2025 年全国大学生软件系统安全赛

作品报告

作品名称： SecFed：针对横向联邦学习系统的信任评估和异常检测系统

电子邮箱： 194993236@qq.com

提交日期： 2024.12.15

填写说明

1. 所有参赛项目必须为一个基本完整的设计。作品报告书旨在能够清晰准确地阐述（或图示）该参赛队的参赛项目（或方案）。

2. 作品报告采用A4纸撰写。除标题外，所有内容必需为宋体、小四号字、1.5倍行距。

3. 作品报告中各项目说明文字部分仅供参考，作品报告书撰写完毕后，请删除所有说明文字。(本页不删除)

4. 作品报告模板里已经列的内容仅供参考，作者可以在此基础上增加内容或对文档结构进行微调。

5. 为保证网评的公平、公正，作品报告中应避免出现作者所在学校、院系和指导教师等泄露身份的信息。

目录

摘要

近年来，随着物联网技术和云计算的发展，海量数据的产生和使用成为推动社会智能化的重要驱动力。然而，这也带来了数据隐私泄露、数据孤岛和网络传输效率低下等问题。为应对这些挑战，联邦学习作为一种新兴技术为多方协同数据使用提供了解决方案。

本系统基于联邦学习框架，设计并实现了一种可信的多方协同学习系统，旨在解决异构节点间协作的信任度评估和异常模型检测等核心问题。本系统在技术创新方面具有显著优势，主要体现在细粒度信任评估和隐私保护下的异常检测两个关键领域，并通过全面的方案设计提升了联邦学习系统的可靠性和安全性。在细粒度信任评估方面，系统设计了一种基于多维信任属性的动态融合算法，该算法充分利用参与节点的历史行为数据，从直接信任和推荐信任两个维度对节点进行细粒度建模。通过动态融合信任因素，系统能够精准量化参与节点的可信度，从而实现高效可靠的节点选择，避免低质量或恶意节点参与学习过程，为全局模型的构建提供可靠保障。在隐私保护下的异常检测中，本系统设计了一种无需额外计算开销的拜占庭检测机制。该机制在全局模型聚合阶段，通过精准识别异常节点，确保隐私保护环境下的全局模型训练精确度和鲁棒性。

通过实验验证，本系统在多个关键性能指标上展现了显著优势。在资源利用率方面，系统实现了较高的计算与通信效率，优化了联邦学习过程中节点资源的调度与分配，显著降低了额外计算和通信开销。在抗干扰能力方面，通过设计细粒度的信任评估机制与高效的拜占庭异常检测方法，系统能够精准识别并隔离恶意节点，显著提升了全局模型的鲁棒性。实验结果显示，即使在节点质量参差不齐或存在攻击行为的复杂环境下，系统仍然能够保持稳定的性能。在模型精度方面，本系统在隐私保护的前提下，通过优化全局模型聚合规则，将全局模型的精度稳定提升至传统方法的 15%-20%。

本系统通过创新的信任评估机制和隐私保护的异常检测方法，成功解决了联邦学习中存在的参与方可靠性和模型安全性问题，显著提升了系统的鲁棒性与资源利用效率。未来，本系统将进一步优化核心算法，拓展应用范围，致力于构建更加安全、高效的多方协同学习体系，加速万物互联时代下的智能化发展。

Abstract

In recent years, with the development of Internet of Things (IoT) technology and cloud computing, the generation and use of massive amounts of data has become an important driving force to promote social intelligence. However, it also brings problems such as data privacy leakage, data silos and inefficient network transmission. To cope with these challenges, federated learning as an emerging technology provides a solution for multi-party collaborative data usage.

Based on the federated learning framework, this work designs and implements a trustworthy multi-party collaborative learning system, aiming to solve the core problems of trust assessment and anomaly model detection for collaboration among heterogeneous nodes. This system has significant advantages in terms of technological innovation, mainly in the two key areas of fine-grained trust assessment and anomaly detection under privacy protection, and enhances the reliability and security of the federated learning system through a comprehensive program design. In terms of fine-grained trust assessment, the system proposes a dynamic fusion algorithm based on multi-dimensional trust attributes, which makes full use of the historical behavioral data of the participating nodes, and models the nodes at a fine-grained level from the dimensions of direct trust and recommended trust. By dynamically fusing trust factors, the system is able to accurately quantify the trustworthiness of participating nodes, thus realizing efficient and reliable node selection, avoiding low-quality or malicious nodes from participating in the learning process, and providing a reliable guarantee for the construction of the global model. In the anomaly detection under privacy protection, this system designs a Byzantine detection mechanism without additional computational overhead. This mechanism ensures the accuracy and robustness of global model training under privacy-preserving environment by accurately identifying anomaly nodes in the global model aggregation phase.

Through experimental validation, this system demonstrates significant advantages in several key performance metrics. In terms of resource utilization, the system achieves high computational and communication efficiency, optimizes the scheduling and allocation of node resources in the federated learning process, and significantly reduces the additional computation and communication overhead. In terms of anti-interference ability, by designing a fine-grained trust assessment mechanism and an efficient Byzantine anomaly detection method, the system is able to accurately identify and isolate malicious nodes, which significantly improves the robustness of the global

model. Experimental results show that the system is able to maintain stable performance even in complex environments with uneven node quality or the presence of attack behaviors. In terms of model accuracy, this system stably improves the accuracy of the global model to 15%-20% of the traditional method by optimizing the global model aggregation rules under the premise of privacy protection.

This system successfully solves the problems of participant reliability and model security in federated learning through the innovative trust assessment mechanism and privacy-protected anomaly detection method, which significantly improves the robustness and resource utilization efficiency of the system. In the future, this system will further optimize the core algorithm, expand the application scope, and commit to building a safer and more efficient multi-party collaborative learning system to accelerate the intelligent development in the era of Internet of Everything.

第一章 项目概述

1.1 项目背景

近年来，网络设备数量和产生的数据都呈现出急速增长的趋势。根据公开有效数据显示，2020 年将有超过 500 亿台机器、设备通过网络进行互联，超过 2000 亿个联网传感器产生海量数据；根据 IDC 预测，2025 年，全球数据总量将大于 20 泽字节（ZB）。这一趋势带来了前所未有的海量数据，在云计算平台的巨大算力以及人工智能先进算法的支持下，这些海量客户端节点产生的海量数据可以被转化为有效信息，产生更高的价值，让各行各业的效率更高、发展更具活力，从而推进整个社会的智能升级。

1.1.1 领域痛点

面对异构设备产生的海量数据的共享智能化需求与数据隐私、数据孤岛间的矛盾，学术界与工业界均希望通过技术手段，确保多方在进行机器学习模型训练的同时，能做到数据无需共享、隐私不被泄露、数据使用行为可控。2016 年，谷歌首次提出了联邦学习概念，其主要思想为利用分布在多个设备上的数据训练机器学习模型，同时确保数据不出设备本地，防止数据泄露，保护数据安全与隐私。微众银行的杨强教授团队将“联邦学习”概念扩展为所有隐私保护、去中心化协作机器学习技术的一般概念，其核心思想为：存在多个数据拥有方F_i（$i \in [1,n]$），拟利用各自拥有的数据集D_i联合训练机器学习模型，其中D_i =（I,X,Y），I表示样本ID空间、X表示特征空间、Y表示标签空间，其中各方的特征空间和样本空间可能不完全相同，根据数据在特征空间和样本ID空间中的分布情况可将联邦学习分为横向联邦学习与纵向联邦学习。在训练过程中任一方均不向其他方共享其数据，同时保证联合训练得到的模型M_fed的效果与将各方数据整合后进行训练得到的模型M_sum的效果的差距足够小。蚂蚁金服公司也提出了共享学习的概念，其主要思想为在多方参与且各数据提供方与平台方互不信任的场景下，能够聚合多方信息并保护参与方数据隐私的学习范式，这种学习模式主要采用了 TEE（Trusted Execution Environment）硬件可信执行环境与安全多方计算

（Security Multi-Party Computation, SMC）技术来保护数据隐私。

1.1.2 传统模式弊端

传统利用人工智能技术训练模型的模式为一方收集数据后进行清理融合，另一方利用融合数据构建模型以供其他方使用。例如，收集摄像头产生的海量视频图像，进行人脸、人体、车辆、非机动车的识别与融合，将视频、图形数据转换为人/车/人脸/物等全目标结构化数据，通过智能挖掘技术为上层应用（例如案件提前预防和精确打击，交通路径优化等）提供数据支撑以及数据治理等服务；通过收集零售门店海量的用户消费行为并标签化，结合金融机构数据进行机器学习，重构零售人-货-场生态，助力零售企业精准营销、金融机构风险防范；通过对大量患者的医学检验和影像检查结果的深度学习，为智慧医疗服务提供支撑，提升诊断效率与精度，降低患者医疗费用。

在上述模式中，所有联网设备产生的数据需要首先通过网络传输到云计算中心，利用云计算中心超强的计算能力对所有数据进行集中式训练。这种集中式数据处理的模式在当前的智能世界的构建范式中面临着诸多挑战：

（1）海量数据由大量边缘、异构设备产生，将这些数据全部传输至云端处理将造成网络带宽的过大压力，同时会由于传输时延而无法满足对实时性要求高的应用；

（2）在多数领域（尤其是金融、医疗等行业），由于商业利益相关、数据融合成本高等因素使得打破数据孤岛、实现数据互通共享仍是一个难题；

（3）网络设备产生的数据中敏感数据与隐私数据越来越多，基于云的集中式计算模式会增加泄露数据隐私的风险。《通用数据保护条例》（General Data Protection Regulation，GDPR）、《中华人民共和国网络安全法》以及《2018年加州消费者隐私法案》（The California Consumer Privacy Act of 2018，CCPA）等法案对数据安全与隐私保护做出了严格的规定。随着这些法案的陆续出台，粗放式地收集用户数据将面临巨大的法律风险。

1.1.3 现有协同学习问题

在下文，我们将以上提到的这些隐私保护、分布式、异构多方之间进行协同

学习的框架统称为协同学习。

尽管不同的协同学习框架与模型被提出，并且基于这些框架与模型的应用也越来越多。然而，新技术的出现往往是一把“双刃剑”，异构多方之间的协同学习也不例外。目前，有关异构多方之间进行隐私保护的协同学习的相关工作大多集中于全局模型聚合规则、隐私保护方法等方面，然而协同学习系统的可信性是其在关键领域落地的前提。针对构建可信协同学习系统这一需求，目前还有一些问题被学术界与工业界视为公开难题，亟需深入研究。

（1）如何提升协同学习系统的可靠性，使其适用于安全关键场景是协同学习面临的重要问题之一。不可靠的 AI 不能够胜任任何政治、经济或安全攸关的关键性场景。参与方之间无信任关系，参与方的原始数据无法共享，并且本地训练模型只能在加密状态下被传输与访问，这使得协同学习系统的可靠性面临着更大的挑战。为了提高系统的可靠性，需要研究学习参与方的可信度评估问题。

（2）如何在协同学习的每一轮迭代过程中发现参与方生成的异常本地模型，提高聚合模型的精确度也是协同学习系统面临的挑战之一。现有的协同学习模型通常假设各参与方均以有效训练样本作为学习算法的输入，并诚实执行学习算法得到本地模型，并未采取安全机制保障参与全局模型聚合的本地模型是真实、正确的。然而在实际中这一假设很难成立，无论是恶意参与方随意生成恶意模型上传至聚合服务器，还是可靠参与方由于设备故障等原因生成错误的本地模型，都会导致全局模型精度降低，甚至得到错误的全局模型，在未来的预测任务中出现预测偏差或预测错误，造成重大损失。为了保证全局模型的高精确度，需要研究本地模型的异常检测与安全的模型聚合规则等问题。

如果上述问题得不到有效解决，那么多方协同学习系统参与方的可靠性与模型的安全性就无法得到保障，从而无法构建可信的多方协同学习系统，进而严重阻碍其在关键性领域的发展和实际应用。

1.2 研究目标

研究成果将为构建异构设备间安全、可信的协同学习系统提供理论和技术支撑，解决人工智能在关键领域的应用障碍，促进其在金融、政治、军事等领域的应用，为政府和企业解决数据孤岛和隐私保护的难题，加快万物智能互联时代的

到来。

1.3 场景和价值

在这种协同学习模式中，参与学习的各方多数情况下为异构网络节点，即各节点拥有的数据数量、数据质量、计算能力、通信能力与安全能力均不同。参与学习的联网终端设备利用自身拥有的数据进行机器学习模型训练，接着，各参与方将学习结果通过聚合服务器（或平台）进行模型聚合，形成全局模型，或者参与方之间遵循交互协议进行数据交互形成全局模型，从而避免将海量数据传输至云端集中训练，解决了当前的数据隐私保护和数据孤岛的挑战。本研究所产生的成果能够提高异构设备间协同学习系统参与方的可靠性与模型的安全性，构建可信的联邦学习系统。研究成果将为构建异构设备间安全、可信的协同学习系统提供理论和技术支撑，解决人工智能在关键领域的应用障碍，促进其在金融、政治、军事等领域的应用，为政府和企业解决数据孤岛和隐私保护的难题，加快万物智能互联时代的到来。

1.4 创新性说明

本研究以目前多方协同学习架构中的联邦学习为代表，从选择高信任度节点参与协同学习、抵御恶意模型参与全局模型聚合两方面入手，进行了联邦学习系统的深入研究，重点关注横向联邦学习系统中参与方的信任度评估以及隐私保护的异常检测两个内容。本章结合仿真实验结果与现有的保证联邦学习系统参与方的可靠性与模型的安全性的方案作对比，对本作品的创新性进行说明：

- **抵抗恶意行为的攻击方面表现好，识别恶意参与方准确度高，具有良好的鲁棒性**

针对现有联邦学习参与方信任度评估方法中普遍考虑的信任因素较为单一的问题，本研究设计了一种新的单一上下文中联邦学习参与方的细粒度信任评估方案。该方案基于多种信任属性对参与方进行行为建模，提出了直接信任信息和推荐信任信息的动态融合算法，实现对参与方信任度的细粒度评估。该算法能够准确评估出不同行为模式的参与方，且较基于主观逻辑的信任评估方法在抵抗具有不同行为模式的恶意参与方方面表现出色，为评估参与方的可靠性与可信度提

供了有效的参考方案。

- **抗干扰能力强，未产生额外的计算开销，提高了资源利用率**

针对横向联邦学习参与方发起投毒攻击破坏全局模型完整性的问题，本研究提出了一种专注于在 IID 场景下隐私保护的拜占庭节点检测方案。该方案提出一种在聚合阶段无需 Shamir 秘密共享的新掩码机制。结合类似 Krum 的安全聚合方案，能够准确检测出 AD1 型和 AD2 型敌手，防止恶意节点对整体模型训练效果产生不良影响，且所提拜占庭检测方案不受节点的本地模型结构的影响，可以在隐私保护环境下进行全局聚合。由一系列仿真实验结果知生成的全局模型精度基本在 70%-80%之间，远高于 AvgFed 和基于自编码器的异常检测机制下生成的全局模型，合理设置的门限参数取值，使方案能降低良性节点被误判为拜占庭节点的概率，减少节点验证模型的时间开销。同时该方案在聚合服务器端几乎未产生额外的计算开销。

第二章 作品设计与实现

2.1.1 总体框架

本研究立足于万物智能互联浪潮下，构建可信的异构多方协同学习系统的实际需求，以及相关研究刚刚起步的现状，以目前多方协同学习模型中的联邦学习为代表，从抵御网络中不可靠节点参与学习和防御异常模型参与全局模型更新两方面入手，针对联邦学习系统参与方的可信度评估、恶意模型检测方面存在的技术挑战，**重点开展以下两个方面的研究：（1）横向联邦学习系统中参与方的信任度评估；（2）隐私保护的联邦学习系统的异常检测。主要解决：（1）横向联邦学习系统中参与方行为数据的演化性以及时空关联性；（2）不同数据分布情况的联邦学习系统中，隐私保护的正常模型与异常模型的统计差异性及其可检测条件两个关键科学问题。从可靠参与方选择和异常数据检测两方面提升、优化联邦学习系统的可靠性与安全性。**

2.1.2 关键技术

建立可信的联邦学习系统，具体将通过提高学习系统参与方的可靠性以及全

局模型的准确性来实现该目标：

（1）研究横向联邦学习系统中参与方的信任度评估模型与方法，基于参与方的历史行为对其信任度（在未来学习过程中提供良好行为的概率）进行评估，使得服务器可以选择高信任度的节点参与学习，提高学习参与方行为的可靠性。

（2）研究隐私保护的联邦学习系统的异常检测，针对横向联邦学习研究隐私保护的本地模型的异常检测方法，针对纵向联邦学习研究主动方计算结果的完整性验证方法，从而确保异常模型与数据不参与全局模型的迭代更新，保证全局模型的安全性。

本研究的主要研究内容如图 1 研究内容所示。

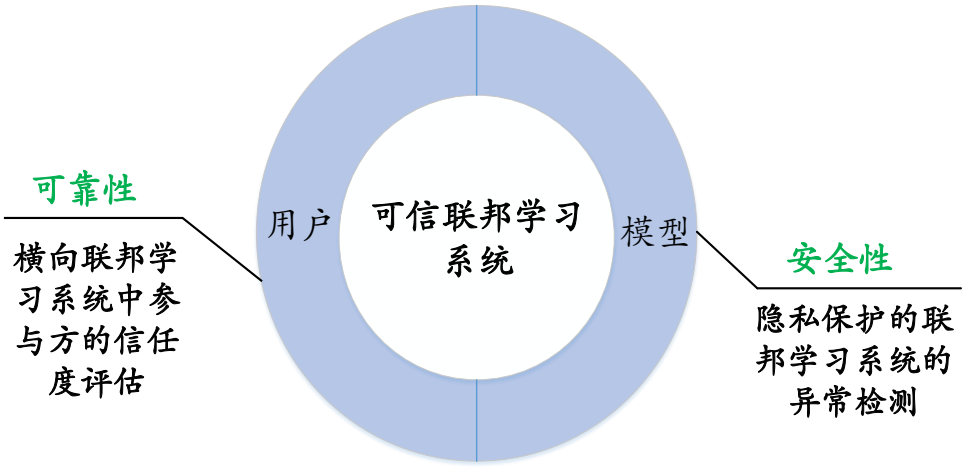


图 1 研究内容

2.1.3 系统实现

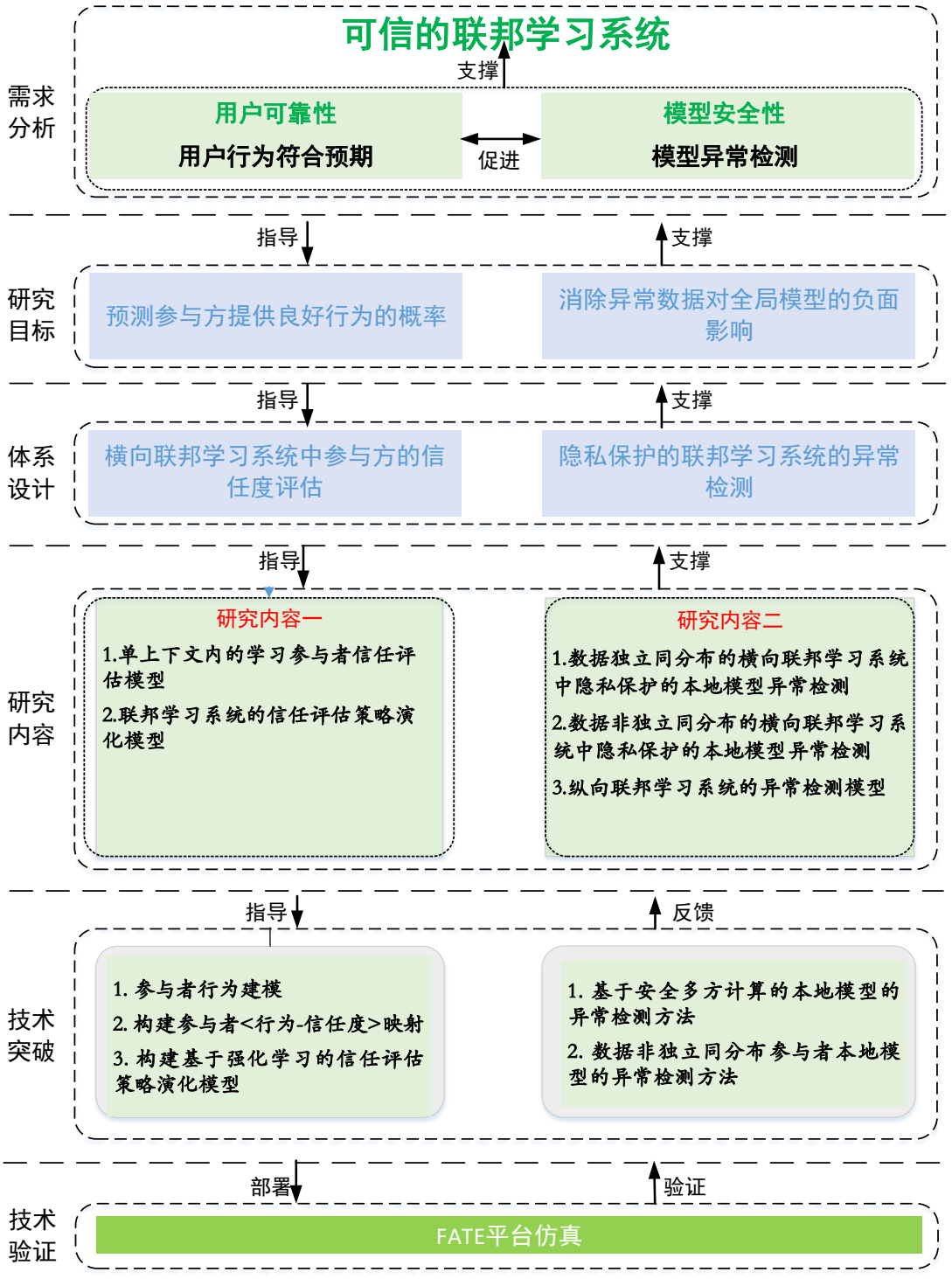


图 2 系统实现流程图

2.2 单一上下文联邦学习参与方信任度评估方案

2.2.1 功能介绍

针对现有联邦学习参与方信任度评估方法中普遍考虑的信任因素较为单一的问题，本研究设计了一种**单一上下文中联邦学习参与方的细粒度信任评估方案**，该方案基于多种信任属性对参与方进行细粒度的行为建模，计算各参与方的直接信任度，根据系统中其他聚合服务器关于参与方的推荐信任信息评估参与方的推荐信任度。同时，设计了直接信任度与推荐信任度的动态融合方法，实现对参与方综合信任度的评估。

本研究将对所提方案的系统假设、用户行为建模方法以及具体的信任评估方法依次进行介绍。

2.2.2 系统假设及用户行为建模

系统假设：本研究所考虑的联邦学习系统由用户、聚合服务器、参与方、主聚合服务器和推荐服务器组成，对本研究各类实体做如下假设：

（1）用户：能够申请加入联邦学习任务并进行模型训练，本系统中用户使用集合$U = \{U_1, U_2, \cdots, U_N, \cdots\}$表示。

（2）聚合服务器：能够发布联邦学习任务并进行模型聚合，同时能够对参与学习任务的各用户进行信任度评估，并且根据评估结果做出相应决策，本系统中聚合服务器使用集合$S = \{S_1, S_2, \cdots, S_j, \cdots, S_M\}$表示，假设本系统中共有$M$个聚合服务器，聚合服务器均是值得信赖的。

（3）参与方：参与方是针对某一具体的联邦学习任务而言，表示实际参与某次联邦学习任务的用户。一个联邦学习系统中多个参与方使用集合u表示，$u \subseteq U$，其中$u_i \in u$表示参与某次联邦学习任务的某一参与方。

（4）主聚合服务器：主聚合服务器针对某一具体的联邦学习任务而言，是聚合服务器集合S中的一个元素，负责完成某次联邦学习任务发布以及模型聚合，一个联邦学习系统中只有一个主聚合服务器，主聚合服务器在本研究中也直接称为聚合服务器。

（5）推荐服务器：推荐服务器针对某一具体的联邦学习任务而言，把系统中除主聚合服务器之外的其余$M-1$个聚合服务器均称为推荐服务器，推荐服务器负责完成其他联邦学习任务，同时根据自身经验向主聚合服务器提供各参与方的推荐信任信息，推荐服务器对于参与方的推荐信任信息是准确的，一个联邦学习系统中有多个聚合服务器被称为推荐服务器，为了便于区分，推荐服务器使用RS表示，记为$RS=\{rs_1,rs_2,\cdots,rs_{M-1}\}$，其中$RS\subseteq S$。

本研究中与信任相关的概念，包括：信任、信任空间、信任属性、信任证据、上下文的相关定义描述如下：

（1）信任：聚合服务器对本次联邦学习系统中参与方的实际行为与预期行为相符程度的预测，对参与方实际行为与预期行为相符程度的预测越高，参与方越容易在联邦学习过程中表现出良好行为，此时聚合服务器对参与方的信任程度也越高[1][2]，反之亦然。信任评估是聚合服务器对本次联邦学习系统中的参与方可信程度的评估过程[3]。

（2）信任空间：信任空间定义为封闭的连续空间$[0,1]$，即信任度的取值是$[0,1]$中任意一个实数，其中 0 表示聚合服务器对参与方完全不信任，1 表示聚合服务器对参与方完全信任，0.5 表示聚合服务器对参与方的信任程度无法评判。

（3）信任属性：信任属性由影响联邦学习系统中参与方信任度的两大因素组成，参与方上传本地模型的时延信息以及异常度信息，具体细节将在下文介绍。

（4）信任证据：信任证据由完成信任评估所需要的相关信任信息组成，对于信任度的评估结果具有决定性作用，本研究中的信任证据均存储在信任证据数据库中。

（5）上下文：上下文指与用户行为相关的信息，需要根据具体的应用场景来确定。本研究对于联邦学习系统中参与方的信任度评估，均在相同上下文c中进行，因此不对上下文进行具体的形式化表示。

本研究所提联邦学习信任评估方案的相关参数以及含义如表 1 所示。

表 1 单一上下文信任评估方案相关参数

c:本联邦学习系统所处上下文	α_i:u_i的异常因子
S:聚合服务器集合	β_i:u_i的时延因子
RS:推荐服务器集合	θ:遗忘因子，用于α_i和β_i的计算
U:用户集合	R_c^i:上下文c中u_i的行为可靠性

u:参与方集合	S_c^i:上下文c中u_i的行为稳定性
M:聚合服务器数量	$T_c^j(i)$:上下文c中u_i的综合信任度
N:本联邦学习系统参与方数量	$dir_{T_c^j(i)}$:上下文c中u_i的直接信任度
P:历史交互信息队列HL长度	$recom_{T_c^j(i)}$:上下文c中u_i的推荐信任度
Q:推荐信任信息队列RL长度	$hist_{T_c^j(i)}$:上下文c中u_i的历史信任度
CL_c^i:上下文c中u_i的当前交互信息	$curr_{T_c^j(i)}$:上下文c中u_i的当前信任度
HL_c^i. :上下文c中u_i的历史信任信息队列	ω_{dir}:直接信任度$dir_{T_c^j(i)}$的权重
RL_c^i:上下文c中u_i的推荐信任信息队列	ω_{recom}:推荐信任度$recom_{T_c^j(i)}$的权重
$h_c^{i,j}$:上下文c中u_i与S_j的总交互次数	ω_{hist}:历史信任度$hist_{T_c^j(i)}$的权重
$TE_c^{i,j}$:上下文c中u_i的信任证据	ω_{curr}:当前信任度$curr_{T_c^j(i)}$的权重
$list_{t_c^i}$:HL_c^i中元素t_{hl}形成的集合	σ:本地信任信息与推荐信任信息相对质量
$list_{r_c^i}$:HL_c^i中元素R_c^{hl}形成的集合	χ_i:u_i直接交互频繁度度量因子
$list_{rt_c^i}$:RL_c^i中元素rt形成的集合	γ_i:u_i推荐信任量度量因子
$list_{rh_c^i}$:RL_c^i中元素$h_c^{i,rs}$形成的集合	$f_\omega(\sigma)$：整形函数，用于ω_{dir}的计算
b_r:CL_c^i中存储的第r条当前交互信息	δ：整形函数$f_\omega(\sigma)$的阈值
hl_k:HL_c^i中存储的第k条历史信息	$\Omega_{i,j}$:u_i与S_j的熟悉程度
rl_d:RL_c^i中存储的第d条推荐信息	ϕ:$\Omega_{i,j}$的调节因子
acc_i^r:第r轮迭代交互u_i的行为异常度	φ_p:时间衰减因子
acc_i:u_i的异常度信息集合	$g(h)$:当前信任调节函数
d_i^r:第r轮迭代交互u_i的时延	ε:当前信任调节因子
d_i:u_i的时延信息集合	H:交互阈值

本研究所提方案在一个联邦学习系统中进行，该系统由一个负责发布联邦学习任务和完成模型聚合的主聚合服务器S_j与N个参与方组成，共同完成特定的联邦学习任务。在S_j对外发布某联邦学习任务后，S_j基于信任度高低从系统所有用户集合U中选择部分用户作为本次联邦学习的参与方参与学习任务，协作完成联邦学习全局模型训练，参与方集合记为$u = \{u_1, u_2, \cdots, u_i, \cdots u_N\}$。此外，系统中其余$M-1$个聚合服务器作为推荐服务器向主聚合服务器$S_j$提供各参与方的推荐信任信息。下面会给出一次联邦学习过程中聚合服务器S_j对参与方u_i的信任行为建模以及信任度计算方法。

用户行为建模是信任评估的基础，通过对用户行为信息的分析，能够发掘用户的行为特征[4][5]。假设聚合服务器S_j发布的本次联邦学习任务共需要进行m轮

迭代，对应本次学习的m轮迭代事件$\{e_1, e_2, \cdots, e_m\}$，其中$e_k$发生的时间比$e_{k+1}$发生的时间早。在此$m$轮迭代过程中，依据参与方$u_i$与聚合服务器$S_j$的直接交互行为，$S_j$依次收集$u_i$当前交互信息，当前交互信息$CL_c^i$的形式化表示如式(1)所示。

$$CL_c^i = \{b_r =< u_i, c, acc_i^r, d_i^r >_r, r \in [1, m]\} \tag{1}$$

式(1)中，b_r为CL_c^i中存储的由四元素组成的第r条信息，表示u_i第r轮迭代产生的行为信息，由u_i第r轮迭代交互的上下文c信息、行为异常度acc_i^r和时延d_i^r共同组成。为了便于后面的计算，对应m轮迭代事件$\{e_1, e_2, \cdots, e_m\}$，把$u_i$的异常度信息组成长度为$m$的集合$acc_i = \{acc_i^1, acc_i^2, \cdots, acc_i^r, \cdots, acc_i^m\}$，$u_i$的时延信息同样组成长度为$m$的集合$d_i = \{d_i^1, d_i^2, \cdots, d_i^r, \cdots, d_i^m\}$。其中$S_j$对$u_i$在本次联邦学习中的行为可靠性的评判随着$acc_i^r$取值的减小和$d_i^r$取值的增大而不断减弱。由于具有高质量本地模型的用户能够使得本地损失函数和全局损失函数的快速收敛，同时，可靠高质量的通信环境可以减少参与方上传本地模型的时间，从而显著提高联邦学习效率，使得聚合服务器对该用户更信赖。因此，本研究使用异常度和时延作为联邦学习参与方的信任行为影响参与方信任度的评估。其中异常度可以通过聚合服务器对本地模型的异常检测方法检测得到，而时延同样可以通过聚合服务器检测得到，这里，时延指参与方上传本地模型的时间，包括本地模型的训练时间以及模型参数的传输时间。

参与方u_i参与聚合服务器S_j历史发布的联邦学习任务产生的信任信息，存储在历史信任信息队列HL_c^i中，HL_c^i的形式化表示如式(2)所示。hl_k为HL_c^i中存储的由三元素组成的第k条历史信息，其中t_{hl}为此条历史信息交互发生的时间，R_c^{hl}为u_i在上下文c中的行为可靠性，HL_c^i中最多存储P项信息。

$$HL_c^i = \{hl_k =< u_i, t_{hl}, R_c^{hl} >_k, k \in [1, P]\} \tag{2}$$

同时，S_j将系统中其他聚合服务器RS发布的关于u_i的推荐信任信息存储在推荐信任信息队列RL_c^i中，RL_c^i的形式化表示如式(3)所示。rl_d为RL_c^i中存储的由四元素组成的第d条推荐信息，其中rs为发布此条信息的推荐服务器，rt为推荐信任度，t_{rl}为推荐服务器发布此条推荐信任信息的时间，$h_c^{i,rs}$为上下文c中u_i与rs之间的历史交互次数，RL_c^i最多存储Q项信息。

$$RL_c^i = \{rl_d =< rs, rt, t_{rl}, h_c^{i,rs} >_d, d \in [1, Q]\} \quad (3)$$

为了便于后面信任度的计算，按照时间先后依次提取HL_c^i中存储的P项历史信任信息中的第二个元素t_{hl}形成长度为P的集合$list_{t_c^i} = \{t_1^{hl}, t_2^{hl}, \cdots, t_P^{hl}\}$，第三个元素$R_c^{hl}$形成长度为$P$的集合$list_{r_c^i} = \{R_1^{hl}, R_2^{hl}, \cdots, R_P^{hl}\}$；提取$RL_c^i$中存储的$Q$项推荐信息中的第二个元素$rt$形成长度为$Q$的集合$list_{rt_c^i} = \{rt_1, rt_2, \cdots, rt_Q\}$，第四个元素$h_c^{i,rs}$形成长度为$Q$的集合$list_{rh_c^i} = \{h_1^{i,rs}, h_2^{i,rs}, \cdots, h_Q^{i,rs}\}$。

上文所述的关于u_i的当前交互信息CL_c^i、历史信任信息队列HL_c^i、推荐信任信息队列RL_c^i，形成u_i在上下文c的信任证据$TE_c^{i,j}$。同时，信任证据$TE_c^{i,j}$中还存储由S_j记录的u_i与S_j的总交互次数$h_c^{i,j}$。信任证据$TE_c^{i,j}$形式化表示如式(4)所示。

$$TE_c^{i,j} =< CL_c^i, HL_c^i, RL_c^i, h_c^{i,j} > \quad (4)$$

此信任证据$TE_c^{i,j}$存储在聚合服务器S_j维护的信任证据数据库中，其中CL_c^i、HL_c^i和RL_c^i队列均按照时间顺序进行排序，最早入队的信任信息存储在队头位置。CL_c^i的长度根据本次学习的迭代次数而定，CL_c^i的长度为m。由于存储空间的限制，本研究假设HL_c^i最大长度为P，RL_c^i最大长度为Q，当存储的信任信息超过队列最大存储限度，此时认为队列已满，在收到新的信任信息时，需要将当前队列中处于队头位置的信息删除，即最早进队的信息，同时在队尾位置插入新的信息进行队列的更新。$h_c^{i,j}$由聚合服务器S_j记录。此外，此信任证据数据库中还存储参与过本聚合服务器S_j发布的联邦学习任务的其他各参与方的信任信息，同样包括当前交互信息CL、历史信任信息队列HL、推荐信任信息队列RL，和其他参与方与S_j的总交互次数。

依据u_i在信任证据数据库中存储的信任证据$TE_c^{i,j}$，对其进行行为分析和建模。首先，分别由u_i的异常度信息集合$acc_i = \{acc_i^1, acc_i^2, \cdots, acc_i^r, \cdots, acc_i^m\}$和时延信息集合$d_i = \{d_i^1, d_i^2, \cdots, d_i^r, \cdots, d_i^m\}$对$u_i$的异常度因子$\alpha_i$和时延因子$\beta_i$进行计算。具体计算方法如公式(5)和公式(6)所示。

$$\alpha_i = f_\alpha(\{acc_i\}) = 1 - e^{-\sum_{r=1}^{m}(acc_i^r \times \theta^{\Delta t_r})} = 1 - e^{-(acc_i^1 \times \theta^{\Delta t_1} + acc_i^2 \times \theta^{\Delta t_2} + \cdots + acc_i^m \times \theta^{\Delta t_m})} \quad (5)$$

$$\beta_i = f_\beta(< d_i >) = e^{-\sum_{r=1}^{m}(d_i^r \times \theta^{\Delta t_r})} = e^{-(d_i^1 \times \theta^{\Delta t_1} + d_i^2 \times \theta^{\Delta t_2} + \cdots + d_i^m \times \theta^{\Delta t_m})} \quad (6)$$

其中，$\alpha_i \in [0,1]$，$\beta_i \in [0,1]$，θ称为遗忘因子，$\theta \in [0,1]$。Δt_r表示当前时间和第r轮迭代事件e_r发生时间的时间间隔。某一轮迭代事件发生的时间距离当前时间越长，那么证明该轮迭代事件发生的时间越早，u_i该轮迭代事件的行为信息在u_i本次学习的异常因子α_i和时延因子β_i中所占的比重会越小。因此，α_i和β_i具有时间衰减性。

本研究根据异常因子α_i和时延因子β_i对u_i在上下文c的行为可靠性R_c^i和行为稳定性S_c^i进行计算。R_c^i反映u_i的行为是良好的还是恶意的，S_c^i反映u_i的行为是否始终保持稳定，$R_c^i \in [0,1]$，$S_c^i \in [0,1]$。如果u_i的行为可靠性R_c^i越接近 1，那么认为其行为表现越好；如果u_i的行为稳定性S_c^i越接近 1，则认为其行为表现越稳定。

$$R_c^i = sigmoid(\alpha_i) \cdot \beta_i = \frac{1}{1 + e^{-\alpha_i}} \cdot \beta_i \quad (7)$$

参与方u_i在上下文c中的行为可靠性R_c^i计算方法如公式(7)所示。其中，对u_i行为可靠性R_c^i产生影响的两个主要因素为异常因子α_i和时延因子β_i。R_c^i随着α_i和β_i取值的递减均逐渐减小。

公式(8)给出了u_i在上下文c中的行为稳定性S_c^i的计算方法。式中$R_p^{\hbar l}$表示集合$list_r_c^i = \{R_1^{\hbar l}, R_2^{\hbar l}, \cdots, R_P^{\hbar l}\}$中的第$p$个元素，$P$为集合$list_r_c^i$中的总元素数量。如果相邻时刻$u_i$行为可靠性的值越接近，那么认为$u_i$的行为在不同时刻的波动范围越小，行为稳定性$S_c^i$的值便越高。因此，无论是参与方始终保持良好行为还是始终保持恶意行为，均认为此参与方行为较稳定，对此参与方行为稳定性的评判也均处于较高水平。因而，本研究使用行为可靠性和行为稳定性两个因素对参与方进行行为建模，以此来对联邦学习参与方的行为可信程度进行度量。

$$S_c^i = \sum_{p=1}^{P-1} \frac{R_{p+1}^{\hbar l} - R_p^{\hbar l}}{P - 1} \quad (8)$$

基于u_i的行为可靠性R_c^i和行为稳定性S_c^i，对u_i行为模型进行形式化表示，如式(9)所示的四元组。

$$< u_i, c, R_c^i, S_c^i > \quad (9)$$

2.2.3 信任度评估方案

本节将依次介绍所提的联邦学习参与方直接信任度、推荐信任度和综合信任度的计算方法。首先，本研究所提出的针对单一上下文给出信任评估模块的基本框架如图 3 所示。

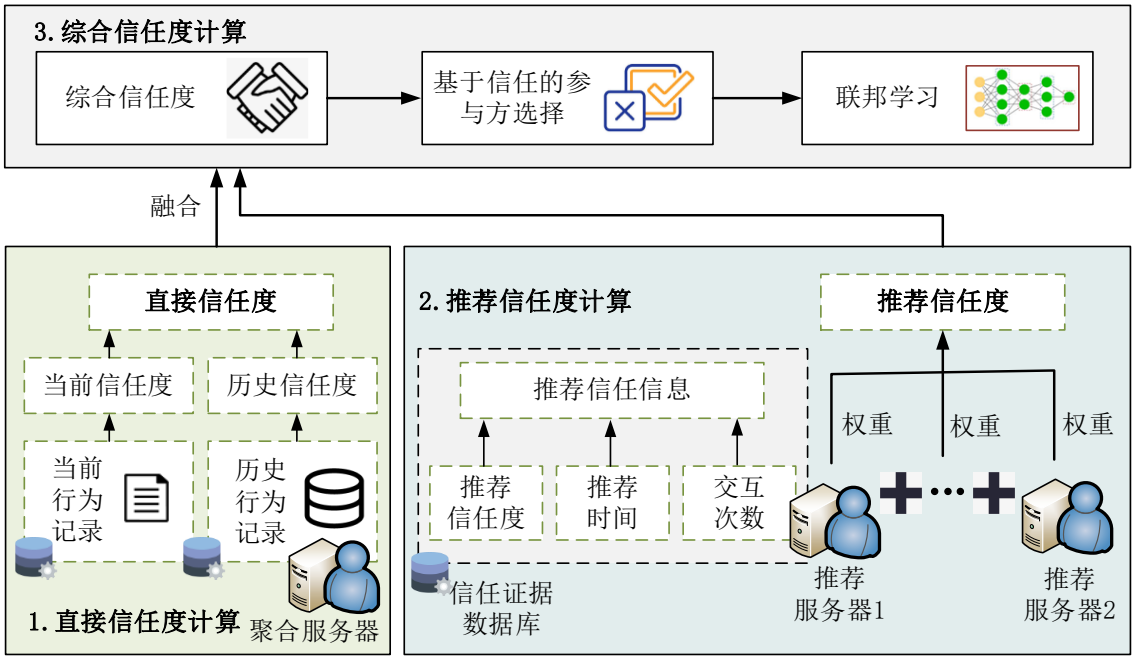


图 3 单一上下文联邦学习信任管理系统框架

该模块由直接信任度计算、推荐信任度计算和综合信任度计算三个部分组成：

（1）直接信任度计算。直接信任度计算的数据来源于参与方与聚合服务器直接交互产生的行为信息，包括当前行为记录以及历史行为记录，当前行为记录为信任证据数据库中存储的参与方的当前交互信息CL，历史行为记录为信任证据数据库中存储的参与方的历史信任信息队列HL，直接信任度由历史行为记录产生的历史信任度和当前行为记录产生的当前信任度综合得到；

（2）推荐信任度计算。推荐信任度由系统中其他多个聚合服务器推荐的关于参与方的信任信息加权计算得到，推荐信任信息存储在信任证据数据库中参与方的推荐信任信息队列RL中，包括推荐信任度、推荐时间以及推荐服务器与参与方的交互次数等信息；

（3）综合信任度计算。对于参与方综合信任度的计算重点需要考虑如何通过直接信任度和推荐信任度的动态的权重变化实现两者的动态融合。依据参与方

信任评估结果，聚合服务器能够在下一次发布联邦学习任务时完成参与方的选择以及联邦学习过程。

在上下文c中，S_j对u_i的直接信任度$dir_T_c^j(i)$计算方法如公式(10)所示，其中，ω_{hist}为历史信任度$hist_T_c^j(i)$的权重，ω_{curr}为当前信任度$curr_T_c^j(i)$的权重，具体计算方法如公式(11)和(12)所示。

$$dir_{T_c^j}(i) = \omega_{hist} \cdot hist_{T_c^j}(i) + \omega_{curr} \cdot curr_{T_c^j}(i) \tag{10}$$

$$\omega_{hist} = \Omega_{i,j} \cdot S_c^i \tag{11}$$

$$\omega_{curr} = 1 - \omega_{hist} \tag{12}$$

$$\Omega_{i,j} = \begin{cases} 1 - \dfrac{1}{\sqrt[\phi]{e^{h_c^{i,j}-1}+1}}, h_c^{i,j} \geq 1 \\ 0, else \end{cases} \tag{13}$$

参与方u_i的历史信任度$hist_T_c^j(i)$和当前信任度$curr_T_c^j(i)$的权重ω_{hist}和ω_{curr}随着聚合服务器S_j与u_i的熟悉程度$\Omega_{i,j}$以及u_i的行为稳定性S_c^i而动态变化，$\Omega_{i,j}$取决于S_j与u_i的交互次数，交互次数越多，那么认为S_j与u_i越熟悉，$\Omega_{i,j}$的取值就越大。$\Omega_{i,j}$具体计算方法如公式(13)所示，当S_j与u_i之间没有发生过历史交互行为时，$h_c^{i,j} = 0$，此时$\Omega_{i,j} = 0$。其中，ϕ为$\Omega_{i,j}$的调节因子，用于控制$\Omega_{i,j}$趋于 1 的速度，其取值范围为满足$\phi \geq 2$的任意常数，ϕ取值越大，$\Omega_{i,j}$趋于 1 的速度越慢。如果聚合服务器由于与参与方历史发生过多次直接交互而对参与方较为熟悉，同时此参与方具有较高的行为稳定性，那么此参与方的历史信任度便会在直接信任度的计算中占有较高的权重。

参与方u_i历史信任度$hist_T_c^j(i)$的计算方法如公式(14)所示。对u_i历史信任度$hist_T_c^j(i)$的计算综合历史信任信息队列HL_c^i中的所有信任信息，HL_c^i队列中的信任信息越早发生，此信任信息在u_i的历史信任度$hist_T_c^j(i)$中将占有越小的比重。因此，对于$hist_T_c^j(i)$的计算考虑时间衰减性，其中φ_p为时间衰减因子，t为当前交互发生的时间。式中R_p^{hl}为集合$list_r_c^i = \{R_1^{hl}, R_2^{hl}, \cdots, R_P^{hl}\}$中的元素，$t_p^{hl}$为

集合$list_t_c^i = \{t_1^{hl}, t_2^{hl}, \cdots, t_P^{hl}\}$中的元素。

$$hist_{T_c^j(i)} = \sum_{p=1}^{P} \left(R_p^{hl} \cdot \varphi_p\right), \varphi_p = 2^{-\left(t - t_p^{hl}\right)} \tag{14}$$

参与方u_i在上下文c的当前信任度$curr_T_c^j(i)$的计算方法如公式(15)所示，式中R_c^i为u_i与S_j本次交互过程中u_i的行为可靠性。$g(h)$为当前信任调节函数，具体计算方法如公式(16)所示，式中h表示参与方与聚合服务器的实际交互次数，u_i与S_j的实际交互次数表示为$h = h_c^{i,j}$。ε为当前信任调节因子，如公式(17)所示，式中的R_{curr}表示参与方的当前行为可靠性，u_i的当前行为可靠性$R_{curr} = R_c^i$。式(16)中λ为区间$[0,1]$的实数，用于调节本系统交互阈值H_{th}的大小，交互阈值H_{th}的计算如式(18)所示，交互阈值H_{th}的取值随着λ值的不断减小而增大。如果u_i的历史交互次数$h_c^{i,j} < H_{th}$，此时需要对u_i的当前信任度进行处理，$curr_T_c^j(i)$将趋向 0.5；如果$h_c^{i,j} \geq H_{th}$，此时由于交互次数达到阈值而不需要对当前信任度进行任何操作。当$R_c^i \leq 0.5$时，$curr\_T_c^j(i) = 0$。

$$curr_{T_c^j(i)} = \left(0.5 + g\left(h_c^{i,j}\right) \cdot \left(R_c^i - 0.5\right)\right) \cdot \varepsilon \tag{15}$$

$$g(h) = \begin{cases} \lambda h^2 + \dfrac{1}{2}, 0 \leq h \leq \dfrac{1}{\sqrt{2\lambda}} \\ 1, else \end{cases} \tag{16}$$

$$\varepsilon = \begin{cases} 1, R_{curr} > 0.5 \\ 0, else \end{cases} \tag{17}$$

$$H_{th} = \left\lfloor \frac{1}{\sqrt{2\lambda}} \right\rfloor \tag{18}$$

在对参与方当前信任度的计算过程中，以不确定性信任度 0.5 为基准，如果参与方与聚合服务器的历史交互次数小于本系统交互阈值H_{th}，那么认为此参与方的可信程度由于历史交互次数太少而具有较高不确定性，基于参与方本次学习的行为可靠性无法完全评判此参与方是否可信，因此，此参与方的信任度将会更趋向于不确定，即 0.5；而当双方历史交互次数达到本系统交互阈值H_{th}时，将不再对参与方本次的行为可靠性进行处理，此时参与方的当前信任度即为其当前的行为可靠性值，即，随着参与方与聚合服务器的交互次数的不断增大，聚合服务器对参与方本次学习的行为可信程度也不断增高，如果参与方本次的行为可靠性

低于0.5，认为此参与方本次行为较差，此时直接认为该参与方当前信任度为0。因此，当前信任调节函数$g(h)$和当前信任调节因子ε的设定是对参与方基于当前信任度的动态分析，对参与方当前信任度这样的设计满足信任度缓慢增长、快速下降的特点，以便于抵抗恶意参与方的攻击。

在上下文c中，u_i的推荐信任度$recom_T_c^j(i)$的计算方法如公式(19)所示。u_i推荐信任度的计算结合推荐信任列表RL_c^i中的所有推荐信息，根据推荐服务器与u_i的交互次数，对各推荐信息进行权重的分配，如果推荐服务器与参与方具有较高的历史交互次数，那么此推荐服务器推荐的信任信息将在推荐信任度中占有更大的权重。式中，$h_q^{i,rs}$为集合$list_rh_c^i = \{h_1^{i,rs}, h_2^{i,rs}, \cdots, h_Q^{i,rs}\}$中的元素，$H_c^{i,RS}$为集合$list_rh_c^i$中所有元素的和，计算方法如式(20)所示，$rt_q$为集合$list_rt_c^i = \{rt_1, rt_2, \cdots, rt_Q\}$中的元素。

$$recom_{T_c^j(i)} = \sum_{q=1}^{Q}\left(\frac{h_q^{i,rs}}{H_c^{i,RS}} \times rt_q\right) \tag{19}$$

$$H_c^{i,RS} = \sum_{q=1}^{Q} h_q^{i,rs} \tag{20}$$

聚合服务器S_j在上下文c中对于u_i的综合信任度计算方法如公式(21)所示，ω_{dir}为直接信任度$dir_T_c^j(i)$的权重，ω_{recom}为推荐信任度$recom_T_c^j(i)$的权重，关于ω_{dir}和ω_{recom}的具体计算方法如公式(22)和(23)所示。

$$T_c^j(i) = \omega_{dir} \cdot dir_{T_c^j}(i) + \omega_{recom} \cdot recom_{T_c^j}(i) \tag{21}$$

$$\omega_{dir} = \frac{\chi_i \cdot \frac{f(\sigma_i)}{f(\sigma_i)+1}}{\chi_i \cdot \frac{f(\sigma_i)}{f(\sigma_i)+1} + \gamma_i \cdot \frac{1}{f(\sigma_i)+1}} \tag{22}$$

$$\omega_{recom} = 1 - \omega_{dir} \tag{23}$$

$$\chi_i = \frac{h_c^{i,j}}{\frac{1}{N}\sum_{u_k \in u} h_c^{k,j}} \tag{24}$$

$$\gamma_i = \frac{H_c^{i,RS}}{\frac{1}{N}\sum_{u_k \in u} H_c^{k,RS}} \tag{25}$$

$$\sigma_i = \frac{h_c^{i,j}}{H_c^{i,RS}} \tag{26}$$

$$f_\omega(\sigma) = \begin{cases} \delta, \sigma \geq \delta \\ \sigma, else \end{cases} \tag{27}$$

由于在信任评估中不能保证总是有可用的可信本地交互信息或推荐信任信息，因此需要在信任评估函数中同时考虑本地交互信息和推荐信任信息。ω_{dir}和ω_{recom}的值随着u_i本地信任信息和推荐信任信息的相对数量而动态变化，依此确定S_j在对u_i进行信任计算时，能够在多大程度上依赖u_i的本地交互信息和推荐信任信息。ω_{dir}和ω_{recom}的值由参数χ_i、γ_i以及函数$f_\omega(\sigma_i)$决定，χ_i和$f_\omega(\sigma_i)$的值越大，直接信任度$dir_T_c^j(i)$的权重越大。χ_i的计算方法如公式(24)所示，反映了S_j与u_i直接交互次数和与$u = \{u_1, u_2, \cdots, u_i, \cdots u_N\}$内所有参与方之间直接交互的平均次数的比例；$\gamma_i$的计算方法如公式(25)所示，表示推荐服务器$RS$与$u_i$之间的直接交互次数和与$u = \{u_1, u_2, \cdots, u_i, \cdots u_N\}$内所有参与方之间的平均直接交互次数之比。$\sigma_i$表示$u_i$的本地信任信息与其推荐信任信息的相对质量，如式(26)所示。当u_i与S_j没有直接交互时，χ_i和ω_{dir}的值将为 0；当u_i与其他推荐服务器RS之间没有直接交互时，γ_i和ω_{recom}的值将为 0，ω_{dir}的值将为 1。函数$f_\omega(\sigma_i)$的计算方法如公式(27)所示，其中δ是函数$f_\omega(\sigma_i)$中的阈值，防止由于σ_i取值过大而产生不合理的数值.

2.3 IID 场景下隐私保护的拜占庭节点检测方案

2.3.1 系统检测思路

本章提出一种 IID 环境下隐私保护的拜占庭节点检测方案，该方案在文献[7]的基础上提出一种在聚合阶段无需 Shamir 秘密共享[8]的新掩码机制，可以在隐私保护环境下进行全局聚合。然后，聚合服务器基于应答-挑战机制检测拜占庭节点，其检测思路如下：首先，聚合服务器向所有节点发送一个挑战；接下来，节点根据计算结果向聚合服务器做出应答；最后，聚合服务器根据各节点的应答结果判断节点的异常情况。

2.3.2 模型构成

如图 4 所示，本报告方案的系统模型由三个部分组成，分别为：信任权威（Trust Authority，TA）、聚合服务器（Aggregation Server，AS）和节点集。TA 主要负责系统初始化以及公钥、私钥的配发。AS 主要有两个任务，分别为：（1）聚合节点上传的经过掩码加密的模型更新集合；（2）基于挑战-应答机制检测拜占庭节点。节点集由良性节点和拜占庭节点共同组成。良性节点遵从联邦学习协议，诚实地参与联邦学习流程。而拜占庭节点则伺机上传恶意模型更新以破坏联邦学习的正常训练流程。

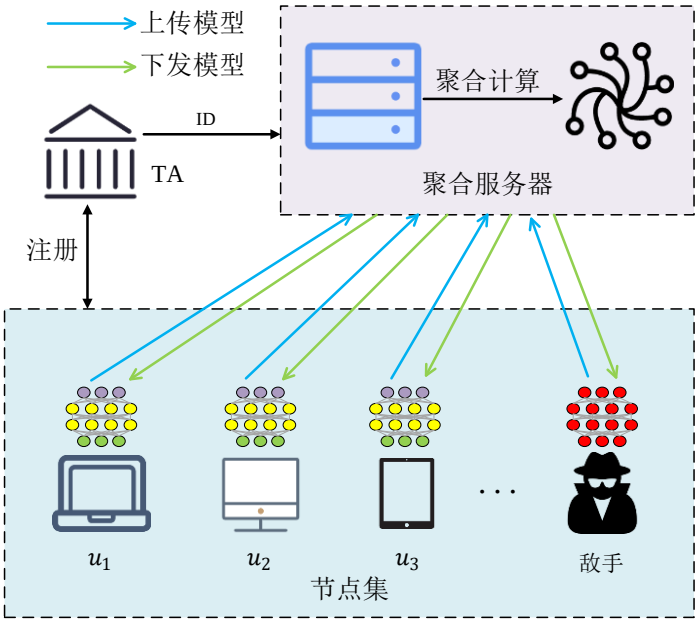


图 4 横向联邦学习系统模型图

根据有无本地数据集，本方案将拜占庭节点分为两类，如表 5 所示。AD1 型敌手持有本地训练集，试图攻破挑战-应答机制破坏全局模型。AD1 型敌手发动“搭便车”攻击[9]的动机不强，因为“搭便车”的目的是为窃取可用的模型，而 AD1 型敌手本就持有训练集，其完全可以正常参与联邦学习以获得更高性能的全局模型。AD2 型敌手没有本地训练集，攻击目的明确，即破坏全局模型或“搭便车”以窃取可用模型。

表 2 挑战-应答方案敌手类型

敌手类型	本地训练集	攻击目的
AD1	有	破坏全局模型

AD2	无	破坏全局模型或搭便车

为进一步描述所提方案，作如下系统假设：

1) 联邦学习系统中共有n个节点，用集合$\{u_i\}_{i=1}^{n}$表示；

2) TA 负责生成 DH 密钥协商协议所需的(p,g)密钥对以及协助节点交换中间加密参数；

3) 节点间不可共谋，且各节点的本地训练集独立同分布；

4) 聚合服务器诚实且好奇。

2.3.3 方案执行流程

在上述联邦学习系统中进行一次模型训练的大致流程如下：

（1）TA 协助节点进行密钥协商，此密钥为生成掩码的伪随机数生成器的随机数种子；（2）节点在本地进行模型训练并得到模型更新，然后用掩码加密模型更新得到掩码模型更新集合并将其上传聚合服务器；（3）聚合服务器接收各节点上传的掩码模型更新集合并生成混淆模型更新列表，然后向所有节点发起挑战；（4）节点针对挑战做出应答并将应答结果上传聚合服务器；（5）聚合服务器验证节点的应答结果以此检测拜占庭节点；（6）聚合服务器根据检测结果进行全局聚合得到全局模型并将其下发给良性节点；（7）返回步骤（2）开始新的一轮训练，直到模型收敛或者达到终止条件。

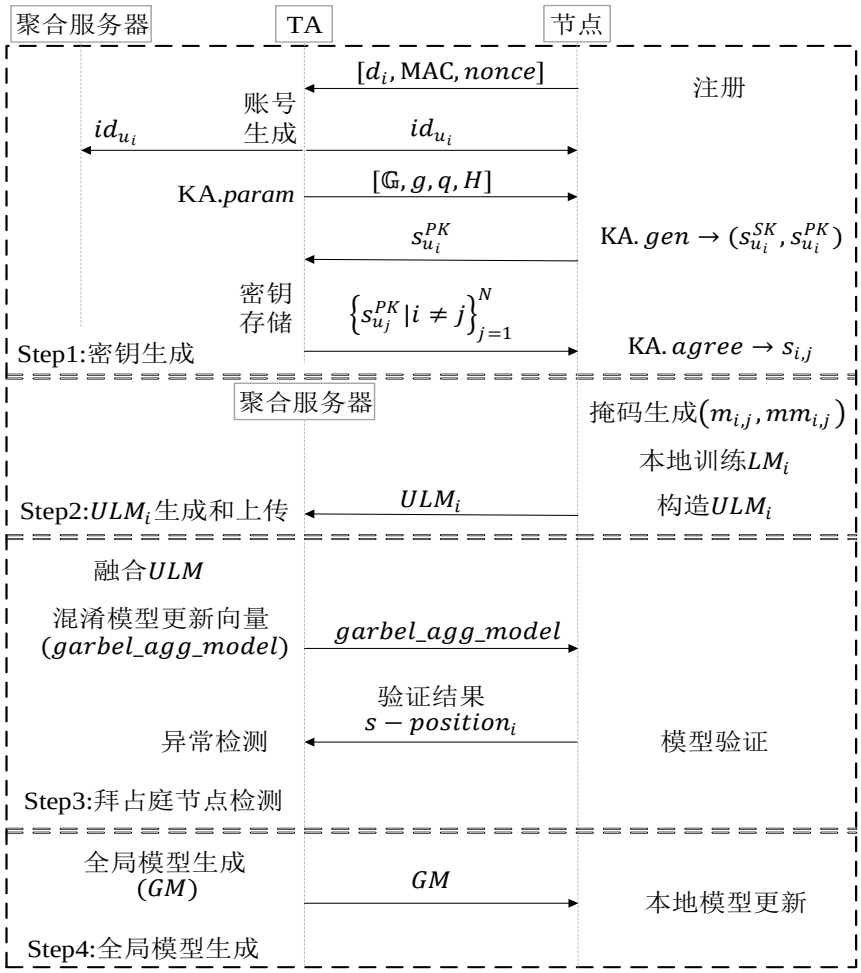


图 5 挑战-应答方案系统流程图

如图 5 所示，本报告所提拜占庭检测方案的执行流程大致可分为以下四个步骤：

1）密钥生成：各节点进行密钥协商生成相应的密钥，该密钥作为伪随机数生成器的随机数种子；

2）掩码模型更新集合生成和上传：节点在本地进行模型训练获得模型更新，然后生成掩码加密的模型更新集合，最后将掩码模型更新集合上传聚合服务器；

3）拜占庭节点检测：聚合服务器根据各节点上传的掩码模型更新集合生成混淆模型更新列表并将其作为挑战信息分发给各节点，然后验证节点的应答内容，从而检测节点的异常情况；

4）全局模型生成：聚合服务器根据拜占庭节点检测结果进行模型聚合生成全局模型更新。

下面对上述四个步骤进行详细阐述，其中涉及的相关符号的含义见表 6 所示。

表 3 挑战-应答方案符号说明表

符号	含义	符号	含义
n	节点数量	$MLM_{i,j}$	经掩码$m_{i,j}$处理后的本地模型
th	预定义门限	MLM_i	经掩码$mm_{i,j}$处理后的本地模型
$s_{u_i}^{SK}$	节点u_i的私钥	ULM_i	u_i的经过掩码加密的模型更新集合
$s_{u_i}^{PK}$	节点u_i的公钥	MMA	聚合服务器生成的模型更新矩阵
d_i	节点u_i的数据量	agg_model	聚合服务器生成的模型更新列表
m	拜占庭节点比例	$garble\_agg\_model$	混淆模型更新列表
$s_{i,j}$	u_i和u_j的共享密钥	$garble\_model\_len$	混淆模型更新列表的长度
$m_{i,j}$,$mm_{i,j}$	节点u_i的两类掩码	$o-position$	记录模型更新列表中的元素在混淆模型更新列表中的索引位置的集合
LM_{u_i}	节点u_i的本地模型	$s-position_i$	节点u_i对挑战的应答
L	agg_model的长度	$inter-position$	$o-position$和$s-position$的交集

（1）密钥生成

密钥协商算法由密钥生成和密钥协商两部分组成。密钥生成包含KA.param和KA.gen两个算法，密钥协商由KA.agree算法实现，下面对这三个算法进行具体描述。

KA.param$(k)\rightarrow pp$：KA.param算法使用保密参数k生成一些公共参数$pp=(\mathbb{G},g,q,H)$，其中q是一个大素数，g是q的本原根，$\mathbb{G}$是阶为q的群，H是哈希函数。

KA.gen$(pp)\rightarrow(s_u^{SK},s_u^{PK})$：任意节点$u$可通过KA.gen算法生成一个私钥-公钥密钥对。例如，u随机选择$x\leftarrow\mathbb{Z}_q$作为私钥s_u^{SK}，然后计算公钥$s_u^{PK}=g^x\mathrm{mod}(q)$，最终得到密钥对$(s_u^{SK},s_u^{PK})$。

KA.agree$(s_u^{SK},s_v^{PK})\rightarrow s_{u,v}$：任意节点$u$和$v$可通过KA.agree算法协商共享密钥$s_{u,v}=H\left((s_v^{PK})^{(s_u^{SK})}\right)$，进一步有$H\left((s_v^{PK})^{(s_u^{SK})}\right)=H\left((g^{(s_v^{SK})})^{(s_u^{SK})}\right)$，显然有KA.agree$(s_v^{SK},s_u^{PK})=$KA.agree$(s_u^{SK},s_v^{PK})$，即$s_{u,v}=s_{v,u}$。KA.agree算法本质是节点$u$和$v$交换各自的公钥，然后计算共享密钥。

密钥生成阶段的主要任务为 TA 协助节点注册账号和协商密钥，具体执行步骤如下所示：

1）节点u向 TA 发送一个三元组$(d, \mathrm{MAC}, nonce)$信息，其中d表示u本地训练集的数据量，MAC表示u的硬件地址，$nonce$是一个随机数；

2）针对每个节点，TA 根据三元组信息生成一个唯一的身份标识id_i并将其发送给u_i和聚合服务器；

3）TA 根据$\mathrm{KA.param}$算法生成公共参数$pp = (\mathbb{G}, g, q, H)$并在联邦学习系统内进行广播；

4）节点u接收pp后基于$\mathrm{KA.gen}$算法生成私钥s_u^{SK}和公钥s_u^{PK}，然后将s_u^{PK}发送给 TA；

5）TA 接收各节点的公钥，然后再将其在联邦学习系统内进行广播。

上述密钥协商流程完成后，节点u根据本地私钥s_u^{SK}和其它节点的公钥s_v^{PK}计算共享密钥$s_{u,v} = \mathrm{KA.agree}(s_u^{SK}, s_v^{PK})$。

（2）**掩码模型更新集合生成和上传**

节点u_i首先向其本地模型更新LM_i添加掩码来实现隐私保护，接着生成模型更新集合ULM_i，然后将其上传聚合服务器，具体流程如下：

1）节点u_i在本地进行模型训练得到本地模型LM_i；

2）针对节点$u_j, i \neq j$，节点u_i根据公式(28)生成掩码$m_{i,j}$，然后再根据公式(29)生成掩码$mm_{i,j}$，其中$s_{i,j}$是u_i和u_j的共享密钥，公式(30)决定掩码的符号，$m_{i,j}$和$mm_{i,j}$与LM_i的结构相同；

$$m_{i,j} = \begin{cases} \Delta_{i,j} \cdot \mathrm{PRG}(s_{i,j}), i \neq j \\ 0, \quad i = j \end{cases} \tag{28}$$

$$mm_{i,j} = \begin{cases} \Delta_{i,j} \cdot \mathrm{PRG}(s_{i,j} + 1), i \neq j \\ 0, \quad i = j \end{cases} \tag{29}$$

$$\Delta = \begin{cases} 1, & i > j \\ -1, & i < j \end{cases} \tag{30}$$

3）节点u_i根据公式(31)生成集合$\{MLM_{i,j}\}_{j=1}^{n}$，然后再根据公式(32)生成MLM_i，其中$j \in [1, n]$且$i \neq j$；

4) 节点u_i根据公式(33)将$\{MLM_{i,j}\}_{j=1}^{n}$和MLM_i合并得到掩码模型更新集合ULM_i，然后将其上传聚合服务器。

$$MLM_{i,j} = LM_i - m_{i,j} \tag{31}$$

$$MLM_i = LM_i - \sum_{j=1}^{N} mm_{i,j} \tag{32}$$

$$ULM_i = \left\{\{MLM_{i,j}\}_{j=1}^{N}, MLM_i\right\} \tag{33}$$

（3）拜占庭节点检测

聚合服务器首先将各节点上传的ULM融合并生成混淆模型更新列表，然后向节点发起挑战并等待节点应答，最后根据节点的应答结果检测拜占庭节点，具体流程如下：

1) 聚合服务器根据公式(34)将各节点上传的模型更新集合ULM融合生成一个n行n列的模型更新矩阵MMA，然后根据算法 1 生成模型更新列表。算法 1 的输入为MMA，第 2-7 行是在计算任意两个节点的模型更新的平均值（需要计算C_n^2次），从而避免单个节点的原始模型更新暴露，第 8 行是在计算全部节点的平均模型，最后输出聚合服务器生成的模型更新列表；

2) 聚合服务器生成一个随机模型列表$random_model$，其中每个随机模型的维度与聚合服务器生成的模型更新列表中模型的维度一致，且前者的长度是后者的整数倍，然后将$random_model$和聚合服务器生成的模型更新列表合并且将其中元素顺序打乱后得到混淆模型更新列表，最后将其作为挑战下发给各节点；

$$MMA = \begin{pmatrix} MLM_{1,2} & MLM_{1,3} & \cdots & MLM_{1,n} & MLM_1 \\ MLM_{2,1} & MLM_{2,3} & \cdots & MLM_{2,n} & MLM_2 \\ \vdots & \vdots & & \vdots & \vdots \\ MLM_{n,1} & MLM_{n,2} & \cdots & MLM_{n,n-1} & MLM_n \end{pmatrix} \tag{34}$$

3) 聚合服务器计算生成的模型更新列表和混淆模型更新列表的交集生成集合$o - position$，其存储的信息为聚合服务器生成的模型更新列表中的元素在混淆模型更新列表中的位置索引；

4) 节点用本地训练集依次测试混淆模型更新列表中的模型，挑选精度最高的$C_n^2 + 1$个模型，然后按照精度从高到低的顺序记录这$C_n^2 + 1$个模型在混淆模

型更新列表中的位置索引得到节点u_i对挑战的应答，最后将节点u_i对挑战的应答作为应答内容上传聚合服务器；

5) 针对任一节点u_i，聚合服务器计算记录模型更新列表中的元素在混淆模型更新列表中的索引位置的集合和节点u_i对挑战的应答的交集。若任一节点的此交集都为空集，则联邦学习遭受严重攻击，训练流程终止；反之，聚合服务器根据不等式(35)成立与否依次判断每个节点的异常情况。不等式(35)表示只要交集的长度大于一定阈值，u_i便被判断为良性节点，否则视为拜占庭节点，其中$|\cdot|$表示集合的长度，th为预定义的门限。

$$|inter - position| > th \times |agg\_model| \tag{35}$$

公式(35)所示的拜占庭节点判定方式只能检测 AD2 型敌手，因为 AD1 型敌手有本地数据，可区分混淆模型更新列表中的真实模型参数和随机生成的模型参数。针对 AD1 型敌手可采用类似 Krum 的安全聚合方案，例如先选取各节点的记录模型更新列表中的元素在混淆模型更新列表中的索引位置的集合和节点u_i对挑战的应答的交集中前$0.5 \times th \times |agg_model|$的元素组成一个新的集合，然后统计其中出现频次最高的一个或者几个元素进行加权平均生成全局模型。

算法 1 模型更新列表agg_model生成算法

输入：模型更新矩阵MMA

输出：聚合服务器生成的模型更新列表

$k \leftarrow 0,\ agg\_model \leftarrow []$
for $i \leftarrow 1 \sim (n-1)$ **do**:
    for $j \leftarrow (i+1) \sim n$:
        $agg\_model[k] = \frac{MLM_{i,j}+MLM_{j,i}}{2}$
        $k++$
    end for
end for
$agg\_model[k] = (\sum_{t=1}^{n} MLM_t)/n$
return agg_model

（4）全局模型生成

聚合服务器根据公式(36)选择良性节点的本地模型更新以数据量为权值进行加权聚合，得到第t轮的全局模型GM^t，其中$bs = \{u_i | u_i$是良性节点$\}$表示良性

节点集，lbs表示bs中节点数量，公式(37)表示良性节点的数据量总和。

$$GM^t = \frac{\sum_{u_i,u_j \in bs, i \neq j} \left(\frac{d_{i+} d_j}{2 \times dbs} \times MLM_{i,j} \right)}{C_{lbs}^2} \tag{36}$$

$$\mathrm{dbs} = \sum\nolimits_{u \in bs} d_u \tag{37}$$

2.4 本章小结

本研究旨在建立可信的联邦学习系统，通过评估参与方的可靠性以及生成全局模型的准确性来实现，本章针对信任度评估提出新的单一上下文中联邦学习参与方的细粒度信任评估方案以及针对隐私保护的异常检测提出基于挑战-应答方式的拜占庭节点检测方案：

（1）本研究所提出的细粒度信任评估方案专为联邦学习系统设计，旨在全面评估参与方的信任度。该方案基于多种信任属性分析用户行为信息，通过设置异常因子和时延因子进行调节，再从多个维度挖掘参与方在参加联邦学习模型训练的每轮迭代过程中的行为特征以对其进行细粒度的行为建模，计算参与方的行为可靠性和行为稳定性。在每轮联邦学习模型训练过程中，基于行为模型计算参与方的直接信任度和推荐信任度，设计两者的动态的权重变化实现两者的动态融合方法，进而评估其综合信任度。

（2）本研究所提出的 IID 场景下隐私保护的拜占庭节点检测方案，基于安全多方计算提供精度无损的隐私保护环境，因此聚合服务器在整个训练流程中无法访问节点的原始模型更新信息，可以在隐私保护环境下进行全局聚合。首先，各节点协商密钥并向本地模型更新添加掩码实现隐私保护，将其上传至聚合服务器，接着，聚合服务器基于应答-挑战机制检测拜占庭节点，即聚合服务器将混淆模型更新列表作为挑战下发给各节点，而节点需要识别混淆模型更新中的真正的模型更新，并将识别结果作为应答内容上传聚合服务器，最后聚合服务器根据各节点的应答结果判断节点的异常情况，选择良性节点的本地模型更新进行加权聚合，生成全局模型。

第三章 作品测验与分析

3.1 信任度评估方案测试

为了验证本研究所提方案的有效性，采用 PyCharm 平台搭建了联邦学习仿真实验环境，使用本研究所提信任评估方案分析具有不同行为模式的参与方信任度变化趋势，同时考察本方案相关参数的设定对信任评估结果的影响，并对本研究所提信任评估方案与现有的被广泛应用的基于主观逻辑的信任评估方法进行了对比。

3.1.1 测试设备及参数设置

实验环境见表 2 所示，实验过程中的相关仿真参数及取值见表 3 所示。

表 4 单一上下文信任评估方案实验环境

操作系统	Windows 10
CPU	Intel Core i5-10400
内存	16.00GB（RAM）
编程环境	Python 3.8.8，anaconda 4.10.1

系统中部署了 100 个聚合服务器与 100 个参与方，其中聚合服务器S_1作为发布任务的主聚合服务器完成模型聚合以及参与方的信任度评估，其余$\{S_2, S_3, \cdots, S_{100}\}$共 99 个聚合服务器作为推荐服务器向$S_1$提供关于参与方的推荐信任信息，所有参与方的初始信任度为不确定信任度 0.5。此外，本实验认为信任度处于 0.8-1 的行为为良好行为，信任度处于 0-0.4 的行为为恶意行为，其余为行为不确定，无法给出直接定论。同时，对于联邦学习系统中参与方可能出现六种不同的行为模式总结如表 4 所示。为了跟踪具有不同行为模式的联邦学习参与方在与聚合服务器的多次交互过程中信任度的变化趋势，本研究假设无论参与方是否恶意，均参与每次学习过程。

表 5 单一上下文信任评估方案实验参数设置

参数	取值	描述

M	100	聚合服务器数量（个）
N	100	参与方数量（个）
ϕ	10	熟悉度函数中的调节因子
H_{th}	70	交互阈值
λ	0.00010	$g(h)$ 调节函数中的调节因子

表 6 单一上下文信任评估方案参与方行为模式

行为模式	描述
行为一	诚实参与方，始终保持良好行为
行为二	恶意参与方，始终保持恶意行为
行为三	恶意参与方，10 次良好行为，5 次恶意行为交替进行
行为四	恶意参与方，10 次良好行为，10 次恶意行为交替进行
行为五	恶意参与方，1 次良好行为，1 次恶意行为交替进行
行为六	恶意参与方，5 次良好行为，10 次恶意行为交替进行

3.1.2 实验结果及分析

首先跟踪具有六种不同行为模式的参与方在与聚合服务器的 100 次交互学习周期的信任度变化趋势，参与方与聚合服务器的一个交互学习周期为一次联邦学习过程。结果如图 6 所示，可以看到，具有行为一的诚实参与方在与聚合服务器短短几次交互之后，其信任度能够迅速上升到 0.8，这是因为诚实参与方的推荐信任度很高，推荐信任度会在短时间之内将该参与方信任度提高到一定程度，然后随着交互次数的增加而缓慢增加。而除具有行为一的诚实参与方之外，具有其他行为（行为二至行为六）的恶意参与方的信任度均在短时间之内降至 0.4 以下。因此，本研究所提方案能够使聚合服务器准确识别恶意参与方。更具体的说，具有行为二至行为六的恶意参与方也会因为行为模式的不同具有不同的变化趋势。具有行为二的参与方由于始终执行恶意行为而具有较低的信任度，在与聚合服务器的一次交互之后便可以被识别。对于交替执行良好行为和恶意行为的参与方（行为三、行为四和行为六），在交替周期中信任度随着执行良好行为次数比例的增高而增大。但比较具有行为四和行为五的参与方，虽然执行良好行为次数和恶意行为次数占比是相同的，但由于良好行为和恶意行为交替进行的时间间隔不断减小，参与方的信任度会有一个小幅度的增高趋势，但即便如此，具有行为五的参与方在与聚合服务器的几次交互之后仍然能够被识别为恶意用户。因此，

具有不同行为模式的参与方的可信程度均能够被准确评估，从而为联邦学习系统参与方的选择提供有效的决策支持，以此提高联邦学习系统的可靠性。

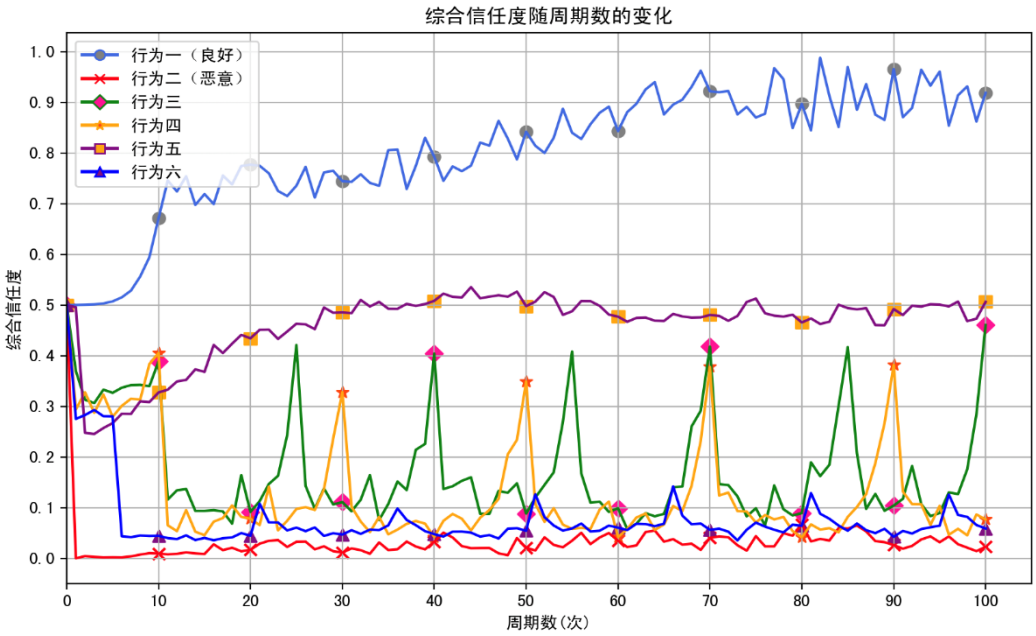


图 6 六种行为模式参与方信任度变化趋势

为了验证诚实参与方突发的偶然故障对其信任度的影响，模拟诚实参与方在交互过程中突发故障的场景，跟踪其在与聚合服务器 100 次的交互学习周期过程中信任度的变化趋势，实验结果如图 7 所示。可以看到，在第 43 次交互学习周期时，由于突发的偶然故障导致该参与方信任度的急剧下降。然而，随着随后与聚合服务器持续的良好行为交互，其信任度首先缓慢上升，然后在累积几次良好交互之后，其信任度很快达到之前的高度。因此，如果诚实参与方的信任度由于偶尔的故障问题而降低，那么在连续几次良好交互之后，该参与方的信任度仍然可以恢复到发生故障之前的相同水平，并不会因此被聚合服务器判断为恶意参与方。在整个交互过程中，其信任度始终没有低于 0.4，因此不会被聚合服务器评判为恶意参与方。由此可以看出本研究所提出的方案具有良好的鲁棒性。

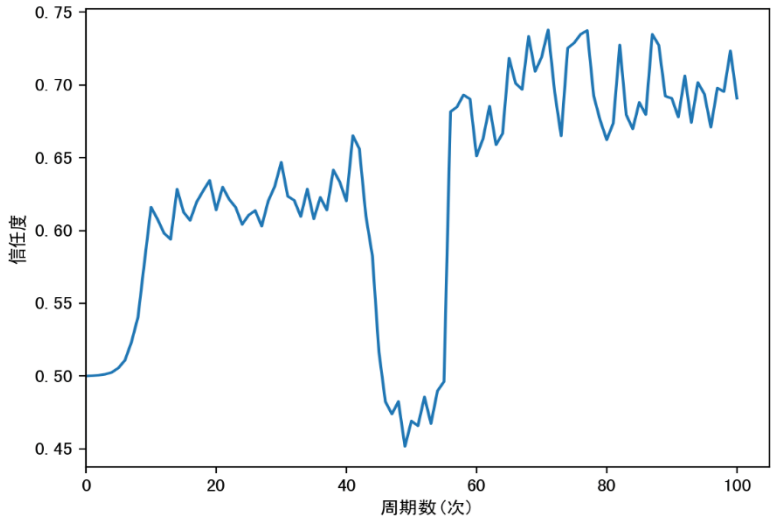


图 7 诚实参与方交互过程中突发故障信任度变化趋势

接下来，为研究参与方与聚合服务器之间的熟悉度Ω对其信任水平的影响，假设系统中的所有 100 个参与方均为诚实参与方，并将他们平均分为两组。其中组 1 包含 50 个参与方，编号 1-50，组 2 包含 50 个参与方，编号 51-100。在聚合服务器发布联邦学习任务后，首先组 1 中 50 个参与方参与前 50 次交互学习周期，此时组 2 中参与方始终保持初始状态 0.5；组 2 中的参与方则在 50 次交互学习周期之后开始参与学习，以第 51 个交互学习周期为基准，此时组 1 中参与方与聚合服务器历史交互次数为 50 次，组 2 中参与方与聚合服务器历史交互次数为 0 次，不同的历史交互次数导致参与方与聚合服务器的熟悉度Ω不同。在不同熟悉度Ω的前提下，组 1 和组 2 中所有参与方参与接下来的 50 个交互学习周期，跟踪参与方信任度变化趋势，结果如图 8 所示，可以看到，随着参与方与聚合服务器的历史交互次数增多，参与方与聚合服务器之间的熟悉度Ω增大，从第 51 次交互学习周期开始，组 1 中的参与方由于聚合服务器对其更熟悉而使其信任度更加接近真实值，而组 2 中参与方由于历史没有和聚合服务器产生过交互而使其熟悉度为 0，但随着交互次数增加，组 2 中参与方的信任度也会越来越接近组 1 中参与方信任度。

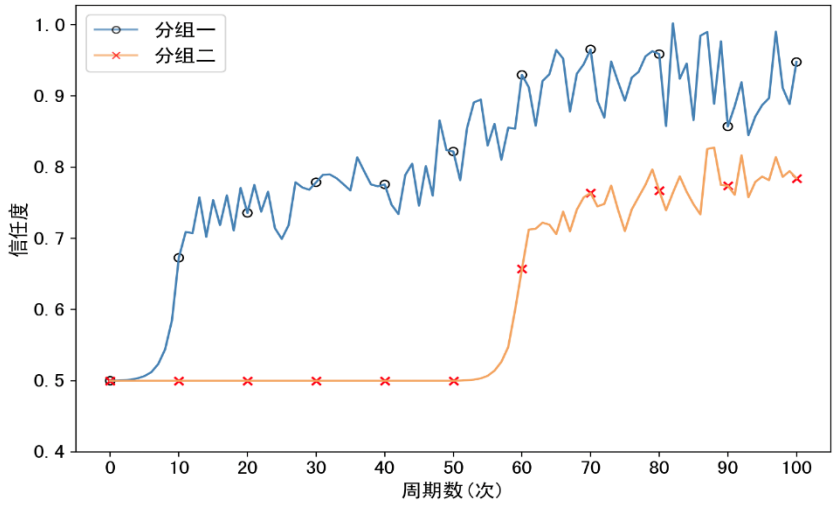


图 8 熟悉度对信任度演化的影响

为研究分析本方案时间衰减因子φ_p对信任度变化的影响，对比存在时间衰减因子φ_p和不存在时间衰减因子φ_p某参与方信任度的变化趋势。对于存在时间衰减因子φ_p的情况，对参与方的信任度计算即为本方案的信任度计算方法；对于不存在时间衰减因子φ_p的情况，认为在历史信任度计算中，无论历史交互发生时间距离当前多久，历史信任信息队列HL中存储的所有历史行为可靠性具有相同的权重，这种情况下，参与方的历史信任度为该参与方所有历史行为可靠性的均值。此外，设计一种行为恶意参与方，该参与方为了积累自己的信任度能够保持在较高水平，在与聚合服务器的前 88 次交互学习周期中始终保持良好行为，从第 89 次学习开始执行恶意行为，此参与方使用mu表示，跟踪mu在存在时间衰减因子φ_p和不存在时间衰减因子φ_p时信任度的变化趋势，结果如图 9 所示，可以看到，当使用时间衰减因子φ_p时，mu的信任度可以从第 89 次交互学习周期开始迅速降至 0.4 以下，聚合服务器能够很快发现mu是恶意参与方。而对于不考虑时间衰减因子φ_p的情况，虽然mu的信任度同样也会下降，但是由于下降幅度太小，即使在已经执行了 10 次恶意行为的情况下，其信任度仍高于 0.5，聚合服务器在短时间之内很难发现此恶意用户mu。

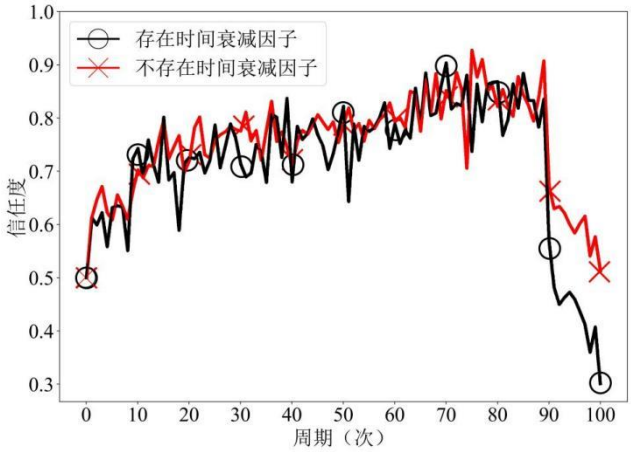
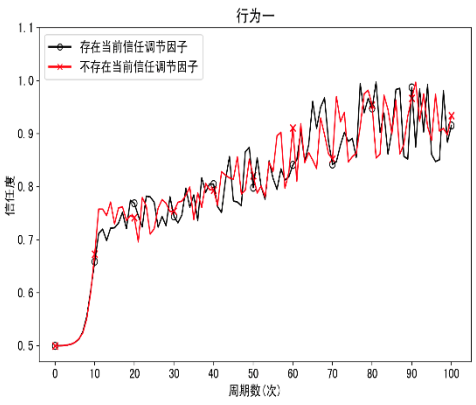
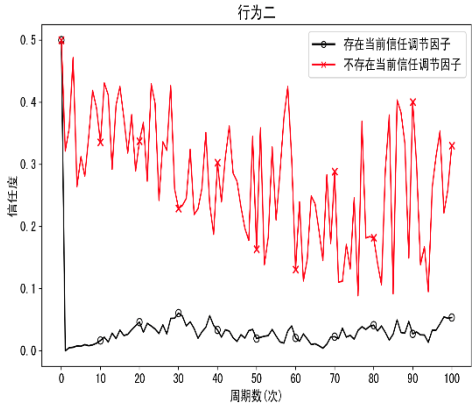


图 9 时间衰减因子对信任度演化的影响

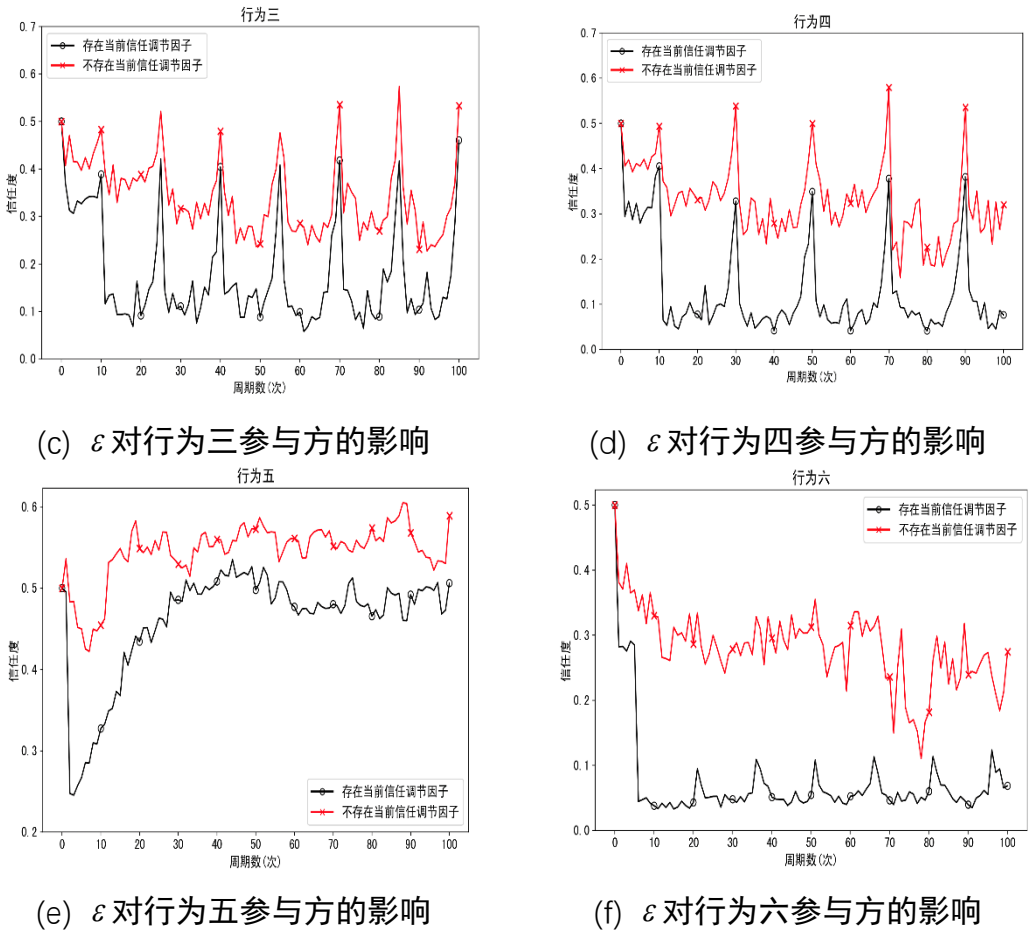
此外，为研究分析当前信任调节因子ε对信任度变化的影响，在本实验中对比$\varepsilon = 1$和ε为公式(17)所示时具有不同行为模式的参与方信任度变化趋势，当$\varepsilon = 1$时，可以认为在对参与方进行信任度评估时ε不会对评估结果产生影响，对比这样两种情况下，六种不同行为模式参与方的信任度变化趋势，结果如图 10 所示，可以看到，对于具有行为一的诚实参与方（图 10 (a)），ε对信任度的评估结果几乎没有影响，这是因为此诚实参与方一直保持当前行为可靠性大于 0.5，所以公式(17)的结果总是为 1；而对于恶意参与方（图 10 (b, c, d, e, f)），无论是持续执行恶意行为的参与方还是交替执行恶意行为的参与方，ε的存在均可以抑制恶意行为参与方信任度的增长，这是当因为恶意参与方行为可靠性在某一交互学习周期低于 0.5 时，将对该恶意参与方的当前信任度置为 0，这会降低其综合信任度。



(a) ε 对行为一参与方的影响



(b) ε 对行为二参与方的影响



(c) ε对行为三参与方的影响　(d) ε对行为四参与方的影响

(e) ε对行为五参与方的影响　(f) ε对行为六参与方的影响

图 10 当前交互信任调节因子对信任度演化的影响

3.2 拜占庭节点检测方案测试

3.2.1 实验环境与参数设置

实验环境与参数设置如表 7 所示，本实验基于 MPI 编程实现一个并发的机器学习架构来仿真联邦学习环境。联邦学习的训练集使用 MNIST 手写数据集，其训练集有 60000 条样本，测试集有 10000 条样本。MNIST 数据集的每张图片都是28×28比特的灰度图，标签为 0-9 之间的数字。联邦学习的本地训练模型选择逻辑回归（Logistic Regression，LR）模型和神经网络（Neural Network，NN）模型。下面将从所提方案的可行性、性能和资源开销等方面开展相关实验并对实验结果进行分析，然后将现有方案同本方案进行比较以说明本方案的优越性。

表 7 实验环境与参数设置

操作系统	Ubuntu 18.04	n	10, 20
CPU	2*Intel 4210R	th	0.3, 0.4, 0.5
内存	5*32G	$garble\_model\_length$	$2L$, $3L$, $4L$
仿真环境	Python 3.6.5 Tensorflow 1.12.0 mpi4py 3.0.3	m	0.1, 0.2, 0.3 0.4, 0.5, 0.6

3.2.2 方案结果及分析

（1）掩码模型更新集合生成和上传：

节点u_i首先向其本地模型更新LM_i添加掩码来实现隐私保护，接着生成模型更新集合ULM_i，然后将其上传聚合服务器，具体流程如下：

5) 节点u_i在本地进行模型训练得到本地模型LM_i；

6) 针对节点$u_j, i \neq j$，节点u_i根据公式(28)生成掩码$m_{i,j}$，然后再根据公式(29)生成掩码$mm_{i,j}$，其中$s_{i,j}$是u_i和u_j的共享密钥，公式(30)决定掩码的符号，$m_{i,j}$和$mm_{i,j}$与LM_i的结构相同；

$$m_{i,j} = \begin{cases} \Delta_{i,j} \cdot \mathrm{PRG}(s_{i,j}), i \neq j \\ 0, \quad i = j \end{cases} \tag{28}$$

$$mm_{i,j} = \begin{cases} \Delta_{i,j} \cdot \mathrm{PRG}(s_{i,j} + 1), i \neq j \\ 0, \quad i = j \end{cases} \tag{29}$$

$$\Delta = \begin{cases} 1, & i > j \\ -1, & i < j \end{cases} \tag{30}$$

7) 节点u_i根据公式(31)生成集合$\{MLM_{i,j}\}_{j=1}^{n}$，然后再根据公式(32)生成MLM_i，其中$j \in [1,n]$且$i \neq j$；

8) 节点u_i根据公式(33)将$\{MLM_{i,j}\}_{j=1}^{n}$和MLM_i合并得到掩码模型更新集合ULM_i，然后将其上传聚合服务器。

$$MLM_{i,j} = LM_i - m_{i,j} \tag{31}$$

$$MLM_i = LM_i - \sum_{j=1}^{N} mm_{i,j} \tag{32}$$

$$ULM_i = \left\{\left\{MLM_{i,j}\right\}_{j=1}^{N}, MLM_i\right\} \tag{33}$$

（1）性能分析：

所提拜占庭检测方案的检测性能主要由拜占庭节点比例m，门限th和混淆模型更新列表的长度这三个因素决定。下面通过实验来具体分析这三个因素对所提方案性能的影响，其中节点的本地训练模型使用逻辑回归模型。

图 11 给出了$n = 10$，$th = 0.5$，混淆模型更新列表的长度为$3L$条件下不同拜占庭节点比例对全局模型精度的影响，其中拜占庭节点比例分别设置为 0.1、0.2、0.3、0.4、0.5 和 0.6。实验结果表明，当$m = 0.6$时，全局模型精度在第 12 轮仍然可以接近 80%，与$m = 0.1$时的全局模型精度的区别不大。不论拜占庭节点在联邦学习系统中的占比有多高，最终的全局模型精度总是在 70%-80%之间，这表明所提检测方案成功检测出了所有拜占庭节点，从而避免全局模型精度大幅降低。

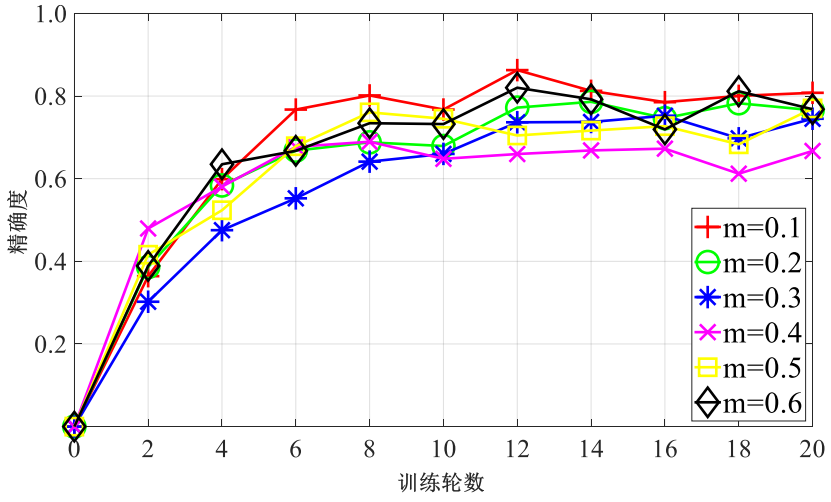


图 11 n=10, th=0.5, 混淆模型更新列表的长度为 3L 时 m 对全局模型精度的影响

对全局模型精度的影响，其中门限分别设置为 0.3、0.4 和 0.5。从图 12 中可知，门限并非越严格越好。这是因为节点的本地训练集不同和模型收敛前精度不稳定导致节点验证模型的结果（节点针对聚合服务器发出的挑战所做的应答）不是特别准确，因此存在良性节点被误判为拜占庭节点的问题。

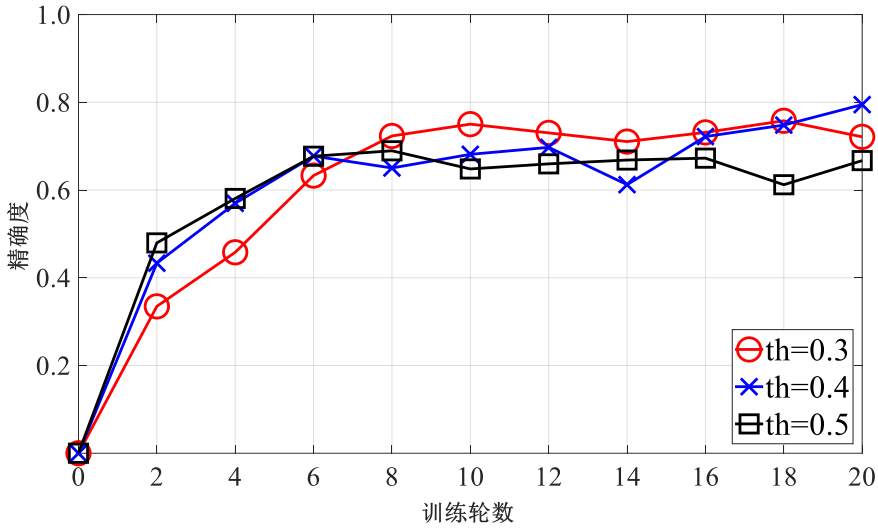


图 12　n=10, m=0.4, 混淆模型更新列表的长度为 3L 时 th 对全局模型精度的影响

图 13 给出了$n = 10$，$th = 0.4$，$m = 0.4$条件下不同混淆模型更新列表的长度对模型精度的影响，其中混淆模型更新列表的长度的取值分别为$2L$、$3L$和$4L$。如图 13 所示，大致在取值为$3L$时模型的精度最高，混淆模型更新列表的长度的取值太小会增加拜占庭节点攻击成功的概率，反之则会提高良性节点被误判的概率，还会增加节点验证模型的时间开销。

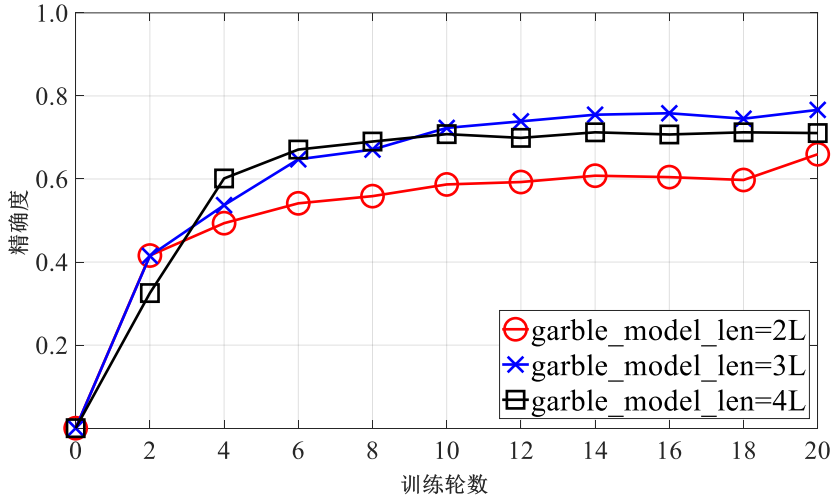


图 13 n=10, th=0.4, m=0.4 时，混淆模型更新列表的长度对全局模型精度的影响

（2）方案可行性分析：

为验证所提方案的可行性，本实验分别使用逻辑回归模型和神经网络模型进行联邦学习模型训练，然后通过观察记录有拜占庭节点和无拜占庭节点这两种联邦学习环境中全局模型精度的变化情况来分析所提方案能否有效检测出拜占庭节点。

图 14 给出了$th = 0.5$，混淆模型更新列表的长度为$3L$条件下逻辑回归模型

的精度随训练轮数增加的变化过程。在图 12 中，$m = 0.2$时全局模型的精度增长趋势同$m = 0$（即不存在拜占庭节点）时全局模型的精度增长趋势大体一致，且两种情况下最终得到的模型精度相差无几，由此证明所提方案可有效检测出拜占庭节点，从而抵消其对全局聚合的负面影响。在同样的检测机制且$m = 0.2$条件下，当$n = 20$时的精度高于$n = 10$时的精度，这是因为前者的联邦学习系统中的良性节点数量多于后者。

图 15 给出了$n = 10$，$th = 0.5$，混淆模型更新列表的长度为$3L$条件下神经网络模型的精度随训练轮数增加的变化过程。在图 13 中，$m = 0.2$时的全局模型精度与$m = 0$时的全局模型精度最终都接近 0.8，证明了所提拜占庭检测方案的有效性。从图 12 和 13 的实验结果可以看出，所提拜占庭检测方案不受节点的本地模型结构的影响。

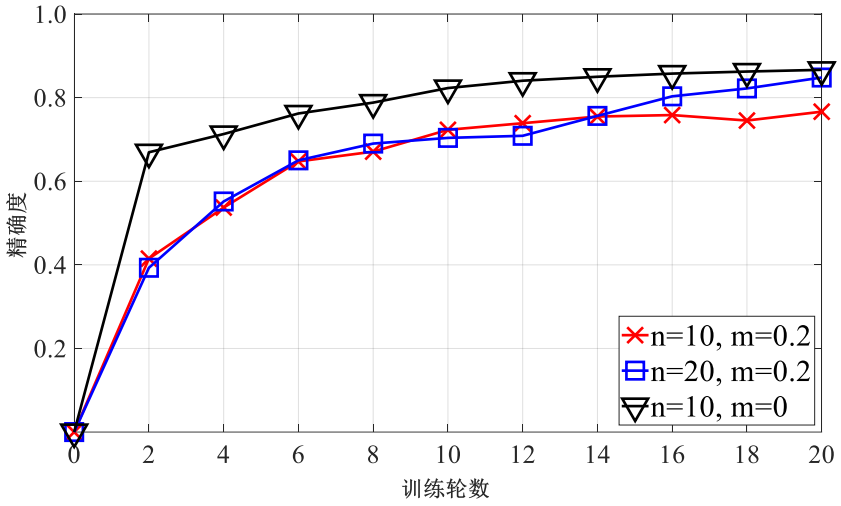


图 14 th=0.5, 混淆模型更新列表的长度为 3L 时逻辑回归模型的全局模型精度

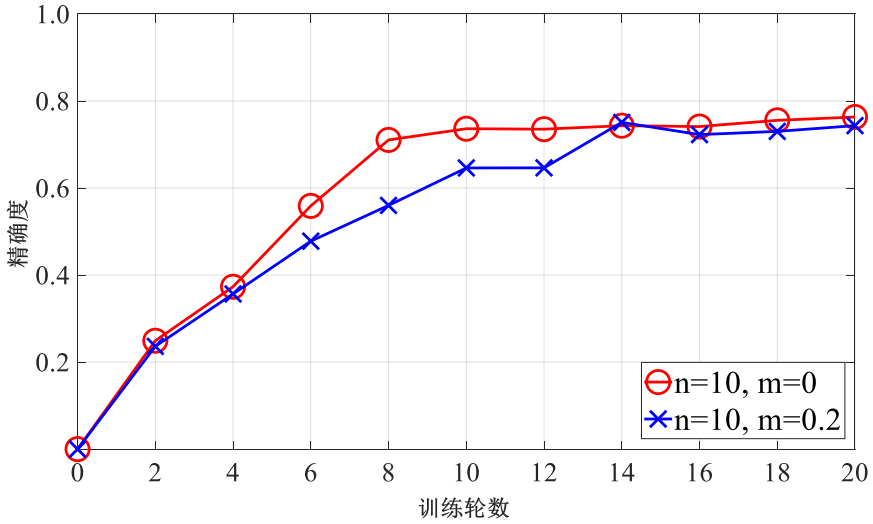


图 15 n=10,th=0.5, 混淆模型更新列表的长度为 3L 时神经网络模型的全局模型精度

（3）方案资源开销分析：

在本节中，首先从理论上分析聚合服务器端和节点端的计算复杂度和通信复杂度，然后以逻辑回归作为训练模型进行联邦学习训练以分析聚合服务器端和节点端的计算复杂度和通信复杂度。

针对计算复杂度，**表 8** 给出了所提方案中聚合服务器和节点每一步执行流程的计算复杂度。从**表 8** 中可知，聚合服务器端和节点端的整体计算复杂度都为$O(n^2)$，其中n是联邦学习系统中的节点个数。

表 8 聚合服务器端和节点端的计算复杂度

阶段		计算复杂度	
		聚合服务器端	节点端
Step1		$O(n)$	$O(n)$
Step2		/	$O(n)$
Step3	模型聚合	$O(n^2)$	/
	混淆模型更新列表生成	$O(1)$	/
	模型验证	/	$O(n^2)$
	拜占庭节点验证	$O(n^2)$	/
Step4		$O(n^2)$	$O(1)$
整体		$O(n^2)$	$O(n^2)$

针对通信复杂度，可从图 10 中分析得出。Step1 期间，聚合服务器端的通信复杂度为$O(n^2)$，节点端的通信复杂度为$O(1)$。Step2 期间，节点需将掩码模型更新集ULM上传聚合服务器，因此聚合服务器端的通信复杂度为 0，节点端的通信复杂度为$O(n)$。Step3 期间，聚合服务器向每个节点发送混淆模型更新列表，因此聚合服务器端的通信复杂度为$O(\text{混淆模型更新列表的长度}) = O(L) = O(n^2)$。另外，节点将对挑战的应答上传聚合服务器，因此节点端的通信复杂度为$O(inter - position) = O(C_n^2 + 1) = O(n^2)$。Step4 期间，聚合服务器端和节点端的通信复杂度均为$O(1)$。整体上看，聚合服务器端和节点端的通信复杂度均为$O(n^2)$。接下来通过具体实验来验证所提方案的资源开销。

图 16 给出了$n = 10$，$th = 0.5$，$m = 0.4$条件下不同长度的混淆模型更新列表对计算开销的影响。从图 16 可以看出，不论是整体的计算开销还是 Step3 阶段的计算开销，均随着混淆模型更新列表的长度的增加而呈线性增长的趋势。这是因为随着混淆模型更新列表中混淆模型数量的增加，节点为生成对挑战的应答需要验证的模型就越多，自然耗时就会增加。

图 17 给出了$th = 0.5$，$m = 0.4$，混淆模型更新列表的长度为$3L$条件下，不同节点数量n对计算开销的影响，其纵坐标为 Step3 阶段的计算开销占整体计算开销的比例。从图 17 可以看出，不论聚合服务器端还是节点端，Step3 阶段的计算开销在整体计算开销中的占比基本保持不变。此外，相比于节点端，聚合服务器端 Step3 阶段的计算开销在整体计算开销中的占比更高。

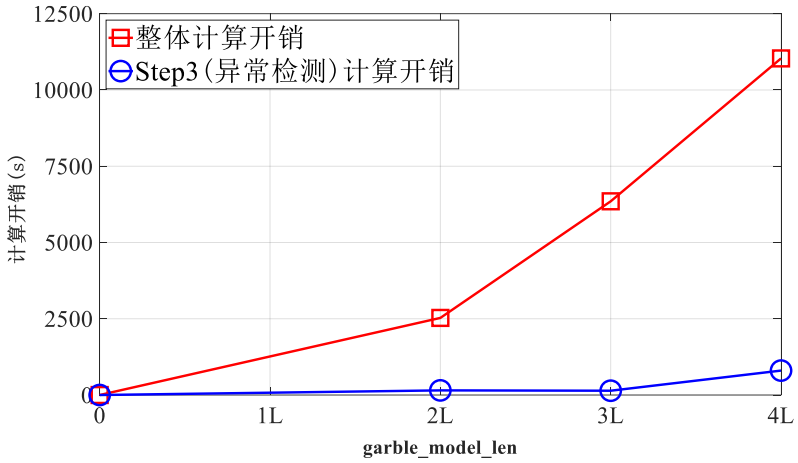


图 16n=10, th=0.5, m=0.4 时，混淆模型更新列表的长度对计算开销的影响

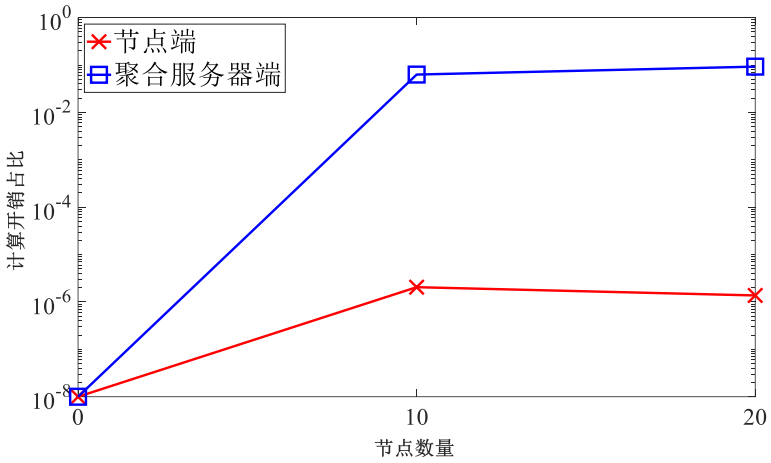


图 17 th=0.4, m=0.4, 混淆模型更新列表的长度为 3L 时 n 对拜占庭检测检的计算开销的影响

图 18 和图 19 分别给出了$th = 0.5$，$m = 0.4$和不同节点数量n和不同混淆模型更新列表的长度条件下节点端和聚合服务器端通信开销。因为在不同的系统环境下，数据占用的内存空间大小可能不同，因此本实验将待传输数据的数据量作为通信开销的衡量指标。从图 18 中可知，Step3（即模型验证）阶段节点端的通信开销随着节点数量增加呈指数增长的趋势。此外，当给定用户数量时，Step3 阶段节点端的通信开销随着混淆模型更新列表的长度呈线性增长的趋势。因此，

所提方案中节点端的整体通信开销主要由节点数量决定，基本与混淆模型更新列表的长度无关。从图 19 中可知，聚合服务器端主要的通信开销来自于处理混淆模型更新列表时期。所提方案的整体和挑战-应答时期的通信开销均随着节点数量增加呈指数增长的趋势。此外，当给定用户数量时，整体和挑战-应答阶段的通信开销随着混淆模型更新列表的长度呈线性增长的趋势。

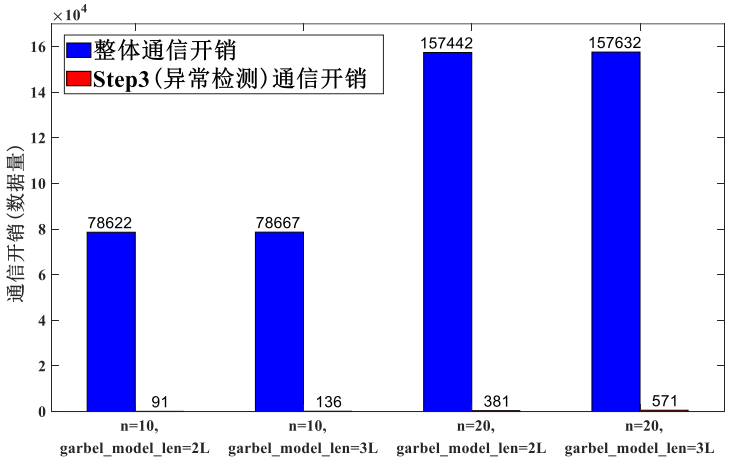


图 18 th=0.5, m=0.4 时节点端的通信开销

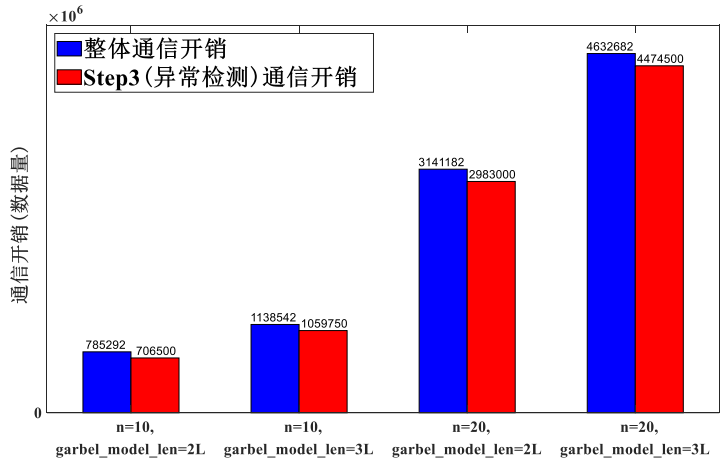


图 19 th=0.5, m=0.4 时服务器端的通信开销

3.3 本章小结

本实验旨在解决不同数据分布情况的联邦学习系统中，隐私保护的正常模型与异常模型的统计差异性及其可检测条件两个关键科学问题。从可靠参与方选择和异常数据检测两方面提升、优化联邦学习系统的可靠性与安全性。

为验证本研究所提出的单一上下文中信任评估方案的性能，设计了相关实验并且进行仿真。通过系统假设构建理想化环境，并建立相应的信任评估模型。方

案针对联邦学习系统中参与方可能出现六种不同的行为模式，采用综合计算方法，将参与方的直接信任度、推荐信任度动态结合，计算出综合信任度，对参与方进行精确评估。实验结果表明，随着交互次数的增加，参与方的信任度逐步提升，同时模拟诚实参与方在交互过程中突发故障的场景，验证了本方案在应对具有多种行为模式的恶意参与方时的优越表现。实验表明该方案为复杂动态交互环境中参与方可靠性与可信度的评估提供了有价值的参考。

本实验中拜占庭检测方案利用拜占庭节点比例m，门限th和混淆模型更新列表的长度garble_model_len这三个因素通过控制变量法开展了三轮实验，每轮实验均成功检测出所有拜占庭节点，证明了该方案能够有效避免全局模型精度的大幅下降。为验证方案的可行性，实验分别采用逻辑回归模型和神经网络模型进行联邦学习训练，并通过比较存在拜占庭节点与无拜占庭节点两种环境下全局模型精度的变化，分析方案的检测效果。实验结果表明，基于挑战-应答机制的方案在 IID 数据场景和隐私保护环境下能够实现高精度的拜占庭节点检测，同时在聚合服务器端几乎未增加额外计算开销，展现了其在隐私保护的优越性能。

第四章 总结和展望

（1）总结：

本研究为解决如何提升协同学习系统的可靠性，使其适用于安全关键场景的难题，相应提出了单一上下文联邦学习参与方信任度评估的方案。针对现有联邦学习参与方信任度评估方法中普遍考虑的信任因素较为单一的问题，本研究设计了一种单一上下文中联邦学习参与方的细粒度信任评估方案。本方案通过系统假设，设计了一个理想环境并建立了相应的模型。并根据所提的单一上下文联邦学习参与方信任度评估的方案采用了联邦学习参与方直接信任度、推荐信任度和综合信任度的综合计算方法。在实验过程中，参与方的信任度随着交互次数的不断增多也随之提升，验证了该方案能在面对具有不同行为模式的恶意参与方时表现出色，为在复杂多变的交互环境中判断参与方的可靠性与可信度提供了参考方案。

本研究为解决如何在协同学习的每一轮迭代过程中发现参与方生成的异常本地模型，提高聚合模型的精确度的问题，提出了 IID 场景下隐私保护的拜占庭节点检测方案。该方案旨在提高异构设备间协同学习系统的参与方可靠性和模型安全性，以构建一个可信的联邦学习系统。为验证该方案的性能，所提拜占庭检测方案利用拜占庭节点比例m，门限th和混淆模型更新列表的长度这三个因素进行实验。该方案通过控制变量法进行了三轮实验，每轮实验结果均成功检测出了所有拜占庭节点，证明该方案能避免全局模型精度的大幅降低。为验证所提方案可行性，本实验分别使用逻辑回归模型和神经网络模型进行联邦学习模型训练，然后通过观察记录有拜占庭节点和无拜占庭节点这两种联邦学习环境中全局模型精度的变化情况来分析所提方案能否有效检测出拜占庭节点。从图 12 和图 13 的实验结果可以看出，所提拜占庭检测方案不受节点的本地模型结构的影响，证明了该方案的可行性。为计算该方案的开销，从理论上分析聚合服务器端和节点端的计算复杂度和通信复杂度。不论聚合服务器端还是节点端，该方案的计算开销在整体计算开销中的占比基本保持不变。

而整体和挑战-应答时期的通信开销均随着节点数量增加呈指数增长的趋势。此外，当给定用户数量时，整体和挑战-应答阶段的通信开销随着混淆模型更新列表的长度呈线性增长的趋势。

种种实验结果表明，基于挑战-应答的方案专注于在 IID 场景以及隐私保护

下进行拜占庭节点检测， 该方案可以在隐私保护环境下提供高准确度的拜占庭节点检测，同时在聚合服务器端几乎未产生额外的计算开销。

（2）展望

联邦学习的核心是保护数据隐私，但现有方法只在参与方的信任度评估和恶意节点的检测两个方面进行改进，因此方案的性能仍存在不足，应从更多角度进行优化升级，以实现在未来进一步优化隐私保护算法、提高计算效率、增强对恶意攻击的抗性。本研究还应将联邦学习与强化学习、迁移学习等技术结合，以进一步提升其在多样化任务中的应用潜力。

此外，本研究将解决人工智能在关键领域应用的瓶颈问题，推动其在金融、政治、军事等领域的广泛应用。同时，本研究将为政府和企业解决数据孤岛和隐私保护问题提供有效方案，加速智能互联时代的到来。本研究还希望将该方案与各个领域进行融合应用，通过技术手段，确保多方在进行机器学习模型训练的同时，能做到数据无需共享、隐私不被泄露、数据使用行为可控，保障更多用户的隐私安全。

图 目录

表 目录

参考资料

[1] Sun W, Lei S, Wang L, et al. Adaptive Federated Learning and Digital Twin for Industrial Internet of Things[J]. IEEE Transactions on Industrial Informatics, 2021, 17(8): 5605-5614.

[2] Kang J, Xiong Z, Niyato D, et al. Reliable Federated Learning for Mobile Networks[J]. IEEE Wireless Communications, 2020, 27(2): 72-80.

[3] Ibrahim A, Ahmed A, Hamid S, et al. Trust and Reputation for Internet of Things: Fundamentals, Taxonomy, and Open Research Challenges[J]. Journal of Network and Computer Applications, 2019, 145(11): 51-62.

[4] 杨明, 胡学先, 张启慧,等. 基于信誉评估机制和区块链的移动网络联邦学习方案[J]. 网络与信息安全学报, 2021, 7(6): 99-112.

[5] Aivaloglou E, Gritzalis S, Skianis C. Trust Establishment in Sensor Networks: Behaviour-based, Certificate-based and A Combinational Approach[J]. International Journal of System of Systems Engineering, 2008, 1(1-2): 128-148.

[6] Kang J, Xiong Z, 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.

[7] Bonawitz K, Ivanov V, Kreuter B, et al. Practical Secure Aggregation for Privacy-Preserving Machine Learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 1175-1191.

[8] Bonawitz, Keith et al. “Practical Secure Aggregation for Privacy-Preserving Machine Learning.” Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017.

[9] Fung, Clement et al. “The Limitations of Federated Learning in Sybil Settings.” International Symposium on Recent Advances in Intrusion Detection, 2020.

[10] McMahan H B, Moore E, Ramage D, et al. Federated Learning of Deep Networks Using Model Averaging[J]. arXiv preprint arXiv:1602.05629, 2016, 2.

[11] Li S, Cheng Y, Liu Y, et al. Abnormal Client Behavior Detection in Federated Learning[J]. arXiv preprint arXiv:1910.09933, 2019.

[12] Zhu, Huafei. “On the relationship between (secure) multi-party computation and (secure) federated learning.” ArXiv abs/2008.02609, 2020.

[13] Jiaqi Zhao,Hui Zhu,Wei Xu,Fengwei Wang,Rongxing Lu,Hui Li: SGBoost: An Efficient and Privacy-Preserving Vertical Federated Tree Boosting Framework.IEEE Trans. Inf. Forensics Secur.18:1022-1036(2023).

[14] Songze Li, Duanyi Yao, Jin Liu:FedVS: Straggler-Resilient and Privacy-Preserving Vertical Federated Learning for Split Models. ICML 2023: 20296-20311.

[15] Runhua Xu, Nathalie Baracaldo, Yi Zhou, Annie Abay, Ali Anwar:Privacy-Preserving Vertical Federated Learning. Federated Learning 2022: 417-438.