

# 基于多层注意力网络的可解释认知追踪方法

孙建文<sup>1,2</sup> 周建鹏<sup>1,2</sup> 刘三女牙<sup>1,2</sup> 何绯娟<sup>3</sup> 唐 云<sup>4</sup>

<sup>1</sup>(华中师范大学人工智能教育学部 武汉 430079)  
<sup>2</sup>(教育大数据应用技术国家工程实验室(华中师范大学) 武汉 430079)  
<sup>3</sup>(西安交通大学城市学院计算机系 西安 710018)  
<sup>4</sup>(华中师范大学心理学院 武汉 430079)  
(sunjw@ccnu.edu.cn)

## Hierarchical Attention Network Based Interpretable Knowledge Tracing

Sun Jianwen<sup>1,2</sup>, Zhou Jianpeng<sup>1,2</sup>, Liu Sannüya<sup>1,2</sup>, He Feijuan<sup>3</sup>, and Tang Yun<sup>4</sup>

<sup>1</sup>(Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079)  
<sup>2</sup>(National Engineering Laboratory for Educational Big Data (Central China Normal University), Wuhan 430079)  
<sup>3</sup>(Department of Computer, Xi'an Jiaotong University City College, Xi'an 710018)  
<sup>4</sup>(School of Psychology, Central China Normal University, Wuhan 430079)

**Abstract** Knowledge tracing is a data-driven learner modeling technology, which aims to predict learners' knowledge mastery or future performance based on their historical learning data. Recently, with the support of deep learning algorithms, deep learning-based knowledge tracing has become a current research hotspot in the field. Aiming at the problems that deep learning-based knowledge tracing models generally have ‘black-box’ attributes, the decision-making process or results lack interpretability, and it is difficult to provide high-value education services such as learning attribution analysis and wrong cause backtracking, a Hierarchical Attention network based Knowledge Tracing model (HAKT) is proposed. By mining the multi-dimensional and in-depth semantic association between questions, a network structure containing three-layer attention of questions, semantics and elements is established, where graph attention neural network and self-attention mechanism are utilized for question representation learning, semantic fusion and questions retrieve. A regularization term to improve model interpretability is introduced into the loss function, with which a trade-off factor is incorporated to balance predictive performance and interpretability of model. Besides, we define an interpretability measurement index for the prediction results—fidelity, which can quantitatively evaluate the model interpretability. Finally, the experimental results on 6 benchmark datasets show that our method effectively improves the model interpretability.

收稿日期:2021-10-11;修回日期:2021-11-16  
基金项目:国家科技创新 2030 新一代人工智能重大项目(2020AAA0108804);国家自然科学基金项目(62077021,61977030,61937001,61807011);陕西省自然科学基金基础研究计划项目(2020JM-711);陕西省教育科学“十三五”规划课题(SGH20Y1397);西安交通大学城市学院课程思政专项研究项目(KCSZ01006);华中师范大学研究生教学改革研究项目(2020JG14)  
This work was supported by the Major Program of National Science and Technology Innovation 2030 of China for New Generation of Artificial Intelligence (2020AAA0108804), the National Natural Science Foundation of China (62077021, 61977030, 61937001, 61807011), the Natural Science Basic Research Program of Shaanxi Province (2020JM-711), the Shaanxi Provincial Education Science Regulations “Thirteenth Five-Year” Plan Project (SGH20Y1397), the Special Research Project of Xi'an Jiaotong University City College (KCSZ01006), and the Teaching Reform Research Project for Postgraduates of Central China Normal University (2020JG14).  
通信作者:何绯娟(hfj@mail.xjtu.edu.cn)

**Key words** knowledge tracing; interpretability; hierarchical attention; question semantic; fidelity

**摘 要** 认知追踪是一种数据驱动的学习主体建模技术,旨在根据学生历史答题数据预测其知识掌握状态或未来答题表现.近年来,在深度学习算法的加持下,深度认知追踪成为当前该领域的研究热点.针对深度认知追踪模型普遍存在黑箱属性,决策过程或结果缺乏可解释性,难以提供学习归因分析、错因追溯等高价值教育服务等问题,提出一种基于多层注意力网络的认知追踪模型.通过挖掘题目之间多维度、深层次的语义关联信息,建立一种包含题目元素、语义和记录等3层注意力的网络结构,利用图注意神经网络和自注意力机制等对题目进行嵌入表示、语义融合和记录检索.特别是在损失函数中引入提升模型可解释性的正则化项与权衡因子,实现对模型预测性能与可解释强度的调控.同时,定义了预测结果可解释性度量指标——保真度,实现对认知追踪模型可解释性的量化评估.最后,在6个领域基准数据集上的实验结果表明:该方法有效提升了模型的可解释性.

**关键词** 认知追踪;可解释性;多层注意力;题目语义;保真度

**中图法分类号** TP391

“利用现代技术加快推动人才培养模式改革,实现规模化教育与个性化培养的有机结合”是《中国教育现代化2035》的战略任务之一.云计算、大数据、人工智能等技术的发展,正推动教育从数字化、网络化向智能化加速跃升,智慧教育成为新一代技术环境下的教育信息化新范式<sup>[1]</sup>,为突破个性化学习技术瓶颈,实现“因材施教”的千年梦想提供了历史机遇.教育情境可计算、学习主体可理解、学习服务可定制是实现个性化学习面临的三大挑战<sup>[2]</sup>,学习主体是教育的核心要素,对学习主体的精准洞察是开展“因材施教”的前提.认知追踪(knowledge tracing, KT)作为一种数据驱动的学习主体建模技术,在大规模开放在线课程(massive open online courses, MOOC)、智能导学系统(intelligent tutoring system, ITS)等数字学习平台蓬勃发展、海量学习过程数据爆发式增长等多重效应的加持下,成为近年国内外智能教育领域的研究热点<sup>[3-8]</sup>.

认知追踪的思想源于美国著名心理学家 Atkinson<sup>[9]</sup>,1995年被美国卡耐基梅隆大学的 Corbett 等人<sup>[10]</sup>引入智能导学系统,并提出贝叶斯认知追踪方法(Bayesian knowledge tracing, BKT),其任务是根据学生过去的答题记录,对学生的知识掌握状态进行建模,目标是预测学生答对下一道题目的概率.2015年,美国斯坦福大学的 Piech 等人<sup>[11]</sup>首次将深度神经网络技术用于认知追踪,提出一种基于循环神经网络的深度认知追踪方法(deep knowledge tracing, DKT),在模型预测性能上取得显著提升.DKT的提出顺应了人工智能的技术发展趋势,吸引了多个领域学者的研究兴趣,先后

提出 DKVMN<sup>[12]</sup>, SKVMN<sup>[13]</sup>, HMN<sup>[14]</sup>, SAKT<sup>[15]</sup>, KQN<sup>[16]</sup>, GKT<sup>[17]</sup>, AKT<sup>[18]</sup>等一系列新模型.纵观认知追踪的整个发展历程,KT模型从技术上可分为3类<sup>[8]</sup>:基于概率的模型<sup>[10,19]</sup>、基于逻辑函数的模型<sup>[5,20-21]</sup>和基于深度学习的模型(以下称之为深度认知追踪模型)<sup>[11-12,15,17,22]</sup>.深度学习具有强大的拟合非线性函数和特征提取能力,使其适合用于建模复杂的认知过程,相比于概率类和基于逻辑函数的模型往往具有更高的预测性能,尤其是对于海量数据集其优势更加明显<sup>[23]</sup>.但目前大多数深度认知追踪模型均采用具有“黑盒”性质的神经网络技术进行建模,使其预测过程或结果的可解释性较差,难以满足教育领域强调归因分析的需求.

近年来,深度认知追踪模型缺乏可解释性的问题开始受到研究者的重视.其中比较有代表性的工作是将注意力机制应用于认知追踪以提升模型预测结果的可解释性<sup>[15,18,22,24]</sup>.其基本思想在于:学生的历史答题记录反映了当前答题表现,而不同历史答题记录对当前答题的影响是不同的;通过注意力机制使模型学习当前题目与历史答题记录的相关性权重,从而根据相关历史记录的题目信息以及答题情况为模型的预测结果提供一定的解释.但是目前利用注意力机制的认知追踪模型只关注当前题目与历史记录的浅层相关性信息,而忽略了当前题目与历史题目之间的多语义深层关联.由于这些模型仅引入题目-技能关系,只能将模型所学注意力权重归因于题目在技能维度的相似性.然而,能够反映当前题目答题表现的不仅仅是技能相同题目的答题记录,还有其他类型的相似题目,如协同相似<sup>[25]</sup>(即从

学生-题目交互数据中挖掘的相似)、模板相似<sup>[25]</sup>、难度相似<sup>[26]</sup>等。

题目之间的深层语义关联可从 2 个层面进行挖掘:1)在哪些语义维度具有关联以及不同语义维度关联强度如何;2)在特定语义维度由哪些元素进行关联以及不同元素的作用如何。“语义”指衡量题目具有关联的方面,如“具有相同技能的题目是有关联的”和“具有相同难度的题目是有关联的”属于不同的语义。“元素”指在特定语义维度关联题目的实体,如题目通过技能进行关联,则技能被称为元素,题目通过学生进行关联,则学生被称为元素。因此,为了建模当前题目与历史答题记录的相关性及其多语义深层关联,提出了一种多层注意力网络,包含记录级注意力、语义级注意力和元素级注意力,如图 1 所示(相比于其他基于注意力机制的认知追踪模型,增加了语义级和元素级注意力)。记录级注意力通过历史题目和当前题目的向量表示计算历史记录的相关性权重,然后按照权重综合历史记录的答题信息对当前答题做预测。语义级注意力能够计算不同语义维度对题目最终向量表示的重要性权重,并根据权重将不同类型的语义信息融合。元素级注意力旨在学习特定语义下不同元素反映题目特征的重要性权重,并按照权重将元素信息聚合到题目上。通过融合多层注意力,不仅可以得到哪些答题记录对当前答题预测具有更高的权重信息,还能获得在计算这些权重时哪些语义信息起到了更大的作用,以及在特定语义下哪些元素更能反映题目的特征。由此,可以结合多层注意力权重分布对模型完整的决策过程进行可视化分析与呈现,详见 3.5.1 节决策过程分析。

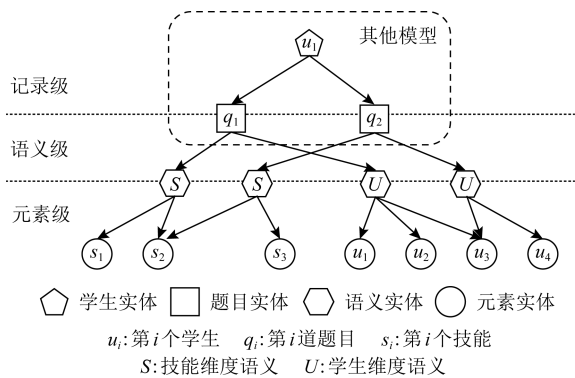


Fig. 1 Hierarchical attention structure of HAKT  
图 1 HAKT 多层次注意力结构

本文提出一种基于多层注意力网络的认知追踪模型,能够为模型的决策过程和预测结果提供解释。

具体地,1)将认知追踪问题域中不同类型实体及其关系表示为异质图,并从中抽取对应不同语义的二部图;2)基于图注意神经网络<sup>[27]</sup>构建元素级注意力,学习不同二部图中题目节点嵌入;3)利用语义级注意力将多种语义维度的题目嵌入融合成最终题目嵌入;4)基于自注意力机制<sup>[28]</sup>构建记录级注意力,检索并融合相关历史答题信息,进而预测当前题目的答对概率。

本文贡献主要体现在 3 个方面:

- 1) 针对当前深度认知追踪模型对预测结果普遍缺乏可解释性或仅通过建模题目之间浅层关系提供解释的问题,提出了一种通过多层注意力网络挖掘题目多语义深层关联信息的方法,能够显著提升模型对预测结果的可解释性。
- 2) 提出了评估认知追踪模型预测结果是否具有可解释性的基本思想,由此设计了提升模型可解释性的损失函数以及预测结果可解释性度量指标。
- 3) 在多个基准数据集上进行了充分实验,并与基于题目嵌入或注意力机制等相关的认知追踪模型进行了比较分析,验证了所提模型在预测性能上的有效性以及预测结果的可解释性。

1 相关工作

1.1 基于深度学习的认知追踪

目前的深度认知追踪模型主要包括 DKT<sup>[11]</sup>, DKVMN<sup>[12]</sup>, SAKT<sup>[15]</sup> 及其改进模型。DKT 利用循环神经网络来建模学生答题序列,并使用高维连续的隐向量表示认知状态。但是,DKT 仅使用题目对应的技能编号作为输入,忽略了其他题目信息。EKT<sup>[22]</sup>, DHKT<sup>[29]</sup>, PEBG<sup>[30]</sup>, GIKT<sup>[31]</sup>, CoKT<sup>[25]</sup> 等模型在 DKT 的基础上扩展了题目嵌入模块,利用题目文本信息或题目与学生、技能之间的交互信息增强题目表示。DKVMN 基于记忆网络建模学生答题序列,利用键、值矩阵分别表示题目的隐藏技能和技能的掌握状态。针对 DKVMN 无法建模学习过程中长期依赖的问题,SKVMN<sup>[13]</sup> 将 LSTM 与 DKVMN 结合,提出 HOP-LSTM 机制来解决。针对 DKVMN 仅使用单层记忆网络的不足,HMN<sup>[14]</sup> 引入层次记忆网络分别建模工作记忆和长期记忆,并设计了相应的划分机制和衰减机制。SAKT 通过自注意力网络建模学生答题序列,显式地建模当前答题与历史答题记录的相关性。RKT<sup>[24]</sup> 和 AKT<sup>[18]</sup> 分别通过引入题目文本信息、题目上下文信息等进一步改进了 SAKT,

并基于学到的注意力权重对预测结果进行解释性分析.总的来说,当前大多数深度认知追踪模型在预测性能上可取得较好的效果,但可解释性普遍较低.

## 1.2 可解释深度认知追踪

深度认知追踪模型已然成为认知追踪领域的研究热点,其可解释性研究则方兴未艾.随着可解释性日益成为人工智能领域的研究热点<sup>[32-33]</sup>,认知追踪领域也越来越关注模型的可解释性研究,并形成初步成果.根据解释的对象,可将相关研究分成2类:面向学生认知状态的解释和面向模型预测过程的解释.面向认知状态的解释旨在显式地建模认知状态,即模型内部存在可理解的参数对应每个技能的掌握状态(标量值).由于深度认知追踪模型均使用隐向量表示认知状态,而向量内部的参数难以解释.为使深度认知追踪模型输出可理解的认知状态,目前通常的做法是将其与经典的认知诊断模型结合(如IRT模型)<sup>[34-37]</sup>,其主要方式是将深度模型的输出作为IRT模型的输入,进而利用IRT模型做预测.由于IRT模型的约束,深度模型的输出可以对应IRT模型中可解释的参数(比如可理解的认知状态、学生能力、题目难度等).另外,一些研究者将深度模型已训练好的答题预测模块中的题目嵌入更换为技能嵌入并将对应题目嵌入的部分参数置零,由此将新的输出视为技能的掌握状态值<sup>[12,22]</sup>.面向预测过程的解释旨在解释模型为什么做出这样的预测.目前常用的方法主要是基于注意力机制的解释,包括EKT<sup>[22]</sup>,SAKT<sup>[15]</sup>,RKT<sup>[24]</sup>和AKT<sup>[18]</sup>等.其共同点都是通过计算当前答题与历史记录的相关性,进而解释模型在预测时关注哪些记录.然而,这些模型仅建模题目之间的浅层关联信息,忽略了题目之间的多语义深层关联.本文所提模型除了运用自注意力机制建模答题序列,还引入了另外2层注意力机制建模题目之间的多语义深层关联.其优点在于不仅能够挖掘更深层的信息,而且使模型的整个预测过程都具备可解释性.除此之外,也有研究者使用分层相关性传播方法(LRP)对认知追踪模型进行事后可解释分析,计算模型输出与输入的相关性来解释其预测结果<sup>[38]</sup>.

## 1.3 基于注意力机制的神经网络

由于注意力机制的有效性,其在涉及序列建模的任务中得到广泛应用.其基本思想是:在预测输出时,重点关注输入的相关部分.因此,它在一定程度上为模型提供了可解释性,因为人们可以通过模型所学权重理解模型在进行预测时更关注输入数据的

哪些部分.2017年,谷歌团队提出的自注意力机制<sup>[28]</sup>更是成为目前大规模预训练语言模型的基础.研究者将注意力机制用于图神经网络提出了GAT<sup>[27]</sup>,HAN<sup>[39]</sup>等图表示学习模型,能够学习不同邻居节点对中心节点的重要性.

# 2 模 型

本节首先对认知追踪问题进行形式化定义,介绍相关概念的符号表示,然后整体描述模型框架,最后依次介绍模型的各个组成模块以及损失函数.

## 2.1 问题定义

一方面,认知追踪问题域中的多种实体及其关系组成了异质图 $G=(E,R)$ ,其中 $E$ 和 $R$ 分别表示实体和关系集合.实体主要包括学生、题目以及题目的属性标签(即技能、模板等),关系主要为学生-题目和题目-属性等.通过异质图可以挖掘题目之间的多语义深层关联.另一方面,一个学生的答题序列由若干时间步对应的答题记录按照时序关系排列而成,时间步 $t$ 的答题记录可以表示为 $x_t=(q_t, a_t)$ ,  $q_t$ 表示题目编号而 $a_t$ 表示答题正确性(0表示答错,1表示答对).在引入题目多语义关联后,认知追踪任务可以被形式化表述为:已知异质图 $G$ ,当给定学生历史答题序列 $X=\{x_1, x_2, \dots, x_{t-1}\}$ 和当前题目 $q_t$ ,要求预测学生正确回答 $q_t$ 的概率,即 $P(a_t=1|G, X, q_t)$ .

## 2.2 整体框架

本文提出基于多层注意力网络的认知追踪模型,包含元素级、语义级和记录级3层注意力.图2展示了模型的整体框架,包括题目嵌入模块、知识检索模块和答题预测模块.

1) 题目嵌入模块.首先,从异质图 $G$ 中抽取不同语义对应的二部图并构建相应的邻接矩阵;然后,通过可学习嵌入层为每个节点生成初始特征向量;接着,基于图注意神经网络分别构建不同语义维度的元素级注意力,学习不同元素的重要性权重,并按照所学权重将邻居元素节点的特征向量聚合到题目节点得到题目嵌入.最后,利用语义级注意力学习不同语义对题目相关性计算的重要性权重,并融合不同语义维度的题目嵌入获得最终题目嵌入.

2) 知识检索模块.基于自注意力机制构建记录级注意力,显式地建模当前题目与历史记录的相关性,并根据相关性权重融合不同历史答题信息获得学生状态向量.

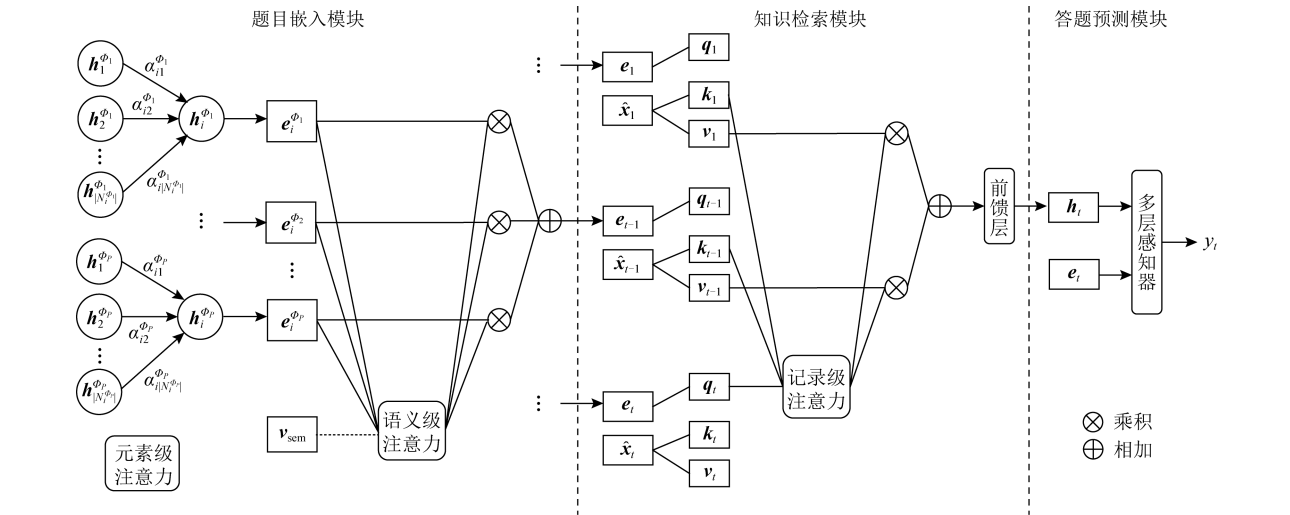


Fig. 2 The overall framework of HAKT  
图2 HAKT 模型框架图

3) 答题预测模块.使用多层感知器(multilayer perception, MLP)建模题目嵌入和学生状态向量的交互过程并预测当前题目的答对概率.

2.3 题目嵌入模块

2.3.1 语义抽取

在异质图中,题目间通过不同元素关联隐含不同的语义信息,例如“题目-技能-题目”表示具有相同技能的题目.从异质图中选取 4 种关系,每种关系对应一个二部图,每个二部图对应一个邻接矩阵.

1) 学生  $\longleftrightarrow$  题目.由训练集学生与其回答正确的题目构成的二部图,隐含的语义(记为  $U_c$ )为:被同一学生均回答正确的题目是有关联的.

2) 学生  $\longleftrightarrow$  题目.由训练集学生与其回答错误的题目构成的二部图,隐含的语义(记为  $U_w$ )为:被同一学生均回答错误的题目是有关联的.

3) 题目  $\longleftrightarrow$  技能.由所有题目与其关联的技能构成的二部图,隐含的语义(记为  $S$ )为:具有相同技能的题目是有关联的.

4) 题目  $\longleftrightarrow$  模板.由所有题目与其关联的模板构成的二部图,隐含的语义(记为  $T$ )为:具有相同模板的题目是有关联的.

2.3.2 元素级注意力

在二部图中聚合邻居元素节点的特征向量获取题目嵌入时,不同邻居的作用是不同的,因此设计元素级注意力来学习不同邻居的重要性.首先,通过可学习嵌入层生成所有节点的初始特征向量(节点  $i$  在语义  $\Phi$  下的初始特征向量记为  $h_i^\Phi$ ).然后,给定语

义  $\Phi$  对应二部图中的节点对  $(i, j)$  (其中  $i$  表示中心节点,  $j$  表示邻居节点),利用注意力机制学习不同邻居节点  $j$  对中心节点  $i$  的归一化权重  $\alpha_{ij}^\Phi$ :

$$\alpha_{ij}^\Phi = \text{attn}_{\text{elem}}(h_i^\Phi, h_j^\Phi; \Phi) \tag{1}$$

其中,  $\text{attn}_{\text{elem}}$  表示元素级注意力权重的具体计算过程.首先将中心节点  $i$  与其所有邻居  $j$  的特征向量各自拼接并通过非线性变换求得权重值;然后通过 softmax 函数将其归一化.其完整的计算过程为

$$\alpha_{ij}^\Phi = \frac{\exp(w_{ij})}{\sum_{k \in N_i^\Phi} \exp(w_{ik})}, \tag{2}$$

$$w_{ik} = \sigma(v_\Phi^\top \cdot [h_i^\Phi \parallel h_k^\Phi])$$

其中,  $\sigma$  表示激活函数,  $\parallel$  表示向量拼接符号,  $v_\Phi$  是对应语义  $\Phi$  的可学习注意力向量,  $N_i^\Phi$  表示节点  $i$  的邻居节点集合.

对于语义  $\Phi$ ,按照归一化权重  $\alpha_{ij}^\Phi$  聚合邻居节点的特征向量获得中心节点的嵌入  $e_i^\Phi$ :

$$e_i^\Phi = \sigma\left(\sum_{j \in N_i^\Phi} \alpha_{ij}^\Phi h_j^\Phi\right), \tag{3}$$

给定语义集合  $\{\Phi_1, \Phi_2, \dots, \Phi_P\}$ ,求得题目  $i$  在不同语义维度的嵌入  $\{e_i^{\Phi_1}, e_i^{\Phi_2}, \dots, e_i^{\Phi_P}\}$ .

2.3.3 语义级注意力

为了求得最终题目嵌入,需要将不同语义维度的题目嵌入进行融合.对于同一道题目,不同语义的重要性可能是不同的;对于不同的题目,相同语义的重要性也可能是不同的.由此,提出语义级注意力学习不同语义对特定题目的重要性.将元素级注意力所学不同语义维度的题目嵌入作为输入,语义级注意力计算不同语义对题目  $i$  的归一化权重:

$$\beta_i^{\Phi_1}, \beta_i^{\Phi_2}, \dots, \beta_i^{\Phi_P} = \text{attn}_{\text{sem}}(e_i^{\Phi_1}, e_i^{\Phi_2}, \dots, e_i^{\Phi_P}), \quad (4)$$

其中,  $\text{attn}_{\text{sem}}$  表示语义级注意力权重的具体计算过程. 首先将不同语义维度的题目嵌入分别进行非线性变换; 然后将变换后的题目嵌入与可学习注意力向量  $\mathbf{v}_{\text{sem}}$  的内积作为权重; 最后利用 softmax 函数将其归一化. 语义  $\Phi_j$  对题目  $i$  的权重的计算过程为

$$\beta_i^{\Phi_j} = \frac{\exp(\omega_i^{\Phi_j})}{\sum_{k=1}^P \exp(\omega_i^{\Phi_k})},$$

$$\omega_i^{\Phi_k} = \mathbf{v}_{\text{sem}}^T \cdot \tanh(\mathbf{W}e_i^{\Phi_k} + \mathbf{b}), \quad (5)$$

其中,  $\mathbf{W}$  是权重矩阵,  $\mathbf{b}$  是偏差向量,  $\mathbf{v}_{\text{sem}}$  是语义级注意力向量, 均为模型可学习参数.  $\beta_i^{\Phi_j}$  越高, 表明在衡量题目关联程度时语义  $\Phi_j$  的作用更大.

最后, 按照所求注意力权重将题目  $\beta_i^{\Phi_j}$  在不同语义维度的嵌入加权求和, 得到题目的最终嵌入  $\mathbf{e}_i$ :

$$\mathbf{e}_i = \sum_{j=1}^P \beta_i^{\Phi_j} \mathbf{e}_i^{\Phi_j}. \quad (6)$$

## 2.4 知识检索模块

### 2.4.1 交互向量生成

通过题目嵌入模块, 可以获得任意一道题目  $i$  的向量表示  $\mathbf{e}_i$ . 对于第  $t$  个时间步的题目嵌入, 记为  $\mathbf{e}_t$ . 为了融入答题情况  $a_t$ , 将题目嵌入  $\mathbf{e}_t$  与相同  $d$  维的零向量拼接, 获得交互向量  $\mathbf{x}_t \in \mathbb{R}^{2d}$ :

$$\mathbf{x}_t = \begin{cases} [\mathbf{e}_t \parallel \mathbf{0}], & a_t = 1, \\ [\mathbf{0} \parallel \mathbf{e}_t], & a_t = 0. \end{cases} \quad (7)$$

相比于循环神经网络, 自注意力机制在序列建模中具有更好的灵活性和有效性, 且模型内部生成的注意力权重可以为模型预测结果的解释提供基础. 因此, 与 SAKT<sup>[15]</sup> 一样, 本文使用自注意力机制建模学生的答题序列. 为了将答题交互的相对位置信息编码进模型, 定义位置嵌入矩阵  $\mathbf{P} \in \mathbb{R}^{l \times 2d}$  来引入答题序列的位置信息,  $l$  为最大序列长度. 在加入位置信息后, 交互向量表示为  $\hat{\mathbf{x}}_t = \mathbf{x}_t + \mathbf{P}_t$ ,  $\mathbf{P}_t$  表示位置嵌入矩阵中第  $t$  行, 即第  $t$  个时间步的位置嵌入.

### 2.4.2 记录级注意力

记录级注意力计算当前题目与历史答题记录的相关性权重, 进而通过聚合历史交互向量得到当前的学生状态向量. 将当前时间步的题目嵌入  $\mathbf{e}_t$  ( $t$  表示当前时间步) 映射为查询向量 (query), 将历史交互向量  $\hat{\mathbf{x}}_i$  ( $i=1, 2, \dots, t-1$ ) 映射为键向量 (key) 和值向量 (value). 计算当前题目与时间步  $i$  的历史答题记录的注意力权重  $\gamma_i$ :

$$\gamma_i = \frac{\exp(\omega_i)}{\sum_{k=1}^{t-1} \exp(\omega_k)},$$

$$\omega_i = \frac{(\mathbf{W}^Q \mathbf{e}_t)^T \cdot \mathbf{W}^K \hat{\mathbf{x}}_i}{\sqrt{d}},$$

其中,  $\mathbf{W}^Q \in \mathbb{R}^{d \times d}$  和  $\mathbf{W}^K \in \mathbb{R}^{2d \times d}$  分别是自注意力机制中 query 和 key 的映射矩阵.

最后, 将历史交互向量按照注意力权重加权求和, 得到记录级注意力模块的输出向量  $\mathbf{o}_t \in \mathbb{R}^d$ , 即与当前答题相关的学生状态向量:

$$\mathbf{o}_t = \sum_{i=1}^{t-1} \gamma_i \mathbf{W}^V \hat{\mathbf{x}}_i, \quad (9)$$

其中,  $\mathbf{W}^V \in \mathbb{R}^{d \times d}$  是自注意力机制中 value 的映射矩阵.

### 2.4.3 前馈层

为了进一步增强特征的交互能力和模型的拟合能力, 将记录级注意力的输出通过 point-wise 前馈层进行非线性变换. 前馈层由 2 个线性变换组成, 中间包含 ReLU 激活函数, 其计算过程为

$$\mathbf{h}_t = \mathbf{W}^{(2)} \text{ReLU}(\mathbf{W}^{(1)} \mathbf{o}_t + \mathbf{b}^{(1)}) + \mathbf{b}^{(2)}, \quad (10)$$

其中,  $\mathbf{W}^{(1)} \in \mathbb{R}^{d \times d}$  和  $\mathbf{W}^{(2)} \in \mathbb{R}^{d \times d}$  为可学习权重矩阵,  $\mathbf{b}^{(1)} \in \mathbb{R}^d$  和  $\mathbf{b}^{(2)} \in \mathbb{R}^d$  为可学习偏差向量.

## 2.5 答题预测模块

预测模块拟合知识检索模块的输出向量  $\mathbf{h}_t$  与题目嵌入  $\mathbf{e}_t$  之间的交互函数, 预测学生正确回答当前题目的概率. 使用多层感知器 (MLP) 实现:

$$\mathbf{z}_t^l = \text{ReLU}(\mathbf{W}^{(l)} \mathbf{z}_t^{l-1} + \mathbf{b}^{(l)}), \quad \mathbf{z}_t^0 = [\mathbf{h}_t \parallel \mathbf{e}_t], \quad (11)$$

$$y_t = \text{Sigmoid}(\mathbf{W}^{(L)} \mathbf{z}_t^{L-1} + \mathbf{b}^{(L)}), \quad (12)$$

其中,  $l$  表示多层感知器的第  $l \in \{1, 2, \dots, L-1\}$  层,  $\mathbf{W}^{(l)} \in \mathbb{R}^{d \times d}$  和  $\mathbf{b}^{(l)} \in \mathbb{R}^d$  分别是第  $l$  层的权重矩阵和偏差向量,  $y_t$  是预测概率.

## 2.6 损失函数及改进策略

尽管在基于注意力机制的认知追踪模型中, 相关性权重可以为预测结果的解释提供基础, 但是有权重并不意味着一定能够产生易于人们理解的解释过程. 例如, 如果模型赋予权重较大的几个历史记录的答案情况都是错误 (正确) 的, 但是模型却预测当前题目会答对 (答错), 那么所学权重就很难对预测结果形成合理的解释. 由此本文提出评估预测结果是否具有可解释性的基本思想: 模型预测结果与其所关注历史记录的真实答题结果的一致性反映了预测结果的可解释性.

基于该思想在损失函数中引入可解释性正则化项, 使得模型在训练时兼顾预测性能和可解释性.

具体地,假设模型预测第  $t$  个时间步的答题,历史答题情况为  $\{a_1, a_2, \dots, a_{t-1}\}$ , 计算得到前  $t-1$  个时间步的相关性权重为  $\{\gamma_1, \gamma_2, \dots, \gamma_{t-1}\}$ . 将历史答题情况加权求和得到  $s_t = \sum_{i=1}^{t-1} \gamma_i a_i$ , 则  $s_t$  反映了历史记录的综合答题情况(在考虑当前题目与历史记录相关性的情况下). 基于上述评估预测结果可解释性的思想可知: 对于当前时间步  $t$  的答题预测, 若  $s_t$  与模型预测值  $y_t$  越接近, 则预测结果的可解释性越好. 因此, 本文将模型的损失函数分成 2 部分: 1) 旨在降低模型预测值  $y_t$  与真实值  $a_t$  的二元交叉熵, 提升模型的预测性能; 2) 旨在减小模型预测值  $y_t$  与历史答题结果加权值  $s_t$  的根均方差, 提升模型预测结果的可解释性. 同时使用超参数  $\lambda$  来权衡模型的预测性能与可解释性, 则损失函数的整体计算过程:

$$\mathcal{L} = (1 - \lambda) \mathcal{L}_{\text{pred}} + \lambda \mathcal{L}_{\text{exp}}, \tag{13}$$

$$\mathcal{L}_{\text{pred}} = - \sum_{i \in I_B} (a_i \log y_i + (1 - a_i) \log (1 - y_i)), \tag{14}$$

$$\mathcal{L}_{\text{exp}} = \sqrt{\sum_{i \in I_B} (y_i - s_i)^2 / |I_B|}, \tag{15}$$

其中,  $I_B$  表示一个批次(batch)中的所有交互,  $|I_B|$  是总交互数, 即一个批次中所有序列长度之和. 为了训练模型参数, 使用 Adam 梯度下降算法最小化上述损失函数  $\mathcal{L}$ .

### 3 实验

本节首先对实验所用数据集、对比模型和评价

指标等进行介绍, 然后对各模型预测性能进行对比分析, 最后对模型可解释性分别进行定性和定量分析.

#### 3.1 数据集及对比模型介绍

本文在认知追踪领域 6 个常用的数据集上进行实验, 分别是 ASSIST09, ASSIST12, ASSIST17, EdNet, Statics2011 和 Eedi. ASSIST 系列数据集是由 ASSISTments 在线辅导平台收集, 其中 ASSIST09 是目前认知追踪领域最常用的基准数据集; Statics2011 收集于某大学静力学课程的辅导系统; EdNet 是由在线辅导平台 Santa 自 2017 至 2019 年收集的数据集; Eedi 是 NeurIPS2020 教育数据挖掘挑战赛使用的数据集, 由在线教育平台 Eedi 自 2018 至 2020 年收集.

参照现有研究工作, 本文对数据集进行预处理: 由于 ASSIST12, EdNet 和 Eedi 数据集太大, 从中随机抽取 5 000 名学生的数据进行实验<sup>[31]</sup>. 对于 ASSIST 系列数据集, 删除脚手架问题关联的记录<sup>[40]</sup>. 对于所有数据集, 删除技能标签为空的记录<sup>[18]</sup>. 对于以技能编号而不是题目编号为输入的模式(即 DKT, DKVMN 和 SAKT), 将一道题目的多个技能组合成一个新技能作为输入<sup>[40]</sup>. 对于 Statics2011, 将原题目编号和步骤编号合成新的题目编号作为输入, 且对同一题目连续多次作答的情况只保留第一次作答记录. 将 80% 的答题序列作为训练集, 其余 20% 作为测试集<sup>[31]</sup>. 最后, 从测试集中删除训练集中未出现题目的相关记录. 预处理后数据集的统计信息如表 1 所示. 加载数据时, 删除长度小于 3 的答题序列; 同时考虑到运行效率问题, 将长度超过 200 的答题序列拆分成多个长度为 3~200 的序列<sup>[18]</sup>.

Table1 Summary Statistics of Processed Datasets

表 1 预处理后的数据集统计信息

统计指标	数据集					
	ASSIST09	ASSIST12	ASSIST17	EdNet	Statics2011	Eedi
学生数	4 160	5 000	1 706	5 002	331	5 000
题目数	15 680	36 056	1 150	11 775	633	26 706
技能数	167	242	86	1 837	106	1 050
模板数	655	5 272				
答题总次数	259 105	940 179	248 989	1 658 820	112 921	751 989
语义 $U_c$ 对应二部图中边的条数	137 403	505 958	79 096	773 164	70 143	380 968
语义 $U_w$ 对应二部图中边的条数	70 256	211 230	74 228	383 090	20 990	216 156
语义 $S$ 对应二部图中边的条数	189 370	36 056	1 150	26 345	733	29 464
语义 $T$ 对应二部图中边的条数	15 680	36 056				
每道题目的平均回答次数	16.5	26.1	216.5	140.9	178.4	28.2
每个技能的平均练习次数	1 551.5	3 885.0	2 895.2	903.0	1 065.3	716.2

由于本文所提模型结合题目嵌入和注意力机制,因此为了验证模型在预测性能和可解释性方面的有效性,本文从3方面选取对比模型:经典类模型(DKT<sup>[11]</sup>,DKVMN<sup>[12]</sup>)、题目嵌入类模型(DHKT<sup>[29]</sup>,PEBG<sup>[30]</sup>,GIKT<sup>[31]</sup>)和注意力类模型(SAKT<sup>[15]</sup>,AKT<sup>[18]</sup>).其中,DKT和DKVMN分别使用循环神经网络和键值记忆网络建模学生序列,是深度认知追踪方向目前最常用的2个基准模型,其均以题目对应的技能编号作为输入.DHKT,PEBG和GIKT是目前预测性能较好的基于题目嵌入的深度认知追踪模型,均以题目-技能关系挖掘题目相似性并学习题目嵌入,再结合DKT或DKVMN做预测.SAKT和AKT均基于自注意力机制,SAKT使用技能编号作为输入,AKT以题目编号作为输入并结合Rasch模型和答题序列的上下文信息学习题目嵌入.SAKT和AKT均只有记录级注意力,本文模型则进一步扩展了语义级注意力和元素级注意力,用于建模多语义深层关联以增强模型可解释性.除了上述对比模型,本文还将输入技能编号的DKT,DKVMN和SAKT分别拓展为输入题目编号的DKT-Q,DKVMN-Q和SAKT-Q.DKT-Q与DKT的不同之处仅在于输入部分由独热编码换成可学习嵌入层,而DKVMN-Q和SAKT-Q在结构上未作改动.

3.2 评价指标

3.2.1 预测性能评价指标

认知追踪任务可以看成是一个二值分类问题,即预测题目回答的正确性(正确或错误).因此,参照绝大多数现有研究工作,本文使用AUC作为衡量模型预测性能的指标.

3.2.2 可解释性评价指标

为了进一步量化模型的可解释性,基于2.6节关于预测结果可解释性评估的基本思想,提出可解释性度量指标:保真度(Fidelity).保真度指可解释模型在输出结果上多大程度上与复杂模型相近,被广泛用于度量机器学习模型的可解释性<sup>[33,41]</sup>.首先定义可解释的预测结果:对于时间步 $t$ 的预测,若模型预测结果 $y_t$ 与历史答题情况的加权值 $s_t$ 的差距小于等于指定阈值 $\theta$ ,则认为该预测结果是可解释的,否则认为不可解释.进一步定义保真度:在所有的预测结果中,可解释的预测结果所占的比例.因此,保真度的计算:

$$Fidelity = \frac{1}{n} \sum I(t),$$
$$I(t) = \begin{cases} 1, & |s_t - y_t| \leq \theta, \\ 0, & |s_t - y_t| > \theta, \end{cases} \quad (16)$$

其中, $n$ 表示总预测次数,即测试集中所有序列长度之和. $Fidelity$ 越大,表明模型可解释性越好.

3.3 实验设置

对于对比模型:PEBG<sup>①</sup>和GIKT<sup>②</sup>的官方代码均从网上获取,其余模型代码根据其论文描述进行复现.其中,PEBG公开代码中在计算题目属性值时使用了测试集数据,存在数据泄露的问题,因此在本文中使用了修正后的版本.所有对比模型的超参数或者采用其论文中的最佳设置,或者在验证集(训练集的10%)上进行最优超参数搜索.本文模型HAKT<sup>③</sup>部分超参数设置为:题目嵌入模块中,题目嵌入维度为128,图注意力网络的注意力头数目为4;知识检索模块中,知识状态向量维度为128,自注意力网络的注意力头数目为8;答题预测模块中,多层感知机层数为2、中间层维度为128;在模型训练阶段,学习率设为0.001,批大小 $batch\ size = 32$ .其余超参数(包括解释性正则化项权衡因子 $\lambda$ )在不同数据集取值不一,均通过超参数搜索确定.另外,2.3.1节中介绍的4种语义( $U_c, U_w, S, T$ )并非全部用于实验.由于数据集特征(如除了ASSIST09和ASSIST12,其他数据集不包含“模板”特征,对应语义 $T$ )以及数据分布差异,其对应的最佳语义组合是不同的.通过实验确定每个数据集的最佳语义组合为ASSIST09和ASSIST12; $S \& T$ ,其余数据集: $U_c \& U_w \& S$ .

3.4 预测性能分析

表2展示了各个模型在6个数据集上的AUC值(取5次重复实验的均值).分析表2数据可得:

1) 对比以技能编号为输入的模型DKT,DKVMN,SAKT及其对应的以题目编号为输入的变体模型DKT-Q,DKVMN-Q,SAKT-Q发现,仅输入题目编号或仅输入技能编号的模型均不能稳定地占有优势.以DKT和DKT-Q为例,在ASSIST17,EdNet和Statics2011这3个数据集上,DKT-Q显著优于DKT,而在其他3个数据集上则相反.其原因是这3个数据集中题目的平均交互次数明显更少(参照表1),表明学生-题目的交互数据很稀疏,从而导致以题目编号为输入的DKT-Q表现更差.

① <https://github.com/lyf-1/PEBG>  
② <https://github.com/Rimoku/GIKT>  
③ <https://github.com/john1226966735/HAKT>

Table 2 Prediction Performance on 6 Datasets  
表 2 在 6 个数据集上的预测性能比较

模型分类	数据集上的 AUC 值						
	模型名称	ASSIST09	ASSIST12	ASSIST17	EdNet	Statics2011	Eedi
经典类	DKT	0.767 6	0.730 5	0.718 6	0.701 5	0.840 3	0.762 4
	DKT-Q	0.714 1	0.705 0	0.754 1	0.736 7	<u>0.885 3</u>	0.740 2
	DKVMN	0.776 2	0.722 0	0.715 9	0.684 6	0.835 0	0.746 6
	DKVMN-Q	0.683 1	0.693 6	0.736 4	0.733 8	0.862 5	0.667 6
题目嵌入类	DHKT	0.769 9	0.755 8	0.766 1	<b>0.765 8</b>	0.873 8	<b>0.791 1</b>
	PEBG	0.770 7	0.752 0	0.761 2	0.763 7	0.834 1	0.776 1
	GIKT	0.785 9	0.759 4	<b>0.772 3</b>	0.764 0	<b>0.890 7</b>	0.788 1
注意力类	SAKT	0.757 7	0.726 2	0.707 7	0.689 5	0.825 2	0.748 6
	SAKT-Q	0.749 4	0.738 2	0.752 5	0.760 3	0.864 9	0.778 9
	AKT	<u>0.788 8</u>	<u>0.766 5</u>	0.762 0	<u>0.764 7</u>	0.871 8	<u>0.789 0</u>
	HAKT	<b>0.790 2</b>	<b>0.767 7</b>	<u>0.766 5</u>	0.761 9	0.883 4	0.784 5

注:最优模型由黑体数字标记,次优模型由下划线标记.

2) 通过题目-技能关系同时引入题目和技能信息的模型(即 DHKT,PEBG,GIKT,AKT 和 HAKT)总体上比上述仅使用技能信息或仅使用题目信息的模型预测性能更好.以 DKT,DKT-Q 和 DHKT 为例,DHKT 在 6 个数据集上的 AUC 均值分别比 DKT 和 DKT-Q 高 3.3 个百分点和 3.1 个百分点.这说明,通过同时引入题目和技能信息,保留题目之间的差异性和相似性,能够使模型的预测更准确.

3) 本文模型 HAKT 在 SAKT 的基础上扩展题目嵌入模块,相比 SAKT 取得了显著的提升,并且总体上也略优于其他基于题目嵌入的模型,尤其是 DHKT 和 PEBG.这表明,相比于后者单纯引入技能维度关联,HAKT 通过引入题目的多语义关联使得模型可以更准确地挖掘题目之间的语义关系,从而学到更优的题目嵌入.在 ASSIST09 和 ASSIST12 数据集上 HAKT 表现最佳,其他数据集上与最优模型的差距约 0.5 个百分点,这说明 HAKT 在提升模型可解释性的同时也具有较高的预测性能.

4) 本文实现的 SAKT 在所有数据集上预测性能均低于 DKT,这一结果与 AKT<sup>[18]</sup>一文中的结果一致.其可能的原因是,一方面认知追踪领域的的数据量较小,另一方面该领域数据集中相似题目往往依次出现,使得题目之间的依赖距离较短.因此,自注意力机制无法发挥其特有的优势.同样,相比于 DHKT 和 GIKT 主要基于循环神经网络,HAKT 完全基于自注意力机制,这可能也是 HAKT 在部分数据集预测性能略低于两者的原因.

3.5 可解释性分析

3.5.1 决策过程分析

HAKT 能够通过其内部计算的注意力权重分布对预测过程和结果提供可解释性分析.相比基于单层注意力机制的模型(如 EKT,RKT,AKT 等),HAKT 整合多层注意力的权重分布能够生成更精确、完整的模型决策过程.具体地,HAKT 不仅能检索出哪些历史题目与当前题目关联,还能得到它们是如何进行关联的.通过案例来分析 HAKT 的可解释性.

从测试集中随机选择一名学生(记为  $u_1$ ),图 3 对模型预测学生回答第 20 道题目(即  $q_{20}$ )时生成的多层注意力权重分布以及题目之间的关联图进行可视化.为了方便展示,仅呈现权重较大的部分历史题目(即  $q_{18}$  和  $q_{19}$ )及权重较大的部分相关语义( $S$  表示技能维度, $U_c$  表示学生维度)和元素(如  $s_1$  和  $u_1$  等).分析图 3 得:

1) 根据记录级注意力权重分布,模型在预测当前答题时  $q_{18}$  和  $q_{19}$  被赋予最大的权重且其真实答题情况均为“正确”;模型预测当前答题正确的概率为 0.75(即答对概率较高).这说明模型当前预测结果与权重较大的历史记录的答案情况一致,即模型能从历史记录中找到与当前答题相关的记录,并综合历史答题情况对当前答题做预测.

2) 尽管通过记录级注意力能够检索出哪些历史记录对当前答题影响更大,但是这些历史题目与当前题目是如何进行关联的无法得知.进一步,通过

语义级注意力权重分布可知,对于相关历史题目 $q_{19}$ ,语义 $U_c$ 的权重较大.这说明它们的关联主要源于其具有相似的答题交互记录.那么与哪些学生的交互体现了它们的关联呢?通过元素级注意力权重分布可知,学生 $u_3$ 和 $u_4$ 的权重较大,这说明由于这些学生均答对了 $q_{19}$ 和 $q_{20}$ ,使得模型认为这2道题目具有关联.

3) 由此,综合3层注意力可得模型预测当前答题的决策过程,可由图3中实线呈现,同时实线部分也表明了历史记录与当前题目的语义关联:由于当前题目与历史题目 $q_{19}$ 具有相似的交互记录且与 $q_{18}$ 具有相同的技能,且 $q_{18}$ 和 $q_{19}$ 均回答正确,因此模型预测当前题目答对的概率较高.

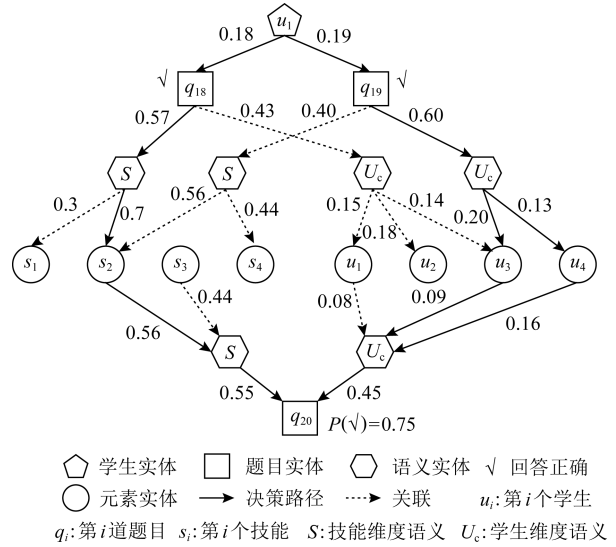


Fig. 3 Visualization of model's decision process  
图3 模型决策过程可视化

此外,为了验证模型在决策可解释性方面的可靠性,本文进一步探究:在所有答题预测中,权重较大的历史记录与当前题目存在语义关联的比例有多大?首先,定义题目的语义关联性:若2道题目至少在一个语义层面具有相同的元素,则它们具有语义

关联性,即关联性标签为1,否则关联性标签为0.然后,定义单次预测中较大权重历史记录与当前题目是否具有相关性:若权重最大的前 $k$ 个历史记录中至少有一个与当前预测题目具有语义关联性,则当次预测命中相关题目,记为1,否则为0;最后,计算命中相关题目的预测数占所有预测数的比例(命中率).以ASSIST09数据集为例,所选语义组合为S&T,即技能维度和模板维度.取 $k=3$ ,命中率为90.5%(若随机选择3个历史记录则命中率为43.7%),即在90%以上的预测中,模型赋予权重最大的3个历史记录,至少有1个历史记录与当前预测题目有显式的语义关联(即有相同技能或相同模板).因此,HAKT不仅能够捕获题目之间的相关性,并且对模型决策过程的解释具有较高的可靠性.

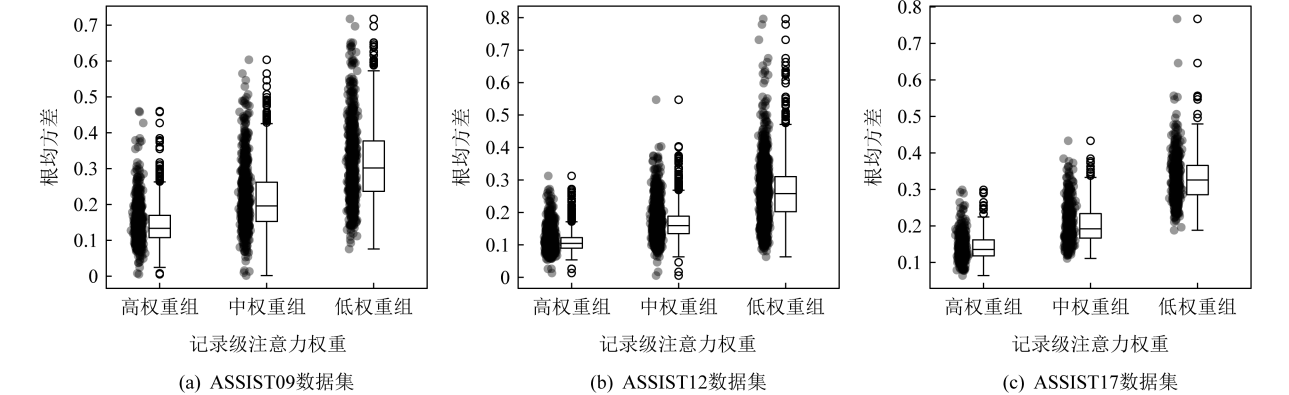
3.5.2 一致性分析

为了进一步验证案例分析中的观察(即:模型当前的预测结果与相关性权重较大的历史记录的答案情况具有更高的一致性),对模型预测值和历史相关题目答题结果进行一致性分析.参考EKT<sup>[22]</sup>:

1) 对于某个学生在某一时间步的预测,首先计算历史答题记录对应的注意力权重,然后将这些答题记录按照注意力权重大小等分成高、中、低3组,最后将每一组的答题得分(正确为1,错误为0)各自按照注意力权重进行加权求和.

2) 对于每一组,计算该学生所有时间步对应的加权求和值与模型预测值之间的根均方差(表示在考虑当前题目的情况下历史答题情况与当前预测值的一致性).

3) 将所有学生对应的根均方差以散点图和盒式图的形式展示出来,如图4所示.从图4中可以看出,在所有数据集中,高注意力权重组对应的根均方差均值明显小于其他组(即其答题情况与预测结果的一致性最好),且中注意力权重组对应的根均方差均值同样明显小于低注意力权重组.



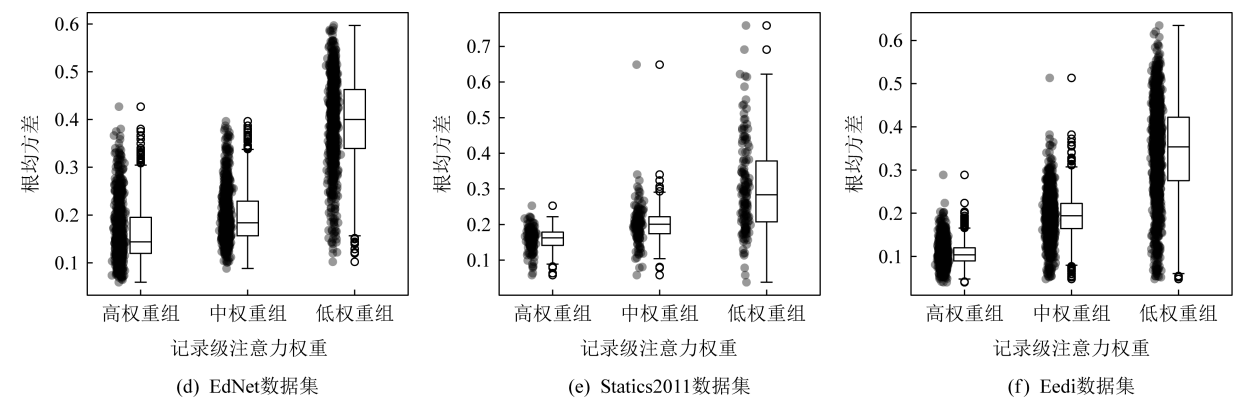


Fig. 4 Analysis of the consistency between historical answers and predicted probability

图 4 历史答题情况与预测概率的一致性分析

进一步,采用双样本  $T$  检验对 3 组的差距进行显著性分析.对高、中组和 中、低组分别进行检验分析得  $p$  值均远远小于显著性基准值 0.01,说明 3 组的差距是显著的.这一发现表明,从统计意义上来说相关权重更大的历史答题与模型当前的预测结果具有更高的一致性.这说明不同历史记录对当前答题预测的影响是不一样的,所以引入注意力机制能有效利用这一规律,从而提升模型的预测性能和可解释性.

3.5.3 可解释性度量

为了进一步评估和对比各模型对于预测结果的可解释性,在所有数据集上实验并绘制模型保真度随可解释性阈值  $\theta$  的变化曲线,如图 5 所示.其中,  $\text{HAKT}(\lambda)$  表示解释性正则化项中权衡因子为  $\lambda$  的模型,  $\text{HAKT}(0)$  则表示无解释性正则化项的版本; SAKT 和 AKT 是对比模型.从图 5 中可看到:

1) 在其中 4 个数据集上,  $\text{HAKT}(0)$  的保真度均高于 SAKT 和 AKT,说明在不加解释性正则化

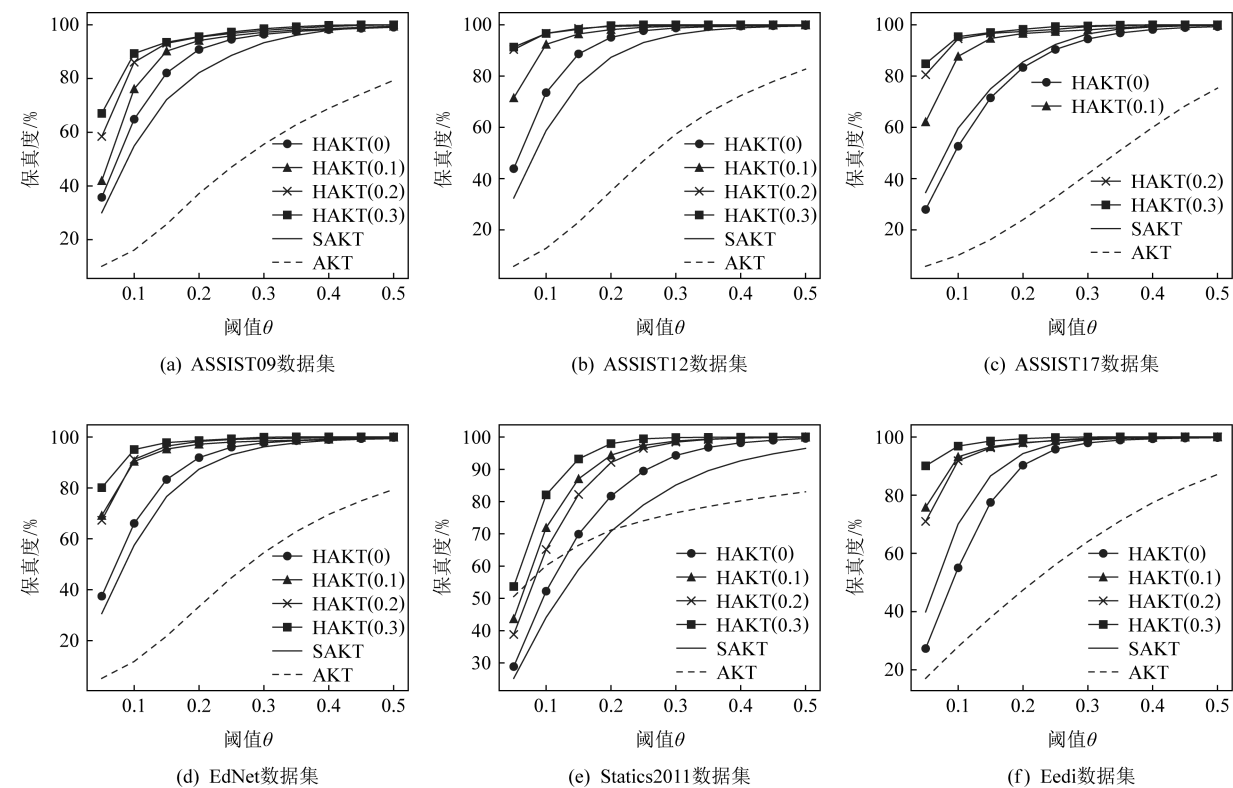


Fig. 5 Variation of fidelity of each model with interpretability threshold

图 5 模型保真度随可解释性阈值的变化

项的情况下,本文所提模型通过建模题目的多语义关联一定程度上提升了模型的可解释性.

2) 当加入解释性正则化项且权衡因子  $\lambda$  增大时,保真度也随之明显上升,这说明引入解释性正则化项显著提升了模型对预测结果的可解释性.注意到 AKT 的保真度总体上较低,潜在原因是 AKT 相比 SAKT 和 HAKT 具有更深的自注意力层和更复杂的结构,由此降低了可解释性.

为了进一步观察 HAKT 中权衡因子  $\lambda$  如何影响预测性能和可解释性,将不同  $\lambda$  值(固定  $\theta=0.20$ )对应的 AUC 和保真度以散点图的形式呈现(如图 6 所示).从图 6 中可以观察到:总体上看,随着  $\lambda$  增

大,AUC 减小而保真度增大.以 ASSIST09 数据集为例,当  $\lambda$  从 0 增大到 0.3 时,AUC 从 79.02%降至 78.81%(下降 0.2 个百分点),而保真度从 89.71%升至 95.52%(上升 5.71 个百分点).这说明模型可解释性与预测性能难以同时提升,两者需要平衡. HAKT 可以通过改变权衡因子调控模型的预测性能和可解释性,在模型预测性能略有下降的情况下,显著提升其可解释性.

在本文所用数据集中,同时考虑模型预测性能和可解释性的情况下,权衡因子  $\lambda$  的较优取值处于 0.05~0.2 之间,此时预测性能略有下降而保真度有较明显的提升.

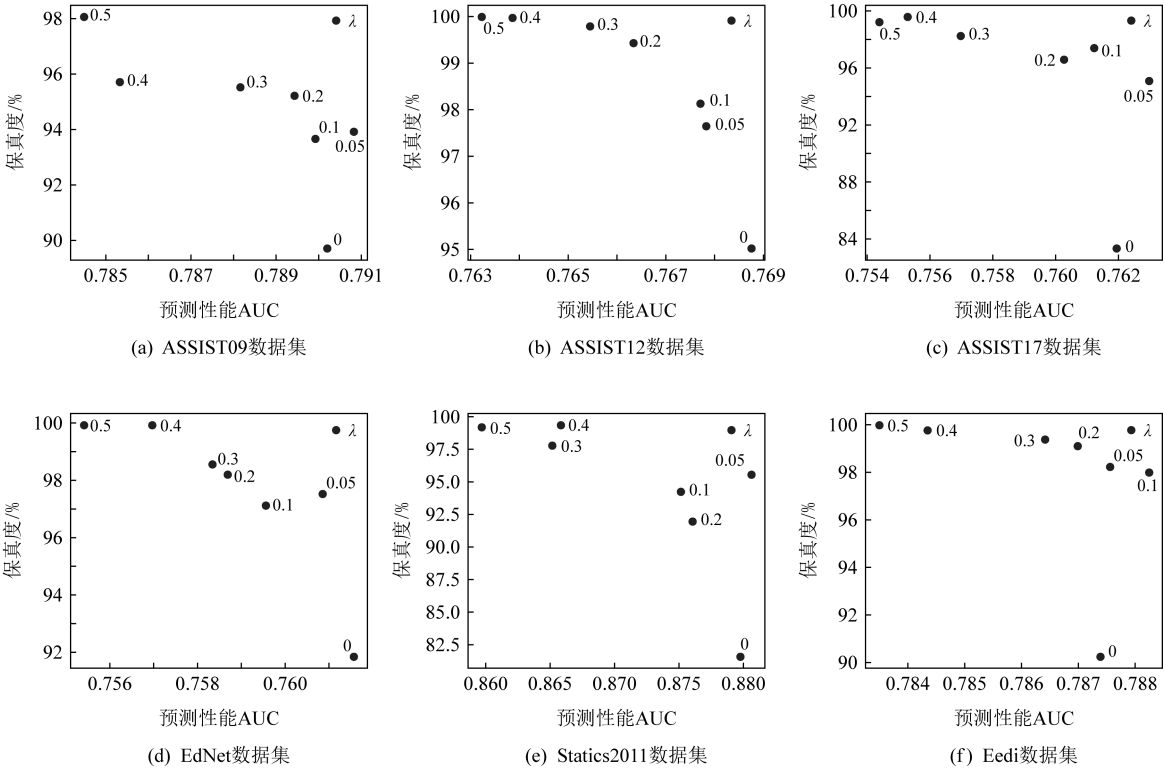


Fig. 6 The impact of trade-off factor  $\lambda$  on predictive performance and interpretability

图 6 权衡因子  $\lambda$  对预测性能和可解释性的影响

4 总 结

针对现有深度认知追踪模型对预测结果普遍缺乏可解释性或者部分模型仅建模题目之间浅层关系的不足,本文提出一种基于多层注意力网络挖掘题目多语义关联的认知追踪模型,进一步提升了模型的预测性能和可解释性.在损失函数中引入解释性正则化项及权衡因子,在预测性能略有下降的情况

下明显提升了其可解释性.设计了评估模型预测结果可解释性的量化指标.在 6 个领域基准数据集上进行了预测性能对比实验和可解释性分析,验证了本文所提模型同时具有较高的预测性能和可解释性.

深度认知追踪是当前国内外智能教育领域的研究热点之一,可有效支撑学习者建模、学习路径规划、学习资源适配等个性化服务.未来,人工智能技术的持续、快速发展,不断为深度认知追踪方法创新提供新的动能.比如旨在突破人工智能非线性瓶颈

的下一代人工智能——精准智能<sup>[42]</sup>,为处理复杂对象可解释性、泛化性与可复现性等难题提供了可能,也为进一步改进深度认知追踪技术带来新的机遇。

**作者贡献声明:**孙建文提出研究问题,设计研究框架,撰写和修改论文,管理研究过程;周建鹏主要负责数据处理、形式化建模、实验设计分析与论文起草;刘三女牙主要提供研究思路与方法指导;何绯娟主要提供研究经费支持,完善研究思路与框架,指导论文修改等;唐云主要提供文献调研、认知建模理论与实验结果分析等指导性支持。

参 考 文 献

[1] Zheng Qinghua, Dong Bo, Qian Buyue, et al. The state of the art and future tendency of smart education [J]. Journal of Computer Research and Development, 2019, 56(1): 209–224 (in Chinese)  
(郑庆华, 董博, 钱步月, 等. 智慧教育研究现状与发展趋势 [J]. 计算机研究与发展, 2019, 56(1): 209–224)

[2] Yang Zongkai. The challenges of personalized learning and their solutions [J]. Chinese Science Bulletin, 2019, 64(Z1): 493–498 (in Chinese)  
(杨宗凯. 个性化学习的挑战与应对 [J]. 科学通报, 2019, 64(Z1): 493–498)

[3] Hu Xuegang, Liu Fei, Bu Chenyang. Research advances on knowledge tracing models in educational big data [J]. Journal of Computer Research and Development, 2020, 57(12): 2523–2546 (in Chinese)  
(胡学钢, 刘菲, 卜晨阳. 教育大数据中认知跟踪模型研究进展 [J]. 计算机研究与发展, 2020, 57(12): 2523–2546)

[4] Li Xiaoguang, Wei Siqi, Zhang Xin, et al. LFKT: Deep knowledge tracing model with learning and forgetting behavior merging [J]. Journal of Software, 2021, 32(3): 818–830 (in Chinese)  
(李晓光, 魏思齐, 张昕, 等. LFKT: 学习与遗忘融合的深度认知追踪模型 [J]. 软件学报, 2021, 32(3): 818–830)

[5] Vie J J, Kashima H. Knowledge tracing machines: Factorization machines for knowledge tracing [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2019: 750–757

[6] Cully A, Demiris Y. Online knowledge level tracking with data-driven student models and collaborative filtering [J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 32(10): 2000–2013

[7] Shen Shuanghong, Liu Qi, Chen Enhong, et al. Learning process-consistent knowledge tracing [C] //Proc of the 27th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2021: 1452–1460

[8] Liu Qi, Shen Shuanghong, Huang Zhenya, et al. A survey of knowledge tracing [J]. arXiv preprint, arXiv:2105.15106, 2021

[9] Atkinson R C, Paulson J A. An approach to the psychology of instruction [J]. Psychological Bulletin, 1972, 78(1): 49–61

[10] Corbett A T, Anderson J R. Knowledge tracing: Modeling the acquisition of procedural knowledge [J]. User Modeling and User-Adapted Interaction, 1995, 4(4): 253–278

[11] Piech C, Bassen J, Huang J, et al. Deep knowledge tracing [C] //Proc of the 29th Int Conf on Neural Information Processing System (NIPS). Cambridge, MA: MIT, 2015: 505–513

[12] Zhang Jiani, Shi Xingjian, King I, et al. Dynamic key-value memory networks for knowledge tracing [C] //Proc of the 26th Int Conf on World Wide Web (WWW). New York: ACM, 2017: 765–774

[13] Abdelrahman G, Wang Qing. Knowledge tracing with sequential key-value memory networks [C] //Proc of the 42nd Int Conf on Research and Development in Information Retrieval (SIGIR). New York: ACM, 2019: 175–184

[14] Liu Sannüya, Zou Rui, Sun Jianwen, et al. A hierarchical memory network for knowledge tracing [J/OL]. Expert Systems with Applications, 2021 [2021-09-01]. <https://doi.org/10.1016/j.eswa.2021.114935>

[15] Pandey S, Karypis G. A self-attentive model for knowledge tracing [C] //Proc of the 12th Int Conf on Educational Data Mining (EDM). Worcester, MA: IEDMS, 2019: 384–389

[16] Lee J, Yeung D Y. Knowledge query network for knowledge tracing: How knowledge interacts with skills [C] //Proc of the 9th Int Conf on Learning Analytics & Knowledge (LAK). New York: ACM, 2019: 491–500

[17] Nakagawa H, Iwasawa Y, Matsuo Y. Graph-based knowledge tracing: Modeling student proficiency using graph neural network [C] //Proc of the 18th Int Conf on Web Intelligence (WI). Piscataway, NJ: IEEE, 2019: 156–163

[18] Ghosh A, Heffernan N, Lan A S. Context-aware attentive knowledge tracing [C] //Proc of the 26th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2020: 2330–2339

[19] Kaser T, Klingler S, Schwing A G, et al. Dynamic bayesian networks for student modeling [J]. IEEE Transactions on Learning Technologies, 2017, 10(4): 450–462

[20] Cen Hao, Koedinger K, Junker B. Learning factors analysis—A general method for cognitive model evaluation and improvement [C] //Proc of the 8th Int Conf on Intelligent Tutoring Systems (ITS). Berlin: Springer, 2006: 164–175

[21] Pavlik P, Cen Hao, Koedinger K. Performance factors analysis: A new alternative to knowledge tracing [C] //Proc of the 14th Int Conf on Artificial Intelligence in Education (AIED). Berlin: Springer, 2009: 531–538

- [22] Huang Zhenya, Yin Yu, Chen Enhong, et al. EKT: Exercise-aware knowledge tracing for student performance prediction [J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(1): 100-115
- [23] Khajah M, Lindsey R V, Mozer M C. How deep is knowledge tracing [C] //Proc of the 9th Int Conf on Educational Data Mining(EDM). Worcester, MA: IEDMS, 2016: 94-101
- [24] Pandey S, Srivastava J. RKT: Relation-aware self-attention for knowledge tracing [C] //Proc of the 29th ACM Int Conf on Information & Knowledge Management (CIKM). New York: ACM, 2020: 1205-1214
- [25] Sun Jianwen, Zhou Jianpeng, Zhang Kai, et al. Collaborative embedding for knowledge tracing [C] //Proc of the 14th Int Conf on Knowledge Science, Engineering and Management (KSEM). Berlin: Springer, 2021: 333-342
- [26] Zhang Nan, Du Ye, Deng Ke, et al. Attention-based knowledge tracing with heterogeneous information network embedding [C] //Proc of the 13th Int Conf on Knowledge Science, Engineering and Management (KSEM). Berlin: Springer, 2020: 95-103
- [27] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks [C/OL] //Proc of the 6th Int Conf on Learning Representations (ICLR), 2018 [2021-09-01]. <https://openreview.net/pdf?id=rJXMpikCZ>
- [28] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] //Proc of the 31st Int Conf on Neural Information Processing Systems(NIPS). Cambridge: MIT Press, 2017: 6000-6010
- [29] Wang Tianqi, Ma Fenghong, Gao Jing. Deep hierarchical knowledge tracing [C] //Proc of the 12th Int Conf on Educational Data Mining(EDM). Worcester, MA: IEDMS, 2019: 671-674
- [30] Liu Yunfei, Yang Yang, Chen Xianyu, et al. Improving knowledge tracing via pre-training question embeddings [C] //Proc of the 29th Int Joint Conf on Artificial Intelligence (IJCAI). San Francisco: Morgan Kaufmann, 2020: 1577-1583
- [31] Yang Yang, Shen Jian, Qu Yanru, et al. GIKT: A graph-based interaction model for knowledge tracing [C] //Proc of the Joint European Conf on Machine Learning and Knowledge Discovery in Databases (ECML PKDD). Berlin: Springer, 2020: 299-315
- [32] Zhang Yu, Tiño P, Leonardis A, et al. A survey on neural network interpretability [J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2021, 5(5): 726-742
- [33] Chen Kerui, Meng Xiaofeng. Interpretation and understanding in machine learning [J]. Journal of Computer Research and Development, 2020, 57(9): 1971-1986 (in Chinese)  
(陈珂锐, 孟小峰. 机器学习的可解释性[J]. 计算机研究与发展, 2020, 57(9): 1971-1986)
- [34] Yeung C K. Deep-irt: Make deep learning based knowledge tracing explainable using item response theory [C] //Proc of the 12th International Conference on Educational Data Mining(EDM). Worcester, MA: IEDMS, 2019: 683-686
- [35] Converse G, Pu Shi, Oliveira S. Incorporating item response theory into knowledge tracing [C] //Proc of the 22nd Int Conf on Artificial Intelligence in Education (AIED). Berlin: Springer, 2021: 114-118
- [36] Wang Fei, Liu Qi, Chen Enhong, et al. Neural cognitive diagnosis for intelligent education systems [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2020: 6153-6161
- [37] Su Yu, Cheng Zeyu, Luo Pengfei, et al. Time-and-concept enhanced deep multidimensional item response theory for interpretable knowledge tracing [J/OL]. Knowledge-Based Systems, 2021 [2021-09-01]. <https://doi.org/10.1016/j.knosys.2021.106819>
- [38] Lu Yu, Wang Deliang, Meng Qinggang, et al. Towards interpretable deep learning models for knowledge tracing [C] //Proc of the 21st Int Conf on Artificial Intelligence in Education(AIED). Berlin: Springer, 2020: 185-190
- [39] Wang Xiao, Ji Houye, Shi Chuan, et al. Heterogeneous graph attention network [C] //Proc of the 28th Int Conf on World Wide Web Conference(WWW). New York: ACM, 2019: 2022-2032
- [40] Xiong Xiaolu, Zhao Siyuan, Van Inwegen E G, et al. Going deeper with deep knowledge tracing [C] //Proc of the 9th Int Conf on Educational Data Mining(EDM). Worcester, MA: IEDMS, 2016: 545-550
- [41] Peake G, Wang Jun. Explanation mining: Post hoc interpretability of latent factor models for recommendation systems [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2018: 2060-2069
- [42] Zheng Zhiming, Lü Jinhu, Wei Wei, et al. Refined intelligence theory: artificial intelligence regarding complex dynamic objects [J]. SCIENTIA SINICA Informationis, 2021, 51(4): 678-690 (in Chinese)  
(郑志明, 吕金虎, 韦卫, 等. 精准智能理论: 面向复杂动态对象的人工智能[J]. 中国科学: 信息科学, 2021, 51(4): 678-690)



**Sun Jianwen**, born in 1982. PhD, associate professor, PhD supervisor. Member of CCF. His main research interests include educational data mining, intelligent tutoring system, and smart education.

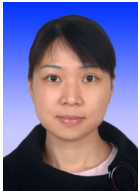
孙建文, 1982年生. 博士, 副教授, 博士生导师. CCF会员. 主要研究方向为教育数据挖掘、智能导学系统和智慧教育。



**Zhou Jianpeng**, born in 1996. Master candidate. His main research interests include knowledge tracing, recommender system, and smart education.  
**周建鹏**, 1996 年生. 硕士研究生. 主要研究方向为认知追踪、推荐系统和智慧教育.



**Liu Sannüya**, born in 1973. PhD, professor, PhD supervisor. His main research interests include computer application, artificial intelligence, and educational information technology.  
**刘三女牙**, 1973 年生. 博士, 教授, 博士生导师. 主要研究方向为计算机应用、人工智能和教育信息技术.



**He Feijuan**, born in 1977. MS, associate professor. Her main research interests include educational data mining and text mining.  
**何绯娟**, 1977 年生. 理学硕士, 副教授. 主要研究方向为教育数据挖掘和文本挖掘.



**Tang Yun**, born in 1982. PhD, associate professor. Her main research interests include educational data mining, self-regulated learning, and the self-determination theory of motivation.  
**唐云**, 1982 年生. 博士, 副教授. 主要研究方向为教育数据挖掘、自我调节学习和自我决定动机理论.

## 2022 年《计算机研究与发展》专题(正刊)征文通知

### ——数据挖掘前沿进展

数据挖掘旨在利用智能数据分析技术,从海量数据中提取或挖掘潜在的知识和规律,为决策任务提供有效支撑.在大数据时代背景下,数据挖掘技术已在工业、金融、医疗、教育、交通、媒体等领域取得广泛应用.然而,大数据的复杂多样性也为数据挖掘研究在理论、方法、应用等多个层面提出了新的挑战.

为及时反映国内同行在数据挖掘方面的前沿研究成果,《计算机研究与发展》将于 2022 年出版“数据挖掘前沿进展”专题.欢迎数据挖掘及相关研究领域的专家学者、科研人员踊跃投稿.此外,专题组稿与第九届中国数据挖掘会议(CCDM 2022)合作,所有专题录用稿件均需在 CCDM 2022 会议报告交流.

#### 征文范围(但不限于)

##### 1) 数据挖掘理论与方法

分类、聚类、排序、集成学习、强化学习、关联分析、链接分析、频繁模式挖掘、动态数据挖掘、交互式与联机挖掘、并行与分布式挖掘、大规模数据挖掘、数据预处理、数据挖掘语言等.

##### 2) 特定类型数据挖掘与分析

关系数据挖掘、图模式挖掘、空间与时序数据挖掘、数据流与增量挖掘、多媒体数据挖掘、社交网络分析与挖掘、文本挖掘、隐私保护数据挖掘、生物信息数据挖掘、推荐系统、数据仓库等.

##### 3) 数据挖掘技术应用

面向工业、金融、医疗、教育、交通、旅游、管理、电商、电信等领域的数据挖掘技术及应用.

#### 征文要求

1) 论文应属于作者的科研成果,数据真实可靠,具有重要的学术价值与推广应用价值,且未在国内外公开发行的刊物或会议上发表,不存在一稿多投问题.作者在投稿时,需向编辑部提交版权转让协议.

2) 论文一律用 word 格式排版,论文格式体例参考近期出版的《计算机研究与发展》的要求(<http://crad.ict.ac.cn/>).

3) 论文须通过期刊网站(<http://crad.ict.ac.cn/>)投稿,投稿时提供作者的联系方式,留言中务必注明“数据挖掘 2022 专题”(否则按自由来稿处理).

4) 论文预录用后,至少有一位作者注册参加 CCDM 2022 会议并做口头报告.否则,论文视为退出专刊.

#### 重要日期

论文截稿日期:2022 年 1 月 8 日

录用通知日期:2022 年 2 月 28 日

最终稿提交日期:2022 年 3 月 10 日

论文出版日期:2022 年 8 月

#### 特邀编委

张长水 教授 清华大学 zcs@mail.tsinghua.edu.cn

杨博 教授 吉林大学 ybo@jlu.edu.cn

钱超 副教授 南京大学 qianc@nju.edu.cn

#### 联系方式

编辑部:crad@ict.ac.cn, 010-62620696, 010-62600350

通信地址:北京 2704 信箱《计算机研究与发展》编辑部

邮编:100190