联盟学习在生物医学大数据隐私保护中的原理与应用*

窦佐超 陈 峰 邓杰仁 陈如梵

（1 四川大学华西医学院系统遗传研究所 成都 610041
2 杭州锘崴信息科技有限公司 杭州 310053）

（杭州锘崴信息科技有限公司 杭州 310053）

（1 杭州锘崴信息科技有限公司 杭州 310053
2 昆士兰大学医学院 澳大利亚 昆士兰州布里斯班 4006）

郑 灏 孙 琪 谢 康 沈百荣

（杭州锘崴信息科技有限公司 杭州 310053）

（中华人民共和国公安部第三研究所信息网络安全国家重点实验室 上海 200041）

（四川大学华西医学院系统遗传研究所 成都 610041）

王 爽

（1 四川大学华西医学院系统遗传研究所 成都 610041 2 杭州锘崴信息科技有限公司 杭州 310053
3 美国印第安纳大学信息计算和工程学院 美国印第安纳州伯明顿 47495
4 同济大学上海普陀区人民医院 上海 200060）

〔摘要〕 结合2020 年初新型冠状病毒疫情探讨数据分享与联合分析的必要性和紧迫性，系统介绍联盟学习技术原理和适用范围，根据不同数据类型特点详细阐述当前联盟学习技术在生物医疗大数据隐私保护中的应用及其与深度学习技术的结合。

〔关键词〕 生物医学大数据；隐私保护；联盟学习；隐私计算；新型冠状病毒
〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2020.05.001

Principles and Applications of Federated Learning in Biomedical Big Data Privacy Protection *DOU Zuochao, CHEN Feng,*

〔收稿日期〕 2020－02－18

〔作者简介〕 窦佐超，博士，研究员，发表论文8 篇；通讯作者：王爽，博士，研究员，教授，发表论文100 余篇。

〔基金项目〕 企业级科研业务费项目“杭州锘崴信息科技有限公司医疗大数据隐私保护的研究”（项目编号：NV202001）；医院级科研业务费项目“华西医院系统遗传研究所生物大数据隐私研究”（项目编号：201906WCH）；公安部第三研究所信息与网络安全重点实验室基金“多中心大数据分享和分析中的隐私保护计算与标准”（项目编号：C19609）。

1Institutes for Systems Genetics, West China Hospital, Chengdu 610041, China, 2Hangzhou Nuowei Information Technology Co., Ltd., Hangzhou 310053, China; DENG Jieren, Hangzhou Nuowei Information Technology Co., Ltd., Hangzhou 310053, China; CHEN Rufan, 1Hangzhou Nuowei Information Technology Co., Ltd., Hangzhou 310053, China, 2School of Medicine, University of Queensland, Brisbane, QLD 4006, Australia; ZHENG Hao, SUN Qi, Hangzhou Nuowei Information Technology Co., Ltd., Hangzhou 310053, China; XIE Kang, Key Lab of Information Network Security, The Third Research Institute of Ministry of Public Security, Shanghai 200041, China; SHEN Bairong, Institutes for Systems Genetics, West China Hospital, Chengdu 610041, China; WANG Shuang, 1Institutes for Systems Genetics, West China Hospital, Chengdu 610041, China, 2Hangzhou Nuowei Information Technology Co., Ltd., Hangzhou 310053, China, 3Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47495, USA, 4Shanghai Putuo Hospital, Tongji University, Shanghai 200060, China

〔**Abstract**〕 Combined with COVID－19 epidemic in early 2020, the paper discusses the necessity and urgency of data sharing and conjoint analysis. It introduces the principles and application scope of federated learning technology systematically, and elaborates on the current application of federated learning technology in biomedical big data privacy protection and its combination with deep learning technology according to the characteristics of different data types in detail.

〔**Keywords**〕 biomedical big data; privacy protection; federated learning; privacy computing; COVID－19

1 引言

1.1 研究背景

在生物医疗大数据分析研究过程中数据样本的规模对于分析结果的适用性和准确性有很大影响。近年来全球暴发了各种新型传染病疫情，如2020年初暴发的新型冠状病毒（COVID－19）、2012年的中东呼吸综合征（MERS）以及2003年的重症急性呼吸综合征（SARS）等，这些病毒都具有非常强的传染性和致病性。人体在感染这些病毒后会产生非常严重的后果，如COVID－19感染者中主要的病症是往往伴随着肺炎及其他并发症[1]，如急性呼吸窘迫综合征（呼吸困难）、败血症、急性心脏损伤和其他感染，严重者甚至会死亡。不仅如此病毒传播和扩散会危害公共安全，同时造成不可估量的经济损失。在新型冠状病毒疫情初期，大量研究对其进行病毒学、病理学、药物、流行病学分析，提出各种假说和推断。但受限于样本数据的规模无法得出确切结论。Chan 等根据一个6口之家的聚集性发病案例推测COVID－19存在人际传播的可能性，但需要更多研究和样本来进一步确认[2]。De Wit 等强调目前仍不清楚COVID－19是否会像SARS一样使免疫细胞渗透入肺部从而导致严重的肺部损伤。因此理解肺部微循环和免疫系统对于该病毒的反应非常重要且迫切，而这可能会为确认患者病情程度以及理清发病机理带来突破点[3]。Huang 等以及Chen 等同时报告COVID－19患者的临床症状及流行病学特征[1,4]，且结果高度相似，其样本量分别为41和99。值得注意的是两份样本都显示呼吸衰竭为导致患者死亡的主要原因，但是目前并未有研究报告死亡患者的肺部病理情况，对于有关信息或研究的需求应引起更多重视。因此有关新型冠状病毒的病理性和流行病学研究对遏制其传播和暴发有重大意义。Zulma 等指出由于有关COVID－19特效药的情况目前尚不明朗，且研制新药可能需要几年的时间，应将目光投向现有、已知安全的药物或是疗法上。目前全球已有多项药物研究正在进行中，研究人员正尝试从已有的抗病毒药物中发现可能存在的特效药，包括瑞德西韦（Remdesivir）在内的若干种药物被寄予厚望，且很有可能有效抑制病毒[5－7]。此外宿主定向疗法及细胞疗法等放大或促进患者自身免疫系统活动的方法也被认为能有效对抗该病毒[8]。

1.2 数据共享与联合分析必要性

除传统的针对病毒本身和其引发疾病的研究外，还有一个方面应该引起人们的注意，且在必要

时提高至最高优先级，那就是数据的共享和整合[9-10]。例如 Heymann 在国际著名医学杂志《柳叶刀》（*Lancet*）发表的文章指出有关病毒学、病理学、药物、流行病学等方面的研究和数据等于是拼图的碎片[9]。只有将所有信息整合在一起才能使这些碎片拼凑起来成为完整的图画，而这幅完整的拼图是成功控制此次疫情的关键。如何才能快速、安全、高效地将这些碎片拼起来则是即将要解决的难题。针对新型冠状病毒疫情，将零散信息进行有效共享并联合运用起来才能使疫情防控工作变得更加高效和准确。当新的患者被确诊时相关部门可以立刻对其活动范围进行历史轨迹分析（手机运营商位置、交通运输购票、购物、金融消费信息等），以便找出密切接触者及可能被感染的人群，从而控制传染源，避免疫情进一步扩散。另一方面，当发现潜在感染对象但缺乏进一步诊断技术手段情况下（基层医院基础设施不足，缺乏有经验的医生），可以根据这些整合的信息快速建立传染模型，通过多维度的数据分析发现更简单有效的诊断方法。同时结合后期的医疗记录分析，以数据为基础形成更好的治疗应对方案。更进一步，通过大数据技术可以建立传染关系，分析不同传染路径下病毒变异程度和病情监控，建立知识图谱，挖掘传染关联关系，分析流行病学传染特性，精细化病毒变异关联关系以及基础病关联关系和治疗方案，为药物研究、未来风险防范奠定坚实的数据基础。此外基于大数据分析可以对整个社会的传染风险进行分析。识别病毒毒株的特性，识别出高中低不同风险等级的地区、交通路线、人员行为。各单位可以通过这些信息评估自身状况，在决策部门统一决策的基础上根据自身业务特点采取针对性的应对措施，尽早复工。

综上所述，现代生物医学研究中只有大量的多方、多源大数据才能支持研究模型预测的高适用性和高准确性。然而目前的生物医疗数据和个人隐私息息相关，各个国家和地区均对个人隐私数据的收集、传播和使用进行严格立法。例如我国的《中华人民共和国网络安全法》，《信息安全技术个人信息去标识化指南》以及《信息安全技术个人信息安全规范》，美国的《医疗电子交换法案》，欧盟的《通用数据保护条例》以及较早的《数据保护指令》等。如何合理地保护个人敏感信息，在隐私信息不被泄露前提下又能有效地进行生物医疗数据分享、联合分析以及多元医疗数据融合，成为目前医学信息领域重点研究课题之一。

本文将结合信息科学中的大数据处理、联盟学习、信息分享，计算机科学中的安全计算、隐私保护算法，以及生物医学领域的电子病例、基因和图像数据分析处理和多中心合作的实际问题，介绍具有隐私保护的基于深度学习的多中心合作方法。其中包括利用联盟学习通过分享中间结果（如本地数据统计值、模型参数等）而不分享原始数据的特点实现基于隐私保护的跨机构大数据分析合作；利用隐私计算进行多机构间加密统计数据的分享和分析，保证预测的准确性以及对各方数据的隐私安全保护；采用深度学习技术进行样本测序数据特性分析。

2 联盟学习原理

2.1 结构模式

2.1.1 概述 联盟学习通过分享中间结果（如本地数据统计值、模型参数等）而不分享原始数据的特点实现基于隐私保护的跨机构大数据分析合作。支持大量基于隐私计算和分析的分布式算法，十分契合现代生物医疗大数据跨机构联合分析计算的需求[11-14]，同时保证分析预测的准确性以及对患者数据的隐私安全保护。一般认为联盟学习有两种基本结构：客户端/服务器和去中心化。

2.1.2 客户端/服务器模式 用来预测全局模型参数并进行统计学检验。客户端基于本地数据计算中间结果发送到服务器，服务器根据所有客户端中间结果不断迭代优化全局模型参数，见图 1。该模式下患者敏感数据一直保存在客户端本地，只有中间结果被共享到服务器进行安全分析和计算。所有客户端之间不直接通信和数据传输，客户端和服务器之间只传输中间结果，中间结果不涉及任何原始数据信息，其传输可使用安全传输层协议进一步保

护。值得注意的是联盟学习模式下对于全局模型参数的预测准确性和传统数据集中处理方式一致。联盟学习支持分布式或多类别逻辑回归，模型参数和方差－协方差矩阵可以通过迭代的中间结果推算出来。以基于最大似然估计的分布式二分逻辑回归算法为例，每个客户端计算并发送到服务器的中间结果为通过本地数据计算的海森矩阵（Hassian Matrix，全局水平分割数据）或核矩阵（Kernel Matrix，全局垂直分割数据）。结合模型参数和方差－协方差矩阵，联盟学习模型可进一步预测各种逻辑回归算法的统计学参数，如置信区间、标准误差、Z－检验以及 p 值检验等。在生物医学领域出于隐私保护考虑，患者原始数据和分析预测结果不能泄露。在联盟学习模式下以拟合优度检验为例，需要交换的中间数据只包括每个客户端的总体数据量和每 10 人中阳性患者数量[15]，从而有效保护患者隐私。

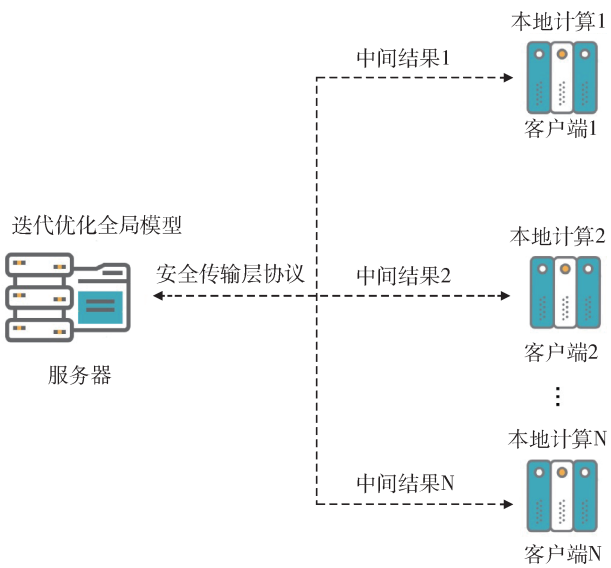


图 1　联盟学习的客户端/服务器模式

2.1.3　去中心化模式　无需中心服务器，所有客户端只和相邻客户端进行通信，与客户端/服务器模式相比节省部分通信消耗。通信分享内容同样为中间结果，保证原始数据安全性。具体来讲，每次迭代计算过程中每个客户端根据本地原始数据和相邻客户端发送的中间结果计算本次中间结果，见图 2。去中心化联盟学习模式下相邻客户端通过不断交换本地模型参数进而获得全局模型参数，避免原始数据（包含患者敏感隐私数据）泄露。去中心化模式被广泛应用到各种分布式算法中，如稀疏线性回归、主成分分析以及向量支持机等[16－18]。

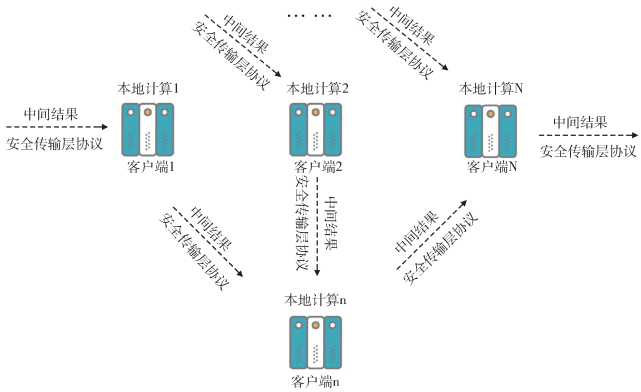
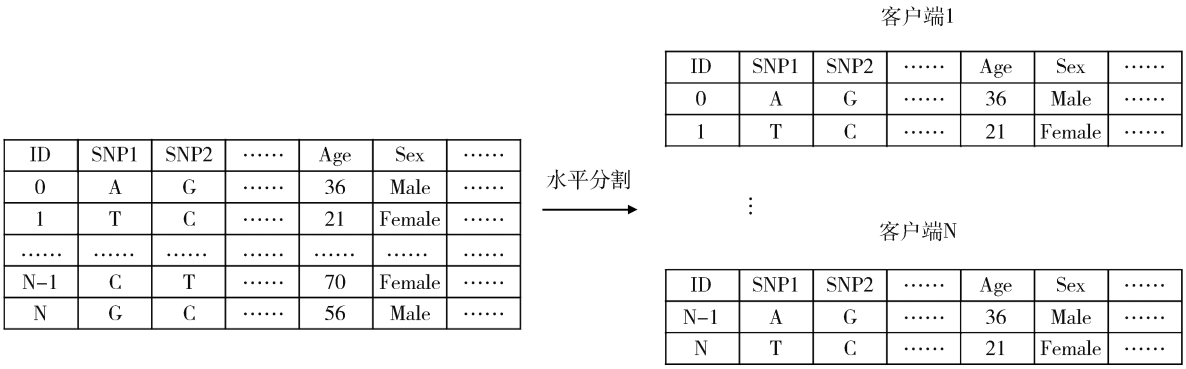


图 2　联盟学习的去中心化模式

2.2　数据分割模式

联盟学习框架下多中心数据分割分为两种模式：水平分割和垂直分割。使用患者生物基因数据的水平分割和垂直分割多中心数据模型，见图 3、图 4。在水平分割模式下每个客户端拥有不同患者所有类型生物基因数据（SNP 数据，年龄、性别等），适用于不同地域多中心生物医疗大数据联合分析的应用场景，每个客户端拥有不同的患者群体，每个患者具有相同的数据类型和属性列表。多个数据客户端联盟学习增大数据总量，研究结果更具有普遍性。而在垂直分割模式下每个客户端拥有所有患者部分数据。该模式适合不同客户端拥有相同患者不同数据的情景，联盟学习可以融合患者各种不同数据，使得联合分析预测的自变量域更加丰富，结果更加准确。例如联盟计算在 PCORnet 临床数据研究网络中被广泛采用[19]，其网络覆盖医疗福利部门、保险公司、健康系统以及退伍军人事务部，各个部门拥有超过 3 100 万患者的不同类型数据。

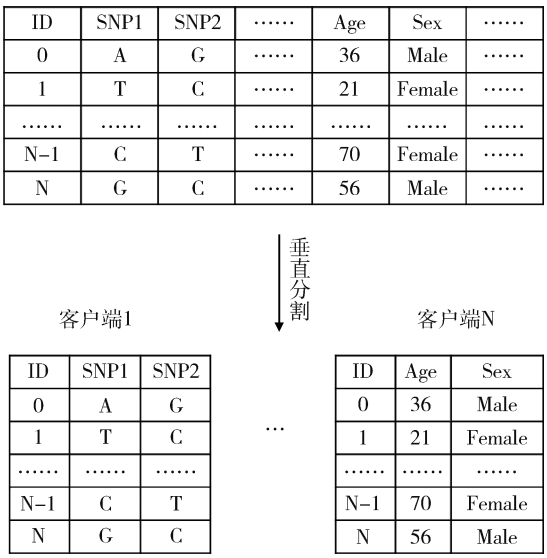


ID	SNP1	SNP2	……	Age	Sex	……
0	A	G	……	36	Male	……
1	T	C	……	21	Female	……
……	……	……	……	……	……	……
N-1	C	T	……	70	Female	……
N	G	C	……	56	Male	……

ID	SNP1	SNP2	……	Age	Sex	……
0	A	G	……	36	Male	……
1	T	C	……	21	Female	……

ID	SNP1	SNP2	……	Age	Sex	……
N-1	A	G	……	36	Male	……
N	T	C	……	21	Female	……

图3 水平分割多中心数据模型



ID	SNP1	SNP2	……	Age	Sex	……
0	A	G	……	36	Male	……
1	T	C	……	21	Female	……
……	……	……	……	……	……	……
N-1	C	T	……	70	Female	……
N	G	C	……	56	Male	……

ID	SNP1	SNP2
0	A	G
1	T	C
……	……	……
N-1	C	T
N	G	C

…

ID	Age	Sex
0	36	Male
1	21	Female
……	……	……
N-1	70	Female
N	56	Male

图4 垂直分割多中心数据模型

3 联盟学习应用

3.1 概述

联盟学习被广泛应用于生物医疗大数据分析、计算和隐私保护中，基于不同数据类型如结构化与非结构化电子病历、基因和图像数据对各种不同算法和模型在联盟学习框架下的应用进行介绍和分析，见表1。

表1 基于联盟学习架构的研究应用

文献号	应用方法	应用场景	数据类型	数据分割
[20]	牛顿-拉弗森方法	二元逻辑回归	结构化数据	水平分割
[21]	牛顿-拉弗森方法	多元逻辑回归	结构化数据	水平分割
[22]	牛顿-拉弗森方法	二元逻辑回归	结构化数据	垂直分割
[14]	牛顿-拉弗森方法	生存分析模型	结构化数据	水平分割
[13]	贝叶斯方法	贝叶斯逻辑回归	结构化数据	水平分割
[23]	ADMM	逻辑回归	结构化数据	水平/垂直分割
[18]	ADMM	支持向量机器	结构化数据	水平分割
[11]	多方安全计算	二元逻辑回归	结构化数据	水平分割
[24]	Word2Vec	文本预测	非结构化数据	水平分割
[25]	哈希编码	相似性度量	非结构化数据	水平分割
[26]	多方安全计算	基因频率统计	基因数据	水平分割
[27]	加密硬件计算	族谱学研究	基因数据	水平分割
[28]	加密硬件计算	罕见病统计假设检测	基因数据	水平分割
[29]	加密硬件计算	数据外部托管和查询	基因数据	水平分割
[30]	随机梯度下降	基于梯度的神经网络优化	医学图像数据	水平分割
[31]	随机梯度下降	基于梯度的神经网络优化	医学图像数据	水平分割

3.2 结构化电子病历数据

3.2.1 概述 结构化电子病历使用标准的数据格式对所有患者采集相同信息，将这些信息通过标准化的数据模型进行存储，包括但不限于个人生物信息、过敏信息、药物使用、吸烟情况、家庭健康史和化验测试结果等。结构化电子病历有利于患者信息的标准化收集、传播和使用，不同医疗健康机构

可以高效地进行跨机构数据联合分析。

3.2.2 基于牛顿-拉弗森方法 现代计算中涉及大量的工程优化计算任务，这些任务往往无法直接采用常见公式进行求解。利用计算机最擅长的数字计算优势，引入一种基于数值的方法——牛顿-拉佛森方法求解优化问题的近似解。牛顿-拉佛森方法的基本思想是利用迭代点处的一阶导数（梯度）和二阶导数（Hessen 矩阵）对目标函数进行二次函数近似，然后将二次模型的极值点作为新迭代点，不断重复这一过程直到满足精度的近似极值。早期的联盟式二元逻辑回归模型[20]通过分享模型但不分享患者数据保护数据安全性和隐私性，通过 Hosmer-Lemeshow 检验和 AUC 检验其准确性与传统中心化的模型相比并无差异。通过扩展研究，联盟式多元线性回归[21]，即使因变量是多种类别，牛顿-拉佛森方法在联盟学习框架下可重新设计并取得良好效果，相比于传统中心化的模型具有同样的准确性。相对于上述两种逻辑回归方式是在水平分割的数据上应用，即各个分布式的数据中心提供的是不同患者同一属性数据，垂直分割数据的联盟式逻辑回归[22]则是应用牛顿-拉佛森方法对各个分布式的数据中心的同一批患者不同属性数据进行训练。同样，该方法通过优化表现出和传统中心化逻辑回归一样的准确性。除逻辑回归外，生物医学上经常出现的生存分析也可以通过牛顿-拉佛森方法进行分布式应用。半参比例风险模型便是生成分析中被广泛关注和应用的模型[14]。针对水平分割的患者数据，基于服务器-客户端架构的 WebDISCO 系统通过服务器和客户端之间交换不敏感的中间值，不仅保护患者隐私，而且和其他生存分析对比展现出等同于传统中心化生存模型的准确性。

3.2.3 贝叶斯方法 受应用牛顿-拉佛森法的联盟式逻辑回归启发，针对不同隐私保护应用场景，一系列相关扩展和提升方法被广泛研究。期望传播逻辑回归模型就是其中的一项成果[13]。该模型基于统计学领域的贝叶斯方法研究并拓展，通过期望传播去最大化一个后验概率的预测，期望传播逻辑回归基于部分后验函数不断更新其参数。为避免中间结果信息泄露，期望传播逻辑回归模型的本地客户端和全局服务器之间交换的是加密的后验分布系数而不是模型参数预测的中间结果。因此不仅在模型训练时过程安全可靠，敏感患者数据也不会被泄露。

3.2.4 ADMM（交替方向乘子）方法 ADMM 是一种增广拉格朗日式的变体，用来解决凸函数的优化问题。它可以应用于多种场景，包括逻辑回归、Lasso 回归、支持向量机、线性回归、递归最小二乘法、主成分分析以及循环神经网络。在回归模型场景中基于 ADMM 的分布式带逻辑回归[23]可分别应用于水平分割和垂直分割的患者数据。针对水平分割的数据，将通过最小化每个数据中心的数据 L_2 正则化对数似然函数来分别更新总体的模型参数。如果数据是垂直分割在各个数据中心，基于 ADMM 框架的 Lasso 回归将合并来自各个数据中心的中间结果，在 L_2 正则化逻辑损失函数中衍生出一个辅助变量。通过合并所有数据中心的模型中间结果和平均化辅助变量来更新双变量的方式，ADMM 框架下的分布式逻辑回归可以防止本地数据中心从更新的中间结果来推论模型的方式以达到高效保护隐私的效果。在分类模型场景中基于 ADMM 开发的分布式支持向量机分类器[18]同样可以应用于水平分类的数据中心。通过引入辅助变量，线性支持向量机分类器在共识约束下被分解为一系列的凸函数的子函数优化问题。在 ADMM 框架下全局支持向量机分类器通过每个节点交换其估计的模型参数而不是其训练参数来达到隐私保护的效果。为处理连续且异步的学习任务，线性分布式支持向量机分类器支持基于时间变化的数据线上更新，该分类器可以通过部分更新的方式来适应从训练数据集中增加或移除的部分数据样本。

3.2.5 基于多方安全计算方法 基于多方安全计算的逻辑回归模型保护的是在模型迭代训练期间传递的中间结果[11]。该方法通过在最大化似然估计时使用固定的海森矩阵来呈现一种安全的矩阵求和、求积和求逆的协议用于模型训练，是一种提供在特定情况下为防止数据和中间结果泄露的解决方案，具有可靠的准确性。但是该方法在数据计算和传输方面成本较高，不适用于有限网络带宽下的大数据的大规模计算。

3.3 非结构化电子病历数据

除结构化数据外，医学上还有很多数据是非结构化的，如医嘱病历的文本数据。非结构化电子病历是指医护人员在记录患者信息过程中未遵循统一的数据格式和记录标准，通常包含患者个人相关的细节和医护人员的非疾病相关记录。非结构化电子病历有利于医护人员对患者情况进行快速、准确的回顾和判断，但是不利于数据的传播和二次使用。具有隐私保护的联盟学习下的文本预测模型[24]通过Word2Vec（词转换成向量）对来自不同数据源的数据进行上下文标记的转换，然后匹配相对应的一系列锚点融合进入不同的向量模型。通过对比结构化和非结构化的数据来预测患者最有可能的疾病。通过联盟学习基于多中心数据建立的全局融合模型较只基于各个中心独立数据建立的模型具有更高的预测准确性。另外对于患者相似性的诊断也是一个重大难题。建立一名患者信息的哈希结构框架[25]，用哈希编码的方式表示来自不同机构的患者信息，患者之间的相似性能被该框架有效地计算出来，可以避免逆向工程中的安全攻击。此外该算法框架分别在平衡数据集和不平衡数据集中的准确性为91.56%和80.12%，兼顾隐私安全和系统效率。

3.4 基因数据

近年来随着测序技术的发展，基因数据被有效地应用于生物医学研究领域，其在医学诊断和治疗方面更为精准医疗奠定重要基础[26]。然而由于基因数据包含很多敏感信息，如身份[27]、疾病[28]、面部特征[29]信息等，所以在合理利用基因数据的同时保护数据拥有者的隐私十分必要。文献[30]基于软件加密电路设计一套多方安全计算的基因分析框架来发现疾病与基因的相关性，但由于软件加密电路的开销巨大，该方法只适用于非常小尺度数据的情况。文献[31]设计一套基于安全计算的联邦机器学习框架来帮助人们确定自己的祖先，该方法对预筛选的SNPs 数据基于EM 算法进行联合估计并最终得到每个混血个体的祖先组成。文献[32]基于加密硬件设计一套高效的跨国基因数据联合分析系统，该系统对来自3 个不同国家的695 784 条SNPs 基因数据进行分析，从而找出统计学上排名前10 位的SNPs 来自于五段基因，为川崎病的诊断和治疗提供一定依据。文献[33]设计一套安全的基因数据比对分析系统，该系统用加密硬件来托管基因数据库并提供比对分析服务，客户通过安全通道和加密硬件进行通信，得到个人基因数据和数据库基因数据的比对结果。

3.5 医学图像数据

在医学影像领域，深度学习的卷积神经网络技术可以迅速、准确地帮助医生识别图像中的相关信息，极大降低误诊率，节约时间成本，有助于医生及时并准确地做出诊断，降低患者健康风险。例如CT 可以作为有效的临床依据来帮助确诊COVID－19。然而卷积神经网络依赖于准确的网络参数，而准确的网络参数依赖于大数据的支持。传统方法是假定所有数据汇总到同一方统一进行计算。然而出于个人数据隐私保护和法律规章制度的考虑，这种集中式计算方式在实际应用中面临很大困难。但是如果仅依赖个体或少数医院的医疗影像数据计算出模型参数，可能由于样本量不够而降低准确率，一个没有充分计算的卷积网络甚至会出现大相径庭的结果，不仅不能很好地服务于广大医务人员，还会提高误诊率。联邦学习很好地解决卷积神经网络训练医学图像数据量和数据拥有者之间的隐私问题。通用的卷积神经网络训练方法，见图5。假定 W 为模型参数，N 为数据拥有者个数，K 为迭代步骤，W 上标表示不同的迭代步数，下标表示不同的数据拥有者，另外含有下标的参数表示局部模型参数，否则其为全局模型参数，其具体步骤如下：全局服务器初始化模型参数 W^0 并发送给各个数据拥有者 $\{1,\ldots,N\}$；各个局部数据拥有者在得到全局模型参数 W^k 后初始化本地模型，并基于本地数据训练出局部模型 W_i^k；所有局部数据拥有者局部模型发往全局服务器进行合成；全局服务器在得到局部参数 $\{W_0^k, W_1^k, \ldots, W_N^k\}$ 后更新参数得到全局模型参数 $\{W^{k+1}\}$，如果 W^{k+1} 达到精度要求或 $k+1$ 达到最大迭代次数 K 则停止模型训练并输出模型结果，否则继续步骤2。现有的大部分联邦深度学习算法都

是基于以上框架或者其变体。在医学图像方面也有一些应用。例如美国宾夕法尼亚大学和英特尔合作，基于以上框架提出联邦深度学习网络用于对脑部肿瘤分割的方案[34]，训练的模型是基于 U-Net[35]，为保护隐私其提出可以采用英特尔的加密芯片来辅助计算。英伟达和英国国王学院合作开发一种类似的联邦学习下网络用来进行脑部肿瘤分割[36]，为进一步保护数据提出基于查分隐私的方案，用过加入拉普拉斯噪声[37]到局部参数模型的策略来提高本地数据的保护级别，但会很大程度地降低模型准确性。

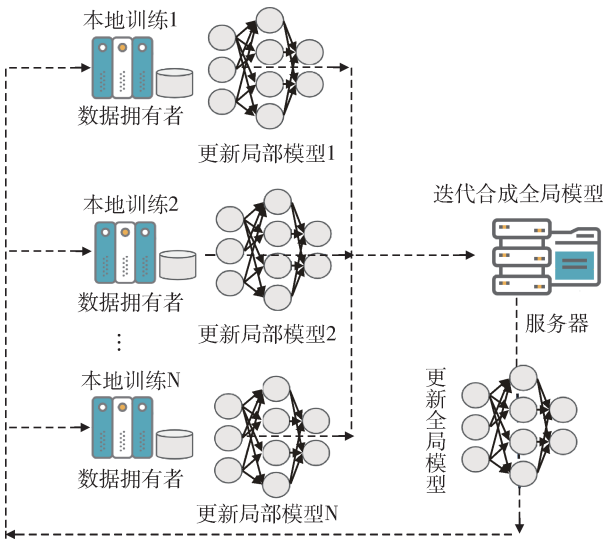


图5 通用的卷积神经网络训练方法

4 结语

本文结合当前新型冠状病毒疫情，阐述生物医疗大数据基于隐私保护的数据共享和联合分析的必要性，具体介绍联盟学习技术特点和适用性以及当前针对各种不同数据类型基于联盟学习框架的机器学习和深度学习的实际应用。面对疫情，积极迎战，充分运用现代化工具，建设合规、稳定、长期、充分利用全社会资源的有效疫情监控、跟踪和分析系统，及时把握整体情况，进行准确的预警跟踪分析，尽早恢复正常的生产生活。

（致谢：本项研究是由杭州锘崴信息科技有限公司科研基金、华西医院系统遗传研究所和公安部第三研究所信息与网络安全重点实验室基金（C19609）共同资助完成。）

参考文献

1 Huang C, Wang Y, Li X, et al. Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China [J]. Lancet, 2020, 395 (10223): 497-506.

2 Chan JF, Yuan S, Kok KH, et al. A Familial Cluster of Pneumonia Associated with the 2019 Novel Coronavirus Indicating Person-to-Person Transmission: a study of a family cluster [J]. Lancet, 2020, 395 (10223): 514-523;.

3 De Wit E, Van Doremalen N, Falzarano D, et al. SARS and MERS: recent insights into emerging coronaviruses [J]. Nat Rev Microbiol, 2016, 14 (8): 523-534.

4 Chen N, Zhou M, Dong X, et al. Epidemiological and Clinical Characteristics of 99 Cases of 2019 Novel Coronavirus Pneumonia in Wuhan, China: a descriptive study [J]. Lancet, 2020, 395 (10223): 507-513.

5 Chu CM, Cheng VCC, Hung IFN, et al. Role of Lopinavir/Ritonavir in the Treatment of SARS: initial virological and clinical findings [J]. Thorax, 2004, 59 (3): 252-256.

6 Holshue ML, DeBolt C, Lindquist S, et al. First Case of 2019 Novel Coronavirus in the United States [EB/OL]. [2020-01-31]. https://www.nejm.org/doi/full/10.1056/NEJMoa2001191.

7 Morse JS, Lalonde T, Xu S, et al. Learning from the Past: possible urgent prevention and treatment options for severe acute respiratory infections caused by 2019-nCoV [EB/OL]. [2020-01-27]. https://chemrxiv.org/articles/Learning_from_the_Past_Possible_Urgent_Prevention_and_Treatment_Options_for_Severe_Acute_Respiratory_Infections_Caused_by_2019-nCoV/11728983.

8 Zumla A, Hui DS, Azhar EI, et al. Reducing Mortality from 2019-nCoV: host-directed therapies should be an option [J]. Lancet, 2020, 395 (10224): e35-e36.

9 Heymann DL. Data Sharing and Outbreaks: best practice exemplified [J]. Lancet, 2020, 395 (19223): 469-470.

10 Wang C, Horby PW, Hayden FG, et al. A Novel Coronavirus Outbreak of Global Health Concern [J]. Lancet, 2020, 395 (19223): 470-473.

11 Shi H, Jiang C, Dai W, et al. Secure Multi-pArty Computation Grid LOgistic REgression (SMAC-GLORE) [J]. BMC Med Inform Decis Mak, 2016, 16 (Suppl 3): 89.

12 Jiang W, Li P, Wang S, et al. WebGLORE: a web service for Grid LOgistic REgression [J]. Bioinformatics, 2013, 29 (24): 3238 - 3240.

13 Wang S, Jiang X, Wu Y, et al. EXpectation Propagation LOgistic REgRession (EXPLORER): distributed privacy - preserving online model learning [J]. J Biomed Inform, 2013, 46 (3): 480 - 496.

14 Lu C - L, Wang S, Ji Z, et al. WebDISCO: a web service for distributed cox model learning without patient - level data sharing [J]. J Am Med Inform Assoc, 2015, 22 (6): 1212 - 1219.

15 Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression [M]. USA: Wiley Series in Probability and Statistics, 2013: 127.

16 Mateos G, Bazerque JA, Giannakis GB. Distributed Sparse Linear Regression [J]. IEEE Transactions on Signal Processing, 2010, 58 (10): 5262 - 5276.

17 Schizas ID, Aduroja A. A Distributed Framework for Dimensionality Reduction and Denoising [J]. IEEE Transactions on Signal Processing, 2015, 63 (23): 6379 - 6394.

18 Forero PA, Giannakis GB. Consensus - based Distributed Support Vector Machines [J]. Journal of Machine Learning Research, 2010 (11): 1663 - 1707.

19 Ohno - Machado L, Agha Z, Bell DS, et al. pSCANNER: patient - centered scalable national network for effectiveness research [J]. J Am Med Inform Assoc, 2014, 21 (4): 621 - 626.

20 Wu Y, Jiang X, Kim J, et al. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data [J]. J Am Med Inform Assoc, 2012 (5): 758 - 764.

21 Wu Y, Jiang X, Wang S, et al. Grid Multi - category Response Logistic Models [J]. BMC Med Inform Decis Mak, 2015, 15 (1): 1 - 10.

22 Li Y, Jiang X, Wang S, et al. VERTIcal Grid lOgistic Regression (VERTIGO)[J]. J Am Med Inform Assoc, 2016, 23 (3): 570 - 579.

23 Boyd S, Parikh N, Chu E, et al. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers [J]. Foundations and Trends in Machine Learning, 2011, 3 (1): 1 - 122.

24 Huang Y, Lee J, Wang S, et al. Privacy - preserving Predictive Modeling: harmonization of contextual embeddings from different sources [J]. JMIR Med Inform, 2018, 6 (2): e33.

25 Lee J, Sun J, Wang F, et al. Privacy - preserving Patient Similarity Learning in a Federated Environment [J]. JMIR Medical Informatics, 2018, 6 (2): e20.

26 Collins FS, Varmus H. A New Initiative on Precision Medicine [J]. N Engl J Med, 2015, 372 (9): 793 - 795.

27 Gymrek M, McGuire AL, Golan D, et al. Identifying Personal Genomes by Surname Inference [J]. Science, 2013, 339 (6117): 321 - 324.

28 McGuire AL, Fisher R, Cusenza P, et al. Confidentiality, Privacy, and Security of Genetic and Genomic Test Information in Electronic Health Records: points to consider [J]. Genet Med, 2008, 10 (7): 495 - 499.

29 Claes P, Liberton DK, Daniels K, et al. Modeling 3D Facial Shape from DNA [J]. PLoS Genet, 2014, 10 (3): e1004224.

30 Jagadeesh KA, Wu DJ, Birgmeier JA, et al. Deriving Genomic Diagnoses without Revealing Patient Genomes [J]. Science, 2017, 357 (6352): 692 - 695.

31 Chen F, Dow M, Ding S, et al. PREMIX: privacy - preserving estiMation of individual admiXture [J]. AMIA Annu Symp Proc, 2016 (2016): 1747 - 1755.

32 Chen F, Wang S, Jiang X, et al. PRINCESS: privacy - protecting rare disease international network collaboration via encryption through software guard extensions [J]. Bioinformatics, 2017, 33 (6): 871.

33 Chen F, Wang C, Dai W, et al. PRESAGE: privacy - preserving gEnetic testing via software guard extension [J]. BMC Med Genomics, 2017, 10 (2): 48.

34 Sheller MJ, Anthony Reina G, Edwards B, et al. Multi - institutional Deep Learning Modeling without Sharing Patient Data: a feasibility study on brain tumor segmentation [M]. Switzerland: Springer, 2019: 92 - 104.

35 Ronneberger O, Fischer P, Brox T. U - Net: convolutional networks for biomedical image segmentation [M]. Switzerland: Springer, 2015: 234 - 241.

36 Li W, Milletarì F, Xu D, et al. Privacy - preserving Federated Brain Tumour Segmentation [M]. Switzerland: Springer, 2019: 133 - 141.

37 Dwork C. Differential Privacy in Bugliesi M, Preneel B, Sassone V, Wegener I (eds) Automata, Languages and Programming [M]. Berlin: Springer, 2006: 1 - 12.