



基于联邦学习的社区数字化学习数据挖掘研究*

崔东浩^{1,2} 张华亮^{1,2}

(1.上海市长宁区业余大学, 上海 200336)

(2.上海开放大学航空运输学院, 上海 200336)

摘要: 对社区数字化学习数据的挖掘有助于促进社区居民个性化学习、提升社区教育管理决策水平和提高社区教育学习成果认证的准确性。然而, 传统机器学习方法基于数据集中训练进行数据挖掘, 面临个人隐私和数据安全相关的监管, 道德、法律及技术挑战。研究表明, 联邦学习作为强调数据安全和隐私保护的分布式机器学习技术, 可以通过数据不动模型动、数据可用不可见的方式, 针对区域内、跨区域以及跨教育类型的数字化学习数据, 分别采用纵向、横向和迁移这三种不同类型的联邦学习方法进行数据挖掘。因此, 基于联邦学习, 可以分析相似相关数据, 解决新用户冷启动问题; 深度分析用户学习行为, 推荐个性化学习资源; 利用正规教育管理和学习数据, 提升社区教育管理决策水平; 联通各级各类教育数据, 助力社区教育学习成果认证。

关键词: 联邦学习; 数字化; 社区教育; 数据挖掘; 人工智能; 机器学习

中图分类号: G720 **文献标识码:** A **文章编号:** 1674-9154(2023)06-0100-11

DOI: 10.16851/j.cnki.51-1728/g4.20231203.001

一、问题提出

数字化学习是指学习者在数字化学习环境中, 利用数字化学习资源, 以数字化方式进行学习的过程^[1]。社区数字化学习是指在社区居民以数字化的方式进行自主学习、协作学习的一种学习过程^[2]。社区数字化学习数据是社区居民进行数字化学习过程中所产生的数据, 其数据来源覆盖全国各级各类社区数字化学习平台。大数据环境下, 通过数据挖掘深入探究学习行为以及学习行为数据背后隐藏的行为模式和学习绩效关联性, 为社区居民提供更好的个性化学习支持服务, 已经成为未来发展趋势。

2022年11月, OpenAI公司发布的ChatGPT上线两个月活跃用户数量破亿, 被称为有史以来最好的人工智能聊天机器人^[3]。近年来人工智能技术的

发展进入了新时代, 机器学习、深度学习、学习分析等人工智能技术, 已经逐渐融入教育各领域, “人工智能+教育”成为人工智能技术应用的主要形态之一^[4]。机器学习作为大数据驱动的人工智能技术, 已成为数据挖掘和分析的重要方法^[5]。传统机器学习方法是基于数据的集中训练进行数据挖掘^[6], 而集中收集和保存不同来源的分布式数据, 还要面临个人隐私和数据安全相关的监管, 道德、法律及技术挑战。有没有一种新兴机器学习技术适合社区数字化学习数据挖掘? 能否从技术层面应对个人隐私和数据安全方面的挑战? 社区数字化学习数据挖掘有哪些步骤? 数据挖掘成果如何应用到社区教育领域等问题, 亟须深入探讨和研究。

早在2016年, 谷歌首次提出联邦学习(Federated Learning)理论^[7]。作为机器学习的新范式,

*基金项目: 全国教育科学“十四五”规划2021年度教育部重点课题“教育现代化视域下社区教育体系研究”(编号: DKA210393)。

作者简介: 崔东浩(1972—), 上海市长宁区业余大学(长宁区社区学院), 上海开放大学航空运输学院信息中心副主任, 讲师, 硕士, 研究方向: 社区教育, 成人教育, 教育管理及教育技术; 张华亮(1981—), 上海市长宁区业余大学(长宁区社区学院), 上海开放大学航空运输学院社区教育指导中心主任, 副教授, 硕士, 研究方向: 成人教育, 教育管理, 社区教育。

收稿日期: 2023-06-05





联邦学习可以化解大数据发展所面临的数据安全、数据隐私、数据孤岛等难题^[8],为社区数字化学习数据挖掘提供了新颖的解决方案。目前,联邦学习已初步应用于政府数据运营^[9]、教育数据挖掘^[10]、个性化推荐^[11]、学生综合素养评价^[12]等领域。相较而言,联邦学习在社区教育领域却鲜有研究。本研究阐明社区数字化学习数据挖掘的意义,分析联邦学习内涵、核心优势、分类及应用场景,并基于联邦学习理论,探索社区数字化学习数据挖掘路径以及联邦学习在社区教育中的具体应用,为社区数字化学习数据挖掘提供新的思路。

二、社区数字化学习数据挖掘的意义

社区数字化学习数据承载着社区居民进行数字化学习的全部学习行为和学习成果,数据背后隐藏着社区居民多样化的学习需求和学习偏好、学习行为模式和普遍学习规律,具有重要的数据价值,有必要进行深入的数据挖掘和分析。

(一) 促进社区居民个性化学习

大数据环境下,社区居民的学习需求呈现多样化、多元化发展态势。社区不同人群学习需求存在差异性和倾向性,例如青少年学习者更关注个体发展需要,在职人员更关注工作与生活方面的学习,而老年人更关注未来生活的学习^[13]。用户进入学习平台后,如未能及时得到满意的学习资源推荐,未能感受有效的个性化服务,用户对学习平台的使用体验则会受到影响,可能导致用户放弃对学习平台的使用。通过数据挖掘和分析,更好地掌握社区居民多样化的学习需求和学习偏好,能更好地为社区居民提供有针对性的学习资源推荐和学习指导服务,促进社区居民个性化、精准化和智慧化的学习。

(二) 提升社区教育管理决策水平

传统的管理决策往往基于直觉和经验,受决策者的知识结构、成长环境、性格秉性等主观因素的影响,而大数据决策遵从将万物数据化的逻辑^[14],在海量、完整、多维度的数据基础上,通过数据挖掘找出数据背后隐藏的规律,有助于做出更加科学有效的决策。基于新兴技术的社区数字化数据挖

掘,能够从技术层面解决数据来源少、覆盖范围不足等问题,能够体现社区数字化学习数据整体的价值,为社区教育决策者提供更加深入、广泛和可信的数据支撑,提升社区教育管理决策水平。

(三) 提高社区教育学习成果认证准确性

我国有着全球规模最大最复杂的多元化学习成果^[15]。2019年2月,中共中央、国务院印发《中国教育现代化2035》,提出建立学习成果认证制度、国家学分银行制度、国家资历框架^[16]。其中,学习成果认证是实施国家资历框架的关键,是建设学分银行的基础。社区教育作为非正式教育形态,其学习成果来源多样且复杂,认证过程存在效率低、成本高、信息不对称和监管缺失等问题^[17]。社区数字化学习数据记录着学习者的学习行为和学习结果,是社区教育成果认证的重要依据,其数据挖掘可以提高学习成果认证的准确性和可信度,帮助社区教育管理者更加科学地进行学习成果认证。

三、基于联邦学习的社区数字化学习数据挖掘理论基础

当前的人工智能技术,尤其是机器学习,为数据挖掘提供了有效的技术手段,而传统的机器学习方法基于数据集中训练的方式进行数据挖掘,不应用于分散在不同地区的社区数字化学习数据的挖掘上。联邦学习方法突破数据集中训练的瓶颈,采用分布式机器学习方法进行数据训练,能够更好地实现社区数字化学习数据的挖掘和分析。

(一) 联邦学习的定义及内涵

1. 联邦学习提出的背景

人工智能的力量来自于大数据,2016年AlphaGo在30万盘人类对局棋谱的基础上训练模型,击败了顶尖的人类围棋职业选手。在传统机器学习训练模型中,通常所有参与计算的数据拥有方都会被集中收集到中心服务器或者上传到云上进行处理。然而在大多数行业中数据往往以孤岛的形式存在,由于行业竞争、隐私安全、复杂的管理流程、高昂的成本等原因,打破数据壁垒、整合分散在全国各地的数据几乎是不可能完成。

集中收集和保存不同来源的数据,面临着诸多

挑战。其一,个人隐私保护已成为个人用户和政府都高度关注的内容,《中华人民共和国网络安全法》《数据安全管理办法(征求意见稿)》《中华人民共和国个人信息保护法(草案)》都针对数据的安全与隐私保护提出了更严格的要求,集中收集和保存用户数据,会面临与个人隐私和数据安全相关的监管,道德、法律及技术的挑战。其二,在大多数行业,数据的产生、管理和维护需要花费大量的时间、精力和费用,这样的数据具有重要的业务价值,使得这些数据不太可能被自由共享。相反,数据拥有方通常对他们管理的数据保持细致的控制,数据仍以孤岛的形式存在,很难打通数据壁垒实现集中训练。其三,机器学习所需的训练数据很容易就达到TB级,数据本身是分布式的且数据量太大而不太可能集中存储在单个机器上,集中式解决方案甚至不可取^[18]。为了解决上述问题,联邦学习的概念应运而生。

2. 联邦学习定义

联邦学习是一种带有隐私保护、安全加密技术的分布式机器学习框架,旨在让分散的各参与方在满足不向其他参与方披露隐私数据的前提下,协作进行机器学习模型训练^[19]。联邦学习强调的核心理念是“数据不动模型动,数据可用不可见”,保证数据在不出本地的前提下,各参与方之间协同构建训练模型。一方面,数据不出本地可以很好地保护用户隐私和数据安全;另一方面,能充分利用各参与方的数据协同训练模型。

在联邦学习中,各参与方首先从中心服务器下载初始共享模型参数,基于本地数据进行训练后把更新的模型参数上传至中心服务器,由中心服务器聚合各方的参数并形成全局模型并把全局模型参数再度下发出去,如此循环,直到全局模型收敛至停止条件^[20]。

设当前有N个数据拥有方共同参与训练,记为 $\{P_1, P_2, P_3, \dots, P_N\}$,其各自拥有的本地数据集为 $\{D_1, D_2, D_3, \dots, D_N\}$,传统的训练方法将所有参与方的数据收集起来存储在中心服务器中,并使用 $D=D_1 \cup D_2 \cup \dots \cup D_N$ 作为训练数据集训练机器学习模型 M_{SUM} 。联邦学习不需要收集各参与方所拥有的数

据,在参与方本地便能协同训练出模型 M_{FED} 。设 V_{SUM} 和 V_{FED} 分别为集中学习模型 M_{SUM} 和联邦学习模型 M_{FED} 的性能精度,设 δ 为任意一个非负实数,在满足以下条件时,称联邦学习模型 M_{FED} 具有 δ 的精度损失。

$$|V_{FED} - V_{SUM}| < \delta. \quad (1)$$

式(1)表示联邦学习模型 M_{FED} 的性能将以 δ 的精度损失近似于传统机器学习模型 M_{SUM} 的性能,当 $\delta=0$ 时,则表示联邦学习模型与传统机器学习模型的效果等价^[21]。

3. 联邦学习算法

FedAvg (Federated Averaging) 算法是目前最常用的联邦学习优化算法^[22]。与常规的优化算法不同,其本质思想是对数据持有方的局部随机梯度下降进行单机优化,并在中央服务器上聚合操作,其训练目标是找出最佳模型参数 ω ,使总训练损失 $f(\omega)$ 最小。FedAvg算法的目标函数定义如下:

$$f(\omega) = \sum_{k=1}^K \frac{n_k}{n} F_k(\omega), \quad (2)$$

$$F_k(\omega) = \frac{1}{n_k} \sum_{i \in P_k} f_i(\omega), \quad (3)$$

$$f_i(\omega) = l(x_i, y_i, \omega). \quad (4)$$

式(2)和式(3)中,K为参与方数量, P_k 为第k个参与方本地样本集, n_k 为第k个参与方本地样本数量即 $n_k=|P_k|$,n为所有参与方样本数量总和, $F_k(\omega)$ 为第k个参与方本地目标函数。式(4)中 $f_i(\omega)$ 为模型参数 ω 对 P_k 中的第i个样本 (x_i, y_i) 产生的损失函数。FedAvg是一种比较基础的联邦优化算法,部署相对简单,应用领域很广泛。FedAvg的算法如下:

(中央服务器方)

- 1 Server executes:
- 2 initialize ω_0
- 3 for each round $t = 1, 2, \dots$ do
- 4 $m \leftarrow \max(C \cdot K, 1)$
- 5 $St \leftarrow$ (random set of m clients)
- 6 for each client $k \in St$ in parallel do
- 7 $\omega_{t+1}^k \leftarrow \text{ClientUpdate}(k, \omega_t)$



$$8 \quad \omega_{t+1} \leftarrow \sum_{k=1}^K \frac{n_k}{n} \omega_{t+1}^k$$

(参与方)

9 ClientUpdate (k, ω) : // 第 k 个参与方更新

10 $B \leftarrow$ (split P_k into batches of size B)

11 for each local epoch i from 1 to E do

12 for batch $b \in B$ do

13 $\omega \leftarrow \omega - \eta \nabla l (\omega; b)$

14 return ω to server

其中, K 个客户端由 k 索引, B 为本地小批量数据集大小, E 为本地数据集遍历总次数, η 为学习率。大部分联邦优化算法是在 FedAvg 的基础上发展而来的, 例如 FedProx、FedPer 等。

(二) 联邦学习核心优势

基于联邦学习的社区数字化学习数据挖掘, 能够有效克服传统机器学习集中式数据训练的缺点, 打破数据孤岛、保障数据安全的同时达到预期模型性能效果, 使其成为社区数字化学习数据挖掘的理想解决方案。

1. 打破数据孤岛

通过“数据不动模型动”的方式, 学习平台之间无须共享原始数据, 让模型在学习平台之间沟通交流。整个训练过程均在学习平台本地进行, 并在中心服务器协调下完成全局模型的聚合, 在此过程中仅交换模型特征参数和梯度等^[23]。“联邦”就是数据联盟, 通过联邦学习方法, 各个学习平台以同等的身份和地位参与模型训练, 打破数据孤岛协同构建数据模型。

2. 保障数据安全

联邦学习技术实现“数据可用不可见”, 即数据都保留在本地, 别人看不见你的数据, 你也看不见别人的数据, 从而最大限度地减少数据隐私泄露问题。传输模型参数的明文信息也不安全, 也会存在信息泄露的风险, 攻击者通过窃取这些参数信息来还原或者部分还原原始数据, 从而导致数据隐私的泄露^[24]。联邦学习采用同态加密、差分隐私和安全多方计算等安全策略, 在联合训练时与模型相关的信息以加密方式在各方之间传输和交换, 任何一个参与方在接收到这些信息时都不能推测出对方的原始数据, 更好地保障数据的安全与隐私。

3. 提升模型性能

大数据是人工智能的基础, 模型的准确率与训练数据量成正比。联邦学习从技术上解决了数据拥有方所担心的隐私、所有权和数据位置的基本问题, 使大多数数据拥有方的参与成为可能。相较于传统集中式训练, 联邦学习拥有更多的参与方及训练数据, 训练出来的模型更准确。即便在数据量相同的情况下, 联邦学习技术可使分布式训练模型性能与传统集中式训练模型性能相差无几, 训练出的全局模型几乎是无损的。基于联邦学习的数据训练能够提升训练模型性能效果, 各参与方能够共同获益。

(三) 联邦学习分类及应用场景

设 D_i 表示第 i 个参与方本地训练数据集, 通常 D_i 以矩阵的形式存在, D_i 的每一行表示训练数据样本。每一条样本都有样本 ID, 将样本 ID 空间设为 I; 每一列表示一个具体的数据特征, 将特征空间设为 X; 同时, 一些数据集还可能包含标签数据, 将标签空间设为 y。样本 ID 空间 I、特征空间 X 和标签空间 y 组成一个训练数据集 $D_i: (I, X, y)$ 。根据各个参与方的样本 ID 空间 I 和数据特征空间 X 的重叠关系不同, 将联邦学习划分为纵向联邦学习、横向联邦学习和联邦迁移学习^[25], 分别适用于不同的应用场景。社区数字化学习数据训练样本中, 用户 ID 组成样本 ID 空间, 学习资源、学习行为、学习结果等信息共同组成特征空间。

1. 纵向联邦学习

纵向联邦学习适用于参与方的数据样本重叠较多而数据特征重叠较少的场景, 即参与方之间的数据样本是对齐的, 但在数据特征上有所不同。纵向联邦学习 (见图 1), 类似于在表格视图中垂直划分数据的情况。

纵向联邦学习: 特征划分, 数据特征增加

ID	X1	X2	X3	X4	X5	X6	ID
u1							
u2							u2
u3							u3
							u4

参与方 1 参与方 2

图 1 纵向联邦学习示意图

应用场景：某一个地区有一个直播教学网站和一家数字图书馆，为当地社区居民提供学习服务，分别以视频资源和数字图书为学习内容。两个学习平台提供不同的学习服务，但在用户群体上有着非常大的交集，可以联合双方的数据进行建模，以此来提升双方对用户的学习服务能力。例如，数字图书馆中要预测新用户对某一本书的阅读概率，但数字图书馆缺乏新用户的阅读数据，而直播教学网站上有该用户的视频观看记录，这部分信息的数据特征能够很好地体现该用户的学习偏好，如果能将这部分特征补充到建模中，无疑将有效提升模型预测的能力。两个学习平台具有重叠度很高的当地用户群体，但两者提供的学习内容不同，从纵向联邦的角度来说两者的特征不同（一方是视频的特征，另一方是书籍的特征），通过纵向联邦学习在各自有差异的数据特征空间上协同训练，可以得到性能更优的机器学习模型。

2. 横向联邦学习

横向联邦学习适用于参与方的数据特征重叠较多而数据样本重叠较少的场景，即数据特征在参与方之间是对齐的，但是参与方拥有的数据样本有所不同。横向联邦学习（见图2），类似于在表格视图中将数据进行水平划分的情况。



图2 横向联邦学习示意图

应用场景：两个直播教学网站分别为各自所属地区的社区居民提供直播教学服务。通常社区居民选择当地的网站进行学习，各地区直播教学网站的用户重合度比较低。两个网站都提供直播教学服务，有着相似的学习资源、学习行为、学习结果等数据特征。这两个网站的用户群体重叠部分较小，但在数据特征维度上的重叠部分较大。通过横向联邦学习，联合双方的学习数据，扩大训练的样本空

间，从而提升模型的准确度和泛化能力。

3. 联邦迁移学习

联邦迁移学习适用于参与方的数据样本和数据特征都很少重叠的场景，联邦迁移学习是对横向联邦学习和纵向联邦学习的补充。联邦迁移学习（见图3），联邦迁移学习特别适合处理异构数据的联邦问题。

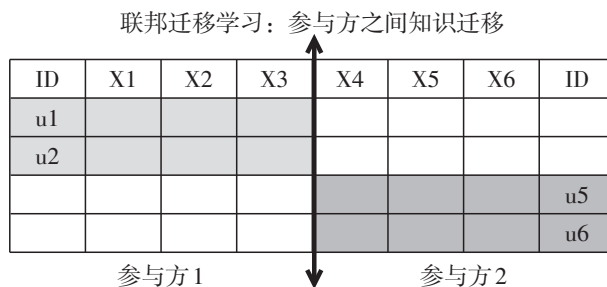


图3 联邦迁移学习示意图

应用场景：一个网站为社区居民提供社区教育课程资源服务，另一个网站为中小學生提供基础教育课程资源服务。虽然都为用户提供学习服务，但因为教育类型不同，两个网站所服务的用户群体基本不重叠，学习资源、学习行为、学习结果等数据特征也大不相同。利用联邦迁移学习，可以找到两个网站服务内容的相似性和差异性。相较于社区教育，基础教育的课程资源更全面，学习服务更系统，用户的学习行为数据也更完整。通过联邦迁移学习的知识迁移，社区教育课程资源网站可以学习基础教育中完善的学习数据，扩充自身的数据特征，提升模型的性能效果，更好地为社区居民提供学习服务。

四、基于联邦学习的社区数字化学习数据挖掘路径探索

全国各地的数字化学习普及程度不一，导致不同区域数字化学习平台的种类和数量有较大差异，同时不同区域乃至不同教育类型的数字化学习用户群体也截然不同，带来学习者用户数据以及数字化学习过程中产生的学习行为数据有着较大的差异。针对区域内、跨区域以及跨教育类型的数字化学习数据，须采用不同类型的联邦学习方法分别进行数据挖掘和分析。



（一）基于纵向联邦学习挖掘区域内社区数字化学习数据

随着社区教育的发展,全国各地都在建设各类社区数字化学习平台,如社区直播教学网站、社区数字图书馆、社区慕课在线平台等,为当地居民提供各类学习服务。社区居民通过当地社区数字化学习平台进行学习,并留下学习行为数据,这些数据背后隐藏着人们的学习习惯、学习方式和学习需求。但实际上这些学习平台上的学习数据,其数据规模还不够大,数据质量参差不齐,单个学习平台上进行机器学习往往得不到预期的训练效果。联邦机器学习能够联合当地不同类型的多个学习平台协同训练,参与训练的数据量成倍地增长,能够有效克服数据质量的差异,训练出性能优越的数据模型,能够更好地体现本区域社区居民的学习需求和学习偏好。这里所指的区域,其行政范围可以根据实际需要划定为区县、城市或省份。因为提供不同类型的学习服务,这些学习平台之间学习资源、学习行为、学习结果等数据特征有着较大的区别,但都有共同的当地用户群体,具有高度重叠的用户ID,因此适合采用纵向联邦学习的方法进行模型训练。

基于纵向联邦学习的区域内社区数字化学习数据模型训练示意图(见图4)。具体步骤如下:①全局模型(区域模型)参数加密后下发。中心服务器

确定全局模型的参数、训练过程和加密方式,并将全局模型参数加密后下发给区域内各参与方,如社区直播教学网站、社区数字图书馆、社区慕课在线平台等。②本地模型更新与训练。各参与方接收到全局模型参数后进行解密并更新本地模型,分别利用本地数据进行训练。③本地模型参数加密后上传。各参与方将训练好的本地模型的参数加密后回传给中心服务器。④全局模型聚合与更新。中心服务器对收到的各个本地模型参数解密,然后聚合参数信息并更新全局模型。更新后的全局模型的参数加密后再次下发给各参与方,如此重复以上训练步骤,循环迭代更新全局模型,直到全局模型收敛至停止条件,精度损失减少到可接受的程度。

（二）基于横向联邦学习挖掘跨区域社区数字化学习数据

社区数字化学习平台大都以当地社区居民为服务对象,不同区域学习平台的用户群体基本上没有重叠。相较于某区域单个学习平台上的模型训练,从分散在全国各地的学习平台中选择具有代表性的同类型学习平台进行联合训练,训练出来的数据模型具有更好的性能效果。一方面,参与训练的学习数据成倍地增长,在此大数据基础上训练出来的数据模型更准确更具有代表性。另一方面,不同区域的社区居民所处的生活环境、工作环境都有较大的差别,居民的学习习惯和学习方式,关注的学习内

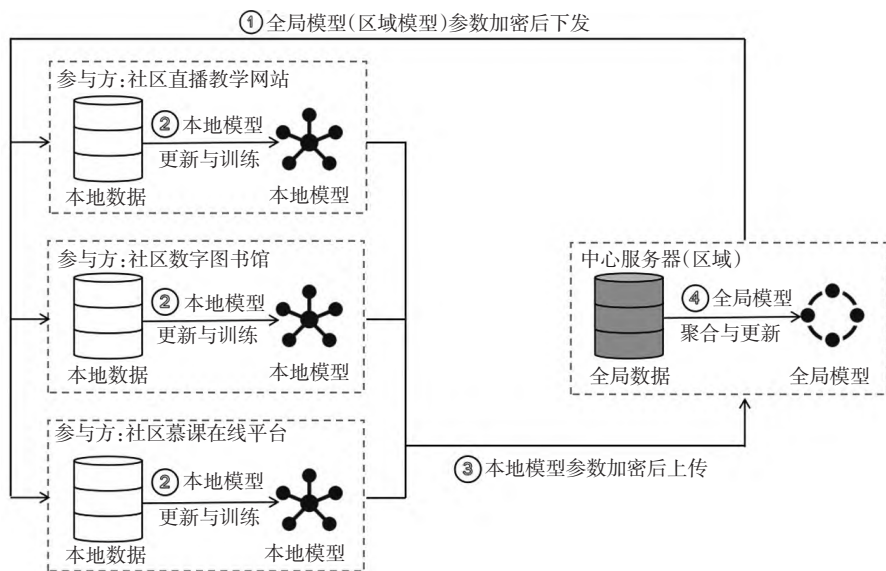


图4 基于纵向联邦学习的区域内社区数字化学习数据模型训练示意图

容和社会热点也会有较大的差异,因此基于跨区域学习平台训练出来的数据模型更具有普适性,从整体上体现不同区域社区居民的普遍学习需求和学习偏好。因为服务的区域不一样,这些学习平台之间用户ID基本上不会重叠,但因为选择同类型学习平台进行训练,其学习资源、学习行为、学习结果等数据特征基本类似,因此采用横向联邦学习方法构建跨区域社区居民的学习行为模型。

基于横向联邦学习的跨区域数字化学习数据模型训练示意图(见图5),基本流程与纵向联邦学习类似。具体步骤如下:①全局模型(跨区域模型)参数加密后下发。跨区域中心服务器确定全局模型的参数、训练过程和加密方式,并将全局模型参数加密后下发给各区域参与方,如区域A学习网站、区域B学习网站、区域C学习网站等。②本地模型更新与训练。各区域参与方接收到全局模型参数后解密并更新本地模型,分别利用本地数据进行训练。③本地模型参数加密后上传。各区域参与方将新的本地模型参数加密后回传给跨区域中心服务器。④全局模型聚合与更新。跨区域中心服务器接收到各区域本地模型参数后解密,然后聚合参数信息并更新全局模型。更新后的全局模型的参数加密后再次下发给各区域参与方,如此重复以上训练步骤,循环迭代更新全局模型,直到全局模型收敛至停止条件,精度损失减少到可

接受的程度。

(三) 基于联邦迁移学习挖掘跨教育类型数字化学习数据

联邦迁移学习方法将数据模型的训练扩展到其他的教育类型,如职业教育、高等教育和基础教育,共同构建跨教育类型数字化学习行为模型。从不同教育类型学习平台中,筛选出具有代表性的学习平台作为参与方进行联合模型训练,突破社区教育范围,最终训练出来的学习行为模型,可以从更高的维度反映不同教育类型受教育者的学习情况,有利于掌握受教育者的普遍学习需求。通过本地化的数据模型训练,联邦迁移学习不要求学习数据集中统一存放,使得不同教育类型学习数据的模型训练成为可能。不同教育类型的用户群体大不相同,学习资源、学习行为、学习结果等数据特征也有较大的差异,因此采用联邦迁移学习的方法构建跨教育类型数字化学习行为模型。

基于联邦迁移学习的跨教育类型数字化学习数据模型训练示意图(见图6)。具体步骤如下:①全局模型(跨教育类型)参数加密后下发。跨教育类型中心服务器确定全局模型的参数、训练过程和加密方式,并将全局模型参数加密后下发给各教育类型参与方,如社区教育学习网站、职业教育学习网站、高等教育学习网站和基础教育学习网站。②本地模型更新与训练。各教育类型参与方接收到全局

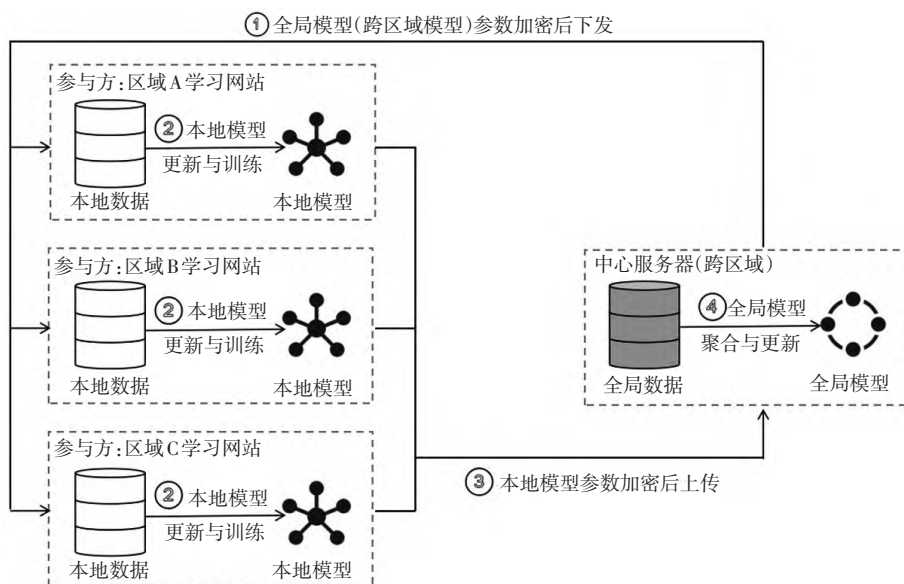


图5 基于横向联邦学习的跨区域社区数字化学习数据模型训练示意图

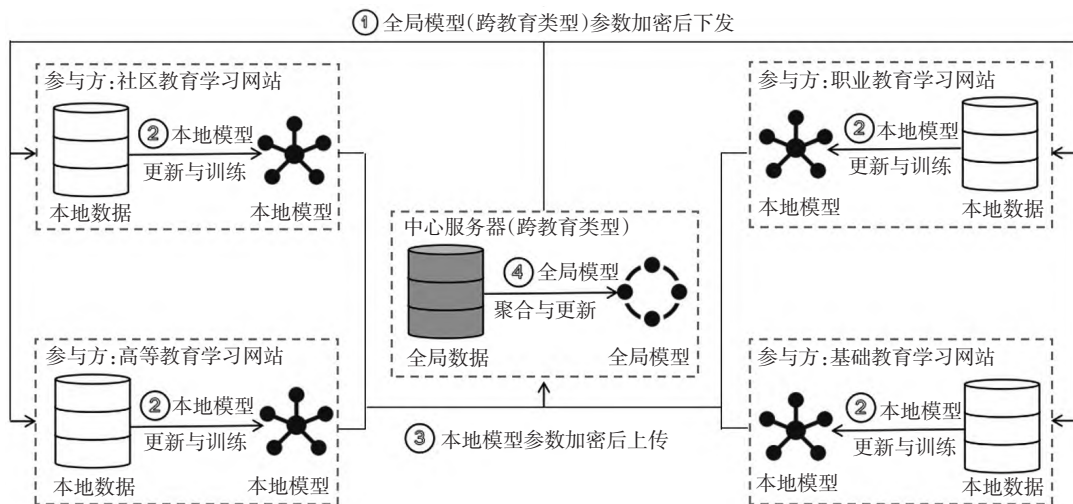


图6 基于联邦迁移学习的跨教育类型数字化学习数据模型训练示意图

模型参数后进行解密并更新本地模型,分别利用本地数据进行训练。③本地模型参数加密后上传。各教育类型参与方将新的本地模型参数加密后回传给跨教育类型中心服务器。④全局模型聚合与更新。跨教育类型中心服务器接收到各教育类型本地模型参数后解密,然后聚合参数信息并更新跨教育类型全局模型。更新后的全局模型的参数加密后再次下发给各教育类型参与方,如此重复以上训练步骤,循环迭代更新全局模型,直到全局模型收敛至停止条件,精度损失减少到可接受的程度。

五、基于联邦学习的社区数字化学习数据挖掘应用探索

基于联邦学习的分布式机器学习数据模型的训练,能够联合更多的参与方进行数据训练,适合跨平台跨区域跨教育类型的数字化学习数据的挖掘和分析,并能训练出更加精细和完整的数字化学习行为模型,更好地应用于社区教育的各种场景。

(一) 分析相似相关数据,解决新用户冷启动问题

当新用户登录到学习平台时,由于该用户在平台中缺少历史学习行为数据,学习平台不能根据传统算法预测其学习偏好,无法对新用户进行个性化推荐,由此产生新用户冷启动问题^[26]。用户的首次体验与学习平台提供的学习服务息息相关,良好的首次体验既能让用户在轻松愉快的心境中学习到有

用的知识,又能建立起用户预期与网站学习服务之间的桥梁。目前,针对新用户冷启动问题有较多的研究,有学者提出基于偏好的推荐算法,根据领域相关度、评价相似度从所有用户中筛选与新用户偏好相近的用户,并依据这些用户的偏好为新用户推荐学习资源^[27]。也有学者基于除用户偏好数据以外的其他相关数据,如性别、年龄、职业等用户的人口统计学数据,为新用户进行推荐^[28]。

基于纵向联邦学习的区域内社区数字化学习行为模型为解决新用户冷启动问题提供参考依据。一方面,依据新用户在其他各类学习网站上的学习数据推荐学习资源。新用户可能在学习网站上注册并学习,纵向联邦学习联合不同类型当地学习网站协同训练,了解新用户常去哪些当地的学习网站,喜欢看哪些类型的学习资源,以此来推荐本学习平台的类似学习资源。另一方面,依据其他用户的学习数据推荐学习资源。用户不是独立存在的,用户之间具有社会关系,那些现实中与该用户社会地理距离较近且行为习惯相近者对用户的偏好决策影响较大,地理信息和社会信息在用户个性化推荐中被广泛使用^[29]。将新用户所属区域的不同类型学习平台进行联合训练,结合新用户注册时填写的性别、年龄层次、职业类型、居住区域等粗粒度个人信息,找出同类型用户的学习偏好,有针对性地推荐新用户最有可能感兴趣的学习内容,有效解决冷启动的问题。



（二）深度分析用户学习行为，推荐个性化学习资源

目前各级各类社区教育学习平台中的有效学习用户数量远没有达到覆盖全体社区居民的程度，学习平台都面临着发展和维护用户的迫切需要。让更多的社区居民参与到终身学习中来，是每一个学习平台的首要任务。面对不同人群学习需求的多样化和多元化，平台应采取更加有效的措施提供个性化学习支持服务，提高用户满意度。学习平台中融入个性化推荐系统，能显著提高用户满意度和对学习平台的黏性。

纵向和横向联邦学习结合起来为社区居民提供个性化学习资源推荐。一方面，当前各类学习平台存在大量非结构化数据，如果能够精准分析这些庞杂的数据，形成精准的用户画像，学习平台的个性化推荐会得到质的飞跃。但是，形成精准用户画像的基础是建立用户数据标签体系，从基础数据到衍生指标，再到立体化数字画像标签体系，均依赖于用户属性、兴趣偏好、行为特征、学习习惯等不同维度的大量数据，而这些数据分散在不同类型的学习平台上。纵向联邦学习能够联合不同类型的学习平台进行协同训练，获取相同用户群体的不同维度的大量数据，有助于构建精准的用户画像，基于学习者画像可以构建学习预警模型，为学习者提供个性化的教学干预^[30]。另一方面，机器学习已经成为网站提供个性化推荐的一项重要技术，为了获得准确的机器学习模型，在其训练过程中需要大量的样本数据进行特征映射和参数调整。然而学习平台受本地数据样本不足的限制，而且跨区域数据共享又存在隐私泄露的风险。通过横向联邦学习技术，学习平台根据中央协调及异构平台的联合建模方式，实现跨区域模型参数的流通和共享，从而打破数据孤岛，在有效保护隐私的同时获得更高精度的数据模型，更好地实现个性化学习资源推荐。

（三）利用正规教育管理和学习数据，提升社区教育管理决策水平

相较于正规教育，社区教育缺乏统一规范的管理机制，其数据构成比较复杂。正规学校教育对学生的管理具有规范性、强制性，新生入学注册必须

正确填写个人信息并统一分配学号，毕业时根据学号就能查询到个人在校期间所有学习记录，包括成绩、学分等信息。社区教育具有非正规性，对社区居民的管理是松散的、学习是非强制的，其管理和学习数据没有统一标准。居民的学习具有随意性、自发性，学习网站上注册的个人信息未必准确和完整，社区教育课程大多数也没有统一的学习要求和标准，成绩认定、学分获得都有困难。管理和学习数据的不规范不完整将影响社区教育管理决策，影响社区教育发展。

不同教育类型的学习平台，因其管理体制差异、个人隐私保护等原因，其管理和学习数据基本上不可能集中在一起进行训练，联邦迁移学习借助本地训练数据的方式能够实现不同教育类型管理和学习数据的联合训练。基于联邦迁移学习的跨教育类型数字化学习行为模型，其数据样本既来自社区教育领域，也来自其他教育领域，能够利用正规学校教育的规范完整的管理和学习数据特征，填补社区教育不规范不完整的数据结构，以此来优化社区教育数据模型的性能效果。机器学习是一个从数据中自动生成知识的过程，借助联邦迁移学习的知识迁移，使社区教育管理者能够运用正规教育管理和学习数据中生成的知识，有效提升社区教育管理决策水平。

（四）联通各级各类教育数据，助力社区教育学习成果认证

学习成果是指学习者在完成一段时间的学习后，被期望知道、理解和能够做什么的陈述，学习成果重在描述学习者的成就，是一个人成功经历某一类型教育而取得的结果^[31]。但是，对于发生在学校系统之外的社区教育非正规和非正式学习活动，其形式多样，结构较为松散，过程不可预控，效果难以确定。面对终身学习背景下学习成果认证的各种现实需求，偏重正规教育的传统教育评价体系已难以相适应，迫切需要针对非正规教育和非正式学习成果的认证，以满足整个社会全民终身学习的需要^[32]。

基于联邦学习的数字化学习行为模型为学习成果认证提供数据支撑，提高认证的准确性和可信





度。一方面,社区数字化学习是产生社区居民学习成果的重要渠道,基于纵向和横向联邦学习的社区数字化学习行为模型,可通过对学习者的学习路径、学习记录、测试结果等学习数据的分析,为学习成果认证提供参考依据。另一方面,不同教育类型的学习成果认证需要学习成果之间具有可比性、等效性和可转换性,基于联邦迁移学习的跨教育类型数字化学习行为模型,对不同教育类型学习者的学习数据进行对比,将社区教育中的非正式学习成果与其他各级各类学习成果相互联通。数字化学习行为模型有助于社区教育学习成果的客观评估和鉴定,有助于不同教育类型学习成果获得实质等效的资历和学分,为建设终身教育资历框架和学分银行提供数据支撑。

参考文献:

- [1]李克东.数字化学习(上)——信息技术与课程整合的核心[J].电化教育研究,2001(8):46-49.
- [2]宋亦芳.论我国社区数字化学习的形成与发展[J].职教论坛,2013(9):60-65.
- [3]令小雄,王鼎民,袁健.ChatGPT爆火后关于科技伦理及学术伦理的冷思考[J].新疆师范大学学报(哲学社会科学版),2023,44(4):123-136.
- [4]陈颖博,张文兰.国外教育人工智能的研究热点、趋势和启示[J].开放教育研究,2019,25(4):43-58.
- [5]仲玮,李志琛.基于机器学习的网络教育系统研究[J].通信学报,2018,39(S1):135-140.
- [6]杨庚,王周生.联邦学习中的隐私保护研究进展[J].南京邮电大学学报(自然科学版),2020,40(5):204-214.
- [7]McMahan B, Moore E, Ramage D, et al.Communication-Efficient Learning of Deep Networks from Decentralized Data[EB/OL].(2020-10-11)[2023-10-15].<http://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf>.
- [8]李少波,杨磊,李传江,张安思,罗瑞士.联邦学习概述:技术、应用及未来[J].计算机集成制造系统,2022,28(7):2119-2138.
- [9]胡业飞,陈美欣,张怡梦.价值共创与数据安全的兼顾:基于联邦学习的政府数据授权运营模式研究[J].电子政务,2022(10):2-19.
- [10]李默妍.基于联邦学习的教育数据挖掘隐私保护技术探索[J].电化教育研究,2020,41(11):94-100.
- [11]李康康,袁萌,林凡.联邦个性化学习推荐系统研究[J].现代教育技术,2022,32(2):118-126.
- [12]郭利明,朱迁踏,郑勤华,王怀波.基于联邦学习的学生综合素养评价数据共享与安全保护研究[J].中国电化教育,2022(10):56-63.
- [13]宋亦芳.社区学习者终身学习需求的特征分析——基于2020年上海社区三类人群学习需求调研[J].中国职业技术教育,2021(18):82-88.
- [14]郑蕴铮,郑金洲.教育行动研究:成效、问题与改进[J].教育发展研究,2020,40(4):18-23.
- [15]张伟远.构建“资历框架为标准、学习成果认证为保障”的学分银行制度[J].中国职业技术教育,2020(24):5-7.
- [16]中华人民共和国教育部.中共中央、国务院印发《中国教育现代化2035》[EB/OL].(2019-02-23)[2023-10-15].http://www.gov.cn/xinwen/2019-02/23/content_5367987.htm.
- [17]韦书令.融合区块链技术的社区教育学习成果认证模型及实现机理研究[J].成人教育,2021,41(8):71-77.
- [18]薄列峰,黄恒,顾松庠,陈彦卿等.联邦学习:算法详解与系统实现[M].北京:机械工业出版社,2022:10.
- [19]王健宗,李泽远,何安珣.深入浅出联邦学习:原理与实践[M].北京:机械工业出版社,2021:11.
- [20]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.
- [21]杨强,刘洋等.联邦学习(Federated Learning)[M].北京:电子工业出版社,2020:5.
- [22]McMahan, H. Brendan, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data[EB/OL].(2023-01-26)[2023-10-15].<https://arxiv.org/abs/1602.05629>.
- [23]Hard A, Rao K, Mathews R, et al. Federated Learning for Mobile Keyboard Prediction[EB/OL].(2019-02-28)[2023-10-15].<https://arxiv.org/abs/1811.03604>.
- [24]杨强,黄安埠,刘洋,陈天健等.联邦学习实战[M].北京:电子工业出版社,2021:15.
- [25]杨强.联邦学习:人工智能的最后一公里[J].智能系统学报,2020,15(1):183-186.
- [26]王国霞,刘贺平.个性化推荐系统综述[J].计算机工程与应用,2012,48(7):66-76.
- [27]朱锐,王怀民,冯大为.基于偏好推荐的可信服务选择[J].软件学报,2011,22(5):852-864.
- [28]Stuart E. Middleton and Nigel R. Shadbolt and David C. De Roure. Ontological user Profiling in Recommender Systems[J].

ACM Transactions on Information Systems (TOIS), 2004, 22 (1):54-88.

[29]安敬民,李冠宇,蒋伟,等.基于用户活动轨迹和个性化区域划分的兴趣点推荐[J].计算机学报,2022,45(6):1176-1194.

[30]成亚玲,谭爱平.基于学习者画像的在线开放课程学习预警研究[J].当代职业教育,2022(4):102-112.

[31]严中华.国外职业教育核心理念解读:学习成果导向职业教育课程开发理论与实践[M].北京:清华大学出版社,2017:7.

[32]王海东.学习成果认证制度研究[M].北京:中国人民大学出版社,2017:2.

责任编辑 黎恩 乔兴媚

Research on Mining of Community Digital Learning Data Based on Federated Learning

CUI Donghao^{1,2}, ZHANG Hualiang^{1,2}

(1.Shanghai Changning Spare-time University, Shanghai 200336)

(2.Shanghai Open University Air Transport College, Shanghai 200336)

Abstract: The mining of community digital learning data helps to promote personalized learning for community residents, enhance the management decision-making ability of community education, and improve the accuracy of community education learning achievement certification. However, traditional machine learning methods based on centralized data training face regulation, ethical, legal, and technical challenges related to personal privacy and data security. Research has shown that federated learning, as a distributed machine learning technology that emphasizes data security and privacy protection, can use three different types of federated learning methods, namely vertical, horizontal, and transfer, to mine digital learning data within regions, across regions, and education types through data immovable but model movable, data available but invisible way. Therefore, based on federated learning, similar related data can be analyzed to solve the cold-start issue of new users, user learning behavior can be deeply analyzed to recommend personalized learning resources, formal education management and learning data can be used to improve the management decision-making ability of community education, data of various types of education at all levels can be connected to assist community education learning achievement certification.

Keywords: federated learning; digitalization; community education; data mining; artificial intelligence; machine learning