

以咳嗽为第一主诉的儿童疾病医疗数据共享模型构建^{*}

朱志文 齐国强 杨恺熠 赵永根

余 卓

(1 浙江大学医学院附属儿童医院数据信息部 杭州 310052

(慧影医疗科技(北京)股份有限公司

2 浙江-芬兰儿童健康人工智能联合实验室 杭州 310052

北京 100192)

3 国家儿童健康与疾病临床医学研究中心 杭州 310052)

俞 刚

(1 浙江大学医学院附属儿童医院数据信息部 杭州 310052

2 浙江-芬兰儿童健康人工智能联合实验室 杭州 310052

3 国家儿童健康与疾病临床医学研究中心 杭州 310052)

〔摘要〕 分析以咳嗽为第一主诉的儿童疾病背景及其医疗数据应用情况和发展困境,结合联邦学习技术提出多中心横向联邦学习的数据安全共享框架,介绍3类联邦学习基础概念,从框架实施目的、系统架构组成、应用流程以及系统优点等方面详细阐述模型框架及构建路径。

〔关键词〕 儿童咳嗽;横向联邦学习;数据安全;人工智能

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2023.03.011

Construction of Children Cough as Chief Complaint Data Sharing Model ZHU Zhiwen, QI Guoqiang, YANG Kaiyi, ZHAO Yonggen, 1Department of Data and Information, The Children's Hospital, Zhejiang University School of Medicine, Hangzhou 310052, 2Sino-Finland Joint AI Laboratory for Child Health of Zhejiang Province, Hangzhou 310052, 3National Clinical Research Center for Child Health, Hangzhou 310052, China; YU Zhuo, Huiying Medical Technology (Beijing) Co. Ltd., Beijing 100192, China; YU Gang, 1Department of Data and Information, The Children's Hospital, Zhejiang University School of Medicine, Hangzhou 310052, 2Sino-Finland Joint AI Laboratory for Child Health of Zhejiang Province, Hangzhou 310052, 3National Clinical Research Center for Child Health, Hangzhou 310052, China

〔修回日期〕 2022-09-04

〔作者简介〕 朱志文,硕士;通信作者:俞刚,高级工程师,发表论文46篇。

〔基金项目〕 国家重点研发计划“以咳嗽为第一主诉症状的儿童疾病智能分析平台及应用”(项目编号:2019YFE0126200);国家自然科学基金面上项目“基于多模态融合的儿童慢性咳嗽人工智能辅助诊断技术研究”(项目编号:62076218);浙江省公益技术应用研究项目“数字化改革中儿童医疗数据智能分级方法研究及示范应用”(项目编号:LGF22H180004)。

[Abstract] The paper analyzes the background of children diseases with cough as chief complaint and the applications of its medical data as well as the dilemma of development. Combined with federated learning technology, it proposes a data security sharing framework of multi-center horizontal federated learning, introduces the basic concepts of three types of federal learning, and elaborates the model framework and construction path from the implementation purpose, system architecture composition, application process, system advantages, etc.

[Keywords] children cough; horizontal federated learning; data security; artificial intelligence (AI)

1 引言

以咳嗽为第一主诉的儿童呼吸系统疾病是儿童高发疾病之一, 占儿科总就诊率的 35%。诱发以咳嗽为第一主诉的儿童疾病病因十分复杂, 主要包括上呼吸道感染、下呼吸道感染、细菌性支气管炎、百日咳、气道发育异常、哮喘、异物吸入、胃食道返流以及囊性纤维化等数十种^[1]。其诊断复杂, 加上医疗资源分布不均、数据孤岛现象严重、医患交互困难等影响因素, 快速准确地诊断治疗该疾病已成为国内外儿童医学的热点和难点。目前, 对于以咳嗽为第一主诉的儿童疾病诊断比较常规的方法包括: 超敏 C-反应蛋白 (C-reactive protein, CRP) 测定、血常规、生化五类、胸部 X 片、胸部电子计算机断层扫描 (computed tomography, CT) 检查、肺通气功能检查、呼出气一氧化氮检测 (fractional exhaled nitric oxide, FeNO)、过敏原检测、支气管镜检查等, 产生的医疗数据高度异构、种类繁多^[2]。

随着电子病历的推广, 以咳嗽为第一主诉的儿童疾病医疗数据量快速增长, 出现两大挑战。一是大部分以咳嗽为第一主诉儿童疾病的医疗数据以孤立形式存在于各个医疗机构, 缺乏一致性、质量参差不齐。二是随着数据的爆炸式增长, 需要加强医疗数据安全和隐私保护。欧盟于 2018 年 5 月正式颁布实施的《通用数据保护法案》强调对儿童信息和个人生物学数据的重点保护^[3-4]。我国也在不断修正和完善信息数据管理方面的法律法规, 先后颁布《数据安全法 (草案)》和《信息安全技术个人信息安全规范》^[5]。横向联邦学习可以在不侵犯患者数据隐私的前提下, 根据各个医疗机构数据的分布特点, 对来自多方的医疗数据进行模型训练, 迭代

计算收敛全局损失函数直至最优。因此, 本文提出基于多中心横向联邦学习的方法, 以期为解决以咳嗽为第一主诉的儿童疾病医疗数据孤岛及安全保护等问题提供参考。

2 研究现状

2.1 研究对象

本研究以浙江大学医学院附属儿童医院 (国家儿童健康与疾病临床医学研究中心及国家儿童区域医疗中心依托单位) (以下简称浙大儿院) 为主要依托单位开展相关研究。浙大儿院小儿呼吸专业是国家临床重点专科, 为落实分级诊疗政策、提高传染病精准防控能力、建设现代化智慧医疗体系, 以浙大儿院为主中心、省内其他各级医疗机构为分中心, 基于多中心横向联邦学习技术, 探究以咳嗽为第一主诉的儿童疾病医疗数据共享模型。

2.2 联邦学习

2.2.1 联邦学习技术 是一种新兴的机器学习框架, 能有效帮助各机构在遵守法律法规、不交换彼此原数据的前提下进行联合训练建模^[6]。相比传统机器学习方法, 联邦学习能更有效地解决数据安全、数据孤岛以及用户隐私等问题^[7]。微众银行人工智能团队在 2018 年打造了首个工业级的联邦学习技术框架 (federated AI technology enabler, FATE)^[8], 并将联邦学习技术应用到金融、工业、医疗等行业形成联邦生态, 为下一代可信、可判断、可预测、可决策的安全人工智能系统奠定基础。根据数据源分布特点, 联邦学习可分为 3 类: 纵向联邦学习、联邦迁移学习和横向联邦学习。

2.2.2 纵向联邦学习 根据参与者特性不同进行

分布式机器学习，适用于各中心数据集相同或者相似但其特征不同的情况。具体来讲，同一地点不同领域的企业或者组织之间数据空间存在重叠，但是其数据的特征空间交叉较少，纵向联邦学习通过联合计算不同企业或组织数据不同特征来保护用户数据隐私的方式收敛全局损失函数、建立全局模型，见图 1（a）。

2.2.3 联邦迁移学习 是纵向联邦学习中的一种特殊情况。纵向联邦学习与联邦迁移学习都适用于各中心的数据特征重叠较少的情况，但联邦迁移学习更适用于参与者之间没有相同或者相似的样本，且其特征存在较少重叠的情况。具体来讲，在不同领域数据里面找到相同的变量，是进行迁移学习的核心。所以联邦迁移学习目的是在不同数据集中寻找共同点，见图 1（b）。

2.2.4 横向联邦学习 根据样本的不同进行分布式机器学习，适用于参与者拥有相同或者相似特征

空间，具有不同样本空间的数据情况，见图 1（c）。针对个人安卓手机终端，美国谷歌公司在训练本地用户数据解决方案的模型时，首次提出联邦学习概念^[9]，利用客户端-服务器基础框架，聚合分散数据进行模型训练，是横向联邦学习的雏形。手机客户端在中心服务器的协作下，回传模型参数进行集中联合训练，而手机客户端无需上传或共享数据到中心服务器，始终持有本地数据的支配权。相比传统机器学习方式，一方面加强了数据的安全性和私密性，另一方面减少汇聚数据所造成的巨额通信开销。横向联邦学习主要有 4 个步骤^[10]。第 1 步：各分中心分别从中心服务器下载模型，并将模型参数发送给中心服务器。第 2 步：中心服务器聚合分中心模型参数。第 3 步：中心服务器将聚合模型参数回传给分中心。第 4 步：分中心从中心服务器下载结果，迭代更新模型。

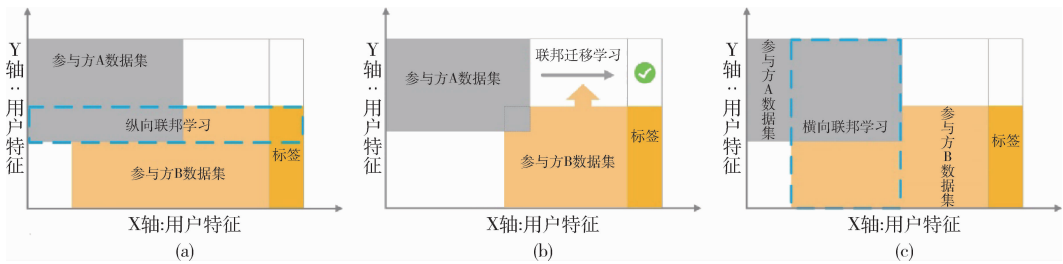


图 1 3 类联邦学习

2.3 儿童咳嗽医疗数据问题

以咳嗽为第一主诉的儿童疾病是儿科门诊就诊率最高的疾病类型。除病种多样、病情复杂造成的诊断困难之外，我国儿科医疗资源严重不足，尤其是高水平儿科医生匮乏也加剧了疾病误诊率较高、医疗资源浪费严重、患儿及其家庭负担沉重等问题^[11]。近年来，随着全国各地电子病历系统的推广，医疗数据被广泛记录在各医疗机构中。通过检查、检验产生的大量结构化、半结构化和非结构化数据^[12]存储在各医疗机构的本地服务器中，导致大部分咳嗽数据逐渐成为数据孤岛。即使在多家医疗机构联合成立以咳嗽为第一主诉的儿童疾病的研究中心，并达成协议确保患者隐私安全的前提下，想

要共享彼此数据来训练模型，依然面临各数据平台大小不一，咳嗽数据缺乏统一标准等问题^[13]。同时，不同检验、检查设备的多样性进一步加剧训练数据集的困难。为了解决该困境，本文提出横向联邦学习，其可以将所有独立的医疗机构及检验、检查设备的数据采用分布式存储结构联合起来，共享有隐私保障的临床数据，实现数据共享模型更加精确、泛化、安全。

3 基于多中心横向联邦学习的以咳嗽为第一主诉的儿童疾病数据共享框架

3.1 模型选择

多中心横向联邦学习数据共享模型可以实现在

保障各医疗机构数据不出院的情况下，打破数据壁垒，获得海量医疗数据，并进行高质量的医学数据运用。目的是提高诊疗准确性，更早进行医疗干预，从而减少儿童慢性咳嗽的发病率，提供高质量的多中心医疗数据。多中心医疗数据是指各种医疗机构（医院、卫生院、诊所、学校医务室等）各自职能和功能差异造成就医者分布不同、患者重叠率较低，进而形成具有不同样本空间的数据。若把以咳嗽为第一主诉当作患者的特征，将患者年龄限定在 0~18 岁，那么在不同医疗机构的数据集中就会出现较大的特征重叠。鉴于以上情况，相比纵向联邦学习和联邦迁移学习，横向联邦学习更加适合以咳嗽为第一主诉的儿童疾病医疗数据共享模型。

3.2 系统模型组成及训练流程（图 2）

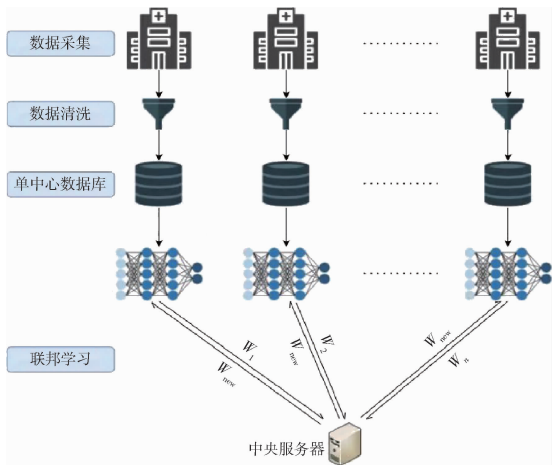


图 2 多中心横向联邦学习机制

3.2.1 多中心数据基层 多中心数据基层主要由各医疗机构组成。所有开展疾病诊断、治疗活动的卫生机构都能成为以咳嗽为第一主诉的儿童疾病模型数据提供方。每个医疗机构负责采集、接收患者就诊信息，包括患者电子病历、基本信息、病史情况、处方信息、检查数据、影像数据等原始数据。独立医疗机构扮演本地数据中心的角色，以便更广泛地收集年龄在 0~18 岁以咳嗽为第一主诉的儿童疾病临床数据并进行单独训练。

3.2.2 数据清洗层 以咳嗽为第一主诉的儿童疾病数据质量关系到最终模型精确度。因此，数据清

洗层最为关键的环节在于每个医疗机构根据统一清洗规则，对原始数据进行整合，清除错误、重复和缺失数据，进而提高数据质量。首先在各医疗机构中提取以咳嗽为第一主诉的患者信息，再通过年龄阈值设定获取 0~18 岁患儿电子病历、患者基本信息、病史、处方信息、病理报告、影像、检验结果等相关诊疗数据。对错误、重复、缺失数据进行统一判断，删除脏数据。数据清洗使多维异构数据结构实现标准化和归一化，提升基层数据整体质量，为高效采集和利用打下基础^[12]，清洗后的数据存储于各医疗机构单中心数据库中。

3.2.3 横向联邦学习层 该层主要为医疗机构之间数据共享提供安全保障。联邦学习层中的中央服务器提供共享模型参数并分配给各医疗机构，各医疗机构依据本地以咳嗽为第一主诉的儿童疾病医疗数据进行学习。经过初步训练各医疗机构将得到各自新的模型参数，并回传至中央服务器。上述过程中，医疗机构单中心数据库原数据无需上传至中央服务器即可共同训练全局共享模型。模型训练过程分为 3 步。第 1 步：中央服务器根据临床实际应用下发所训练的模型参数。第 2 步：医疗机构接收到训练任务后，根据各自采集的医疗数据进行本地学习，寻找各自的模型参数，在此过程中，模型参数会产生新值 W_n 。由于各医疗机构数据不同，因此得到的模型参数 W_n 也不同。第 3 步：中央服务器端收集各医疗机构模型参数，并根据更新的模型参数 $W_1, W_2, \dots, W_n$ 进行算术平均得到 W_{new} ，中央服务器再将聚合后的模型参数 W_{new} 回传给每个医疗机构，医疗机构用 W_{new} 开始新一轮学习。训练重复此过程，直到最终联邦学习模型准确性达到期望值为止。

3.3 系统优点

3.3.1 泛化性强 在传统统计学理论中，模型参数越多泛化能力会越差。模型如果被训练得太过拟合，往往泛化性差，甚至无法泛化。通过多中心共建以咳嗽为第一主诉的儿童疾病医疗数据横向联邦学习共享模型，可以避免全病种模式下模型参数冗杂的问题，大幅提升机器学习模型泛化性^[14]。

3.3.2 扩展性高 系统基于多中心横向联邦学习

场景,大量医疗数据能通过扩展数据接口从不同医疗机构获取,除儿童专科医院外还包括大型综合医院以及社区医院等,数据源呈现多样性,提升了系统的扩展性。

3.3.3 安全、稳定性良好 系统中的单中心数据库都是相互独立的。联邦学习全流程中,中央服务器端、医疗机构之间都不存在原始数据的传输与共享,互不干扰、互不联系。所以全局模型不会因医疗机构的数据量不平衡而受到影响。同时,医疗机构对自身原始数据拥有绝对控制权,只要加强本地网络安全建设,无需承担数据泄露风险。

4 结语

本文针对以咳嗽为第一主诉的儿童疾病健康数据孤岛和数据隐私安全问题,提出基于多中心横向联邦学习的数据共享模型。该模型通过多中心化方式联合解决数据孤岛问题,运用横向联邦学习技术打通数据壁垒。最终实现不受地域限制,即患者数据不出本地,仅通过聚合模型参数并迭代更新,就得出更为准确的全局模型。在智慧医疗领域中,由于横向联邦学习还处于起步阶段,前期容易受到医疗机构参与意愿、本地计算能力、量级数据储存受限及医疗机构网络安全环境等因素的影响。对于上述问题,后续可以设计相应的医疗机构数据贡献激励机制^[15]。各医疗机构在参与使用本地以咳嗽为第一主诉的儿童疾病数据联邦学习时,能够反馈其在联邦学习机制上的成效,并获得一定收益。例如对医疗机构的数据质量和数量进行评估,并合理分配收益。合理的数据贡献评估能激励更多医疗机构加入横向联邦学习,构建以咳嗽为第一主诉的儿童疾病数据模型“共同富裕”联盟。未来将在本共享模型的基础上,拓展构建儿童医疗数据共享模型,助力卫生健康数字化改革中儿童医疗数据的安全共享。

参考文献

- 王晓丽,郑兴厂,管栋,等. 青岛地区儿童慢性湿性咳嗽病因分析 [J]. 中国小儿急救医学, 2021, 28 (4): 304 - 307.
- 罗征秀,李卫国,李沁原,等. 中国儿童咳嗽诊断与治疗临床实践指南 (2021 版) 计划书 [J]. 中华儿科杂志, 2021, 59 (1): 10 - 13.
- 廖博宇. 欧盟《通用数据保护法案》对数据主体权利的保护及其影响 [J]. 中德法学论坛, 2019 (2): 3 - 22.
- EU. Genral date protection regulation [EB/OL]. [2021 - 05 - 23]. <https://gdpr-info.eu/>.
- 冯诗宇,叶泳琪,郑承华. 移动互联网应用中个人信息安全规范探析 [J]. 网络安全技术与应用, 2021 (11): 82 - 83.
- 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.
- 冯霖,蔡其志,姜远. 联邦学习下对抗训练样本表示的研究 [J]. 中国科学 (信息科学), 2021, 51 (6): 900 - 911.
- 杨强. AI 与数据隐私保护: 联邦学习的破解之道 [J]. 信息安全研究, 2019, 5 (11): 961 - 965.
- MCMAHAN B, MOORE E, RAMAGE D, et al. Communication - efficient learning of deep networks from decentralized data [C]. Fort Lauderdale: International Conference on Artificial Intelligence and Statistics, 2017.
- KONEN J, MCMAHAN H B, YU F X, et al. Federated learning: strategies for improving communication efficiency [EB/OL]. [2022 - 05 - 04]. <https://arxiv.org/pdf/1602.05629v3.pdf>.
- 马光举. 医疗机构儿科医疗资源现状与对策 [J]. 中国农村卫生, 2017 (18): 71.
- 张靓,杨志清,王海星,等. 基于 Hadoop 的医疗科研大数据平台的构建与应用 [J]. 中国数字医学, 2020, 15 (10): 82 - 84.
- 邢丹,徐琦,姚俊明. 边缘计算环境下基于区块链和联邦学习的医疗健康数据共享模型 [J]. 医学信息学杂志, 2021, 42 (2): 33 - 37.
- 陈瑞锋,谢在鹏,朱晓瑞,等. 一种异步联邦学习聚合更新算法 [J]. 小型微型计算机系统, 2021, 42 (12): 2473 - 2478.
- 叶东东. 分层边缘联邦学习: 一种基于契约论的双层激励机制 [J]. 工业控制计算机, 2022, 35 (1): 104 - 105, 108.