基于协同过滤和认知诊断的试题推荐方法

齐 斌 邹红霞 王 宇 李冀兴
(航天工程大学航天信息学院 北京 101416)

摘 要 智能教育中,试题推荐方法是数据挖掘在教育测量领域的新运用,是自适应测试的智能化和个性化程度的重要体现,目前主流的试题推荐方法有两类,分别是协同过滤试题推荐方法和认知诊断试题推荐方法,前者忽略了独立个体的知识属性,后者缺乏对种群的共性评估。针对上述问题,为提高试题推荐的精确度和效率,综合考虑独立被试者的知识属性和类环境群体的知识共性,文中提出了基于协同过滤和认知诊断的试题推荐方法。首先,设计了基于多级属性评分的认知诊断模型,并利用该模型对被试者的答题情况进行建模;然后,将被试者的知识属性掌握模式用于概率矩阵分解,预测被试者的潜在答题情况;最后,根据信息量指标向被试者动态地推荐合适的试题。试题推荐方法综合考虑了个体的个性特征和群体的共性特征,提高了解释性和可靠性。实验结果表明,相比单协同过滤试题推荐算法和认知诊断选题策略,所提方法的测试效率分别提升了 20.35%和 2.5%。

关键词 协同过滤,认知诊断,试题推荐,认知诊断模型,信息量,数据挖掘
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/jsjkx.180901827

Questions Recommendation Based on Collaborative Filtering and Cognitive Diagnosis

QI Bin ZOU Hong-xia WANG Yu LI Ji-xing
(School of Space Information, Space Engineering University, Beijing 101416, China)

Abstract The question recommendation method isthe new application of data mining on the Education Measurement, which is an important performance of the intelligence and personalization in the intelligent education, particularly. At present, there are two types of mainstream test recommendation methods, including the question recommendation based on collaborative filtering and the question recommendation based on cognitive diagnosis. However, the former ignores the knowledge attribute of independent individuals, the latter is lack of the common evaluation. In order to improve the accuracy and efficiency of the question recommendation, comprehensive considering the knowledge attributes of the independent testing subjectand the knowledge commonality of the environment-like groups, this paper proposed a testing recommendation method based on collaborative filtering and cognitive diagnosis. Firstly, the proposed method designs a cognitive diagnosis model based on multi-level attributes scoring, which is used to model the subject's answer. Then, the subject's knowledge attribute is used for probabilistic matrix factorization to predict the potential answers. Finally, the appropriate questions are recommended to the subjects according to the information value. The testing recommendation comprehensively improves the interpretability and reliability that the experiment shows the method improves the efficacy by 20.35% and 2.5% respectively compared with collaborative filtering and cognitive diagnosis.

Keywords Collaborative filtering, Cognitive diagnosis, Questions recommendation, Cognitive diagnosis model, Information value, Data mining

1 引言

大数据时代促进了智能教育的发展,准确、高效地量化学生真实的能力水平是定制化教育的基础和重点研究方向。然而,由于试题的资源数量庞大,如何在限定的时间内抽取合适的试题并准确、全面地测试学生的知识水平,是智能教育的重要研究领域,即试题推荐方法。

试题推荐[1]是数据挖掘在教育测量领域的新运用,在教育测量领域中,尤其是计算机自适应测试[2]中是一个非常重要的研究方向,在计算机自适应测试中通常将试题推荐称为"选题策略"[3]。限于认知诊断模型的发展,基于认知诊断理论的试题推荐还有很大的改进空间,由于试题推荐同推荐系统存在一定的相似之处,因此将认知诊断和协同过滤相结合对试题推荐进行改进以提高测试的精度和准确性,是数据科

到稿日期:2018-09-30 返修日期:2018-12-13 本文受国家 863 计划项目(2015AAxxx2078),省部级科技创新工程(ZYX14030011)资助。
齐 斌(1994-),男,硕士,主要研究方向为智能教育、网络空间安全;邹红霞(1968-),女,硕士,副教授,主要研究方向为数据挖掘,E-mail:xiahongzou@sina.com(通信作者);王 宇(1971-),男,博士,教授,CCF 会员,主要研究方向为保密技术;李冀兴(1993-),男,硕士,主要研究方向为网络空间安全。

学和认知科学的一次交叉运用，目前在此方面的研究较少。

近年来，部分学者借鉴推荐系统中的协同过滤算法[4]解决了学生试题的得分预测问题，从而进行试题推荐和精准组卷。在各类推荐算法中，基于模型的协同过滤方法在可扩展性方面表现出了较好的性能，如奇异值分解、非负矩阵分解及概率矩阵分解(Probabilistic Matrix Factorization，PMF)[5]等模型，在一定程度上也改善了数据稀疏性的问题。因 PMF 模型具有易解释和易操作等良好特性，在推荐系统中的应用较为广泛，文献[6]表明矩阵分解已经可以取得较好的推荐精度，并不断有学者提出 PMF 在不同场景的改进方法，Chen 等[7]在 PMF 模型的基础上建立多属性相关性，基于内容和用户将项目相关性整合到 PMF 模型中，提高了推荐质量。但在试题推荐领域，被试者的知识属性维度和试题的多重维度相比推荐系统的用户和内容而言更加客观、具体，对推荐的准确度和效率的要求也更高，因此不仅需要重视相似群体的共性特征，还需要关注被试个体的知识属性掌握模式，而现有的 PMF 及其改进模型分解得到的潜在向量解释性弱，推荐效果存在局限性，难以满足智能教育中精准的试题推荐[8]。

随着心理测量学的快速发展，基于认知诊断理论(Cognitive Diagnosis，CD)的计算机自适应测试[9]成为了智能教育领域的新思路，系统可根据被试者当前的作答情况自适应地动态调整下一道试题，实时反馈被试者的诊断信息，因此被广泛用于评估考生的能力水平。自适应测试中的试题推荐算法和认知诊断模型直接反映了测试的智能化和个性化程度，因此部分学者[10]试图将协同过滤的试题推荐算法和认知诊断相结合以提高试题推荐的准确度。朱天宇等[11]提出了一种基于学生知识点掌握程度的协同过滤试题推荐方法，利用答题情况和知识点的关联对学生实际知识水平进行建模，并将掌握水平用于概率矩阵以分解预测作答状况，从而进行相应的试题推荐，一定程度上提高了解释性和可靠性。

但由于当前学者[10-11]采用的认知诊断 DINA 模型是典型的 0-1 得分模型[12]，不满足多维、多属性的复杂试题处理要求，因此在工程实践的应用中普适性较低。而目前大多数的试题推荐方法为静态推荐，无法通过测试准确地评估被试者的真实知识水平，减弱了智能教育的诊断评估意义。

为提高试题推荐的准确性、可解释性和普适性，本文在自适应测试环境下提出了基于协同过滤和认知诊断的试题推荐方法，延伸了 DINA 模型在多级属性评分上的应用规则，拓展了试题推荐的应用场景。

2 认知诊断模型设计

实际测试中，人们因认知结构倾向和涉及的认知属性成分不同，在实际测试中往往需要评估属性级别。因此，为符合实际的复杂测试要求与测试环境，提高诊断评价的准确性和计算效率，本文将 P-DINA 模型[13]拓展为基于多级属性评分的认知诊断模型，记为 PH-DINA (Polytomous Hierarchical DINA)。

为便于读者理解后续相关模型和框架，首先定义多级属性测试题型的表达方式，因存在一道测试题可考核多个知识点的客观现象，即每一道试题至少包含一个知识属性，则试题 q 可表示为 $q_j=(q_{j1},q_{j2},\cdots,q_{jk})$。其中，$j$ 为测试题目的编号；k 为题目 j 待考核知识点的最大数目；$q_{jk}=\{0,1,2,\cdots,n\}$ 表示第 k 个测试属性具有 n 个级别，若 $q_{jk}=n\geqslant 1$ 则代表考查难度为 n 的第 k 个知识属性，反之，$q_{jk}=0$ 则代表不考查该知识属性。

2.1 PH-DINA 模型

P-DINA[14]是典型的非补偿模型，即要求被试者必须掌握待测的全部技能或知识属性 $\boldsymbol{\alpha}_i$ 才可被认定正确作答，项目所考查的技能或属性则全部被包含在待测项目 q_j 中，项目反应函数为：

$$P(Y_{ij}=t\mid\boldsymbol{\alpha}_i)=P^*(Y_{ij}=t\mid\boldsymbol{\alpha}_i)-P^*(Y_{ij}=t+1\mid\boldsymbol{\alpha}_i) \tag{1}$$

$$P^*(Y_{ij}=t\mid\boldsymbol{\alpha}_i)=(1-s_{jt})^{\eta_{ij}}\cdot g_{jt}^{1-\eta_{ij}} \tag{2}$$

$$\eta_{ij}=\prod_{K=1}^{K}\alpha_{ik}^{q_{jk}} \tag{3}$$

其中，$P(Y_{ij}=t\mid\boldsymbol{\alpha}_i)$表示被试者 i 在项目 j 上得 t 分的概率；$P^*(Y_{ij}=t\mid\boldsymbol{\alpha}_i)$表示被试者 i 在项目 j 上得 t 分及以上的概率；$\eta_{ij}\in\{0,1\}$表示被试者在理想情况下(不考虑猜测和失误的情况)作答的结果得分；$s_j=P(Y_{ij}=1|\eta_{ij}=1)$是项目 j 的失误参数，指被试者在掌握项目 j 考核的属性下仍答错的概率；$g_j=P(Y_{ij}=1|\eta_{ij}=0)$为项目 j 的猜测参数，指被试者在尚未完全掌握项目 j 考核的属性下答对的概率。

式(2)的猜测参数和失误参数需满足要求：$s_{jt}\leqslant s_{jt+1}$，即对于需要掌握项目 j 考核属性的被试者而言，得 t 分的失误概率要小于得 $t+1$ 分的失误概率；$g_{jt}\geqslant g_{jt+1}$，即对未全部掌握项目 j 考核属性的被试者而言，猜对 t 分的概率要大于猜对 $t+1$ 分的概率，从而保证了被试者答对的概率恒不为负。

式(3)中，K 表示测试属性的数量，$\boldsymbol{\alpha}_i=(\alpha_{i1},\alpha_{i2},\cdots,\alpha_{ik}\mid\alpha_{ik}=\{0,1\})$表示被试者 i 对各个属性的掌握情况，$\alpha_{ik}=1$ 说明被试者 i 掌握了 k 属性，反之为 0 表示未掌握；$q_j=(q_{j1},q_{j2},\cdots,q_{jk}\mid q_{jk}=\{0,1\})$表示项目 j 对各个属性的考查，$q_{jk}=1$ 说明项目 j 考查了属性 k，反之为 0 表示未考查。

为适用于复杂知识结构和实际多属性试题，本文引入属性层级(Hierarchical)的概念。属性多级的 $\boldsymbol{Q}$ 矩阵可以进行任意整数赋值，如 $q_{jk}=3$ 代表项目 j 考查指标为 3 的 k 属性，$\alpha_{ik}=2$ 代表被试者 i 掌握了 k 属性的第二层次。如果被试者要正确作答项目，则需要掌握考核属性指标水平及其以上的层次，如项目 j 测量的属性 $p_j=(1,3,2)$，属性 A_1,A_2,A_3 分别具有 2,4,3 种层次，则被试者掌握属性模式 $\alpha_i=\{(1,3,2)|(2,3,2)|(1,3,3)|(2,3,3)|(3,3,2)|(3,3,3)\}$才可能答对项目 j。

对于属性多级模型，α_{ik}和 q_{jk}的取值共有 L_k 种，即属性 k 的层级计有 $L\geqslant 2$ 种，因此如果属性 k 数值为非 0-1 元素，则理想反应得分 η_{ij}和项目反应函数不再适用，且增加了参数估计的难度和计算量。为了保持认知诊断模型的简洁性和易解释性，需要通过 Discriminant 函数将多级 $\boldsymbol{\alpha}$，q 转换为 0-1 元素。Discriminant 函数如下：

$$\alpha'_{ik}=\begin{cases}1, & \alpha_{ik}\geqslant q_{jk}\\ 0, & \alpha_{ik}<q_{jk}\end{cases}$$

$$q'_{jk}=\begin{cases}1, & q_{jk}\geqslant 1\\ 0, & q_{jk}=0\end{cases} \tag{4}$$

此时模型虽然实现了属性多级化的计算处理,满足了多级属性的客观考查要求,但观察得分同理想得分仍然无法对应,因此为进一步描述被试者掌握属性模式对项目的真实反馈,利用 Weight 函数将式(3)拓展为多级理想得分函数:

$$\eta_{ij}^{*}=round(\sum_{k=1}^{K}\rho_{ijk}\cdot\omega_{jk}\cdot mf_j) \tag{5}$$

$$\rho_{ijk}=\begin{cases}\dfrac{\alpha'_{ik}}{q'_{jk}}, & q'_{jk}=1\\ 0, & q'_{jk}=0\end{cases} \tag{6}$$

其中,ω_{jk}是项目 j 考查属性中 k 属性所占的权重,$\dfrac{\alpha'_{ik}}{q'_{jk}}$为被试者 i 在项目 j 上掌握属性的比例,mf_j 是第 j 题目的满分值。为便于参数估计,降低模型的计算规模,结合上述改进方案将式(2)转化为:

$$P^{*}(Y_{ij}=t\mid\boldsymbol{\alpha}_i)=(1-s_{jt})^{\varphi_{ijt}}\cdot g_{jt}^{1-\varphi_{ijt}} \tag{7}$$

$$\varphi_{ijt}=\begin{cases}1, & \eta_{ij}^{*}\geqslant t\\ 0, & \text{otherwise}\end{cases} \tag{8}$$

式(1),式(5),式(7)即为 PH-DINA 模型的项目反应概率函数。相比 P-DINA 模型,PH-DINA 模型不仅增加了被试者和项目多级属性指标的运算,还拓展了理想反应得分的计算方法,增加了项目的实际考查范围和反馈的信息。

2.2　参数估计

计算机自适应测试的参数估计一般包括被试者知识属性条件估计和项目参数条件估计,为优先保障项目参数的准确性,本文结合 PH-DINA 模型的知识属性参数对极大似然估计算法做了相应的改进,但在一定程度上增加了时间复杂度,仍需要对其进行进一步优化。

假设 $L(Y_i\mid\boldsymbol{\alpha})$是被试者 i 在多级属性评分下的似然函数,则有:

$$\mathrm{L}(\mathrm{Y}_i\mid\boldsymbol{\alpha})=\prod_{j=1}^{J}\prod_{t=0}^{mf_j}P(Y_{ij}=t\mid\boldsymbol{\alpha})^{u_{ijt}} \tag{9}$$

$$u_{ijt}=\begin{cases}1, & Y_{ij}=t\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

因此,PH-DINA 模型的似然函数为:

$$L(Y_i\mid\boldsymbol{\alpha})=\prod_{j=1}^{J}\prod_{t=0}^{mf_j}\left[(1-s_{jt})^{\varphi_{ijt}}\cdot g_{jt}^{1-\varphi_{ijt}}-(1-s_{jt+1})^{\varphi_{ijt+1}}\cdot g_{jt+1}^{1-\varphi_{ijt+1}}\right]^{u_{ijt}} \tag{11}$$

其中,$u_{ijt}=\{0,1\}$指被试者 i 在项目 j 上得 t 分的事实判断,则被试者 i 的知识属性的极大似然估计的计算式为:

$$\hat{\alpha}_i=\arg\max\{L(Y_i\mid\alpha_c)\} \tag{12}$$

对于项目参数估计中的 s,g 而言,精确度是首要考虑的要素,且项目参数的精确与否将会直接影响被试者知识属性的判准率,因此选用 MCMC 算法对其进行估计。项目属性参数引入了层级属性结构,模型的先验概率分布为 $g_j\sim 4-Beta(0,0.6,1,2)$,$1-s_j\sim 4-Beta(0.4,1,2,1)$,$\alpha_{ik}\sim U(0,L_k)$。

根据 Beyes 定理,待估参数的近似条件概率分布为:

$$P(s,g\mid\boldsymbol{Y},\boldsymbol{\alpha})\propto L(s,g\mid\boldsymbol{\alpha})P(s)P(g)$$

因此,从均匀分布 $U(s_j^{n}-\delta_s,s_j^{n}+\delta_s)$中随机抽取$\{s_j^{n+1}\}$,从均匀分布 $U(g_j^{n}-\delta_g,g_j^{n}+\delta_g)$中随机抽取$\{g_j^{n+1}\}$,现假定 $\delta_s=\delta_g=0.1$,则参数转移概率公式为:

$$p(\{s_j^{n},g_j^{n}\},\{s_j^{n+1},g_j^{n+1}\})=$$

$$\min\{\frac{L(s_j^{n+1},g_j^{n+1}\mid\alpha_{ik}^{n+1})P(s_j^{n+1})P(g_j^{n+1})}{L(s_j^{n},g_j^{n}\mid\alpha_{ik}^{n+1})P(s_j^{n})P(g_j^{n})},1\} \tag{13}$$

由于仅需要估计 PH-DINA 模型的项目 j 参数,假设 N 为参与测试的总人数,因此有效似然函数为:

$$L(s,g\mid\boldsymbol{\alpha})=\prod_{i=1}^{N}\left[\{(1-s_j)^{\eta'_{ij}}g_j^{1-\eta'_{ij}}\}^{Y_{ij}}\cdot\{1-(1-s_j)^{\eta'_{ij}}g_j^{1-\eta'_{ij}}\}^{1-Y_{ij}}\right] \tag{14}$$

3　试题推荐方法

直接应用基于认知诊断的自适应选题和基于协同过滤的试题推荐方法均存在一定的不足,前者对内在因素衡量的不确定性会提高参数估计的计算量,后者忽略了被试个体的独立性,导致可解释性弱、个性化差[11]。因此,本文在计算机自适应测试的应用场景下,提出了基于概率矩阵分解和认知诊断的试题推荐方法,简记为 CDPMF。

文献[11]所述的 PMF-CD 框架实质上是 DINA 模型和 PMF 模型的先后运用,主要依赖于已有的全部答题情况和基于 PMF 模型的得分预测,是一种静态输入、静态输出的模型框架,最终结果是输出限定条件下的一组试题。

本文所述试题推荐方法 CDPMF 是动态试题推荐,虽然同 PMF-CD 框架一样都是协同过滤和认知诊断理论的结合应用,但是本文主要以 PH-DINA 认知诊断模型为主要得分预测框架,PMF 模型仅作为其中的校正参数来提高估计得分概率的准确性,并最终通过信息量函数来实现自适应测试过程中的动态试题推荐,从而达到被试者每做一题便根据作答情况推荐符合其相应水平的试题的目的。该方法的具体内容如下。

3.1　得分预测

根据式(11)和式(12),可准确估计出被试者的认知属性结构,包括具体掌握的属性等级指标,直接用于概率矩阵分解。将同领域同岗位 n 个被试者的答题情况构建成得分矩阵 $\boldsymbol{S}$,S_{ij} 指被试者 i 对项目 j 的作答情况,从而提出特征参数 b_{ij} 作为 PMF 的先验信息,矩阵 $\boldsymbol{S}$ 和特征参数的计算式如下:

$$\boldsymbol{S}=\begin{bmatrix}S_{11} & \cdots & S_{1j}\\ \vdots & & \vdots\\ S_{i1} & \cdots & S_{ij}\end{bmatrix}$$

$$b_{ij}=b_i+b_j=\frac{1}{J}\times\sum_{j=1}^{J}S_{ij}+\frac{1}{I}\times\sum_{i=1}^{I}S_{ij} \tag{15}$$

其中,b_i 表示被试者 i 的知识得分先验程度,描述了被试者间知识掌握程度的差异性,即矩阵 $\boldsymbol{S}$ 第 i 行的平均值;b_j 表示试题 j 的先验得分,描述了项目之间的属性级别差异性,即矩阵 $\boldsymbol{S}$ 第 j 列的平均值。

在引入被试者和项目的先验参数 b_{ij} 后,通过在概率矩阵分解中加入被试者的认知属性掌握模式,可以使 PMF 分解出低维度的潜在因子被试特征矩阵 $\boldsymbol{M}$、测试项目特征矩阵 $\boldsymbol{N}$,其中,$\boldsymbol{M}\in\boldsymbol{Z}^{L\times I}$,$\boldsymbol{N}\in\boldsymbol{Z}^{L\times J}$,且 $\boldsymbol{M}_i$ 和 $\boldsymbol{N}_j$ 分别表示特定被试者 i 和试题 j 的潜在特征向量,用于刻画被试者及试题在低维空间下的表现,则被试者对项目的得分 R_{ij} 应满足条件分布:

$$p(\boldsymbol{R}\mid\boldsymbol{M},\boldsymbol{N},\sigma^2)=\prod_{i=1}^{I}\prod_{j=1}^{J}[N(R_{ij}\mid\boldsymbol{M}_i^{\mathrm{T}}\boldsymbol{N}_j,\sigma^2)] \tag{16}$$

将被试者的知识属性特征矩阵与测试项目特征矩阵作为

校正参数融入到认知诊断模型中，主要针对式(5)进行优化，以提高理想得分的预测准确性。CDPMF 得分预测模型的示意图如图 1 所示。

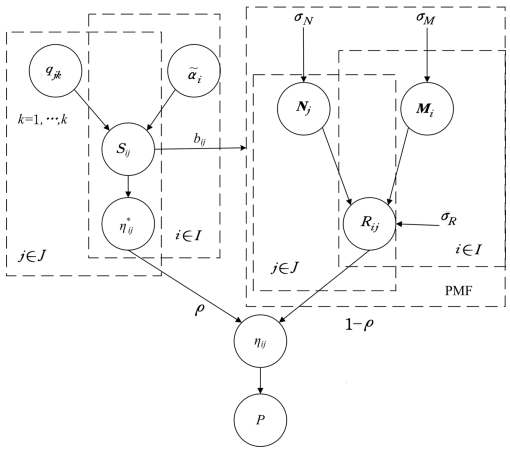

图 1 CDPMF 得分预测模型

Fig. 1 CDPMF scoring prediction model

在该模型框架下被试者的潜在作答情况可由式(17)表示：

$$\eta_{ij}=\rho\eta^*_{ij}+(1-\rho)\boldsymbol{M}_i^{\mathrm{T}}\boldsymbol{N}_j \tag{17}$$

其中，$\rho\in[0,1]$表示共性特征和被试者独立属性掌握模式之间的比例参数，ρ 值越大，预测得分受认知诊断模型的影响就越大，ρ 值越小，预测得分受共性特征的影响就越大。在模型应用过程中，应根据实际数据和测试情况选取 ρ 值。

预测模型中被试者的得分由理想得分、试题先验、被试先验和被试-试题 4 个部分交互组成，每一部分都可对观察值的某一属性进行解释。其中，$N(x\mid\mu,\sigma^2)$是均值为 μ、方差为 σ^2的高斯分布概率密度函数，则被试者的特征向量和项目的特征向量也应同时满足均值为 0 的高斯分布 $p(\boldsymbol{M}\mid\sigma^2)=\prod_{i=1}^{I}N(\boldsymbol{M}_i\mid 0,\sigma^2 I)$、$p(\boldsymbol{N}\mid\sigma^2)=\prod_{j=1}^{J}N(\boldsymbol{N}_j\mid 0,\sigma^2 I)$。通过贝叶斯推论，特征向量的后验概率应为：

$$\begin{aligned}&p(\boldsymbol{M},\boldsymbol{N}\mid\boldsymbol{R},\sigma_{\boldsymbol{R}}^2,\sigma_{\boldsymbol{M}}^2,\sigma_{\boldsymbol{N}}^2)\\&\propto p(\boldsymbol{R}\mid\boldsymbol{M},\boldsymbol{N},\sigma^2)p(\boldsymbol{M}\mid\sigma^2)\ p(\boldsymbol{N}\mid\sigma^2)\\&=\prod_{i=1}^{I}\prod_{j=1}^{J}[N(R_{ij}\mid g(\boldsymbol{M}_i^{\mathrm{T}}\boldsymbol{N}_j),\sigma^2)]^{I_{ij}^{R}}\times\prod_{i=1}^{I}N(\boldsymbol{M}_i\mid 0,\sigma^2 I)\times\\&\quad\prod_{j=1}^{J}N(\boldsymbol{N}_j\mid 0,\sigma^2 I)\end{aligned} \tag{18}$$

为便于 CDPMF 模型在试题推荐中的实际应用，其优化目标可变为最小化函数公式，λ 是模型的正则化系数：

$$E=\sum_{i=1}^{I}\sum_{j=1}^{J}I_{ij}(R_{ij}-\eta_{ij})^2+\lambda_{\boldsymbol{M}}\parallel\boldsymbol{M}\parallel_{Fro}^2+\lambda_{\boldsymbol{N}}\parallel\boldsymbol{N}\parallel_{Fro}^2 \tag{19}$$

另外，为解决协同过滤中存在的“冷启动”问题，避免因新被试者和新试题的出现干扰试题推荐，可通过 $\rho=1$ 规避 PMF 模型的影响，当共性特征无法提取时，则完全利用认知诊断的选题策略进行试题推荐。基于认知诊断的自适应测试通过新被试者或新试题的实际作答概况估计被试者的能力参数和试题的属性参数，具体算法详见 2.2 节。为进一步确保参数估计的有效性，可根据用户需求和实际情况在冷启动测试后补充同等参数条件下的试题对被试者的能力估计水平进行校正。

3.2 试题推荐

计算机自适应测试中，试题推荐方法可根据被试者潜在的作答情况 η^*_{ij} 向不同的被试者推荐合适的试题。智能教育中自适应测试的目的不仅是评估被试者的真实能力水平，更需要诊断其知识短板以便于及时查漏补缺或自我提升，因此试题的推荐不同于商品等项目可以按照被试者的兴趣或者难度等级进行推荐，而是选择能够快速高效地反馈被试者真实能力水平的试题。因此，只有当被试者的知识属性水平恰好在适应性范围内大于或等于项目所考查的级别时，提供的信息量才最大，这就是自适应测试中试题推荐的理论依据[15]。

知识属性往往是非连续性的，考虑到 PH-DINA 模型的参数多维性，本文选用 KL(Kullback Leibler)信息量[16]作为试题推荐指标，信息量越大越能证明被试者的认知状态属于估计的属性模式，即选择同知识属性掌握模式相当的试题推荐。HKL 选题策略[17]因为对属性掌握模式有良好的区分度在自适应测试中得到了广泛的应用，但因为原始公式面向的是二值函数，不适用于多级属性模式的计算，因此结合 PH-DINA 模型对其进行拓展，记作 PH-HKL 试题推荐。将参数估计所得的被试者能力水平参数与试题属性参数作为输入，选择能够使 PH-HKL 的信息量值最大的试题项目作为最优试题推荐。

PH-KHL 信息量不仅考虑了后验概率加权，而且进一步考虑了被试者之间知识属性的相似性，其计算公式为：

$$\begin{aligned}PHHKL_j(\hat{\boldsymbol{\alpha}})=\sum_{c=1}^{2^k}\sum_{t=0}^{mf_j}\frac{1}{d(\boldsymbol{\alpha}_c,\hat{\boldsymbol{\alpha}})}\cdot\{[\log(\frac{P(Y_j=t\mid\hat{\boldsymbol{\alpha}})}{P(Y_j=t\mid\boldsymbol{\alpha}_c)})\\P(Y_j=t\mid\hat{\boldsymbol{\alpha}})]\pi(\boldsymbol{\alpha}_c\mid Y_j)\}\end{aligned} \tag{20}$$

其中，$P(Y_j=t\mid\hat{\boldsymbol{\alpha}})$，$P(Y_j=t\mid\hat{\boldsymbol{\alpha}}_c)$指不同属性状态的被试者在项目上得分的反应概率，$\pi(\boldsymbol{\alpha}_c\mid Y_j)$是指知识属性为 $\boldsymbol{\alpha}_c(c=1,2,\cdots,2^K)$的后验概率，记 $\rho(\boldsymbol{\alpha}_c)$为知识状态 $\boldsymbol{\alpha}_c$ 的先验概率，则后验的概率公式为：

$$\pi(\boldsymbol{\alpha}_c\mid Y_j)=\frac{p(\boldsymbol{\alpha}_c)L(Y_j\mid\boldsymbol{\alpha}_c)}{\sum_{c=1}^{2^K}p(\boldsymbol{\alpha}_c)L(Y_j\mid\boldsymbol{\alpha}_c)} \tag{21}$$

$d(\boldsymbol{\alpha}_c,\hat{\boldsymbol{\alpha}})$指不同被试者的知识状态间的相似性，具体描述如下：

$$d(\boldsymbol{\alpha}_c,\hat{\boldsymbol{\alpha}})=\sqrt{\sum_{k=1}^{K}(\alpha_{ck},\hat{\alpha}_k)^2} \tag{22}$$

综上，基于协同过滤和认知诊断的试题推荐方法在进行试题抽取时，综合了被试者的认知属性掌握模式的个性和同类型被试人员知识属性的共性，体现了被试者当前的认知掌握模式等学习状态，符合复杂试题类型和多维知识属性的试题推荐环境，提高了试题推荐的准确度。

4 模拟实验分析与论证

4.1 评价指标

试题推荐的效果评价采用自适应测试的选题策略评价指标，包括属性判准率和非约束指标两类，判准率通常采用平均属性边际判准率(Average Attribute Match Ratio，AAMR)、模式判准率(Pattern Match Ratio，PMR)两类评价指标。非

约束类指标主要包括题库的安全性和测试效率(Testing Efficiency,TE),其中安全性包括题目曝光率(Exposure Ratio,ER)、测验重叠率(Testing Overlap Ratio,TOR)。

α_i 假设自适应测试中对 N 个被试者的 K 个属性进行认知诊断测验,是第 i 个被试者的真实知识状态,$\hat{\alpha}_i$ 是估计的被试属性掌握模式,具体的计算公式及含义阐述如下。

$$AAMR=\sum_{k=1}^{K}MMR(k)/K$$

其中,若被试者 i 的第 k 个属性判别正确则记为 $g_{ik}=1$,否则记为 $g_{ik}=0$,AAMR 是实验中全部属性的平均判准率,通过测试结果判断试题推荐的准确度。

$PMR=\sum_{i=1}^{N}h_i/N$,若 $\alpha_i=\hat{\alpha}_i$,则记 $h_i=1$,否则记 $h_i=0$,验证判断每一个被试者的知识属性掌握模式是否全部预测正确。

ER 反映了试题库的曝光程度,一般采用卡方类计算指标:

$$\chi^2=\sum_{j=1}^{M}\frac{[ER_j-E(ER_j)]^2}{E(ER_j)} \tag{23}$$

其中,$ER_j=f_j/N$ 是第 j 题的曝光率,f_j 是第 j 题被抽取的次数,$E(ER_j)$是试题 j 期望曝光率,ER_j 越小则曝光率越低,安全性越高。测试中的理想情况是所有试题都可被均匀抽取,即 $E(ER_j)=L/M$,L 是平均测验的长度,M 是题库的试题总数。因此,χ^2 用来统计观察曝光率与期望曝光率间的距离,χ^2 值越小,说明题库的安全性越高。

TOR 是反映不同被试者抽取相同试题的重叠情况,因此其计算式与题目曝光率、测验长度和被试人数有直接关系,重叠率越高,测试项目的安全性越低,公式如下:

$$\overline{\hat{T}}=\frac{N\times\sum_{j=1}^{M}ER_j^2}{(N-1)\times L}-\frac{1}{N-1} \tag{24}$$

TE 是综合评定测试效能比的指标,指在相同测量精度下平均耗用的试题数量,TE 值越低,效率就越高。其中 L_i 是测试中被试者 i 平均耗用的题目数量。

$$TE=\sum_{i=1}^{N}L_i/N \tag{25}$$

4.2　实验数据

为进一步研究和验证基于协同过滤和认知诊断的试题推荐方法的可行性,本文采用蒙特卡洛法进行模拟实验。参考高中数学教学大纲,本次模拟实验设置了 5 个相对独立的知识认知属性,每一个属性都将包含 4 个级别,即 $k=\{0,1,2,3\}$,分别对应大纲要求的"了解、理解、掌握、灵活运用"。根据大纲要求设计知识属性结构,如表 1 所列,有 $4\times3\times3\times3\times4=432$ 种可能存在的知识属性模式,即针对这 5 个知识点的考核至多存在 432 种评估结果。

表 1　知识属性指标

Table 1　Knowledge property indicators

属性	α_1	α_2	α_3	α_4	α_5
层级	4	3	3	3	4
值域	0,1,2,3	0,1,2	0,1,2	0,1,2	0,1,2,3

本文实验将从这 432 种知识属性掌握模式中随机抽取某一种对一名模拟考生进行答题测试,模拟考生的知识属性模式标记为 $\boldsymbol{\alpha}_i=\{\alpha_{i1},\alpha_{i2},\alpha_{i3},\alpha_{i4},\alpha_{i5}\}$,$i=\{1,2,\cdots,n\}$。实验模拟考生答题时,考生的任一知识属性参数值只有不小于测试题目的考查参数才能被认为是答题正确,即 $\alpha_{ik}\geqslant q_{jk}$,实验不考虑真实情况下可能存在的由于马虎等因素造成的误答问题。为提高实验结果的可信度,保证模拟考生的知识属性模式的覆盖范围,实验随机选取 $n=5000$ 名模拟考生进行答题测试。

测试试题同考生模拟方法相似,试题的模拟参数从均匀分布中随机产生,$s_{jt}\sim U(0,0.3)$,$g_{jt}\sim U(0,0.3)$,严格控制 $s_{jt}\leqslant s_{jt+1}$,$g_{jt}\geqslant g_{jt+1}$,且所有试题均采用 $mf_j=3$ 的评分方式,即要求试题参数 $q_{jk}\leqslant4$。

4.3　结果与分析

本研究基于实际测试选用两因素对比实验设计方法,两因素分为 3 种测量精度和 5 种试题推荐算法,包括随机推荐的选题策略 Random、不采用协同过滤的认知诊断试题推荐算法 HKL 改进型 PHHKL、仅使用 PMF 模型的试题推荐算法、PMF-CD[11] 试题推荐框架和本文所述的 CDPMF 试题推荐算法,实验结果如表 2 所列。需要说明的是,由于 PMF-CD 推荐方法是静态推荐,对考生过去的测试结果的依赖性强且存在冷启动等问题,在自适应测试环境下难以直接应用,因此在保证该模型主体方法不变和参数 $\rho=0.4$ 的情况下进行了适当的改进以方便实验结果的对比。由于实验数据量较大,借鉴文献[11]的参数调整和经验,本文所述的 CDPMF 推荐方法的参数 ρ 值选取为 0.6,以提高认知诊断模型的主体作用。

表 2　测试结果的指标数据对比

Table 2　Comparison of testing results

测量精度	推荐方法	AAMR	PMR	ER	TOR	TE
$p=0.75$	Random	0.922	0.805	0.18	0.07	22.81
	PHHKL	0.955	0.875	85.36	0.35	8.74
	PMF	0.947	0.870	46.27	0.24	10.36
	PMF-CD	0.952	0.886	66.32	0.29	8.88
	CDPMF	0.957	0.892	54.22	0.26	8.52
$p=0.80$	Random	0.955	0.868	0.18	0.07	24.02
	PHHKL	0.973	0.915	87.77	0.35	9.23
	PMF	0.966	0.913	47.85	0.24	10.96
	PMF-CD	0.970	0.927	69.02	0.29	9.37
	CDPMF	0.974	0.933	56.33	0.27	9.02
$p=0.85$	Random	0.963	0.899	0.20	0.07	25.98
	PHHKL	0.978	0.932	88.01	0.35	10.15
	PMF	0.971	0.928	48.99	0.24	11.68
	PMF-CD	0.975	0.936	70.36	0.29	10.18
	CDPMF	0.980	0.940	57.58	0.27	9.88

表 2 中的数据表明,CDPMF 在固定精度的条件下,其属性边际判准率和模式判准率优于其他试题推荐方法。以 PMR 测试结果数据曲线为例,如图 2 所示(由于 Random 策略的 PMR 结果明显低于其他方法,因此不在图中进行比较),CDPMF 的判准率在 3 个精度下均明显高于其他方法,但同时也表现出在较高精度下差异性逐渐减小的趋势。在测试效率的对比上,CDPMF 平均使用了(8.52+9.02+9.88)/3=9.14 题目,相比于 PMF 的(10.36+10.96+11.68)/3=11.00 题目、PHHKL 的(8.74+9.23+10.15)/3=9.37 题目

以及 PMF-CD 框架的(8.88+9.37+10.18)/3=9.477 题目分别提高了 20.35%,2.5%和 3.69%的效能。

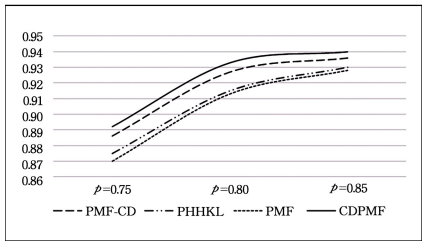

图 2 PMR 结果数据对比

Fig. 2 Comparison of PMR

此外,针对非约束指标的 χ^2 和 TOR,CDPMF 因需要权衡考虑被试者抽取试题的信息量以及同种群被试者的共性特征,使得在试题曝光率的控制上并不如单协同过滤的选题策略,但相比于 PHHKL 试题推荐方法仍有显著的进步。由于 PMF-CD 是限定条件范围内的试题推荐,虽然准确率较 PMF 有了明显的提高,但是试题的重叠率和曝光程度却比较严重,综合而言,CDPMF 的试题推荐效果优于该框架。

结束语 本文改进了多级多属性评分的认知诊断模型 PH-DINA,给出了同模型相适应的 PHHKL 选题策略以及参数估计方法,弥补了现有的自适应测试在处理多级属性评分数据上的不足,利用概率矩阵分解 PMF 模型辅助 PHHKL 信息量函数设计了试题的动态推荐方案,提出了一种基于协同过滤和认知诊断的试题推荐方法框架 CDPMF,解决了自适应测试中应用限制的问题,进一步提高了测评的智能化和个性化。实验证明,试题推荐方法综合考虑了个体的个性特征和群体的共性特征,提高了解释性和可靠性,为智能教育的诊断评估测试试题推荐提供了参考。

参 考 文 献

[1] WANG L,WU B,YANG J,et al. Personalized recommendation for new questions in community question answering[C]// IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. San Francisco, CA, USA: IEEE Press, 2016:901-908.

[2] VELDKAMP B P,LINDEN W J V D. Designing Item Pools for Computerized Adaptive Testing[M] // Computerized Adaptive Testing:Theory & Practice. Springer,2000:149-162.

[3] FIVES,BARNES H,NICOLE. Informed and Uninformed Naïve Assessment Constructors' Strategies for Item Selection [J]. Journal of Teacher Education,2017,68(1):85-101.

[4] SALEHI M,KAMALABADI I N. Personalized recommendation of learning material using sequential pattern mining and attribute based collaborative filtering[J]. Education & Information Technologies,2014,19(4):713-735.

[5] THAI-NGHE N,DRUMOND L,HORVÁTH T,et al. Matrix and Tensor Factorization for Predicting Student Performance [C]// Proceedings of the International Conference on Computer Supported Education(CSEDU 2011). Netherlands:IEEE Press, 2011:69-78.

[6] REN X,SONG M,HAIHONG E,et al. Context-aware probabilistic matrix factorization modeling for point-of-interest recommendation[J]. Neurocomputing,2017,241(6):38-55.

[7] CHEN G,ZHU F,HENG P A. Large-Scale Bayesian Probabilistic Matrix Factorization with Memo-Free Distributed Variational Inference[J]. ACM Transactions on Knowledge Discovery from Data,2018,12(3):1-24.

[8] SALEHI M. Application of implicit and explicit attribute based collaborative filtering and BIDE for learning resource recommendation[J]. Data & Knowledge Engineering,2013,87(9):130-145.

[9] KAPLAN M,TORRE J D L,BARRADA J R. New Item Selection Methods for Cognitive Diagnosis Computerized Adaptive Testing[J]. Applied Psychological Measurement,2015,39(3):167-188.

[10] SHAN R T,LUO Y C,SUN Y. Collaborative Filtering Algorithm Based on Cognitive Diagnosis[J]. Computer Systems Applications,2018,27(3):136-142. (in Chinese)
单瑞婷,罗益承,孙翼. 基于认知诊断的协同过滤试题推荐[J]. 计算机系统应用,2018,27(3):136-142

[11] ZHU T Y,HUANG Z Y,CHEN E H,et al. Cognitive Diagnosis Based Personalized Question Recommendation[J]. Chinese Journal of Computers,2017,40(1):176-191. (in Chinese)
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,2017,40(1):176-191.

[12] CHIU C Y,KÖHN H F. Consistency of Cluster Analysis for Cognitive Diagnosis: The Reduced Reparameterized Unified Model and the General Diagnostic Model[J]. Psychometrika, 2016,81(3):585-610.

[13] CAI Y,MIAO Y,TU D B. The polytomously scored cognitive diagnosis computerized adaptive testing[J]. Acta Psychologica Sinica,2016,48(10):1338-1346. (in Chinese)
蔡艳,苗莹,涂冬波. 多级评分的认知诊断计算机化适应测验[J]. 心理学报,2016,48(10):1338-1346.

[14] CAI Y,ZHAO Y,LIU S C,et al. An Extended Polytomous Cognitive Diagnostic Model[J]. Journal of Psychological Science, 2017,40(6):1491-1497. (in Chinese)
蔡艳,赵洋,刘舒畅,等. 一种优化的多级评分认知诊断模型[J]. 心理科学,2017,40(6):1491-1497.

[15] JING Y J,LI X,LIU T H. Using Maximum Information Selection Strategy to Computer Adaptive Test[J]. Advanced Materials Research,2014,1022(9):282-285.

[16] YANEZ F,BACH F. Primal-dual algorithms for non-negative matrix factorization with the Kullback-Leibler divergence[C]// IEEE International Conference on Acoustics,Speech and Signal Processing. New Orleans, LA, USA: IEEE Press, 2017: 2257-2261.

[17] HSU C L,WANG W C,CHEN S Y. Variable-Length Computerized Adaptive Testing Based on Cognitive Diagnosis Models [J]. Applied Psychological Measurement,2013,37(7):563-582.