

基于迭代协作学习框架的信誉医学参与方选择

陆 枫^{1,2,3,4} 李 炜^{5,6,7} 顾 琳^{1,2,3,4} 刘 帅^{1,2,3,4} 王润衡^{1,2,3,4} 任宇飞⁸ 戴小海^{1,2,3,4} 廖小飞^{1,2,3,4} 金 海^{1,2,3,4}

¹(大数据技术与系统国家地方联合工程研究中心(华中科技大学) 武汉 430074)

²(服务计算技术与系统教育部重点实验室(华中科技大学) 武汉 430074)

³(集群与网格计算湖北省重点实验室(华中科技大学) 武汉 430074)

⁴(华中科技大学计算机科学与技术学院 武汉 430074)

⁵(中澳能源信息学和需求响应技术联合研究中心(悉尼大学) 悉尼 2006)

⁶(分布式与高性能计算中心(悉尼大学) 悉尼 2006)

⁷(悉尼大学计算机学院 悉尼 2006)

⁸(华中科技大学同济医学院附属同济医院计算机中心 武汉 430030)

(lufeng@hust.edu.cn)

Selection of Reputable Medical Participants Based on an Iterative Collaborative Learning Framework

Lu Feng^{1,2,3,4}, Li Wei^{5,6,7}, Gu Lin^{1,2,3,4}, Liu Shuai^{1,2,3,4}, Wang Runheng^{1,2,3,4}, Ren Yufei⁸, Dai Xiaohai^{1,2,3,4}, Liao Xiaofei^{1,2,3,4}, and Jin Hai^{1,2,3,4}

¹(National Engineering Research Center for Big Data Technology and System (Huazhong University of Science and Technology), Wuhan 430074)

²(Services Computing Technology and System Lab (Huazhong University of Science and Technology), Ministry of Education, Wuhan 430074)

³(Cluster and Grid Computing Lab (Huazhong University of Science and Technology), Wuhan 430074)

⁴(School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074)

⁵(The Australia-China Joint Research Centre for Energy Informatics and Demand Response Technologies (The University of Sydney), Sydney 2006)

⁶(Centre for Distributed and High Performance Computing (The University of Sydney), Sydney 2006)

⁷(School of Computer Science, The University of Sydney, Sydney 2006)

⁸(Department of Computer Center, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030)

Abstract Collaborative learning technologies such as federated learning and swarming learning can fully use distributed data to deeply mine the knowledge contained in the data while protecting data privacy. It has a broad application prospect, especially in the medical and health field, where privacy practices and ethical constraints are emphasized. Collaborative efforts always require reliable participants. The performance of the global model in collaborative learning largely depends on participant selection. However, the existing studies need to pay more attention to the heterogeneity of medical participants' data. As a result, the performance of the global model, including stability, is difficult to be guaranteed. We propose to solve this problem from the perspective of reputation. Through iterative collaborative learning, reputation participants are selected as much as possible to obtain a stable and reliable

收稿日期: 2023-04-03; 修回日期: 2023-12-11

基金项目: 国家自然科学基金重点项目(62232012); 湖北医疗健康大数据分析平台与智能服务项目

This work was supported by the Key Program of the National Natural Science Foundation of China (62232012) and the Hubei Big Data Analysis Platform and Intelligent Service Project for Medical and Health.

通信作者: 戴小海(xhdai@hust.edu.cn)

high-performance global model in collaborative learning. We first propose an AI medical promise (AMP) to describe a medical institution's data quality and help form a good AI ecosystem in the medical field. Secondly, an iterative collaborative learning framework based on backward selection (colback-learning) is established. The backward selection method is used to iteratively calculate a well-performing and stable global model in polynomial time complexity to complete AMP calculation and accumulation in a single collaborative learning task. In calculating AMP, a scoring function that comprehensively considers global performance indicators is formulated to guide the training of the global model in the medical field. Finally, using real-world medical data to simulate various collaborative learning scenarios, we have shown that the colback-learning can select reliable participants to obtain a global model with good performance. The model's performance stability is 1.3 to 6 times higher than that of the state-of-the-art methods. The interpretability of the global model maintains a high consistency with centralized learning.

Key words collaborative learning; federated learning; participant selection; data contribution; block chain; neural networks

摘要 联邦学习和群智学习等协作学习技术,能够在保护数据隐私的条件下充分利用分布在各地的分布式数据深度挖掘数据中所蕴含的知识,拥有非常广阔的应用前景,尤其是在强调隐私惯例和道德约束的医疗健康领域.任何协作工作都需要选择可靠的参与方,协作学习中全局模型的性能在很大程度上取决于参与方的选择.然而,现有研究在选择参与方时都没有对不同机构医疗数据中存在的异质性加以直接关注,导致包含稳定性在内的全局模型的性能难以得到保障.提出了从信誉的角度尝试探索求解该问题.以迭代协作学习的方式尽可能挑选出具有良好信誉的参与方进行协作学习,以获得稳定可靠的高性能全局模型.首先,提出了一个描述医疗机构数据质量的AI信誉值指标AMP(AI medical promise),以帮助其在医疗领域中形成良好的AI生态.其次,建立了一个基于后向选择的迭代协作学习(colback-learning)框架.在单次协作学习任务中,利用后向选择方法以多项式时间复杂度迭代计算出性能良好且稳定的全局模型,完成AMP计算和积累.在AMP信誉值计算中,制定了一个综合考虑全局性能指标的评分函数,以针对医疗领域更有效地指导全局模型的训练.最后,使用真实医疗数据模拟多样化的协作学习场景.实验表明,colback-learning能够选择可靠参与方训练得到性能良好的全局模型,模型的性能稳定性比现有最好的参与方选择方法提高了1.3~6倍.全局模型的可解释性与集中式学习保持了较高的一致性.

关键词 协作学习;联邦学习;参与方选择;数据贡献;区块链;神经网络

中图法分类号 TP391

联邦学习^[1-4]和群智学习等协作学习(collaborative learning, CL)^[5]技术为人工智能(artificial intelligence, AI)在强调隐私惯例和道德约束的领域中提供了一种有效应用途径^[6-8].对电子病历系统和各类监护仪器得以广泛普及的医疗领域产生了极其重要的影响.因为它们使得人工智能方法可以在数据共享法规(例如,通用数据保护条例(general data protection regulation, GDPR)^[9])的严格限制下,对医疗机构中每天都产生的数以亿计的海量数据进行建模分析,有效支持医疗健康领域问题的探索 and 发现.

协作学习中全局模型的性能在很大程度上取决于参与方的选择.现有研究会基于一些模型指标来选择参与方^[5,10].比如Cho等人^[11]提出了一种偏向于选择局部损失较高的参与方的选择策略.Zhang等人^[12]利用模型损失来检验参与方局部模型的质量.Tahir

等人^[13]在每轮训练中选取训练时间相似的参与方,不同轮次间随机选取网络或计算条件不同的参与方,减少收敛时间,保证数据的异构性.

现有研究都尚未对不同机构医疗数据中天然存在的数据异质性问题加以明确关注.这些异质性源于不同机构患者群体的分布、环境、疾病的亚型、专家的不一致认知和标注,以及受时空影响的病程发展等.例如,门诊手术数据显示,不同年龄组的女性和男性在统计学上有显著差异^[14].在睡眠阶段监测分析中,不同评分者对非快速眼动睡眠阶段1(non rapid eye movement sleep stage 1, N1)和非快速眼动睡眠阶段3(non rapid eye movement sleep stage 3, N3)的标注的一致性低至63%和67%^[15].这些数据异质性不仅会影响模型性能,甚至会造成模型具有很高的偏倚性和随机性.比如当我们纳入3个客户端的1~5岁肺炎

患者,对其胸部X线影像用联邦学习进行建模分析时,模型达到了97%的准确率.而当添加另一个参与方后,全局模型的准确率降低了2个百分点, $F1$ 分数下降5个百分点.究其原因是第4个参与方纳入了成年患者.目前尚未发现有模型相关指标能够对医疗机构间复杂的数据异质性进行评估或提示.医学研究通常依据临床指南和实施标准等制定严格的研究对象纳入准则,以确保模型的“正确性”和可靠性.然而,由于数据异质性源于太多杂乱因素,在数据不可见条件下,即便采用共同严格的纳入准则,也很难避免.比如在数据不可见条件下,基本无法发现标注不一致现象.

本文提出从信誉的角度尝试探索求解该问题.尽可能选择具有良好信誉的参与方进行协作学习,以获得稳定可靠的高性能全局模型.通常而言,参与方的医疗数据质量在很大程度上受机构的长期整体质量水平影响.任何合作工作,都希望选择具有良好信誉的参与方.而长期可靠的信誉建立基础是参与方数据在单次学习模型中的表现.受现有Shapley方法^[16]和留一法(leave-one-out, LOO)^[17]的启示,我们基于参与方对模型的贡献来评估参与方在单次学习模型中的信誉表现并给予积累.为了解决阈值设定困难和Shapley方法暴力求解计算复杂度高等问题,我们基于后向选择的思路,迭代排除单次贡献度排序最差的参与方,在多项式时间复杂度下获得性能良好且稳定的全局模型,积累参与方的信誉值.此外,由于协作学习场景中参与方的分布式和互不信任特点,参与方信誉的建立需依赖去中心化方式进行.对此,本文借助区块链技术进行数据的去中心化存储,并基于区块链系统中的智能合约进行信誉值的实时更新,保证了信誉建立的安全性和可靠性.主要贡献包括4点:

1) 提出了一个描述医疗机构数据质量的AI信誉值指标AMP(AI medical promise).在强调去中心化和互不信任的联邦学习和群体智能场景中,天然地基于区块链技术设立医疗领域协作学习生态与场景,并具体利用智能合约和参与方的AMP指标来进行参与方信誉的评估、积累和选择,以帮助在医疗领域中形成良好的AI生态.

2) 在医疗领域协作学习场景中建立了一个基于后向选择的迭代协作学习(colback-learning)框架.在单次协作学习任务中,利用后向选择方法以多项式时间复杂度迭代计算出性能良好且稳定的全局模型,并基于贡献度排序给出参与方的单次信誉值指标

a^2mp (alone AI medical promise)用于信誉值积累.

3) 在信誉值计算中,制定了一个综合考虑全局性能指标的评分函数,以针对医疗领域更准确地评估全局模型的性能,指导参与方信誉值的计算和全局模型的训练.

4) 基于规模化真实医疗数据的模拟实验表明,我们的方法能够选择可靠的参与方,得到性能良好的全局模型,其稳定性比现有参与方选择方法提高1.3~6倍,全局模型的可解释性与集中式学习保持了较高的一致性.

1 相关工作

协作学习已经在医疗领域开始应用研究^[18-25].比如2021年5月26日的Nature封面文章给出了4个利用不同医疗机构数据的群智学习案例^[5].Kumar等人^[24]将联邦学习应用于基于CT图像的新冠肺炎检测.Roth等人^[25]将联邦学习应用到乳腺BI-RADS分类辅助诊断.协作学习能有效解决医疗领域中“数据孤岛”问题,拥有巨大的应用价值和发展潜力.

1.1 参与方选择

自2016年谷歌公司提出联邦学习以来,协作学习中的参与方选择一直是研究热点之一^[4].最受欢迎的FedAvg算法提出采用随机选择策略选择参与方,确保模型的鲁棒性^[26].一些参与方选择策略以加速联邦学习模型收敛或者提高联邦学习模型性能为目的.例如,Cho等人^[11]提出了一种偏向于选择局部模型损失较高的参与方的选择策略.局部损失越大,代表模型改进的潜力越大,因此该方法可以提高全局模型的收敛速度.Zhang等人^[12]引入了一个损失阈值来确定是否接受局部模型.

后续研究考虑的是在资源与时间约束下,如何鼓励尽可能多的参与方参与学习,或者在计算成本与模型性能之间取得平衡.比如,Zeng等人^[27]提出了采用多维拍卖激励机制的FMore机制,鼓励更多数据质量高且计算成本低的参与方参与联邦学习,最终提高了联邦模型的性能.

在多方医疗数据协作场景中,相较于计算成本与训练时长,模型性能和医疗数据质量更受人们关注.高质量的参与方应该不惜代价地被纳入,而这恰好被现有协作学习参与方选择策略所忽略甚至相悖.比如,Nishio等人^[28]提出的在资源限定下选择参与方的FedCS协议只考虑了参与方的通信条件和计算资源.由于训练时间和传输延迟,高质量、数据量大的

参与方经常被排除.

1.2 参与方评估

在协作学习中,各参与方的数据都保留在本地,不能被共享、交换和直接访问,无法直接通过训练数据本身来评估其数据质量.很多研究考虑评估参与方在协作学习中的贡献,并据此设计相关的激励机制和分配机制.例如, Jia 等人^[29]通过使用 Shapley 值^[16]来评估数据的贡献. Sim 等人^[30]研究如何根据每个客户的训练数据的 Shapley 值设计激励机制. Song 等人^[31]通过训练不同参与方组合的模型来评估参与方贡献.留一法^[17]是一种类似于 Shapley 值的删除-再训练方法. Wang 等人^[32]使用删除法和留一法的方法来衡量每个参与方对横向联邦学习中的贡献.它可以传递每个参与方对模型的影响,但无法区分影响是积极的还是消极的.

文献 [17, 29–32] 方法在评估参与方贡献时都需要评价模型性能.但机器学习领域最常用的模型性能评价指标不能单独在医疗领域 AI 中使用.比如,大多数医疗数据都存在严重的阴阳性样本不平衡问题^[10].此时,仅仅用准确率(Accuracy)来评估模型就

会出现偏差,需要更多的考虑召回率(Recall)和特异度(Specificity)等^[5].而一些医学模型也会更多地关注灵敏度(Sensitivity).

2 医疗领域协作学习生态

我们首先描述医疗领域协作学习生态.如图 1 所示,医疗领域的协作学习生态由多个协作学习场景(社区)构成.一个社区由拥有特定医疗数据和计算资源的各机构共同组成.任何包含医疗数据与计算资源的机构都可以是社区内任务的请求者或参与方.在特定时刻,由任务请求者向社区发布一个协作学习任务,请求具有相关医疗数据和计算资源的机构加入,形成一个独立的多方医疗数据协作场景.经过审核以后,被选中的机构成为参与方集合 SP .任务请求者与各参与方一起在该场景下进行协作学习,训练出 AI 全局模型.全局模型及其解释结果由各参与方共享,以多方医疗数据促进对医学临床问题的认知和研究,促进疾病的早期预警与辅助诊断,以及进行疾病的辅助治疗等.

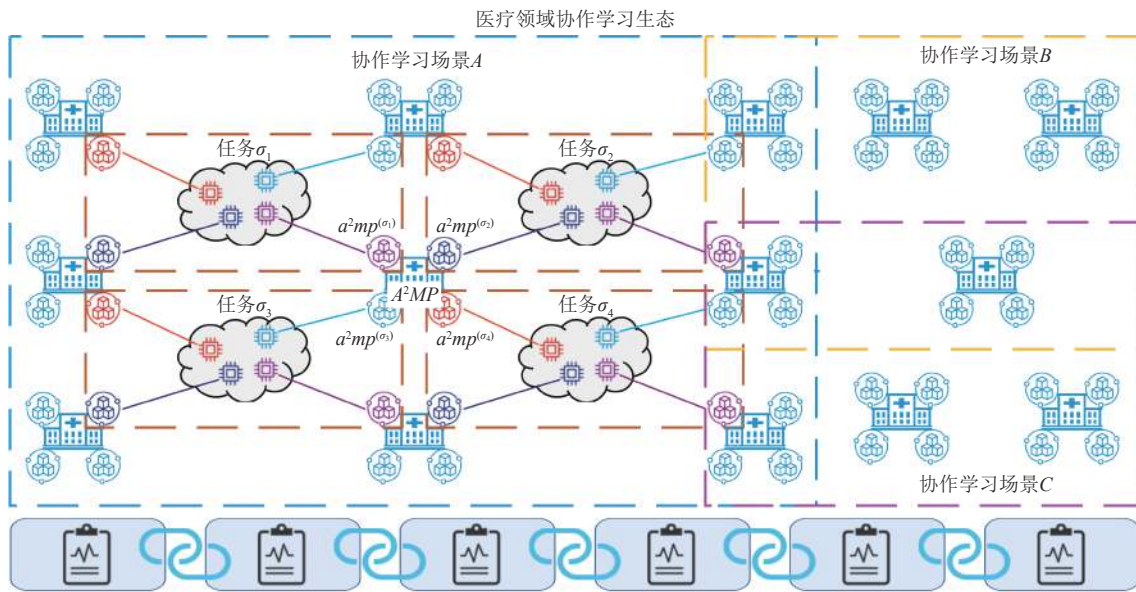


Fig. 1 Collaborative learning ecology and AMP in medical field

图 1 医疗领域协作学习生态与 AMP

2.1 参与方信誉值

参与方信誉值是医疗领域协作学习生态中反映医疗机构拥有的医疗数据可靠性与质量的指标.这里,信誉值分为单次任务信誉值 a^2mp 和累积信誉值 A^2MP (accumulative AMP).单次任务信誉值 a^2mp 表示在单次协作任务中参与方的数据对协作任务的贡献

度.累积信誉值 A^2MP 表示医学机构(参与方)在多个协作学习任务中的累积信誉影响. a^2mp 和 A^2MP 共同形成医疗机构的 AMP, 它们均存储在区块链中.区块链是公开透明并不可篡改的,在强调隐私保护和道德约束的医疗领域中非常适用.表 1 总结了本文中常用的一些符号和定义.

Table 1 Notations and Their Descriptions

表 1 符号及其描述

符号	描述
N	参与方总数
T	模型训练的总轮次
$w(t)$	第 t 轮中的全局模型参数
$w_i(t)$	第 t 轮中参与方 i 的局部模型参数
q_i	参与方 i 的本地训练数据数量
$a^2mp_i^{(\sigma)}$	任务 σ 评估的参与方 i 的单次任务信誉值
$A^2MP_i^{(\sigma)}$	任务 σ 时候选参与方 i 的累积信誉值
$Score(\cdot)$	用于评估模型性能的评分函数
$\mathcal{M}^{(t)}$	迭代轮次 t 中获得的全局模型
$S^{(t)}$	迭代轮次 t 中获得的全局模型的评价分数
$Con_i^{(t)}$	参与方 i 在迭代轮次 t 中的数据贡献
C_i	参与方 i 在多轮迭代中的平均数据贡献
$Rank_i^{(t)}$	迭代轮次 t 中参与方 i 的排名
\mathcal{R}_i	每个任务中参与方 i 的平均排名
SP	参与方集合
SR	移除的参与方集合
SL	本地数据集
SG	全局模型集
SC	数据贡献集

2.2 任务 workflow

协作学习任务中存在 2 类成员：一类是任务请求者，即发布协作学习任务的机构，一般仅有 1 个；另一类是参与方，即参与协作学习的机构，可以有多个。

任务请求者在向社区发布任务请求后，会依据 A^2MP 以及申请者提供的资源描述选择符合要求的机构成为参与方。经过基于后向选择的迭代协作学习框架，一方面获得性能良好的全局模型，另一方面计算出参与方在本次协作学习中的贡献影响，即 a^2mp 。由 a^2mp 更新计算 A^2MP ，具体有 7 个步骤：

步骤 1. 任务请求者向社区广播任务，由机构提出参与申请后，依据存储在链上的 A^2MP 和提供的资源描述选择出参与方集合 SP ，参与协作学习任务。

步骤 2. 进行迭代式协作学习。每次迭代中，任务请求者将初始全局模型 $w(t)$ 分发给参与方集合 SP 中的每个参与方 i ，并对参与方集合 SP 执行协作学习，以训练全局模型 $w(t+1)$ 。

步骤 3. 任务请求者计算参与方集合中所有参与方的贡献 $Con_i^{(t,\sigma)}$ 。

步骤 4. 任务请求者对所有参与方的贡献 $Con_i^{(t,\sigma)}$ 进行排序，并从参与方集合中删除贡献最小的参与方。

步骤 5. 任务请求者将任务迭代的信息上传到区块链，其中包括任务 id、任务请求者 id、迭代 id、参与方 id 以及参与方在该轮任务迭代中的贡献和表现。

然后我们重复步骤 2~5，直到参与方集中只剩下 1 个参与方。

步骤 6. 在区块链系统中，智能合约利用步骤 2~5 中的数据进行参与方信誉值 a^2mp 的去中心化计算，并基于所有模型在任务请求者的医疗数据上的性能对在前面步骤中训练的所有模型进行排序。

步骤 7. 表现最好的全局模型（包括参与方的信誉值 a^2mp 、模型排序结果 $\mathcal{R}_i^{(\sigma)}$ 和模型解释结果）被自动维护在区块链中，供用户在后实时调用。对于医疗场景来说，在迭代过程步骤中注意所有模型的解释结果是有意义的。

基于上述步骤，面向单个协作学习任务，本文可设计基于后向选择的迭代协作学习框架，并将在第 3 节详细地介绍框架的总体设计与具体实现。

3 基于后向选择的迭代协作学习框架

我们提出一个基于后向选择的迭代协作学习框架，在单次协作学习任务中挑选出在本次训练中表现良好的信誉参与方，获得在包含任务请求者在内的各参与方医疗数据上表现良好的全局模型 $\mathcal{M}^{(t,\sigma)}$ ，并将各参与方在本次任务中对模型贡献的信誉值 a^2mp 积累到 A^2MP 上。如图 2 所示，colback-learning 由任务初始化、参与方后向选择以及信誉值累积 3 个模块组成。任务初始化模块完成单次协作学习任务的初始化、对应 2.2 节中的步骤 1。参与方后向选择模块迭代构建全局模型，并计算各参与方在本次任务迭代中的贡献和表现，对应 2.2 节中的步骤 2~5。最后，由信誉值累积模块完成各方的信誉值计算和累积，对应 2.2 节中的步骤 6 和步骤 7。

3.1 任务初始化

任务初始化模块中最重要工作就是选择具有良好信誉值的机构（候选参与方）参与单次协作学习任务。这里，用 $a^2mp_i^{(\sigma)}$ 表示候选参与方 i 在编号为 σ 的任务中得到的信誉值。使用 $A^2MP_i^{(\sigma)}$ 表示候选参与方 i 的累积信誉值。假设区块链中任务已经存储到编号为 σ 的任务中，则候选方 i 的累积信誉值可表示为 $A^2MP_i^{(\sigma)}$ ，它由对候选参与方 i 的所有参与任务中的信誉值进行滑动平均后计算得到。也即，在当前的编号为 σ 的任务中，我们使用 $\{a^2mp_i^{(\theta)}\}_{\theta=\sigma_0}^{\sigma-1}$ 计算获得

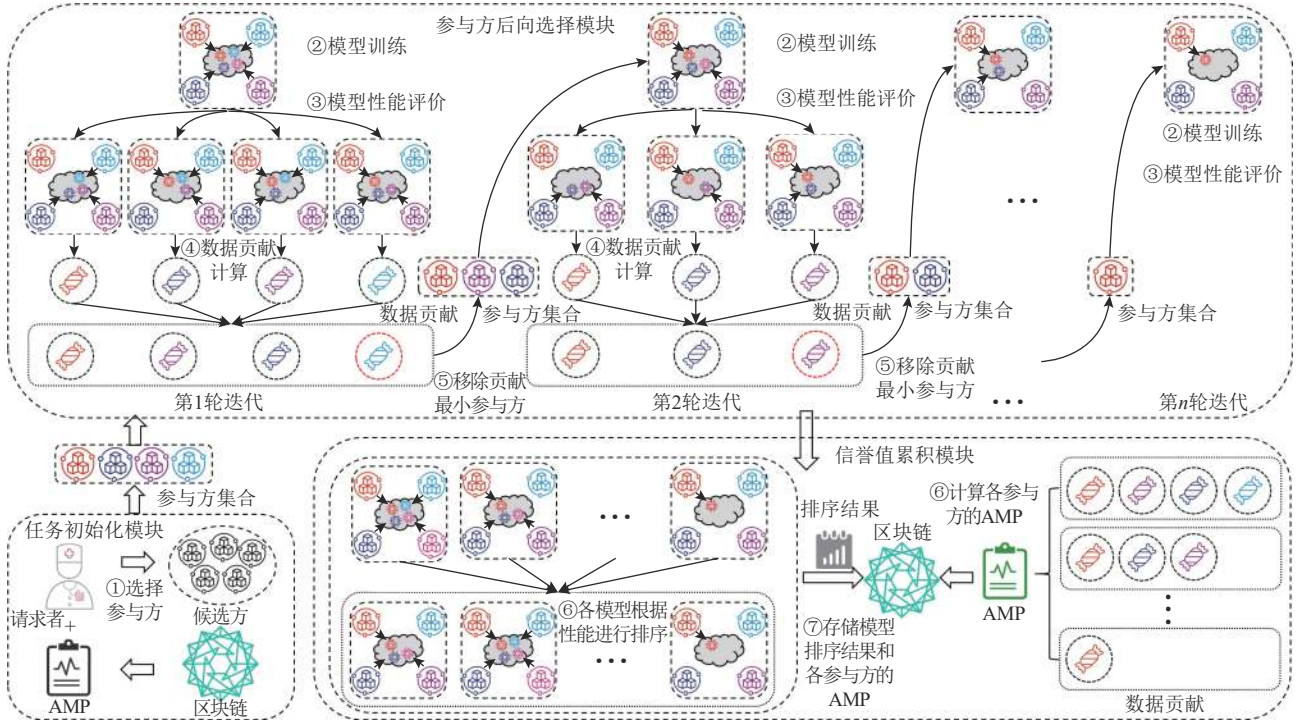


Fig. 2 Architecture of iterative collaborative learning based on backward selection

图2 基于后向选择的迭代协作学习架构

$A^2MP_i^{(\sigma-1)}$. 在 σ 任务完成后, 得到 $a^2mp_i^{(\sigma)}$ 和 $A^2MP_i^{(\sigma)}$ 用于第 $\sigma+1$ 个任务. 其计算过程如式 (1) 所示:

$$A^2MP_i^{(\sigma)} = \beta A^2MP_i^{(\sigma-1)} + (1-\beta)a^2mp_i^{(\sigma)}, \quad (1)$$

其中 $A^2MP_i^{(\sigma)}$ 为集合 $\{a^2mp_i^{(\theta)}\}_{\theta=\sigma_0}^{\sigma}$ 的滑动平均结果, β 为衰减因子. 在用式 (1) 计算所有候选对象的 A^2MP 后, 得到所有 $A^2MP_i^{(\sigma)}$. 对候选对象进行非降序排序, 然后选择前 N 位的候选对象组成参与方集合 SP .

3.2 参与方后向选择

使用信誉值 A^2MP 选择最初参与任务的参与方集合 SP 是我们工作的第 1 步. 虽然考虑了信誉值, 但仍然无法保证入选参与方对本次协作学习任务的可信性和医疗数据质量, 所以还需要进一步筛选参与方. 然而, 在协作学习场景中, 为保护数据隐私, 不可能通过直接评估参与方的医疗数据质量来筛选出可靠的参与方. 我们的工作基于参与方参与和不参与协作学习之间的模型性能差异, 即数据贡献, 来间接评估参与方医疗数据的质量. 此时, 存在一个问题, 即数据贡献达到多少才能算是可靠的参与方. 也即我们很难确定数据贡献的阈值. 阈值过大或过小的情况下, 选择出的参与方协作训练得到的协作学习模型都有可能这样或那样的问题.

针对上述问题, 我们改进了现有的后向选择方法, 在不需要任何阈值的情况下挑选出可靠的参与方. 在后向选择中, 我们首先使用参与方集合 SP 中的

所有参与方进行协作学习来训练一个全局模型 $M^{(1,\sigma)}$, 即 σ 任务中第 1 轮联邦学习获得的全局模型. 然后, 我们计算每个参与方在第 1 轮联邦学习的数据贡献 $Con_i^{(1,\sigma)}$, 并从参与方集合中去除贡献最小的参与方 x . 重复上述步骤, 直到参与方集合的大小为 1. 后向选择也是一个迭代式的协同学习训练过程, 包括模型训练、模型性能评价、贡献计算和淘汰方确定等步骤.

3.2.1 模型训练

跟经典联邦学习一样, 在协作学习第 t 轮 (这里 $t=1, 2, \dots, T$) 中, 任务请求者首先将全局模型参数 $w(t-1)$ 分配给参与方集合 SP 中的每个参与方, 然后参与方 i 使用其本地医疗数据 D_i 按照式 (2) 训练其本地模型:

$$w_i(t) = w(t-1) - \eta \nabla F_i(w_i(t-1)), \quad (2)$$

其中 η 为步长, $F(w)$ 为损失函数.

经过本地训练后, 参与方将其本地模型参数 $w_i(t)$ 发送给任务请求者. 与 FedAvg^[4] 类似, 在收集到所有参与方发来的本地模型参数后, 任务请求者将根据式 (3) 对所有本地模型参数进行加权平均, 得到新的全局模型参数 $w(t)$:

$$w(t) = \frac{\sum_{i=1}^N q_i w_i(t)}{\sum_{i=1}^N q_i}, \quad (3)$$

其中 q_i 表示参与方 i 的医疗数据 D_i 的样本数量.

当 t 达到预定义的训练轮数或模型达到早停法停止条件时, 模型训练过程结束.

3.2.2 模型性能评价

在参与方集合 SP 协作训练得到协作学习模型之后, 我们需要使用任务请求者持有的测试集来评估模型的性能和泛化能力.

在大多数研究中, 准确率通常被用作评价模型性能的常用指标. 但在医疗领域中, 单独的准确率指标无法进行全面评估, 因为医疗数据中普遍存在阳性样本与阴性样本不平衡的现象. 以抑郁状态检测^[33]为例, 患者不会一直处于抑郁状态, 而只是在某些时期处于抑郁状态, 这就导致医疗数据的阴性样本(如非抑郁状态)远远多于阳性样本(如抑郁状态)^[10]. 一个模型即便只能识别出少量的阳性样本, 但由于能够正确识别出绝大多数的阴性样本, 其模型的准确率也会非常高. 然而此时, 我们更为关注的阳性样本, 其模型表现并不好.

我们分析发现, 很多的指标都在医学领域中被关注. 比如, 医疗领域很在意漏诊率和误诊率. 漏诊率和误诊率越低, 意味着模型性能越好. 而漏诊率和查全率相关, 误诊率和特异性指标相关. 查全率代表了真阳性率, 也经常被称为灵敏度, 即被正确识别为阳性的阳性样本在所有真正的阳性样本中的比例, 计算同召回率. 特异度代表了真阴性率, 即所有现有阴性样本中被正确识别为阴性样本的阴性样本的比率. 希望漏诊率和误诊率都低时, 我们还需要考虑 $F1$ 分数的表现.

受试者工作特征(receiver operating characteristic, ROC)曲线下的面积 AUC (area under the curve)也是医学领域中得到广泛关注的指标. 据此, 我们在综合考虑准确率、召回率、特异度、 AUC 、灵敏度等指标的基础上, 设计了模型性能评价函数 $Score(m)$, m 指被评估的模型. $Score(m)$ 是一个拟线性函数, 其计算如式(4)所示:

$$Score(m) = p(m) + \max\{s_1(m), s_2(m)\}, \quad (4)$$

其中

$$p(m) = \alpha_1 Accuracy(m) + \alpha_2 AUC(m) + \alpha_3 F1(m), \quad (5)$$

$$s_1(m) = \omega_1 Recall(m) + \omega_2 Precision(m), \quad (6)$$

$$s_2(m) = \omega_3 Sensitivity(m) + \omega_4 Specificity(m), \quad (7)$$

其中, α 和 ω 是相关系数. 如果这些指标同样重要, 可以将它们设置为相同的值, 比如 1. 简单来说, 可以认为 $Score(m)$ 越高, 模型 m 的性能就越好.

精确率(Precision)通常伴随灵敏度一起出现, 即

被正确识别为阳性的阳性样本在所有被预测为阳性的样本中的比例. $s_1(m)$ 是灵敏度和精确率的加权平均, 灵敏度和精确率是一对有些矛盾的指标, 一般灵敏度高一点, 精确率就会低一点. $s_2(m)$ 是灵敏度和特异度的加权平均, 这 2 个指标是医学上比较关心的指标. 由于 $s_1(m)$ 与 $s_2(m)$ 比较接近, 其中都包含灵敏度, 因此取其中最大值即可.

准确率代表模型正确预测出样本真实类别的能力, $F1$ 综合考虑了灵敏度和精确率的影响, 是灵敏度和精确率的调和平均值. AUC 与预测阈值无关, 是检测准确性的综合代表. $p(m)$ 作为这 3 个综合指标的加权平均, 可以很好地代表模型综合性能. 如同根据多个指标判断深度学习模型性能, $Score(m)$ 通过聚合多个评价指标, 可以获得对模型更加全面准确的评价.

3.2.3 贡献计算

在完成利用参与方集合 SP 中所有参与方协作训练得到模型 M , 并使用函数 $Score(\cdot)$ 对其进行性能评价得到评价分数 S 后, 我们可以来计算参与方 k 的数据贡献. 首先, 需要将参与方 k 从参与方集合中移除. 在去除需要评估的参与方 k 后, 其余参与协作学习的参与方生成一个全局模型 M_{-k} , 并评估获得的模型性能 S_{-k} . S 和 S_{-k} 之间的差异就是参与方 k 的数据贡献 Con_k , 它可以间接反映 k 的医疗数据质量. 算法 1 给出了贡献计算的具体步骤. 其基于这样一个事实, 即如果有一个具有优质医疗数据的参与方参与协作学习的训练过程, 则所获得的模型的性能应该得到改善或保持.

算法 1. 各参与方的数据贡献的计算.

输入: 协作学习模型 M 及其评价分数 S , 当前迭代轮次参与方总数 K , K 个参与方的本地训练数据的集合 $\{D_1, D_2, \dots, D_k\}$;

输出: Con_k for $k = 1, 2, \dots, K$. /* K 个参与方的数据贡献 */

① for all $k = 1, 2, \dots, K$ do

② 将 D_k 从训练集中移除;

③ 再次训练获得 M_{-k} ;

④ 任务请求者根据式(4)对 M_{-k} 进行性能评价得到评价分数 S_{-k} ;

⑤ $Con_k \leftarrow S - S_{-k}$;

⑥ end for

⑦ return Con_k for $k = 1, 2, \dots, K$.

3.2.4 基于淘汰的迭代工作机制

首先进行淘汰方确定. 在数据贡献计算完成后, 我们将参与方按其贡献大小进行非降序排序, 排名

范围为 $1 \sim N$. 为进一步通过后向选择方法明确更多的可靠参与方, 我们从参与方集合中删除贡献值最小的参与方. 参与方集中的其余参与方进入该任务的下一轮迭代. 重复上述步骤, 直到参与方集合中只有 1 个参与方.

在淘汰方确定后, 我们记录每轮迭代中各参与方的数据贡献 $Con_i^{(t,\sigma)}$ 和排名 $Rank_i^{(t,\sigma)}$. 参与方一旦在此轮迭代中被淘汰, 其数据贡献将被设置为 0, 其排名在后续迭代轮次中保持此轮迭代的排名.

在最后一轮迭代中, 参与方集合中只剩下 1 个参与方. 此时, 我们只需要进行模型训练和性能评估, 不再需要贡献计算和淘汰方确定.

后向选择的伪代码在算法 2 中给出. 算法 2 的时间复杂度为多项式时间复杂度 $O(n^2)$, 小于暴力搜索的时间复杂度 $O(2^n)$. 对于医疗研究问题, 更关注的是结果的准确率. 医疗数据是非常宝贵的, 特别是标注数据. 数据的收集和获取所需时间以月、年为单位. 与此相比, 计算所需的时间和代价是可以承受的. 而且大部分情况下, 数据分析并非是实时的. 我们的方法在任务初始化时根据信誉值可以选出较为优秀的参与方, 这可以提前筛掉一些质量较低的参与方, 较好地控制参与方数量, 提高计算效率.

算法 2. 基于后向选择的迭代式协作学习.

输入: 协作学习模型的初始参数 $w(0)$, 参与方集合 $SP = \{1, 2, \dots, N\}$, 被移除的参与方集合 $SR = \{\}$, 各参与方的本地训练数据集 $SL = \{D_1, D_2, \dots, D_N\}$;

输出: 协作学习模型集合 SG , 数据贡献集合 SC .

- ① for each iteration $t = 1, 2, \dots, T$
do /*每一轮迭代*/
- ② $\mathcal{M}^{(t,\sigma)} \leftarrow \text{ModelTrain}(w(t-1), SP)$;
- ③ 任务请求者根据式 (4) 对 $\mathcal{M}^{(t,\sigma)}$ 进行性能评价得到评价分数 $\mathcal{S}^{(t,\sigma)}$;
- ④ $SG \leftarrow SG \cup \{(\mathcal{M}^{(t,\sigma)}, \mathcal{S}^{(t,\sigma)})\}$;
- ⑤ if $|SP| > 1$ then
- ⑥ set $SP' = SP$;
- ⑦ for each participant $i \in SP$ do
- ⑧ 将参与方 i 从 SP' 中移除, 此时集合 SP' 可称为集合 SP'_{-i} ;
- ⑨ $\mathcal{M}_{-i}^{(t,\sigma)} \leftarrow \text{ModelTrain}(w(t-1), SP'_{-i})$;
- ⑩ 任务请求者根据式 (4) 对 $\mathcal{M}_{-i}^{(t,\sigma)}$ 进行性能评价得到评价分数 $\mathcal{S}_{-i}^{(t,\sigma)}$;
- ⑪ $SG \leftarrow SG \cup \{(\mathcal{M}_{-i}^{(t,\sigma)}, \mathcal{S}_{-i}^{(t,\sigma)})\}$;
- ⑫ $Con_i^{(t,\sigma)} \leftarrow \mathcal{S}^{(t,\sigma)} - \mathcal{S}_{-i}^{(t,\sigma)}$;

- ⑬ $SP' \leftarrow SP' \cup \{i\}$;
- ⑭ end for
- ⑮ 任务请求者对 $\{Con_i^{(t,\sigma)} | i \in SP\}$ 进行非降序排序并据此为参与方分配排名
 $\{Rank_i^{(t,\sigma)} | i \in SP\}$;
- ⑯ $Con_x^{(t,\sigma)} \leftarrow \min \{Con_i^{(t,\sigma)} | i \in SP\}$;
- ⑰ for each participant $i \in SR$ do
- ⑱ $Con_i^{(t,\sigma)} \leftarrow 0$; /*数据贡献置为 0*/
- ⑲ $Rank_i^{(t,\sigma)} \leftarrow Rank_i^{(t-1,\sigma)}$;
- ⑳ end for /*注意: 第 1 轮迭代中, 此时 SR 为空集, 不执行此 for 循环*/
- ㉑ $SC \leftarrow SC \cup \{(Con_i^{(t,\sigma)}, Rank_i^{(t,\sigma)}) | i \in SP \cup SR\}$;
- ㉒ 从参与方集合 SP 中移除贡献最小参与方 x ;
- ㉓ $SR \leftarrow SR \cup \{x\}$;
- ㉔ end if
- ㉕ end for
- ㉖ return SG, SC .

3.3 信誉值累积

经过参与方后向选择, 我们根据所有获得的模型的性能排序结果得到最优协作学习模型. 与该模型相对应的参与方满足我们的性能可靠性要求, 并具有较高的医疗数据质量. 最后, 根据各参与方在任务中的表现, 计算其本次任务的信誉值 a^2mp , 并更新其累积信誉值 A^2MP . 信誉值的更新依赖于区块链系统中的智能合约. 具体而言, 将本节中的计算用智能合约实现, 并利用交易进行智能合约的触发和信誉值的更新.

单次协作学习任务中各参与方的信誉值与其在该任务中的贡献和表现以及在迭代过程中参与模型训练的次数相关联. 其参与协作学习次数越多, 参与方就越重要. 此外, 参与方的数据贡献代表了它对所构建的全局协作学习模型的影响程度, 参与方的排名代表了它在当前任务的每次迭代中的表现如何. 具体而言, 参与方的贡献 C 是当前任务多次迭代中参与方数据贡献的平均值, 其计算过程如式 (8) 所示:

$$C_i^{(\sigma)} = \frac{\max \left\{ 0, \sum_t Con_i^{(t,\sigma)} \right\}}{|\{Con_i^{(t,\sigma)} \neq 0\}|}. \quad (8)$$

类似地, 参与方的排名 \mathcal{R} 是当前任务多次迭代中参与方排名的平均值. 由于最后一轮迭代中, 参与方集合 SP 中只有 1 个参与方, 没有进行数据贡献计算, 也没有排名. 因此, 参与方的任务排名取其在 $N-1$ 轮迭代中排名的平均值. 计算过程如式 (9) 所示:

$$\mathcal{R}_i^{(\sigma)} = \frac{\sum_t \text{Rank}_i^{(t,\sigma)}}{T-1}. \quad (9)$$

不同协作学习任务要求的参与方数量是不同的,致使不同任务中参与方的数据贡献及排名的取值范围也不同.因此,需要对参与方的任务数据贡献 \mathcal{C} 和任务排名 \mathcal{R} 做归一化处理.参与方 i 的相对任务数据贡献 c_i ,是基于任务数据贡献最大值计算的相对值,计算方式为

$$c_i^{(\sigma)} = \frac{\mathcal{C}_i^{(\sigma)}}{\max_i (\mathcal{C}_i^{(\sigma)})}. \quad (10)$$

参与方 i 的相对任务排名 r_i 是基于任务最高排名的相对值.例如, N 个参与方的协作学习任务的最高排名是 N . r_i 计算过程如式(11)所示:

$$r_i^{(\sigma)} = \frac{\mathcal{R}_i^{(\sigma)}}{\mathcal{R}_{\max}^{(\sigma)}}. \quad (11)$$

我们使用冈珀茨函数(Gompertz function)^[34]来衡量参与方被选择参与协作学习的次数对评价其信誉值的影响.冈珀茨函数适用于对个体交互中的信任概念建模^[35].参与方 i 的冈珀茨函数如式(12)所示:

$$y_i = ae^{be^{-\gamma_i}}, \quad (12)$$

其中 a, b, c 是参数.

冈珀茨函数的输入 γ_i 定义如式(13)所示:

$$\gamma_i = \frac{\varepsilon m_i^{\text{sele}} - (1-\varepsilon)m_i^{\text{remo}}}{\varepsilon m_i^{\text{sele}} + (1-\varepsilon)m_i^{\text{remo}}}, \quad (13)$$

其中 m_i^{sele} 表示参与方 i 迭代过程中参与协作学习的次数, m_i^{remo} 表示参与方 i 未被选中参与协作学习的次数,即被移除次数.不能只关注被选中次数而忽视被移除次数,因此,权重 ε 的取值范围可设为 $(0, 0.5]$.

在计算得到参与方的相对任务数据贡献 c_i 、相对任务排名 r_i ,以及迭代过程中参与协作学习的次数 y_i 之后,式(14)展示了如何评价参与方 i 的医疗数据对本次任务的影响,即计算参与方 i 的信誉值(a^2mp_i):

$$a^2mp_i = y_i c_i r_i. \quad (14)$$

基于智能合约可以自动、可信地进行以上计算.所有参与方的信誉值,即 A^2MP , a^2mp ,将自动维护在区块链系统中,供用户进行实时轻量化地查询.

我们的框架将所有中间处理结果存储在区块链中,它是公开透明的,不能被篡改.区块链通过以下特性和技术,保证数据在区块链上传过程中的公平性和真实性,并防止参与方恶意篡改数据.

1)加密技术.使用加密算法对数据进行加密,确保数据安全.除非破解此种加密算法或者获得对应的密钥,才能篡改数据.

2)分布式存储.区块链去中心化地在多个节点上分散存储同一份数据,每个节点都有完整的数据副本,这些节点可以相互验证和协作,以确保数据的公正性和不可篡改.

3)区块链共识机制.通过共识机制,如工作量证明(proof of work, PoW)、权益证明(proof of stake, PoS)等,来约束所有参与方达成共识、共同验证数据,并确保数据经过多次确认之后才被写入到区块链中,从而保证数据的真实性和不可篡改.

4)智能合约.智能合约^[36]是预设好的计算机程序,它可以自动执行区块链网络中的具体操作.智能合约根据指定的规则检查工作量证明或权益证明共识机制,确保只有通过合适的计算和验证的操作才能修改数据.

5)公开透明.在区块链上的数据都是公开透明的,每个参与者都可以看到对于数据的改动,并且通过记录元数据信息来检查该更改是否由一个合法的委托者发起.并且所有的更改都必须得到其他节点的确认,从而增加破坏数据完整性和真实性的复杂度.

上述措施与技术,保证了上链后数据的真实性,但是如何保证上链前数据的真实性、防止任务请求者恶意上传不真实数据上链,是区块链应用过程中的一个关键问题.

区块链主要保证了上链数据的去中心存储和难以篡改,上链前数据的真实性还需要其他技术和措施的帮助.如参与方通过相似任务贡献结果存在较大出入或协作学习中淘汰结果不合理对任务请求者提出质疑,由可信的第三方进行复现检测.由于区块链上数据难以清除或篡改,并且在医疗协作学习生态中,参与机构是非匿名的,恶意提交将会对其所在机构的信誉造成巨大的伤害.高昂的造假成本,可以有效吓阻恶意行为.

4 对比实验与结果分析

本节介绍我们基于规模化真实医疗数据进行的对比实验设计和结果分析.不失一般性,我们称这里的协作学习为联邦学习.

4.1 场景描述

我们使用术中低血压(intraoperative hypotension, IOH)相关数据集模拟联邦学习场景进行实验. IOH事件是许多术后并发症的独立危险因素,如死亡率的急剧增加、心肌损伤、急性肾损伤、术后精神错乱

等神经系统并发症^[37]. 然而, 对 IOH 事件的干预主要是反应性的, 而且总是延迟的. 已经有许多研究^[38-40]开始开发基于机器学习的 IOH 事件预警系统.

我们的 IOH 相关数据集有 3 个来源. 一是来自世界上最大的生物信号公共数据集 VitalDB^[41]. 它收集了首尔国立大学医院 2016 年 6 月至 2017 年 8 月的 6 388 名外科患者的手术数据, 包括采样频率为 1~7 s 的高质量术中生物信号数据. 另外 2 个真实数据集 TJ30sDB 和 TJ5minDB, 均来自武汉同济医院. TJ30sDB 以 30 s 的数据采样频率涵盖了 2015 年 10 月至 2021 年 4 月的 387 291 名外科患者的手术数据. TJ5minDB 以 5 min 的采样频率收集了 2016 年 1 月至 2019 年 12 月的 174 434 名外科患者的手术数据.

4.1.1 数据选取及预处理

我们在上述 3 个 IOH 相关数据集中选取了适合的数据来模拟联邦学习场景. 它们是:

1) Vital. 从 VitalDB 中选取特征相对完整的 2 872 例患者, 包括 1 782 例阳性样本 (即 IOH) 和 1 090 例阴性样本 (即非 IOH). 数据采样频率为 2 s.

2) TJ30s. 从 TJ30sDB 中选取特征比较完整的患者 25 817 例, 其中阳性 18 049 例, 阴性 7 768 例. 数据采样频率为 30 s.

3) TJ5min. 从 TJ5minDB 中选取特征完整的患者 60 180 例, 其中阳性样本 26 834 例, 阴性样本 33 346 例. 数据采样间隔为 5 min.

我们将这 3 个数据集的采样间隔统一为 1 min. 当数据的采样间隔大于 1 min (如 TJ5min 的采样间隔为 5 min) 时, 我们根据术中监测数据的实际生理特征对数据进行样条插值.

4.1.2 联邦学习场景设置

在实验中, 为了模拟 colback-learning 框架和对比算法的联邦学习场景, 我们假设 colback-learning 中初始选择模块选择的参与方数量为 5. 任务请求者有 1 个验证集, 其由 10% 的 Vital 患者和 10% 的 TJ30s 患者组成. 任务请求者的测试集包括 20% 的 Vital 和 20% 的 TJ30s. Vital 和 TJ30s 中的剩余数据作为训练数据. Vital 中的训练数据被划分为参与方 P_1 和 P_2 . P_1 包括 144 520 个样本, P_2 包括 70 413 个样本. 阳性样本与阴性样本的比例均为 1 : 9. TJ30s 中的训练数据也被划分为参与方 P_3 和 P_4 , 分别包括 276 978 和 138 942 个样本. 阳性样本与阴性样本的比例均为 1 : 3. TJ5min 用于模拟参与方 P_5 , 包含 383 118 个样本. 阳性样本与阴性样本的比例为 1 : 28. 表 2 给出了 5 个参与方的数据描述.

Table 2 Data Description of Five Participants in the Simulating Federated Learning Scenarios

表 2 模拟联邦学习场景中 5 个参与方的数据描述

参与方编号	数据来源数据集	样本数量	比例
P_1	Vital	144 520	1 : 9
P_2	Vital	70 413	1 : 9
P_3	TJ30s	276 978	1 : 3
P_4	TJ30s	138 942	1 : 3
P_5	TJ5min	383 118	1 : 28

注: 比例表示阳性样本与阴性样本的比例.

4.1.3 模型介绍

我们使用 15 min 的滑动窗口来分割时间序列数据. 首先, 使用 GRU 模型来有效地捕获时间序列数据中潜在的依赖关系^[42-43]. 这里使用了 2 层 GRU 单元, 并在每层 GRU 之后添加了 1 个 Dropout 层. 同时, 使用了具有 3 个卷积层、1 个全局池化层和 1 个全连接层的 FCN 网络^[44-45]. 我们在全局池化层之后添加了 1 个 Dropout 层. 其次, 我们选择 ResNet^[46-47] 来分析静态特征, 包括术前静态特征和时序统计特征. 这里, 主要采用了 2 层残差块与 1 层 Dropout 层的结合. 最后, 将 3 种网络的输出相结合, 并通过 1 层全连接层得到预测结果. 实验结果取 10 次重复实验结果的平均值.

3 个 IOH 相关数据集模拟的联邦学习场景较真实地模拟了医疗领域的协作学习生态场景. 首先, 各数据集的样本量十分充足. 其次, 各数据集存在严重的阴阳性样本比例不平衡. 由于 IOH 事件的特殊性, 阳性样本和阴性样本的比例甚至可低于 0.1. 同时, 模拟的参与方在数据数量及质量上存在明显差异. 最后, 用于预测时间序列数据的神经网络足够复杂.

4.2 对比实验设置

4.2.1 对比方法

为了更好地验证我们提出的 colback-learning 的有效性, 对比了多种参与方选择方法, 包括:

1) AccBS. 使用单一的准确率评价模型性能, 指导数据贡献的计算, 进而展开迭代的 colback-learning.

2) BacBS. 使用单一的平衡准确率 (balanced accuracy, BACC) 评价模型性能, 指导计算数据贡献, 进而展开迭代的 colback-learning.

3) BaseFL. 使用选择所有参与方参与训练的 FedAvg^[4].

4) MQDFL. 使用 RRAFL^[12] 中提出的模型质量检测 (MQD) 方法替代随机选择策略的 FedAvg.

这里, AccBS 和 BacBS 相当于消融实验. 为了进行一致性评价, 我们进行了统一.

4.2.2 评价指标

实验中使用了深度学习中常用的标准评价度量：

1) 准确率. 衡量正确预测结果与总样本数之比的标准性能指标.

2) 平衡准确率. 平衡准确率通过正样本和负样本的数量使真阳性和真阴性预测标准化. 如果大多数数据集是不平衡的, 则建议使用平衡准确率, 而不是准确率.

3) 灵敏度. 它代表真实阳性率, 也经常被称为敏感度, 和召回率相同, 它是所有现有阳性样本中被正确识别为阳性样本的比率.

4) 特异度. 它表示真阴性率, 即所有现有阴性样本中被正确识别为阴性样本的比率.

5) *AUC*. 它是衡量学习模型性能的重要指标. ROC 曲线是以真阳性率(true positive rate, TPR)为纵坐标轴, 以假阳性率(false positive rate, FPR)为横坐标轴的曲线. *AUC* 是 ROC 曲线下的面积. *AUC* 越大, 意味着分类模型性能越好. 无论预测阈值如何改变, ROC 曲线都不会变, *AUC* 也不会改变. *AUC* 是真正意义上与阈值无关的评价指标.

6) *F1*. 通过计算灵敏度和精确率的调和平均值来衡量模型性能. 其计算公式为 $F1 = (2 \times Recall \times Precision) / (Recall + Precision)$.

4.2.3 实验参数设置

FedAvg 中的随机选择策略是最基础的参与方选择算法, 在模型训练过程的每轮迭代中, 它会随机选取一定比例的参与方参与训练. 在 IOH 预测的协作场景中, 参与方数量为 5, 因此, 本文考虑使所有参与方参与联邦学习, 即相当于将随机选择策略的比例系数设置为 100%. MQDFL 中, 通过检验参与方参与和不参与的 2 个模型在测试集上的损失值差值是否超过设定的损失阈值, 决定该参与方的本地模型是否参与模型聚合计算. 实验中我们的损失阈值设为 -0.01. 实验中, 深度学习相关模型训练时, 采用的 batch 大小为 1 024, 本地训练最大迭代次数为 5, 联邦学习全局迭代次数为 10. 性能评价时, 我们认为选取的指标都很重要, 因此将模型性能评价函数 *Score(m)* 中的各相关系数均设为 1.

4.3 实验结果与验证分析

4.3.1 参与方可靠性对比

在我们的实验中, colback-learning 使用模型性能评价函数 *Score(·)* 对模型性能进行评价, 最终选择了参与方 P_1, P_3, P_4 . AccBS 使用准确率对模型性能进行评价, 最终选择了参与方 P_2, P_3, P_4 . BacBS 使用 BACC

对模型性能进行评价, 最终选择了参与方 P_3 与 P_4 . BaseFL 选择所有参与方, 而 MQDFL 在模型质量检测后最终选择了参与方 P_1, P_2, P_3, P_4 参与模型聚合. 表 3 列出了不同算法选择的参与方.

Table 3 Results Selected by Participants in Five Methods

表 3 5 种方法的参与方选择结果

方法	P_1	P_2	P_3	P_4	P_5
colback-learning (本文)	✓		✓	✓	
AccBS		✓	✓	✓	
BacBS			✓	✓	
BaseFL	✓	✓	✓	✓	✓
MQDFL	✓	✓	✓	✓	

如表 2 所示, 参与方 P_1 和 P_2 数据来源于 Vital 数据集, 数据采样间隔为 2 s, 两者的阳性、阴性样本比例均为 1 : 9. 参与方 P_3 和 P_4 的训练数据都来自于 TJ30s 数据集, 数据采样间隔为 30 s, 其阳性与阴性比例均为 1 : 3. 参与方 P_5 的训练数据来自于 TJ5min 数据集, 其采样间隔为 5 min, 阳性、阴性样本比例低至 1 : 28, 阴阳性样本比例严重失衡, 且采样间隔很短(不到 1 min). 因此, 即使 P_5 拥有大规模医疗数据样本, 相较于其他参与方, 其数据质量显然较差. Vital 数据集和 TJ30s 数据集的参与方, 如 P_1, P_2, P_3, P_4 , 理论上可以提高验证集和测试集的模型性能, 而 P_5 加入协作后可能会为模型带来不利的影响. 此外, P_1 和 P_3 分别比 P_2 和 P_4 具有更多的数据. P_3 的数据质量优于 P_2 . 因此, P_3 被选择的概率最高, P_1 次之; P_4 优于 P_2 , 而 P_5 最差.

据此分析, 选择 P_1 和 P_3 对于提高模型在验证集和测试集上的性能是合理的. 此外, P_4 的数据量接近 P_1 , 阴阳性样本比例优于 P_1 . 因此, P_4 加入联邦学习后, 将有助于模型性能的提升. 而 P_2 的数据量是所有参与方中最小的, 阳性、阴性样本数量比值接近 0.1. 因此, P_2 相比 P_1 和 P_4 对模型性能提升的贡献不大. 换句话说, P_2 不是一个可靠的参与方. 对于验证和测试来说, 它没有较好的数据质量. 在这种情况下, 应该避免选择数据量小且数据很不平衡的参与方参与联邦学习. 综上所述, 基于对参与方数据数量及数据质量的分析, P_1, P_3, P_4 对于联邦学习是更可靠和更有价值的. 结果表明, 我们的 colback-learning 的参与方选择比其他算法更合理.

4.3.2 全局模型的性能比较

如图 3 给出的箱型图所示, 本文所提的 colback-learning 训练的全局模型的性能, 特别是 *AUC* 的性能

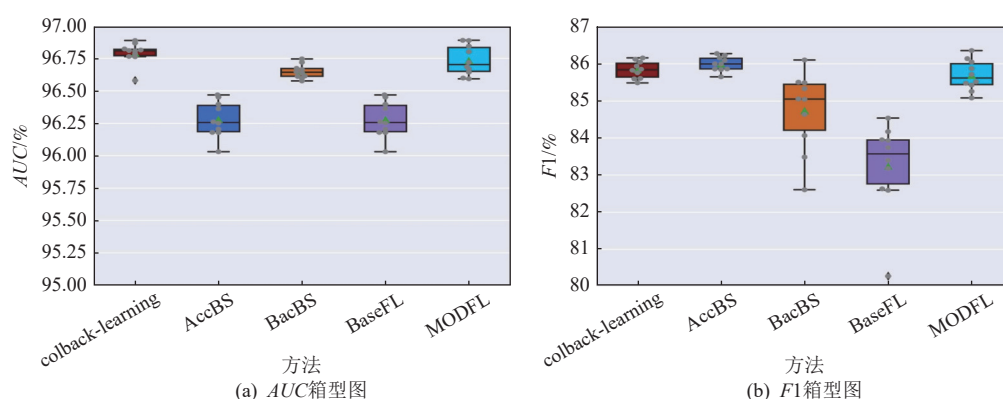


Fig. 3 AUC & F1 of the global models trained by colback-learning, AccBS, BacBS, BaseFL and MQDFL

图3 colback-learning 与 AccBS、BacBS、BaseFL、MQDFL 训练所得全局模型的 AUC 和 F1

优于其他对比算法. colback-learning、AccBS、BacBS 训练的全局模型的 AUC 和 F1 得分远高于 BaseFL, 接近 MQDFL.

在图 3 中, 方法散点的集中分布意味着 10 次实验得到的模型具有相似的性能. 相反, 如果散点分布是分散的, 则模型每次的表现都不一样. 也就是该方法训练的模型是不稳定的. 例如, colback-learning 训练模型的 F1 得分达到 $85.8\% \pm 0.3\%$, AccBS 训练模型的 F1 得分达到 $85.9\% \pm 0.3\%$, BacBS 训练模型的 F1 得分达到 $85.5\% \pm 0.9\%$, MQDFL 训练模型的 F1 得分达到 $85.7\% \pm 0.7\%$, BaseFL 训练模型的 F1 得分达到 $82.4\% \pm 2.1\%$. 分析 F1 得分情况可知, colback-learning 训练模型的稳定性是 MQDFL 训练模型的稳定性的 2.3 倍, 更是 BaseFL 训练模型稳定性的 7 倍.

经过 10 次实验, colback-learning 训练的全局模型的 AUC 和 F1 评分均较高且更稳定. AccBS 的 F1 分数和 BacBS 的 AUC 较稳定, 但 AccBS 的 AUC 和 BacBS 的 F1 分数不稳定. BaseFL 的 AUC 和 F1 分数都不稳定且较低, MQDFL 的评分较高但不稳定. 结果表明, 我们的 colback-learning 与 $Score(\cdot)$ 可以学习获得一个性能良好且稳定的全局模型.

继续从全局模型的其他性能指标来观察. 图 4 给出了 colback-learning, AccBS, BacBS, BaseFL, MQDFL 在 10 次实验中训练的全局模型的各项性能指标的增强箱型图. 图 4 中可以看到, 与 BaseFL 和 MQDFL 相比, colback-learning 获得的全局模型的各项性能指标比 BaseFL 好很多, 非常接近目前最先进模型质量检测策略 MQDFL. 结果同时表明, colback-learning 训练的全局模型在 10 次重复实验中是很稳定的. colback-learning 模型波动最大的性能指标(此处为灵敏度)的波动小于 3%, 而 MQDFL 模型波动最大的性能指标(此处也是灵敏度)的波动超过 8%.

使用 MQDFL, 在损失阈值设置为 -0.01 时, 模型质量检测判断来自 P_2 的局部模型是良好的. 因此, MQDFL 选择参与方 P_2 参与联合学习. 然而, P_2 并没有提高模型的性能, 有时还会降低灵敏度. 这些结果进一步证实, 我们的 colback-learning 和性能评价模块可以通过选择具有正确医疗数据质量的可靠参与方, 获得在特定数据上表现良好且稳定的模型.

4.3.3 消融实验

从图 4 可以看出, 使用我们提出的评分函数得到的灵敏度和 AUC 均高于 AccBS 和 BacBS 这 2 个只使用准确率或平衡准确率对全局模型进行评分的方法. 这证实了评分函数更适合医疗领域中的不平衡数据集. 从结果来看, AccBS 选择了 P_2 , 但排除了具有相同数据质量但更多数据的 P_1 . 同样, BacBS 选择了 P_2 , 但排除了 P_4 , 即使 P_4 包含了更多更高质量的数据. 这表明, 仅使用单独的性能指标对模型进行评分, 很难找到具有较好医疗数据质量的可靠参与方.

4.3.4 全局模型的可解释性分析

图 5 给出了使用 SHAP(Shapley additive explanation) 方法^[48]对以不同训练方式学习得到的模型进行分析得到的解释结果. 解释结果中, 给出了各特征对模型的影响, 这在医疗领域尤为关键. 因为医生需要知道模型判断的依据, 并据此寻找医学证据. 从可解释性的角度看, 我们的协作学习训练的模型与集中训练的模型找到的危险因素几乎是一致的. 而众所周知, 由于数据选择偏倚性的存在, 单点训练模型找出的危险因素大多都具有偏倚性. 从图 5 给出的结果来看, 也基本如此.

4.3.5 信誉值评价的合理性分析

表 4 显示了 colback-learning 中具有不同数据配置的参与方的标准化贡献、标准化排名和信誉值. 其最终选择的具有较好数据质量的参与方是 P_1, P_3, P_4 .

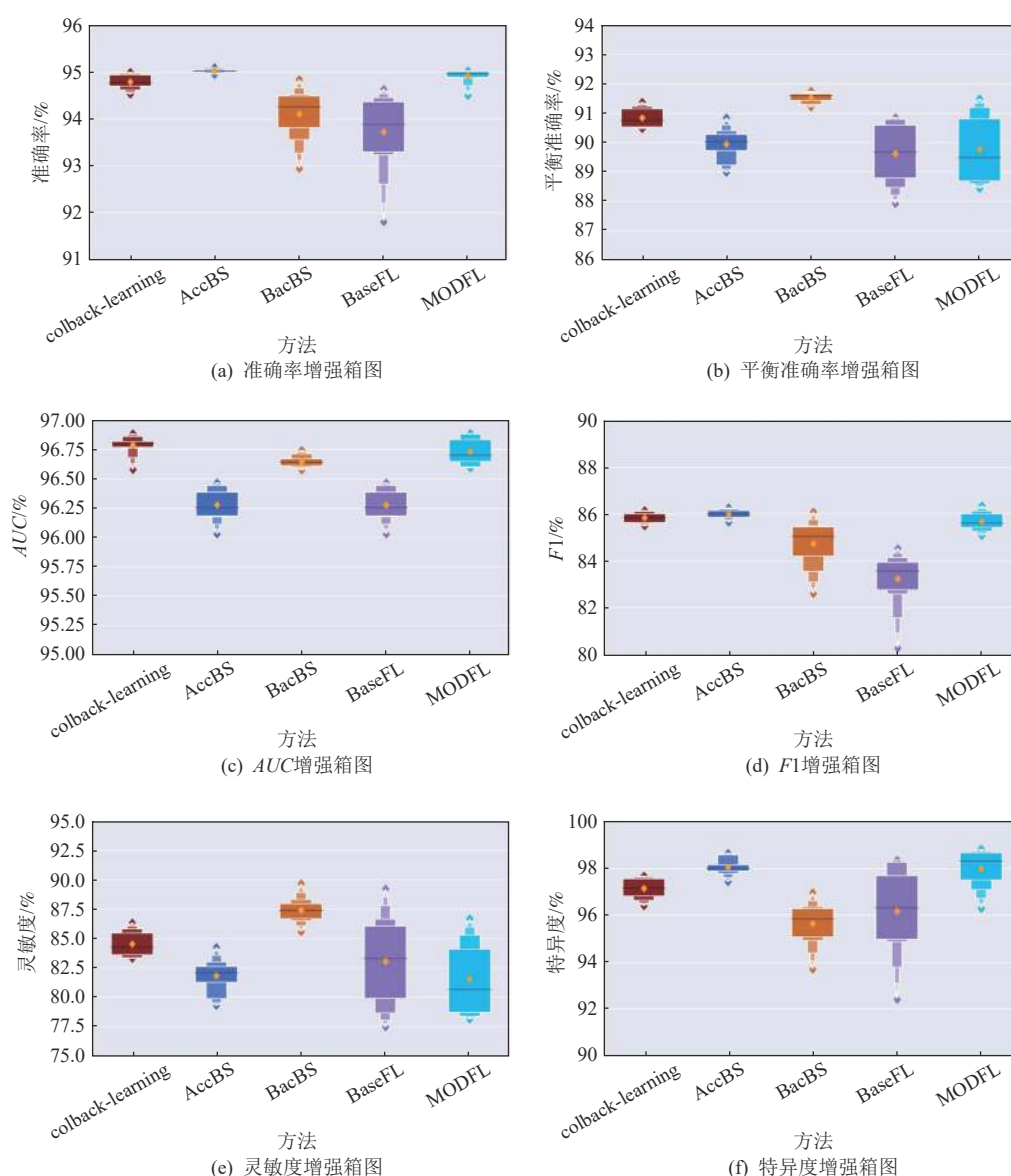


Fig. 4 Performance metrics of the global models trained by colback-learning, AccBS, BacBS, BaseFL and MQDFL

图4 colback-learning 与 AccBS, BacBS, BaseFL, MQDFL 训练所得全局模型的性能指标

P_5 由于极低的采样频率和阳性、阴性样本比例极不平衡, 导致其数据质量低下. 因此, 尽管 P_5 的数据量很大, 但 P_5 一开始就被淘汰了. P_2 被淘汰的原因是它的数据量小, 阳性、阴性样本数量比值接近 0.1. 因此, P_2 和 P_5 对于模型性能的改进没有帮助.

相比之下, P_1, P_3, P_4 有助于提高模型性能. P_3 是所有参与方中数据最多的, 它的采样频率达到 30 s, 阳性、阴性样本数量比率高于 0.3. 基于这些, P_3 在所有参与方中得分最高. P_1 包含来自 Vital 的大量数据, 这是验证和测试所必需的. 即使 P_1 和 P_4 在协作学习中具有相似的性能, 但 P_1 和 P_3 对模型性能有很大的影响, P_4 对模型的性能改进几乎没有影响, 导致在该任务中评估的 P_4 的信誉值小于 P_1 .

在 10 个实验中, P_1 和 P_3 总是被 colback-learning 选中, P_4 有时不被选中. 这些结果表明, 我们的 colback-learning 可以合理地衡量每个参与方在训练过程中对 AI 模型和性能的影响. 因此, 信誉值在一定程度上间接反映了参与方可靠性和医疗数据质量.

5 结 论

本文提出了基于后向选择的迭代协作学习 (colback-learning) 框架, 以选择具有高质量医疗数据的可靠参与方, 获得在特定数据上表现较好的全局模型. 我们提供了信誉值数据评价指标体系来估算机构拥有的数据质量, 并基于区块链智能合约进行

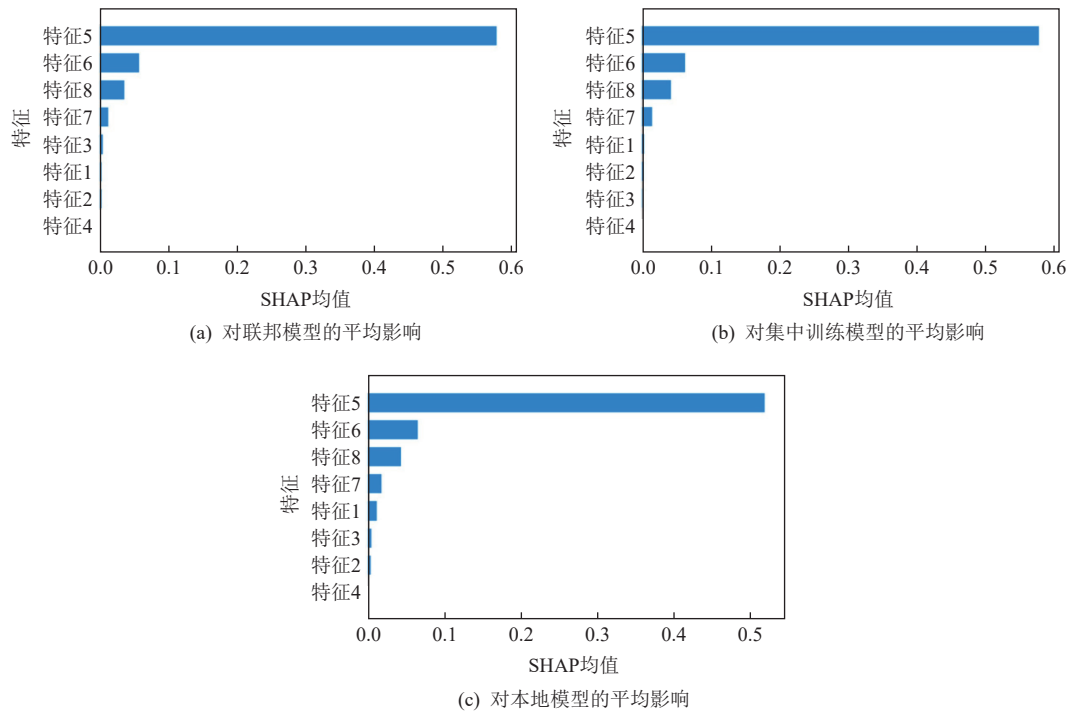


Fig. 5 SHAP analysis results of models obtained by different training methods

图5 不同训练方式获得模型的 SHAP 分析结果

Table 4 Contribution, Rank and AI Medical Promise of Each Participant

指标	P_1	P_2	P_3	P_4	P_5
c_i	18.3	0	100.0	22.2	0
r_i	75.0	40	90.0	75.0	20
$a^2mp_i(\varepsilon=0.4)$	13.2	0	89.6	11.9	0
$a^2mp_i(\varepsilon=0.5)$	12.6	0	89.6	6.1	0

去中心化更新计算. 在3个真实数据集上的比较结果证明了我们的性能评价函数 $Score(\cdot)$ 的有效性. $Score(\cdot)$ 由多个适用于医疗领域的性能指标组成, 可以更全面地评估模型的性能. 真实医疗数据实验结果证实了我们的 colback-learning 框架的稳定性, 能够选择出可靠的参与方并有助于提高全局模型性能, 最终获得的全局模型表现优秀且稳定.

未来, 我们将专注于提高框架的效率, 并考虑创建奖励策略, 以吸引和鼓励更多具有高质量医疗数据的数据方参与协作学习任务.

作者贡献声明: 陆枫提出了算法思路和实验方案; 李炜指导并修改了算法思路; 顾琳指导了算法思路并修改论文; 刘帅和王润衡负责完成实验并撰写论文; 任宇飞提供了数据并对数据分析和处理提供指导意见; 戴小海指导了区块链技术和设计思路; 廖

小飞和金海提出指导意见并修改论文.

参 考 文 献

- [1] Liu Yang, Huang Anbu, Luo Yun, et al. Fedvision: An online visual object detection platform powered by federated learning[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 13172–13179
- [2] Hard A, Rao K, Mathews R, et al. Federated learning for mobile keyboard prediction[J]. arXiv preprint, arXiv: 1811.03604, 2018
- [3] Li Jun, Wei Kang, Ma Chuan, et al. DP-GenFL: A local differentially private federated learning system through generative data[J]. *Science China Information Sciences*, 2023, 66(8): 189303
- [4] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Proc of the 20th Int Conf on Artificial Intelligence and Statistics. Brookline, MA: Microtome Publishing, 2017: 1273–1282
- [5] Warnat-Herresthal S, Schultze H, Shastri K L, et al. Swarm learning for decentralized and confidential clinical machine learning[J]. *Nature*, 2021, 594(7862): 265–270
- [6] Yang Qiang, Liu Yang, Chen Tianjian, et al. Federated machine learning: Concept and applications[J]. *ACM Transactions on Intelligent Systems and Technology*, 2019, 10(2): 1–19
- [7] Zhang Chen, Xie Yu, Bai Hang, et al. A survey on federated learning[J]. *Knowledge-Based Systems*, 2021, 216: 106775
- [8] Adnan M, Kalra S, Cresswell J C, et al. Federated learning and

- differential privacy for medical image analysis[J]. *Scientific Reports*, 2022, 12(1): 1953
- [9] Voigt P, Von-Dem-Bussche A. The EU General Data Protection Regulation (GDPR)[M]. Berlin: Springer, 2017: 152–166
- [10] Zhang D Y, Kou Ziyi, Wang Dong. Fedsens: A federated learning approach for smart health sensing with class imbalance in resource constrained edge computing[C]//Proc of the 40th IEEE Conf on Computer Communications. Piscataway, NJ: IEEE, 2021: 9488776
- [11] Cho Y J, Wang Jianyu, Joshi G. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies[J]. arXiv preprint, arXiv: 2010.01243, 2020
- [12] Zhang Jingwen, Wu Yuezhou, Pan Rong. Incentive mechanism for horizontal federated learning based on reputation and reverse auction[C]//Proc of the 30th Web Conf 2021. New York: ACM, 2021: 947–956
- [13] Tahir A, Chen Yongzhou, Nilayam P. FedSS: Federated learning with smart selection of clients[J]. arXiv preprint, arXiv: 2207.04569, 2022
- [14] Hall M J, Schwartzman A, Zhang Jin, et al. Ambulatory surgery data from hospitals and ambulatory surgery centers: United States, 2010[J]. *National Health Statistics Reports*, 2017(102): 1–15
- [15] Stephansen J B, Olesen A N, Olsen M, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy[J]. *Nature Communications*, 2018, 9(1): 5229
- [16] Ghorbani A, Zou J. Data Shapley: Equitable valuation of data for machine learning[C]//Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 2242–2251
- [17] Koh P W, Liang P. Understanding black-box predictions via influence functions[C]//Proc of the 34th Int Conf on Machine Learning. New York: ACM, 2017: 1885–1894
- [18] Yang Qiang. AI and data privacy protection: The way to federated learning[J]. *Journal of Information Security Research*, 2019, 5(11): 961–965
- [19] Chen Shaoqi, Xue Dongyu, Chuai Guohui, et al. FL-QSAR: A federated learning-based QSAR prototype for collaborative drug discovery[J]. *Bioinformatics*, 2020, 36(22/23): 5492–5498
- [20] Sheller M J, Edwards B, Reina G A, et al. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data[J]. *Scientific Reports*, 2020, 10(1): 1–12
- [21] Chen Yiqiang, Qin Xin, Wang Jindong, et al. Fedhealth: A federated transfer learning framework for wearable healthcare[J]. *IEEE Intelligent Systems*, 2020, 35(4): 83–93
- [22] Kim Y, Sun Jimeng, Yu H, et al. Federated tensor factorization for computational phenotyping[C]//Proc of the 23rd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2017: 887–895
- [23] Sheller M J, Reina G A, Edwards B, et al. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation[C]//Proc of the 4th Int Brain Lesion Workshop. Berlin: Springer, 2019: 92–104
- [24] Kumar R, Khan A A, Kumar J, et al. Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging[J]. *IEEE Sensors Journal*, 2021, 21(14): 16301–16314
- [25] Roth H R, Chang Ken, Singh P, et al. Federated learning for breast density classification: A real-world implementation[C]//Proc of the 23rd Int Conf on Medical Image Computing and Computer Assisted Intervention. Berlin: Springer, 2020: 181–191
- [26] McMahan H B, Moore E, Ramage D, et al. Federated learning of deep networks using model averaging[J]. arXiv preprint, arXiv: 1602.05629, 2016
- [27] Zeng Rongfei, Zhang Shixun, Wang Jiaqi, et al. FMore: An incentive scheme of multi-dimensional auction for federated learning in MEC[C]//Proc of the 40th IEEE Int Conf on Distributed Computing Systems. Piscataway, NJ: IEEE, 2020: 278–288
- [28] Nishio T, Yonetani R. Client selection for federated learning with heterogeneous resources in mobile edge[C/OL]//Proc of the 53rd IEEE Int Conf on Communications. Piscataway, NJ: IEEE, 2019[2023-02-26]. <https://ieeexplore.ieee.org/abstract/document/8761315>
- [29] Jia Ruoxi, Dao D, Wang Boxin, et al. Towards efficient data valuation based on the Shapley value[C]//Proc of the 22nd Int Conf on Artificial Intelligence and Statistics. Brookline, MA: Microtome Publishing, 2019: 1167–1176
- [30] Sim R H L, Zhang Yehong, Chan M C, et al. Collaborative machine learning with incentive-aware model rewards[C]//Proc of the 37th Int Conf on Machine Learning. New York: ACM, 2020: 8927–8936
- [31] Song Tianshu, Tong Yongxin, Wei Shuyue. Profit allocation for federated learning[C]//Proc of the 7th IEEE Int Conf on Big Data. Piscataway, NJ: IEEE, 2019: 2577–2586
- [32] Wang Guan, Dang C X, Zhou Ziyue. Measure contribution of participants in federated learning[C]//Proc of the 7th IEEE Int Conf on Big Data. Piscataway, NJ: IEEE, 2019: 2597–2604
- [33] Garcia-Ceja E, Riegler M, Jakobsen P, et al. Depresjon: A motor activity database of depression episodes in unipolar and bipolar patients[C]//Proc of the 9th ACM Multimedia Systems Conf. New York: ACM, 2018: 472–477
- [34] Gompertz B. On the nature of the function expressive of the law of human mortality and on a new mode of determining the value of life contingencies[J]. *Philosophical Transactions of the Royal Society of London*, 1825, 115: 513–583
- [35] Huang K L, Kanhere S S, Hu Wen. On the need for a reputation system in mobile phone based sensing[J]. *Ad Hoc Networks*, 2014, 12(SI): 130–149
- [36] He Haiwu, Yan An, Chen Zehua. Survey of smart contract technology and application based on blockchain[J]. *Journal of Computer Research and Development*, 2018, 55(11): 2452–2466 (in Chinese)

(贺海武, 延安, 陈泽华. 基于区块链的智能合约技术与应用综述[J]. *计算机研究与发展*, 2018, 55(11): 2452–2466)

- [37] Wesselink E M, Kappen T H, Torn H M, et al. Intraoperative hypotension and the risk of postoperative adverse outcomes: A systematic review[J]. *British Journal of Anaesthesia*, 2018, 121(4): 706–721
- [38] Kendale S, Kulkarni P, Rosenberg A D, et al. Supervised machine-learning predictive analytics for prediction of postinduction hypotension[J]. *Anesthesiology*, 2018, 129(4): 675–688
- [39] Hatib F, Jian Zhongping, Buddi S, et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis[J]. *Anesthesiology*, 2018, 129(4): 663–674
- [40] Wijnberge M, Geerts B F, Hol L, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: The HYPE randomized clinical trial[J]. *The Journal of the American Medical Association*, 2020, 323(11): 1052–1060
- [41] Lee H C, Jung C W. Vital recorder—A free research tool for automatic recording of high-resolution time-synchronised physiological data from multiple anaesthesia devices[J]. *Scientific Reports*, 2018, 8(1): 1527
- [42] Che Zhengping, Purushotham S, Cho K, et al. Recurrent neural networks for multivariate time series with missing values[J]. *Scientific Reports*, 2018, 8(1): 6085
- [43] Giles C L, Lawrence S, Tsoi A C. Noisy time series prediction using recurrent neural networks and grammatical inference[J]. *Machine Learning*, 2001, 44(1/2): 161–183
- [44] Karim F, Majumdar S, Darabi H, et al. LSTM fully convolutional networks for time series classification[J]. *IEEE Access*, 2017, 6: 1662–1669
- [45] Ismail-Fawaz H, Forestier G, Weber J, et al. Deep learning for time series classification: A review[J]. *Data Mining and Knowledge Discovery*, 2019, 33(4): 917–963
- [46] Tan Qingxiong, Ma A J, Deng Huiqi, et al. A hybrid residual network and long short-term memory method for peptic ulcer bleeding mortality prediction[C]//Proc of the 15th American Medical Informatics Association Annual Symp. Bethesda, MD: American Medical Informatics Association, 2018: 998–1007
- [47] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770–778
- [48] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[J]. *Advances in Neural Information Processing Systems*, 2017, 30: 4765–4774



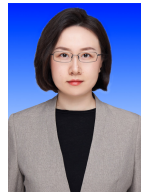
Lu Feng, born in 1971. PhD, associate professor. Her main research interests include parallel computing, big data, and blockchain.

陆 枫, 1971 年生. 博士, 副教授. 主要研究方向为并行计算、大数据、区块链.



Li Wei, born in 1983. PhD, PhD supervisor. His main research interests include distributed computing, sustainable computing, decision making, scheduling, and applied computing.

李 炜, 1983 年生. 博士, 博士生导师. 主要研究方向为分布式计算、可持续计算、决策、调度论、应用计算.



Gu Lin, born in 1985. PhD, associate professor, PhD supervisor. Her main research interests include cloud native computing, edge computing, and resource scheduling for the cloud-edge platforms.

顾 琳, 1985 年生. 博士, 副教授, 博士生导师. 主要研究方向为云原生计算、边缘计算以及云边融合平台的资源调度.



Liu Shuai, born in 1999. Master. His main research interests include big data, artificial intelligence, and federated learning.

刘 帅, 1999 年生. 硕士. 主要研究方向为大数据、人工智能、联邦学习.



Wang Runheng, born in 1998. Master. Her main research interests include medical big data and federated learning.

王润衡, 1998 年生. 硕士. 主要研究方向为医疗大数据、联邦学习.



Ren Yufei, born in 1985. Master, senior engineer. His main research interests include hospital information system, health information standard, and medical big data processing.

任宇飞, 1985 年生. 硕士, 高级工程师. 主要研究方向为医院信息系统、卫生健康信息标准、医疗大数据处理.



Dai Xiaohai, born in 1992. PhD. His main research interests include blockchain scalability and Byzantine consensus protocols.

戴小海, 1992 年生. 博士. 主要研究方向为区块链的可扩展性和拜占庭共识协议.



Liao Xiaofei, born in 1978. PhD, professor, PhD supervisor. His main research interests include system software, P2P system, cluster computing, and streaming services.

廖小飞, 1978 年生. 博士, 教授, 博士生导师. 主要研究方向为系统软件、P2P 系统、集群计算、流服务.



Jin Hai, born in 1966. PhD, professor, PhD supervisor. Fellow of CCF. His main research interests include computer architecture, virtualization technology, cluster computing, cloud computing, peer-to-peer computing, network storage, and network security.

金 海, 1966 年生. 博士, 教授, 博士生导师. CCF 会士. 主要研究方向为计算机体系结构、虚拟化技术、集群计算、云计算、P2P 计算、网络存储、网络安全.