文章编号:1000-5862(2013)06-0652-05

各层分布近似的计算机化自适应测验分层选题策略

章沪超,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:选题策略是计算机化自适应测验重要的组成部分,其好坏直接关系到测验的准确性、安全性、效率和测验信度,而分层法又是其中极其重要的一种方法.针对在分层法中按区分度分层(a-STR)和按最大信息量分层(MIS)的曝光率依然较大的缺点,提出了按区分度近似分布分层法(A-SDS)和按最大信息量近似分布分层法(MI-SDS)2 种方法.通过 Matlab 模拟实验表明:在测验精度和效率与原方法接近的情况下,新方法比 a-STR 和 MIS 方法较明显地降低了项目的曝光率.

关键词:计算机化自适应测验;选题策略;按区分度近似分布分层法;按最大信息量近似分布分层法
中图分类号:B 841.7;TP 301.6 **文献标志码**:A

1 研究背景

计算机化自适应测验(CAT)与传统纸笔测验截然不同.在传统的纸笔测验模式下,每个人作答相同的一套试题,而 CAT 给每一位被试一份量身定制的试卷.具体的做法是依据被试的答题情况,从题库中连续选取最接近被试实际水平的项目让其作答,使得被试不必作答过多与自身能力相差较远的题目.因此,CAT 通过较少的考题就能对被试水平做出更为有效的测量,这是它较之传统纸笔测验的最大优势.

CAT 包括 6 个基本组成部分:所采用项目的反应模式、题库、初始项目的选择、选题策略、特质估计方法和测验终止规则[1].其中选题策略是非常重要的一个环节,关系到测验准确性、安全性、效率和信度.

CAT 选题不仅仅要考虑统计优化问题,非统计约束条件也同样重要.统计优化主要指根据被试的反应,选择最适合其作答的项目以提高被试能力估计的精度.非统计约束包括项目曝光度控制、测验中各个内容域的恰当比例、正确答案选项的平衡分布、项目的长短适当、被试反应时间均衡等[2].由于统计优化是所有 CAT 首要目标,因此按 CAT 是否要求非统计约束条件,将选题策略划分为提高测量准确性的选题策略,如最大 Fisher 信息量(MFI)方法[3]、最小期望后验标准差(MEPSD)方法[4]、最大全局信息量(MGI)方法[5]等;以及具有非统计约束的选题策略[6],主要包括以下几类:(ⅰ)随机方法.这种方法的题库利用率较高,项目曝光率近似均匀分布,但是能力估计的准确性很差;(ⅱ)分层方法.由于 MFI 方法趋向于选取区分度(a)较大的项目对被试作答.为解决这个问题,Chang Huahua 等[7]提出了按 a 分层方法(a-STR),该方法提高了低区分度项目的使用率,却不能明显降低项目曝光率[8].M. S. Wingersky 等[8]发现项目区分度和难度(b)通常存在正相关.鉴于此,Chang Huahua 等[9]提出了按 b 分块按 a 分层的方法(AS-B),该方法相比 a-STR 较好地平衡了项目的曝光率,同时还提高了测量的准确性.考虑到猜测参数(c)在 3 参数逻辑斯蒂克模型(3PLM)中的作用,Juan Ramón Barrada 等[10]提出了最大信息量分层(MIS)方法.(ⅲ)曝光率参数控制法.如程小扬等[11]提出了引入曝光因子(ecf)的方法等.

2 新的选题策略

针对分层方法的不足,本文提出了 2 种新的选

收稿日期:2013-08-17
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)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: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、应用及教育和心理测量方面的研究.

题策略,按区分度近似分布分层选题策略(A-SDS)与按最大信息量近似分布分层选题策略(MI-SDS).

2.1　按区分度近似分布分层选题策略(A-SDS)

分层步骤如下:(ⅰ)将项目按区分度非递减排序;(ⅱ)从题库的第 1 个项目开始每间隔 n 选取一个项目组成一层,以此类推组成 n 层.

每层中选取的项目 i_0 满足条件

$$i_0 = \arg\min_{j \in R_{\alpha h}} |\hat{\theta} - b_j|, \quad (1)$$

其中 $\hat{\theta}$ 为被试的能力估计值,$R_{\alpha h}$ 表示被试 α 在当前第 h 层尚未作答的项目集.

2.2　按最大信息量近似分布分层选题策略(MI-SDS)

对于 3PLM:

$$P_i(\theta) = c_i + (1 - c_i) / (1 + e^{-Da_i(\theta - b_i)}), \quad (2)$$

其项目信息函数为

$$I_i(\theta) = \frac{D^2 a_i^2 (1 - c_i)}{[c_i + e^{Da_i(\theta - b_i)}][1 + e^{-Da_i(\theta - b_i)}]^2}. \quad (3)$$

当项目信息函数值到达最大值时,被试水平值为 $\theta_{\max}$ 为

$$\theta_{\max} = b_i + (\ln[1 + \sqrt{1 + 8c_i}] - \ln 2) / (1.7a_i), \quad (4)$$

$\theta_{\max}$ 对应的最大信息量为 $I_i(\theta)_{\max}$ 为

$$I_i(\theta)_{\max} = D^2 a_i [1 - 20c_i - 8c_i^2 + (1 + 8c_i)^{3/2}] / (8(1 - c_i)^2). \quad (5)$$

分层步骤与 A-SDS 方法类似,只要将区分度换成(5)式的最大信息量即可.

每层中选取的项目 i_0 满足条件

$$i_0 = \arg\min_{j \in R_{\alpha h}} |\hat{\theta} - \theta_{\max}|. \quad (6)$$

由本文 2.1 和 2.2 节的 2 种分层方法可以使得各层的区分度(最大信息量)服从分布的近似,故做相应的命名.

3　CAT 的模拟过程

3.1　被试及题库模拟

由于被试能力真值未知,项目参数的真值也未知,所以要评价一种新的 CAT 选题策略对能力估计的影响目前只能用 Monte Carlo 模拟实验. 本文用 Monte Carlo 模拟方法进行试验,模拟的数据结构[12]如下:(ⅰ)用 $N(\mu, \sigma^2)$ 表示均值为 μ,方差为 σ^2 的正态分布,$b \sim U(a, b)$ 表示在区间(a, b)上的均匀分布,$c \sim \beta(a, b)$ 表示参数为 a,b 的 Beta 分布. 按项目参数分布的不同模拟产生 4 个题库,每个题库题量为 1 000 并分成 4 层;(ⅱ)模拟生成 1 000 个被试,且被试的能力真值 $\theta \sim N(0, 1)$.

3.2　施测过程

变长和定长是测验的 2 种形式. 对于变长测验而言,各层终止规则满足 $I_k = I_{总}(k/T)^2$,$I_{总}$ 为总信息量固定为 16(即测验标准误为 0.25),T 为总层数(本实验分 4 层),k 为当前所在的层数,各层不设最大测验长度. 对于定长测验,每层测验 10 题,总测验长度为 40.

随机选 3 个项目给被试作答以计算其能力初始值[11],此后进入精估阶段,由(1)式或(6)式选取项目作答,确定被试在该项目上的得分,再用贝叶斯后验期望法(EAP)估计被试的能力. 重复该步骤,直到满足测验结束条件.

测验过程通过以下方法模拟被试得分: 根据被试能力真值 θ 和当前所选择的项目 i 的参数,由(2)式计算被试的在第 i 个项目上的答对概率$P_i(\theta)$,再产生一个随机数$r(0 \leqslant r \leqslant 1)$,若$r \leqslant P_i(\theta)$,则该被试在第 i 个项目上得 1 分,否则得 0 分.

3.3　评价指标

本文采用能力估计的准确性(*ABS*)和能力估计标准差(*SD*)这 2 项指标来评价能力估计情况; 用人均用题数和测验效率这 2 项指标来评价效率; 用项目调用的均匀性、χ^2 检验统计量和测试重叠率这 3 项指标来评价项目曝光率[12-13].

4　实验结果与分析

测验分为定长和变长 2 种测验形式,每种形式都分为加和不加曝光因子,实验过程重复 30 次,题库的生成和被试的模拟均由 Matlab 实现. 在表 1 ~ 表 4 中,+ecf 表示使用曝光因子.

由表 1 ~ 表 4 可知,对于变长测验,不管是否使用曝光因子,在人均用题数、测验效率、均匀性、卡方和重叠率这 5 项指标上 A-SDS 和 MI-SDS 都要好过同等条件下的 *a*-STR 和 MIS,特别是在表 1 和表 2 中优势尤为明显; 在表 1 中未使用曝光因子的 A-SDS 和 MI-SDS 在这 5 项指标上也要明显好过使用曝光因子的 *a*-STR 和 MIS; 在表 2 中未使用曝光因子的 A-SDS 和 MI-SDS 在人均用题数、测验效率、均匀性和重叠率这 4 项指标上也要好过使用曝光因子的 *a*-STR 和 MIS; 在表 3 中未使用曝光因子的

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

	策略	*ABS*	*SD*	人均用题数	测验效率	项目均匀性	卡方	测试重叠率
变长	*a*-STR	0.195 1	0.234 4	48.835 8	0.348 2	81.589 5	128.430 2	0.190 5
	MIS	0.195 2	0.235 0	49.434 7	0.344 7	78.323 8	117.003 0	0.178 8
	A-SDS	0.198 7	0.238 4	34.939 0	0.481 1	49.207 6	63.828 2	0.109 5
	MI-SDS	0.198 6	0.238 9	33.469 9	0.498 7	47.606 9	62.148 1	0.106 5
	a-STR + ecf	0.195 7	0.234 4	47.838 0	0.354 9	67.857 5	90.581 3	0.149 4
	MIS + ecf	0.194 4	0.233 1	47.294 4	0.357 6	68.136 3	92.313 6	0.150 7
	A-SDS + ecf	0.200 4	0.242 6	34.281 8	0.487 2	37.448 4	37.618 5	0.080 4
	MI-SDS + ecf	0.196 0	0.235 1	34.033 8	0.490 9	38.813 0	40.681 0	0.083 6
定长	*a*-STR	0.171 3	0.207 8	40.000 0	0.576 1	61.129 5	86.908 4	0.138 7
	MIS	0.1716	0.207 3	40.000 0	0.563 8	60.979 3	86.481 3	0.138 3
	A-SDS	0.186 2	0.225 3	40.000 0	0.496 9	51.450 4	61.564 9	0.111 4
	MI-SDS	0.181 5	0.219 1	40.000 0	0.507 3	49.705 5	57.460 1	0.107 0
	a-STR + ecf	0.175 8	0.211 4	40.000 0	0.566 0	50.750 7	59.902 0	0.109 7
	MIS + ecf	0.167 4	0.200 5	40.000 0	0.605 0	49.260 1	56.436 3	0.105 9
	A-SDS + ecf	0.187 5	0.224 4	40.000 0	0.502 1	41.644 3	40.334 7	0.088 6
	MI-SDS + ecf	0.185 8	0.223 7	40.000 0	0.502 3	42.928 3	42.858 6	0.091 3

表 2 $\ln a \sim N(0,1)$, $b \sim N(0,1)$, $c \sim \beta(5,17)$ 的实验结果

	策略	*ABS*	*SD*	人均用题数	测验效率	项目均匀性	卡方	测试重叠率
变长	*a*-STR	0.196 6	0.236 2	51.517 0	0.330 1	51.621 6	50.032 3	0.105 9
	MIS	0.198 2	0.237 7	48.572 3	0.349 8	51.085 2	51.926 7	0.104 9
	A-SDS	0.198 3	0.239 8	36.447 1	0.462 2	35.822 3	33.875 8	0.073 7
	MI-SDS	0.198 4	0.237 6	37.519 2	0.445 1	36.696 5	34.255 9	0.076 1
	a-STR + ecf	0.196 7	0.235 4	54.170 6	0.314 9	37.598 3	25.199 8	0.083 3
	MIS + ecf	0.195 1	0.233 8	45.778 4	0.370 2	37.798 0	29.723 9	0.080 7
	A-SDS + ecf	0.198 1	0.240 2	38.303 2	0.439 6	25.690 3	17.015 9	0.055 6
	MI-SDS + ecf	0.200 2	0.241 1	38.767 2	0.434 2	26.085 6	17.180 7	0.057 1
定长	*a*-STR	0.166 9	0.202 7	40.000 0	0.601 4	41.687 7	40.417 1	0.088 7
	MIS	0.171 2	0.207 0	40.000 0	0.572 8	42.251 2	41.517 7	0.089 9
	A-SDS	0.187 3	0.227 0	40.000 0	0.478 5	35.164 3	28.785 0	0.076 1
	MI-SDS	0.188 3	0.227 0	40.000 0	0.479 7	36.018 2	30.171 7	0.077 7
	a-STR + ecf	0.172 9	0.209 9	40.000 0	0.556 5	28.879 1	19.398 7	0.066 1
	MIS + ecf	0.170 0	0.204 5	40.000 0	0.572 1	30.093 6	21.063 0	0.067 9
	A-SDS + ecf	0.189 4	0.229 0	40.000 0	0.470 8	20.085 1	9.383 7	0.055 3
	MI-SDS + ecf	0.188 2	0.227 2	40.000 0	0.488 1	22.451 5	11.724 1	0.057 8

表 3 $a \sim U(0.2,2.5)$, $b \sim U(-3,3)$, $c \sim \beta(5,17)$ 的实验结果

	策略	*ABS*	*SD*	人均用题数	测验效率	项目均匀性	卡方	测试重叠率
变长	*a*-STR	0.197 9	0.240 7	24.062 8	0.720 9	46.390 5	79.526 4	0.118 9
	MIS	0.197 9	0.238 9	26.390 2	0.658 1	49.810 6	84.422 9	0.125 8
	A-SDS	0.201 0	0.242 4	20.640 2	0.821 6	37.902 2	60.772 6	0.095 6
	MI-SDS	0.199 6	0.240 8	20.369 8	0.836 1	36.102 7	55.775 5	0.089 7
	a-STR + ecf	0.199 1	0.241 6	25.735 7	0.676 7	40.435 3	56.904 0	0.094 6
	MIS + ecf	0.196 9	0.237 7	26.634 8	0.650 1	41.390 7	57.813 5	0.096 3
	A-SDS + ecf	0.199 8	0.242 2	22.113 9	0.771 3	26.481 1	27.924 9	0.059 2
	MI-SDS + ecf	0.198 2	0.239 6	21.355 5	0.794 5	24.975 7	25.613 4	0.055 9
定长	*a*-STR	0.138 4	0.168 5	40.000 0	0.899 1	54.568 6	69.252 5	0.119 7
	MIS	0.135 5	0.165 3	40.000 0	0.923 4	56.256 4	73.603 2	0.124 4
	A-SDS	0.143 9	0.176 5	40.000 0	0.844 5	46.985 9	51.346 0	0.100 4
	MI-SDS	0.136 5	0.166 0	40.000 0	0.930 4	46.547 0	50.388 7	0.099 4
	a-STR + ecf	0.141 3	0.171 8	40.000 0	0.907 6	46.298 0	49.851 5	0.098 8
	MIS + ecf	0.136 4	0.164 7	40.000 0	0.946 4	45.482 1	48.111 2	0.097 0
	A-SDS + ecf	0.144 6	0.175 2	40.000 0	0.844 3	39.722 7	36.697 6	0.084 7
	MI-SDS + ecf	0.143 1	0.172 3	40.000 0	0.867 8	39.200 8	35.740 9	0.083 7

表 4　$a \sim U(0.2, 2.5)$,$b \sim N(0, 1)$,$c \sim \beta(5, 17)$ 的实验结果

	策略	*ABS*	*SD*	人均用题数	测验效率	项目均匀性	卡方	测试重叠率
变长	*a*-STR	0.197 9	0.241 0	28.650 1	0.606 2	34.953 9	39.442 5	0.075 2
	MIS	0.198 3	0.240 7	30.230 8	0.584 4	33.912 7	37.716 0	0.068 1
	A-SDS	0.200 7	0.244 0	24.181 7	0.703 3	30.771 7	36.027 8	0.066 8
	MI-SDS	0.203 3	0.244 9	27.579 6	0.631 2	30.908 8	35.720 0	0.059 8
	a-STR + ecf	0.199 1	0.240 8	27.549 7	0.631 6	20.790 9	14.402 5	0.047 4
	MIS + ecf	0.201 1	0.242 2	29.119 1	0.601 3	21.259 9	14.988 6	0.045 8
	A-SDS + ecf	0.200 7	0.243 6	24.061 0	0.706 3	17.654 6	11.977 7	0.040 1
	MI-SDS + ecf	0.205 7	0.249 0	31.217 9	0.559 6	18.968 0	12.447 7	0.037 5
定长	*a*-STR	0.137 6	0.168 2	40.000 0	0.901 6	37.763 4	33.166 5	0.080 9
	MIS	0.132 5	0.161 7	40.000 0	0.957 7	39.337 7	35.989 7	0.083 9
	A-SDS	0.141 8	0.174 8	40.000 0	0.863 3	33.268 5	25.741 6	0.072 9
	MI-SDS	0.142 1	0.172 5	40.000 0	0.878 2	33.168 2	25.585 9	0.072 7
	a-STR + ecf	0.137 9	0.168 5	40.000 0	0.904 8	25.476 9	15.098 1	0.061 4
	MIS + ecf	0.135 7	0.164 7	40.000 0	0.943 9	25.324 0	14.916 2	0.061 2
	A-SDS + ecf	0.145 0	0.177 3	40.000 0	0.839 6	18.962 7	8.364 5	0.054 2
	MI-SDS + ecf	0.142 8	0.174 4	40.000 0	0.832 5	18.505 3	7.965 1	0.053 8

A-SDS 和 MI-SDS 在人均用题数,测验效率,均匀性这 3 项指标上也要比使用曝光因子的 *a*-STR 和 MIS 好.

对于定长测验,不管是否使用曝光因子,2 种新方法在测验精度和策略效率与 *a*-STR 和 MIS 稍差一点的情况下,均匀性、卡方和重叠率这 3 项指标上 A-SDS 和 MI-SDS 都要明显好过同等条件下的 *a*-STR 和 MIS; 在表 1 和表 3 中未使用曝光因子的 A-SDS 和 MI-SDS 在在均匀性、卡方和重叠了这 3 项指标上和使用曝光因子的 *a*-STR 和 MIS 几乎相当.

总结表 1 ~ 表 4 可以看出相同条件下在测验精度和效率与 *a*-STR 和 MIS 方法接近的情况下 A-SDS 和 MI-SDS 方法较明显的降低了项目的曝光率. 此外,在变长测验时人均用题数、测验效率这 2 项指标也有不同程度的改进. 所以在以后的 CAT 测验中,A-SDS 和 MI-SDS 可以作为一种比 *a*-STR 和 MIS 更安全经济的选题策略来进行施测.

5　小结与展望

这 2 种选题策略有 3 个特点:(ⅰ)分层简单,新策略只需对区分度或最大信息量分层即可;(ⅱ)使用范围广,适合于不同的逻辑斯蒂克模型;(ⅲ)在测验精度和效率与 *a*-STR 和 MIS 方法接近的情况下更好地平衡了项目的曝光率.

本文的方法仅仅是按区分度和最大信息量分层,没有考虑区分度和难度是否相关,如果它们相关则该如何改进? 此方法能否应用到多级评分中,以期达到理想的效果? 题库划分为 4 层是不是最优,如果不是,划分为几层最为合适? 题库按不均衡分层到达的效果能否比均衡分层更为理想? 以上这几点都有待于做进一步探讨.

6　参考文献

[1] Weiss D J, Kingsbury G G. Application of computerized adaptive testing to educational problems [J]. Journal of Educational Measurement, 1984, 21(4): 361-375.

[2] Wim J van der Linden, Cees A W Glas. Computerized adaptive testing: theory and practice [M]. Dordrecht: Kluwer Academic Publishers, 2000: 27-52.

[3] Lord F M. A broad-range tailored test of verbal ability [J]. Applied Psychological Measurement, 1977(1): 95-100.

[4] Owen R J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing [J]. Journal of the American Statistical Association, 1975, 70(350): 351-356.

[5] Chang Huahua, Ying Zhiliang. A global information approach to computerized adaptive testing [J]. Applied Psychological Measurement, 1996, 20(3): 213-229.

[6] 毛秀珍,辛涛. 计算机化自适应测验选题策略述评 [J]. 心理科学进展, 2001, 19(10): 1552-1562

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

[8] Wingersky M S, Lord F M. An investigation of methods for reducing sampling error in certain IRT procedures [J]. Applied Psychological Measurement, 1984, 8(3): 347-364.

[9] Chang Huahua, Qian Jianhe, Ying Zhiliang. a-stratified multistage computerized adaptive testing with *b* blocking [J]. Applied Psychological Measurement, 2001, 25(4): 333-341.

[10] 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.
[11] 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.
[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] 汤楠,丁树良,余丹. 结合优先级指标和曝光因子的多级评分选题策略 [J]. 江西师范大学学报: 自然科学版, 2011, 35(6): 646-650.

Similar Distributed Stratification Method for Computerized Adaptive Testing

ZHANG Hu-chao, DING Shu-liang*
(College of Computer Information Engineer, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: Item selection method is a key part of computerized adaptive testing, it will have a direct impact on accuracy, safety, efficiency and reliability of the testing. Besides, stratified method is a most important part of item selection method. For the relatively high item exposure rate shortcoming against a-stratified(a-STR) and maximum information stratification(MIS) method, two kinds of methods are proposed: similar distributions of discrimination parameters/maximum information stratification(A-SDS/MI-SDS). The results of the Matlab simulation study show that the new item selection methods obtain lower average exposure rates than a-STR and MIS methods while maintaining the approximate accuracy and efficiency of testing.

Key words: computerized adaptive testing; item selection method; similar distributions of discrimination parameters stratification; similar distributions of maximum information stratification

(责任编辑: 冉小晓)

(上接第 651 页)

A Hybrid Simplex Search and Particle Swarm Optimization Algorithm with Immune Evolutionary

MIAO Chen, LIU Guo-zhi
(Department of Fundamental Teaching, Yingkou Institute of Technology, Yingkou Liaoning 115014, China)

Abstract: The hybrid NM-IEPSO algorithm is proposed based on the Nelder-mead(NM) simplex search method and particle swarm optimization algorithm with immune evolutionary(IEPSO) for unstrained optimization. NM-IEPSO is very easy to implement in practice since it does not require gradient computation, and intends to produce faster and more accurate convergence. The main propose is to demonstrate how the IEPSO can be improved by incorporating a hybridization strategy. In a suit of 6 test function problems taken from the literature, computational results, show that the hybrid NM-IEPSO approach outperforms five relevant search techniques(i.e., IEPSO, PSOPC, GSPSO, LSPSO and CPSO) in terms of solution quality and convergence rate. In a later part of the comparative experiment, the NM-IEPSO algorithm is compared to three hybrid algorithms procedures appearing in the literature. The comparison report still largely favors the NM-IEPSO algorithm in the performance of accuracy, robustness and function evaluation. As evidenced by the overall assessment based on two kinds of computational experience, the new algorithm has demonstrated to be extremely effective and efficient at locating best-practice optimal solutions for unstrained optimization.

Key words: simplex search method; particle swarm optimization; unstrained optimization; immune evolutionary

(责任编辑: 曾剑锋)