软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(11):5126−5142 [doi: 10.13328/j.cnki.jos.006724]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

多知识点融合嵌入的深度知识追踪模型*

琚生根, 康 睿, 赵容梅, 孙界平

(四川大学 计算机学院, 四川 成都 610065)

通信作者: 孙界平, E-mail: sunjieping@scu.edu.cn

摘 要: 知识追踪任务是根据学生历史答题记录追踪学生知识状态的变化, 预测学生未来的答题情况. 近年来, 基于注意力机制的知识追踪模型在灵活性和预测性能上都明显优于传统知识追踪模型. 但是现有深度模型大多只考虑了单一知识点题目的情况, 无法直接处理多知识点题目, 而智能教育系统中存在着大量的多知识点题目. 此外, 如何提高可解释性是深度知识追踪模型的关键挑战之一. 为了解决这些问题, 提出一种多知识点融合嵌入的深度知识追踪模型. 所提模型考虑涉及多知识点的题目中知识点之间的关系, 提出两种新颖的多知识点嵌入方式, 并且结合教育心理学模型和遗忘因素提升预测性能和可解释性. 实验表明所提模型在大规模真实数据集上预测性能上优于现有模型, 并验证各个模块的有效性.

关键词: 教育数据挖掘; 知识追踪; 注意力机制; 深度神经网络

中图法分类号: TP18

中文引用格式: 琚生根, 康睿, 赵容梅, 孙界平. 多知识点融合嵌入的深度知识追踪模型. 软件学报, 2023, 34(11): 5126–5142. http://www.jos.org.cn/1000-9825/6724.htm

英文引用格式: Ju SG, Kang R, Zhao RM, Sun JP. Deep Knowledge Tracing Model Based on Embedding of Fused Multiple Concepts. Ruan Jian Xue Bao/Journal of Software, 2023, 34(11): 5126–5142 (in Chinese). http://www.jos.org.cn/1000-9825/6724.htm

Deep Knowledge Tracing Model Based on Embedding of Fused Multiple Concepts

JU Sheng-Gen, KANG Rui, ZHAO Rong-Mei, SUN Jie-Ping

(College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: The task of knowledge tracing is to trace the changes in students' knowledge state and predict their future performance in learning according to their historical learning records. In recent years, knowledge tracing models based on attention mechanisms are markedly superior to traditional knowledge tracing models in both flexibility and prediction performance. Only taking into account exercises involving single concept, most of the existing deep models cannot directly deal with exercises involving multiple concepts, which are, nevertheless, vast in intelligent education systems. In addition, how to improve interpretability is one of the key challenges facing deep knowledge tracing models. To solve the above problems, this study proposes a deep knowledge tracing model based on the embedding off used multiple concepts that considers the relationships among the concepts in exercises involving multiple concepts. Furthermore, the study puts forward two novel embedding methods for multiple concepts and combines educational psychology models with forgetting factors to improve prediction performance and interpretability. Experiments reveal the superiority of the proposed model over existing models in prediction performance on large-scale real datasets and verify the effectiveness of each module of the proposed model.

Key words: educational data mining; knowledge tracing (KT); attention mechanism; deep neural network (DNN)

智能教育系统如 Coursera、Udacity 等平台提供了大量的在线课程和练习, 吸引了越来越多的学习者和教学工作者. 此类智能教育系统收集了大量的学习者详细的学习行为信息, 这些信息对于教育学研究来说有着重要意

* 基金项目: 国家自然科学基金 (62137001)
收稿时间: 2022-02-21; 修改时间: 2022-04-20, 2022-05-25; 采用时间: 2022-06-12; jos 在线出版时间: 2023-04-27
CNKI 网络首发时间: 2023-04-28

义. 从教育服务角度来说, 研究人员可以基于这些信息, 利用计算机技术、人工智能技术等为学习者提供个性化导学服务.

知识追踪 (knowledge tracing, KT) 是个性化导学中的关键任务[1], 是教育心理学和数据挖掘领域的研究热点之一. 知识追踪任务是对学生知识状态进行建模, 换言之就是根据学生的历史学习轨迹和其他相关信息来追踪学生的知识状态随时间的变化, 预测学习者在未来的学习互动中将如何表现, 通常是用于预测回答准确率. 此外还可以帮助实现个性化题目推荐等服务. 在计算机技术、深度学习技术等支持下, 搭建大型个性化智能教育系统是未来教育和教育数据挖掘的重要任务和未来发展的必然趋势. 作为个性化导学中的关键问题, KT 任务不仅注重结果, 即预测的准确性, 还注重可解释性.

知识追踪模型于 1972 年由 Pardos 等人首次提出, 由 Corbett 等人引入智能教育领域[2,3], 应用于智能教育系统. 早期研究工作主要依赖于一阶马尔可夫模型, 其中贝叶斯知识追踪是经典代表模型, 也是研究最多的模型之一. 这种基于贝叶斯的方法指导了这个领域长达十余年之久[4], 此外还提出了很多相关模型如基于矩阵分解的知识追踪模型等. 随着深度学习在自然语言处理、数据挖掘、计算机视觉等领域的应用取得了巨大成果, Piech 等人首次将深度学习技术引入 KT 领域, 提出了基于循环神经网络的深度知识追踪 (deep knowledge tracing, DKT)[5]. 受 DKT 的成功启发, 研究人员基于深度学习技术提出了基于记忆网络的 KT 模型[6]等多种模型. 一些学者[7,8]还提出融入遗忘行为来提升预测性能. 近年来, 受 Transformer[9]在自然语言处理任务上的出色表现启发, 一些学者提出了基于注意力机制的 KT 模型[10,11], 相较之前的深度知识追踪模型表现了更好的预测性能. Su 等人提出的 EERNN 是 KT 领域第 1 次引入注意力机制, 通过考虑题目相似性来帮助预测学生未来表现[12]. Choi 等人提出的 SAINT[10] 是第 1 个具有编码器-解码器结构的知识追踪模型, 对 Transformer 结构进行改进以适应 KT 任务, 有效地捕捉了习题和回答之间复杂的关系. AKT 使用单调注意力机制对学生的遗忘行为进行建模, 并使用 IRT 对题目进行建模, 相比以往模型有了很大的改进[13]. Pandey 等人提出的关系感知的自注意力知识追踪模型[14]尝试用核函数对学生遗忘行为建模. 对于基于深度学习的知识追踪方法, 如何对题目建模以及如何提高其可解释性, 一直是研究者关注的重中之重.

现有主流的题目建模方式分为两种: 一种是传统的以题目为中心的建模, 这种方式带来了过度参数化的问题; 另一种以知识点为核心的题目建模考虑了知识点的重要性, 解决了过度参数化等问题, 但是绝大多数工作忽略了多知识点的情况. 然而, 在实际智能教育系统中存在着大量的综合性题目, 这些题目往往涉及多个知识点. 因此, 如何处理多知识点题目是实际应用中迫切需要解决的问题. 此外, 虽然现有一些研究工作考虑了多知识点的情况, 对多个知识点信息采取平均池化的方式进行融合[15], 但忽略了题目与所涉及的多个知识点之间的关系.

知识追踪任务中的可解释性主要是指对预测过程或者结果的可解释性, 以便其能够在实际教学中提供服务, 例如学习能力分析、题目推荐等下游应用. 传统 KT 模型如贝叶斯模型通过专家设计的函数能够从用户猜测、学习等方面对用户水平进行建模, 从而提供可解释的预测结果. 然而, 深度 KT 模型普遍存在可解释性不足的问题, 主要原因是很难将深度模型的抽象决策映射到最终用户能够容易理解的目标领域[16]. 虽然一些工作[11,13,14]通过引入教育心理学理论对改进对题目的建模取得了一定成效, 但是这种改进不能满足教育领域中对归因分析的需求. 还有一部分研究者结合深度方法改进传统认知诊断模型, 但是在捕捉交互中复杂关系的能力不足.

针对上述问题, 本文提出了多知识点融合嵌入的深度知识追踪模型 (MCAKT). 本文工作的主要贡献和创新如下.

(1) 针对题目涉及多个知识点的情况, 本文提出了两种新的多知识点嵌入融合方式, 即基于注意力机制和基于 SENet 的多知识点融合方式; 以及一种多知识点融合的题目嵌入方式, 引入了题目相关特征以保留题目本身信息.

(2) 本文考虑随时间的遗忘行为对学生知识状态的影响对传统认知诊断模型进行改进, 用于预测学生的答题情况, 通过对学生能力水平的追踪建模增强预测性能和可解释性. 此外, 本文对基于 Transformer 的编码器结构进行修改以适用于知识追踪任务.

(3) 在两个真实的在线教育数据集上进行了一系列实验. 实验表明在大规模数据集上, MCAKT 预测性能优于现有工作.

1 相关工作

现有知识追踪方法主要可以分为 3 类: 认知诊断模型、传统知识追踪模型和基于深度学习的知识追踪模型.

传统认知诊断模型随教育测量领域的发展而被提出, 其核心思想源自一些经典测验理论. 项目反应理论 (item response theory, IRT) 是用于评估被试者对某一项目或某类项目的潜在特质, 引入知识追踪领域后用于评估学习者对知识点的掌握程度[1]. 其中最具代表性也最常使用的是双参数模型 (two parameter logistic model, 2PL):

$$P(\theta)=\frac{1}{1+\mathrm{e}^{-1.7a(\theta-b)}} \tag{1}$$

其中, θ 代表被试者也就是答题者对于题目考核内容的掌握程度, a 代表题目的区分度, b 代表题目的难度. 近年来, 还有一些学者提出结合深度神经网络来学习 IRT 模型参数的框架[16,17]. 虽然认知诊断模型拥有很高的可解释性, 但是在捕捉智能教育系统内学生交互中复杂的关系的能力上还有所不足.

传统知识追踪模型中, 贝叶斯知识追踪 (Bayesian knowledge tracing, BKT)[3]是最流行的 KT 模型之一. 这种模型在提出之后的很长一段时间在 KT 领域都具有主导地位[4]. BKT 将学生知识掌握情况用一组二进制变量表示, 每个变量分别表示对某一个知识点是否掌握. BKT 基于隐马尔可夫模型更新学生知识状态, 进而预测学生的表现. 这使得贝叶斯模型能够拥有比较好的可解释性. 但是, BKT 的泛化和扩展能力有限, 比如传统 BKT 的建模只针对某一个知识点. 对于新特征需要专家进行设计, 不能自动理解利用. 此外, BKT 也没有考虑遗忘的情况, 即当学习者掌握了某知识点之后就不会遗忘, 其对应的变量不再更新. 在这之后, 研究人员也不断开发了基于 BKT 的多个变体, 比如引入学习者参数或者知识信息等[18−20].

深度学习模型具有更高的通用性和灵活性, 因为它不需要专家设计的函数来模拟学生表现和知识状态之间的关系. Piech 等人提出了 DKT[5], 使用 RNN 对学习者的学习过程进行建模. 自 DKT 被提出后, 更多深度学习技术应用于 KT 领域中. 在深度学习的发展中, 注意力机制最初被用于机器翻译, 近年来在各个领域取得了不错的成绩. 依赖于注意力机制的 Transformer 模型在特征提取和依赖捕获方面具有优越的能力, 同时保持较高的计算效率[9]. 一些基于注意力机制和基于 Transformer 结构的知识追踪模型展现出了优异的性能.

值得注意的是, 在深度知识追踪模型中, 主流的题目建模方法分为两种. 第 1 种是以题目为中心的建模, 但这种方法存在数据稀疏和过度参数化问题. 因为在智能教育系统中一个学生往往只与小部分的题目有过交互, 而且相比相对固定的知识点, 题目数量往往在不断增加. 因此, 第 2 种题目建模方法以知识点为中心, 考虑了知识点中蕴含着至关重要的信息. 一些工作简单地用知识点代替题目, 但这种解决方法忽略了题目与知识点、知识点之间复杂的关系以及题目其他特征. 而题目其他特征能有效帮助预测, 例如两道同样涉及一元二次方程的题目在难度上的不同会导致在同一时刻同一学生的答题表现很可能截然相反. 针对这一问题, 有一些工作提出了解决思路. EERNN 采用题目文本来获取题目嵌入[12], 但文本描述在实际工作中不易获得. Minn 等人[21]引入题目难度来区分不同的问题. 但现有工作多知识点的题目建模还未充分考虑, 而此类题目是教学中十分常见的题目类型. 实际上, 已有传统模型[22]考虑了多知识点题目, 扩展了贝叶斯模型的多个参数以便直接适用多知识点题目. 但是在深度 KT 模型中大多受限于结构, 对于多知识点题目往往采用将一条多知识点题目交互记录拆分为多条单一知识点交互记录或者视为新知识点[23]. 这种处理方式不仅不能体现题目对于多知识点之间的关系, 还会造成数据冗余. 还有一些深度模型对多知识点信息进行平均池化[15], 但如何体现知识点与题目的不同关系仍是一个亟待解决的问题. 除此之外, 深度模型虽然表示出良好的性能, 但在可解释性上仍不如传统认知诊断模型和贝叶斯模型. 这是由于深度模型具有“黑箱”特性, 不如此类传统模型结构本身具有可解释性. 一些深度模型通过结合 IRT 对题目进行建模从而获得有意义的题目表示, 但是如何对预测过程和结果进行解释仍是知识追踪领域的研究重点.

为了解决上述问题, 本文提出了一种多知识点融合的题目建模方式, 并且结合 IRT 和遗忘因素提高模型可解释性.

2 多知识点融合嵌入的深度知识追踪模型

本节主要给出知识追踪任务的形式化定义, 并且详细介绍本文所提出的多知识点融合嵌入的深度知识追踪模

型 MCAKT. 图 1 是 MCAKT 的总体框架, 由 3 个部分构成, 分别是多知识点融合的题目嵌入、基于注意力机制的历史状态编码器和基于 IRT 和遗忘因子的预测层. 首先, 为了构造多知识点融合的题目嵌入, 提出了两种多知识点嵌入融合的方式, 即基于注意力机制的融合方式和基于 SENet 的融合方式. 其次, 在基于注意力机制的历史状态编码器模块中, 使用改进后的 Transformer 编码器结构对历史交互嵌入进行编码. 最后, 为了增强模型的可解释性, 提出了基于 IRT 和遗忘因子的预测层, 做出最终的预测.

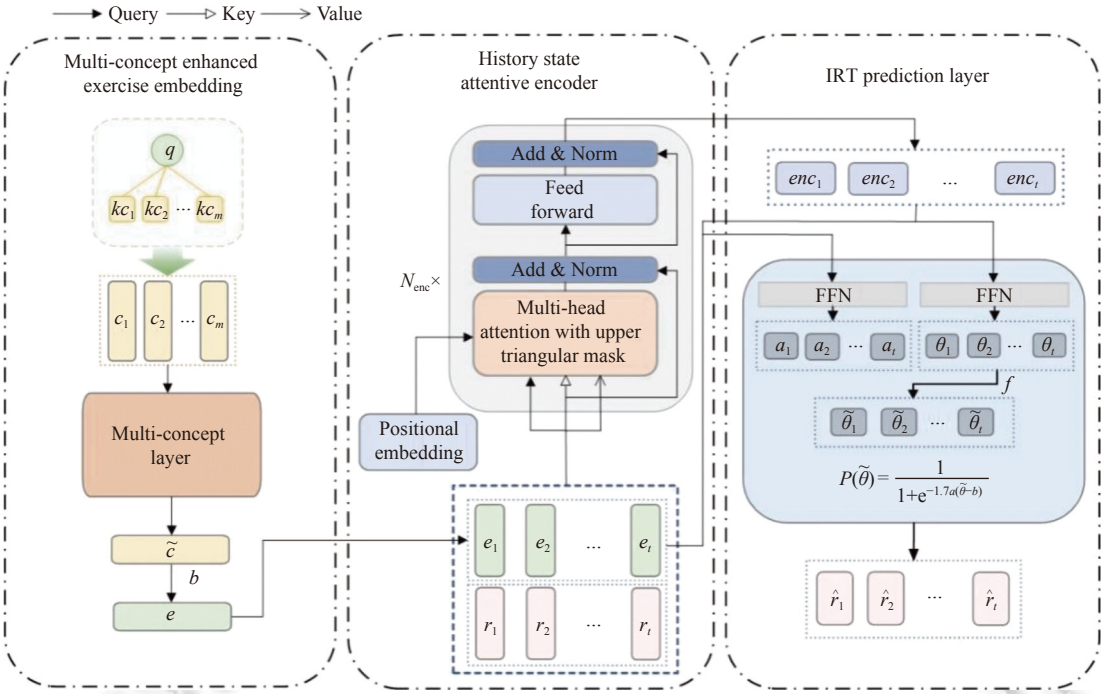


图 1 MCAKT 框架

2.1 问题定义

在智能教育系统中, 用户会利用系统提供的题目进行练习, 而每个用户的历史学习轨迹都会被系统收集记录. 用户在一段时间内的历史学习轨迹由该用户在该时间内的所有交互序列组成, 其中交互通常使用题目问答交互来表示. 知识追踪任务是根据学生的历史学习轨迹来追踪学生的知识状态, 以便能够准确地预测学生在未来的学习中的表现. 在本文中, 本文定义 Q 为题目集合, KC 为知识点集合, $kc_q \subseteq KC$ 表示题目 $q \in Q$ 涉及的知识点集合. $X_{t-1}=\{x_1,x_2,\ldots,x_{t-1}\}$ 为一个学生截止 t 时刻的历史交互记录, 其中 $x_t=(q_t,r_t)$ 代表该学生在 t 时刻的问答交互, q_t 为该生在 t 时刻回答的题目, $r_t \in \{0,1\}$ 表示该生回答该题正确与否. 其中, 本文引入题目难度系数 d 作为题目特征, 即题目 q 可以表示为 $q=\left(kc_q,d_q\right)$, 其中 d_q 表示题目 q 的难度系数. 因此, 知识追踪任务目标可以总结为两个方面:

(1) 预测学生的未来表现, 即该生在 t 时刻回答题目 q_t 正确概率 $P(r_t=1 \mid X_{t-1},q_t)$.

(2) 提供可解释性, 即通过追踪学生知识水平的变化等方式解释预测结果或过程.

此外, 学生的知识水平和学习表现都会受各方面的影响. 第 1, 在同一时刻下, 学生在不同题目的表现是不同的. 这是因为即使在知识状态一样的条件下, 不同题目考察的知识是有所不同的; 同样, 面对同一道题目, 学生在不同时刻的表现也有所不同. 第 2, 根据心理学家 Ebbinghaus 提出的遗忘曲线[24]描述的人对知识随时间遗忘的规律, 体现了随时间造成的遗忘对学生掌握情况的影响. 基于上述两点, 本文提出以下两个合理假设.

(1) 学生对题目考核内容的掌握程度与所测题目以及历史学习情况有关.

(2) 学生的掌握程度会随与上次练习之间的间隔时间增长而逐渐衰减.

2.2 多知识点融合的题目嵌入

考虑实际中多知识点题目的情况, 本文提出使用知识点与题目特征结合构造多知识点融合的题目嵌入方式. 对于题目涉及的多个知识点, 本文通过构建融合了多个知识点信息的低维向量来将这些丰富的信息同时融入题目嵌入. 受特征交叉研究工作启发, 本文提出了基于注意力机制的多知识点融合方式和基于 SENet 的多知识点融合方式.

2.2.1 基于注意力机制的多知识点融合方式

考虑题目往往涉及多个知识点的情况, 如何捕捉多知识点之间复杂的高阶关系是一大挑战. 而关于解决如何构建高阶关系的问题, AutoInt[25]提供了一种基于多头注意力机制的方法, 该工作将不同域中的特征嵌入映射到多个空间中进行特征交叉. 受 AutoInt 的成功启发, 本文提出如图 2 的基于注意力机制的多知识点嵌入融合机制, 利用多头注意力机制计算知识点间相似性来构造不同子空间中的特征组合, 并且通过堆叠这种注意力层来实现捕捉高阶特征.

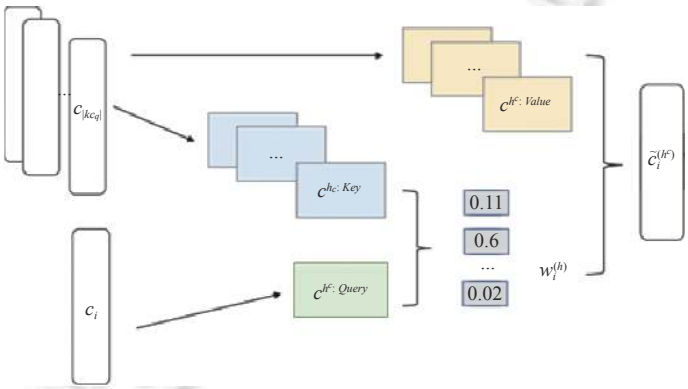


图 2 基于注意力机制的多知识点融合嵌入

首先, 通过与知识点嵌入矩阵 $M^{KC}\in\mathbb{R}^{|KC|\times dim}$ 相乘, 题目 q 涉及的 KC 中所有知识点能够分别得到对应的低维表示 $c\in\mathbb{R}^{dim}$, 即得到题目 q 对应的一组知识点嵌入 $\{c_1,c_2,\ldots,c_{|kc_q|}\}$. 然后, 将该组知识点嵌入作为多知识点融合层的输入, 得到最终的融合了题目多个知识点信息的多知识点嵌入 $\tilde{c}\in\mathbb{R}^{dim}$.

多知识点嵌入融合层由多个多头注意力层堆叠构成, 通过这种堆叠来实现不同阶数的特征组合. 图 2 为 c 在子空间 h^c 中的表示的计算示意. 先将每个知识点嵌入映射到查询、键和值中分别得到 3 个表示向量 $c^{h^c:Query}$、$c^{h^c:Key}$ 以及 $c^{h^c:Value}$, 并计算每个知识点嵌入 c_i 与其他知识点嵌入的相关程度:

$$\varphi^{(h^c)}\left(c_i,c_j\right)=<c_i^{h^c:Query},c_j^{h^c:Key}> \tag{2}$$

$$w_{ij}^{(h^c)}=\frac{\exp\left(\varphi^{(h^c)}\left(c_i,c_j\right)\right)}{\sum_{l=1}^{|kc_q|}\exp\left(\varphi^{(h^c)}\left(e_i,e_l\right)\right)} \tag{3}$$

其中, $\varphi(\cdot,\cdot)$ 通过计算向量内积来定义两者之间的相似性. 然后通过得到的权重对 c_i 映射到 $Value$ 空间的表示进行加权求和得到 c_i 在子空间 h^c 中的表示 $\tilde{c}_i^{(h^c)}\in\mathbb{R}^{dim/|H^c|}$, 其中, $|H^c|$ 表示注意力机制下多知识点融合方式中的子空间集合.

$$\tilde{c}_i^{(h^c)}=\sum_{j=1}^{|kc_q|}w_{ij}^{(h^c)}c^{h^c:Value} \tag{4}$$

本文将所有子空间下的表示进行连接:

$$\tilde{c}_i=\tilde{c}_i^{(h_1^c)}\oplus\tilde{c}_i^{(h_2^c)}\oplus\ldots\oplus\tilde{c}_i^{(h_{|H^c|}^c)} \tag{5}$$

为了显式建模低阶特征[25], 本文使用残差连接保留知识点原始信息减少了信息损失:

$$\tilde{c}_i^{Res} = ReLU\left(\tilde{c}_i + W_{Res}^c c_i\right) \tag{6}$$

其中, W_{Res}^c 是与 c_i 维度相匹配的投影矩阵. 本文可以用前一层的输出作为下一层的输入来堆叠多个这样的层, 从而捕捉知识点之间的高阶关系. 最后, 本文对题目的每个知识点表示 $\tilde{c}_i$ 相加得到最终的多知识点融合嵌入 $\tilde{c}$:

$$\tilde{c} = \sum_{i=1}^{|kc_q|} \tilde{c}_i^{Res} \tag{7}$$

2.2.2 基于 SENet 的多知识点融合方式

不同于基于多头注意力来捕捉高阶信息, FiBiNET[26]强调特征本身的重要程度, 动态学习不同特征的重要性权重并进行特征交互. 受 FiBiNET[26]的工作启发, 本文利用 SENet[27]模块通过学习不同知识点嵌入的权重, 对多个单一知识点嵌入进行融合, 以便强化题目中重要知识点的影响. 具体如图 3 所示, 由压缩和激励两个阶段组成[26], 在压缩阶段对每个知识点的表示进行数据压缩和信息提取, 在激励阶段学习知识点相应权重, 通过加权的方式融合不同重要程度的知识点信息.

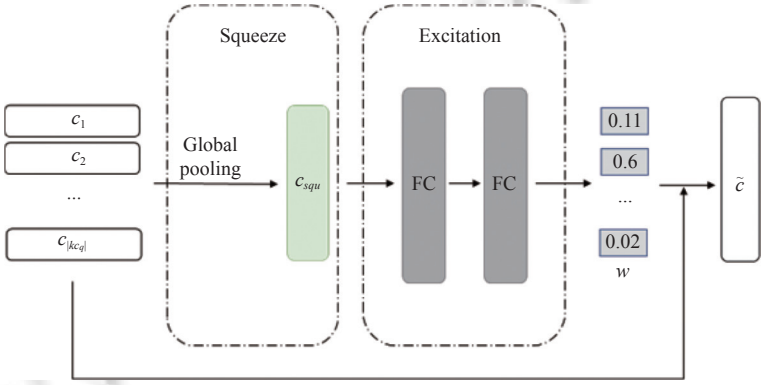


图 3 基于 SENet 的多知识点融合嵌入

首先, 在压缩阶段对题目 q 对应的知识点嵌入 $\{c_1, c_2, \ldots, c_{|kc_q|}\}$ 进行压缩得到知识表示 $c_{squ} \in \mathbb{R}^{|kc_q|}$:

$$c_{squ}^i = \frac{1}{dim} \sum_j^{dim} c_i^{(j)} \tag{8}$$

$$c_{squ} = c_{squ}^1 \oplus c_{squ}^2 \oplus \ldots \oplus c_{squ}^{|kc_q|} \tag{9}$$

具体来说, 通过对每个知识点 c_i 进行平均池化得到表示该知识点信息的标量 c_{squ}^i, 从而得到压缩后蕴含所有知识点信息的知识表示 c_{squ}. 其次在激励阶段, 将 c_{squ} 作为两层全连接层输入, 其中第 1 层对维度进行缩减进行特征交叉, 第 2 层对维度进行扩展得到每个知识点嵌入的权重, 进行重新加权得到最终的多知识点融合嵌入 $\tilde{c}$:

$$w^{\text{SENet}} = Sigmoid\left(W_2^{ex}\left(W_1^{ex} c_{squ}\right)\right) \tag{10}$$

$$\tilde{c} = \sum_{i=1}^{|kc_q|} w_i^{\text{SENet}} c_i \tag{11}$$

其中, w_i^{SENet} 表示权重向量 w^{SENet} 中的第 i 个元素; $W_1^{ex} \in \mathbb{R}^{|kc_q| \times \frac{|kc_q|}{2}}$、$W_2^{ex} \in \mathbb{R}^{\frac{|kc_q|}{2} \times |kc_q|}$.

2.2.3 题目嵌入

虽然知识点蕴含了题目中最重要的信息, 但是不能为了解决数据稀疏问题而忽略题目本身的特征. 题目本身特征体现了相同知识点下题目之间的区别. 因此, 不同于直接进行题目嵌入, 本文提出一种多知识点融合的题目嵌入构造方式, 在解决数据稀疏性问题同时结合题目的其他特征. 本文引入题目难度系数 $d \in [0,1]$ 作为题目特征, 定义如下:

$$d = \frac{N_{\text{fail}}}{N_{\text{ans}}} \tag{12}$$

其中, N_{ans} 表示回答了题目的学生总数, N_{fail} 表示没有答对该题的人数. 题目难度系数越高, 代表题目越难. 本文将难度系数作为输入, 通过线性变换得到难度因子 b:

$$b = W_2^b(W_1^b d + b_1^b) + b_2^b \tag{13}$$

其中, $W_1^b \in \mathbb{R}^{1\times dim}$ 、 $W_2^b \in \mathbb{R}^{dim\times 1}$ 、 $b_1^b \in \mathbb{R}^{dim}$ 、 $b_2^b \in \mathbb{R}^1$. 本文提出题目嵌入 $e \in \mathbb{R}^{dim}$ 构造方式如下:

$$e = \tilde{c} + b \tag{14}$$

2.3 基于注意力机制的历史状态编码器

基于注意力机制的历史状态编码器整体结构基于 Transformer 中的编码器, 并且为了适应于知识追踪任务, 挖掘历史交互中潜在的复杂关系, 对自注意力机制进行了一些调整, 以便对历史状态进行编码. 首先, 正如 LANA[28] 以及其他研究[29]认为, 对于深度的模型往往会对第 1 层输入的信息有所损失, 本文将表示序列信息的位置嵌入直接输入到注意力模块中. 其次, 对于知识追踪任务来说, 学生的历史交互序列是一个时间序列, 意味着未来的交互不会对以前的交互有所影响. 因此, 在注意力层需要通过一个上三角的掩码防止未来信息的泄露.

具体来说, 首先对输入的题目嵌入序列 $E = \{e_1, e_2, \ldots, e_t\}$ 与回答嵌入序列 $R = \{r_{emb}^1, r_{emb}^2, \ldots, r_{emb}^t\}$ 依次连接, 作为注意力模块的键和值的输入, 其中 r_{emb}^i 是题目 e_i 对应回答 r_i 通过回答嵌入矩阵 $M^R \in \mathbb{R}^{3\times dim}$ 得到的回答嵌入. 值得注意的是, 为了防止未来信息泄露, 将题目序列与 $r_{\text{padding}} \in \mathbb{R}^{dim}$ 进行拼接作为查询的输入, 其中 r_{padding} 是通过 M^R 得到的填充嵌入. 编码器对题目-回答嵌入序列 $\{I_1, I_2, \ldots, I_t\}$ 进行编码, 首先通过 $W^Q \in \mathbb{R}^{2dim\times dim_q}$ 、 $W^K \in \mathbb{R}^{2dim\times dim_k}$ 和 $W^V \in \mathbb{R}^{2dim\times dim_v}$ 进行如下计算, 其中 dim_q 、 dim_k 和 dim_v 分别代表投影到查询、键和值空间后的向量维度.

$$I^r = e \oplus r_{emb} \tag{15}$$

$$I^{wor} = e \oplus r_{\text{padding}} \tag{16}$$

$$I^Q = I^{wor}W^Q,\ I^K = I^rW^k,\ I^V = I^rW^V \tag{17}$$

学生的历史交互通常以序列的形式展现, 本文为了融入交互间的序列信息, 不同于 Transformer 中的位置编码, 本文使用可学习的位置嵌入使模型能够像 RNN 一样对序列的顺序进行编码. 序列中每个交互的序列表示 p 通过位置嵌入矩阵 $M^P \in \mathbb{R}^{len\times dim}$ 获得, 其中 len 表示序列长度. 将相应的位置嵌入 p 分别通过线性变换得到 $p^Q \in \mathbb{R}^{dim_q}$ 、 $p^K \in \mathbb{R}^{dim_k}$ 输入注意力模块, 并且对注意力权重的计算进行修改:

$$\begin{cases} \alpha = Softmax\left(Mask\left(\dfrac{I^QI^{K^\top}}{\sqrt{dim_k}}\right)\right) + \alpha_p \\ \alpha_p = Softmax\left(\dfrac{p^{Q_p}p^{K_p^\top}}{\sqrt{dim_k}}\right) \end{cases} \tag{18}$$

其中, $Mask(\cdot)$ 代表上三角矩阵掩码. 根据多头注意力机制得到输出如下:

$$head = \alpha I^V \tag{19}$$

$$\begin{cases} heads = head^{(h_1^e)} \oplus head^{(h_2^e)} \oplus \ldots \oplus head^{(h_{|H^e|}^e)} \\ MHead^{\text{out}} = headsW^{\text{out}} \end{cases} \tag{20}$$

其中, H^e 表示该多头注意力下的子空间集合. 参考 Transformer[9]的工作, 本文将输入全连接前馈网络, 并且在注意力模块和前馈网络之后都采用残差连接和层归一化机制. 值得注意的是, 历史状态编码器由 N_{enc} 个上述结构的编码器堆叠组成, 每一层经全连接前馈网络的输出都会作为下一层编码器的输入. 最后, 经多层堆叠的历史状态编码器得到对应每个时刻的状态编码 $\{enc_1, enc_2, \ldots, enc_t\}$.

2.4 基于 IRT 模型和遗忘因子的预测层

可解释性不足是大多数基于深度学习的 KT 模型存在的问题, 其中一个解决方案是将心理测量模型和深层网络相结合. 本文利用神经网络试图学习 IRT 模型中可解释的参数, 结合遗忘因素对学生掌握程度的影响, 对 2PL 模型进行修改并用于最终的预测.

首先, 将题目嵌入通过线性变换和非线性激活函数得到题目的区分度参数 a:

$$a = ReLU(W^a(e) + b^a) \tag{21}$$

其中, $W^a \in \mathbb{R}^{dim\times 1}$、$b^a \in \mathbb{R}^1$.

根据假设 1: 学生对题目考核内容的掌握程度与所测题目以及历史学习情况有关, 本文提出根据历史状态 enc 和题目嵌入 e 计算学生掌握程度的方式. 具体来说, 历史状态 enc 和题目嵌入 e 作为输入, 通过全连接前馈神经网络学习学生每个时刻的掌握程度 $\{\theta_1,\theta_2,\ldots,\theta_t\}$, 对于其中每个标量元素 θ 计算如下:

$$\theta = W_2^{\Theta}\left(ReLU\left(W_1^{\Theta}(enc\oplus e)+b_1^{\Theta}\right)\right)+b_2^{\Theta} \tag{22}$$

其中, $W_1^{\Theta} \in \mathbb{R}^{3dim\times dim}$, $W_2^{\Theta} \in \mathbb{R}^{dim\times 1}$, $b_1^{\Theta} \in \mathbb{R}^{dim}$, $b_2^{\Theta} \in \mathbb{R}^1$.

根据假设 2: 学生的掌握程度会随与上次练习之间间隔时间的增长而逐渐衰减, 并且通过 b_{base}^{Θ} 来表示学生基本的掌握程度, 本文提出受遗忘影响的掌握程度计算方式:

$$\tilde{\theta} = (1-\mu_f)\theta+\mu_f\cdot f+b_{\text{base}}^{\Theta} \tag{23}$$

其中, b_{base}^{Θ} 表示从训练过的学生记录中学习到学生的基础能力, 代表了该系统中学生的普遍水平, 在面对没有历史记录的学生时也能做出预测. 此外, μ_f 是可学习的遗忘权重, f 为交互间隔时间相关的遗忘因子, 定义如下:

$$\begin{cases} f_{ls} = \exp(-|lag_s|) \\ f_{lm} = \exp(-|lag_m|) \\ f_{ld} = \exp(-|lag_d|) \\ f = \mu_{ls}\cdot f_{ls}+\mu_{lm}\cdot f_{lm}+\mu_{ld}\cdot f_{ld} \end{cases} \tag{24}$$

其中, μ_{ls}、μ_{lm} 以及 μ_{ld} 都是可学习的权重, lag_s、lag_m 以及 lag_d 分别表示两次交互在秒级、分钟级和天级 3 种时间尺度下衡量的相对间隔时间离散化嵌入[30]. 根据公式 (1) 的 2PL 函数, a 和 b 分别为学习得到的题目区分度和题目难度因子参数, 考虑遗忘因素对其进行修改, 做出最终预测即答对概率 $\hat{r}\in[0,1]$:

$$\hat{r} = Sigmoid(-1.7a(\tilde{\theta}-b)) \tag{25}$$

本文模型中需要更新的全部参数主要来自上述 3 个部分, 即题目嵌入层、编码器层以及预测层中的参数. MCAKT 通过最小化输出的预测结果 $\hat{r}$ 和实际得分 r 之间的二进制交叉熵损失:

$$loss = -\sum_{n=1}^{|X|}(r_n\log(\hat{r}_n)+(1-r_n)\log(1-\hat{r}_n)) \tag{26}$$

其中, X 表示训练集中所有的交互记录. 本文使用 Xavier 方法对模型参数进行初始化[31], 并通过 RAdam 方法进行参数优化[32].

3 实验设置及结果分析

知识追踪任务的目标在于追求预测性能和可解释性. 针对知识追踪目标, 本节通过一系列实验验证本文方法的有效性, 从预测性能和可解释性两个方面具体分析实验结果. 首先, 第 3.1 节和第 3.2 节介绍了实验使用的数据集、评估指标和对比方法. 其次, 为了验证本文方法相比现有工作在预测性能方面的优势, 在第 3.3 节中进行了对比实验及分析, 并且在第 3.4 节中通过冷启动实验进一步体现 MCAKT 面对冷启动问题的有效性. 此外, 第 3.5 节中通过消融实验进一步分析 MCAKT 各模块对预测性能的影响. 最后, 为了体现本文方法的可解释性, 第 3.6 节从 3 个方面总结了 MCAKT 的可解释性, 并且通过可视化详细分析具体案例.

3.1 数据集及评估指标

本文在知识追踪任务中广泛使用的两个真实的在线教育数据集上进行了实验.

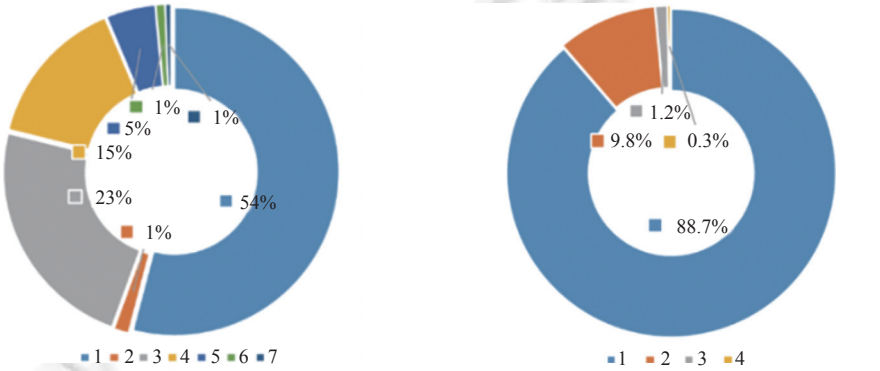
EdNet[33]是 Santa 收集的学生交互数据集, Santa 为韩国超过 78 万的用户提供在线学习辅导服务. EdNet 数据集由自 2017 年收集的学生问题解答日志组成. 该大规模数据集包含了每个用户的交互历史记录, 每个记录都包含了用户接收每个题目的时间戳、每个题目的 ID 以及每个题目所属类别等.

ASSISTment2009 包含了 2009 学年至 2010 学年内在 ASSISTments 上的学生交互信息. ASSISTments 是一个非盈利的在线教育系统, 为数学教学提供智能化的导学服务. ASSISTment2009 数据集包含了用户交互记录的基本数据.

这两个数据集的详细统计信息见表 1. 如表 1 所示, EdNet 数据集收集了更大规模的学生数据, EdNet 数据集中平均每个学生的记录数也更多. 此外, 表 1 中也表明 ASSISTment2009 数据集虽然有更多题目, 但是标注的知识却更少, 这意味着同样的知识点下有更多数量的题目. 图 4 中分析了数据集中多知识点题目的分布情况, 图例说明中的数字代表了题目所标注的知识点个数. 尤其值得注意的是, 如图 4 所示, EdNet 数据集中存在近一半的题目涉及 2 至 7 个知识点, 而 ASSISTment2009 数据集中绝大部分题目仅涉及 1 个知识点.

表 1 数据集统计信息

标签	EdNet	ASSISTment2009
# 学生	784 309	4 217
# 题目	18 143	26 688
# 知识点	219	123
# 交互记录	95 293 926	346 860
# 平均记录	121.50	82.25



(a) EdNet 数据集中多知识点题目分布 (b) ASSISTment2009 数据集中多知识点题目分布

图 4 数据集中的多知识点题目占比

本文使用 AUC (area under the curve) 和 ACC (accuracy) 作为评估指标来评估模型在每个数据集上的预测性能. 这两个指标都广泛应用在度量各种知识追踪方法中.

3.2 对比方法

为了评估本文模型的有效性, 将本文中的 MCAKT 模型与以下基线模型进行比较:

DKT[5]是第 1 个使用深度神经网络建模学生知识状态的知识追踪模型.

DKVMN[6]是一种基于记忆网络的知识追踪模型, 通过键矩阵、值矩阵分别存储潜在的题目表示和知识状态.

SAKT[34]利用 Transformer 自注意力机制解决知识追踪任务. SAKT 结构基于 Transformer 编码器, 根据过去的相关练习进行预测.

AKT[13]提出了一种单调注意力机制应用于知识追踪任务, 并且结合教育测量模型对题目建模.

GIKT[35]是一种基于图的知识追踪模型, 利用图卷积网络捕捉题目和知识点之间的相关性.

SSAKT[11]通过多维项目反应理论对题目建模, 并且在自注意层使用 LSTM 进行位置编码. 此外, 还设计了一个上下文模块来捕获上下文信息.

3.3 对比实验

关于实验环境设置, 所有实验都是在安装了 Nvidia Tesla V100S GPU 的 Linux 服务器上进行. 对于本文提出的 MCAKT 模型, 在两个数据集下的超参数设置如表 2 所示. 其中, # Attentive Layers 是 MCAKT 在基于注意力机制的多知识点融合方式下的多头注意力层的层数.

在对比实验中, 模型输入的交互序列长度 len 统一为 200, 对于长度大于 200 的序列将其分为多个长度为 200 的序列. 本文将数据集中 10% 的数据作为对比实验的测试集, 将其他数据分别以 80% 和 20% 的比例作为训练集和验证集的划分.

表 2 MCAKT 超参数设置

参数	EdNet	ASSISTment2009
dim	128	128
Batch size	256	64
Learning rate	1E–03	5E–05
# Attentive Layers	3	1
# Heads	8	8
# Encoder	5	4

为了比较 MCAKT 与现有方法在预测性能上的表现, 对不同知识追踪方法在两个数据集上进行了对比实验, 结果分别如表 3 所示. 表 3 中基线模型在两个数据集上的实验结果引用自相关论文[11,13,35]. 由于 ASSISTment2009 数据集中缺少交互时间戳信息, 在 ASSISTment2009 数据集下实验模型忽略时间相关的遗忘因子, 即公式 (23) 中 μ_f 置零. 值得注意的是, 两个数据集中的题目都提供了由人工标注的所涉及知识点 ID 集合, 在实验中将知识点集合与题目特征作为题目数据, 而不使用题目 ID 作为输入.

实验结果表明, 在预测学生未来表现方面, 在 EdNet 数据集上 MCAKT 模型性能明显优于对比模型, 在 ASSISTment2009 数据集上 MCAKT 模型与现有模型性能相当. 如表 3 所示, SAKT 在所有数据集上表现最差. 在 Zhang 等人[11]的工作中将其归因于 SAKT 使用了可学习的位置嵌入, 没有明确地对遗忘行为建模, 因此无法在数据集中学习有效的位置表示.

值得注意的是, 不同于 EdNet 数据集下的结果, 在 ASSISTment2009 数据集中 SSAKT 性能略胜 MCAKT, 这是由于 SSAKT 在进行题目嵌入时使用了更丰富的题目信息. 在该数据集下题目数量更多而标注知识点数量却更少, 这导致了没有直接对题目进行嵌入的 MCAKT 无法像 SSAKT 能更好区别题目. 但是 MCAKT 能很好地处理新题目具有更好的灵活性, 而 SSAKT 则需要进行重新训练. 而且, 在 EdNet 数据集下 GIKT 的性能也优于 SSAKT, 这可以归因于在大规模数据集中 GIKT 能更好挖掘知识点之间的关系, 有助于预测性能的提高. 如表 3 所示, MCAKT-SE 和 MCAKT-A 性能相当且在 EdNet 数据集中都优于其他模型, 验证了两种融合方式的有效性.

表 3 中基线模型对多知识点题目情况都采取拆分的处理, 即将涉及某道多知识点题目的交互记录分割成多个单一知识点题目交互记录. 而 MCAKT-SE 和 MCAKT-A 分别是采用基于 SENet 和基于注意力机制的多知识点融合方式, 无须拆分记录. 本文认为这种拆分的数据处理方式会造成数据的冗余. 为了对比这两种处理方式, 进一步分析多知识点融合对模型预测性能的提升, 本文在 ASSISTment2009 数据集上设计了实验如表 4 所示, 其中对比模型的实验结果引自已有工作[13]. DKT-M、DKVMN-M、SAKT-M、AKT-M 和 MCAKT-M 是对应模型的一种变体, 对题目的多个知识点嵌入进行平均池化作为题目知识点嵌入, 即对多知识点题目交互记录不进行分割.

表 3 在两个数据集上的实验结果

方法	EdNet	ASSISTment2009
DKT (2015)	0.6928	0.8170
DKVMN (2017)	0.6937	0.8093
SAKT (2019)	0.6919	0.7520
AKT (2020)	0.7454	0.8346
GIKT (2020)	0.7523	0.7896
SSAKT (2021)	0.7510	**0.8432**
MCAKT-SE	0.7752	0.8416
MCAKT-A	**0.7770**	0.8419

表 4 对多知识点题目使用平均池化的嵌入方式

模型	AUC
DKT-M	0.7616
DKVMN-M	0.7556
SAKT-M	0.7432
AKT-M	0.7866
MCAKT-M	**0.8410**

通过对比表 3 与表 4 可以观察到, 在两种不同的数据处理方式下, 将多知识点题目拆分后的实验结果有所提高, 这是数据被重复利用造成的, 这一点在已有的一些工作[13,23]中也得到了验证. 在 GIKT、SSAKT 等以往采用拆分或视多知识点为新知识点方式的工作中, 数据重复利用的问题同样存在.

综上, 与现有模型相比, MCAKT 在大规模数据集上预测表现优异, 并且具有更好的灵活性. 并且, MCAKT 中不存在数据重复利用的问题.

3.4 冷启动实验

冷启动是智能导学系统经常遇到的问题. 对于应用了知识追踪的智能导学系统, 冷启动主要涉及两种挑战, 即对新学生和新题目做出预测. 而多知识点融合的题目建模方式能很好地处理该问题, 为了验证 MCAKT 在面对冷启动问题时的预测性能, 本文分别设计了关于学生和题目的冷启动实验. 表 5 展示了 MCAKT 在测试集中面对没有历史记录的学生的预测表现. 表 6 展示了 MCAKT 在测试集中面对在训练集中未出现过的新题目的预测表现. 为了在测试集上拥有足够的新题目样本, 在题目冷启动实验中, 本文对数据集进行了重新划分, 划分出 20% 的数据集作为测试集. 值得注意的是, 虽然题目没有在训练集中出现过, 但是其涉及的知识点已出现在训练阶段, 只是可能会出现新的知识点交叉, 即在训练集中题目没有出现过的知识点组合.

表 5 学生的冷启动实验

方法	EdNet				ASSISTment2009			
	新学生		测试集		新学生		测试集	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
MCAKT-SE	**0.6906**	**0.7580**	0.7263	0.7752	0.7544	0.8175	0.7747	0.8416
MCAKT-A	0.6900	0.7578	**0.7270**	**0.7770**	**0.7719**	**0.8203**	**0.7735**	**0.8419**

表 6 题目的冷启动实验

方法	EdNet				ASSISTment2009			
	新题目		测试集		新题目		测试集	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
MCAKT-SE	0.7080	0.7560	**0.7178**	0.7620	**0.7519**	0.8174	0.7663	0.8282
MCAKT-A	**0.7144**	**0.7595**	0.7176	**0.7634**	0.7422	**0.8234**	**0.7685**	**0.8320**

对新学生的冷启动实验结果表明, 即使在没有任何历史记录的条件下, MCAKT 也能通过模型在训练阶段学习到的非个性化先验进行预测. 类似地, 对于新题目 MCAKT 也能根据他涉及的知识点做出预测. 如表 6 所示, MCAKT 在新题目的预测表现有所下降. 这可以归因于新题目缺少题目信息, 以及新题目可能涉及新的知识点组合. 值得注意的是, 即使面对不断增加的新题目, MCAKT 不需要重新训练就能够自然地处理冷启动问题.

3.5 消融实验

为了研究模型中不同模块的有效性, 本文在两个数据集上进行了几项消融实验, 将本文模型与几种变体进行比较. 基于 MCAKT 的变体总结如表 7 所示. 对比的模型使用一致的超参数, 并且输入序列长度都设置为 100.

为了验证多知识点融合层的有效性, 本文设计了 MCAKT-M, 即通过对多个单一知识点嵌入平均池化来进行融合.

为了验证将位置信息融入注意力计算的有效性, 本文设计了 MCAKT-RP, 将公式 (18) 中的 α_p 置零, 即 SAKT 中注意力权重的计算方式.

为了验证基于 IRT 模型和遗忘因子的预测层的有效性, 本文设计了 MCAKT-RF 和 MCAKT-R2PL 两种变体. MCAKT-RF 在 MCAKT 基础上去掉了遗忘因子, 即在公式 (23) 中将 μ_f 置零. MCAKT-R2PL 在 MCAKT 基础上去掉了基于 IRT 模型的预测层, 即使用一层线性层代替原有的预测层输出最后的结果.

如表 8 所示, 消融实验结果验证了 MCAKT 各模块的有效性. 如表 8 所示, 基于 SENet 和基于注意力的融合方式下的模型都较 MCAKT-M 有所提升, 验证了多知识点融合嵌入通过捕捉题目与知识点之间不同的关系, 能帮助对模型表现的提升. MCAKT-RP 与 MCAKT-A 相比性能有所下降, 验证了时序信息在 KT 任务中的重要性, 验

证了将序列信息融入注意力计算能帮助基于 Transformer 结构的知识追踪模型的序列预测. MCAKT-RF 和 MCAKT-R2PL 与 MCAKT-A 相比都有所下降, 且 MCAKT-R2PL 下降明显, 这验证了基于 IRT 模型和遗忘因子的预测层除了有助模型实现内置可解释性外, 对于模型预测性能的提升也同样有益.

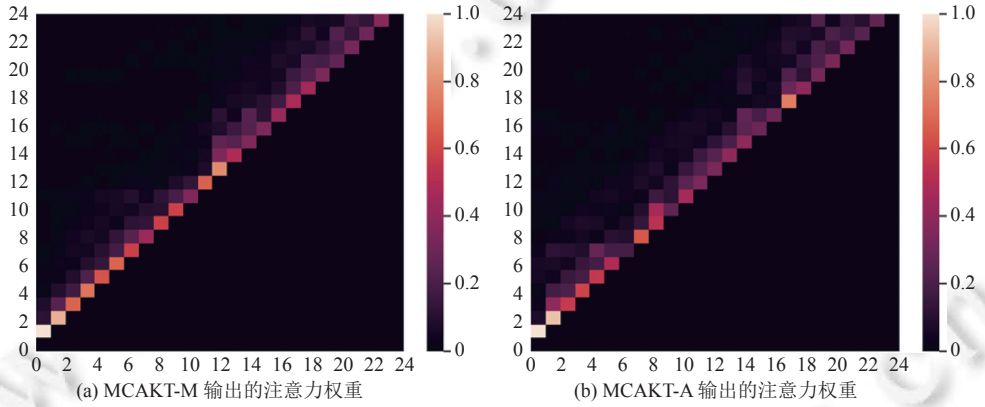
表 7 MCAKT 变体

方法	知识点嵌入	位置相关注意力	遗忘因子	基于2PL预测层
MCAKT-M	Mean	√	√	√
MCAKT-RP	Attentive	×	√	√
MCAKT-RF	Attentive	√	×	√
MCAKT-R2PL	Attentive	√	×	×
MCAKT-SE	SENet	√	√	√
MCAKT-A	Attentive	√	√	√

表 8 消融实验

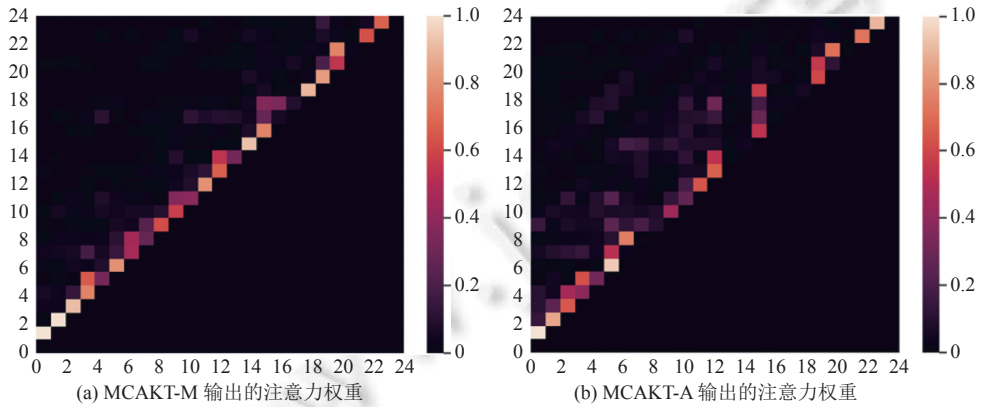
方法	EdNet		ASSISTment2009	
	ACC	AUC	ACC	AUC
MCAKT-M	0.726 0	0.774 7	0.774 3	0.841 0
MCAKT-RP	0.725 9	0.775 0	0.773 0	0.841 5
MCAKT-RF	0.726 2	0.774 8	—	—
MCAKT-R2PL	0.723 7	0.771 1	0.769 5	0.832 6
MCAKT-SE	0.726 3	0.775 2	**0.774 7**	0.841 6
MCAKT-A	**0.727 0**	**0.777 0**	0.773 5	**0.841 9**

为了进一步分析多知识点融合方式的有效性, 本文在 EdNet 和 ASSISTment2009 数据集中各自截取了一个学生的一段学习记录. 图 5 是将 MCAKT-A 及其变体 MCAKT-M 在 EdNet 数据集上输出的注意力权重可视化结果, 图 6 是将 MCAKT-A 及其变体 MCAKT-M 在 ASSISTment2009 数据集上输出的注意力权重可视化结果.



(a) MCAKT-M 输出的注意力权重 (b) MCAKT-A 输出的注意力权重

图 5 EdNet 数据集下注意力权重可视化



(a) MCAKT-M 输出的注意力权重 (b) MCAKT-A 输出的注意力权重

图 6 ASSISTment2009 数据集下注意力权重可视化

本文有以下 3 点观察. 首先, 从图 5、图 6 的 4 张注意力图可以观察到注意力主要分布在近期交互. 这可以归因于学生的遗忘行为以及学生通常在相近时间内进行一类题目练习的学习习惯. 其次, 将图 5(a) 与图 6(a) 对比,

将图 5(b) 与图 6(b) 对比, 可以观察到 EdNet 数据集下重点信息更集中在近期交互. 这可以归因于在 ASSISTment2009 数据集下没有考虑遗忘因素的影响; 以及 ASSISTment2009 数据集虽然有更多题目但是知识点却更少, 导致对不同题目的区别没有 EdNet 数据集下的明显. 最后, 如图 5 和图 6 所示, 可以观察 MCAKT-M 和 MCAKT-A 输出的注意力权重. 值得注意的是, 图 5(b) 中第 24 时刻代表对题目 q8120 的交互行为, 其在多知识点融合机制下对第 19 时刻的题目 q7338 的注意力权重有所增加. 但是题目 q7338 与题目 q8120 所标注的知识点并不相同. 由此看出, 基于注意力机制的融合方式不仅能帮助捕捉近期中相同知识点题目, 还能有助于捕捉知识点潜在关系.

3.6 案例分析

在知识追踪任务中, 除了希望能准确预测学生未来做题表现外, 还希望能够对预测结果具有一定的可解释性. MCAKT 的可解释性主要体现在 3 个方面, 第 1 是对历史记录的关注程度, 第 2 是对学生能力水平的建模, 第 3 是对题目的建模. 从这 3 个方面, MCAKT 能够在一定程度上对预测结果做出解释. 第 1 点是注意力知识追踪方法共有的特性, 如图 5、图 6 所示, 模型通过注意力机制计算历史交互之间的相似程度来学习重要的历史记录从而做出预测, 这种决策过程是易于理解的. 第 2 点和第 3 点是在模型结合 IRT 和遗忘因素进行构建的基础上体现的, 所学习的参数如掌握程度、题目区分度等可以为教学中所需的归因分析提供参考. 因此, 本文设计了一些实验, 通过可视化来分析具体案例. 关于学生能力水平的建模, MCAKT 能够从题目层面和知识点层面进行建模. 首先, 对学生关于具体某道题目的解题能力, 可以通过公式 (23) 建模. 本文截取了 EdNet 数据集中一位学生的一段学习记录, 图 7 是对该学生在学习过程中两个不同阶段输出的部分题目的解题能力的可视化结果. 其中 seq_0 表示的是 MCAKT 根据训练集中学生记录学习到的在 5 道题目上的非个性化先验解题能力. 可以观察到, 相比完成 50 道题目时的能力水平, 该生在经过 100 道题目训练后, 对题目的解题能力发生了变化. 经过一段时间的学习, 该生对题目 q4259、q6722、q6947 解题能力都有了提升, 但是对题目 q5135 的解题能力有所下降. 下降的原因可能是随时间的遗忘, 也可能是在学习其他知识点过程中对以前学习过的内容有一定干扰[36].

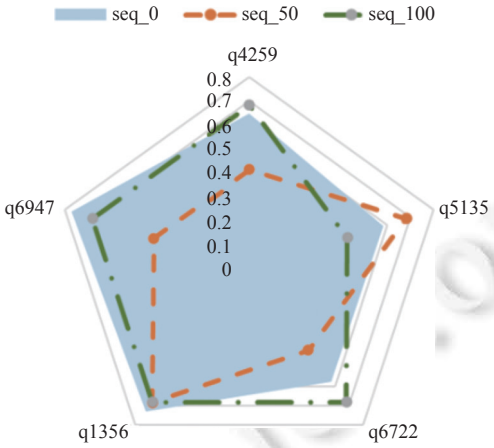


图 7 某学生在不同时刻下对 5 道题目的解题能力

其次, 关于追踪学生知识水平的变化, 可以将公式 (23) 对 θ 的计算进行调整如下, 能够输出学生在不同时刻对不同知识点的掌握程度:

$$\theta = W_2^{\Theta}\left(ReLU\left(W_1^{\Theta}(enc \oplus \tilde{c}) + b_1^{\Theta}\right)\right) + b_2^{\Theta} \tag{27}$$

其中, $\tilde{c}$ 可以通过公式 (7)、公式 (11) 得到. 如图 8 所示, 本文截取了 EdNet 数据集中一位学生的一段学习记录, 观察该学生在学习过程中输出的部分知识水平变化. 其中图 8 的横坐标为每个时刻的学习记录 $\left(q, kc_q, r\right)$, 分别代表题目、知识点集合及其回答情况, 纵坐标代表了 5 种不同的知识点. 实验结果显示, 在学生第 0 和第 13 时刻对涉

及知识点 86 的题目回答正确后, 该生在下一时刻对该知识点的掌握水平都有所上升. 在学生第 2–11 时刻对涉及知识点 74 的题目回答错误后, 该生在下一时刻对知识点 74 的掌握水平都有所下降. 这说明 MCAKT 能从学生回答情况中捕捉到对应的知识水平的变化. 如图 8 所示, 在第 1–7 时刻和第 7–13 时刻, 该生分别对 7 道题目进行了先后两次交互, 在第 2 次交互后相较第 1 次交互时的知识水平整体上都有所提升. 值得注意的是, 其中第 1 列表示的是 MCAKT 根据训练集中学生记录学习到的在 5 个知识点上的非个性化先验知识水平.

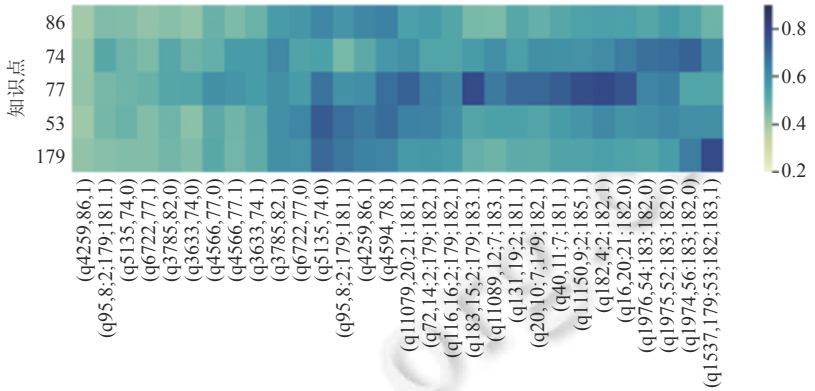


图 8 某学生在不同时刻对 5 种知识点的知识水平

此外, 如图 8 所示, 在第 0–29 时刻之间学生并未与涉及知识点 53 的题目进行交互, 在第 1–12 时刻之间学生并未与涉及知识点 179 的题目进行交互, 但是对应知识水平并不会一成不变. 可见, MCAKT 除了考虑了遗忘因素外, 还在学生交互过程中对潜在相关的知识点水平进行更新, 但这也会导致知识水平变化的波动.

关于题目的建模, MCAKT 根据公式 (1)、公式 (13)、公式 (21), 对题目特征进行学习. 图 9 展示了经由 MCAKT 学习后, EdNet 数据集中 5 道题目在区分度、难度上体现出的不同.

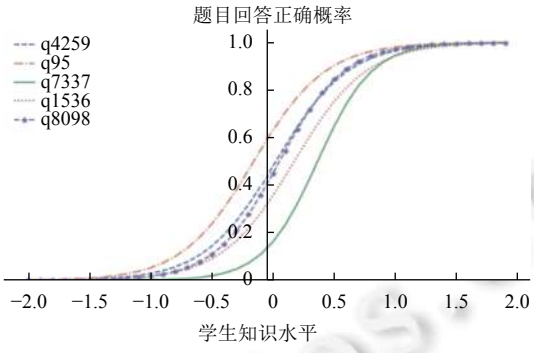


图 9 基于 2PL 建模的 5 道题目

综上, 针对知识追踪对预测性能和可解释性的两个要求, 本文通过实验验证了 MCAKT 的有效性. 第一, 本文方法能在预测学生未来做题表现上有所助力, 在大规模真实数据集上优于现有方法. 第二, 本文方法从 3 个方面提升深度模型的可解释性, 为理解模型决策和归因分析提供一定参考. 这也验证了结合教育心理学理论来提升深度 KT 模型可解释性的有效性.

4 总结及未来工作

本文提出了一种多知识点融合嵌入的深度知识追踪模型. 该方法重点研究了对涉及多知识点题目的建模, 提出了两种多知识点融合方式. 并且提出一种基于注意力机制的深度知识追踪模型, 结合 IRT 模型以及遗忘因素提

高预测效果以及增加模型可解释性. 在两个真实数据集上的实验验证了该方法的有效性.

未来将从以下两点展开进一步的研究工作: 一是引入更多的题目信息以改进题目嵌入方法, 比如题目文本等; 二是将本文结合的 IRT 模型扩展为多维项目反应理论模型[37]以改进对具体每个知识点掌握程度的建模.

References:

[1] Liu HY, Zhang TC, Wu PW, Yu G. A review of knowledge tracking. Journal of East China Normal University (Natural Science), 2019, (5): 1–15 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-5641.2019.05.001]

[2] Pardos ZA, Heffernan NT. Modeling individualization in a Bayesian networks implementation of knowledge tracing. In: Proc. of the 18th Int'l Conf. on User Modeling, Adaptation, and Personalization. Big Island: Springer, 2010. 255–266. [doi: 10.1007/978-3-642-13470-8_24]

[3] Corbett AT, Anderson JR. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-adapted Interaction, 1994, 4(4): 253–278. [doi: 10.1007/BF01099821]

[4] Ravikiran M. What's happened in MOOC posts analysis, knowledge tracing and peer feedbacks? A review. arXiv:2001.09830, 2020.

[5] Piech C, Bassen J, Huang J, Ganguli S, Sahami M, Guibas L, Sohl-Dickstein J. Deep knowledge tracing. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Cambridge: MIT Press, 2015. 505–513.

[6] Zhang JN, Shi XJ, King I, Yeung DY. Dynamic key-value memory networks for knowledge tracing. In: Proc. of the 26th Int'l Conf. on World Wide Web. Perth: Int'l World Wide Web Conf. Steering Committee, 2017. 765–774. [doi: 10.1145/3038912.3052580]

[7] Li XG, Wei SQ, Zhang X, Du YF, Yu G. LFKT: Deep knowledge tracing model with learning and forgetting behavior merging. Ruan Jian Xue Bao/Journal of Software, 2021, 32(3): 818–830 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6185.htm [doi: 10.13328/j.cnki.jos.006185]

[8] Nagatani K, Zhang Q, Sato M, Chen YY, Chen F, Ohkuma T. Augmenting knowledge tracing by considering forgetting behavior. In: Proc. of the World Wide Web Conf. San Francisco: Association for Computing Machinery, 2019. 3101–3107. [doi: 10.1145/3308558.3313565]

[9] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[10] Choi Y, Lee Y, Cho J, Baek J, Kim B, Cha Y, Shin D, Bar C, Heo J. Towards an appropriate query, key, and value computation for knowledge tracing. In: Proc. of the 7th ACM Conf. on Learning @ Scale. Association for Computing Machinery, 2020. 341–344. [doi: 10.1145/3386527.3405945]

[11] Zhang XL, Zhang JT, Lin NZ, Yang XD. Sequential self-attentive model for knowledge tracing. In: Proc. of the 30th Int'l Conf. on Artificial Neural Networks. Bratislava: Springer, 2021. 318–330. [doi: 10.1007/978-3-030-86362-3_26]

[12] Su Y, Liu QW, Liu Q, Huang ZY, Yin Y, Chen EH, Ding C, Wei S, Hu G. Exercise-enhanced sequential modeling for student performance prediction. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence and the 30th Innovative Applications of Artificial Intelligence Conf. and the 8th AAAI Symp. on Educational Advances in Artificial Intelligence. New Orleans: AAAI Press, 2018. 2435–2443.

[13] Ghosh A, Heffernan N, Lan AS. Context-aware attentive knowledge tracing. In: Proc. of the 26th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. Association for Computing Machinery, 2020. 2330–2339. [doi: 10.1145/3394486.3403282]

[14] Pandey S, Srivastava J. RKT: Relation-aware self-attention for knowledge tracing. In: Proc. of the 29th ACM Int'l Conf. on Information & Knowledge Management. Association for Computing Machinery, 2020. 1205–1214. [doi: 10.1145/3340531.3411994]

[15] Wang TQ, Ma FL, Gao J. Deep hierarchical knowledge tracing. In: Proc. of the 12th Int'l Conf. on Educational Data Mining. Montreal: Int'l Educational Data Mining Society, 2019. 667–670.

[16] Lu Y, Wang DL, Meng QG, Chen PH. Towards interpretable deep learning models for knowledge tracing. In: Proc. of the 21st Int'l Conf. on Artificial Intelligence in Education. Ifrane: Springer, 2020. 185–190. [doi: 10.1007/978-3-030-52240-7_34]

[17] Cheng S, Liu Q, Chen EH, Huang Z, Huang ZY, Chen YY, Ma HP, Hu GP. DIRT: Deep learning enhanced item response theory for cognitive diagnosis. In: Proc. of the 28th ACM Int'l Conf. on Information and Knowledge Management. Beijing: Association for Computing Machinery, 2019. 2397–2400. [doi: 10.1145/3357384.3358070]

[18] Baker RSJD, Corbett AT, Aleven V. More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In: Proc. of the 9th Int'l Conf. on Intelligent Tutoring Systems. Montreal: Springer, 2008. 406–415. [doi: 10.1007/978-3-540-69132-7_44]

[19] Pelánek R. Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. User Modeling and

User-adapted Interaction, 2017, 27(3): 313–350. [doi: 10.1007/s11257-017-9193-2]

[20] Yudelson MV, Koedinger KR, Gordon GJ. Individualized Bayesian knowledge tracing models. In: Proc. of the 16th Int'l Conf. on Artificial Intelligence in Education. Memphis: Springer, 2013. 171–180. [doi: 10.1007/978-3-642-39112-5_18]

[21] Minn S, Desmarais MC, Zhu FD, Xiao J, Wang JZ. Dynamic student classification on memory networks for knowledge tracing. In: Proc. of the 23rd Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Macao: Springer, 2019. 163–174. [doi: 10.1007/978-3-030-16145-3_13]

[22] Xu MK, Wu WJ, Zhou X, Pu YJ. Research on knowledge tracing model for multiple knowledge points and visualization. E-education Research, 2018, 39(10): 53–59 (in Chinese with English abstract). [doi: 10.13811/j.cnki.eer.2018.10.008]

[23] Xiong XL, Zhao SY, Van Inwegen E, Beck J. Going deeper with deep knowledge tracing. In: Proc. of the 9th Int'l Conf. on Educational Data Mining. Raleigh: Int'l Educational Data Mining Society, 2016. 545–550.

[24] Murre JMJ, Dros J. Replication and analysis of Ebbinghaus' forgetting curve. PLoS One, 2015, 10(7): e0120644. [doi: 10.1371/journal.pone.0120644]

[25] Song WP, Shi CC, Xiao ZP, Duan ZJ, Xu YW, Zhang M, Tang J. Autoint: Automatic feature interaction learning via self-attentive neural networks. In: Proc. of the 28th ACM Int'l Conf. on Information and Knowledge Management. Beijing: Association for Computing Machinery, 2019. 1161–1170. [doi: 10.1145/3357384.3357925]

[26] Huang TW, Zhang ZQ, Zhang JL. FiBiNET: Combining feature importance and bilinear feature interaction for click-through rate prediction. In: Proc. of the 13th ACM Conf. on Recommender Systems. Copenhagen: Association for Computing Machinery, 2019. 169–177. [doi: 10.1145/3298689.3347043]

[27] Hu J, Shen L, Albanie S, Sun G, Wu EH. Squeeze-and-excitation networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011–2023. [doi: 10.1109/TPAMI.2019.2913372]

[28] Zhou YH, Li XH, Cao YB, Zhao XM, Ye Q, Lv JC. LANA: Towards personalized deep knowledge tracing through distinguishable interactive sequences. In: Proc. of the 14th Int'l Conf. on Educational Data Mining. Int'l Educational Data Mining Society, 2021. 602–608.

[29] Shiv VL, Quirk C. Novel positional encodings to enable tree-based transformers. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 12081–12091.

[30] Shin D, Shim Y, Yu H, Lee S, Kim B, Choi Y. SAINT+: Integrating temporal features for EdNet correctness prediction. In: Proc. of the 11th Int'l Learning Analytics and Knowledge Conf. Irvine: Association for Computing Machinery, 2021. 490–496. [doi: 10.1145/3448139.3448188]

[31] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proc. of the 13th Int'l Conf. on Artificial Intelligence and Statistics. Sardinia: JMLR, 2010. 249–256.

[32] Liu LY, Jiang HM, He PC, Chen WZ, Liu XD, Gao JF, Han JW. On the variance of the adaptive learning rate and beyond. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[33] Choi Y, Lee Y, Shin D, Cho J, Park S, Lee S, Baek J, Bae C, Kim B, Heo J. EdNet: A large-scale hierarchical dataset in education. In: Proc. of the 21st Int'l Conf. on Artificial Intelligence in Education. Morocco: Springer, 2020. 69–73. [doi: 10.1007/978-3-030-52240-7_13]

[34] Pandey S, Karypis G. A self attentive model for knowledge tracing. In: Proc. of the 12th Int'l Conf. on Educational Data Mining. Montreal: Int'l Educational Data Mining Society, 2019. 384–389.

[35] Yang Y, Shen J, Qu YR, Liu YF, Wang KR, Zhu YM, Zhang WN, Yu Y. GIKT: A graph-based interaction model for knowledge tracing. In: Proc. of the 2020 Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. Ghent: Springer, 2020. 299–315. [doi: 10.1007/978-3-030-67658-2_18]

[36] Ausubel DP, Robbins LC, Blake Jr E. Retroactive inhibition and facilitation in the learning of school materials. Journal of Educational Psychology, 1957, 48(6): 334–343. [doi: 10.1037/h0043524]

[37] Reckase MD. The past and future of multidimensional item response theory. Applied Psychological Measurement, 1997, 21(1): 25–36. [doi: 10.1177/0146621697211002]

附中文参考文献:

[1] 刘恒宇, 张天成, 武培文, 于戈. 知识追踪综述. 华东师范大学学报(自然科学版), 2019, (5): 1–15. [doi: 10.3969/j.issn.1000-5641.2019.05.001]

[7] 李晓光, 魏思齐, 张昕, 杜岳峰, 于戈. LFKT: 学习与遗忘融合的深度知识追踪模型. 软件学报, 2021, 32(3): 818–830. http://www.jos.

org.cn/1000-9825/6185.htm [doi: 10.13328/j.cnki.jos.006185]
[22] 徐墨客, 吴文峻, 周萱, 蒲彦均. 多知识点知识追踪模型与可视化研究. 电化教育研究, 2018, 39(10): 53–59. [doi: 10.13811/j.cnki.eer.2018.10.008]

琚生根(1970－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为数据挖掘, 自然语言处理, 知识图谱.

赵容梅(1996－), 女, 博士生, CCF 学生会员, 主要研究领域为推荐系统.

康睿(1998－), 女, 硕士生, 主要研究领域为数据挖掘, 知识追踪.

孙界平(1962－), 男, 副教授, 主要研究领域为数据挖掘, 智慧教育.