第49卷第1期 中国科学技术大学学报 Vol. 49, No. 1
2019年1月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jan. 2019

文章编号:0253-2778(2019)01-0021-10

学生得分预测:一种基于知识图谱的卷积自编码器

苏 喻[1],张 丹[2],刘青文[2],张英杰[2],陈玉莹[3],丁宏强[1]

(1.安徽大学计算机科学与技术学院,安徽合肥 230039;2.科大讯飞股份有限公司,安徽合肥 230088;
3.大数据分析与应用安徽省重点实验室,中国科学技术大学,安徽合肥 230027)

摘要:在线个性化学习系统能够根据学生的学习历史,为学生提供个性化的学习资源,辅助学生高效学习.要提供精准的个性化诊断报告和个性化资源推荐,首先要对学生进行学业能力评估,其中一个基础性任务为得分预测.对于得分预测任务,现有的研究和方法存在如下不足:①不能充分利用大数据提升预测精度,②无法解决实际应用场景中常见的冷启动问题,③预测结果不可解释.为此提出并实现了一种基于知识图谱的自编码模型(knowledge-aware auto-encoder model,KAEM)用于学生得分预测.首先介绍了含有教育专家先验知识的一种知识图谱,称之为锚题图谱;然后KAEM采用深度学习自编码技术,将教研对锚题图谱的先验理解作为自编码器的正则化项加入模型中,有效地解决冷启动问题.此外,此类模型的预测结果还可以解释化,为实际个性化学习推荐等应用场景提供教研依据.KAEM已经在国内某在线教育系统上运行,取得了良好的效果;在大规模数据上也实验验证了KAEM的有效性.

关键词:个性化学习;知识图谱;自编码;冷启动;得分预测

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2019.01.004

引用格式:苏喻,张丹,刘青文,等.学生得分预测:一种基于知识图谱的卷积自编码器[J].中国科学技术大学学报,2019,49(1):21-30.

SU Yu, ZHANG Dang, LIU Qingwen, et al. Student score prediction: A knowledge-aware auto-encoder model[J]. Journal of University of Science and Technology of China, 2019,49(1):21-30.

Student score prediction: A knowledge-aware auto-encoder model

SU Yu[1], ZHANG Dang[2], LIU Qingwen[2], ZHANG Yingjie[2], CHEN Yuying[3], DING Chris[1]

(1. *Department of Computer Science and Technology, Anhui University, Hefei* 230039, *China*;
2. *iFLYTEK Co., Ltd., Hefei* 230088, *China*;
3. *Anhui Province Key Lab. of Big Data Analysis and Application, University of Science and Technology of China, Hefei* 230027, *China*)

Abstract: To reduce study burden and boost efficiency, online education systems offer personalized learning experience for students. In such systems, ability assessment is a fundamental task as reflected by a basic task, named score prediction. The main drawbacks of existing prediction methods are: ① Inability unable to fully exploit the potential of big data, ② cold start problem, ③ lack of reasonable explanations. A novel knowledge-aware auto-encoder model (KAEM) is proposed to address these issues. Specifically, an exercise-knowledge-graph with education experts' prior knowledge is introduced. Then students' performance is modeled using auto-encoders with the combination of information in knowledge graph as regularization item. By encoding and integrating the experts' prior knowledge, KAME can improve both

收稿日期:2018-01-09;**修回日期**:2018-06-30

基金项目:国家自然科学基金(61672483和61572030),国家基础研究发展(973计划)(2015CB351705)

作者简介:苏喻(通讯作者),男,1984年生,博士生.研究方向:自然语言处理、数据挖掘.E-mail:yusu@iflytek.com

prediction accuracy and model robustness and deal with the cold start problem well. Furthermore, reasonable explanations for recommendations can be generated using this model. KMAE has been applied to a famous online education system. Extensive experiments on large-scale real data clearly demonstrate its effectiveness.

Key words: personalized learning, knowledge graph, auto-encoder, cold start problem, score prediction

0 引言

随着互联网和人工智能的发展,教育领域中的个性化学习系统也得到了长足的发展[1]. 个性化学习系统通过大数据和人工智能算法,提供单个学生或者班级的个性化诊断报告,便于学生、老师和家长全面了解个体的学习状况;另外,个性化学习还可以为学生提供精准的个性化资源推荐(如试题、学习视频等),改变大班教学中千面一人的教学和学习方法,让学生的学习更加有针对性[2-4].

要提供精准的个性化诊断报告和个性化资源推荐,首先要对学生进行学业能力评估. 学生在所有测试题集合上的得分情况,即可代表该学生在该学科上的学业能力[5]. 传统教育流程中,教育专家会设计一套封闭的试题集合,比如一套专项学习的或者针对某一个学科学段(如高一数学)的题集. 若要准确地评估学生的学业能力,则需要学生在该试题集合上尽可能多地做题. 考虑到测试完所有试题会耗费学生过多的时间和精力,我们期望学生在试题集合上做少量的试题,就能够预测该学生在剩下试题上的得分,以正确评估该学生的学业能力,因此如何预测一个学生在某一试题集合上的得分,是个性化学习的基础性任务.

国内外在得分预测这个任务上已经有大量的研究,这些方法有教育学领域类模型,如认知诊断类模型中的项目反应理论(item response theory, IRT)[6],结合贝叶斯推断和隐马尔可夫模型的基于贝叶斯的知识跟踪模型(knowledge tracing)[7],机器学习类的概率矩阵分解类模型[8],基于用户的协同过滤[9],深度学习类模型等[10]. 这些方法有如下不足:①大部分模型参数较少,很难对海量的学生数据进行充分训练,导致模型预测精度不够. ②这些模型无法解决冷启动问题,在一些数据极度稀疏的情况下,失去了可用的预测精度. ③这些模型不具有可解释性,学生、老师、家长无法看到诊断和推荐结果的充分证据. 如何利用更多的学生学业数据以及如何增加更直接的试题间关系信息,以提升预测模型的预测精度、健壮性和可解释性,是本文要解决的问题.

首先,实际场景中面临的得分日志数据,是海量并且高度稀疏的(因为海量的学生在试题集合上做题,但是大部分学生只能做这个题集中的若干题),这导致许多建模方法失去了预测精度[11]. 本文设计了一种能够进行多层信息提取的模型,利用大数据模糊处理数据的缺失部分;此外,本文将教育专家对试题的先验理解,融入预测模型中,既解决了数据极度稀疏情况下模型的鲁棒性,又能增加了预测结果的可解释性.

自编码技术在图像去噪[12]及推荐领域[13]中有着非常成功的应用案例. 自编码技术可以将某个对象从原始空间中转换到一个低秩空间中,并且能够将对象再映射回原始空间,在这个数据二次映射(数据重构)过程中,原始对象的一些空缺元素会被填充上. 受此启发,本文将深度学习自编码技术应用到教育领域中学生学业能力的建模上,即对学生在试题得分的空间进行非线性转换,得到学生能力的低秩编码,再将低秩的学生能力解码,重构到试题得分空间,填补学生在未做题目的得分. 此外,教育领域中的学业知识有很多结构化方法,知识图谱就是其中之一[14],教育领域的知识图谱是将一个学科的知识间的关系用加权有向图的形式表示. 试题是学科知识最好的表现形式,因此可以用典型试题来表示对应知识,称作锚题[15]. 锚题之间的关系用前驱后继进行定义. 比如,对于锚题 A、B,若要学会 B,则必须已经学会 A,则称 A 为 B 的前驱,B 为 A 的后继. 用有向箭头代表前驱后继关系,箭头上的权重代表锚题和锚题间的难度差. 由锚题和锚题之间的关系构成的加权有向图称为锚题图谱. 基于锚题图谱的个性化学习推荐可以让学生的学习更加系统化,学习目标更清晰[16]. 可以将图谱间试题的关系,作为教育专家先验知识的载体,融入预测模型中,解决冷启动和预测结果可解释化的问题.

综上所述,本文提出并实现了一种基于知识图谱的自编码器用于学生得分预测,称为基于知识图

谱的自编码模型(knowledge-aware auto-encoder model,KAEM).具体而言,首先,本文采用自编码技术对学生答题日志进行多层信息压缩和提炼,以提升模型预测精度.此外,本文提出的 KAEM 融合了锚题图谱中的试题难度偏序,即本文的模型既是基于数据的,也是基于领域专家先验知识的,因此 KAEM 能够很好地解决数据稀疏问题,表现出良好的鲁棒性.最后,KAEM 的预测结果具有可解释性.在实际应用中,KAEM 不仅仅告诉用户(老师、学生等)精确的学生学业诊断结果,还可以告诉用户模型是基于什么信息进行诊断的,使得用户能有针对性地进行下一步的教学或学习.总之,本文提出的 KAEM 的主要贡献总结如下:

(Ⅰ)KAEM 能够有效地对学生的学业能力进行建模,在得分预测任务上有较高的精度.

(Ⅱ)该模型在数据极度稀疏下仍然能够保持良好的预测精度,有效解决了大部分应用场景下都会遇到的冷启动问题.

(Ⅲ)该模型的预测结果具有可解释的特性.该特性能够让研究者深入理解该模型的内部机理,并且为实际个性化学习推荐等应用场景提供教研依据.

1　相关工作

对于得分预测这个任务,国内外也有相关学者做了大量的研究.本节将从教育领域类模型、机器学习类模型及深度学习类模型 3 个方面,总结和介绍针对得分预测的相关研究和技术.

1.1　教育领域类模型

教育心理学家在经典测量理论(classical test theory,CTT)[17]的基础上,基于现代认知心理学、计量心理学、现代统计学和计算机科学,演进成一套较完备的认知诊断模型(cognitive diagnosis model,CDM)[18],以项目反应理论(IRT)[6]模型应用最为广泛.IRT 自从创立以来就被应用于各类教育学分析应用,通过引入猜测系数等参数,并结合大量测试结果训练这些参数,可以诊断每个学生的潜在能力.一旦获得了学生的能力参数和试题参数(如难度和区分度),则可以预测该学生在该试题上的得分.另外,基于贝叶斯的知识跟踪模型[1]能够对学生学业能力进行建模.该模型假设学生对知识点的掌握情况被表示为一个二元变量,即学生处于"已掌握该知识点"和"未掌握该知识点"两种状态之一.虽然学生的知识点掌握情况不能被直接观察,但是可以通过观察学生回答该知识点对应的练习题的正误来推测隐藏变量的概率分布.该模型可以对每个学生的知识点掌握情况进行预测,从而推断该学生在相应试题上的得分.

1.2　机器学习类模型

一些数据低秩重构类模型可以应用在这个任务上.比如矩阵分解类算法(probabilistic matrix factorization, PMF)[8],可以对大量的学生学业数据进行低秩分解,获得关于学生和试题的隐信息,利用隐信息重构学业数据以补全空缺值,作为学生对应试题的预测得分.此外,传统协同过滤算法也可以应用到教育领域的学生学业数据构建中[9].该方法在海量学生集合中找到和当前预测学生最相似的 k 个学生,以这些相似学生集合在某试题上的得分作为依据,预测该学生在该试题上的得分.

1.3　深度学习类模型

当前,深度学习技术在不同的领域中取得了丰硕的成果,比如,语音识别领域[19]、图像识别领域[20]、自然语言处理领域[21]等.得分预测中,能收集到的学生做题规模是海量的,因此这个研究任务也非常适合引入深度学习技术.近些年,也有一些学者利用深度学习技术做出了一些成果.DKT(deep knowledge tracing)[10]模型是利用深度学习中的递归神经网络(RNN)的 LSTM 模型,对学生的答题记录进行建模.该模型引入了试题知识点信息,假设该学生在某时刻做了一道试题,输入的是学生在该题对应知识点上的得分情况.一旦某位学生有相当数量的答题日志,则模型可以输出该学生试题空间中所有试题的掌握情况,即获得该学生在这些试题上的预测得分.

2　基于知识图谱的自编码模型

2.1　任务描述和定义

定义学生集合 $S=\{s_1,s_2,\cdots,s_m\}$,锚题图谱集合 $T=\{t_1,t_2,\cdots,t_n\}$,学生答题记录集 $R=\{r_{ij}\}$,其中 $0<i<m+1$,$0<j<n+1$,r_{ij} 代表学生 s_i 在试题 t_j 上的得分,对于客观题来说,得分为 0 和 1,对于主观题来说,将用该题的满分做归一化.本文将日志 R 视为一个得分矩阵,其中 r_{ij} 可能为空缺值,表示对应学生没有做相关的题目.空缺值部分占矩阵所有元素比例越高,说明该日志 R 的稀疏程度越高.对于不同的应用场景,获得的学生答题日志 R

的稀疏程度是不同的. R 越稀疏,建模的难度越高.

本文的任务是训练一个得分预测模型,该模型有如下能力:给定一个学生 s_k 在锚题图谱集合上的部分得分记录 $R_k=\{r_{kj}\},j\in X$,预测该学生在其他试题上的表现 $\hat{R}_k=\{r_{kj}\},j\in Y$,其中,$t_i\neq t_j,i\in X,j\in Y$,$X$ 代表学生 s_k 已经做过的题目下标集合,Y 代表该生没有做过的题目下标集合.

本文依托某在线学习系统,获得一定量的学生答题日志 R. 此外,教育专家会对锚题图谱的偏序关系 G 进行先验估计. 我们将答题日志 R 和图谱偏序关系 G 作为输入信息来训练模型,对学生学业能力进行建模.

2.2 KAEM 整体模型框架

图 1 是 KAEM 的整体模型框架. 模型可以分为两个部分. 第一个部分为自编码过程. 学生原始的得分日志矩阵 $\boldsymbol{R}$(m 个学生,n 道试题)通过一系列的编码转换,会得到一个 $m\times n$ 维的重构矩阵 $\hat{\boldsymbol{R}}$,本文要求重构矩阵 $\hat{\boldsymbol{R}}$ 能够和原始得分日志矩阵 $\boldsymbol{R}$ 更相近,具体细节见 2.3 节;第二个部分为图谱关系约束,本文希望重构矩阵 $\hat{\boldsymbol{R}}$ 的一些特性,能够符合教育专家的先验知识,具体细节见 2.4 节. 这两部分会作为一个整体联合建模.

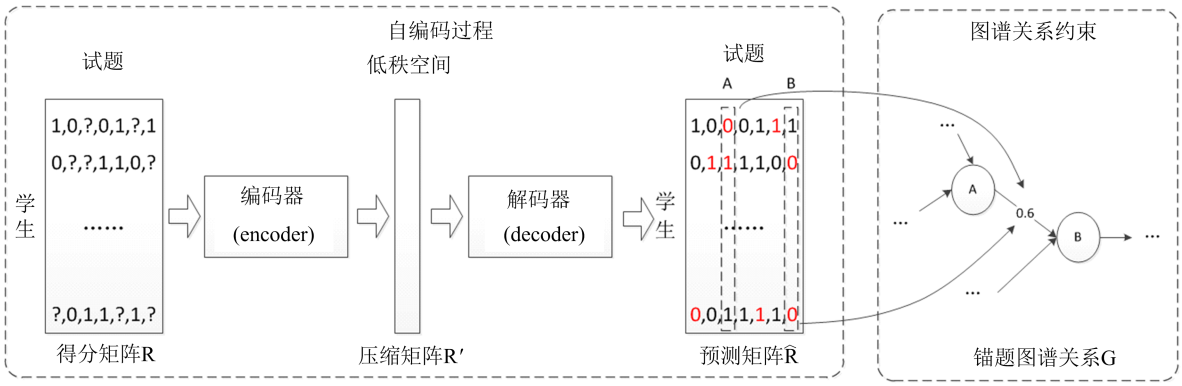

图 1 KAEM 整体模型框架

Fig. 1 The framework of KAEM

一旦获得了某个应用场景下的具体模型,将一个待预测学生的得分日志 $\boldsymbol{R}_k$ 作为模型的输入,就可以预测该学生在图谱集合 T 中的所有题目的得分率 $\hat{\boldsymbol{R}}_k$.

2.3 自编码过程

如图 1 所示,学生原始的得分日志矩阵 $\boldsymbol{R}$,会先通过一个 encoder(编码器),将每个学生的特征向量从 n 维降为 n' 维,其中 $n'\ll n$. 对于学生日志 $\boldsymbol{R}$ 来说,即 $\boldsymbol{R}'=\mathrm{EC}(\boldsymbol{R})$. 意味着每个学生由具象的答题结果($n$ 维),转换为由更加抽象的 n' 维特征(有可能是隐含了学生更深层的抽象学业能力)表示,降维编码过程使得模型学习到了题与题之间在得分上的关系. 之后,模型再通过 decoder(解码器),将每个学生压缩后的特征向量从 n' 维恢复成 n 维. 对于矩阵 $\boldsymbol{R}'$ 来说,即 $\hat{\boldsymbol{R}}=\mathrm{DC}(\boldsymbol{R}')$. 在解码过程中,原始学生日志中的缺失值能够补全,因为模型通过对大量学生日志的信息提取,学习到了学生能力和题目得分之间的关系. 整个二次转换过程称为学生得分信息的自编码过程. 最后,自编码模型的目标是比较原始 $\boldsymbol{R}$ 和重构矩阵 $\hat{\boldsymbol{R}}$ 之间的差异. 差异越小,则意味着自编码过程的信息损失越小,模型性能越强. 目标函数可以形式化的表示为

$$\min \boldsymbol{I}\parallel \boldsymbol{R}-\hat{\boldsymbol{R}}\parallel_F^2.$$

其中,$\parallel\cdot\parallel_F^2$ 为 Frobenius norm,$\boldsymbol{I}$ 是一个 $m\times n$ 维的指示矩阵,若 $\boldsymbol{R}$ 中的 r_{ij} 是非空缺值,则对应的 I_{ij} 为 1,否则为 0.

2.3.1 数据初始化

如果 r_{ij} 为空缺值,本文以 r_{ij} 所在位置的行列平均值,作为其初始化值,填充在 $\boldsymbol{R}$ 中,即

$$r_{ij}=\alpha\times\mathrm{mean}(r_{i,})+\beta\times\mathrm{mean}(r_{,j}).$$

式中,$r_{i,}$ 代表矩阵 $\boldsymbol{R}$ 中第 i 行所有非空缺值元素的集合,其均值物理意义为第 i 个学生的历史平均水平;$r_{,j}$ 代表矩阵 $\boldsymbol{R}$ 中第 j 列所有非空缺值元素的集合,其均值物理意义为第 j 道题的平均得分率(试题难度). α 和 β 为先验权重,其和为 1. 一般 $\alpha<\beta$,因为试题平均得分率的统计量比较稳定.

值得注意的是,也可以对空缺值进行随机初始化,但是这会导致模型训练收敛变慢.

2.3.2 自编码策略

本文采取了两种不同的自编码策略,这两种自

编码策略在图像、文本处理中都有着不错的应用效果,这是第一次把该策略引入到教育领域的得分预测任务中.

一种策略是栈式自编码(stacked auto-encoder, SAE)[22],对于学生 i 的得分日志 $R_i=\{r_{ij}\}$,$0<j<n+1$(学生在每道题上都有原始得分或者初始化的得分),本策略对 R_i 进行多层编码,即

$$z(l)=W(l)a(l-1)+b(l),$$
$$a(l)=\mathrm{ReLu}(z(l)).$$

式中,$W(l)$ 和 $b(l)$ 是第 l 层的编码器的参数,$z(l)$ 是第 l 层编码后的结果,$a(l)$ 是 $z(l)$ 经过 ReLu(·)激活函数后的结果,也是第 l 层的最终输出. 当 $l=1$ 时,$a(l-1)=R_i$,最后一层编码器,不需要再使用 ReLu(·)激活函数. 编码器的层数视锚题的个数而定,一般锚题的个数越多,编码器的层数设置越多.

接下来,对编码后的低秩向量进行解码,假设编码器有 p 层,那么

$$z(p+l)=W(p+l)a(p+l-1)+b(p+l),$$
$$a(p+l)=\mathrm{ReLu}(z(p+l)).$$

对于解码器的最后一层,本文使用 Tanh(·)激活函数(ReLu(·)函数在 0 点处没有梯度,导致收敛不好). 由于 Tanh(·)激活函数不能保证最终解码出的 $\hat{\boldsymbol{R}}_i$ 中的元素是大于 0 的,不符合分数非负的物理意义,因此还会加上非负的约束,即

$$\min\| \hat{\boldsymbol{R}}_- \|_F^2.$$

式中,$\hat{\boldsymbol{R}}_-$ 代表 $\hat{\boldsymbol{R}}$ 中元素为负数的部分.

栈式自编码结构简单、易于实现. 当模型层数较深时,会有模型参数较多、容易过拟合等缺陷. 为此本框架将采用图像处理中常见的卷积自编码策略(convolutional auto-encoder, CAE). 由于卷积自编码中每层的卷积核是复用的,因此模型参数较少,不容易过拟合,在实际应用中更受青睐.

卷积自编码[23]同样地,对学生 i 的得分日志 R_i 进行多层的卷积编码,即

$$z(l)=\mathrm{Cov}(a(l-1);W(l),b(l)),$$
$$h(l)=\mathrm{ReLu}(z(l)),$$
$$a(l)=\mathrm{Pool}(h(l)).$$

式中,$W(l)$ 和 $b(l)$ 是第 l 层的编码器的卷积核参数,Cov(·)为卷积操作,$z(l)$ 是第 l 层编码后的结果,$h(l)$ 是 $z(l)$ 经过 ReLu(·)激活函数后的结果,$a(l)$ 是 $h(l)$ 经过池化后的结果,也是 l 层的最后输出. 本文使用的池化方法为 max Pooling. 当 $l=1$ 时,$a(l-1)=R_i$.

假设编码器有 p 层,卷积自编码的解码过程为

$$z(p+l)=\mathrm{DCov}(a(p+l-1);$$
$$W(p+l),b(p+l))a(p+l)=\mathrm{ReLu}(z(p+l)).$$

式中,DCov(·)操作为反卷积操作. 对于解码器的最后一层,本策略使用的是 Tanh(·)激活函数,因此也要做非负的约束.

本文把对学生日志 $\boldsymbol{R}$ 做自编码的过程,简写成 $\hat{\boldsymbol{R}}=\mathrm{AE}(\boldsymbol{R})$,其中 AE(·)函数代表自编码整体函数,于是上一节的公式,可以改写为

$$\min\boldsymbol{I}\| \boldsymbol{R}-\mathrm{AE}(R) \| \|_F^2.$$

自编码技术增加了模型的复杂度,但能够充分利用大数据的优势,提升预测精度.

2.4　图谱关系约束

本小节将讨论如何将偏序关系引入如上的深度学习框架中.

大多数实际应用场景中,针对一个固定的锚题图谱集合 T,每一个学生只会在少量的题目上有答题记录,这样导致答题日志矩阵 $\boldsymbol{R}$ 有很多的缺失值,即这个矩阵十分稀疏,因此仅仅利用 $\boldsymbol{R}$ 中的信息,很难对学生学业能力进行泛化建模. 由于锚题图谱的偏序关系 G 中,蕴含了大量的教研先验,因此可以将偏序关系 G 作为预测模型一个有效的正则化项,对自编码后得出的 $\hat{\boldsymbol{R}}$ 进行约束.

具体的,锚题图谱的偏序关系 G 是如下元素的集合:

$$P_{ij}=D_i-D_j.$$

式中,D_i 代表第 i 道题的先验难度. 由于试题的难度与试题本身以及测试学生相关,所以教育专家很难预估这个先验难度[24-25]. P_{ij} 代表第 i 题和第 j 题的相对难度差. 教育专家是能够对 P_{ij} 进行先验估计的,因为图谱上部分相邻题都有明确的前驱后继,或者难度偏序关系,教育专家可以对这种相对难度差,进行大致的先验估计. 在本文中,教育专家对 P_{ij} 分为 7 档,从 0.3~0.9,步长为 0.1;0.3 以下的偏序关系,教育专家一般认为是非稳定的,不能作为锚题图谱的偏序关系 G 的元素. 根据重构后得出的 $\hat{\boldsymbol{R}}$,获得了试题 A 和 B 的预测难度,并且期望它们的难度差能够符合锚题图谱中的先验偏序关系.

自编码重构后得出的 $\hat{\boldsymbol{R}}$,试题 i 的难度定义为 $\hat{D}_i$,因此 $\hat{P}_{ij}=\hat{D}_i-\hat{D}_j$,为从 $\hat{\boldsymbol{R}}$ 中得出的偏序关系. 模型希望 $\hat{P}_{ij}$ 能和先验的锚题图谱的偏序关系 G 中的 P_{ij} 更加接近,即

$$\min \frac{1}{N}\sum(\hat{P}_{ij}-P_{ij})^2, P_{ij}\in G, N=|G|.$$

这个限制会促使自编码重构出来的 $\hat{\boldsymbol{R}}$，更加符合真实的教研规律，即对同一个学生来说，预测其做对难题的概率与预测其做对简单题的概率之差，要符合教育专家对这两道题偏序关系的先验理解.

融合先验的偏序关系一方面可以将原本独立的学生答题数据关联起来. 在数据极度稀疏的情况下，两个学生群体所做的试题往往没有交叠部分，而试题的偏序关系可以将这两部分答题数据联系起来统一建模，保证了一定的预测精度，解决常见的冷启动问题；另一方面会约束预测结果和先验的偏序关系保持一致，结合锚题图谱的结构后，可以对预测结果进行解释化，增加实际的应用价值.

将上述几个约束合并，即可得出 KAEM 整体的损失函数，即

$$\min_{\theta} I \parallel \boldsymbol{R}-\mathrm{AE}(\boldsymbol{R})\parallel_F^2+\lambda\frac{1}{N}\sum(\hat{P}_{ij}-P_{ij})^2,$$
$$P_{ij}\in G, N=|G|.$$

式中，θ 为自编码器的模型参数，λ 为模型预设的超参数，一般设为 0.2 左右.

3 实验

本节将从得分预测的效果、冷启动情况下的效果对比、得分预测的可解释化 3 个方面，来说明本文所提 KAEM 的有效性和实用性.

3.1 数据集介绍

本文所使用的实验数据，来自于国内某知名在线学习系统[26]. 该系统采取基于锚题图谱的推荐方式，为了便于检验各种方法的预测效果，本实验取答题日志中最稠密的，即缺失值最少的部分. 该部分为高中数学解析几何相关的锚题图谱片段，共有 110 个锚题节点. 数据集统计指标细节如表 1 所示.

表 1 实验数据集介绍

Tab. 1 Experimental dataset summary of KAEM

统计指标	值
日志 $\boldsymbol{R}$ 中的元素个数	2 047 744
日志 $\boldsymbol{R}$ 中所涉及的题目数（即锚题图谱节点个数）	110
日志 $\boldsymbol{R}$ 中所涉及的学生数	19 432
平均学生答题数	105.38
学生答题平均正确率	0.61
日志 $\boldsymbol{R}$ 中缺失值占比	4.2%
偏序关系 G 的元素个数	138

3.2 对比方法介绍

为了检验本文 KAEM 的效果，首先采用 KAEM 的两种不同自编码策略的模型：基于知识图谱的栈式自编码方案（knowledge-aware stacked auto-encoder，KSAE）和基于知识图谱的卷积自编码方案（knowledge-aware convolutional auto-encoder，KCAE）. 实验对比去除教育专家先验经验的卷积自编码模型，一方面用来验证用深度学习中自编码方案做学生得分预测是否合理；另一方面对比说明，在数据极度稀疏情况下融入教育专家先验经验确实会增加模型的鲁棒性. 此外，实验还将对比若干种流行和先进的学生建模及得分预测模型，有传统机器学习类模型协同过滤方法和 PMF，有教育类模型 IRT 和 BKT（Bayesian knowledge tracing）以及深度学习类模型 DKT. 特别需要注意的是 BKT 和 DKT 模型需要对试题进行知识点（考点）的标注. 对比方法的介绍如下：

（Ⅰ）IRT[6]：项目反应理论是一种认知诊断模型，通过对学生在试题上的得分建模，得到学生以及试题的相关参数，可用于得分预测任务.

（Ⅱ）BKT[7]：基于贝叶斯的知识跟踪模型，根据学生在练习过程中的答题表现，跟踪其知识掌握情况.

（Ⅲ）协同过滤方法（collaborative filtering，CF）[9]：根据答题日志 $\boldsymbol{R}$，以学生的答题得分作为特征，计算学生间的相似度（基于用户的协同过滤）. 用待预测学生的相似学生集合的对应的试题得分均值，作为该学生该题的预测得分.

（Ⅳ）PMF[8]：概率矩阵分解方法，会将答题日志 $\boldsymbol{R}$ 进行低秩分解，再将分解后的子矩阵合并后，得到重构后的矩阵 $\hat{\boldsymbol{R}}$. 以重构后 $\hat{\boldsymbol{R}}$ 的元素 $\hat{r}_{ij}$，作为学生 i 在第 j 题的预测得分.

（Ⅴ）DKT[10]：深度知识跟踪模型利用深度学习中的一种递归神经网络 LSTM 模型，对学生的答题记录进行建模.

（Ⅵ）卷积自编码（CAE）[23]：用卷积自编码策略，对得分预测矩阵 $\boldsymbol{R}$ 进行重构. 以重构后 $\hat{\boldsymbol{R}}$ 的元素 $\hat{r}_{ij}$，作为学生 i 在第 j 题的预测得分. 该模型不会融入锚题图谱偏序关系的约束.

3.3 得分预测效果对比

为了检验得分预测的准确性，本实验将得分日志 $\boldsymbol{R}$ 分为训练集 R_s 和测试集 R_t. 具体的会随机地将每个学生的得分日志中抽取比例为 C 的日志作为测试集 R_t，其余部分作为训练集 $\boldsymbol{R}_s$. 比如当

取 C 为 0.8 时,即表示会用每个学生得分日志 $\boldsymbol{R}$ 中 20%的数据做模型训练,去预测该学生在其余 80% 的试题得分.

本实验用如下两个指标来衡量不同方法的预测精度[27-28]. 一种是从回归的角度来衡量,因为所有方法预测学生的得分是一个 0 到 1 之间的浮点数,所以本实验用均方根误差(RMSE)来衡量 R_t 和预测出的 $\hat{R}_t$ 之间的差异. RMSE 值越小,说明该方法预测的精度越高. 另一种是从分类的角度来衡量. 为了更直观地理解得分预测的效果,还可以用阈值 0.5 将得分率划分成 0 或 1,即错误和正确这两种情况. 这样就可以用预测准确率(ACC)来直观地比较 R_t 和预测出的 $\hat{R}_t$ 之间的差异. ACC 值越高,说明该方法预测的精度越高.

本实验取 C 为 0.8,用 3.2 节所提到的各种方法,得到如图 2 和图 3 的对比实验结果.

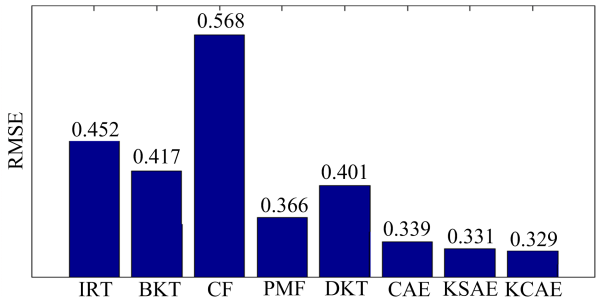

图 2　各种方法在 RMSE 指标上的结果

Fig. 2　Overall results on RMSE

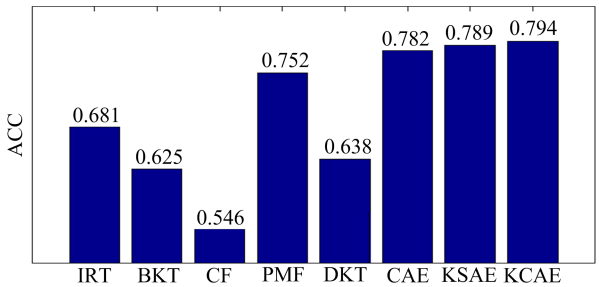

图 3　各种方法在 ACC 指标上的结果

Fig. 3　Overall results on ACC

从图 2 和图 3 中可看出,采用深度学习自编码技术的 3 个模型,CAE,KAE 和 KCAE,无论是在 RMSE 还是 ACC 两个指标上,都优于其他模型(PMF 在相对稠密数据上,表现也较好). 这说明了自编码技术可以利用海量的学生答题数据,从中提取学生能力和试题得分的关系,有助于提升得分预测的精度. 另外,增加锚题图谱偏序关系的两个模型 KAE 和 KCAE,在性能上略优. 这说明了教育专家的先验知识会对试题间的偏序关系进行有效补充,而这部分信息有可能从学生日志中没有充分体现.

3.4　冷启动情况下的效果对比

在一些推荐应用场景中,会遇到冷启动问题. 特别是教育类个性化学习场景中,当学习系统刚上线使用时,大多数学生只会做过锚题图谱中的极少部分试题,即学生日志 $\boldsymbol{R}$(m 维远高于 n 维)的稀疏比例非常高,甚至可能高过 90%. 我们将稀疏参数 C 从 0.1 逐渐增大到 0.99,来检验不同方法在面对训练数据逐渐减少的情况下 RMSE 和 ACC 的表现,如图 4 和 5 所示. 其中较粗的曲线为 KAEM 类模型的变化趋势.

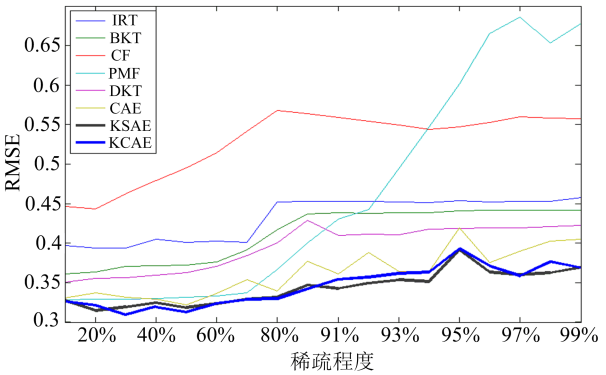

图 4　随着稀疏程度增加,RMSE 的变化趋势

Fig. 4　The curves with different sparse degree on RMSE

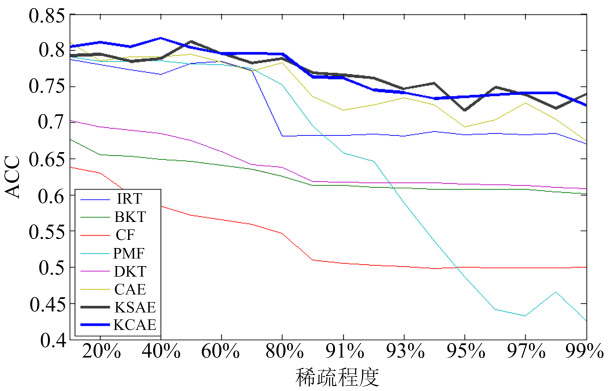

图 5　随着稀疏程度增加,ACC 的变化趋势

Fig. 5　The curves with different sparse degree on ACC

由图 4 和图 5 可知:①随着数据稀疏程度的增加,所有方法的预测性能都会下降,这是因为随着数据稀疏程度的增加,各种模型获得的有效信息在逐渐减少. ②KAEM 类模型由于增加了先验的锚题图谱的偏序关系,鲁棒性得到了极大的增强. 在一些极端情况下(甚至稀疏参数 C 达到了 0.99),KAE 和 KCAE 模型仍然可以保持良好的预测性能. ③其他方法的性能下降的趋势较为明显,特别是当稀疏参

数 C 大于 0.9 后，大多数方法的性能会下降到几乎随机. 特别地，在数据相对稠密情况下表现良好的 PMF 模型，随着数据稀疏度增高，性能下降的尤其显著. 总之，相比于传统模型，增加了锚题图谱偏序关系的 KAEM 类模型，更擅长处理冷启动情况.

3.5 得分预测的可解释化

由于 KAEM 类模型增加了先验的锚题图谱的偏序关系 G，因此可以使得该类模型的得分预测结果具有可解释性. 一方面，可以直观地说明为什么该模型在数据极度稀疏下，仍然能够保持良好的预测性能；另一方面，可解释性对某些应用场景来说非常重要. 比如在个性化学习中，具有可解释性的学生诊断报告，能够指导老师或者学生进行有针对性的教学或学习.

3.5.1 可解释化对比实验

为了检验模型的可解释性是符合教育专家的认知，本文进一步设计了对比实验. 具体地，第一组采用本文提出的模型 KAE 和 KCAE，对照组采用 PMF 和 CAE 模型(其他模型在预测精度上已经与 KAEM 类模型差异较大). 在同一个测试集上，本实验获取模型的一个预测得分 $\hat{r}_{ij}$，并且让老师观察学生 i 在 j 题前驱后继题目的实际表现. 如果教育专家认为预测得分 $\hat{r}_{ij}$ 能够由该学生在其他相关题的表现中推断出来，则预测得分 $\hat{r}_{ij}$ 是符合教育专家认知的. 本实验在稀疏程度 C 为 80%的情况下，随机抽取了 50 个学生的预测结果，分别由 5 名教育专家对每个模型的预测结果进行合理性判断. 本实验以合理比例(预测结果合理的个数占预测结果总数的比例)作为验证指标，并取 5 名教育专家的合理比例的平均值作为某一模型最终的“合理比例”，结果如图 6 所示.

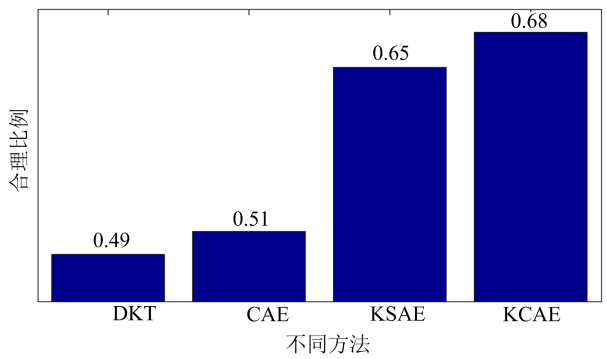

图 6 不同方法预测结果的合理比例

Fig. 6 Overall results on reasonable proportion

由图 6 可知，结合锚题图谱的偏序关系 G 的两个模型，在预测结果可解释性的合理比例上，远高于其他模型. 实际应用中，KAEM 类模型产生的预测结果，能够给老师和学生带来更多的帮助.

3.5.2 可解释化示例

图 7 为 KAEM 可解释化的一个例子. 这是某个学生在一道试题上的预测结果展示. $A\sim E$ 五个节点分别代表锚题图谱中的五道试题，节点上的数字代表某学生在这五道题上的掌握程度(即答题结果或预测结果)，其中 C 节点上为模型预测值. 节点间的有向边代表锚题图谱的偏序关系，箭头的方向是前驱到后继的方向，边上的权重为教育专家预设的难度差 P_{ij}. 比如锚题 A 为锚题 C 的前驱题，锚题 C 为锚题 A 的后继题，并且教育专家认为锚题 C 比锚题 A 的难度要高，它们的平均得分率的差值应该在 0.4 左右. $A\sim E$ 五个节点的试题题面见表 2，五道题均为选择题，此表中忽略这五题的选项.

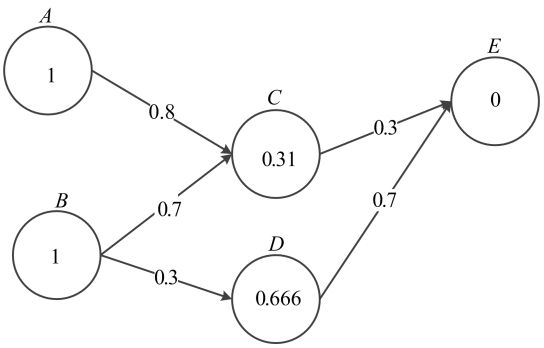

图 7 可解释化示例

Fig. 7 Visualization case of KAEM

由图 7 和表 2 可知，模型对该学生在锚题 C 预测的得分(掌握程度)为 0.31 是比较合理的. 因为，首先 C 的得分率是介于两个前驱 A、B 和一个后继 E 之间的；另外，与 C 有共同前驱(B)的 D 的得分率为 0.666，在 B 的两个后继中，C 的综合程度更高，得分率应比 D 略低. 通过日志的查询，该学生在锚题 C 的实际得分率为 0.33，也验证了模型预测具有较高的准确性.

上例说明了模型为何在数据极度稀疏的情况下，仍然有较好的表现. 因为图 7 中锚题 C 的出入度都较高，所以锚题 C 是一个关键点. 假设在数据极度稀疏的情况下，某学生仅仅有锚题 C 的实际得分，模型可以通过图谱偏序关系 G，合理地推断出该学生在锚题 A、B、E 的得分，并且通过锚题 B 和 E 的得分，推断出锚题 D 的得分. 锚题图谱的偏序关

系 G,可以将稀疏数据中的各个有值部分关联起来,并结合深度学习的自编码方法,将日志 R 中的有效信息融入预测模型中.

另外,该学生做错了锚题 E,学习系统可以根据图谱偏序关系和预测结果,推荐该学生去学习锚题 C,因为很有可能是该生在锚题 C 的相关知识没有掌握,导致其在后驱锚题 E 上失分. 同样地,老师也可以根据大量学生在图谱上的汇总掌握程度,去指导其日常的教学.

表 2　可解释化示例中的试题信息

Tab. 2　The exercises in visualization case of KAEM

节点号	题面	知识点(考察点)
A	过点(3,−2)且与椭圆 $4x^2+9y^2-36=0$ 有相同焦点的椭圆方程是	椭圆标准方程
B	已知椭圆 $\frac{x^2}{32}+\frac{y^2}{16}=1$ 内有一点 $B(2,2)$,F_1、F_2 是其左、右焦点,M 为椭圆上的动点,则 $\lvert\vec{MF_1}\rvert+\lvert\vec{MB}\rvert$ 的最小值为	椭圆的定义
C	已知椭圆 $\frac{x^2}{4}+y^2=1$ 的左、右焦点分别为 F_1,F_2,点 P 在椭圆上,当$\triangle F_1PF_2$ 的面积为 1 时,$\vec{PF_1}\cdot\vec{PF_2}=$	椭圆的焦点三角形
D	已知椭圆的长轴是 8,离心率是 $\frac{3}{4}$,此椭圆的标准方程为	椭圆的离心率
E	过椭圆 $\frac{x^2}{16}+\frac{y^2}{12}=1$ 的左顶点 A 作斜率为 $k(k\neq0)$的直线 l 交椭圆于点 C,交 y 轴于点 D,P 为 AC 中点,定点 Q 满足:对于任意的 $k(k\neq0)$都有 $OP\perp DQ$,则 Q 点的坐标为	椭圆与直线的位置关系

5　结论

本文针对得分预测问题,提出了一种融合锚题图谱偏序关系的得分预测模型,称为基于知识图谱的自编码模型(KAEM). 此类模型采用了深度学习自编码技术,在预测精度上超过了现有教育领域类和传统机器学习类等模型. 另外,本文利用了教育领域下试题间特有的难度偏序关系,将教研对锚题图谱的先验理解加入到模型中,能够有效地解决大多数应用场景下会遇到的数据极度稀疏和冷启动问题. 本文的 KAEM 除了具有较高预测精度和鲁棒性外,还可以解释化,让研究者能够深入理解该模型的内部机理,为实际个性化学习推荐等应用场景提供教研依据.

在未来的工作中,还可以在如下几点对模型框架进行优化和改进:①融合更多的试题特征. 本文当前只是融合了锚题图谱的偏序关系,下一步还可以结合试题的知识点标注,将试题知识点间的关联性,甚至试题的题面文本[29]等,融入模型中;②融合更多的学生特征. 学生的在线答题时长,学生的先验学习水平等,都可以作为模型的输入特征,增加模型的鲁棒性;③多学科建模. 当前的模型仅仅是对学生的单学科进行建模. 学生的深层抽象能力应该会影响学生在不同学科上的表现. 后续可以对同一个学生群体的不同学科的答题表现进行统一建模,并结合教育专家的教学经验,从学生的学业大数据中挖掘出更加深刻的学生综合建模信息. 此外,还可以尝试将该模型用于其他的应用场景中. 比如在电影推荐系统刚上线时,用户对电影的打分记录比较稀疏,我们可以利用电影票房等电影间的先验偏序关系作为约束,利用 KAEM 框架建模和预测用户对电影的打分情况.

参考文献(References)

[1] HONG C M, CHEN C M, CHANG M H, et al. Intelligent web-based tutoring system with personalized learning path guidance [C]// IEEE International Conference on Advanced Learning Technologies. Niigata, Japan: IEEE, 2007:787-814.

[2] 黄振亚, 苏喻, 吴润泽,等. 一种面向教育评估的智能教育辅助平台[J]. 中国科学技术大学学报, 2015, 45(10): 846-854.
HUANG Zhenya, SU Yu, WU Runze, et al. An intelligent tutoring platform for educational assessment [J]. Journal of University of Science and Technology of China, 2015, 45(10): 846-854.

[3] ROMERO C, VENTURA S. Educational data mining: a review of the state of the art [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2010, 40(6): 601-618.

[4] 朱天宇, 黄振亚, 陈恩红,等. 基于认知诊断的个性化

试题推荐方法[J]. 计算机学报，2017(1)：176-191.
[5] KISOR Y. The state of educational data mining in 2009：A review and future visions[J]. Computer Communications，2009，6(2)：82-87.
[6] RASCH G. On general laws and the meaning of measurement in psychology[C]// Proceedings of the 4th Berkeley Symposium on Mathematical Statistics. Berkeley，USA：IEEE，1961：321-333.
[7] 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.
[8] SALAKHUTDINOV R，MNIH A. Probabilistic matrix factorization[C]//Advances in Neural Information Processing Systems. Vancourver，Canada：Curran Associates Inc.，2007：1257-1264.
[9] TÖSCHER A，JAHRER M. Collaborative filtering applied to educational data mining[J]. Journal of Machine Learning Research，2010：1-11.
[10] PIECH C，BASSEN J，HUANG J，et al. Deep knowledge tracing[C]// Advances in Neural Information Processing Systems. 2015：arXiv：1506.05908.
[11] BERGNER Y，DROSCHLER S，KORTEMEYER G，et al. Model-based collaborative filtering analysis of student response data：Machine-learning item response theory[C]// 5th International Conference on Educational Data Mining. Chania，Greece：ERIC Press，2012：1-8.
[12] XIE J Y，XU L L，CHEN E H. Image denoising and inpainting with deep neural networks[C]// International Conference on Neural Information Processing System. Lake Tahoe，USA：Curran Associates Inc.，2012：341-349.
[13] WANG H，WANG N，YEUNG D Y. Collaborative deep learning for recommender systems[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney，Australia：ACM，2015：1235-1244.
[14] NOVAK J D. Learning，Creating，and Using Knowledge：Concept Maps as Facilitative Tools in Schools and Corporations[M]. Taylor & Francis Group，3ed，2010.
[15] ZAHORIAN S A，LAKDAWALA V K，GONZALEZ O R，et al. Question model for intelligent questioning systems in engineering education[C]// Frontiers in Education Conference. Reno，USA：IEEE Computer Society，2001：T2B(7-12).
[16] 方圆媛. 翻转课堂在线支持环境研究——以可汗学院在线平台为例[J]. 远程教育杂志，2014(6)：41-48.
[17] DEVELLIS RF. Classical test theory[J]. Medical Care，2006，44(3)：50-59.
[18] RUPP A A，TEMPLIN J，HENSON R A. Diagnostic measurement：Theory，methods，and applications[A]. Methodology in the Social Sciences[M]. New York：Guiford Press，2010：78-79.
[19] GRAVES A，MOHAMED A R，HINTON G. Speech recognition with deep recurrent neural networks[C]// IEEE International Conference on Acoustics，Speech and Signal Processing. Vancouver，Canada：IEEE，2013：6645-6649.
[20] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM，2017，60(6)：84-90.
[21] MIKOLOV T，SUTSKEVER I，CHEN K，et al. Distributed representations of words and phrases and their compositionality[C]// International Conference on Neural Information Processing Systems. Lake Tahoe，USA：Curran Associates Inc.，2013：3111-3119.
[22] BENGIO Y，LAMBLIN P，POPOVICI D，et al. Greedy layer-wise training of deep networks[C]// International Conference on Neural Information Processing Systems. British Columbia：MIT Press，2006：153-160.
[23] MASCI J，MEIER U，CIREŞAN D，et al. Stacked convolutional auto-encoders for hierarchical feature extraction[C]// International Conference on Artificial Neural Networks. Espoo，Finland：Springer-Verlag，2011：52-59.
[24] FUCHS L S，FUCHS D. Effects of expert system consultation within curriculum-based measurement，using a reading maze task[J]. Exceptional Children，1992，58(5)：436-450.
[25] HUANG Z Y，LIU Q，CHEN E H，et al. Question difficulty prediction for READING problems in standard tests[C]// The 31st AAAI Conference on Artificial Intelligence. San Francisco，USA：AAAI Press，2017：1352-1359.
[26] 智学网：http://www.zhixue.com/.
[27] FOGARTY J，BAKER R S，HUDSON S E. Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction[C]// Proceedings of Graphics Interface. British Columbia，Canada：DBLP，2005：129-136.
[28] WU R Z，LIU Q，LIU Y P，et al. Cognitive modelling for predicting examinee performance[C]// International Conference on Artificial Intelligence. Buenos Aires，Argentina：AAAI Press，2015：1017-1024.
[29] DIBELLO L V，ROUSSOS L A，STOUT W. 31A Review of Cognitively Diagnostic Assessment and a Summary of Psychometric Models[J]. Handbook of Statistics，2006，26(6)：979-1030.