

文章编号: 1000-5862(2014)03-0270-04

计算机化自适应认知诊断测验按模式分层选题策略

艾国金, 甘登文*, 丁树良

(江西师范大学计算机信息工程学院 江西 南昌 330022)

摘要: 提出了一种题库按属性模式分层并结合项目 Fisher 信息量、曝光控制因子和项目优先级的新的选题策略. 与已有方法比较, 该方法不仅提高了题库测验精度, 而且在项目使用均衡性上也表现优异.

关键词: 计算机化自适应认知诊断测验; 按模式分层; 选题策略; 题库安全; 测验精度

中图分类号: B 841.7; TP 301.6

文献标志码: A

0 引言

认知诊断是建立在传统测验理论基础之上, 针对个体认知过程、加工技能或知识结构(统称为属性)的诊断评估. 随着素质教育的不断推进, 对学生学业成绩的评定不仅仅依赖于成绩排名, 教育认知诊断可为家长、老师和学生提供诊断信息, 为因材施教提供参考和指导, 因此备受国内外研究者和应用者的青睐. 计算机化自适应认知诊断测验(CD-CAT)是认知诊断与计算机化自适应测验相结合的产物. 在测验中, 运用能够体现“因人施教、量体裁衣”的认知诊断 CAT 选题策略, 根据被试当前的状态自适应匹配项目进行测验, 获得被试对项目的反应, 通过反应快速、准确地诊断出被试对测验所涉及属性的掌握情况^[1].

CD-CAT 较传统 CAT, 在“自适应”方面有较大差异, 因此传统的 CAT 的选题策略难以直接照搬到 CD-CAT. 随着题库的不断更新, 规模的不断扩大, 对测验精度、速度、题库的安全性等方面的要求也越来越高. “快速、准确、安全”的 CD-CAT 选题策略成为了国内外研究者共同追求的目标.

1 预备知识

CD-CAT 主要用于诊断被试在测验所测属性上的掌握水平, 发现学生学习中存在的问题, 通过诊断

获得反馈信息, 为老师和学生进行针对性的教学提供参考和指导. 通常认为该测验属于低风险测验. 传统 CD-CAT 要求快速、准确地获得测验结果, 但是忽视了项目的曝光率和使用均衡性等问题, 导致有些项目曝光过高、有些内容使用次数过多, 致使题库安全性受到威胁, 而有些项目的曝光不足、有些内容使用次数过低, 又导致题库资源的浪费. 目前, 国内外关于 CD-CAT 项目曝光控制方法的研究并不多, Xu Xueli 等^[2]在基于简化的融合模型(FM or RUM)下通过随机法对 Shannon 熵(SHE)选题策略和 Kullback-Leibler(KL)信息量选题策略的曝光率进行了比较研究, 结果发现 KL 信息量选题策略在项目曝光率控制方面略好于 SHE 选题策略; 毛秀珍等^[3]在基于 DINA 模型下对分层多阶段(stratified multistage, SM)方法、项目合格(item eligibility, IE)方法、SM 与 IE 相结合得到的 SMIE 方法、最大修正优先指标(maximum modified priority index, MMPI)方法、限制阈值(restrictive threshold, RT)方法和限制进度(restrictive progressive, RPG)方法的曝光率进行了比较研究, 发现各方法项目曝光均匀性的优劣依次为: MMPI、RPG、SMIE、RT、SM 和 IE; 陈平等^[4]在基于 DINA 模型下对随机化选题策略(RANDOM, RD)、KL 信息量方法、SHE 方法、后验加权的 KL(posterior weighted kullback-leibler, PWKL)信息量方法和结合后验加权与距离加权的混合 KL(hybrid kullback-leibler, HKL)信息量方法的曝光率进行了比较研究, 结果发现 KL、SHE、PWKL 和 HKL 在

收稿日期: 2014-01-09

基金项目: 国家自然科学基金(30860084, 31160203, 31100756, 31360237), 国家社会科学基金(12BYY055)和江西省教育厅科技计划(GJJ13207, GJJ13226, GJJ13227, GJJ13208, GJJ13209)资助项目.

通信作者: 甘登文(1966-), 男, 江西奉新人, 教授, 主要从事计算机辅助教学和统计应用方面的研究.

项目曝光控制和题库使用均匀性方面均不尽如人意。在文献[5]提出按 $(s+g)$ 进行分层,在各个分层内选题时选取 $(s+g)$ 值最小的项目方法来控制曝光率;文献[6]在基于DINA模型下对RD选题策略、KL信息量方法、SHE方法和最大期望信息量方法(maximum expect information strategy, MEIS)的曝光率进行了比较研究,结果发现KL信息量、SHE和MEIS等3种选题策略的项目调用均匀性都较差。另外,兼顾能力与知识状态估计方面的研究也非常少, M. K. McGlohen等^[7]根据被试能力估计值 $\hat{\theta}$ 和被试估计的知识状态 $\hat{\alpha}$ 提出了3种设计方案并进行了比较,发现考察对象($\hat{\theta}$ 或 $\hat{\alpha}$)不同,得到的设计效果也不同; de la Torre^[8]认为在认知诊断中,作为知识状态的属性间可能存在相关,它们与一般智力或一般能力相关,从而提出了高阶DINA模型,并给出了属性与能力 θ 间的数学关系式,为描述被试对属性的掌握情况及掌握属性与被试能力之间的关系提供了丰富的信息。

针对上述问题,本文设计了一个新的基于按模式分层CD-CAT选题策略。在新的选题策略中,引入Fisher信息量,实现对不同被试在相同知识状态下其本身能力水平的区分,引入曝光控制因子和项目优先级,实现对项目曝光度的控制和均衡题库项目的使用频次,保证题库的安全性和项目使用的均匀性。

2 引入Fisher信息量、曝光因子和项目优先级的选题策略

2.1 DINA模型

K. K. Tasuoka^[9-10]提出的 Q 矩阵理论用 Q 矩阵来描述诊断内的属性和项目(知识状态)的关联性,经修正及拓展得以完善^[11]。包含 K 个属性和 m 个项目的 Q 矩阵表示为 $Q_{K \times m}$, $Q_{K \times m}$ 的每一列都代表一个项目类,该列的第 i 行为1或0,分别表示该项目是否包含了第 i ($i=1, 2, \dots, K$)个属性。本文沿用文献[12-13]的做法,将每一列称作一种模式,在测验开始时将题库按项目模式分层,计算每层项目失误参数和猜测参数的均值(即 \bar{s} 和 \bar{g}),在测验时统计被试对各层题库项目的反应,并计算每层项目将不同模式下的2类被试正确区分的概率,根据期望区分函数得到期望判准率最大的模式层,该层即为对被试施测项目所在的层。最后在层内选题时,采用引入Fisher信息量、曝光控制因子和项目优先级的新选

题策略进行选题,在小幅降低测验速度的前提下大幅提高了模式判准率和项目的利用率。

项目的期望区分函数(expected discrimination function, EDF)^[12-13]是被试对某一项目反应后,通过参数估计将该被试正确归类的概率,在本文中又称作期望判准率。在确定对被试施测项目所在层之前,本文通过EDF获得项目对被试的区分度,EDF最大的项目所在的层作为测验被试最佳的模式层。

在DINA模型^[8]下,区分函数 $f(\alpha_i, \alpha_j; t)$ 为

$$f(\alpha_i, \alpha_j; t) = \begin{cases} 0.5, & \text{若 } E_{tp} = (0, 0) \text{ 或 } (1, 1), \\ 1 - (\bar{s}_t + \bar{g}_t)/2, & \text{若 } E_{tp} = (1, 0) \text{ 或 } (0, 1), \end{cases} \quad (1)$$

其中 E_{tp} 表示期望反应模式(Expected Response pattern)。对区分函数取期望,得到项目期望区分函数 $y(t, h)$ 为

$$y(t, h) = \sum_i \sum_j P(\alpha_i; h) P(\alpha_j; h) f(\alpha_i, \alpha_j; t). \quad (2)$$

在(1)式中 α_i, α_j 为任意2种知识状态, \bar{s}_t, \bar{g}_t 分别为第 t 层子题库的失误参数和猜测参数均值。在(2)式中 $P(\alpha_i; h)$ 为知识状态 α_i 发生的概率,当 $h=0$ 时, $P(\alpha_i; 0)$ 为 α_i 的先验分布,当 $h>0$ 时, $P(\alpha_i; h)$ 为 α_i 的后验分布。

2.2 层内选题新方法

在文献[12-13]中只考虑了被试 i 掌握了第 j 个属性,在项目 k (失误参数为 s_k ,猜测参数为 g_k)上正确反映的概率,没有讨论属性模式相同的不同被试其本身能力水平如何区分,对项目内容平衡问题也没有关注,考虑到不同被试在拥有相同知识状态情况下其本身能力水平存在不同,同一层子题库中的项目的使用曝光度和先后顺序等问题,针对以上3个问题本文分别采用引入Fisher信息量、曝光控制因子和项目优先级,为层内选题提供参考依据。

受文献[14-15]启发,以1、0记分3参数Logistic模型^[17]的测验信息函数 $I_t(\hat{\theta}) = \sum_{j=1}^t I_j(\hat{\theta})$,其中 $I_t(\hat{\theta})$ 表示能力值为 $\hat{\theta}$ 的被试对第 t 层子题库的信息函数为基础,将题库中的第 j 个项目的曝光控制因子^[15]记为 $\lambda = m_j/\bar{m}$,其中 m_j 为第 j 个被试参加测验时,前 $(i-1)$ 个被试使用第 j 个项目的总次数, \bar{m} 为前 $(i-1)$ 个被试使用总题库的平均次数,即 $\bar{m} = \sum_{j=1}^M m_j/M$ (若 \bar{m} 等于0,则令其等于一个充分小的正数 ε),并引入第 j 个项目的区分度参数 a_j 的幂函数

$a(j, L, l) = a^{2(L-l)/L}$, 其中 L 为测验设定长度, l 为被试在测验第 j 个项目时的当前测验长度, 得到新选题策略中项目 j 的控制函数

$$U_j = I_i(\hat{\theta}) / [m_j / \bar{m} \cdot a(j, L, l)].$$

本文采用 $s + g$ 作为选题的参考, 得到新选题策略中项目 j 的控制函数 $W_j = u_i - (s_j + g_j) m_j$, 其中 u_i 表示第 i 个被试参加测验时, 前 $i - 1$ 个被试使用第 t 个模式层的总次数.

考虑到函数 U_j 与 W_j 的值可能不在同一个量级上, 根据刘珍等^[17]统一量纲的方法, 将获得的函数 U_j 和 W_j 的值作统一量纲处理 $f_j = U_j / \max_{j \in R_t} U_j + W_j / \max_{j \in R_t} W_j$, 其中 R_t 为第 t 层子题库的项目集. 通过模拟实验发现, 在选题时仅仅选取使函数 f_j 值最大的项目 j_0 , 项目的均衡性会受到影响, 结合曝光控制因子^[14]的思想, 本文引入层内项目优先级函数 $\omega_j =$

$1 - m_j / \bar{m}$ (若 \bar{m} 等于 0, 则令其等于一个充分小的正数 ε), 在层内选题时, 选择使函数 $f_j \omega_j$ 值最大的项目 j_0 , 即 $j_0 = \arg \max_{j \in R_t} (f_j \omega_j)$.

由于被试的模式(知识状态)种类众多, 本文使用模式判准率作为考察指标. 模式判准率(pattern match ratio, PMR), 即被试掌握模式并判准的人数占总人数的百分比, 计算公式为 $P_{mr} = N_p / N$, 其中 N_p 为被试掌握模式并判对的人数, N 为总人数.

3 CD-CAT 模拟实验

为验证新方法, 采用 Matlab8.0(R2012b) 为工具进行 Monte Carlo 模拟实验. 实验中共考察了 8 个属性, 分为 4 种结构: 线型、收敛型、发散型和无结构型. 如图 1 所示, 依次为 D_1 、 D_2 、 D_3 、 D_4 .

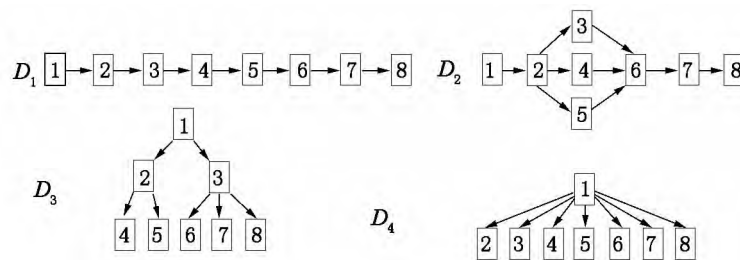


图 1 4 种属性层级结构图

被试人数设为 10 000 人, 被试能力真值服从标准正态分布, 记为: $\theta \sim N(0, 1)$, 测验长度 $L = 30$, 题库中每类模式的项目个数为 100, 项目的失误参数和猜测参数均服从均匀分布 $U(0.05, 0.25)$, 项目区分度参数 a 服从对数正态分布且 a 介于 0.2 和 2.5 之间, 难度参数 b 服从标准正态分布且 b 介于 -3 和 3 之间, 猜测参数 c 服从 a 为 5 和 b 为 17 的贝塔分布, 记为 $\ln a \sim N(0, 1)$, $b \sim N(0, 1)$, $c \sim \beta(5, 17)$. 基于 DINA 模型下的 Monte Carlo 模拟测验中, 由于被试能力初始值是人为给定的, 在施测过程中, 以 3 参数 Logistic 模型^[16]为基础, 参考陈平等^[18]能力估计方法, 对被试能力进行修正, 得到较精确的被试能力初始值.

在实验中, 将 RD、SHE、文献[13]等选题策略作为参照选题策略, 利用 Monte Carlo 模拟测验并重复 20 次求平均值的方法, 得到 4 种结构下不同选题策略的模式判准率, 如表 1 所示.

表 1 4 种结构下各选题策略的模式判准率

属性层级结构	RD	SHE	文献[13]	新方法
D_1	0.470 2	0.916 7	0.980 9	0.999 2
D_2	0.479 9	0.826 2	0.981 0	0.991 8
D_3	0.559 9	0.959 4	0.977 0	0.988 1
D_4	0.420 1	0.943 7	0.927 9	0.957 9

在实验中 D_1 、 D_2 、 D_3 、 D_4 结构对应的模式(包括全 0 列)分别为 9 种、31 种、46 种、129 种. 在被试人数为 10 000 人, 测验长度 $L = 30$, 题库中每类模式的项目个数为 100 的情况下, 不同选题策略在 4 种结构下的 Monte Carlo 模拟实验各重复 20 次, 每次实验中 10 000 人、每人 30 题的平均耗时如表 2 所示; 每次实验中每人、每题的平均耗时如表 3 所示.

表 2 4 种结构下各选题策略模拟实验每次实验平均耗时 s

属性层级结构	RD	SHE	文献[13]	新方法
t_{D_1}	27.820	1 337.810	57.875	118.891
t_{D_2}	28.740	10 894.023	360.109	643.440
t_{D_3}	29.560	29 962.340	943.880	1 516.710
t_{D_4}	38.060	641 450.140	21 305.320	26 486.230

表 3 4 种结构下各选题策略模拟实验每次每人每题的平均耗时

属性层级结构耗时	RD	SHE	文献 [13]	新方法
t_{D_1}	0.000 092 33	0.004 458 73	0.000 192 30	0.000 395 66
t_{D_2}	0.000 095 29	0.036 313 29	0.001 200 00	0.002 144 14
t_{D_3}	0.000 097 95	0.099 873 95	0.003 145 60	0.005 055 26
t_{D_4}	0.000 126 65	2.138 166 61	0.071 017 20	0.088 286 70

4 结果与讨论

从表 1 可看出 4 种结构下新方法的模式判准率比其他参照选题策略要高一些. RD 选题策略在选题时不需要对项目进行任何比较和计算, 直接从题库中随机选取下一题, 因此它的耗时最短; SHE 选题策略需要对题库中所有题目的信息量进行计算和比较, 选取区分能力最强的项目作为下一题, 在比较上花费太多时间, 因此它的耗时最长; 文献 [13] 方法和新方法在测验前已经按模式将题库进行了分层, 通过 EDF 获得不同被试最佳的测验模式层, 在层内按照各自方法选取下一题, 与 SHE 选题策略相比省去了计算和比较其他模式层中项目信息量的时间, 自然耗时较 SHE 选题策略短, 由于文献 [13] 方法和新方法在层内选题时需要计算和比较项目间的信息量, 因此又比 RD 选题策略耗时要长, 新方法比文献 [13] 方法多进行了项目 Fisher 信息量、层内优先级和统一量纲等处理, 因此耗时略长于文献 [13] 方法. 由于不同结构的模式数量不同, 从表 2 也可看出, 不同选题策略在 D_1 、 D_2 、 D_3 、 D_4 结构下, 随着模式数量的增加选题耗时也越来越多.

新方法是基于 DINA 模型下的一种选题策略, 该方法在选题中加入了对项目 Fisher 信息量、曝光控制因子和项目优先级 3 个因子, 在小幅降低测验速度的前提下, 大幅提高了测量精度, 使题库项目使用趋向均匀. 但新方法只探讨了在 DINA 模型下对项目曝光率的控制. 如果改 DINA 模型为其他模型, 新方法是否可用. 另外, 在新方法验证实验中对项目的失误参数和猜测参数都假定为服从均匀分布 $U(0.05, 0.25)$, 而现实中还存在其他情况, 对不同项目失误参数和猜测参数, 该方法的效果又如何, 这些都值得进一步探讨.

5 参考文献

[1] Wen Jianbing. Application of the rule space model in computerized adaptive testing for diagnostic assessment [D]. Hong Kong: The Chinese University of Hong Kong ,

2003.

[2] Xu Xueli ,Chang Huahua ,Jeff Douglas. A simulation study to compare CAT strategies for cognitive diagnosis [EB / OL]. [2013-02-19]. <http://iacat.org/sites/default/files/biblio/xu03-01.pdf>.

[3] 毛秀珍, 辛涛. 认知诊断 CAT 中项目曝光控制方法的比较 [J]. 心理学报, 2013, 45(6): 694-703.

[4] 陈平, 李珍, 辛涛. 认知诊断计算机化自适应测验的题库使用均匀性初探 [J]. 心理学报, 2011, 9(6): 125-132.

[5] 陈平. 认知诊断计算机化自适应测验的项目增补: 以 DINA 模型为例 [D]. 北京: 北京师范大学, 2011.

[6] 汪文义. 计算机化自适应测验选题策略研究: 以 GRM 和 DINA 模型为例 [D]. 南昌: 江西师范大学, 2009.

[7] McGlohen M K. The application of cognitive diagnosis and computerized adaptive testing to a large-scale assessment [D]. Austin: University of Texas, 2004.

[8] de la Torre J. DINA model and parameter estimation: a didactic [J]. Journal of Educational and Behavioral Statistics, 2009, 34(1): 115-130.

[9] Tatsuoaka K K. Architecture of knowledge structure and cognitive diagnosis: a statistical pattern recognition and classification approach [C]. Erlbaum: Hillsdale, 1995: 327-361.

[10] Tatsuoaka K K. Cognitive assessment: an introduction to the rule space method [C]. New York: Taylor & Francis Group, 2009.

[11] 丁树良, 杨淑群, 汪文义. 可达矩阵在认知诊断测验编制中的重要作用 [J]. 江西师范大学学报: 自然科学版, 2010, 34(5): 490-494.

[12] 尚志勇, 丁树良. 认知诊断自适应测验选题策略探新 [J]. 江西师范大学学报: 自然科学版, 2011, 35(4): 418-421.

[13] 卫芳娜, 甘登文, 丁树良. 认知诊断计算机化适应性测验按模式分层选题策略 [J]. 江西师范大学: 自然科学版, 2013, 37(5): 445-448.

[14] 程小扬, 丁树良, 严深海, 等. 引入曝光因子的计算机化自适应测验选题策略 [J]. 心理学报, 2011, 43(2): 203-212.

[15] 李萍, 甘登文, 丁树良. 自动控制区分度作用的选题策略 [J]. 江西师范大学: 自然科学版, 2013, 37(1): 101-105

(下转第 299 页)

The Synthesis of Analcime Zeolite Membranes from Local Kaolin and Diatomite

DING Min-zheng ZHANG Xiao-liang^{*}, WANG Liu-jie, LU Zhang-hui, HU Na, GUI Tian, CHEN Xiang-shu^{*}

(Jiangxi Inorganic Membrane Materials Engineering Centre, College of Chemistry and Chemical Engineering,
Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: The analcime(ANA) zeolite membranes were synthesized on the tubular mullite supports by the secondary growth method using local cheap kaolin and diatomite as silica and aluminum sources. The X-ray diffraction (XRD), scanning electron microscope(SEM) and gas permeation tests were used to characterize and evaluate the as-synthesized membranes. Several parameters were systematically investigated to evaluate their influence on crystallization and gas separation performance of the membranes, including the alkalinity, synthesis temperature and crystallization time. The results showed that pure ANA zeolite membranes with high crystallinity and good permeation performance were synthesized at 150 °C for 6 h under the optimized $n(\text{H}_2\text{O})/n(\text{Na}_2\text{O})$ ratio of 200 in the synthesis sols. The hydrogen permeance and ideal selectivity of H_2/CO_2 are achieved to $1.8 \times 10^{-8} \text{ mol} \cdot \text{m}^{-2} \cdot \text{s}^{-1} \cdot \text{Pa}^{-1}$ and 15, respectively, which is higher than their Knudsen diffusion values of 4.7 at room temperature.

Key words: kaolin; diatomite; analcime zeolite membrane; secondary growth; hydrogen separation

(责任编辑: 刘显亮)

(上接第 273 页)

[16] 漆书青,戴海琦,丁树良. 现代教育与心理测量学原理 [M]. 北京: 高等教育出版社, 2002.

[17] 刘珍,丁树良,林海菁. 基于 GPCM 的计算机自适应测验选题策略比较 [J]. 心理学报, 2008, 40(5): 618-625.

[18] 陈平,丁树良,林海菁,等. 等级反应模型下计算机化自适应测验选题策略 [J]. 心理学报, 2006, 38(3): 461-467.

The Exploration of p -STR Item Selection in Computerized Adaptive Testing for Cognitive Diagnosis

AI Guo-jin, GAN Deng-wen^{*}, DING Shu-liang

(College of Computer Information Engineering, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: A new item selection strategy of computerized adaptive testing for cognitive diagnosis is proposed. It partitioned an item pool into some sub-pools according to the attribute pattern(briefly p -STR) in item and combined with Fisher information, exposure-control factor and item priority. Compared with pre-existing methods, the new strategy not only improves the measurement accuracy of item pool but also works well in balance of using items.

Key words: CD-CAT; pattern-stratified; item selection strategy; securer item pool; measurement accuracy

(责任编辑: 冉小晓)