

学习者学习行为建模： 一种基于预训练模型的可解释性知识追踪模型

周涛, 李艳婷, 杨月婷, 任俊霖, 琚生根, 师维
四川大学计算机学院, 四川 成都 610065

摘要

知识追踪旨在追踪学习者在学习过程中的知识状态变化, 从而提升学习效率。当前, 知识追踪因其在教育领域的重要意义而备受关注。然而, 基于深度学习的知识追踪模型大多致力于追求学习者成绩预测的高准确性, 在可解释方面存在明显不足。如何在确保知识追踪模型可解释能力的同时, 尽可能提升模型性能, 仍然是一个挑战。基于此, 首先基于IRT和Transformer架构提出一个精心设计的高性能可解释知识追踪模型TIKT; 然后使用经微调的预训练模型, 预测题目文本中隐含的题目难度等级, 以提升数据的可解释能力, 并将题目难度等级引入对比学习框架中, 从而增强知识追踪模型的整体性能。在3个真实世界基准数据集上的实验充分证明, 所提出的模型在可解释性与性能方面相较其他模型均有一定提升。

关键词

知识追踪; 对比学习; 预训练模型

中图分类号: TP399

文献标志码: A

doi:10.11959/j.issn.2096-0271.2025055

Learner behavior modeling: an interpretable knowledge tracking model based on pretrained model

ZHOU Tao, LI Yanting, YANG Yueting, REN Junlin, JU Shenggen, SHI Wei
College of Computer Science, Sichuan University, Chengdu 610065, China

Abstract

Knowledge tracing aims to track the transformation of learners' knowledge states during the learning process, thereby enhancing learners' learning efficiency. Recently, due to its significant importance in education, knowledge tracing has attracted considerable attention from researchers. However, deep learning-based knowledge tracing models primarily focus on achieving high accuracy in predicting learners' performance, but they significantly lack interpretability. Balancing the interpretability of knowledge tracing models while enhancing their performance remains a challenge. Based on this, a meticulously designed, high-performance, interpretable knowledge tracing model TIKT, was firstly proposed using item response theory (IRT) and Transformer. Subsequently, a fine-tuned pretrained model to predict the implicit difficulty levels of question texts to enhance data-side interpretability was used, and incorporated these

difficulty levels into a contrastive learning framework to boost the overall performance of the knowledge tracing model. Experiments conducted on three real-world benchmark datasets conclusively demonstrate that our proposed model achieves improvements in both interpretability and performance relative to other models.

Key words

knowledge tracing, contrastive learning, pretrained model

0 引言

随着信息技术的迅速发展,传统的线下教育正不断向在线智能教育迈进,在线学习已经成为用户获取知识的重要途径之一。尽管在线学习提供了巨大的便利,但其通常缺乏专家的专业指导和合理的课程规划,使用户在面向大量的学习资源时难以有效选择和利用,这大大降低了用户的学习积极性。智能教学系统(intelligent tutoring system, ITS)作为智能教育的重要组成部分,从收集的海量信息中进一步挖掘学生的相关信息,动态建模学习者的知识水平和学习偏好。知识追踪^[1]则是实现该目标的重要方法,是构建智能教学系统的关键技术,为学习路径规划和题目推荐等相关研究奠定了基础。知识追踪使用一系列机器学习方法和深度学习方法,从相关教育数据中挖掘学生的动态知识状态。如今,知识追踪被广泛应用于ITS中,并受到越来越多的关注。

知识追踪基于ITS所收集到的历史交互数据,对学生的知识水平进行建模,从而预测学生未来的答题情况。深度学习技术在数据挖掘、自然语言处理等领域取得了不错的成效。Piech等^[2]首次提出运用深度神经网络来构建知识追踪模型,即基于循环神经网络的深度知识追踪(deep knowledge tracing, DKT)。Zhang等^[3]

受记忆增强神经网络的启发,构建了基于记忆网络的知识追踪模型,以提高捕捉时间序列中历史信息的能力。由于Transformer架构在自然语言处理任务上的卓越表现,一些学者尝试将注意力机制引入知识追踪领域^[4],使模型能够在一系列交互中学习问题的相对重要性,进而预测正确回答下一问题的概率。Ghosh等^[5]提出上下文感知的注意力知识追踪(attentive knowledge tracing, AKT)模型,使用单调注意力机制和Rasch模型来生成上下文感知的问题和知识嵌入,从而有效地预测用户的知识状态。

基于深度学习的知识追踪模型通常缺乏可解释性,这些模型决策过程的不透明限制了它们的进一步应用。鉴于此,Yeung^[6]提出了用项目反应理论解释基于深度学习的知识追踪(deep learning based knowledge tracing explainable using item response theory Deep-IRT),其使用动态键值对记忆网络(dynamic key-value memory network, DKVMN)模型来处理学习者的学习序列,据此估计题目难度和随时间变化的学习者能力,再使用IRT模型估计学生正确回答题目的概率。Sun等^[7]提出基于多尺度状态的可解释知识跟踪(multiscale-state-based interpretable knowledge tracing, MIKT)模型,解决了在知识状态表示上的粒度失衡与可解释性差问题。

题目难度对学习者的实际答题交互表

现具有重大影响。经典测量理论(classical test theory, CTT)和IRT是计算测试项难度的常用方法,如AKT嵌入层使用的Rasch模型也被称为单参数项目反应理论。Tian等^[8]在双向编码表示知识追踪(deep knowledge tracing with bidirectional encoder representations from transformers, BEKT)模型中采用CTT计算题目的困难度特征 d_i ,并用于时间步 t 输入元组对 $(q_i, c_i, r_i, d_i, p_i)$ 。Shen等^[9]提出难度匹配知识追踪(difficulty matching knowledge tracing, DIMKT)模型,从学习者的历史答题序列中挖掘知识概念难度和问题难度,并嵌入知识追踪模型中。王士进等^[10]将知识概念难度和题目难度用于学习者的知识画像建模。此外,梁祥等^[11]提出融合题目难度与作答经验的深度知识追踪模型(deep knowledge tracing model integrating problem difficulty and answering experience, DKT-DE),将题目难度预测转变为分类任务,并将预测题目难度等级嵌入DKT中。

尽管这些研究取得了不错的学习成效,但仍存在以下挑战:(1)当前题目难度主要基于CTT或将题目难度作为可训练参数进行计算,然而,题目的难度与其文本内容直接相关,但很少有研究基于题目文本内容来预测题目难度;(2)题目难度作为重要的可解释参数,如何将其引入知识追踪下的对比学习中,以进一步提升模型的整体性能;(3)考虑可解释性的知识追踪模型通常会牺牲性能,如何设计高性能可解释知识追踪模型,仍是一个巨大挑战。

本文提出了一种基于Transformer架构和预训练模型的可解释知识追踪新方法TIKT,旨在弥补上述不足。首先,将使用CTT得到的题目难度作为预训练模型的

微调目标,将题目的文本特征作为输入,通过微调后的预训练模型得到题目难度等级预测值,并将题目难度等级引入对比学习模块中,辅助模型训练;其次,为确保可解释性,TIKT引入IRT预测模块来解释模型预测过程。

本文的主要贡献如下。

- 所提出的TIKT模型性能不仅优于基于深度学习的知识追踪模型,同时能有效推断学习者的能力和题目难度,进一步增强模型的可解释性。
- 本文通过对预训练语言模型进行微调,使用微调后的模型预测题目文本中隐含的难度等级,从而在数据层面提供辅助解释;随后,这些难度等级被整合到一个融合题目难度的对比学习模块中,进一步提升了TIKT模型的性能。
- 本文在3个广泛使用的公共数据集上进行实验,并与3个最先进的知识追踪模型进行比较,实验结果表明,TIKT模型在预测学习者表现与可解释性方面表现出色。

1 相关工作

1.1 上下文感知的注意力知识追踪

上下文感知的注意力知识追踪(attentive knowledge tracing, AKT)基于学习者的知识记忆随时间推移逐渐遗忘的假设,提出单调注意力机制(缩放点积注意力的单调版本)。由于绝对时间距离度量 $d(t, \tau) = |t - \tau|$ 不带有上下文感知能力,并且忽略了每位学习者过去练习知识概念间的相互关系,为了反映该关系,AKT提出一种新的基于上下文感知的距离度量,表示如下。

$$d(t, \tau) = |t - \tau| \cdot \sum_{t'=\tau+1}^t \gamma_{t,t'} \quad (1)$$

$$\gamma_{t,t'} = \frac{\exp\left(\frac{\mathbf{q}_t^\top \mathbf{k}_{t'}}{\sqrt{D_k}}\right)}{\sum_{1 \leq \tau' \leq t} \exp\left(\frac{\mathbf{q}_t^\top \mathbf{k}_{\tau'}}{\sqrt{D_k}}\right)}, \forall t' \leq t$$

其中, t 与 τ 表示时间步, $\mathbf{q}_t$ 与 $\mathbf{k}_\tau$ 分别表示在时间 t 与 τ 上对应的查询与键值, D_k 表示 $\mathbf{k}$ 的维度。

根据过去练习的知识点与当前练习知识点的相关性, 以及二者之间的相对时间步数计算注意力值, 得到上下文感知的距离度量, 并引入乘法指数衰减函数, 最终得到单调注意力机制, 其计算式如下所示。

$$\alpha_{t,\tau} = \frac{\exp(s_{t,\tau})}{\sum_{\tau'} \exp(s_{t,\tau'})} \quad (2)$$

$$s_{t,\tau} = \frac{\exp(-\theta \cdot d(t, \tau)) \cdot \mathbf{q}_t^\top \mathbf{k}_\tau}{\sqrt{D_k}}$$

其中, $\theta > 0$ 为可学习的衰减率参数。对于 AKT 中的问题和知识编码器, 满足 $\tau \leq t$; 对于知识检索器, 满足 $\tau < t$ 。

1.2 可解释知识追踪

可解释知识追踪 (interpretable knowledge tracing, IKT) 是指在知识追踪任务中, 不仅能评估学生的知识掌握情况, 还能提供清晰、直观的解释, 以阐明模型是如何得出预测结果的, 从而促进个性化教学并制定有效的学习策略。MIKT 中, IKT 被分为如下 3 类。

(1) 后验局部解释: 通过分析输入特征与输出结果之间的关系, 确定哪些特征对预测结果的贡献最大, 以解释模型为何做出此预测或决策。例如, Lu 等^[12]使用逐层相关性传播 (layer-wise relevance

propagation, LRP) 技术^[13]将相关性分数从模型的输出层传播到输入层, 以解释 KT 模型。

(2) 具有可解释结构的全局可解释性: 通过设计一种具有内在可解释能力的模型结构, 从整体上理解和解释知识状态建模的过程。例如, Wang 等^[14]在模型中嵌入了一个可解释的认知框架, 以理解学生知识状态建模的过程。

(3) 具有可解释参数的全局可解释性: 通过设计和使用模型派生的可解释参数, 从整体上解释其预测。在深度知识追踪领域, 提升模型可解释的一种常用方式是与项目反应理论 (item response theory, IRT) 等经典方法进行融合, 以解决 IRT 无法动态建模知识状态的问题。这些经典理论的内在参数通常具有可解释性, 如问题中心可解释知识追踪 (question-centric interpretable KT, QIKT) 导出学生的 3 个可解释能力评分, 并在输出层使用这些评分进行预测^[15]。单参数 IRT 的项目反应函数如下所示。

$$\hat{r}_{t+1} = p(r_{t+1} = 1 | \theta) = \frac{1}{1 + e^{-D(\theta - \beta)}} \quad (3)$$

其中, 学习者正确回答题目的概率 $\hat{r}_{t+1}$ 由学习者能力 θ 与题目难度 β 共同决定, D 为常量。当 D 取值为 1.702 时, 此函数的概率密度与正态肩型曲线的差异小于 0.01。

1.3 对比学习

对比学习 (contrastive learning, CL) 是自监督学习的一个特殊分支^[16], 在众多研究领域取得了很好的效果, 如计算机视觉^[17]、自然语言处理^[18]和推荐系统^[19]。CL 是一种通过比较正样本和负样本来学习表示的方法^[20]。研究表明, 针对特定任务采用合适的数据增强方法对于成功

使用CL至关重要。当前已经有各种增强方法被使用,如计算机视觉领域的图像裁剪、旋转和颜色失真,以及自然语言处理和推荐系统的掩蔽单词或特征等。在知识追踪领域,已有学者使用对比学习进行了相关研究,如Lee等^[21]提出了基于对比学习的知识追踪(contrastive learning for knowledge tracing, CL4KT)模型,在对比学习中使用多种数据增强技术来生成正样本,并通过反转响应来构造负样本。

2 基于预训练模型的可解释知识追踪模型

本节详细介绍基于Transformer和预训练模型的可解释知识追踪模型TIKT,首先,阐述了整个模型架构与知识追踪问题定义;随后,全面概述了TIKT中的3个模块(响应预测模块、对比学习模块、基于预训练语言模型与题目文本特征的题目难度等级预测模块);最后,概述了模型的整个训练过程。

2.1 问题定义

在知识追踪任务中,定义学习者与系统的历史交互序列为 $s=(s_1, s_2, \dots, s_t)$ 。其中, $s_t=(q_t, c_t, d_t, r_t)$ 是一个交互四元组,该四元组由表示特定题目的题目索引 $q_t \in N^+$ 、表示题目涉及核心概念的概念索引 $c_t \in N^+$ 、表示题目难易程度的题目难度索引 $d_t \in N^+$ 以及表示学习者对该题目的回答正确与否的响应结果 $r_t \in \{0, 1\}$ 组成。

本研究旨在利用历史学习交互序列,结合特定时间步下的题目、知识概念和题目难度,构建一个知识追踪模型,如下所示。

$$\hat{r}_{t+1}=p(r_{t+1}=1|s_1, s_2, \dots, s_t, q_{t+1}, c_{t+1}, d_{t+1}) \quad (4)$$

其中, $\hat{r}_{t+1}$ 表示学习者正确回答下一个问题 q_{t+1} 的概率。

2.2 模型架构概述

模型整体架构如图1所示。学习者在与习题交互过程中,习题难度与学习者能否正确回答问题有紧密的联系,并能从数据侧辅助解释模型预测结果。因此,本文设计了基于预训练模型Bert和题目文本特征的题目难度等级预测模块,将题目难度预测值作为辅助参数,用于解释模型预测结果;同时,作为对比学习模块的输入参数用于辅助响应预测模型训练,得到更优的题目难度表征。本文提出的知识追踪模型由响应预测模块和对比学习模块组成,其中响应预测模块包括嵌入层、可解释参数预测层和IRT预测层;对比学习模块则由3个相同的基于Transformer的子模块构成,用于辅助模型训练。

2.3 响应预测模块结构

响应预测模块旨在分析学习者的历史交互数据,准确预测该学习者是否能正确回答即将到来的问题。响应预测模块由嵌入层、可解释参数预测层和IRT预测层构成。

2.3.1 嵌入层

嵌入层负责将题目、知识概念、学习者响应编码为嵌入向量。定义题目嵌入矩阵 $E^Q = \mathbb{R}^{N \times d}$ 、概念嵌入矩阵 $E^C = \mathbb{R}^{K \times d}$ 、响应预测模块可训练题目难度嵌入矩阵 $E^D = \mathbb{R}^{N \times d}$ 、对比学习模块难度等级嵌入矩阵 $E^{DL} = \mathbb{R}^{M \times d}$,以及学习者对题目响应嵌入矩

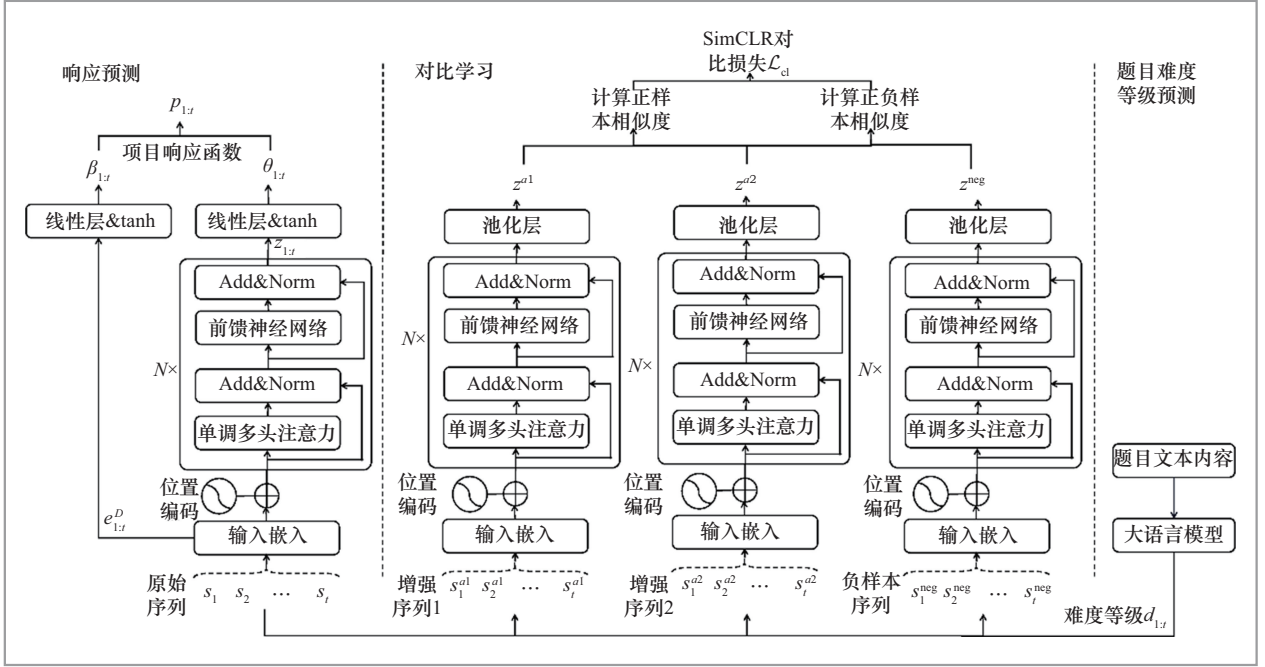


图1 模型整体架构

阵 $\mathbf{E}^R = \mathbb{R}^{O \times d}$ 。其中， N 、 K 、 M 和 O 分别表示题目数量、知识概念数量、题目难度等级数与响应类别数， d 为嵌入向量维度。为学习到更丰富且精准的题目难度表征，本文在对比学习框架中引入了基于难度等级的嵌入向量，辅助响应预测模块中可训练题目难度嵌入矩阵 $\mathbf{E}^D$ 的学习。具体而言，通过对比学习模块的难度等级嵌入 $\mathbf{E}^{DL}$ ，能够为响应预测模块提供更多维度的难度信息，帮助其捕捉同一难度层级下不同题目之间的微妙差异，从而学到更具区分度和泛化能力的难度表示，并在模型整体性能上取得提升。本文默认将题目难度分为100个等级，响应有正确或错误两种类别。

对于给定的交互四元组 $s_t = (q, c, d, r_t)$ ，其中 d_t 来自 Bert 模型难度预测。定义题目嵌入向量 $\mathbf{e}_t^O = \mathbf{E}_q^O \in \mathbb{R}^d$ 、概念嵌入向量 $\mathbf{e}_t^C = \mathbf{E}_c^C \in \mathbb{R}^d$ 、题目难度嵌入向量 $\mathbf{e}_t^D = \mathbf{E}_{d_t}^D \in \mathbb{R}^d$ 、题目难度等级嵌入向量 $\mathbf{e}_t^{DL} = \mathbf{E}_{d_t}^{DL} \in \mathbb{R}^d$ 以及学

习者响应嵌入向量 $\mathbf{e}_t^R = \mathbf{E}_{r_t}^R \in \mathbb{R}^d$ 。对比学习模块嵌入层通过式 (5) 融合题目、知识概念、题目难度与学习者响应结果信息，并输出到下个模块，其中 $\text{avg}()$ 表示计算平均值。

$$\mathbf{e}_t = \text{avg}(\mathbf{e}_t^O, \mathbf{e}_t^C, \mathbf{e}_t^{DL}, \mathbf{e}_t^R) \quad (5)$$

响应预测模块使用公式 (6) 融合题目、知识概念与学习者响应结果，用于预测学习者随时间变化的能力值。

$$\mathbf{e}_t = \text{avg}(\mathbf{e}_t^O, \mathbf{e}_t^C, \mathbf{e}_t^R) \quad (6)$$

2.3.2 可解释参数预测层

本文提出的 KT 模型基于 Transformer 编码器架构，并融合 AKT 模型单调多头注意力机制。输入融合题目、知识概念与学习者响应的嵌入向量序列 $\mathbf{e}_{1:t}$ 到 Transformer 编码器中，输出一个编码后的序列 $\mathbf{z}_{1:t}$ 。由于输入包含学习者作答题目、相关知识概念、题目响应结果，因此，

$z_{1:t}$ 可以用来推断学习者在题目 q_t 上的能力。在传统方法中,题目难度往往只由离散的难度等级标签来表征。这些标签或许可以区分整体的难度层次,但难以刻画同一难度等级下不同题目之间的细微差异。为进一步提升模型能力,本文不使用事先预测的静态难度等级,而是引入可训练的难度嵌入矩阵,在对比学习框架下为每道题目学习到更具区分度的难度表示。基于该可训练题目难度嵌入矩阵表征的题目难度,使用前馈神经网络来预测。根据神经网络的用途,将这两个网络分为学习者能力网络和题目难度网络。两个网络使用单层全连接结构,其表达式如下所示。

$$\theta_{1:t} = \tanh(W_{\theta,1:t} z_{1:t} + b_{\theta,1:t}) \quad (7)$$

$$\beta_{1:t} = \tanh(W_{\beta,1:t} e_{1:t}^D + b_{\beta,1:t}) \quad (8)$$

其中, $\theta_{1:t}$ 与 $\beta_{1:t}$ 分别表示时间步1到 t 对应的学习者能力值与题目难度值。

2.3.3 IRT 预测层

IRT 预测层负责预测学习者正确回答问题的概率。预测层将预测的 $\theta_{1:t}$ 与 $\beta_{1:t}$ 传入项目反应函数式(9),以计算学习者正确回答时间步1到 t 对应题目的概率值。

$$p_{1:t} = \sigma(\alpha \times \theta_{1:t} - \beta_{1:t}) \quad (9)$$

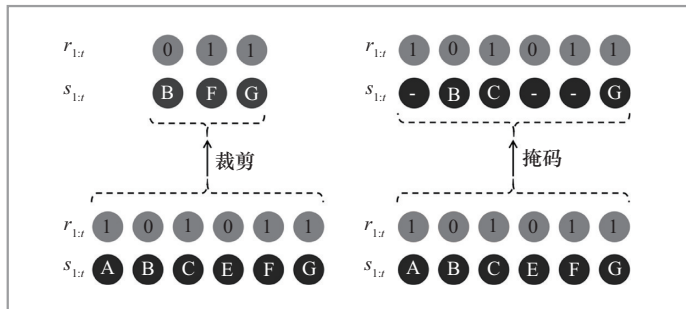


图2 数据增强模块的具体操作

其中, $p_{1:t}=(\hat{r}_1, \hat{r}_2, \dots, \hat{r}_t)$ 表示时间步1到 t 对应的正确概率序列, α 为一个固定常数,对学习者的能力进行缩放。如果不对学生能力进行缩放,最大输出概率仅为 $\sigma(1-(-1))=0.8808$ 。

对于整个响应预测模块,本文采用了二值交叉熵(binary cross-entropy, BCE)作为损失函数进行训练,损失函数如下所示。

$$\mathcal{L}_{\text{main}} = \sum_t -(r_t \ln(\hat{r}_t) + (1-r_t) \ln(1-\hat{r}_t)) \quad (10)$$

2.4 对比学习框架

CL是一种有效的机器学习技术,它通过比较不同样本之间的相似性和差异性来学习区分性的特征表示。这种方法致力于让特征空间中相似样本的表示更加紧密,同时使不相似样本的表示相距更远,从而提高模型的泛化能力和鉴别力。本文通过如下数据增强技术来生成对比学习所需的正样本,数据增强模块的具体操作如图2所示。

(1) 知识概念和题目掩码:通过随机掩盖部分知识概念或题目序列内容,迫使模型在部分信息缺失的情况下仍能正确理解和处理数据。

(2) 裁剪:从序列中裁剪出部分内容,以增强模型对关键信息的聚焦能力。

以上方法生成的正样本对 (s_t^{a1}, s_t^{a2}) 能有效保留原始样本的主要特征和语义联系,而负样本 s_t^{neg} 则分别通过 $d_t^{\text{neg}} = M - d_t$ 和 $r_t^{\text{neg}} = 1 - r_t$ 反转题目难度等级和响应结果来生成。将 s_t^{a1} 、 s_t^{a2} 和 s_t^{neg} 分别传入3个编码模块,得到正样本表示 h_t^{a1} 与 h_t^{a2} ,以及负样本表示 h_t^{neg} 。最后,通过式(11)计算整个序列的表示:

$$\begin{aligned} z^{a1} &= \text{pool}(h_{1:t}^{a1}) & z^{a2} &= \text{pool}(h_{1:t}^{a2}) \\ z^{\text{neg}} &= \text{pool}(h_{1:t}^{\text{neg}}) \end{aligned} \quad (11)$$

其中, $\text{pool}(\cdot)$ 表示平均池化。最后的输出 $z^{a1} \in \mathbb{R}^d, z^{a2} \in \mathbb{R}^d$ 与 $z^{\text{neg}} \in \mathbb{R}^d$ 用于对比学习损失值计算。

$$\mathcal{L}_{\text{cl}} = -\ln \frac{e^{\text{sim}(z^{a1}, z^{a2})}}{e^{\text{sim}(z^{a1}, z^{a2})} + \sum e^{\text{sim}(z^{a2}, z^{\text{neg}})}} \quad (12)$$

2.5 基于预训练模型习题困难度预测

在知识追踪领域, 准确评估题目难度是实现个性化学习与有效学习路径优化的关键, 题目难度会显著影响学习者的答题表现, 以及从数据侧为知识追踪模型预测结果提供解释。题目文本隐含丰富的信息, 能直接反映题目难度。基于此认识, 本文从习题文本内容及其相关知识视角出发, 对题目难度进行建模。

定义题目难度估计训练数据集 $D_{\text{train}} = (q_{\text{text}}^i, c_{\text{text}}^i, d^i)_{i=1}^N$, 其中, N 为训练集总题目数, q_{text} 与 c_{text} 分别表示题目文本与相关联的知识概念文本, d^i 为训练数据集中错误回答问题 i 的比率。从 HuggingFace 下载预训练中文 Bert 模型, 并在 D_{train} 上进行微调, 从而使微调后的 Bert 模型能够根据题目文本与相关概念估计题目难度, 模型表达式如下所示。

$$\hat{d}^i = \text{Bert}(q_{\text{text}}^i, c_{\text{text}}^i) \quad (13)$$

其中, $\text{Bert}(\cdot)$ 表示预训练模型, $\hat{d}^i$ 为 Bert 依据题目和知识概念文本预测的题目难度。选择均方根误差函数作为预训练模型微调的损失函数, 具体表达式如下所示。

$$\mathcal{L}_{\text{diff}} = \frac{1}{N} \sum_{i=1}^N (d^i - \hat{d}^i)^2 \quad (14)$$

最后得到微调后的预训练模型 $\text{Bert}_{\hat{d}}$, 并将其用于数据集的题目难度等级预测。

2.6 模型训练

TIKT 总体目标损失函数被定义为响应预测模块损失 $\mathcal{L}_{\text{main}}$ 与对比学习模块损失 $\mathcal{L}_{\text{cl}}$ 的结合。整个训练过程的损失函数如下所示。

$$\mathcal{L} = (1 - \lambda) \times \mathcal{L}_{\text{main}} + \lambda \times \mathcal{L}_{\text{cl}} \quad (15)$$

其中, 控制超参数 $\lambda \in [0, 1]$ 。

3 实验与结果分析

本节通过在 3 个现实世界数据集上的实验评估 TIKT 模型以及 TIKT-LLM 模型的性能。

3.1 数据集

实验所用的 3 个现实世界数据集相应的统计信息见表 1。

ASSIST09 和 ASSIST12 都是从 ASSISTments 在线辅导平台收集的。对于这两个数据集, 实验删除了缺失知识概念和题目号的记录, 还删除了记录少于 3 条的用户。经过预处理, ASSIST09 数据集

表1 现实世界数据集相应的统计信息

数据集	学生数量/名	题目数量/个	知识概念数/个	交互数/个	题目文本	概念文本
ASSIST09 ^[22]	3 841	15 911	123	190 320	×	✓
ASSIST12 ^[23]	27 405	47 104	265	1 867 167	×	✓
XES3G5M ^[24]	18 066	7 652	865	5 549 635	✓	✓

由 3 841 名学生回答的 123 项知识概念和 15 911 个问题组成, 共包含 190 320 条记录。ASSIST12 数据集包含 265 项知识概念、47 104 个问题以及 27 405 名学生产生的 1 867 167 条记录。XES3G5M 是从好未来教育集团收集的数据集, 该数据集涵盖三年级学生的数学学习成绩, 记录了 500 多万次互动, 并提供了丰富的辅助信息, 包括问题的文本内容、知识成分关系、问题类型和答案分析等。

3.2 评估指标

实验采用五折交叉检验进行性能评估, 此外, 留出 10% 的训练集作为验证集。验证集用于调整超参数以及确定早期停止点。本文比较了所有方法的受试者工作特征曲线下面积 (area under curve, AUC) 和均方根误差 (root mean square error, RMSE)。AUC 与 RMSE 计算式分别如下所示。

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (r_i - \hat{r}_i)^2} \quad (16)$$

$$AUC = \frac{\sum_{i \in \text{position}} \text{rank}_i - N_{\text{pos}}(1 + N_{\text{pos}})/2}{N_{\text{pos}} \times N_{\text{neg}}} \quad (17)$$

其中, N 表示学习者交互记录数, N_{pos} 表示正例的数量, N_{neg} 表示负例的数量。

3.3 基线方法与实验设置

为了验证模型的有效性, 本文与以下 3 种近几年提出的知识追踪模型进行比较。

(1) AKT 模型: 该模型提出一种新颖的单调注意力机制, 通过指数衰减和融合知识相似性的上下文感知的相对距离来计算注意力权重, 并使用 Rasch 模型构建知

识概念和问题嵌入。

(2) CL4KT 模型: 由于学生与问题间的交互稀疏, 知识状态的隐藏表示很容易过拟合, 常无法捕获学生的知识状态, 因此提出对比学习下的知识追踪。

(3) MIKT 模型: 该模型从多尺度考虑结合学习者知识状态, 使用 IRT 模块预测模型输出, 具有极高的性能和可解释性。

本文使用 PyTorch 实现 TIKT, TIKT 中的超参数 λ 设为 0.2, 嵌入大小设为 128, 使用 Adam 对模型进行优化, 批量大小设为 64, 学习率为 0.001, 数据增强中的掩码率以及裁剪率分别设为 0.2 与 0.3, 项目反应函数中的缩放因子 α 设为 3。最后, 所有模型的最大序列长度被设置为 100。所有实验在带有 2.00 GHz Intel (R) Xeon (R) CPU、64 GB 内存与双卡 Nvidia RTX4090 GPU 的 Linux 服务器上进行。

3.4 模型可解释

从 XES3G5M 数据集中随机挑选一条学习者作答序列, 使用 TIKT 模型预测学生正确作答的概率。为了不引入其他知识点带来的干扰, 本文保留原作答序列相同知识点的相对顺序, 按知识点从原序列提取对应知识点相关子序列数据, 并选择序列长度最长的子知识点序列作为分析对象, 知识过滤如图 3 所示。

3.4.1 数据侧可解释

知识点正确响应概率预测如图 4 所示。图 4 展示了某学习者在知识点 78 上的正确作答概率, 横坐标为时间步, 纵坐标为对应的数值; Bert 模型预测的题目难度等级被缩放到 $[0, 1]$ 区间, 数值越大, 难度越高; “+” 节点线条表示 TIKT 预测的正确回答概率, 并通过节点颜色表示真值是正确或

错误，其中黑色节点表示题目回答正确，灰色节点表示回答错误。

从图4中的第8~10以及第12~15时间步不难看出，随着题目难度的降低，学习者正确回答概率也随之增长，反之，正确回答概率则下降。当问题难度超过0.4时，学习者在时间步14出现回答错误，但由于学习者之前的练习都正确，尽管模型仍然输出较高的正确回答概率，却明显低于之前的输出。由于题目难度降低，在时间步15和16时，模型输出比时间步14相对高的概率，但在时间步18，由于学习者在时间步17出现第二次失败并且难度等级达到最高，因此模型输出最低的正确作答概率，这也与实际情况相契合。

响应预测模块与Bert模型预测的题目难度变化趋势如图5所示。图5展示了响应预测模块估计的题目难度与Bert模型预测的题目难度等级的变化趋势。由于响应预测模块输出的题目难度在 $[-1, 1]$ 区间，因此，将Bert模型预测的题目难度缩放到相同区间上。从图5可以看出，两者的变化趋势基本相同，除极个别时间步外，响应预测模块预测的题目难度与Bert模型预测的基本一致。

3.4.2 模型预测可解释

为了进一步证明TIKT的可解释性，本文对随机挑选的知识点78进行了更深入的分析。预测过程解释如图6所示。图6给出了3条线，分别是响应预测模块预测某学习者在知识点78上的正确回答概率、学习者能力值，以及时间步对应的题目难度值。在正确回答概率中，黑色“+”节点表示真值正确，灰色“+”节点表示真值错误。从图6可以清晰看出，当能力值高于题目难度时，学习者有较高的概率回答正确；所回答题目难度显著高于学习者

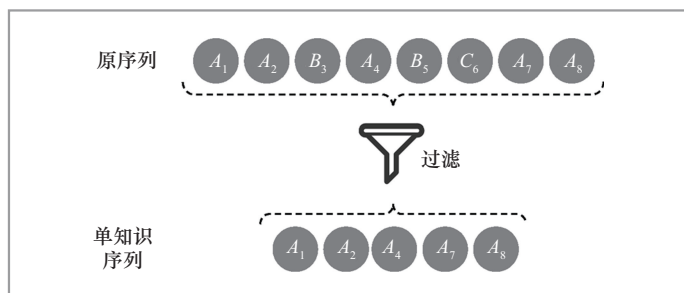


图3 知识过滤

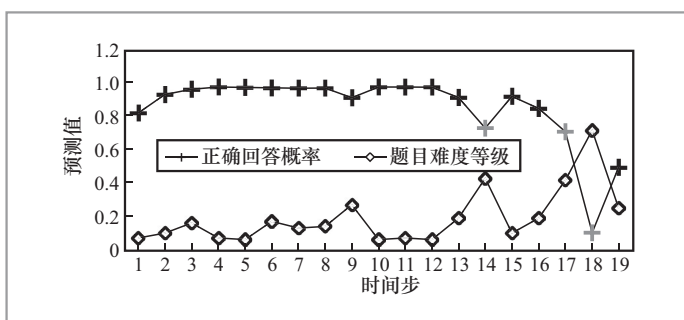


图4 知识点正确响应概率预测

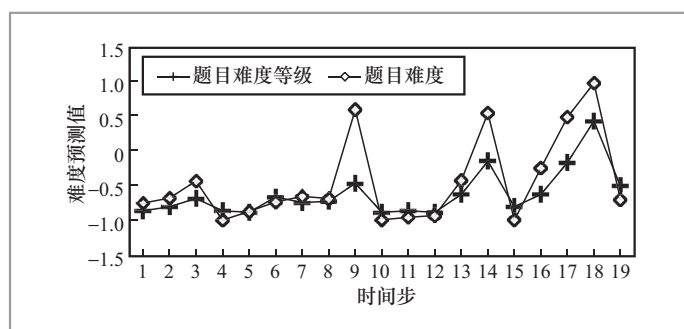


图5 响应预测模块与Bert模型预测的题目难度变化趋势

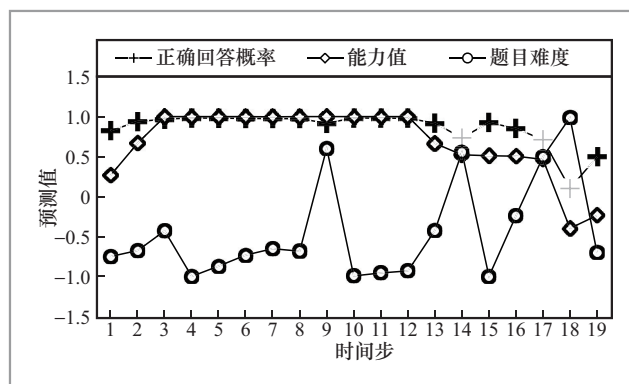


图6 预测过程解释

能力时，正确回答概率急剧下降。显然，TIKT的决策过程是完全透明和可解释的。如果学生回答错误，TIKT可以提供以下解释：（1）学生的能力值相对较低，这可能是由于学生整体水平较低或对特定概念/问题掌握不足；（2）这个问题对该学生来说相对有挑战性。总的来说，学生的能力值低于问题的难度，这使得学生很难正确回答问题。

3.5 模型性能比较

本文基于3个基准数据集来评估各知识追踪模型的整体性能。**表2**展示了各个模型的性能表现，其中最好的性能已用粗体标出。TIKT中对比学习模块默认使用CTT计算的题目难度等级参数，而TIKT-LLM表示在对比学习模块中使用Bert模型预测的题目难度等级参数。由于ASSIST09与ASSIST12缺乏题目文本数据，因此TIKT-LLM未在这两个数据集上进行实验。从**表2**不难看出，TIKT在3个数据集上均取得最好的表现，这表明本文提出的知识追踪模型具有极强的竞争力。即使相较于目前最先进的MIKT算法，TIKT在3个数据集上的各指标均有提升。同时，基于状态顺序和状态图的MIKT需要逐步处理数据，并行度低，训练慢。然

而，基于Transformer架构的TIKT使用注意力机制能同时处理整个输入序列，在保证性能的同时，训练速度大幅提升。在XES3G5M上，MIKT的表现非常差，而TIKT的表现最优；在ASSIST09和ASSIST12上，相较于MIKT，TIKT的AUC分别提升了0.68%和0.7%，而RMSE分别下降了0.001 1与0.005 3；相较于AKT，AUC分别提升了4.36%和0.82%；相较于CL4KT，AUC分别提高了3.29%和0.63%。在XES3G5M上，相较于TIKT，由于TIKT-LLM能生成更优的题目难度特征，从而更好地区分学习者能否回答问题，因此在AUC指标上提升0.18%，虽然提升了正样本预测概率，但其对RMSE指标的贡献较小，导致其提升不明显。这表明，相较于CTT，微调后的Bert模型能更有效地分析题目文本内容，并据此预测题目难度等级，同时，Bert模型的预测结果能从数据侧为TIKT的结果解释提供支撑。

3.6 消融实验

消融实验结果见**表3**。为了验证对比学习模块及题目难度等级各自对TIKT模型的影响，本文比较了从对比学习模块移除题目难度等级以及整个对比学习模块后的

表2 模型性能比较

数据集	指标	AKT	CL4KT	MIKT	TIKT	TIKT-LLM
ASSIST09	AUC	0.746 1	0.756 8	0.7829	0.7897	—
	RMSE	0.440 3	0.435 2	0.4155	0.4144	—
ASSIST12	AUC	0.749 3	0.751 2	0.7505	0.7575	—
	RMSE	0.424 7	0.4152	0.4172	0.4119	—
XES3G5M	AUC	0.862 1	0.858 1	0.7878	0.8827	0.8845
	RMSE	0.325 1	0.326 9	0.3768	0.3120	0.3114

表3 消融实验性能比较

数据集	指标	ASSIST09	ASSIST12	XES3G5M
TIKT(\Diff)	AUC	0.7838	0.7564	0.8825
	RMSE	0.4151	0.4123	0.3101
TIKT(\CL)	AUC	0.7821	0.7543	0.8812
	RMSE	0.4166	0.4137	0.3127

表现。TIKT(\.)表示移除对应模块，其中Diff表示题目难度等级，CL表示对比学习。**表3**的结果表明，题目难度与对比学习对提升模型整体性能具有益处。由于移除对比学习模块也会移除题目难度等级，因此不难理解移除前者会比仅移除后者性能稍差。结合**表2**可以看出，移除对比学习模块后的TIKT性能也优于3个最先进的KT模型，这表明本文提出的响应预测模块不仅具有可解释能力，同时也未牺牲模型性能。

3.7 训练时间成本分析

训练时间成本对比见**表4**。为了验证训练时间，本文设置训练epoch为100，**表4**记录了每个epoch的平均运行时间，从**表4**中可以明显看出，MIKT的epoch时间远大于其他对比基线算法。这是由于MIKT是基于前一个状态的顺序训练，缺乏并行处理数据的能力，而AKT、CL4KT与TIKT则基于Transformer架构，能并行处理数据，从而极大地提升了训练速度。由于CL4KT与TIKT引入了对比学习方法，使得其训练时间稍大于经典的AKT模型。

4 结束语

本文以题目难度为切入点，通过 Bert

表4 训练时间成本对比

数据集	不同模型的epoch 平均运行时间/min			
	AKT	CL4KT	MIKT	TIKT
ASSIST09	0.034	0.074	1.573	0.102
ASSIST12	0.325	0.692	5.336	0.742
XES3G5M	0.393	1.127	11.514	1.412

模型预测题目文本内容中隐含的题目难度等级，增强了数据的可解释性，提出基于Transformer架构的高性能可解释KT模型，将模型预测过程分解为题目难度预测与学习者能力预测，IRT模块基于这两个解释参数预测学习者正确回答概率。为进一步提升模型性能，本文引入对比学习方法，并将Bert模型预测的题目难度引入其中。性能比较实验与消融实验结果表明，本文提出的TIKT模型在具有可解释能力的同时，拥有更好的性能表现。模型可解释实验结果表明，Bert模型能有效对题目难度等级进行预测，并且与响应预测模块估计的题目难度一致。未来的研究将聚焦于多知识概念下不同知识概念难度与题目难度的关系，通过题目文本等辅助信息预测出不同知识概念难度，为学习者提供更深入的指导，同时，分析学习者在掌握该知识点后却连续出现错误的原因并建立预警模型，也是后续的研究重点。

参考文献:

- [1] ABDELRAHMAN G, WANG Q, NUNES B. Knowledge tracing: a survey[J]. ACM Computing Surveys, 2023, 55(11): 1–37.
- [2] PIECH C, SPENCER J, HUANG J, et al. Deep knowledge tracing[J]. Advances in Neural Information Processing Systems, 2015, 3(3): 19–23.
- [3] ZHANG J N, SHI X J, KING I, et al. Dynamic key-value memory networks for knowledge tracing[C]//Proceedings of the 26th International Conference on World Wide Web. New York: ACM, 2017: 765–774.
- [4] PU S, YUDELSON M, OU L, et al. Deep knowledge tracing with transformers[M]//Artificial Intelligence in Education. Cham: Springer International Publishing, 2020: 252–256.
- [5] GHOSH A, HEFFERNAN N, LAN A S, et al. Context-aware attentive knowledge tracing[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2020: 2330–2339.
- [6] YEUNG C K. Deep-IRT: make deep learning based knowledge tracing explainable using item response theory[EB]. arXiv preprint, 2019, arXiv: 1904.11738v1.
- [7] SUN J W, YU F H, WAN Q, et al. Interpretable knowledge tracing with multi-scale state representation[C]//Proceedings of the ACM Web Conference 2024. New York: ACM, 2024: 3265–3276.
- [8] TIAN Z J, ZHENG G C, FLANAGAN B, et al. BEKT: deep knowledge tracing with bidirectional encoder representations from transformers[C]//Proceedings of the 29th International Conference on Computers in Education. Bangkok: APSCE, 2023.
- [9] SHEN S H, HUANG Z Y, LIU Q, et al. Assessing student's dynamic knowledge state by exploring the question difficulty effect[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2022: 427–437.
- [10] 王士进, 吴金泽, 张浩天, 等. 可信的端到端深度学生知识画像建模方法[J]. 计算机研究与发展, 2023, 60(8): 1822–1833.
- WANG S J, WU J Z, ZHANG H T, et al. Trustworthy end-to-end deep student knowledge portrait modelling method[J]. Journal of Computer Research and Development, 2023, 60(8): 1822–1833.
- [11] 梁祥, 刘梦赤, 胡婕, 等. 融合习题难度和作答经验的深度知识追踪模型[J]. 华南师范大学学报(自然科学版), 2023, 55(4): 81–86.
- LIANG X, LIU M C, HU J, et al. Deep knowledge tracing model by integrating problem difficulty and answering experience[J]. Journal of South China Normal University (Natural Science Edition), 2023, 55(4): 81–86.
- [12] LU Y, WANG D L, MENG Q G, et al. Towards interpretable deep learning models for knowledge tracing[M]//Artificial Intelligence in Education. Cham: Springer International Publishing, 2020: 185–190.
- [13] BACH S, BINDER A, MONTAVONG G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PLoS One, 2015, 10(7): e0130140.
- [14] WANG F, LIU Q, CHEN E H, et al. Neural cognitive diagnosis for intelligent education systems[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 6153–6161.
- [15] CHEN J H, LIU Z T, HUANG S Y, et al. Improving interpretability of deep sequential knowledge tracing models with

- question-centric cognitive representations[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(12): 14196–14204.
- [16] LIU X, ZHANG F J, HOU Z Y, et al. Self-supervised learning: generative or contrastive[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(1): 857–876.
- [17] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations [C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM, 2020: 1597–1607.
- [18] GAO T Y, YAO X C, CHEN D Q. SimCSE: simple contrastive learning of sentence embeddings[EB]. arXiv preprint, 2021, arXiv: 2104.08821.
- [19] XIE X, SUN F, LIU Z Y, et al. Contrastive learning for sequential recommendation[C]//Proceedings of the 2022 IEEE 38th International Conference on Data Engineering. Piscataway: IEEE Press, 2022: 1259–1273.
- [20] WANG T Z, ISOLA P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere[EB]. arXiv preprint, 2020, arXiv: 2005.10242.
- [21] LEE W, CHUN J, LEE Y, et al. Contrastive learning for knowledge tracing[C]//Proceedings of the ACM Web Conference 2022. New York: ACM, 2022: 2330–2338.
- [22] FENG M Y, HEFFERNAN N, KOEDINGER K. Addressing the assessment challenge with an online system that tutors as it assesses[J]. User Modeling and User-Adapted Interaction, 2009, 19(3): 243–266.
- [23] SAN PEDRO M O Z, BAKER R S J D, GOWDA S M, et al. Towards an understanding of affect and knowledge from student interaction with an intelligent tutoring system[M]//Artificial Intelligence in Education. Berlin, Heidelberg: Springer, 2013: 41–50.
- [24] LUI Z T, LIU Q Q, GUO T, et al. XES3G5M: a knowledge tracing benchmark dataset with auxiliary information [C]//Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS). New York: Curran Associates Inc., 2023: 32958–32970.

作者简介



周涛 (1995–), 男, 四川大学计算机学院博士生, 主要研究方向为知识追踪等。



李艳婷 (2001–), 女, 四川大学计算机学院硕士生, 主要研究方向为知识追踪。



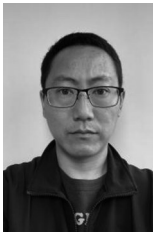
杨月婷（2002–），女，四川大学计算机学院硕士生，主要研究方向为推荐系统等。



任俊霖（2000–），男，四川大学计算机学院硕士生，主要研究方向为自然语言处理、代码智能等。



琚生根（1970–），男，博士，四川大学计算机学院教授，主要研究方向为数据挖掘、自然语言处理、代码智能。



师维（1985–），男，四川大学计算机学院工程师，主要研究方向为智慧教育。

收稿日期: 2024-10-29
通信作者: 师维, shiwei_851210@scu.edu.cn
基金项目: 国家自然科学基金项目(No.62137001)
Foundation Item: The National Natural Science Foundation of China (No.62137001)