文章编号:1000-5862(2014)02-0119-05

基于抽样原理的计算机化自适应测验选题策略

章沪超,丁树良*,戴 勰,关潮辉
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要:沿用曝光控制因子的同时,基于抽样原理,引入区分度分布因子,按区分度的分布情况来选取测验中的项目.以 $\ln a \sim N(0,1)$ 为例,Monte Carlo 模拟结果表明:该方法在估计精度、效率和安全性等指标上表现得比较优异.

关键词:抽样原理;计算机化自适应测验;分布因子
中图分类号:B 841.7;TP 301.6 **文献标志码**:A

0 引言

随着计算机的发展,计算机为计算机化自适应测验的编制、心理和教育测量的实施提供了不可或缺的技术支持[1].现在,美国教育测验服务社(ETS)对研究生资格考试(GRE)等实行了计算机化自适应测验(CAT).随着计算机和项目反应理论等的发展,他们更把普遍推行 CAT,全面实现考试的"无纸化"作为自己的一项世纪工程[2].同样地,国内 CAT 的研究与应用也在蓬勃发展,如江西师范大学、第四军医大学等单位编制的一系列 CAT 等[3].

CAT 致力于给每个考生"量身"定做一个能真实反映其能力的测验.题库、选题策略、能力估计方法以及终止条件是 CAT 主要的组成部分,而其中最为关键的是怎样从题库中选取最优项目作答[4].正是因为它的重要性,其好坏直接影响到测验效率、题库的安全性和经济性.Lord 提出的 MIC 法以及 Chang Huahua 等[5-6] 提出的按 a 分层法(a-STR)是比较常用的几种方法[5-6].在他们的基础上文献[7]提出引入曝光控制因子(*ecf*)的选题策略,文献[8]又在文献[9]基础上提出自动控制区分度的选题策略.MIC 的优点是测验效率很高,能力估计准确.但是,由于 MIC 原理上是选取信息量大的题目,而信息量与区分度的平方成正比,所以区分度大的项目被选到的概率远远大于区分度小的项目,其结果必然导致测验的曝光率大大增加,这对于题库的安全性就是一大威胁.因为考生在考前可以对高曝光度的试题做好充分准备,从而降低了试题的真实难度,这对于考试的公平性也是一大挑战.而 a-STR 针对 MIC 做出重大修改,其原理是让区分度非递减排序,然后按序分成若干层,施测时先从低区分度的层中选取项目,满足一定条件后再选取区分度更高层的项目,以此类推,直到满足终止条件退出测验.该选题策略在能力粗估阶段被强制选取区分度低的项目,能力精估阶段选取高区分度的项目.但此选题策略也有不足之处,比如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.文献[7-8]所提出的选题策略是在 MIC 的基础上引入 1 个项目 j 的控制曝光因子和区分度的幂函数作为信息函数的分母,有效地兼顾了项目调用均匀性和测验效率,而文献[8]相对于文献[7]的优点在于文献[8]的方法看似没有对区分度分层,其实质上分层更细,每个作答的项目即为 1 层.

纵观以上各种方法可以发现,它们均未考虑到题库中区分度的分布情况,即区分度是服从正态分布还是均匀分布抑或是其他分布.试想,当题库中 $\ln a$ 服从标准正态分布时,$\ln a$ 等于 0 周围,即 a 等于 1 周围的项目密度最大(a 在相等的取值范围内,对应的项目数最多).很自然地想到,如果项目数多的区间内题目多选取一些,项目数少的区间内少选取一些,自然也会起到均匀选题的效果.于是,为了起到选题的均匀效果,本文以最常见的区分度的自然对数服从标准正态分布,即 $\ln a \sim N(0,1)$ 为例,

收稿日期:2013-10-17
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(12BYY055,13BYY087)和江西省教育厅科技计划(GJJ3207,GJJ13226,GJJ13227,GJJ13208,GJJ13209)资助项目.
通信作者:丁树良(1949-),男,江西樟树人,教授,博士生导师,主要从事计算辅助教学及教育和心理测量方面的研究.

提出在曝光因子的基础上引入区分度分布因子的新的选题策略.

1 预备知识

1.1 3 参数逻辑斯蒂克模型

3 参数逻辑斯蒂克模型,其函数为

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}},$$

其中 $D = 1.7$,为量表因子[2].

1.2 3PLM 项目信息函数

3PLM 项目信息函数为

$$I_i(\theta) = \frac{D^2 a_i^2 (1 - c_i)}{[c_i + e^{Da_i(\theta - b_i)}][1 + e^{-Da_i(\theta - b_i)}]^2},$$

根据局部独立性,假设测验信息函数可以累加,即可表示为项目信息函数之和[2]

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta).$$

1.3 曝光因子

曝光因子为

$$ecf(i) = m_i / \overline{m},$$

其中 m_i 表示项目 i 被前 $m - 1$ 个被试调用的次数,$\overline{m}$ 表示前 $m - 1$ 个被试调用题库中所有项目的平均次数,即 $\overline{m} = \sum_{i=1}^{M} \frac{m_i}{M}$,$M$ 为题库项目总数.

2 引入区分度分布因子的选题策略

考虑到当 $\ln a \sim N(0,1)$ 时题库中区分度的密度函数为 $e^{-(\ln a)^2/2} / \sqrt{2\pi}$,很明显当 $\ln a = 0$ 时,$e^{-(\ln a)^2/2} / \sqrt{2\pi}$ 取得最大值 $1/\sqrt{2\pi}$. 注意到信息量表达式中区分度 a 的指数为 2,为了能更好地减少 a 对信息量函数的影响,在区分度 a 的指数上再乘以 $e^{-(\ln a)^2/2}$,修改成 $2e^{-(\ln a)^2/2}$ 于是,在不分层的前提下,定义区分度分布因子(discrimination-distribution factor),记为 $ddf(i)$,即

$$ddf(i) = a_i^{2e^{-(\ln a_i)^2/2}}, \quad (1)$$

其中 a_i 为第 i 个项目的区分度. 为了按 a 的分布情况均匀地选取项目,即在区间长度相等的范围内,落入该区间的项目多则多取,落入的项目少则少取,故制定如下新的选题策略:

$$i_0 = \arg\max_{i \in R\alpha} f_i, \quad (2)$$

其中 $R\alpha$ 为尚未对该被试施测的项目集,也可以称为该被试的剩余题库[10-11],且

$$f_i = I_i(\hat{\theta}) / [ecf(i) \cdot ddf(i)].$$

仔细考察(1)式可以发现 a_i 的指数当 $a = 1$ 时达到最大值,此时区分度分布因子 $ddf(i)$ 的指数为 2,当 a 的取值向 1 两边变动时,$ddf(i)$ 的指数由 2 逐渐趋向于 0. 由(2)式可以看到 $ddf(i)$ 作为分母,起到的作用是削弱区分度对信息量函数的影响.

本文把新方法与文献[7-9]的方法中区分度对信息量函数影响的变化情况进行比较. 由于信息量函数比较复杂,涉及到区分度、难度、猜测度、能力等参数,而这 3 种方法除了区分度指数的不同,其余基本相同,故本文只挑出区分度来对比,即对比 $a^2/a(j,i)$($a(j,i)$ 是文献[7-9]方法中 a 的指数函数,其从 a^2 变化到 a^0,故可以简单表示成 $a^2/a(j,i) = a^x$,x 从 0 变到 2)和 $a^2/ddf(i)$,以期得出哪种方法更好的结论. 由于每次选题时,文献[7-9]的方法中区分度上的指数均会发生变化,而它们的区别仅仅在于区分度指数变化的平滑度不同,文献[8]相对于文献[7]的方法更平滑. 为了简单起见,以 0.3 为间隔,挑取 0.3,0.6,0.9,1.2,1.5,1.8 作为 a^x 的指数部分,借用 Matlab 画出区分度对信息量函数影响的变化情况,如图 1 所示.

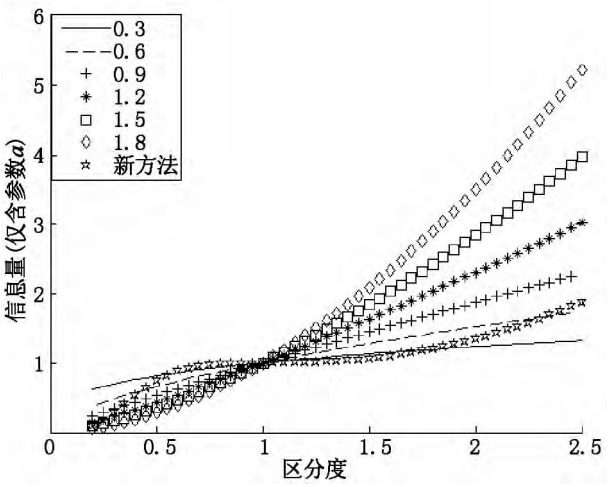

图 1 区分度对信息量函数的影响

从图 1 可以看出按照文献[7-8]的方法被试作答的项目信息量并没有按区分度分布的变化而变化,而新方法挑选的项目的信息量分布很明显可以看出在 $a = 1$ 左右曲线斜率接近于 0,变化很慢,而越往两边斜率越大,变化越快,这很好符合了区分度的分布情况. 此外当 a 达到 2.5 上限时,文献[7-8]的曲线对应的信息量已升到 5 以上,而新方法对应的信息量还不到 2. 可以看出,新方法在削弱区分度 a 对信息量的影响方面,比文献[7-8]的方法应该

更好.

3　CAT 的模拟实验

文本用 $N(\mu,\sigma^2)$ 表示均值为 μ,方差为 σ^2 的正态分布,$b \sim U(-3,3)$ 表示在区间 $(-3,3)$ 上的均匀分布,$c \sim \beta(5,17)$ 表示参数为 5,17 的 Beta 分布. 本文中所有试验均采用 Monte Carlo 模拟. 在实验过程中模拟生成 1 000 个被试,且所有被试的能力真值都服从标准正态分布,记为 $\theta \sim N(0,1)$,并模拟生成 2 种题库,每个题库均生成 1 000 个项目且都满足 $\ln a \sim N(0,1)$,$a \in [0.2,2.5]$ 和 $c \sim \beta(5,17)$,但在题库 1 中 $b \sim N(0,1)$;在题库 2 中 $b \sim U(-3,3)$.

3.1　CAT 的模拟

本文模拟定长和不定长 2 种测验: (i) 定长测验: 分为 5 层,每层 6 个题目,测验题数为 30 题; (ii) 不定长测验: 累积信息量达到 25 时结束测验[12-13].

3.2　评价指标

本文分别用能力估计的准确性(ABS)、能力估计标准差(SD) 指标来反映能力估计情况; 用人均用题数、测验效率指标反映效率; 用项目调用的均匀性和 χ^2 统计量以及测试重叠率指标来评价项目曝光率.

(i) 能力估计的准确性

$$Recovery = \frac{1}{C}\sum_{j=1}^{C}\left(\sum_{i=1}^{N}|\theta_i - \hat{\theta}_{ij}|/N\right)$$,其中 N 表示被试人数,C 表示模拟重复的次数,θ_i 为被试 i 的能力真值,$\hat{\theta}_{ij}$ 代表被试 i 在第 j 次模拟得到的能力估计值.

(ii) 能力估计标准差

$$SD = \frac{1}{N}\sum_{i=1}^{N}\left(\sqrt{\sum_{j=1}^{C}(\hat{\theta}_{ij} - \bar{\hat{\theta}}_i)^2/C}\right)$$,其中 $\bar{\hat{\theta}}_i = \sum_{j=1}^{C}\hat{\theta}_{ij}/C$,表示被试 i 在 C 次模拟中得到的能力估计平均值.

(iii) 用人均用题数

$$Nf = \frac{1}{C}\sum_{j=1}^{C}\left(\sum_{i=1}^{N} r_{ij}/N\right)$$,其中 r_{ij} 表示被试 i 在第 j 次模拟中的作答项目数.

(iv) 测验效率

测验效率 $= \sum_{i=1}^{N}\inf_i / \sum_{i=1}^{N} L_i$,其中 $\inf_i$ 为被试 i 测验的信息总量,L_i 为被试 i 的测试长度.

(v) 用项目调用的均匀性

$$SE = \frac{1}{C}\sum_{j=1}^{C}\left(\sqrt{\sum_{i=1}^{M}(m_{ij} - \bar{m}_j)^2/M}\right)$$,其中 M 为题库中项目总数,m_{ij} 表示项目 i 在第 j 次模拟中被调用的次数,$\bar{m} = \sum_{i=1}^{M}\frac{m_{ij}}{M}$ 表示在第 j 次模拟中项目被调用的平均次数.

(vi) χ^2 检验统计量

$$\chi^2 = \sum_{j=1}^{M}\left\{[A_j - (\sum_{j=1}^{M}A_j/M)]^2/(\sum_{j=1}^{M}A_j/M)\right\}$$,其中 A_j 是第 j 题曝光率,A_j 的计算公式为 A_j = 第 j 题被使用的次数 /N.

(vii) 测试重叠率

$$Rt = 2TO_{总} / [(N-1)\sum_{i=1}^{N} L_i]$$,其中 $TO_{总}$ 是考生的试题重叠总数,计算方法为 $TO_{总} = \sum_{j=1}^{M} C_{M_j}^2$,$M_j$ 是试题库中第 j 题使用次数.

对于上述 7 个指标,除测验效率是越大越好之外,其它 6 个指标都是越小越好.

4　实验结果与分析

本文模拟 3 种实验方法与新的方法进行比较,MIC 是在按 a 分层出现前比较经典的选题策略,文献[7-8]的方法是基于按 a 分层的思想,是目前业内综合各项指标都较好的选题策略,而新的方法与这 3 种方法进行对比,能较好地反映出此方法的优劣,实验结果见表 1 ~ 表 4.

表 1　不定长测验 $\ln a \sim N(0,1)$,$b \sim U(-3,3)$,$c \sim \beta(5,17)$ 时实验结果

策略	ABS	SD	项目调用均匀性	人均用题数	测验效率	卡方统计量	测试重叠率
MIC	0.161 6	0.184 9	66.265 0	19.298 7	1.338 7	269.413 1	0.240 7
文献[7]方法	0.227 1	0.256 6	16.359 1	29.801 6	0.517 7	9.985 5	0.032 2
文献[8]方法	0.226 8	0.254 3	13.804 9	29.752 4	0.505 6	7.123 9	0.029 6
新策略	0.210 8	0.187 4	16.569 9	29.652 3	0.567 4	10.301 7	0.032 3

表 2 不定长测验 $\ln a \sim N(0,1)$, $b \sim N(0,1)$, $c \sim \beta(5,17)$ 时实验结果

策略	*ABS*	*SD*	项目调用均匀性	人均用题数	测验效率	卡方统计量	测试重叠率
MIC	0.238 5	0.228 8	77.078 9	23.336 5	0.720 1	292.148 1	0.271 7
文献[7]方法	0.329 6	0.285 3	12.929 3	28.874 5	0.418 6	6.460 8	0.028 1
文献[8]方法	0.341 6	0.288 8	14.037 6	29.286 2	0.369 6	7.496 7	0.029 5
新策略	0.317 3	0.269 5	13.583 5	28.524 2	0.424 9	7.229 2	0.028 4

表 3 定长测验 $\ln a \sim N(0,1)$, $b \sim U(-3,3)$, $c \sim \beta(5,17)$ 时实验结果

策略	*ABS*	*SD*	项目调用均匀性	测验效率	卡方统计量	测试重叠率
MIC	0.132 5	0.143 0	78.369 7	1.279 6	245.676 1	0.225 0
文献[7]方法	0.196 6	0.222 8	21.150 5	0.616 8	17.894 1	0.034 9
文献[8]方法	0.186 0	0.210 5	22.965 9	0.694 9	19.534 7	0.041 0
新策略	0.219 6	0.231 4	16.026 5	0.527 5	9.513 1	0.032 0

表 4 定长测验 $\ln a \sim N(0,1)$, $b \sim N(0,1)$, $c \sim \beta(5,17)$ 时实验结果

策略	*ABS*	*SD*	项目调用均匀性	测验效率	卡方统计量	测试重叠率
MIC	0.243 4	0.214 9	82.876 0	0.880 2	274.748 1	0.249 2
文献[7]方法	0.302 9	0.266 3	14.111 1	0.405 8	7.965 5	0.026 7
文献[8]方法	0.286 6	0.262 0	14.818 9	0.480 5	8.133 8	0.030 8
新策略	0.313 0	0.268 9	13.538 4	0.420 2	6.788 7	0.029 5

在测验为不定长情况下,从表 1 和表 2 可以看出,新方法在能力估计的准确性(*ABS*)、能力估计标准差(*SD*)、人均用题量和测验效率这 4 项指标上要优于文献[7-8]的方法. 在项目均匀性,卡方和重叠率这 3 项指标上新选题策略要明显优于 MIC 方法,与文献[7]的方法相接近,但相对于文献[8]的方法还有一定的差距. 由于 MIC 方法测验效率较高,能力估计准确,但其为了选取信息量大的题目,常常会选择区分度大的项目,带来的结果就是区分度高的项目的曝光度大大提高,区分度小的项目很少甚至不被调用,使得均匀性,卡方和重叠率指标数值很高,在日常测验中为了兼顾试题的安全性一般不会采用 MIC 方法.

在测验为定长情况下,从表 3 中可以看出新方法在项目调用均匀性、卡方统计量和测试重叠率上都要优于文献[7-8]的方法,但在能力估计的准确性(*ABS*)、能力估计标准差(*SD*)和测验效率方面相对稍差. 从表 4 中可以看出项目调用均匀性和卡方统计量要优于文献[7-8]的方法,在能力估计的准确性(*ABS*)、能力估计标准差(*SD*)和文献[7]的比较接近,在测验效率和测试重叠率上 3 种方法各有优劣.

综合来看,在不定长情况下,既要兼顾曝光度,又要有较高的测验精度和效率,新方法不失为一种较理想的选题策略. 在定长情况下,如要较好控制曝光度又要有较准确的测验精度,新方法将是一种较好的选题策略.

5 讨论

本文对于 3PLM 在 MIC 的基础上引入了区分度分布因子是按照 a 的分布情况进行了研究,不同程度地减少区分度 a 对信息量函数的影响,使得 a 的密度较大处,就让其多抽取一些项目,密度较小处则少抽取一些. 从以上 4 张表中的各项数据看,新选题策略并不是每一项指标都优于文献[7-8]的方法,故还有继续改进的空间. 由于区分度的分布情况不仅仅只有 $\ln a \sim N(0,1)$,还可能服从其他分布,能否以本文为参考,找出针对不同分布情况的区分度分布因子?此外,如果 $\ln a \sim N(0,1)$,但 a 的取值范围可以有不同的选取范围,比如 a 可以在区间 $[0.2, 2.5]$ 或者在区间 $[0.5,2]$ 中均匀分布等,这时区分度分布因子是否也应做出适当的修改以期达到最好的测验效果?以上几个问题都值得进一步探讨研究.

6 参考文献

[1] Weiss D J. New horizons in testing-latent trait test theory and computerized adaptive testing [M]. New York: Aca-

demic Press,1983: 237-254.

[2] 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.

[3] 汪文义,丁树良. 2PLM 下 CAT 选题策略比较[J]. 考试研究,2009,5(3): 60-70.

[4] 张华华,程莹. 计算机化自适应测验(CAT)的发展和前景展望[J]. 考试研究,2005,1(1): 12-24.

[5] Chang Huahua, Ying Zhiliang. A global information approach to computerized adaptive testing[J]. Applied psychological Measurement,1996,20(2): 213-219.

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

[9] 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.

[10] 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

[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.

[12] 罗照盛,欧阳雪莲,漆书青,等. 项目反应理论等级反应模型项目信息量[J]. 心理学报,,2008,40(11): 1212-1220.

[13] 游晓锋,丁树良,刘红云. 计算机化自适应测验中原始题项目参数的估计[J]. 心理学报,2010,42(7): 813-820.

The New Item Selection Strategy for Computerized Adaptive Testing Based on Sampling Principle

ZHANG Hu-chao, DING Shu-liang*, DAI Xie, GUAN Chao-hui

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

Abstract: The exposure-control factors has been followed, at the same time, a discrimination-distribution factor has been introduced. According to the distribution of discrimination, the new item selectisn srategy to balance item usage exposure rate has been proposed. Suppose that $\ln a \sim N(0,1)$ and according to Monte Carlo simulation, the results show that the new approach has more preferably performance comparing with other approaches on several assessment indexes.

Key words: sampling principle; computerized adaptive testing; distribution factor

(责任编辑: 冉小晓)