可解释学习者模型: 可信个性化学习的技术关键

□ 江 波 丁莹雯 魏雨昂

[摘 要]基于人工智能的个性化学习技术一直是智能教育领域的研究热点,其技术挑战在于如何构建全面精准的学习者模型。在非常注重公平、伦理和责任的教育领域,人工智能的"黑箱"本质可能会阻碍人对机器决策的信任,因此构建透明和可解释的学习者模型尤为重要。通过对学习者模型特征、结构和决策结果的解释,可以让教育关益者理解其动机,接纳其决策,实现更好地人机协作。该研究将可解释人工智能的技术理念延伸至个性化学习中,通过分析其研究现状,阐明实现可解释学习者模型的必要性,并剖析现有的可解释学习者模型实例的技术原理,最后提出可解释学习者模型的基本框架,旨在将可解释性作为学习者建模的关键原则。对可解释学习者模型的研究可为个性化学习系统的设计、开发、应用到评估的整个周期实现可解释性提供借鉴及参考,驱动可信个性化学习成为可能。

[关键词]学习者模型;个性化学习;可解释性;可信人工智能

[中图分类号]G420 [文献标识码]A [文章编号]1672-0008(2023)02-0048-10

[DOI] 10.15881/j.cnki.cn33-1304/g4.2023.02.005

一、引言

人工智能的"黑盒"模型会产生结果不可解释、无法追责和歧视偏见等风险(孔祥维,等,2022)。因此,联合国教科文组织在《人工智能伦理建议书》(2021)中提出,在设计与使用人工智能技术时需要考虑其"透明度和可解释性(Transparency and Explainability)"。我国《新一代人工智能伦理规范》(国家新一代人工智能治理专业委员会,2021)也强调,在设计、实现、应用等环节中,需要提升算法透明性、可解释性、可理解性、可靠性和可控性。可解释性旨在让人工智能模型过程可以被追溯、决策机制可以被解释、决策结果可以被理解,是可信人工智能的关键维度和重要原则之一(孔祥维,等,2022; Liu,et al.,2022)。

教育是人工智能应用的重要领域之一,人工智能赋能教育最主要的目的之一是实现个性化学习(刘斌,等,2021),学习者模型(Learner Model)是个性化学习系统的核心组件。学习者模型是指从知识、认知、情感等方面对学习者进行画像,是当前智能教育领域的研究热点问题(戴静,等,2022)。从技术的角度看,无论是早期的智能导学系统(Intelligent Tutoring System)还是如今的自适应学习系统(Adaptive Learning System),其技术路线均是通过采集学习者的学习过程数据,再利用人工智能算法建立学习者模型,驱动个性化学习路径和资源的生成。当前学习者模型研究侧重于提升模型预测性能,对于模型可解释性重视不足。例如,在被广泛研究的知识追踪问题中,基于长短期神经网络、卷积神经网络和图神经网络等各种深度学习算法被广泛用于知识追踪模型(Khajah,et al.,2016; Nakagawa,et al.,2019; Pandey,et al.,2019; Piech,et al.,2015; Zhang,et al.,2017),促使模型的预测性能得到显著提升。再如,在学习者情绪建模中,大量研究在获取脑电、皮肤电等内在生理数据以及表情、动作、话语等外显行为数据的基础上,再结合深度学习技术进行建模,实现了对复杂学习情绪的精准预测(李云,等,2019; Martinez,et al.,2013; Rouast,et al.,2019)。然而,模型精度上的提升

基金项目:本文系 2019 年度国家自然科学基金面上项目"面向图形化编程项目式学习的自动评价研究及应用"(项目编号:61977058)、2020 年度上海市自然科学基金面上项目"中小学信息科技核心素养自动评价研究"(项目编号:23ZR1418500)、2020 年度上海市科技创新行动计划"人工智能"专项"教育数据治理与智能教育大脑关键技术研究及典型应用"(项目编号:20511101600)的研究成果。

作者简介:江波,博士,副教授,华东师范大学教育信息技术学系(上海 200062);丁莹雯,在读硕士研究生,华东师范大学教育信息技术学系(上海 200062);魏雨昂,在读博士研究生,华东师范大学计算机科学与技术学院、上海智能教育研究院(上海 200062)。

引用信息:江波,丁莹雯,魏雨昂,2023.可解释学习者模型:可信个性化学习的技术关键[J].远程教育杂志,41(2):48-57.

往往以牺牲其可解释性为代价。在深度知识追踪中，尽管模型预测结果的准确度得到了提升，但教师和学生却难以从模型的表征结构和输出结果中发现影响知识掌握的内外因素。因此，近年来越来越多的学者开始重视对于提升学习者建模可解释性的研究。例如，在深度知识追踪中使用注意力机制提升模型结果的可解释性（刘坤佳，等，2021）；在 IRT 的基础上引入学习率和学习次数进行知识追踪（Cen，2009）；从练习材料中获取关于知识点的信息，再采用马尔可夫性质实现可解释的知识追踪（Liu，et al.，2019）；利用因果结构方法对学习者的 MOOC 学习过程进行建模（Koedinger，et al.，2015）。

总的来看，国内外研究已开始探索可解释的学习者模型，但主要停留在知识追踪等微观层面的可解释，尚未从宏观上揭示可解释学习者模型的基本特征与关键要素。为此，本研究探讨智能教育视角下的可解释学习者模型，辨析构建可解释学习者模型的重要性和必要性，梳理和分析可解释学习者模型的现有理论研究及应用实践案例，最后提出可解释学习者模型（Explainable Learner Model，xLM）的基本框架，旨在从多个角度阐述构建可解释学习者模型的基本特征与关键要素，为个性化学习的技术实现提供理论参考。

二、教育视角下的可解释人工智能

（一）智能教育的可解释性

在人工智能的发展过程中，必须遵循公平、责任、透明和道德（Fairness，Accountability，Transparency，Ethics，FATE）原则[①]，其中，透明是实现模型可解释性的途径之一。透明原则体现在：模型的可模仿性（Simulatability）、可分解性（Decomposability）和算法透明（Algorithmic Transparency）（Barredo Arrieta，et al.，2020），即模型的各个决策单元内部可被解释、输出可被理解、计算步骤可被模仿。解释是人类相互沟通并获取信任的途径，可解释人工智能（Explainable Artificial Intelligence，xAI）通过与用户交互，向用户传递信息，帮助用户理解模型，从而建立人机信任。因此，可解释性是实现人工智能可信的核心，也是决定人工智能能否被广泛应用的关键因素（Confalonieri，et al.，2021）。xAI 是帮助人类理解、解释、信任算法学习过程及最终决策的工具和框架、过程和方法[②]。美国国防部高级研究计划局（DARPA）制定了 xAI 规划，帮助使用者从六个角度来理解模型：（1）为什么这么做；（2）为什么不那样做；（3）何时成功；（4）何时失败；（5）何时能信任模型；（6）为什么出现错误（Gunning，2017）。该规划旨在为开发者和研究者在设计可解释模型的具体解释内容时提供指导。虽然，足够优秀的人工智能模型可解决其算法精度与可解释性的冲突（Gunning，et al.，2021），但在实际应用中，仍不可忽视模型的准确性与可解释性的共存问题。统计研究表明（如图 1 所示），机器学习模型的预测性能与其可解释性总体成反比关系（Gunning，et al.，2019），因此，如何平衡二者是xAI 亟待解决的问题。

在教育领域中，尊重学生差异、遵循学生个性化发展的教育理念不断得到重视（李广，等，2005），教育中的人工智能：改变学习的速度（*AI in Education: Change at the Speed of Learning*）（Duggan，et al.，2020）中强调，从开发阶段就要关注智能教育的伦理问题，通过人工智能技术的透明度和监督机制能够确保教育的优质、包容和公平。以数据支持的教育决策是实现学生个性化发展的重要手段（张学波，等，2022）。因此，当人工智能技术辅助教学关益者做出对应决策时，需要确保模型结果准确性、解释可靠性和解释相关性，充分利用学科知识、教学经验、教育心理学等理论知识，让关益者透彻地分析模型细节、理解输出结果，从而以人机协同的形式做出科学、合理的教育干预。

从技术的角度，模型可解释性分为事前（Ante-hoc）可解释性和事后（Post-hoc）可解释性。事前可解释性模型称为“白盒模型”，此类模型结构简单，其推理机制、决策过程等易于被使用者理解，如线性回归模型、决策树模型等。事后可解释性模型即“黑盒模型”，其预测精度高但结构复杂，故使用者无法直接理解其内部的推理决策过程。对于本身不具备可解释性的模型来说，通常可以采用特征相关性计算、样例解释、局部解释、可视化对比解释和反事实解释等技术对其进行分析（Khosravi，et al.，2022）。事前可解释性在建模阶段关注参数设置、模型结构，而事后可解释性从结果出发，通过可解释的方法探究其决策

① https://www.microsoft.com/en-us/research/theme/fate/.
② https://www.ibm.com/cn-zh/watson/explainable-ai.

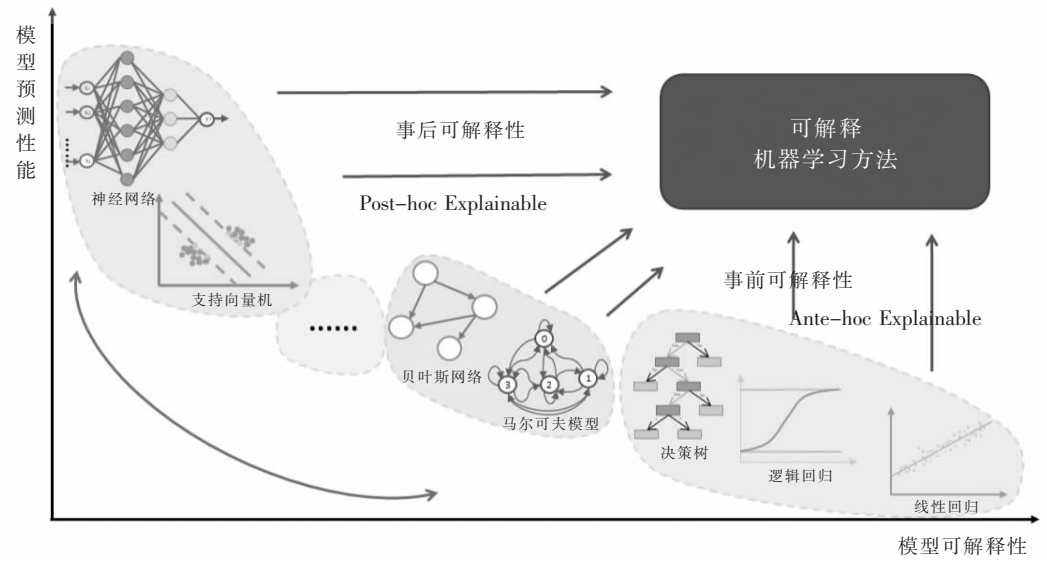

图 1 模型的性能与可解释性之间的关系

背后的原因，即事前可解释性建模、事后可解释分析。以个性化学习技术的发展历程为例,早期的教学决策模型具有良好的可解释性，如 SOPHIE-I(Nicolosi,1988)、BUGGY(Brown,et al.,1978)等,但其数据利用有限、预测性能低。随着人工智能技术的进步，一些学习者模型在预测性能上有所提升并保留了可解释性。例如基于概率图的贝叶斯知识追踪(Bayesian Knowledge Tracing,BKT)模型(Corbett,et al.,1994)便是具有代表性的可解释学习者模型之一。近年来，具有强大预测性能的深度学习技术受到教育研究者的追趋逐者，基于该技术的深度知识追踪(Deep Knowledge Tracing,DKT)应运而生(Khajah,et al.,2016; Nakagawa,et al.,2019; Pandey,et al.,2019; Piech,et al.,2015; Zhang,et al.,2017)。随之而来的问题便是模型可解释性的大幅下降,因此,探寻高性能的可解释教育教学决策模型便成为当前研究热点问题。

(二)学习者模型的可解释性

从 20 世纪 50 年代开始，计算机技术和网络辅助的教学系统打破了传统课堂教学在时间和空间上的约束,但依旧缺少情景支持和适应性支持、学生与系统的协同支持(Xu,et al.,2002)。因此,随着人工智能与教育的逐步融合，支持个性化学习的智能导学系统（Intelligent Tutoring Systems,ITS）成了研究热点,使得教辅系统真正走向“智能”。时至今日,教育人工智能技术应用已衍生到自适应学习、自动测评、课堂评价、数据决策等场景中(杨晓哲,等,2021)。

智能教育场景的本质是人工智能辅助教育关益者明晰学习的原理与机制、开发智能学习工具、提供适应性学习环境(闫志明,等,2017)。尽管学习者模型服务的主体是学习者，但是系统采集学习者的多维特征来构建模型,再结合教学模型和领域模型,形成“学-教-评”的“内循环”,输出状态评估、资源推荐、教学策略等信息反馈给“学-教-管-评-研”流程中对应的关益者,以数据驱动、人机交互的方式循序渐进地提升学生学习状态。同时“研”和“管”的关益者结合实际形势和教育理论提出相对宏观的优化建议。如图 2 所示,整个系统通过微观调整的“学-教-评”和宏观把控的“管-研”达到“外平衡”。智能教育始终坚持“以人为本”的教育理念（祝智庭，等，2021),而学习者模型是调节“内循环”和“外平衡”的关键，因而实现该模型的可解释性是提升教育主体理解和信任的核心。构建可解释学习者模型势必能够改善教学环节中的 FATE 问题，促使智能教育教学场景中的每一位关益者都受益于学习者模型的可解释性,帮助关益者洞察学习者学习状态、披露学习者发展规律、模拟学习演变过程或推演教学情境(Rosé,et al.,2019),从而做出相对公平、合理的教育干预和教学决策。

学习者模型是对学习者特征的抽象表示和描述,是实现个性化学习的核心(岳俊芳,等,2017),也是实现学习管理系统、智能导学系统、教育机器人、自适应学习系统等智能教育产品研发的关键（王一岩,等,2022)。学习者特征和建模方法是学习者模型

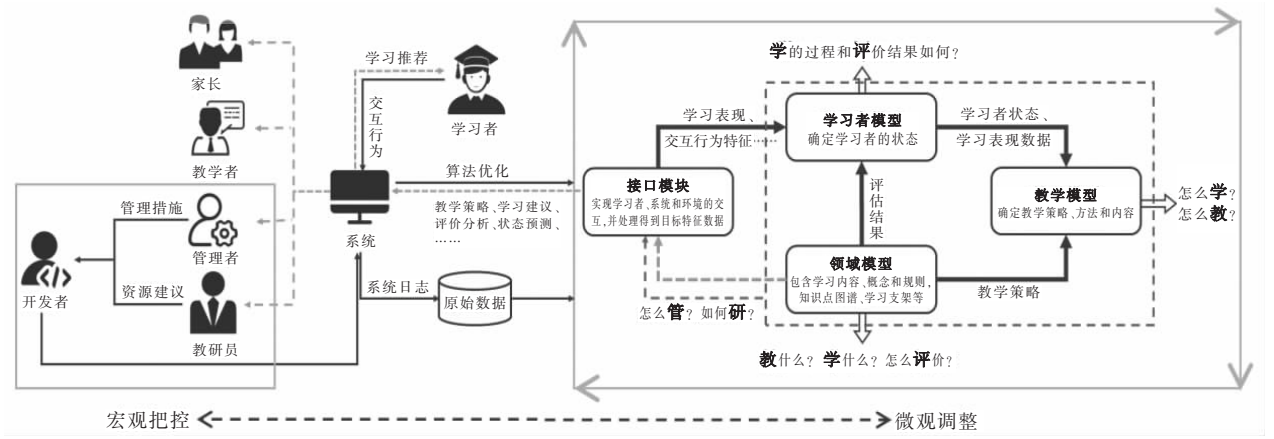

图 2 “学–教–评–管–研”个性化学习系统架构

的关键要素。对学习者全方位、多层次的建模分析，能够为个性化学习、智慧化教学、精准化管理的开展提供多元支持(黄涛，等，2020)，因此学者们对学习者特征的选择从单一认知维度转向了认知维度与非认知维度相结合。例如克里萨菲亚迪等(Chrysafiadi，et al.，2013)总结了学习者建模方法，认为学习者特征维度包括知识水平、技能、学习偏好、动机、情感、认知和元认知等方面。这些特征又可以归纳为智力和非智力因素(武法提，等，2019)，前者包括学习者的知识状态、认知能力；后者则指向社会与情感能力等。黄涛等(2020)强调除学习者的基础数据外，更要深度挖掘不同应用场景中学习者的知识、认知、情感、交互等方面的发展情况。张涛等(2022)根据学习者本体、内部心理和外部行为特征提出了由本体、认知、知识、行为和情感构成的学习者通用模型。面对复杂的学习情境需要选择多维特征进行建模，才能取得理想的模型效果。王小根等(2021)从学习者的特征、模型构建和建模技术三个维度对学习者建模研究进展的分析发现，学习者建模正朝着特征选择的多维化、数据收集的聚集化等方向演进。可以看出，面对数据聚集化、多维化的转变和模型可视化、精准化的需求，学习者模型变得越发复杂，因而实现学习者模型的可解释性变得尤为重要。

三、可解释学习者模型的分类及应用案例

本研究从技术方法的角度将学习者建模大致划分为描述建模、预测建模和归因建模这三类。这三类建模方法在准确性和可解释性上呈逐步递进的关系：描述模型通常采用统计、聚类的方法；预测模型主要通过回归、分类来实现，这两类模型具有部分可解释性；归因模型是以因果模型为基础的可解释模型，通过因果结构、处理效应异质性、反事实推理等方法建模。

(一) 描述建模：基于统计描述分析数据可解释性

描述型建模主要通过对历史数据的处理来展示其总体特征和结构，即“已经发生了什么”，通常由降维、特征提取等方法实现建模。在教育领域进行描述建模的主要目的是通过对学习者行为、生理等数据的采集和统计来构建学习者画像，从而进一步做出教学干预。

现有研究通过对学习者数据的采集和统计，从而实现对学习者能力的描述和刻画，再依据学习者能力进行资源推荐。例如，萨布莱罗尔斯等(Sablayrolles，et al.，2022) 提出了通过语义本体表示知识、技能和能力(Knowledge–Skill–Competency，KSC)，构建学生核心素养的描述模型，进而实现基于核心素养的个性化学习推荐，并且具有可解释性。在图 3 所示的可解释图例中，系统向学习者解释了其在相关素养、知识和技能上的掌握情况，同时向学习者解释了不同的素养、知识和技能之间的关系。例如，在图 3 中所展示的某个学习者的 Shell 编程这门课程的掌握情况中，编写交互式脚本 (Write Interactive Scripts)这一素养是操纵变量(Manipulate Variable)的先修素养，而操纵变量包含 2 项知识，分别是了解变量的概念(Know the Notion of Variable)和了解变量的环境(Know the Environment of Variables)；2项技能分别为获得变量值 (Get the Value Assigned to A Variable)和显示变量值(Display the Value of A Variable)。同时，系统通过该学习者在了解变量所处的

环境(Know the Environment of Variables)这一技能的学习资源上的交互数据发现，该学习者在这个知识点上处于未掌握状态(Not Mastered)。

科纳蒂等(Conati,et al.,2021)等也在个性化学习系统通过描述建模来解释干预。作者设计开发了一个面向算法学习的个性化学习系统，该系统能为学习者提供自适应提示、反馈和对应的解释。该系统通过采集学习者的学习交互数据，利用聚类分析方法识别出学习行为模式相似的学习者，再通过关联规则方法挖掘学习行为模式与学习成绩之间的关联性,来找到导致不佳学习成绩的行为模式。最后,学困生会收到系统自适应干预，告知其下一步的学习策略,并向其详细解释为何进行干预。研究证实提供解释提升了学习者对个性化学习中干预的信任、感知提示的有用性和再次使用提示的意愿。从这两个典型案例可以看出，利用描述建模对个性化学习系统做出的决策行为进行充分解释，有助于提高学习者对于自身学习情况的理解,促进学习者的学习。

(二)预测建模:基于特征结构洞察模型事前可解释性

描述建模是对已经发生的行为进行分析回顾，而预测建模则是一种前瞻性的建模，旨在通过对学习者的可观测数据进行建模来预测他们的潜在特征,如知识掌握、认知水平和情感状态。当前,知识追踪是预测建模中最受关注的方向之一，其中以基于可解释的贝叶斯概率模型的贝叶斯知识追踪模型(Bayesian Knowledge Tracing,BKT)(Corbett,et al.,1994)、基于可解释的逻辑回归模型的可加性因素模型(Additive Factor Model,AFM)(Cen,2009)和基于不可解释的深度神经网络的深度知识追踪模型(Deep Knowledge Tracing,DKT)(Piech,et al.,2015)最具代表性。BKT 模型主要有影响学习结果的初始掌握概率 P(L)、学习概率 P(T)、失误率 P(S)和猜测率 P(G)四个关键参数;AFM 则在 IRT 的基础上融合了学习过程学习率、练习次数的动态特征,对学习掌握状态进行动态估计，这两种模型都通过参数估计确定最优解。反观 DKT 模型,尽管它具备优秀的预测性能,但由于深度神经网络算法复杂、内部结构不透明，因此在教育情境中无法直接得到关益者的信任,也同样难以证明其算法的公平性和伦理性。

为了满足学习者模型具备可解释性和高预测性能的需求，国内外学者分别基于 BKT 和 AFM 模型进行改进。如对模型本身的参数进行个性化建模,帕多斯(Pardos,et al.,2010)考虑到每位学习者初始的先验知识 P(L)和本身具备的学习能力 P(T)存在差异，因此将这两个参数做个性化处理。尤德尔森(Yudelson,et al.,2013)将初始掌握概率和学习率归属为学生参数，将失误率和猜测率则归属为技能参数。其针对学生参数的个性化实验发现,个性化 P(T)模型比个性化 P(L)模型准确率更高,由此推断学习者的后天学习差异比初始掌握差异更为重要。基于与题目难度相关的技能参数也会随学习者习得过程发生变化,帕多斯(Pardos,et al.,2011)将题目难度拓扑到 BKT 模型建立了项目难度效应模型 KT-idem。

除了对模型本身的参数进行处理外，也有研究将模型与学习过程和教学情境相关的其他参数进行

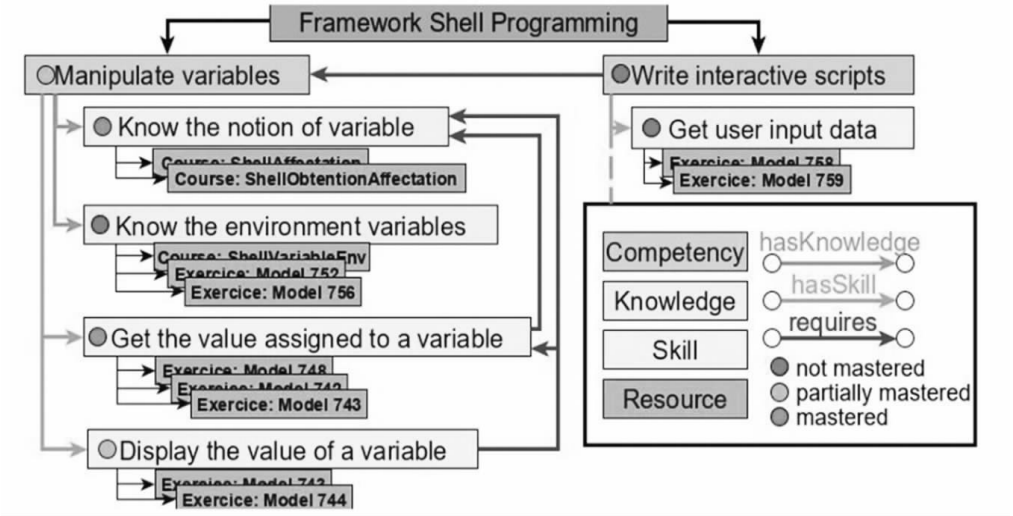

图 3 基于编程知识构建的能力框架案例

融合。学习者的情绪会随题目难易、知识点掌握情况、学习时间等因素发生波动，斯波尔丁(Spaulding, et al.,2015) 提出了考虑学习者的情绪状态的情感BKT模型。此外，艾宾浩斯遗忘曲线表明，学习者的练习时间间隔、练习次数会导致不同程度的遗忘，有学者(Qiu,et al.,2011)发现BKT模型高估了学习者间隔一天后的学习表现，提出KT-遗忘模型。黄诗雯等(2021)对智能导学系统采集到的行为数据分类成积极和消极行为，融入遗忘因素，最终形成BF-BKT模型。在逻辑回归模型中，很多研究对传统AFM模型中的参数做出了优化和调整。例如巴甫利克(Pavlik,et al.,2009)提出将AFM中的练习次数划分为作对和作错次数，构建PFA模型(Performance Factor Analysis)，麦克莱伦等 (MacLellan,et al., 2015)又将失误参数纳入上述两个模型。

(三)归因建模:基于因果分析分析模型事后可解释性

因果推理是当前实现模型事后可解释的主流技术，是提升人工智能可信度的充要条件 (黄闪闪, 2021)，也是一种有效的事后可解释方法。目前，大多数机器学习模型都是基于关联统计的，事实上变量之间的关联性可能无法被解释，比如婴儿出生率与白鹳种群之间的关联性(Matthews,2000)。关联性分为因果、混淆、样本选择偏差，用不可解释的关联特征构建的模型，从根本上无法保证模型的可解释性，而因果关系兼顾了稳定性和可解释性 (Cui,et al., 2022)。因此，通过剖析学习者模型中的因果结构，可以实现学习者模型的可解释性。稳定的因果关系分为关联、干预和反事实推理三个层次，分别借由观察、行动和想象三种方式来确定关系 (Pearl,et al., 2018)。在教学场景中，通过这三种因果推理方式，可以分析不同的行为特征对学习效果产生的影响。在未知模型中寻找特征的因果关系是实现模型的可解释性和稳定性的有效手段，即一个具备因果关系的模型结构被认为是可解释的。为证明这一观点，现有研究通过对不同学习场景下的具体学习行为进行归因，并通过具体数据与实验进行分析验证。

在个性化学习场景中，江波等(2023)利用学习者在某个性化学习平台中产生学习过程数据，基于数据驱动的因果分析方法探究影响学习收益的学习行为。分析得到的关键学习行为与学习收益的因果关系模型如图4所示。由图4可以看出，学习者是否采纳该系统的个性化学习建议对于学习收益有着最为显著的正向影响关系 (影响系数为159.0266)，采纳该系统的个性化建议通过减少学生做题数量 (影响系数为-10.5013)，而学习者在学习过程中频繁的寻求提示则会给降低学习收益 (系数为-10.5260)。该结果不仅表明该个性化学习系统是否有效以及如何有效，还发现了学习者典型学习行为对于学习收益的直接影响。在MOOC学习的场景中，科丁格等(Koedinger,et al.,2015)聚焦于学习者的前测、课程活动、测验得分和期末考试，挖掘这些特征之间是否存在潜在因果关系。因果分析发现，主动学习行为能够显著提升学习收益。在编程学习场景中，江波等(Jiang,et al.,2021)对学习者的参与度进行归因，通过因果推断发现，发现只有项目的部分计算思维特征会影响项目的流行度。在数字教材阅读的场景中，顾美俊等(2022)分析了阅读者的阅读行为对成绩的影响，发现读者的若干回看行为都不同的程度地影响后测成绩。

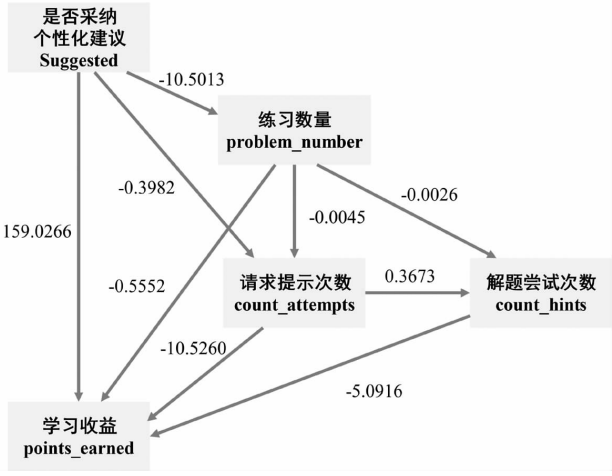

图4 因果关系模型图

上述研究中的学习者模型的可解释性由因果关系而得出，其主要目的是挖掘数据背后潜在的因果关系，从而为教学关益者提供可信的指导和帮助。虽然，这些基于理论层面的研究尚未构建完整的个性化学习系统，但是这些模型对个性化学习系统的开发和优化提供了新思路，以此评估系统是否真正有益于提升学习者的学习效果，也能够对学习者的行为表现做出科学合理的决策。

四、可解释学习者模型的基本框架

通过已有案例可以发现，当模型的解释能够具

备教育可读性时，才能让教育关益者更好地接纳和信任模型。这就要求模型的可解释性需要贯穿其整个生命周期,教育技术开发人员在设计、开发可解释学习者模型时,应当要考虑具体的教育情境、使用对象等因素，能够在模型的应用阶段向非技术人员做出解释。基于此,本研究提出的可解释学习者模型的基本框架(Explainable Learner Model,xLM),旨在凝练可解释学习者模型需要考量的关键要素，如图5所示。框架主要由三部分组成:应用场景、通用方法和基本要求。在建模过程中,首先,根据特定的应用场景展开解释内容;再次,选择合适的解释技术并遵循解释三要素;最后,基于五大原则向教育关益者做出解释,最终输出模型的决策结果及其解释,并反馈给"学-教-管-评-研"环节中的对应对象。

(一)可解释学习者模型的应用场景

xAI-ED框架(Khosravi,et al.,2022)显示,构建可解释教育模型必须充分考虑各个教学环节中教育关益者的需求,这里结合具体教育应用场景将解释划分为"学-教-管-评-研"五个场景。从图2所示的个性化学习系统架构中不难看出学习者模型的输出直接影响了学习者的学习,它向学习者提供个性化资源推荐、路径推荐和学情报告等,同时也与其他环节息息相关,尽管可解释学习者模型并不能完全覆盖系统的所有解释，但其影响了其他教学流程的推进和优化。因此在构建可解释学习者模型过程中,充分考虑对应教学情境做出解释是必要的,比如通过解释,帮助学习者理解学情分析的准确性和推荐资源的合理性;帮助教学者理解教学模型给出的教学策略和建议;向教学管理者解释数据诊断的过程、提出治理决策的原因;向关益者阐明学习者的评价指标和流程;让教研员信任教学方案设计和资源分配决策。

(二)可解释学习者建模的通用方法

可解释学习者建模的通用方法包括解释技术和解释要素。其中,解释技术划分为解释范围和方法。解释方法按模型结构划分为事前可解释模型和事后可解释模型。事前可解释除了线性回归、决策树等简单机器学习模型外,也包括借助知识图谱、注意力机制等方法实现事前可解释性。对于不可通过模型自身得到可解释性的，则需要通过部分依赖图PDP(Partial Dependence Plot)、LIME (Local Interpretable Model-Agnostic Explanations)、SHAP (Shaply Value)等技术来实现事后可解释性。在教育场景中,事前可解释性学习者模型能够帮助教育技术开发人员和教研员清楚地了解模型内部结构，结合领域知识及时做出干预,以确保模型构建的合理性;而事后可解释性学习者模型主要是向参与智能学习系统的学习者、教学者等教育参与者解释模型,进而做出决策原因,通过解释让他们能够信任模型的决策,充分利用

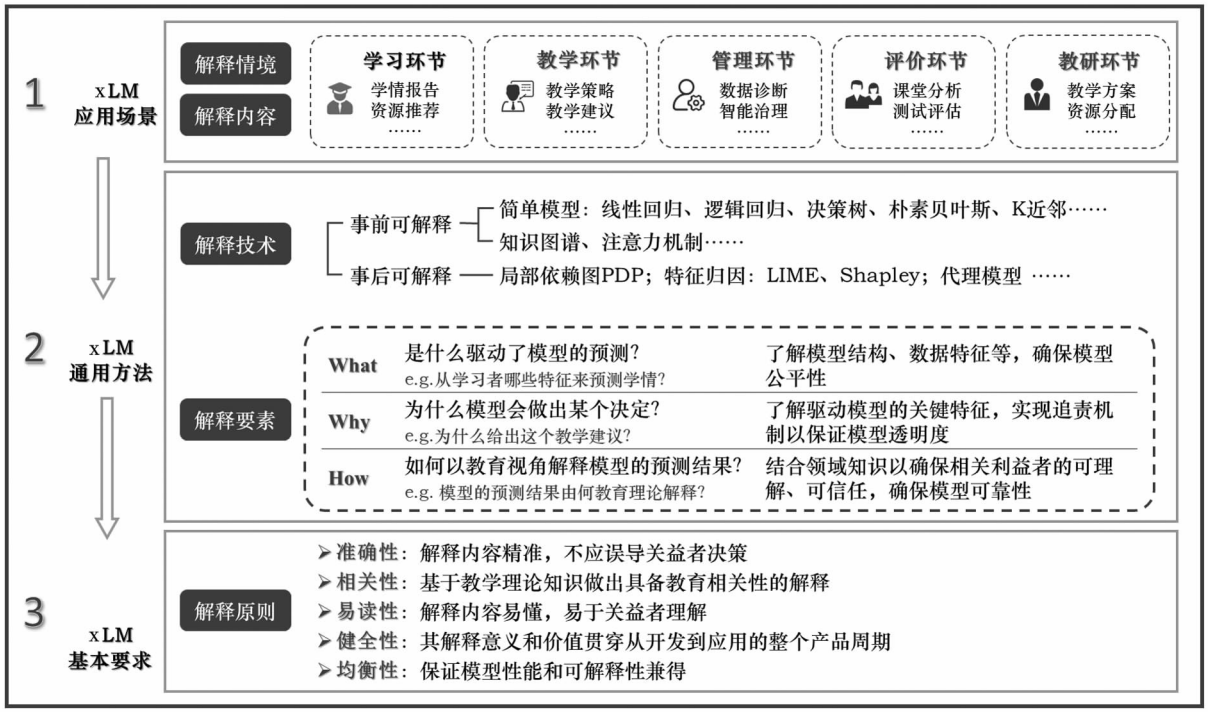

图5 可解释学习者模型(xLM)框架

系统资源提升学习收益。

解释要素围绕“What”“Why”“How”三个关键词展开。“What”表示向教育关益者解释是什么驱动了模型的预测，例如说明学习者建模过程中需要考虑的学习者特征，其目的是了解模型的结构、数据特征，并确保模型公平性。“Why”表示向教育关益者解释模型做出决策的原因，例如对模型给出的某一推荐策略或者教学建议做出分析，旨在帮助教育关益者了解驱动模型的关键特征，及时诊断模型的有效性，实现追责机制，并提高模型透明度。“How”表示以教育视角向教育关益者解释模型的预测结果，例如模型的决策过程和结果符合哪个具体的教育理论知识，通过结合教育领域知识确保教学关益者的可理解、可信任，由此确保模型可靠性。

(三)可解释学习者模型的基本要求

可解释学习者模型需要遵循教育场景下的解释基本要求，具体将解释原则概括为准确性、相关性、易读性、健全性和均衡性。

解释内容的准确性需要保证解释内容精准，避免非技术人员因对解释内容产生误解而影响决策。相关性体现在当技术应用于某一特定领域时，需要遵循该领域内的理论基础和伦理准则，即可解释学习者模型包含但不限于对算法本身做出解释，它必须结合教育原理、遵循教育伦理、运用领域知识，体现教育相关性，并且能够易于教育主体理解。易读性要求模型的解释内容能被非技术人员读懂并接纳。健全性体现在既要考虑开发阶段实现技术层面的模型可解释性，又要确保在实际运用过程中教育层面的模型可解释性，进而减少教育者提出解释需求，并进一步回溯迭代的成本。需要指出的是，均衡性旨在权衡模型性能和可解释性的优先级，一味追求模型的可解释性而放弃某些性能优秀的算法也是不可取的。

五、结语

本研究聚焦于个性化学习技术中的学习者模型，从描述建模、预测建模、归因建模这三个角度，阐述了可解释学习者模型的建模思路以及可解释性的实现方法。通过分析和梳理，提出了通用的可解释学习者模型框架 xLM，期望能够使教育工作者意识到学习者模型的可解释性的重要性，并为其研究、设计、应用与评估可解释学习者模型提供科学有效的方法，通过技术路径助力可信个性化学习的实现。本研究提出的 xLM 框架从教育和技术两个角度考虑了建立可解释学习者模型过程中需要考虑的多个要素。此外，模型可解释性的优劣主要取决于其是否与使用者达成有效沟通、做出合理解释。然而，现阶段对于模型可解释性的评估方法主要以质性评估为主，尤其在个性化学习领域对于模型可解释性的评估方法仍较少，没有形成比较严谨、统一的评价体系。因此，本研究提出的 xLM 框架，为可解释学习者模型评估阶段提供了可参考的具体维度。可解释学习者模型才刚刚起步，这也是构建可信教育人工智能的第一步，仍有很多先进的人工智能可解释技术值得智能教育领域工作者继续探究、创新与应用，也还有更多未能充分开发的应用场景亟待研究，未来会有更加多样的解释模型的技术工具、方法手段融入智能教育模型。

[参考文献]

戴静, 顾小清, 江波, 2022. 殊途同归：认知诊断与知识追踪——两种主流学习者知识状态建模方法的比较[J]. 现代教育技术(4): 88–98.

顾美俊, 江波, 殷成久, 2022. 数字教材阅读中回看行为与学习效果的关系[J]. 现代教育技术(5): 49–58.

国家新一代人工智能治理专业委员会,2021. 新一代人工智能伦理规范[EB/OL]. [2023–03–02]. http://www.safea.gov.cn/kjbgz/202109/t20210926_177063.html.

黄闪闪, 2021. 因果推理研究的认知进路:基于人工智能视角[J]. 湖南科技大学学报(社会科学版)(2): 44–50.

黄诗雯, 刘朝晖, 罗凌云, 等, 2021. 融合行为和遗忘因素的贝叶斯知识追踪模型研究[J]. 计算机应用研究(7): 1993–1997.

黄涛, 王一岩, 张浩, 等, 2020. 智能教育场域中的学习者建模研究趋向[J]. 远程教育杂志(1): 50–60.

江波,章恒远,魏雨昂, 2023.如何判定自适应学习系统的有效性——基于因果结构分析框架[J].现代远程教育研究(2): 95–101.

孔祥维, 王子明, 王明征, 等, 2022. 人工智能使能系统的可信决策:进展与挑战[J]. 管理工程学报(6): 1–14.

李广, 姜英杰, 2005. 个性化学习的理论建构与特征分析[J]. 东北师大学报(3): 152–156.

李云, 阚威, 2019. 基于 LSTM 的脑电情绪识别模型[J]. 南京大学学报(自然科学)(1): 110–116.

联合国教科文组织,2021.人工智能伦理问题建议书[EB/OL]. [2023–03–02]. https://unesdoc.unesco.org/ark:/48223/pf0000381137_chi.

刘斌, 王孟慧, 2021. 人工智能时代的个性化学习:内涵、技术

支持与实现路径[J]. 教育探索(7): 80–83.
刘坤佳, 李欣奕, 唐九阳, 等, 2021. 可解释深度知识追踪模型[J]. 计算机研究与发展(12): 2618–2629.
王小根, 吕佳琳, 2021. 从学习者模型到学习者孪生体——学习者建模研究综述[J]. 远程教育杂志(2): 53–62.
王一岩, 郑永和, 2022. 基于情境感知的学习者建模:内涵、特征模型与实践框架[J]. 远程教育杂志(2): 66–74.
武法提, 黄石华, 殷宝媛, 2019. 基于场景感知的学习者建模研究[J]. 电化教育研究(3): 68–74.
闫志明, 唐夏夏, 秦旋, 等, 2017. 教育人工智能(EAI)的内涵、关键技术与应用趋势——美国《为人工智能的未来做好准备》和《国家人工智能研发战略规划》报告解析[J]. 远程教育杂志(1): 26–35.
杨晓哲, 任友群, 2021. 教育人工智能的下一步——应用场景与推进策[J]. 中国电化教育(1): 89–95.
岳俊芳, 陈逸, 2017. 基于大数据分析的远程学习者建模与个性化学习应用[J]. 中国远程教育(7): 34–39.
张涛, 张思,2022. 教育大数据挖掘的学习者模型设计与计算研究[J]. 电化教育研究(9): 61–67.
张学波, 林书兵, 2022. 数据驱动的差异化教学决策:症结、逻辑与机制[J]. 现代远程教育研究(3): 48–57.
祝智庭, 韩中美, 黄昌勤, 2021. 教育人工智能(eAI):人本人工智能的新范式[J]. 电化教育研究(1):5–15.
BARREDO ARRIETA A, DÍAZ–RODRÍGUEZ N, DEL SER J, et al., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI[J]. Information Fusion,58: 82–115.
BROWN J S, BURTON R R,1978. Diagnostic models for procedural bugs in basic mathematical skills[J]. Cognitive Science, 2(2): 155–192.
CEN H, 2009. Generalized learning factors analysis: Improving cognitive models with machine learning[D]. Carnegie Mellon University.
CHRYSAFIADI K, VIRVOU M, 2013. Student modeling approaches: A literature review for the last decade[J]. Expert Systems with Applications, 40(11): 4715–4729.
CONATI C, BARRAL O, PUTNAM V, et al., 2021. Toward personalized xAI: A case study in intelligent tutoring systems[J]. Artificial Intelligence, 298: 103503.
CONFALONIERI R, COBA L, WAGNER B, et al., 2021. A historical perspective of explainable Artificial Intelligence[J]. WIREs Data Mining and Knowledge Discovery, 11 (1): e1391.
CORBETT A T, ANDERSON J R, 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge[J]. User Modeling and User–Adapted Interaction, 4(4): 253–278.
CUI P, ATHEY S, 2022. Stable learning establishes some common ground between causal inference and machine learning[J]. Nature Machine Intelligence, 4(2): 110–115.
DUGGAN T, CORPORATION T, 2020. AI in education: Change at the speed of learning[EB/OL].[2022–11–29]. https://iite.unesco.org/wp–content/uploads/2021/05/Steven_Duggan_AI–in–Education_2020–2. pdf.
GUNNING D, 2017. Explainable artificial intelligence (xAI)[R]. Arlington: Defense Advanced Research Projects Agency (DARPA).
GUNNING D, STEFIK M, CHOI J, et al. , 2019. xAI–Explainable artificial intelligence[J]. Science Robotics, 4(37): eaay 7120.
GUNNING D, VORM E, WANG J Y, et al., 2021. DARPA's explainable AI(xAI) program: A retrospective[J]. Applied AI Letters, 2(4): e61.
JIANG B, ZHAO W, GU X, et al., 2021. Understanding the relationship between computational thinking and computational participation: A case study from Scratch online community[J]. Educational Technology Research and Development, 69.
KHAJAH M, LINDSEY R V, MOZER M C, 2016. How deep is knowledge tracing?[J/OL].[2022–12–28]. https://arxiv.org/pdf/1604.02416.pdf.
KHOSRAVI H, BUCKINGHAM SHUM S, CHEN G, et al., 2022. Explainable artificial intelligence in education[J]. Computers and Education:Artificial Intelligence,3: 100074.
KOEDINGER K R, KIM J, JIA J Z, et al., 2015. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC[C]//Association for Computing Machinery.Proceedings of the Second (2015)ACM Conference on Learning @ Scale, Vancouver, BC, Canada: 111–120.
LIU H, WANG Y, FAN W, et al., 2022. Trustworthy AI: A computational perspective[J].ACM Transactions on Intelligent Systems and Technology (TIST).
LIU Q, HUANG Z, YIN Y, et al., 2019. Ekt: Exercise–aware knowledge tracing for student performance prediction[J]. IEEE Transactions on Knowledge and Data Engineering, 33 (1): 100–115.
MACLELLAN C, LIU R, KOEDINGER K, 2015. Accounting for slipping and other false negatives in logistic models of student learning[C]//The 8th International Conference on Educational Data Mining. Madrid, Spain.
MARTINEZ H P, BENGIO Y, YANNAKAKIS G N, 2013. Learning deep physiological models of affect[J]. IEEE Computational Intelligence Magazine, 8(2):20–33.
MATTHEWS R, 2000. Storks deliver babies(p=0.008)[J]. Teaching Statistics, 22(2): 36–38.
NAKAGAWA H, IWASAWA Y, MATSUO Y, 2019. Graph–based knowledge tracing: Modeling student proficiency us-

ing graph neural network[C]//Proceedings of the 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI). Thessaloniki, Greece.

NICOLOSI S L, 1988. Sophie: An expert system for the selection of Hazard evaluation procedures [M].Artificial Intelligence and Other Innovative Computer Applications in the Nuclear Industry. Boston, MA: Springer US: 817–818.

PANDEY S, KARYPIS G, 2019. A self–attentive model for knowledge tracing[J]. ArXiv Preprint ArXiv: 190706837.

PARDOS Z A, HEFFERNAN N T, 2010. Modeling individualization in a bayesian networks implementation of knowledge Tracing[C]// Proceedings of User Modeling, Adaptation & Personalization, International Conference, Umap, Big Island, HI, USA.

PARDOS Z A, HEFFERNAN N T, 2011. KT–IDEM: Introducing item difficulty to the knowledge tracing model[C]//Proceedings of the 19th international conference on User modeling, adaption, and personalization, Girona, Spain.

PAVLIK JR P I, CEN H, KOEDINGER K R, 2009. Performance factors analysis: A new alternative to knowledge tracing[M]// DIMITROVA V, MIZOGUCHI R,BOULAY B D, et al. (Eds.), 2009. Artificial Intelligence in Education. Brighton, England: IOS Press:531–538.

PEARL J, MACKENZIE D, 2018. The book of why: The new science of cause and effect[J]. Journal of MultiDisciplinary Evaluation,14(31):47–54.

PIECH C, SPENCER J, HUANG J, et al., 2015. Deep knowledge tracing[J]. Computer Science, 3(3):19–23.

QIU Y, QI Y, LU H, et al., 2011. Does time matter? Modeling the effect of time with bayesian knowledge tracing[C]//Proceedings of the EDM 2011–Proceedings of the 4th International Conference on Educational Data Mining.

ROSÉ C P, MCLAUGHLIN E A, LIU R, et al., 2019. Explanatory learner models: Why machine learning (alone) is not the answer[J]. British Journal of Educational Technology, 50(6): 2943–2958.

ROUAST P V, ADAM M T, CHIONG R, 2019. Deep learning for human affect recognition: Insights and new developments[J]. IEEE Transactions on Affective Computing, 12(2): 524–543.

SABLAYROLLES L, LEFEVRE M, GUIN N, et al., 2022. Design and evaluation of a competency: Based recommendation process[C]//Proceedings of the Intelligent Tutoring Systems, Cham. Springer International Publishing.

SPAULDING S, BREAZEAL C, 2015. Affect and inference in bayesian knowledge tracing with a robot tutor[C]//Proceedings of the Tenth Annual ACM/IEEE International Conference on Human–Robot Interaction.

Xu D, Wang H, Su K, 2002. Intelligent student profiling with fuzzy models[C]// System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on. IEEE, 2002.

YUDELSON M V, KOEDINGER K R, GORDON G J,2013. Individualized bayesian knowledge tracing models[C]//International Conference on Artificial Intelligence in Education. Springer, Berlin, Heidelberg: 171–180.

ZHANG J, SHI X, KING I, et al., 2017. Dynamic key–value memory networks for knowledge tracing[C]// Proceedings of the 26th International Conference on World Wide Web: 765–774.

收稿日期:2022 年 12 月 8 日
责任编辑:陈　媛

Explainable Learner Models: Technical Keys to Trustworthy Personalized Learning

Jiang Bo, Ding Yingwen & Wei Yuang

【Abstract】 Personalized learning technology based on artificial intelligence has always been a research hotspot in the field of intelligent education, and its technical challenge lies in how to build a comprehensive and accurate learner model. In the field of education, where fairness, ethics, and accountability are very much in focus, the "black box" nature of AI may affect the user's trust in the machine's decision, so building transparent and explainable learner models is particularly important.By clarifying the characteristics, structure and decision result of learner models, the education stakeholders can understand its motivation and accept its decision, so as to achieve better human–machine cooperation. This paper extends the technical concept of explainable artificial intelligence to personalized learning. By analyzing its research status, this paper clarifies the necessity of realizing explainable learner model, analyzes the technical principles of existing examples of explainable learner model, and finally proposes the basic framework of explainable learner model, aiming to take explainability as the key principle of learner modeling. The research on the explainable learner models can provide reference for the whole cycle of the design, development, application and evaluation of personalized learning system to realize the explainability, and lay a foundation for the realization of trustworthy personalized learning.

【Keywords】 Learner Model; Personalized Learning; Explainability; Trustworthy Artificial Intelligence