文章编号:1000-5862(2013)05-0445-04

# 认知诊断计算机化自适应测验按模式分层选题策略

卫芳娜,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**:提出一种题库按属性模式分层,并结合项目曝光度控制的新方法. 蒙特卡洛模拟研究显示:该方法在题库使用均匀程度上表现优异,并保有较高测量精度.

**关键词**:计算机化自适应认知诊断测验;题库分层;选题策略;题库安全
**中图分类号**:B 841.7;TP 301.6　　**文献标志码**:A

## 0 引言

认知诊断评估是针对被试在解决问题时所需使用的知识、技能和策略(统称为认知属性),开发出相应的测试项目,然后利用被试在测试项目上的反应达到对内在各方面的认知属性和认知过程进行推断的目的[1]. 具有认知诊断功能的计算机化自适应测验(Computerized Adaptive Testing for Cognitive Diagnosis,CD_CAT)根据 CAT 思想,运用智能的选题策略,依据对被试知识状态的估计,选择恰当的与所估出的知识状态相适应的项目,并依据被试对所选项目的反应,快速而准确地诊断被试知识状态[2]. CD_CAT 既具有认知诊断功能,能对被试的心理特质做结构化的细致测量,同时又兼具 CAT 的测验形式,可以提高诊断测量的准确性和效率. 因此,近些年来,CD_CAT 在教育测量领域备受关注.

通常认为认知诊断测验是低风险测验,所以传统的认知诊断 CAT(CD_CAT)在判定选题策略时,存在重测量精度轻平衡项目曝光度的倾向,即选题策略存在项目曝光不均匀问题,使某些项目过度曝光,而某些项目曝光不足. 然而,在实施认知诊断过程中,如果将诊断结果作为某种考核指标,引起某些学校、老师对认知诊断测验的过度重视,这必然会使诊断测验的风险加大. 项目曝光不均匀,使用频数高的项目过度曝光会给测验题库的安全性带来威胁,有些项目使用频数低甚至根本不使用而导致测验题库资源的浪费. 在目前文献中,关于 CD_CAT 项目曝光控制方法的研究并不多,只有 Xu Xueli 等[3] 以及汪文义[4] 对 CD_CAT 选题策略的曝光率进行过比较研究,陈平[5] 以香农熵选题(SHE)策略为例,提出了 3 种项目曝光控制的方法,分别是按猜测参数和失误参数相加的值($g+s$)分层的方法,修改的最大优先指标法,按 $g+s$ 分层和修改的最大优先指标法二者相结合的方法,这 3 种方法在不同程度上提高了项目使用的效率,但后 2 种方法在测量精度上受到一些损失.

针对上述问题,本文建立了一个新的选题策略. 在新的选题策略中引入了曝光因子,实现对项目的曝光控制,均衡整个题库中项目的曝光率,尽量使每个项目的使用频次接近于题库中项目使用频次的平均水平,使题库的使用更均匀.

## 1 引入曝光控制的选题策略

### 1.1 DINA 模型下,期望区分函数简介

K. K. Tasuoka 提出的 $Q$ 矩阵理论[6],引入 $Q$ 矩阵表达属性和项目之间的关联,包含 $K$ 个属性、$n$ 个项目的 $Q$ 阵为 $Q_{K\times n}$. $Q_{K\times n}$ 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 $i$ 个属性($i=1,2,\cdots,K$). 将每一列称作一个项目属性模式. 考虑到项目属性模式在评估被试知识状态中所起的特殊作用,沿用尚志勇等[7] 的做

收稿日期:2013-03-16
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(1956-),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.

法,将题库按项目属性模式分层,并计算各层题库中项目参数的均值,依据最大期望判准率选择项目所在的层,最后在层内选题时未采用原方法随机选题施测的做法,转而采用函数控制项目曝光度选题的策略,在小幅降低测量精度前提下大幅提高题库的利用率.

整个 CD_CAT 的过程可概括为,根据对被试知识状态的估计值去选择最适合当前被试的项目,然后又根据被试对所选项目的反应去估计被试真实的知识状态,如此反复,直到测验符合终止条件为止,如测验长度达到要求等. 项目的期望区分函数(Expected Discrimination Function,EDF)[7],也即期望判准率,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一个值. 在新选题策略中,将期望判准率作为评价项目对被试的区分能力. *EDF* 值最大的项目被认为是当前测试被试最恰当的项目.

假设 $\alpha_i$ 为被试真实模式,$\alpha_j$ 为干扰模式,则某个项目对被试的期望判准率是根据被试的期望反应来估计被试的模式后,将被试归类正确的一个值. 当然,被试的干扰模式可能不止一种,而是一个集合,且这个集合中的所有模式在测验前有个先验分布概率、在测验当中有个后验分布概率,以用来反应被试属于各种模式的可能性大小. 这时,可以计算项目在各种干扰模式下的期望判准率,将这些期望判准率加权求和为该项目的期望判准率. 在真实测验中,被试的真实模式也是不知道的,它也是服从某个概率分布的集合,所以 $\alpha_i$ 和 $\alpha_j$ 是对称的,可以不再区分真实模式和干扰模式.

由此,DINA 模型的期望区分函数(EDF)构造过程如下:若已知 $\alpha_i$ 为被试真实模式,$\alpha_j$ 为干扰模式时,对任意的知识状态为 $\alpha_i$、$\alpha_j$ 这 2 类被试在第 $t$ 个项目上的期望反应模式(Expected Response Pattern,ERP)只能是以下 4 种:(0,0),(0,1),(1,0),(1,1). DINA 中规定 $1-s_t > g_t$,故 $1 > s_t + g_t$,即 $0.5 > (s_t + g_t)/2$,也即 $0.5 \leqslant 1-(s_t+g_t)/2$. 如果反应结果为(0,0)或(1,1),说明项目 $t$ 不能区分 $\alpha_i$ 和 $\alpha_j$ 两知识状态,此时令区分函数 $f(\alpha_i,\alpha_j;t) = 0.5$. 如果反应结果为(0,1)或(1,0),说明项目 $t$ 能区分 $\alpha_i$ 和 $\alpha_j$ 两状态,此时令 $f(\alpha_i,\alpha_j;t) = 1-(s_t+g_t)/2$. 因此,DINA 模型下的区分函数可定义为

$$f(\alpha_i,\alpha_j;t) = \begin{cases} 0.5, & \text{若 } ERP = (0,0) \text{ 或 } (1,1), \\ 1-(s_t+g_t)/2, & \text{若 } ERP = (0,1) \text{ 或 } (1,0). \end{cases} \quad (1)$$

因为 $\alpha_i$、$\alpha_j$ 可以是任意的知识状态,其出现的概率为

$$P(\alpha_i;X_1,X_2,\cdots,X_h) \doteq P(\alpha_i;h), \quad (2)$$

其中当 $h=0$ 时为 $\alpha_i$ 的先验分布;当 $h>0$ 时为 $\alpha_i$ 的后验分布.

对区分函数取期望,得到期望区分函数为

$$g(t,h) = \sum_i \sum_j P(\alpha_i;h)P(\alpha_j;h)f(\alpha_i,\alpha_j;t). \quad (3)$$

根据上述原理,首先将整个题库按项目属性模式进行分层,项目属性模式相同的项目放在同一层,并计算每层项目参数的均值(此处为 $s$ 均值和 $g$ 均值). 接着计算每层项目将相异模式的两类被试正确区分的概率,从期望区分函数值最大的那一层选择所要施测的项目.

### 1.2 层内选题方法

香农熵(SHE)方法的优点是模式判准率高,缺点是公式复杂,且趋向于选择 $s+g$ 值小的项目[8],使题库的使用很不均匀. 受此启发,本文虽然也采用 $s+g$ 值作为选题的参考,但是为了克服其缺陷,在其基础上加入调节因子[9] $\lambda_j = m_j/\bar{m}$,层内的选题方法为

$$f_j = (s_j+g_j)m_j/\bar{m}. \quad (4)$$

当对第 $m$ 个考生施测时,$m_j$ 表示项目 $j$ 被前 $m-1$ 个考生调用的次数,$\bar{m}$ 表示前 $m-1$ 个考生调用题库中所有项目的平均次数,即 $\bar{m} = \sum_{j=1}^{M} \frac{m_j}{M}$,$M$ 为题库中项目总数,$s_j$,$g_j$ 分别为项目 $j$ 的失误参数和猜测参数.

在选题时,选择使 $f_j$ 值最小的项目 $j_0$,即

$$j_0 = \arg\min_{j\in R_\alpha} f_j, \quad (5)$$

其中 $R_\alpha$ 为对于当前被试尚未施测的项目的集合,也称为剩余题库[9].

由(4)式可得,当 $m_j$ 增大时,则 $\lambda_j$ 随之增大,而 $f_j$ 也随之增大,而选题的目标是选择使 $f_j$ 值小的项目,由此说明,该项目被调用的次数越多,则其在之后的测试中被选到的概率就降低. 反之,项目被调用的次数越少,则在之后测验中被使用的概率就增加. 以此来达到既与香农熵(SHE)方法选择 $s+g$ 值小的项目的目的,同时又提高了项目调用的均匀性. 当然,(4)式是理论上的公式,在实际操作过程中,为避免出现 $m_j$ 或 $\bar{m}$ 为 0 的情况,有 2 种处理方式:(i)分别在 $m_j$ 和 $\bar{m}$ 上加上一个足够小的正数 $\varepsilon$;(ii)将其转化为与之等价的方法. 本文采取(ii)方式,做如

下处理:

$$j_0 = \arg\max_{j \in R_\alpha}(u_t - (s_j + g_j)m_j), \quad (6)$$

其中 $u_t$ 为第 $t$ 个子题库的使用次数,当选定项目所在的层后,$u_t$ 为固定值,此时上式选出来的项目就是与(5)式相符的项目.

## 2　CD_CAT 的模拟实验

### 2.1　被试及题库模拟

实验共考察 8 个属性($K = 8$),4 种属性层次结构,线型、收敛型、发散型、无结构型,如图 1 所示,依次为 $D_1$、$D_2$、$D_3$、$D_4$.

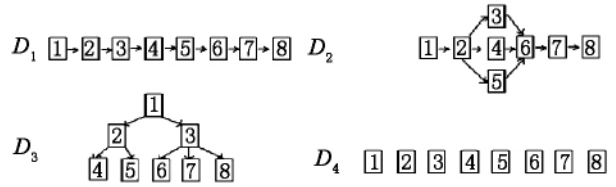


图 1　4 种属性层级结构图

实验题库中项目的失误参数和猜测参数均服从均匀分布 $U(0.05, 0.25)$,每类项目模拟 100 个. 被试人数 $N = 10\,000$,被试的知识状态服从均匀分布,测试长度 $L = 30$.

测验一开始,被试的属性掌握情况是未知的,假设每位被试有 50% 的概率掌握每个属性,以此随机生成被试的初始知识状态 $\alpha_0$,然后基于 $\alpha_0$ 根据选题策略选择第 1 个项目[10].

### 2.2　评价指标

本文选用模式判准率,$\chi^2$ 检验统计量和测试重叠率[9,11]来评价各方法的质量.

(i) 模式判准率(Pattern Match Ratio,PMR)

若被试的测验估计值 $\hat{\alpha}$ 和真实值 $\alpha$(在测验前的被试知识状态模拟值)2 个向量相等,则认为对被试的模式判断正确,否则认为对被试的模式判断错误. 每次实验后,统计测验模式被正确判断的被试个数 $PN$,再除以被试总人数 $N$,即得模式判准率,即 $PMR = PN/N$.

(ii)$\chi^2$ 检验统计量

$$\chi^2 = \sum_{j=1}^{M}\left(A_j - \left(\sum_{j=1}^{M} A_j/M\right)\right)^2 \Big/ \left(\sum_{j=1}^{M} A_j/M\right),$$

其中 $A_j$ 为第 $j$ 题曝光率,且 $A_j$ = 第 $j$ 题被使用的次数 $/N$,$M$ 为题库中项目总数.

(iii) 测试重叠率

$$Rt = 2TO_{总}/\left((N-1)\sum_{i=1}^{N} L_i\right),$$

其中 $TO_{总}$ 是考生的试题重叠总数,且 $TO_{总} = \sum_{j=1}^{M} C_M^2$,其中 $M_j$ 为题库中第 $j$ 题使用次数,$N$ 为被试总数,$L_i$ 为第 $i$ 个被试的测试长度.

### 2.3　实验结果分析

将随机选题策略(RD)、加权 KL 选题策略、香农熵选题策略(SHE)作为参照选题策略,把新方法同香农熵选题做精度对比以及同随机选题做项目曝光度对比,结果如表 1 和表 2 所示. 从表 1 可看出,在 4 种类型的属性层级结构下新方法仍旧保持着较高的模式判准率,而表 2 显示,加入了曝光控制后的新方法在题库的安全性上表现较为优异,无论是 $\chi^2$ 检验统计量还是测试重叠率,新方法都有所提高,项目的使用趋于均匀.

表 1　各方法在 4 种结构下的模式判准率

| 属性层级结构 | RD | 加权 KL | SHE | 新方法 |
|---|---|---|---|---|
| $D_1$ | 0.461 9 | 0.902 4 | 0.917 8 | 0.981 4 |
| $D_2$ | 0.482 5 | 0.746 2 | 0.827 3 | 0.980 5 |
| $D_3$ | 0.567 6 | 0.763 1 | 0.961 5 | 0.976 8 |
| $D_4$ | 0.423 9 | 0.567 3 | 0.942 0 | 0.928 3 |

表 2　各方法的题库使用均匀情况

| | RD | 加权 KL | SHE | 新方法 |
|---|---|---|---|---|
| $\chi^2$ 检验统计量 | 0.036 76 | 151.228 40 | 148.953 20 | 3.873 40 |
| 测试重叠率 | 0.033 10 | 0.471 50 | 0.468 70 | 0.041 30 |

## 3　讨论

本文新方法是一种基于 DINA 模型的选题策略,在选题中加入了对项目曝光控制的因子,在确保测量精度的前提下,能够使得项目的使用趋于均匀. 但新方法只探讨了 DINA 模型下项目曝光控制的方法,对于其他认知诊断模型并未做探讨,而且采用不同的认知诊断模型,期望区分函数(EDF)的计算公式也是不一样的,本文所介绍的公式,仅适用于 DINA 模型这个特例上. 当期望区分函数应用到其他模型时,具体计算公式如何定义、效果如何仍有待探讨. 同时,对于项目的内容平衡问题本文新方法也没有予以关注,在后续的研究中将一并考虑.

## 4　参考文献

[1] Leighton J P, Gierl M J. Cognitive diagnostic assessment

for education: theory and applications [M]. Cambridge: Cambridge University Press, 2007.

[2] 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.

[3] 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.

[4] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] Cheng Ying. When cognitive diagnosis meets computerized adaptive testing [J]. Psychometrika, 2009, 74(4): 619-632.

[11] 刘珍, 丁树良, 林海菁. 基于 GPCM 的 CAT 选题策略比较 [J]. 心理学报, 2008, 40(5): 618-625.

# *p*-STR Item Selection Strategy of Computerized Adaptive Testing for Cognitive Diagnosis

WEI Fang-na, GAN Deng-wen*, DING Shu-liang

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

**Abstract:** Partition of an item pool into some sub-pools according to the attribute pattern (briefly, *p*-STR) in items is made and a new item selection strategy is proposed. The four types of the attribute hierarchies of linear, convergent, divergent and unstructured structure are considered and each type corresponds to an item pool and tests are fixed-length. The results of Monte Carlo simulations indicate that compared to other methods, the new strategy is effective in improvement of the security of item pool, meanwhile with high measurement accuracy.

**Key words:** CD_CAT; stratification of item pools; item selection strategy; the security of item pool

(责任编辑: 冉小晓)