基于区块链共识激励机制的新型联邦学习系统

米 波[1], 翁 渊[1], 黄大荣[1], 刘 洋[1]

[1]重庆交通大学 信息科学与工程学院 重庆 中国 400074

摘要 随着云存储、人工智能等技术的发展, 数据的价值已获得显著增长。但由于昂贵的通信代价和难以承受的数据泄露风险迫使各机构间产生了“数据孤岛”问题, 大量数据无法发挥它的经济价值。虽然将区块链作为承载联邦学习的平台能够在一定程度上解决该问题, 但也带来了三个重要的缺陷: 1) 工作量证明(Proof of Work, POW)、权益证明(Proof of Stake, POS)等共识过程与联邦学习训练过程并无关联, 共识将浪费大量算力和带宽; 2) 节点会因为利益的考量而拒绝或消极参与训练过程, 甚至因竞争关系干扰训练过程; 3) 在公开的环境下, 模型训练过程的数据难以溯源, 也降低了攻击者的投毒成本。研究发现, 不依靠工作量证明、权益证明等传统共识机制而将联邦学习与模型水印技术予以结合来构造全新的共识激励机制, 能够很好地避免联邦学习在区块链平台上运用时所产生的算力浪费及奖励不均衡等情况。基于这种共识所设计的区块链系统不仅仍然满足不可篡改、去中心化、49%拜占庭容错等属性, 还天然地拥有 49%投毒攻击防御、数据非独立同分布(Not Identically and Independently Distributed, Non-IID)适应以及模型产权保护的能力。实验与论证结果都表明, 本文所提出的方案非常适用于非信任的机构间利用大量本地数据进行商业联邦学习的场景, 具有较高的实际价值。

关键词 联邦学习; 区块链; 共识算法; 模型产权保护; 投毒攻击

中图法分类号 TP309.2 DOI 号 10.19363/J.cnki.cn10-1380/tn.2024.01.02

A Novel FL System Based on Consensus Motivated Blockchain

MI Bo[1], WENG Yuan[1], HUANG Darong[1], LIU Yang[1]

[1] School of Information and Engineering, Chongqing Jiaotong University, Chongqing 400074, China

Abstract With the advancement of technologies such as cloud storage and AI (artificial intelligence) in recent years, the value of data has experienced significant growth. However, the exorbitant costs associated with communication and the intolerable risks of data leakage have given rise to a pervasive issue of “data isolation” among institutions, rendering a substantial portion of data unable to realize its full economic potential. Although using blockchain as a platform for federated learning can solve this problem to a certain extent, it also brings three primary shortcomings: 1) traditional consensus processes like PoW (proof of work) and PoS (proof of stake) remain largely disconnected from the federated learning training process, resulting in substantial wastage of computational power and bandwidth; 2) nodes may decline to participate actively in the training process or even disrupt it due to self-interest considerations, driven by competitive dynamics; 3) in open environments, data traceability during the model training process is challenging to establish, consequently diminishing the cost of attack for potential malevolent actors. Our study manifested that, instead of relying on traditional consensus mechanisms such as PoW and PoS, combining federated learning and model watermarking technology can make the consensus algorithm more fair and reliable. It can avoid the waste of computing power and unbalanced rewards thanks to federated learning, and the innovative consensus mechanism not only retained the properties of immutability, decentralization, and 49% byzantine fault tolerance but also naturally resisted 49% poisoning attack, adapted Non-IID (not independent and identically distributed) dataset and protected intellectual property. Both experimental and empirical evidence unequivocally demonstrate that the proposed solution in this study is exceptionally well-suited for scenarios involving non-trusting institutions collaboratively leveraging large volumes of local data for commercial federated learning, thereby holding substantial practical value.

Key words federated learning; blockchain; consensus algorithm; intellectual property protection; poison attack

通讯作者: 翁渊, Email: wengyuan980930@mails.cqjtu.edu.cn。

本课题得到中国国家自然基金(No. 61903053), 重庆市科教委项目(No. KJCX2020033), 上海市信息安全综合管理技术重点实验室开放课题(No. AGK2020006)资助。

收稿日期: 2022-05-05; 修改日期: 2022-08-20; 定稿日期: 2023-09-26

1 引言

大数据驱动的人工智能技术有助于在整体上生成高精度泛化模型, 但在实际应用过程中却往往存在着数据来源不足的状况[1-2]。作为一种新兴的机器学习框架, 联邦学习(Federated Learning, FL)可以在节点数据孤立的情况下实现分布式模型训练, 在一定程度上解决机器学习过程中的数据稀缺问题。此外, 由于这种方案[3]能够在人工智能模型的训练过程中将数据离线, 因而也具有数据隐私保护和节省带宽的能力。随着智能边缘设备的普及和性能提升, 移动网络的计算能力不断增强, 联邦学习在智慧交通[4]、智慧城市[5]、商业数据挖掘[6-7]等领域都得到了广泛的应用。目前联邦学习已经与很多行业相融合, 且在区块链、模型水印等技术的促进下不断赋予新的功能[8], 对实际生活产生了良好的经济效益和社会价值。

在信息化时代, 大数据背景下的数据隐私问题愈来愈受到人们的关注。由于数据与生活、生产的关联性日益增强, 隐私泄露问题必然会遭到社会的广泛抵制, 信息价值开发和敏感数据保护之间的矛盾正不断显现[9]。例如, 2020 年 12 月, “明星健康宝照片泄露”事件中大量用户个人数据被非法贩卖, 引起我国公安机关的高度警觉和公众的广泛讨论。2017 年 6 月 1 日起实施的《中华人民共和国网络安全法》指出不得泄露、篡改用户数据, 且自 2020 年以来《数据安全法》、《个人信息保护法》相继出台, 这也充分说明了国家对数据隐私保护的重视。

针对机器学习中存在的数据安全风险, 学者提出了一系列的隐私保护方案, 主要包括联邦学习、多方安全计算(Secure multiparty computation, SMPC)[10-11]、同态加密(Homomorphic encryption, HE)[12-13]和差分隐私(Differential privacy, DP)[14-15]这几类主流技术, 其中联邦学习采用的分布式离线训练方法能够在隐私保护的同时有效节省通信及计算资源, 非常适用于数据量大、数据源分布广、信息敏感度高的场景。

联邦学习的概念最初出现于文献[16], 逐步演化为纵向联邦学习[17]、横向联邦学习[18]和联邦迁移学习[19]三种基本框架。其中, 纵向联邦学习主要适用于参与方数据记录大量重合的场景, 而横向联邦学习主要考虑节点间数据特征基本相同的情况, 当参与方的样本空间有部分重叠但特征不尽相同时联邦迁移学习则更为适合。在算力不均衡的可信任环境中, 上述三类方案往往采用 C/S(客户/服务器, Client/Server)模式予以实现。正是因为充分利用了吞吐量高、性能优异的设备作为中心节点, C/S 模式相较于分布式学习具有训练效率更高、利益分配更均衡、本地数据更安全等优势。然而, 在非信任环境下, C/S 模式的联邦学习方法极易遭受身份伪造、数据篡改、拒绝服务(Denial of Service, DoS)等攻击的威胁。为解决这些信任问题, 文献[20]提出一种基于区块链的联邦学习方案, 将抽象的可信服务节点实例化为分布式的共识激励机制; 文献[21]将联邦学习中的梯度作为一部分贡献, 结合 Algorand 共识协议提升了激励的公平性。文献[22]中通过降低联邦学习中的交互参数以保证用户的匿名性从而降低收到攻击的风险。

图 1 展示了基于链上共识的联邦学习整体框架。该框架中的节点可同时或分别扮演数据提供者和区块挖掘者两种角色。所有参与者在本地数据集上完成子模型的训练, 随后将其上传至随机选择或投票选举出来的矿工。矿工负责对所有本地模型进行验证与融合, 然后根据 PoW 或 PoS 共识机制产生新的区块。这些区块要负责记录矿工的挖矿奖励和数据提供者的贡献奖励, 并存储模型更新后的参数。随后, 参与者将聚合后的模型再次下载, 不断地重复上述过程直至得到满意的全局机器学习模型。由此可见, 这种机器学习方法的本质在于间接的数据共享和有效的合作激励, 因此共识算法的可靠性和奖励机制的公平性会直接影响整个系统的性能。

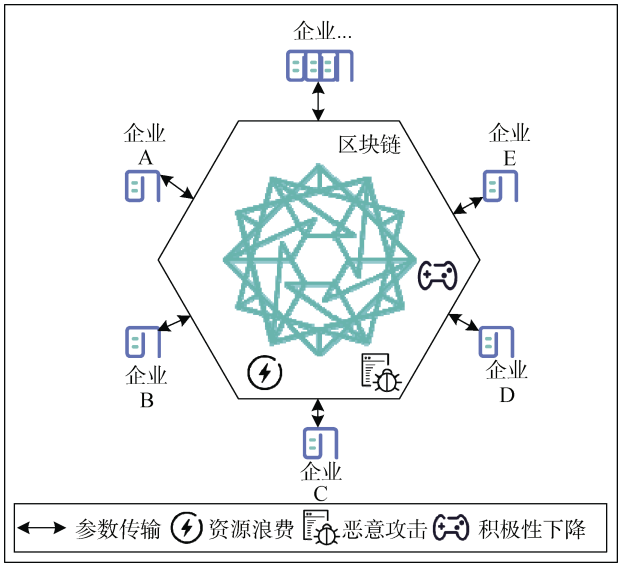


图 1 基于区块链的联邦学习框架

Figure 1 A federated learning framework based on blockchain

尽管基于共识的联邦学习方法有助于建立起参与节点间的广泛信任, 但现有方案仍普遍存在着以下三方面的缺陷:

1) 资源浪费问题。文献[23]指出, 将区块链作为联邦学习过程中数据和模型的载体, 主要是为了保证相关信息能够被可靠地记录及追溯。然而, 由于PoW[24]、PoS[25]等“挖矿”行为与联邦学习过程的收敛性并无直接关联, 共识机制的引入会直接导致大量算力和带宽被浪费。

2) 节点活性问题。在实际生产环境中, 节点数据和计算资源都是具有一定经济价值的。在某一节点发起联邦学习的模型训练后, 其他节点可能会因为利益的考量而拒绝或消极合作, 甚至会因为竞争关系投入虚假数据对模型进行干扰, 最终导致全局模型无法使用或训练过程无法收敛。

3) 攻击手段的多样性问题。尽管联邦学习领域正不断引入各种新的机制来对抗日益多样化的攻击手段, 但大都针对片面的安全目标[26]。与传统机器学习所面临的威胁类似, 模型攻击[27]、投毒攻击[28]、后门攻击[29]、推理攻击[30]等方法在联邦学习中也主要是对数据隐私和全局模型进行破坏。事实上, 联邦学习在一定程度上具有数据隐私保护的特性。因此, 安全机制的实现不应当以攻击手段为驱动, 而需要将数据保密性和模型准确性作为根本目的。

联邦学习的商业场景往往具有参与节点数量少、合作关系松散耦合的特点。此外, 非信任分布式环境的物理脆弱性和攻击来源的多样性极有可带来节点丢失、数据污染、模型篡改等隐患, 从而导致训练过程因无法准确收敛而失败。为此, 本文将针对节点数量有限、数据吞吐量大、互信程度低的跨企业分布式场景, 结合区块链及水印技术来构造一种全新的共识激励机制, 从而解决联邦学习中算力浪费、奖励不均以及鲁棒性弱的问题。总体而言, 其基本思想是借助区块链的一致性记录能力以及模型水印的版权保护机制, 将模型训练分发到多个节点上并行执行, 每轮结束后多个矿工将分别对收集到的本地模型进行聚合, 并根据评价准则在链上达成模型准确度和参与者贡献度的共识, 由此产生新的区块, 不断迭代直至获得期望的全局模型。在具体的实施过程中, 参与训练的节点会将自身的水印嵌入到梯度模型中用于证明所做出的贡献。为了争夺写入权限, 所有融合节点将利用所接收到的梯度构造一个能够让大多数节点都认可的全局模型。最终, 达成共识的全局模型将会由它的创造者写入区块。基于上述策略, 本文将 Paxos 共识协议[31]中的投票理念与联邦学习相结合, 构造出一种新型共识协议 Paxos Federated Consensue(PFconsensue), 并通过高鲁棒性水印融合算法的设计, 最终形成一套可证明完备的联邦学习共识激励机制。

本文的贡献主要在以下几个方面:

1) 基于联邦学习的共识协议。将联邦学习的训练过程作为节点“挖矿”环节, 使消耗的资源转换成具有经济价值的人工智能模型。同时, 模型聚合采用去中心化与性能投票的方式进行, 克服了联邦学习中 Non-IID[32]与投毒攻击所造成的全局模型性能下降的缺点, 实现了联邦学习与区块链技术的优势互补。

2) 公平的区块链共识激励机制。为提高联合训练的参与度, 依靠高鲁棒性模型水印技术和参数距离算法, 实现了公平的节点贡献度分配, 可以更好地刺激节点参与模型训练过程。在模型聚合环节, 将区块的写入权奖励给最优模型的创造者, 也能够充分地保证节点积极参与模型聚合。可见, 该区块链系统在本地训练和模型聚合两方面均保证了参与节点的活性。

3) 系统的整体完备性证明。从理论上了证明了共识算法的正确性, 并通过形式化方式分析了共识算法在拜占庭环境下的容错能力。同时, 通过实际数据的分布情况抽象出相应的约束条件, 分别讨论了该系统组成部分在实际环境中运行的有效性与稳定性。此外, 对系统的整体安全性也进行了充分的证明。

4) 实验仿真及分析。利用计算机模拟验证了共识协议的有效性。根据实际采集的“重庆市实时交通流”数据在多台设备间部署共识决策环境, 验证了本方案在现实环境中的可行性及准确性。此外, 基于系统性的区块链仿真, 进一步展示了本方案对联邦学习中潜在威胁的抵抗力。

2 系统整体模型

由于区块链具有不可篡改、易追溯和去中心化等优势, 与联邦学习相结合能够极大程度地克服联邦学习中所潜在的风险。对此, 本章节将基于 PFconsensue 协议、模型水印等技术构造整体的区块链系统, 并给出实际环境中的安全性形式化定义。

2.1 系统框架设计

当前已有部分研究人员将区块链用于解决联邦学习在非信任环境中的安全协同训练问题。文献[33]中选取区块链上的可靠节点来参与联邦学习, 并通过差分隐私技术以保证训练数据的安全。文献[34]则将联邦学习过程中的全局数据组织成“全局模型状态树”, 作为交易内容存储到区块链中。而文献[35]也类似地利用区块链存储联邦学习过程中的各种模型参数, 该方案还可以借助其他边缘设备来分担训

练能耗。然而, 由于以上方案皆未考虑模型所具有的知识产权特性, 可能产生模型盗用现象, 也将导致参与方发生产权纠纷。另一方面, 依附于区块链的联邦学习会因为共识过程而造成大量的资源浪费, 导致节点参与度下降。为了解决上述两个问题, 本文设计了图 2 所示的联邦区块链结构。在该结构中, 链上记录的数据主要包括: (1) 上一个区块的 Hash; (2) 融合后的模型参数; (3) 构造融合模型所使用的局部梯度集合; (4) 基于评价准则的产权奖励; (5) 下一轮训练的优化目标。

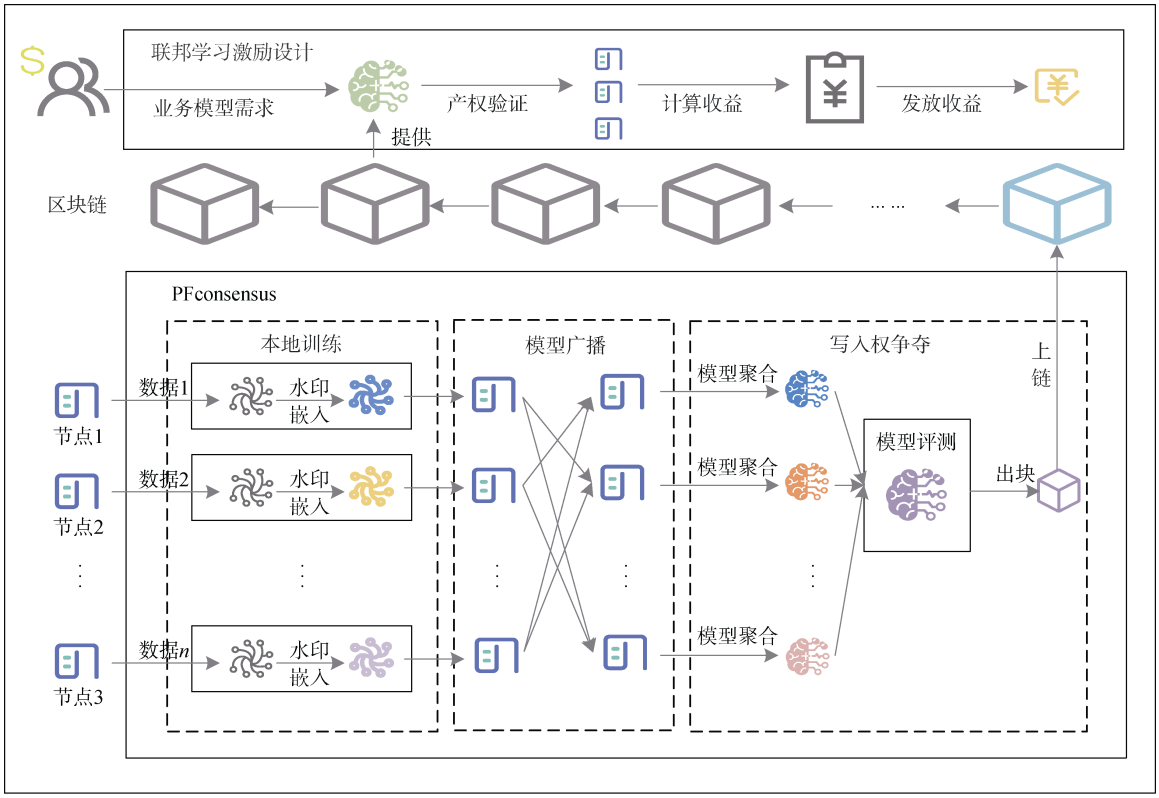


图 2　本文区块链系统结构

Figure 2　The structure of the blockchain system in this paper

在协议开始时, 参与节点将会从区块链上获取公开发布的初始模型及训练目标, 并在本地训练出包含水印的梯度模型。随后, 节点会将梯度模型通过 Gossip 协议[36]进行广播, 并在收到足够的梯度信息后尝试通过聚合算法得到聚合模型。最后, 聚合模型会传送至各个节点进行评测, 投票产生的最优模型和下一轮协议的优化目标将被同时写入新的区块。考虑到数据的防篡改问题, 除分布式存储外还将借助 Hash 链式结构和最长链原则[37]来确保区块链的持久性。值得一提的是, 本方案在设计区块数据结构时将各个节点的梯度模型一并记录在区块上,这样可以确保聚合模型的可信度。

就节点活性而言, 由于区块链上的聚合模型保留有各参与方的梯度模型水印, 他们可以据此对调用该模型的第三方收取知识产权费。与此同时, 高鲁棒水印融合技术的使用还能够有效防止公开模型被盗用。可见, 该方案能够充分激励各个节点参与联邦学习过程。

更进一步地, 本文对上述区块链的整体构架进行如图 3 所示的逻辑刻画和分层设计。节点之间主要负责构造区块链数据服务, 而第三方只需通过 API 接口发布模型需求或对模型进行调用。

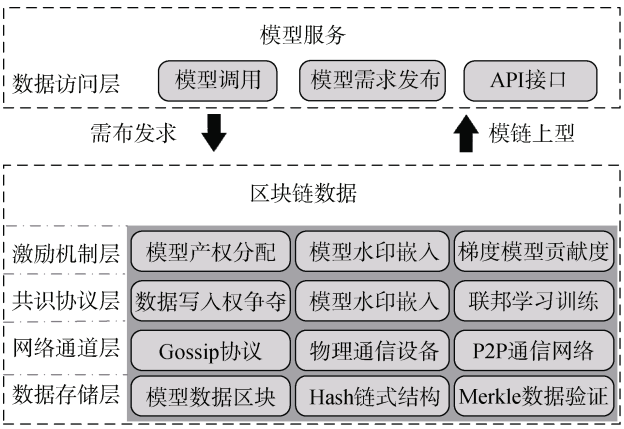


图 3　本文区块链框架设计

Figure 3　The blockchain framework of this paper

2.2 攻击模型及安全定义

区块链能够解决联邦学习的中心化问题, 联邦学习则实现了区块链上的数据隐私保护。为确保本文设计的方案能够可靠运行, 首先对其性能与安全进行形式化定义, 后面章节也将围绕这些定义进行阐述及论证。

本文提出的联邦学习共识算法PFconsensus主要用于解决分布式环境下的数据一致性问题。PFconsensus 协议的攻击环境和安全性定义如下。

定义 1. 拜占庭攻击环境.

设参与第 j 轮共识的节点集合为 $\mathbb{G}_j$, 对 $\forall k \in \mathbb{G}_j$, k 具有概率多项式时间(probabilistic polynomial-time, PPT)的计算能力, 且拥有所有节点的数字签名验签密钥集合 $PK = \left\{ pk_i \in i \mid i \in \mathbb{G}_j \right\}$, 用于验证其他节点传输数据的真实性。节点 k 自己上传的梯度模型 $\boldsymbol{W}_k$ 具有以下特征:

1.梯度模型 $\boldsymbol{W}_k$ 本身需满足水印验证 $V_w\left(\boldsymbol{W}_k, \left(B_k, \theta_k\right)\right) = True$;

2. 聚合节点 d 在融合过程中利用 $\boldsymbol{W}_k$ 所得到的聚合模型 $\boldsymbol{N}_d$ 能够通过水印验证, 即 $V_w\left(\boldsymbol{N}_d, \left(B_k, \theta_k\right)\right) = True$;

3. 如果聚合节点 d 在融合过程中未利用 $\boldsymbol{W}_k$, 那么所得到的全局 $\boldsymbol{N}_d$ 不能满足水印校验, 即 $V_w\left(\boldsymbol{N}_d, \left(B_k, \theta_k\right)\right) = False$;

在攻击环境下存在着部分拜占庭节点, 本文记这些节点所构成的集合为 $\mathbb{A}$, 对任意 $a \in \mathbb{A}$, 它具有伪造该集合中其他节点数字签名和模型水印的能力, 并可能发起选择性通信、延时通信、通信乱码等攻击。而对于诚实节点 $c \in \mathbb{G}_j - \mathbb{A}$ 而言, 它们将按照 PFconsensus 协议正常运行, 且在联合训练期间一直在线, 不存在延时通信、通信乱码的情况。

定义 2.待融合梯度模型.

记正常传输梯度模型所需的时间为 $\overline{t^w}$, 节点 k 训练梯度模型的耗时为 t_k^{train}, 进行模型聚合的耗时为 t_k^{avg}。在 n 个参与节点中选取 m 个梯度模型进行融合, 要求其执行过程满足以下条件:

$$\frac{\sum_{k=0}^{n} t_k^{train}}{n} = \overline{t^{train}}; \tag{1}$$

$$\max(t_k^{train}) + \overline{t^w} = m\overline{t^{train}}; \tag{2}$$

$$\overline{t^{train}} \geqslant \overline{t^w} \gg \overline{t^{avg}}; \tag{3}$$

$$\sqrt{\frac{\sum_{k=0}^{n}(t_k^{avg} - \overline{t^{avg}})^2}{n}} \cong 0。 \tag{4}$$

这四个条件能够保证在拜占庭环境下至少存在一个诚实节点正确地执行共识, 从而避免因性能差异或共谋等原因将所有诚实节点排除在共识过程之外。其中, 式(2)能够保证诚实节点在承诺打分阶段至少接收到 m 个正确的梯度模型, 而式(3)保证了聚合后的模型集合中必然包含一个正常的聚合模型。具体分析将在后面给出。

定义 3.拜占庭环境下共识协议的安全性.

在攻击环境下满足以下两个条件则表明共识协议是安全的, 其中 $Card(\mathbb{Q})$ 表示集合 $\mathbb{Q}$ 中的元素个数:

(1) 当 $Card(\mathbb{A}) < Card(\mathbb{G}_j)/2$ 时, PFconsensus 协议能够完成;

(2) 当 $Card(\mathbb{A}) < Card(\mathbb{G}_j)/2$ 时, 非拜占庭节点能够得到相同的结果。

本文将联邦学习算法作为模型训练的基本框架, 但为保证去中心化后仍然能够正常工作, 还需考虑如下额外因素及需求。

定义 4.投毒攻击节点.

对于去中心化环境中的拜占庭节点 $k \in \mathbb{A}$, 它能够发布恶意梯度 $\boldsymbol{W}_k'$, 使任意聚合了 $\boldsymbol{W}_k'$ 的聚合模型 $\boldsymbol{N}'$ 性能下降。

定义 5.去中心化环境中联邦学习算法的有效性.

针对区块链上联邦学习算法的有效性问题, 本文方案需满足以下性质:

(1) 当 $Card(\mathbb{A}) < Card(\mathbb{G}_j) - m$ 时, 最终上链的聚合模型以可忽略的概率包含投毒梯度模型(其中 m 为常数);

(2) 最终上链的聚合模型与中心化联邦学习方案在准确性方面的差异可忽略。

最后, 对区块链的整体安全性做如下定义:

定义 6.拜占庭环境下区块链的整体安全性.

(1) 当 $Card(\mathbb{A}) < Card(\mathbb{G}_j)/2$ 时, 区块链上的数据被拜占庭节点所篡改的概率可忽略;

(2) 基于区块链的联邦学习共识激励机制对于任意拜占庭节点 $k \in \mathbb{A}$, 盗取其他诚实节点贡献度的概率可忽略。

3 基于区块链的联邦学习共识激励机制

针对上述对拜占庭攻击环境的定义, 本章节将先引入模型水印技术来保证联邦学习过程中的模型产权证明。进一步的, 将详细介绍 PFconsensue 协议

的运行过程。最终, 通过上链模型数据和模型水印设计了一种公平的激励机制。该机制能够在保证节点数据隐私的同时维持参与节点的训练积极性。

3.1 FedIPR 模型水印

在设计 PFconsensue 共识算法时, 需确保网络中数据传输的可靠性, 并维护模型版权对融合过程的鲁棒性, 为此需要构造适应的数字签名和模型水印方案。由于在共识激励的过程中需要对联邦学习产生的梯度模型进行交叉验证, 本文考虑结合 FedIPR 模型水印与数字签名算法来保证模型的唯一性。此外, 在对聚合模型进行产权证明时, FedIPR 算法也能提供一个可信的结果来保证激励机制的公平。FedIPR 算法最初由 Fan 等人[38]提出, 它能够通过调整模型的目标函数, 同时植入白盒水印与黑盒水印, 本文构造类似的水印植入过程如下:

(1) 对于节点 $k \in \{1, ..., K\}$, 其密钥生成算法为 $I: I() \to B_k, \theta_k, \boldsymbol{T}_k$, 其中白盒水印部分的签名内容为 B_k, 签名提取参数为 $\theta_k = \{S_k, E_k\}$, 而黑盒水印的后门数据集为 $\boldsymbol{T}_k = \{(X_1, Y_1), ..., (X_J, Y_J)\}$, X 和 Y 分别表示后门数据的特征及标签。

(2) 节点 k 对联邦模型进行训练的过程中将后门 $\boldsymbol{T}_k$ 及签名 B_k 加入其目标函数:

$$Opt = \overbrace{L_{D_k}(\boldsymbol{W}_k^t)}^{\text{mian task}} + \alpha_k \overbrace{L_{T_k}(\boldsymbol{W}_k^t)}^{\text{trigger sign}} + \beta_k \overbrace{R_{B_k,\theta_k}(\boldsymbol{W}_k^t)}^{\text{feature sign}}, \quad (5)$$

则, 对于节点在第 t 轮的梯度计算过程将按照公式 $CilentUpdata(n, \boldsymbol{W}^t) = \boldsymbol{W}^{t-1} - \eta \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{W}}$。

(3) 模型聚合算法将采用梯度平均策略 (Federated Averaging):

$$\boldsymbol{W}^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \boldsymbol{W}_k^{t+1}, \quad (6)$$

其中, $\boldsymbol{W}_k^{t+1} \leftarrow CilentUpdate(n, \boldsymbol{W}^t)$。

模型验证包括黑盒与白盒两种方式。就白盒水印而言, 若节点 k 需证明其对聚合模型的贡献, 可以通过提取算法 $\widehat{B_k} = MLsign(S_k(\boldsymbol{W}), E_k)$ 从聚合模型 $\boldsymbol{N}[]$ 的权值中恢复出近似签名信息 $\widehat{B_k}$, 之后通过汉明算法 $H(B_k, \widehat{B_k})$ 计算 B_k 与 $\widehat{B_k}$ 之间的距离, 并通过判断距离大小来证明其知识产权的有效性, 即

$$V_w(W, (B_k, \theta_k)) = \begin{cases} True, & \text{if } H(B_k, \widehat{B_k}) \leqslant \varsigma_H \\ False, & \text{otherwise} \end{cases}。 \quad (7)$$

而对于节点 k 植入的黑盒水印, 它可以通过将后门数据 $\boldsymbol{T}_k$ 输入模型 $\boldsymbol{N}[]$ 并判断输出的准确性来予以证明:

$$V_B(\boldsymbol{N}, \boldsymbol{T}_k) = \begin{cases} True, & \text{if } E_{T_n}(I(Y_j \neq \boldsymbol{N}[X_i])) \leqslant \varsigma_y \\ False, & \text{otherwise} \end{cases}。 \quad (8)$$

上述算法的正确性与鲁棒性已经在文献[38]中得到了验证。本文中实验也表明该算法在分布式联邦学习环境下具有很好的鲁棒性, 能够满足定义 1 中对签名算法的要求。

为保证模型在传输过程中的真实性并降低其通信轮数, 本方案也将使用数字签名算法来进行可靠的数据传输。对于任意节点 $k \in \mathbb{G}_j$, 传输消息 $m \leftarrow \{0, 1\}^n$ 时, 将同时计算 $\sigma \leftarrow Sign(sk, Hash(m))$, 最终打包得到 (σ, m), 并将其发送到目标节点。

3.2 PFconsensus 联邦共识算法

最早提出的联邦学习算法是一种基于 C/S 框架的中心化服务, 每个用户需要传输各自的梯度模型给服务器, 而服务器会利用他们的梯度模型进行聚合并返回给客户端进行迭代训练。该结构极易导致拒绝服务攻击, 因而有学者通过结合区块链中的智能合约, 将其改造为去中心化方案。为避免无谓的能耗, 本文将联邦学习算法本身作为共识, 并结合区块链与水印技术在一定程度上解决联邦学习中的 Non-IID 及投毒问题。该算法的核心在于通过对比聚合模型的性能来达成一致, 主要可以概括为模型性能筛选和共识写入两个部分。

本文将参与共识过程的角色分为三种: proposer、acceptor 以及 learner(acceptor 和 learner 的角色互斥)。其中, proposer 是数据的产生和发送者, acceptor 表示数据的接收者和模型性能的裁决者, learner 作为数据的最终写入者。

1) 模型筛选阶段:

a. 本地模型训练: proposer 会发布模型的基本结构, 并初始化参数。随后, acceptor 会利用本地数据对 proposer 的初始模型进行训练并在植入模型水印后将其广播。

b. 模型聚合: proposer 在收集到梯度模型后, 利用 Federated Averaging 算法对模型进行聚合并微调, 得到聚合模型后将其广播。

c. 模型打分承诺: 此时 acceptor 集合中的成员会对得到的聚合模型进行准确性验证, 将第一个收到的模型标记为 $\boldsymbol{N}_k$, 而随后的模型标记为 $\boldsymbol{N}_a$, 其中 a 表示该模型的发布者。按接收到的顺序对比 $\boldsymbol{N}_k$ 与 $\boldsymbol{N}_a$ 的性能, 如果 $\boldsymbol{N}_a$ 性能较好则计算承诺 $(\boldsymbol{N}_a, vote_k)$, 同时不再接收性能小于 $\boldsymbol{N}_a$ 的模型, 并

令 $\boldsymbol{N}_k = \boldsymbol{N}_a$; 否则向 a 返回消息 $(\boldsymbol{N}_k, k)$。

d.优秀模型筛选: proposer a 在收到 $(\boldsymbol{N}_k, k)$ 后也会将自己的模型 $\boldsymbol{N}_a$ 和 $\boldsymbol{N}_k$ 进行对比, 若 $\boldsymbol{N}_k$ 性能优于 $\boldsymbol{N}_a$, 则令 $\boldsymbol{N}_a = \boldsymbol{N}_k$, 并向 $\boldsymbol{N}_k$ 模型的生成者返回承诺 $(\boldsymbol{N}_k, vote_a)$。

2) 共识阶段:

e. 共识请求: 当 proposer 收到一半以上对聚合模型 $\boldsymbol{N}$ 的承诺后, 便开始广播包含 $\boldsymbol{N}$ 哈希值的一致性消息 *message*。

f. 共识认定: acceptor 在收到了 proposer 的 *message* 数据后, 对比自己的 $\boldsymbol{N}_k$ 是否等于 $\boldsymbol{N}$, 如果是则转换为 learner 成员从而退出模型打分承诺阶段, 广播 $(message, revotes_k)$; 否则, 不回复。

g. 区块写入: 当 learner 收到 $(message, revotes_i)$ 消息后, 核对其中相同的 *message* 是否已经超过了总节点数量的 1/2, 若是, 则将写入区块, 同时广播 $(end, message)$; 在收到全体成员 1/2 以上的 $(message, revotes_i)$ 后, 如果相同的 *message* 未超过总体成员的 1/2, 那么该 learner 将撤回其 learner 身份, 从新转换成 acceptor。

h. 协议结束: 当 learner 收到 $(end, message)$ 时, 本轮协议结束。

本文所提出的 PFconsensus 协议利用了聚合模型的性能优劣来选取最终的共识内容, 聚合过程的随机性与模型性能的有界性使得一段时间内只存在一个 proposer。相比于原始 paxos 协议中产生多个 proposer 的现象, 本文方案避免了活锁的出现。此外, 在保证去中心化特性的同时, 该算法也高效地利用了分布式设备的资源来优化训练及验证过程。各个节点利用本地数据集对聚合模型进行性能评测的方式, 促使被投毒、普适性弱的聚合模型难以被大多节点所接受, 在一定程度上解决了 Non-IID 问题及模型投毒问题。

以上述方案为基础, 下面进一步对 PFconsensus 协议在区块链环境中的实现进行了更为详细的设计。首先, 为了保证 PFconsensus 正常运行, 需根据区块链应用的实际情况定义以下函数(1 表示是, 0 表示否):

$\boldsymbol{W}_k \leftarrow Train_k(\boldsymbol{W}_{init}, ID_k)$: 表示节点 k 利用本地数据对发布的初始模型进行训练, 并通过 FedIPR 算法嵌入模型水印;

$\boldsymbol{N}_k \leftarrow FedAvg(\mathbb{M}_k, ID_k)$: 表示节点 k 从待选梯度集合 $\mathbb{M}_k$ 中随机选择 m 个及以上模型, 并通过联邦聚合算法得到聚合模型 $\boldsymbol{N}_k$;

$0 \text{ or } 1 \leftarrow VS(\boldsymbol{W}_k, \sigma_k)$: 表示验证 $\boldsymbol{W}_k$ 的水印和数字签名是否同时合法;

$0 \text{ or } 1 \leftarrow FC_k(\boldsymbol{W}_i)$: 表示节点 k 的待选梯度集合 $\mathbb{M}_k$ 中是否包含节点 i 的梯度模型;

$0 \text{ or } 1 \leftarrow V(\boldsymbol{N}_i, \sigma_i)$: 验证由节点 i 产生的 $\boldsymbol{N}_i$ 是否聚合了 m 个以上的梯度模型, 以及数字签名和模型水印签名是否能够通过验证;

Gossip(*message*): 表示利用 Gossip 协议广播消息 *message*;

$Append_k(\boldsymbol{W}_i)$: 表示节点 k 在他的待选梯度集合 $\mathbb{M}_k$ 中添加 $\boldsymbol{W}_i$;

根据 PFconsensus 协议, 区块链的运行主要包括模型生成和竞争写入算法两个部分。首先对模型的生成算法进行设计:

算法 1. $Block_{j+1}$ 上的模型训练与模型聚合算法。

定义第 j 个区块为 $Block_j$, 本轮参与节点所构成的集合记为 $\mathbb{G}_{j+1}$, 每一个节点都有各自的签名密钥对, $\boldsymbol{W}_{init}^{j}$ 为本轮优化目标的模型初始权值。

输入: $Block_j$, $\mathbb{G}_{j+1}$, $\boldsymbol{W}_{init}^{j}$, 各个节点的密钥对。

输出: 聚合模型 $\boldsymbol{N}$。

1 FOR $k \in \mathbb{G}_{j+1}$:

2 训练 $\boldsymbol{W}_k \leftarrow Train_k\left(\boldsymbol{W}_{init}^{j}, ID_k\right)$;

3 计算 $\sigma_k \leftarrow Sign\left(sk_k, \boldsymbol{W}_k\right)$;

4 运行 Gossip$\left(\left(\boldsymbol{W}_k, \sigma_k\right)\right)$;

5 定义 $num = 0$;

6 当 k 节点收到 i 节点的 $\left(\boldsymbol{W}_i, \sigma_i\right)$ 时:

7 **IF** $VS\left(\boldsymbol{W}_i, \sigma_i\right)$ 为 1:

8 则 $Append_k\left(\boldsymbol{W}_i\right)$;

9 **IF** $FC_k\left(\boldsymbol{W}_i\right)$ 为 0:

10 则 $num = num + 1$;

11 **IF** $num \geqslant m$:

12 计算 $\boldsymbol{N}_k \leftarrow FedAvg\left(\mathbb{M}_k, ID_k\right)$;

13 计算 $\sigma_k \leftarrow Sign\left(sk_k, \boldsymbol{N}_k\right)$;

14 运行 Gossip$\left(\left(\boldsymbol{N}_k, \sigma_k\right)\right)$;

15 END FOR

在梯度模型生成后, 各个节点需要对网络中的聚合模型进行评价, 并筛选出全局性能最好的模型。

最终, 生成该最优聚合模型的节点将指定下一轮协议的模型初始参数。算法 2a、2b 将具体描述如何实现筛选并达成共识。

算法 2a.模型评价及写入权限争夺。

定义由节点 k 生成的聚合模型为 $\boldsymbol{N}_k$, 集合 $\mathbb{G}_{j+1}$ 中共有 G 个节点。所有 proposer 节点构成集合 $\mathbb{G}_{j+1}$ 的子集 $\mathbb{P}_{j+1}$, 而 learner 组成的集合记为 $\mathbb{L}_{j+1}$, 显然 $\mathbb{L}_{j+1}$ 与 $\mathbb{P}_{j+1}$ 的交集为空。

输入: $\boldsymbol{N}_k$, $\mathbb{P}_{j+1}$, $\mathbb{L}_{j+1}$。

1 FOR $k \in \mathbb{P}_{j+1}$:

2 令 $\boldsymbol{BN}_k = \boldsymbol{N}_k$;

3 $votes = 0$;

4 当收到 i 节点的 $\left(\boldsymbol{W}_i, \sigma_i\right)$ 时:

5 **IF** $V\left(\boldsymbol{N}_i, \sigma_i\right)$ 和 $Cont\left(\boldsymbol{BN}_k, \boldsymbol{N}_i\right)$ 均 1:

6 令 $\boldsymbol{BN}_k = \boldsymbol{N}_i$, 并发送 $vote_k$ 给 i;

7 当收到 i 节点的 $vote_i$ 时:

8 $votes = 1 + votes$;

9 **IF** $votes \geqslant G/2$:

10 构造下一轮模型初始权值 $\boldsymbol{W}_{init}^{j+1}$;

11 计算 $\sigma_k \leftarrow Sign\left(sk_k, \left(\boldsymbol{N}_k, \boldsymbol{W}_{init}^{j+1}\right)\right)$;

12 令 $messages_k = \left(\boldsymbol{N}_k, \boldsymbol{W}_{init}^{j+1}, \sigma_k\right)$;

13 运行 $\text{Gossip}\left(messages_k\right)$;

14 当收到 i 节点的 $message_i$ 时:

15 **IF** $V\left(messages_i\right) = 1$且$\boldsymbol{BN}_k == \boldsymbol{N}_i$:

16 离开组群 $\mathbb{P}_{j+1}$, 加入 $\mathbb{L}_{j+1}$;

17 运行 $\text{Gossip}\left(revote_i\right)$;

18END FOR

算法 2a 表明当节点作为一个 proposer 时, 会进行对聚合模型的评价并争夺区块的写入权限。算法 2a 可以与算法 1 在节点 k 上并发执行, 当 proposer 在算法 2a 中转换为 learner 角色后, 将停止执行这两个, 其目的在于确保模型数据写入区块时唯一。

算法 2b.共识及区块写入算法。

输入: $\mathbb{L}_{j+1}$。

1 FOR $k \in \mathbb{L}_{j+1}$:

2 在本地初始化所有 $revotes_i = 0$;

3 当收到对 i 节点的认定 $revotes_i$:

4 $revotes_i = 1 + revotes_i$;

5 **IF** $revotes_i > G/2$:

6 令 $messages_{end} = messages_i$;

7 运行 $\text{Gossip}\left(messages_{end}\right)$;

8 END FOR

算法 2b 的目的是确定大多数节点皆认为 是最优模型并已做好写入区块链中的准备。从算法 2a、2b 可以发现, 若节点由 proposer 转化成 learner 角色, 将会失去发布新模型和竞争写入权限的能力, 可能因时延而导致模型无法得到一半以上节点的认定, 从而协议失败。为此, 需要设计活性算法以解决该问题。

算法 3.活性算法。

输入: $\mathbb{P}_{j+1}$, $\mathbb{L}_{j+1}$。

1 FOR $k \in \mathbb{L}_{j+1}$:

2 令 $revotes = 0$;

3 FOR $i \in \mathbb{G}_{j+1}$:

4 计算 $revotes = revotes_i + revotes$;

5 **END FOR**

6 **IF** $revotes_i \leqslant G/2$ 且 $revotes > G/2$:

7 则离开组群 $\mathbb{L}_{j+1}$, 加入 $\mathbb{P}_{j+1}$;

8 END FOR

当节点转换成一个 learner 角色后将运行算法 3, 这可以防止由于大部分节点被激发成为 learner 后所出现的死锁现象。

3.3 参与者贡献度评价算法

通常评价联邦学习中梯度模型对最终模型的贡献度时, 往往会将其梯度模型从最终模型中剔除, 将剔除后模型在测试集上的性能差异作为贡献值。然而, 由于在分布式的竞争环境下往往不存在普遍认可的测试集, 因此本文将设计一种利用梯度模型和最终模型参数距离来换算贡献度的方法。该方法能够在各节点数据隐私得到保护的同时获取一个令人信服的贡献度指标。

记聚合模型为 $\boldsymbol{N}_{end}$, 聚合方式采用 Federated Averaging 算法:

$$\boldsymbol{N}_{end} \leftarrow \sum_{K=1}^{K} \frac{n_k}{n} \boldsymbol{W}_k, \tag{9}$$

其中, $\boldsymbol{W}_k \leftarrow CilentUpdate(n, \boldsymbol{W})$。

那么评价一个梯度模型 $\boldsymbol{W}_k$ 的贡献程度 C_k 需要

采用以下两个步骤:

$$\theta_k = \frac{\langle \boldsymbol{N}_{end}, \boldsymbol{W}_k \rangle}{|\boldsymbol{N}_{end}| \cdot |\boldsymbol{W}_k|}, \tag{10}$$

$$C_k = \frac{\theta_k}{\sum_{i=0}^{n} \theta_i}。 \tag{11}$$

通过计算聚合模型和不同梯度模型之间的夹角大小即可衡量它对整体的贡献度。

4 系统正确性及安全性分析

如果上文所设计的系统能够满足 2.2 节中所给出的安全性和有效性定义, 则说明本文的整体系统在实际运行中是安全有效的。在本章节中, 将先结合共识协议的活性对第三章中所设计的PFconsensus协议进行正确性分析。此后, 将围绕2.2节中的攻击模型和安全性定义对本文所设计的区块链系统进行安全性和有效性的形式化证明。

4.1 PFconsensus 协议正确性分析

PFconsensus 协议在本质上是基于聚合模型的性能来争夺写入权, 主要包括模型性能筛选和共识达成两方面的内容。下面将围绕该协议在攻击环境下的节点活性及共识结果的唯一性来进行讨论。

设定 1.设本轮参与节点构成的集合为 $\mathbb{G}$, $\mathbb{G}$ 的子集 $\mathbb{L}$ 为所有 learner 节点。

设定 2. $\mathbb{R} \subseteq \mathbb{L}$ 表示本轮认定聚合模型 $\boldsymbol{N}$ 的节点集合, 当集合 $\mathbb{R}$ 中节点个数超过 $\mathbb{G}$ 中节点数量的一半时 $\boldsymbol{N}$ 为本轮最终模型 $\boldsymbol{N}_{end}$。

显然算法 2a、2b 可满足以上设定, 但为了保证在当前轮得到唯一的聚合模型, 需要在 PFconsensus 协议中引入下面的约束条件:

约束 1.任意 learner 节点只能认定唯一的聚合模型 $\boldsymbol{N}$。

断言 1.在约束条件 1 下, 每一轮协议中不存在两个不同的 $\boldsymbol{N}_{end}$。

证明. 假设某轮协议产生出多个 $\boldsymbol{N}_{end}$, 那么至少有一个节点同时认定了两个或以上不同的 $\boldsymbol{N}$, 与约束 1 矛盾, 因此断言 1 成立。

约束 1 只能确保在每一轮协议结束后至多获得一个 $\boldsymbol{N}_{end}$, 但实际运行过程中还须保证至少获得一个 $\boldsymbol{N}_{end}$。因此, 需要进一步对约束条件 1 进行加强。

约束 1a.在时间段 t 内, 所有节点都能收到相同的待选聚合模型集合 $\mathbb{N} = \{\boldsymbol{N}_i \mid i \in \mathbb{G}\}$, 该集合中最优模型 $\boldsymbol{N}_{best}$ 存在且唯一。

约束 1b.任意节点在时间段 t 内节点只能认定一个性能最优模型 $\boldsymbol{N}_{best}$。

断言 2.若一轮协议在时间段 t 内完成, 那么在同时满足约束条件 1a、1b 的情况下不可能产生出多个不同的 $\boldsymbol{N}_{best}$。

证明. 假设某一轮协议产生出多个 $\boldsymbol{N}_{end}$, 那么必然存在至少一个节点同时对两个以上的不同聚合模型进行了认定, 与约束 1a、1b 矛盾, 因此断言 2 成立。

断言 3.若一轮协议在时间段 t 内完成, 在满足约束条件 1a、1b 的情况下, 必然能够得到唯一的 $\boldsymbol{N}_{end}$。

证明. 假设某一轮协议未产生出任何的 $\boldsymbol{N}_{end}$, 则存在一半以上的节点选择了不同的性能最优模型, 这与约束 1a 矛盾, 因此断言 3 成立。

至此, 已经证明了当约束条件 1a 和 1b 同时成立时, 协议必然能够得到唯一的 $\boldsymbol{N}_{end}$。但在分布式环境下, 由于联邦学习节点的本地数据不同, 需要进一步分析约束 1a 是否成立。

首先, 将约束条件 1a 转换成: 在时间段 t 内, 存在一半以上的节点的候选聚合模型集合 $\mathbb{N}$ 中包含相同的最优模型 $\boldsymbol{N}_{best}$。

在联邦学习过程中, 大多聚合模型性能评测机制都只考虑了数据独立同分布(Independently Identically Distribution, IID)的情况[39], 而本文采用的是本地评测方案, learner 节点 k 将利用本地数据 $data_k$ 作为测试集来评价聚合模型 $\boldsymbol{N}$ 的准确率、召回率等。这种评测方案在数据分布为 IID 的条件下具有较好的客观性, 筛选过程中大部分的节点评测结果也比较相近, 可确保约束 1a 成立。然而, 当分布为 Non-IID 的情况, 由于本地模型对本地数据的预测情况会更好, 必会导致不同数据分布下的节点间评测结果存在较大差异。考虑到大部分节点对 $\boldsymbol{N}$ 的评测结果存在较大差异时也说明 $\boldsymbol{N}$ 的普适性较差, 因此在 Non-IID 条件下若存在一个令一半以上节点都认可的聚合模型 $\boldsymbol{N}_{best}$, 则 $\boldsymbol{N}_{best}$ 可认为具有较好的普适性。

假设 PFconsensus 协议在对模型进行训练的过程中节点间的数据分布主要存 IID 和 Non-IID 这两种情况, 下面分别对这两种条件下 $\boldsymbol{N}_{best}$ 的产生概率进行分析。

设定 3.记第 j 轮参与节点集合为 $\mathbb{G}_j$, $\mathbb{G}_j$ 中节点个数为 G, 对 $\forall k \in \mathbb{G}_j$ 其本地数据 $data_k$ 服从分布 χ_k, 由该数据训练得到的梯度模型为 $\boldsymbol{W}_k$, 当采用 Federated Averaging 算法对 $\{\boldsymbol{W}_k \mid k \in \mathbb{G}_j\}$ 进行聚合时, 根据选取出的梯度模型集合可获得一个相应的聚合

模型 $\boldsymbol{N}_k$。

该设定可由算法 1 满足, 据此有模型 $\boldsymbol{N}_k$ 被一半以上节点认定为 $\boldsymbol{N}_{best}$ 的如下结论。

定理 1.记聚合模型 $\boldsymbol{N}_k$ 在节点 i 的数据集 $data_k$ 上的性能评价为 $F(\boldsymbol{N}_k, i)$, 并约定当 $\boldsymbol{N}_k$ 聚合了 $\boldsymbol{W}_i$ 时 $F(\boldsymbol{N}_k, i)=1$, 否则, $F(\boldsymbol{N}_k, i)=0$, 对于 $\forall a, b\in\mathbb{G}_j$, $\chi_a\neq\chi_b$, 即 non-IID 的极端情况下, 模型 $\boldsymbol{N}_k$ 被一半以上节点认为是 $\boldsymbol{N}_{best}$ 的概率为:

$$P(\boldsymbol{N}_k)=1-\sum_{i=0}^{G/2}\binom{G-1}{i}\left(\frac{1}{4}\right)^i\left(\frac{3}{4}\right)^{G-i-1}。\tag{12}$$

证明. 对节点 i 而言, 当 $F(\boldsymbol{N}_i, i)<F(\boldsymbol{N}_k, i)$ 时, 性能比较函数 $Cont(\boldsymbol{N}_i, \boldsymbol{N}_k)=1$, 可以发现在节点收到模型进行比较的过程中有:

$$\begin{aligned}P(\boldsymbol{N}_k)&=P((\sum_{i=0}^{G-1}Cont(\boldsymbol{N}_i, \boldsymbol{N}_k))\geqslant\frac{G}{2})\\&=1-F(\frac{G}{2};\ G-1,\ P(i\in\boldsymbol{N}_k,\ i\notin\boldsymbol{N}_i))\\&=1-F(\frac{G}{2};\ G-1,\ \frac{1}{4})\\&=1-\sum_{i=0}^{G/2}\binom{G-1}{i}\left(\frac{1}{4}\right)^i\left(\frac{3}{4}\right)^{G-i}。\end{aligned}\tag{13}$$

其中 $\binom{n}{k}=\prod_{i=1}^{k}\frac{n-(k-1)}{i}$, $F(x;\ n,\ p)$ 为二项分布的累积概率函数。由此得证。

定理 2.记聚合模型 $\boldsymbol{N}_k$ 在节点 i 的数据集 $data_i$ 上的性能评价为 $F(\boldsymbol{N}_k, i)$, 当节点之间数据分布为 IID 时, 即 $\forall a, b\in\mathbb{G}_j$, $\chi_a=\chi_b$, $\boldsymbol{N}_k$ 被一半以上节点认为是 $\boldsymbol{N}_{best}$ 的概率为:

$$P(\boldsymbol{N}_k)=1-\sum_{i=0}^{\left|\frac{G}{2}\right|}\binom{G-1}{i}\left(\frac{1}{2}\right)^{G-1}。\tag{14}$$

证明. 当节点数据为同分布时, 每个节点对 $\boldsymbol{N}_k$ 的评价相同, 即满足:

$$F(\boldsymbol{N}_k, i)=F(\boldsymbol{N}_k, k),\ k\in\mathbb{G}_j,\tag{15}$$

此时 $P(Cont(\boldsymbol{N}_i, \boldsymbol{N}_k)=1)\approx 1/2$。一半及以上节点认为 $\boldsymbol{N}_k$ 不是 $\boldsymbol{N}_{best}$ 的概率可写为累积概率二项分布:

$$F(\frac{G}{2};\ G-1,\ \frac{1}{2})=\sum_{i=0}^{G/2}\binom{G-1}{i}\left(\frac{1}{2}\right)^i\left(\frac{1}{2}\right)^{G-i-1}。\tag{16}$$

由此得证。

通过对节点数据集概率分布的极端关系进行分析, 已经得到两种不同的 $P(\boldsymbol{N}_k)$ 结果, 下面进一步对联邦学习场景中实际存在的数据分布情况进行讨论。联邦学习场景下主要存在五种 Non-IID 的数据分布情况[40]。为此, 需要采用更加详细地定义各节点对模型的性能评价公式。设节点 a 和节点 b 的数据集分别服从 $p(x_a\mid y_a)\sim\chi_a$ 和 $p(x_b\mid y_b)\sim\chi_b$ 分布, 那么两节点数据概率分布的交叉熵函数定义为:

$$\begin{aligned}&\mathcal{H}(p_a,\ q_b)=\mathcal{H}(p(x_a\mid y_a),\ p(x_b\mid y_b))\\&=\mathcal{H}(p(x_a\mid y_a))+D_{KL}(p(x_a\mid y_a)\|\ p(x_b\mid y_b))\end{aligned}。\tag{17}$$

其中, $\mathcal{H}(\rho(x_a\mid y_a))$ 表示 $p(x_a\mid y_a)$ 的熵值, 而 $D_{KL}(p(x_a\mid y_a)\|\ p(x_b\mid y_b))$ 表示 $p(x_a\mid y_a)$ 相对于 $p(x_b\mid y_b)$的相对熵值。假定当 $\partial\leqslant\varphi=\boldsymbol{H}(p_a,\ q_b)$ 时, 可认为 a 和 b 的数据独立同分布, 并定义 a与$b\in\mathbb{k}_l$, 其中 $\mathbb{k}_l$ 为数据独立同分布的节点所构成的集合。由此, 进一步定义性能评价函数:

$$F(\boldsymbol{N}_k,\ a)=\underbrace{\alpha(\varphi)\cdot f(x_k^l)\cdot\theta}_{\boldsymbol{N}_k\text{ include }\mathbb{k}_l\text{'s }\boldsymbol{W}}+\underbrace{(1-\theta)\cdot f(x_k^l)}_{\boldsymbol{N}_k\text{ not include }\mathbb{k}_l\text{'s }\boldsymbol{W}}。\tag{18}$$

其中, $\alpha(\varphi)$ 表示一个与 φ 负相关的函数, 其取值范围为[0, 1], θ 取{0, 1}用于表明 $\boldsymbol{N}_k$ 是否聚合了集合 $\mathbb{k}_l$ 内的梯度模型, $f(x_k^l)$ 表示 x_k^l 个 $\mathbb{k}_l$ 集合中的梯度模型对准确率的整体贡献大小。根据 $F(\boldsymbol{N}_k,\ a)$, 可以进一步得到该评价值的概率函数 $P(F(\boldsymbol{N}_k,\ a))$。为进一步简化, 本文将 $\alpha(\varphi)$ 固定为常数, 从而在性能评价时, 模型在第一个节点(记为 a)的评价值概率函数为:

$$P_1(X_1=F(\boldsymbol{N}_k,\ a))=P_1(Y_1=x_k^a)。\tag{19}$$

当模型被第二个节点(记为 b)进行评价, 由于梯度模型的选取方式要求评价结果与前一个节点相关, 因此得到:

$$\begin{aligned}&P_2(X_2=F(\boldsymbol{N}_k,\ b)\mid F(\boldsymbol{N}_k,\ a))\\&=P_2(Y_2=x_k^b\mid Y_1=x_k^a)\end{aligned}。\tag{20}$$

依此类推, $\boldsymbol{N}_k$ 在下一个节点上的评价结果会受前面节点的影响, 而这也正好反映了算法 1 中所描述的随机模型聚合方式。为便于论证, 下面先假定计算 $\boldsymbol{N}_k$ 的梯度模型选取方式已知, 据此可求解在一般的 Non-IID 情况下 $P(\boldsymbol{N}_k)$ 概率, 并自然地引申出任意 Non-IID 情况下的 $P(\boldsymbol{N}_k)$ 边界。

假设梯度模型的选取方式已知, 相应地在算法 1 中引入如下设定。

设定 3. $\mathbb{G}_j$ 中共有 $2l$ 组独立同分布的节点集合 $\mathbb{k}_1, \mathbb{k}_2, ..., \mathbb{k}_{2l}$, 且每个集合的元素个数满足 $Card(\mathbb{k}_1)=...=Card(\mathbb{k}_{2l})=n$。对 $\boldsymbol{N}_k$ 而言, 它需要聚合 G/2 个梯度模型 $\{\boldsymbol{W}_i\mid i\in\mathbb{G}_j\}$, 令 $g_1, g_2, ..., g_{2l}$ 分别

表示 $\boldsymbol{N}_k$ 从每组独立同分布节点集合中选取的梯度模型个数。

定理 3.根据设定 3, $\boldsymbol{N}_k$ 被一半以上节点认为是 $\boldsymbol{N}_{best}$ 的概率满足:

$$P(\boldsymbol{N}_k)\geqslant 1-\sum_{i=0}^{G/2}\binom{G-1}{i}\left(\frac{1}{4}\right)^i\left(\frac{3}{4}\right)^{G-i},n>2。\quad(21)$$

证明. 在设定 3 中, 若 $\boldsymbol{N}_k$ 通过随机均匀地从每个集合 $\mathbb{k}_z(1\leqslant z\leqslant 2l)$ 选取 $n/2$ 个梯度模型来生成, 则可以得到 $Cont(\boldsymbol{N}_i, \boldsymbol{N}_k)$, 其值等价于 x_i^z 与 x_k^z 的大小比较。已知 $\boldsymbol{N}_k$ 聚合了任意分布中的 $n/2$ 个梯度模型, 那么 $P(Cont(\boldsymbol{N}_i, \boldsymbol{N}_k)=1)=P(x_i^z<(n/2))$, 而该概率服从超几何分布 $X\sim HD(G/2, n, G)$, 因此可以进一步细化每一个节点上的性能对比函数:

$$P(Cont(\boldsymbol{N}_i, \boldsymbol{N}_k)=1)=F(\frac{n}{2}-1;\ \frac{G}{2}, n, G)$$

$$=1-\frac{\binom{G/2}{n/2}\binom{G/2}{n/2}}{\binom{G}{n}}\,{}_3F_2\left[\begin{matrix}1, & -\frac{n}{2},\frac{n}{2}-\frac{G}{2}\\ \frac{n}{2}+1, & \frac{G}{2}+\frac{n}{2}+1\end{matrix};1\right]。\quad(22)$$

其中 $F(n/2-1; G/2, n, G)$ 表示超几何分布的概率累积函数, ${}_pF_q$ 代表广义超几何函数。根据二项分布可以推测模型 $\boldsymbol{N}_k$ 被一半以上节点认为是 $\boldsymbol{N}_{best}$ 的概率为:

$$P(\boldsymbol{N}_k)=1-F(\frac{G}{2};\ G,\ F(\frac{n}{2}-1;\ \frac{G}{2}, n, G)),\quad(23)$$

进一步可以得到:

$$P(\boldsymbol{N}_k)\geqslant 1-\sum_{i=0}^{G/2}\binom{G}{i}\left(\frac{1}{4}\right)^i\left(\frac{3}{4}\right)^{G-i},\ n>2。\quad(24)$$

从而发现, 对定理 3 中的 $2l$ 值进行缩放可构造 Non-IID 的所有数据分布情况, 此时, 最终 $P(\boldsymbol{N}_k)$ 的取值位于定理 2 和定理 1 之间。由上述结论可知, 本文的方案符合约束 1a。并且函数 $P(\boldsymbol{N}_k)$ 的大小与节点个数 G、认定票数 u 在一定程度上表现为阶乘级关系。此时, 可以通过控制参数 G 和 u 来调整全局最优模型的生成概率。

就通信性能而言, 本文将对协议的模型训练与筛选两个阶段进行分析。模型训练阶段包括本地数据训练与梯度发布两个过程, 由此可得该阶段的通信复杂度为 $T_1=O(G^2)$。而在筛选阶段, 可由公式(24)推导出通信复杂度为:

$$T_2=O((1-\alpha(\varphi))(\frac{1}{2}\frac{G!}{(G-u)!}))+\omega,\quad(26)$$

其中 $\alpha(\varphi)$ 表示一个与 φ 负相关的函数, u 为模型上链需要的认定票数, ω 为网络常数。以至得到协议的整体通信复杂度为:

$$T=T_1+T_2。\quad(27)$$

至此, 证明完成了协议的正确性并推导出协议执行的通信复杂度。

4.2 PFconsensus 安全性分析

PFconsensus 协议是否安全关键在于数据的一致性及协议运行的活性。针对参与共识的节点集合 $\mathbb{G}$, 首先根据定义 3 给出 PFconsensus 协议在拜占庭环境下的安全性命题:

命 题 1. 在 $Card(\mathbb{A})<Card(\mathbb{G})/2$ 的情况下, PFconsensus 协议能在有限次通信内得到 $\boldsymbol{N}_{end}$, 且对任意节点 $k\in\mathbb{G}, k\notin\mathbb{A}$, 它们得到的最终模型 $\boldsymbol{N}_{end}$ 相同, 则表明 PFconsensus 协议满足安全性。

为便于讨论, 下面将拜占庭节点从整体上划分为宕机节点与恶意扰乱节点两种身份。显然, 某一节点不可能扮演两种身份。

定义 7.宕机节点.

对宕机节点 $k\in\mathbb{A}$, 它们均不回复所有诚实节点 $c\notin\mathbb{A}$ 所发送消息, 也不自主发送任何消息至 c。

断言 4.在拜占庭节点扮演宕机身份时, PFconsensus 协议满足安全性。

证明. 由宕机节点的定义可知, 参与共识的最终节集合为 $\mathbb{B}=\mathbb{G}-\mathbb{A}$, 而完成共识的通信复杂度是:

$$T=O(Card(\mathbb{B})^2)+O((1-\alpha(\varphi))(\frac{1}{2}\frac{Card(\mathbb{B})!}{(G/2)!}))。\quad(28)$$

由于 $Card(\mathbb{B})<G$, 因此必然能够在有限次通信内得到 $\boldsymbol{N}_{end}$。而又由于宕机节点未发送任何消息至诚实节点, 因而诚实节点的最终模型 $\boldsymbol{N}_{end}$ 必然相同。断言 4 成立。

在实际情况下, 拜占庭节点更可能扮演恶意扰乱节点。因此, 下面将考虑拜占庭节点扮演恶意扰乱节点时的安全问题。

定义 8.恶意扰乱节点.

所有恶意扰乱节点 $k\in\mathbb{A}$, 它们必须回复诚实节点 $c\notin\mathbb{A}$ 所发送询问, 并可以随意发送任意信息到 c。

断言 5.在拜占庭节点扮演恶意扰乱身份时, PFconsensus 协议满足安全性。

证明. 假设 PFconsensus 协议无法完成, 这意味

着不存在模型 $\boldsymbol{N}_k$ 被一半以上的节点认定。但由于 $Card(\mathbb{A}) < Card(\mathbb{G})/2$, 而诚实节点又以不可忽略概率 $P(\boldsymbol{N})$ 得到了一致的最终模型 $\boldsymbol{N}_{end}$, 因而与上述假设矛盾, 所以协议必然能够完成。此外, 根据协议的要求, 任意节点 $k \in \mathbb{G}$ 上均存在一个认定票数集合: $\{Votes_1, ..., Votes_G\}$, 其中 $Votes_a$ 表示聚合模型 $\boldsymbol{N}_a$ 所获得的认定票数。根据算法 3, 在 $\sum_{i=1}^{G} Votes_i > G/2$ 时, 节点会重新对比所收到的聚合模型而再次认定。如果 PFconsensus 协议完成时诚实节点之间的最终模型 $\boldsymbol{N}_{end}$ 不同, 则必然存在节点 $k \in \mathbb{G}$ 对多个模型进行了认定, 且存在至少两个的诚实节点 n_1和$n_2 \notin \mathbb{A}$ 分别对不同的模型 $\boldsymbol{N}_a$ 与 $\boldsymbol{N}_b$ 进行了认定。但由于诚实节点 n_1 需要按照协议正常运行, 节点 n_2 必然能够收到节点 n_1 的认定内容, 此时 n_2 会对 n_1 认定的模型进行性能对比。而根据待融合梯度模型筛选式(2)和(3)所限制的时间条件, 必然能够保证 n_1 与 n_2 同时收到了 $\boldsymbol{N}_a$ 和 $\boldsymbol{N}_b$, 这意味着其中一个必然聚合了恶意梯度 $\boldsymbol{W}_k'$。此时由于算法 3 的条件, n_1 和 n_2 会重新对比 $\boldsymbol{N}_a$ 和 $\boldsymbol{N}_b$ 的性能并只会认定未包括 $\boldsymbol{W}_k'$ 的聚合模型, 从而导致 n_1 与 n_1 的认定相同, 与假设矛盾。因此, 诚实节点的最终模型 $\boldsymbol{N}_{end}$ 唯一, 断言 5 成立。

在上面的论证中, 拜占庭环境中的节点被分割成两种互补类型进行论证, 且已证明本文协议在这两种条件下皆满足定义 6。结合这两种攻击情况, 可以得到如下结论。

断言 6.在任意拜占庭环境下, PFconsensus 协议能满足定义 6 中的安全性要求。

证明. 记所有诚实节点组成得集合为 $\mathbb{H}$, 满足 $\mathbb{G} = \mathbb{A} + \mathbb{H}$。由于拜占庭节点 $k \in \mathbb{A}$ 无法同时扮演宕机节点和恶意扰乱节点, 此时, 记扮演宕机节点的拜占庭节点集合为 $\mathbb{A}_{I1}$, 扮演恶意扰乱节点的拜占庭节点集合为 $\mathbb{A}_{I2}$。显然集合 $\mathbb{G}_1 = \mathbb{A}_1 + \mathbb{H}$ 满足断言 4, $\mathbb{G}_2 = \mathbb{A}_2 + \mathbb{H}$ 满足断言 5, 且 $\mathbb{G}_1 \cap \mathbb{G}_2 = \mathbb{H}$。因此, $\boldsymbol{N}_{end}$ 必然唯一, 断言 6 成立。

4.3 联邦学习算法有效性分析

联邦学习算法在嵌入到 PFconsensus 协议后, 应保证模型性能不低于对应的中心化方案, 并确保去中心化场景下抵抗投毒攻击的能力。因此, 下面将对本文联邦学习算法在去中心化环境中的有效性进行分析和论证。

记节点 k 产生的梯度模型为 $\boldsymbol{W}_k$, 且在满足水印验证 $V_w(\boldsymbol{W}_k, (B_k, \theta_k)) = True$ 和签名验证 $Vrfy(\sigma, H(\boldsymbol{W}_k), pk) = True$ 时模型合法。同样, 记恶意梯度为 $\boldsymbol{W}_k'$, 在满足上述验证要求时同样合法。

断言 7. 若投毒攻击节点个数满足, $Card(\mathbb{A}) < Card(\mathbb{G}) - m$ 且 $Card(\mathbb{G})$ 有限, 那么最终的上链模型 $\boldsymbol{N}_{end}$ 以 1-p 的概率含有恶意梯度 $\boldsymbol{W}'$。

证明. 记含有恶意梯度的聚合模型为 $\boldsymbol{N}'$, 根据公式(3), 任意一个诚实节点必然收到至少 m 个正常梯度模型 $\boldsymbol{N}$。定义均匀随机聚合 $G/2$ 个梯度模型得到的聚合模型为 $\boldsymbol{N}$。假设未含有恶意梯度的聚合模型被其他诚实节点认定的概率为 $Pr(\boldsymbol{N})$, 而含有恶意梯度的聚合模型被其他诚实节点认定的概率为 $Pr(\boldsymbol{N}')$, 那么根据定义 4 与算法 2a 中的性能对比情况, 恶意梯度模型对聚合模型的影响满足:

$$Pr(\boldsymbol{N}') < Pr(\boldsymbol{N}), \tag{29}$$

此时, 在 $Card(\mathbb{G}) - Card(\mathbb{A})$ 个诚实节点中以概率 p 产生了模型 $\boldsymbol{N}$。若产生了 $\boldsymbol{N}$, 则由断言 6 表明模型 $\boldsymbol{N}_{end} \neq \boldsymbol{N}'$。因此可得:

$$\begin{aligned} &\max(P(\boldsymbol{N}_{end} = \boldsymbol{N}')) = 1 - p \\ &= 1 - (Card(\mathbb{G}) - Card(\mathbb{A})) \frac{m!(Card(\mathbb{G}) - m)!}{Card(\mathbb{G})!}, \end{aligned} \tag{30}$$

当 $m \leqslant Card(\mathbb{A})$ 且 $Card(\mathbb{G})$ 有限时, 全局模型 $\boldsymbol{N}$ 以以 1-p 的概率含有恶意梯度 $\boldsymbol{W}'$, 断言 7 成立。

为降低上述概率, 下面在算法 1 的基础上引入额外的约束条件。

约束 7.对任意诚实节点 $k \notin \mathbb{G} - \mathbb{A}$, 在式(2)中的 $\overline{mt^{train}}$ 时间内只会接收一次由节点 $a \in \mathbb{G}$ 所发布的梯度模型, 且诚实节点在生成聚合模型 $\boldsymbol{N}$ 后将预先评价它在本地数据集上的性能。

断言 8.根据约束条件 7, 当投毒攻击节点个数满足 $Card(\mathbb{A}) < Card(\mathbb{G}) - m$ 且 $Card(\mathbb{G})$ 有限时, 最终上链的全局模型 $\boldsymbol{N}_{end}$ 以可忽略的概率包含恶意梯度 $\boldsymbol{W}'$。

证明. 由定义 2 可知每个节点 k 必然收到至少一个其他节点 $j \in \mathbb{G}, j \neq k$ 所发布的梯度模型, 且 $\overline{t^{train}} \geqslant \overline{t^{w}} \gg \overline{t^{avg}}$。此时令:

$$\frac{\overline{t^{w}}}{\overline{t^{avg}}} > \frac{(1-\varepsilon)G!}{m!(G-m)!}, \tag{31}$$

其中ε为任意小实数, 则 $\max(P(\boldsymbol{N}_{end} = \boldsymbol{N}')) < \varepsilon$。假设断言 8 不成立, 那么必然有 $Pr(\boldsymbol{N}') \geqslant Pr(\boldsymbol{N})$, 与定义 4 得到的式(29)矛盾, 因此断言 8 成立。

上述过程证明了本文方案中的上链全局模型 $\boldsymbol{N}_{end}$ 以可忽略的概率含有恶意梯度，即在满足 $Card(\mathbb{A}) < Card(\mathbb{G}) - m$ 条件时，投毒攻击难以对联邦学习造成威胁。

就模型性能而言，本文在协议设计的过程中采用式(9)作为聚合算法，该算法与常用的联邦学习算法[1]相同。因此，在 FedIPR 水印的使用不影响全局模型性能的条件下，本文方案能够在去中心化环境中满足联邦学习算法的准确性要求。

4.4 联邦学习区块链系统安全性分析

记链上的正常区块数据为 $Block$，拜占庭节点伪造的区块数据为 $Block'$，满足 $Block' \neq Block$。定义第 i 块区块被篡改的概率为 $P(VB(Block_i') = True)$。其中 $VB(Block_i') = True$ 表示 $Block_i'$ 被认定合法。其中 $Block$ 主要包含的数据如图 4 所示，即

(1) 上一个区块的哈希;

(2) 上链模型 $\boldsymbol{N}_{end}$;

(3) 下一轮协议的任务目标初始模型 $\boldsymbol{N}_{init}$;

(4) 本轮聚合模型 $\boldsymbol{N}_{end}$ 中所包含的梯度模型集合 $\mathbb{M}$;

(5) learner 节点的签名集合 $\mathbb{S}$。其中签名集合 $\mathbb{S} = \{s_1, s_2, ..., s_k\}$。

下面首先对区块的合法性判断进行定义。

定义 9.合法区块.

$VB(Block_i) = Ture$ 需同时满足以下 5 个条件:

(1) $\forall \boldsymbol{W}_k \in \mathbb{M}, V_w(\boldsymbol{W}_k, (B_k, \theta_k)) = V_w(\boldsymbol{N}_{end}, (B_k, \theta_k)) = True$;

(2) $\exists j \in \mathbb{G}, V_w(\boldsymbol{N}_{init}, (B_j, \theta_j)) = True$;

(3) $\forall s_a \in \mathbb{S}, Vrfy(s_a, Hash(\boldsymbol{N}_{end} \mid \boldsymbol{N}_{init} \mid \mathbb{M} \mid H_{i-1}), pk_a) = True$，其中 $Vrfy()$ 表示签名验证算法，$Hash()$ 表示哈希函数，“|”表示级联符号，当 $i \neq 0$ 时，则 $H_{i-1} = Hash(Block_{i-1})$，否则 $H_{i-1} = 0$;

(4) $Card(\mathbb{S}) > Card(\mathbb{G})/2$

(5) 所在链满足最长链原则。

从系统的整体结构而言，所采用的签名机制、散列算法、水印技术必然要满足以下给定的几个条件:

条件 1. $\forall e \in \{0, 1\}^b, \exists e' \in \{0, 1\}^a, e \neq e'$，其中 $\{0, 1\}^b$ 表示 b 位长的字符串，ε为任意小可忽略实数，由于散列函数的非碰撞性，必然满足 $P(Hash(e) = Hash(e')) < \varepsilon$。

条件 2.设本文所采用的签名方案 $\pi = (Gen, Sign, Vrfy)$ 在适应性消息攻击下满足不可伪造性[41]。

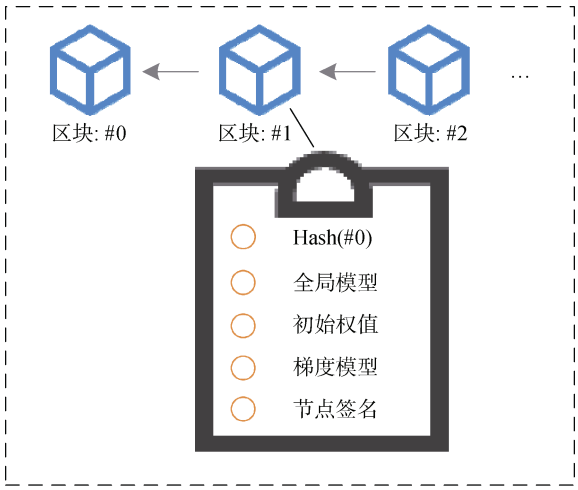


图 4 区块数据

Figure 4 Data on the block

条件 3.本系统采用的模型水印算法具有较高的鲁棒性，强行去除后会损坏模型的保真度。

条件 4.聚合模型的生成速度与节点性能成正比，并记所有拜占庭节点的算力之和为 Pw'，而所有诚实节点的算力之和为 Pw。即使拜占庭节点在伪造最后一个区块 $Block_n'$ 后产生了分叉，但大部分诚实节点仍然将训练由它们所构造的 $Block_n$ 上的初始模型。

为清晰描述拜占庭环境下联邦学习区块链的整体安全性要求，文中方案的安全性将被拆分为下面几个断言来进行论证。

断言 9. 设一条合法的区块链 $\mathbb{C}$ 是由区块 $Block_0, Block_i, ..., Block_n$ 所构成的集合，取其中一个区块 $Block_i \in \mathbb{C}/\{Block_n\}$。在拜占庭攻击环境下，若 $Card(\mathbb{A}) < Card(\mathbb{G})/2$，则由拜占庭节点伪造的区块能够满足定义 9 中所有条件的概率为 $P(VB(Block_i') = True) < \varepsilon$，其中$\varepsilon$是可忽略的任意小实数。

证明. 假设断言 9 为假，即伪造的区块能满足 $VB(Block_i') = True$，那么根据定义 9 必然存在 $Block_{i+1}$ 记录了 $H_i = Hash(Block_i)$。由于 $Block_i \neq Block_i'$，若 $\exists H_i' \neq H_i$，且该 H_i' 有效，必满足 $Vrfy\left(s_a, \left(Hash\left(\boldsymbol{N}_{end} \mid \boldsymbol{N}_{init} \mid \mathbb{M} \mid H_i'\right)\right), pk_a\right) = True$，这与条件 1 矛盾; 同样在 $Block_i \neq Block_i'$ 的情况下伪造的签名将与条件 2 矛盾。断言 9 成立。

断言 10. 设一条合法的区块链 $\mathbb{C}$ 是由区块 $Block_0, Block_i, ..., Block_n$ 所构成的集合，取其中任

意 d 个连续的区块组成集合 $\mathbb{B}=\{Block_x, ..., Block_y\}$。在拜占庭攻击环境下, 若 $Card(\mathbb{A})<Card(\mathbb{G})/2$, 那么拜占庭节点伪造的区块满足定义 9 的概率将满足 $\exists Block_i \in \mathbb{B}, P(VB(Block_i')=True)<\varepsilon$, 其中$\varepsilon$为任意小的可忽略实数。

证明. 假设断言 10 为假, 则存在两种情况: 1) $\mathbb{B}$ 中有最后一块区块; 2) $\mathbb{B}$ 中不含最后一个区块。当出现情况 1)时, 由于 $Card(\mathbb{A})<Card(\mathbb{G})/2$, 即满足 $Pw'<Pw$, 由条件 4 可知诚实节点生成下一个区块的速度比拜占庭节点快。在生成下一个正常区块 $Block_{n+1}$ 后若 $Block_n'$ 依然合法, 则由断言 9 得知其与条件 1 或条件 2 矛盾。当出现情况 2)时, 设最后一个伪造区块为 $Block_i' \neq Block_i$, 满足 $VB(Block_i)=Ture$。但由于 $Block_{n+1}$ 也在 $Hash$ 链表上, 若 $Block_i'$ 合法则与条件 1 矛盾。断言 10 得证。

断言 11.依据式(11), 记节点的贡献度大小为 $C_k=f(\boldsymbol{N}_{end}, \boldsymbol{W}_k)$。对于已确定的 $\boldsymbol{N}_{end}$, 且当 $C_a=0$ 时, 以可忽略的概率存在函数 $g(\boldsymbol{W}_k, \boldsymbol{W}_a)$ 使得 $C_a>0$。

证明. 记 $\boldsymbol{N}_{end} \in \{Block_0, ..., Block_n\}$, 由于 $\boldsymbol{N}_{end} \in Block_i$, $C_a=0$, 因此 $\boldsymbol{W}_a \notin Block_i$。此时, 若断言 11 不成立, 则存在 $g(\boldsymbol{W}_k, \boldsymbol{W}_a)$ 使得 $C_a>0$, 从而可构造 $\boldsymbol{W}_a \in Block_i'$ 满足 $VB(Block_i')=Ture$。与上文的条件 1、2 皆矛盾。断言 11 成立。

断言 12.对于由诚实节点 k 产生的梯度模型 $\boldsymbol{W}_k$ 满足 $V_W(\boldsymbol{W}_k, (B_k, \theta_k))=True$, 以可忽略的概率存在函数 $y(\boldsymbol{W}_k, a)$ 使得 $V_w(\boldsymbol{W}_k, (B_a, \theta_a))=True$。

证明. 若断言 12 不成立, 则存在函数 $y(\boldsymbol{W}_k, a)$ 使 $V_w(\boldsymbol{W}_k, (B_a, \theta_a))=True$, 并以不可忽略的概率满足 $V_w(\boldsymbol{W}_k, (B_a, \theta_a))=V_w(\boldsymbol{W}_k, (B_k, \theta_k))=True$。但依据条件 3 中对于模型水印的鲁棒性需求, 可得到 $V_w(\boldsymbol{W}_k, (B_a, \theta_a))=V_w(\boldsymbol{W}_k, (B_k, \theta_k))=True)<\varepsilon$, 其中$\varepsilon$为任意小实数。因此与条件 3 矛盾, 断言 12 成立。

5 仿真结果

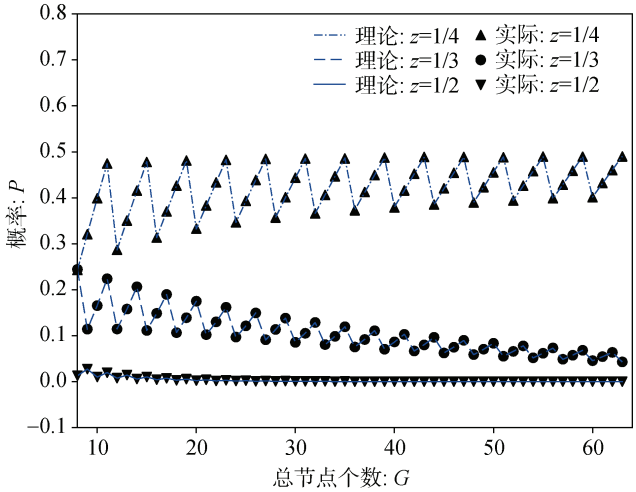
为证明本文的方案在实际应用中具有较高的可信性, 下面将采用了单机并发模拟的形式进行联邦学习区块链系统仿真。运行环境如下: CPU 为 Intel i7-6700HQ, 内存为 16GB, python 版本为 3.6.12, Pytorch 版本为 1.7.1+cu110。数据集采用“重庆市主干道交通数据流”中 16 条主干道数据, 包括 640000 个样本, 训练二分类模型用于预测路段的拥堵情况。联邦学习共识激励机制通过 Pytorch 上的 CNN 模型来进行构造, 训练本地模型的学习率为 α=0.001, 隐藏层包含 400 个神经元, 采用 ReLU 激活函数, epoch 为 20。本实验整体将涵盖协议一致性验证、联邦学习精度、系统整体运行性能等内容。为叙述清晰, 表 1 对实验与图例中重复提及的关键词做出了说明。

表 1 关键词说明

Table 1 Key words

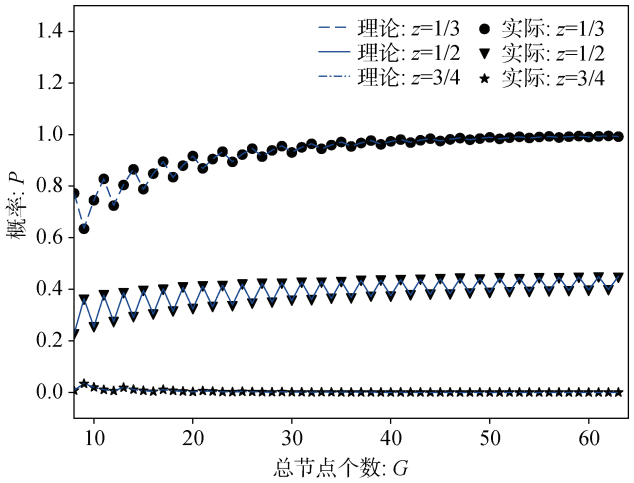
关键词	关键词含义
N_B	表示 $\boldsymbol{N}$ 被一半以上的节点认为是 $\boldsymbol{N}_{best}$
$\boldsymbol{N}_{end}$	表示共识完成后上链的模型
u	表示上链模型需得到的认定票数
G	表示参与协议的节点数量
z	表示 u 与 G 的比值, 即 $z=u/G$
P	表示各个节点能够争夺到打包权限的概率

在定理 1 和 2 中, 本文推导出了各节点产生出的 $\boldsymbol{N}$ 被一半以上节点认为是 $\boldsymbol{N}_{best}$ 的概率, 记为 P。相应地, 下面依照定理 1 和 2 中的性能评价函数 $F(\boldsymbol{N}_k, i)$ 来筛选模型, 通过重复 100000 轮 PFconsensus 协议的写入权限争夺过程, 记录其中某一随机节点产生出 $\boldsymbol{N}_B$ 的次数, 进而换算为 P 值。从图 5a 可以看出, 当节点间的数据为 non-IID 的极端情况下时, z 的增大会降低 P 值, 并在 z 大于 0.33 的情况下随着参与节点个数 G 的增大而减小。图 5b 则表明节点之间数据分布为 IID 时, P 依然在 z 大于 0.5 的情况下随着 G 增大而减小。与此同时, 因节点个数和票数在实际中只



(a) 数据为 Non-IID 的极端情况下时 z 与 G 对 P 的影响

(a) The influence of z and G on P when data distribution is extreme case of Non-IID



(b) 数据分布为 IID 时 z 与 G 对 P 的影响

(b) The influence of z and G on P when data distribution is IID

图 5 不同数据分布下 z 与 G 对 P 的影响

Figure 5 The influence of z and G on P by different data distribution

能取整, 这导致 P 会随着 G 的增加产生周期性的变化。以上结果均表明本文理论分析与实际情况一致。

将同样的评价函数用于计算机器学习模型的精确度指标(ACC 与 F1 的加权和), 可以进一步分析该系统在实际应用中的节点打包情况。

图 6 表示重复 100 轮写入权限争夺过程后某随机节点产生 N_B 的次数。可见, 实际生产环境中 z 值在大于 0.5 时, 各节点产生 N_B 的难度会随着 G 增大而略微提高, 但整体难度符合定理 3 的结论, 其任意节点的 P 介于定理 1 与定理 2 之间。在图 6 中, 任意节点产生了一个 N_B 后需要进行认定, 最终完成上链。为此, 本文进一步记录了在在z=0.5 的情况下, 产生出新区块所需的通信次数与 G 值的对应关系。

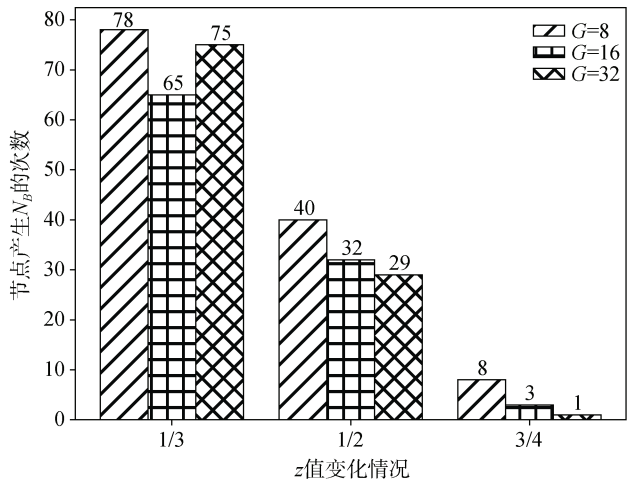


图 6 在实际模型性能对比条件下节点产生 N_B 的情况

Figure 6 The situation of any node generating N_B in the case of actual model performance comparison

通过观察图 7 可以发现, 区块链出块所需通信次数符合式(27)的规律。以上结论反映了第四章理论分析的正确性, 但不足以说明本文联邦学习区块链的可行性能。

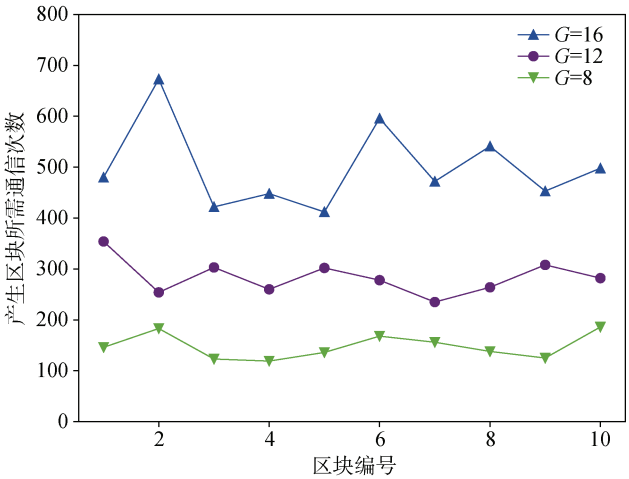


图 7 产生新区块的通信次数变化情况

Figure 7 The times of communications required to generate a new block

图 8 描述了各个节点模型在共同测试集下准确率与 F1 的加权和变化情况。可以发现在同一段时间内, 争夺到上链权限的聚合模型性能表现最为优秀。该实验结果说明本方案在实际运行过程中符合 PFconsensus 协议中对上链模型的性能假设与协议的正确性。

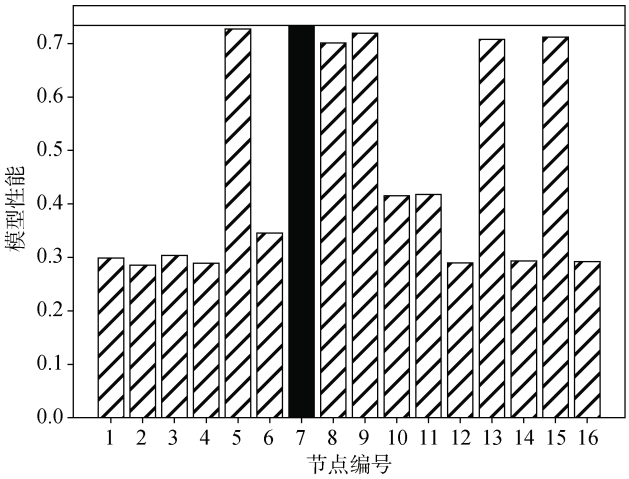


图 8 各节点产生模型性能对比(全黑表示上链模型)

Figure 8 Performance comparison of models generated by each node (The all black bar indicate winding models)

进一步的, 本文对比了本方案中联邦学习的效率性能与 C/S 方案下的效率性能。得到图 9 所示的模型性能情况。

图 9 中分别记录了 PFconsensus 协议方案与 C/S 模式方案下的联邦学习在准确率和 F1 上的差别, 其中 PFconsensus 协议方案针对统一个目标任务连续迭代了 9 次模型, 而 C/S 模式方案下同样也持续迭代了 9 次模型。从图 9 中可以看出本文方案与基于 C/S 模式的联邦学习方法在性能上相近, 符合本文对区块链模式下的训练性能要求。同时, 为了证明本文区块链方案能够符合定义 5, 本文测试了其在投毒节点个数为 0 到 $G/2$ 条件下的性能变化, 并对比在 C/S 方案下的表现。

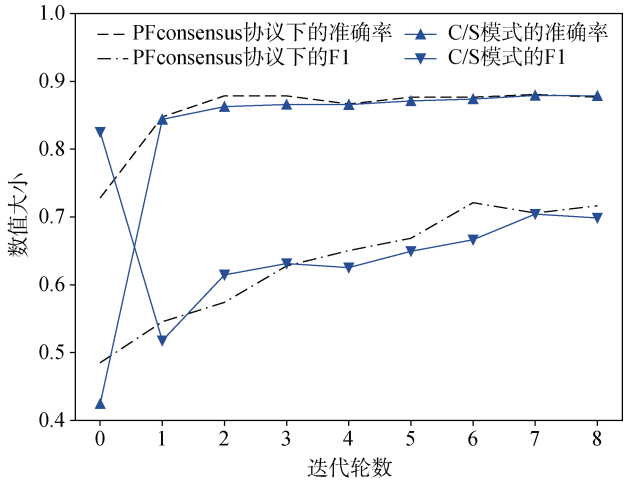


图 9　本文方案与 C/S 方案下联邦学习过程的性能变化对比

Figure 9　The performance change of the federated learning process under the scheme in this paper and the C/S scheme

图 10 表明了本文的模型投票筛选机制在投毒节点少于 $G/2$ 时依然能够得到不包含投毒梯度的聚合模型, 而在 C/S 方案下的聚合模型性能则会受投毒节点个数的增长导致性能直线下降。

同时, 本文对比了文献[20]方案的能源消耗和抵御投毒攻击情况。本文在 MNIST 数据下设计的本地卷积模型学习率为 α=0.001, 包含 2 个卷积层, 采用 ReLU 激活函数。本方案的投毒样本与文献[20]中采用的方式一样, 将训练样本的部分标签进行打乱, 并作为投毒样本分配给指定的攻击者。具体而言, 即将 MNIST 数据集中样本中的源标签“1”改为目标标签“8”。

图 11 表示该卷积模型在应对投毒攻击时, 其协议在迭代后依然能够完成收敛, 并能够得到与无投毒环境下相近的准确率。在投毒攻击下, 该方案比文献[20]中的方案具有更强抵御能力, 并在能源的利用上更为环保。

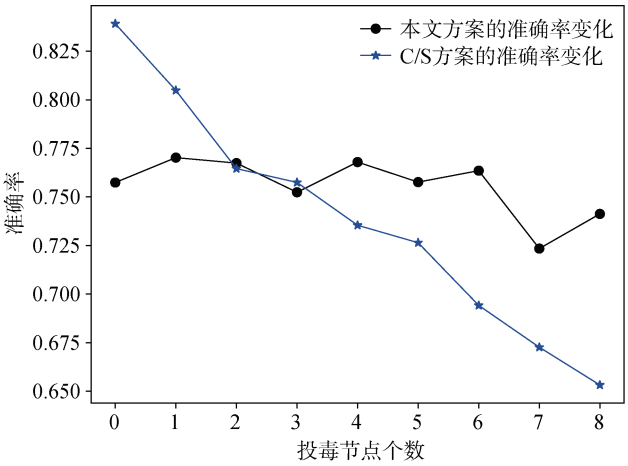


图 10　本文方案与 C/S 方案下聚合模型性能受投毒节点个数的影响(其中 G=16)

Figure 10　The performance of the aggregation model under this scheme and the C/S scheme is affected by the number of poisoned nodes (G=16)

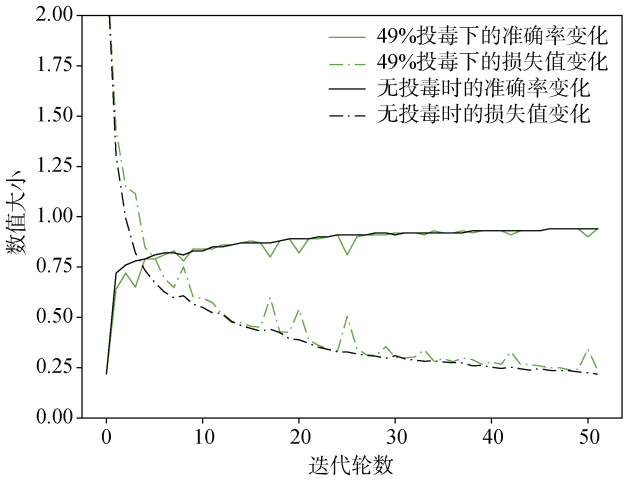


图 11　本文方案在投毒环境下与无毒环境下训练 MNIST 数据模型的情况

Figure 11　The situation of training the MNIST data model in the poisoning environment and the non-toxic environment

表 2　本文方案与文献[20]方案对比

Table 2　The scheme in this paper is compared with the scheme in the Ref.[20]

	文献[20]	本文方案
能源利用率(训练能耗/共识+训练能耗)	15.2%～40%	100%
最大投毒占比	30%	49%

通过仿真实验可以看到, 本文所提出的联邦学习共识激励机制比直接将区块链作为联邦学习平台资源利用率更高。基于本地数据进行全局模型的性能进行评测不仅能够有效保护参与者的数据隐私性,

也能够更好地解决联邦学习中的Non-IID问题。此外, 本文所采用的水印融合方案能够使模型在上链后帮助完成节点的行为追溯和贡献度分配, 充分激励不同的数据持有者参与训练。将联邦学习本身融入区块链共识激励机制可以获得众多优势, 包括: (1) 去中心化架构能够有效地保证联邦训练过程的稳健性; (2) 基于本地数据的模型性能评价能够很地解决Non-IID 问题; (3) 模型水印技术能够有效地对节点行为和贡献进行记录; (4) 利用模型训练替代传统挖矿算法能够充分缓解资源浪费的问题; (5) 对上链模型采用协同性能筛选的方式可以抵御包括投毒攻击在内的各种联邦学习威胁。

6 结论

本文为商业领域普遍存在的“数据孤岛”问题提供了一个完备的解决方案。联邦学习虽然能够在一定程度上保证商业数据的隐私性与可用性, 但采用中心化的架构极易招致拒绝服务、推理攻击、模型投毒等威胁。本文将联邦学习嵌入到区块链共识协议当中, 并借助水印融合技术, 克服了模型协同训练过程中所存在的资源浪费与消极产与等问题。为验证本文方案的可行性及安全性, 专门针对分布式联邦学习场景下的数据分布和系统规模进行了理论分析, 其结果也通过了以最优模型产生概率、共识难度、模型性能、通信复杂度以及知识产权配额等为指标的实验验证。

后续研究将针对实验过程中所发现的缺陷进行, 包括: (1) 由于各个节点都会参加训练过程, 而只有一半节点能够得到最终模型的知识产权, 因而需要进一步优化该系统的奖励制度; (2) 针对推理攻击, 如果在训练过程中对模型梯度进行差分隐私保护可能会影响传输效率及最终模型的准确率, 所以有必要更进一步引入梯度隐私保护机制, 并在解决链上数据可追溯性与机密性之间的矛盾。

参考文献

[1] Yang Q A, Liu Y, Cheng Y, et al. Federated Learning[J]. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2019, 13(3): 1-207.

[2] Li W, Chai Y B, Khan F, et al. A Comprehensive Survey on Machine Learning-Based Big Data Analytics for IoT-Enabled Smart Healthcare System[J]. *Mobile Networks and Applications*, 2021, 26(1): 234-252.

[3] Rathore M M, Shah S A, Shukla D, et al. The Role of AI, Machine Learning, and Big Data in Digital Twinning: A Systematic Literature Review, Challenges, and Opportunities[J]. *IEEE Access*, 2021, 9: 32030-32052.

[4] Yuan H T, Li G L. A Survey of Traffic Prediction: From Spatio-Temporal Data to Intelligent Transportation[J]. *Data Science and Engineering*, 2021, 6(1): 63-85.

[5] Jiang J C, Kantarci B, Oktug S, et al. Federated Learning in Smart City Sensing: Challenges and Opportunities[J]. *Sensors*, 2020, 20(21): 6230.

[6] Hosseini S, Sardo S R. Data Mining Tools -a Case Study for Network Intrusion Detection[J]. *Multimedia Tools and Applications*, 2021, 80(4): 4999-5019.

[7] Lee E, Jang Y, Yoon D M, et al. Game Data Mining Competition on Churn Prediction and Survival Analysis Using Commercial Game Log Data[J]. *IEEE Transactions on Games*, 2019, 11(3): 215-226.

[8] Zeng S Q, Huo R, Huang T, et al. Survey of Blockchain: Principle, Progress and Application[J]. *Journal on Communications*, 2020, 41(1): 134-151.
(曾诗钦, 霍如, 黄韬, 等. 区块链技术研究综述: 原理、进展与应用[J]. *通信学报*, 2020, 41(1): 134-151.)

[9] Papadimitriou P, Garcia-Molina H. Data Leakage Detection[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2011, 23(1): 51-63.

[10] Bogetoft P, Christensen D L, Damgård I, et al. Secure Multiparty Computation Goes Live[C]. *International Conference on Financial Cryptography and Data Security*, 2009: 325-343.

[11] Evans D, Kolesnikov V, Rosulek M. A Pragmatic Introduction to Secure Multi-Party Computation[J]. *Foundations and Trends in Privacy and Security*, 2018, 2(2/3): 70-246.

[12] Wood A, Najarian K, Kahrobaei D. Homomorphic encryption for machine learning in medicine and bioinfor-matics[J]. *ACM Computing Surveys*, 2020, 53(4): 1-35.

[13] Kuang F, Mi B, Li Y, et al. Multiparty Homomorphic Machine Learning with Data Security and Model Preservation[J]. *Mathematical Problems in Engineering*, 2021, 2021: 166-179.

[14] Hassan M U, Rehmani M H, Chen J J. Differential Privacy Techniques for Cyber Physical Systems: A Survey[J]. *IEEE Communications Surveys & Tutorials*, 2020, 22(1): 746-789.

[15] Sen A A A, Eassa F A, Jambi K, et al. Preserving Privacy in Internet of Things: A Survey[J]. *International Journal of Information Technology*, 2018, 10(2): 189-200.

[16] Kairouz P, McMahan H B, Avent B, et al. Advances and Open Problems in Federated Learning[J]. *Foundations and Trends® in Machine Learning*, 2021, 14(1/2): 1-210.

[17] Liu Y, Kang Y, Xing C P, et al. A Secure Federated Transfer Learning Framework[J]. *IEEE Intelligent Systems*, 2020, 35(4): 70-82.

[18] Huang Y T, Chu L Y, Zhou Z R, et al. Personalized Cross-Silo Federated Learning on Non-IID Data[EB/OL]. 2020: arXiv: 2007.03797. https://arxiv.org/abs/2007.03797.pdf

[19] Zhu H Y, Zhang H Y, Jin Y C. From Federated Learning to Federated Neural Architecture Search: A Survey[J]. *Complex & Intelligent Systems*, 2021, 7(2): 639-657.

[20] Zhu J M, Zhang Q N, Gao S, et al. Privacy Preserving and Trustworthy Federated Learning Model Based on Blockchain[J]. *Chi-*

nese Journal of Computers, 2021, 44(12): 2464-2484.
(朱建明, 张沁楠, 高胜, 等. 基于区块链的隐私保护可信联邦学习模型[J]. *计算机学报*, 2021, 44(12): 2464-2484.)

[21] Zhang J H, Li X W, Zeng X, et al. Cross Domain Authentication and Key Agreement Protocol Based on Blockchain in Edge Computing Environment[J]. *Journal of Cyber Security*, 2021, 6(1): 54-61.
(张金花, 李晓伟, 曾新, 等. 边缘计算环境下基于区块链的跨域认证与密钥协商协议[J]. *信息安全学报*, 2021, 6(1): 54-61.)

[22] Zhao B, Fan K, Yang K, et al. Anonymous and Privacy-Preserving Federated Learning with Industrial Big Data[J]. *IEEE Transactions on Industrial Informatics*, 2021, 17(9): 6314-6323.

[23] Pokhrel S R, Choi J. Federated Learning with Blockchain for Autonomous Vehicles: Analysis and Design Challenges[J]. *IEEE Transactions on Communications*, 2020, 68(8): 4734-4746.

[24] Qu X D, Wang S L, Hu Q, et al. Proof of Federated Learning: A Novel Energy-Recycling Consensus Algorithm[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2021, 32(8): 2074-2085.

[25] Li Y Z, Chen C, Liu N, et al. A Blockchain-Based Decentralized Federated Learning Framework with Committee Consensus[J]. *IEEE Network*, 2021, 35(1): 234-241.

[26] Wang Y C, Tian Y Y, Yin X Y, et al. A Trusted Recommendation Scheme for Privacy Protection Based on Federated Learning[J]. *CCF Transactions on Networking*, 2020, 3(3): 218-228.

[27] Luo Z P, Zhao S Q, Lu Z, et al. Adversarial Machine Learning Based Partial-Model Attack in IoT[C]. *The 2nd ACM Workshop on Wireless Security and Machine Learning*, 2020: 13-18.

[28] Chacon H, Silva S, Rad P. Deep Learning Poison Data Attack Detection[C]. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence*, 2020: 971-978.

[29] Liu Y F, Ma X J, Bailey J, et al. Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks[EB/OL]. 2020: arXiv: 2007.02343. https://arxiv.org/abs/2007.02343.pdf

[30] Rahman M A, Rahman T, Laganière R, et al. Membership Inference Attack Against Differentially Private Deep Learning Model[J]. *Trans Data Priv*, 2018, 11: 61-79.

[31] Lamport L. Paxos Made Simple[J]. *ACM SIGACT News*, 2001, 32(4): 51-58.

[32] Sattler F, Wiedemann S, Müller K R, et al. Robust and Communication-Efficient Federated Learning from Non-I.i.d. Data[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, 31(9): 3400-3413.

[33] Lu Y L, Huang X H, Dai Y Y, et al. Blockchain and Federated Learning for Privacy-Preserved Data Sharing in Industrial IoT[J]. *IEEE Transactions on Industrial Informatics*, 2020, 16(6): 4177-4186.

[34] Majeed U, Hong C S. FLchain: Federated Learning via MEC-Enabled Blockchain Network[C]. *2019 20th Asia-Pacific Network Operations and Management Symposium*, 2019: 1-4.

[35] Peng Z, Xu J L, Chu X W, et al. VFChain: Enabling Verifiable and Auditable Federated Learning via Blockchain Systems[J]. *IEEE Transactions on Network Science and Engineering*, 2022, 9(1): 173-186.

[36] Handelman D. Gossip in encounters: The transmission of information in a bounded social setting[J]. *Man*, 1973, 8(2): 210-227.

[37] Bonneau J, Miller A, Clark J, et al. SoK: Research Perspectives and Challenges for Bitcoin and Cryptocurrencies[C]. *2015 IEEE Symposium on Security and Privacy*, 2015: 104-121.

[38] Fam L, Li B, Gu H, et al. FedIPR: Ownership verification for federated deep neural network models[EB/OL]. 2021: ArXiv Preprint ArXiv:2109.13236.

[39] Zhu H Y, Jin Y C. Multi-Objective Evolutionary Federated Learning[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, 31(4): 1310-1322.

[40] Li X C, Zhan D C. FedRS: Federated Learning with Restricted Softmax for Label Distribution Non-IID Data[C]. *The 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021: 995-1005.

[41] Poupard G, Stern J. Security Analysis of a Practical "on the Fly" Authentication and Signature Generation[M]. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998: 422-436.

米波 博士, 重庆交通大学信息科学与工程学院教授, 博士生导师。研究领域包括密码学、区块链、智能交通、车载自主式网络等。Email: mi_bo@163.com

翁渊 于 2019 年在重庆交通大学计算机通信专业获得学士学位。现在重庆交通学校计算机与科学专业攻读硕士学位。研究领域为密码学、人工智能。研究兴趣包括区块链、同态加密。Email: wengyuan980930@mails.cqjtu.edu.cn

黄大荣 博士, 重庆交通大学信息科学与工程学院教授, 博士生导师。研究领域包括车联网安全容错控制、交通系统可靠性控制。Email: drhuang@cqjtu.edu.cn

刘洋 博士, 重庆交通大学信息科学与工程学院副教授, 研究生导师。研究领域包括形式化验证、信息安全和数据处理等。Email: liuyang13@cqjtu.edu.cn