公平联邦学习及其设计研究综述

古天龙[1)] 李 龙[1)] 常 亮[2)] 李晶晶[1)]

[1)](暨南大学可信人工智能教育部工程研究中心 广州 510632)

[2)](桂林电子科技大学广西可信软件重点实验室 广西 桂林 541004)

摘 要 联邦学习是由多个客户端协作开展模型训练的一种分布式机器学习解决方案.在联邦学习架构下,公平性被赋予了更加丰富的内涵:一方面,联邦学习中不同参与者对模型训练的贡献可能会有很大不同,能够公平反映每个参与者贡献的奖惩激励机制是联邦学习生态可持续发展的关键;另一方面,无论发送给各个参与方的全局模型是被直接用于结果预测还是用于优化参与方的个性化模型,各个参与方所使用的模型在最终的预测性能或精准度上应该具有公平性.具有某一个或多个方面公平性的联邦学习称为公平联邦学习.通过系统梳理和全面剖析近年来的研究工作,对联邦学习的公平性概念、定义及度量进行了阐释;从公平联邦学习生命周期的不同阶段出发,分别对与公平联邦学习设计相关的公平客户端选择、公平模型优化、公平贡献评估、公平激励机制等进行了综述;从可信人工智能及可信联邦学习的角度,对联邦学习公平性与隐私性、鲁棒性的综合设计进行了讨论;立足于区块链与联邦学习的不同耦合方式,即完全耦合、柔性耦合和松散耦合,对基于区块链的联邦学习框架结构进行了分类阐述,进一步从框架结构、公平性、鲁棒性及隐私保护功能等方面对相关研究工作进行了述评;最后,从公平性定义及度量、公平联邦学习方法、鲁棒公平联邦学习及符合伦理联邦学习等四个方面,给出了公平联邦学习及其设计所面临的主要问题、挑战及研究热点.

关键词 联邦学习;公平性;隐私保护;鲁棒性;区块链;人工智能伦理

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2023.01991

Fair Federated Machine Learning and Its Design: A Comprehensive Survey

GU Tian-Long[1)] LI Long[1)] CHANG Liang[2)] LI Jing-Jing[1)]

[1)](*Engineering Research Center of Trustworthy AI, Ministry of Education, Jinan University, Guangzhou* 510632)

[2)](*Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi* 541004)

Abstract Federated learning is a distributed machine learning solution in which multiple clients train models under the coordination of a central server. Fairness is endowed with richer connotation under the federal learning framework. The fairness of federal learning has roughly two meanings: cooperative fairness and model fairness. Federated learning with one or more aspects of fairness is called fair federated learning or fairness-aware federated learning. Firstly, this paper systematically reviews and comprehensively analyzes the research work in recent years, and the concept, definition and metrics of fairness in federated learning are explained. Federated learning requires many different clients to cooperatively undertake model training. Fairness is not only related with sensitive attributes or protected groups, but also affected by different clients and their interactions. On the view of the clients, the global model for each client may have different accuracy, the model fairness should guarantee that the clients with similar local data have the compar-

收稿日期:2022-06-20;在线发布日期:2023-01-11.本课题得到国家自然科学基金(No.U22A2099, 62172350)资助.**古天龙**,博士,教授,中国计算机学会(CCF)杰出会员(07007D),主要研究领域为形式化方法、可信人工智能、人工智能伦理、数据治理等. E-mail: gutianlong@jnu.edu.cn. **李 龙**,博士,讲师,中国计算机学会(CCF)会员(50487M),主要研究领域为人工智能安全、公平机器学习、逻辑程序设计等.**常 亮**,博士,教授,主要研究领域为知识图谱、知识表示、形式化方法等.**李晶晶**,博士,讲师,主要研究领域为可解释机器学习、基于区块链的联邦学习等.

ative prediction results (i. e. the individual fairness), or different client groups should have considerable model prediction accuracy (i. e. the group fairness); Secondly, the fair federal learning design methods are thoroughly surveyed. Fairness enhancement methodology in machine learning, or fair machine learning design, includes pre-processing, in-processing, and post-processing. Fairness enhancement for federated learning, or fair federated learning design, can also be categorized as these three approaches. Most of the existing fair federated learning designs focus on in-processing approaches. From the life cycle stages of fair federation learning, they can be roughly divided into client selection, model optimization, contribution evaluation, incentive mechanism and so on; Meanwhile, in light of trustworthy artificial intelligence, the integrated design of fairness, privacy and robustness of federated learning is discussed, and the fair federated learning framework based on block chain is illustrated in detail. Fairness and privacy are complementary ethical concepts. Many application scenarios require both privacy protection and fairness. Data sensitivity is the key factor of both fairness and privacy, and federated learning may be used in privacy-protected data scenarios, where the privacy of sensitive data and fairness of sensitive attribute groups need to be guaranteed. The fairness of federated learning has become a new target of adversary attack. Malicious adversary can influence the consistency of model performance distribution by data poisoning. The backdoor attack and cooperative fairness attack of fair federated learning are also emerging. The robustness of fair federation learning requires defense mechanisms to deal with these fairness attacks, and fairness in anomaly detection should also be considered. Fair blockchain federated learning is the combination of blockchain and fair federated learning, whose main combination methods include blockchain smart contract or consensus mechanism for fair client selection and fair incentive mechanism design, and blockchain's distributed ledger to store information related to fairness design and decision making; Finally, the main problems, challenges and research topics in the field of fair federated learning are proposed from the definition and measurement of fairness in federated learning, methods of fair federated learning, robust and fair federated learning, and ethically aligned federated learning for the healthy and sustainable development of artificial intelligence ecology.

Keywords federated machine learning; fairness; privacy protection; robustness; blockchain; artificial intelligence ethics

1 引 言

人工智能(Artificial Intelligence,AI)在带来巨大机遇的同时,也蕴含着一系列风险和挑战,如:算法安全导致应用风险、黑箱模型导致算法不透明、数据歧视导致智能决策偏见、数据滥用导致隐私泄露风险等.这些问题直接影响着社会和公众对人工智能的信任,影响着人工智能技术的应用及其系统的部署.面对人工智能引发的信任焦虑,发展可信人工智能(Trustworthy Artificial Intelligence)成为了全球共识[1].从学术研究角度,可信人工智能研究包含了安全性、可解释、公平性、隐私保护等多方面内容.机器学习是一种实现人工智能的重要方法,可信机器学习是建立可信人工智能系统的核心技术[2].

公平性(Fairness)概念的提出和探讨始于 20 世纪 60 年代[3].能够确保每个人都有平等的机会获得一些利益的行为,称为公平的行为,或者称这样的行为具有公平性.不能够确保每个人平等地获得一些利益,损害弱势群体的利益的行为,称为不公平的行为,或者称这样的行为具有不公平性.歧视和偏见是与不公平相关联的概念,不公平的行为又称为具有偏见的行为或者歧视的行为.如果机器学习的预测或决策能够确保每个人或群体都有平等的机会获得一些利益,则称该机器学习具有公平性,并称之为公平机器学习[4-5].

联邦学习(Federated Learning,FL)又称联邦机器学习(Federated Machine Learning,FML),是

由多个客户端(用户)在中央服务器(聚合器)的协调下协作开展模型训练的一种分布式机器学习解决方案[6-7].联邦学习概念最早由 McMahan 等领衔的 Google 研究院团队提出[8],用以解决面向个体用户的键盘输入优化问题.联邦学习面向的是分散或分布式的多用户(客户端)场景,每个客户端拥有用户自己的数据集.传统的机器学习需要将这些客户端的数据收集在一起,通过数据集完成模型的集中式训练.联邦学习则无需汇总收集各个客户端的数据,而是由参与的用户分别在本地训练各自的本地模型,同时将本地模型参数上传至服务器,服务器则聚合形成全局模型参数(根据不同的联邦学习架构,可以配置由独立的服务器,也可以由某个客户端来承担服务器的模型参数聚合任务),并发布给各个参与方共享使用.联邦学习较好地解决了数据孤岛和数据隐私问题,引起了学术研究和工业应用人员的极大关注[6-7,9-10],并被大致划分为跨设备(cross-device)FL 和跨孤岛(cross-silo)FL 两类,其中跨设备 FL 是指参加学习的客户端是大量的移动设备或物联网设备,而跨孤岛 FL 则指参加学习的客户端是组织、机构或数据孤岛.近年来,伴随着对可信人工智能关注度的不断提升,可信联邦学习(Trustworthy Federated Learning)的概念被提出.可信联邦学习是在传统联邦学习分布式联合训练的基础上,加入安全可信机制保证数据隐私及模型安全,同时满足模型性能可使用、效率可控、决策可监督、模型可监管以及普惠等要求,是一种增强型的联邦学习.由其内涵可以发现,可信联邦学习更多聚焦安全性、隐私保护及实用性方面的内容[11-12].

鉴于在数据安全和隐私保护方面所具备的明显优势,FL 在医疗、金融、保险等领域均具有较好的应用前景[6-7].通过高效利用多来源医疗数据,比如电子病历、诊断记录、治疗方案,FL 可以协助医疗组织在不侵犯患者隐私的情况下,提升在病情相似性学习、疾病早期诊断、康养方案建议、诊疗效果预测等方面的能力,达到提升疾病诊断及治疗方案的科学性与合理性的效果.通过融合销售记录、纳税情况、产能数据、征信报告等多个数据源,FL 可实现更为精准的企业风险评估,类似地也可协助提升个人信贷风险评估、金融欺诈监测的效果.通过融合保险购买记录、历史出险数据等多个数据源,FL 可实现更为合理的保险定价及智能理赔等.通过对智能终端上的行为数据进行学习,FL 可建立用户行为模型,提升下一个单词预测、人脸检测、语音识别等方面的效果,而不会泄漏个人数据.例如,谷歌将联邦学习应用于手机键盘(谷歌 Gboard 系统)的输入预测,大大提升了手机输入法预测的准确度.随着《通用数据保护条例》(General Data Protection Regulation,GDPR)、《健康保险携带和责任法案》(Health Insurance Portability and Accountability Act,HIPPA)、《个人信息保护法》等法律法规的严格实施,以人们对数据安全及隐私保护重视程度的提升,FL 必将被越来越多地用于解决各行各业的实际问题.

机器学习的广泛应用给人们生活带来了诸多深刻改变,其中的公平性也因此逐渐受到了广泛关注,产业界和学术界从机器学习公平性的定义及度量指标、公平数据集构造、公平机器学习算法设计等多个维度展开了探讨,取得了一定的研究进展[4-5].作为一种新型机器学习架构,联邦学习不同于传统的集中式学习(针对传统机器学习公平性的主要研究主题如表 1 所示)[4-5].因此,联邦学习如何处理传统机器学习中的公平性?联邦学习是否会引发新的公平性?这些是联邦学习必须面对的问题[6].公平性在联邦学习架构下,被赋予了更加丰富的内涵.联邦学习的公平性具有协作公平性和模型公平性两个方面的含义[13-14].联邦学习是由多方用户参与的协作学习,不同参与者对学习过程的贡献可能会有很大不同,参与者的贡献会受到数据规模、数据质量、通信开销等多个因素影响.例如,银行、政府和科技巨头等具有较大数据规模和更广数据类型范围,这些参与者对学习模型的效果影响更大.此外,一些参与者可能会“搭便车”,只是利用模型,并不对模型学习提供任何贡献;有些甚至是攻击者,可能对学习过程有恶意的负面影响.因此,一个能够公平反映每个参与者贡献的奖惩机制对于联邦学习是非常必要的,奖惩激励系统是联邦学习生态健康、可持续发展的关键.现有的联邦学习系统通常允许每个参与者访问基于所有参与者的数据训练的同一个全局模型,缺乏协作公平性.为此,必须构建衡量每个参与者贡献的度量指标,并设计相应的激励方案,使贡献更多的参与者获得更高的回报,以吸引更多的用户参与模型训练.

联邦学习的全局模型会发送到各个参与方的客户端进行预测或决策.一方面,全局模型对于各个客户端在预测性能或精准度上应该具有公平性[11-12].首先,具有相似数据和相同全局模型的客户端,应该具有大致相当的模型预测结果,即,联邦学习客户端

的个体公平性;其次,多个客户端站点的群体之间应当具有相当的模型预测精度,即,联邦学习客户端的群体公平性.另一方面,全局模型对于具有敏感属性或受保护群体的预测,应当满足如下公平性[4-6]:个体公平性,相似属性的数据具有相似的模型预测结果;群体公平性,受保护群体和非受保护群体具有相同的正类预测率,或者,模型预测与敏感属性无关;反事实公平性,假设个体具有某些其他属性或者属于其他群体,在考虑了所有因果相关的途径后该个体仍能获得一致的结果,即,预测的结果不受属性变化的影响.具有某一个或多个方面公平性的联邦学习称为公平联邦学习(Fair Federated Learning,FFL),或者公平感知联邦学习(Fairness-aware FL,FAFL)[6].

在 FL 的核心功能实现方面,Kairouz、Yang、Li 等从基本概念、数据分布、机器学习模型、优化技术、信息理论、通信结构、通信效率、隐私机制、资源管理、应用领域等角度出发,论述了 FL 的研究进展[6-7,9-10,15-17].在 FL 激励机制设计方面,Zhan、Zeng、Stackelberg 等从数据价值评估、最优资源分配、基于合约理论/博弈论/Shapley 值/强化学习/区块链/拍卖理论的 FL 激励机制设计等方面出发,论述了已有研究和值得开展的工作[18-22].Liu 等进一步从 FL 模型质量改进[23]、区块链 FL[24]、个性化 FL[25]、模型融合和学习范式[26]等角度全面分析和论述了相关研究工作.Lyu、Yin、Mothukuri 等从 FL 安全性和隐私保护的角度,对与 FL 相关的投毒攻击和推理攻击技术[27]、安全威胁与隐私泄露风险[28-29]、隐私攻击技术及其防御方法[28-30]等相关的研究工作进行了综述.关于 FL 在边缘计算及物联网中的应用方面,Lim、Khan、Nguyen 等分别从资源开销[31]、性能增强[32]、业务实现[33]、硬件需求[34]等角度对现有研究工作进行了综述讨论.针对联邦学习在医疗领域的应用,Nguyen、Shyu、Antunes 等分别从医疗功能实现(如电子病历管理、远程健康监控、医学图像分析)[35]及医疗数据处理及应用(如医学影像数据、重症监护数据的数据分割、数据分布、数据隐私保护)[36-38]的角度综述了 FL 在智慧医疗中的应用研究.Liu 等对分类、推荐、语音识别、医学文本挖掘等自然语言处理任务的 FL 算法进行了综述,并介绍了相关的评测及工具[39].Yu 等结合教育、医护等应用中的数据挖掘分析了 FL 的优势,讨论了 FL 数据挖掘面临的挑战和问题[40].Gadekallu 等讨论了大数据获取、存储、分析和隐私保护等业务应用中的 FL 技术[41].Agrawal 等从入侵检测系统的部署结构、异构异常检测、分布式拒绝服务攻击(Distributed Denial of Service,DDoS)攻击检测等方面综述了 FL 在入侵检测的应用研究[42].以上工作所关注的主要内容如表 2 所示.

表 1 针对传统机器学习公平性的主要研究主题

研究维度	研究内容或技术	补充说明
公平性概念及度量	基于预测结果的公平性统计与度量,基于预测和真实结果的公平性统计与度量,基于预测概率和真实结果的公平性统计与度量等	FL 架构的特殊性,使得其公平性涵义更为丰富,进一步导致需要更加多样化的公平性分析及设计技术.
机器学习公平性分析	关联规则挖掘方法,k 最邻近分类方法,概率因果网络方法等	
公平机器学习的设计	预处理,中间处理,后处理	
公平性与隐私性	多方安全计算,差分隐私,同态加密,安全聚合协议等	当前应用于传统机器学习的隐私保护技术基本适用于 FL.但 FL 对隐私保护的要求更高、公平性含义也更加多样.
公平性与鲁棒性	面向不同计算环境(如分布漂移、数据不足等)的鲁棒性,面向不同攻击意图(如数据投毒、模型窃取等)的鲁棒性	FL 的实现依赖于分布式客户端的联合学习,架构的复杂性使其更容易收到来个多方面的攻击或干扰,如恶意客户端投毒攻击、多客户端串谋攻击等.
区块链 FL 的公平性	/	区块链的分布式特征使其能够较好地协助 FL 训练,并能够利用其激励机制、智能合约等功能协助提升 FL 的公平性.

注:"/"表示文献中缺少相关内容。

尽管如此,目前尚缺乏联邦学习的公平性及其设计的全面性综述讨论.联邦学习架构下公平性的概念和度量如何有别于传统公平性以及公平机器学习?公平联邦学习的训练数据、学习算法、隐私保护、激励机制的特点及其设计方法和技术有哪些?公平联邦学习的应用场景以及应用中面临的问题和挑战有哪些?这些都是需要系统梳理和亟待解决的问题.为此,本文从联邦学习的公平性定义和度量、公平联邦学习设计方法、联邦学习的公平性和隐私性、联邦学习的公平性和鲁棒性、基于区块链的公平联邦学习等方面,对公平联邦学习及其设计进行了分析和综述,同时讨论了公平联邦学习面临的挑战和进一步研究方向.本文主要内容及相互间关系如图 1 所示.

表 2　联邦学习相关的综述论文

综述内容分类	主要文献	涉及公平性文献
综合性 FL 技术	Kairouz et al., 2021[6];Yang et al., 2019[7];Li et al., 2020[9];AbdulRahman et al., 2020[10];Zhang et al., 2021[15];Li et al., 2021[16];Lo et al., 2021[17]	Kairouz et al., 2021[6];Li et al., 2021[16]
FL 的激励机制	Zhan et al., 2021[18];Zhan et al., 2022[19];Zeng et al., 2021[20];Ali et al., 2021[21];Tu et al., 2021[22]	Zeng et al., 2021[20]
FL 的其他技术	Liu et al., 2020[23];Wang et al., 2021[24];Tan et al., 2021[25];Ji et al., 2021[26]	Ji et al., 2021[26]
FL 隐私与安全	Lyu et al., 2020[27];Yin et al., 2021[28];Mothukuri et al., 2021[29];Liu Yi-Xuan et al., 2022[30]	
边缘计算 FL	Lim et al., 2020[31];Khan et al., 2021[32];Nguyen et al., 2021[33];Abreha et al., 2022[34]	
医疗健康 FL	Nguyen et al., 2022[35];Shyu et al., 2021[36];Antunes et al., 2022[37];Pfitzner et al., 2021[38]	
FL 其他应用	Liu et al., 2021[39];Yu et al., 2022[40];Gadekallu et al., 2021[41];Agrawal et al., 2021 [42]	

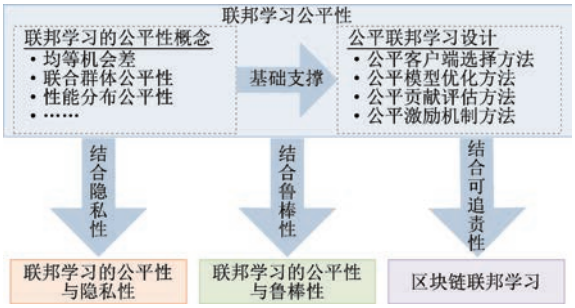

图 1　本文整体架构

本文的组织结构如下:第 2 部分对联邦学习的公平性概念、定义及度量进行了阐释;第 3 部分从客户端选择、模型优化、贡献评价、激励机制等几个联邦学习环节角度,对公平联邦学习的设计方法进行了综述;第 4 部分讨论了联邦学习的公平性和隐私性之间的联系,以及联邦学习的公平性和隐私性综合设计;第 5 部分剖析了联邦学习的公平性和鲁棒性,介绍了鲁棒公平联邦学习设计方法;第 6 部分介绍了区块链架构的公平联邦学习结构模式及设计方法;第 7 部分给出了公平联邦学习领域面临的问题、挑战及研究热点.

2　联邦学习的公平性概念

公平性是一个交叉学科概念,许多学科领域都对其进行了研究[4,43].在传统机器学习领域,已经提出了多种公平性的概念定义及度量[4-5,44].从公平性的观测对象角度,这些概念大致可分为群体公平性和个体公平性两类[4-5].群体公平性强调的是具有敏感属性或受保护群体应该受到与其他群体相当的待遇,要求机器学习模型的性能表现不应该受到敏感属性影响,受保护群体的模型表现应当和其他群体相当;个体公平性要求相似的个体应该受到相似的待遇,对于机器学习而言,特征或性质相似的个体的模型预测结果应该保持一致.从公平性的度量方式角度,这些概念大致可分为统计度量、相似度量和因果推理等三类[4-5].总体上,传统机器学习是基于数据训练模型的决策和推理,受这一固有特性影响,已有机器学习的公平性概念基本上都是基于反映个体或群体特征的数据的敏感属性而定义和度量.

有别于传统机器学习,联邦学习需要协调不同的客户端协同完成模型训练,公平性不仅仅体现在敏感属性或受保护群体,而且还要考虑不同客户端及其相互作用.从敏感属性或受保护群体角度,具有某些敏感属性(受保护群体)的数据可能分散在不同的客户端,公平性的评测需要对模型的预测精度进行衡量,联邦学习的服务器并没有客户端的原始数据,因此难以获取公平性评测的敏感属性信息.从联邦学习的客户端角度,全局模型对于各个客户端可能具有不同的精确度,模型的公平性应该使得拥有相似数据的客户端具有相当的模型预测结果(客户端的个体公平性),或者,多个客户端的群体之间应当具有相当的模型预测精度(客户端的群体公平性)[6,45].

均等机会差(Equal Opportunity Difference, EOD)[46]:对于具有敏感属性 A 和非敏感属性 Y 的数据样本,如果模型(二分类任务)的预测 $\hat{Y}$ 满足 $\Pr(\hat{Y}=1 \mid A=1, Y=1)=\Pr(\hat{Y}=1 \mid A=0, Y=1)$,则称模型满足均等机会[4-5].均等机会可以采用均等机会差进行度量:

$$\mathrm{EOD}=\Pr(\hat{Y}=1 \mid A=0, Y=1)-\Pr(\hat{Y}=1 \mid A=1, Y=1).$$

在联邦学习下,均等机会差有全局公平性和局部公平性两方面含义:全局公平性需要考虑所有客户端的全部数据,局部公平性则考虑单一客户端的局部数据.全局公平性的 EOD 度量为

$$F_{\mathrm{global}}=\Pr(\hat{Y}=1 \mid A=0, Y=1)-$$

$$\Pr(\hat{Y}=1 \mid A=1, Y=1).$$

客户端 k 的局部公平性的 EOD 度量为

$$F_k=\Pr(\hat{Y}=1 \mid A=g, Y=1, C=k)-\Pr(\hat{Y}=1 \mid A=g, Y=1, C=k),$$

其中，$C=k$ 表示第 k 个客户端.

联合群体公平性(Unified Group Fairness)[47]：对于由变量 A' 划分为 M 个群体 $\{D_i^g \mid i=1, 2, \cdots, M\}$ 的数据集 $D=\{D_1^g, D_2^g, \cdots, D_M^g\}$(设数据 D_i^g 对应的分布为 P_i^g)，群体 D_i^g 的联合群体公平性定义为

$$Disparity=\sqrt{\frac{1}{M}\sum_{i=1}^{M}(Acc(D_i^g)-Avg\_Acc)^2},$$

其中，$Acc(D_i^g)$ 是群体 D_i^g 的预测精度，$Avg\_Acc=\frac{1}{M}\sum_{i=1}^{M}Acc(D_i^g)$. 变量 A' 的不同含义对应不同的公平性问题：如果 A' 表示客户端编号，则对应客户端级公平性；如果 A' 表示敏感属性或受保护属性，则对应属性级公平性；如果 A' 表示未知分布的潜在客户端编号，则对应不可知分布公平性.

最小最大帕累托公平性(Minmax Pareto Fairness)[48]：对于含有 n 个独立样本 $\{x_i, y_i, a_i\}\sim P(X, Y, A)$ 的数据集 $D=\{(x_i, y_i, a_i)\}_{i=1}^n$(其中，$x_i\in X$ 为数据特征，$y_i\in Y$ 为预测变量，$a_i\in A$ 为敏感属性或群体成员身份)以及群体期望损失函数 $\boldsymbol{r}(h)$，如果假设 h^* 具有最小最差群体风险，即

$$h^*=\arg\min\max_{h\in P_{A,H},a\in A} r_a(h)=\arg\min_{h\in P_{A,H}}\|\boldsymbol{r}(h)\|_\infty,$$

则称假设 h^* 是最小最大帕累托公平的，或者具有最小最大帕累托公平性，其中 $P_{A,H}$ 是帕累托前沿.

罗尔斯最大最小公平性(Rawlsian Max-Min Fairness)[49]：对于假设 $h\in H$ 及其在群体 $a\in A$ 上的期望效用 $U_{D_A}(h)$，如果假设 h^* 最大化最差群体效用，即

$$h^*=\arg\max\min_{h\in H,a\in A}U_{D_A}(h),$$

则称假设 h^* 是罗尔斯最大最小公平的，或者具有罗尔斯最大最小公平性.

最小最大群体公平性(Minmax Group Fairness)[50]：如果模型 h^* 能够在不损害其他群体的性能下实现最差群体的性能优化，则称模型 h^* 是最小最大群体公平的，或者具有最小最大群体公平性.

有界群体损失公平性(Bounded Group Loss Fairness)[51]：如果假设 h 对于所有群体(敏感属性) $a\in A$ 满足 $U_{(x,y,a)\sim P}[l(h(x), y)\mid A=a]\leqslant\delta$，则称假设 h 是 δ 级有界群体损失的，或者具有 δ 级有界群体损失公平性，其中，$l(h(x), y)$ 表示样本 x 的预测 $h(x)$ 与其真实值 y 之间的损失.

性能分布公平性(Fairness of Performance Distribution)：对于模型 w 和 $\widetilde{w}$，如果 m 个客户端的性能 $\{a_1, a_2, \cdots, a_m\}$ 在模型 w 上较之于模型 $\widetilde{w}$ 上更一致或均匀，则称模型 w 较模型 $\widetilde{w}$ 更加公平[46]，其中，性能 a_k 通常为模型 w 在设备端 k 的测试数据上的精确度，性能的一致性可以采用性能分布的均值和/或方差来度量.

性能分布公平性是客户端级公平性，传统基于敏感属性的公平性是敏感属性级公平性(如，统计公平、精度平等). 在某些特殊情形下，客户端可能会自然形成某些敏感属性的群体，如果单一客户端对应敏感属性群体，二者定义吻合. 在一般情形下，客户端可能含有多个敏感属性群体，或者某敏感属性群体分散在多个客户端.

善意公平性(Good-Intent Fairness)：善意公平性是指将受保护客户端(类)的最大损失降至最低，以避免牺牲其他客户端性能而过拟合任何特定模型[47].

善意公平性是针对联邦学习的各个客户端可能存在不同数据分布而提出的概念，客户端数据的异构性决定了全局模型的泛化需要应对不可知的数据分布. 善意公平性能够更加客观和精确地评测全局模型的损失.

从联邦学习的客户端协作角度，客户端的梯度更新用于全局模型聚合的频次需要满足一定指标下的公平性，客户端的数据质量、数据规模、训练效率等相关的贡献应该得到相称的奖励回报，这些属于联邦学习的协作公平性[13,45].

选择公平性(Selection Fairness)：选择公平性是指通过增加代表性缺失客户端的参与机会来减轻联邦模型中的客户端偏见[45,48].

联邦学习中客户端选择的常用方式是设置阈值(如，通信延迟)，这可能导致某些计算或通信资源有限的客户端没有机会参与模型训练，引发此类客户端数据代表性缺失的模型偏见. 确保选择公平性的方式之一是设置采样约束，如，客户端的参与训练频次不低于某下限值(长期公平性)[49].

贡献公平性(Contribution Fairness)：贡献公平性是一种分配公平性[45,50]，是指客户端的回报与它对联邦学习模型的贡献成正比，而与联邦学习模型

的预测精度无关.客户端的贡献可以依据客户端的数据规模、数据质量、参与训练的频次等进行衡量,客户端得到的回报可能是货币、性能相称的模型等.

遗憾分布公平性(Regret Distribution Fairness):遗憾分布公平性是指客户端等待收到激励支付而产生的遗憾的一致性[51].遗憾是数据所有者迄今为止收到的与他应该收到的回报之间的差额,同时考虑等待收到全额回报的时间.

预期公平性(Expectation Fairness):预期公平性旨在最大限度地减少客户端在奖励发放过程中不同时间点上的不公平[51].遗憾分布公平性和预期公平性适用于联邦学习模型的未来收益的激励,并且随着收益的产生,参与联邦学习的客户将逐渐得到补偿.

模型公平性和协作公平性是联邦学习的公平性的不同视角,模型公平性旨在改善各个参与方所使用模型的性能一致性,协作公平性的目的在于吸引更多的客户端参与联邦学习.从模型性能改善角度,二者都有助于一定程度上改善全局模型的性能.但是通过分析以上公平性定义的内涵发现,不同的公平性定义侧重于不同的公平性视角(参见表 3),大多难以同时从多个视角对公平性进行度量.因此,联邦学习的敏感属性公平性、模型公平性和协作公平性的综合公平性的定义和度量是值得探讨的问题[52].最小最大群体公平性实现了无需访问数据敏感属性对敏感属性公平性的度量,为敏感属性公平联邦学习的设计探索了可行路径,需要研究在此框架下协作公平性和模型公平性的综合度量[53].在缺乏敏感属性数据信息的情况下,度量和抑制敏感属性不公平性是公平联邦学习研究面对的挑战.随着联邦学习应用的推广,越来越多的敏感属性不可知的用户会不断加入用户群体,独立于敏感属性以确保公平模型性能的联邦学习方法,是未来联邦学习需要解决的问题.

3　公平联邦学习设计

传统机器学习的公平性增强,或者公平机器学习设计,包括预处理、中间处理和后处理等三类方法[4-5,44].基于此,联邦学习的公平性增强,或者公平联邦学习设计,也可以采取这三种方式.已有的公平联邦学习设计大都集中于中间处理方法.从公平联邦学习的生命周期角度,大致可分划为:客户端选择[54-57]、模型优化[58-59]、贡献评价[60]、激励机制[61]等方法[45].

表 3　公平性定义侧重衡量的公平性类型

公平性定义	公平性视角
均等机会差	属性级公平性
联合群体公平性	属性级公平性 客户端级公平性(模型公平性)
最小最大帕累托公平性	属性级公平性
最小最大群体公平性	属性级公平性
有界群体损失公平性	属性级公平性
性能分布公平性	客户端级公平性(模型公平性)
善意公平性	客户端级公平性(协作公平性)
选择公平性	客户端级公平性(模型公平性)
贡献公平性	客户端级公平性(协作公平性)
遗憾分布公平性	客户端级公平性(协作公平性)
预期公平性	客户端级公平性(协作公平性)

3.1　公平客户端选择方法

联邦学习的客户端不公平的一个重要诱因是客户端选择方法.许多联邦学习方法聚焦于全局模型,即重视 FL 服务器的利益(如,提高收敛速度[54]或改善模型精度[55]),而忽视了 FL 客户端的利益.这些工作通常使用基于阈值的方法来选择 FL 客户端,不满足阈值(如,传输速度、带宽、本地精度等)的客户端将被过滤掉、不允许其参加相应轮次的全局模型参数聚合.例如,FedCS(Federated Learning with Client Selection)以受限资源下尽可能多的客户端参与聚合模型训练为优化目标,依据客户端的计算资源和通信带宽来选择客户端[56],难免造成计算或通信性能较差的客户端始终无法参与模型参数训练.

FL 客户端选择的不公平可分为三类:代表过度、代表不足和代表缺失.例如,移动边缘 FL 系统通常对网络传输速率敏感.在这种情况下,传输速率差的数据客户端不太可能被选中参与 FL 训练(即,代表不足),而传输率较高的客户端更有可能被选中(即,代表过度).同时,可能永远不会选择信道条件一直较差的数据客户端(即,代表缺失).

值得注意的是,公平的客户端选择并不意味着等概率地选择所有数据客户端.客户端之间的异构性也是需要考虑的因素.为了实现客户端选择的公平性,公平联邦学习方法需要在 FL 服务器和 FL 客户端的利益之间取得平衡.下面将介绍现有的公平联邦学习的客户端选择方法.这些工作可分为两类:客户端的参与频次控制和客户端模型架构定制.

3.1.1　客户端的参与频次控制

在跨设备联邦学习中,可能存在大量的可以参与训练的客户端,有限的通信带宽限制了服务器和

所有客户端之间的模型参数分发和梯度更新上传，需要对参与训练的客户端进行选择. Huang 等将 FL 客户端选择描述为长期公平约束下的最小平均模型交换时间问题，基于 Lyapunov 优化框架，将原离线问题转化为在线优化问题，通过排队动力学方法优化 FL 客户的参与频次，其中的长期公平约束确保每个客户的平均参与频次不低于客户选择的预期频次阈值[56]. 实验结果表明，公平 FL 客户选择策略可以在牺牲训练效率的前提下提高模型的准确性.

Yang 等将 FL 客户端选择问题描述为一个组合多臂 Bandit 问题[62]，其中，每个臂代表一个客户端，超级臂代表所有 FL 客户端的集合. 超级臂奖励是参与的单臂奖励的非线性组合. 选择频次较高的客户端被视为更受信任并获得更高的回报，但通过使用 N 轮中每个客户端至少有 1 次被选择的策略，为参与频次较低的客户端提供了参与 FL 训练的机会.

客户端参与频次控制可以促进 FL 客户端选择的公平性，但是许多不公平性的重要诱因，如，客户端数据质量、数据规模、训练质量等并未考虑在内，需要建立多因素综合的客户端参与频次控制的设计方法.

3.1.2 客户端模型架构定制

联邦学习公平客户端选择的另一类方法是模型架构定制. 客户端的系统异构性和数据异构性是 FL 应用部署中的两个关键挑战. FL 通常给所有客户端分发相同的初始训练模型，因此，计算能力较低的客户就需要更长的时间来完成训练. 在这种情况下，FL 服务器可能会将这些客户端视为掉队者，并在后续的客户端选择中将其剔除. 基于客户端能力动态调整 FL 模型框架，使得代表性不足或代表缺失的客户端能够参与 FL 训练，是公平性增强的一类有效方法.

(1)定制客户端模型

为了避免联邦学习中客户端的掉队或丢弃，Caldas 等将丢弃技术(Dropout)用于裁剪 FL 模型[63]，以获得适配于客户端计算资源和通信能力的局部模型，客户端对接收到的服务器发送的裁剪模型进行训练，并上传裁剪模型训练所得模型梯度更新至服务器. 对于经过设计的客户端裁剪模型，客户端具有足够的计算和通信能力对其进行训练，从而克服了基于阈值的客户端选择所引发的客户端缺失的不公平性. 有损压缩技术是一种适用于卷积神经网络的简单方法，可以将其应用于客户端梯度更新的有损压缩，以进一步降低通信成本. 联邦模型裁剪则通常采取如下方式：对于完全连接的层，丢弃一定数目的激活节点；对于卷积层，丢弃固定比率的过滤器.

模型裁剪中激活节点的随机丢弃，忽视了丢弃操作所导致的模型内部结构的变化. 为此，Bouacida 等提出了自适应联邦丢弃(Adaptive Federated Dropout，AFD)[64](参见图 2)，通过维护一个激活评分图(每一个激活的训练过程重要性和影响力的对应数值)，来选择激活节点，裁剪生成最适合每个客户端的子模型，从而加快收敛速度，减少精度损失.

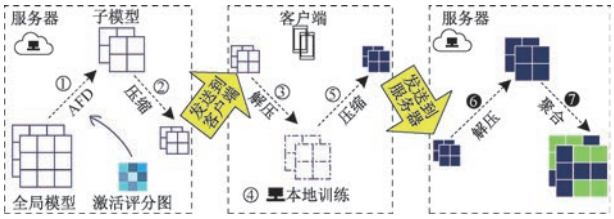

图 2 自适应联邦丢弃

(2)定制客户端工作负荷

避免联邦学习中客户端的掉队或被丢弃的另一种方式是：为计算能力较低的客户端分配较少的工作，从而使他们能够通过客户端选择阈值. Li 等提出了允许各个客户端根据其可用计算资源条件进行部分训练的 FedProx 方法[65]，该方法有助于减轻系统异构的不良影响，但对于异构数据的情况，客户端更新可能会导致模型发散. 对此，为了改善训练过程，引入了减轻客户端模型更新影响的策略：其一，通过限制更接近全局模型的局部更新，鼓励更高质量的局部更新；其二，安全地聚合不同计算工作量产生的局部模型更新.

(3)定制客户端网络连接

客户端选择也可能受到客户端通信能力(例如，比特率、数据包丢失)的影响. 当客户端在 FL 期间向服务器发送其本地更新时，一些数据包可能会丢失. 在检测到丢失的数据包时，服务器向客户端发送重新传输请求，以尝试恢复丢失的数据包. 然而，对于网络速度较慢的客户端，重传可能会导致 FL 模型训练的额外延迟. 因此，通信信道较差的客户的模型更新就不太可能参与最终模型的聚合，从而导致模型偏见.

为了解决上述问题，Zhou 等提出了一个容错 FL 框架 ThrowRightAway(TRA)[55]. 其主要思想

在于,数据包丢失可能并不总是有害的.基于这个假设,TRA 通过忽略一些丢失的数据包来加速 FL 训练.开始时,所有参与的 FL 客户端都会向 FL 服务器指示其网络状况.根据客户端的报告,服务器将客户端分为能力充足和能力不足两种类型.然后,服务器随机选择要向其发送全局模型的客户端,并等待其模型更新.当检测到数据包丢失时,服务器仅对能力充足的客户端发送重新传输请求.否则,它只会将丢失的数据包记录为零.客户端完成上传后,TRA 根据丢包记录重新计算聚合.这种方法的有效性很大程度上取决于准确地将客户分端划分为不同的类型.TRA 隐含地假设 FL 客户端能够准确地评估自己的网络状况,并诚实地向 FL 服务器报告这些信息.这种假设具有较难的实际可操作性.

3.2　公平模型优化方法

除了在 FL 模型训练之前进行的客户端选择外,FL 模型训练期间的优化过程也可能引发 FL 全局模型出现偏见.这些偏见可能会导致对某些受保护群体的歧视,或以牺牲其他客户端的利益为代价来过度满足某些客户端,由此,全局模型可能会在客户端之间表现出不公平.例如,在图像识别场景中,FL 服务器已经访问了许多由青年用户群体使用的移动设备,但这些设备只有少数其他年龄组人员使用,该模型可能在青年年龄组的设备上表现良好,但在其他年龄组的设备上表现不佳.FL 的公平模型优化方法可以大致分为两种:基于目标函数的方法和基于梯度的方法.

3.2.1　基于目标函数的方法

在 FL 模型训练中,处理公平性问题的常见方法是更改目标函数的损失函数或增加正则项,以满足公平性约束.现有方法主要关注保持相当的全局模型精确度下减少客户端模型精确度的方差,以实现 FL 客户端的预测精度的分布一致性,即精度平等或性能分布公平性[46].

不可知联邦学习(Agnostic Federated Learning,AFL)是 FL 公平性的较早研究工作之一[54],AFL 通过防止模型过度拟合任何特定客户端而牺牲其他客户端的利益,实现了善意公平的理念.联邦学习的各个客户端数据可能呈现非独立同分布,AFL 针对不同分布的客户端形成的任何(不可知)目标分布进行了优化,只要不增加性能最差客户端的损失,就不会对其他客户端的模型性能产生负面影响.然而,AFL 仅适合于客户端数量较少的场景,对于大量客户端的联邦学习,模型的泛化难以得到有效保证.

为了克服 AFL 的泛化能力不足,受无线通信网络中公平资源分配方法的启发,Li 等提出了 q-FFL(q-Fair Federated Learning)方法[46],q-FFL 通过使用参数 q 来重新加权总损耗,为损耗较高的客户端分配更高的权重,反之亦然.与 AFL 相比,q-FFL 更为灵活,因为可以通过调节 q 来调整公平性程度.在 q 设置为大数值的情形下,q-FFL 的性能与 AFL 相当.Li 等还进一步提出了一种基于 q-FFL 的通信效率更高的 FL 聚合方法 q-FedAvg.q-FFL 比 AFL 具有更低的精度方差(即更高的公平性)和更快的收敛速度,在 FL 客户端上实现了更均匀的精度分布,即更强的精度平等.然而,q-FFL 和 AFL 都不能抵抗敌手攻击.例如,如果客户恶意增加损失,可能会导致 q-FFL 和 AFL 的性能恶化.

AFL 和 q-FFL 通常假设数据分布是静态的,不能应对动态变化的数据分布问题.客户端的数据集随时间推移出现的漂移可能会带来数据分布的变化.对于测试数据分布不同于训练数据分布或者测试数据分布未知的情形,Du 等提出了 AgnosticFair(Fairness-aware Agnostic Federated Learning Framework)方法[66],该方法通过重新加权函数,为损失函数和公平约束中的每个数据样本分配一个单独的重新加权值,以实现统计公平.在数据漂移或者测试数据分布未知的情况下,可以实现良好的准确性和公平性.然而,重新加权值的确定需要分布的先验知识和客户端的敏感信息,此举会泄露客户端的数据分布信息.

传统机器学习的群体公平性概念[4-5],如,不平等影响(Disparate Impact)、统计平等(Demographic Parity)、均等机率(Equal Odds)、均等机会(Equal Opportunities),是基于模型的预测精度来度量的.然而,模型的预测精度是一个非光滑不可微函数,将这些公平性度量表述为联邦学习模型优化算法的约束条件极其困难,甚至不可行.为此,研究人员提出了一系列将基于预测精度的公平性度量转换为光滑代理约束的解决方案[66-70].但是,训练模型的公平性性能受到了代理约束和公平性度量之间的近似误差的限制[68].

Cui 等将每一个客户端的预测精度损失添加到目标函数,构建了约束多目标优化框架以及公平一致性联邦学习方法 FCFL(Fair and Consistent Federated Learning)[71],通过优化包含所有客户端损失项的代理函数来最大化具有公平性约束的客户端最

差性能.

Chu 等通过引入广义均等机会差(Difference of Generalized Equal Opportunities, DGEO)约束条件,提出了公平联邦学习方法 Fedfair[72],将 DGEO 计算转换为局部联邦估计问题,避免了客户端的信息泄露. Zhang 等基于联合公平性概念定义新的损失函数,实现了客户端性能公平性、敏感属性公平性和未知分布客户端公平性的一体化设计[48].

Papadaki 等研究了客户端数据中可能含有部分敏感属性的群体公平性联邦学习问题,在这一情形下,客户端公平并不一定意味着群体公平,为此引入了最小最大群体公平性[51,73],旨在最小化最大群体精度损失,并给出了群体公平性联邦学习的最小最大目标函数及其求解算法 Federated Minimax (FedMinMax)[74]. Yue 等通过在目标函数中添加正则项来惩罚精确度的损失,在每一通信轮次对性能差的个体或群体进行动态权重调整,提出了群体和个体公平联邦学习方法 GIFAIR-FL(Group and Individual Fairness to Federated Learning)[75].

3.2.2 基于梯度的方法

基于梯度的方法是模型优化实现 FL 公平性的另一种方式. Wang 等指出差别较大的更新梯度冲突也是导致 FL 不公平的一个诱因,为了改善梯度差异很大的客户端的性能,全局模型可能会牺牲一些客户端的模型准确性(这类情况通常会导致 FL 训练期间的不公平). 为此,提出了公平联邦学习 FedFV(Federated Fair Averaging)方法[76],在对客户端梯度更新求平均值之前,通过余弦相似度来检测冲突的梯度更新,采用梯度投影来减少选定客户端之间的内部冲突和选定客户端与未选定客户端之间的外部冲突,对梯度更新的方向和幅值进行修改以消除梯度冲突. FedFV 在公平性、精确度和效率之间取得了平衡. 然而,用于缓解外部冲突的梯度估计方法可能并不始终可靠,因为基于前几轮的估计梯度可能会过时,并且与最新梯度更新不兼容. 将梯度的估计直接应用于最新的全局更新可能会导致模型偏见. 为了解决这个问题,需要将可靠的梯度估计方法与 FL 客户选择方法结合使用.

Huang 等设计了将客户端预测精度和参与训练频次用于服务器梯度聚合的加权策略[77],以增强客户端性能的公平性. Kanaparthy 等给出了基于客户端梯度更新的公平性和预测精度评测结果进行模型梯度聚合的启发式方式[78]:FairBest、FairAvg、FairAccRatio 和 FairAccDiff. 不足之处在于,服务器需要预先配置标准测试数据集.

梯度稀疏(Gradient Sparsification, GS)是提升跨设备联邦学习中梯度传输效率的有效方式. Han 等提出了联邦学习中服务器与客户端之间梯度传输的公平双向 top-k 梯度稀疏方法[79],通过自适应调整参数 k 确保不同客户端具有相当数量的梯度更新.

受客户端数据隐私保护的限制,具有敏感属性公平性的联邦学习设计方法非常有限. Ezzeldin 等通过 EOD 来度量全局模型和客户端的敏感属性公平性,并将全局模型和客户端模型的 EOD 差 $|F_k - F_{global}|$ 用于计算梯度聚合的权重,提出了公平联邦学习方法 Fairfed(Fairness-aware Federated Learning)[47].

3.3 公平贡献评估方法

贡献评估是无需访问客户端数据对每个客户端贡献的重要性所进行的评估. 常用方法是评估每个客户端对 FL 聚合模型性能的影响. 反过来,客户端的贡献可用于 FL 的客户端选择和奖励分配. 因此,公平评估 FL 客户端的贡献至关重要. 现有的 FL 贡献评估方法大致可分为[45]:(1)自我报告信息;(2)个体评估;(3)效用博弈;(4)Shapley 值和(5)经验方法.

3.3.1 自我报告信息

自我报告信息评估方法基于客户端的自我报告信息进行客户端的贡献评估. 自我报告信息涉及客户端(本地)数据集的质量、数量和收集成本,以及客户端向 FL 承诺的计算和通信能力等.

Zhang 等提出了分级公平联邦学习方法 HFFL (Hierarchically Fair Federated Learning)[80],要求客户端向 FL 服务器报告其本地数据集的可公开验证信息(例如,数据质量、数据量、数据收集成本等). 然后,服务器使用报告的信息为客户端的贡献进行评级,具有相同级别贡献的客户端会从服务器收到相同的模型参数,贡献较多的客户端会从服务器收到性能更好的模型. Kang 等采用了类似的方法,利用自我报告信息(客户端数据质量、数据可靠性等)根据合约理论设计了一个 FL 激励方案[81]. FL 服务器设计合约并将其发布给数据所有者(客户端),每个合约都包含客户端的奖励及数据相关信息. 每个客户端选择适配其贡献类型的最理想合约,作为参与 FL 的承诺,没有按照合约规定完成训练任务的客户端将被列入黑名单并不予奖励.

Sarikaya 和 Ercetin 提出了基于 Stackelberg 博弈的 FL 激励方案[82],在这种博弈场景下,服务器扮

演领导者身份,客户端承担随从者角色,服务器向客户端购买训练服务,客户端作为卖方提供服务,Stackelberg 均衡解是完成单次随机梯度下降的平均时间,客户端的最佳策略是向 FL 服务器(任务发布者)如实报告一个单位 CPU 的期望价格. Le 等将自我报告信息用于基于拍卖的 FL 激励机制设计[83],服务器是拍卖者,客户端是买家,客户端报告其竞价(包括所需资源量、客户端精度和能源成本等信息),FL 服务器使用投标信息来衡量每个客户的潜在贡献以确定获胜(参与训练)客户端以及获胜客户端的奖励.

值得注意的是,以上方法假设客户端能够可靠地评估自己的情况并如实报告信息. 在实践中,这种假设可能会制约应用的部署.

3.3.2　个体评估

个体评估方法根据客户端在特定任务中的表现来衡量其贡献. 声誉机制被广泛用于追踪 FL 参与客户端的历史贡献,可用于客户端选择和奖励分配方案设计. Lyu 等将声誉机制引入了客户端-服务器模式的 FL[60],FL 服务器管理客户端的声誉列表,每个客户端的声誉值由服务器根据客户端模型的预测精度来确定并不断更新,客户端只能将与其声誉相当的部分更新梯度上传至服务器. 为了实施该方案,服务器需要一个准确且平衡的测试数据集.

Zhang 等提出了分散式 FL 系统的声誉机制[84],每个任务发布者(承担参数聚合任务的客户端)根据参与者(客户端)在每一轮的局部模型梯度更新来衡量参与者的声誉,声誉记录被存储在区块链以公开共享,因此任何一方都不能篡改声誉评分. 任务发布者需要在训练期间保存所有本地模型和全局模型,随着客户端数量和/或训练迭代次数的增加,可能会导致较高的存储开销.

除了声誉机制,还有其他 FL 客户贡献的个体评估方式. Zeng 等人提出的 FL 客户端选择的拍卖方案中[85],基于每个客户端的出价和资源质量(数据、计算能力、带宽、CPU)进行评分,通过多维竞拍来择优选择多个客户端参与 FL 训练.

个体评估方法通常采用如下两种假设:(1)FL 服务器和 FL 客户端都是可信的;(2)客户端模型与其他参与者的模型(或全局模型)相似程度越高,则该参与者的贡献更大. 这两个假设在实际应用中可能并不总是成立. 对于假设(1),FL 服务器和 FL 客户端可能自私且行为不端;对于假设(2),在非独立同分布下,参与者通常持有具有异构分布的数据集,在这种情况下,来自具有不同数据分布的参与者的不同模型更新可以为提高 FL 模型性能提供更有价值的知识补充. 如果处理不当,这些因素可能会对贡献评估的公平性产生负面影响.

3.3.3　效用博弈

基于效用博弈的 FL 贡献评估方法与利润分享计划密切相关,即建立参与者效用与其相应奖励间的映射规则. 有三种广泛采用的利润分享方案:(1)平等主义:团队产生的任何效用在团队成员之间平均分配;(2)边际收益:参与者的收益等于参与者加入团队导致团队获得的增加效用;(3)边际损失:参与者的回报等于该参与者离开团队带来的效用损失.

对于边际收益和边际损失,每个客户端收到的回报金额取决于其加入的顺序,因为回报计划通常旨在激励客户端尽早加入. FL 中最常用的方案是边际损失方案. Wang 等采用边际损失来衡量 FL 中各方的贡献[86],基于删除某参与方来衡量对模型的贡献的思想,使用近似算法实现客户端影响的度量. Nishio 等给出了通过单个 FL 训练过程的边际损失来评估每个客户端贡献的方法[87],该方法有效减少了客户端贡献评估的通信和计算开销.

简单边际损失适用于公平评估 FL 模型的给定客户端群体中指定客户端的贡献. 这是一种相对评估(即取决于其他参与客户的贡献),并不能反映客户端本地数据的实际价值.

3.3.4　Shapley 值

基于 Shapley 值(Shapley Value,SV)的 FL 贡献评估方法引起了广泛关注. SV 是一种基于边际贡献的方法,为了解决合作博弈问题,该概念于 1953 年提出[88]. 某一客户端的 SV 反映了该客户端对 FL 模型的贡献,其数值可以通过不包含该客户端的其他所有客户端群体上的边际贡献之和的平均来计算,该数值仅取决于客户端拥有的局部数据,而与它加入联邦的顺序无关. SV 的计算复杂度为 $O(2^n)$,因此为了提高 SV 计算的效率,在传统的机器学习中提出了许多启发式方法,例如截断蒙特卡罗 Shapley 和梯度 Shapley[89]. 已有基于 SV 的 FL 客户端贡献评估方法包括:基于客户的方法和基于特征的方法.

(1)基于客户端的方法

Song 等提出了两种基于梯度的 SV 方法[90]:①单轮重建(One-round,OR)和②多轮重建(Multi-rounds,MR). 这两种方法都从 FL 客户端收集梯度

更新来重建 FL 模型，而无需对不同的客户端子集进行重新训练. OR 收集所有训练回合中的所有梯度更新. 然后，在最后一轮中重建所有子集的模型. OR 在最后一轮中使用重建的模型只计算一次 SV. 相比之下，MR 在每一轮训练中计算一组 SV，然后将其聚合，以计算基于 SV 的最终贡献值. Wei 等对 MR 进行了扩展，提出了截断多轮（Truncated Multi-Rounds，TMR）方法[91]. TMR 对 MR 进行了两个方面改进：首先，它为具有更高精度的训练回合分配更高的权重；其次，它消除了不必要的模型重建，提高了效率.

通过利用这些基于梯度的 SV 估计方法，可以显著提高评估 FL 客户端贡献的效率. 然而，仍然需要评估每轮训练中不同客户端子集的子模型. 为了进一步降低计算成本，Wang 等提出了两种有效的近似方法[92]：①基于置换采样的近似方法和②基于群体测试的近似方法. 这些方法旨在提高每轮 SV 计算的效率. Fan 等通过对所有客户端不同子集的可能贡献的矩阵的补全，提出了补全联邦 Shapley 值，提升了 Shapley 值的公平性[93].

（2）基于特征的方法

纵向联邦学习参与者的数据集在特征空间中几乎没有重叠，但在样本空间中有显著重叠[6]，这对客户端贡献评估提出了新的挑战. Wang 等将 Shapley 值用于计算纵向联邦学习的特征重要性[86]. 由于直接使用 SV 评估每个预测可以揭示潜在的敏感特征，该方法对特征组而不是每个单独的特征进行 SV 计算. 然而，这种方法的计算成本仍然很高，因为计算成本随着训练数据的大小呈指数增长. Fan 等提出了纵向联邦 Shapley 值（Vertical Federated Shapley Value，VerFedSV）的贡献评估概念，该概念不仅满足公平性的期望性质，而且计算效率得到提升[94].

Shapley 值已成为 FL 贡献评估中得到认可的方法. 然而，提高效率所带来的高额计算成本和不精确估计限制了此类方法的可扩展性.

3.3.5 经验方法

作为基于理论的 FL 客户端贡献评估方法的替代方案，人们开展了客户端贡献评估的经验方法. Shyn 等人提出了经验评估方法 FedCCEA（Federated Client Contribution Evaluation through Accuracy Approximation）[95]，通过构造具有采样数据规模的精度近似模型（Accuracy Approximation Model，AAM）来学习每个客户端的数据质量. 该方法通过采样数据规模有效地近似客户端的贡献，并通过设置用于 FL 模型训练的本地数据的所需规模，允许客户端部分参与. FedCCEA 由模拟器和评估器组成（参见图 3），模拟器通过运行一轮的 FL 分类任务，获得 AAM 的输入和目标. 输入即采样数据规模，假设 n 个客户端每次分别完成 R 轮 FL 训练，则第 r 轮采样数据规模为 $[X_r^{(1)}, X_r^{(2)}, \cdots, X_r^{(n)}]$. 目标则为 Round-wise 精度. 然后，评估器使用存储的输入（由 R 轮训练汇总得来）和目标优化来自 AMM 的权重向量 ω. 在模型收敛后，提取第一层共享权重，学习数据规模对每个客户的重要性. 然而，由于 AAM 是一个非常简单的神经网络结构，FedCCEA 目前仅限于简单的 FL 任务，因此不太适合实际应用.

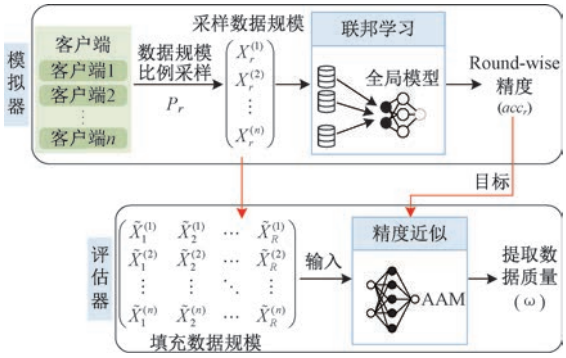

图 3 FedCCEA 框架

3.4 公平激励机制方法

激励机制的设计旨在鼓励客户端（数据所有者）参与联邦模型训练，公平的激励机制是将奖励“公平”地分配给 FL 客户端（例如，基于他们的贡献）的方式，对于奖励分配的成功至关重要. 近年来，人们提出了许多针对 FL 的激励机制[19,96]，但并不是所有工作都关注于提高公平性. 例如，在基于 Stackelberg 博弈的 FL 场景中[82,97]，服务器和客户端竞争以优化自己的效用. 博弈的均衡解实现了双方的权益平衡，但可能不公平. 这里将重点讨论考虑公平的 FL 激励机制设计，这些激励机制大致可以分为两类：(1)货币激励机制；(2)非货币激励机制.

3.4.1 货币激励机制

货币激励机制以货币收益或回报来奖励 FL 客户端. Zeng 等将多维拍卖用于激励高质量数据的客户端参与 FL 训练[85]，通过使用博弈论来推导客户端之间分配利润的最优策略（考虑了数据所有者之间的竞争），并在贡献评估时考虑了公平性，该方案被命名为 FMore. 在投标阶段，服务器通过客户端的评分规则发布其要价. 在收到要价后，客户端将根

据其可用资源决定是否出价.在收集到足够的出价后,服务器根据客户端的估计贡献值选择客户端.由于采用了相同的评分规则,客户端可以确定他们是否得到了公平对待.FMore 的一个缺陷是没有考虑客户选择的公平性.为此,他们提出了 FMore 的扩展版本 ψ-FMore,以应对客户选择的公平性.ψ-FMore 为每个客户分配一个概率 ψ,用于增加低分客户获胜概率,降低高分客户获胜概率.然而,这项工作假设服务器和客户端都真实地出价,如果存在行为不端的 FL 参与者,训练性能会恶化.此外,评分计算需要增强隐私保护,以免泄露敏感信息.

Cong 等提出了基于 VCG(Vickrey-Clarke-Groves)机制的公平 VCG(Fair-VCG,FVCG)[98].FVCG 鼓励客户端贡献高质量数据和如实报告训练成本.服务器通过为所有数据所有者设置相同的数据质量单价,将奖励分配给所有数据所有者.然而,FVCG 在实践中的可行性难以保证,因为无法保证客户端自我报告信息的真实性.

为了克服这一不足,Zhang 等利用声誉间接反映客户端的数据质量和可靠性[84],而不要求他们进行自我报告.针对分散式 FL 系统,提出了基于声誉和反向拍卖的 RRAFL(Federated Learning Based on Reputation and Reverse Auction)机制.FL 任务发布者(服务器)负责追踪客户的历史行为,以获得直接声誉评分.此外,可以共享来自多个 FL 任务发布者的声誉记录,以获得间接声誉评分.任务发布者根据单位声誉出价对 FL 参与者(客户端)进行排序,并选择单位声誉出价最低的前 k 个参与者加入 FL.然后,任务发布者使用第 $k+1$ 个参与者的单位声誉出价来确定每个选定参与者的报酬,因为任务发布者提供的最终单价将高于所有选定参与者的单价,因此,可以有效地激励所选择的参与者加入 FL.然而,基于声誉的方法依赖于客户端历史性能记录的可用性.

契约理论也被用来设计 FL 激励机制.在移动网络中,任务发布者和移动设备之间存在信息不对称,即任务发布者不知道每个移动设备的数据质量、数据规模和可用计算资源.如果要监控这些信息,可能会产生高昂的成本.为了减少信息不对称的影响,Kang 等应用契约理论设计了一种有效的激励机制[81],以吸引具有高质量数据的移动设备加入 FL.它将本地数据质量参数定义为契约模型的类型.任务发布者使用上一轮的观察结果,为具有不同数据质量的数据所有者设计不同的合约.因此,具有更高质量数据和更多计算资源贡献的客户端可以从任务发布者那里获得更高额的回报.类似地,Ye 等使用契约理论为车辆边缘计算环境中的 FL 设计了一种激励机制[99],它引入了一个二维合同,根据客户端的数据质量和计算能力来确定适当的奖励.这些方法的一个缺点是,它们针对的是只有一个 FL 任务发布者的垄断场景.作为垄断者,任务发布者只为每个客户端提供有限的合同选择,这对移动设备的利润产生了负面影响.在实践中,可能会有许多 FL 任务发布者竞相吸引 FL 客户端.目前,现有的 FL 激励机制尚未考虑此类情况.

上述 FL 激励机制隐含地假设激励预算已经预先确定.在某些情况下,FL 模型训练时的激励预算是不可用的,因此参与者希望获得 FL 模型训练完成甚至被使用后的收入.为了处理这种情况,Diana 等定义了遗憾分布公平性和预期公平性[51],这两个概念对 FL 的长期可持续运营也很重要,遗憾分布公平要求根据客户等待获得激励奖励的时间来公平对待客户.由于 FL 模型的训练和商业化需要时间,服务器可能没有足够的收入在早期阶段补偿参与者.这导致客户端的贡献与迄今为止获得的回报之间暂时不匹配.为了克服这个问题,他们提出了一个动态回报分享方案,该方案通过最大化集体效用和最小化数据所有者的遗憾和等待时间之间的不平等,将回报动态地分配给客户端.它确保那些贡献了更多高质量数据并等待了更长时间的客户,将在接下来的训练中获得更多收入.在客户奖励的逐步支付过程中,预期公平的概念用来确保客户的遗憾值尽可能公平地减少,以相称于他们的贡献.

3.4.2　非货币激励机制

另一类 FL 激励机制不是基于货币奖励,而是根据客户端的贡献为他们分配不同性能的 FL 模型以激励客户端.这类方法适用于以下应用场景:(1)不能使用货币激励预算;(2)利用 FL 模型产生的未来收入来奖励客户不可行,或者(3)FL 客户端之间存在竞争,如果高贡献客户端收到与低贡献客户端相同的 FL 模型,则会导致高贡献客户端感到受到不公平对待.

Zhang 等人提出的分级公平联邦学习框架 HFFL 通过向拥有更高质量数据的客户端提供更高质量的模型更新,确保客户端之间激励的公平性[80].HFFL 根据客户端的特点(如数据质量、数据规模)将其分为不同的级别,然后训练多个 FL 模型(每个级别一个).为了训练较低级别的模型,来自较高级

别的客户端只提供与来自较低级别的客户端相同的数据规模.当训练更高级别的 FL 模型时,来自较低级别的客户端需要贡献他们所有的本地数据.因此,更高级别的客户端会收到性能更好的 FL 模型.然而,HFFL 有如下不足:首先,同一级别的客户端可能没有相同数量的数据,因为在对客户端进行分类时会考虑其他因素;其次,用于对客户端进行分级的一些客户端特征是通过自我报告获得的,这使得这种方法容易受到虚假报告的影响.

Sim 等给出了基于客户端数据 Shapley 值评估的模型奖励机制[100],通过对聚合模型参数注入与 Shapley 值相对应强度的高斯噪声来调节客户端的模型质量,以实现奖励机制的公平性.Lyu 等提出了一个分布式公平隐私保护深度学习框架 FPPDL (Fair and Privacy-Preserving Deep Learning)[101],每个客户端根据自己的贡献获得性能不同的 FL 模型.在 FPPDL 中,本地声誉体系通过任意两个客户端之间的相互评估来维持,每个客户端根据其本地声誉和承诺水平获得一定数量的交易积分,客户端可以使用他们的交易积分从其他客户端那里下载梯度.每个客户端都可以在没有协作的情况下获得改进的局部模型,并且每个客户端获得的模型改进与其相应的贡献成正比.

对于公平联邦学习设计,客户端数据及其数据生成过程的偏差是联邦学习公平性增强不容忽视的问题,零数据增强(Zero-Shot Data Augmentation)[102]、客户端数据重新加权和全局数据差分隐私重新加权[103]等预处理方法是联邦学习公平性增强的有益尝试.将事后纠偏处理(后处理方法)用于从可能存在偏差的训练数据中习得的联邦学习模型也是有价值的未来工作.综合实现模型公平性、协作公平性和属性公平性的联邦学习设计解决方案是亟待解决的问题.以上公平联邦学习研究工作间的对比见表 4.

4 联邦学习的公平性与隐私性

隐私和非歧视(公平)是相关但不同的概念[104],二者都有助于遵守普适社会价值观来公平对待他人.然而,隐私通常被视为保护信息不被泄露,非歧视则被认为是禁止基于已知受保护属性信息(如,性别、民族)的偏见行为.隐私和非歧视之间可能存在正向或负向的影响.例如,在求职申请中不披露性别可能有助于防止歧视.此外,非歧视可能会降低隐私关注.例如,禁止医疗保险公司歧视已有疾病患者降低了重病患者对隐私的关注程度.公平性和隐私性是互补的伦理概念:许多应用场景既需要隐私保护也需要满足公平性.数据的敏感性是公平性和隐私性所共同关注的焦点[6],联邦学习可能用于大量的隐私保护数据场景,在这里敏感数据的隐私性和敏感属性群体的公平性都需要得到保证.

公平性与隐私性存在某种程度上的不一致:差分隐私通常追求个体身份特征的模糊化,而公平性通常需要了解敏感群体或个体的身份特征.例如,为了确保贷款发放算法决策不对女性产生歧视,若不了解个体或群体的性别属性,公平机器学习模型的设计是相当困难的.Du 等人提出的 AgnosticFair 方法[66]也印证了这一点.AgnosticFair 方法能够在数据漂移或者测试数据分布未知的情况下实现良好的准确性和公平性,但重新加权值的确定需要分布的先验知识和客户端的敏感信息,从而引发客户端隐私泄露.

Ekstrand 等讨论了综合考虑公平性和隐私性的意义,指出差分隐私为某些场景下实现隐私保护和公平性增强提供了一个数学框架[105].Cummings 等探讨了机器学习同时实现差分隐私和统计公平性(假阳性率平等、假阴性率平等)的可能性,研究发现:维持模型预测精度不可能同时满足差分隐私和精确公平性,只能满足隐私和近似公平性[106].Pujol 等开展了差分隐私机制对公平性影响的实验验证研究,结果表明,差分隐私会对数据总体都会带来误差,但是 ϵ-差分隐私对不同群体的数据具有不同的误差,带来了明显的不公平性[107].Bagdasaryan 等分析了差分隐私随机梯度下降对模型预测精度的影响,发现差分隐私对于代表性不足群体的预测精度下降更为明显,降低了模型的公平性[108].

考虑到模型的隐私性和公平性通常以其准确性为代价,Zhang 等对梯度裁剪和噪声添加对深度神经网络的影响进行了一系列理论和实证分析.同时证明,由于 DP-SGD(Differential Privacy Stochastic Gradient Descent)使训练稳定性变差,深度学习中训练周期的设定对准确性、公平性和隐私性之间的平衡变得至关重要,并进一步提出了两种用于确定停止模型训练最佳时期早期停止标准.停止标准一是在训练收敛后,当准确性相对增加而公平性相对降低时停止训练;停止标准二更多依据歧视水平的变化情况确定是否停止训练,当歧视水平在某一阈值内连续下降时断定此时获得最优公平性,并将满

表 4　公平联邦学习设计的相关研究

研究内容分类		主要文献	公平性视角	度量方式
公平客户端选择	客户端参与频次控制	Huang et al. 2020[56]	客户端级公平性	选择公平性
		Yang et al. 2021[62]	客户端级公平性	贡献公平性
	客户端模型架构定制	Zhou et al. 2021[55]	客户端级公平性	最大最小公平性
		Li et al. 2020[65]	客户端级公平性	贡献公平性
公平模型优化	基于目标函数	Li et al. 2019[46]	客户端级公平性	贡献公平性、性能分布公平性
		Zhang et al. 2021[48]	属性级公平性、客户端级公平性	联合群体公平性
		Du et al. 2021[66]	属性级公平性	均等机会差
		Cui et al. 2021[71]	属性级公平性	均等机会差、有界群体损失公平性
		Chu et al. 2021[72]	属性级公平性	均等机会差
		Papadaki et al. 2022[74]	属性级公平性	最小最大群体公平性
		Yue et al. 2021[75]	属性级公平性	统计公平性
	基于梯度	Ezzeldin et al. 2021[47]	属性级公平性、客户端级公平性	均等机会差、统计公平性
		Wang et al. 2021[76]	属性级公平性、客户端级公平性	统计公平性
		Huang et al. 2022[77]	客户端级公平性	贡献公平性
		Kanaparthy et al. 2021[78]	属性级公平性、客户端级公平性	均等机会差、贡献公平性
		Han et al. 2020[79]	客户端级公平性	选择公平性
公平贡献评估	自我报告信息	Zhang et al. 2020[80]	客户端级公平性	贡献公平性
		Sarikaya et al. 2019[79]	客户端级公平性	贡献公平性
		Le et al. 2021[83]	客户端级公平性	贡献公平性
	个体评估	Lyu et al. 2020[60]	客户端级公平性	贡献公平性
		Zhang et al. 2021[84]	客户端级公平性	贡献公平性
		Zeng et al. 2020[85]	客户端级公平性	贡献公平性
	效用博弈	Wang et al. 2019[86]	客户端级公平性	贡献公平性
		Nishio et al. 2020[87]	客户端级公平性	贡献公平性
	Sharply 值	Wang et al. 2019[86]	客户端级公平性	贡献公平性
		Song et al. 2019[90]	客户端级公平性	贡献公平性
		Wei et al. 2020[91]	客户端级公平性	贡献公平性
		Wang et al. 2020[92]	客户端级公平性	贡献公平性
		Fan et al. 2021[93]	客户端级公平性	贡献公平性
		Fan et al. 2022[94]	客户端级公平性	贡献公平性
	经验方法	Shyn et al. 2021[95]	客户端级公平性	贡献公平性
	货币激励机制	Diana et al. 2021[51]	属性级公平性	统计公平性
		Kang et al. 2019[81]	客户端级公平性	贡献公平性
		Zhang et al. 2021[84]	客户端级公平性	贡献公平性
		Zeng et al. 2020[85]	客户端级公平性	贡献公平性
		Cong et al. 2020[98]	客户端级公平性	贡献公平性
		Ye et al. 2020[99]	客户端级公平性	贡献公平性
	非货币激励机制	Zhang et al. 2020[80]	客户端级公平性	贡献公平性
		Sim et al. 2020[100]	客户端级公平性	贡献公平性
		Lyu et al. 2020[101]	客户端级公平性	贡献公平性

足停止标准二的多个训练周期中的首个训练周期作为停止周期[109].

Jagielski 等通过增加 γ-均等机率(Equalized Odds)和 ε-差分隐私的量化约束,给出了综合优化精确度、公平性和敏感属性差分隐私的公平机器学习方法[110]. Tran 等提出了差分隐私公平深度学习的拉格朗日对偶方法[111],将受保护群体的均等机率、统计公平和精度平等的公平性度量作为优化的约束条件,同时利用了差分隐私随机梯度(L2 范数裁剪和添加高斯噪声)下降方法,前者用于抑制不公平性,后者用于增强隐私性.

Pentyala 等认为减轻偏见本质上需要使用所有用户的敏感属性值,而联邦学习的目标是通过不授予用户数据访问权限来保护隐私,因此联邦学习中实现群体公平具有挑战性. 为了平衡联邦学习中公平性和隐私之间的冲突,Pentyala 等将联邦学习与安全多方计算和差分隐私相结合,提出了一种隐私保护下在跨设备联邦学习中训练具备群体公平性模型的方法[112]. So 等通过聚焦服务器能否使用多轮聚合模型来重构单个局部模型(即长期隐私安全),提出了用于度量联邦学习长期隐私的指标,进一步证明即使在每一轮训练中都使用了安全聚合,联邦

学习中传统随机客户端选择策略也会导致局部模型在一定轮数(与客户端数量线性相关)内泄漏,并由此断言客户端持续多轮参与联邦学习训练会导致严重的隐私泄露.为了在考虑客户端选择公平性的同时保证长期隐私,该研究提出了能够保障客户端多轮隐私(长期隐私)的安全聚合框架[113].

Rodríguez-Gálvez 等将差分乘子法用于以群体公平性(敏感属性)为约束条件的约束优化问题的求解,通过差分隐私随机梯度下降实现客户端梯度更新的隐私保护,给出了公平隐私联邦学习方法 FPFL(Fair and Private Federated Learning)[114]. Lyu 等给出了贡献与模型性能相称的公平隐私联邦学习方法 FPPDL[101,115],其中,隐私保护包括两个阶段:(1)初始阶段,每个客户端利用局部数据训练出差分隐私对抗生成网络,得到采样数据并进行发布;(2)训练阶段,客户端采用三层洋葱式加密更新梯度.协作公平机制为:每个客户端对其他客户端的信息交互行为进行声誉评分,并发送与声誉排序相称的梯度更新至其他客户端,公平性用不同客户端贡献与其对应模型精准度的相关系数来度量.

Padala 等给出了两阶段客户端公平隐私保护联邦学习方法[116],首先,拉格朗日乘子法求解满足统计平等和均等机会的客户端公平联邦学习模型 Fair-SGD;其次,客户端训练模型 Fair-SGD 的差分隐私近似模型 DP-SGD.客户端将近似模型 DP-SGD 的梯度更新参数上传至服务器,实现了原始数据和梯度参数的隐私保护.不足之处在于,客户端的模型公平性并不一定能保证全局模型的公平性.

为了平衡公平性与准确性,并应对信息受限和协调受限等问题给提升公平性带来的挑战,Zhang 等给出了公平联邦学习框架 FairFL[117]. FairFL 主要有两部分组成,其一是基于团队马尔可夫博弈(Team Markov Game for Client Selection,TMGCS)的多智能体协同强化学习,允许客户协作决定是否参与本地更新过程;其二是安全聚合协议,旨在解决 TMGCS 的近视视图问题,即允许客户端在不违反隐私约束的情况下收集有关所有客户端状态的信息.

基于 Zhang 等人关于 DP-SGD 中剪裁边界将极大影响公平性的论断[117],Gu 等人借助于梯度裁剪、局部差分隐私、全局差分隐私实现了联邦学习中准确率、公平性及隐私性间的平衡,并进一步探讨了不同梯度裁剪边界和隐私噪声对联邦学习的影响.结果表明,更严格的隐私保护会削弱公平性,但在梯度分布不均匀时,局部差分隐私可以通过扰动梯度的方式增强公平性[118].

针对纵向联邦学习(Vertical Federated Learning,VFL)的公平性增强与隐私保护,Qi 等提出了一个公平纵向联邦学习框架 FairVFL[119]. FairVFL 的核心思想是以隐私保护的方式,基于分散的特征字段学习样本的统一和公平表示.具体而言,具有用于目标任务且公平不敏感特征(fairness-insensitive features)的客户端独立从本地特征中学习数据表示,然后将这些本地表示上传到服务器聚合为目标任务的统一表示.为了学习与公平相关的统一表示,各客户端训练得到的数据表示被发送到含有公平敏感特征(fairness-sensitive features)的平台,以应用对抗性学习消除从有偏数据中继承的偏见.在隐私保护方面,则借助于新提出的对比对抗学习方法,在将隐私信息发送到含有公平敏感特征的平台之前从统一表示中删除隐私信息. Salem 等则研究了时空联邦学习(spatial-temporal FL)中的位置隐私保护与公平性[120].考虑到这一场场景中公平性的有效度量,该工作假设具有相似熵的用户轨迹应该获得相似的隐私增益,因此提出基于香农熵(更大的熵代表轨迹具备更大的无序性和不可预测性)量化隐私保护程度及公平性,并进一步通过对比联邦学习前后用户轨迹的熵变情况判定时空联邦学习的公平性.实验表明,在应用联邦学习之后,具备低前熵的用户后熵变化较小,具备高前熵的用户后熵变化较大,即联邦学习引发了不公平性.不足之处是,该工作聚焦于从实验角度证明隐私保护联邦学习会影响公平性,并未对其作用机理进行探究,也未提出富有针对性的解决方案.

联邦学习的公平性与隐私性涉及两个方面,其一,隐私保护联邦学习的公平性;其二,公平联邦学习的隐私性.隐私保护联邦学习采取了加密类技术(如,多方安全计算、同态加密、秘密共享)和数据扰动技术(如,差分隐私)[28-30].加密类技术对公平性的影响,以及基于加密类技术隐私保护联邦学习的公平性设计值得探索.许多已有的公平联邦学习设计方法的模型梯度参数传递仍然缺乏隐私保护机制,同样值得改进.此外,当前对于联邦学习隐私性及公平性的研究大多聚焦于横向联邦学习,对纵向联邦学习等其他联邦学习范式的关注相对较少.

5 联邦学习的公平性与鲁棒性

对于样本微小的变化,模型的预测不会产生巨

大变化,从而保证预测结果的稳定性,这样的模型称为鲁棒(性)模型,或者具有鲁棒性[27].投毒攻击、后门攻击、拜占庭攻击、搭便车攻击等是 FL 鲁棒性的主要威胁[27,121],其目的在于降低或破坏 FL 模型预测的准确性.公平性和鲁棒性是 FL 应用部署期望兼备的性能.

Wang 等对均等机率公平机器学习的数据投毒攻击进行了研究,分析发现:投毒攻击不仅降低模型的预测精度,而且会带来明显的公平性损失;机器学习模型存在明显的公平性和鲁棒性冲突,增强公平性会降低针对投毒攻击的鲁棒性[122].Chang 等对 FL 的边缘样本后门攻击(Edge-case Backdoors)进行了分析[123],发现此类后门攻击对公平性有较大影响,同时,后门攻击防御技术(如,Krum、Multi-Krum、弱差分隐私等)会降低诚实客户的性能,使其受到不平等对待,因此,需要折衷考虑 FL 的鲁棒性和公平性.

Hu 等将 FL 描述为多梯度下降算法求解的多目标优化问题[124],提出了满足公平性和鲁棒性的 FL 方法 FedMGDA+(Multiple-Gradient Descent Algorithm,MGDA),该方法通过客户端预测误差的方差来度量 FL 的公平性(客户端精准度的一致性),模型梯度参数聚合对客户端梯度更新进行了归一化预处理,防止了恶意客户端梯度的肆意"放大".Li 等认为训练数据的统计异构性(非独立同分布)是导致 FL 模型的准确性、公平性和鲁棒性之间相互冲突的根源,提出了全局正则化联邦多任务学习方法 Ditto,实现了客户端精确度分布的公平性和投毒攻击(标签投毒、随机梯度更新、模型替换)的鲁棒性的综合设计[125].

Xu 等提出了鲁棒公平联邦学习框架 RFFL(Robust and Fair Federated Learning)[126],用本地模型的精度来衡量参与者的贡献,协作公平性采用参与者贡献和回报模型性能的皮尔逊相关系数来度量,参与者的声誉用参与者的模型梯度和声誉加权聚合梯度的余弦相似度来计算.RFFL 根据参与者的贡献对其分配相当性能的模型,并检测和剔除贡献低甚至无贡献的客户端,通过声誉评分聚合全局模型参数,同时考虑了协作公平性和拜占庭攻击的鲁棒性.

ur Rehman 等通过计算客户端精度 A_i 的均值 μ_i 和方差 σ_i,并依据统计控制域 $A_i \in \{\sigma_i \pm \mu_i\}$ 来检测异常客户端,不满足统计控制域的异常客户端的梯度更新将不能参与服务器的模型参数聚合,既实现了客户端模型性能的公平性又防护了恶意客户端的攻击[127].

Wang 等研究了无人驾驶飞行器场景的鲁棒公平联邦学习问题[128],通过设计训练数据规模、数据质量、通信时间、声誉和补偿等的合约,将合约理论用于设计满足合约公平性(参与公平性和报酬公平性)的最优合约,将鲁棒聚合规则 Multi-Krum 用于过滤低质量的局部模型更新(异常值),并根据客户端的学习记录和行为动态估计客户端的长期声誉,长期声誉用于公平分配模型的预期效益以吸纳可信客户参与训练.

联邦学习的公平性已成为敌手攻击的新目标,恶意敌手可以通过数据投毒来影响模型性能分布的一致性[129],公平联邦学习的后门攻击、协作公平性攻击也是值得研究的问题.公平联邦学习的鲁棒性需要应对这些公平性攻击的防御机制[130],同时也要考虑异常检测中的公平性[131].此外,已有的公平联邦学习的鲁棒性研究,基本上都是都基于恶意用户攻击防御角度的鲁棒性,实现联邦学习的公平性和鲁棒性的综合量化设计是值得关注的研究[132].

6　区块链联邦学习的公平性

6.1　区块链概念

区块链作为比特币的底层支撑技术[133],能在去中心化的环境下以安全可验证且不可篡改的方式有效地记录各方之间的交易.在区块链中,网络中的任何节点都可以进行事务的验证和转发,所有节点共同维护包含事务的区块有序链接的分类账.除了数字货币领域外,区块链已广泛应用于智慧城市、物联网、医疗保健等领域的交易记录、消息传递、身份认证和访问控制管理等,区块链技术正在改变着各种无信任环境下的交易模式[134-135].

区块链具有去中心化、可追溯、匿名性和不可篡改性等特点[133-134]:区块链利用对等(Peer to Peer,P2P)网络,不需要第三方或单个中心节点来协助网络传播,所有节点都是对等的;区块链的特殊结构使得区块链上的数据可以追溯其来源;虽然链上的数据是公开的,但可以通过加密用户的隐私信息来防止他人获得;区块链结构存储的数据很难更改.

区块链是一个分布式账本,每个矿工在本地保留一个完整分类账的副本,并竞争赢得生成包含交易包的新区块的机会.比特币系统是公共的,这意味着每个人都可以在没有权限要求的情况下加入或离

开.而其他基于区块链的系统是私有的,只允许认证用户参与.通常,区块链可以大致分为三类,即私有区块链、联盟区块链和公共区块链.

私有区块链简称为私有链,与公有链相比,私有链上的节点处于监管之下,这意味着只有授权节点才能加入该网络并访问共享账本.同时,私有链上的节点对其他节点是公共的,允许该区块链上的所有行为或活动都是可追踪的.然而,私有链在某种程度上并不是完全分散的.

联盟区块链简称为联盟链,联盟链是部分分散,并由几个预定义或选定的节点(即有权生成新区块的机构)控制.联盟链是具有不同权限机制的私有链.

公共区块链简称为公有链,在公有链中,每个人都可以在没有权限的情况下加入或离开,并参与共识过程和访问公共分类账.比特币和以太坊就是公有链.公有链是完全分散的,没有控制网络的中央机构,公有链上的记录保持不变.然而,在公有链上进行交易的速度有限,因为该链上有大量用户,需要处理的交易量很大.

私有链和联盟链也被称为许可链,因为它们都需要用户在注册到区块链网络之前获得许可.在区块链的应用中,应该采用什么样的区块链取决于用途及目的.

6.2 区块链联邦学习

区块链与联邦学习的结合产生了新的联邦学习范式:区块链联邦学习(Blockchain Federated Learning,BCFL)[24,135-136],BCFL解决了传统FL面临的一些问题.首先,去中心化可以在FL中部署的区块链得以实现,这意味着中央聚合器可以由点对点区块链系统来代替,全局模型的聚合工作可以由区块链节点来完成,从而克服了集中式服务器的单点故障所带来的整个FL系统的不可靠性[137].此外,区块链还可以通过交易验证为FL提供验证机制,从而实现在全局模型聚合之前删除不合格甚至恶意的本地模型更新[138].此外,区块链可以有效地实施FL客户端的奖励分配,以鼓励客户端诚实地积极参与模型训练[139].

BCFL至少具有如下一些优点[140]:(1)区块链代替中央聚合器可以避免单点故障,BCFL系统的模型聚合可以由多个客户端执行;(2)不可靠的数据可以通过验证机制得到过滤,在聚合客户端模型更新之前,检测出不可靠的数据,并且有效数据才能参与全局模型的聚合;(3)激励机制可以吸引更多参与者和计算资源,货币激励(例如,加密货币)鼓励更多客户端参与模型训练,还鼓励客户端遵守规则使用数据;(4)相关数据或信息可以在分布式账本上存储和共享,一旦数据记录在分布式账本上,就很难被篡改.同时,授权客户可以访问分布式账本检索公共数据,从而提高训练效率.

区块链和联邦学习的结合主要体现为三种形式的框架结构[24]:完全耦合BCFL(Fully Coupled BCFL,FuC-BCFL)、柔性耦合BCFL(Flexible Coupled BCFL,FlC-BCFL)和松散耦合BCFL(Loosely Coupled BCFL,LoC-BCFL).

(1)完全耦合BCFL

区块链网络嵌入FL客户端的框架结构称为完全耦合区块链联邦学习,换言之,客户端不仅训练本地模型,还验证更新并生成新的区块.完全耦合BCFL的拓扑结构如图4所示[24].在完全耦合BCFL中,FL模型是分散的,区块链上的每个节点/客户端都有机会参与本地模型训练和全局模型聚合,中央聚合器的角色由区块链来承担.

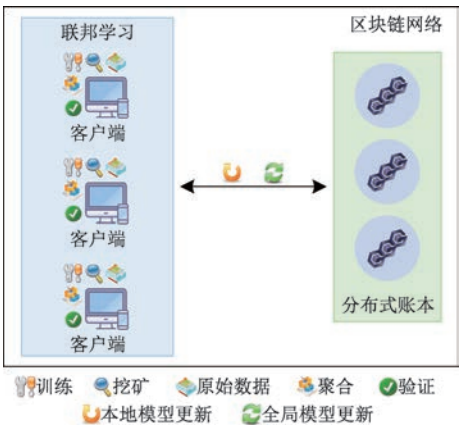

图4 完全耦合联邦学习拓扑结构

在完全耦合BCFL中,全局模型参数聚合有两种方式:①选定的客户端(节点)收集经过验证的客户端部模型更新,然后执行聚合算法;②所有客户端都可以参与全局模型聚合.分布式账本存储经验证的本地模型更新、全局模型更新和学习过程中产生的其他数据.完全耦合BCFL的大致工作流程为:客户端收集数据并在当地训练模型→(选定的)客户端验证客户端模型更新→(选定的)客户端收集已验证的客户端更新并聚合新的全局模型→已验证模型更新的新区块被添加存储到分布式账本→根据激励机制分配奖励给参与者.

完全耦合BCFL具有如下优点:①每个客户端(节点)都有一份分布式账本,因此可以有效避免单

点故障;②无需将数据传输到任何中央服务器,避免了数据隐私泄露,降低了通信成本.其不足之处在于:①由于区块链和 FL 的操作都在同一网络上运行,客户不仅要进行本地训练,还要整合全局模型,因此需要更多的计算资源;②区块链网络的通信带宽有限,因此通信延迟可能对完全耦合 BCFL 的部署构成挑战.

(2)柔性耦合 BCFL

区块链和 FL 系统处于不同的网络的框架结构定义为柔性耦合区块链联邦学习.在该结构中,FL 的客户端不是区块链的节点(矿工).柔性耦合 BCFL 的拓扑结构如图 5 所示[24].从拓扑结构中可以看出,客户负责本地数据收集和训练,而本地模型更新验证将由区块链上的矿工完成.在柔性耦合 BCFL 中,区块链存储模型更新,区块链上的矿工也可以聚合全局模型,该系统中没有中央聚合器.

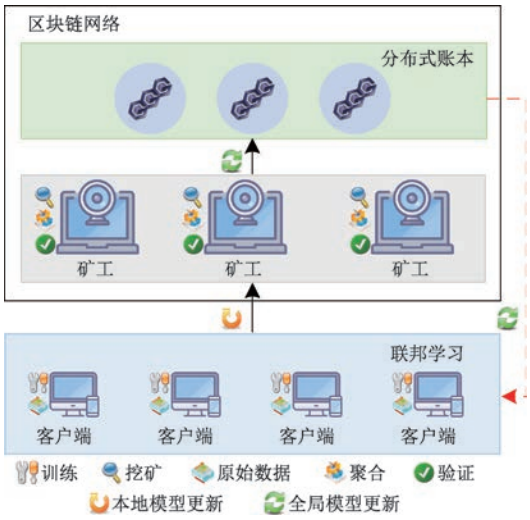

图 5　柔性耦合联邦学习拓扑结构

柔性耦合 BCFL 的大致工作流程为:客户端收集本地数据并训练本地模型并将本地模型更新上传到区块链→区块链上的矿工执行验证机制(只有经过验证的更新才能用于更新全局模型)→全局模型聚合数据存储在分布式账本→根据参与者的表现分配奖励.

柔性耦合 BCFL 具有如下优势:①FL 和区块链在不同的网络和设备上运行,减少了通信压力和延迟;②原始数据保留在客户端,降低了区块链网络被恶意攻击所导致的数据泄漏风险;③区块链可以为 FL 提供数据共享,比传统的 FL 更高效.柔性耦合 BCFL 存在如下不足:①区块链和 FL 属于两个不同的系统,因此很难协调它们的管理;②如果存在中央聚合器时,仍会发生单点故障.

(3)松散耦合 BCFL

FL 采用服务器-客户端模式,区块链用于验证模型更新和管理客户端的声誉,这种框架结构称为松散耦合联邦学习.在松散耦合 BCFL 中,只有声誉相关的数据才保留在分布式账本,验证更新和声誉管理是激励机制的一部分,以确保参与者能够诚实行事.松散耦合 BCFL 的拓扑结构如图 6 所示[24].

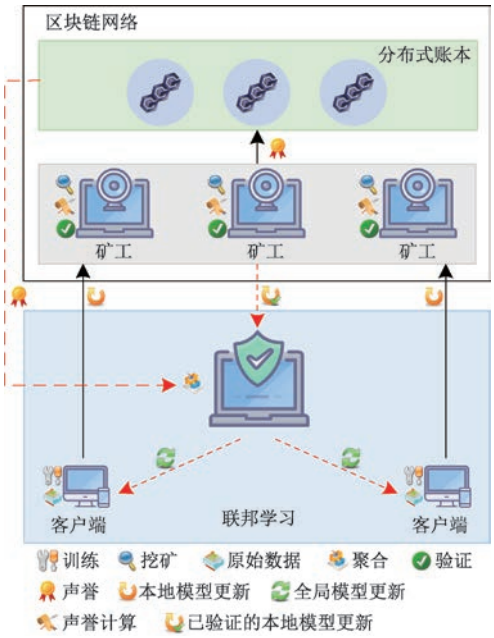

图 6　松散耦合联邦学习拓扑结构

松散耦合 BCFL 的工作流程如下:客户训练本地模型并将本地模型更新上传到区块链→矿工验证当地模型更新并评估客户的声誉→矿工们竞争生成包含声誉相关数据的新区块(新区块将被添加到分布式账本)→聚合器收集经过验证的更新并执行全局模型聚合算法→依据客户的信息实施奖励和惩罚.

松散耦合 BCFL 具有如下优势:①区块链和 FL 完全独立,FL 在其客户端上更好地保留其数据;②声誉管理机制能够更好地管理参与者,确保模型训练期间提交的数据质量,提高了模型的准确性,还可以防止恶意参与者攻击系统.其不足之处在于:①区块链很少参与 FL 过程,只负责验证和声誉管理,因此 FL 模型没有分散,隐私数据泄露和单点故障等风险仍然存在;②独立维护区块链和 FL,导致资源利用效率低下.

公有链由于其分散性和透明性,在 BCFL 系统中被广泛使用.公有链上的节点可以是任何愿意且有足够能力参与学习过程的设备,无需进一步认证.与公有链相比,许可链只对授权客户可用.在 BCFL

系统中,设备在 FL 中注册之前,需要根据其计算资源、参与意愿和历史性能进行选择.

公有链 BCFL 具有如下优点:①可以吸引更多的数据资源和计算能力,协同训练一个通用模型,从而实现大规模 FL 任务;②公有链是完全分散和透明的,因此学习过程是可追溯和可审计的.其不足之处在于:①对所有设备开放可能导致难以阻止的低质量数据和恶意行为;②通常需要复杂的共识机制,以验证模型更新并创建新的区块,从而导致计算资源的大量消耗.

许可链 BCFL 的优点在于:①为轻量级共识协议提供了一个平台,在保持系统安全的同时减少资源消耗;②将未经授权的设备排除在模型训练之外,减少了系统遭受恶意攻击的风险;③许可链内的评估方案可以使授权节点的性能得到约束,从而保证模型的准确性.许可链 BCFL 具有如下缺陷:①对设备和计算资源的吸引力不如公有链;②由于用户访问门槛,系统适用性降低.

区块链的类型本质上决定了 BCFL 系统用户的数量和质量.由于在某些计算环境中需要更多的计算资源和更多的参与者,这种任务的 FL 系统需要选择公有链.然而,如果 FL 模型的训练需要小规模实施,则可以选择许可链.

6.3 公平区块链联邦学习

公平区块链联邦学习是区块链与公平联邦学习的结合.主要结合方式包括:(1)区块链智能合约或共识机制用于公平的客户端选择、公平激励机制设计;(2)区块链的分布式账本用于存储公平性设计和决策相关的信息.

Kuo 等提出了公平区块链联邦学习 GloreChain(Grid Binary Logistic Regression on Permissioned BlockChain)[141],客户端轮流承担全局模型参数聚合任务,通过平等证明(Proof of Equity,PoE)共识机制来确定客户端进行全局参数聚合的公平轮叫(Round-robin)次序,区块链上存储了模型的部分信息.GloreChain 不能防御恶意攻击,也不具有客户端之间梯度信息传递的隐私保护机制.

Weng 等给出了 DeepChain(Deep Learning with Blockchain-based Incentive)方法[142],客户端训练本地模型并进行梯度更新交易,矿工下载梯度更新进行模型参数聚合,最快完成聚合任务的矿工被选作为领袖,领袖获得创建区块的权利并将其聚合梯度用于更新模型参数,创建区块的领袖会获得一定的代币奖励.客户端之间的梯度信息传送采用了 Paillier 同态加密算法,实现了梯度信息的隐私保护.激励机制由区块链的两种安全机制组成:可信时钟机制确保合约的操作在规定时间内完成,货币惩罚机制对梯度收集和合谋破译进行处罚.这一激励机制增强了 FL 的鲁棒性和协作公平性.DeepChain 难以保障客户端模型性能一致的公平性,也无法应对敏感属性公平性.

Bao 等给出的 Flchain(Federated Learning Blockchain)方法中[143],客户端训练本地模型,并从聚合器获得全局模型更新,获得全局模型需要支付一定的模型使用费用,客户端支付费用构成了 FL 的收益,购买和使用模型的客户端上传模型的评价,客户端的可靠性也随之更新.客户端梯度的交互传送采用了秘密共享机制.区块链存储训练信息、验证训练过程,提供按照客户端的贡献和可靠性进行收益分配的激励机制,并鼓励客户端检测其他客户端的不诚实行为.

Toyoda 等提出的区块链联邦学习框架中[144],客户端在区块链注册登记,具有任务要求的客户端在区块链上发布任务,所有登记的客户端可以决定是否同意参与该任务训练.发布任务的客户端选择训练的轮次以及参与训练的客户端.在每一训练轮次,智能合约随机选择固定数目的客户端,这些客户端从区块链下载模型参数,并上传模型梯度更新,这些客户端会选择上一轮客户端提交的前 k 个梯度更新用以更新自己的本地模型.模型参数或梯度更新采用了公钥加密技术.在每一轮次训练后,参与训练的客户端对前 k 个模型进行投票,智能合约计算出所有前面训练轮次中各个模型的排序得分总和并用于奖励分配.

Kang 等利用多权重主观逻辑模型(Multi-weight Subjective Logic Model)方法计算客户端的声誉,通过联盟区块链管理声誉,基于声誉来选择客户端以公平度量客户端的可靠性进而防御不可靠客户端的模型梯度攻击,设计了客户端声誉与合约理论结合的奖励机制,鼓励具有高声誉和高质量数据的客户端参与模型训练[145].服务器发布学习任务、计算客户端声誉、设计合约条款、选择参与学习的客户端、聚合模型参数、评估客户端模型质量、更新声誉得分并上传至区块链.客户端训练本地模型并发送梯度更新到任务发布服务器.该区块链联邦学习方法中缺乏梯度更新的隐私保护机制,也没有考虑客户端性能一致的公平性.

TrustFed 方法是一个 LoC-BCFL 框架结构[14]

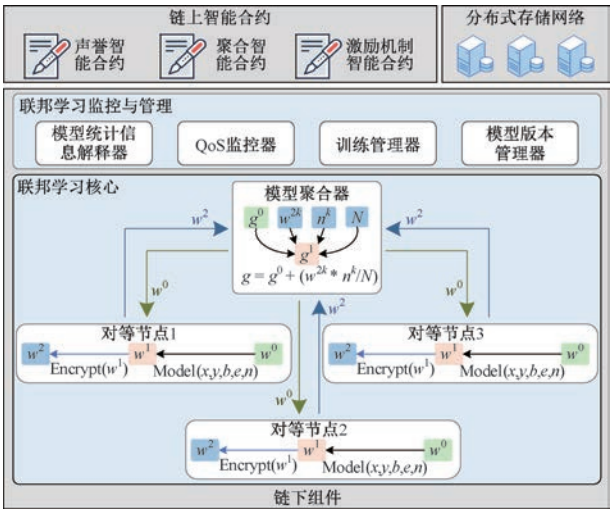

图 7　TrustFed 拓扑结构

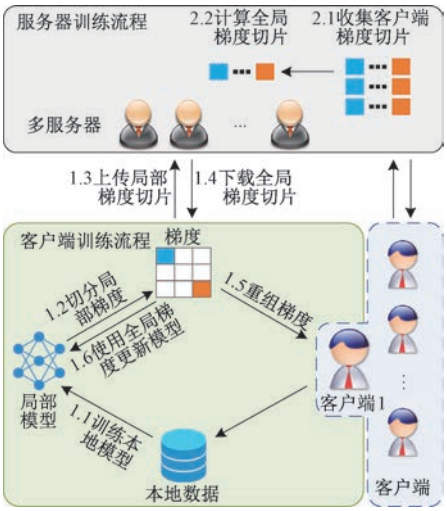

图 8　FGFL 拓扑结构

(参见图 7). TrustFed 中联邦学习模型的训练及聚合主要由对等节点和模型聚合器两类实体实现:对等节点即客户端,可以负责训练本地模型(基于模型聚合器下发的模型 w^0,利用 Model 函数对本地数据进行训练得到新模型 w^1,并利用 Encrypt 函数对其进行加密得到 w^2.),也可以承担全局模型参数聚合任务,还可以担任本地模型训练和聚合器双重角色;模型聚合器依据区块链上的声誉记录随机选择一组超过声誉阈值要求的客户端参与本地模型训练,并基于前一全局模型对各客户端上传的模型进行聚合,得到新的全局模型. 区块链上设计了 3 个智能合约:声誉智能合约维护一个社区驱动的分散式声誉系统,客户端的声誉存储在分布式账本,任何客户端都不能篡改区块链上的声誉记录,新客户端可以自行注册、检查和更新所有其他设备的声誉;激励机制智能合约帮助聚合器发布每次任务的服务质量(Quality of Service,QoS)要求,并通过数字货币代币奖励客户端;聚合智能合约根据聚合器报告的加权矩阵和模型参数进模型参数的链上聚合. TrustFed 缺乏客户端之间梯度信息交互的隐私保护机制,也没有考虑客户端性能分布一致的公平性以及敏感属性公平性.

Gao 等给出了公平区块链联邦学习方法 FGFL (Blockchain-based incentive Governor for Federated Learning)[146](参见图 8),通过多中心联邦学习结构网络来协调客户端训练. 在该模型框架结构中,所有客户端都参与训练本地模型,部分可靠的客户端既要训练本地模型训练又要承担模型参数聚合任务(不可靠的客户端只能训练本地模型,不能承担模型参数聚合任务),客户端与承担聚合任务的客户端之间梯度更新信息传递采用了数字签名方式;区块链存储客户端的声誉和贡献,承担聚合任务的客户端的签名和信息,并通过智能合约来管理公平激励系统. 智能合约由五个主要功能模块组成:攻击检测模块、声誉模块、贡献模块、激励模块和激励分配模块. 攻击检测模块接收客户端的梯度更新,并依据客户端梯度与聚合模型梯度的相似度剔除有害的客户端梯度更新(梯度相似度取负值). 声誉模块采用主观逻辑模型(Subjective Logic Model,SLM)方法根据客户端检测的历史记录来计算客户端的声誉指数. 贡献模块基于梯度相似度计算客户端贡献. 激励模块使用声誉和贡献的乘积来确定客户端奖励分配份额,综合考虑客户端声誉和贡献,增强了公平性. 激励反馈模块用于根据客户端激励值(声誉和贡献的乘积)来确定客户端的最佳训练策略. FGFL 是一种 LoC-BCFL 结构,具有隐私保护性和鲁棒性. 然而,FGFL 的奖励分配是一种间接行为,没有直接影响客户端模型性能,也没有考虑客户端性能分布一致公平性以及敏感属性公平性.

Lo 等设计了公平可追责联邦学习方案[147],根据全局模型测试数据集的类分布,动态调整客户端各类数据样本的采样权重,以实现训练数据样本分布的公平性. 在每一训练轮次,客户端训练局部模型,并将局部模型参数和数据版本的哈希值上传给数据模型注册智能合约;服务器聚合模型参数,同样将全局模型参数的哈希值上传给数据模型注册智能合约. 局部模型参数的哈希值还经过了对称加密. 区块链记录数据、局部和全局模型版本的哈希值,实现数据-模型的可追溯. 此外,Lo 等进一步基于区块链实现了可信联邦学习体系结构[148],以同步增强联

邦学习系统的可追责和公平性.在这一体系结构中,区块链承担数据模型来源注册中心的角色,并基于智能合约实现问责制.公平性增强则由加权公平数据采样器算法实现,本质上是通过预处理技术实现对训练数据的公平性增强.不同于以安全、隐私为侧重点的可信联邦学习研究,该研究更多关注的是联邦学习的可追责和公平性,且模型整体性能较好.

Rückel 等将零知识证明、局部差分隐私用于区块链联邦学习,以增强联邦学习的公平性、完整性和隐私性[149].零知识证明用于客户端验证其他客户端是否根据他们之前承诺的隐私数据真实训练了他们所提交的模型更新,这一过程无需任何客户端隐私数据.为了确保客户端的模型更新不会泄露其数据的模式信息,利用局部差分隐私对每个客户端的模型更新添加了拉普拉斯噪声.区块链的智能合约基于客户端的实际参数(即无局部差分隐私噪声)衡量每个客户端对全局模型性能的贡献,并给予相应的奖励.该方法对客户端的贡献评测需要公共测试数据集,同时缺乏对恶意客户端攻击的防御机制,也没有考虑客户端性能分布一致公平性以及敏感属性公平性.

区块链实现了公平联邦学习的可追责性,为可信联邦学习研究探索了一条可行途径.从公平性角度,现有工作大都侧重于激励公平(参见表 5),缺乏对客户端模型一致公平性、敏感属性公平的研究.同时,区块链带来新的计算和通信资源需求,这些和联邦学习训练的资源需求如何折衷考虑?区块链和公平联邦学习需要协调运作,模型的训练成本不仅关系到全局模型的性能,而且影响到整个模型的可追责性,考虑计算成本下的公平联邦学习是需要进一步研究的问题.

表 5 公平区块链联邦学习的相关研究

研究文献	BCFL 框架结构	公平性	鲁棒性	隐私保护
Kuo et al. 2019[141]	FuC-BCFL	客户端选择	×	×
Weng et al. 2019[142]	FuC-BCFL	激励机制	√	同态加密
Bao et al. 2019[143]	FuC-BCFL	激励机制	√	秘密共享
Toyoda et al. 2019[144]	FlC-BCFL	激励机制	×	公钥加密
Kang et al. 2019[145]	LoC-BCFL	客户端选择、激励机制	√	×
ur Rehman et al. 2021[14]	LoC-BCFL	客户端选择、激励机制	√	×
Gao et al. 2022[146]	LoC-BCFL	贡献评估、激励机制	√	数字签名
Lo et al. 2022[147]	LoC-BCFL	平衡训练数据	×	公钥加密
Rückel et al. 2022[149]	FlC-BCFL	激励机制	×	零知识证明、差分隐私

7 进一步研究工作展望

公平联邦学习是为了适应可信人工智能应用部署的需要而提出的机器学习解决方案.虽然近些年产出了不少的研究工作,但是无论从联邦学习的公平性定义及度量、公平联邦学习方法、鲁棒公平联邦学习,还是从面向人工智能生态的健康持续发展要求的符合伦理联邦学习,尚需要进一步的探索和研究.

7.1 公平性定义及度量

客户端数据的本地训练,实现了对原始数据的隐私保护,是联邦学习的突出优点.然而,敏感属性公平性的度量通常需要数据样本的属性信息,尽管客户端局部模型的公平性可以通过本地数据的属性信息进行度量,但是局部模型的公平性并不代表全局模型的公平性,缺失原始数据属性信息的公平性度量是公平联邦学习面临的挑战[6].分布式鲁棒优化(Distributionally Robust Optimization,DRO)通过优化训练过程中任意群体的最差个体结果来增强公平性[150],该优化过程并不需要群体及其成员的信息,基于 DRO 构建无需属性信息的联邦学习公平性定义和度量是值得开展的研究.多校准(Multi-calibration)是群体的多个子群体上满足校准的群体公平性[151],目前尚没有在联邦学习中得到应用,建立联邦学习的多校准公平性概念及度量是有前景的技术路径.客户端本地模型公平性度量经过聚合也可以获得全局模型的公平性度量[50],需要研究敏感属性公平性的复合和聚合机制,建立聚合公平性框架的定义及度量.

偏见是引发歧视和导致不公平的主要来源.针对偏见的来源及其特征建立公平性的定义及度量具有重要的意义.联邦学习不仅存在传统机器学习的偏见[4-5],而且存在客户端选择、数据异构性、聚合算法等可能引发的偏见[103,152].从单个客户端角度,FL 的本地模型训练类似于传统集中式机器学习,传统

机器学习存在的偏见都会在 FL 中出现,各个客户端的这些偏见通过梯度更新带入全局模型将会引起更为复杂的公平性问题;客户端参与全局模型的训练需要通过选择,由于各种原因或设计机制(频次控制、采样、丢弃、掉队等),客户端参与训练的机会并不平等,客户端缺失的训练轮次可能与敏感属性有关,从而引发新的偏见;不同客户端拥有不同的数据,客户端数据的异构性是联邦学习面临的重要挑战[6],即使各个客户端平等地参与了模型训练,这种异构性依然会诱发预测偏见;聚合算法对于各个客户端梯度更新的不同加权会导致全局模型产生新的不公平性[153].对于这些联邦学习的偏见特殊性,需要研究适合于联邦学习的公平性定义及度量.

敏感属性公平性、模型公平性、协作公平性分别从不同的视角对联邦学习的公平性进行了定义和度量.敏感属性公平性侧重于受保护群体的模型预测精度的平等无偏见;模型公平性聚焦于各个参与方所使用的模型在预测性能上的一致性;协作公平性着重于通过公平的激励分配机制吸引客户端参与联邦学习训练,进而促进模型性能的改善.尽管如此,这些公平性的终极目标应该是有助于联邦学习模型提供精确且无偏的预测或决策.敏感属性公平性是属性级公平性度量,模型公平性是客户端级公平性度量,属性级和客户端级的一体化综合公平性度量是值得研究的问题[47-48,74].协作公平性从客户端的数据质量、数据收集成本或模型性能改进等贡献给予奖励分配,这些贡献并没有考虑属性级或客户端级的性能公平性,需要建立能够实现属性级或客户端级性能公平性的协作公平性定义及度量[52].公平性定义及度量是静态的,然而模型决策的公平性具有动态性和时滞效应[3,154-155],研究联邦学习公平性的动态演化机理、动态公平性定义及度量显得非常必要[156].

7.2　公平联邦学习方法

训练数据预处理是集中式机器学习公平性增强的一类重要方法[4-5],通过对训练数据进行预先修改(如,改变敏感属性或类标签、加权采样),以消除训练数据中的不公平.联邦学习客户端数据具有明显的数据集漂移和非独立同分布(non-Independent and Identically Distributed,non-IID)特征[6],客户端会受到参与训练的时刻影响而带来训练数据的时间漂移,计算能力较强的客户端会有更多机会参与训练而带来训练数据代表过度,数据规模较小的客户端需要较少训练时间也会有更多机会参与训练而带来训练数据代表过度,受地理位置或通信能力影响的客户端会有较少机会参与训练而带来训练数据代表不足.在联邦学习的局部和全局加权预处理方法研究中[152],这些训练数据特征必须得到充分考虑.根据全局模型测试数据集的类分布,动态调整客户端各类数据样本的采样权重,以实现平衡的训练数据样本分布,是一个可行的技术路径[147].公平表示方法无需改变训练数据[157],能够保持数据的完整性,公平联邦表示学习是值得探索的研究方向[158].

模型压缩不仅有助于联邦学习的高效率通信和模型训练以及公平客户端选择,而且也是联邦学习的公平非货币激励(奖励不同性能的模型)的需要[159-160].梯度量化通过将梯度向量的每个元素量化为有限位低精度值来对梯度向量进行有损压缩[161],而梯度稀疏化是通过选择性地传输部分梯度向量来减少通信消耗[162].模型剪枝是删除一定剪枝率的模型最小绝对值权值,模型剪枝通过减小模型的规模大小来降低传输时的通信成本[163].这些技术可用于服务器端,也可以用于客户端.为了实现高效率和高性能的模型压缩,可以对这些技术组合使用[164-166].梯度稀疏化和梯度量化如何影响客户端模型性能的一致性?如何实现通信效率和客户端模型性能一致性的折衷处理?梯度稀疏化和梯度量化是否会引发新的偏见?模型的剪枝与模型性能的对应关系以及对各个客户端性能影响是否一致?如何设计基于模型压缩技术的联邦学习奖励机制?综合通信效率、模型性能一致性、激励机制的模型压缩技术及公平联邦学习设计方法是需要进一步研究的问题.

联邦学习中客户端数据集的非独立同分布问题是影响联邦学习训练效果的重要因素[6],联邦学习实际应用的客户端数据质量和分布是不可控的,无法要求客户端数据满足独立同分布,因此联邦学习算法需要对于非独立同分布数据具有良好的性能表现.如果将每个客户端的训练视为一个单独的任务,联邦学习就可以看作多任务联邦学习[167-168],多任务学习的训练结果是每个客户端任务对应一个模型,这使得此类技术与跨场景 FL 应用相关,难以应用于跨设备 FL.多任务联邦学习中客户端模型性能一致的公平性如何衡量?多任务联邦学习的激励机制、公平多任务联邦学习方法都是需要研究的问题.联邦元学习方法将元学习(Meta Learning)算法用于联邦学习[169-170],为每个客户端训练个性化的模

型,缩小模型在不同客户端上表现的差异,有助于增强客户端模型性能一致公平性.然而,全局模型性能和客户端模型性能之间的关系缺乏量化,依据全局模型的性能训练的模型可能会损害客户端的后期性能[171].具有激励机制的公平元联邦学习方法是值得进一步研究的问题.

7.3 鲁棒公平联邦学习

联邦学习的攻击分为隐私攻击和安全攻击两类[6,27,121].隐私攻击的对象一般是用户的数据集、训练模型的参数等,目的在于非法获取或窃取用户隐私信息.安全攻击则是通过一些恶意样本对模型的预测结果产生负面影响,目的在于恶化模型的预测性能(准确度),即模型的鲁棒性.公平联邦学习同样会受到类似方式的攻击,然而,安全攻击的目的不仅可能恶化模型预测的准确度而且还可能降低模型预测的公平性.为了表述方便,将以模型预测的准确度为攻击目的的安全攻击称为鲁棒性攻击,将以模型预测的公平性为攻击目的安全攻击称为公平性攻击.联邦学习的公平性具有敏感属性公平性、模型公平性、协作公平性等多个维度[45].敏感属性公平性与数据敏感属性有关,敏感属性公平性攻击可能需要隐私攻击和鲁棒性攻击的双重作用.模型公平性攻击要达到客户端模型性能(预测精度)分布的恶化,鲁棒性攻击无法实现这一任务.协作公平性涉及贡献评估和效益分配等因素,攻击协作公平性就更为复杂.因此需要对公平性攻击机理开展探究.

数据投毒攻击通过修改或者注入挑选的训练数据内容,以降低模型预测准确度[6,172].数据投毒攻击可以改变训练数据的局部分布,实现公平性攻击[129,173-175].对抗样本攻击是一种作用于模型预测阶段让模型产生错误分类的攻击方式[176].模型投毒攻击通过控制客户端传递给服务器的模型更新参数以影响模型的预测准确度[177].后门攻击是攻击者在模型的训练过程中隐藏后门,以使模型输出变为攻击者预先指定的标签以实现攻击者的意图.联邦学习模式下,恶意参与者很容易完成后门攻击[178-179],攻击者可以在本地模型更新数据发送到服务器之前对更新数据进行投毒,也可以在全局模型中植入供攻击者使用的后门[180].此外,如果允许敌手共谋[181-182],模型更新投毒攻击的有效性可能会大幅提高,这类攻击可以让敌手创建更有效、更难以检测的模型更新攻击[183-184].这些对鲁棒公平联邦学习设计带来了新的挑战,需要开展进一步的研究.联邦学习公平性的模型投毒攻击、后门攻击、共谋攻击、组合攻击[185],以及各类公平性攻击的防御技术亟待研究[186-188].

全面理解联邦学习的隐私性、公平性和鲁棒性之间的关系,对于鲁棒公平联邦学习设计具有重要的意义[6,189-191].许多联邦学习的已有研究,通常集中在鲁棒性(单一攻击方式)、隐私性和公平性中的单一方面.综合设计具有鲁棒性、隐私性和公平性的联邦学习系统,是一个挑战性课题.综合设计方法可以从不同但互补的机制中实现对多种模型性能的同步改善.隐私保护的差分隐私机制可以减轻数据推理攻击,并有助于提升防御数据投毒攻击的鲁棒性,同时还能发现新的机器学习漏洞[192-193].在独立于某些敏感属性的数据表示场景,隐私性和公平性可以同时得到满足,并能够保持模型预测的准确度,这可以通过对隐私性(处理数据隐藏敏感属性)和公平性(表示学习的模型具有敏感属性公平性)两个方面的综合设计实现.在集中式机器学习中,对抗训练是实现这种表示学习的可行方法[194-196],联邦学习的客户端可以对本地数据进行类似处理,以强制或改善 FL 的隐私性和/或公平性.然而,设计这种 FL 的类似数据处理技术(满足隐私性和/或公平性约束)还是一个尚待研究的问题.

7.4 符合伦理联邦学习

国际电气电子工程师协会发布了《符合伦理设计:人工智能和自主系统促进人类福祉的远景》[197-198],目的在于推动人工智能和自主系统伦理的公开讨论,促进人工智能和自主系统朝着造福于人类的方向发展.欧盟委员会的欧洲人工智能高级别专家组撰写并发布了《可信赖 AI 的伦理指导原则》[199],从尊重人的自主性、预防伤害、公平性、可解释性等四个方面提出了可信赖 AI 的伦理原则,从受人类监管、技术的稳健性和安全性、隐私和数据管理、透明度、非歧视性和公平性、社会和环境福祉、问责制等七个方面提出了可信赖 AI 应当满足的条件.联合国教科文组织通过的《人工智能伦理建议书》中提出了包括公平性和非歧视性、隐私权和数据保护、透明度和可解释性等在内的十项 AI 原则.可解释、可追责、隐私保护、公平性等被列为可信人工智能的基本伦理属性,也被认为是符合伦理的可信人工智能的重要维度.联邦学习是实现人工智能的重要方法,在人类决策中发挥着愈来愈重要作用,联邦学习应用的推广有赖于人们对其信任程度的提高,符合伦理的联邦学习是必然的发展方向[1-2,200-201],可解释、可追责、隐私保护、公平联邦学习设计是值得

探索的研究[200,202].

随着机器学习技术在人类社会和生活应用中的广度和深度不断拓展,人工智能系统的责任界定与追责问题愈来愈引起了关注.因此,在与符合伦理机器学习相关的研究和讨论中,可追责性也被普遍认为是其中的重要维度之一[203-204].作为典型机器学习范式的联邦学习,是实现人工智能的重要方法,并且拥有数据分布式存储、局部模型本地化训练、全局模型个性化使用(主要体现在个性化联邦学习中)等传统机器学习所不具备的功能或特征,使得其责任界定与追责问题进一步突出.区块链分布式账本能够真实记录交易详情且交易记录不可被篡改、智能合约能够实现中立且可信的自动程序执行,使得区块链能够可靠地实现对历史交易的可追溯性,这些区块链固有特征为解决联邦学习的可追责性提供了一条可行的技术路径[202].例如,基于区块链智能合约实现联邦学习中恶意客户端的自动检测及追责[205-206],基于区块链实现对局部模型及全局模型的验证、审计与追责[150,207].尽管区块链公平联邦学习具有诸多方面的优势,然而仍然存在制约其设计及应用的一些问题[24,148-149].区块链联邦学习应该在鲁棒性、隐私性、公平性和高效性等方面得到良好平衡.首先,鲁棒性和隐私性对于区块链公平联邦学习非常重要,BCFL 的客户端可以根据区块链的公开信息获得公共地址等身份信息实现相互沟通,增加了客户端之间的共谋风险,需要研究这些攻击的防御机制和鲁棒性设计方法.其次,无论 BCFL 的数据或模型参数是在客户端之间还是通过区块链的矿工进行验证都需要耗费一定的时间,FL 和区块链网络都存在通信延迟,这些因素影响着 FL 的训练效率,需要研究考虑网络通信和区块链计算延迟的公平联邦学习设计方法.最后,BCFL 不仅需要本地模型训练、模型聚合和更新,还需要数据验证和块生成,这些活动消耗了大量的计算资源,增加了训练模型的成本,这些是未来研究需要解决的问题.

可解释是符合伦理联邦学习的另一个重要维度[6,16,208-209].公平联邦学习的可解释性研究还非常有限.Haffar 等构建了黑盒 FL 模型的随机森林代理模型,将有限深度随机森林用于可解释的黑盒 FL 模型的(错误)预测,并利用随机森林代理模型来检测针对 FL 模型训练的安全和隐私攻击[210].Chen 等提出了一个基于反事实解释的可解释纵向联邦学习框架,通过特征的重要性评估来实现特征的解释[211].Wang 等给出了联邦学习模型的 Shapley 解释方法,Shapley 值可以客观反映客户端对整体全局模型贡献[212].Salim 等将 SHAP(SHapley Additive exPlanations)可解释性方法用于输入特征重要性的可视化,通过 Shapley 值对模型进行各种特征组合测试,获得每个特征的相关性以解释模型的预测结果[213-214].Raza 等在心电图(Electrocardiography,ECG)数据的迁移联邦学习模型上建立了一个可视化可解释模块,以帮助解释模型的预测结果,并做出快速和可靠的决策[215].然而,公平联邦学习的可解释性尚未开展研究.公平联邦学习的可解释性不仅有助于说明和理解模型预测结果,而且可用于模型的安全和隐私攻击的诊断和分析[216],还可用于公平性相关的偏见发现,这些都是需要进一步研究的问题.

8　结束语

随着可信人工智能、可信联邦学习、公平机器学习等概念的不断提出,联邦学习的公平性逐渐引起了学术界和产业界的关注,并诞生了一系列研究成果.本文从联邦学习的公平性概念、公平联邦学习设计、联邦学习公平性与隐私性及鲁棒性的同步增强、区块链联邦学习的公平性等角度出发,对已有研究工作进行了分类梳理,系统地分析了当前的研究进展及存在的主要问题,并进一步从公平性定义及度量、公平联邦学习方法、鲁棒公平联邦学习、符合伦理联邦学习等角度,对未来的研究工作进行了展望.公平联邦学习及其设计存在大量的值得探讨和尚待解决的问题,亟待开展进一步的深入研究.

参　考　文　献

[1] Kaur D, Uslu S, Rittichier K J, et al. Trustworthy artificial intelligence: A review. ACM Computing Surveys, 2022, 55(2): 1-38

[2] Eshete B. Making machine learning trustworthy. Science, 2021, 373(6556): 743-744

[3] Chouldechova A, Roth A. A snapshot of the frontiers of fairness in machine learning. Communications of the ACM, 2020, 63(5): 82-89

[4] Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning. ACM Computing Surveys, 2021, 54(6): 1-35

[5] Gu Tian-Long, Li Long, Chang Liang, et al. Fair machine learning: Concepts, analysis, and design. Chinese Journal of Computers, 2022, 46(5): 1018-1051(in Chinese)

(古天龙，李龙，常亮等. 公平机器学习：概念、分析与设计. 计算机学报，2022，46(5)：1018-1051)

[6] Kairouz P，McMahan H B，Avent B，et al. Advances and open problems in federated learning. Foundations and Trends® in Machine Learning，2021，14(1-2)：1-210

[7] Yang Q，Liu Y，Chen T，et al. Federated machine learning：Concept and applications. ACM Transactions on Intelligent Systems and Technology，2019，10(2)：1-19

[8] Konečny J，McMahan H B，Yu F X，et al. Federated learning：Strategies for improving communication efficiency. arXiv：1610. 05492，2016

[9] Li T，Sahu A K，Talwalkar A，et al. Federated learning：Challenges，methods，and future directions. IEEE Signal Processing Magazine，2020，37(3)：50-60

[10] Abdulrahman S，Tout H，Ould-Slimane H，et al. A survey on federated learning：The journey from centralized to distributed on-site learning and beyond. IEEE Internet of Things Journal，2020，8(7)：5476-5497

[11] Zhang X，Gu H，Fan L，et al. No free lunch theorem for security and utility in federated learning. arXiv：2203.05816，2022

[12] Yang Z，Shi Y，Zhou Y，et al. Trustworthy Federated Learning via Blockchain. IEEE Internet of Things Journal，2023，10(1)：92-109

[13] Zhou Z，Chu L，Liu C，et al. Towards fair federated learning//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York，USA，2021：4100-4101

[14] ur Rehman M H，Dirir A M，Salah K，et al. TrustFed：A framework for fair and trustworthy cross-device federated learning in IIoT. IEEE Transactions on Industrial Informatics，2021，17(12)：8485-8494

[15] Zhang C，Xie Y，Bai H，et al. A survey on federated learning. Knowledge-Based Systems，2021，216：106775

[16] Li Q，Wen Z，Wu Z，et al. A survey on federated learning systems：Vision，hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering，2021，PP(99)：1-1

[17] Lo S K，Lu Q，Wang C，et al. A systematic literature review on federated machine learning：From a software engineering perspective. ACM Computing Surveys，2021，54(5)：1-39

[18] Zhan Y，Li P，Guo S，et al. Incentive mechanism design for federated learning：Challenges and opportunities. IEEE Network，2021，35(4)：310-317

[19] Zhan Y，Zhang J，Hong Z，et al. A survey of incentive mechanism design for federated learning. IEEE Transactions on Emerging Topics in Computing，2022，10(2)：1035-1044

[20] Zeng R，Zeng C，Wang X，et al. A comprehensive survey of incentive mechanism for federated learning. arXiv：2106.15406，2021

[21] Ali A，Ilahi I，Qayyum A，et al. Incentive-driven federated learning and associated security challenges：A systematic review. TechRxiv：14945433.v1，2021

[22] Tu X，Zhu K，Luong N C，et al. Incentive mechanisms for federated learning：From economic and game theoretic perspective. arXiv：2111. 11850，2021

[23] Liu Y，Zhang L，Ge N，et al. A systematic literature review on federated learning：From a model quality perspective. arXiv：2012. 01973，2020

[24] Wang Z，Hu Q. Blockchain-based federated learning：A comprehensive survey. arXiv：2110. 02182，2021

[25] Tan A Z，Yu H，Cui L，et al. Towards personalized federated learning. arXiv：2103. 00710，2021

[26] Ji S，Saravirta T，Pan S，et al. Emerging trends in federated learning：From model fusion to federated x learning. arXiv：2102. 12920，2021

[27] Lyu L，Yu H，Yang Q. Threats to federated learning：A survey. arXiv：2003. 02133，2020

[28] Yin X，Zhu Y，Hu J. A comprehensive survey of privacy-preserving federated learning：A taxonomy，review，and future directions. ACM Computing Surveys，2021，54(6)：1-36

[29] Mothukuri V，Parizi R M，Pouriyeh S，et al. A survey on security and privacy of federated learning. Future Generation Computer Systems，2021，115：619-640

[30] Liu Yi-Xuan，Chen Hong，Liu Yu-Han，et al. Privacy-preserving techniques in federated learning. Journal of Software，2022，33(3)：1057-1092(in Chinese)
(刘艺璇，陈红，刘宇涵等. 联邦学习中的隐私保护技术. 软件学报，2022，33(3)：1057-1092)

[31] Lim W Y B，Luong N C，Hoang D T，et al. Federated learning in mobile edge networks：A comprehensive survey. IEEE Communications Surveys & Tutorials，2020，22(3)：2031-2063

[32] Khan L U，Saad W，Han Z，et al. Federated learning for internet of things：Recent advances，taxonomy，and open challenges. IEEE Communications Surveys & Tutorials，2021，23(3)：1759-1799

[33] Nguyen D C，Ding M，Pathirana P N，et al. Federated learning for internet of things：A comprehensive survey. IEEE Communications Surveys & Tutorials，2021，23(3)：1622-1658

[34] Abreha H G，Hayajneh M，Serhani M A. Federated learning in edge computing：A systematic survey. Sensors，2022，22(2)：1-45

[35] Nguyen D C，Pham Q V，Pathirana P N，et al. Federated learning for smart healthcare：A survey. ACM Computing Surveys，2022，55(3)：1-37

[36] Shyu C R，Putra K T，Chen H C，et al. A systematic review of federated learning in the healthcare area：From the perspective of data properties and applications. Applied Sciences，2021，11(23)：1-35

[37] Antunes R S, da Costa C A, Küderle A, et al. Federated learning for healthcare: Systematic review and architecture proposal. ACM Transactions on Intelligent Systems and Technology, 2022, 13(4): 1-23

[38] Pfitzner B, Steckhan N, Arnrich B. Federated learning in a medical context: A systematic literature review. ACM Transactions on Internet Technology, 2021, 21(2): 1-31

[39] Liu M, Ho S, Wang M, et al. Federated learning meets natural language processing: A survey. arXiv: 2107.12603, 2021

[40] Yu B, Mao W, Lv Y, et al. A survey on federated learning in data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2022, 12(1): 1-20

[41] Gadekallu T R, Pham Q V, Huynh-The T, et al. Federated learning for big data: A survey on opportunities, applications, and future directions. arXiv: 2110.04160, 2021

[42] Agrawal S, Sarkar S, Aouedi O, et al. Federated learning for intrusion detection system: Concepts, challenges and future directions. arXiv: 2106.09527, 2021

[43] Hutchinson B, Mitchell M. 50 years of test (un) fairness: Lessons for machine learning//Proceedings of the Conference on Fairness, Accountability, and Transparency. Atlanta, USA, 2019: 49-58

[44] Liu Wen-Yan, Shen Chu-Yun, Wang Xiang-Feng, et al. Survey on fairness in trustworthy machine learning. Journal of Software, 2021, 32(5): 1-24(in Chinese)
(刘文炎, 沈楚云, 王祥丰等. 可信机器学习的公平性综述. 软件学报, 2021, 32(5): 1-24)

[45] Shi Y, Yu H, Leung C. A survey of fairness-aware federated learning. arXiv: 2111.01872, 2021

[46] Li T, Sanjabi M, Beirami A, et al. Fair resource allocation in federated learning. arXiv: 1905.10497, 2019

[47] Ezzeldin Y H, Yan S, He C, et al. FairFed: Enabling group fairness in federated learning. arXiv: 2110.00857, 2021

[48] Zhang F, Kuang K, Liu Y, et al. Unified group fairness on federated learning. arXiv: 2111.04986, 2021

[49] Martinez N, Bertran M, Sapiro G. Minimax pareto fairness: A multi objective perspective//Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria, 2020: 6755-6764

[50] Lahoti P, Beutel A, Chen J, et al. Fairness without demographics through adversarially reweighted learning//Proceedings of the 34th Conference on Neural Information Processing Systems. 2020: 1-13

[51] Diana E, Gill W, Kearns M, et al. Minimax group fairness: Algorithms and experiments//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, USA, 2021: 66-76

[52] Hu S, Wu Z S, Smith V. Provably fair federated learning via bounded group loss. arXiv: 2203.10190, 2022

[53] Donahue K, Kleinberg J. Models of fairness in federated learning. arXiv: 2112.00818, 2021

[54] Mohri M, Sivek G, Suresh A T. Agnostic federated learning//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 4615-4625

[55] Zhou P, Fang P, Hui P. Loss tolerant federated learning. arXiv: 2105.03591, 2021

[56] Huang T, Lin W, Wu W, et al. An efficiency-boosting client selection scheme for federated learning with fairness guarantee. IEEE Transactions on Parallel and Distributed Systems, 2020, 32(7): 1552-1564

[57] Nishio T, Yonetani R. Client selection for federated learning with heterogeneous resources in mobile edge//Proceedings of the IEEE International Conference on Communications. Shanghai, China, 2019: 1-7

[58] Wang H, Kaplan Z, Niu D, et al. Optimizing federated learning on non-IID data with reinforcement learning//Proceedings of the IEEE Conference on Computer Communications. Toronto, Canada, 2020: 1698-1707

[59] Yoshida N, Nishio T, Morikura M, et al. Hybrid-FL for wireless networks: Cooperative learning mechanism using non-IID data//Proceedings of the IEEE International Conference on Communications. Dublin, Ireland, 2020: 1-7

[60] Lyu L, Xu X, Wang Q, et al. Collaborative fairness in federated learning//Yang Q, Fan L, Yu Heds. Federated Learning. Cham: Springer, 2020: 189-204

[61] Yu H, Liu Z, Liu Y, et al. A fairness-aware incentive scheme for federated learning//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, USA, 2020: 393-399

[62] Yang M, Wang X, Zhu H, et al. Federated learning with class imbalance reduction//Proceedings of the 29th European Signal Processing Conference. Dublin, Ireland, 2021: 2174-2178

[63] Caldas S, Konečny J, McMahan H B, et al. Expanding the reach of federated learning by reducing client resource requirements. arXiv: 1812.07210, 2018

[64] Bouacida N, Hou J, Zang H, et al. Adaptive federated dropout: Improving communication efficiency and generalization for federated learning//Proceedings of the IEEE Conference on Computer Communications Workshops. 2021: 1-6

[65] Li T, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks//Proceedings of the Machine Learning and Systems. Austin, USA, 2020: 429-450

[66] Du W, Xu D, Wu X, et al. Fairness-aware agnostic federated learning//Proceedings of the 2021 SIAM International Conference on Data Mining. Virtual Event, 2021: 181-189

[67] Wu Y, Zhang L, Wu X. On convexity and bounds of fairness-aware classification//Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 3356-3362

[68] Donini M, Oneto L, Ben-David S, et al. Empirical risk minimization under fairness constraints. arXiv: 1802.

08626, 2018

[69] Xu R, Cui P, Kuang K, et al. Algorithmic decision making with conditional fairness//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 2125-2135

[70] Cotter A, Jiang H, Gupta M R, et al. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. Journal of Machine Learning Research, 2019, 20(172): 1-59

[71] Cui S, Pan W, Liang J, et al. Addressing algorithmic disparity and performance inconsistency in federated learning. arXiv: 2108. 08435, 2021

[72] Chu L, Wang L, Dong Y, et al. Fedfair: Training fair models in cross-silo federated learning. arXiv: 2109. 05662, 2021

[73] Shekhar S, Fields G, Ghavamzadeh M, et al. Adaptive sampling for minimax fair classification. arXiv: 2103. 00755, 2021

[74] Papadaki A, Martinez N, Bertran M, et al. Minimax demographic group fairness in federated learning. arXiv: 2201. 08304, 2022

[75] Yue X, Nouiehed M, Kontar R A. Gifair-fl: An approach for group and individual fairness in federated learning. arXiv: 2108. 02741, 2021

[76] Wang Z, Fan X, Qi J, et al. Federated learning with fair averaging. arXiv: 2104. 14937, 2021

[77] Huang W, Li T, Wang D, et al. Fairness and accuracy in horizontal federated learning. Information Sciences, 2022, 589: 170-185

[78] Kanaparthy S, Padala M, Damle S, et al. Fair federated learning for heterogeneous face data. arXiv: 2109. 02351, 2021

[79] Han P, Wang S, Leung K K. Adaptive gradient sparsification for efficient federated learning: An online learning approach//Proceedings of the 40th International Conference on Distributed Computing Systems. Singapore, 2020: 300-310

[80] Zhang J, Li C, Robles-Kelly A, et al. Hierarchically fair federated learning. arXiv: 2004. 10386, 2020

[81] Kang J, Xiong Z, Niyato D, et al. Incentive design for efficient federated learning in mobile networks: A contract theory approach//Proceedings of the 2019 IEEE VTS Asia Pacific Wireless Communications Symposium. Singapore, 2019: 1-5

[82] Sarikaya Y, Ercetin O. Motivating workers in federated learning: A stackelberg game perspective. IEEE Networking Letters, 2019, 2(1): 23-27

[83] Le T H T, Tran N H, Tun Y K, et al. An incentive mechanism for federated learning in wireless cellular networks: An auction approach. IEEE Transactions on Wireless Communications, 2021, 20(8): 4874-4887

[84] Zhang J, Wu Y, Pan R. Incentive mechanism for horizontal federated learning based on reputation and reverse auction// Proceedings of the Web Conference. Ljubljana, Slovenia, 2021: 947-956

[85] Zeng R, Zhang S, Wang J, et al. Fmore: An incentive scheme of multi-dimensional auction for federated learning in MEC//Proceedings of the 40th International Conference on Distributed Computing Systems. Singapore, 2020: 278-288

[86] Wang G, Dang C X, Zhou Z. Measure contribution of participants in federated learning//Proceedings of the 2019 IEEE International Conference on Big Data. Los Angeles, USA, 2019: 2597-2604

[87] Nishio T, Shinkuma R, Mandayam N B. Estimation of individual device contributions for incentivizing federated learning//Proceedings of the 2020 IEEE Globecom Workshops. Taipei, China, 2020: 1-6

[88] Shapley L S. Stochastic games. Proceedings of the National Academy of Sciences, 1953, 39(10): 1095-1100

[89] Ghorbani A, Zou J. Data shapley: Equitable valuation of data for machine learning//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 2242-2251

[90] Song T, Tong Y, Wei S. Profit allocation for federated learning//Proceedings of the 2019 IEEE International Conference on Big Data. Los Angeles, USA, 2019: 2577-2586

[91] Wei S, Tong Y, Zhou Z, et al. Efficient and fair data valuation for horizontal federated learning//Yang Q, Fan L, Yu H eds. Federated Learning. Cham: Springer, 2020: 139-152

[92] Wang T, Rausch J, Zhang C, et al. A principled approach to data valuation for federated learning//Yang Q, Fan L, Yu H eds. Federated Learning. Cham: Springer, 2020: 153-167

[93] Fan Z, Fang H, Zhou Z, et al. Improving fairness for data valuation in federated learning. arXiv: 2109. 09046, 2021

[94] Fan Z, Fang H, Zhou Z, et al. Fair and efficient contribution valuation for vertical federated learning. arXiv: 2201. 02658, 2022

[95] Shyn S K, Kim D, Kim K. FedCCEA: A practical approach of client contribution evaluation for federated learning. arXiv: 2106. 02310, 2021

[96] Hu M, Wu D, Zhou Y, et al. Incentive-aware autonomous client participation in federated learning. IEEE Transactions on Parallel and Distributed Systems, 2022, 33(10): 2612-2627

[97] Feng S, Niyato D, Wang P, et al. Joint service pricing and cooperative relay communication for federated learning//2019 International Conference on Internet of Things and IEEE Green Computing and Communications and IEEE Cyber, Physical and Social Computing and IEEE Smart Data. Atlanta, USA, 2019: 815-820

[98] Cong M, Yu H, Weng X, et al. A VCG-based fair incentive mechanism for federated learning. arXiv: 2008. 06680, 2020

[99] Ye D, Yu R, Pan M, et al. Federated learning in vehicular edge computing: A selective model aggregation approach. IEEE Access, 2020, 8: 23920-23935

[100] Sim R H L, Zhang Y, Chan M C, et al. Collaborative machine learning with incentive-aware model rewards. arXiv: 2010. 12797, 2020

[101] Lyu L, Yu J, Nandakumar K, et al. Towards fair and privacy-preserving federated deep models. IEEE Transactions on Parallel and Distributed Systems, 2020, 31(11): 2524-2541

[102] Hao W, El-Khamy M, Lee J, et al. Towards fair federated learning with zero-shot data augmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3310-3319

[103] Abay A, Zhou Y, Baracaldo N, et al. Mitigating bias in federated learning. arXiv: 2012. 02447, 2020

[104] Datta A, Sen S, Tschantz M C. Correspondences between privacy and nondiscrimination: why they should be studied together. arXiv: 1808. 01735, 2018

[105] Ekstrand M D, Joshaghani R, Mehrpouyan H. Privacy for all: Ensuring fair and equitable privacy protections//Proceedings of the Conference on Fairness, Accountability and Transparency. New York, USA, 2018: 35-47

[106] Cummings R, Gupta V, Kimpara D, et al. On the compatibility of privacy and fairness//Proceedings of the 27th Conference on User Modeling, Adaptation and Personalization. Larnaca, Cyprus, 2019: 309-315

[107] Pujol D, McKenna R, Kuppam S, et al. Fair decision-making using privacy-protected data//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona, Spain, 2020: 189-199

[108] Bagdasaryan E, Poursaeed O, Shmatikov V. Differential privacy has disparate impact on model accuracy//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 15479-15488

[109] Zhang T, Zhu T, Gao K, et al. Balancing learning model privacy, fairness, and accuracy with early stopping criteria. IEEE Transactions on Neural Networks and Learning Systems, 2021, PP(99): 1-13

[110] Jagielski M, Kearns M, Mao J, et al. Differentially private fair learning//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 3000-3008

[111] Tran C, Fioretto F, Van Hentenryck P. Differentially private and fair deep learning: A lagrangian dual approach. arXiv: 2009. 12562, 2020

[112] Pentyala S, Neophytou N, Nascimento A, et al. PrivFairFL: Privacy-preserving group fairness in federated learning. arXiv: 2205. 11584, 2022

[113] So J, Ali R E, Guler B, et al. Securing secure aggregation: Mitigating multi-round privacy leakage in federated learning. arXiv: 2106. 03328, 2021

[114] Rodríguez-Gálvez B, Granqvist F, van Dalen R, et al. Enforcing fairness in private federated learning via the modified method of differential multipliers. arXiv:2109. 08604, 2021

[115] Lyu L, Yu H, Ma X, et al. Privacy and robustness in federated learning: Attacks and defenses. arXiv: 2012. 06337, 2020

[116] Padala M, Damle S, Gujar S. Federated learning meets fairness and differential privacy//Proceedings of the International Conference on Neural Information Processing. Sanur, Indonesia, 2021: 692-699

[117] Zhang D Y, Kou Z, Wang D. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models//Proceedings of the 2020 IEEE International Conference on Big Data. 2020: 1051-1060

[118] Gu X, Tianqing Z, Li J, et al. Privacy, accuracy, and model fairness trade-offs in federated learning. Computers & Security, 2022, 122: 1-10

[119] Qi T, Wu F, Wu C, et al. FairVFL: A fair vertical federated learning framework with contrastive adversarial learning. arXiv: 2206. 03200, 2022

[120] Salem A B, Khalfoun B, Mokhtar S B, et al. Quantifying fairness of federated learning LPPM models//Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services. Portland, USA, 2022: 569-570

[121] Zhou Chun-Yi, Chen Da-Wei, Wang Shang, et al. Research and challenge of distributed deep learning privacy and security attack. Journal of Computer Research and Development, 2021, 58(5): 927-943(in Chinese)
(周纯毅, 陈大卫, 王尚等. 分布式深度学习隐私与安全攻击研究进展与挑战. 计算机研究与发展, 2021, 58(5): 927-943)

[122] Wang H, Sreenivasan K, Rajput S, et al. Attack of the tails: Yes, you really can backdoor federated learning//Proceedings of the 34th Conference on Neural Information Processing Systems. 2020:1-13

[123] Chang H, Nguyen T D, Murakonda S K, et al. On adversarial bias and the robustness of fair machine learning. arXiv: 2006. 08669, 2020

[124] Hu Z, Shaloudegi K, Zhang G, et al. Fedmgda+: Federated learning meets multi-objective optimization. arXiv: 2006. 11489, 2020

[125] Li T, Hu S, Beirami A, et al. Ditto: Fair and robust federated learning through personalization//Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria, 2021: 6357-6368

[126] Xu X, Lyu L. Towards building a robust and fair federated learning system. arXiv: 2011. 10464, 2020

[127] ur Rehman M H, Dirir A M, Salah K, et al. FairFed: Cross-device fair federated learning//Proceedings of the 2020 IEEE Applied Imagery Pattern Recognition Workshop. Washington, USA, 2020: 1-7

[128] Wang Y, Su Z, Luan T, et al. Federated learning with fair incentives and robust aggregation for UAV-aided crowdsensing. IEEE Transactions on Network Science and Engineering, 2022, 19(3): 2608-2618

[129] Van M H, Du W, Wu X, et al. Poisoning attacks on fair machine learning. arXiv: 2110. 08932, 2021

[130] Hu H, Lan C. Inference attack and defense on the distributed private fair learning framework//Proceedings of the AAAI Workshop on Privacy-preserving Artificial Intelligence. New York, USA, 2020: 1-7

[131] Singh A K, Blanco-Justicia A, Domingo-Ferrer J, et al. Fair detection of poisoning attacks in federated learning//Proceedings of the 32nd International Conference on Tools with Artificial Intelligence. Baltimore, USA, 2020: 224-229

[132] Li T, Beirami A, Sanjabi M, et al. Tilted empirical risk minimization. arXiv: 2007. 01162, 2020

[133] Böhme R, Christin N, Edelman B, et al. Bitcoin: Economics, technology, and governance. Journal of economic Perspectives, 2015, 29(2): 213-38

[134] Shao Qi-Feng, Jin Che-Qing, Zhang Zhao, et al. Blockchain: Architecture and research progress. Chinese Journal of Computers, 2018, 41(5): 969-988(in Chinese)
(邵奇峰, 金澈清, 张召等. 区块链技术: 架构及进展. 计算机学报, 2018, 41(5): 969-988)

[135] Qu Y, Uddin M P, Gan C, et al. Blockchain-enabled federated learning: A survey. ACM Computing Surveys, 2022, PP(99): 1-33

[136] Lu Y, Huang X, Dai Y, et al. Blockchain and federated learning for privacy-preserved data sharing in industrial IoT. IEEE Transactions on Industrial Informatics, 2019, 16(6): 4177-4186

[137] Ramanan P, Nakayama K. BAFFLE: Blockchain based aggregator free federated learning//Proceedings of the 2020 IEEE International Conference on Blockchain. Rhodes, Greece, 2020: 72-81

[138] Kim Y J, Hong C S. Blockchain-based node-aware dynamic weighting methods for improving federated learning performance//Proceedings of the 20th Asia-Pacific Network Operations and Management Symposium. Matsue, Japan, 2019: 1-4

[139] Liu Y, Ai Z, Sun S, et al. Fedcoin: A peer-to-peer payment system for federated learning//Yang Q, Fan L, Yu H. Federated Learning. Cham: Springer, 2020: 125-138

[140] Kang J, Xiong Z, Jiang C, et al. Scalable and communication-efficient decentralized federated edge learning with multi-blockchain framework//Proceedings of the International Conference on Blockchain and Trustworthy Systems. Singapore, 2020: 152-165

[141] Kuo T T, Gabriel R A, Ohno-Machado L. Fair compute loads enabled by blockchain: Sharing models by alternating client and server roles. Journal of the American Medical Informatics Association, 2019, 26(5): 392-403

[142] Weng J, Weng J, Zhang J, et al. Deepchain: Auditable and privacy-preserving deep learning with blockchain-based incentive. IEEE Transactions on Dependable and Secure Computing, 2019, 18(5): 2438-2455

[143] Bao X, Su C, Xiong Y, et al. FLChain: A blockchain for auditable federated learning with trust and incentive//Proceedings of the 5th International Conference on Big Data Computing and Communications. Qingdao, China, 2019: 151-159

[144] Toyoda K, Zhang A N. Mechanism design for an incentive-aware blockchain-enabled federated learning platform//Proceedings of the 2019 IEEE International Conference on Big Data. Los Angeles, USA, 2019: 395-403

[145] Kang J, Xiong Z, Niyato D, et al. Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. IEEE Internet of Things Journal, 2019, 6(6): 10700-10714

[146] Gao L, Li L, Chen Y, et al. FGFL: A blockchain-based fair incentive governor for federated learning. Journal of Parallel and Distributed Computing, 2022, 163: 283-299

[147] Lo S K, Liu Y, Lu Q, et al. Towards trustworthy AI: Blockchain-based architecture design for accountability and fairness of federated learning systems. IEEE Internet of Things Journal, 2022,10(4):3276-3284

[148] Rückel T, Sedlmeir J, Hofmann P. Fairness, integrity, and privacy in a scalable blockchain-based federated learning system. Computer Networks, 2022, 202: 1-18

[149] Hashimoto T, Srivastava M, Namkoong H, et al. Fairness without demographics in repeated loss minimization//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 1929-1938

[150] Lo S K, Liu Y, Lu Q, et al. Blockchain-based trustworthy federated learning architecture. arXiv: 2108. 06912, 2021

[151] Hébert-Johnson U, Kim M, Reingold O, et al. Multicalibration: Calibration for the (computationally-identifiable) masses//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 1939-1948

[152] Mandhala V N, Bhattacharyya D, Midhunchakkaravarthy D. Need of mitigating bias in the datasets using machine learning algorithms//Proceedings of the 2022 International Conference on Advances in Computing, Communication and Applied Informatics. Chennai, India, 2022: 1-7

[153] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2017: 1273-1282

[154] Liu L T, Dean S, Rolf E, et al. Delayed impact of fair ma-

chine learning//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 3150-3158

[155] D'Amour A, Srinivasan H, Atwood J, et al. Fairness is not static: deeper understanding of long term fairness via simulation studies//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, USA, 2020: 525-534

[156] Williams J, Kolter J Z. Dynamic modeling and equilibria in fair decision making. arXiv: 1911. 06837, 2019

[157] Zemel R, Wu Y, Swersky K, et al. Learning fair representations//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA, 2013: 325-333

[158] Liang P P, Liu T, Ziyin L, et al. Think locally, act globally: Federated learning with local and global representations. arXiv: 2001. 01523, 2020

[159] Shahid O, Pouriyeh S, Parizi R M, et al. Communication efficiency in federated learning: achievements and challenges. arXiv: 2107. 10996, 2021

[160] Gao Han, Tian Yu-Long, Xu Feng-Yuan, et al. Survey of deep learning model compression and acceleration. Journal of Software, 2021, 32(1): 68-92(in Chinese)
(高晗, 田育龙, 许封元等. 深度学习模型压缩与加速综述. 软件学报, 2021, 32(1): 68-92)

[161] Reisizadeh A, Mokhtari A, Hassani H, et al. FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization//Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. 2020: 2021-2031

[162] Sun H, Ma X, Hu R Q. Adaptive federated learning with gradient compression in uplink NOMA. IEEE Transactions on Vehicular Technology, 2020, 69(12): 16325-16329

[163] Jiang Y, Wang S, Valls V, et al. Model pruning enables efficient federated learning on edge devices. IEEE Transactions on Neural Networks and Learning Systems, 2022, PP (99): 1-13

[164] Sattler F, Wiedemann S, Müller K R, et al. Robust and communication-efficient federated learning from non-IID data. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(9): 3400-3413

[165] Rothchild D, Panda A, Ullah E, et al. FetchSGD: Communication-efficient federated learning with sketching//Proceedings of the 37th International Conference on Machine Learning. 2020: 8253-8265

[166] Jiang P, Agrawal G. A linear speedup analysis of distributed deep learning with sparse and quantized communication. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 2530-2541

[167] Smith V, Chiang C K, Sanjabi M, et al. Federated multi-task learning. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 4427-4437

[168] Zhang Y, Yang Q. A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering, 2022, 34 (12): 5586-5609

[169] Chen F, Luo M, Dong Z, et al. Federated meta-learning with fast convergence and efficient communication. arXiv: 1802. 07876, 2018

[170] Fallah A, Mokhtari A, Ozdaglar A. Personalized federated learning: A meta-learning approach. arXiv: 2002. 07948, 2020

[171] Jiang Y, Konečný J, Rush K, et al. Improving federated learning personalization via model agnostic meta learning. arXiv: 1909. 12488, 2019

[172] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines. arXiv: 1206. 6389, 2012

[173] Solans D, Biggio B, Castillo C. Poisoning attacks on algorithmic fairness//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Ghent, Belgium, 2020: 162-177

[174] Jo C, Sohn J, Lee K. Breaking fair binary classification with optimal flipping attacks. arXiv: 2204. 05472, 2022

[175] Mehrabi N, Naveed M, Morstatter F, et al. Exacerbating algorithmic bias through fairness attacks. arXiv: 2012. 08723, 2020

[176] Yuan X, He P, Zhu Q, et al. Adversarial examples: Attacks and defenses for deep learning. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824

[177] Bhagoji A N, Chakraborty S, Mittal P, et al. Analyzing federated learning through an adversarial lens//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 634-643

[178] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning//Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. 2020: 2938-2948

[179] Chen Da-Wei, Fu An-Min, Zhou Chun-Yi, et al. Federated learning backdoor attack scheme based on generative adversarial network. Journal of Computer Research and Development, 2021, 58(11): 2364-2373(in Chinese)
(陈大卫, 付安民, 周纯毅等. 基于生成式对抗网络的联邦学习后门攻击方案. 计算机研究与发展, 2021, 58(11): 2364-2373)

[180] Wu C, Yang X, Zhu S, et al. Mitigating backdoor attacks in federated learning. arXiv: 2011. 01767, 2020

[181] Byrd D, Mugunthan V, Polychroniadou A, et al. Collusion resistant federated learning with oblivious distributed differential privacy. arXiv: 2202. 09897, 2022

[182] So J, Güler B, Avestimehr A S. Byzantine-resilient secure federated learning. IEEE Journal on Selected Areas in Com-

munications, 2020, 39(7): 2168-2181

[183] Fung C, Yoon C J M, Beschastnikh I. Mitigating sybils in federated learning poisoning. arXiv: 1808. 04866, 2018

[184] Xiao X, Tang Z, Li C, et al. SCA: Sybil-based collusion attacks of IIoT data poisoning in federated learning. IEEE Transactions on Industrial Informatics, 2022, 19(3): 2608-2618

[185] Schwarzschild A, Goldblum M, Gupta A, et al. Just how toxic is data poisoning? A unified benchmark for backdoor and data poisoning attacks//Proceedings of the 38th International Conference on Machine Learning. 2021: 9389-9398

[186] Huang Y, Gupta S, Song Z, et al. Evaluating gradient inversion attacks and defenses in federated learning//Proceedings of the 35th Conference on Neural Information Processing Systems. 2021: 1-10

[187] Zhao Y, Chen J, Zhang J, et al. PDGAN: A novel poisoning defense method in federated learning using generative adversarial network//Proceedings of the 19th International Conference on Algorithms and Architectures for Parallel Processing. Melbourne, Australia, 2019: 595-609

[188] Goldblum M, Tsipras D, Xie C, et al. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 1563-1580

[189] Benz P, Zhang C, Karjauv A, et al. Robustness may be at odds with fairness: An empirical study on class-wise accuracy//Proceedings of the NeurIPS 2020 Workshop on Pre-registration in Machine Learning. 2021: 325-342

[190] Nanda V, Dooley S, Singla S, et al. Fairness through robustness: Investigating robustness disparity in deep learning//Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021: 466-477

[191] Xu H, Liu X, Li Y, et al. To be robust or to be fair: Towards fairness in adversarial training//Proceedings of the 38th International Conference on Machine Learning. 2021: 11492-11501

[192] Avent B, Korolova A, Zeber D, et al. BLENDER: Enabling local search with a hybrid differential privacy model//Proceedings of the 26th USENIX Conference on Security Symposium. Vancouver, Canada, 2017: 747-764

[193] Song L, Shokri R, Mittal P. Privacy risks of securing machine learning models against adversarial examples//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. London, UK, 2019: 241-257

[194] Kairouz P, Liao J, Huang C, et al. Censored and fair universal representations using generative adversarial models. arXiv: 1910. 00411, 2019

[195] Feutry C, Piantanida P, Bengio Y, et al. Learning anonymized representations with adversarial neural networks. arXiv: 1802. 09386, 2018

[196] Madras D, Creager E, Pitassi T, et al. Learning adversarially fair and transferable representations//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 3384-3393

[197] How J P. Ethically aligned design. IEEE Control Systems Magazine, 2018, 38(3): 3-4

[198] Bryson J, Winfield A. Standardizing ethical design for artificial intelligence and autonomous systems. Computer, 2017, 50(5): 116-119

[199] Floridi L. Establishing the rules for building trustworthy AI. Nature Machine Intelligence, 2019, 1(6): 261-262

[200] Gu Tian-Long, Li Long. Artificial moral agents and their design methodology: retrospect and prospect. Chinese Journal of Computers, 2021, 44(3): 632-651 (in Chinese)
(古天龙, 李龙. 伦理智能体及其设计: 现状和展望. 计算机学报, 2021, 44(3): 632-651)

[201] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nature Machine Intelligence, 2019, 1(9): 389-399

[202] Yang Q. Toward responsible AI: An overview of federated learning for user-centered privacy-preserving computing. ACM Transactions on Interactive Intelligent Systems, 2021, 11(3-4): 1-22

[203] Smith H. Clinical AI: opacity, accountability, responsibility and liability. AI & Society, 2021, 36(2): 535-545

[204] Porter Z, Zimmermann A, Morgan P, et al. Distinguishing two features of accountability for AI technologies[J]. Nature Machine Intelligence, 2022, 4(9): 734-736

[205] Bimal Desai H, Safa Ozdayi M, Kantarcioglu M. BlockFLA: Accountable federated learning via hybrid blockchain Architecture. arXiv: 2010.07427

[206] Zhu Jian-Ming, Zhang Qin-Nan, Gao Sheng, et al. Privacy preserving and trustworthy federated learning model based on blockchain. Chinese Journal of Computers, 2021, 44(12): 2464-2484(in Chinese)
(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. 计算机学报, 2021, 44(12): 2464-2484)

[207] Peng Z, Xu J, Chu X, et al. Vfchain: Enabling verifiable and auditable federated learning via blockchain systems. IEEE Transactions on Network Science and Engineering, 2021, 9(1): 173-186

[208] Samek W, Wiegand T, Müller K R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv: 1708. 08296, 2017

[209] Majumdar S. Fairness, explainability, privacy, and robustness for trustworthy algorithmic decision making//Basak S, Vračko M. Big Data Analytics in Chemoinformatics and Bioinformatics. Amsterdam: Elsevier, 2022: 1-33

[210] Haffar R, Sánchez D, Domingo-Ferrer J. Explaining predictions and attacks in federated learning via random forests. Applied Intelligence, 2022: 1-17

[211] Chen P, Du X, Lu Z, et al. EVFL: An explainable vertical

federated learning for data-oriented artificial intelligence systems. Journal of Systems Architecture, 2022, 126: 1-12

[212] Wang G. Interpret federated learning with shapley values. arXiv: 1905. 04519, 2019

[213] Salim S, Turnbull B, Moustafa N. A blockchain-enabled explainable federated learning for securing Internet-of-Things-based social media 3. 0 Networks. IEEE Transactions on Computational Social Systems, 2021, PP(99): 1-17

[214] Lundberg S M, Lee S I. A unified approach to interpreting model predictions//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 4768-4777

[215] Raza A, Tran K P, Koehl L, et al. Designing ECG monitoring healthcare system with federated transfer learning and explainable AI. Knowledge-Based Systems, 2022, 236: 1-12

[216] Ji Shou-Ling, Li Jin-Feng, Du Tian-Yu, et al. Survey on techniques, applications and security of machine learning interpretability. Journal of Computer Research and Development, 2019, 56(10): 2071-2096(in Chinese)
(纪守领, 李进锋, 杜天宇等. 机器学习模型可解释性方法、应用与安全研究综述. 计算机研究与发展, 2019, 56(10): 2071-2096)

GU Tian-Long, Ph. D., professor. His research interests mainly include formal methods, trustworthy artificial intelligence, artificial intelligence ethics, and data governance.

LI Long, Ph. D., lecturer. His research interests mainly include artificial intelligence security, fair machine learning and logic programming.

CHANG Liang, Ph. D., professor. His research interests mainly include knowledge graph, knowledge representation and formal methods.

LI Jing-Jing, Ph. D., lecturer. Her research interests mainly include interpretable machine learning and blockchain-based federated learning.

Background

This paper is the frontier research in the field of machine learning(ML)and trustworthy artificial intelligence (AI). AI not only brings great opportunities, but also contains a series of risks and challenges, such as application risk by algorithm security, algorithm opacity by black-box model, decision bias by data discrimination, and privacy disclosure by data abuse. These problems directly affect the confidence of the society and the public on AI, and restrict the deployment of AI applications. From the perspective of academic research, the research of trusted AI includes security, explainability, fairness, privacy protection and so on. ML is an important method to realize AI, and trusted ML is the core technology to build trusted AI system. Federated learning (FL) is a distributed ML solution for model training coordinated by multiple clients (users) in the central server (aggregator), where each client has its own data set for the user. Traditional ML involves gathering data from these clients together and training the model from the aggregated data set. In FL, there is no need to collect the data of each client. Instead, the participating users train their local models locally, and upload the local model parameters to the server, which then aggregates the global model parameters (according to different FL architectures, and publish them to all participants for the clients. FL is a good solution to data silos and data privacy problems, which has attracted great attention from academic researchers and industrial applications. Influenced by the nature and technical characteristics of ML itself, the prediction and decision-making of ML will inevitably produce bias or unfairness, which has gradually attracted the attention of scientific researchers, industrial practitioners and the public. In the decision-making process, fairness refers to the absence of any prejudice, preference, discrimination or injustice based on the inherent or acquired characteristics of individuals or groups. Therefore, an unfair algorithm is one whose decisions are biased against individuals or specific groups, which leads to unfair treatment of the individual or disadvantaged groups and damages the interests of them. Fairness is endowed with richer connotation under the FL framework. The fairness of FL has two meanings: cooperative fairness and model fairness. FL with one or more aspects of fairness is called fair federated learning or fairness-aware FL. Through a systematic review and comprehensive analysis of recent research work, this paper explains the concepts, definitions and metrics of fairness in FL; In the framework of the client selection, model optimization, contribution evaluation and incentive mechanism of fair FL, the design methods of fair FL are expounded; From the perspective of trusted AI, the hybrid design of fairness, privacy and robustness of FL is discussed, and the architectures of blockchain enabled fair FL are illustrated. Finally, the main problems, challen-

ges and research hotspots regarding fair FL are presented. Recently, there are several review articles regarding FL, but they only focus on some aspects, and the scope of review is not comprehensive enough. The work of this paper benefits from the research experiences of the NSFC general projects and key projects hosted by the author in recent years. These projects have carried out a lot of research on formal methods, artificial intelligence, machine learning, knowledge engineering, big data of urban governance, and big data of education. The author has published some works, such as "formal method of software development" and "ordered binary decision graph and application", and some academic papers. Researchers can fully understand the research status of fair federated machine learning at home and abroad from the work of this paper, and it is helpful to guide interested researchers in this field to realize the state of the art of fair federated machine learning and to grasp the topics of further research.