联邦学习的公平性研究综述

朱智韬[1,2]，司世景[1]，王健宗[1]，程宁[1]，孔令炜[1]，黄章成[1]，肖京[1]
1. 平安科技（深圳）有限公司，广东 深圳 518063；
2. 中国科学技术大学，安徽 合肥 230026

摘要
联邦学习使用来自多个参与者提供的数据协同训练全局模型，近年来在促进企业间数据合作方面发挥着越来越重要的作用。另外，联邦学习训练范式常常面临数据不足的困境，因此为联邦学习参与者提供公平性保证以激励更多参与者贡献他们宝贵的资源是非常重要的。针对联邦学习的公平性问题，首先依据公平目标不同，从模型表现均衡、贡献评估公平、消除群体歧视出发进行了联邦学习公平性的3种分类；然后对现有的公平性促进方法进行了深入介绍与比较，旨在帮助研究者开发新的公平性促进方法；最后通过对联邦学习落地过程中的需求进行剖析，提出了未来联邦学习公平性研究的5个方向。

关键词
联邦学习；公平性；表现均衡；贡献衡量

中图分类号：F49，F270．7,TP399 文献标志码：A doi：10.11959/j.issn.2096-0271.2022088

A survey on the fairness of federated learning

ZHU Zhitao[1,2], SI Shijing[1], WANG Jianzong[1], CHENG Ning[1], KONG Lingwei[1], HUANG Zhangcheng[1], XIAO Jing[1]
1. Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518063, China
2. University of Science and Technology of China, Hefei 230026, China

Abstract
Federated learning uses data from multiple participants to collaboratively train global models and has played an increasingly important role in recent years in facilitating inter-firm data collaboration. On the other hand, the federal learning training paradigm often faces the dilemma of insufficient data, so it is important to provide assurance of fairness to motivate more participants to contribute their valuable resources. This paper illustrates the issue of fairness in federated learning. Firstly, three classifications of fairness based on different equity goals, from model performance balance, contribution assessment equity, and elimination of group discrimination are proposed, and then we provide in-depth introduction and comparison of existing fairness promotion methods, aiming to help researchers develop new fairness promotion methods. Finally, by dissecting the needs in the process of federal learning implementation, five directions for future federated learning fairness research are proposed.

Key words
federated learning, fairness, balance in performance, measure of contributions

0 引言

由于现实世界中诸多政府法规及行业公约的限制，许多数据集本质上是分散的，传统的机器学习方法难以突破数据实体间的重重阻隔将数据样本聚集到一个中央存储库中，因此学习能力通常受到单一数据持有方的限制。作为一种新兴的分布式机器学习框架，联邦学习（federated learning，FL）允许多个用户协作训练一个共享的全局模型，而无须将数据从本地设备集中至中心化存储器，保证了数据的隐私性与安全性[1]。

联邦学习以全局模型的精度或收敛速度为优化考量，这一原则有利于那些能够快速响应或有助于提高最终模型性能的客户，但是使弱势用户无法得到契合本地数据的全局模型，损害了全局模型的泛化性能。这类由训练过程中的偏倚引起的问题，被称为公平性问题，不公平的联邦学习导致了全局模型对部分数据实体的实质性歧视，使得参与者无法平等享受到联邦学习带来的好处；不合理地分配了协同训练的全局收益，严重损害了数据持有者的参与积极性；同时引入了对特定群体的偏见，会产生恶劣的社会影响。

现有文献中关于联邦学习公平性的综述或偏向某一方面，如文献[2-3]侧重于对公平激励机制的总结，文献[4]侧重于介绍实现公平性的技术手段；或仅机械套用传统机器学习的公平性，未见有对联邦学习公平性从产生原因到干涉环节再到方法分析的全面综述。因此本文针对联邦学习范式下可能产生的公平性问题进行分析，着重针对客户端表现均衡、客户端贡献衡量、全局模型公平这3个方面的公平定义进行详细说明，并有针对性地总结了相关研究的进展，旨在帮助有志于研究相关问题的学者快速了解当前研究现状与富有发展可能的研究方向，以期对减轻联邦学习中的不公平现象、促进联邦学习进一步发展与普及提供一定的帮助。

1 联邦学习公平性定义与分类方法

机器学习的公平性已经积累了相当多的研究[5-6]，不同的公平性概念之间往往不具有可比性，因为它们服务于不同的设计目的或特定利益相关者群体。回顾公平联邦学习系统提出的背景动机：①联邦学习生态系统的可持续发展要求不打击参与者的积极性；②联邦学习的许多参与者是自利的，要求获得激励；③社会伦理要求部署的模型要求不歧视某些个人或群体。

从上述不同的公平性动机出发，本文依据公平性的优化目标将现有的联邦学习中的公平性概念在**表1**中总结为3类，通过阐述它们的动机与利益相关者来说明它们各自如何服务于不同的联邦学习改善方向。其中，对于前两个动机，初始的联邦学习设置遵循的是过程公平：各方根据数据量、计算力、网络质量等条件获得相应的参与机会与聚合权重，最终获得同一全局模型[7]。但现实中各方客户端出于自身利益考虑，往往更看重结果公平：弱势客户端要

表1 联邦学习中的公平性目标分类及实现

公平性分类	分类标准	相关工作
表现均衡公平性	客户端选择	[8-15]
	权重分配	[16-25]
	个性化模型	[26-31]
贡献评估公平性	资源条件	[32-40]
	效用影响	[41-47]
	验证精度	[48-52]
模型公平性	不公平来源	[53-60]
	联邦学习优势	[61-64]
	公平与隐私	[18,65-73]

求全局模型在该方取得的效用不得显著低于其他方，强势客户端则倾向于获得与自己付出相匹配的模型效用或其他激励。前者要求模型在客户端之间表现均衡、不偏向某一数据方，笔者将相关公平指标总结为表现均衡公平性，并在后文具体阐述相关公平指标；后者则允许模型有表现差异的存在，但追求对各方贡献的准确评估并用来指导后续利益分配，笔者将相关公平指标总结为贡献评估公平性。针对第三个动机的实现则更加困难，需要跨越客户端间阻隔，保证模型对分散于所有客户端上的特定属性样本群体都没有歧视，笔者将其总结为模型公平性。

2 表现均衡公平性

虽然联邦学习已经证明了对保护用户数据隐私的效力[74]，但数据上的隔绝带来的一个挑战是统计异质性[57,75]——客户端之间的数据分布可能表现出显著差异，这种异质性决定了全局模型很难获得所有客户端上的最优性能，而只能趋近于集中训练模型的最优性能[76]。因此，现实中的训练结果往往受到数据强势方的主导，训练出的最终模型在数据强势方表现良好，在少部分数据提供者处则表现不佳，这种对少部分数据提供方的歧视促使了业界提出基于客户端间模型表现均衡的公平性。这种公平概念旨在通过测量模型在联邦学习客户端设备上的测试损失一致性程度来实现一致的模型效用，常用的衡量指标有测试损失（实际应用中可使用准确率、召回率等多种指标）的标准差与方差、基尼系数、Jain公平指数等。

（1）表现均衡公平性评价标准1——标准差

对于任意两个模型 w_1 与 w_2，比较二者在N个客户端上的测试损失的标准差，若有 $\mathrm{std}\left(\{L_n(w_1)\}_{n\in[N]}\right)<\mathrm{std}\left(\{L_n(w_2)\}_{n\in[N]}\right)$，则可认为模型 w_1 比 w_2 更公平。

（2）表现均衡公平性评价标准2——基尼系数

对于任意两个模型w_1与w_2，比较二者在N个客户端上的测试损失的基尼系数，若有$\mathrm{Gini}\left(\{L_n(w_1)\}_{n\in[N]}\right)<\mathrm{Gini}\left(\{L_n(w_2)\}_{n\in[N]}\right)$，则可认为模型 w_1 比 w_2 更公平，其中

$$\mathrm{Gini}\left(\{L_n(w_1)\}_{n\in[N]}\right)=\frac{\sum_{i=1}^{N}\sum_{N}^{j=1}|L_i(w_1)-L_j(w_1)|}{2N^2\mu},$$

$\mu=\frac{1}{N}\sum_{N}^{n=1}L_n(w_1)$。

（3）表现均衡公平性评价标准3——Jain公平指数

若将联邦学习中的全局模型作为一种资源，其在N个参与者本地数据集上的准确率分布组成向量$\boldsymbol{x}$，则其Jain公平指数（Jain's Index）可由$J(x)=\left(\sum_{N}^{i=1}x_i\right)^2/\left(N\sum_{N}^{i=1}x_i^2\right)$计算得到。

在明晰了评价标准之后，针对模型表现均衡公平性定义，如**图1**所示，已经有一些工作分别从联邦学习的3个环节入手加以解决：考虑客户端选择机制，关注掉队的客户端代表不足或从未代表的问题；考虑聚合更新环节，依照公平约束分配更新权重；考虑本地更新环节，训练个性化本地模型以提升全体客户端模型表现。本文在**图1**中根据时间关系与演进情况对相关工作进行了整理，并在**表2**中从公平指标、优势、局限性等因素逐个进行比较。

2.1 基于客户端选择机制的表现均衡公平性

在联邦学习的4个主要环节中，客户

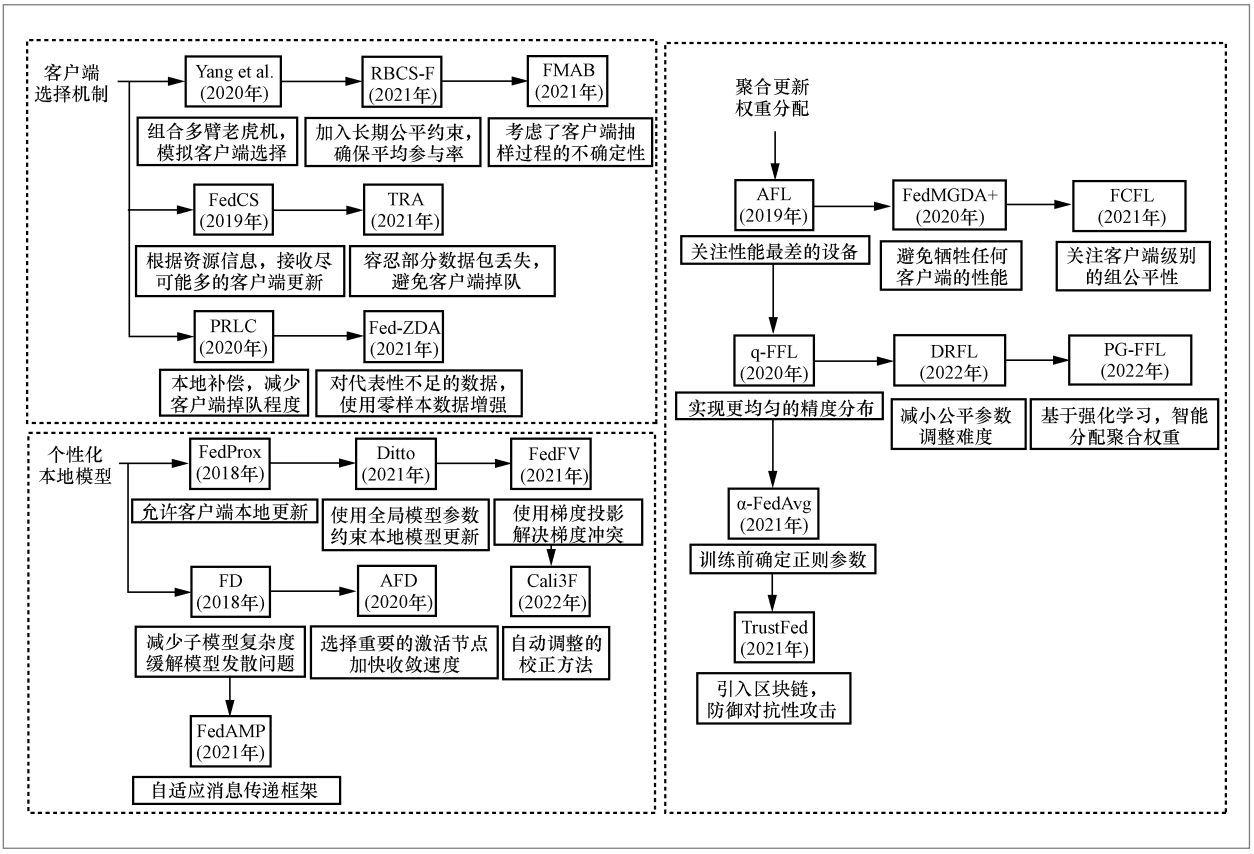

图1 基于客户端间表现均衡的公平性实现方法

端抽样环节决定了客户端在训练中的参与度，从而决定了最终模型对不同数据的偏向程度。现有的客户端选取过程优化方法通常使用基于阈值的方法来过滤掉不合格的客户端，例如针对传输速度、带宽、本地精度等做出限制[8]。但这些方法有些是为了保证总收益的最大化（例如提高全局模型准确性），而忽略了部分FL客户端的利益；有些是以最小化总交换时间为优化目标（如加快收敛速度），因此具有高计算能力和良好信道的设备有较高的优先级被选择[77]，这就导致条件较差的客户端被剥夺了参与协作训练的机会，也导致数据的多样性与模型的泛化性无法得到保证。此外，我们希望选取出的客户端能够最大化地代表整体数据分布，但实际的客户端选取过程中存在着3种不公平现象：过度代表；代表不足；忽略数据[4]。

为了减轻对计算能力较低或数据集较小的FL客户端的偏见，Yang等人[9]研究了采样约束的设计，以解决公平的客户端选择问题。他们将参与频率考虑在内，允许不经常被选择的客户更频繁地参与培训。这种方法通过让代表不足或从未代表过的客户参与FL来促进公平。其具体做法考虑了利用-探索（exploit-explore，EE）问题，将联邦学习的客户端选择问题描述为一个组合多臂老虎机（combinatorial multi-armed bandit，CMAB）问题，每只臂代表一个客户，超级臂代表所有FL客户的集合。超级臂的奖励是参与的单臂奖励的非线性组合。虽然经常被选中的客户被认为更值得信任，并获得更高的回报，但该方案也为不经常参与的客户提供了加入

表 2 客户端表现均衡方法的比较

参考文献	技术方案	公平指标	优势	局限性	帕累托最优
Yang et al.[9]	组合多臂老虎机	测试准确度	为弱势客户端提供参与机会	仅考虑参与频率	
RBCS-F[10]	上下文组合多臂老虎机	选择分布参数	实现了长期公平约束	牺牲训练效率	√
FMAB[11]	多臂老虎机	–	考虑客户端抽样过程的不确定性	在探测与通信时产生损耗	
FedCS[8]	估计等待时间指导客户选择	聚合更新客户端数	考虑了设备异质性问题	依赖诚实且有评估能力的客户端	
TRA[12]	容忍丢包损失	测试准确度方差	保证了一定丢包比例下的个性化和公平性	依赖诚实且有评估能力的客户端	
PRLC[14]	本地补偿	拉动操作数	有更好的扩展性	容易遭遇模型发散问题	√
Fed-ZDA[15]	本地补偿	测试准确度方差	零样本数据增强	耗费更多的算力与时间	
AFL[16]	最小化最大期望损失	验证准确度标准差	防止模型对任何特定客户的过度拟合	仅适用于小规模客户，缺乏灵活性	
FedMGDA+[17]	多目标优化	训练损失平均方差	不牺牲任何参与者性能	难以识别恶意客户端	√
FCFL[18]	多目标优化	人口均等、机会均等	考虑所有客户端优化目标	只关注客户端级别的群体公平性	√
q-FFL[19]	放大高损失客户端的损失	测试准确度方差	更均匀的精度分布	需要多次试验确定最佳权重分配参数且对异构数据效果差	
α-FedAvg[20]	提升较低准确度客户端的权重	Jain's 指数	可通过算法在训练之前确定参数α的取值	无法抵御膨胀损失攻击	
DRFL[21]	由客户端损失动态调整权重	准确度均匀程度	更方便地调整公平参数	没有解决公平参数对不同数据集的适应问题	
PG-FFL[22]	强化学习	基尼系数	自适应学习聚合权重	训练成本较高	
TrustFed[24]	维持信誉分数剔除异常设备	–	引入区块链抵御对抗性攻击	增加传输数据量	
FedProx[26]	允许客户端训练个性化模型	收敛稳定性	有助于减少系统异质性的负面影响	–	
Ditto[78]	多任务学习	测试准确度标准差	插件式设计方便改造现有模型	增加计算量	
FedFV[30]	使用梯度投影缓和梯度冲突	测试准确度方差	避免牺牲部分客户端模型准确性	过时的梯度估计可能导致模型发散	√
Cali3F[31]	梯度校正技术与更新共享	NDCG标准差	避免本地模型发散，提高推荐性能均匀性	在推荐性能上非最优	
FD[27]	随机丢弃子模型激活节点	–	降低了通信和本地计算成本	需要多次试验选取最佳暂退率	√
AFD[28]	维护激活分数图生成子模型	–	自动选取激活比例	–	√
FedAMP[29]	自适应消息传递框架	测试准确度	增量优化	效果依赖于深度神经网络提高了计算量	

FL的机会，以探索他们对FL模型训练的贡献。而在文献[10]中，作者提出的RBCS-F算法引入了一个长期的公平约束来实现公平的FL客户端选择，即每个客户端被选中的概率长期来看不能低于一个阈值。由于联邦学习的客户端选择阶段可以同时做多个选择，并且每次选择的报酬非固定而是服从线性的随机公式，所以可以将客户端选择问题抽象为一个上下文组合多臂老虎机问题（contextual combinational multi-arm bandit，C^2MAB）。由于时间耦合的调度问题基本不可能线下求解，作者提出了一个基于Lyapunov理论的优化框架，将原来的离线长期优化问题转化为在线优化问题，通过排队动力学方法优化FL客户的参与率。研究结果表明，由于更多的约束条件和更公平的FL客户选择策略使得更多客户端有机会参与训练，整体而言使用到的数据更多样，可以提高最终的准确性，但会导致训练效率上的牺牲。此外，针对联邦多臂老虎机中客户端抽样与臂抽样的区别，文献[11]提出了一个通用的联邦多臂老虎机框架，阐述了客户端抽样过程的不确定性。

尽管以上3项工作都考虑了优化参与频率以促进FL客户选择的公平性，但许多其他重要因素（例如数据质量、培训质量、等待时间等）并未被考虑在内。FedCS[8]考虑到了联邦学习的设备异质性问题[79]，认为一部分客户端可能拥有更多数据但处于严重不良的通信条件下，简单的设置等待时间阈值会导致网络带宽使用效率低下，并降低部分客户端的代表性。因此FedCS在初始客户端选择阶段要求随机数量的客户端告知运营商其资源信息（如无线信道状态、是否有可用CPU和GPU、与任务相关的数据大小等），而后运营商确定哪些客户端在一定期限内可以完成后续步骤（模型分发、更新及聚合），以估计分发与上传更新步骤所需的时间。在此基础上以接受尽可能多的客户端更新为客户端选择的目标，达到综合考虑客户端机会公平和结果公平的目的。此外，考虑到网速较慢的客户频繁重传可能会导致FL模型训练的额外延迟，使得来自通信渠道较差的客户的模型更新不太可能被聚合到最终的模型中，从而导致模型偏差，文献[12]提出了一个可容损的FL框架——ThrowRightAway （TRA）。其主要思想是，网络限制的挑战在某些情况下可能被夸大了，丢包并不总是有害的[13]。基于这个假设，TRA通过有意忽略一些数据包丢失来接纳低带宽设备的数据上传，加速FL训练。在选择开始时，每个客户端报告自身的网络状况，服务器根据报告将候选客户分为充足组和不足组，然后不考虑所属组地随机选择一些客户端，并发送全局模型等待它们的更新。服务器检测到丢包时，如果客户端属于充足组，则发送重传通知，否则直接将丢失的数据设置为零，其余过程皆遵循经典的联邦学习流程。客户端上传完成后，TRA会根据丢包记录重新计算聚合。结果表明，通过适当地与其他算法集成，TRA可以保证面对一定比例（10%~30%）的丢包时的个性化和公平性能。不过上述两种方法的有效性在很大程度上依赖于客户端的资源类型划分情况，其使用到的两种隐性假设（①FL客户端能够准确地评估自己的资源状况；②客户端总是诚实的）在实践中可能无法满足。

避免欠采样客户端代表不足的另一个思路是本地补偿。Wang等人[14]提出了一种带本地补偿的新方法PRLC（pulling reduction with local compensate），它基于联邦学习达成端到端的通信，主要思想是在每轮迭代中只有部分设备参与全局模型更新，不参与更新的设备通过PRLC在本地更新以减少与全局模型之间的差距。PRLC的收敛率被证明与强凸性和非凸性情况下的非压缩方法一样，并且有更好的拓

展性。Hao等人[15]则对代表性不足的数据采用零样本数据增强来减轻统计异质性。

2.2 基于客户端间权重分配的表现均衡公平性

传统的联邦学习聚合更新环节根据服务器数据量来分配权重，这导致了模型偏倚对数据量偏倚产生继承关系。基于此，部分研究者探索了调整不同代表性客户端的聚合权重来改进公平性的多种依据。文献[16]通过关注性能最差的设备提出基于min-max损失函数的AFL方法，防止模型对任何特定客户的过度拟合而牺牲其他客户的利益来实现公平概念。作者认为，在标准横向联邦学习中采用的客户分布，可能与每个客户的目标分布不一致。因此，AFL算法着眼于训练出一个权重分布，使得全局模型针对由各个客户端分布聚合而成的任一目标分布进行优化，以最小化在所有可能的数据-标签联合分布中造成的最大期望损失。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。但由于仅针对最差客户的性能进行优化，AFL仅适用于小规模的客户且缺乏灵活性。文献[17]同样关注性能最差的设备，但为了兼顾公平性和鲁棒性，提出了FedMGDA+方法，将多目标优化推广到联邦学习场景下，通过修改参与者梯度合并的权重来改进联邦模型公平性。FedMGDA+方法相比于联邦平均更关注联邦模型在全局损失的结果，因此会牺牲某个参与者的表现来提升全局结果，多目标优化更关心当前模型中所有参与者的结果，它使用帕累托稳定（Pareto-stationary）解决方案，为所有选定的客户找到一个共同的下降方向（意味着任何参与用户的目标函数都只能下降），以避免牺牲任何客户的性能。此外，FedMGDA+还引入了两种技术——梯度归一化和受AFL中Chebyshev方法启发的内置鲁棒性方法，以增强对膨胀损失攻击（inflated loss attack）的鲁棒性。然而，如何准确识别恶意客户端仍然是一个有待解决的问题。随后Cui等人[18]提出了改进的多目标优化（multi-objective optimization）框架FCFL。FCFL使用了一个考虑所有客户端优化目标的平滑代理最大值函数，而非仅考虑最差客户端，由此使客户端之间的性能更加一致。Li等人[19]同样对AFL的公平性表现进行了改进，提出了q-FFL方法来赋予损失较高的客户端数据更大的权重，实现了更均匀的精度分布。它将目标函数的加权平均更改为目标函数q+1次方项的加权平均，引入了一个新的参数q来重新衡量总损失，通过放大损失来增加损失高的客户端的惩罚。与AFL相比，q-FFL更灵活，其可以通过调优q来调整公平程度。当q-FFL将q设置为一个较大的值时，其性能与AFL类似。但该方法无法提前确定最佳的参数q值来达到公平性和有效性的平衡，这个q值在不同数据集之间的差距巨大，且算法在本地数据异构较强时较难收敛。

针对q-FFL无法提前确定最佳参数的问题，田家会等人[20]提出了α-FedAvg算法，引入Jain's指数和α-公平概念来研究FL模型公平性和有效性的平衡。该算法可以在保持FL模型整体性能不损失的情况下，有效减小各参与方准确率分布的方差，使准确率分布更均衡。与q-FFL相比，他们的方法更简单，并且可通过算法在训练之前确定参数α的取值，而不需要使用多个参数值训练获得全局模型后，验证模型性能再从中选择表现较好的参数值。Zhao等人[21]提出将q-FFL中的损失放大机制替换为简单的权重再分配机制，通过为损失高的客户端分配更大的权重来增加对这些客户端的惩罚，减小了q值的调整难度。Sun

等人[22]则提出PG-FFL方法，基于基尼系数定义了一种具有规模不变性特点的客户端粒度公平概念，并引入一个基于强化学习算法的公平性调整插件，可以自适应学习客户端聚合权重。与q-FFL相比，该方法的公平性插件可以用于各种算法，更具有普适性，但是由于强化学习优化较慢，训练成本很高，文献[23]则是通过经验风险最小化调整设备的权重，以实现灵活的公平性/准确性权衡。

上述的AFL及其改进方法都假设参与者是诚实的，对于来自客户端侧膨胀损失攻击的防御能力都不强，例如，如果客户端恶意夸大其损失，可能会导致全局模型的性能下降[4]。对于这一问题，Rehman等人[24]提出将区块链作为训练网络中分散的训练实体，提出一个完全去中心化的跨设备FL系统TrustFed，使用以太坊区块链和智能合约技术来实现去中心化，并维护诸多客户端与服务器的信誉分数，将异常的设备剔除出模型聚合更新阶段，防止恶意客户端通过发起对抗性攻击来串通和污染某些设备上的训练模型，导致不公平地训练出低质量的FL模型。文献[25]同样利用了区块链的去中心化、难以篡改性以及智能合约的特点，利用区块链的共识机制选择信用值最高的区块进行模型聚合，降低参数在传输过程中面临的安全风险。

2.3 基于个性化本地模型的表现均衡公平性

导致FL中模型性能不均匀的根本原因是数据的异质性，因此在客户端级别训练个性化模型优于训练全局模型[78]。Li等人[26]提出的FedProx允许每个客户端根据其可用资源执行部分训练。尽管这有助于减少系统异质性的负面影响，但大量的本地更新可能导致模型发散问题的产生，因此FedProx引入了一个由本地模型与全局模型之间距离平方项构成的约束项，通过鼓励局部更新趋向于全局模型来获取更高质量的局部更新，提高训练的稳定性。

另外，针对较复杂的模型在以较大的梯度进行本地训练时容易引发难以约束的发散的情况，可以尝试通过减少子模型复杂度来缓解。文献[27]提出了联邦暂退（federated dropout，FD）方法，根据每个FL客户端本地计算资源量的不同来设计不同的简化版子模型（对于全连接层，通常会放弃许多激活节点；对于卷积层，通常会放弃固定比例的滤波器，并且会进一步使用有损压缩来降低通信量），客户端更新的模型由FL服务器重新构建并聚合形成全局模型。更小的子模型减少了计算和通信成本，使能力较低的客户也可以通过训练一个自定义修剪的子模型来加入FL，在机会公平层面实现对异构客户的公平对待。该方法的实现难点在于确定合适的联邦暂退率过程中会产生额外的成本。为解决这一问题，文献[28]提出了自适应联邦暂退（adaptive federated dropout，AFD）法，它记录并维护一个激活分数图，用于选择一组重要的激活节点来生成最适合每个客户端的子模型，从而加快收敛速度并减少准确性损失[4]。

与FedDropout随机放弃激活节点以构

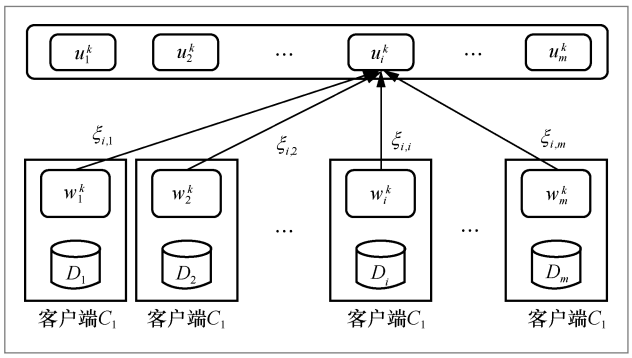

图 2 FedAMP 的消息传递机制

建较小子模型不同，文献[29]提出了如图2所示的自适应消息传递框架FedAMP，根据相似度为每个模型定制更新，训练过程基于增量优化方法。

同样为了防止模型聚合发散，Li等人[78]提出的Ditto则采用了多任务学习的思想，在以FedAvg方法训练的全局模型之外，每轮每个被选中的客户端都会使用本地数据，并且使用与FedProx类似的本地模型与全局模型的距离平方作为近端正则项来额外训练一套本地模型，在保证了模型不发散的同时，能够获得较具代表性的个性化模型。由于可以附加在FedAvg方法上，Ditto可以较为方便地应用于现有模型的改造。并且由于直接使用本地模型进行推理，Ditto能够有效抵御恶意客户端对全局模型造成的污染问题，增强了方法的鲁棒性，其弊端是同时更新全局与众多本地模型带来了计算开销上的增加。

Ditto的本地模型更新方法在梯度层面上可以视作为本地模型附加一个指向全局模型上一轮状态的辅助更新梯度，与基于目标函数的方法相比，基于梯度的方法仍处于发展的早期阶段。Wang等人[30]提出的FedFV从另一个角度实现了FL中的公平性。作者认为，全局模型可以牺牲一些客户的模型准确性，以提高那些梯度差异很大的客户的性能（这种情况往往导致FL训练中的不公平）。FedFV的梯度投影聚合机制如图3所示，FedFV使用了梯度投影的方法在进行梯度平均之前减轻各客户端梯度之间的潜在冲突。首先使用余弦相似度来检测梯度冲突，然后修改梯度的方向和大小来迭代地消除这种冲突。实验证明，FedFV能有效地缓解梯度冲突问题并收敛到Pareto平稳解。然而，FedFV基于历史梯度的梯度估计方法过于简单，在梯度变化较剧烈的时候可能会导致模型发散。Zhu等人[31]提出的Cali3F联邦推荐模型使用指向全局模型参数的、大小正比于本地更新大小的校正梯度，能够自适应地调整校正力度，在降低客户端间推荐性能标准差的同时防止个性化模型发散，并利用了基于聚类的参数更新共享方法实现快速收敛。

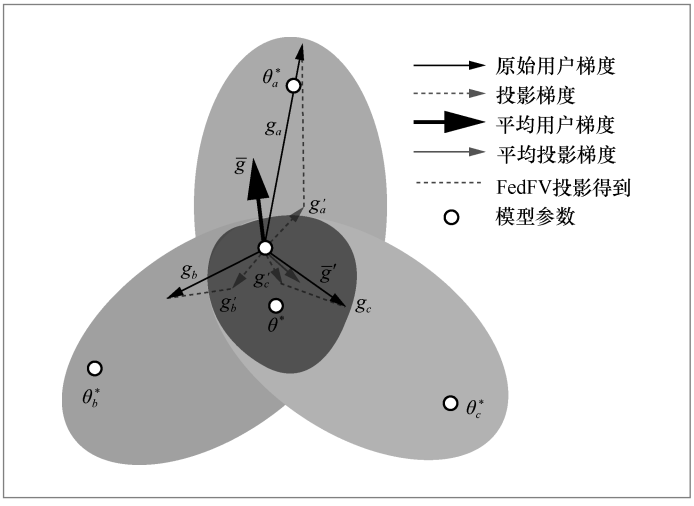

图 3 FedFV 的梯度投影聚合机制

3 贡献评估公平性

前文所述的表现均衡公平性更多地注重结果公平——要求最终模型在不同客户端上的表现尽可能相近，而不考虑搭便车问题[49]——不同贡献客户端收到相同的FL模型，这会引起高贡献客户端的不满，影响联邦学习的可持续性经营。要想解决这一问题，就需要在无法获取客户端本地数据的前提下，公平地衡量每个客户端对全局模型的贡献程度，形成多方认可的激励分配机制。

近年来已经提出了许多FL的激励机制[3,41]，但并非所有方法都专注于提高公平性，本节重点介绍考虑公平性的FL激励机制设计。本节对现有的FL贡献公平评估方法从图4所示的3个维度加以总结，分别为衡量依据、技术方案和分配机制。此外

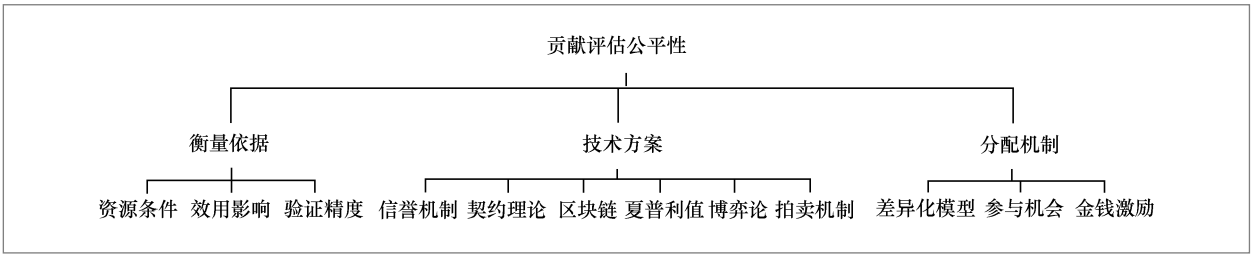

图 4 贡献评估公平性工作分类依据

对于实现相关公平目标过程中采用的技术以及在客户端之间分配的对象亦有归类，并且指明了各方法对诚实客户端的限制以及对膨胀损失攻击的鲁棒性，具体对比见表3。

3.1 基于客户端资源条件的贡献评估公平性

对于联邦学习而言，如何吸引更多参与方持续地参与到联邦学习的进程中，是长期成功的关键。在新的客户端加入联邦学习时，其为最终全局模型带来的真实增益无法立即得到，在这种情况下，客户的贡献值可以根据其资源条件信息进行预先评估，例如本地数据集的数量、质量和收集成本，以及联邦学习过程中约定付出的算力与流量份额。在文献[32]中，作者运用熵权法定义用户数据质量计算方法，结合用户数据的质量和数量计算综合得分作为用

表 3 贡献衡量方法的比较

参考文献	关键技术	衡量标准	分配结果	要求诚实客户端	抗膨胀攻击
[33]	分级别训练	本地数据条件	分级模型	√	
[34]	信誉机制	可信度、承诺水平	分级模型	√	
[35]	契约理论	本地数据	参与机会	√	
[36]	Stacklberg博弈	CPU功率	参与机会	√	
[37]	拍卖理论	多种资源	参与机会	√	
[38]	拍卖理论	学习质量	参与机会	√	
[39]	拍卖理论	相互评估	参与机会	√	√
[40]	VCG机制	计算成本、数据质量	金钱激励	√	
[41]	边际损失	特征重要性	–		√
[42,80]	边际损失	边际损失	–		√
[43,45-47]	夏普利值	梯度信息	–		√
[48]	采样测试方法	采样数据大小	–	√	√
[49]	动态收益共享	期望损失、等待时间	金钱激励	√	
[50]	信誉机制	验证数据集上准确度	–		√
[51]	信誉机制	本地计算时间和数据集大小	–		√
[52]	信誉机制	验证数据集上准确度	–		√

户的贡献值，并赋予相应的聚合权重。

为了向不同资源条件的客户端提供差异化的联邦模型，Zhang等人[33]提出了分层公平联邦学习（HFFL）框架，通过为贡献较多的客户端提供更高质量的数据和更高质量的模型更新来确保客户端之间的公平性。HFFL首先根据客户端自我报告的数据特征将客户端分为不同的级别，随后为每个级别训练一个FL模型。在训练一个较低级别的模型时，较高级别的客户端只提供与较低级别的客户端相同数量的数据；在协作训练更高级别的FL模型时，则要求较低级别的客户端提供其所有本地数据。因此，更高级别的客户端会收到性能更好的FL模型。然而，HFFL也存在一些缺点。首先，因为在划分级别时不仅考虑数据的数量，对于质量与成本也需要做出考量，所以同一级别的客户端可能拥有不同数量的数据；其次，客户端的数据特征是通过它们自我报告的信息获得的，这导致该方法容易受到虚假报告的影响[49]。

不同于为每个级别训练一个模型的HFFL，文献[34]提出了一种分散的公平和隐私保护深度学习（FPPDL）框架，其中每个参与者都会基于自己的贡献收到最终全局模型的不同等级变体模型。FPPDL通过任意两个参与者之间的相互评估来维护本地的信誉体系，参与者可以使用它们的交易点从其他参与者那里下载梯度。如此，每个参与者在没有协作的情况下都可以获得与它们的独立模型相比改进的本地模型，并且每个参与者获得的改进与其相应的贡献成正比。

在文献[35]中，作者基于契约理论[81]构建了客户端贡献衡量方案。服务器设计契约项目并将其广播给数据所有者，每个项目都包含奖励和有关客户端本地数据的信息，每个客户端选择最需要的项目作为参与FL的承诺。在文献[36]中，作者提出了一种基于Stackelberg博弈的FL贡献衡量方案。在这种博弈论设置下，客户的最佳策略是如实向FL任务发布者报告他们对一个CPU功率单位的期望价格。自我报告的信息也用于基于拍卖的FL激励机制设计，因为这种机制允许数据所有者经常报告他们的成本。如**图5**所示，在文献[37]中，作者基于拍卖理论对客户端资源条件做出评估并

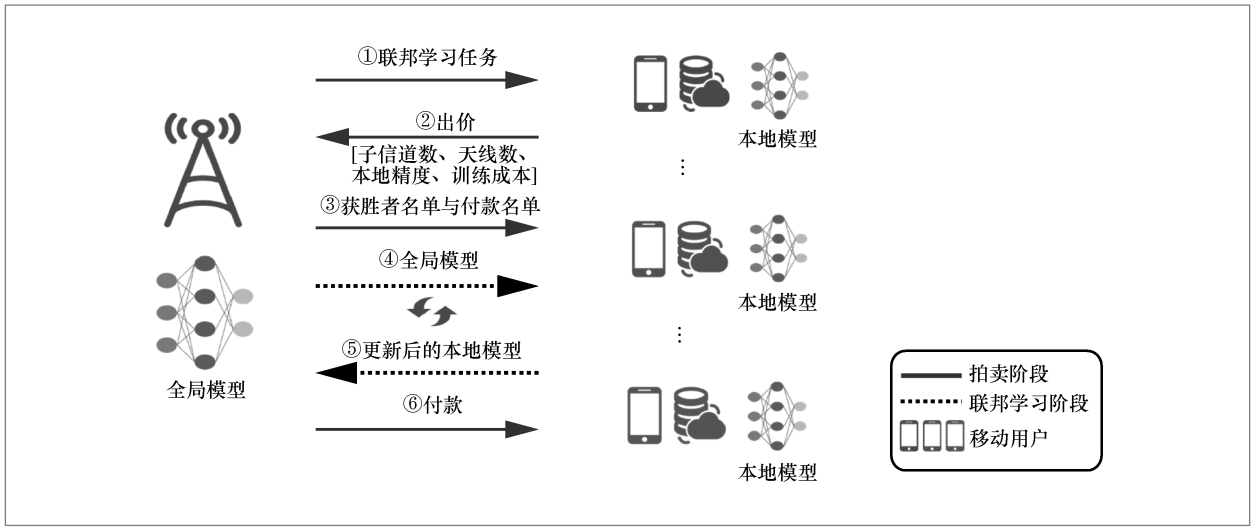

图 5 拍卖理论[37]

给出相应激励，首先以服务器作为购买方发起并广播一个FL任务，当每个移动用户收到任务信息时，各自决定参与模型训练所需的资源量并向服务器提交一个投标，其中包括所需的资源量、本地精度和相应的能源成本。之后由服务器使用投标信息来衡量每个客户的潜在贡献，确定获胜者参与本轮更新，并为获胜的移动用户进行结算。

Deng等人[38]提出了一种质量感知拍卖方案，将获胜者选择问题描述为一个NP难的学习质量最大化问题，并基于迈尔森定理设计了一个贪婪算法来执行实时任务分配和报酬分配。Zeng等人[39]进一步扩展了多维采购拍卖理论，提出了基于每个客户端的出价和资源质量的评分函数，在激励更多高质量数据所有者加入FL的同时最大限度地降低总成本。在投标阶段，客户端资源的质量向量（即本地数据、计算能力、带宽、CPU周期等）经过本地差分隐私保护[82]，并且通过应用相同的评分规则，客户端可以确定他们是否受到公平对待，这可以保证客户端的参与积极性。

除了准确评估每个客户的贡献之外，公平的激励机制还需要确保每个客户根据其对FL模型的贡献得到公平的报酬。Cong等人[40]提出了基于VCG机制的Fair-VCG（FVCG）。FVCG激励数据所有者如实报告他们的成本和数据质量。然后，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。

需要注意的是，前述的贡献评估方法均假设客户有能力并值得信赖，因此他们可以可靠地评估自己的情况并如实报告信息。然而在实践中，这种假设可能不成立，在存在行为不端的FL参与者时，模型训练性能会下降。此外，在发送资源条件信息阶段需要在客户端与服务器之间使用安全多方计算等手段进行加密计算，加强隐私保护，避免泄露敏感信息。

3.2 基于对全局模型效用影响的贡献评估公平性

基于客户端资源条件直接估计贡献虽然计算简单，可操作性强，但可能无法准确反映来自不同客户端的数据对全局模型的真实影响，因为某客户端的资源条件与该客户端对最终全局模型的效用不一定都是成正比的。

为了更精准地探知这种影响，一部分学者提出了基于效用博弈的FL贡献评估方法，此类方法与利润分享计划密切相关[80]，即将参与者产生的效用映射到相应奖励的规则。FL中最常用的利润分享方案是边际损失方案：参与者的收益等于参与者离开团队时损失的效用。Wang等人[41]采用边际损失法来衡量HFL中各方的贡献，通过每次删除某一方提供的实例，重新训练模型，计算新模型与原始模型之间预测结果的差异，并使用这个差异措施来决定该方的贡献。作者提出了近似算法来高效测量单个客户端数据损失对于全局模型效用的影响。Nishio等人[42]同样采用边际损失方案来评估每个客户端在单个FL训练过程中的贡献，以减少通信和计算开销。简单的边际损失方案适用于公平地评估给定客户在协作训练FL模型的一组给定客户中的贡献，但这是一个相对评估（即取决于其他参与的客户贡献了多少），并不反映客户本地数据的实际价值，因此夏普利值已被用来解决这个缺点。

近年来，基于夏普利值（Shapley value，SV）的FL贡献评估方法引起了广泛的研究和关注。SV是一种基于边际贡献的方案，于1953年作为合作博弈论中的解决方案概念引入。考虑具有数据集$D_1, D_2, \cdots, D_n$

的n个客户端，机器学习算法$\mathcal{A}$和标准测试集T，D_S是一个多重集（multi-set），其中$S \subseteq N = \{1,2,\cdots,n\}$。通过算法$\mathcal{A}$在$D_s$上训练得到的模型记为$M_s$，模型$M$在标准测试集$T$上的评估表现记为$U(M,T)$，则用于计算客户端$i$的贡献的夏普利值$\phi_i$表示为：

$$\phi_i = C\sum_{S\subseteq N\setminus\{i\}} \frac{U\left(M_{S\cup\{i\}}\right)-U\left(M_S\right)}{\binom{n-1}{|S|}} \quad (1)$$

通过对不包含i的所有D子集的边际贡献的总和进行平均，SV反映了i对FL模型的贡献，因为它仅是其本地数据的结果，而不管其加入联邦的顺序如何，SV可以产生更公平的客户贡献评估。然而，计算SV的计算复杂度是指数级的$O(2^n)$，随着参与者的数量或特征数量的增加，相关方法的计算成本急剧增加[83]。为了提高SV计算的效率，传统的机器学习中已经提出了许多启发式方法，如截断的TMC-Shapley（Monte Carlo Shapley）方法和Gradient Shapley方法。受这些方法的启发，基于SV的FL客户贡献评估方法正在兴起。

Song等人[43]提出了两种基于梯度的SV方法：单轮重建（OR）和多轮重建（MR）。这两种方法都从FL客户端收集梯度更新以重建FL模型，而不是使用不同的客户端子集进行重新训练。OR收集所有训练轮次的所有梯度更新，然后为最后一轮中的所有子集重建模型。OR在最后一轮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，然后将它们聚合以计算最终基于SV的贡献值。Wei等人[44]扩展了MR以提出截断多轮方法（TMR）。TMR在两个方面改进了MR。首先，它为具有更高准确度值的训练轮分配更高的权重。其次，TMR消除了不必要的模型重建以提高效率。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，但是，我们仍然需要在每一轮训练中评估不同客户子集的子模型。为了进一步降低计算成本，受文献[46]的启发，Wang等人[45]提出了两种有效的近似方法：基于置换采样的近似；基于组测试的近似。这些方法旨在提高每轮SV计算的效率。文献[47]则采用了比例因子近似方法获得了计算效率上的提升，但它没有解决在不需要额外代价的前提下向多个联邦贡献同一数据集的问题。

纵向联邦学习（VFL）[84]参与者的数据集在特征空间中几乎没有重叠，但在样本空间中有很大的重叠，这对贡献评估提出了新的挑战。在文献[41]中，作者利用了Shapley值来计算VFL中的特征重要性。由于直接使用SV来评估每个预测可以揭示潜在的敏感特征，因此作者建议对特征组执行SV计算，而不是对每个单独的特征进行计算。然而，这种方法在计算上仍然成本很高，因为计算成本随着训练数据的增加呈指数增长。

3.3 基于少部分数据上验证精度的贡献评估公平性

夏普利值等基于全体数据的反事实贡献评价手段存在计算成本高、估计精度不高等缺点，限制了其可扩展性。因此，基于少部分数据验证精度的贡献评估方法已经被用于替代反事实方法。

文献[48]提出了FedCCEA，它通过构造一个带有采样数据大小的准确性近似模型（accuracy approximation model，AAM）来学习每个客户端的数据质量。该方法通过使用采样数据的大小，稳健而有效地逼近客户的贡献，并通过设置用于FL

模型训练的期望的本地数据大小，允许客户的部分参与。FedCCEA由一个模拟器和一个评估器组成。模拟器通过运行单历元FL分类任务获得AAM的输入（即采样数据大小）和目标（即全向精度）。然后，评估器利用存储的输入和目标，对AAM中的权重向量进行优化。模型收敛后，提取第一层的共享权值，学习数据大小对每个客户端的重要性。然而，由于AAM是建立在一个非常简单的神经网络架构之上的，所以FedCCEA目前仅限于简单的FL任务，这使得它不太适合实际应用。

上述FL激励机制隐含的假设激励预算已经预先确定。在FL模型训练时，存在激励预算不可用的情况。相反，参与者希望在以后可以获得FL模型产生的收入奖励[57]。为了解释这种情况，Yu等人[49]定义了除了贡献公平之外对FL的长期可持续运行很重要的两个公平标准概念：期望损失分配公平和期望公平。期望损失分配的公平性要求基于客户获得奖励的等待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来补偿早期的参与者。这导致客户的贡献与他们迄今为止获得的奖励之间出现暂时的不匹配。为了克服这个问题，作者提出了一种动态收益共享方案联邦学习激励器FLI——通过最大化集体效用和最小化数据所有者的期望损失和等待时间之间的不平等，动态地将奖励分配给客户。它确保提供更多高质量数据并等待更长时间获得全额回报的客户将在后续轮次中获得更多收入。在客户奖励的逐步支付过程中，期望公平的概念用于确保客户的期望损失值尽可能公平地减少，以代替他们的贡献。

若是引入信誉机制来反映参与者的可靠性与贡献，则可以通过采样选出一个准确且平衡的验证数据集，基于客户端在特定任务上的表现来衡量客户的贡献。文献[50]在集中式FL设置下引入信誉机制，根据每个客户端在公共验证数据集上的准确度更新客户端的信誉分数。而后文献[51]与[52]将其扩展到了分布式FL设置下，并结合区块链技术防止恶意方篡改信誉分数。在信誉分数的计算规则方面，前者根据参与者贡献的数据量与恶意行为来衡量信誉分数，后者则根据客户端在每一轮训练中的本地模型更新梯度来衡量。

采样测试方法经常采用两个假设：①可信服务器/客户端；②拥有与其他参与者（或全局模型）更相似的本地模型的客户端被认为贡献更多。它们在实践中可能并不总是成立的[4]，联邦学习的参与方可能不诚实或具有恶意，其数据分布也存在异质性，且不同数据对于全局模型性能提高的价值不同。如果处理不当，这些因素会对贡献评估的公平性产生负面影响。

4 模型公平性

4.1 模型公平性和不公平性含义

不同于以客户端为利益主体的前两种公平性，全局模型公平性强调全局模型对分散于全部客户端上的某类用户的公平性。如果具有相似特征的实例从同一模型接收到不同模式的结果，那么这违反了个人公平的标准[85]；如果某些敏感群体（特定种族、性别等）收到不同模式的结果，如不同的假阴性率，这可能会违反群体公平标准[86]，关于模型公平性的更详细的相关定义可参见文献[85,87]。由于无法跨越客户端之间的阻隔统计特定类别用户收到的模型推理结果，联邦学习对模型公平性研究提出了新的严峻挑战，但同时也提供了

许多机会。

4.2 模型不公平性来源

统计异质性导致的准确性损失不仅可能在客户端之间造成不平等，还可能跨越客户端导致某类用户遭受到不公平对待。具有较不常见类别样本数据的客户端往往会收到更差的模型性能，这可能是由于灾难性遗忘导致的，来自亚群之外的客户端倾向于忘记在他们自己的数据中找不到的特征，并且在聚合过程中，当模型权重被平均时，代表性不足的客户端学习到的特征可能会被淹没[15]。在现实世界中，这些特征可能代表性别[53]、年龄[54]、种族[55]、使用语言[56]、外貌特征、患病情况等。无法应对统计异质性导致了潜在的不公平算法。由于联邦学习的设计要求数据的私有性与本地存储，隔绝了客户端之间的数据交换，客户端无法访问未在己方数据中得到较好表示的群体的数据，导致用户群体在客户端上的不均衡分布演变为用户群体间的不公平。此外，联邦学习中使用的数据访问过程可能引入的数据集移位和非独立性，也存在引入偏差的风险[57]。

数据生成过程中的偏差也会导致从这些数据中学习到的模型不公平[58-59]，如何识别或减轻数据生成过程中的偏差是联邦学习研究和更广泛的ML研究的关键问题。先前的有限研究已经提出了在联邦环境中识别和纠正已收集数据中的偏见的方法（例如文献[60]中的对抗性方法），但需要在该领域进行进一步研究。将事后公平校正应用于从可能有偏差的训练数据中学习的模型的方法也是未来工作中一个有价值的方向[57]，例如通过存储一小部分均衡样本作为缓冲区以供模型微调，防止灾难性遗忘。

在集中式训练公平模型方面已经涌现了大量工作[86,88]，但这些方法中的大多数假设有一个统一的可用训练数据集，这侵犯了数据隐私，因此在现实场景中常常陷入不可用的境地。利益相关者可能由于隐私保护法规而无法与其他方共享其原始数据。例如出于保护患者隐私的目的，各个医疗机构只被允许使用自己的电子健康记录或临床图像，而不是跨机构汇集数据和模型[61-62]，而训练数据缺乏代表性和多样性已被证明会导致模型性能不佳[57]，如遗传疾病模型[63]和图像分类模型[64]。通过有效的分散训练协议、模型结果的隐私和不可识别性保证，联邦学习提供了一个通过分布式训练利用多样化数据集的机会。更重要的是，联邦学习可以通过跨孤岛联合可能与敏感属性相关的数据来提高全局模型的公平性。

4.3 减少模型不公平性典型方法

因为联邦学习最有可能部署在需要隐私和公平性的敏感数据环境中，所以检视FL如何能够解决现有的关于机器学习公平性的担忧，以及FL是否会带来新的公平性问题非常重要[57]。联邦学习的模型公平性研究面临着如下困难[89]。

- 准确评估模型的公平性需要访问各方的数据，但联邦学习禁止访问各方的私人数据。
- 由于缺乏每个参与方的数据，在本地衡量模型公平性是不准确的，且在每个客户端本地应用公平约束可能引发冲突，导致较差的公平性能或无效的解决方案。
- 为训练找到合适的公平约束是困难的。部分约束需要访问所有数据；部分公平约束非平滑、不可微分，不适用于训练；部分公平约束以近似误差为代价转换为平滑可微。

现有的公平模型训练方法由于侵犯

数据隐私而无法直接扩展，因此现有的联邦学习训练方法大多数不考虑训练公平模型，一部分工作进行了尝试但无法同时实现数据隐私与公平[16]，例如针对未知的测试数据分布情况，文献[65]提出了AgnosticFair来为每个数据样本分配一个单独的权值，并包含一个不可知论的公平约束，以实现人口均等的公平概念。在数据迁移的情况下，该算法可以实现良好的准确性和公平性。然而，该方法需要先验知识来确定评估函数。这限制了它在系统环境不断变化的动态系统中的应用，有侵犯隐私的风险。

在不违反隐私政策的情况下，从分散的数据中学习公平模型的一种简单的方法是训练本地公平模型并集成得到最终结果，这比任何数据共享方案或联邦学习方法都更加私密，因为只有经过充分训练的模型进行了一次共享。但是如果数据是高度异构的，那么即使组合模型在本地是公平的，本地训练模型的集成模型也不会是公平的。Zeng等人[66]证明了简单地将联邦平均方法与公平模型训练方法共同使用，通过中央服务器定期计算本地训练模型参数的加权平均值，可能会通过频繁的通信找到一个全局更公平的模型，但这种方法是以牺牲隐私为代价的，并且会导致性能有较大的下降。基于此，他们修改了现有的联邦平均算法，使其能够有效地模仿集中式公平学习，并提出了基于去中心化数据的私有公平学习算法FedFB，适用于多种定义下的模型公平性。

同期工作中，文献[18]旨在训练一个同时满足多个定义于客户端本地数据之上的公平性的联邦公平模型，但不能保证全局级别的群体公平性；文献[65,67-68]通过非常频繁地为每次本地更新而非每轮本地训练交换信息来模仿集中式公平学习设置，后两者仅适用于人口均等定义下的模型公平性；文献[69]将联邦平均算法与重加权相结合，使联邦公平模型的性能得到了一定程度的改善，但仍低于FedFB。此外，最近的证据表明，个性化学习可能会对敏感的子群体产生不同的影响[70-72]，缓解隐私和公平之间紧张关系的潜在解决方案可能是应用个性化和混合差分隐私——部分用户贡献的数据要求较少的隐私保证[58]。

5 未来研究方向

（1）数据分级工作

基本的联邦学习方法往往简单地假设客户端将现有数据全部贡献出来，这点在前期协商阶段会带来顾虑：在无法预知己方数据与其他参与方数据的质量差异之前，数据实体往往倾向于做出保守决策。因此亟待发展数据质量评估方法，对持有数据进行分级，督促参与者付出与等级相符的数据。之后可尝试结合课程学习[90]（curriculum learning）或持续学习（continual learning）技术，构建循序渐进的联邦学习策略，允许参与方在训练过程中根据当前收益情况决定是否继续参与联邦训练以及是否分阶段追加投入数据。

（2）针对恶意客户端的反制措施

出于隐私保护的要求，当前诸多公平性措施的衡量对象出于客户端的自我报告数据。特别是许多方法都隐性地包含了诚实客户端或者诚实可信第三方假设，极易受到膨胀损失攻击或串谋攻击，使恶意客户端获得与其数据质量及成本不匹配的模型收益或金钱收益。通过将区块链、零知识证明等技术应用于贡献衡量机制中，可实现在不刺探客户端隐私信息前提下的报告信息真实性保证。

（3）联邦遗忘学习

当联邦学习公平性难以实现或力度无法达到参与者要求时，需要合理的退出机制保证联邦学习客户端的“反悔权”，即在不重新训练的前提下保证某参与方数据的完全退出，从当前模型中剔除该方用户的隐知识。当前已经涌现了一些保证小部分数据安全退出的遗忘学习方法[91]，但大批量的数据退出机制仍然值得积极探索。

（4）服务器之间的市场化竞争

现有公平性方法绝大多数仅考虑多个客户端之间的合作与竞争，但在市场化经济的背景下，可能出现多个聚合服务提供商（即联邦学习中心服务器），这会给各方成本带来变化，如服务器与客户端间的不同通信条件可能导致客户端训练成本的逆转。因此，可以尝试探索多个服务器场景下的竞价机制。

（5）对模型公平性的更好实现

现有的模型公平性实现方法大多假设有一个统一的可用训练数据集，这侵犯了数据隐私，因此在联邦学习框架中不可用。并且由于准确评估模型的公平性需要访问各方的数据，但联邦学习保护数据隐私，禁止访问各方的私人数据。因此如何在不侵犯数据隐私的前提下，实现诸多公平模型训练方法的联邦化，以及如何在联邦学习环境下评价模型公平性仍是亟待攻克的难题。

6 结束语

本文首先对近年来联邦学习中的公平性问题及现存解决方案做了系统整理，并按照3种公平目标进行了详细的分类介绍，从联邦学习公平性的本质出发，介绍相关技术的应用现状，最后依据当前联邦学习落地过程中的难点与热点提出了相应的研究建议。联邦学习是一个非常有前景的研究领域，目前已经吸引了众多学者进行相关领域的研究，也取得了一系列重要研究成果。但联邦学习技术的发展还处于初级阶段，与公平机器学习方法的结合仍然有许多问题尚待解决。在未来工作中，需要继续研究联邦学习领域的公平性问题，加快研究和发展相关安全与隐私保护技术，促进联邦学习的进一步发展。

参考文献:

[1] 王健宗，孔令炜，黄章成，等. 联邦学习算法综述[J]. 大数据，2020，6(6)：64-82.
WANG J Z，KONG L W，HUANG Z C，et al. Research review of federated learning algorithms[J]. Big Data Research，2020，6(6)：64-82.

[2] ZENG R F，ZENG C，WANG X W，et al. A comprehensive survey of incentive mechanism for federated learning[J]. ArXiv e-Prints，2021，arXiv：2106.15406.

[3] ZHAN Y F，ZHANG J，HONG Z C，et al. A survey of incentive mechanism design for federated learning[J]. IEEE Transactions on Emerging Topics in Computing，2022，10(2)：1035-1044.

[4] SHI Y，YU H，LEUNG C. A survey of fairness-aware federated learning[J]. CoRR，2021：abs/2111.01872.

[5] CATON S，HAAS C. Fairness in machine learning：a survey[J]. ACM Computing Surveys，2023，arXiv：2010.04053.

[6] MEHRABI N，MORSTATTER F，SAXENA N，et al. A survey on bias and fairness in machine learning[J]. ACM Computing Surveys，2022，54(6)：1-35.

[7] KONG L W，TAO H T，WANG J Z，et al. Network coding for federated Learning Systems[M]//Neural Information Processing. Cham：Springer，2020：546-557.

[8] NISHIO T, YONETANI R. Client selection for federated learning with heterogeneous resources in mobile edge[C]//Proceedings of the ICC 2019 - 2019 IEEE International Conference on Communications. Piscataway: IEEE Press, 2019: 1-7.

[9] YANG M, WANG X M, ZHU H B, et al. Federated learning with class imbalance reduction[C]//Proceedings of the 2021 29th European Signal Processing Conference. Piscataway: IEEE Press, 2021: 2174-2178.

[10] HUANG T S, LIN W W, WU W T, et al. An efficiency-boosting client selection scheme for federated learning with fairness guarantee[J]. IEEE Transactions on Parallel and Distributed Systems, 2021, 32(7): 1552-1564.

[11] SHI C, SHEN C. Federated multi-armed bandits[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, 33rd Conference on Innovative Applications of Artificial Intelligence, The 11th Symposium on Educational Advances in Artificial Intelligence. New York: AAAI Press, 2021: 9603-9611.

[12] ZHOU P, FANG P, HUI P. Loss tolerant federated learning[J]. CoRR, 2021: abs/2105.03591.

[13] XIA J C, ZENG G X, ZHANG J X, et al. Rethinking transport layer design for distributed machine learning[C]//Proceedings of the 3rd Asia-Pacific Workshop on Networking. New York: ACM, 2019: 22-28.

[14] WANG H Z, QU Z H, GUO S, et al. Intermittent pulling with local compensation for communication-efficient distributed learning[J]. IEEE Transactions on Emerging Topics in Computing, 2022, 10(2): 779-791.

[15] HAO W T, EL-KHAMY M, LEE J, et al. Towards fair federated learning with zero-shot data augmentation[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2021: 3305-3314.

[16] MOHRI M, SIVEK G, SURESH A T. Agnostic federated learning[C]//Proceedings of the 36th International Conference on Machine Learning. [S.l.:s.n.], 2019: 4615-4625.

[17] HU Z, SHALOUDEGI K, ZHANG G, et al. FedMGDA+: federated learning meets multi-objective optimization[J]. CoRR, 2020: abs/2006.11489.

[18] CUI S, PAN W, LIANG J, et al. Addressing algorithmic disparity and performance inconsistency in federated learning[C]//Proceedings of Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021. [S.l.:s.n.], 2021: 26091-26102.

[19] LI T, SANJABI M, BEIRAMI A, et al. Fair resource allocation in federated learning[EB]. arXiv preprint, 2019, arXiv: 1905.10497.

[20] 田家会, 吕锡香, 邹仁朋, 等. 一种联邦学习中的公平资源分配方案[J]. 计算机研究与发展, 2022, 59(6): 1240-1254.
TIAN J H, LÜ X X, ZOU R P, et al. A fair resource allocation scheme in federated learning[J]. Journal of Computer Research and Development, 2022, 59(6): 1240-1254.

[21] ZHAO Z Y, JOSHI G. A dynamic reweighting strategy for fair federated learning[C]//Proceedings of the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2022: 8772-8776.

[22] SUN Y Q, SI S J, WANG J Z, et al. A fair federated learning framework with reinforcement learning[C]//Proceedings of the 2022 International Joint Conference on Neural Networks. Piscataway: IEEE

Press, 2022: 1-8.

[23] LI T, BEIRAMI A, SANJABI M, et al. Tilted Empirical Risk Minimization[C]// Proceedings of the 9th International Conference on Learning Representations. [S.l.:s.n.], 2021.

[24] REHMAN M H U, DIRIR A M, SALAH K, et al. TrustFed: a framework for fair and trustworthy cross-device federated learning in IIoT[J]. IEEE Transactions on Industrial Informatics, 2021, 17(12): 8485-8494.

[25] 陈乃月, 金一, 李浥东, 等. 基于区块链的公平性联邦学习模型[J]. 计算机工程, 2022, 48(6): 33-41.
CHEN N Y, JIN Y, LI Y D, et al. Federated learning model with fairness based on blockchain[J]. Computer Engineering, 2022, 48(6): 33-41.

[26] LI T, SAHU A K, ZAHEER M, et al. Federated optimization in heterogeneous networks [EB]. arXiv preprint, 2018, arXiv: 1812.06127.

[27] CALDAS S, KONE\VCNY J, MCMAHAN H B, et al. Expanding the reach of federated learning by reducing client resource requirements[J]. ArXiv e-Prints, 2018: arXiv: 1812.07210.

[28] BOUACIDA N, HOU J H, ZANG H, et al. Adaptive federated dropout: improving communication efficiency and generalization for federated learning[C]// Proceedings of the IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops. Piscataway: IEEE Press, 2021: 1-6.

[29] HUANG Y, CHU L, ZHOU Z, et al. Personalized cross-silo federated learning on non-IID data[C]//Proceedings of 35th AAAI Conference on Artificial Intelligence, AAAI 2021, 33rd Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The 11th Symposium on Educational Advances in Artificial Intelligence. New York: AAAI Press, 2021: 7865-7873.

[30] WANG Z, FAN X L, QI J Z, et al. Federated learning with fair averaging[C]// Proceedings of the Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2021: 1615-1623.

[31] ZHU Z T, SI S J, WANG J Z, et al. Cali3F: calibrated fast fair federated recommendation system[C]//Proceedings of the 2022 International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2022: 1-8.

[32] 杨秀清, 彭长根, 刘海, 等. 基于数据质量评估的公平联邦学习方案[J]. 计算机与数字工程, 2022, 50(6): 1278-1285.
YANG X Q, PENG C G, LIU H, et al. Fair federated learning based on data quality evaluation[J]. Computer & Digital Engineering, 2022, 50(6): 1278-1285.

[33] ZHANG J F, LI C, ROBLES-KELLY A, et al. Hierarchically fair federated learning[EB]. arXiv preprint, 2020, arXiv: 2004.10386.

[34] LYU L J, YU J S, NANDAKUMAR K, et al. Towards fair and privacy-preserving federated deep models[J]. IEEE Transactions on Parallel and Distributed Systems, 2020, 31(11): 2524-2541.

[35] KANG J W, XIONG Z H, NIYATO D, et al. Incentive design for efficient federated learning in mobile networks: a contract theory approach[C]//Proceedings of the 2019 IEEE VTS Asia Pacific Wireless Communications Symposium. Piscataway: IEEE Press, 2019: 1-5.

[36] SARIKAYA Y, ERCETIN O. Motivating workers in federated learning: a stackelberg game perspective[J]. IEEE Networking Letters, 2020, 2(1): 23-27.

[37] THI LE T H, TRAN N H, TUN Y K, et al. An incentive mechanism for federated learning in wireless cellular networks: an

auction approach[J]. IEEE Transactions on Wireless Communications, 2021, 20(8): 4874–4887.

[38] DENG Y H, LYU F, REN J, et al. FAIR: quality-aware federated learning with precise user incentive and model aggregation[C]//Proceedings of the IEEE INFOCOM 2021 – IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2021: 1–10.

[39] ZENG R F, ZHANG S X, WANG J Q, et al. FMore: an incentive scheme of multi-dimensional auction for federated learning in MEC[C]//Proceedings of the 2020 IEEE 40th International Conference on Distributed Computing Systems. Piscataway: IEEE Press, 2020: 278–288.

[40] CONG M S, YU H, WENG X, et al. A VCG-based fair incentive mechanism for federated learning[EB]. arXiv preprint, 2020, arXiv: 2008.06680.

[41] WANG G, DANG C X, ZHOU Z Y. Measure contribution of participants in federated learning[C]//Proceedings of the 2019 IEEE International Conference on Big Data. Piscataway: IEEE Press, 2019: 2597–2604.

[42] NISHIO T, SHINKUMA R, MANDAYAM N B. Estimation of individual device contributions for incentivizing federated learning[C]//Proceedings of the 2020 IEEE Globecom Workshops. Piscataway: IEEE Press, 2020: 1–6.

[43] SONG T S, TONG Y X, WEI S Y. Profit allocation for federated learning[C]//Proceedings of the 2019 IEEE International Conference on Big Data. Piscataway: IEEE Press, 2019: 2577–2586.

[44] WEI S Y, TONG Y X, ZHOU Z M, et al. Efficient and fair data valuation for horizontal federated learning[M]//YANG Q, FAN L, YU H. Federated Learning. Cham: Springer, 2020: 139–152.

[45] WANG T H, RAUSCH J, ZHANG C, et al. A principled approach to data valuation for federated learning[M]//YANG Q, FAN L, YU H. Federated Learning. Cham: Springer, 2020: 153–167.

[46] JIA R, DAO D, WANG B, et al. Towards efficient data valuation based on the Shapley value[C]//Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS. [S.l.:s.n.], 2019: 1167–1176.

[47] RADANOVIC G, FALTINGS B, JURCA R. Incentives for effort in crowdsourcing using the peer truth serum[J]. ACM Transactions on Intelligent Systems and Technology, 2016, 7(4): 1–28.

[48] SHYN S K, KIM D, KIM K. FedCCEA: a practical approach of client contribution evaluation for federated learning[EB]. arXiv preprint, 2021, arXiv: 2106.02310.

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

[50] LYU L J, XU X Y, WANG Q, et al. Collaborative fairness in federated learning[M]//Lecture notes in computer science. Cham: Springer, 2020: 189–204.

[51] KANG J W, XIONG Z H, NIYATO D, et al. Reliable federated learning for mobile networks[J]. IEEE Wireless Communications, 2020, 27(2): 72–80.

[52] ZHANG J W, WU Y Z, PAN R. Incentive mechanism for horizontal federated learning based on reputation and reverse auction[C]//Proceedings of the Proceedings of the Web Conference 2021. New York: ACM, 2021: 947-956.

[53] ZHU Z W, HU X, CAVERLEE J. Fairness-aware tensor-based recommendation[C]//Proceedings of the Proceedings of the 27th ACM International Conference on

Information and Knowledge Management. New York: ACM, 2018: 1153-1162.

[54] DÍAZ M, JOHNSON I, LAZAR A, et al. Addressing age-related bias in sentiment analysis[C]//Proceedings of the Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2019: 6146-6150.

[55] KLARE B F, BURGE M J, KLONTZ J C, et al. Face recognition performance: role of demographic information[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(6): 1789-1801.

[56] GU J T, HASSAN H, DEVLIN J, et al. Universal neural machine translation for extremely low resource languages[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg: Association for Computational Linguistics, 2018: 344-354.

[57] KAIROUZ P, MCMAHAN H B, AVENT B, et al. Advances and open problems in federated learning[J]. Foundations and Trends® in Machine Learning, 2021, 14(1/2): 1-210.

[58] ECKHOUSE L, LUM K, CONTI-COOK C, et al. Layers of bias: a unified approach for understanding problems with risk assessment[J]. Criminal Justice and Behavior, 2019, 46(2): 185-209.

[59] RICHARDSON R, SCHULTZ J, CRAWFORD K. Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice[R]. New York University Law Review Online, 2019.

[60] LIAO J C, KAIROUZ P, HUANG C, et al. Learning generative adversarial representations under fairness and censoring constraints[EB]. arXiv preprint, 2019, arXiv: 1910.00411.

[61] BRISIMI T S, CHEN R D, MELA T, et al. Federated learning of predictive models from federated Electronic Health Records[J]. International Journal of Medical Informatics, 2018, 112: 59-67.

[62] CHANG K, BALACHANDAR N, LAM C, et al. Distributed deep learning networks among institutions for medical imaging[J]. Journal of the American Medical Informatics Association, 2018, 25(8): 945-954.

[63] MARTIN A R, KANAI M, KAMATANI Y, et al. Current clinical use of polygenic scores will risk exacerbating health disparities[J]. bioRxiv, 2019, doi: 10.1101/441261.

[64] BUOLAMWINI J, GEBRU T. Gender Shades: intersectional accuracy disparities in commercial gender classification[C]//Proceedings of Conference on Fairness, Accountability and Transparency. [S.l.:s.n.], 2018: 77-91.

[65] DU W, XU D P, WU X T, et al. Fairness-aware agnostic federated learning[M]// DEMENICONI C, DAVIDSON I, eds. Proceedings of the 2021 SIAM international conference on data mining. Philadelphia: Society for Industrial and Applied Mathematics, 2021: 181-189.

[66] ZENG Y C, CHEN H X, LEE K. Improving fairness via federated learning[EB]. arXiv preprint, 2021, arXiv: 2110.15545.

[67] RODRÍGUEZ-GÁLVEZ B, GRANQVIST F, VAN DALEN R, et al. Enforcing fairness in private federated learning via the modified method of differential multipliers[EB]. arXiv preprint, 2021, arXiv: 2109.08604.

[68] CHU L Y, WANG L J, DONG Y J, et al. FedFair: training fair models in cross-silo federated learning[EB]. arXiv preprint, 2021, arXiv: 2109.05662.

[69] EZZELDIN Y H, YAN S, HE C Y, et al. FairFed: enabling group fairness in federated learning[EB]. arXiv preprint,

2021, arXiv: 2110.00857.

[70] CUMMINGS R, GUPTA V, KIMPARA D, et al. On the compatibility of privacy and fairness[C]//Proceedings of the UMAP'19 Adjunct: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization. New York: ACM, 2019: 309-315.

[71] CUMMINGS R, KREHBIEL S, LAI K A, et al. Differential privacy for growing databases[C]//Proceedings of the Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 8878-8887.

[72] JAGIELSKI M, KEARNS M J, MAO J, et al. Differentially private fair learning[C]// Proceedings of the 36th International Conference on Machine Learning. [S.l.:s.n.], 2019: 3000-3008.

[73] AVENT B, KOROLOVA A, ZEBER D, et al. BLENDER: enabling local search with a hybrid differential privacy model [EB]. arXiv preprint, 2017, arXiv: 1705.00831.

[74] WANG J Z, HUANG Z C, KONG L W, et al. Modeling without sharing privacy: federated neural machine translation[M]// Web Information Systems Engineering - WISE 2021. Cham: Springer, 2021: 216-223.

[75] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[76] XUE A S, ZHU X H, WANG J Z, et al. Diversified point cloud classification using personalized federated learning[C]// Proceedings of the 2021 International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2021: 1-8.

[77] ZENG Q S, DU Y Q, HUANG K B, et al. Energy-efficient radio resource allocation for federated edge learning[C]//Proceedings of the 2020 IEEE International Conference on Communications Workshops. Piscataway: IEEE Press, 2020: 1-6.

[78] LI T, HU S Y, BEIRAMI A, et al. Ditto: fair and robust federated learning through personalization[EB]. arXiv preprint, 2020, arXiv: 2012.04221.

[79] 朱智韬, 司世景, 王健宗, 等. 联邦推荐系统综述[J]. 大数据, 2022, 8(4): 105-132.
ZHU Z T, SI S J, WANG J Z, et al. Survey on federated recommendation systems[J]. Big Data Research, 2022, 8(4): 105-132.

[80] GOLLAPUDI S, KOLLIAS K, PANIGRAHI D, et al. Profit sharing and efficiency in utility games[C]//Proceedings of 25th Annual European Symposium on Algorithms. [S.l.:s.n.], 2017: 1-14.

[81] BOLTON P, DEWATRIPONT M. Contract theory[M]. Cambridge: MIT Press, 2004.

[82] CORMODE G, JHA S, KULKARNI T, et al. Privacy at scale: local differential privacy in practice[C]//Proceedings of the Proceedings of the 2018 International Conference on Management of Data. New York: ACM, 2018: 1655-1658.

[83] ZHAO J, ZHU X H, WANG J Z, et al. Efficient client contribution evaluation for horizontal federated learning[C]// Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2021: 3060-3064.

[84] YANG Q, LIU Y, CHEN T J, et al. Federated machine learning: concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 12.

[85] DWORK C, HARDT M, PITASSI T, et al. Fairness through awareness[C]// Proceedings of the Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. New York: ACM, 2012: 214-226.

[86] ZAFAR M B, VALERA I, GOMEZ-

RODRIGUEZ M, et al. Fairness constraints: mechanisms for fair classification[C]//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. [S.l.:s.n.], 2017: 962-970.

[87] MITCHELL S, POTASH E, BAROCAS S. Prediction-based decisions and fairness: a catalogue of choices, assumptions, and definitions[EB]. arXiv preprint, 2018, arXiv: 1811.07867.

[88] ROH Y, LEE K, WHANG S E, et al. FairBatch: batch selection for model fairness[EB]. arXiv preprint, 2020, arXiv: 2012.01696.

[89] ZHOU Z R, CHU L Y, LIU C X, et al. Towards fair federated learning[C]// Proceedings of the Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York: ACM, 2021: 4100-4101.

[90] BENGIO Y, LOURADOUR J, COLLOBERT R, et al. Curriculum learning[C]//Proceedings of the Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM, 2009: 41-48.

[91] LIU G Y, MA X Q, YANG Y, et al. FedEraser: enabling efficient client-level data removal from federated learning models[C]//Proceedings of the 2021 IEEE/ACM 29th International Symposium on Quality of Service. Piscataway: IEEE Press, 2021: 1-10.

作者简介

朱智韬(1996-),男,中国科学技术大学硕士生,中国计算机学会会员,现任平安科技(深圳)有限公司算法工程师,主要研究方向为人工智能、联邦学习和推荐系统等。

司世景(1988-),男,英国帝国理工学院博士,深圳市海外高层次人才,美国杜克大学人工智能博士后,中国计算机学会会员,现任平安科技(深圳)有限公司资深算法研究员,中国科学技术大学研究生企业导师。至今累计发表机器学习、大数据和人工智能领域国际核心论文20余篇。

王健宗(1983-),男,博士,平安科技(深圳)有限公司副总工程师,资深人工智能总监。美国佛罗里达大学人工智能博士后,中国计算机学会(CCF)理事、杰出会员,CCF大数据专家委员会委员,曾任美国莱斯大学电子与计算机工程系研究员,主要研究方向为联邦学习、深度学习、云计算、物联网和元宇宙。

程宁（1981– ），男，中国科学院自动化研究所博士，中国科学院计算技术研究所博士后，现任平安科技（深圳）有限公司数据挖掘专家，清华大学、中国科学技术大学研究生企业导师。主持广东省自然科学基金项目一项，至今累计发表语音识别和语音合成领域国际核心论文50余篇。

孔令炜（1995– ），男，平安科技（深圳）有限公司联邦学习团队算法工程师，中国计算机学会会员，主要研究方向为联邦学习系统和安全通信等。

黄章成（1990– ），男，平安科技（深圳）有限公司联邦学习团队资深算法工程师，人工智能专家，中国计算机学会会员，主要研究方向为联邦学习、分布式计算及系统和加密通信等。

肖京（1972– ），男，博士，美国卡耐基梅隆大学博士，国家特聘专家。国家新一代普惠金融人工智能开放创新平台技术负责人、深圳市政协委员、深圳市决策咨询委员会委员，兼中国计算机学会深圳分部副主席、广东省人工智能与机器人学会副理事长、深圳市人工智能行业协会会长、深圳市人工智能学会副理事长，清华大学、上海交通大学、同济大学等客座教授。肖京长期从事人工智能与大数据分析挖掘相关领域研究，先后在爱普生美国研究院及美国微软公司担任高级研发管理职务，现任平安集团首席科学家，技术研究院院长，负责人工智能技术研发及在金融、医疗、智慧城市等领域的应用，带领团队树立了多项传统行业智能化经营的标杆。肖京已发表学术论文249篇，美国授权专利101项，中国发明专利155项，参与及承担国家级项目8项。

收稿日期： 2023-01-10

通信作者： 王健宗，jzwang@188.com

基金项目： 广东省重点领域研发计划“新一代人工智能”重大专项（No.2021B0101400003）

Foundation Item: The Key Research and Development Program of Guangdong Province (No.2021B0101400003)