https://crad.ict.ac.cn
62(2): 364−373, 2025

知识点相关性与遗忘程度融合的深度知识追踪模型

贾 瑞[1] 董永权[1,2,3] 刘 源[1] 陈 成[1]

[1](江苏师范大学计算机科学与技术学院 江苏徐州 221116)

[2](江苏省教育信息化工程技术研究中心(江苏师范大学) 江苏徐州 221116)

[3](徐州云计算工程技术研究中心 江苏徐州 221116)

(614891727@qq.com)

Deep Knowledge Tracing Model with the Integration of Skills Relation and Forgetting Degree

Jia Rui[1], Dong Yongquan[1,2,3], Liu Yuan[1], and Chen Cheng[1]

[1] (*College of Computer Science and Technology*, *Jiangsu Normal University*, *Xuzhou*, *Jiangsu* 221116)

[2] (*Jiangsu Education Informatization Engineering Technology Research Center*, *Xuzhou*, *Jiangsu* 221116)

[3] (*Xuzhou Cloud Computing Engineering Technology Research Center*, *Xuzhou*, *Jiangsu* 221116)

Abstract Knowledge tracing is a pivotal technique for modeling students' knowledge level, and typically relies on their past learning interactions to predict their future performance on exercises. These interactions represent a student's process of answering a sequence of questions. Current knowledge tracing methods ignore times a skill has been practiced when modeling student's forgetting behaviors. Also, few models consider the relation between skills and its influence on performance prediction. To address these questions, we propose a deep knowledge tracing model with the integration of skills relation and forgetting degree. Firstly, a relation matrix is constructed using statistical methods to capture the relation between skills. Secondly, the time intervals between interactions and the times a student practices the same skill are used to compute the forgetting degree of each skill for better modeling of students' forgetting behaviors. Finally, skills relation and forgetting degrees are integrated into an attention module to obtain the influence of each past interaction on future performance prediction. Based on new attention weights, students' performance on future exercises and knowledge level can be predicted. Experiments on two real-world online education datasets, algebra2005-2006 and ASSISTment2012, demonstrate that the proposed model achieves better prediction results compared with existing mainstream methods.

Key words knowledge tracing; deep knowledge tracing; skills relation; forgetting degree; intelligent education

摘 要 知识追踪(knowledge tracing, KT)是对学习者知识水平建模的一种技术,根据学习者过去的学习交互预测其未来答题表现,这些交互代表了学习者回答一个习题序列的过程.当前知识追踪方法在建模学习者遗忘行为时缺乏考虑知识点重复练习次数,并且少有模型考虑知识点相关性对答题预测的影响.基于此,提出了一个融合知识点相关性和遗忘程度的深度知识追踪模型.首先使用统计方法构建了一个关联矩阵,以捕获知识点之间的相关性.其次,利用交互之间的时间间隔和学习者学习同一知识点的次数

收稿日期: 2023-08-23; **修回日期**: 2024-01-08

基金项目: 国家自然科学基金项目(61872168); 江苏省教育科学"十四五"规划项目(D/2021/01/112); 江苏师范大学研究生科研与实践创新项目(2022XKT1549)

This work was supported by the National Natural Science Foundation of China (61872168), the Project of Jiangsu Province Education Science "14th Five-Year Plan" (D/2021/01/112), and the Graduate Research and Practice Innovation Program of Jiangsu Normal University (2022XKT1549).

通信作者: 董永权(tomdyq@163.com)

来计算知识点的遗忘程度，以更好地拟合学生的遗忘行为. 最后，将知识点相关性和遗忘程度整合到一个注意力模块中，以获得过去的每个交互对未来答题的影响，据此预测学习者的答题结果. 在真实的在线教育数据集 algebra2005-2006 和 ASSISTment2012 上的实验表明，相较于已有主流方法，所提模型取得了更好的预测结果.

关键词 知识追踪；深度知识追踪；知识点相关性；遗忘程度；智慧教育

中图法分类号 TP18

DOI: 10.7544/issn1000-1239.202330697 **CSTR:** 32373.14.issn1000-1239.202330697

中国大学 MOOC, Coursera 等智能导学平台提供了大量的在线课程和练习，为学习者的学习带来了便利和更多的选择. 知识追踪（knowledge tracing, KT）是智能导学平台的重要技术支撑，通过平台收集到的学习记录，知识追踪模型可评估学习者对知识点的掌握程度[1]. 知识追踪技术的深入研究可应用于学习者习题推荐、学习路径选择等，有助于推进个性化学习的发展. 知识追踪的任务主要包括：1）根据学习者的学习历史预测其在未来学习上的答题表现；2）实时追踪学习者的知识水平（即对各知识点的掌握程度）[2−4].

传统的知识追踪技术主要包括基于隐马尔可夫模型设计的贝叶斯知识追踪[5]（Bayesian knowledge tracing, BKT）及其变体. 随着深度学习的迅速发展，研究者们将其应用于知识追踪任务. 相比于传统知识追踪模型，深度学习方法具有更好的预测能力. 2015 年，Piech 等人[6] 利用循环神经网络（recurrent neural network, RNN）和长短期记忆网络（long short-term memory, LSTM）研究知识追踪. 在此基础上，动态键值记忆网络[7]（dynamic key-value memory network, DKVMN）和基于自注意力机制的知识追踪（self-attentive knowledge tracing, SAKT）模型[8] 相继提出. DKVMN 和 SAKT 都是通过捕捉知识点之间的关联性预测未来答题结果，二者的实验证明了将知识点间关系纳入知识追踪模型可以有效地提升模型的性能.

上述模型都忽视了知识点遗忘程度的影响. 教育心理学中的艾宾浩斯遗忘曲线理论指出[9]，学习者会遗忘学习到的知识，表现为对知识点掌握程度的下降. 知识点的学习时间间隔和重复学习知识点的次数会影响知识点的遗忘程度. 目前，在深度知识追踪（deep knowledge tracing, DKT）领域，只有少数的研究者考虑了遗忘因素，深度知识追踪的扩展工作[10] 表明学习者的遗忘与最近一次学习重复知识点所间隔的时间、学习次数以及最近一次学习的时间间隔 3 个要素有关. 然而该研究只考虑了最近一次学习和重复学习知识点 2 个情况，忽略了过去每一次学习都会对预测结果产生影响，同时缺乏对知识点相关性的考虑.

Pandey 等人[11] 在 SAKT 基础上提出了关系感知自注意力知识追踪（relation-aware self-attention for knowledge tracing, RKT）模型，设计了一个包含了习题相关性和学习者遗忘行为的自注意力层计算过去每一次学习对于预测结果的影响. RKT 相较于 SAKT 预测的准确性更高，这说明了遗忘行为对预测学习者表现有正向作用. 但是 RKT 对影响遗忘的因素考虑不全，忽视了重复练习次数的影响. 综上所述，知识点相关性和遗忘程度均有助于提高模型的预测性能. 但是现有模型并没有同时融合这 2 种因素.

针对上述问题，本文提出了一种融合知识点相关性和遗忘程度的深度知识追踪模型（deep knowledge tracing model with the integration of skill relation and forgetting degree, RFKT）. 本文的主要创新和贡献有 3 点：

1）考虑了知识点相关性和遗忘程度对知识追踪的影响，其中知识点相关性来自于先验知识，遗忘程度考虑了知识点的学习时间间隔和重复练习次数 2 个因素.

2）提出了一个融合知识点相关性和遗忘程度的深度知识追踪模型. 通过将知识点相关性和遗忘程度纳入注意力层，捕捉过去每一次学习对未来答题的影响，使模型能够更好地预测学习者的答题情况.

3）在真实在线教育数据集 algebra2005-2006 和 ASSISTment2012 上进行了实验，结果表明 RFKT 能够有效地对学习者的学习过程进行建模，其预测性能优于现有模型.

1 相关工作

现有的知识追踪模型主要分为 2 类：基于隐马尔可夫模型的知识追踪模型和基于深度学习的知识追踪模型[12].

BKT 是隐马尔可夫模型在知识追踪领域应用的经典模型，1995 年由 Corbett 等人[5]引入教育领域. BKT 将学习者的潜在知识状态建模为一组二元变量，每一组二元变量表示学习者对 1 个知识点的掌握情况. 在 BKT 的基础上，通过考虑各种因素，研究者们提出了许多改进方案. Yudelson 等人[13]考虑了个别学习者的先验知识对预测学习者答题结果的影响. Baker 等人[14]考虑了每个知识点的做错概率和猜错概率对预测学习者答题结果的影响. Pardos 等人[15]考虑了习题困难度对预测学习者答题结果的影响. 这类模型将学习者对各知识点的掌握程度分别建模，但是忽略了各知识点之间的关系.

DKT[6]是第一个将深度学习引入知识追踪的模型. DKT 利用 RNN 处理学习者的学习历史序列，RNN 的隐藏状态向量包含了学习历史中的所有信息. 但是 DKT 存在 2 个问题：一个是随着时间的推移，模型预测的知识状态具有波动性和不稳定性；另一个是 DKT 模型缺乏可解释性，无法解释预测结果与历史习题之间的关系. 为了使得预测的学生知识状态趋于平稳，Yeung 等人[16]提出了 DKT+，该模型通过添加正则化项对损失函数进行扩充，用以平滑预测结果. 对于 DKT 在可解释性方面的不足，Zhang 等人[7]提出了 DKVMN 模型，该模型引入了一个外部记忆模块来存储知识概念，并更新学习者相应的知识点掌握情况，可以较为准确地建模学习者对各个知识点的掌握情况. 但是 DKVMN 只能通过学习者最近一次学习记录模拟学习者的知识水平，未能捕捉到学习过程中的长期依赖.

Transformer[17]的提出使得许多深度学习模型可以使用自注意力机制来捕获学习历史序列中的依赖关系. 基于此，Pandey 等人[8]提出了 SAKT 模型. 该模型的主体为一个自注意力层，利用注意力机制捕捉学习历史与未来答题结果之间的关系，预测学习者的答题结果，该方法证明了过去每一次学习对学习者未来的答题表现都具有影响.

DKT, DKVMN, SAKT 在预测时仅用到了学习记录中的习题编号和习题答案，但是忽略了学习记录中存在的其他因素，例如学习时间间隔和习题相关性等. 因此，Pandey 等人[11]在 SAKT 基础上又提出了 RKT 模型. RKT 分别利用习题文本内容和答题情况分析习题间关系，利用做题时间建模学习者遗忘行为，将二者融入自注意力模型中，提升模型性能. 但是 RKT 只针对习题间的相关性进行建模，随着习题数目的增多，其计算所消耗的时间也随之增加. 同时，RKT 只考虑了学习时间间隔对遗忘程度的影响，而对重复练习次数并未进行关注. 除此之外，Zhou 等人[18]提出了 LANA（leveled attentive knowledge tracing，LANA）模型，将时间因素作为一种特征用于建模，而没有将其与学习者的遗忘行为进行关联.

在上述方法中，DKT, DKVMN, SAKT, LANA 的输入都是习题所对应的知识点，输出为学习者对各知识点的掌握程度. 而 RKT 的输入为习题，不包含知识点信息，该模型无法输出学习者对各知识点的掌握情况. 因此，本文使用习题所对应的知识点信息作为输入. 为方便表述，下文中提及的“习题”与“知识点”均为习题所对应的知识点.

总的来说，现有方法可以在一定程度上预测学习者的答题结果，但或者忽略了历史记录中包含的知识点间相关性信息，或者忽略了学习者的遗忘过程，或者对于学习者的遗忘行为建模不全. 本文提出的 RFKT 模型综合考虑到了知识点相关性和遗忘程度，并从 2 个方面建模学习者的遗忘行为，预测学生未来答题表现.

2 知识点相关性与遗忘程度融合的深度知识追踪模型

本节主要介绍知识追踪的形式化定义、知识点相关性和遗忘程度的计算过程及所提模型的各个模块.

2.1 问题定义

知识追踪任务中，将学习者的 1 次学习称之为 1 个交互. 第 i 个交互表示为 $\boldsymbol{x}_i=(e_i,r_i,t_i)$. 其中 $e_i\in\{1,2,\cdots,E\}$表示第 i 个交互的知识点 id，E 表示知识点总数；$r_i\in\{0,1\}$表示第 i 个交互的回答正确性，“0”表示回答错误，“1”表示回答正确；$t_i\in\mathbb{R}^+$是第 i 个交互的开始时间. 知识追踪根据学习者过去的 $n-1$ 个交互序列 $\boldsymbol{X}=(\boldsymbol{x}_1,x_2,\cdots,\boldsymbol{x}_{n-1})$，预测学习者在第 n 个交互中的答题结果 r_n.

本文提出了一个深度知识追踪模型，该模型将知识点相关性和遗忘程度作为上下文信息进行整合，并将其传播到使用自注意力机制计算的注意力权重中，以此代表过去的每一次交互对未来交互的影响. 本文使用的数学符号如表 1 所示.

2.2 知识点相关性计算

本文从学习者的交互数据中计算所有知识点彼此之间的相关性，表示为关系矩阵 $\boldsymbol{\varphi}$，$\varphi_{i,j}$代表了知识点 j 对知识点 i 的重要性.

φ是基于卡方的关联度量，主要用于测量2个二元变量之间相关性，具有良好的解释性，故而本文使用φ系数表示知识点间相关性. 知识点相关性的计算过程为：

首先，通过仅考虑知识点 i 和 j 的相关性构建一个列联表，如表2所示，表2中知识点 j 在交互序列中出现在知识点 i 之前，取最近3次知识点 j 的答题结果来判断学习者是否掌握了该知识点，如果3次答案的平均值大于0.5，则认为学习者掌握了知识点 j，反之则认为学习者未掌握知识点 j.

Table 1　Symbol Representation

表1　符号表示

符号	含义
E	知识点数量
$\boldsymbol{x}_i$	学习者的第 i 个交互
e_i	第 i 个交互中的知识点 id
r_i	第 i 个交互中学习者回答e_i的正确性
t_i	第 i 个交互的开始时间
$\boldsymbol{rel}_i$	e_i与过去交互中每一个知识点之间的相关性
$\boldsymbol{f}_i$	学习者对过去交互中每个知识点的遗忘程度
$\boldsymbol{M}^{\mathrm{I}}$	交互嵌入矩阵
$\boldsymbol{M}^{\mathrm{K}}$	知识点嵌入矩阵
p_i	KT模型预测学习者对e_i的掌握程度
$\boldsymbol{\varphi}$	知识点相关性矩阵

Table 2　Columnar Table of Skills *i* and *j*

表2　知识点 i 和 j 的列联表

知识点 j	知识点 i		总计
	正确	错误	
未掌握	n_{00}	n_{01}	a_1
掌握	n_{10}	n_{11}	a_2
总计	b_1	b_2	n

其次，根据列联表计算2个知识点的φ系数，计算过程如式(1)(2)(3)所示：

$$\varphi_{i,j}=\frac{n_{11}n_{00}-n_{01}n_{10}}{\sqrt{a_2a_1b_2b_1}}, \tag{1}$$

$$a_1=n_{00}+n_{01},\ a_2=n_{10}+n_{11}, \tag{2}$$

$$b_1=n_{00}+n_{10},\ b_2=n_{01}+n_{11}, \tag{3}$$

其中 n_{00} 表示在未掌握知识点 j 的情况下，知识点 i 回答正确的次数；n_{01} 表示在未掌握知识点 j 的情况下，知识点 i 回答错误的次数；n_{10} 则表示在掌握知识点 j 的情况下，知识点 i 回答正确的次数；而 n_{11} 表示在掌握知识点 j 的情况下，知识点 i 回答正确的次数. $\varphi_{i,j}\in[-1,1]$，$\varphi_{i,j}$的绝对值越大，表示知识点 j 对于知识点 i 的影响性越大，正值代表2个知识点正相关，负值代表2个知识点负相关.

最后，可以根据知识点相关性矩阵计算得到e_i与前$i-1$个交互中知识点之间的相关性$\boldsymbol{rel}_i=(\varphi_{e_i,e_1},\varphi_{e_i,e_2},\cdots,\varphi_{e_i,e_{i-1}})$，计算过程如图1所示.

2.3　学习者遗忘程度计算

艾宾浩斯发现，随着时间的流逝，人们会在一定程度上遗忘之前学习过的知识，因此提出了遗忘曲线. 除此之外，艾宾浩斯也指出，通过多次的重复可以减弱人们对于知识的遗忘程度. 为了更好地模拟学习者遗忘行为，本文将一个核函数设计成随时间呈指数衰减的曲线，以贴合艾宾浩斯遗忘曲线. 然后将知识点重复练习次数加入到函数中，以更接近学

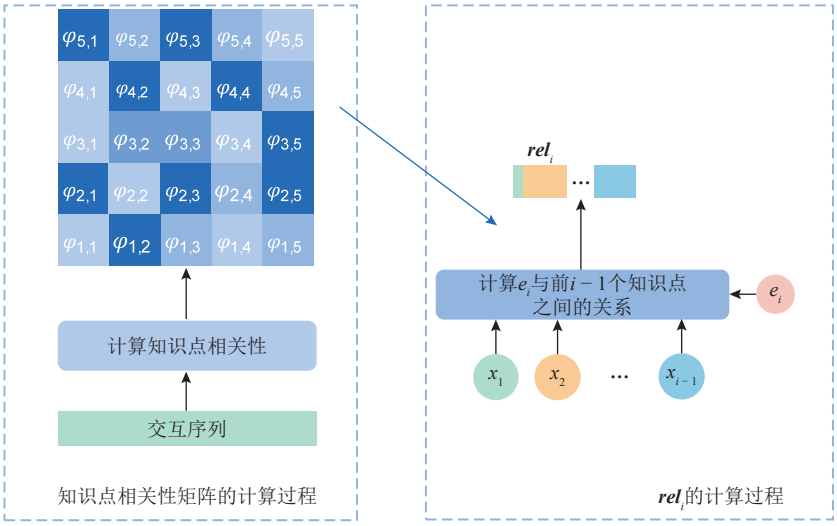


Fig. 1　The process of calculating the skills relation

图1　知识点相关性计算过程

习者的遗忘行为. 本文计算e_i和e_j之间的相对时间间隔为$\Delta_j=|t_i-t_j|$. 接下来计算在时刻t_i之前e_j的重复练习次数c_j. 最后, 根据式(4)计算得到学习者在时刻t_i的遗忘程度$\boldsymbol{f}_i$. 计算过程如图 2 所示, 每个圆点对应一个知识点, 相同的色度代表相同的知识点. 学习时间间隔、知识点重复练习次数和遗忘程度均表示为 1 维数组, 通过向右补 0 而成为固定长度数组.

$$\boldsymbol{f}_i=\left(\exp\left(\frac{-\Delta_{i,1}}{c_1}\right),\exp\left(\frac{-\Delta_{i,2}}{c_2}\right),\cdots,\exp\left(\frac{-\Delta_{i,i-1}}{c_{i-1}}\right)\right). \tag{4}$$

2.4 RFKT 模型

本文提出了知识相关性与遗忘程度融合的深度知识追踪模型 RFKT, 具体框架如图 3 所示. 模型主要分为输入层、注意力层和预测层.

2.4.1 输入层

模型的输入分为 2 个部分: 1)时刻 i 之前学习者的交互序列$\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{i-1})$, 为便于模型计算, 将交互$\boldsymbol{x}_j$表示为一个实数$y_j=e_j+r_j\times E$, $j\in\{1,2,\cdots,i-1\}$. 2)待预测习题e_i. RFKT 训练了 2 个嵌入矩阵$\boldsymbol{M}^{\mathrm{I}}\in\mathbb{R}^{2E\times d}$和$\boldsymbol{M}^{\mathrm{K}}\in\mathbb{R}^{E\times d}$, 分别将交互$y_j$和习题$e_i$映射到连续向量空间, 得到交互嵌入$\boldsymbol{M}^{\mathrm{I}}_{y_j}$和习题嵌入$\boldsymbol{M}^{\mathrm{K}}_{e_i}$.

学习者对知识点的掌握程度受到答题顺序的影响, 因此, RFKT 模型为每一个交互添加一个位置编

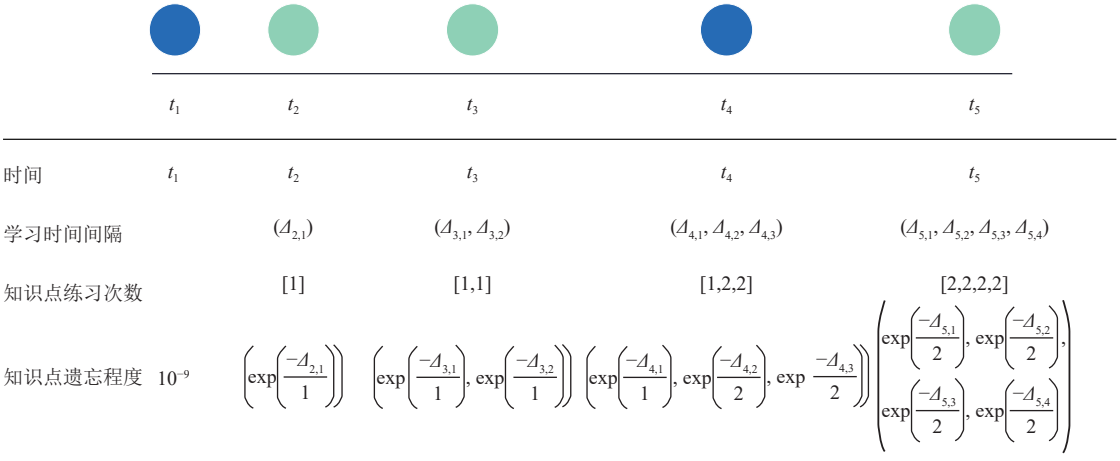


Fig. 2 Calculation process of learners' forgetting degree

图 2 学习者遗忘程度计算过程

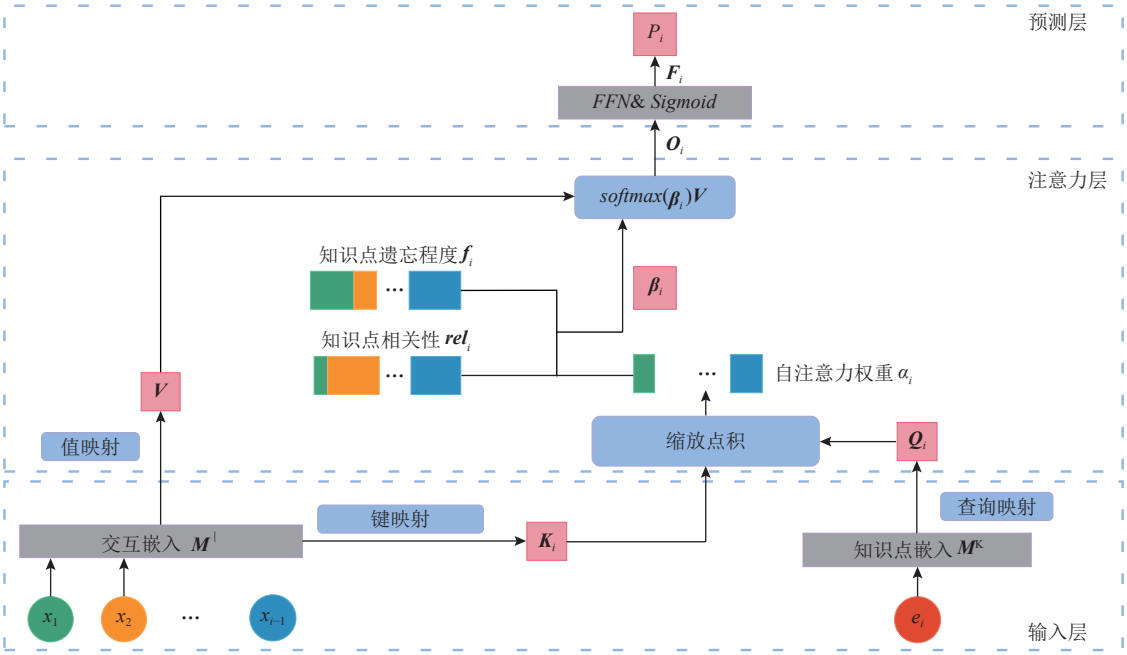


Fig. 3 RFKT framework

图 3 RFKT 框架

码$\boldsymbol{Pos}\in\mathbb{R}^{(i-1)\times d}$，$\boldsymbol{Pos}$是一个可训练的矩阵. 输入层的输出是交互嵌入矩阵$\hat{\boldsymbol{M}}^{\mathrm{I}}$和习题嵌入矩阵$\hat{\boldsymbol{M}}^{\mathrm{K}}$：

$$\hat{\boldsymbol{M}}^{\mathrm{I}}=\begin{pmatrix}\boldsymbol{M}^{\mathrm{I}}_{y_1}+\boldsymbol{Pos}_1\\ \boldsymbol{M}^{\mathrm{I}}_{y_2}+\boldsymbol{Pos}_2\\ \vdots\\ \boldsymbol{M}^{\mathrm{I}}_{y_{i-1}}+\boldsymbol{Pos}_{i-1}\end{pmatrix},\ \hat{\boldsymbol{M}}^{\mathrm{K}}=\begin{pmatrix}\boldsymbol{M}^{\mathrm{K}}_{e_2}\\ \boldsymbol{M}^{\mathrm{K}}_{e_3}\\ \vdots\\ \boldsymbol{M}^{\mathrm{K}}_{e_i}\end{pmatrix}. \tag{5}$$

2.4.2 注意力层

注意力层的作用是计算历史交互与待预测习题之间的相关权重. 在本文模型中，重新设计了注意力权重计算公式，使得更多的关注点放在由知识点相关性和遗忘程度确定的相关权重上. 计算过程如下：

设$\alpha_{i,j}$表示第j个交互与习题e_i的相关权重，利用自注意力机制计算得到，如式(6)(7)所示：

$$\boldsymbol{Q}_i=\hat{\boldsymbol{M}}^{\mathrm{K}}_i\boldsymbol{W}^{\mathrm{Q}},\ \boldsymbol{K}_j=\hat{\boldsymbol{M}}^{\mathrm{I}}_j\boldsymbol{W}^{\mathrm{K}}, \tag{6}$$

$$\alpha_{i,j}=\frac{\exp\left(\gamma_{i,j}\right)}{\sum_{j=1}^{i-1}\exp\left(\gamma_{i,j}\right)},\ \gamma_{i,j}=\frac{\boldsymbol{Q}_i\boldsymbol{K}_j^{\mathrm{T}}}{\sqrt{d}}, \tag{7}$$

其中$\boldsymbol{W}^{\mathrm{Q}}\in\mathbb{R}^{d\times d_k}$，$\boldsymbol{W}^{\mathrm{K}}\in\mathbb{R}^{d\times d_k}$分别表示查询矩阵和键映射矩阵，将向量线性映射到不同的空间，d为映射矩阵的维度.

之后，本文模型将$\alpha_{i,j}$与知识点相关性$rel_{i,j}$和遗忘程度$f_{i,j}$三者相加得到新的注意力权值$\beta_{i,j}$：

$$\beta_{i,j}=\lambda_1\left(\lambda_2\alpha_{i,j}+(1-\lambda_2)rel_{i,j}\right)+(1-\lambda_1)f_{i,j}, \tag{8}$$

其中$rel_{i,j}$代表知识点j与知识点i的相关性，$f_{i,j}$表示在时刻i对知识点j的遗忘程度. 为了避免计算量的明显增加，本文模型选择加法运算，λ_1和λ_2是2个可训练参数.

注意力层的第i个输出$\boldsymbol{O}_i$计算如式(9)所示，其中$\boldsymbol{\beta}_i$为前$i-1$个交互与习题e_i之间的相关权重，$\boldsymbol{V}$为前$i-1$个交互经过一个值矩阵$\boldsymbol{W}^{\mathrm{V}}\in\mathbb{R}^{d\times d_k}$映射得到的矩阵.

$$\boldsymbol{O}_i=softmax(\boldsymbol{\beta}_i)\boldsymbol{V},\ \boldsymbol{V}=\sum_{j=1}^{i-1}\hat{\boldsymbol{M}}^{\mathrm{I}}_j\boldsymbol{W}^{\mathrm{V}}. \tag{9}$$

值得注意的是，在预测第i个答题结果时，只需要考虑前$i-1$个交互，所以本文在注意力层中为键映射矩阵$\boldsymbol{W}^{\mathrm{K}}$添加一个上三角的掩码矩阵掩盖未来交互信息. 算法1给出了RFKT的工作流程：

算法 1. RFKT.

输入：学习者交互序列$\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{i-1})$，其中$\boldsymbol{x}_j=\left(e_j,r_j,t_j\right)$，待预测的习题$e_i$，知识点相关性矩阵$\boldsymbol{rel}$，知识点遗忘程度$\boldsymbol{f}$，真实标签$a$；

输出：学习者对e_i的掌握程度P_i.

① 初始化习题嵌入矩阵$\boldsymbol{M}^{\mathrm{K}}\in\mathbb{R}^{E\times d}$和交互嵌入矩阵$\boldsymbol{M}^{\mathrm{I}}\in\mathbb{R}^{2E\times d}$，定义一个位置嵌入矩阵$\boldsymbol{Pos}\in\mathbb{R}^{(i-1)\times d}$；

② for j =1 to $i-1$ do

③ $y_j=e_j+r_j\times E$; /*将交互元组$\boldsymbol{x}_i$表示为一个实数*/

④ $\boldsymbol{M}^{\mathrm{I}}_{y_j}=\boldsymbol{M}^{\mathrm{I}}\left(y_j\right)$; /*$y_j$映射到连续向量空间*/

⑤ $\boldsymbol{M}^{\mathrm{K}}_{e_i}=\boldsymbol{M}^{\mathrm{K}}(e_i)$; /*$e_i$映射到连续向量空间*/

⑥ $\hat{\boldsymbol{M}}^{\mathrm{I}}_j=\boldsymbol{M}^{\mathrm{I}}_{y_j}+\boldsymbol{P}_j$; /*添加位置编码 */

⑦ $\hat{\boldsymbol{M}}^{\mathrm{K}}_i=\boldsymbol{M}^{\mathrm{K}}_{e_i}$;

⑧ $\boldsymbol{K}_j=\hat{\boldsymbol{M}}^{\mathrm{I}}_j\boldsymbol{W}^{\mathrm{K}}$; /*将交互嵌入经过映射得到键向量*/

⑨ $\boldsymbol{Q}_i=\hat{\boldsymbol{M}}^{\mathrm{K}}_i\boldsymbol{W}^{\mathrm{Q}}$; /*将习题嵌入经过映射得到查询向量*/

⑩ $\alpha_{i,j}=\frac{\exp\left(\gamma_{i,j}\right)}{\sum_{j=1}^{i-1}\exp\left(\gamma_{i,j}\right)},\ \gamma_{i,j}=\frac{\boldsymbol{Q}_i\boldsymbol{K}_j^{\mathrm{T}}}{\sqrt{d}}$; /*计算自注意力权重*/

⑪ $\beta_{i,j}=\lambda_1\left(\lambda_2\alpha_{i,j}+(1-\lambda_2)rel_{i,j}\right)+(1-\lambda_1)f_{i,j}$; /*将$rel_{i,j}$和$f_{i,j}$融入自注意力权重，$\beta_{i,j}$作为新的注意力权值*/

⑫ end for

⑬ $\boldsymbol{V}=\sum_{j=1}^{i-1}\hat{\boldsymbol{M}}^{\mathrm{I}}_j\boldsymbol{W}^{\mathrm{V}}$; /*将前$i-1$个交互嵌入经过映射得到值矩阵*/

⑭ 注意力层的输出：$\boldsymbol{O}_i=Softmax(\boldsymbol{\beta}_i)\boldsymbol{V}$;

⑮ 预测：$\boldsymbol{F}_i=ReLU(\boldsymbol{O}_i\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2$;

⑯ $P_i=Sigmoid(\boldsymbol{F}_i\boldsymbol{W}_3+\boldsymbol{b}_3)$;

⑰ 用标签a评估预测值P_i;

⑱ 通过反向传播更新所有模型参数.

2.4.3 预测层

为了在模型中加入非线性，并考虑不同潜在维度之间的相互作用，本文模型使用前馈网络，如式(10)所示：

$$\boldsymbol{F}_i=FFN(\boldsymbol{O}_i)=ReLU(\boldsymbol{O}_i\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2, \tag{10}$$

其中$\boldsymbol{W}_1\in\mathbb{R}^{d_k\times d}$，$\boldsymbol{W}_2\in\mathbb{R}^{d_k\times d}$，$\boldsymbol{b}_1\in\mathbb{R}^d$，$\boldsymbol{b}_2\in\mathbb{R}^d$是训练中学习到的参数.

最后，将上述得到的$\boldsymbol{F}_i$通过*Sigmoid*激活函数全连接网络来预测学习者的答题结果.

$$P_i=Sigmoid(\boldsymbol{F}_i\boldsymbol{W}_3+\boldsymbol{b}_3), \tag{11}$$

其中P_i为一个标量，表示学习者答对习题e_i的概率，$\boldsymbol{W}_3\in\mathbb{R}^{d\times d}$，$\boldsymbol{b}_3\in\mathbb{R}^d$是训练中学习到的参数.

3 实验与结果分析

本节首先介绍实验的基本设置, 包括数据集、测评指标和对比方法. 接着进行参数实验, 已获得最优的模型参数. 随后展示本文模型和各对比方法在各个数据集的预测情况. 最后进行消融实验以验证知识点相关性和遗忘程度的有效性.

3.1 数据集

为了评估 RFKT 的有效性, 使用 2 个真实的数据集 algebra2005-2006 和 ASSISTment2012. 对于所有的数据集, 删除了知识点有缺失值的数据和重复的记录.

algebra2005-2006 是在 2010 年 KDD 杯竞赛中发布的一个开放数据集. 该数据集包括 574 个学习者、436 个知识点、1 084 个习题以及 424 918 条学习记录.

ASSISTment2012 数据集由 ASSISTment 在线辅导平台提供, 广泛用于知识追踪任务. 该数据集包括 28 834 个学习者、198 个知识点、50 983 个习题和 2 629 870 条学习记录.

本文将数据集的 10% 作为测试集. 在其余 90% 的数据集中, 取 20% 作为验证集, 剩余 80% 作为训练集. 每个实验重复 5 轮, 将结果的平均值作为实验结果.

3.2 测评指标

对学习者答题结果的预测是一个二元分类问题, 即回答正确或错误. 因此, 本文使用曲线下面积(area under curve, AUC)和精度(accuracy, ACC)指标来比较模型性能. 一般来说 AUC 或 ACC 值为 0.5 时代表随机猜测时的预测结果, 2 个指标的值越大越好.

3.3 对比方法与参数设置

为了评估 RFKT 模型的性能, 本文选择了 DKT, DKT+, DKVMN, SAKT, RKT, LANA 作为对比方法. 所有方法的批处理大小为 256, 使用 Adam 作为优化器, 学习率为 0.001. 其他参数设置如下.

1)DKT[6]: 嵌入层和隐藏层维度分别为 100;

2)DKT+[16]: 嵌入层和隐藏层维度分别为 100, 3 个正则化参数为 0.01, 0.003, 3.0;

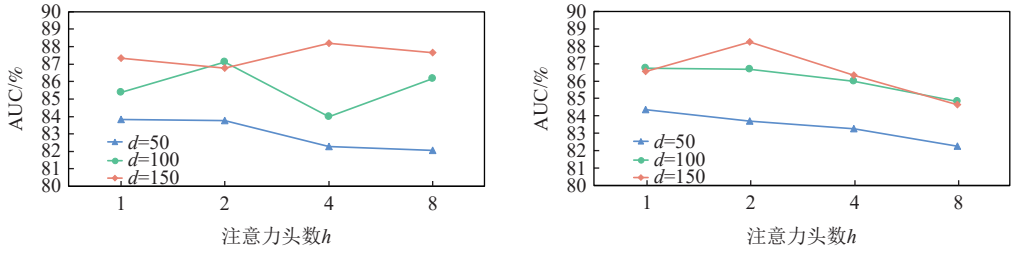
3)DKVMN[7]: 记忆矩阵列数为 50, 嵌入层维度为 50, 隐藏层维度为 20;

4)SAKT[8]: 嵌入层维度为 100, 隐藏层维度为 200, 注意力维度为 32, 注意力头数为 8, 随机失活率为 0.2;

5)RKT[11]: 嵌入层维度为 100, 隐藏层维度为 200, 注意力维度为 32, 注意力头数为 4, 随机失活率为 0.2;

6)LANA[18]: 编码层和解码层数均为 2, 注意力维度为 32, 注意力头数为 8, 随机失活率为 0.1;

7)RFKT: 隐藏层维度为 300, 注意力维度为 64, 随机失活率为 0.2. 对于嵌入层维度 d 和注意力头数 h, 本文通过在测试数据集上进行参数实验进行选取, 测试结果如图 4 所示, 在 algebra2005-2006 数据集中, 当 d=150, h=4 时, 平均 AUC 值为 88.19, 高于其他超参数设置情况, 在 ASSISTment2012 数据集中, 当 d=150, h=2 时, 平均 AUC 值为 88.23, 高于其他超参数设置情况. 因此, 对比实验中 RFKT 根据参数实验的最优结果分别设置 2 个数据集的嵌入层维度和注意力头数.



(a) algebra2005-2006 (b) ASSISTment2012

Fig. 4 Comparison of AUC values of RFKT on two datasets with different hyperparameter settings

图 4 不同超参数设置下 RFKT 在 2 个数据集上的 AUC 值对比

3.4 对比实验及结果分析

本节对 RFKT 与 6 个对比方法的实验结果进行分析, 以观察 RFKT 在不同数据集上的预测性能.

如表 3 所示, RFKT 在 algebra2005-2006 和 ASSISTment2012 数据集上都优于所有对比方法. 具体来说, 在 algebra2005-2006 上, RFKT 的 AUC 和 ACC 分别比次优方法 RKT 高 5.6% 和 5.2%. 在 ASSISTment2012 上, RFKT 的 AUC 和 ACC 分别比 RKT 高 11.3% 和 12.1%.

DKT, DKT+, DKVMN, SAKT 这 4 个模型都只简单地考虑了知识点 id 及学习者答题情况, 输入的信息较少, 所以它们对学习者未来答题表现的预测能力相近. LANA 与 RKT 都考虑了时间因素, LANA 将时间戳数据放入模型, RKT 则是将时间戳数据加工为学习时间间隔, LANA 的性能并没有提升, 而 RKT

Table 3 Prediction Results of Different Models on Knowledge Tracing Tasks

表 3 不同模型在知识追踪任务上的预测结果 %

模型	algebra2005-2006		ASSISTment2012	
	AUC	ACC	AUC	ACC
DKT[6]	80.34	80.25	72.75	73.51
DKT+[16]	80.52	80.27	72.04	73.29
DKVMN[7]	79.17	79.62	72.20	73.20
SAKT[8]	79.78	80.01	72.35	73.27
RKT[11]	83.46	79.70	79.30	71.90
LANA[18]	78.17	79.11	72.54	73.38
RFKT（本文）	**88.19**	**83.86**	**88.23**	**80.63**

注：黑体数值表示最优值，下划线数值表示次优值.

有较优的性能，证明了对数据进行加工的必要性. RKT 对学习者遗忘行为的建模缺乏对知识点重复练习次数的考虑，RFKT 的预测性能优于 RKT，说明知识点重复练习次数有利于模型建模学习者遗忘行为.

3.5 消融实验

本节对 RFKT 进行了消融实验，以深入了解该模型的每个部分对整体的贡献. 通过移除模型中的一些模块，得到了该模型的 5 个变体.

1）变体 1. 注意力层中移除知识点相关性，即仅保留传统注意力权重和知识点遗忘程度，简写为 R.

2）变体 2. 遗忘程度的计算过程中，移除知识点重复练习次数，即仅考虑学习时间间隔的遗忘行为，简写为 C.

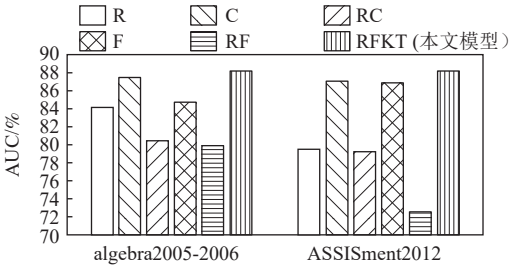
3）变体 3. 注意力层中移除知识点相关性的同时在遗忘程度计算过程中移除知识点重复练习次数，即仅保留传统注意力权重和只考虑了学习时间间隔的知识点遗忘程度，简写为 RC.

4）变体 4. 注意力层中移除知识点遗忘程度，即仅保留传统注意力权重和知识点相关性，简写为 F.

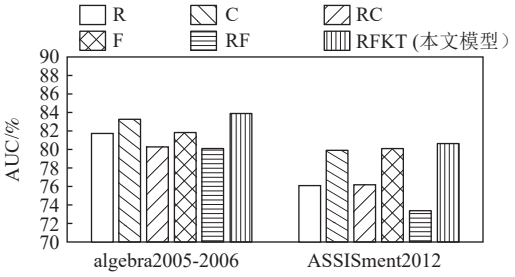
5）变体 5. 注意力层中移除知识点相关性和遗忘程度，即仅保留传统注意力权重，简写为 RF.

图 5(a)(b)分别展示了本文模型与 5 个变体在 2 个数据集上 AUC 和 ACC 的预测结果. 从实验结果看，移除知识点相关性、遗忘程度、知识点重复练习次数后模型性能有不同程度的下降，说明三者均对模型的预测能力有提升作用.

对于变体 1(R)，由于缺少考虑知识点相关性，难以构建知识点之间的联系，从而造成了模型预测性能的下降. 对于变体 2(C)，忽略了知识点重复练习次数在遗忘过程中的作用，模型预测 AUC 下降，证明知识点重复练习次数有助于提升模型的预测能力.



(a) RFKT与各变体的 AUC预测结果对比



(b) RFKT与各变体的 ACC预测结果对比

Fig. 5 Comparison of prediction results of RFKT and its variants

图 5 RFKT 与各变体的预测结果对比

对于变体 3(RC)，考虑了传统注意力权重和学习时间间隔，其性能高于变体 5(RF)（仅考虑传统注意力权重），证明了学习时间间隔的重要性. 对于变体 4(F)，忽略了学习者的遗忘行为，模型预测性能下降，说明模拟学习者的遗忘行为对模型预测能力的提升发挥了重要作用. 对于变体 5，同时移除了知识点相关性和遗忘程度，性能大幅度下降，且其性能低于变体 1（仅移除知识点相关性）和变体 4(仅移除知识点遗忘程度）证明了知识点相关性和遗忘程度 2 种因素相互作用，比单使用其中一种因素能更好的预测学习者未来答题表现.

除此之外，变体 5 的预测性能与 DKT, DKT+, DKVMN, SAKT 等方法相近，这是由于 RFKT 移除知识点相关性和遗忘程度后与这 4 种模型所用的特征一致，模型性能受到特征的限制.

RKT 与本文模型 RFKT 最为接近，二者之间存在 2 个区别：一是 RFKT 考虑了知识点相关性，而 RKT 考虑的是习题相关性；二是 RFKT 考虑了知识点的重复练习次数，而 RKT 没有考虑该因素. 为了进一步研究这 2 个因素的影响，设计了一个 RKT 变体，将知识点重复练习次数纳入 RKT 的遗忘程度计算过程，表示为 RKT+C. 该变体在 algebra2005-2006 数据集上的 AUC 结果如图 6 所示. 对比 RKT+C 和 RFKT，二者计算遗忘程度的方法相同，但考虑的相关性不同，RFKT 的性能优于 RKT+C，这说明相比于习题相

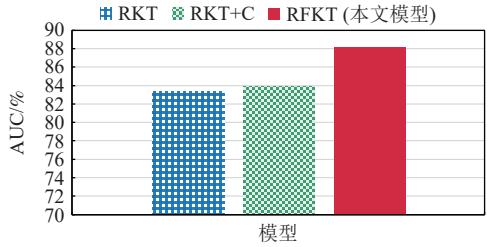


Fig. 6 Comparative study on algebra2005-2006 dataset of RKT, RKT+C, and RFKT

图 6 RKT, RKT+C, RFKT 在 algebra2005-2006 数据集上的对比实验

关性, 知识点相关性对于知识追踪模型预测能力的提升效果更好. 对比 RKT+C 与 RKT, 二者均使用了习题相关性, 但 RKT+C 在遗忘程度的计算过程中考虑了知识点重复练习次数, 其预测性能高于 RKT, 证明了知识点重复练习次数对于模型建模学习者遗忘行为有正向作用.

3.6 知识追踪结果分析

知识追踪的任务之一是实时输出学习者对于各个知识点的掌握程度, 本文进行以下实验验证 RFKT 在此任务上的有效性.

本文截取了数据集 algebra2005-2006 中一位学习者在一段时间内的学习记录, 并使用 RFKT 追踪学习者对 5 个知识点的掌握程度变化, 如图 7 所示. t_i 对应的雷达图表示时刻 i 学习者对 5 个知识点的掌握程度, 箭头上方的元组 (e_i,r_i) 表示学习者的学习记录, e_i 表示学习的知识点, r_i 表示答题情况.

实验结果显示, 在时刻 t_0, 学习者对知识点 22 答题错误后, RFKT 模型评估的知识点 22 掌握程度有所下降; 在时刻 t_3, 学习者对知识点 44 答题正确后, RFKT 模型评估的知识点 44 的掌握程度有所提高. 以上结果说明: RFKT 在获得学习者答题结果后, 都会根据其更新对应知识点的掌握程度.

对比时刻 t_0 学习知识点 22 到时刻 t_4 第 2 次学习知识点 22 期间, RFKT 模型显示学习者对于知识点 22 的掌握程度一直在下降, 这是由于随着时间间隔的增加, 学习者对知识点的掌握程度会下降, 证明了 RFKT 可以有效模拟学习者的遗忘行为.

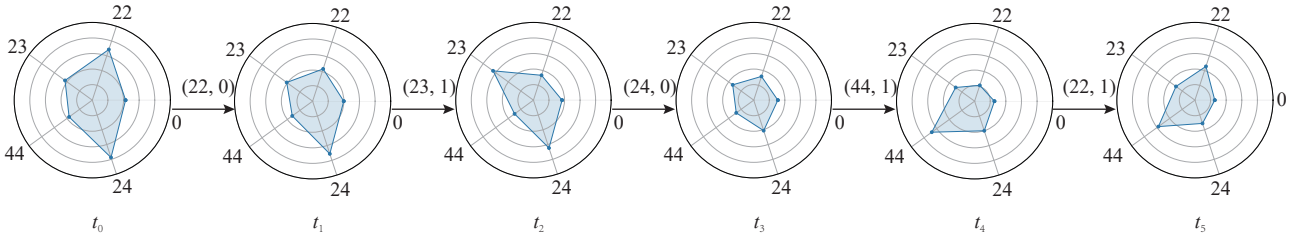


Fig. 7 Knowledge level output result of RFKT

图 7 RFKT 知识水平输出结果

4 总结与展望

本文提出了一个融合知识点相关性和遗忘程度的深度知识追踪模型 RFKT, 在预测学习者未来答题表现方面有良好的准确性. 这表明了在知识追踪中同时考虑知识点相关性和遗忘程度的必要性.

由于学习者有不同的认知能力和学习风格, 知识追踪过程的个性化可以使学习者的成绩得到更准确的预测, 未来将探索使用迁移学习和元学习等技术来开发个性化的知识追踪模型, 进一步研究学习者的个体差异对知识追踪的影响.

作者贡献声明: 贾瑞设计了初步的模型框架、实现了本文的所有实验并撰写论文初稿; 董永权对论文框架和实验设计提出了建设性意见, 并对论文进行了细致修改; 刘源参与了论文框架设计; 陈成对实验设计提出了建设性意见.

参 考 文 献

[1] Liu Tieyuan, Chen Wei, Chang Liang, et al. Research advances in the knowledge tracing based on deep learning[J]. Journal of Computer Research and Development, 2022, 59(1): 81−104 (in Chinese)
(刘铁园, 陈威, 常亮, 等. 基于深度学习的知识追踪研究进展[J]. 计算机研究与发展, 2022, 59(1): 81−104)

[2] Liu Kunjia, Li Xinyi, Tang Jiuyang, et al. Interpretable deep knowledge tracing model[J]. Journal of Computer Research and Development, 2021, 58(12): 2618−2629 (in Chinese)
(刘坤佳, 李欣奕, 唐九阳, 等. 可解释深度知识追踪模型[J]. 计算机研究与发展, 2021, 58(12): 2618−2629)

[3] Wang Yu, Zhu Mengxia, Yang Shanghui, et al. Review and performance comparison of deep knowledge tracing models[J]. Journal of Software, 2023, 34(3): 1365−1395 (in Chinese)
(王宇, 朱梦霞, 杨尚辉, 等. 深度知识追踪模型综述和性能比较[J].

软件学报, 2023, 34(3): 1365−1395)
[4] Yeung C K, Yeung D Y. Incorporating features learned by an enhanced deep knowledge tracing model for STEM/Non-STEM job prediction[J]. International Journal of Artificial Intelligence in Education, 2019, 29(3): 253−278
[5] Corbett A T, Anderson J R. Knowledge tracing: Modeling the acquisition of procedural knowledge[J]. User Modelling and User-Adapted Interaction, 1995, 4(4): 253−278
[6] Piech C, Spencer J, Huang J, et al. Deep knowledge tracing[J]. arXiv preprint, arXiv: 1506.05908, 2015
[7] Zhang Jiani, Shi Xingjia, King I, et al. Dynamic key-value memory networks for knowledge tracing[C]//Proc of the 26th Int Conf on World Wide Web. New York: ACM, 2017: 765−774
[8] Pandey S, Karypis G. A self-attentive model for knowledge tracing[J]. arXiv preprint, arXiv: 1907.06837, 2019
[9] Choffin B, Popineau F, Bourda Y, et al. DAS3H: Modeling student learning and forgetting for optimally scheduling distributed practice of skills[J]. arXiv preprint, arXiv: 1905.06873, 2019
[10] Nagatani K, Zhang Qian, Sato M, et al. Augmenting knowledge tracing by considering forgetting behavior[C]//Proc of the 19th World Wide Web Conf. New York: ACM, 2019: 3101−3107
[11] Pandey S, Srivastava J. RKT: Relation-aware self-attention for knowledge tracing[C]//Proc of the 29th ACM Int Conf on Information & Knowledge Management. New York: ACM, 2020: 1205−1214
[12] Liu Qi, Shen Shuanghong, Huang Zhenya, et al. A survey of knowledge tracing[J]. arXiv preprint, arXiv: 2105.15106, 2021
[13] Yudelson M V, Koedinger K R, Gordon G J. Individualized Bayesian knowledge tracing models[C]//Proc of the 16th Int Conf on Artificial Intelligence in Education. Berlin: Springer, 2013: 171−180
[14] Baker R S J, Corbett A T, Aleven V. More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing[G]//LNCS 5091: Proc of the 9th Int Conf on Intelligent Tutoring Systems (ITS 2008). Berlin: Springer, 2008: 406−415
[15] Pardos Z A, Heffernan N T. Modeling individualization in a Bayesian networks implementation of knowledge tracing[G]//LNCS 6075: Proc of the 18th Int Conf on User. Berlin: Springer, 2010: 255−266
[16] Yeung C K, Yeung D Y. Addressing Two problems in deep knowledge tracing via prediction-consistent regularization[J]. arXiv preprint, arXiv: 1806.02180, 2018
[17] 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. San Dieg, California: Curran Associates, 2017: 5998−6008
[18] Zhou Yuhao, Li Xihua, Cao Yunbo, et al. LANA: Towards personalized deep knowledge tracing through distinguishable interactive sequences[J]. arXiv preprint, arXiv: 2105.06266, 2021



Jia Rui, born in 1999. Master candidate. Her main research interests include data mining, knowledge tracing, and machine learning.
贾　瑞, 1999 年生. 硕士研究生. 主要研究方向为数据挖掘、知识追踪、机器学习.



Dong Yongquan, born in 1979. PhD, professor, master supervisor. His main research interests include deep learning, data mining, and education informatization.
董永权, 1979 年生. 博士, 教授, 硕士生导师. 主要研究方向为深度学习、数据挖掘、教育信息化.



Liu Yuan, born in 1997. Master candidate. His main research interests include course recommendation and machine learning.
刘　源, 1997 年生. 硕士研究生. 主要研究方向为课程推荐、机器学习.



Chen Cheng, born in 1999. Master candidate. His main research interests include data mining, knowledge tracing, and machine learning.
陈　成, 1999 年生. 硕士研究生. 主要研究方向为数据挖掘、知识追踪、机器学习