文章编号:1000-5862(2013)06-0657-04

结合影子题库的选题策略

戴 勰,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:高效安全的选题策略是计算机化自适应测验追求的目标. 最大 Fisher 信息量选题测验效率高、能力估计准确,但项目调用不均匀,影响考试的安全;而增设影子题库能较好地平衡项目调用的均匀性. 根据上述 2 种选题策略的优缺点,在 0-1 评分模型下,结合影子题库得到一种新的选题策略,并在按 a 分层、按最大信息量分层中引入了新的选题方法. 计算机模拟实验显示:新的选题方法效果比较理想.

关键词:计算机化自适应测验;影子题库;按 a 分层;按最大信息量分层
中图分类号:B 841.7;TP 301.6 **文献标志码**:A

0 引言

计算机化自适应测验(Computerized Adaptive Test,CAT)[1]是用项目反应理论建立题库,并由计算机根据被试能力水平自动选择测题,最终对被试能力做出估计的一种新型测验. 它既不同于传统的纸笔测验,也不同于一般的计算机化测验. 它根据被试对试题的不同作答,能自动选择最适宜的试题让被试作答,最终达到对被试能力做出最恰当的估计. 因此,CAT 是因人而异的测验. 目前 ETS 采用 GRE 的 General Test 和 GMAT 计算机考试则属于这种形式. 但令人震惊的是,2002 年 8 月,亚洲部分地区 GRE 考试的安全性出现问题,重新启用纸笔测试方式,取代 CAT. 从此 CAT 在实际应用中遭到了测验安全问题的质疑.

李铭勇等[2]认为国内外学者主要从 2 种研究思路提出了测验安全控制的方法:(ⅰ)控制项目的最大曝光率,沿着这个思路发展出来的方法有 SH 法、项目合格方法(item eligibility method)、多重最大曝光率法等;(ⅱ)改进选题策略,沿着这个思路发展的方法主要是按 a 分层法及其变式. 其中项目合格方法是 Wim J van der Linden 等[3-4]在 2004 年提出的,他们认为项目曝光的控制不是在项目选择之后,而是在被试参加测试之前. 也就是说,不是要决定选择的项目是否安排给被试,而是在项目选择之前就决定题库中的哪些题目对于被试来说是合格的,如果项目是合格的,那它就留在子题库中(或者叫做影子题库),否则就从子题库中移除. 有了影子题库(shadow bank)概念后,在多级评分模型中应用了影子题库[5]. 按 a 分层法(a-STR)[6]既能够使得题库中项目曝光相对均匀,又能使被试能力估计有较高的准确性,同时测验效率也较高. 考虑到在执行 a-STR 时,区分度不能按照指定的规则跟随能力估计精度的变化而逐题做比较细微的变化,这可能是影响该方法测验效率的一个原因,故提出将曝光控制因子引入到分层的 CAT 选题策略中,有效地兼顾了项目调用均匀性和测验效率[7]. 此外,按最大信息量分层选题策略[8],其效果比按 a 分层要好;在此基础上国内学者提出在等级评分模型下的最大信息量分层选题策略[9]和引入曝光因子的最大信息量动态分层法[10]等.

本文提出的选题策略是先选一批满足条件的项目作为影子题库,再从中选出一个 Fisher 信息量最大的作为被试作答的下一项目. 该策略能使得更好地权衡项目调用均匀性、能力估计准确性、测验的安全性和效率,并将此方法引入按 a 分层、按最大信息量分层.

1 新选题策略

1.1 3PLM 模型及其信息函数的简从

0-1 评份 3PLM[1]的第 j 个项目对能力为 θ_i 的被试的信息函数为

$$I_j(\theta_i) = \frac{D^2 a_j^2 (1 - c_j)}{[c_j + e^{Da_j(\theta_i - b_j)}][1 + e^{-Da_j(\theta_i - b_j)}]^2},$$

收稿日期:2013-09-12
基金项目:国家自然科学基金(30860084,31160203,31100756,31360237,31300876),国家社会科学基金(12BYY055)和江西省教育厅科技计划(GJJ13207,GJJ13227,GJJ13226,GJJ13208,GJJ13209,13JY01)资助项目.
通信作者:甘登文(1956-),男,江西奉新人,教授,主要从事计算辅助教学和统计应用方面的研究.

项目 j 的最大信息量($I_j^{\max}$)为

$$I_j^{\max}=\frac{D^2a_j^2}{8(1-c_j^2)}\left[1-20c_j-8c_j^2+(1+8c_j)^{3/2}\right].$$

当项目 j 的信息函数达到最大值时被试水平值($\theta_j^{\max}$)为

$$\theta_j^{\max}=b_j+\left(\ln\left(1+(1+8c_j)^{1/2}\right)-\ln 2\right)/(Da_j),$$

其中 a_j 为区分度参数,b_j 为难度参数,c_j 为猜测度参数,θ_i 为被试 i 的潜在特质(也称能力),D 为常量(常取 1.72).

1.2 最大信息量分层策略(MIS)

Juan Ramón Barrada 等[8]考虑到在 a-STR 中并未涉及猜测度参数 c,为了将 c 参数引入到 a 分层法中,对 a-STR 进行了 2 处重要的修改:(ⅰ)使用题目最大信息量 $I_j^{\max}$ 代替区分度 a;(ⅱ)用题目信息函数达到最大值时对应的能力值 $\theta_j^{\max}$ 代替难度 b.

1.3 曝光控制因子

曝光因子[7]记为 $ecf(j)$,此处 $ecf(j)=m_j/\bar{m}$ 的调节因子记为 $\lambda(j)$,$\lambda(j)$ 主要用来控制 $ecf(j)$ 在施测过程中对项目选择的影响力度(本文中 $\lambda(j)$ 取值均为 1).当第 i 个考生参加考试时,m_j 表示前 $i-1$ 个考生使用第 j 个项目的总次数,$\bar{m}$ 表示题库中所有项目被前 $i-1$ 个考生使用的平均次数,因此,$\bar{m}=\sum_{j=1}^{M}m_j/M$,$M$ 为题库中的项目总个数.

1.4 结合影子题库的新选题策略

在新选题策略中被试选择的作答项目 j 服从下列规则:(ⅰ)连续选出满足公式 $j=\arg\min_{j\in B_q}|\hat{\theta}-b_j|$ 的 35 个项目构成影子题库;(ⅱ)从影子题库中选择一个满足公式 $j=\arg\max_{j\in B_q}I_j(\theta_i)/ecf(j)$ 的项目作为被试下一个作答的目;其中 B_q 为每一阶段剩下可供选择的项目集合,$\hat{\theta}$ 为能力估计值,b_j 为所在层项目的难度值,$I_j(\theta_i)$ 为项目 j 对被试能力为 θ_i 的信息量函数值,$ecf(j)$ 为曝光控制因子.

在多级评分模型中用到的影子题库是出于降低项目调用均匀性的考虑[5],其做法是根据被试能力估计值从剩余题库中抽取一批最适合的项目,然后随机抽取一题作为被试的作答项目.本文的做法不仅考虑到项目调用均匀性和测验的安全性,而且还要兼顾能力估计准确性和效率.能力估计值与真值有偏差,仅仅考虑难度与估计能力差最小的那个项目不一定最好,或许它附近某个才是最佳项目,所以需抽取一组项目(规定一组 35 个);然后选择组内 Fisher 信息量最大的项目,可以又快又准地估计出能力值;但只选择信息量最大的又会导致曝光度较高,故沿用曝光因子.揉合这些方法是为了更好地权衡项目调用均匀性、能力估计准确性、测验的安全性和效率.

1.4.1 *引入新方法的按 a 分层* 若下文中 $ecf(j)=0$,则令 $ecf(j)=\varepsilon$,ε 是事先指定的一个足够小的正数.将新方法引入按 a 分层中,具体步骤为:(ⅰ)题库中所有项目按区分度参数 a 升序排列;(ⅱ)把题库划分成 K 层,每层项目数量大致相等,第 1 层 a 值平均最小,第 K 层 a 值平均最大;(ⅲ)被试先进入第 1 层选题,连续选出 35 个使得 $|\hat{\theta}-b_j|$ 最小的项目组成一组,再从组内选择一个使得 $I_j(\theta_i)/ecf(j)$ 最大的项目给被试作答;(ⅳ)当满足该层结束条件时进入下一层,一直作答到第 K 层,测验结束.

1.4.2 *引入新方法的按最大信息量分层* 该方法与上述做法基本相同,不同之处是用 $\theta_j^{\max}$ 代替 b_j,用 $I_j^{\max}$ 代替 a_j.

2 实验设计

2.1 被试及其题库的模拟

采用 Monte Carlo 方法模拟生成 1 000 个被试,所有被试的能力参数 θ 服从标准正态分布 $N(0,1)$,记为 $\theta\sim N(0,1)$.在实验过程中模拟生成 4 种题库,每种题库均生成 1 000 个项目:在题库 1 中:$\ln a\sim N(0,1)$,$b\sim N(0,1)$;在题库 2 中:$b\sim N(0,1)$,区分度参数 a 服从 0.2 到 2.5 的均匀分布,记为 $a\sim U(0.2,2.5)$;在题库 3 中:$\ln a\sim N(0,1)$,$b\sim U(-3,3)$;在题库 4 中:$a\sim U(0.2,2.5)$,$b\sim U(-3,3)$.以上 4 种题库猜测度参数 c 均服从 α 为 5 和 β 为 17 的贝塔分布,记为 $c\sim Beta(5,17)$.

2.2 模拟 CAT 的施测过程

本文在 CAT 施测时,参与比较的选题策略有:(ⅰ)随机法(RAN);(ⅱ)难度匹配法(b-match)[8];(ⅲ)最大 Fisher 信息量法(MFI)[11];(ⅳ)陈平影子题库(shadow)[5];(ⅴ)新方法(mix);(ⅵ)按 a 分层(a-STR);(ⅶ)程小扬引入曝光因子的按 a 分层(modi-a-STR)[7];(ⅷ)陈平影子题库的按 a 分层(shadow-a-STR);(ⅸ)引入新方法的按 a 分层(mix-a-STR);(ⅹ)按最大信息量分层(MIS);(ⅺ)引入曝光因子的按最大信息量分层(modi-MIS);(ⅻ)陈平影子题库的按最大信息量分层(shadow-MIS);(ⅹⅲ)引入新方法的按最大信息量分层(mix-MIS).实验重复 30 遍.

2.2.1 *能力初估阶段 CAT 的模拟* 本测验为 3PLM 模型下的 0-1 评分测验.在测验能力初始估计阶段,随机选 3 个题目给被试作答,被试的能力估计

值为被试作答正确与错误题目个数之比的自然对数. 若被试全部作答正确,则初估其能力值为 3; 若被试全部作答错误,则初估被试的能力值为 - 3; 被试在初估阶段所作答的项目不计入定长测验时被试的测验长度,在初始阶段所作答的项目信息量不计入不定长测验时被试的累积信息量.

2.2.2　能力精估阶段CAT的模拟　模拟定长和不定长 2 种测验: (ⅰ) 定长测验: 设定测验长度为 40,需要分层的统一分 4 层,每层 10 题,共 40 题[12]; (ⅱ) 不定长测验: 测验在被试累积信息量达到 16 时,则结束(不限定测验最大长度),需要分层的统一分 4 层,各层累积信息量分配比例是 1 :4 :9 :16[5].

2.3　评价指标

本文用以下 7 个指标[13]: 能力估计准确性(*ABS*)、能力估计标准差(*SD*)、测验效率(*Eff*)、项目调用均匀性(*Se*)、卡方统计量(χ^2)、测试重叠率(*Rt*)、人均用题数(*Nf*)来评价 CAT 的质量. 综合指标(*Y*)采用统一量纲[13],本实验假设 7 个评价指标同等重要,即加权系数均设定为 1.

3　实验结果分析

限于篇幅,只列题库 $\ln a \sim N(0,1)$, $b \sim N(0,1)$, $C \sim \beta(5,17)$ 的实验数据,见表 1 和表 2,其他题库结果类似.

从表 1 可以看出,当测验为定长测验时,新方法比陈平的方法[5]在能力估计准确性、能力估计标准差、测验效率方面表现要好,几乎接近最大 Fisher 信息量法,同时项目调用均匀性、测验重叠率、卡方这类指标明显比最大 Fisher 信息量法低很多,但稍差于陈平的方法. 当把新方法引入按 a 分层、按最大信息量分层时,效果更明显. 在按 a 分层中,新方法比按 a 分层、程小扬引入曝光因子的按 a 分层策略[7]各方面指标都要好,比陈平的方法在项目调用均匀性、测验重叠率、卡方指标差不多的情况下,能力估计准确性、测验效率提高颇多; 在按最大信息量分层策略中,结果亦如此.

表 1　$\ln a \sim N(0,1)$, $b \sim N(0,1)$, $C \sim \beta(5,17)$, 定长

选题策略	*ABS*	*SD*	*SE*	*Eff*	χ^2	*Rt*	*Y*
RAN	0.266 7	0.304 2	6.333 5	0.213 0	0.953 1	0.041 2	4.012 0
b-match	0.184 9	0.222 5	35.365 0	0.472 4	29.090 0	0.071 2	2.329 6
MFI	0.113 1	0.134 9	92.641 0	1.474 4	199.590 0	0.241 8	3.243 6
shadow	0.179 8	0.216 9	9.124 0	0.495 2	1.938 5	0.044 0	3.710 3
mix	0.126 0	0.153 4	44.114 0	1.019 7	45.258 0	0.087 3	3.105 1
a-STR	0.169 5	0.206 0	40.883 0	0.543 8	38.874 0	0.081 0	2.379 8
modi-*a*-STR	0.172 4	0.207 1	25.601 0	0.540 9	15.243 0	0.057 3	2.704 0
shadow-*a*-STR	0.189 0	0.227 3	19.115 0	0.445 8	8.499 7	0.050 6	2.753 4
mix-*a*-STR	0.159 5	0.193 7	20.063 0	0.603 3	9.361 9	0.051 4	3.033 9
MIS	0.166 6	0.201 6	41.027 0	0.549 1	39.147 0	0.081 2	2.406 7
modi-MIS	0.171 2	0.207 2	29.717 0	0.532 7	20.541 0	0.062 6	2.591 0
shadow-MIS	0.179 6	0.215 5	22.416 0	0.486 6	11.688 0	0.053 7	2.716 9
mix-MIS	0.159 1	0.193 3	22.108 0	0.614 3	11.367 0	0.053 4	2.967 5

从表 2 中结果可以看出,当测验为不定长测验时,由于不定长测验是规定测验总信息量达到一定值则终止测试,所以各策略的能力估计准确性和能力估计标准差相差无几,故主要看剩余 6 项指标. 比较可知,当不分层时,新方法在效率和人均用题量上比陈平的方法要好很多,但卡方、项目调用均匀性比陈平的方法稍差,测验重叠率指标与陈平的方法不相上下. 把新方法引入按 a 分层、按最大信息量分层后,新方法在按最大信息量分层中比其他方法各项指标都要好,在按 a 分层中人均用题数、测验效率比陈平方法要好,卡方、测验重叠率、项目调用均匀性比陈平的方法略差,但比其他方法的各项指标要好.

表 2　$\ln a \sim N(0,1)$, $b \sim N(0,1)$, $C \sim \beta(5,17)$, 不定长

选题策略	*ABS*	*SD*	*SE*	*Nf*	*Eff*	χ^2	*Rt*	*Y*
RAN	0.195 0	0.233 4	8.343 8	84.705 0	0.195 5	0.879 2	0.074 1	4.848 9
b-match	0.198 3	0.239 0	30.333 0	35.118 0	0.480 2	26.204 0	0.060 4	3.716 0
MFI	0.201 3	0.247 0	52.865 0	14.562 0	1.191 7	192.250 0	0.206 0	4.260 8
shadow	0.198 4	0.239 7	8.752 1	39.716 0	0.422 4	1.933 4	0.040 7	5.020 6
mix	0.198 5	0.241 7	19.504 0	19.905 0	0.857 3	19.130 0	0.038 0	4.872 7
a-STR	0.196 8	0.236 3	50.063 0	54.213 0	0.315 2	47.199 0	0.097 4	3.087 1
modi-*a*-STR	0.195 4	0.234 7	35.238 0	54.009 0	0.316 7	24.172 0	0.071 0	3.336 1
shadow-*a*-STR	0.196 6	0.235 8	26.512 0	58.269 0	0.291 8	12.811 0	0.063 0	3.463 7
mix-*a*-STR	0.195 2	0.234 9	29.672 0	49.957 0	0.341 8	17.865 0	0.065 3	3.483 0

续表 2

选题策略	*ABS*	*SD*	*SE*	*Nf*	*Eff*	χ^2	*Rt*	*Y*
MIS	0.196 2	0.234 7	50.006 0	51.662 0	0.328 7	49.813 0	0.096 3	3.125 1
modi-MIS	0.195 3	0.233 6	42.886 0	56.504 0	0.300 8	33.622 0	0.084 7	3.177 5
shadow-MIS	0.195 2	0.233 3	35.991 0	52.255 0	0.324 2	24.929 0	0.075 6	3.320 3
mix-MIS	0.197 4	0.236 4	27.246 0	41.135 0	0.416 9	18.110 0	0.058 0	3.689 7

从综合指标更能清晰地看出,在不分层的情况下,当项目参数 b 服从均匀分布时,新方法比陈平的方法结果要好,当 b 服从正态分布时,结果略差. 新方法在 2 种分层策略中,实验结果比其他策略都要理想. 总之,若要兼顾测验安全性及测验效率,那么新方法可认为是一种可供选择的选题策略.

4 讨论

本文在 3PLM 的 0-1 评分模型下,提出一种新的选题策略,实验中在选择一组项目时,通过模拟研究,讨论了组内到底包含多少项目为好的问题,最后认为组内个数规定为 35 比较好,是不是其他数值会有更好的效果,或者说组内项目个数或许不该是个定值而应是个变值,这个问题值得研究. 文中所提新策略都只是在 0-1 评分模型中探讨,如何迁移至多级评分模型中,还有待研究.

5 参考文献

[1] 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.

[2] 李铭勇,张敏强,简小珠. 计算机自适应测验中安全控制方法评述[J]. 心理科学进展,2010,18(8):1339-1348.

[3] Van der Linden W J,Bernard P Veldkamp. Constraining item exposure in computerized adaptive testing with shadow tests[J]. Journal of Educational and Behavioral Statistics,2004,29(3):273-291.

[4] Van der Linden W J,Bernard P Veldkamp. Conditional item-exposure control in adaptive testing using item-ineligibility probabilities[J]. Journal of Educational and Behavioral Statistics,2007,32(4):398-418.

[5] 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.

[6] Chang Huahua,Ying Zhiliang. *a*-stratified multistage computerized adaptive testing[J]. Applied Psychological Measurement,1999,23(3):211-222.

[7] 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.

[8] Juan Ramón Barrada,Paloma Mazuela,Julio Olea. Maximum information stratification method for controlling item exposure in computerized adaptive testing[J]. Psicothema,2006,18(1):156-159.

[9] 程小扬,丁树良,朱隆尹,等. 等级评分模型下的最大信息量分层选题策略[J]. 江西师范大学学报: 自然科学版,2012,36(5):446-451.

[10] 詹沛达,王立君,扬卫敏. 引入曝光因子的最大信息量动态分层法[J]. 中国考试,2013(2):12-20.

[11] 戴海琦,陈德枝,丁树良,等. 多级评分题计算机自适应测验选题策略比较[J]. 心理学报,2006,38(5):778-783.

[12] 程小扬,丁树良. 子题库题量不平衡的按 a 分层选题策略[J]. 江西师范大学学报: 自然科学版,2011,35(1):5-9.

[13] 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.

New Item Selection Method Combining with Shadow Bank

DAI Xie,GAN Deng-wen*,DING Shu-liang

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

Abstract: Computerized Adaptive Testing(CAT) has been in pursuit of the goal is to develop both efficient and safe item selection strategies. It is well known that there is a typical selection strategy called Maximum Fisher Information (MFI). However,this strategy has its advantages together with its downsides. On the one hand,MFI method can obtain high efficiency and accurate estimation of ability; on the other hand,its uneven item selection may lead to the insecurity of examination. Meanwhile,though shadow bank can be a good method of the item called evenly,it may result in the inefficiency of the test. Taking the advantages and disadvantages of the two selection strategies in the 0-1 scored CAT into consideration,a new item selection strategy is proposed in this paper,and pull this new method in *a*-Stratification(*a*-STR) and Maximum Information Stratification(MIS). The computer simulation shows that the new method works ideally.

Key words: computerized adaptive testing; shadow bank; *a*-STR; MIS

(责任编辑: 冉小晓)