# 边缘计算中针对中毒攻击的差分私有联邦学习模型

## 摘要

联邦学习越来越受欢迎，因为它允许我们通过使用一个或多个数据所有者/来源的数据来训练全局模型来规避由于数据孤岛（例如，获取原始数据集的挑战，特别是当数据由多个数据所有者拥有时）造成的挑战。然而，在边缘计算部署中，不能假设资源受限的终端设备足够安全。换句话说，我们必须考虑这些设备被破坏的可能性，并可能被滥用以促进中毒攻击。隐私保护是处理终端设备上的敏感用户数据时要考虑的另一个重要属性。大多数现有方法只考虑防御中毒攻击或支持隐私，而不是同时考虑这两种属性。在本文中，我们提出了一种针对中毒攻击的差分私有联邦学习模型，专为边缘计算部署而设计。首先，我们设计了一种基于权重的算法对边缘节点终端设备上传的参数进行异常检测，仅使用小规模验证数据集提高检测率，并最大限度地减少通信成本。然后，利用差分隐私技术来保护边缘计算设置中数据和模型的隐私。我们还在随机和定制的恶意终端设备存在的情况下评估检测性能，并将其性能与IEEE Transactions on Departendable和Secure Computing发布的另外两种竞争方法的性能进行了比较，在攻击弹性和通信和计算成本方面。实验结果表明，我们的方案可以在安全性、效率和准确性之间实现最佳折衷。

## 1.概述

数据孤岛和隐私泄露是人工智能(AI)模型训练和应用的两个关键挑战。例如，基于联邦学习 (FL) [1]、[2]、[3]、[4]、[5] 的分布式机器学习方法可以克服局限性，并在电子医疗保健系统 [6] 和 VANET [7] 中被广泛采用。一般来说，FL 方法允许参与者根据本地收集的数据（例如，在他们自己的组织中）训练他们的本地模型，然后将参数传输到 FL 服务器。然后后者将执行统计融合，将聚合的参数返回给参与者，最终得到令人满意的全局模型。

不幸的是，集中式 FL 服务器容易受到单点故障的影响，并导致无法容忍的延迟，而许多用户需要在实时系统中反馈。因此，边缘计算中的 FL 将是令人信服的替代方案，其中每个边缘节点负责聚合来自其自身区域中用户的参数，并进一步将它们传输到 FL 服务器以进行最终聚合。一系列工作研究了FL在边缘计算中的应用情况，如决策、疾病诊断和财务趋势分析[8]、[9]、[10]、[11]、[12]、[13]、[14]。例如，利用分布式交通数据集在VANETs中进行实时导航，训练神经网络预测未来时间段的道路状况。路边单元 (RSU) 可以从配备在其邻域内行驶的车辆上的船上单元 (OBU) 收集和聚合本地训练的参数，车辆云将从所有 RSU 计算全局参数。

### 1.1问题背景

尽管联邦学习方法在边缘计算部署方面很受欢迎，但仍有几个具有挑战性的安全和隐私威胁。首先，由于边缘计算中的数据源，资源受限的物联网设备容易受到攻击和中毒攻击，恶意参与者打算通过污染训练数据集来操纵机器学习模型的预测。第二个担忧与隐私披露有关。虽然数据集没有直接传输，但[15]等研究成果表明，通过参与者的更新参数可以提取几乎相同的模型，从而获得局部数据集的统计特性。换句话说，直接使用联邦学习技术无法为诚实设备上的用户敏感数据提供足够强大的隐私保护。同时，全局模型还需要在多个场景中得到很好的保护，例如交易秘密（即知识产权），并且它直接暴露于参与者和开放渠道可能对模型所有者具有显着的财务、声誉和/或法律影响。

虽然文献中已经提出了许多解决方案，但观察到了几个限制。例如，Tolpegin等人[16]利用主成分分析(PCA)等降维技术来识别恶意更新。不幸的是，它在非 IID 设置中效果不佳，检测效果受参与者数量的影响。此外，它不能提供任何隐私保护。Zhao[17]提出了一种通过客户端交叉验证检测异常更新的方案，并结合差分隐私来保护敏感数据。然而，检测方法成本高昂，因此不适合部署在边缘节点上。此外，他们假设连接的设备不会相互勾结。

### 1.2主要贡献

为了解决具有挑战性的安全和隐私威胁，本文提出了一种针对边缘计算中毒攻击的差分私有联邦学习模型。具体来说，我们提出的方法旨在支持对端边缘设备（即确保模型精度）上传的参数进行异常检测，并保护终端设备和整体模型中敏感数据的隐私。为了抵抗恶意设备的串通攻击，我们使用了一个小的验证数据集，并动态调整添加的噪声规模，以保护数据隐私不被泄露。我们提出的方法总结如下。

* 我们为边缘网络设置设计了一个安全且保护隐私的联邦学习模型，其中终端设备联合训练准确的神经网络模型，而不会泄露数据中毒攻击下的敏感数据。
* 所提出的模型可以通过基于权重的检测方案抵抗中毒攻击。我们的方案使边缘节点能够使用小型验证数据集检测和过滤恶意终端设备上传的异常参数。根据检测结果，边缘节点为参数设置适当的权重，以消除虚假参数对模型的影响。因此，它确保联邦学习能够在边缘网络设置中抵抗中毒攻击。
* 所提出的模型可以在诚实设备上保护用户敏感数据的隐私和边缘计算设置中的整体模型。改进了边缘网络中联邦学习的传统差分隐私技术。具体来说，在我们的方案中动态添加噪声。只有当终端设备添加的当前噪声不满足隐私要求时，边缘节点和云服务器才会添加额外的噪声，从而最大限度地减少模型的干扰，提高精度。
* 我们在随机和定制的恶意终端设备存在的情况下在三个常见数据集上评估我们的方案。结果表明，与攻击弹性、通信成本和计算成本方面的先前工作相比，我们的方案能够以高概率检测异常参数，并在安全性、效率和准确性之间实现最佳权衡。例如，我们的方案在准确性方面平均提高了 12%，以及减少了2-10倍的计算小号。

在下一节中，我们将简要回顾相关文献。

## 2.相关工作

### 2.1防御投毒攻击

目前，有许多防御机制来抵抗中毒攻击。Nelson等人[18]提出了RONI方案来检测有毒样本点。其核心思想是使用可信数据集训练标准模型，然后将训练好的模型与可信数据集和不可信数据集的并集进行比较，从而筛选出有毒的样本点。然而，这种方法需要基于不同的训练数据集进行多次训练和比较，这需要大量的资源和时间，不适合边缘计算中的联邦学习。Baracaldo等人[19]使用用户身份和位置等元信息来检测训练集中的恶意样本点。该方案将训练数据集划分为多个不相交的子集，基于每个子集训练模型，然后比较性能，以识别异常训练数据子集。然而，这种检测方法需要提供数据样本点出现的设备的附加信息，包括用户ID、防火墙版本等。一旦这些信息泄露，会造成严重的隐私威胁，检测结果将暴露设备的安全级别。Liu等人[20]通过整合改进的鲁棒低秩矩阵逼近和鲁棒主成分回归，提出了一种有效的防御技术。上述中毒防御方法主要针对数据中毒。在联邦学习中，针对模型中毒的防御措施也很重要。

为了抵抗模型中毒攻击，Blanchardet等人[21]提出了一种拜占庭容忍分布式随机梯度下降算法Krum，该算法通过消除距离太远的向量来避免拜占庭敌人的攻击。为了测量向量之间的距离，Krum 计算所有其他向量的向量 Vi 和 V jðj 6 1/4iÞ 之间的两个范数距离，然后选择最小 n f 2 范数距离之和作为其分数，最后选择得分最高的那个进行聚合。作者证明，对于 n d 维向量集，Krum 的复杂度是 Oðn2 dÞ。有许多设备连接到边缘网络中的边缘节点。Krum 将消耗大量的计算资源，也会导致较大的延迟。Krum 基于拜占庭敌人的假设，即恶意参与者的数量小于 33%，这在边缘计算中无法确认。FoolsGold [22] 可以根据任意数量的敌人的贡献相似性抵抗巫攻击。然而，对于单个攻击者，相似性检测的方式不能发挥有效的作用，对于边缘网络中的各种攻击，FoolsGold 也很难处理。此外，[23]在基于自动编码器的联邦学习中执行异常检测，并使用预训练的自动编码器模型根据模型参数的重构误差率计算聚合权重。然而，由于对预训练模型的依赖，这种方法很难实现广泛的检测应用。Tolpegin 等人。[16] 使用降维技术，例如 PCA 来识别恶意更新。在他们方案的每一轮训练中，聚合器计算参与者更新的模型与全局模型之间的差异，并通过 PCA 提取参数的相关特征。但它在非 IID 设置中效果不佳，检测效果受参与者数量的影响。此外，它不能提供任何隐私保护。Zhao等人[17]专注于精度检测。在他们的方法中，检测任务被交给其他参与者以促进检测，参数的真实性由投票决定，投票用作设置权重的重要指标。然而，通信成本很高，该方法需要收集参与者的附加信息，这显然会导致通信成本和隐私影响。

### 2.2防止隐私泄露

加密原语[24]，[25]，[26]，[27]，[28]，[29]，[30]，[31]，[32]和差分隐私[33]，[34]，[35]，[36]，[37]，[38]，[39]可用于加密数据和参数，以保护模型训练过程的隐私。例如，Hesamifard 等人。 [24] 开发了 CryptoDL，它使用近似多项式而不是原始激活函数来训练卷积神经网络，并基于密文数据实现训练过程。然而，这个过程会产生显着的通信和计算开销，在联邦学习场景中，终端设备不太可能拥有足够大的计算能力。Bonawitz等人[25]设计了一种基于Shamir秘密共享的聚合方案。由于该方法旨在保证在存在诚实和好奇的服务器的情况下更新参数的安全性，因此它不会防止串通攻击。Aono等人[26]提出了一种使用同态加密的深度学习系统，以防御诚实和好奇的服务器，而不会损失准确性。然而，在大规模数据分析过程中，模型训练需要多次迭代训练。加密过程所需的通信资源增加了网络的负担，降低了每个组件的计算效率和性能。

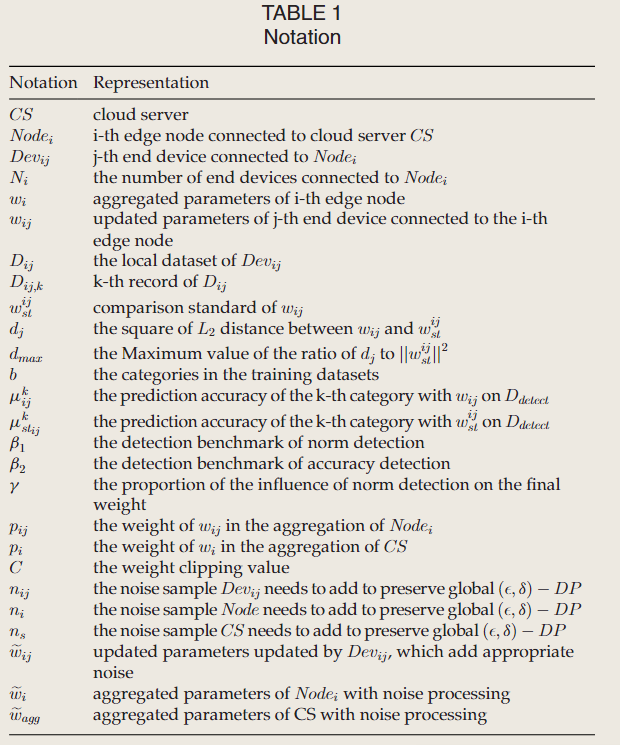
Geyer等人[33]提出了一种差分隐私随机梯度下降算法，该算法将数据样本随机划分为小块，在聚合过程中加入高斯噪声，实现差分隐私保护。然而，直接在训练数据上使用差分隐私技术会添加太多的噪声，从而降低整体模型的准确性。Matthew等人[34]提出了一种新的局部差分隐私(LDP)技术来实现联邦学习中的隐私保护。将差分隐私技术应用于本地数据集，在本地实现隐私保护。LDP实现了更严格的隐私定义。因为它更符合分布式场景，LDP已广泛应用于联邦学习场景。Sunet等人[35]提出了LDP-FL来分割和破坏参与者上传的模型参数，增加数据保护，减少引入噪声的方差。Truex等人[36]给出了局部多轮迭代的正式差分隐私保证，并实现了一系列选择和过滤技术来破坏参数服务器的选择。魏等人。 [37] 提出了一种基于 Global-DP 概念的框架，在该框架中，服务器和客户端都添加了适当的噪声，使得参数在上游通道和下游通道中都实现了差分隐私。提供了理论证明。然而，该方案只涉及服务器和客户端之间的直接交互，不适合边缘网络设置中的多级网络结构。

与上面提到的工作不同，在我们的方案中，边缘节点的可信检测数据集用于实现终端设备上传的参数异常检测。参数的贡献将根据检测结果动态调整，以抵抗中毒攻击对整体模型的侵权。因此，我们的方案非常适合边缘计算中的联邦学习，特别是对于多层网络。此外，时间和资源消耗大大减少，使该方案更加实用。最后，毒药防御和隐私保护的综合组合使得联邦学习在边缘计算中的训练过程更加安全。

## 3.解决方案

### 3.1系统模型

符号表示：



### 3.1.2系统框架

该方案的总体框架如图1所示。边缘网络主要包括(i)云服务器CS (ii)边缘节点节点i (iii)终端设备Devij。假设有m个边缘节点参与联邦学习训练，每个边缘节点Nodei连接Ni终端设备Dev ij (i=1，，m;j=1，;Ni)，其中有一个本地数据集Dij。基于边缘网络的联邦学习训练过程主要分为两个阶段。

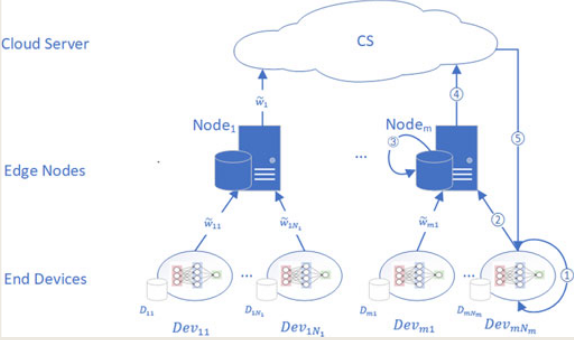


图1 系统交互流程图

1)预训练阶段。云服务器 CS 将网络超参数和随机初始化的模型参数发送到边缘节点和终端设备。此外，CS 需要将方案中的统一必要参数发送到相应的接收器，例如差分隐私预算、权重裁剪边界 C 等。

2)训练阶段。如图1所示，主要有以下步骤:

* 每个终端设备Dev ij基于局部训练集Dij训练局部模型，得到更新后的模型参数wij，并根据预设的隐私规则添加适当的噪声(详见4.2节)，最终得到噪声ew ij的更新参数。
* 终端设备 Dev ij 将处理后的模型参数 ew ij 发送到最近的边缘节点 Node i。
* 每个边缘节点节点 i 对参与者 Dev ijðj 1/4 1 上传的参数 ewij 执行异常检测；.; N iÞ 根据我们基于权重的方案中的预设检测规则（详见 4.1 节），并将它们聚合以获得局部聚合参数 wi，然后判断额外的噪声是否需要添加并根据预设的隐私规则确定噪声尺度（详见第 4.2 节），最终得到局部聚合参数 ewi。
* 边缘节点Nodei将本地聚合参数与噪声处理ewi发送到云服务器CS。
* 云服务器CS聚合边缘节点上传的局部聚合参数ewi，得到整体聚合参数wagg，然后根据预设的隐私规则判断额外噪声是否需要添加和确定噪声尺度(详见4.2节)，并将噪声处理ewagg的聚合参数返回给终端设备，开始下一轮迭代训练。

### 3.1.3假设条件

根据我们的方案的系统框架，我们做了以下两个假设：

1)假设边缘节点可以收集安全级别较高的数据作为检测数据集，因为这部分数据用于上传模型参数的异常检测，以减少不可信参数的贡献，保证模型的准确性。正如[9]中假设的那样，检测数据集不需要非常大，这可能来自一些(默认)本地用户，这些用户通过身份验证或安全验证，应该与训练数据集区分开来进行异常检测。

2)为了保证模型的收敛性，我们假设当每个终端设备训练局部模型时，损失函数为[37]也采用了凸函数。在这种情况下，我们可以通过梯度下降算法逐步获得模型最准确的信息。

### 3.2威胁模型

#### 3.2.1威胁模型

我们提出了一种边缘网络下的联邦学习训练模型，主要面临以下两个安全威胁。

* 中毒攻击：中毒攻击涉及训练中的恶意参与者或重新训练过程中污染训练数据集来操纵机器学习模型的预测 [40]。在联邦学习中，攻击者通常有两种方法来执行中毒攻击：数据中毒和模型中毒。数据中毒意味着攻击者污染训练集中的样本，例如添加错误的标签或有偏差的数据以降低数据的质量，从而影响最终训练的局部模型。模型中毒意味着恶意参与者发送错误的参数或损坏的模型，以在全局聚合[41]期间破坏学习过程。在两种中毒方法中，数据中毒更难被预防和检测，模型中毒显然对整体模型有更大的攻击影响。因此，该解决方案主要通过检测参与者的异常上传参数来抵抗中毒攻击。
* 隐私泄露:联邦学习过程中的隐私安全通常容易受到两种类型的攻击:模型提取攻击和模型反向攻击[40]。通过模型提取攻击，攻击者试图窃取模型的参数和超参数，破坏模型的机密性，侵犯模型所有者的隐私安全。通过模型反向攻击，攻击者试图从参与者上传的参数信息中获取训练数据集的私人信息，从而获得用户的私人信息，例如成员是否包含在训练过程中，或训练数据集的一些统计特征。为了保护模型隐私和安全和敏感数据，而不会造成过多的通信和计算消耗，该解决方案应用差分隐私技术将适当的噪声添加到上行链路和下行链路信道中的参数中，以防止中间参数的多次迭代造成的隐私泄露，并确保模型在培训和应用过程中的隐私和安全性。

#### 3.2.2潜在的对手

整个过程中有四个主要的潜在对手或攻击者：

* 恶意终端设备：物联网中有无数的设备，其安全级别较低。该方案考虑了恶意终端设备的存在，将虚假参数发送给边缘节点。这些参数要么是随机生成的，要么是故意处理的，旨在破坏整体模型的收敛性能，要么是导致模型对某些类型的数据样本进行错误分类。
* 诚实和好奇的边缘节点：该方案认为边缘节点是半可信的，即在遵守协议的前提下，可以获得尽可能多的附加信息。在这种情况下，如果诚实的终端设备直接上传本地参数，诚实和好奇的边缘节点可以通过模型推理攻击获得参与者训练数据集的统计特征或其他信息，从而破坏他们的隐私安全。
* 外部窃听者：外部窃听者只能通过监控通信通道来获取信息。如果对传输的参数没有采取保护措施，外部窃听者也可以获取参数信息，这可能会导致潜在的隐私泄露。
* 模型用户：在这个方案中，模型用户也被视为潜在的对手。通过连续预测分析，模型用户可以从中推断出整个模型。众所周知，整个模型通常需要机密性，因此这种信息泄漏会给模型所有者带来很大的损失，因此需要隐私保护。

## 4.详细描述

### 4.1异常检测规则

对于3.2节中描述的两种类型的中毒攻击，恶意节点可以采用以下攻击方法[41]。

* 可用性：恶意参与者会随机或其他方式生成有毒模型参数，并将它们上传到边缘节点，影响整体模型的准确性和收敛性能。
* 完整性：为了对特定类别的错误分类，恶意参与者可以通过修改数据集标签并根据数据集生成有毒且难以检测的模型参数来破坏模型的完整性。

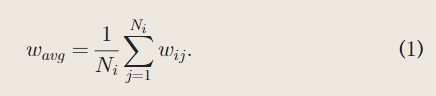
尽管有毒参数可能被发送到边缘节点，但由于诚实参与者的存在，影响是有限的，这将占多数。为了提高有毒参数的影响，一种常见的方法是扩展局部有毒模型参数。为了检测异常参数，该方案主要采用以下方法：

* 规范检测：基于水平联邦学习，每个参与者将是网络结构的统一模型和预训练阶段的初始模型参数。我们在本文中不讨论稀疏参数，每个参与者上传的默认参数是完整的参数数组。如果每个终端设备的训练数据是独立且同分布的，则参数差异应在一定范围内；如果来自每个终端设备的数据不遵循独立的共分布，则参数之间的差异将更大，但不会偏离标准太远。即使来自诚实设备的一些参数是独特的，我们也可以为它们独特的数据特征减少它们的权重。比较参数的范数并调整权重而不是直接聚合潜在的不可信参数将更合理。
* 精度检测:范数检测方法可以在一定程度上消除破坏模型可用性的异常参数，但在检测破坏模型完整性的参数方面不能发挥重要作用，因为对手会训练一个看起来尽可能“正常”的模型，除了一般某种类型的样本分类的误差。因此，该方案提供了第二种检测:精度检测。众所周知，边缘节点还将从数据生产者收集数据，这些数据生产者不用于模型训练。同时，由于边缘节点的安全级别较高，边缘节点直接收集的数据具有较高的可信度。因此，这部分数据非常适合准确性检测。

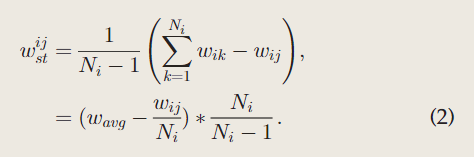
基于这两种检测方法，我们提出了基于权重的检测方案，该方案可以集成检测结果以确定每个参数的权重。

权重越小，相应参数的贡献或影响越小。暂时忽略噪声处理，边缘节点Nodei接收参数w ijðj 1/4 1;.; N iÞ 由每个终端设备 Dijðj 1/41;.; N iÞ 并计算平均聚合参数 w avg，如等式所示。

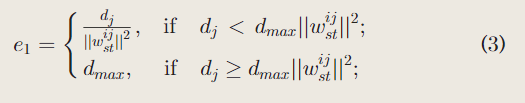
(1)。wavg计算的目的是降低计算每个参数的比较标准的复杂性



为了确定 wij 的权重，Node i 首先计算除 w ij 之外的所有参数的平均聚合，用作由 wij st 表示的比较标准，如等式所示。(2)

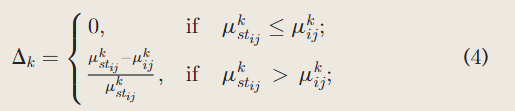


计算wij st的值很方便，即wij的比较标准，常数时间w avg。然后我们基于wij st和wij之间的差异对参数w ij进行异常检测。异常检测主要结合了两种检测方法:范数检测和精度检测。1)执行范数检测。对于终端设备Devij加载的参数wijup，Nodei计算其L2距离与比较标准dj 1/4 jjwij stj2ðj 1/4 1;。; N iÞ。定义检测基准b1，如果差异dj b1jwij stj2，则wij的权重，记为p ij不受影响;如果差异dj > b1jwij stjj2, pij需要减小，偏差越大，p ij越小。因此，我们定义参考值 e1 1/4 dijw ij stjj2。请注意，e1 的值范围为 1/20； þ1푡。在计算 e1 之前，我们需要限制值范围。假设在合理的范围内，w ij st 和 w ij 之间的 L2 距离的平方不超过 jjwij stj2 的 dmax 倍，那么我们有 Eqn。(3)



然后将参考值e1与检测基准b1进行比较，得到速率范数ij 1/4 1 maxð0;e1 b1Þ。2)执行精度检测。边缘节点Nodei使用本地收集的数据集D i检测进行精度检测。类似地，将每个参数wij与比较标准wij st进行比较。假设训练数据集中有 b 个类别。以wij st为模型参数，对di进行精度检验，得到各类别m1stij的预测精度;m2stij;.;mb stij;以wij为模型参数，对Di检测进行精度检验，得到各类别m1ij的预测精度;m2ij;.;mb ij。

考虑使用mk stij mk ijmk stijðk 1/4 1;2;。; bÞ 来衡量每个类别的准确度差异，记录为 Dkðk 1/4 1; 2;.; bÞ。为了防止除除数 0 外，在计算 Dk 之前判断 mk st ij 的值。当 mk st ij mk ij 时，对于类别 k，参数 w ij 与比较标准 mk st ij 具有相同的或更好的预测性能，因此 Dk 1/4 0；当 mk st ij > mk ij 时，必须有 mk st ij 6 1/40，Dk 计算为 Dk 1/4 mk st ijmk st ij1/4 1 mk ijmk st ijmk st ijmk st ij。此时，由于 mk st ij &gt; mk ij ，Dk 的值范围为 [0,1]，在合理的范围内。根据上述分析，对于每个类别 kðk 1/4 1; 2;.; bÞ，计算 Dkðk 1/4 1; 2;.; bÞ 根据方程式。(4)



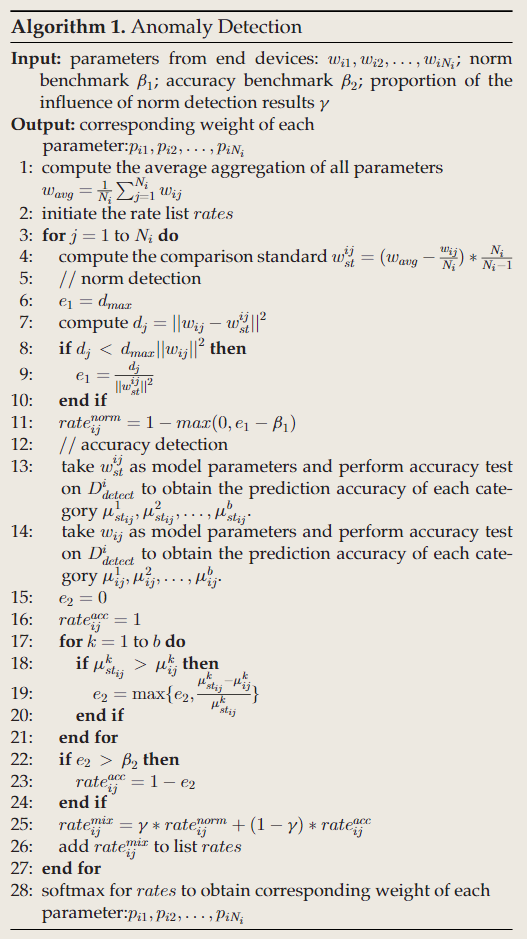
根据上述计算，如果Dk等于零，即对于检测数据集中类别k的样本，wij至少有相同的预测性能，因此，我们不应该改变该参数的原始权重;如果Dk大于0，则与wij st相比，wij在D i检测上的准确率较低。在这种情况下，需要进一步的判断。我们定义参考值 e2 1/4 maxk21/2b푡Dk。e2 表示所有类别的 wij st 和 wij 之间的准确度差异最大。作为聚合的结果，必须有一些参数，其准确率可能低于聚合后的准确率。因此，需要定义检测基准b2。当 e2 大于 b2 时，我们有rateacc ij 1/4 1 e2；否则，rateacc ij 1/4 1。

我们定义g来描述范数检测在混合检测中的影响程度。如果 ratenorm ij 、rateacc ij 、ratemix ij 分别用于描述分别通过范数检测、准确度检测和混合检测后参数的可靠性，则 ratemix ij 1/4 g ratenorm ij þ ð1 gÞ rateacc ij ，速率应与最终聚合中参数的权重成正比。考虑到权重之和为 1 的特征，所有速率都可以由 softmax 统一处理，最终得到 pi1； pi2；.; piN i ，分别表示聚合中 wij 的权重。

合理性分析。对于范数检测的一部分，速率范数 ij 1/4 1 maxð0； e1 b1Þ。当 e1 1/4 jjw ij w ij stjj2jjw ij stj2 不超过基准 b1 时，wij 的最终权重不需要修改速率范数 ij 1/4 1，即成功传递检测；当 e1 大于 b1 的边界时，速率范数 ij < 1，权重将减少，距离越远，权重减少。

对于精度检测的一部分，如果上传参数wij预测某一类别k0 ðk0 2 1/21的样本的精度;D i检测数据集mk0ij上的b푡Þ明显低于与其他类别相比的比较标准wij st，则e2 1/4 maxk21/2b푡Dk 1/4 Dk0，如果e2 > b2,rate acc ij 1/4 1 e2 < 1，差异越大，rateacc ij越小。相应地，如果对于所有类别kðk 2 1/21;b푡Þ，有Dk < b2，则e2 < b2，则该参数的权重不受精度检测、速率acc ij 1/4 1的影响。

该方案结合了两种检测算法的优点，形成混合检测算法。通过g的调整，混合检测算法可以对不同的敌人发挥更好的防御作用。当g接近0时，范数检测的比例较大，对随机中毒攻击的防御效果更好;当g接近1时，准确率检测的比例较大，专业中毒攻击的检测效果较好。对于其他攻击，混合检测有望很好地工作。在训练开始时，每个参数的准确率相对较低，差异很小，因此范数检测将作为基本滤波器，控制一些异常参数的权重；随着迭代次数的增加，准确率检测结果的距离将成为主要判断因素，提高整个检测效果。



我们观察到，当攻击者打算逃避检测时，我们的方案也可以有效地处理这种情况。例如，为了抵抗攻击者想要使用对抗样本干扰检测的中毒攻击，已经有一系列令人信服的解决方案，例如 [42] 和 [43] 来解决这个具有挑战性的问题。具体来说，根据[42]，服务器可以创建额外的预测模型来推断某个梯度属于哪个参与者，这有助于通过比较局部梯度之间的差异来识别有毒的本地更新。[43] 使用 GAN 构建了一个模拟模型来模拟由干净样本训练的目标模型的行为。然后服务器能够通过比较模拟模型和目标模型之间的预测差异来区分中毒样本和干净样本。尽管如何处理对抗样本的中毒攻击超出了这项工作，但幸运的是，我们发现 [42] 和 [43] 中提出的这两种技术都可以应用于我们的建议，以使其对旨在逃避检测的攻击者更加稳健。

### 4.2隐私规则

该方案采用差分隐私技术来保护用户数据的隐私安全性和整体模型。

定义1:((Adjacent Databases[44]):如果两个数据库x, y，则jx yjj1 1，则数据库x和y称为相邻数据库。

相邻数据库之间最多有一条记录差异。我们的方案旨在保护用户数据的隐私，因此本文中的相邻数据库最多相差一个数据样本。

定义2:(;dÞ DP[44]):随机机制M: X !R域X和范围R满足ð;dÞ DP;如果所有可测集SR和任意两个相邻数据库Di;D 0i 2 X



该方案主要添加高斯噪声来满足差分隐私。下面介绍了隐私敏感性。

定义3:((l2灵敏度[44]):函数f: ðN !RkÞ定义为D2f 1/4 maxx;y2njX j kfðxÞ fðyÞk2，其中x和y是相邻的数据库。

灵敏度包括全局灵敏度和局部灵敏度。局部灵敏度是函数对给定数据库D及其相邻数据库D0的敏感性，记录为Df D;全局灵敏度是函数f在任何相邻数据库上的灵敏度，记录为Df。两者之间的关系可以表示为Df 1/4 maxDDf D。

定义4:((高斯机制[44]):设inð0;1Þ，对于任何函数f，定义算法M 1/4 fðxÞ þ n，其中n服从高斯分布n N ð0;s2Þ。当参数 s cD2f 时，算法 M 1/4 fðxÞ þ n 满足ð ; dÞ 差分隐私，其中常数 C 满足 c2 > 2lnð1:25=dÞ。

使用高斯机制，我们可以通过添加适当的噪声来确保算法满足差分隐私。然而，在系统中，我们需要保证该方案在多个算法的联合作用下满足全局差分隐私。下面介绍了差分隐私的组合定理。

定理1:(DP串行定理[36]):设随机机制Mi: X !Ri满足所有i 2 1/2k푡的i DP。那么如果 M 1/2k푡 : NjX j ！ Qk i1/41 Ri 定义为 M 1/2k푡 ðxÞ 1/4 ðM1ðxÞ; M2ðxÞ;.; MkðxÞÞÞ，那么 M 1/2k푡 是 Pk i1/41 i 差分私有的。

定理2:(DP并行定理[36]):设随机机制Mi: X !Ri满足所有i 2 1/2k푡的i DP。如果数据库 x 被划分为 k 个不相交的子集，即 x 1/4S k i1/41xi ; xiT xj 1/4。ð8i 6 1/4 jÞ。然后，如果 M1/2k푡 : NjX j ！ Qk i1/41 Ri 定义为 M 1/2k푡 ðxÞ 1/4 ðM1ðx1Þ; M2ðx2Þ;.; Mkðx kÞÞ ismaxk i1/41 i 差分私有。

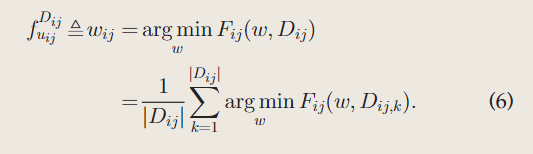
为了实现联邦学习中的差分隐私保护，NbAFL [37] 可以满足全局 ð ； dÞ 通过将适当的噪声添加到不同的实体并证明了模型的收敛性来差分隐私。然而，主要存在两个缺点：1）NbAFL 只考虑联邦学习中服务器和客户端之间的交互，不再适用于多层网络结构； 2）NbAFL假设参与联邦学习的所有设备的训练数据集相等，因此平均聚合算法用于模型聚合。事实上，考虑到恶意参与者的存在和检测方案，训练集的大小可能不同，聚合中的权重不会相同。我们开发了NbAFL来克服缺陷。首先，我们展示了如何以三层网络为例，将DP机制扩展到多层网络中，然后根据聚合中的权重动态调整噪声尺度，为设备中的敏感数据提供足够的隐私保障。

如第 3.3 节所述，假设有 m 个边缘节点参与训练，每个边缘节点节点节点 i 连接 Ni 终端设备 Dev ij，其中有一个本地数据集 D ij 和 Dij;k 表示数据集 D ij 中的第 k 个记录。

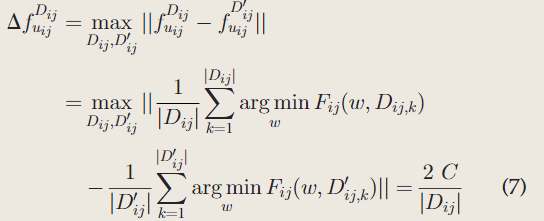
为了限制隐私敏感性的上限，我们定义了裁剪值 C。假设Devij上传的参数已被切割，jjwijj Cði 1/4 1;.; m; j 1/4 1;.; N iÞ，因此我们有 jjwij w0ijj 2 Cði 1/4 1;.; m; j 1/4 1;.; N iÞ。

假设迭代次数为 T，由 DP 串行定理，为了满足全局 DP，设第 t 次迭代满足 t DP，其中 t 1/4 1； 2；.; T ，我们需要确认 PT t1/41 t 1/4。我们平等地划分隐私预算，因此每次迭代的隐私预算为 T。

在边缘节点聚合参与者参数的过程中，假设Devijis Fijð Þ中定义的损失函数，则其训练过程表示如下:



从Dev ij到Node i上行链路的灵敏度可以表示为



从Dev ij到Nodei上行的全局灵敏度可以定义为



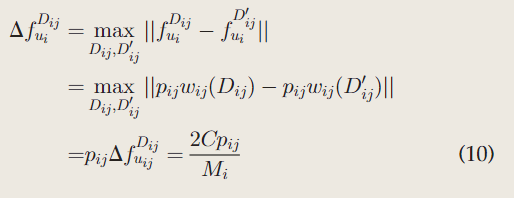
将节点i的局部数据集的最小大小设置为Mi，得到Df uij 1/4 2 C Mi。

为了保证全局ð;dÞ-DP, Dev ij需要添加噪声n ij，满足高斯分布N ð0;s2ijÞ，其中sij 1/4 cT Df u ij。

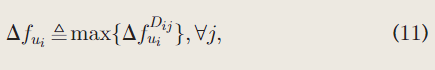
在云服务聚合边缘节点参数的过程中，Nodei的聚合过程可以定义为



从Nodei到云服务器CS上行链路的灵敏度可以表示为



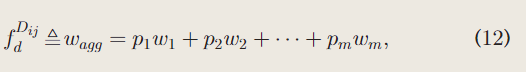
从节点i到云服务器CS的上行链路的全局灵敏度可以定义为



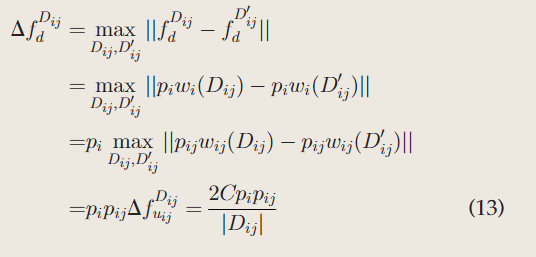
得到Df ui 1/4 2CmaxjfpijgMi。

为了保证全局 ð ; dÞ-DP，Nodei 在将聚合参数 wi 上传到云服务器 CS 之前需要添加噪声 ni，满足高斯分布 N ð0； s2i Þ，其中 si 1/4 cT Df ui。

在云服务器将聚合参数传输到Devij的过程中，云服务器的聚合过程可以定义为



从云服务器CS到Devij的下行链路的灵敏度可以表示为



从云服务器CS到Devij的下行链路的全局灵敏度可以定义为

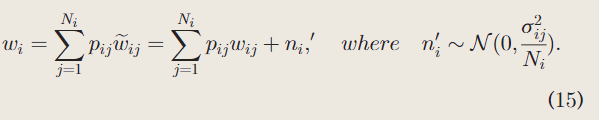


设置M 1/4 minifMig，得到Df d 1/4 2Cmaxi;jfpi pijgM。

为了保证全局 ð ; dÞ-DP，云服务器 CS 在将聚合参数 w ag 返回给终端设备之前需要添加噪声 ns，满足高斯分布 N ð0； s2s Þ，其中 ss 1/4 cT Df d。

为了在整体方案满足差分隐私的前提下尽可能地减少噪声方差，我们进一步分析了每个阶段的噪声尺度，从而计算终端设备、边缘节点和云服务器要添加的噪声方差。假设终端设备中数据集的最小数据集大小为 M，即 M 1/4 M i ; 8i。

1)连接到边缘节点节点i的每个终端设备Devij添加sij 1/4 cLDf uij的噪声nij，其中Df u ij 1/4 2 cm，因此Dev ij上传的参数为ewij 1/4 wij þ n ij。节点i聚合这些参数，得到

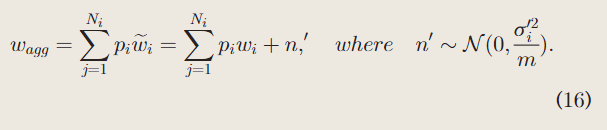


为了保证最终噪声 ni 满足高斯分布 N ð0； s2i Þ，我们需要将 s2ij N i 与 s2i 进行比较。

(i) s2ij N i s2i ，即 N i1/2maxfpijg푡2 1，则 Nodei 不需要额外的噪声；

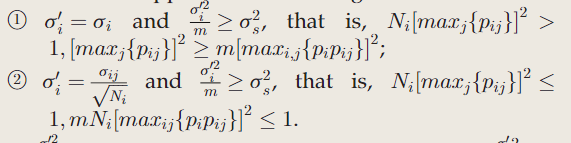
(ii) s2ij N i < s2i，即N i1/2max jfpijg푡2 > 1，则Nodei必须添加额外的噪声bni N ð0;s2i s2ij Ni Þ。

2)每个边缘节点Nodei必须在将聚合参数上传到云服务器CS之前添加足够的噪声。上传的参数可以表示为ew i 1/4wi þ n0i，其中s0i 1/4 maxfsi;sijffiffiffiNip g。云服务器CS聚合这些参数以获得

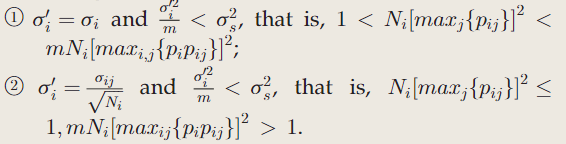


为了保证最终噪声ns满足高斯分布N ð0;s2s Þ，我们需要将s02i m与s2s进行比较。

i) s02i m s2s ，那么 CS 不需要添加额外的噪声。这将在以下情况下发生：



ii) s02i m < s2s ，CS 必须添加噪声 bn s N ð0； s2s s s0i2m Þ 以满足全局 ð ； dÞ-DP。这将在以下情况下发生：



隐私分析。在每一轮迭代中，datasetD主要在训练过程中经过三个过程:Dev ij在本地数据集Dij上训练模型，边缘节点节点i聚合来自Dev ijðj 1/4 1的参数;.; NiÞ 和 CS 从 Node iði 1/4 1; 2;.;米Þ。

在Dev ij的局部训练过程中，添加了尺度为ij 1/4 cT Df uij的噪声。通过高斯机制，该过程满足 ð T ; dT Þ 差分隐私。每个终端设备的训练数据彼此不相交，训练过程相互独立，满足差分隐私并行组合定理。因此，该过程满足 ð T ; dT Þ D 上的差分隐私。同样，Nodeiði 1/4 1 中的过程； 2;.; mÞ 和 CS 满足 ð T ; dT Þ D 上的差分隐私。因此，在一次迭代中，整个训练过程满足ð T;dT Þ D上的差分隐私。由于总共有T次迭代，整个训练过程满足ð;dÞ差分隐私串行组合定理对D的差分隐私。

## 5.实验

### 5.1实验环境

在本节中，我们经验地讨论了参数对整个模型精度的影响，并展示了检测方法在计算成本和通信成本方面的效果。所有实验均在具有运行在 3.00 GHz 处理器和 NVIDIA Tesla P40 GPU 的 Intel Xeon Silver 4114 CPU 的机器上进行。

网络。根据第 3 节，我们构建了一个三层网络来模拟边缘计算框架下的训练过程，该框架由 (i) 云服务器 CS (ii) 边缘节点 (iii) 终端设备组成。假设有5个边缘节点参与联邦学习训练，每个边缘节点Nodeiði 1/4 1;.; 5Þ 连接 5 个终端设备。

数据集。我们使用 MNIST [45]、Fashion-MNIST [46] 和 CIFAR-10 [47] 进行实验，以评估我们的建议在不同复杂度的典型数据集上的性能。MNIST 和 Fashion-MNIST 由 60,000 个训练样本和 10,000 个测试样本组成。每个样本都是一个 28 28 大小的灰度图像。CIFAR-10 包含 10 个对象类中大小为 32 32的 60,000 个彩色图像。从上述网络中，10%的训练样本将构成检测数据集Ddetect，其余样本在Devijði 1/4 1之间平均分配;.; 5; j 1/4 1;.; 5Þ。

模型。对于 MNIST 和 Fashion-MNIST，我们使用一个简单的三层神经网络，具有 784 单元输入层、128 单元隐藏层和 10 单元输出层，以及 10 个具有 MSE 损失的类的 softmax。对于 CIFAR-10，我们使用 ResNet20 模型。

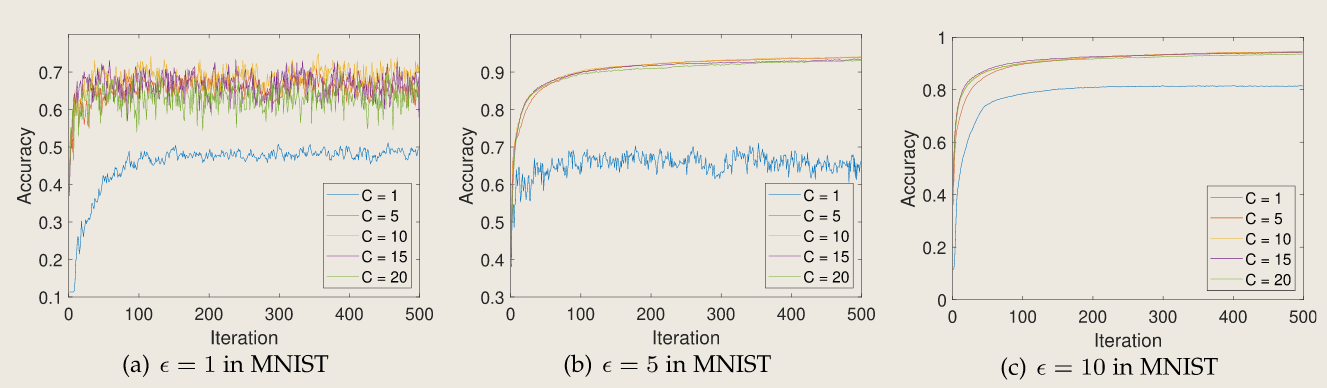
系统参数。我们设置 d 1/4 0:01，学习率等于 1e-1，M 等于 Dev ij 中的数据集的大小。在实验中，我们将 dj 与 jjw ij stj2 的比率的最大值设置为 9。

### 5.2实验结果

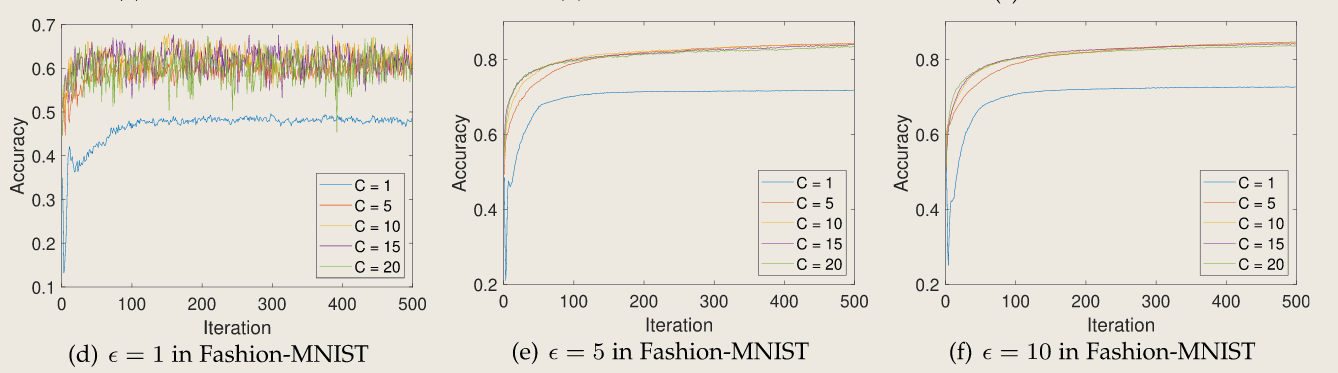
#### 5.2.1准确性

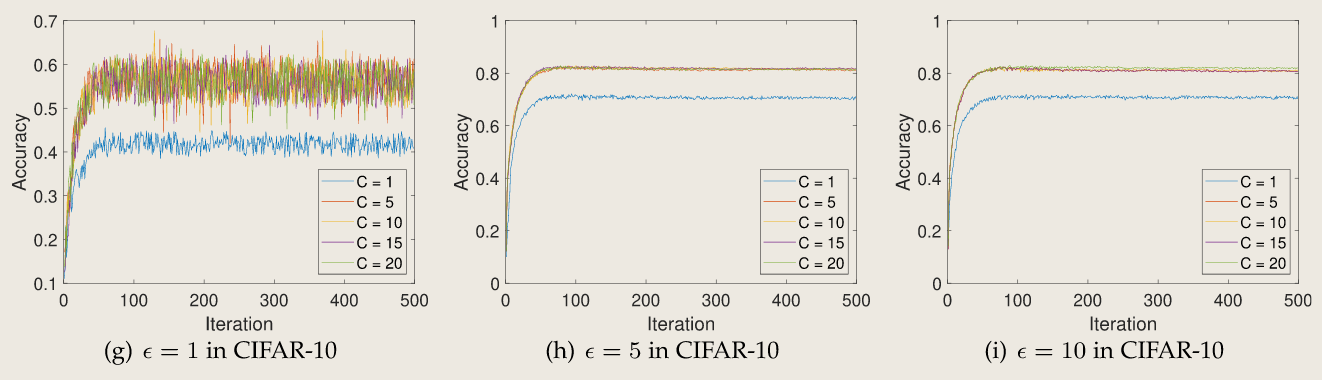
在本节中，我们进行了一系列实验来评估参数的系统性能。

隐私预算的影响。为了评估差分隐私的影响，为了方便起见，我们将噪声方差设置为0.01，并设置1/4 1;1/4 5;1/4 10，MNIST的结果如图2a、2b和2c所示。我们可以发现，随着隐私预算的增加，模型的准确性也会增加。结果与我们预测的结果相同。随着缩小，隐私保护级别变得更高，这导致准确性的牺牲更大。Fashion-MNIST 和 CIFAR-10 的实验显示出相同的趋势。为了在隐私和性能之间取得更好的权衡，我们在实验中设置为 5。

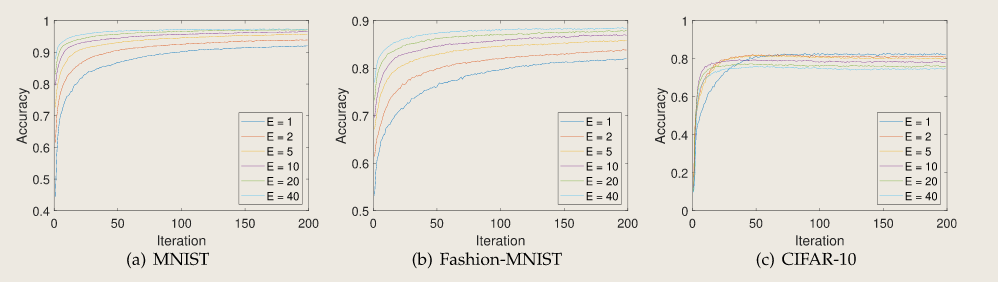


剪裁值 C 的影响。裁剪值 C 的引入为定义训练过程中灵敏度提供了重要的基础。但是，C 的大小会对模型的收敛性和性能产生一定影响。我们研究了剪裁值C对不同隐私预算条件的影响。图 2a、2b、2c 显示了隐私预算为 1/4 1 时全局模型的准确性； 1/4 5 和 1/4 10，裁剪值 C 1/4 1； C 1/4 5； C 1/4 10； MNIST 的 C 1/4 15 和 C 1/4 20。根据实验结果，可以看出C的裁剪值在不同的隐私预算条件下具有相似的影响。当 C 1/4 1 时，全局模型的准确性受到显着影响。这是因为 C 的值太小，这将导致模型参数被大大剪切，改变原始梯度下降方向，破坏模型的预测性能。此外，从实验结果可以看出，当 C 1/4 10 时，全局模型的准确率略好于 C 1/4 5 时，这可能是由于参数的部分剪切造成的。当 C &gt;10 时，模型的精度随着 C 的增加而略有下降。这是因为本文根据差分隐私方案，终端设备添加的噪声标准差为sij 1/4 cT Df uij，其中df uij 1/4 2 cm。因此，当 C 增加时，终端设备会添加更多噪声，从而降低某些模型的准确性。Fashion-MNIST 和 CIFAR-10 的类似结果如图 2d、2e、2f、2g、2h 和 2i 所示。通过以上实验，我们可以看到模型在T 1/4 200轮中达到了收敛。我们设置C 1/4 10;T 1/4 200。





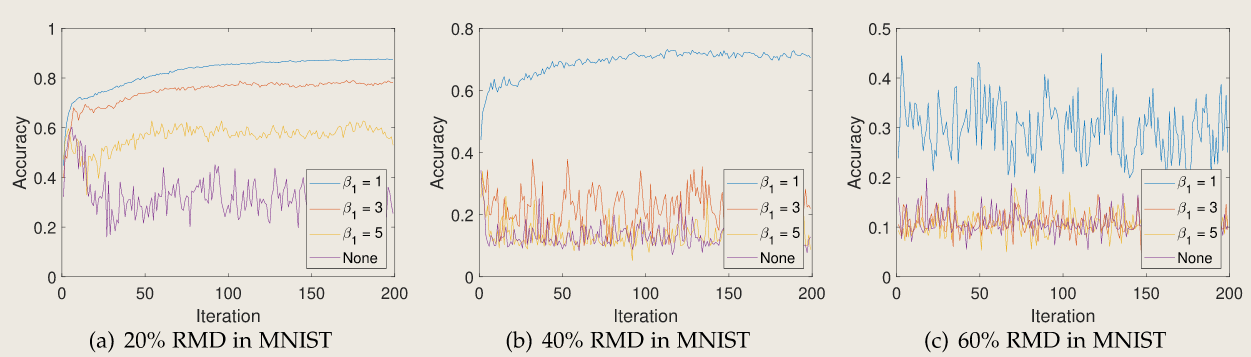
局部圆 E 的影响。本实验研究了局部轮数 E 的影响。我们将局部训练轮数设置为E 1/4 1;E 1/4 2;E 1/4 5;E 1/410;E 1/4 20和E 1/4 40。图3显示了实验结果。随着局部训练轮数的增加，模型的收敛速度加快，模型的精度逐渐增大。然而，对于 CIFAR-10，模型精度随着 E 的增加而略有下降，因为复杂数据集上过多的局部训练轮次会导致过度拟合。通过这个实验，可以证明异常检测方案和差分隐私方案对诚实参与者的参数聚合和模型表达没有影响。当参与训练的每条边节点诚实地训练局部模型并上传时，联邦学习训练过程可以正常执行，全局模型可以有效、准确地获得。

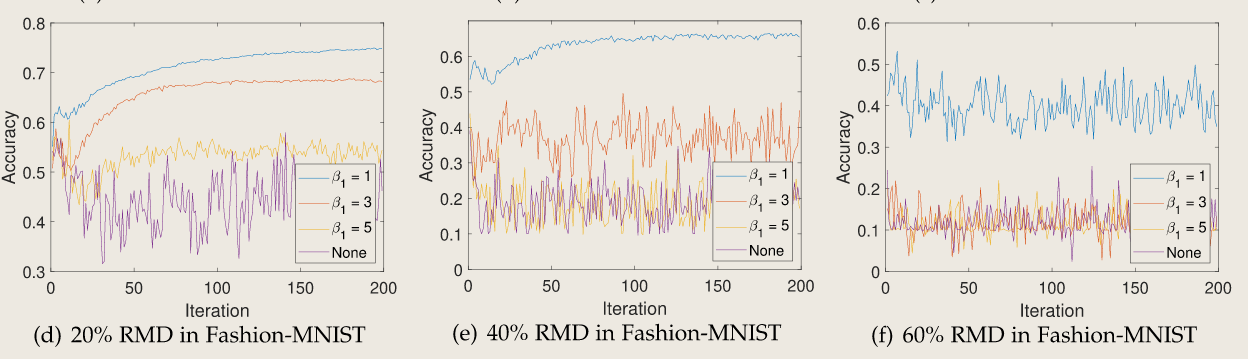


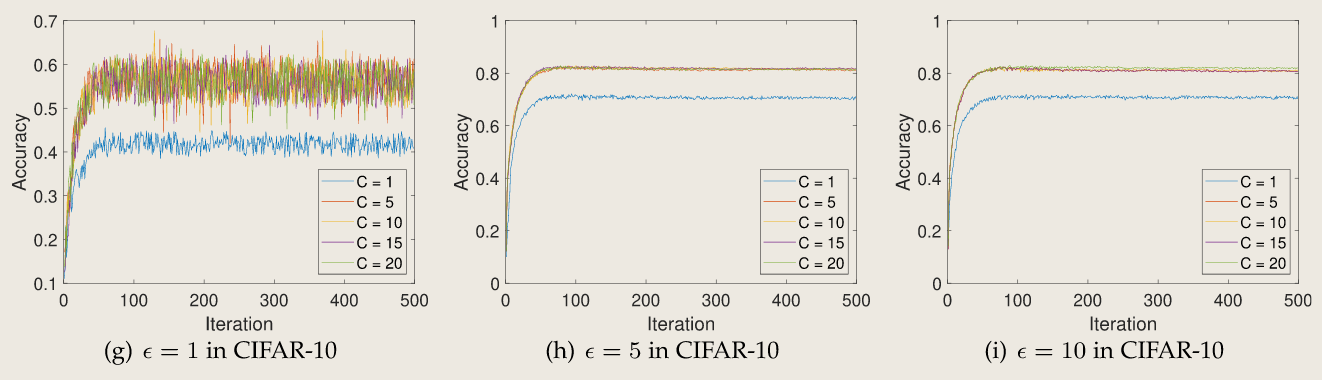
#### 5.2.2检测效率

我们对两种类型的恶意方进行了实验，以测试该方案的检测效果。

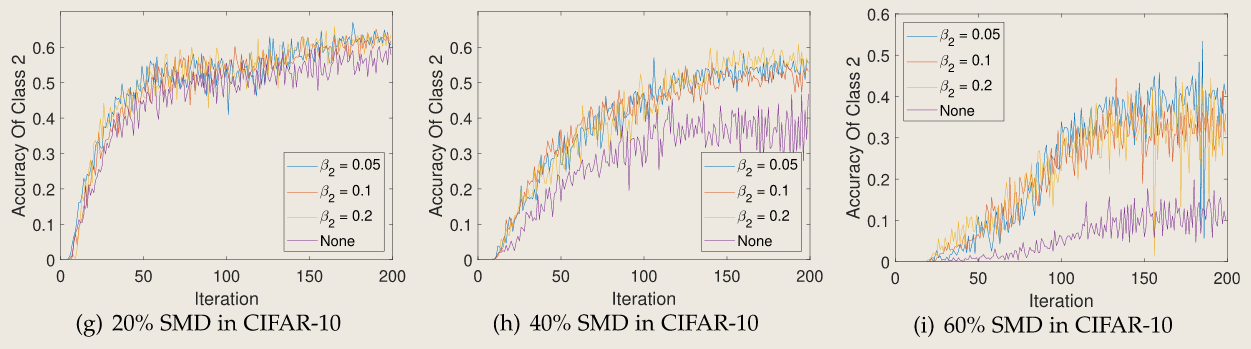
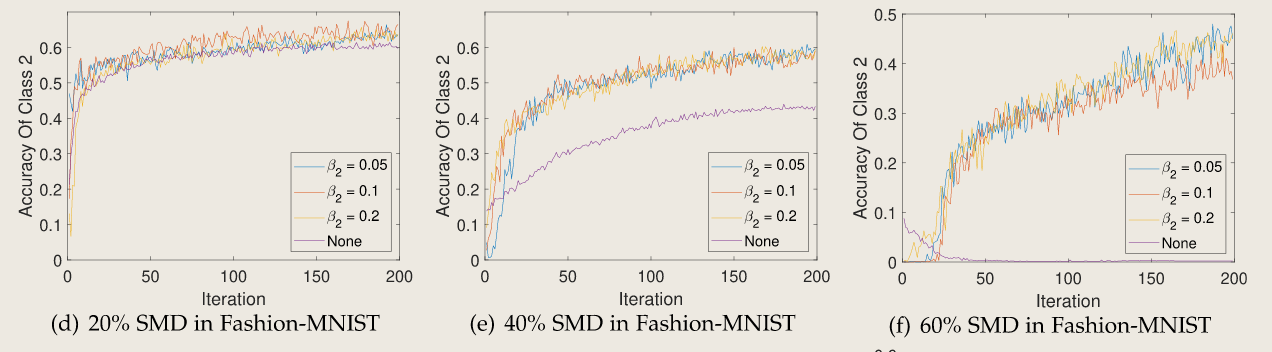
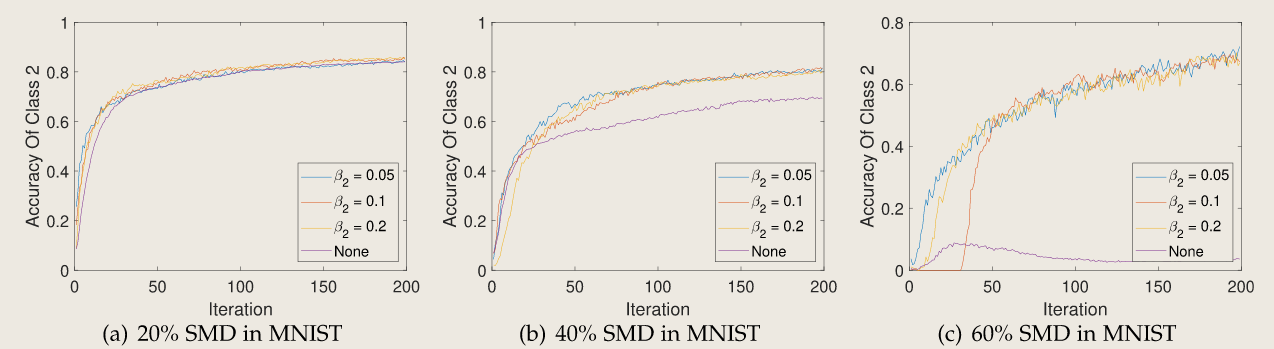
第一类恶意参与者每轮向边缘节点返回一个随机生成的更新参数，从而达到损害模型的准确性的影响，我们称之为随机恶意设备(RMD)。第二类恶意参与者将使用本地数据集来训练专门的参数，即修改某个类别的样本标签，从而降低该类别的整体模型的准确性，我们称之为专门的恶意终端设备 (SMD)。在本实验中，恶意终端设备将修改类2的标签。本地训练轮数越大，参数之间的差异越大，检测异常参数越容易。为了验证对各种训练模型的适应性，我们设置局部训练轮数 E = 1。







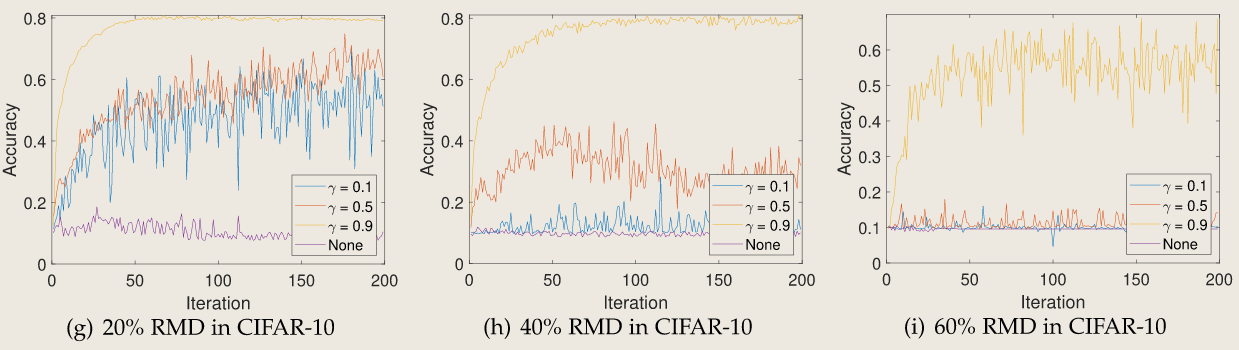
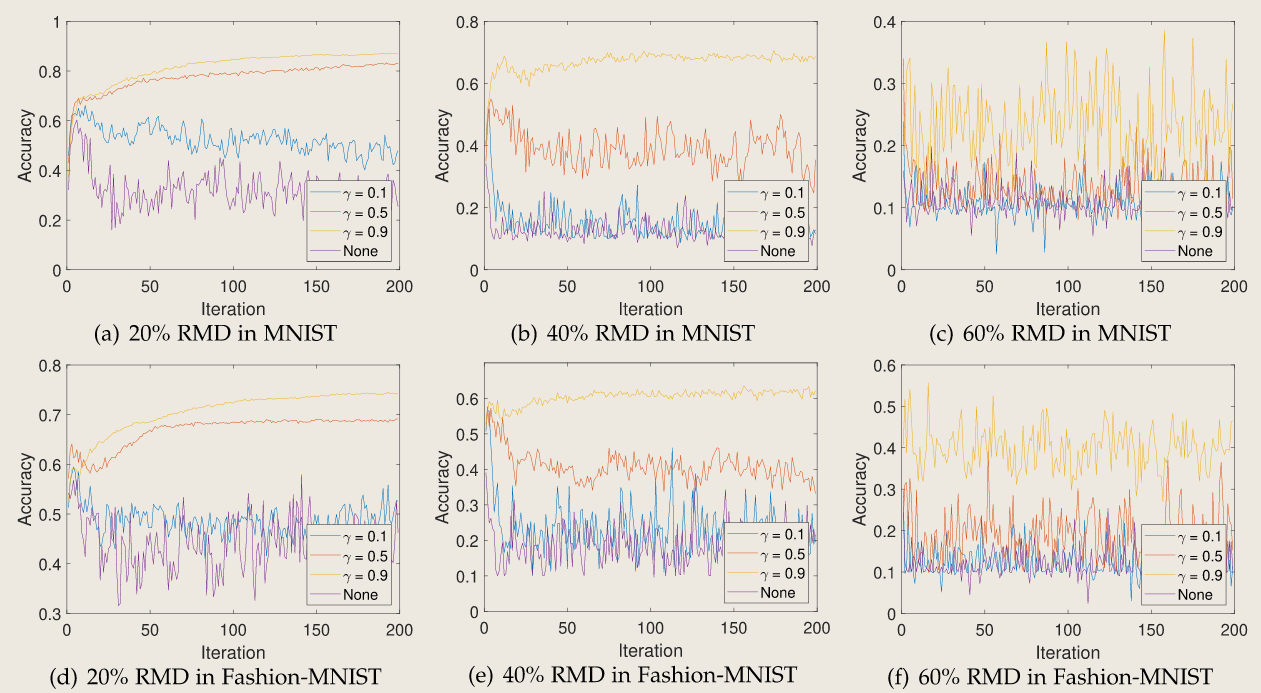
基准 b1 与 RMD 的影响。我们评估了 b1 对 RMD 模型的性能的影响。我们将终端设备中的 RMD 比例分别设置为 20%、40% 和 60%，看看 b1 如何影响准确性。为了评估 b1，我们将 g 设置为 1，那么所有检测度量都是范数检测，并将 b1 的值分别设置为 1,3,5，MNIST 结果如图 4a、4b 和 4c 所示。从结果中，我们可以看到范数检测在检测 RMD 方面可以发挥重要作用。从图4a可以看出，在没有异常检测方案的情况下，只有20%的RMD对模型的性能有很大的影响，破坏模型的收敛性。通过T = 200轮训练，模型的精度仅为0.2544。范数异常检测方案受基准参数b1的影响。随着b1的减小，范数异常检测的检测效果逐渐提高，随着迭代次数的增加，模型的精度逐渐增大。通过 T = 200 轮训练，当 b1 1/4 5 时，模型的精度可以达到 0.5297；当 b1 1/4 3 时，模型的精度可以达到 0.78226；当 b1 1/4 1 时，模型的精度可以达到 0.8756。当 RMD 尺度为 40%（如图 4b 所示）时，设置 b1 1/4 1，模型的精度仍然高达 0.7049。与准确率为 0.1285 的检测方案相比，范数异常检测算法可以获得更好的检测性能。然而，当RMD比为60%时，从图4c的结果可以看出，模型的性能很差，准确率仅为0.2503。事实上，当b1 1/4 1和T 1/4 200时，诚实终端设备的聚合权值约为0.4。通过范数异常检测算法后，RMD的聚合权值降至0.1以下，说明范数异常检测算法具有较高的筛选作用，但过多的恶意设备仍然对模型的训练过程造成很大干扰。图4d、4e和4f中Fashion-MNIST的结果显示了类似的效果。从图 4g、4h 和 4i 中 CIFAR-10 的结果可以看出，由于模型更复杂和参数数量增加，随机生成的参数更容易被检测到。



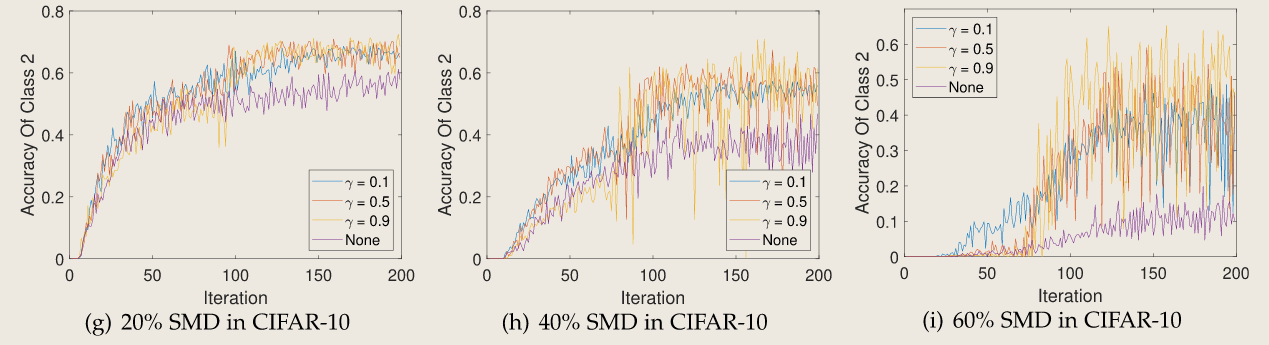
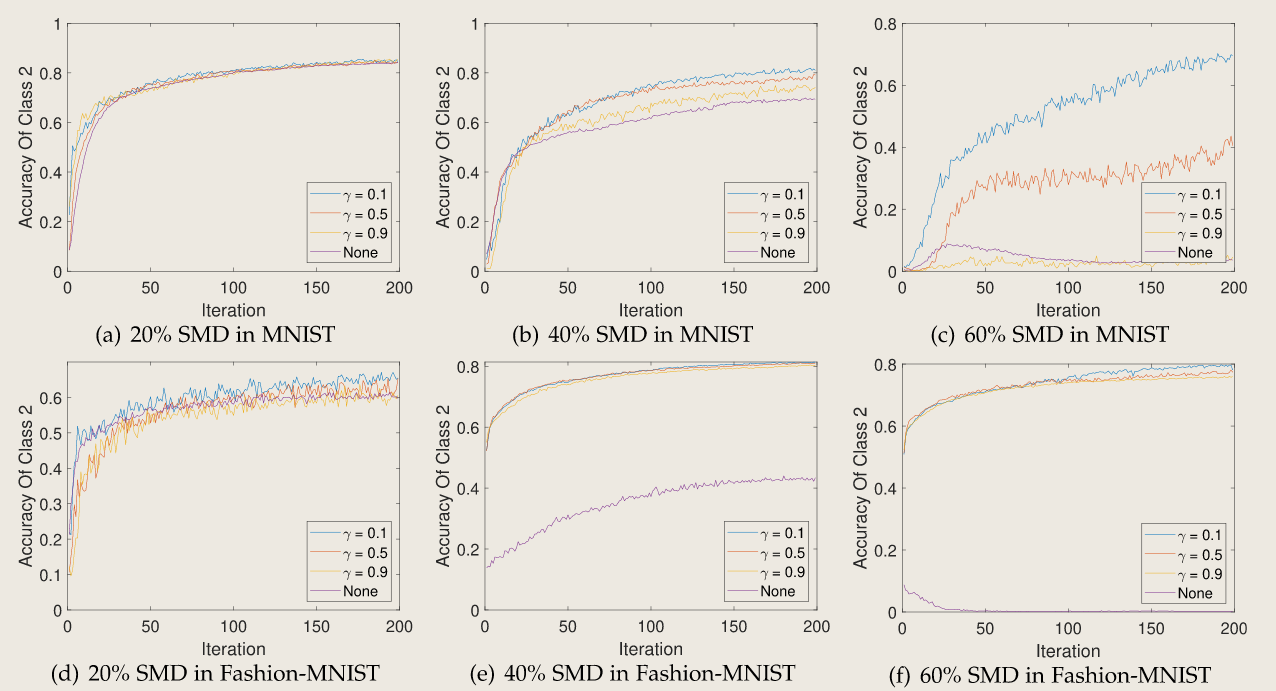
基准 b2 与 SMD 的影响。我们使用 SMD 评估 b2 对模型的性能的影响。我们将终端设备中 SMD 的比例分别设置为 20%、40% 和 60%，看看 b2 如何影响准确性。类似地，我们将 g 设置为 0，并评估 b2 的影响。我们将 b2 的值分别设置为 0.05、0.1、0.2，研究了对准确率检测算法检测性能的影响。为了提高特定类别的误分类率，在本实验中，SMD通过修改2类的标签来实现目标中毒攻击。因此，我们使用2类模型的预测准确率作为检测性能评价指标。实验结果如表5a、5b和5c所示。

如图5a所示，当参与训练的所有终端设备中SMD的比例为20%时，准确率异常检测算法并没有显著提高2类的预测精度。这是因为很难故意毒化目标，20%SMD实现的攻击效果并不明显。如果没有额外的攻击防御，第 2 类的预测准确性仍然可以达到 0.8421。虽然第 2 类的预测准确性略低于没有攻击的 0.8886，但它对模型的性能没有太大影响。精度检测可以加快模型的收敛速度。当SMD的比例为40%时(如图5b所示)，精度异常检测算法开始表现出优越的性能。此时，如果没有采取防御措施，整体模型的准确性仍然高达0.9069，第2类的预测精度降低到0.6938，SMD实现了更好的攻击效果。例如，当 b2 1/4 0:2 时，第 2 类的准确率仍然可以达到 0.8，这取得了很好的检测效果。

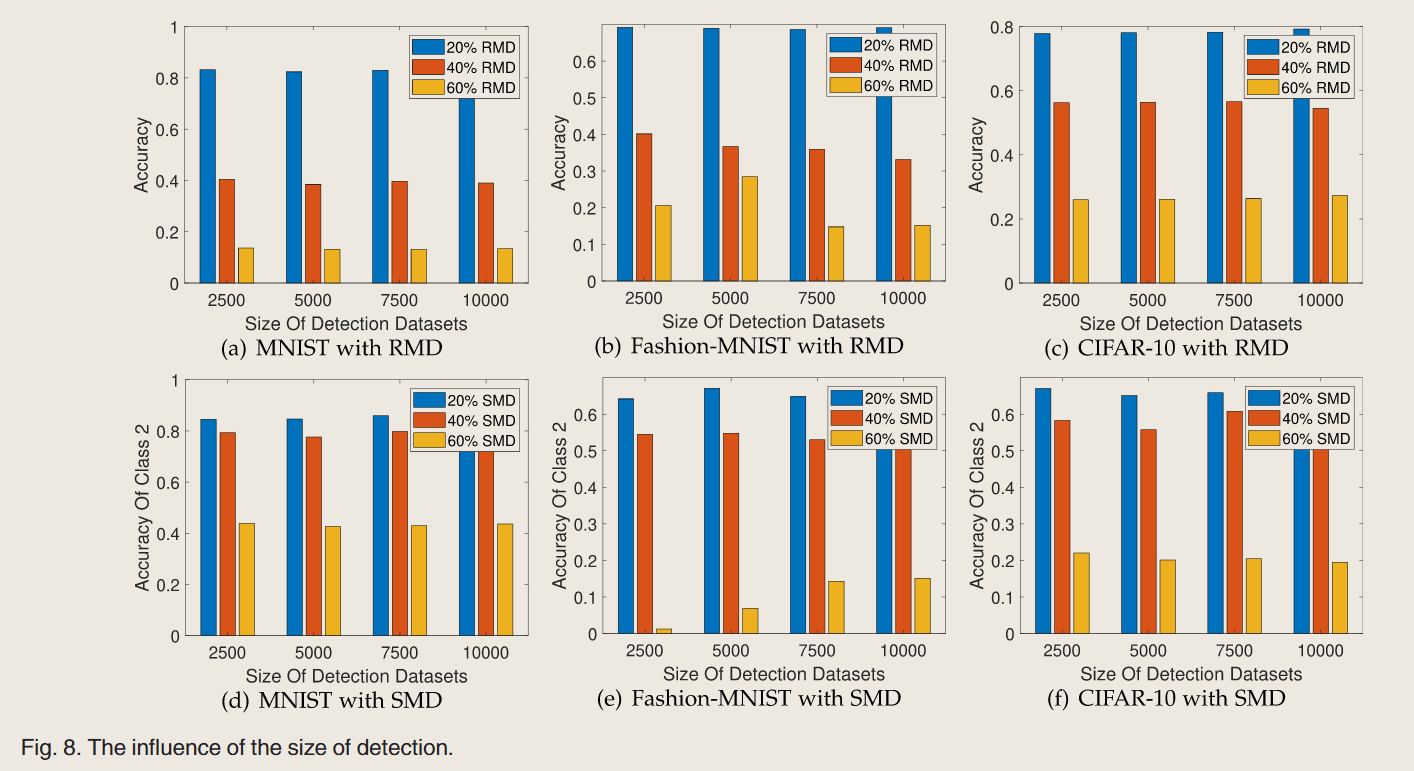
当SMD的比例增加到60%时，如图5c所示，在没有防御的情况下，虽然整体模型的准确性仍然为0.8413，但第2类的预测精度已降至0.0359，不仅实现了良好的攻击效果，而且难以检测。但在精度异常检测算法的影响下，当b2 1/4 0:2时，第2类的准确率仍可达0.6715，能够很好地抵抗目标中毒攻击。从图5c可以清楚地看出，b2对模型收敛速度的影响，在合理范围内b2越小，模型收敛速度越快。此属性还反映在对应于图 5d、5e、5f、5g、5h 和 5i 的 FashionMNIST 和 CIFAR-10 中。



权重比例 g 的影响。混合异常检测算法可以通过调整g的值来抵抗不同类型的攻击。我们通过实验验证了此属性。我们设置b1 1/4 1;b2 1/4 0:05，保证了范数检测和精度检测算法都能发挥很大的检测效果。首先，对于RMD攻击，通过设置g 1/4 0:1;g 1/4 0:5;g 1/4 0:9来研究g对混合检测算法的影响。MNIST的实验结果如图6a、6b和6c所示。当RMD的比例为20%时(如图6a所示)，g越大，混合检测算法的检测性能越好，即范数检测算法结果对混合检测结果的影响越大，检测后模型的精度越高。这是因为范数检测对 RMD 有很好的检测效果，可以快速降低 RMD 的聚合权重。然而，事实上，当 g 1/4 0:5 时，检测到的模型的准确性仍然可以达到 0.8298。此时，该模型仍然可以抵抗 RMD 攻击。然而，当g 1/4 0:1时，混合检测算法主要基于精度检测算法的检测结果，破坏模型的精度收敛，影响较大。主要原因是一开始，由于每个参数的准确率较低，无法很好地筛选恶意参数，导致模型不能很好地收敛。从图6d、6e和图6f和CIFAR-10中Fashion-MNIST的结果可以看出，图6g、6h和6i中Fashion-MNIST的结果，我们仍然可以看到类似的属性。



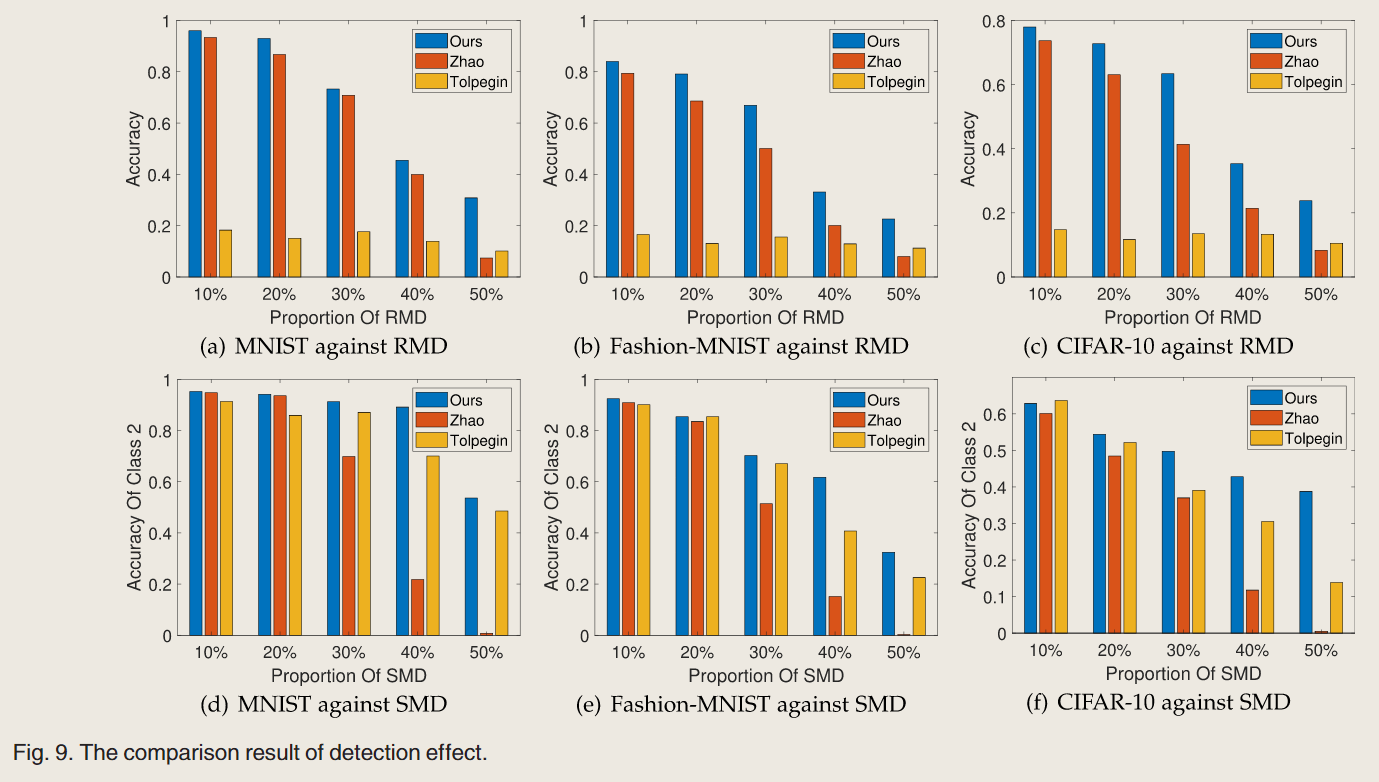
然后，对于SMD攻击，我们设置g 1/4 0:1;g 1/4 0:5;g 1/4 0:9。MNIST的实验结果如图7a、7b和7c所示。当SMD的比例为20%时(如图7a所示)，g对混合检测算法的性能影响不大，因为SMD的中毒攻击效果此时不明显。当SMD的比例增加到40%时，从图7b可以看出，g越小，混合检测算法的检测性能越好，即准确率检测算法结果对混合检测结果的影响越大，检测后模型的精度越高。这是因为SMD的攻击目标是使恶意参数尽可能接近正常参数，范数检测不能起到很好的检测效果。精度检测可以抵抗SMD攻击。因此，g越小，准确率检测结果的贡献越大，混合检测算法的效果越好。当 SMD 的比例高达 60%（如图 7c 所示）时，此属性更加明显。当g太大时，混合检测算法无法抵抗SMD中毒攻击。然而，从图7d、7e和7f所示的Fashion-MNIST的结果来看，较大的g仍然可以很好地工作，但较小的g具有更好的检测效果。对于图7g、7h和7i对应的CIFAR-10的结果，随着数据集和模型变得复杂，检测结果略不稳定，但仍明显优于无检测情况。



检测数据集大小的影响。我们研究了检测数据集大小对检测效果的影响。在 RMD 和 SMD 的条件下，我们将边缘节点的检测数据集的大小分别设置为 2500、5000、7500 和 10000 个样本，并采用混合检测算法，设置 b1 1/41; b2 1/4 0:05; g 1/4 0:5，观察 T 1/4 200 轮训练后模型精度。实验结果如表8所示。可以看出，随着检测数据集规模的增加，模型在相同条件下的准确率仅略有变化。例如，当MNIST中RMD的比例为20%时，模型精度分别为0.8320、0.8234、0.8192、0.898。从Fashion-MNIST和CIFAR10的实验结果来看，可以得到类似的结论。然而，当Fashion-MNIST的SMD比例为60%时，检测性能随着检测数据集的大小的增加而增加，变化范围在0.1以内。这表明检测数据集的大小对检测效果影响不大。这与我们的预期一致，证实了该方案的实用性。边缘节点可以通过小规模数据集实现异常检测，抵抗恶意设备的中毒攻击，保证模型的准确性而不影响模型的训练速度。

#### 5.2.3与其他工作比较

在本节中，我们将我们的方案与 [17] 和 [16] 的工作进行比较，以抵御中毒攻击、通信成本和计算成本。为了显示公平性，我们只考虑连接到它的一个边缘节点和设备之间的交集，并采用训练模型CNN(具有2个卷积层、一个平均池化层和两个完全连接层)，并保持其他超参数相同，我们方案中边缘节点的隐私预算1/4 5和检测大小设置为10000。



我们评估了两种方案在不同比例的恶意设备下的防御效果，以比较两种方案的安全性。对于我们的方案，我们设置b1 1/41;b2 1/4 0:05;g 1/4 0:5。RMD和SMD的实验结果分别如图9所示。结果表明，我们的方案在RMD和SMD的影响下都能达到更高的精度，特别是当恶意设备的比例较高时，由于共谋攻击，检测效果可能会差得多。从这个角度来看，我们的方案在安全性上表现更好。

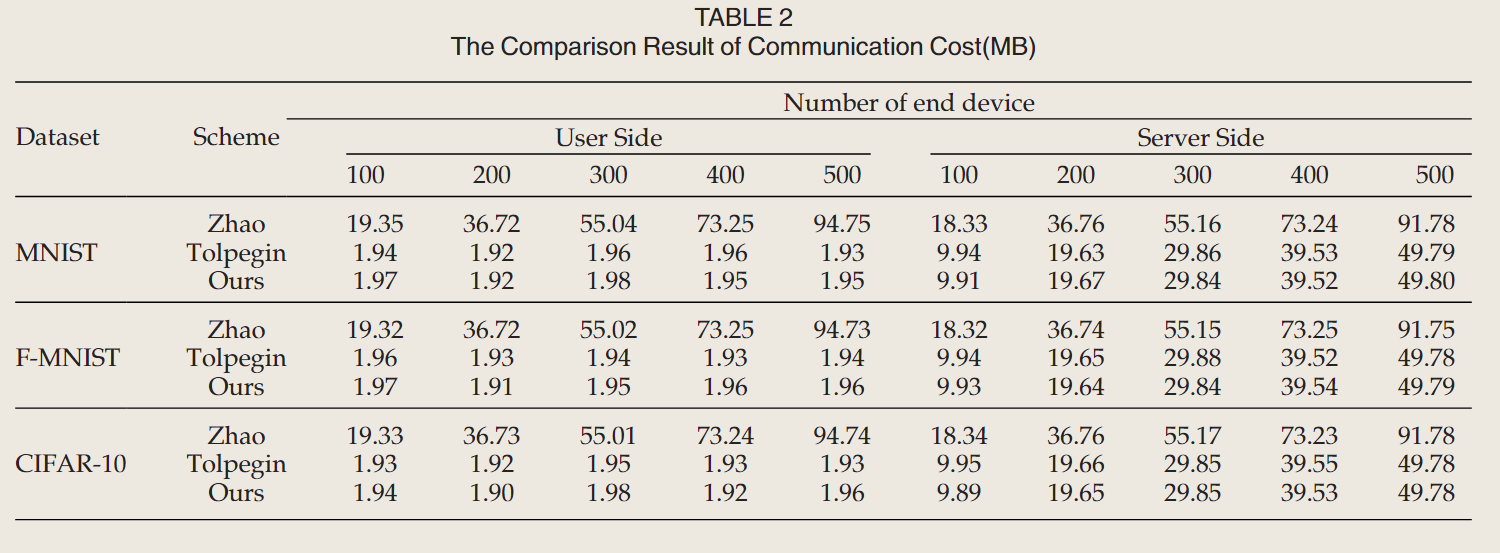
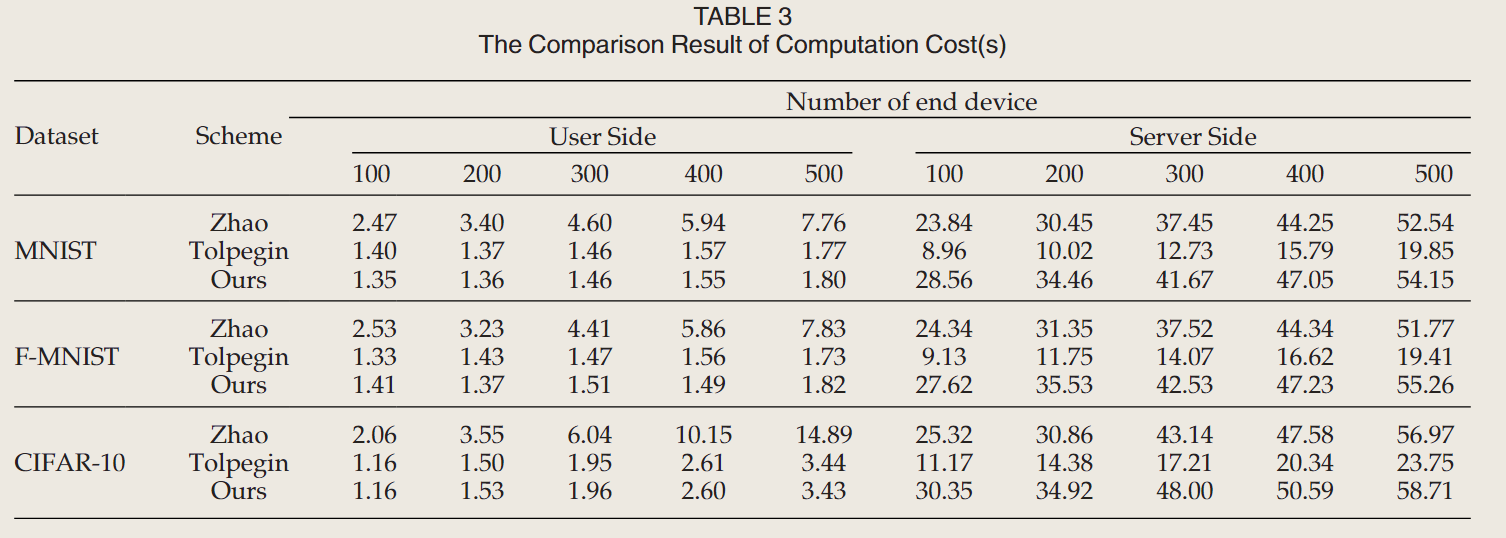


表 2 显示了在不同数据集上进行的三种方案的用户端和服务器端的不同通信成本。我们可以看到我们的方案总共有较低的通信成本。对于我们的方案中的每个用户和Tolpegin[16]，唯一要做的是用数据集局部训练模型并获得参数，添加适当的噪声并将它们发送到服务器。对于Zhao[17]中的每个用户，将有更多的任务来处理，因为检测任务被交给用户。随着用户数量的增加，每个用户都会有更多的通信开销。此外，检测过程中用户和服务器之间存在多轮交互，从而导致更多的通信成本。至于服务器端，我们的方案和Tolpegin[16]中的服务器需要检测异常参数，并根据其他参数和验证数据集计算相应的权重，Zhao[17]中的服务器还需要根据分布式子模型的报告来确定用户参数的权重，分布和报告过程在一定程度上会导致更多的通信消耗。值得注意的是，虽然 Tolpegin [16] 具有与我们的方案几乎相同的通信成本，但我们的方案从图 9 所示的实验结果具有更好的检测效果。如表 2 所示，三个不同数据集上的通信成本几乎相同。这是同一模型和网络设置的结果。



最后，我们分别在用户和服务器方面比较了我们的方案与Zhao[17]和Tolpegin[16]的计算成本。表 3 显示了每轮方案中用户的运行时间统计数据。我们方案中消耗的时间几乎没有设备数量的增加，而运行时间随着 Zhao [17] 中的用户数量的增加而增加，用户端会采取更多额外的检测。然而，Zhao[17]中的服务器的计算成本较低，导致耗时的任务被委托给用户。由于更多的分布和统计工作，随着用户数量的增加，服务器的时间也会增加。相反，在我们的方案中，所有检测任务都将由服务器完成，因此服务器所花费的时间随着用户数量的增加而增长得更快。如表 3 所示，Tolpegin [16] 在服务器中的计算成本要低得多，因为它使用更简单的算法而不进行准确性检测，从而提高性能以牺牲准确性为代价。同样，表 3 所示的三个不同数据集的实验结果几乎是一致的。

此外，我们的方案在训练期间具有较小的 CPU 开销和内存开销。具体来说，对于 CIFAR10，内存开销为 2903.65 MB，CPU 开销为 15.1%，这对于计算资源有限的终端设备是合理的。需要注意的是，资源可以通过运行参数人为分配。

综上所述，在安全性和准确性方面，本文提出的异常检测方案对RMD和SMD攻击具有较好的防御效果，在两种攻击下，模型都能收敛到更高的精度;在效率方面，本文方案带来的服务器与用户之间的通信开销最小。虽然与Zhao[17]和Tolpegin[16]相比，检测过程会在服务器上造成更多的计算开销，但在边缘计算中，服务器的计算资源比用户更强大。因此，权衡是合理的，即牺牲服务器端更多的资源来提高用户端的效率和模型精度。因此，我们的方案可以在安全性、效率和准确性之间实现最佳权衡。

## 6.结论

本文提出了一种针对边缘计算设置的中毒攻击的差分私有联邦学习模型。在该方案中，终端设备与边缘节点交互，在多层网络模型下实现模型训练，允许获得高效率和准确性的预测模型。该方案旨在防御中毒攻击并防止隐私泄露。具体来说，它通过我们提出的基于权重的检测方案抵抗中毒攻击，其中边缘节点为验证数据集中表现不佳的参数分配较低的权重，以减少它们的贡献。该方案通过利用多层网络中的差分隐私来保留数据隐私和模型隐私。实验结果表明，我们的方案是安全的、高效和准确的。在未来的工作中，我们将研究我们的方案对其他 FL 攻击的适用性，例如基于 GAN 的中毒攻击、后门攻击和回归攻击。