采用多头注意力机制的C&RM-MAKT预测算法

王炼红,罗志辉,林飞鹏,李潇瑶
(湖南大学电气与信息工程学院,湖南长沙 410082)

摘 要: 针对深度知识追踪模型中普遍存在知识状态向量可解释性弱、缺失历史序列数据语义特征信息、忽视历史序列数据对预测结果影响程度等问题,本文提出了一种融合认知诊断理论和多头注意力机制的预测模型C&RM-MAKT(Cognitive & Response Model- Multi-head Attention Knowledge Tracing). C&RM-MAKT采用Word2Vec和BiLSTM(Bi-directional Long Short-Term Memory)网络将时序数据变换为低维连续实值向量,引入C&RM训练出的可解释性参数来建模学生学习状态,在模型机理层面将知识状态向量扩展为知识状态矩阵. 最后,C&RM-MAKT使用多头注意力机制计算出历史序列数据对预测结果的影响程度,以提高模型的可解释性与精度. 预测实验结果表明:C&RM-MAKT在HNU_SYS1、HNU_SYS2、Math1和Frcsub四个数据集上都取得了最佳性能结果,尤其在HNU_SYS2中,C&RM-MAKT相较于现有知识追踪模型在AUC(Area Uder the Curve)、ACC(ACCuracy)和F_1(F_1-Measure)指标上分别提升了4.3%、3.6%和5.9%. 此外,HNU_SYS2数据集上的可解释性分析表明:C&RM-MAKT模型内部参数可解释性强,一定程度上缓解了深度模型的"黑箱"特性.

关键词: 预测算法;知识追踪;认知诊断;注意力机制;LSTM网络;时序数据;语义特征
基金项目: 国家重点研发计划(No.2019YFE0105300);中国高等教育学会数字化课程资源专项研究课题(No.21SZYB15)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2023)05-1215-08
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220790

C&RM-MAKT Prediction Algorithm Using Multi-Head Attention Mechanism

WANG Lian-hong, LUO Zhi-hui, LIN Fei-peng, LI Xiao-yao
(*School of Electrical and Information Engineering, Hunan University, Changsha, Hunan* 410082, *China*)

Abstract: To address the problems of weak interpretability of knowledge state vectors, lackness of the semantic feature of historical sequence data, and failure to consider the influence of historical sequence data on performance prediction in existing deep knowledge tracking models, this paper proposes a predictive model C&RM-MAKT (Cognitive & Response Model-Multi-head Attention Knowledge Tracing) integrating cognitive diagnostic theory with multiple attention mechanisms. C&RM-MAKT uses Word2Vec and BiLSTM (Bi-directional Long Short-Term Memory) networks to transform the time series data into low-dimensional continuous real vectors, and applies C&RM to pre-train the interpretable parameters for student state modeling, and extends the knowledge state vectors into a knowledge state matrix at the model mechanism level. C&RM-MAKT utilizes multiheaded attention mechanism to estimate the influence degree of historical exercises on the performance prediction to improve the interpretability and accuracy of the model. The prediction experiment results show that C&RM-MAKT performs the best on datasets HNU_SYS1, HNU_SYS2, Math1, and Frcsub. Especially on dataset HNU_SYS2, C&RM-MAKT improves the existing knowledge tracking models by 4.3%, 3.6%, and 5.9% in terms of AUC (Area Uder the Curve), ACC (ACCuracy), and F_1 (F_1-Measure), respectively. In addition, according to the interpretability analysis on dataset HNU_SYS2, the internal parameters of the C&RM-MAKT model are highly interpretable, which alleviates the "black box" characteristics of the deep model to a certain extent.

Key words: prediction algorithm; knowledge tracking; cognitive diagnosis; attention mechanism; LSTM (Long Short-Term Memory) network; time series data; semantic features
Foundation Item(s): National Key R&D Program of China (No.2019YFE0105300); Special Research Project on Digital Curriculum Resources of China Association of Higher Education (No.21SZYB15)

收稿日期:2022-07-08;修回日期:2022-10-03;责任编辑:覃怀银

1 引言

近年来,知识追踪成为解决学生习题表现预测问题的主流方法. 其中,认知诊断系列静态知识追踪模型[1~3]和贝叶斯知识追踪(Bayesin Knowledge Tracing,BKT)[4~6]的主要优点在于其强大的可解释性. 然而,受限于传统模型的表征能力弱,它们在大规模数据集上的预测表现欠佳,难以对海量的学生、习题、知识点同时进行追踪.

深度知识追踪模型(Deep Knowledge Tracing,DKT)[7]能够在大规模数据集上实现对多个知识点的追踪,充分利用习题的时间序列信息,在不依赖专家标注的同时能够得出知识点间的依赖关系. 继 DKT 后,基于 RNN (Recurrent Neural Network)及其网络变体的深度知识追踪成为国内外学者的研究热点[8~12]. 纵观上述模型主要缺点在于:在深度网络中将神经网络的隐状态视为知识状态过于抽象,追踪结果可解释性较差,模型中的参数在揭示学生知识掌握状态上没有明确的指导意义.

总之,现有的深度知识追踪系列模型主要存在如下问题:(1)深度知识追踪模型的初始知识状态影响知识追踪结果的可解释性和模型收敛速度,通常采用的随机初始化参数手段导致其可解释性差;(2)学生知识状态被建模为整体的知识状态向量,难以揭示学生在各知识点上的掌握水平的变化;(3)现有方法未考虑历史作答习题对预测习题结果的贡献程度,无法关注到能对预测习题有重大影响的历史作答习题.

为了解决上述问题,本文提出了一种基于认知诊断模型和多头注意力机制的学生表现预测算法 C&RM-MAKT(Cognitive & Response Model-Multi-head Attention Knowledge Tracing),旨在融合传统模型参数可解释性良好、深度学习模型表征能力强的优势,提高知识追踪结果的可信度和习题表现预测的精度.

2 相关理论基础

2.1 认知反应模型

认知诊断模型经历了从低阶到高阶的发展过程. 研究表明,高阶模型更符合学习者的真实认知结构,如高阶 DINA(Deterministic Input Noisy And gate)模型[13]、高阶 IRT(Item Response Theory)模型[14]. 有学者在高阶模型的基础上,优化了层级间参数的建模形式[15,16]. 王等[17]扩展了能力层级架构,引入与能力参数具有补偿性质的努力参数,提出认知反应模型(Cognitive & Response Model, C&RM). 本文采用认知反应模型,用于训练深度模型中的初始"知识水平"参数及"习题-知识点考察程度"参数. C&RM 的结构见图 1. 模型参数及其表示含义见表 1.

C&RM 模型引入能力特征参数与努力特征参数的联合补偿机制,基于联合补偿机制建模学生的知识水平,如式(1)~(3)所示.

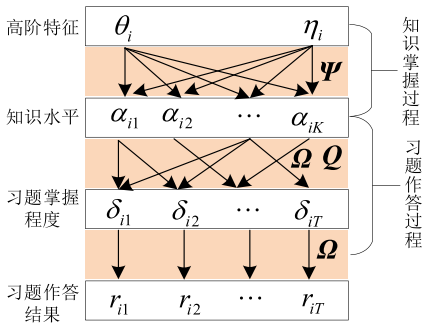


图 1 C&RM 结构图

表 1 C&RM 中参数及含义

参数名	含义
θ_i	学习者 i 的能力特征
η_i	学习者 i 的努力特征
$\boldsymbol{\Psi}=(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2\cdots,\boldsymbol{\psi}_K)$	知识特性
$\boldsymbol{\Omega}=(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_T)$	习题特性
$\boldsymbol{Q}=(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_T)$	习题知识点
δ_{it}	学习者 i 对习题 t 的掌握水平
α_{ik}	学习者 i 对知识点 k 的掌握水平
r_{it}	学习者 i 在习题 t 上的作答结果

$$\alpha_{ik}=\frac{\beta_{\theta k}w_{k1}}{1+\exp[-D(\theta_i-d_{\theta k1})]}+\frac{\beta_{\eta k}w_{k2}}{1+\exp[-D(\eta_i-d_{\eta k1})]} \tag{1}$$

$$\beta_{\theta k}=\frac{1}{1+w_{k2}\exp[-D(\eta_i-d_{\eta k2})]} \tag{2}$$

$$\beta_{\eta k}=\frac{1}{1+w_{k1}\exp[-D(\theta_i-d_{\theta k2})]} \tag{3}$$

其中,w_{k1}、w_{k2}分别表示知识点 k 对学习者能力特征、努力特征的考察权重,$\beta_{\theta k}$、$\beta_{\eta k}$分别为能力特征及努力特征对知识水平作用效果的衰减系数. $d_{\theta k1}$、$d_{\theta k2}$、$d_{\eta k1}$、$d_{\eta k2}$分别表示知识点 k 在能力特征及努力特征上的难度参数与区分度参数. D 为连续型认知诊断模型的一个经验常数,通常取值 1.7.

知识水平至习题掌握层级中,学习者 i 对习题 t 的掌握程度 δ_{it} 依赖于学习者 i 对习题 t 考察的知识点 k 的弱项参数 l_{itk}. 见式(4)和式(5).

$$\delta_{it}=1-\max(l_{it1},\cdots,l_{itK}) \tag{4}$$

$$l_{itk}=\mu_t^{(k)}(1-\alpha_{ik}) \tag{5}$$

其中,$\mu_t^{(k)}$表示知识点 k 对作答习题 t 的重要性,即"习题-知识点考察程度"参数.

学习者对习题的掌握程度与学习者最终的习题作答结果的建模关系见式(6):

$$P(r_{it}=1|\delta_{it},\lambda_{0t},\lambda_{1t})=\frac{1}{1+\exp[-D\lambda_{1t}(\delta_{it}-\lambda_{0t})]} \tag{6}$$

其中,λ_{0t}、λ_{1t}分别表示习题 t 对学习者而言的难度与区

分度. 此处D与式(1)~(3)一致,是经验常数,通常取值为1.7. C&RM中,习题t的特征向量被定义为$\boldsymbol{\omega}_t=(\lambda_{0t},\lambda_{1t},\mu_t)$,知识点$k$的特征向量定义为$\boldsymbol{\psi}_k=(w_{k1},w_{k2},d_{\theta k1},d_{\eta k1},d_{\theta k2},d_{\eta k2})$,学习者$i$的能力特征为$\theta_i$、努力特征为$\eta_i$. 其中,努力特征参数是对测验分数的排序转换计算得到,只需对参数$\boldsymbol{\theta}$、$\boldsymbol{\Omega}$、$\boldsymbol{\Psi}$进行求解,其他参数均可求出.

2.2　深度知识追踪

深度知识追踪模型采用学生作答行为的实时反馈作为建模对象. 有学者指出,基于LSTM(Long Short-Term Memory)架构的深度知识追踪模型在处理长序列数据上有更好的表现[10,12]. 因此,本文采用LSTM作为骨干网络,其单元结构如图2所示.

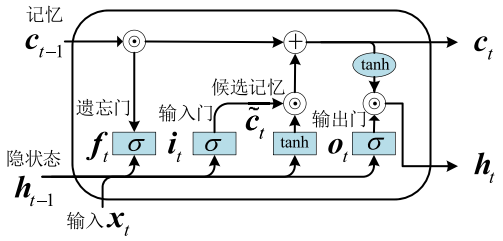


图2　LSTM模型单元结构图

图2中,遗忘门$\boldsymbol{f}_t$决定上一时刻的记忆单元中内容的遗忘程度,$\boldsymbol{i}_t$控制上一时刻记忆单元的内容更新,$\tilde{\boldsymbol{c}}_t$表示记忆单元在t时刻经输入门得到的信息,最终根据$\boldsymbol{c}_t$和输出控制门$\boldsymbol{o}_t$计算LSTM在t时刻的输出$\boldsymbol{h}_t$. LSTM单元结构如式(7)~(12)所示.

$$\boldsymbol{i}_t=\sigma(\boldsymbol{x}_t\boldsymbol{W}_{xi}+\boldsymbol{h}_{t-1}\boldsymbol{W}_{hi}+\boldsymbol{b}_i) \tag{7}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{x}_t\boldsymbol{W}_{xf}+\boldsymbol{h}_{t-1}\boldsymbol{W}_{hf}+\boldsymbol{b}_f) \tag{8}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{x}_t\boldsymbol{W}_{xo}+\boldsymbol{h}_{t-1}\boldsymbol{W}_{ho}+\boldsymbol{b}_o) \tag{9}$$

$$\tilde{\boldsymbol{c}}_t=\tanh(\boldsymbol{x}_t\boldsymbol{W}_{xc}+\boldsymbol{h}_{t-1}\boldsymbol{W}_{hc}+\boldsymbol{b}_c) \tag{10}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\tilde{\boldsymbol{c}}_t \tag{11}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t) \tag{12}$$

3　C&RM-MAKT算法设计

3.1　知识追踪任务问题定义

知识追踪模型的输入为学生的历史习题记录. 学生集合表示为$U=\{u_1,u_2,\cdots,u_i,\cdots,u_I\}$,已作答习题集合表示为$E=\{e_1,e_2,\cdots,e_t,\cdots,e_T\}$. 学生的作答结果集合表示为$R=\{r_1,r_2,\cdots,r_t,\cdots,r_T\}$. 知识追踪模型根据输入数据,解决两大问题:(1)预测学生在未来习题上正确作答的概率r_{T+1};(2)追踪学生每作答一道习题后的知识点掌握水平变化.

3.2　C&RM-MAKT模型整体框架

C&RM-MAKT的整体框架如图3所示. 模型由五个模块组成. 五个模块分别为:习题的嵌入式表示、基于C&RM的参数训练、结合多头注意力机制的习题相似度计算、融合习题相似度和认知诊断参数的学生状态建模、学生习题表现结果预测输出. 如算法1所示.

3.2.1　习题的嵌入式表示

习题的基本组成元素是文本与公式,传统的独热编码等编码方式极易丢失习题的语义特征信息. 为了解决这一问题,本文采用双向长短期记忆网络(Bi-directional Long Short Term Memory,BiLSTM)构建习题的句模型. 习题的嵌入式表示模型如图4所示.

习题文本转化为向量的过程分为两个阶段:(1)词模型训练:使用Word2Vec工具在习题语料库上训练得到第t道习题e_t的向量组表示:$\boldsymbol{N}_t=(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_m,\cdots,\boldsymbol{w}_M)$,其中,$\boldsymbol{w}_m$表示第$m$个分词的词向量;(2)句模型训练:在第$m$个分词上,双向架构中对应该分词的前向和反向隐向量分别建模为$\overrightarrow{\boldsymbol{g}_m}$、$\overleftarrow{\boldsymbol{g}_m}$. 具体计算过程如式(13)和式(14)所示:

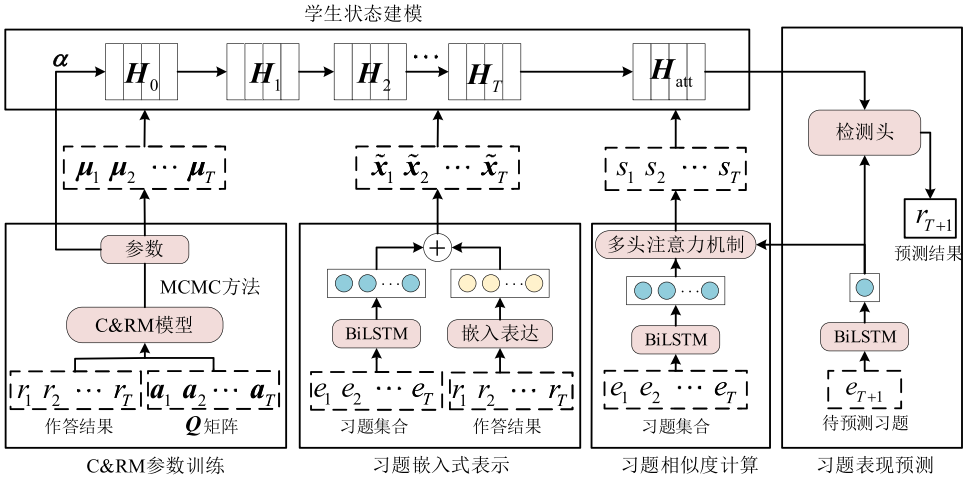


图3　C&RM-MAKT整体框架

算法1 C&RM-MAKT算法

输入:学生作答结果集合R、习题-知识点关联矩阵$\boldsymbol{Q}$、已作答习题集合E、目标预测习题e_{T+1}

输出:预测习题作答结果r_{T+1}

随机初始化参数初始值$\boldsymbol{\Psi}^{(1)}$、$\boldsymbol{\theta}^{(1)}$、$\boldsymbol{\Omega}^{(1)}$

FOR $z=1,\cdots,$ itertion

$\boldsymbol{\Psi}^{(z+1)},\boldsymbol{\theta}^{(z+1)},\boldsymbol{\Omega}^{(z+1)}=\mathrm{MCMC}(\boldsymbol{\Psi}^{(z)},\boldsymbol{\theta}^{(z)},\boldsymbol{\Omega}^{(z)})$

END FOR

$\boldsymbol{\alpha},\boldsymbol{\mu}=\mathrm{C\&RM}(\boldsymbol{\Psi}^{(z+1)},\boldsymbol{\theta}^{(z+1)},\boldsymbol{\Omega}^{(z+1)})$

FOR $e_t\in E$

$\tilde{\boldsymbol{x}}_t=\mathrm{BiLSTM}(e_t)$ //习题的嵌入式表达

$s_t=\text{multi-attention}(e_t)$//计算与预测习题的相似度

END FOR

$\tilde{\boldsymbol{x}}_t^{(k)}=\mu_t^{(k)}\tilde{\boldsymbol{x}}_t$//$\mu_t^{(k)}$为习题$t$对知识点$k$的考察程度

$\boldsymbol{H}_0=\boldsymbol{\alpha}$//C&RM训练出的$\boldsymbol{\alpha}$作为初始知识掌握状态

$\boldsymbol{H}_t=(\boldsymbol{h}_t^{(1)},\boldsymbol{h}_t^{(2)},\cdots,\boldsymbol{h}_t^{(k)},\cdots,\boldsymbol{h}_t{}^{(K)})//t=1,2,3,\cdots,T$

$\boldsymbol{h}_t^{(k)}=\mathrm{LSTM}(\tilde{\boldsymbol{x}}_t^{(k)},\boldsymbol{h}_{t-1}^{(k)})$//追踪知识点$k$的掌握情况

$\boldsymbol{H}_{\text{att}}=\sum s_t\boldsymbol{H}_t$

$r_{T+1}=\mathrm{Predict}(\boldsymbol{H}_{\text{att}},e_{T+1})$//预测习题表现

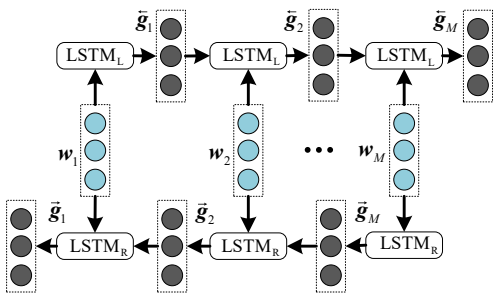


图4 习题嵌入式表示

$$\overrightarrow{\boldsymbol{g}_m}=\sigma(\boldsymbol{w}_m\boldsymbol{W}_{wh}^{(f)}+\overrightarrow{\boldsymbol{g}_{m-1}}\boldsymbol{W}_{hh}^{(f)}+\boldsymbol{b}_h^{(f)}) \tag{13}$$

$$\overleftarrow{\boldsymbol{g}_m}=\sigma(\boldsymbol{w}_m\boldsymbol{W}_{wh}^{(b)}+\overleftarrow{\boldsymbol{g}_{m-1}}\boldsymbol{W}_{hh}^{(b)}+\boldsymbol{b}_h^{(b)}) \tag{14}$$

其中,σ是激活函数,$\boldsymbol{W}_{wh}^{(f)}$、$\boldsymbol{W}_{hh}^{(f)}$、$\boldsymbol{b}_h^{(f)}$为模型前向权重参数和偏置参数,$\boldsymbol{W}_{wh}^{(b)}$、$\boldsymbol{W}_{hh}^{(b)}$、$\boldsymbol{b}_h^{(b)}$为反向权重参数和偏置参数. 前向隐向量和反向隐向量拼接得到$\boldsymbol{w}_m$的隐向量$\boldsymbol{g}_m=(\overrightarrow{\boldsymbol{g}_m},\overleftarrow{\boldsymbol{g}_m})$. 最后,为了获取习题$e_t$的整体语义表示,本文对隐向量采取最大池化操作,得到习题的句模型表示,如式(15)所示.

$$\boldsymbol{x}_t=\mathrm{maxpooling}(\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_m,\cdots,\boldsymbol{g}_M) \tag{15}$$

句模型$\boldsymbol{x}_t$根据作答结果与零向量进行拼接:若作答正确,则拼接在零向量之前;若作答错误,则拼接在零向量之后,得到"习题-作答结果"拼接向量$\tilde{\boldsymbol{x}}_t$.

3.2.2 基于C&RM的参数训练

本文所提出的算法使用C&RM诊断出学生完成前$n(n\ll T)$道习题后的"知识水平"参数$\boldsymbol{\alpha}$,代替深度模型中随机初始化的学生初始状态. 为了提高学生知识状态的可解释性,C&RM-MAKT将认知诊断模型中具有明确可解释性意义的"习题-知识点考察程度"参数$\boldsymbol{\mu}$引入到知识追踪过程中. 为了适应这一参数,C&RM-MAKT将学生的知识状态向量扩展为知识状态矩阵,矩阵的列向量代表对应知识点的掌握程度向量.

参数$\boldsymbol{\alpha}$和参数$\boldsymbol{\mu}$为模型中待求解的未知参数,由式(1)~(6)各参数之间的关系可知,只需要对参数$\boldsymbol{\theta}$、$\boldsymbol{\Psi}$、$\boldsymbol{\Omega}$进行求解,未知参数$\boldsymbol{\alpha}$、$\boldsymbol{\mu}$就可相继求出. 本文采用马尔可夫蒙泰卡罗参数估计算法(Markov Chain Monte Carlo, MCMC)对参数$\boldsymbol{\theta}$、$\boldsymbol{\Psi}$、$\boldsymbol{\Omega}$进行求解[18].

3.2.3 结合多头注意力机制的习题相似度计算

现有的相似度计算模型大多数基于的是距离度量,如Cosine[19],Jaccard[20],忽视了习题语义特征,受到Transformer模型中多头注意力机制启发[21],本文提出结合多头注意力机制计算习题相似度,挖掘习题语义特征信息,区分学习者历史作答习题对于预测习题的贡献程度. 多头注意力机制的结构如图5所示.

本文所计算的习题相似度来自于多个注意力头的汇聚,学习到习题向量对应的查询、键、值的子空间表示,以便捕获序列间长短距离依赖关系. 本文使用式(16)计算向量$\boldsymbol{x}_t$在第n个注意力头对应的查询向量$\boldsymbol{q}_t^{(n)}$、键向量$\boldsymbol{k}_t^{(n)}$、值向量$\boldsymbol{v}_t^{(n)}$.

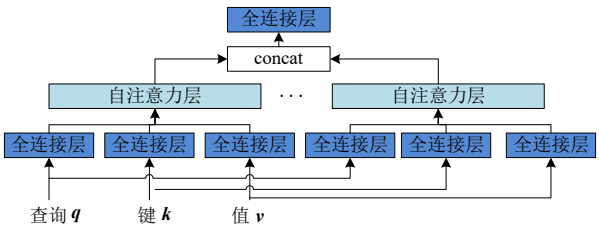


图5 多头注意力机制

$$\boldsymbol{q}_t^{(n)}=\boldsymbol{x}_t\boldsymbol{W}_n^{(q)},\boldsymbol{k}_t^{(n)}=\boldsymbol{x}_t\boldsymbol{W}_n^{(k)},\boldsymbol{v}_t^{(n)}=\boldsymbol{x}_t\boldsymbol{W}_n^{(v)} \tag{16}$$

其中,$\boldsymbol{W}_n{}^{(q)}$、$\boldsymbol{W}_n{}^{(k)}$、$\boldsymbol{W}_n{}^{(v)}$分别为查询向量、键向量、值向量在第n个注意力头中的投影矩阵,它们将各自向量投影到对应的子空间. 其中,自注意力的计算使用式(17)进行缩放点积运算.

$$\mathrm{attention}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v})=\mathrm{softmax}(\frac{\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}}{\sqrt{d}})\boldsymbol{v} \tag{17}$$

本文使用三头注意力机制学习习题在子空间中的相似度信息$\boldsymbol{s}_t^{(n)}$,将学习到的不同子空间下的注意力权重进行拼接,得到多头注意力机制下习题t与预测习题的相似度s_t,见式(18)和式(19).

$$s_t=\mathrm{concat}(\boldsymbol{s}_t^{(1)},\boldsymbol{s}_t^{(2)},\boldsymbol{s}_t^{(3)})\boldsymbol{W}^{(O)} \tag{18}$$

$$\boldsymbol{s}_t^{(n)}=\mathrm{attention}(\boldsymbol{q}_t^{(n)},\boldsymbol{k}_t^{(n)},\boldsymbol{v}_t^{(n)}),n=1,2,3 \tag{19}$$

由于结合多头注意力机制的习题相似度参数是针对习题预测任务的定制优化参数,从习题语义特征的维度挖掘习题在多个子空间下的相似度信息,因此具备更强的表征能力.

3.2.4 融合习题相似度与认知诊断参数的学生状态建模

学生状态建模的网络融入了“习题-知识点考察程度”参数$\boldsymbol{\mu}=(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_t,\cdots,\boldsymbol{\mu}_T)$和习题相似度参数向量$\boldsymbol{s}=(s_1,s_2,\cdots,s_t,\cdots,s_T)$,能在提高学生知识状态的可解释性的同时提升预测性能. 其中,$\boldsymbol{\mu}_t=(\mu_t^{(1)},\mu_t^{(2)},\cdots,\mu_t^{(k)},\cdots,\mu_t^{(K)})^{\mathrm{T}}$,$\mu_t^{(k)}$表示习题$t$对知识点$k$的考察权重,作为先验参数引入到学生状态建模当中. 参数$\mu_t^{(k)}$与拼接作答结果信息的习题向量做乘积,如式(20)所示:

$$\tilde{\boldsymbol{x}}_t^{(k)}=\mu_t^{(k)}\tilde{\boldsymbol{x}}_t \tag{20}$$

融合了知识点考察参数信息的习题向量$\tilde{\boldsymbol{x}}_t^{(k)}$作为LSTM模型的输入. 如式(21)所示. 学生在t时刻的隐状态矩阵为$\boldsymbol{H}_t=(\boldsymbol{h}_t^{(1)},\boldsymbol{h}_t^{(2)},\cdots,\boldsymbol{h}_t^{(k)},\cdots,\boldsymbol{h}_t^{(K)})$. 其中,向量$\boldsymbol{h}_t^{(k)}$表征学生在做完第$t$道习题后在第$k$个知识点上的掌握程度. 式(21)实现了对于学生每做完一道习题后的实时知识状态追踪. 弥补了静态认知诊断模型的不足.

$$\boldsymbol{h}_t^{(k)}=\mathrm{LSTM}(\tilde{\boldsymbol{x}}_t^{(k)},\boldsymbol{h}_{t-1}^{(k)}) \tag{21}$$

在3.2.3节求得习题相似度后,以习题相似度为权重对各阶段的知识状态矩阵$\boldsymbol{H}_t$进行加权求和得到学生作答第t道习题后的隐状态矩阵$\boldsymbol{H}_{\mathrm{att}}$,见式(22). 矩阵的列向量表示对应知识点的掌握向量,即$\boldsymbol{H}_{\mathrm{att}}=(\boldsymbol{h}_{\mathrm{att}}^{(1)},\boldsymbol{h}_{\mathrm{att}}^{(2)},\cdots,\boldsymbol{h}_{\mathrm{att}}^{(k)},\cdots,\boldsymbol{h}_{\mathrm{att}}^{(K)})$.

$$\boldsymbol{H}_{\mathrm{att}}=\sum_{t=1}^{T}s_t\boldsymbol{H}_t \tag{22}$$

3.2.5 学生表现结果预测输出

在线预测阶段中,基于注意力机制的习题文本相似度计算,可以很好地解决新习题的冷启动问题. C&RM模型训练的参数$\boldsymbol{\mu}_{T+1}$保证了预测结果的可解释性. 其中,$\mu_{T+1}^{(k)}$表示待预测的第$T+1$道习题在知识点k上的考察程度,与学生对各知识点的掌握水平向量进行数乘运算得到预测向量$\boldsymbol{p}_{T+1}$,如式(23)所示.

$$\boldsymbol{p}_{T+1}=\sum_{k=1}^{K}\mu_{T+1}^{(k)}\boldsymbol{h}_{\mathrm{att}}^{(k)} \tag{23}$$

预测向量与习题特征向量拼接并输入到双层神经网络中得到预测结果,如式(24)和式(25)所示.

$$\boldsymbol{y}_{T+1}=\mathrm{Re\,Lu}(\boldsymbol{W}_1\cdot[\boldsymbol{p}_{T+1}\oplus\boldsymbol{x}_{T+1}]+\boldsymbol{b}_3) \tag{24}$$

$$\tilde{r}_{T+1}=\sigma(\boldsymbol{W}_2\cdot\boldsymbol{y}_{T+1}+\boldsymbol{b}_2) \tag{25}$$

损失函数采用交叉熵损失函数,采用Adam优化器优化式(26)所示的目标函数.

$$L=-\sum_{t=1}^{T}(r_t\log\tilde{r}_t+(1-r_t)\log(1-\tilde{r}_t)) \tag{26}$$

4 实验结果与分析

本文所有实验均在以下工作环境中进行:Windows10操作系统,主机使用Intel Core i9-10900K 10核CPU,主频为3.7 GHz,内存为32 GB.

4.1 数据集介绍

数据集HNU_SYS1和HNU_SYS2是本文自建数据集,来源于中国大学MOOC平台上收集的湖南大学“信号与系统”课程真实作答记录. 数据分别来源于2018年秋季学期和2019年秋季学期的学习者作答记录,数据清洗过程中剔除了作答记录小于5条的记录. 此外还有中国科学技术大学黄振亚等人采集的分数加减法数据集Frcsub、高中数学作答数据集Math1. 各数据集的详细信息见表2.

表2 数据集信息

数据集	HNU_SYS1	HNU_SYS2	Math1	Frcsub
学生数量	466	770	4 209	536
习题数量	90	221	15	20
知识点数量	6	50	11	8
习题作答记录数	21 450	84 320	63 515	10 720

4.2 评价指标

知识追踪任务的常用评估指标有ROC曲线下与坐标轴围成图形的面积(Area Under the Curve,AUC)、准确率(ACCuracy,ACC)、和F_1值. 计算过程如式(27)~(29)所示. 其中,TP表示预测学生答对,实际答对的样本数,FN表示预测学生答错,实际学生答对的样本数,FP表示预测学生答对,实际学生答错的样本数.

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{27}$$

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{28}$$

$$F_1=\frac{2\times\mathrm{Recall}\times\mathrm{Precision}}{\mathrm{Recall}+\mathrm{Precision}} \tag{29}$$

4.3 消融实验

为了验证习题向量化、认知诊断参数及多头注意力机制对模型预测性能的影响. 本文设计了消融实验,消融实验中,“E”表示习题向量化模块、“CD”表示认知诊断模块,“A”表示多头注意力机制模块. 此外,数据集HNU_SYS1和HNU_SYS2收集了习题的文本信息,Frcsub和Math1数据集未收集习题文本信息. 为了便于开展习题向量化模块的消融实验,本文选择在HNU_SYS2数据集上进行消融实验,具体结果见表3.

从表3中可以看出,本文方法C&RM-MAKT所加的习题向量化模块、认知诊断模块以及多头注意力机制模块均提高了知识追踪模型的性能. 当训练集比例为30%~80%时,添加单一模块的AUC值和ACC值均比基准模型DKT更优.

可以看出,多头注意力机制模块相比其他模块对模型预测性能的提升更明显,例如,在70%训练集下,“DKT+A”方法比“DKT+CD”方法在AUC指标上提升了

表3 HNU_SYS2数据集上消融实验结果

指标	模型	训练集比例					
		30%	40%	50%	60%	70%	80%
AUC	DKT	0.720	0.724	0.729	0.737	0.744	0.732
	DKT+E	0.726	0.728	0.739	0.743	0.749	0.746
	DKT+CD	0.727	0.731	0.735	0.742	0.746	0.747
	DKT+A	0.746	0.740	0.747	0.756	0.769	0.753
	DKT+E+CD+A	**0.755**	**0.757**	**0.759**	**0.763**	**0.787**	**0.786**
ACC	DKT	0.768	0.770	0.764	0.772	0.777	0.769
	DKT+E	0.772	0.778	0.779	0.779	0.782	0.783
	DKT+CD	0.773	0.777	0.780	0.779	0.780	0.782
	DKT+A	0.777	0.783	0.783	0.784	0.786	0.790
	DKT+E+CD+A	**0.786**	**0.789**	**0.797**	**0.806**	**0.813**	**0.818**
F_1	DKT	0.861	0.862	0.860	0.877	0.865	0.858
	DKT+E	0.862	0.863	0.865	0.870	0.876	0.875
	DKT+CD	0.864	0.870	0.869	0.873	0.877	0.875
	DKT+A	0.886	0.881	0.872	0.875	0.896	0.887
	DKT+E+CD+A	**0.890**	**0.901**	**0.917**	**0.923**	**0.924**	**0.917**

2.3%,“DKT+A”方法比“DKT+E”方法提升了2.0%. 此外,相对于只增加了单一模块,“DKT+E+CD+A”方法在ACC、AUC、F_1指标上均取得最佳结果. 这是因为本文模型将DKT中的知识状态向量扩展为知识状态矩阵,提高了知识追踪结果的可解释性,多头注意力机制能够在预测学生表现时自适应地捕获历史作答习题对预测习题的贡献程度.

4.4 对比实验

本文选取了6个具有代表性的动静态知识追踪模型设计了对比实验,考虑的6个对比基准方法分别是PMF[22]、HO-DINA[13]、FuzzyCDM[15]、NeuralCD[16]、DKT[7]、DKVMN[10]. 基准方法的介绍和参数设置如下:

PMF:将学生和习题映射成低维度潜在向量. 其中,学生与习题潜在特征向量设置为10维.

HO-DINA:考虑了学习者自身能力特征对知识水平的作用关系,扩展了认知诊断的层级架构.

FuzzyCDM:将知识能力水平表示为模糊集合的隶属度,考虑了教育学中联结型和补偿型的认知作答模式.

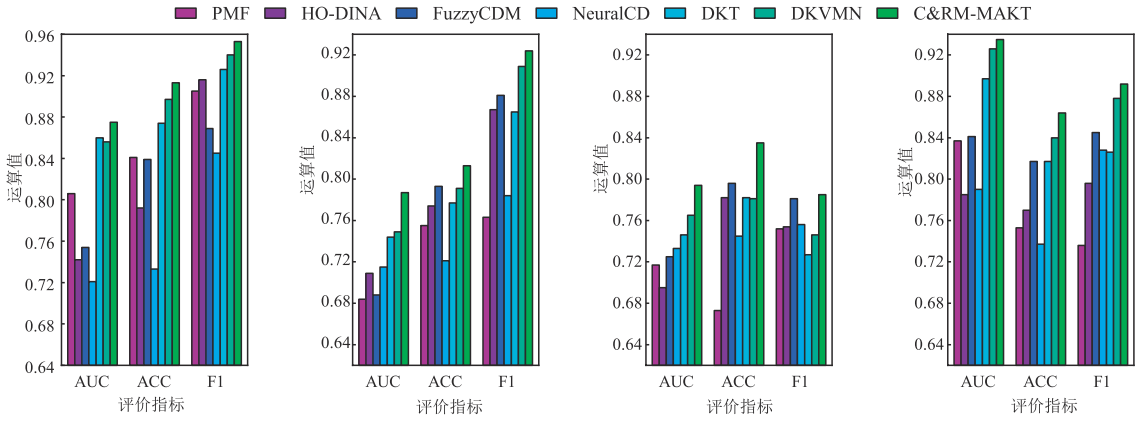
NeuralCD:使用前馈神经网络学习学生作答习题的复杂交互过程.

DKT:使用循环神经网络处理知识追踪任务.

DKVMN:使用动态键值对记忆网络存储并更新学生的知识概念掌握状态.

图6展示了C&RM-MAKT模型与主要的静态知识追踪(认知诊断模型)和动态知识追踪模型在四个数据集上的对比结果.

其中,模型训练过程中,针对Frcsub、Math1数据集,习题向量化采用独热编码方式,针对HNU_SYS1和



(a) HNU_SYS1 (b) HNU_SYS2 (c) Math1 (d) Frcsub

图6 动静态知识追踪模型对比结果

HNU_SYS2数据集,习题向量化采用BiLSTM网络训练. 四个数据集的训练集比例均设置为70%. 其中,NeuralCD、DKT、DKVMN、C&RM-MAKT训练过程中基本参数设置如下:迭代轮次(epoch)为100,批处理数量(batch_size)设为16,学习率(learning rate)为0.002.

从图6结果可以看出,本文所提出的方法在四个数据集上表现最优. C&RM-MAKT的性能远超传统方法. 本文所提出的方法相比DKT在各指标上的性能提升如表4所示.

不难看出C&RM-MAKT预测算法在大数据集上(HNU_SYS系列数据集和Math1数据集)相比于Frcsub数据集取得了更明显的提升,我们认为这是由于Frcsub数据集的习题数量较少,且Frcsub缺乏习题文本向量化的处理,C&RM-MAKT因此在更大规模数据集上表现更优.

表4　C&RM-MAKT相比DKT的性能提升　　单位:%

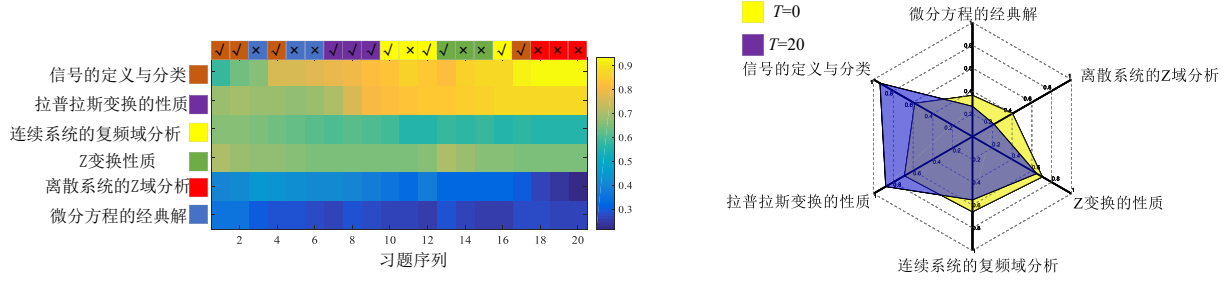
数据集	指标		
	ACC	AUC	F_1
HNU_SYS1	3.9	1.5	2.7
HNU_SYS2	3.6	4.3	5.9
Frcsub	2.7	1.8	1.6
Math1	5.3	1.2	5.5

4.5　知识追踪可解释性分析

为了说明C&RM-MAKT的可解释性,本文从HNU_SYS2数据集中选取了一位学生的习题作答序列,追踪其在习题作答过程中对某些知识的掌握水平变化. 如图7所示. 为了更清晰地进行可视化呈现,本文做了一些处理:首先选择了学生最常练习的6个知识点,若将HNU_SYS2数据集中所有50个知识点均可视化在一张图中很难清晰呈现;其次,热力图所展示的习题均为考察单个知识点的习题,以便准确地对学生知识状态变化进行归因.

如图7(*a*)所示,该学生回答了6个知识点上的20个问题. 热力图的纵坐标为知识点名称,横坐标表示学生作答的习题序列,其习题考察的知识点在图最上方用色块标记:色块中"√"代表实际回答正确,"×"代表实际回答错误. 热力图中色块颜色的深浅代表学生对知识点的掌握程度. 从图中可以明显看出:经C&RM-MAKT模型追踪的知识水平变化较为平滑,符合教育学规律,具备良好的可解释性. 具体来说,当该学生回答一个习题正确(错误)时,该学生对相应概念的知识掌握程度会增加(减少). 可以看出,该学生逐渐掌握了"信号的定义与分类"的概念,却无法理解"离散系统的Z域分析",因为该学生对"信号的定义与分类"的所有练习均作答正确,但"离散系统的Z域分析"的所有练习均作答错误.

当学生作答完20道习题后,可绘制出学生的知识掌握雷达图,如图7(*b*)所示,可以看出:该学生已经很好地掌握了"信号的定义与分类"、"拉普拉斯变换的性质"知识;部分掌握了"Z变换的性质"、"连续系统的复频域分析"知识;但在"微分方程的经典解"、"离散系统的Z域分析"知识方面基本未掌握. 上述分析表明:本文所提出的算法能够较好地分析和解释预测结果,在实际教学应用中具备重要指导意义.



(*a*) 知识追踪结果热力图展示　　(*b*) 知识掌握情况雷达图

图7　C&RM-MAKT算法可解释性分析

5　结论

深度知识追踪模型的"黑箱"特性使得知识追踪结果难以得到高置信度的可解释性意义,且忽视了历史作答习题对预测习题的影响这一因素. 本文针对以上问题设计了基于认知诊断参数和多头注意力机制的C&RM-MAKT模型,在多个公开数据集上的结果表明,习题之间的相似程度影响着最终答题结果,良好的习题嵌入方法也影响着学生表现预测效果. 在未来的研究中,关注学习者、习题、知识点的更高维特征是一个改进的方向,本文模型未考虑习题背后的知识点关联网络的拓扑结构信息,未来需要对知识点嵌入方法进行探索研究.

参考文献

[1] FAN X. Item response theory and classical test theory: An empirical comparison of their item/person statistics[J]. Educational and Psychological Measurement, 1998, 58(3): 357-381.

[2] DE LA TORRE J. DINA model and parameter estimation: A didactic[J]. Journal of Educational and Behavioral Statistics, 2009, 34(1): 115-130.
[3] HARTZ M C. A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality[J]. American Journal of Gastroenterology, 2002, 95(4): 906-909.
[4] CORBETT A T, ANDERSON J R. Knowledge tracing: Modeling the acquisition of procedural knowledge[J]. User Modeling and User-Adapted Interaction, 1994, 4(4): 253-268.
[5] HAWKINS W J, HEFFERNAN N T. Using similarity to the previous problem to improve Bayesian knowledge tracing[C]//Proceedings of the Workshops held at Educational Data Mining 2014 (WSEDM 2014). London: CEUR-WS, 2014: 136-140.
[6] AGARWAL D, BAKER R, MURALEEDHARAN A. Dynamic knowledge tracing through data driven recency weights[C]//The 13th International Conference on Educational Data Mining. Morocco: Open Access, 2020: 725-729.
[7] PIECH C, SPENCER J, HUANG J, et al. Deep knowledge tracing[J]. Computer Science, 2015, 3(3): 19-23.
[8] YEUNG C K, YEUNG D Y. Addressing two problems in deep knowledge tracing via prediction-consistent regularization[C]//Proceedings of the 5th Annual ACM Conference on Learning at Scale. London: ACM, 2018: 1-10.
[9] MINN S, YI Y, DESMARAIS M C, et al. Deep knowledge tracing and dynamic student classification for knowledge tracing[C]//2018 IEEE International Conference on Data Mining. Singapore: IEEE, 2018: 1182-1187.
[10] ZHANG J, SHI X, KING I, et al. Dynamic key-value memory networks for knowledge tracing[C]//Proceedings of the 26th International Conference on World Wide Web. Perth: ACM, 2017: 765-774.
[11] SUN X, ZHAO X, LI B, et al. Dynamic key-value memory networks with rich features for knowledge tracing[J]. IEEE Transactions on Cybernetics, 2022, 52(8): 8239 - 8245.
[12] LIU Q, HUANG Z, YIN Y, et al. EKT: Exercise-aware knowledge tracing for student performance prediction[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(1): 100-115.
[13] TORRE J, DOUGLAS J A. Higher-order latent trait models for cognitive diagnosis[J]. Psychometrika, 2004, 69 (3): 333-353.
[14] DE L, SONG H. Simultaneous estimation of overall and domain abilities: a higher-order IRT model approach[J]. Applied Psychological Measurement, 2009, 33(8): 620-639.
[15] LIU Q, WU R Z, CHEN E H, et al. Fuzzy cognitive diagnosis for modelling examinee performance[J]. ACM Transactions on Intelligent Systems and Technology, 2018, 9(4): 1-26.
[16] WANG F, LIU Q, CHEN E, et al. Neural cognitive diagnosis for intelligent education systems[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 6153-6161.
[17] 王炼红, 刘畅, 周熊, 等. 基于学习者认知反应模型的认知诊断方法: CN202110122198.0[P]. 2021-05-07.
[18] GELFAND A E, HILLS S E, RACINE-POON A. Illustration of Bayesian inference in normal data models using Gibbs sampling[J]. Journal of the American Statistical Association, 1990, 85(412): 972-985.
[19] WEN H, DING G, LIU C, et al. Matrix factorization meets cosine similarity: Addressing sparsity problem in collaborative filtering recommender system[C]//The 16th Asia-Pacific Web Conference. Cham: Springer, 2014: 306-317.
[20] BAG S, KUMAR S K, TIWARI M K. An efficient recommendation generation using relevant Jaccard similarity [J]. Information Sciences, 2019, 483: 53-64.
[21] SUBAKAN C, RAVANELLI M, CORNELL S, et al. Attention is all you need in speech separation[C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021: 21-25.
[22] SALAKHUTDINOV R, MNIH A. Probabilistic matrix factorization[C]//Advances in Neural Information Processing Systems 20 (NIPS 2007). Vancouver: ACM, 2007: 1257-1264.

作者简介

王炼红 女,1971年5月生,湖南宁乡人. 博士,副教授、硕士生导师. 2011年3月至2012年3月,于美国布兰迪斯大学做访问学者. 主要研究方向为信号处理、数据挖掘技术、人工智能.
E-mail: wanglh@hnu.edu.cn

罗志辉 男,1998年9月出生于湖南省永州市. 2020年于南京农业大学获得工学学士学位. 现为湖南大学电气与信息工程学院硕士研究生,主要研究方向为教育数据挖掘和机器学习.
E-mail: luo1998@hnu.edu.cn