应初步聚合阶段的表现就越差），其中$S=\sum_{P_i\in\{P_i'|P_i'\text{为非拜占庭客户端}\}}\text{score}\left(\widetilde{\boldsymbol{w}_i^t}\right)$。$\gamma_i^t$的具体确定方法详见1.4节，计算过程详见4.1节及式(22)~式(24)。

边缘服务器依据下属客户端局部模型的检测结果确定客户端的激励，即通过调整自适应初步聚合中γ_i^t的值，实现激励关联的自适应初步聚合。

3.3　基于指数加权平均的全局聚合

在各边缘服务器完成初步聚合之后，主服务器对所有边缘服务器的初步聚合结果$\boldsymbol{w}_m(m=1,2,\cdots,M)$进行基于指数加权平均的全局聚合。

$$\boldsymbol{w}_G^t\leftarrow\mu\boldsymbol{w}_G^t+(1-\mu)\boldsymbol{w}_m \tag{20}$$

其中，$\boldsymbol{w}_G^t$初始值为$\boldsymbol{w}_1$,μ为全局聚合因子($0<\mu<1$)。

然后，主服务器向$\boldsymbol{w}_G^t$添加噪声，扰动后的全局模型表示为

$$\widetilde{\boldsymbol{w}_G^t}=\boldsymbol{w}_G^t+\arg\min_{\boldsymbol{\eta}_i}\left(\|\boldsymbol{\eta}_i\|\right) \tag{21}$$

接着将扰动后的全局模型$\widetilde{\boldsymbol{w}_G^t}$分发到客户端，以备下一轮协同训练。通过基于指数加权平均的全局聚合，并对全局模型进行最低程度的扰动，达到较高的模型准确度。基于指数加权平均的全局聚合整体过程如算法4所示。

算法4　基于指数加权平均的全局聚合
输入　边缘服务器的初步聚合结果$\{\boldsymbol{w}_m\}_{m=1}^M$
输出　全局聚合模型$\widetilde{\boldsymbol{w}_G^t}$
1)// 全局聚合
2)for m=1 to M do
3)　$\boldsymbol{w}_G^t\leftarrow\mu\boldsymbol{w}_G^t+(1-\mu)\boldsymbol{w}_m$
4)end for
5)// 添加噪声以扰动全局模型
6)$\widetilde{\boldsymbol{w}_G^t}=\boldsymbol{w}_G^t+\arg\min_{\boldsymbol{\eta}_i}\left(\|\boldsymbol{\eta}_i\|\right)$
7)return $\widetilde{\boldsymbol{w}_G^t}$

4　实验评估

4.1　实验设置

本文实验在Windows 11平台上进行，硬件环境为Intel(R) Core(TM) i5-11500 2.70 GHz处理器，16 GB内存。实验数据集采用图片数据集MNIST，共包含60 000个训练样本和10 000个测试样本。训练模型采用卷积神经网络（CNN, convolutional neural network），由2个5×5卷积层、2个全连接层、一个Dropout层（防止过拟合）和一个softmax输出层组成。考虑到直观性、易理解性和计算的便利性，设定ω的取值为10，$N=80$，$M=4$，$\alpha=0.01$，$\beta=0.9$，$\mu=0.75$，$T=50$。对学习率α和动量参数β进行网格搜索，以寻求最优参数组合。设定α的取值空间为{0.01,0.005,0.02}，β的取值空间为{0.8,0.9,0.99}。结果表明，当$\alpha=0.01$，$\beta=0.9$时，α与β为最优参数组合，与本文的参数设置相吻合。

对于非拜占庭客户端局部模型参与初步聚合的权重γ_i^t的确定，考虑的因素W_1,W_2,W_3已在3.2节中说明。权重γ_i^t计算如式(22)所示。

$$\begin{aligned}\gamma_i^t=\vartheta_1W_1+\vartheta_2W_2+\vartheta_3W_3=\\ \vartheta_1\frac{\delta_i^t}{\sum_{P_i\in\{P_i'|P_i'\text{为非拜占庭客户端}\}}\delta_i^t}+\\ \vartheta_2\frac{U_i(t)}{\sum_{P_i\in\{P_i'|P_i'\text{为非拜占庭客户端}\}}U_i(t)}+\\ \vartheta_3\left(1-\frac{\text{score}\left(\widetilde{\boldsymbol{w}_i^t}\right)}{\sum_{P_i\in\{P_i'|P_i'\text{为非拜占庭客户端}\}}\text{score}\left(\widetilde{\boldsymbol{w}_i^t}\right)}\right)\end{aligned} \tag{22}$$

现采用层次分析法来确定γ_i^t。首先，根据W_1,W_2,W_3这3个因素的相对重要程度，构造判定矩阵

$$\boldsymbol{A}=\begin{bmatrix}1 & \frac{1}{2} & 3\\ 2 & 1 & 4\\ \frac{1}{3} & \frac{1}{4} & 1\end{bmatrix} \tag{23}$$

然后，计算权重向量。按算术平均法得权重向量[0.320 24,0.557 14,0.122 62]（保留五位小数，下同）。按几何平均法得权重向量[0.319 61,0.558 43,0.121 96]，按特征值法得权重向量[0.319 65,0.558 41,0.121 94]以及最大特征值$\lambda_{\max}=3.018\,29$。

接下来，对一致性进行检验。首先计算一致性指标$\mathrm{CI}=\frac{\lambda_{\max}-n}{n-1}=0.009\,145$，查表可得3阶矩阵对应的平均随机一致性指标RI = 0.52。则一致性比例$\mathrm{CR}=\frac{\mathrm{CI}}{\mathrm{RI}}=0.017\,59<0.1$，因此，一致性检验通过。

为保证计算结果的稳健性，将上述3种方法的平均值[0.3,0.6,0.1]，作为ϑ_1,ϑ_2,ϑ_3的取值，即

$$\begin{aligned}\gamma_i^t=&0.3W_1+0.6W_2+0.1W_3=\\&0.3\frac{\delta_i^t}{\sum\limits_{P_i\in\{P_i'|P_i'\text{为非拜占庭客户端}\}}\delta_i^t}+\\&0.6\frac{U_i(t)}{\sum\limits_{P_i\in\{P_i'|P_i'\text{为非拜占庭客户端}\}}U_i(t)}+\\&0.1\left(1-\frac{\mathrm{score}\left(\widetilde{\boldsymbol{w}}_i^t\right)}{\sum\limits_{P_i\in\{P_i'|P_i'\text{为非拜占庭客户端}\}}\mathrm{score}\left(\widetilde{\boldsymbol{w}}_i^t\right)}\right)\end{aligned} \tag{24}$$

至此，非拜占庭客户端局部模型参与初步聚合的权重γ_i^t得以确定。

4.2 安全性对比

将本文方案与FedAvg[1]、SecFedDMC[22]、DP-SIGN[23]和F2ED-Learning[24]的安全性进行对比。FedAvg[1]不具备对拜占庭攻击的防御和隐私保护，缺乏对参与客户端的激励。SecFedDMC[22]具备隐私保护（安全多方计算），但不具备激励关联的自适应拜占庭防御。DP-SIGN[23]具备隐私保护（差分隐私），但不具备激励关联的自适应拜占庭防御。F2ED-Learning[24]考虑对拜占庭攻击的防御但不具备隐私保护和激励关联的自适应拜占庭防御。与其他方案相比，本文方案在利用差分隐私进行隐私保护的同时，兼具对拜占庭攻击的防御功能，且具备激励关联的自适应拜占庭防御机制，该机制在其他方案中均不具备。方案功能性对比如表3所示。

表3 方案功能性对比

方案	防御拜占庭攻击	隐私保护	激励关联	自适应拜占庭防御
FedAvg[1]	×	×	×	×
SecFedDMC[22]	√	√	×	×
DP-SIGN[23]	√	√	×	×
F2ED-Learning[24]	√	×	×	×
本文方案	√	√	√	√

4.3 实验结果

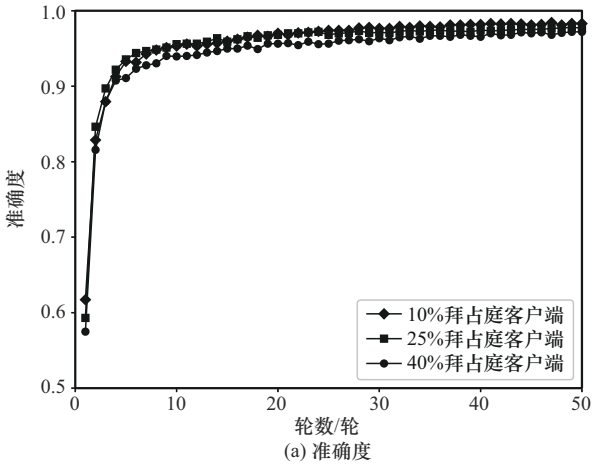
4.3.1 不同拜占庭客户端数量占比下的结果对比

在不同数量占比（10%、25%和40%）的拜占庭客户端场景下，本文方案相关指标表现如图2、表4和图3、表5所示。

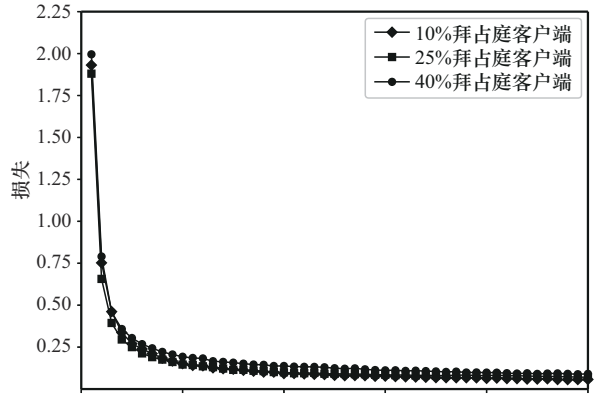
图2和表4展示了拜占庭客户端数量占比分别为10%、25%和40%时全局模型的准确度与损失情况。结果显示，10%、25%和40%拜占庭客户端占比下全局模型的准确度分别达到0 983 1、0.977 2和0.972 0，损失分别达到0.057 3、0.071 0和0.090 6。

表4 不同拜占庭客户端数量占比下模型准确度与损失对比

拜占庭客户端数量占比	指标	轮数/轮				
		10	20	30	40	50
10%	准确度	0.953 1	0.970 1	0.976 6	0.981 8	0.983 1
	损失	0.151 8	0.092 4	0.077 0	0.063 5	0.057 3
25%	准确度	0.955 7	0.968 1	0.971 4	0.974 6	0.977 2
	损失	0.146 2	0.100 0	0.085 6	0.079 4	0.071 0
40%	准确度	0.939 5	0.956 4	0.963 5	0.965 5	0.972 0
	损失	0.192 8	0.137 7	0.111 6	0.098 3	0.090 6

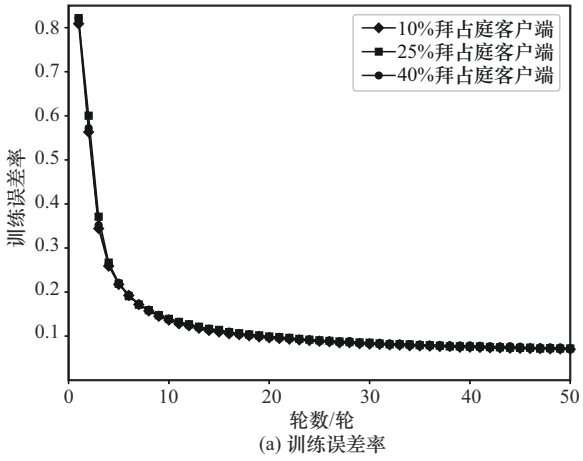


(a) 准确度

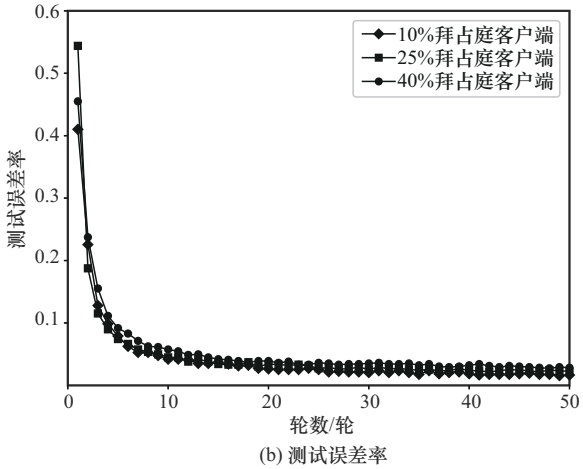


(b) 损失

图2　不同拜占庭客户端数量占比对模型准确度与损失的影响对比



(a) 训练误差率



(b) 测试误差率

图3　不同拜占庭客户端数量占比对模型训练误差率与测试误差率的影响对比

图3和表5展示了拜占庭客户端数量占比分别为10%、25%和40%时全局模型的训练误差率与测试误差率情况。结果显示，10%、25%和40%拜占庭客户端占比下全局模型的训练误差率分别达到0.070 3、0.071 9和0.072 7，测试误差率分别达到0.016 6、0.022 5和0.028 3。可以看出，随着拜占庭客户端数量占比的增加，全局模型的准确度有所下降，误差率有所升高，这表明本文方案能够有效防御拜占庭攻击。

4.3.2　不同方案(方式)对比

与Multi-Krum算法直接丢弃拜占庭客户端局部模型的方式不同，本文方案对不同可信程度的拜占庭客户端局部模型实施不同的惩罚处理力度，以削弱其对全局模型训练的影响，自适应防御拜占庭攻击。

将本文方案与FedAvg[1]、DP-SIGN[23]、F2ED-Learning[24]和Multi-Krum方式进行对比实验（由于本文方案采用差分隐私作为隐私保护手段，故同样

表5　　**不同拜占庭客户端数量占比下模型训练误差率与测试误差率对比**

拜占庭客户端数量占比	指标	轮数/轮				
		10	20	30	40	50
10%	训练误差率	0.136 5	0.096 4	0.081 9	0.075 1	0.070 3
	测试误差率	0.042 0	0.026 4	0.020 5	0.018 6	0.016 6
25%	训练误差率	0.139 0	0.099 0	0.084 9	0.076 2	0.071 9
	测试误差率	0.044 9	0.034 2	0.025 4	0.023 4	0.022 5
40%	训练误差率	0.139 2	0.098 9	0.085 3	0.078 9	0.072 7
	测试误差率	0.057 6	0.039 1	0.035 2	0.032 2	0.028 3

采用差分隐私的DP-SIGN[23]参与对比，而SecFed-DMC[22]的隐私保护手段为安全多方计算，与本文方案不同，故不参与对比）。图4与表6展示了10%拜占庭客户端下不同方案（方式）实验结果对比情况，图5与表7展示了25%拜占庭客户端下不同方案（方式）实验结果对比情况，图6与表8展示了40%拜占庭客户端下不同方案（方式）实验结果对比情况（表格中加粗数值表示本文方案实验结果）。

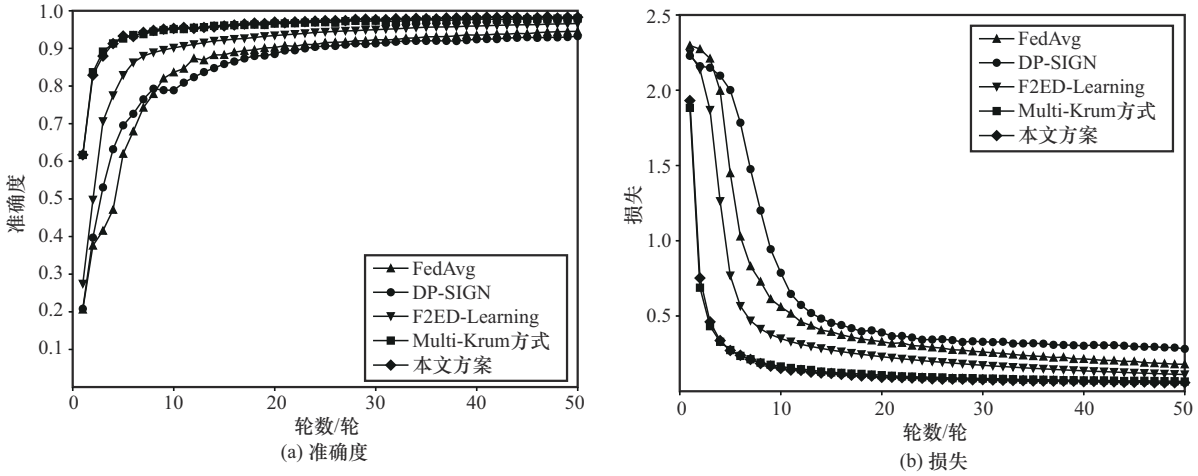


图4　不同方案（方式）实验结果对比（10%拜占庭客户端）

表6　不同方案（方式）实验结果对比（10%拜占庭客户端）

方案（方式）	指标	轮数/轮				
		10	20	30	40	50
FedAvg	准确度	0.836 6	0.902 3	0.923 2	0.935 2	0.945 4
	损失	0.560 5	0.329 7	0.261 0	0.215 2	0.179 3
DP-SIGN	准确度	0.788 8	0.885 1	0.913 2	0.924 2	0.932 6
	损失	0.787 3	0.390 6	0.328 7	0.302 2	0.281 7
F2ED-Learning	准确度	0.901 6	0.934 1	0.949 7	0.959 4	0.966 1
	损失	0.350 6	0.229 6	0.172 8	0.135 6	0.111 3
Multi-Krum方式	准确度	0.951 6	0.965 8	0.973 6	0.977 3	0.979 5
	损失	0.164 5	0.107 0	0.084 2	0.072 3	0.065 1
本文方案	准确度	**0.953 1**	**0.970 1**	**0.976 6**	**0.981 8**	**0.983 1**
	损失	**0.151 8**	**0.092 4**	**0.077 0**	**0.063 5**	**0.057 3**

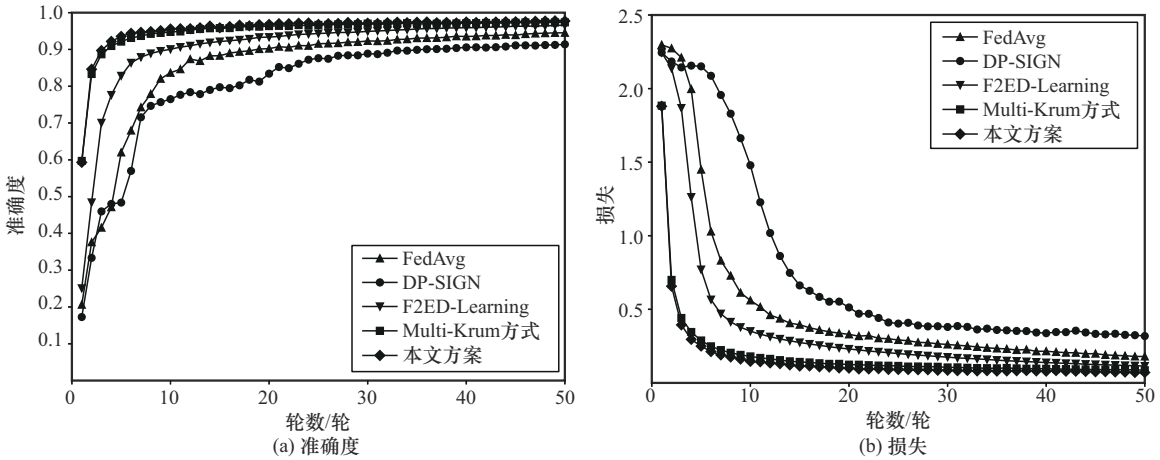


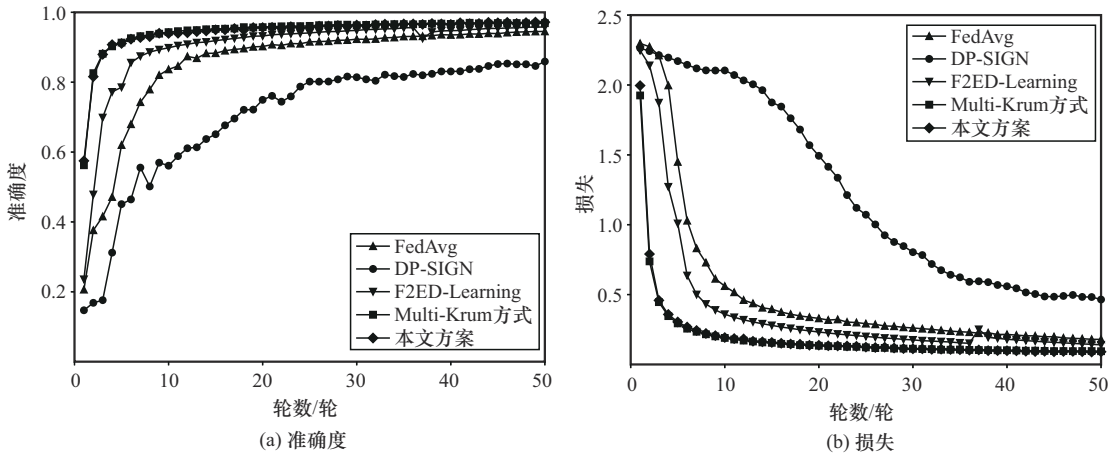
图5　不同方案（方式）实验结果对比（25%拜占庭客户端）

表 7　不同方案(方式)实验结果对比(25% 拜占庭客户端)

方案(方式)	指标	轮数/轮				
		10	20	30	40	50
FedAvg	准确度	0.836 6	0.902 3	0.923 2	0.935 2	0.945 4
	损失	0.560 5	0.329 7	0.261 0	0.215 2	0.179 3
DP-SIGN	准确度	0.764 9	0.834 2	0.888 7	0.905 7	0.913 9
	损失	1.479 2	0.512 8	0.380 2	0.337 5	0.318 3
F2ED-Learning	准确度	0.900 7	0.933 5	0.949 1	0.959 0	0.965 9
	损失	0.351 5	0.230 5	0.175 4	0.138 2	0.114 7
Multi-Krum 方式	准确度	0.947 7	0.963 5	0.970 3	0.971 5	0.973 6
	损失	0.180 0	0.123 9	0.104 9	0.095 8	0.086 4
本文方案	准确度	**0.955 7**	**0.968 1**	**0.971 4**	**0.974 6**	**0.977 2**
	损失	**0.146 2**	**0.100 0**	**0.085 6**	**0.079 4**	**0.071 0**

表 8　不同方案(方式)实验结果对比(40% 拜占庭客户端)

方案(方式)	指标	轮数/轮				
		10	20	30	40	50
FedAvg	准确度	0.836 6	0.902 3	0.923 2	0.935 2	0.945 4
	损失	0.560 5	0.329 7	0.261 0	0.215 2	0.179 3
DP-SIGN	准确度	0.561 0	0.749 5	0.814 3	0.830 8	0.859 2
	损失	2.104 1	1.493 0	0.803 4	0.559 6	0.464 6
F2ED-Learning	准确度	0.898 4	0.932 2	0.948 2	0.947 6	0.957 8
	损失	0.359 5	0.231 9	0.175 7	0.179 9	0.140 8
Multi-Krum 方式	准确度	0.943 4	0.957 6	0.963 1	0.966 8	0.968 9
	损失	0.190 0	0.134 1	0.112 8	0.101 7	0.094 8
本文方案	准确度	**0.939 5**	**0.956 4**	**0.963 5**	**0.965 5**	**0.972 0**
	损失	**0.192 8**	**0.137 7**	**0.111 6**	**0.098 3**	**0.090 6**



(a) 准确度　　(b) 损失

图 6　不同方案(方式)实验结果对比(40% 拜占庭客户端)

由图 4 和表 6 可知，在 10% 拜占庭客户端下，FedAvg、DP-SIGN、F2ED-Learning、Multi-Krum 方式和本文方案模型准确度分别达到 0.945 4、0.932 6、0.966 1、0.979 5 和 0.983 1，损失分别达到 0.179 3、0.281 7、0.111 3、0.065 1 和 0.057 3。由图 5 和表 7 可知，在 25% 拜占庭客户端下，FedAvg、DP-SIGN、F2ED-Learning、Multi-Krum 方式和本文方案模型准确度分别达到 0.945 4、0.913 9、0.965 9、0.973 6 和 0.977 2，损失分别达到 0.179 3、0.318 3、0.114 7、0.086 4 和 0.071 0。由图 6 和表 8

可知，在 40% 拜占庭客户端下，FedAvg、DP-SIGN、F2ED-Learning、Multi-Krum 方式和本文方案模型准确度分别达到0.945 4、0.859 2、0.957 8、0.968 9和0.972 0，损失分别达到0.179 3、0.464 6、0.140 8、0.094 8和0.090 6。可以看出，在10%拜占庭客户端下，本文方案相比其他对比方案（方式）模型准确度分别提升3.77%、5.05%、1.70%和0.36%，损失分别降低0.122 0、0.224 4、0.054 0和0.007 8；在25%拜占庭客户端下，本文方案相比其他对比方案（方式）模型准确度分别提升3.18%、6.33%、1.13%和0.36%，损失分别降低0.108 3、0.247 3、0.043 7和0.015 4；在40%拜占庭客户端下，本文方案相比其他对比方案（方式）模型准确度分别提升2.66%、11.28%、1.42%和0.31%，损失分别降低0.088 7、0.374 0、0.050 2和0.004 2。这表明本文方案能够在防御拜占庭攻击、提供安全隐私保护与激励的同时达到较高的模型准确度。

5 结束语

本文提出了一种基于自适应拜占庭防御的安全联邦学习方案。首先，提出了激励关联的自适应初步聚合算法，有效防御拜占庭攻击，并调动联邦学习参与者的积极性。其次，提出了基于指数加权平均的全局聚合算法，以提高模型准确度。未来将考虑设计基于区块链的安全隐私保护联邦学习方案，增强全局模型训练过程以及相关数据的可审计性与可追溯性。

参考文献：

[1] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[J]. arXiv Preprint, arXiv: 1602.05629, 2016.

[2] LIU W, CHEN L, CHEN Y F, et al. Accelerating federated learning via momentum gradient descent[J]. IEEE Transactions on Parallel and Distributed Systems, 2020, 31(8): 1754-1766.

[3] KANG J W, XIONG Z H, NIYATO D, et al. Incentive mechanism for reliable federated learning: a joint optimization approach to combining reputation and contract theory[J]. IEEE Internet of Things Journal, 2019, 6(6): 10700-10714.

[4] 金歌, 魏晓超, 魏森茂, 等. FPCBC: 基于众包聚合的联邦学习隐私保护分类系统[J]. 计算机研究与发展, 2022, 59(11): 2377-2394.
JIN G, WEI X C, WEI S M, et al. FPCBC: federated learning privacy preserving classification system based on crowdsourcing aggregation[J]. Journal of Computer Research and Development, 2022, 59(11): 2377-2394.

[5] 李瑞琪, 贾春福, 王雅飞. 基于NTRU的多密钥同态代理重加密方案及其应用[J]. 通信学报, 2021, 42(3): 11-22.
LI R Q, JIA C F, WANG Y F. Multi-key homomorphic proxy re-encryption scheme based on NTRU and its application[J]. Journal on Communications, 2021, 42(3): 11-22.

[6] XU P, HU M Q, CHEN T Y, et al. LaF: lattice-based and communication-efficient federated learning[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 2483-2496.

[7] CHU K F, GUO W S. Privacy-preserving federated deep reinforcement learning for mobility-as-a-service[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(2): 1882-1896.

[8] ZHANG L, XU J B, VIJAYAKUMAR P, et al. Homomorphic encryption-based privacy-preserving federated learning in IoT-enabled healthcare system[J]. IEEE Transactions on Network Science and Engineering, 2023, 10(5): 2864-2880.

[9] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2016: 308-318.

[10] 高胜, 袁丽萍, 朱建明, 等. 一种基于区块链的隐私保护异步联邦学习[J]. 中国科学(信息科学), 2021, 51(10): 1755-1774.
GAO S, YUAN L P, ZHU J M, et al. A blockchain-based privacy-preserving asynchronous federated learning[J]. Scientia Sinica (Informationis), 2021, 51(10): 1755-1774.

[11] WU N, FAROKHI F, SMITH D, et al. The value of collaboration in convex machine learning with differential privacy[C]//Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2020: 304-317.

[12] WEI K, LI J, DING M, et al. Federated learning with differential privacy: algorithms and performance analysis[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3454-3469.

[13] WANG S W, HUANG L S, NIE Y W, et al. Local differential private data aggregation for discrete distribution estimation[J]. IEEE Transactions on Parallel and Distributed Systems, 2019, 30(9): 2046-2059.

[14] LANG N, SOFER E, SHAKED T, et al. Joint privacy enhancement and quantization in federated learning[J]. IEEE Transactions on Signal Processing, 2023, 71: 295-310.

[15] GAUTHIER F, GOGINENI V C, WERNER S, et al. Personalized graph federated learning with differential privacy[J]. IEEE Transactions on Signal and Information Processing over Networks, 2023, 9: 736-749.

[16] YIN L H, FENG J Y, XUN H, et al. A privacy-preserving federated learning for multiparty data sharing in social IoTs[J]. IEEE Transactions on Network Science and Engineering, 2021, 8(3): 2706-2718.

[17] ZHU H Y, WANG R, JIN Y C, et al. PIVODL: privacy-preserving vertical federated learning over distributed labels[J]. IEEE Transactions on Artificial Intelligence, 2023, 4(5): 988-1001.

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

[19] FAN X, WANG Y, HUO Y, et al. BEV-SGD: best effort voting SGD against Byzantine attacks for analog-aggregation-based federated learning over the air[J]. IEEE Internet of Things Journal, 2022, 9(19): 18946-18959.

[20] DATA D, DIGGAVI S N. Byzantine-resilient high-dimensional federated

learning[J]. IEEE Transactions on Information Theory, 2023, 69(10): 6639-6670.

[21] HUANG S M, ZHOU Y, WANG T, et al. Byzantine-resilient federated machine learning via over-the-air computation[C]//Proceedings of the 2021 IEEE International Conference on Communications Workshops (ICC Workshops). Piscataway: IEEE Press, 2021: 1-6.

[22] 穆旭彤, 程珂, 宋安霄, 等. 抗拜占庭攻击的隐私保护联邦学习[J]. 计算机学报, 2024, 47(4): 842-861.
MU X T, CHENG K, SONG A X, et al. Privacy-preserving federated learning resistant to Byzantine attacks[J]. Chinese Journal of Computers, 2024, 47(4): 842-861.

[23] LYU L J. DP-SIGNSGD: when efficiency meets privacy and robustness[J]. arXiv Preprint, arXiv: 2105.04808, 2021.

[24] ZHU B, WANG L, PANG Q, et al. Byzantine-robust federated learning with optimal statistical rates and privacy guarantees[J]. arXiv Preprint, arXiv: 2205.11765, 2022.

[作者简介]

周由胜（1979– ），男，湖北利川人，博士，重庆邮电大学教授、博士生导师，主要研究方向为车联网安全、隐私计算、人工智能安全、网络攻防等。

高璟琨（1997– ），男，山东威海人，重庆邮电大学硕士生，主要研究方向为联邦学习、隐私计算等。

左祥建（1990– ），男，湖北监利人，博士，重庆邮电大学讲师、硕士生导师，主要研究方向为安全多方计算、数据安全与隐私保护。

刘媛妮（1982– ），女，河南邓州人，博士，重庆邮电大学教授、博士生导师，主要研究方向为移动群智感知、物联网安全、IP路由技术、复杂网络。