文章编号:1672-3961(2024)01-0100-09 DOI:10.6040/j.issn.1672-3961.0.2023.143

基于交互序列特征相关性的可解释知识追踪

陈成[1],董永权[1,2,3*],贾瑞[1],刘源[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 徐州 221116; 3.徐州市云计算工程技术研究中心, 江苏 徐州 221116)

摘要:为提高知识追踪(knowledge tracing, KT)模型的可解释性,提出适用于 KT 事后可解释性的 Shapley Value 和 ISP 算法以及可解释性评价指标和谐度,以 KT 领域经典的深度学习模型 DKT 为例,计算历史交互与预测结果之间的相关性分数,解释 DKT 的预测结果。Shapley Value 算法计算每次交互对预测结果的贡献,将贡献视为相关性分数;ISP 算法基于原序列和模型自身的推理能力构造伪标签,实现对原序列的扰动,计算相关性分数;基于解释方法计算出的相关性分数,使用和谐度指标评价各方法的解释效果。在试验层面,5 个公开数据集上的试验结果表明,相对于最优的基线方法,本研究提出的方法取得显著的可解释性效果提升;在具体应用层面,利用可解释性挖掘知识点之间的偏序关系,帮助学生探究更加合理的学习顺序。

关键词:机器学习;深度学习;知识追踪;可解释性;特征相关性
中图分类号:TP391 **文献标志码**:A
引用格式:陈成,董永权,贾瑞,等. 基于交互序列特征相关性的可解释知识追踪[J].山东大学学报(工学版),2024,54(1):100-108.
CHEN Cheng, DONG Yongquan, JIA Rui, et al. Interpretable knowledge tracing based on the feature relevance of interaction sequence [J]. Journal of Shandong University (Engineering Science), 2024, 54(1):100-108.

Interpretable knowledge tracing based on the feature relevance of interaction sequence

CHEN Cheng[1], DONG Yongquan[1,2,3*], JIA Rui[1], LIU Yuan[1]
(1. School of Computer Science and Technology, Jiangsu Normal University, Xuzhou 221116, Jiangsu, China; 2. Jiangsu Educational Informatization Engineering Technology Research Center, Xuzhou 221116, Jiangsu, China; 3. Xuzhou Cloud Computing Engineering Technology Research Center, Xuzhou 221116, Jiangsu, China)

Abstract: In order to improve the interpretability of the knowledge tracing (KT) model, the Shapley Value and ISP algorithms suitable for the post-hoc interpretability of KT and the metrics harmony for the interpretability evaluation were proposed. Taking DKT, a classic deep learning model in the field of KT, as an example, the correlation score between history interaction and prediction results were calculated, explaining the prediction results of DKT. The Shapley Value algorithm calculated the contribution of each interaction to the prediction result, and regarded the contribution as a correlation score; the ISP algorithm constructed pseudo-labels based on the original sequence and the inferring ability of the model, realized the perturbation of the original sequence, and calculated the correlation score; based on correlation scores calculated by different explanation methods, harmony was proposed to evaluate the interpretability of each method. At the experimental aspect, the experimental results on five public datasets showed that, compared with the optimal baseline method, the method proposed in this study achieved a significant improvement in interpretability; at the specific application aspect, interpretability was used to mine the partial order relationship between skills, so as to explore a more reasonable learning sequence for students.

Keywords: machine learning; deep learning; knowledge tracing; interpretability; feature relevance

收稿日期:2023-06-25
基金项目:国家自然科学基金资助项目(61872168);江苏省教育科学十四五规划资助项目(d/2021/01/112);江苏师范大学研究生科研与实践创新计划资助项目(2022XKT1527)
第一作者简介:陈成(1999—),男,江苏南京人,硕士研究生,主要研究方向为知识追踪、可解释性。E-mail:15850526434@163.com
*通信作者简介:董永权(1979—),男,江苏宿迁人,教授,硕士生导师,博士,主要研究方向为数据集成、数据挖掘、群体智能、教育信息化。E-mail:tomdyq@jsnu.edu.cn

0　引言

知识追踪(knowledge tracing, KT)是教育数据挖掘领域的一个重要的研究方向,通过建立学生知识状态随时间变化的模型,判断学生对知识的掌握程度,从学生的学习轨迹中挖掘出潜在的学习规律,提供个性化的指导,达到人工智能辅助教育的目的[1]。文献[2]首次介绍 KT 的概念,指出 KT 可以看作是根据学生过去的学习记录,对学生的知识状态进行动态建模的过程,提出基于隐马尔科夫模型(hidden Markov model, HMM)的贝叶斯知识追踪(bayesian knowledge tracing, BKT)模型。BKT 使用二元变量对学生的知识状态进行建模,模型的输出表示是否掌握某个知识点,在每次答题结束后使用 HMM 动态更新学生的知识。文献[3-5]在 BKT 的基础上进一步优化,综合考虑了学生猜测行为、题目难度和知识点关联性等因素。

文献[6]将长短期记忆[7](long short term memory, LSTM)网络应用于 KT 中,提出深度知识追踪(deep knowledge tracing, DKT)模型,首次实现基于深度学习的知识追踪(deep-learning based knowledge tracing, DLKT)。文献[7-8]提出针对 DKT 的一些改进思路。基于 LSTM 的网络模型难以捕捉学习序列中的长距离依赖关系。

为捕捉长距离依赖关系,文献[9-10]提出基于注意力机制的 SAKT(self-attentive knowledge tracing)和 AKT(context-aware attentive knowledge tracing),文献[11-12]提出基于 Transformer 架构的 SAINT(separated self-attentive neural knowledge tracing)和 LANA(leveled attentive knowledge tracing)。为进一步提升预测准确率,文献[13-14]将图卷积网络(graph convolutional networks, GCN)应用于 KT 中,提出 SGKT(session graph knowledge tracing)和 ABKT(ability boosted knowledge tracing)。这些 DLKT 的决策过程缺乏可解释性,难以向用户提供决策依据,导致这些 DLKT 在教育领域的应用受到一定的限制。

针对深度学习的可解释性研究分为事前可解释性和事后可解释性两类,本研究主要讨论事后可解释性。文献[15]以图像分类为例,通过梯度计算出输入特征的显著性图;文献[16-17]通过扰动输入特征并观察预测值的变化量来评估被扰动特征的相关性;文献[18]以图片分类模型为例,基于影响力函数分析各个输入特征对预测值的影响;文献[19]基于 Shapley Value 理论分析 BERT 的底层逻辑;文献[20]将 Shapley Value 运用于文本分类,提出可解释的恶意言论检测模型;文献[21]研究卷积层中的信息丢弃过程,提出与信息丢弃相关的解释方法及评价指标。这些研究主要集中在计算机视觉和自然语言处理领域,针对 KT 领域的事后可解释性研究相对较少。

在 KT 领域,文献[22]将 LRP 算法[23]应用到 KT 中,计算输入特征与模型预测值的相关性分数,但未通过对比试验进行充分的验证;文献[24]深入研究 DKT 模型中的可学习参数在训练前后的差异性,但这样的研究方法仅能针对某个具体的模型作出解释,不具有针对任意模型的通用性;文献[25-26]从不确定性的角度分析模型的行为,基于不确定性的方法只能揭示预测值的波动范围,难以提供直观的解释结果。

Shapley Value 算法和基于扰动的解释方法在行人动作识别[27]、图像分类[28]、文本分类[20]等众多领域的可解释性研究中取得出色的效果,本研究基于这两种方法的解释思路,提出两种事后解释算法对 KT 领域的经典模型 DKT 进行解释。

1　问题定义

1.1　知识追踪

在现实的学习过程中,学生接触到的各类在线或离线的学习资源都遵循循序渐进的性质[29]。很难直接衡量学生在这样的学习过程中的知识状态,KT 模型通常采用一种替代的解决方案[1-2]:根据学生的历史交互 $X=\{x_1,x_2,\cdots,x_{t-1}\}$,预测学生答对下一道问题 q_t 的概率

$$p_t^{t-1}=P(a_t=1\mid X,q_t),$$

式中,$x_i=(q_i,a_i)$ 表示学生在时刻 $i(1\leqslant i\leqslant t-1)$ 的一次交互,q_i 是第 i 道题目的编号,a_i 是学生的作答结果(用 0 表示错误,1 表示正确)。

1.2　相关性分数

在知识追踪中,每一项输入特征都是一次交互。对于题目 q_t,其输入特征是历史交互 $\{x_1,x_2,\cdots,x_{t-1}\}$,为衡量每次历史交互对预测值产生的影响,使用相关性分数来量化这种影响。相关性分数是一个可以表达输入特征与预测值之间相关性的标量,值域为 $[-1,1]$。相关性分数的绝对值越大,输入特征与预测值的相关性越大,对预测值的影响也越明显;相关性分数的符号则表示输入特征对预测值的相关性为正相关或负相关。基于特征相关性的解释结果表示为相关性分数序列 $R=\{r_1,r_2,\cdots,r_{t-1}\}$。

2 Shapley Value 和 ISP 解释方法

2.1 DKT 模型

DKT 模型[6]使用 LSTM 计算学生在不同时刻的知识状态,其结构图如图 1 所示。DKT 将输入序列$\{x_1,x_2,\cdots,x_{t-1}\}$映射为输出序列$\{y_1,y_2,\cdots,y_{t-1}\}$,其中 x_i 表示学生在时刻 i 的交互,y_i 表示学生的知识状态。模型可以根据 y_i 计算出学生答对下一道题 q_{i+1}的概率。

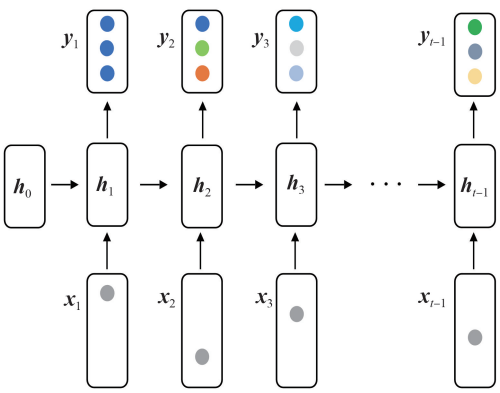


图 1 DKT 模型的结构

Fig.1 Structure of DKT model

LSTM 的计算过程如下:

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{ii}\boldsymbol{x}_t+\boldsymbol{b}_{ii}+\boldsymbol{W}_{hi}\boldsymbol{h}_{t-1}+\boldsymbol{b}_{hi}),\tag{1}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{if}\boldsymbol{x}_t+\boldsymbol{b}_{if}+\boldsymbol{W}_{hf}\boldsymbol{h}_{t-1}+\boldsymbol{b}_{hf}),\tag{2}$$

$$\boldsymbol{g}_t=\tanh(\boldsymbol{W}_{ig}\boldsymbol{x}_t+\boldsymbol{b}_{ig}+\boldsymbol{W}_{hg}\boldsymbol{h}_{t-1}+\boldsymbol{b}_{hg}),\tag{3}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{io}\boldsymbol{x}_t+\boldsymbol{b}_{io}+\boldsymbol{W}_{ho}\boldsymbol{h}_{t-1}+\boldsymbol{b}_{ho}),\tag{4}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\boldsymbol{g}_t,\tag{5}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t),\tag{6}$$

式中,$\boldsymbol{W}_{ii}$、$\boldsymbol{W}_{hi}$、$\boldsymbol{W}_{if}$、$\boldsymbol{W}_{hf}$、$\boldsymbol{W}_{ig}$、$\boldsymbol{W}_{hg}$、$\boldsymbol{W}_{io}$、$\boldsymbol{W}_{ho}$、$\boldsymbol{W}_{yh}$、$\boldsymbol{b}_{ii}$、$\boldsymbol{b}_{hi}$、$\boldsymbol{b}_{if}$、$\boldsymbol{b}_{hf}$、$\boldsymbol{b}_{ig}$、$\boldsymbol{b}_{hg}$、$\boldsymbol{b}_{io}$、$\boldsymbol{b}_{ho}$、$\boldsymbol{b}_y$ 均为可学习的参数。添加一层全连接层,将 $\boldsymbol{h}_t$ 映射为 $\boldsymbol{y}_t$:

$$\boldsymbol{y}_t=\sigma(\boldsymbol{W}_{yh}\boldsymbol{h}_t+\boldsymbol{b}_y),\tag{7}$$

式中:$\boldsymbol{W}_{yh}$、$\boldsymbol{b}_y$ 均为可学习的参数;$\boldsymbol{x}_t$ 是当前时刻交互 $\boldsymbol{x}_t$ 的 one-hot 向量,$\boldsymbol{h}_{t-1}$和 $\boldsymbol{h}_t$ 是上一时刻和当前时刻的隐向量,$\boldsymbol{i}_t$、$\boldsymbol{f}_t$、$\boldsymbol{g}_t$、$\boldsymbol{o}_t$、$\boldsymbol{c}_t$、$\boldsymbol{c}_{t-1}$是 LSTM 的中间计算结果,$\boldsymbol{y}_t$ 是学生在当前时刻的知识向量,σ( ) 和 tanh( )是激活函数。

2.2 基于 Shapley Value 的解释方法

针对 KT 的特点,提出基于 Shapley Value 理论的事后解释方法。Shapley Value 是一种合作博弈论算法,广泛应用在机器学习领域中[30],其主要目标是计算输入的各项特征对模型预测结果的贡献。基于 Shapley Value 理论,可以计算出各个时刻的相关性分数。具体计算过程如算法 1 所示。假设 DKT 进行一个时间步的预测所需时间为 $O(u)$,则本算法的时间复杂度为 $O(ut^2)$。

算法 1 基于 Shapley Value 的解释方法

输入 交互序列 X,目标问题 q_t

输出 相关性分数序列 R

```
1  R={}
2  for i from 1 to t−1
3    if i==1
4      r_i=P(a_t=1|{x_1}, q_t)
5    else
6      p_t^i=P(a_t=1|{x_1,x_2,⋯,x_i}, q_t)
7      p_t^{i−1}=P(a_t=1|{x_1,x_2,⋯,x_{i−1}}, q_t)
8      p_t^{t−1}=P(a_t=1|X,q_t)
9      r_i=(p_t^i − p_t^{i−1})/p_t^{t−1}
10   end if
11   R.append(r_i)
12 end for
13 return R
```

算法 1 的执行过程解释如下:

(1) 算法第 2~10 行建立循环,针对每个时刻 i ($1\leqslant i\leqslant t-1$)的交互,计算相关性分数。

(2) 当 i 等于 1 时,算法第 4 行将$\{x_1\}$作为输入,用 DKT 模型预测 p_t^1,令 $r_1=p_t^1$。

(3) 当 i 大于 1 时,算法第 6~8 行分别将$\{x_1,x_2,\cdots,x_i\}$,$\{x_1,x_2,\cdots,x_{i-1}\}$和 X 作为输入,用 DKT 模型预测 p_t^i,p_t^{i-1} 和 p_t^{t-1}。

(4) 第 9 行计算 p_t^i 与 p_t^{i-1} 相对于 p_t^{t-1} 的变化量 $r_i=(p_t^i-p_t^{i-1})/p_t^{t-1}$。

2.3 基于交互序列扰动的解释方法

本研究提出基于交互序列扰动(interaction sequence perturbation based, ISP)的方法,通过对输入至 DKT 模型中的交互序列进行扰动并观察输出值的变化量来计算相关性分数。KT 中的交互序列具有严格的时序性,历史交互的变化会影响到未来的结果,通过直接扰动某个时刻的历史交互得到的序列不是真实世界中的序列。

在计算机视觉和自然语言处理领域,对输入特征的扰动方法有对图片旋转、添加噪声,在句子中删词、换词等,但是大部分扰动方法难以直接应用到 KT 中。在 KT 中针对交互 $x_i=(q_i,a_i)$ 的扰动方法可以是删除 x_i 或是更改 q_i 或 a_i,但对于后者,无法确定将 q_i 或 a_i 改为何值比较合适。采用直接删除交互的方法作为扰动手段,将 DKT 模型的推理结果输出结果映射至{0,1}内作为伪标签。计算过程如算法 2 所示。假设 DKT 进行一个时间步的预测所需时间为 $O(u)$,则本算法的时间复杂度

为$O(ut^3)$。

算法 2　基于交互序列扰动的解释方法

输入　交互序列 X,目标问题 q_t,DKT 模型精度 A

输出　相关性分数序列 R

```
1  R={}
2  for i from 1 to t−1
3    if i==1
4      r_i=P(a_t=1|{x_1},q_t)
5    else
6      S_i={x_1,x_2,⋯,x_{i−1}}
7      for k from i+1 to t−1:
8        m_k^i=P(a_k=1|S_i,q_k)
9        a'_k=round(m_k^i)
10       for _ from 1 to w:
11         a'_k=randomFlip(a'_k, 1−A)
12       end for
13       x'_k=(q_k, a'_k)
14       S_i.append(x'_k)
15     end for
16     p_t^{t−1}=P(a_t=1|X,q_t)
17     m_t^i=P(a_t=1|S_i,q_t)
18     r_i=(p_t^{t−1}−m_t)/p_t^{t−1}
19   end if
20   R.append(r_i)
21 end for
22 return R
```

算法 2 的执行过程解释如下:

(1) 算法第 2~21 行建立循环,针对每个时刻 i ($1\leqslant i\leqslant t-1$)的交互,计算相关性分数。

(2) 当 i 等于 1 时,算法第 4 行将$\{x_1\}$作为输入,用 DKT 模型预测 p_t^1,令 $r_1=p_t^1$。

(3) 当 i 大于 1 时,算法第 6~15 行构造$\{q_{i+1}, q_{i+2}, \cdots, q_{t-1}\}$的伪标签$\{a'_{i+1}, a'_{i+2}, \cdots, a'_{t-1}\}$。针对时刻 k 的伪标签 a'_k构造方法如下:将序列 $S_i=\{x_1, x_2, \cdots, x_{i-1}\}$ 作为输入,用 DKT 模型预测 $m_k^i=P(a_k=1|S_i,q_k)$;将预测值 m_k^i 映射到$\{0, 1\}$中;将映射结果以 $1-A$ 的概率翻转 w 次,翻转后的值即为 q_k 的伪标签 a'_k。试验中取 $w=1$,3.7 节的试验验证了取 $w=1$ 的有效性。

(4) 记 $x'_k=(q_k, a'_k)$,$S_i=\{x_1, x_2, \cdots, x_{i-1}, x'_{i+1}, \cdots, x'_{t-1}\}$,第 16~17 行分别将 X 和 S_i 作为输入,用 DKT 模型预测 p_t^{t-1} 和 m_t^i。

(5) 第 15 行计算 p_t^{t-1} 与 m_t 相对于 p_t^{t-1} 的变化量:$r_i=(p_t^{t-1}-m_t)/p_t^{t-1}$。

2.4　挖掘知识点之间的偏序关系

由可解释性算法计算出来的相关性系数可以用于挖掘各知识点之间的偏序关系,确定合适的学习顺序。本研究将知识点 u 在知识点 v 之前的偏序关系定义为$\langle u,v\rangle$,其权重记为 J_{uv},通过 J_{uv}可以有效表示出偏序关系$\langle u,v\rangle$的强度:

$$J_{uv}=\bar{r}_i\big|_{1\leqslant i\leqslant t-1,q_i=u,q_t=v}\times 100\%。\tag{8}$$

式中:t 为序列长度,$\bar{r}_i$ 表示所有符合条件的 r_i 平均值。J_{uv}越大,则意味着$\langle u, v\rangle$越有可能是一个合适的偏序关系。本研究在 4.6 节介绍偏序关系的挖掘结果。

3　试验结果

3.1　数据集及对比方法介绍

本研究使用的 5 个数据集是 ASSIST09、ASSIST15、ASSIST17、Statics2011 和 NIPS2020。ASSIST 系列数据集由 ASSISTments 在线辅导平台收集,是目前知识追踪领域最常用的基准数据集;Statics2011 收集于某大学工程力学课程的辅导系统;NIPS2020 来自 Microsoft 和 Eedi 联合举办的教育数据挖掘竞赛。

参照现有研究工作,本研究对数据集进行预处理。对于所有数据集,删除作答结果不为 1 或 0 的交互,删除长度小于 3 的序列,删除问题名称为空的交互,将所有学生按照 8:1:1划分为训练集、验证集和测试集。预处理后的数据集统计信息如表 1 所示。

表 1　预处理后的数据集统计信息

Table 1　Statistics of datasets after preprocessing

数据集	题目数	学生数	交互总次数	学生平均交互次数	每道题平均回答次数
ASSIST09	111	3 862	325 181	84	2 930
ASSIST15	100	19 093	682 797	36	6 828
ASSIST17	101	1 708	864 711	506	8 562
Statics2011	1 224	333	261 947	787	214
NIPS2020	948	4 918	1 382 727	281	1 459

本研究使用预处理后的数据集训练 DKT 模型,在训练过程中将每个序列的最大长度设为 10。训练完成后,利用解释算法对 DKT 在测试集上对最后一个时刻的所有预测成功的结果进行解释。将分

层相关性传播(layer-wise relevance propagation, LRP)[24]和显著性图(saliency map)[17]作为基线方法,在 5 个数据集上进行可解释性对比试验,验证本研究所提方法的有效性。

3.2 评价指标

相关性分数的符号表示历史交互对当前时刻预测结果产生正面或负面的影响,可以为预测结果的解释提供基础。在理想状态下,学生应当答对所有有正面影响的题目,答错所有有负面影响的题目,在答题过程中仅学习与当前时刻题目相关的知识,避免学习无关的知识。本研究提出可解释性量化评估的基本思想为历史交互的回答情况与其相关性分数符号的一致性反映相关性分数的可解释性。

基于这种量化评估的思想,提出和谐度(harmony)作为评估指标。和谐度是指交互序列中的和谐交互所占的比例。和谐度越高,解释效果越好。在此先引入和谐交互的概念。对于交互 $x_i=(q_i,\ a_i)$ 及其相关性分数 r_i,如果 $r_i>0$ 且 $a_i=1$,或 $r_i<0$ 且 $a_i=0$,则交互 x_i 为和谐交互。设 X 中的和谐交互集为 H,则 X 的和谐度

$$h(X)=\frac{\sum\limits_{x_i\in H}|r_i|}{\sum\limits_{x_j\in X}|r_j|}\times 100\%。\tag{9}$$

对于整个数据集,取所有序列的和谐度平均值为该方法在整个数据集上的和谐度。较高的和谐度意味着大多数回答正确的题目对预测结果有着正面影响,回答错误的题目对预测结果有负面影响。

3.3 对比试验

表 2 展示各方法在 5 个数据集上可解释性的性能表现(取 5 次重复试验的均值和标准差)。相较于次优的基线方法,本研究提出的方法在 5 个数据集上的和谐度分别取得显著的性能提升,尤其是在 NIPS2020 数据集上,Shapley Value 相较于 LRP 方法,相对提升率为 33.43%;在 5 个数据集上,ISP 均能取得较高的和谐度;在 ASSIST17 和 Statics2011 数据集上,ISP 能够取得比 Shapley Value 更高的和谐度,这可能是因为这两个数据集的平均序列长度较长。

表 2 本研究方法与对比方法的和谐度对比

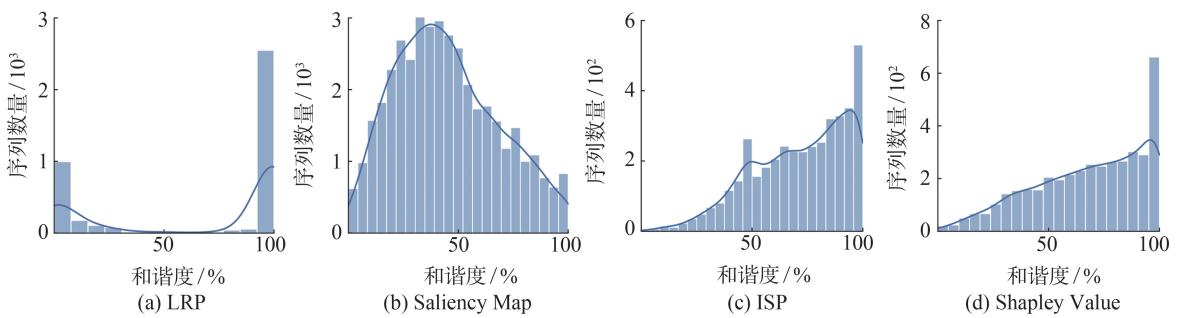
Table 2 Harmony Comparison between methods proposed in this paper and baselines

数据集	平均序列长度	和谐度/%			
		基线方法		本研究方法	
		LRP	Saliency Map	ISP	Shapley Value
ASSIST09	84	66.90±0.79	43.30±0.55	67.89±0.31	**70.83**±0.25
ASSIST15	36	71.17±0.06	29.13±0.52	70.57±0.21	**77.38**±0.36
ASSIST17	506	37.94±0.45	53.14±0.33	**53.48**±0.44	48.10±0.25
Statics2011	787	65.16±4.35	41.98±0.75	**66.39**±0.31	61.44±0.37
NIPS2020	281	52.29±0.18	45.44±0.48	63.40±0.11	**69.77**±0.12

注:相对提升率是指相对于最优基线方法的提升率;黑体数字表示横向最优结果。

3.4 和谐度分布可视化分析

为进一步探究本研究方法的优势之处,本试验以 ASSIST09 和 ASSIST17 数据集为例,展示每种方法的序列和谐度分布情况(如图 2 所示,(a)~(d)为 ASSIST09 的结果,(e)~(f)为 ASSIST17 上的结果)。从图 2 中可以看出,提出的两种方法在宏观层面上能取得较高的总体和谐度,同时在微观层面上分布于各个和谐度区间的序列数量也相对较为可观;相比之下,LRP 方法的和谐度分布情况则较为极端,在这两个数据集上,大量的序列处于和谐度极低或极高的区间内,尤其是在 ASSIST17 数据集上,较多的低和谐度序列拉低总体的和谐度;Saliency Map 的分布较为稳定,但其总体和谐度仍处于较低的水平。



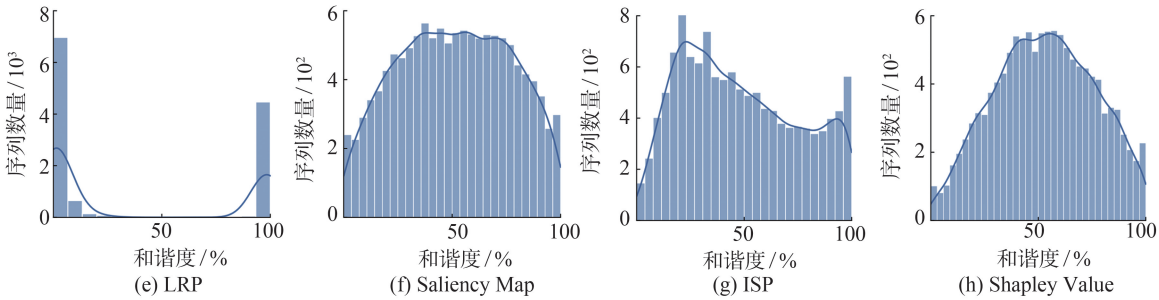


图 2　和谐度分布可视化

Fig.2　Visualization of harmony distribution

3.5　序列长度对可解释性的影响

试验以 ASSIST09 和 ASSIST17 为例,进一步探究序列长度对可解释性的影响,如图 3(a)(c)所示。考虑到过长的序列长度会导致模型的运行效率下降,故所有试验的序列长度均取 10。图 3 的试验结果表明,随着序列长度的增加,本研究所提的两种方法的和谐度基本呈现缓慢下降的趋势;在这两个数据集上,面对不同的序列长度时 Shapley Value 方法始终优于 ISP 方法;Saliency Map 方法随着序列长度的增加呈现缓慢上升的趋势,但其可解释性始终处于相对较低的水平;LRP 方法在长度 22 左右因计算出值为 nan 的相关性系数而导致大幅度的性能下降。nan(Not a Number)是一种特殊的浮点数运算过程(如除以零)导致。图 3(b)(d)展示 LRP 方法随着序列长度的增加而计算出的 nan 序列箱线图,在(x, y)处的箱线越宽,表示序列长度为 x 时含有 y 个 nan 值的序列数量越多。由图 3(b)(d)可以看出,在面对长序列时 LRP 方法的劣势较为明显,而本研究提出的两种方法以及 Saliency Map 方法不会计算出 nan,在面对不同长度的序列时,性能表现均较为稳定。

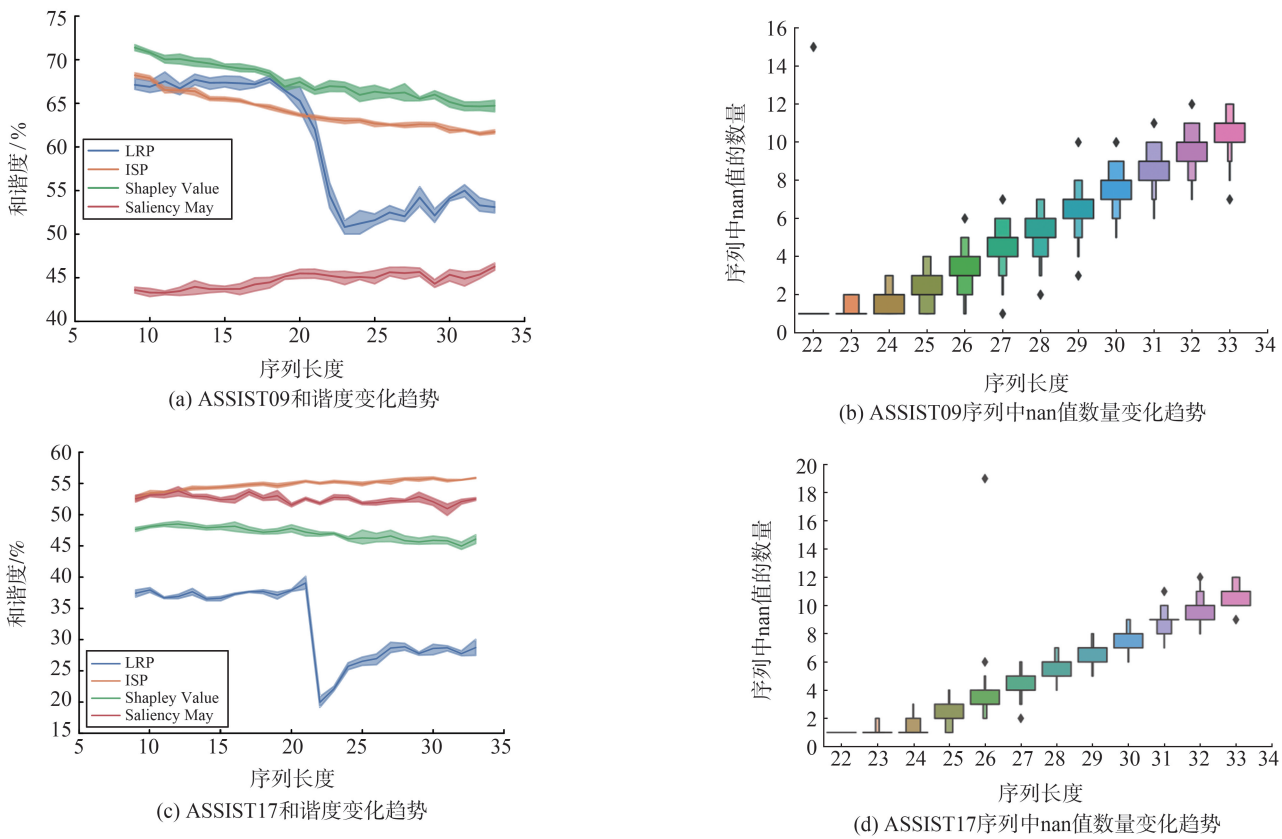


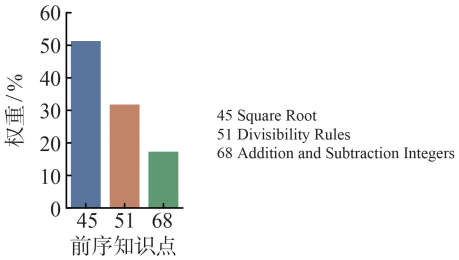
图 3　序列长度对可解释性的影响

Fig.3　Effect of sequence length on interpretability

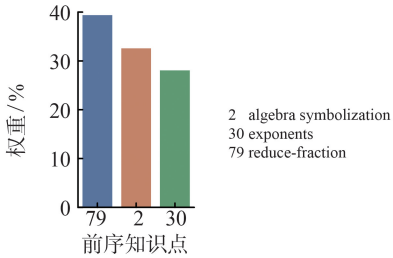
3.6　知识点偏序关系挖掘

本试验在数据集 ASSIST09 和 ASSIST17 上各选取 1 个挖掘案例,如图 4 所示。每个挖掘案例展示目标知识点 v 的所有前序知识点 u 中权重最大的

top3。为便于展示,对图中的权重进行归一化处理[26]。从图 4 中可以看出,挖掘出的 u 多为对目标知识点 v 的学习有促进、互补作用的相似知识点,或是 v 所依赖的重要知识点。例如在图 4(a)中,目标知识点 v 为 52 号知识点“Prime Number”,即质数知识点。学习平方根的知识对质数判定有促进作用,因为在搜索约数的过程中,算术平方根是搜索上界,同时在搜索约数的过程中必然涉及加减运算以及用除法求余数的过程;在图 4(b)中,目标知识点 v 为 14 号知识点“area”,即面积计算。学习面积计算之前必然需要先学习指数、约分、符号代数等重要知识。



(a) 在ASSIST09上的挖掘案例



(b) 在ASSIST17上的挖掘案例

图 4 在 ASSIST09 和 ASSIST17 上的知识点偏序关系挖掘案例

Fig.4 Results of mining skill partial order relationship on ASSIST09 and ASSIST17

3.7 ISP 方法中 w 对可解释性的影响

ISP 方法中 w 对可解释性的影响试验结果如表 3 所示。从表 3 中可以看出,当 $w=1$ 时,ISP 方法取得最佳的效果。由于 DKT 模型的预测精度未达到 100%,其输出的预测值与真实值并非完全匹配,需要将最终生成的伪标签 a'_k 随机反转 1 次以达到更好的效果;但当 w 趋向于正无穷时,会使伪标签的总体期望趋于 0.5,导致伪标签质量的下降,对最终的可解释性造成负面影响。

表 3 w 对可解释性的影响

Table 3 The effect of w on interpretability

数据集	ISP 方法中 w 的值				
	$w=0$	$w=1$	$w=2$	$w=3$	$w=4$
ASSIST09	64.09±0.36	**67.89**±0.31	67.26±0.30	67.54±0.40	67.70±0.31
ASSIST15	66.35±0.09	**70.57**±0.21	70.33±0.20	70.56±0.23	70.55±0.40
ASSIST17	48.65±0.13	**53.48**±0.44	53.08±0.32	53.45±0.16	53.47±0.55
Statics2011	65.87±0.45	**66.39**±0.31	65.95±0.32	66.30±0.33	66.36±0.37
NIPS2020	58.87±0.18	**63.40**±0.11	63.11±0.12	63.38±0.08	63.40±0.11

注:黑体数字为横向最优结果。

3.8 相关性分数可视化分析

为直观地展示本研究所提方法的效果,以及验证本研究所提指标和谐度的合理性,本节试验对 ASSIST09 数据集中 70727 号学生的学习序列以及相关性分数计算结果进行可视化分析,如图 5 所示。图 5(a)和图 5(b)分别展示 ISP 和 Shapley Value 方法对该学生一段学习序列的计算结果可视化。

在图 5(a)中,21 和 29 号知识点分别为“3D 图形网络”和“分数与小数的转换”。在此序列中待预测问题为 3D 图形网络构建的相关问题。而其之前学习过的分数与小数转换的相关知识在 3D 图构建中难以提供帮助,在其答对 29 号知识点时,相关性分数基本上保持在较低的正相关水平;而在答对 21 号知识点的时刻,相关性分数大部分都高于 29 号知识点。

在图 5(b)中,47,55 和 55 号知识点分别为“比率”、“质数”和“找规律”。在此序列中待预测问题与质数有关。对于解决质数问题,“比率”这一知识点也难以提供帮助,在答对“比率”问题的时刻,相关性分数均为负数;而学生在学习最开始的 55 号知识点“找规律”时,其相关性分数较高,这可以解释为“找规律”对质数的学习有一定的促进作用。因为根据质数的定义,学生很快就能发现“除 2 以外的偶数均为非质数”这一明显的质数特征规律;图 5(b)中对于与待预测问题 52 不相关的知识点 47,即使在回答正确的情况下其相关性分数也依然为负,而对于相关的知识点 52 和 55,算法针对答对和答错的情况会计算出正相关性和负相关性,这与实际的生活经验相吻合,从侧面证明和谐度指标的合理性。

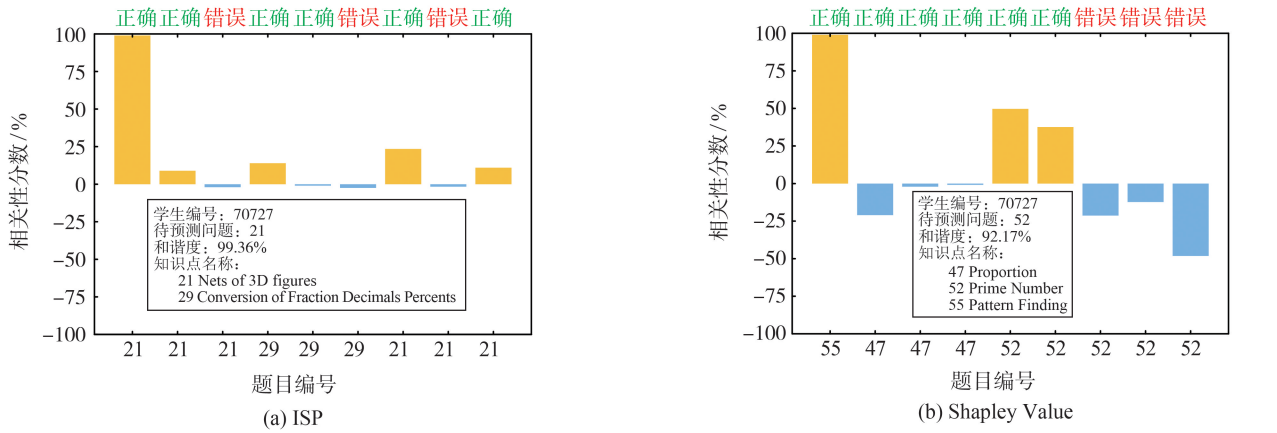


图 5　相关性分数可视化分析

Fig.5　Visualization analysis of relevance score

4　结语

针对现有 KT 模型缺乏可解释性的问题,本研究提出适用于 KT 领域事后可解释性的 Shapley Value 算法和 ISP 算法以及评估指标和谐度,根据算法计算历史交互与预测结果之间的相关性分数,使用和谐度评价各方法的可解释性。试验结果表明,相比于基线方法,两种算法在 5 个数据集上取得显著的性能提升。同时,利用可解释性挖掘知识点之间的偏序关系,帮助学生确定更加合理的学习顺序。

知识追踪在智能教育中有着广阔的应用前景,但也面临着缺乏可解释性的应用瓶颈。未来针对可解释知识追踪的研究包括 3 个方面:(1)探究多知识点 DLKT 模型的可解释性;(2)进一步分析数据集及超参数对可解释性所带来的影响;(3)进一步探究可解释性在教育领域的其它应用。

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.

[2] CORBETT A T, ANDERSON J R. Knowledge tracing: modeling the acquisition of procedural knowledge[J]. User Modeling and User-Adapted Interaction, 1994, 4(4): 253-278.

[3] KÄSER 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.

[4] BAKER R S J D., CORBETT A T, ALEVEN V. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing[C]//Intelligent Tutoring Systems. Berlin, Germany: Springer, 2008: 406-415.

[5] PARDOS Z A, HEFFERNAN N T. KT-idem: introducing item difficulty to the knowledge tracing model[C]//User Modeling, Adaption and Personalization. Berlin, Germany: Springer, 2011: 243-254.

[6] PIECH C, BASSEN J, HUANG J, et al. Deep knowledge tracing[J]. Advances in Neural Information Processing Systems, 2015, 28: 505-513.

[7] MINN S, YU Y, DESMARAIS M C, et al. Deep knowledge tracing and dynamic student classification for knowledge tracing[C]//2018 IEEE International Conference on Data Mining (ICDM). Singapore: IEEE, 2018: 1182-1187.

[8] YEUNG Chun-Kit, YEUNG Dit-Yan. Addressing two problems in deep knowledge tracing via prediction-consistent regularization[C]// Proceedings of the Fifth Annual ACM Conference on Learning at Scale. London, United Kingdom: Association for Computing Machinery, 2018: 1-10.

[9] GHOSH A, HEFFERNAN N, LAN A S. Context-aware attentive knowledge tracing[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: Association for Computing Machinery, 2020: 2330-2339.

[10] PANDEY S, KARYPIS G. A self-attentive model for knowledge tracing[EB/OL]. (2019-07-16)[2022-02-21]. http://arxiv.org/abs/1907.06837.

[11] CHOI Y, LEE Y, CHO J, et al. Towards an appropriate query, key, and value computation for knowledge tracing[C]//Proceedings of the Seventh ACM Conference on Learning @ Scale. New York, USA: Association for Computing Machinery, 2020: 341-344.

[12] ZHOU Y, LI X, CAO Y, et al. LANA: towards per-

sonalized deep knowledge tracing through distinguishable interactive sequences[EB/OL]. (2021-04-20)[2023-03-20]. http://arxiv.org/abs/2105.06266.

[13] WU Z, HUANG L, HUANG Q, et al. SGKT: session graph-based knowledge tracing for student performance prediction[J]. Expert Systems with Applications, 2022, 206: 117681.

[14] LIU S, YU J, LI Q, et al. Ability boosted knowledge tracing[J]. Information Sciences, 2022, 596: 567-587.

[15] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: visualising image classification models and saliency maps[EB/OL]. (2014-04-19)[2023-02-13]. http://arxiv.org/abs/1312.6034.

[16] LI J, MONROE W, JURAFSKY D. Understanding neural networks through representation erasure[EB/OL]. (2017-01-09)[2023-02-13]. http://arxiv.org/abs/1506.06579.

[17] FONG R C, VEDALDI A. Interpretable explanations of black boxes by meaningful perturbation[C]//2017 IEEE International Conference on Computer Vision (ICCV). New York, USA: [S.l.], 2017: 3429-3437.

[18] KOH P W, LIANG P. Understanding black-box predictions via influence functions[C]//International Conference on Machine Learning. Sydney, Australia: PMLR, 2017: 1885-1894.

[19] ZHANG H, XIE Y, ZHENG L, et al. Interpreting multivariate shapley interactions in dnns[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, USA: AAAI Press, 2021: 10877-10886.

[20] WICH M, MOSCA E, GORNIAK A, et al. Explainable abusive language classification leveraging user and network data[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Germany: Springer, 2021: 481-496.

[21] MA H, ZHANG H, ZHOU F, et al. Quantification and analysis of layer-wise and pixel-wise information discarding[C]//Proceedings of the 39th International Conference on Machine Learning. Sydney, Australia: PMLR, 2022: 14664-14698.

[22] LU Y, WANG D, MENG Q, et al. Towards interpretable deep learning models for knowledge tracing[C]//Artificial Intelligence in Education. Berlin, Germany: Springer, 2020: 185-190.

[23] BACH S, BINDER A, MONTAVON G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PLOS ONE, 2015, 10(7): 0130140.

[24] DING X, LARSON E C. Why deep knowledge tracing has less depth than anticipated[C]//Proceedings of the 12th International Conference on Educational Data Mining. Montreal, Canada: ERIC, 2019: 282-287.

[25] DING X, LARSON E C. Incorporating uncertainties in student response modeling by loss function regularization[J]. Neurocomputing, 2020, 409: 74-82.

[26] HU Q, RANGWALA H. Reliable deep grade prediction with uncertainty estimation[C]//Proceedings of the 9th International Conference on Learning Analytics & Knowledge. New York, USA: Association for Computing Machinery, 2019: 76-85.

[27] GUHA R, KHAN A H, SINGH P K, et al. CGA: a new feature selection model for visual human action recognition[J]. Neural Computing and Applications, 2021, 33(10): 5267-5286.

[28] GHORBANI A, ZOU J Y. Neuron shapley: discovering the responsible neurons[J]. Advances in Neural Information Processing Systems, 2020, 33: 5922-5932.

[29] GABRIELLA C, GRILLI L, LIMONE P, et al. Deep learning for knowledge tracing in learning analytics: an overview[C]//CEUR Workshop Proceedings. Foggia, Italy: CEUR-WS, 2021: 1-10.

[30] ROZEMBERCZKI B, WATSON L, BAYER P, et al. The shapley value in machine learning[EB/OL]. (2022-05-26)[2023-03-20]. http://arxiv.org/abs/2202.05594.

(编辑:陈燕)