

基于联邦学习的政务大数据平台应用研究

吴坚平¹, 陈超超¹, 金加和^{2,3}, 吴春明¹

1. 浙江大学计算机科学与技术学院, 浙江 杭州 310027;
2. 浙江省大数据发展中心, 浙江 杭州 310007;
3. 浙江省数据开放融合关键技术研究重点实验室, 浙江 杭州 310007

摘要

当前数字政府建设已进入深水区, 政务大数据平台作为数据底座支撑各类政务信息化应用, 其隐私数据的安全性和合规性一直被业界广泛关注。联邦学习是一类解决数据孤岛的重要方法, 基于联邦学习的政务一体化大数据平台应用具有较高的研究价值。首先, 介绍政务大数据平台及联邦学习应用现状; 然后, 分析政务大数据平台面临的隐私数据的采集、分类分级、共享三大管理挑战; 接着, 阐述基于联邦学习的推荐算法和隐私集合求交技术的解决方法; 最后, 对政务大数据平台隐私数据的未来应用进行了总结和展望。

关键词

政务大数据; 联邦学习; 推荐算法; 隐私集合求交

中图分类号: TP181

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2024032

Research on the application of government big data platform based on federated learning

WU Jianping¹, CHEN Chaochao¹, JIN Jiahe^{2,3}, WU Chunming¹

1. College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
2. Big Data Development Center of Zhejiang Province, Hangzhou 310007, China
3. Key Laboratory of Key Technologies for Open Data Fusion in Zhejiang Province, Hangzhou 310007, China

Abstract

At present, the construction of digital government has entered a deepwater area. The government big data platform, as a data base, supports various government information applications. The security and compliance of its private data has been widely concerned by the industry. Federated learning is an important method to effectively solve data silos, and the application of government big data platforms based on federated learning has high research value. Firstly, the current status of government big data platforms and its federated learning application were introduced. Then this paper analyzed three major management challenges involved in the collection, classification and grading and sharing of privacy data

on government big data platforms. Further, the problem-solving methods of federated learning based recommendation algorithms and privacy intersection techniques were explored. Finally, summaries and prospects were made for the future application of privacy data on government big data platforms.

Key words

government big data, federated learning, recommendation algorithm, private set intersection

0 引言

随着人工智能、隐私计算、大数据技术的融合发展,全国各地掀起了数字政府建设的高潮,政务大数据平台应运而生。政务大数据平台是一个集数据汇聚、交换、治理、分析、应用、安全为一体的统一政务数据共享开放平台。《全国一体化政务大数据体系建设指南》显示,截至2022年9月,我国已有26个省级政务数据平台、257个市级政务数据平台、355个县级政务数据平台^[1],见表1。多地已初步建成省市县一体化的政务大数据平台,这些平台在推进数字政府建设中发挥着重要作用。

随着政务大数据平台的不断落地,大量的数据汇聚到政务大数据平台。平衡数据共享与隐私保护是学术界普遍关注的课题。在技术层面,如果没有强有力的安全管理措施,这些数据一旦泄露,将会是一个巨大灾难。在法律层面,我国越来越重视数据使用的合规性,《中华人民共和国数据安全法》(以下简称《数据安全法》)和《中华人民共和国个人信息保护法》相继出台,进一步增加了政府机构之间、政企之间采集和共享隐私数据的难度。由上可知,推进一体化的政务大数据平台依然面临“数据孤岛”问题。

本文从隐私保护技术的角度对当前政务大数据平台中隐私数据的数据采集、分

表1 我国政务大数据平台建设情况

序号	类型	平台上线量	平台上线率
1	省级政务大数据平台	26	76.47%
2	市级政务大数据平台	257	77.18%
3	县级政务大数据平台	355	12.49%

类分级、数据共享等进行了分析,提出以下研究目标。

- 最小必要。解决政府部门之间数据共享过程中出现的过度归集数据的问题,防止出现超范围归集隐私数据、超权限使用隐私数据、违法违规交易隐私数据等行为。
- 科学分类。基于政务大数据平台数据量大、隐私性强、结构复杂的特点,解决难以对海量多源异构的原始数据进行分类分级管理的问题,保持数据分类分级的一致性以及防护策略的有效性。
- 高质共享。两个垂直机构或水平机构对双方的隐私数据进行融合分析,从而得出决策数据,这种情况较为常见。在保证隐私数据安全的前提下,实现隐私集合求交,获得高质量交集数据。此外,在“数据孤岛”仍然存在的背景下,实现数源部门数据快速有效的供需匹配。

1 研究现状

1.1 政务大数据相关研究

近年来,国家高度重视一体化政务大

数据体系建设,先后出台了《促进大数据发展行动纲要》《关于推进公共信息资源开放的若干意见》等系列文件,明确提出数据是国家基础性战略资源,要求推动建设全国一体化的国家大数据中心,加快新型基础设施建设布局^[2-3],加速推进政务数据开放和应用。从2018年起,全国各地相继成立大数据管理机构,建立政务大数据共享平台,并制定政务数据归集规范,在一定程度上实现了政务数据统筹管理目标^[4],为政务数据共享提供了统一平台,如图1所示。然而,随着我国政府数据的不断开放,数据开放与隐私保护之间的矛盾日益凸显,我国政府数据开放平台在隐私保护方面的整体表现不佳^[5]。由政府数据开放引发的隐私泄露事件不仅侵犯了公民的权益,给政府部门造成经济与声誉上的损失,还严重制约了政府数据开放的发展^[6]。

1.2 联邦学习相关研究

联邦学习是隐私计算的一种主流技术。联邦学习^[7]的概念最早是在2016年由谷歌提出的,它是一种满足隐私保护和数据安全的分布式机器学习框架。联邦学习仅需在数据分散存储的节点上训练模型,由于服务器无法获取原始数据,个人数据隐私得到了有效的保护,有效解决了“数据孤岛”和数据隐私保护的问题^[8-9]。另外,传统的机器学习模型不能直接处理异构数据,利用联邦学习技术,无须处理异构数据即可建立全局数据上的机器学习模型,既保护了数据隐私,又解决了数据异构问题^[10]。在数据隐私问题备受关注的今天,联邦学习在避免数据泄露、中心点数据受到攻击等方面具备显著优势。在实际应用中,按照“孤岛数据”的

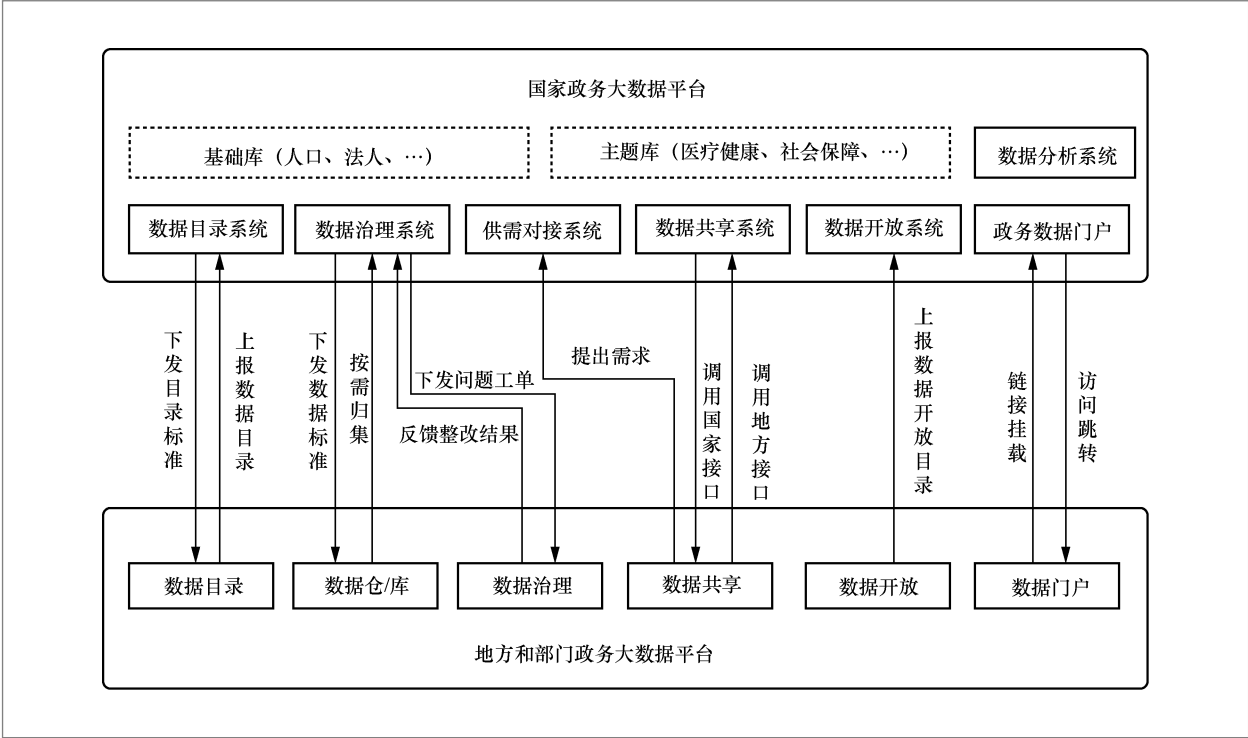


图1 国家平台与地方/部门平台关系^[1]

分布特点，联邦学习可分为横向联邦学习、纵向联邦学习、联邦迁移学习3类。此外，联邦学习架构也可分为需要协调方的客户-服务器架构和不需要协调方的对等网络架构，其中对等网络架构可以进一步确保数据安全性，因为各方无须借助第三方便可通信。

联邦学习在处理数据共享开放和数据隐私保护方面表现突出。到目前为止，联邦学习已在金融风控、广告营销^[9]等方面落地。例如，淘宝、抖音、美团等电商平台推送不同的产品宣传信息，其背后的推荐模型需要大量的用户特征数据^[11-12]。每一家电商企业只拥有用户的一部分特征数据，如果可以接入其他与用户有关联性的数据，补充特征数据，那么广告推荐的效果将大幅提升。然而，这些方法仅限于在线购物、视频浏览等数据维度小的场景，无法处理城市交通综合调度、教育医疗资源需求分配等基于政务大数据平台的大型群体智能应用。

1.3 联邦学习在政务大数据领域相关研究

联邦学习技术快速发展，正在助力政府打破各领域的数据壁垒，已在多个真实场景中发挥了重要作用。在智慧警务领域，文献[13]提出了通过联邦学习技术实现不同地区相同警种间的数据协作，如精准步态识别、有遮挡的人物识别、行人重识别等任务。在智慧医疗领域，联邦学习技术解决了电子健康记录管理、医疗影像分析等隐私保护问题。文献[14]介绍了基于联邦学习的全国首个跨多个省级区域的数字化联防联控平台，解决了医疗大数据在多个地区之间的共享和隐私保护问题。在智慧交通领域，文献[15]提出了一种基于联邦学习的交通资源调度方法，该方法显著

提升了群体的出行效率。在智慧电力领域，文献[16]提出了利用联邦学习预测充电站的能源需求，根据预测结果决策是否在充电站预先存储能源以及存储多少能源。以上4个文献的研究对象均为某一领域的隐私数据分析，但跨领域的多维度数据应用并未取得较大突破，无法利用本领域的数据分析为其他领域的数据预测提供帮助，如智慧医疗领域的联邦学习方法无法对市民的购物等生活数据进行关联分析。然而，国家一体化政务大数据中包含多种跨领域数据，能够挖掘更多的数据价值。本文的研究重点是在国家一体化政务大数据平台的基础上，实现基于联邦学习技术的跨层级、跨领域、跨地区的海量多维隐私数据的融合应用。

2 政务大数据平台隐私数据管理挑战

根据《全国一体化政务大数据体系建设指南》，到2023年年底，全国一体化政务大数据体系初步形成，全国一体化指的是实现国家、省、市、县四级政务大数据平台上下贯通、协同作战。在这个框架体系下，政务大数据平台具有数据链条长、数据维度多、数据价值高的特点，这增加了数据开放共享的安全风险。同时，有些部门以数据安全要求高、仅供特定部门使用为由，在数据供需双方自建共享渠道。针对数据安全的不同生命周期，文献[17]提出了数据安全能力成熟度模型，根据数据的生命周期，数据安全可分为采集安全、传输安全、存储安全、处理安全、交换安全和销毁安全。在基于一体化政务大数据平台的数据全生命周期安全管理中，隐私数据的数据采集、分类分级、数据共享是数据安全领域的研究重点。

2.1 隐私数据安全采集难

数据采集包括外部数据和内部数据的采集，外部数据是指除了组织内部之外的所有数据提供方，采集过程需要保证数据的传输和存储安全，并提供全过程审计的能力。根据采集的数据类型和数据源不同，数据采集可以选择不同的技术工具。目前政务数据主要有数据库数据和系统日志数据两种来源。基于数据库采集源数据^[17]，主要采用直接数据源同步、生成数据文件同步、数据库日志同步3种方式。其中，直接数据源同步是指直接连接业务数据库；生成数据文件同步是指从数据源系统现场生成数据文件，再通过文件系统同步到目标数据库中；数据库日志同步是指基于源数据库的日志文件同步数据。以上技术是非涉密涉敏数据采集的常用手段，在公检法、交通等领域不可行，急需从联邦学习等隐私计算相关技术方法加以突破。

- 数据适用性差。目前，部分地区政务大数据平台主要涉及垂直业务部门和本地同级相关部门的数据采集工作，一般采用必要性和最小化原则进行采集，防止出现超范围采集用户数据、超权限使用用户数据、违法违规交易用户数据及差异化用户决策等违法行为。
- 数据一致性差。根据国家一体化政务大数据建设要求，各地区依托政务数据平台统筹推进本地政务数据的归集工作，实现省市县三级数据汇聚整合^[1]，行业主管部门做好本行业政务数据的归集工作。因数据维度、数据质量等差距，多数据源回归过程普遍存在数据不一致的问题。
- 数据鲜活性低。在国家一体化政务大数据平台建设的背景下，如何通过数据采集来持续提升数据鲜活性，是一项有较

大挑战性的工作。例如，由于人口库数据存在流动人口动态变化及多源管理的问题，维护人口库数据需要从横向和纵向全维度实现反向数据反馈，实现数据“反哺”，以此实现平台数据鲜活性。

2.2 隐私数据分类分级难

为了平衡数据使用的便捷性和数据保护的安全性，促进数据安全能力建设降本增效，政务大数据平台通常会对政务数据资产进行分类分级，以便对数据采用精细化的安全管控手段，避免“一刀切”现象。同时，数据分类分级有助于发现冗余、无关、过时和被遗忘的数据，以便将其从系统中删除。

- 安全合规难。《数据安全法》第二十一条明确规定，国家建立数据分类分级保护制度，对数据实行分类分级保护。这意味着在法律层面，数据分类分级工作已成为政府、企事业单位等开展数据处理活动必须遵守的规定^[18]。随着《数据安全法》的落地，数据分类分级也将成为数据安全合法合规工作的首要任务。然而，目前相当一部分政务大数据平台因起步建设较早、安全工作重视度不够等原因，并没有严格落实数据分类分级，隐私数据安全形势仍然严峻。
- 规则趋同难。国家统筹建国家、省、市、县四级一体化政务大数据平台数据量大、结构复杂，难以对海量多源异构的原始数据进行分类分级管理，如何保持数据分类分级的一致性，是一个技术难题。
- 隐私保护难。目前数据分类分级工作普遍采用机器学习建模进行初筛，再由人工审核的方式。针对隐私数据的分类分级，人工参与程度较深，易造成敏感信息泄露、人力成本高、主观因素大等问题。

2.3 隐私数据充分共享难

为了充分挖掘数据价值，各级政务大数据管理机构会跨系统、跨区域、跨层级进行数据共享，然而数据在共享的过程中可能存在隐私泄露的风险。

- 精准度量难。政务公共数据涉及社会经济发展中各领域的关键数据，在数据开放共享中普遍存在度量难题，一是数源单位没有采用有效的数据分类分级管理办法，存在过度共享导致的隐私数据泄露的问题；二是数源单位开放的数据存在低质或无效问题，难以量化数据合格率。
- 安全管控难。政务公共数据在开放过程中普遍存在涉及链条长、机构多等问题，一是在共享取数过程中，数据存在被篡改或伪造的风险；二是开发过程存在代码不规范等问题，容易造成数据泄露；三是数据二次或多次共享导致数据溯源难。

3 基于联邦学习的政务大数据应用分析

基于国家一体化的政务大数据具备大规模特性，在进行多领域数据关联分析时，常采用隐私集合求交技术实现隐私数据保护。同时，政务大数据应用场景具备数据量大、用户多等特性，缺少个性化信息过滤服务。针对以上问题，本文提出了一种利用联邦推荐系统实时精准定位所需数据的方法。

3.1 隐私集合求交技术及应用

隐私集合求交 (private set intersection, PSI) 也被称为安全求交^[19-20]，它是指多个数源单位分别拥有一个集合，它们

想求多个集合的交集，同时保证最终只会泄露交集的结果，而不会泄露各方非交集的集合元素。该技术被认为是跨机构数据合作的前置步骤，实现了跨源数据间的安全融合。

隐私集合求交技术应用发展迅速，可以分为基于密码学和基于硬件两类^[21]，二者各有优势。基于密码学的技术公开透明，各个参与方是对等的，安全性也有保障，但是在参考方很多的时候，需要两两交互，通信量大。而基于硬件的方法在参与方较多时，所需的通信轮数少。本文只讨论基于密码学的方法，这类方法包括基于公钥体系（如同态加密）的方法、基于电路的方法、基于不经意传输的方法。

在保护数据隐私的前提下解决行业中的数据孤岛问题，是联邦学习价值的核心。目前联邦学习在各种隐私集合求交场景应用广泛，如金融、医疗、交通、旅游等。

以一个区县级旅游目的地的旅游数据共享应用为例，某地旅游主管部门拟开展下一年高铁游精准营销计划，需要建立一个基于游客的高铁旅客数据、景区入园数据、公安住宿数据的游客画像。经调研，当地只掌握游客景区入园数据，而游客在酒店的住宿数据由市级公安部门住宿登记平台管理，通过高铁出行的游客数据由国家铁路部门管理，并且这些数据分别由区县级、市级、国家部级3个不同级别的政务大数据平台管理。在国家一体化政务大数据平台框架下采用联邦学习方案，各数据持有方不需要提交数据，通过隐私集合求交的方式，保证了本地数据的安全，降低了隐私泄露的风险，实现“数据不出域，模型出域”的目标，如图2所示。

3.2 推荐系统技术及应用

随着互联网、大数据技术的迅猛发

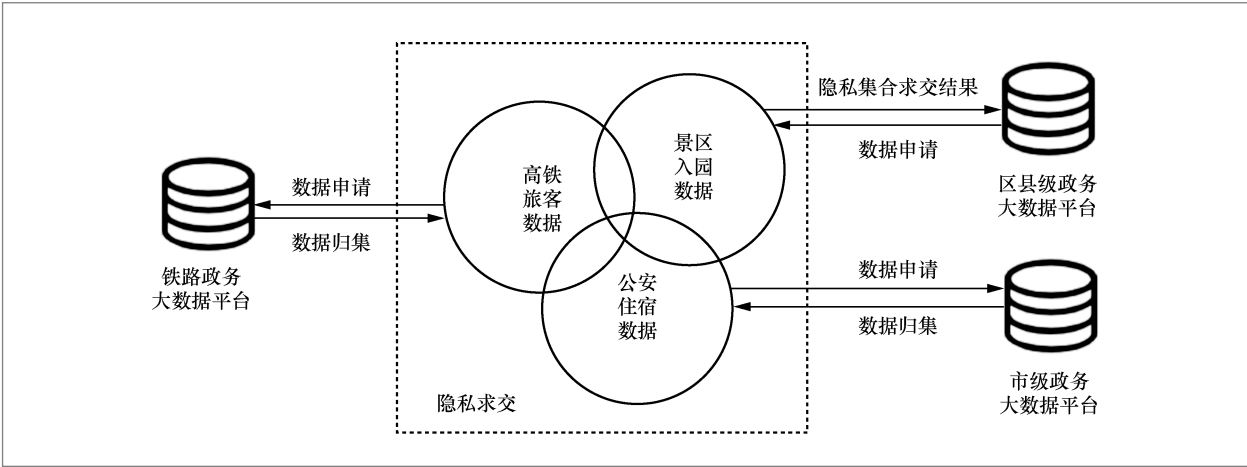


图 2 “联邦学习 + 旅游” 场景中的隐私集合求交技术应用

展，海量信息常常会让用户迷失，无法找到想要的内容。为了解决这类信息过载问题，推荐系统 (recommendation system, RS) 应运而生。推荐系统^[22]利用用户的各种知识和数据生成个性化推荐结果，其构建依赖于大量的用户信息、物品信息以及二者之间的交互信息。目前，推荐模型主要分为协同过滤、基于内容、基于模型、混合推荐这4类^[23-24]。传统的推荐系统以集中式训练为主，即推荐系统的构建者先收集推荐系统所需要的数据（如评分、评论和点击等）至一个集中区域，进而在该区集中式训练推荐模型，达不到保护用户数据隐私的目的。

基于联邦学习的推荐系统是为保护用户隐私而采用去中心化的推荐系统训练方法^[22]，在该方法中，隐私数据依旧由用户保存在各自的终端，不需要上传至集中式的服务器，同时，模型的训练过程是去中心化的，避免了隐私信息的泄露风险。

联邦推荐系统常用的算法有隐私保护矩阵分解^[25]和隐私保护因子分解机^[26]两种。根据所面对的场景，推荐系统可分为跨设备和跨机构两类隐私保护推荐系统。跨设备的隐私保护推荐系统是指在不收

集用户原始隐私信息的前提下构建推荐系统；跨机构的隐私保护推荐系统是指多个机构已经存储了用户的隐私信息，在保护各自数据隐私的前提下，协同构建推荐系统的方法。由于政府的相关职能部门有必要存储用户的一些隐私信息，并且必须保护该信息的安全性，跨机构的隐私保护推荐系统更加符合政务平台的需求。因此，本文以跨机构的隐私保护推荐系统为基础，进行本文方法的探讨。

以推荐系统在政务服务的相关事项应用^[27]为例，为了提高用户办事效率与政府服务水平，通过联邦学习获取用户在社会保障主题库、医疗健康主题库、人口基础库等的多维用户标签信息，构建政务服务用户画像，在推荐计算中结合用户相关属性进行精准推荐，如图3所示。

政务服务联邦推荐过程大致包括以下步骤。

阶段一：通过联邦学习生成用户画像标签信息。

第一步：初始化政务大数据平台（作为聚合服务器）的模型参数，并将模型训练任务发布到社会保障、医疗健康等涉及政务服务的所有专题数据仓和基础库相关的

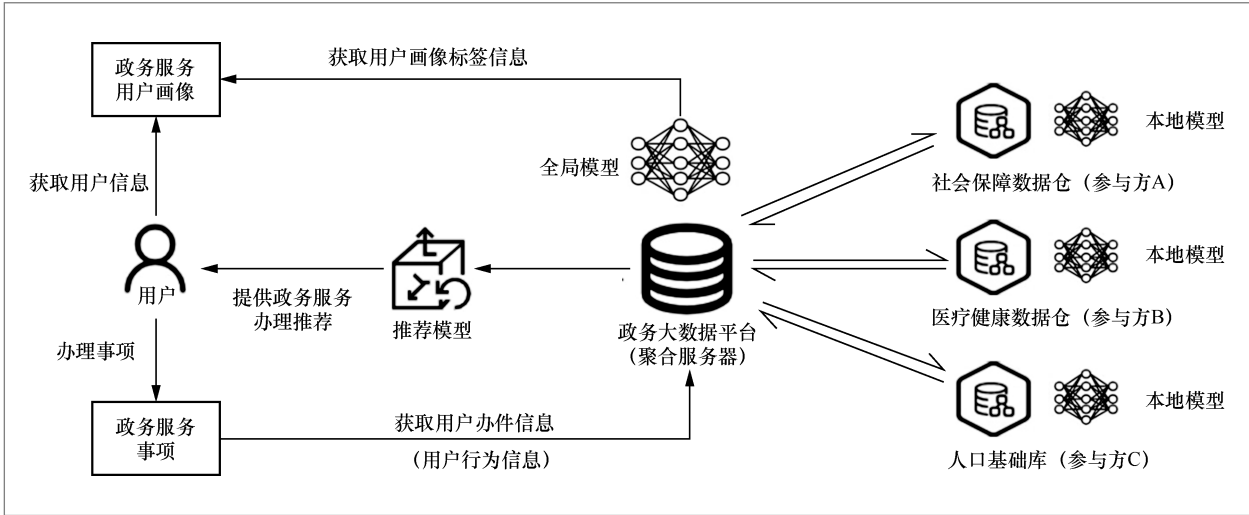


图3 政务服务推荐模型算法流程

隐私计算节点（也称参与方）等。

第二步：各隐私节点通过本地数据仓或基础库数据进行本地模型训练。

第三步：本地模型在训练后，通过安全协议将参数更新传输至聚合服务器，由聚合服务器执行模型聚合计算，迭代生成新的全局模型参数。

第四步：新的全局模型下发至各隐私计算节点。

持续进行上述步骤直至全局模型收敛，最终的训练结果就是全维度用户画像标签信息，为下一阶段在推荐计算中结合用户多维度属性进行精准推荐奠定基础。

阶段二：通过本地模型进行推理预测。

第一步：根据用户提交的登记身份信息和政务服务事项，分别获取用户画像标签信息和用户办件信息。办件信息决定相关的业务数据仓或基础库。

第二步：从用户相似度计算和政务服务事项相似度计算两方面着手，选择一种推荐模型，构建政务服务办理推荐特征矩阵，计算特征值。

第三步：根据政务服务办理特征值实现精准推荐。

3.3 隐私集合求交及联邦推荐系统融合应用探索

本文以联邦推荐系统为技术支撑，对政务大数据平台中的数据采集、分类分级、数据共享3项关键工作进行下一步探索。

- 数据采集阶段。在一体化政务大数据平台环境下，数据需求单位（如市人社局）向数源单位（如省人社厅）申请人口库数据回流，双方通过联邦学习隐私集合求交技术获取相同的数据，为避免数据冗余，该部分数据屏蔽，只取带有该市标签的非已存数据回流。

- 分类分级阶段。通过联邦学习动态建立国家、省、市、区（县）四级统一规范的分分类分级标准和模型，具体做法是服务器节点设置在国家级，省、市、区（县）作为参与方，维护一体化分类分级指标。某参与方在确定需回流数据后，对回流数据打上数源单位分类分级标签。数据申请方收到带标签的数据，及时、高效、准确地进行数据分类分级和归集数据，如图4所示。

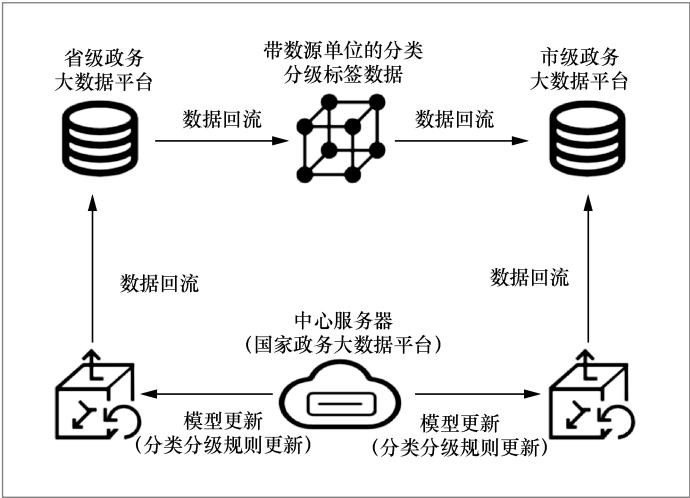


图 4 基于联邦学习的智能化分类分级处理流程

● 数据共享阶段。在国家一体化政务大数据平台的基础上，利用联邦学习融合跨领域信息，构建一体化联邦推荐系统，有效缓解数据稀疏、数据泄露等问题。该系统通过联邦推荐系统的隐私保护因子分解机算法，在确保数据隐私和模型隐私安全的前提下，解决不同政务部门（数源单位）之间的数据供需匹配问题，即对数据拥有单位及数据需求单位进行联邦推荐建模，预测二者的供需匹配度，帮助数据需求单位精准、快速、安全地找到最佳数源单位，然后依托一体化政务大数据平台进行数据高质量共享，如图5所示。

一体化政务大数据平台跨级联邦推荐包括以下步骤。

第一步：区县级数据需求单位向本级政务大数据平台提出数据共享申请。

第二步：区县级政务大数据平台通过对本级社会保障、医疗保障等数据仓和基础库进行联邦推荐，如果推荐数据达到预设可信阈值，直接反馈数据；如果未达到预设可信阈值，向市级政务大数据平台提出数据共享申请。

第三步：市级政务大数据平台通过对本级社会保障、医疗保障数据仓等进行联

邦推荐，如果推荐数据达到预设值，直接反馈数据给区县级政务大数据平台，如果未达到预设值，向省级政务大数据平台提出数据共享申请。

第四步：以此类推，直到推荐模型达到收敛。

联邦聚合服务器将最终的推荐模型进行解密，并将其发送给所有数据拥有者进行模型共享。跨级联邦推荐可以在保护数据隐私和安全的同时，提高推荐效果和准确性，适用于需要在不同级别的数据拥有者之间进行推荐模型训练和共享的场景。下一步深度结合人工智能技术，政务大数据平台将从当前的分析式共享数据转变为生成式共享数据，实现从数据导向向业务导向的变革。

3.4 应用问题分析

目前，联邦学习正处于快速发展的阶段，有大量关于联邦学习的应用场景的研究与讨论。相关研究表明^[28]，基于政务大数据平台的联邦学习还有很多问题亟待解决。

● 隐私保护问题。保证数据的隐私性是联邦学习能快速落地的基石，同时也是联邦学习的核心挑战，政务大数据具备高可信和高价值特性，被网络安全攻击者普遍关注，因此，在政务大数据应用中隐私安全尤为重要。但研究发现，联邦学习架构的隐私保护不足以完全防御外部隐私攻击^[29]。具体来说，在模型训练和模型预测阶段都可能泄露数据隐私，如在横向联邦学习中，梯度信息在未加保护的情况下会泄露隐私训练数据的信息^[30]。目前，联邦学习的隐私保护主要采用差分隐私、同态加密、安全多方计算等隐私保护技术^[31]，并且在一些场景应用中已取得了较好效果。联邦学习系统需要持续增强对各类不良攻击的

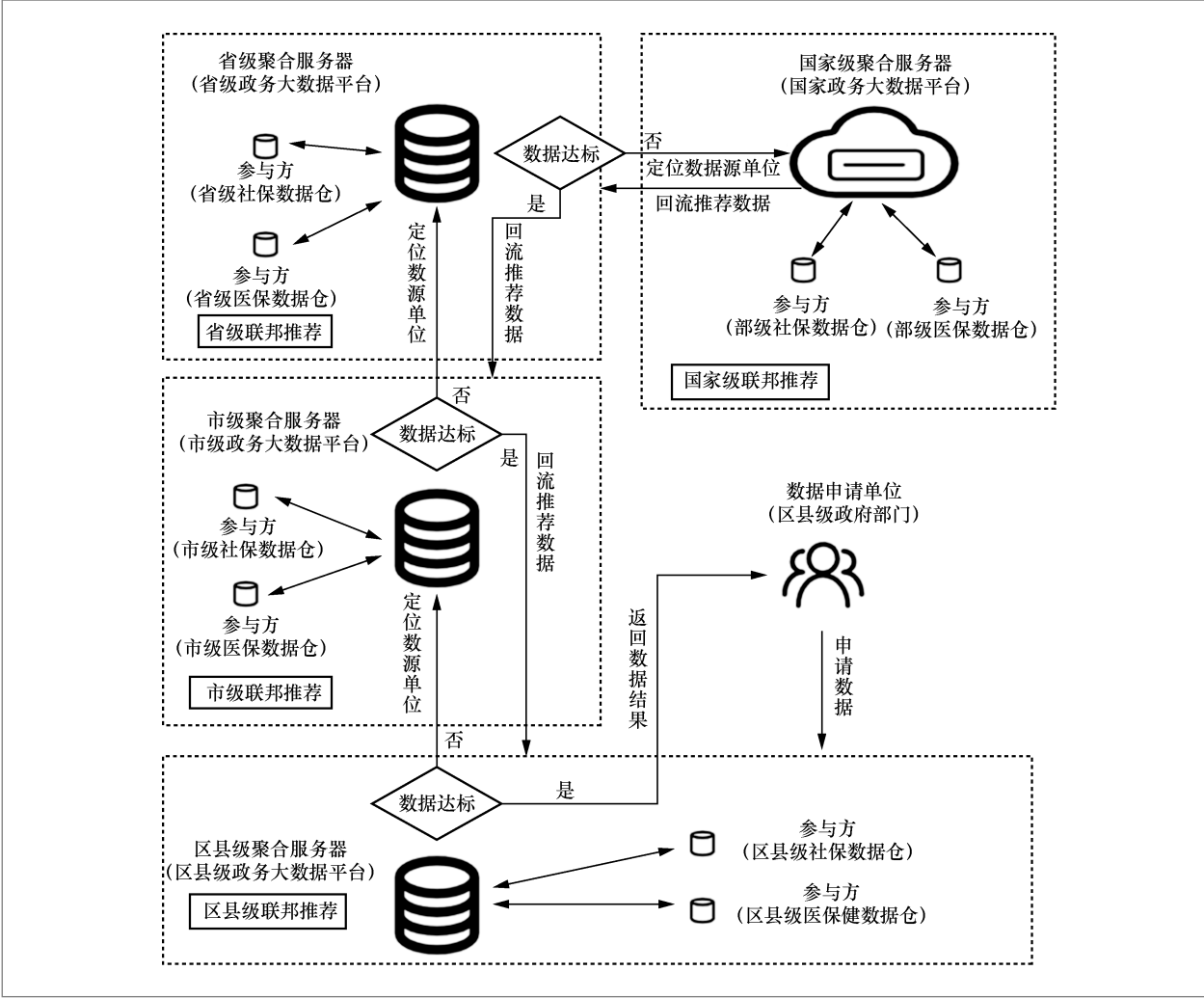


图 5 基于联邦推荐系统的政务公共数据共享流程

防御能力，保障用户数据的隐私。

- **通信效率问题。** 联邦学习易受到参与方设备异构、网络带宽等影响，导致其计算与通信效率成为阻碍相关应用落地的最大挑战^[32]。影响联邦学习算法效率的主要因素是参与方与中央服务器之间需要交换大量模型和参数，会造成较大的通信开销，因此网络中的通信效率会对整体速度有较大的影响。比如不稳定的网络情况、参数上传和下载的过程中速度不一致都会导致整个算法的模型训练成本过大。然而，政务大数据具有多样性，包括结构化

数据（如数据库、表格等）和非结构化数据（如文档、图片、视频等），这些数据来自不同的政府部门，具有不同的格式和类型，对通信效率提出了更高要求。因此，开发通信效率高的方法就显得尤为重要，通常可以从降低传输频率和减少每轮传输的信息量^[33]等方面着手。部分优化以增加参与方的本地计算为代价，部分优化以降低整个模型的准确性为代价。在实际优化的过程中，可以根据实际情况和需求决定采用何种方式降低通信成本。

- **数据异构难题。** 在联邦学习中，训练

数据可以分布在大量的客户端上^[34],比如手机、网络传感器或者各类边缘设备^[35],然后通过服务器聚合成完整的全局模型,而不用传输用户数据,从而保护了用户的隐私信息。然而,在政务大数据平台联邦学习的环境中,各个客户端数据样本量和样本类别都处于变化之中,即数据分布的异构性^[36]。由于参与联邦学习的各个边缘设备所处的实际物理环境不同,这些边缘设备能为联邦学习框架提供的数据各不相同,这种不同的数据分布会导致边缘设备在本地训练过程中产生模型偏移^[37],从而影响联邦学习在政务大数据中的应用成效,特别是使鲁棒性较差的政务应用难以落地。目前,大规模、分布式、多模态多源异构数据采集应用是联邦学习的一项重点研究领域。

4 总结与展望

联邦学习作为破解“数据孤岛”问题实现隐私数据开放共享的新方向,近年来受到学术界和政界的广泛关注。本文从政务大数据平台建设的角度出发,概述了政务领域隐私数据采集、分类分级和共享面临的挑战,进而通过联邦学习隐私集合求交技术和推荐系统进行相关应用探索,并对相关问题进行了总结分析。

国家数据局的成立,将加速跨层级、跨地域、跨部门的国家一体化政务大数据平台的构建,统筹推进数据资源整合共享和开发利用,以尽可能低的成本交换和共享数据。通过强化数据治理,进一步提升数据鲜活度、准确性、实用价值,真正实现畅通的数据资源大循环,促进经济社会高质量发展,是一个长期而具有挑战性的课题。

政务大数据的安全性需持续增强。数据安全性和合规性是制约政务大数据平台

一体化、智能化的关键因素。如何在国家一体化政务大数据平台的基础上,健全国家、省、市、县各级数据统筹管理机构,推动公共数据汇聚整合和开发利用,是一个重要研究方向。其中,在安全合规的前提下,加快推进联邦学习等隐私计算技术的场景应用,实现四级联动,其背后的数据安全工作值得深入研究。

政务大数据的跨域性需持续释放。政务大数据涉及多个领域和部门,蕴含着丰富的信息和价值,可以帮助政府进行政策制定、资源配置、社会管理等方面的决策和优化。通过联邦学习赋能政务大数据平台建设,可有效解决隐私数据跨地域、跨部门、跨层级、跨系统的整合和共享,为政府发现潜在的问题、趋势和机会提供全面、精准的数据支持。

政务大数据的大规模性有助于大模型应用研究。政务大数据通常拥有庞大的数据量,这些数据涵盖了政府的各个方面,如人口数据、交通数据、医疗数据、环境数据等,联邦学习等隐私计算技术打破了领域之间数据、算力、算法共享的壁垒。同时,在ChatGPT引发AI技术新一轮热潮的背景下,数字政府的未来创新也与大模型分不开。下一步围绕政务大数据的大模型应用,如基于算力资源、算法模型、数据处理的超级任务场景等,都将是看得见的机遇。

参考文献:

[1] 国务院办公厅. 关于印发全国一体化政务大数据体系建设指南的通知[Z]. 2022.
General Office of the State Council. Notice on printing and issuing the guidelines for the construction of the national integrated government big data system[Z]. 2022.

[2] 易成岐, 窦悦, 陈东, 等. 全国一体化大数据

- 中心协同创新体系: 总体框架与战略价值[J]. 电子政务, 2021(6): 2-10.
- YI C Q, DOU Y, CHEN D, et al. Collaborative innovation system of national integrated big data center: overall framework and strategic value[J]. E-Government, 2021(6): 2-10.
- [3] 国务院. 关于印发促进大数据发展行动纲要的通知[Z]. 2015.
- The State Council. Notice on issuing the platform for action to promote big data development[Z]. 2015.
- [4] 郑谐维. 隐私计算在政务数据共享中的应用[J]. 上海信息化, 2022(4): 17-21.
- ZHENG X W. Application of privacy computing in government data sharing[J]. Shanghai Informatization, 2022(4): 17-21.
- [5] 张聪丛, 郇颖颖, 赵畅, 等. 开放政府数据共享与使用中的隐私保护问题研究: 基于开放政府数据生命周期理论[J]. 电子政务, 2018(9): 24-36.
- ZHANG C C, GAO Y Y, ZHAO C, et al. Research on privacy protection in open government data sharing and use—based on open government data life cycle theory[J]. E-Government, 2018(9): 24-36.
- [6] 李瀛, 杨芮. 我国政府数据开放的隐私保护困境及管理框架研究[J]. 情报杂志, 2023, 42(1): 152-157.
- LI Y, YANG R. Research on privacy protection dilemma and management framework of open government data in China[J]. Journal of Intelligence, 2023, 42(1): 152-157.
- [7] KONEČNÝ J, MCMAHAN H B, YU F X, et al. Federated learning: strategies for improving communication efficiency[EB]. arXiv preprint, 2016, arXiv: 1610.05492.
- [8] 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.
- YANG Q. AI and data privacy protection: the way to federated learning[J]. Journal of Information Security Research, 2019, 5(11): 961-965.
- [9] 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): 1-19.
- [10] LIU B Y, WANG L J, LIU M, et al. Federated imitation learning: a privacy considered imitation learning framework for cloud robotic systems with heterogeneous sensor data[EB]. arXiv preprint, 2019, arXiv: 1909.00895.
- [11] GONG Y, JIANG Z W, FENG Y F, et al. EdgeRec: recommender system on edge in mobile Taobao[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM, 2020: 2477-2484.
- [12] 王海燕. 抖音的算法推荐特点分析[J]. 新媒体研究, 2018, 4(20): 21-22.
- WANG H Y. Analysis of algorithm recommendation characteristics of Tiktok [J]. NewMedia Research, 2018,4(20): 21-22.
- [13] 李宗辰, 张颐, 叶东, 等. 运用联邦学习技术推动公安大数据的融合运用[J]. 警察技术, 2021(6): 34-37.
- LI Z C, ZHANG Y, YE D, et al. Using federated learning technology to promote the integration and application of public security big data[J]. Police Technology, 2021(6): 34-37.
- [14] 钱学胜, 吴寰宇, 陈诚, 等. 基于联邦学习的新冠肺炎疫情跨区域智慧防控技术: 以上海市实践为例[J]. 科技导报, 2021, 39(24): 96-107.
- QIAN X S, WU H Y, CHEN C, et al. Cross-region smart COVID-19 pandemic management and control based on federated learning: with Shanghai as an example[J]. Science & Technology Review, 2021, 39(24): 96-107.
- [15] SHI D Y, TONG Y X, ZHOU Z M, et al. Learning to assign: towards fair task

- assignment in large-scale ride hailing[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York: ACM, 2021: 3549–3557.
- [16] 戴理朋, 杨鑫, 徐茹枝. 联邦学习在电力数据分析中的应用及隐私保护研究[J]. 电力信息与通信技术, 2022, 20(11): 47–56.
- DAI L P, YANG X, XU R Z. Research on the application of federated learning in power data analysis and privacy protection[J]. Electric Power Information and Communication Technology, 2022, 20(11): 47–56.
- [17] 刘隽良, 王月兵, 覃锦端, 等. 数据安全实践指南[M]. 北京: 机械工业出版社, 2022: 42–82.
- LIU J L, WANG Y B, QIN J D, et al. Data security practice guide[M]. Beijing: China Machine Press, 2022: 42–82.
- [18] 高磊, 赵章界, 林野丽, 等. 基于《数据安全法》的数据分类分级方法研究[J]. 信息安全研究, 2021, 7(10): 933–940.
- GAO L, ZHAO Z J, LIN Y L, et al. Research on data classification and grading method based on data security law[J]. Journal of Information Security Research, 2021, 7(10): 933–940.
- [19] DE CRISTOFARO E, TSUDIK G. Practical private set intersection protocols with linear complexity[C]//SION R. International Conference on Financial Cryptography and Data Security. Berlin: Springer, 2010: 143–159.
- [20] PINKAS B, SCHNEIDER T, ZOHNER M. Faster private set intersection based on OT extension[C]//Proceedings of the 23rd USENIX conference on Security Symposium. New York: ACM, 2014: 797–812.
- [21] 王力, 张秉晟, 陈超超. 隐私保护机器学习[M]. 北京: 电子工业出版社, 2021: 88.
- WANG L, ZHANG B S, CHEN C C. Privacy protection machine learning[M]. Beijing: Electronic Industry Press, 2021: 88.
- [22] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734–749.
- [23] 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.
- [24] 梁锋, 羊恩跃, 潘微科, 等. 基于联邦学习的推荐系统综述[J]. 中国科学: 信息科学, 2022, 52(5): 713–741.
- LIANG F, YANG E Y, PAN W K, et al. Survey of recommender systems based on federated learning[J]. Scientia Sinica (Informationis), 2022, 52(5): 713–741.
- [25] ZHANG Z Q, CHEN C C, ZHOU J, et al. An industrial-scale system for heterogeneous information card ranking in alipay[C]//Proceedings of the International Conference on Database Systems for Advanced Applications. Cham: Springer, 2018: 713–724.
- [26] CHEN C C, ZHENG X L, WANG Y, et al. Context-aware collaborative topic regression with social matrix factorization for recommender systems[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2014, 28(1): 9–15.
- [27] 仇阿根, 张用川, 罗宁, 等. 结合用户特征的政务服务协同过滤推荐方法[J]. 集成技术, 2023, 12(1): 42–55.
- QIU A G, ZHANG Y C, LUO N, et al. Government service collaborative filtering recommendation method based on user characteristics[J]. Journal of Integration Technology, 2023, 12(1): 42–55.
- [28] 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.

[29] 刘俊旭, 孟小峰. 机器学习的隐私保护研究综述[J]. 计算机研究与发展, 2020, 57(2): 346–362.

LIU J X, MENG X F. Survey on privacy-preserving machine learning[J]. Journal of Computer Research and Development, 2020, 57(2): 346–362.

[30] 陈凯, 杨强. 隐私计算[M]. 北京: 电子工业出版社, 2022: 130.

CHEN K, YANG Q. Privacy computing[M]. Beijing: Electronic Industry Press, 2022: 130.

[31] 王腾, 霍峥, 黄亚鑫, 等. 联邦学习中的隐私保护技术研究综述[J]. 计算机应用, 2023, 43(2): 437–449.

WANG T, HUO Z, HUANG Y X, et al. Review on privacy-preserving technologies in federated learning[J]. Journal of Computer Applications, 2023, 43(2): 437–449.

[32] 杨强, 童咏昕, 王晏晨, 等. 群体智能中的联邦学习算法综述[J]. 智能科学与技术学报, 2022, 4(1): 29–44.

YANG Q, TONG Y X, WANG Y S, et al. A survey on federated learning in crowd intelligence[J]. Chinese Journal of Intelligent Science and Technology, 2022, 4(1): 29–44.

[33] 邱鑫源, 叶泽聪, 崔脩龙, 等. 联邦学习通信开销研究综述[J]. 计算机应用, 2022, 42(2): 333–342.

QIU X Y, YE Z C, CUI X L, et al. Survey of communication overhead of federated learning[J]. Journal of Computer Applications, 2022, 42(2): 333–342.

[34] VARGHESE B, WANG N, BARBHUIYA S, et al. Challenges and opportunities in edge computing[C]//Proceedings of the 2016 IEEE International Conference on Smart Cloud (SmartCloud). Piscataway: IEEE Press, 2016: 20–26.

[35] HUANG W K, YE M, DU B. Learn from others and be yourself in heterogeneous federated learning[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 10133–10143.

[36] VOGELS T, HE L, KOLOSKOVA A, et al. RelaySum for decentralized deep learning on heterogeneous data[EB]. arXiv preprint, 2021, arXiv: 2110.04175.

[37] 黄聿辰, 赵彦超, 郝江山, 等. 面向数据异构的联邦学习性能优化研究[J]. 小型微型计算机系统, 2024, 45(4): 777–783.

HUANG Y C, ZHAO Y C, HAO J S, et al. Research on performance optimization of Federated learning for data heterogeneity[J]. Small Micro Computer Systems, 2024, 45(4): 777–783.

作者简介



吴坚平 (1983–), 男, 浙江大学计算机科学与技术学院博士生, 主要研究方向为联邦学习、公共数据安全、推荐系统。



陈超超 (1988-), 男, 博士, 浙江大学计算机科学与技术学院特聘研究员、博士生导师, 主要研究方向为隐私保护机器学习、分布式机器学习、图机器学习和推荐系统等。



金加和 (1965-), 男, 浙江省数据开放融合关键技术研究重点实验室副主任, 浙江省大数据发展中心主任、正高级工程师, 浙江省政务服务标准化技术委员会副秘书长。主要研究方向为数字政府、公共数据融合应用、数据治理、隐私计算等。



吴春明 (1967-), 男, 浙江大学系统结构与网络安全研究所副所长、教授、博士生导师, 主要研究方向为新型网络体系结构、软件定义网络、工业互联网内生安全等。

收稿日期: 2023-12-20

通信作者: 吴春明, wuchunming@zju.edu.cn

基金项目: 浙江省“尖兵”“领雁”研发攻关计划项目 (No.2022C01243)

Foundation Item: Research and Development Plan Project for “Top Soldiers” and “Leading Wild Goose” in Zhejiang Province (No.2022C01243)