第 62 卷　第 5 期
2026 年 9 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 62, No. 5
Sept., 2026

DOI:10.13232/j.cnki.jnju.2026.05.012

面向认知诊断的机器遗忘框架

汪卫星[1],蔡伦忠[2],林铭炜[2*],林佳胤[2]

(1.广东开放大学人工智能学院,广州,510091;2.福建师范大学计算机与网络空间安全学院,福州,350117)

摘　要:认知状态模型需要具备数据遗忘能力,以符合相关法律法规及隐私保护要求.然而,现有的认知诊断模型聚焦于性能优化,忽视了学习者的隐私需求.为此,提出一种面向认知诊断的机器遗忘框架,以解决认知诊断中的机器遗忘问题.首先,基于认知状态的分区方法,根据学习者的认知状态相似度对数据进行划分,并在各个分区内并行训练认知诊断子模型;其次,基于交叉注意力的模型聚合机制,通过融合预测数据与子模型的特征计算注意力权重,提升最终模型在不同学习者群体上的泛化能力.基于四个真实数据集进行了大量的实验,实验结果证明该框架能够有效实现认知诊断模型的数据遗忘,在保证模型性能的同时,大幅降低计算资源消耗和训练时间.

关键词:机器遗忘学习,认知诊断,认知状态,机器学习,隐私保护

中图分类号:TP391　　**文献标志码**:A

A machine unlearning framework for cognitive diagnosis

Wang Weixing[1], Cai Lunzhong[2], Lin Mingwei[2*], Lin Jiayin[2]

(1. School of Artificial Intelligence, Guangdong Open University, Guangzhou, 510091, China;
2. School of Computer and Cyberspace Security, Fujian Normal University, Fuzhou, 350117, China)

Abstract: Cognitive state models must possess data-forgetting capabilities to comply with relevant laws and privacy protection requirements. However, existing cognitive diagnostic models focus on performance optimization while neglecting learners' privacy needs. To address this gap, we propose a machine-forgetting framework for cognitive diagnosis that tackles the problem of data forgetting in this domain. First, leveraging a partitioning approach based on cognitive states, we segment the data according to the similarity of learners' cognitive states and train cognitive diagnostic submodels in parallel within each partition. Second, employing a cross-attention-based model aggregation mechanism, we compute attention weights by fusing prediction outputs with the feature representations of the submodels, thereby enhancing the final model's generalization across diverse learner groups. Extensive experiments conducted on four real-world datasets demonstrate that our framework effectively enables data forgetting in cognitive diagnostic models, significantly reducing computational resource consumption and training time while maintaining model performance.

Keywords: achine unlearning, cognitive diagnosis, cognitive state, machine learning, privacy protection

随着智能教育的快速发展,认知诊断逐渐受到广泛关注,成为教育数据挖掘中的一项核心任务[1−2].认知诊断通过分析学习者的学习行为(如答题记录),评估其在不同知识领域的掌握程度.图1展示了认知诊断的基本流程.通常,学习者在完成课程学习后需要进行练习测试,以评估其对相关知识点的掌握情况.认知诊断模型(Cognitive Diagnosi Models,CDMs)对这些答题记录进

基金项目:广东开放大学人才项目(2021F001),广东开放大学2025年度科研课题重点项目(ZD2502)
收稿日期:2026−07−29
*通信联系人,E-mail:linmwcs@163.com

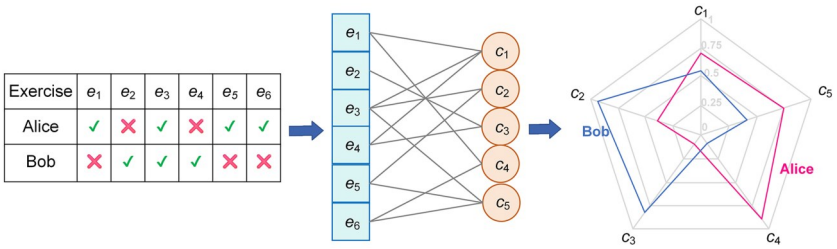


图 1　认知诊断流程

Fig. 1　Cognitive diagnosis process

行分析,生成诊断报告,帮助学习者识别薄弱环节,并为其提供针对性的学习支持和优化建议.

尽管智能教育系统对提升学习效果评估与分析表现卓越,但其广泛应用也引发了严峻的安全隐患与隐私伦理挑战.一方面,平台可能在未经请求情况下过度采集与分析学生数据;另一方面,认知诊断模型在训练过程中可能无意间存储了学习者的敏感属性信息.为了更好地保护个人隐私,世界各地的立法机构都制定了明确的法律,包括欧盟的《通用数据保护条例》(GDPR)[3]、美国加利福尼亚州的《消费者隐私法案》(CCPA)[4]和我国的《个人信息保护法》[5].这些法规都包含"被遗忘权"的条款,即要求企业采取合理的措施消除有关个人的敏感数据[6].随着技术的发展,法规的要求更严格,模型提供方不仅要彻底删除隐私数据,同时还需清除这些敏感数据的影响,这一过程被称为"机器遗忘"(也称数据遗忘)[7].

为了解决"被遗忘权"带来的隐私问题,研究人员提出一项新概念,即机器遗忘学习(Machine Unlearning,MU).作为一种新兴的隐私保护范式,机器遗忘旨在通过特定的算法机制,从模型中精准擦除特定用户的行为数据及其对模型参数的贡献,从根本上阻断隐私泄露的路径.该方法不仅在技术层面实现对敏感信息的有效清除,更在法律层面确保模型对"被遗忘权"的实质性履行,是在保障数据安全合规的前提下维持模型分析效能的关键手段.

近年来,机器遗忘已成为学术界的研究热点,其核心目标在于从机器学习模型中精确移除特定数据的影响.在遗忘的有效性方面,重训练(Retrain)被视为是最彻底的方法[8],即通过在删除目标数据后的数据集上重新训练模型来实现完全遗忘.然而,尽管该方法能保证理论上的遗忘效果,但在面对大规模数据集时,高昂的计算成本与时间开销使其难以满足实际应用中的实时性需求.为了降低计算负担,Bourtoule et al[9]提出代表性的SISA框架,将训练数据划分为多个互不重叠的分片并独立训练子模型,最终通过多数投票机制来聚合预测结果.SISA利用数据分片与并行训练策略,显著降低了重训练的代价.随后,针对推荐系统任务,Chen et al[10]的RecEraser框架引入基于用户-项目交互的自适应数据划分策略,并设计了基于注意力机制的聚合算法,在实现数据遗忘的同时有效保持了推荐模型的性能.然而,现有的机器遗忘框架难以直接应用于认知诊断任务.认知诊断模型的核心在于通过学生的答题序列挖掘其潜在的认知状态,现有方法在进行数据划分与模型聚合时,仅考虑了数据层面的交互特征,没有挖掘学生独特的认知属性,在数据划分阶段破坏了认知诊断所需的特征分布,导致模型在遗忘后的诊断准确性大幅下降.因此,设计一种能够感知认知状态的遗忘框架尤为迫切.

针对现有研究的局限性,本文提出一种面向认知诊断的机器遗忘框架(Unlearning Framework for Cognitive Diagnosis,UFCD),旨在高效实现认知诊断模型的数据遗忘,并有效缓解分区-聚合框架对模型性能的影响.UFCD的基础架构采用典型的分区-聚合框架,通过数据划分与并行训练,大幅降低模型重训练所需的时间与计算成本.然而,该方法不可避免地会影响最终模型的预测性能,为此,提出以下优化方案.

(1)基于认知状态的分区方法(Cognitive State-Based Partitioning,CSP).不同于随机划分,CSP旨在挖掘数据的潜在语义结构.首先,通过

认知诊断提取学习者的认知状态特征向量，构建认知特征矩阵；其次，基于学习者间的认知相似度对答题记录进行聚类划分．该策略确保每个数据分片内部的认知模式一致性，使训练出的子模型能够捕捉特定群体的认知规律，提升模型的个性化适应能力．

（2）基于交叉注意力的模型聚合机制（Cross-Attention-Based Model Aggregation Mechanism，CMAM）．针对传统静态平均或投票机制无法适应异构子模型的聚合问题来设计CMAM，将待预测样本的特征作为查询，各分区的认知状态分布特征作为键．通过综合评估预测样本与各子模型“擅长领域”的匹配度，动态计算注意力权重并进行自适应聚合，显著提升最终模型的预测精度与泛化能力．

为了全面评估所提框架的有效性，在四个真实世界数据集上选取三种具有代表性的认知诊断模型作为主干模型，对其应用不同的机器遗忘方法进行性能基准测试．

本文的主要贡献如下．

（1）填补认知诊断领域的遗忘技术空白．针对现有机器遗忘研究在教育应用场景中的局限性，提出一种专为认知诊断设计的机器遗忘框架（UFCD），创新性地将教育数据的认知属性融入遗忘机制，实现对学生隐私数据的高效遗忘，解决遗忘方法在认知诊断任务中适用性差的问题．

（2）构建认知语义驱动的遗忘机制，创新性地将认知聚类与交叉注意力机制相结合．CSP策略赋予子模型高度的局部语义一致性，CMAM实现了全局视角下的自适应聚合，从根本上缓解了传统分区遗忘框架带来的性能损耗．

（3）实现遗忘效率与诊断精度的平衡，不仅满足数据层面的物理隔离与完全遗忘，同时在特征层面保留高质量的认知表征．广泛的实验表明，UFCD在显著降低重训练开销的同时，其诊断性能逼近完整数据集上训练的原始模型．

1　相关工作

MU旨在高效且精准地从已训练的模型中消除特定数据样本的贡献，在保障模型可用性的同时，严格履行数据隐私法规（GDPR[3]和CCPA[4]）中关于“被遗忘权”的要求．与简单的数据物理删除不同，机器遗忘必须消除目标数据在模型参数中残留的“记忆”[11]．现有的机器遗忘学习主要分为两大类，即近似遗忘（Approximate Unlearning）与完全遗忘（Exact Unlearning）[12]．

1.1　近似遗忘　近似遗忘的核心思想是通过数学优化手段在现有模型参数上施加修正项，以近似模拟模型在剔除目标数据后重新训练的状态[13]．该类方法侧重于用一定的误差边界换取极高的计算效率．

（1）基于梯度更新的方法，通常利用一阶或二阶梯度信息来抵消遗忘数据的贡献．例如，DeltaGrad[14]利用历史梯度缓存与L-BFGS算法[15]快速估计修正参数，显著提升随机梯度下降过程中的遗忘效率．在联邦学习场景下，FedRecover[16]通过积分Hessian矩阵[17]技术结合预热校正策略，实现全局模型的低成本恢复．Down-to-Delete[18]方法引入差分隐私注入高斯噪声，在欧几里得空间内提供不可区分性保障．

（2）基于影响函数（Influence Functions，IF）的方法，致力于从统计学角度量化并消除数据对模型参数的影响．Guo et al[19]的“认证遗忘”框架利用影响函数近似计算单点删除的参数变化，并在强凸模型上提供了理论保障．在此基础上，Sekhari et al[20]进一步优化该框架的存储与计算开销，通过利用充分统计量，使算法能够在不访问完整历史数据集的情况下执行遗忘．针对深层非凸神经网络中Hessian矩阵求逆的计算瓶颈，Mehta et al[21]的基于稀疏参数选择的L-CODEC策略通过识别关键参数子空间并应用分块牛顿更新，将昂贵的二阶计算限制在低维空间中．Cheng et al[22]的GNNDELETE框架强调“删除一致性”与“邻域影响约束”，确保边的移除操作仅在局部子图中传播，从而在阻断特定依赖的同时，最大程度保留图的全局结构完整性．Chien et al[23]聚焦于遗忘的理论边界，针对简单图卷积网络（SGC）提出一种认证图遗忘学习方法CGU，让节点、边及特征遗忘得到严格理论保证．MEGU[24]采用基于学习的方法，通过构造损失函数来微调模型，在修正模型的同时实现推理保护．

1.2　完全遗忘　完全遗忘要求遗忘后的模型参

数分布与从头开始重新训练的模型保持绝对一致,从物理与数学层面提供最严格的隐私保障.目前,主流的完全遗忘技术高度依赖于“分区-聚合”框架.

(1)基础分区框架及其优化.Bourtoule et al[9]的SISA框架将数据划分为互不重叠的分片并独立训练子模型,使遗忘操作的代价被限制在局部子模型内.为了克服SISA随机划分导致的数据异构性,Yan et al[25]的ARCANE框架摒弃了传统的随机平衡划分策略,采用基于类别的划分机制,利用单类分类器独立训练各子模型.这种设计将遗忘操作的影响严格限制在单一类别边界内,有效缓解跨类干扰带来的精度损失.

(2)面向特定数据结构的遗忘.针对具有复杂拓扑结构的图数据与交互数据,传统的随机划分会切断隐式协作信号.在推荐系统领域,Rec-Eraser[10]提出基于用户-项目交互的均衡划分与自适应聚合机制,以保留局部协作信号.Li et al[26]的LASER框架利用超图技术创建平衡的数据划分,优化用户协同信息的保留.在聚合阶段,LASER采用逐步训练策略,从简单到复杂进行训练,以最大化模型对协同信息的利用.SARU模型[6]利用谱聚类结合大语言模型,成功在保证遗忘效率的同时,缓解数据移除导致的推荐准确率下降问题.针对图神经网络,GraphEraser[27]包含“均衡图划分、分片训练、重要性聚合”的三阶段流程,设计了保持图结构完整性的划分算法,生成节点规模均衡的子图分片,在聚合阶段,通过学习各分片模型的重要性权重,优化全局模型的效用.Zhang[28]的GraphRevoker框架旨在解决图划分带来的信息丢失问题,通过设计图属性感知的划分算法以及基于对比学习的子模型聚合策略,成功在加速遗忘过程的同时,显著提升遗忘操作的模型效用及可靠性.此外,对于天然契合SISA思想的树模型(随机森林[29]),Brophy and Lowd[30]的DARE方法进一步优化树内部的更新效率,采用双层节点结构,上层随机节点仅随机选择分裂属性与阈值,降低对数据的依赖,下层贪婪节点基于基尼指数等指标进行精细划分.通过缓存节点统计信息与叶节点数据,DARE仅需更新受影响的子树分支,而非整棵树.

尽管机器遗忘技术在图像视觉和推荐系统等领域已取得显著进展,但将其向认知诊断领域迁移时仍面临诸多挑战,现有的遗忘方法难以直接适配认知诊断模型在数据离散性、学习者群体异质性以及诊断上下文依赖性等方面的特殊需求,迫切需要构建一种专为认知诊断定制的完全遗忘架构.在沿用分区-聚合高效架构的同时,引入基于认知状态的分区方法与动态加权聚合机制,最大化保全甚至提升模型的认知诊断效能.

2 面向认知诊断的机器遗忘框架

2.1 问题概述 定义用户集合$U=\{u_i|1\leqslant i\leqslant N\}$,其中,$N$为用户数量;练习集合$E=\{e_j|1\leqslant j\leqslant J\}$,其中,$J$为练习数量;知识点集合$C=\{c_k|1\leqslant k\leqslant K\}$,其中,$K$为知识点数量.所有用户的认知状态由矩阵$\boldsymbol{A}=(a_{ik})_{N\times K}\in\mathbb{R}^{N\times K}$表示,其中,$a_{ik}\in[0,1]$,$\mathbb{R}$为实数集.矩阵$\boldsymbol{A}$的第$i$行记为$\boldsymbol{a}_i$,表示用户$u_i$的认知状态,元素$a_{ik}$表示用户$u_i$对知识点$c_k$的掌握程度.

本研究实验部分采用的认知诊断模型在执行认知诊断任务时,遵循原始研究中的假设.

假设1 基于单调性假设[31],学习者对某一项练习涉及的知识点掌握得越熟练,对应得分越高.

假设2 学习者的表现主要受练习涉及的知识点影响,不同知识点对练习的影响程度不同[32].练习与知识点之间的关系由矩阵$\boldsymbol{Q}=(q_{jk})_{J\times K}\in\mathbb{Z}^{J\times K}$表示,其中,$q_{jk}\in\{0,1\}$,表示练习$e_j$是否考查知识点$c_k$,$\mathbb{Z}$表示整数集.

通常,用户仅回答练习集合E中的部分练习.因此,定义用户u_i回答过的练习为E_i,对应的答题记录表示为$R_i=\{r_{ij}|0\leqslant r_{ij}\leqslant 1\}$,其中,$r_{ij}$表示用户$u_i$对练习$e_j$的答题情况.用户$u_i$回答过的练习$E_i$被划分为两部分,$E_i^X$和$E_i^Y$,满足$E_i^X\cup E_i^Y=E_i$和$E_i^X\cap E_i^Y=\varnothing$,其中,$E_i^X$用于认知诊断,$E_i^Y$用于得分预测.用户$u_i$在$E_i^X$上的答题记录记作$R_i^X$,在$E_i^Y$上的答题记录记作$R_i^Y$,满足$R_i^X\cup R_i^Y=R_i$.对于任意用户$u_i\in U$,认知诊断基于其回答的练习$E_i^X$及对应的答题记录$R_i^X$进行,从而确定用户$u_i$的认知状态,并最终预测其在练

习 E_i^Y 上的得分 $R_i^{Y'}$. 认知诊断模型的性能由预测的答题记录 $R_i^{Y'}$ 和实际答题记录 R_i^Y 的接近程度来衡量.

在认知诊断系统中，平台基于所有用户的答题记录 $R=\{R_i|1\leqslant i\leqslant N\}$ 训练认知诊断模型 $M(R)$. 随着隐私法规的推进，用户 $u_i\in U$ 有权向平台提交遗忘请求，指定将其部分或全部历史数据从模型记忆中擦除[33]. 设用户 u_i 请求遗忘的数据子集为 R_i^u（其中 $R_i^u\subseteq R_i$）. 本研究旨在提出一种通用的机器遗忘框架，该框架既兼容用户级别的遗忘（即 $R_i^u=R_i$，擦除特定用户的所有记录），也支持更细粒度的数据级别的遗忘（即 $R_i^u\subset R_i$，仅擦除该用户指定的部分作答记录）.

为了评估UFCD的有效性，参考推荐领域中的机器遗忘目标[10]，针对认知诊断模型制定了三个适用于遗忘任务的评估目标.

（1）模型性能保持. 采用适当的遗忘策略，在数据删除后对模型进行重新调整，以最小化性能的显著下降，避免“灾难性遗忘”[34]. 遗忘过程必须确保通过删除特定数据，模型能够保持强大的预测能力和良好的泛化性能.

（2）高效遗忘. 框架应实现高效的遗忘机制，使遗忘操作能够快速执行，减少计算成本，并确保即使在大规模数据集上，过程仍然保持有效.

（3）遗忘完整性. 针对用户的遗忘请求，模型必须确保彻底删除指定数据. 同时，保证无法恢复或重用已删除的数据，彻底消除其对模型参数的影响.

2.2　框架总体架构　UFCD框架遵循“分区-聚合”的设计范式，由三个核心模块组成：（1）认知特征构建模块，负责从异构教育数据中提取学习者的潜在认知状态；（2）子模型训练与遗忘模块，通过并行化训练与特定分区的重训练实现高效遗忘；（3）认知反馈与预测模块，用于输出诊断结果并更新认知状态. 完整工作流如图2所示，各模块的具体功能与交互逻辑详述如下.

（1）认知特征构建模块，主要负责数据的标准化处理与潜在特征提取. 首先，对原始的学习者答题日志及习题-知识点关联数据进行预处理，构

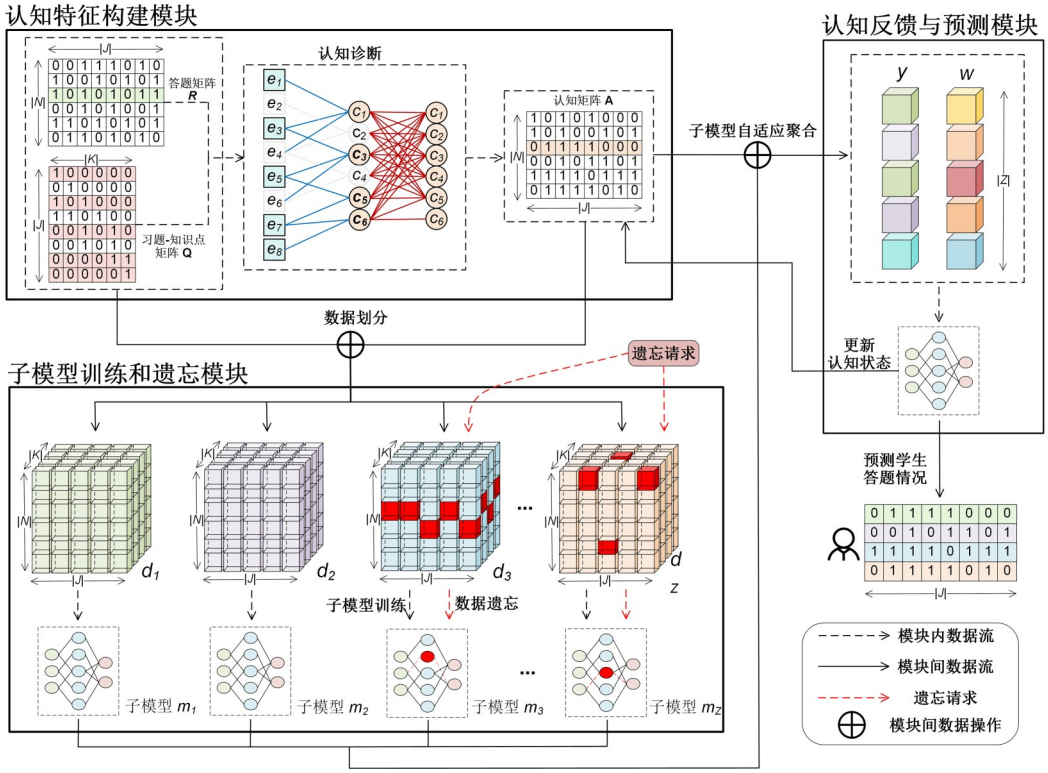


图2　面向认知诊断的机器遗忘框架

Fig. 2　Machine unlearning framework for cognitive diagnosis

建答题矩阵 $\boldsymbol{R}$ 和习题-知识点关联矩阵 $\boldsymbol{Q}$. 随后，利用预训练的认知诊断模型推断学习者的潜在认知状态矩阵 $\boldsymbol{A}$. 基于矩阵 $\boldsymbol{A}$ 中的认知特征向量，引入基于认知状态的分区方法(CSP)，将原始数据集智能地划分为若干具有特定认知特性的独立子集.

(2)子模型训练与遗忘模块. 作为遗忘框架的核心，该模块采用物理隔离策略来平衡遗忘效率与模型性能. 针对一个经 CSP 划分的数据分区，系统独立训练一个认知诊断子模型. 在模型推断阶段，为了解决静态聚合带来的精度损失问题，引入交叉注意力驱动的模型聚合机制(CMAM)，综合考量当前待测样本与各分区认知状态的认知相似度以及各子模型的性能特征、动态计算注意力权重，实现对多子模型的自适应聚合. 这种机制不仅提升整体预测精度，更确保遗忘操作仅局限于特定子模型，无需全局重训练.

(3)认知反馈与预测模块. 聚合后的全局模型对指定学习者的答题表现进行概率预测，并根据预测偏差反向更新学习者的认知状态估计，获得用户最新的认知状态，更精准地预测用户的答题情况.

基于 UFCD 框架详细介绍模型训练流程以及接收到遗忘请求后执行的遗忘流程.

训练过程如图 2 中黑色箭头所示，UFCD 的训练流程主要包含五个步骤.

(1)初始认知状态提取. 采用认知诊断算法对用户集合 U 的答题记录 R 进行初步诊断，获取全局认知状态 $\boldsymbol{A}=\{\boldsymbol{a}_i|1\leqslant i\leqslant N\}$，其中，$\boldsymbol{a}_i=\{a_{ik}|1\leqslant k\leqslant K\}$ 为用户 u_i 在 K 个知识点的认知状态，称为认知向量.

(2)相似度度量. 在基于认知状态的分区方法(CSP)中，利用余弦相似度计算不同用户认知向量 $\boldsymbol{a}_i$ 之间的差异，量化用户间的认知状态相似度.

(3)用户级数据划分. 根据认知相似度对原始数据集进行用户层面的聚类划分，构建若干个代表不同认知模式特征的分区数据集.

(4)子模型独立训练. 针对每个划分后的数据分区，独立训练相应的认知诊断算法，获得专注于特定认知模式的认知诊断子模型.

(5)自适应聚合. 采用交叉注意力驱动的模型聚合方法(CMAM)，计算各子模型对当前预测样本的贡献权重，并通过注意力加权聚合机制生成最终的预测结果.

遗忘过程如图 2 中红色箭头所示，UFCD 的遗忘机制旨在实现高效且彻底的数据移除. 具体包含以下步骤.

(1)请求响应与定位. 当系统接收到用户 $u_i\in U$ 的遗忘请求时，利用请求中的唯一标识符(Record ID)，在框架内精确定位该用户数据所属的特定数据分区.

(2)数据与模型清除. 执行物理删除操作，从目标分区中移除指定用户的答题记录，并同步废弃该分区对应的旧认知诊断子模型，以消除目标数据对模型参数的潜在影响.

(3)子模型重训练. 针对更新后的子数据集(即剔除目标数据后的分区)，采用与初始训练一致的认知诊断算法进行重训练. 这一过程确保新生成的子模型既能适应数据分布的变化，又保持了模型架构的统一性.

(4)动态聚合更新. 子模型重训练完成后，基于其在验证集上的性能表现重新评估并设定初始权重，最后将其重新整合至 CMAM 聚合模块中，完成整体模型的更新.

(5)认知反馈更新. 基于执行遗忘操作后的聚合模型，对用户进行二次诊断，实时修正其潜在认知状态矩阵 $\boldsymbol{A}$，并依据更新后的认知状态调整该用户的数据分布.

值得强调的是，得益于本框架特有的基于认知状态的分区方法(CSP)，特定用户的数据被高度聚集于单一或极少数的特定数据分区中. 因此，在面对遗忘请求时，该操作的影响半径被严格收束，系统通常仅需重训练少数受波及的子模型，无需像传统随机划分(SISA)那样面临大量甚至全部分区被迫重训的风险. 这种基于认知语义的物理隔离，从根本上保证了本框架在处理单一乃至批量遗忘请求时的高效性.

2.3 基于认知状态的分区方法 在现有的完全遗忘框架中，SISA 模型[9]通常采用平均随机划分

的方法来对数据集进行分割．尽管该方法能保证各分片大小绝对均衡，但其在认知诊断场景下面临两个致命缺陷：第一，随机划分割裂了教育数据的结构化关联与学习者群体的潜在认知特征；第二，随机分配使单个用户的历史作答记录散落于多个数据分区中，一旦发生数据遗忘请求，极易触发所有子模型的全局重训练，导致遗忘效率的严重退化．

针对上述问题，完成既能保留认知语义又能收束遗忘影响半径的有效数据划分成为关键．为此，本研究提出一种基于认知状态的分区方法（CSP），将具有相似认知模式的用户数据高度聚类到同一分区，这不仅使子模型能够更专注地拟合特定群体的认知规律，更重要的是，通过物理层面的“语义隔离”，将特定用户的遗忘请求精确限制在单一分区内．即使在多用户并发遗忘的极端场景下，只要这些用户属于同一认知聚类，系统也只需付出单次子模型重训练的代价，在根本上保障了机器遗忘的高效性．

给定用户集 $U=\{u_i|1\leqslant i\leqslant N\}$ 和答题记录集 $R=\{R_i|1\leqslant i\leqslant N\}$，选取适合的认知诊断算法进行训练，获得用户集合 U 对应的认知状态 $\boldsymbol{A}=\{\boldsymbol{a}_i|1\leqslant i\leqslant N\}$．基于认知状态的划分方法，根据用户认知状态的相似度对数据进行分割，如以下算法所示．随机初始化 Z 个中心用户，并构建其对应的用户簇 $L=\{l_1,l_2,\cdots,l_Z\}$，其中，每个簇初始时仅包含一位中心用户．对于剩余用户，将其分配到与其认知状态相似度高的用户簇中．分配完成后，根据最终的用户簇，对答题记录进行划分．划分结束，对数据分区大小进行平衡调整．

算法　基于认知状态的分区方法

输入：答题记录集 $R=\{R_i|1\leqslant i\leqslant N\}$，用户认知状态 $\boldsymbol{A}=\{\boldsymbol{a}_i|1\leqslant i\leqslant N\}$，分区最大阈值 T

输出：数据分区 $D=\{d_i|1\leqslant i\leqslant Z\}$

1. 从 U 中随机抽取 Z 个用户作为各数据分区的中心用户

2. 根据 Z 个中心用户，初始化 Z 个用户簇 $L=\{l_1,l_2,\cdots,l_Z\}$

3. 计算每个用户 $u_i\in U$ 与中心用户 $L_j\in Z$ 之间的认知相似度，得到相似度矩阵 $\boldsymbol{S}$

4. 遍历用户集 U

5.　根据相似度矩阵 $\boldsymbol{S}$，将用户 u_i 分配给与其相似度最高的中心用户所在的用户簇 l_j

6.　更新用户聚合 U

7. 遍历用户簇 L

8.　根据 l_j 中用户的分配情况，将答题记录分配到对应的数据分区 d_i

9. 遍历数据分区 D

10.　若 d_i 的大小超过阈值 T

11.　根据相似度矩阵 $\boldsymbol{S}$，将 d_i 中相似度较低的用户答题记录移入 R

12.　若 d_i 的大小小于阈值 T

13.　根据相似度矩阵 $\boldsymbol{S}$，将 R 中相似度较高的用户答题记录移入 d_i

14. 若 R 中仍剩余元素未分配

15.　优先分配到最小的数据分区

16. 返回数据分区 $D=\{d_i|1\leqslant i\leqslant Z\}$

为了计算用户之间的认知相似度，选取余弦相似度作为度量标准，如式（1）所示．余弦相似度能有效衡量用户知识熟练度的相似性，尤其适用于特征向量稀疏的情况．该方法侧重于用户认知状态的相对模式，而非绝对水平．这一特性使其在面对因数据稀疏导致的特征提取偏差时具有较强的容错性，在一定程度上缓解了基础模型估计不准对分区质量的级联影响，在处理多维稀疏认知数据时能提供更具代表性的相似性度量．

$$\mathrm{sim}(u_i,l_j)=\frac{\boldsymbol{a}_i\cdot\boldsymbol{a}_j}{//a_i//\cdot//a_j//}=\frac{\sum_{k=1}^{K}a_{ik}\cdot a_{jk}}{\sqrt{\sum_{k=1}^{K}(a_{ik})^2\cdot\sum_{k=1}^{K}(a_{jk})^2}}\tag{1}$$

数据分区最大阈值 T 的计算，如式（2）所示：

$$T=\mu\frac{\sum_{i=1}^{N}|R_i|}{Z}\tag{2}$$

其中，$\sum_{i=1}^{N}|R_i|$ 表示所有用户的历史答题记录总数，Z 为数据分区数量，μ 为松弛系数（默认为1.2），用于在保证分区规模相对均衡的前提下，为认知语义聚类提供一定的弹性空间．

在CSP算法中，数据分区最大阈值 T 以及中心用户簇的数量 Z 直接决定数据分片的粒度．理论上，如果将数据集划分得足够小（即 Z 极大，T 极小），每个子模型包含的数据量骤减，重训练该子模型所需的时间开销将趋近极小值，实现极高

的“遗忘效率”. 然而,这种极端的划分会导致严重的“效用崩塌”. 认知诊断模型高度依赖于群体数据间的认知交互信息,若数据分片过小,子模型将陷入严重的数据稀疏困境,无法提取具有泛化能力的认知模式特征,导致最终聚合模型的“诊断性能”大幅衰退. 若分片过大,虽然能保证充足的协同信息以维持高诊断性能,但又会导致单次遗忘请求的重训代价过高. 因此,在本框架中,遗忘性能与遗忘效率之间存在本质的权衡. CSP 正是通过基于认知状态的聚类以及阈值 T 的动态调节,在保障同一分区内认知特征高内聚(维持模型效用)的同时,控制分片规模的上限(保障遗忘效率),实现二者的平衡.

2.4 基于交叉注意力的模型聚合机制 在数据划分阶段,基于用户认知状态的相似度,将答题记录划分为 Z 个数据分区 $D=\{d_i|1\leqslant i\leqslant Z\}$. 每个数据分区自主选择认知诊断算法,并行训练,最终生成 Z 个认知诊断子模型 $M=\{m_1,m_2,\cdots,m_Z\}$. 由于每个数据分区代表具有不同认知状态的用户群体,各子模型在相同的预测任务中会产生不同的贡献. 基于此,本文提出一种基于交叉注意力的认知诊断模型聚合方法,根据子模型性能、数据分区与待预测用户之间的认知相似度,动态调整各个子模型的权重,提升最终模型的预测性能. 与传统的简单加权或平均聚合方法相比,交叉注意力机制能够更灵活地捕捉模型间的异质性,从而提升整体预测的准确性.

子模型注意力权重计算. 首先,如果认知诊断子模型 m_i 性能优于 m_j, 则在预测任务中,m_i 应当被赋予比 m_j 更高的权重;其次,当数据分区 d_i 与预测用户 u 之间的认知相似度较高时,意味着 d_i 能够更准确地代表 u 的认知状态,使 m_i 更适用于该预测任务,在最终预测中应赋予更高的权重. 子模型注意力权重的计算,如式(3)所示.

$$w_i=\frac{acc_i\cdot\mathrm{sim}(u,d_i)}{\sum_{j=1}^{Z}acc_j\cdot\mathrm{sim}(u,d_j)}=\frac{acc_i\cdot\mathrm{sim}(u,l_i)}{\sum_{j=1}^{Z}acc_j\cdot\mathrm{sim}(u,l_j)} \tag{3}$$

其中,子模型的准确率 acc_i 反映了子模型 m_i 的性能. 具体地,考虑到遗忘框架的实际复杂性,在预测用户 u 的分数之前,针对子模型注意力权重的计算分两种情况.

(1)用户 u 存在历史答题记录时,可通过相似度矩阵 $\boldsymbol{S}$ 计算用户 u 与数据分区 d_i 之间的认知相似度 $\mathrm{sim}(u,d_i)$. 对于包含用户 u 历史数据的分区,其对应子模型的权重应在实际情况下适当增加. 此方法充分利用了用户的历史记录,降低了计算资源消耗.

(2)用户 u 不存在历史答题记录(即极端稀疏冷启动场景)时,其初始认知状态的估计往往存在较大偏差. 为了防止这种偏差在聚合阶段传播,将用户 u 与各数据分区 d_i 之间的认知相似度 $\mathrm{sim}(u,d_i)$强制置为 1. 这本质上是一种自适应回退机制,完全依据子模型的客观性能赋予权重,如式(4)所示,有效阻断了前端提取误差对最终预测精度的破坏:

$$w_i=\frac{acc_i\cdot\mathrm{sim}(u,d_i)}{\sum_{j=1}^{Z}acc_j\cdot\mathrm{sim}(u,d_j)}=\frac{acc_i}{\sum_{j=1}^{Z}acc_j} \tag{4}$$

认知诊断子模型的聚合. 在预测过程中,每个认知诊断子模型独立预测用户 u 的得分. 最终,用户 u 的预测得分通过加权聚合各子模型的预测结果生成,其中权重由计算得到的子模型注意力权重决定,如式(5)所示.

$$Y=\sum_{i=1}^{Z}w_i y_i \tag{5}$$

遗忘场景下的模型精度保障机制. 当执行数据遗忘操作时,CMAM 通过“局部认知冗余”与“全局动态补偿”双重机制,有效克服了数据移除对模型预测精度的负面影响.

在局部子模型层面,得益于 CSP 策略形成的高内聚特性,待遗忘数据与同一分区内的剩余数据共享极其相似的认知模式,这种高度相似性实质上为子模型提供了强大的“特征学习冗余度”. 因此,擦除单个用户的交互日志,仅消除了该特定的隐私轨迹,而该群体代表的宏观“认知规律”依然被剩余的相似用户数据完好支撑,避免了局部子模型发生灾难性的认知崩塌. 在全局聚合层面,当执行遗忘并重训受影响的子模型后,CMAM 模块会重新评估其诊断能力,自动微调该子模型的注意力权重,并自适应地提高其他具有一定认知相似度且性能更优的子模型权重. 这种交叉注意力驱动的聚合机制,使全局模型能够在极小的性能扰动下迅速恢复平衡,从根本上保证

遗忘操作前后的模型精度与鲁棒性.

在实际的在线智能教育系统中，数据往往呈高度动态演进的特征，如新用户的持续涌入以及教学内容迭代引发的概念漂移. 由于CSP模块的全局重分区计算代价较高，UFCD在应对动态流式数据时采取“软路由与双频维护”的扩展策略. 一方面，针对新用户，CMAM天然具备动态软路由能力. 通过式(3)与式(4)的自适应权重分配，新用户可在不触发底层重分区的前提下，实时匹配至最优子模型以获取诊断服务. 另一方面，针对概念漂移，系统可周期性地结合在线学习与近似遗忘算法，对局部子模型进行低成本的参数增量更新，仅当监测到全局认知分布发生显著偏移或微调带来的精度损耗达到预设阈值时才触发新一轮的CSP重聚类. 这种微观局部微调与宏观全局重构相结合的范式能在维持长周期诊断精度的同时，有效规避频繁重分区带来的高昂代价.

3　实验分析

3.1　实验设置　为了最大限度地降低随机性带来的实验误差并保证结果的可靠性，所有实验均独立重复运行10次，最终报告结果取其平均值. 实验环境，所有模型训练与测试均部署于搭载双路 Intel(R) Xeon(R) Silver 4208处理器、128 GB内存及NVIDIA GeForce RTX 3090 GPU的服务器上. 在模型训练阶段，统一参考各基线模型原始文献的推荐区间对超参数进行调优. 在统一的最终设定下，模型训练的批次大小设置为32，训练轮数设定为8，学习率为0.02.

3.1.1　数据集　选取认知诊断与教育数据挖掘领域广泛使用的四个公开基准数据集.

ASSIST0910是由在线辅导系统ASSISTments收集的公开数据集，涵盖了2009－2010学年学生的答题记录，为研究人员提供学生在练习中的表现以及相关知识概念的隐含关系. 尽管ASSIST0910提供了宝贵的学习行为数据，但该数据集缺乏知识点之间的交互信息，限制了对知识概念关联的深入分析.

ASSIST2012来源于ASSISTments在线学习平台，在教育数据挖掘和学习分析领域被广泛应用，包含学生对数学题目的详细交互记录，包括答题正确性、尝试次数、时间戳及相关知识点等信息. 作为基准数据集，ASSIST2012支持对学生动态知识状态的评估，常用于开发和验证数据驱动的智能教育系统和预测模型.

ASSIST2017来源于2017 ASSISTments Data Mining Competition，包含2004－2007年的学生答题记录以及练习与知识概念之间的关联关系，为评估学生的学习表现和知识掌握情况提供重要数据支持. 然而，与ASSIST0910类似，它缺少知识点之间的交互数据，这在研究知识概念的复杂关联时带来了一定的挑战.

KDDCUP由PSLC DataShop在KDDCUP 2010竞赛中发布，记录了每道题的多个解题步骤以及学生在各个步骤上的得分. 这种数据结构能够将同一道题目不同步骤的得分视为独立的答题记录，从而提供丰富的数据来分析学生对知识点的逐步掌握情况.

对于缺少知识点交互信息的数据集，采用Gao et al[35]的方法来构建知识点之间的交互关系. 为了确保每位学习者拥有足够的答题记录以进行认知诊断，移除答题记录少于15条的学习者. 数据集预处理之后的统计信息汇总如表1所示.

3.1.2　对比方法　为了全面评估UFCD的性能，选取两类基线方法，一是作为主干模型的经典认知诊断模型，二是用于对比的机器遗忘方法. 通过在不同的主干模型上应用各类遗忘策略，对

表1　数据集汇总

Table 1　Datasets summary

数据集	ASSIST-0910	ASSIST-2012	ASSIST-2017	KDDCUP
学习者	2380	20527	1678	425
练习	16804	52695	2210	101744
答题记录	257585	2654864	351530	163410
知识点	110	265	101	16
知识点交互	0	0	0	47
平均每道练习考查的知识点数量	1.2	1	1.27	1.35
平均每个学生回答的练习数量	108.23	131.06	209.49	384.49
平均每个学生考查的知识点数量	14.87	17.62	52.56	9.75

比分析遗忘操作前后主干模型的各项预测与诊断指标,验证本文所提方法的有效性与优越性.

(1)认知诊断模型.

MIRT[36]是在经典的IRT模型基础上的改进型多维模型,旨在刻画学习者在多个维度上的认知状态.MIRT使用多维向量来表征学习者的认知状态和练习题的因素维度,能够捕捉学习者在多个认知维度上的表现,使MIRT更加适用于复杂的认知诊断场景,深入分析学习者在不同知识维度上的掌握情况.

DINA[37]是认知诊断领域的基础模型之一,采用二元编码表示学习者的知识状态,其中,1表示掌握特定知识点,0表示未掌握.还引入猜测和失误机制,分别用于建模学习者在没有掌握知识点时可能猜对的情况以及在掌握知识点时仍然出错的情况,更精准地反映学习者的答题行为.通过组合多个二元认知状态,DINA能够实现细粒度的认知诊断,提供学习者知识掌握情况的详细分析,因此在个性化学习推荐、教育评估等领域具有重要应用价值.

NCDM[38]是最早将神经网络引入认知诊断的模型之一,通过多层神经网络来建模学习者的认知状态和练习属性,并引入了单调性假设来确保模型具有良好的解释性.NCDM的核心优势在于能够捕捉复杂的非线性关系,同时保证诊断结果的可解释性,使其适用于大规模认知诊断任务.

(2)机器遗忘学习模型.

鉴于当前面向认知诊断模型的机器遗忘研究尚处于探索初期,鲜有可直接迁移的遗忘算法作为现成基准.为了评估UFCD的性能并确保对比实验的公平性,从近几年的研究工作中选取若干具有代表性的通用机器遗忘框架作为基线方法,与UFCD均属于完全遗忘的研究范畴.

Retain是最简单的机器遗忘方法,其核心思想是从数据集中删除需要遗忘的数据,然后在更新后的数据集上重新训练模型.尽管该方法直观且易于实现,但由于模型需要完全重训练,其计算成本较高,因此在大规模数据集上不适用.

SISA[9]是机器遗忘中最具代表性的方法之一,随机将训练数据划分为多个分片,使得可以在每个分片上训练子模型,最终的预测结果通过所有子模型的平均值或多数投票进行聚合.SISA的核心在于通过随机化和聚合技术,增强模型的鲁棒性和泛化能力.

RecEraser[10]是专门为推荐任务设计的通用且高效的机器遗忘框架,将训练集划分为多个碎片,并利用这些碎片训练子模型.为了保持数据之间的协同信息,RecEraser采用一种适用于推荐数据集中的交互信息的数据划分算法,确保训练集能够均匀地划分为碎片,有效地捕捉用户与项目之间的潜在关系.通过引入自适应聚合方法,以高效地整合子模型的输出,提升推荐结果的整体性能.

3.1.3 评估指标 数据集涵盖的习题均为客观题,即作答得分被二值化为0(错误)或1(正确).为了全面、客观地量化各认知诊断模型及遗忘方法的性能差异,选取准确率(Accuracy,ACC)、曲线下面积(Area Under Curve,AUC)、均方根误差(Root Mean Square Error,$RMSE$)以及平均绝对误差(Mean Absolute Error,MAE)四个主流评估指标.各指标的具体定义如下.

ACC直接反映模型整体预测的正确性,即模型对学生答题结果的预测与真实观测值相一致的样本占比,较高的ACC表明模型具备良好的全局预测性能.如式(6)所示:

$$ACC=\frac{1}{\sum_{i=1}^{N}\left|E_i^Y\right|}\sum_{i=1}^{N}\sum_{e_j\in E_i^Y}f\left(y_{ij},\hat{y}_{ij}\right) \tag{6}$$

AUC重点评估模型对正负样本(即答对与答错)的排序与区分能力.AUC的取值范围在0.5~1,越接近1表明模型的区分度越强.该指标在面对样本类别分布不均衡(如海量数据中正确作答远多于错误作答)时,仍能保持极高的客观性与稳定性.

$RMSE$是衡量模型预测得分与学生实际得分之间平方差的均方根.由于计算过程中包含平方项,$RMSE$会对具有较大偏差的异常预测值更敏感,较低的$RMSE$意味着预测分布更贴近真实分布,模型的拟合精度更高.如式(7)所示:

$$RMSE=\sqrt{\frac{1}{\sum_{i=1}^{N}\left|E_i^Y\right|}\sum_{i=1}^{N}\sum_{e_j\in E_i^Y}\left(y_{ij}-\hat{y}_{ij}\right)^2} \tag{7}$$

MAE表示预测值与真实值之间绝对差的平均期望. 相比于$RMSE$，MAE不会过度放大局部大误差的影响，因而提供了平均预测偏差的最直观度量，较低的MAE表明模型整体预测偏差较小. 如式(8)所示：

$$MAE=\frac{\sqrt{\sum_{i=1}^{N}\sum_{e_j\in E_i^v}\left(y_{ij}-\hat{y}_{ij}\right)^2}}{N} \tag{8}$$

3.2　模型间性能对比　验证UFCD框架是否满足遗忘目标一，即“模型性能保持”. 选取MIRT，DINA和NCDM三种经典的认知诊断模型作为主干网络，并将UFCD与两种主流的机器遗忘算法(SISA，RecEraser)以及“重训练(Retrain)”策略进行系统性的性能评估与对比. 鉴于Retrain策略通过物理剔除目标数据并从零开始训练，理论上能够彻底消除数据残留，现有的机器遗忘研究通常将其视为遗忘后模型性能的理想基准[39]. 为了确保实验对比的公平性，所有基于分区-聚合范式的遗忘方法(包含UFCD，SISA以及RecEraser)均将数据集的分片数量统一设定为$S=3$.

实验场景模拟了真实教育系统中的数据遗忘需求，将随机采样数据集中任意用户3%的学习行为记录作为待遗忘数据. 随后，将上述不同的遗忘策略应用于各主干认知诊断模型，并以遗忘操作完成后的模型预测表现作为核心评估依据. 具体实验结果详见表2～4，表中黑体字表示除Retrain外的遗忘算法中的最优表现. 本文以ACC作为核心评估指标.

(1)与重训练基准的对比.

尽管Retrain在理论上代表了模型性能的上限，但UFCD在绝大多数实验设定下均展现出逼近该基准的预测表现. 值得注意的是，在部分特定场景下(如ASSIST2012和KDDCUP数据集上的MIRT以及ASSIST2012和ASSIST2017数据集上的DINA模型)，UFCD的诊断性能甚至实现了对Retrain基线的微弱超越. 这种“超基准”现象从侧面印证了CMAM在模型聚合阶段的有效性，该机制不仅避免了知识遗忘，更在一定程度上发挥了集成学习的优势. 通过动态评估学习者特征与不同子模型之间的匹配度，实现了诊断精度的自适应提升.

表2　MIRT诊断模型遗忘后性能对比

Table 2　Performance of the MIRT diagnostic model after unlearning

数据集	评估指标	Retrain	SISA	RecEraser	UFCD
ASSIST-0910	*ACC*	0.5878	0.5655	0.5852	**0.5853**
	AUC	0.5838	0.5411	0.5615	**0.5684**
	RMSE	0.5942	0.6190	0.5190	**0.5195**
	MAE	0.4161	0.4374	0.4434	**0.4433**
ASSIST-2012	*ACC*	0.5951	0.5714	0.6056	**0.6194**
	AUC	0.5751	0.5443	0.5729	**0.5833**
	RMSE	0.6081	0.6296	0.5137	**0.5070**
	MAE	0.4064	0.4297	0.4263	**0.4143**
ASSIST-2017	*ACC*	0.5946	0.5445	0.5718	**0.5851**
	AUC	0.6237	0.5457	0.5898	**0.6109**
	RMSE	0.5775	0.6317	0.5228	**0.5163**
	MAE	0.4117	0.4580	0.4554	**0.4444**
KDDCUP	*ACC*	0.7059	0.6651	0.7130	**0.7280**
	AUC	0.6154	0.5687	0.5977	**0.6153**
	RMSE	0.4633	0.4986	0.4429	**0.4342**
	MAE	0.3312	0.3633	0.3654	**0.3517**

表3　DINA诊断模型遗忘后性能对比

Table 3　Performance of the DINA diagnostic model after unlearning

数据集	评估指标	Retrain	SISA	RecEraser	UFCD
ASSIST-0910	*ACC*	0.6886	0.5802	0.6591	**0.6706**
	AUC	0.7061	0.6052	0.6575	**0.6674**
	RMSE	0.4715	0.5284	0.4672	**0.4640**
	MAE	0.3325	0.4077	0.4129	**0.4083**
ASSIST-2012	*ACC*	0.6966	0.6064	0.7112	**0.7204**
	AUC	0.7001	0.6246	0.6635	**0.6772**
	RMSE	0.4680	0.5087	0.4500	**0.4460**
	MAE	0.3056	0.3711	0.3696	**0.3564**
ASSIST-2017	*ACC*	0.6188	0.5650	0.6394	**0.6472**
	AUC	0.7005	0.6309	0.6724	**0.6793**
	RMSE	0.5076	0.5428	0.4824	**0.4807**
	MAE	0.3752	0.4177	0.4166	**0.4108**
KDDCUP	*ACC*	0.7904	0.7290	0.7482	**0.7668**
	AUC	0.6619	0.6294	0.6476	**0.6645**
	RMSE	0.4072	0.4487	0.4186	**0.4102**
	MAE	0.2881	0.3352	0.3431	**0.3328**

表 4 NCDM 诊断模型遗忘后性能对比

Table 4 Performance of the NCDM diagnostic model after unlearning

数据集	评估指标	Retrain	SISA	RecEraser	UFCD
ASSIST-0910	*ACC*	0.7273	0.6607	0.6449	**0.6681**
	AUC	0.7614	0.6834	0.5869	**0.6204**
	RMSE	0.4322	0.4799	0.4779	**0.4605**
	MAE	0.3340	0.3630	0.4231	**0.4125**
ASSIST-2012	*ACC*	0.7346	0.6837	0.6703	**0.6866**
	AUC	0.7161	0.6783	0.6012	**0.6097**
	RMSE	0.4314	0.4597	0.4614	**0.4565**
	MAE	0.3444	0.3826	0.4232	**0.4055**
ASSIST-2017	*ACC*	0.6869	**0.6345**	0.6002	0.6105
	AUC	0.7452	**0.6929**	0.6256	0.6196
	RMSE	0.4639	**0.4962**	0.4894	0.4985
	MAE	0.3600	**0.4003**	0.4537	0.4502
KDDCUP	*ACC*	0.7238	0.6717	0.7138	**0.7228**
	AUC	0.7357	0.7029	0.6005	**0.6019**
	RMSE	0.4305	0.4478	0.4435	**0.4360**
	MAE	0.3132	0.3375	0.3892	**0.3813**

(2)与现有遗忘方法的性能对比.

整体上,UFCD 在四个真实教育数据集上的表现优于 SISA 与 RecEraser. 其卓越的性能表现主要源于框架内两大核心机制的协同优化:其一,CSP 机制能够依据学习者的认知状态进行同质性划分,显著降低子模型拟合的数据复杂度并提升局部诊断精度;其二,CMAM 在聚合阶段充分融合子模型的性能先验与局部认知分布信息,实现全局预测结果的最优整合. 各主干模型下的具体表现分析如下.

MIRT 模型. 在 ASSIST2012,ASSIST2017 与 KDDCUP 数据集上,UFCD 的预测性能显著优于 SISA 和 RecEraser,且高度逼近 Retrain 基准. 具体地,和 SISA 和 RecEraser 相比,UFCD 在 ASSIST2012 数据集上的 *ACC* 分别提升 8.40% 和 2.28%;在 ASSIST2017 数据集上分别提升 7.46% 和 2.36%;在 KDDCUP 数据集上分别实现 9.46% 和 2.10% 的显著增益.

DINA 模型. UFCD 在此模型上展现出极强的泛化能力与鲁棒性,在所有测试数据集上均全面领先于对比遗忘方法. 与 SISA 和 RecEraser 相比,其 *ACC* 在 ASSIST0910 数据集上分别跃升 15.58% 和 1.72%;在 ASSIST2012 数据集上分别提升 18.80% 和 1.29%;在 ASSIST2017 数据集上分别提升 14.55% 和 1.22%;在 KDDCUP 数据集上分别获得 5.19% 和 2.49% 的性能提升.

NCDM 模型. 在 ASSIST0910, ASSIST-2012 和 KDDCUP 数据集上,UFCD 同样维持了优异的诊断水准,表现显著优于 SISA 和 RecEraser. 具体地,对比 SISA 和 RecEraser,UFCD 在 ASSIST0910 数据集上的 *ACC* 分别提升 1.12% 和 1.60%;在 ASSIST2012 数据集上分别提升 0.42% 和 2.43%;在 KDDCUP 数据集上分别提升 7.60% 和 1.26%.

3.3 遗忘分析

3.3.1 遗忘效率分析 为了全面评估 UFCD 框架在实际教育场景下的计算优势,以验证其是否满足设定的遗忘目标二“高效遗忘”,从理论计算复杂度与时间开销两个维度对模型的遗忘效率进行系统性分析.

(1)时间复杂度分析. 本框架的遗忘可行性建立在严谨的物理隔离基础上. 假设原始数据集的规模为 N,模型训练所需的迭代轮数为 E. 若采用重训练(Retrain)方法,即直接在全局数据集上定位、移除数据并从头训练整个模型,其计算复杂度高达 $O(N\times E)$. 在 UFCD 框架下,当系统接收到特定用户的遗忘请求时,定位该用户数据所在分区的时间复杂度仅为 $O(1)$. 更关键的是,系统仅需对包含该目标数据的单一子模型进行重训练. 若数据被均衡划分为 Z 个分区,受波及子模型的重训复杂度将大幅骤降至 $O\left(\frac{N}{Z}\times E\right)$. 理论上,本方法可将遗忘的时间开销降低近 Z 倍.

(2)时间开销分析. 将重训练(Retrain)的耗时作为性能下限基线,系统测试在不同数据分区粒度(设定分区数量 $S=\{3,5,10,20\}$)下的遗忘效率. MIRT,DINA 及 NCDM 三种主干模型在各数据集上的具体时间开销对比详见表 5.

3.3.2 遗忘完整性分析 针对遗忘目标三“遗忘完整性”,UFCD 框架遵循严格的分区聚合与完全遗忘范式. 从差分隐私的理论界限来看,完全遗

表5　不同分区数的遗忘时间开销

Table 5　Time consumption for unlearning with different numbers of partitions

	ASSIST0910			ASSIST2012			ASSIST2017			KDDCUP		
	MIRT	DINA	NCDM	MIRT	DINA	NCDM	MIRT	DINA	NCDM	MIRT	DINA	NCDM
Retrain	325.51	288.33	529.91	4538.51	2879.51	4898.91	438.24	462.46	733.62	222.02	207.43	270.47
UFCD (3)	143.67	101.40	300.60	1555.43	921.84	2667.75	222.27	143.72	346.63	79.32	67.12	166.02
UFCD (5)	80.01	74.61	198.89	934.20	740.89	920.51	93.30	88.80	311.27	46.96	43.32	158.43
UFCD (10)	40.06	39.06	151.93	470.96	349.90	502.65	59.98	52.65	197.40	31.96	29.66	51.68
UFCD (20)	25.11	33.35	139.33	241.54	193.37	458.75	30.26	52.07	185.69	13.83	15.13	42.14

忘通过在受影响分区中进行物理级的数据剔除，并从头启动子模型的重训练，确保遗忘后的模型参数分布与从未见过目标数据而直接训练的模型参数分布完全一致．这一过程从根本上阻断了目标数据特征向模型权重的传递路径，在数据存储与模型参数双重层面上保证了遗忘信息的彻底清除，以“完全遗忘”的标准严格达成了遗忘完整性目标[6]．因此，该机制在理论上达到了$\epsilon=0$的差分隐私界限，无需依赖经验性的成员推理攻击即可提供最高级别的遗忘完整性保证．

3.4　参数敏感度分析　在UFCD框架的CSP算法中，簇的数量（即分区数S）与单个分区的最大数据容量（分区大小T）是决定数据划分粒度的核心参数．根据式(2)，在给定数据集与固定松弛系数下，分区数量与分区大小呈现深度的反比耦合关系．因此，探究分区数的变化，本质上即是验证模型对“簇数量及局部数据稀疏度（分区大小）”的敏感性．在四个真实的教育数据集上展开详尽的实验，系统性地探究并对比了多级分区设定（$S=\{3,5,10,20\}$）下，不同机器遗忘算法的性能表现．以Retrain作为性能基准线（以虚线表示），具体实验结果如图3～5所示．

(1)整体趋势分析．实验结果呈现一个共性规律：随着分区数S的递增，所有参与测试的遗忘方法在各数据集上的准确率均表现出不同程度的衰退，总体水平略低于Retrain基准．这一现象的

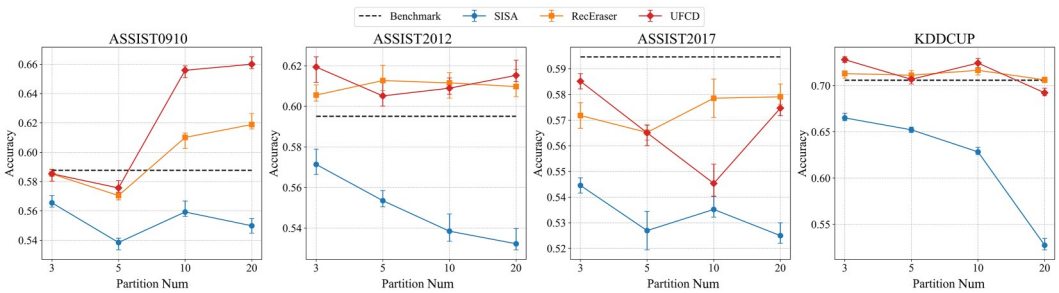


图3　MIRT在不同数据分区数量上的表现

Fig. 3　Performance of MIRT on different numbers of data partitions

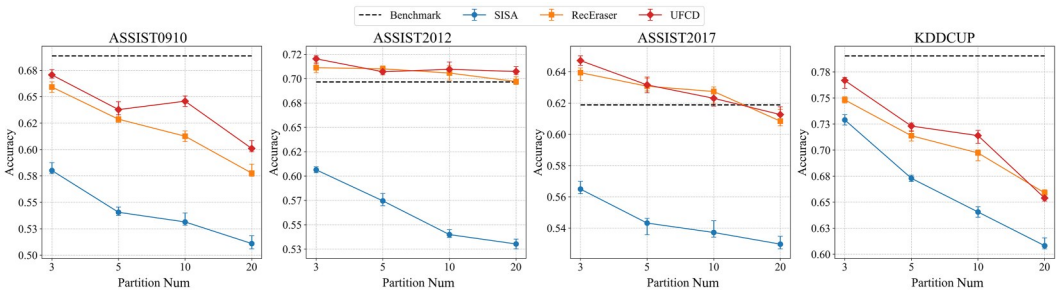


图4　DINA在不同数据分区数量上的表现

Fig. 4　Performance of DINA on different numbers of data partitions

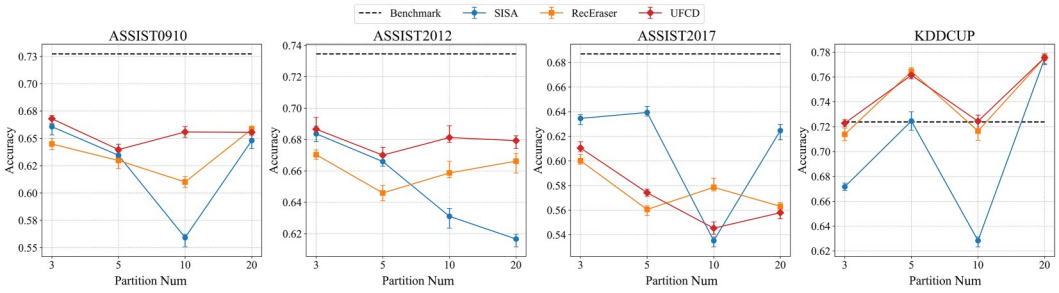


图 5 NCDM 在不同数据分区数量上的表现

Fig. 5 Performance of NCDM on different numbers of data partitions

根本原因在于分区聚合架构引发的“数据稀疏性”问题. 认知诊断模型高度依赖充足的学生交互记录来实现对潜在认知参数的精准估计,过高的分区数会导致单一数据分片内的样本量急剧衰减,严重制约了局部子模型的参数拟合能力.

在高稀疏环境中,UFCD 展现出极强的鲁棒性. 在局部数据大量缺失的冲击下,由于 UFCD 引入了高内聚的认知分区策略(CSP),有效保留了分区内部丰富的局部“认知冗余”. 同时,交叉注意力机制(CMAM)能够对异质化子模型进行动态补偿,使该框架在大幅降低遗忘响应计算开销的同时,有效避免了诊断性能的断崖式下跌,实现了遗忘效率与模型鲁棒性之间的权衡.

(2)各主干模型下的表现差异.

MIRT 模型:UFCD 的整体预测性能显著优于其他基线方法. 特别是在 ASSIST0910, ASSIST2012 和 KDDCUP 数据集上,展现出优异的鲁棒性,部分实验组别的准确率甚至反超重训练基准,印证了其卓越的泛化能力.

DINA 模型:尽管所有评估方法的诊断性能均不可避免地随分区数的增加而下滑,但相较于其他遗忘框架,UFCD 展现出最强的性能衰退缓解能力,在全部四个数据集上均保持持续的领先. 特别地,在 ASSIST2012 和 ASSIST2017 数据集上,其准确率指标再次突破基准线.

NCDM 模型:UFCD 在 ASSIST0910, ASSIST2012 以及 KDDCUP 数据集上均取得了最优的性能表现. 尤其在 KDDCUP 数据集的高分区设定下,其表现再次超越基准.

3.5 消融实验 UFCD 主要由两个核心组件协同驱动,基于认知状态的分区方法(CSP)负责数据的认知分区,基于交叉注意力的模型聚合机制(CMAM)负责子模型的全局自适应整合. 为了剥离并验证各组件的独立贡献与实际效用,设计系统性的消融实验.

3.5.1 基于认知状态的分区方法(CSP)有效性验证 在机器遗忘的分区聚合范式中,数据划分策略的优劣直接决定底层子模型的质量上限. CSP 旨在打破传统随机或均匀划分机制的局限,通过精确计算学习者认知状态的相似度来实现数据的聚类划分,从而高度适配认知诊断模型对数据特征关联性的依赖. 为了客观验证 CSP 的有效性,在四个真实教育数据集上开展横向对比实验. 通过对比“采用 CSP 策略生成的子模型”与“采用 SISA 及 RecEraser 等遗忘框架默认划分策略生成的子模型”的预测表现,直观反映数据划分方法的优劣.

具体实验结果如图 6~8 所示. 由图可见,在 MIRT, DINA 及 NCDM 三种主干模型下,基于 CSP 策略构建的子模型表现均显著优于其他对比基线. 这一结果充分印证了 CSP 的核心价值,即通过精准捕捉用户间的认知同质性,CSP 构建了语义高度一致的训练分区. 和割裂认知特征关联的传统划分策略相比,CSP 显著降低了局部数据的分布复杂度,大幅增强了单一子模型的参数拟合能力与诊断鲁棒性,验证了该机制在教育测量场景下的优越性.

3.5.2 基于交叉注意力的模型聚合机制(CMAM)有效性验证 为了更清晰地剥离并量化基于交叉注意力的模型聚合机制(CMAM)对全局诊断性能的独立贡献,设计针对聚合模块的消融实验. 选取 MIRT 作为主干模型,在维持“基

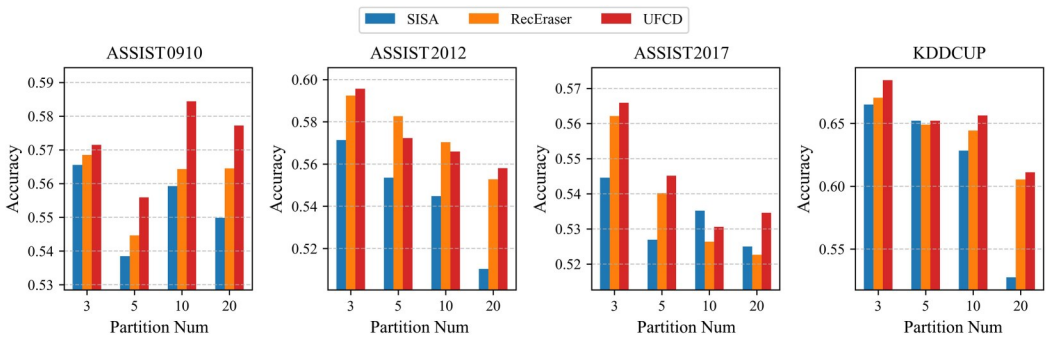


图6　不同遗忘方法下MIRT子模型的性能对比

Fig. 6　Performance of MIRT sub-models under different unlearning methods

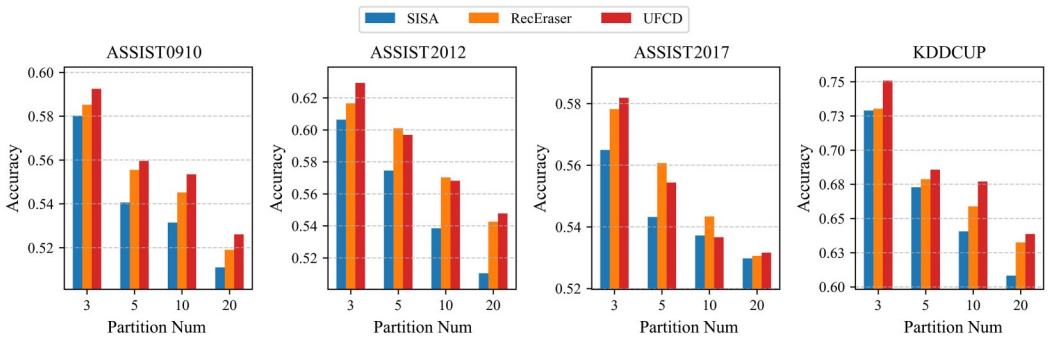


图7　不同遗忘方法下DINA子模型的性能对比

Fig. 7　Performance of DINA sub-models under different unlearning methods

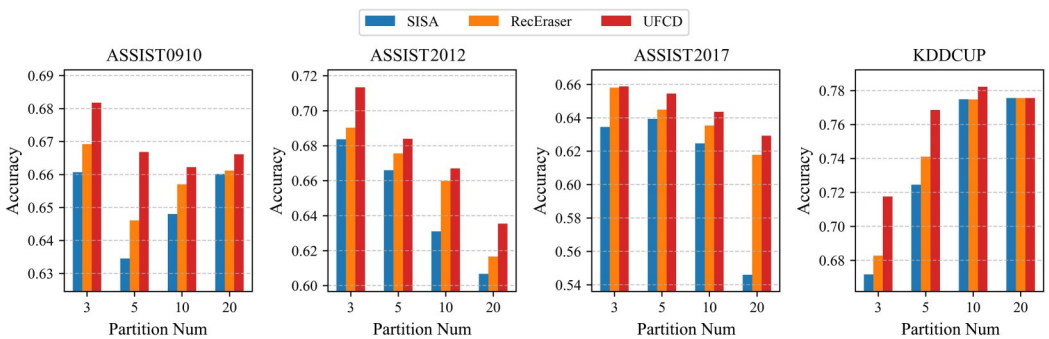


图8　不同遗忘方法下NCDM子模型的性能对比

Fig. 8　Performance of NCDM sub-models under different unlearning methods

于认知状态的分区(CSP)”策略不变的前提下,对比了两种不同的聚合策略,一种为“CSP”(即消融CMAM,预测时采用硬路由策略,仅依赖待测用户所属单一分区的局部子模型输出);另一种为“CSP-CMAM”(即完整UFCD框架,利用交叉注意力对所有子模型进行动态加权聚合). 具体对比结果详见表6.

当框架仅依靠CSP并退化为单一局部子模型预测时,受限于物理数据隔离导致的数据集规模缩减与特征稀疏性,局部子模型缺乏对全局数据分布的感知,其整体诊断性能受到制约. 而CMAM的引入有效弥合了这一“局部与全局”的信息鸿沟. 该机制通过交叉注意力动态计算待预测用户认知状态与各个子模型认知知识域的匹配度,将原本割裂的子模型进行自适应“软聚合”,最大化地激发各子模型的潜能.

实验结果表明,当CSP配合CMAM联合运作时,模型在所有数据集上的表现均获得了显著

表 6 MIRT 诊断模型的 CMAM 消融实验对比

Table 6 Ablation study of CMAM on the MIRT diagnostic model

数据集	对比方法	*ACC*	*AUC*	*RMSE*	*MAE*
ASSIST-0910	CSP	0.5714	0.5521	0.6168	0.4317
	CSP-CMAM	0.5853	0.5684	0.5195	0.4433
ASSIST-2012	CSP	0.5956	0.5636	0.6131	0.4038
	CSP-CMAM	0.6194	0.5833	0.5070	0.4143
ASSIST-2017	CSP	0.5658	0.5826	0.6089	0.4376
	CSP-CMAM	0.5851	0.6109	0.5163	0.4444
KDDCUP	CSP	0.6844	0.5846	0.4842	0.3466
	CSP-CMAM	0.7280	0.6153	0.4342	0.3517

跃升. 具体地,和纯 CSP 模式相比,完整框架的 *ACC* 分别提升 2.43%,3.99%,3.41% 和 6.37%;*AUC* 分别提升 2.95%,3.49%,4.85% 和 5.25%. 此外,*RMSE* 和 *MAE* 指标全面下降,进一步证实 CMAM 能够有效降低预测偏差,显著提升了系统的整体诊断精度.

4 结论

本研究遵循完全遗忘思想,提出一种面向认知诊断的机器遗忘框架(UFCD),解决智能教育系统中隐私数据遗忘效率与诊断性能间的平衡难题.

该框架通过两个关键阶段实现:在数据划分阶段,提出基于认知状态的分区方法(CSP),利用语义聚类将遗忘影响限制在局部,并赋予子模型群体特化能力;在模型聚合阶段,设计基于交叉注意力的聚合机制(CMAM),通过自适应加权融合有效弥补了物理隔离导致的信息缺失,提升全局预测精度.

在四个真实教育数据集上针对三种代表性认知诊断模型的广泛实验表明,UFCD 在快速响应遗忘请求的同时,依然能对异质化学生群体保持优异的诊断性能.

作为率先聚焦认知诊断隐私安全与机器遗忘交叉领域的研究之一,UFCD 为教育数据要素的合规化发展提供了重要参考. 未来主要沿两个维度拓展.

(1)探索更细粒度的认知语义划分标准. 计划引入知识图谱层级结构、多维认知属性等教育学先验知识,以进一步提升数据聚类的语义一致性与底层逻辑的可解释性.

(2)攻坚复杂动态场景下的遗忘鲁棒性. 真实教育系统中的遗忘请求往往呈现高并发与流式特征,针对批量遗忘场景,重点研究在中大规模数据并发撤回时规避聚合权重的剧烈震荡. 探索混合遗忘策略(如结合随机与置零初始化的混合状态重置机制),以平滑大规模数据剔除对局部子模型的冲击. 针对连续遗忘场景,需深入解决子模型在经历多次增量遗忘迭代后可能引发的“模型漂移”与特征空间畸变问题. 致力构建“先解耦、后补偿”的生成式机制,探索引入条件变分自编码器(CVAE)等手段进行底层特征的条件生成与数据重构,在长周期、真实复杂教育场景下全面保障认知诊断模型的适应性与安全性.

参考文献

[1] Anderson A, Huttenlocher D, Kleinberg J, et al. Engaging with massive online courses//Proceedings of the 23rd International Conference on World Wide Web. New York, NY, USA: Association for Computing Machinery, 2014: 687－698.

[2] Nguyen T. The effectiveness of online learning: Beyond no significant difference and future horizons. MERLOT Journal of Online Learning and Teaching, 2015, 11(2): 309－319.

[3] Udovichenko Y G, Nogaeva V U. General data protection regulation. Actual Problems of Russian Law, 2018(8): 221－224.

[4] Illman E, Temple P. California consumer privacy act. The Business Lawyer, 2019, 75(1): 1637－1646.

[5] 国家保密局. 中华人民共和国个人信息保护法. https://www.gjbmj.gov.cn/n1/2021/0903/c409088-32217039.html, 2021－09－03.

[6] 张宇静. 面向推荐任务的谱域聚类自适应聚合模型遗忘方法研究. 硕士学位论文. 武汉:华中科技大学, 2024.

[7] 王依文. 面向医学影像的轻量级数据遗忘技术. 硕士学位论文. 广州:广州大学, 2023.

[8] 黄耀. 面向深度神经网络的数据遗忘研究. 硕士学位论文. 广州:广州大学, 2023.

[9] Bourtoule L, Chandrasekaran V, Choquette-Choo C A, et al. Machine unlearning//2021 IEEE

Symposium on Security and Privacy. San Francisco, CA, USA: IEEE, 2021: 141－159.

[10] Chen C, Sun F, Zhang M, et al. Recommendation unlearning//Proceedings of the ACM Web Conference 2022. New York, NY, USA: Association for Computing Machinery, 2022: 2768－2777.

[11] Garg S, Goldwasser S, Vasudevan P N. Formalizing data deletion in the context of the right to be forgotten//Advances in Cryptology 2020. Cham: Springer, 2020: 373－402.

[12] Li Y Y, Chen C C, Zheng X L, et al. Selective and collaborative influence function for efficient recommendation unlearning. Expert Systems with Applications, 2023, 234: 121025.

[13] Zhang H B, Nakamura T, Isohara T, et al. A review on machine unlearning. SN Computer Science, 2023, 4(4): 337.

[14] Wu Y J, Dobriban E, Davidson S. DeltaGrad: Rapid retraining of machine learning models//Proceedings of the 37th International Conference on Machine Learning. New York, NY, USA: PMLR, 2020: 10355－10366.

[15] Byrd R H, Nocedal J, Schnabel R B. Representations of quasi-Newton matrices and their use in limited memory methods. Mathematical Programming, 1994, 63(1): 129－156.

[16] Cao X Y, Jia J Y, Zhang Z X, et al. FedRecover: Recovering from poisoning attacks in federated learning using historical information//2023 IEEE Symposium on Security and Privacy. San Francisco, CA, USA: IEEE, 2023: 1366－1383.

[17] Thacker W C. The role of the Hessian matrix in fitting models to measurements. Journal of Geophysical Research. Oceans, 1989, 94(C5): 6177－6196.

[18] Neel S, Roth A, Sharifi-Malvajerdi S. Descent-to-delete: Gradient-based methods for machine unlearning//Proceedings of the 32nd International Conference on Algorithmic Learning Theory. New York, NY, USA: PMLR, 2021: 931－962.

[19] Guo C, Goldstein T, Hannun A, et al. Certified data removal from machine learning models//Proceedings of the 37th International Conference on Machine Learning. Brookline, NY, USA: JMLR. org, 2020: 3832－3842.

[20] Sekhari A, Acharya J, Kamath G, et al. Remember what you want to forget: Algorithms for machine unlearning//Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc, 2021: 18075－18086.

[21] Mehta R, Pal S, Singh V, et al. Deep unlearning via randomized conditionally Independent hessians//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans. LA, USA: IEEE, 2022: 10412－10421.

[22] Cheng J L, Dasoulas G, He H, et al. GNNDelete: A general strategy for unlearning in graph neural networks//The 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR, 2023: 1－23.

[23] Chien E, Pan C, Milenkovic O. Efficient model updates for approximate unlearning of graph-structured data//The 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR, 2022: 1－33.

[24] Li X K, Zhao Y L, Wu Z Y, et al. Towards effective and general graph unlearning via mutual evolution. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(12): 13682－13690.

[25] Yan H N, Li X G, Guo Z Y, et al. Arcane: An efficient architecture for exact machine unlearning//Proceedings of the 31st International Joint Conference on Artificial Intelligence Main Track. Vienna, Austria: IJCAI, 2022: 4006－4013.

[26] Li Y Y, Chen C C, Zheng X L, et al. Making recommender systems forget: Learning and unlearning for erasable recommendation. Knowledge-Based Systems, 2024, 283: 111124.

[27] Chen M, Zhang Z K, Wang T H, et al. Graph unlearning//Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. New York, NY, USA: Association for Computing Machinery, 2022: 499－513.

[28] Zhang J H. Graph unlearning with efficient partial retraining//Companion Proceedings of the ACM Web Conference 2024. New York, NY, USA: Association for Computing Machinery, 2024: 1218－1221.

[29] Friedman J H. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 2001,29(5):1189－1232.

[30] Brophy J, Lowd D. Machine unlearning for random forests//Proceedings of the 38th International Conference on Machine Learning. New York, NY, USA:PMLR,2021:1092－1104.

[31] Rosenbaum P R. Testing the conditional independence and monotonicity assumptions of item response theory. Psychometrika, 1984, 49(3): 425－435.

[32] Yang H W, Qi T L, Li J, et al. A novel quantitative relationship neural network for explainable cognitive diagnosis model. Knowledge-Based Systems, 2022, 250:109156.

[33] Ginart A, Guan M, Valiant G, et al. Making ai forget you: data deletion in machine learning. Advances in Neural Information Processing Systems, 2019, 32: 3513－3526.

[34] Chen P H, Wei W, Hsieh C J, et al. Overcoming catastrophic forgetting by Bayesian generative regularization//Proceedings of the 38th International Conference on Machine Learning. New York, NY, USA:PMLR,2021:1760－1770.

[35] Gao L N, Zhao Z Y, Li C, et al. Deep cognitive diagnosis model for predicting students′ performance. Future Generation Computer Systems, 2022, 126: 252－262.

[36] Ackerman T A. Multidimensional item response theory models//Balakrishnan N, Colton T, Everitt B, et al. Wiley StatsRef: Statistics Reference Online: https://doi.org/10.1002/9781118445112.stat06401, 2014－09－29.

[37] De La Torre J. DINA model and parameter estimation: A didactic. Journal of Educational and Behavioral Statistics, 2009, 34(1):115－130.

[38] Wang X Z, Kou L Y, Sugumaran V, et al. Emotion correlation mining through deep learning models on natural language text. IEEE Transactions on Cybernetics, 2021, 51(9):4400－4413.

[39] Nguyen Q P, Oikawa R, Divakaran D M, et al. Markov chain monte Carlo - based machine unlearning: unlearning what needs to be forgotten// Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security. New York, NY, USA: Association for Computing Machinery, 2022:351－363.

(责任编辑 杨可盛)