文章编号:1000-5862(2014)05-0445-04

CAT 分层终止规则探究

胡　姗,丁树良*,程　艳,熊建华
(江西师范大学计算机信息工程学院,江西 南昌　330022)

摘要:分层的不定长 CAT 各分层信息量的控制界点设置问题,到目前为止的研究还是经验性的探索,对此问题提出了新终止规则解决方案. 实验结果表明:新终止规则对于提高题库的安全性及测验效率均有较好的效果.

关键词:计算机自适应测验;终止规则;不定长;分层
中图分类号:B 841.7;TP 301.6　　**文献标志码**:A

0　引言

计算机自适应测验(computerized adaptive test,CAT)是应用项目反应理论(item response theory,IRT)建立题库,并由计算机根据被试能力水平自动选择测试题,最终对被试能力做出估计的一种新型测验方法. 它最大的特点是"量体裁衣",即根据被试的能力从题库中选择难度与能力相匹配的试题施测,获得学生的最大信息,保证能力高的学生不会做到太容易的题,能力低的学生不会做到太难的题. CAT 的实施必须解决以下 4 个方面的问题:(i)题库建设;(ii)选题策略;(iii)参数估计;(iv)测试终止规则. 目前 CAT 常用的选题策略是 Lord 提出的最大信息量选题法,实施这种选题策略会使区分高的项目曝光率过高,而区分度过低的项目则被搁置或极少使用. 这种方法在测验效率方面虽然优点突出(只需较少量题即可测出能力),但在题库安全性方面的缺陷也比较明显. 针对 Lord 选题策略的不足,研究者们提出了按 a 分层法[1]、按 b 分块按 a 层法[2]和按 c 分层法[3]等方法,相关文献均称这些选题策略能较好的增强 CAT 的安全性.

CAT 的终止规则分为定长和不定长. 定长是当施测项目数累加到预设值时即终止测试,这样就违背了自适应的初衷. 不定长则是按照测量标准误差落入预设范围内即终止测试. 由于不同项目所含信息量不同,因此能力不同的被试完成测验所需施测的项目及项目数也有所不同,于是测验的长度就会随着被试的变化而变化,从而更好地体现出了 CAT "因人施测"这个特点,因此不定长 CAT 终止规则得到了研究者们的推崇和青睐.

采用按 a 分层法、按 b 分块按 a 层法和按 c 分层法等方法选题策略后又产生了一个新的问题:分层终止规则如何制定.

要解决各层信息量分配问题,先考察 CAT 中测验信息量的计算方法,用单维 3 参数 Logistic 模型(3PLM),Fisher 信息量相当于抽样标准误平方(即方差)的倒数[4],令 $K_j(\theta) = \{c_j + \exp[Da_j(\theta - b_j)] \cdot \{1 + \exp[-Da_j(\theta - b_j)]\}^2$,则

$$I_j(\theta) = (1 - C_j) D^2 a_j^2 / K_j(\theta) \qquad (1)$$

其中 D 为量表因子,通常取为 1.7;a_j、b_j、c_j 分别为 3PLM 中题目 j 的区分度参数、难度参数和猜测参数;θ 为被试在 CAT 测验中的当前的估计能力;$I_j(\theta)$ 即能力为 θ 的被试在题目 j 上具有的信息量.

目前涉及分层终止规则的分配规则不多见,主要有比较(1:1:1:1)、(1:2:3:4)和(4:3:2:1)3 种信息量分层比例,研究发现(1:2:3:4)较好[5],王茜娟等采用此方法对按 c-分层不定长 CAT 做出了研究[1]. 有人认为各层信息量之比为 $I_1 : I_2 : \cdots : I_k = 1^2 : 2^2 : \cdots : k^2$ 的效果较理想[6-7]. 朱隆尹等[8]给出了

收稿日期:2014-05-16
基金项目:国家自然科学基金(30860084,31160203,31100756),国家社会科学基金教育学青年课题"教育虚拟社区的群集智能化构建方法研究"(CCA110109)和江西省教育厅科技计划(GJJ13207,GJJ13206,GJJ13227,GJJ133208,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,主要从事应用统计及计算机辅助教学的研究.

3PLM 下按 a 分层不定长 CAT 终止规则的 2 个新方案:

$$I_1 : I_2 : \cdots : I_k = 1^2(1-\bar{c}_1) : 2^2(1-\bar{c}_2) : \cdots : k^2(1-\bar{c}_k), \quad (2)$$

$$I_1 : I_2 : \cdots : I_k = 1^2(1-\bar{c}_1)^2 : 2^2(1-\bar{c}_2)^2 : \cdots : k^2(1-\bar{c}_k)^2, \quad (3)$$

其中 $\bar{c}_k$ 表示第 k 层猜测度 c 的平均值. 当 c 等于 0 时,(2) 式(文献[8]中方案 1) 和(3) 式(文献[8]中方案 2) 便等同于文献[5]方案. 显然,文献[8]中是文献[5-6]中在 3PLM 的推广. 然而,有没有比上述更好的方案?

1 新的分层终止规则

张华华等[9]对各层信息量的比例做出讨论,分别为平均、递增、递减,并认为递增的方式较好. 通过对上述终止规则的比较及实验,发现递增的分层确实相对效果更好,而且做实验得出 1∶3∶5∶7 的分层的方式比 1∶2∶3∶4 的分层效果好,文献[5-6]中其实也是扩大分层之间的比例,文献[8]中也是对比例进行调节. 但是并不是越大越好,通过对不同比例的方案进行大量对比试验,得出下面 2 种 k 层信息量分配新方案表现效果更好,具体公式为

$$I_1 : I_2 : \cdots : I_k = 1^2(1-\bar{c}_1) : 3^2(1-\bar{c}_2) : \cdots : (2k-1)^2(1-\bar{c}_k), \quad (4)$$

$$I_1 : I_2 : \cdots : I_k = 1^2(1-\bar{c}_1)^2 : 3^2(1-\bar{c}_2)^2 : \cdots : (2k-1)^2(1-\bar{c}_k)^2. \quad (5)$$

本文把新的终止规则应用到不同分层中,检验实施的可行性.

2 实验模拟

2.1 模拟生成被试

模拟生成一批随机数,其数量为 N,数值均服从标准正态分布,记为: $\theta \sim N(0,1)$,其中 N 为被试总人数,本文均设定 $N = 1\,000$; θ 为被试的能力真值.

2.2 模拟生成题库

用 a、b、c,分别表示 3PLM 中的区分度参数、难度参数和猜测度参数. 若 a 服从对数正态分布,且 $0.2 \leqslant a \leqslant 2.5$,记为 $\ln a \sim N(0,1) \wedge a \in [0.2, 2.5]$; a 服从 0.2 到 2.5 的均匀分布,记为 $a \sim U(0.2, 2.5)$; b 服从标准正态分布,且 $-3 \leqslant b \leqslant 3$,记为 $b \sim N(0,1) \wedge b \in [-3,3]$; b 服从 $-3 \sim 3$ 的均匀分布,记为 $b \sim U(-3,3)$; 猜测参数 c 均服从 α 为 5,β 为 17 的贝塔分布,记为 $c \sim Beta(5,17)$. 模拟生成包含 a、b、c 3 等参数的 4 个题库,依次为题库 1、题库 2、题库 3、题库 4,题量均为 $m = 1\,000$,且 $c \sim Beta(5,17)$,其中: ① 题库 1 中 $a \sim U(0.2, 2.5)$, $b \sim U(-3,3)$; ② 题库 2 中 $a \sim U(0.2, 2.5)$, $b \sim N(0,1) \wedge b \in [-3,3]$; ③ 题库 3 中 $\ln a \sim N(0,1) \wedge a \in [0.2, 2.5]$, $b \sim U(-3,3)$; ④ 题库 4 中 $\ln a \sim N(0,1) \wedge a \in [0.2, 2.5]$, $b \sim N(0,1) \wedge b \in [-3,3]$.

2.3 模拟被试作答

根据当前所选题目 j 的参数和被试 i 的能力真值 θ_i,计算其答对概率 $P_{ij}(\theta)$,其中,$P_{ij}(\theta)$ 的计算公式因为模型的不同而有所不同,如使用 3PLM 时,其值可由下列公式算得:

$$P_{ij}(\theta) = c_j + (1-c_j)\frac{\exp[Da_j(\theta_i - b_j)]}{1+\exp[Da_j(\theta_i - b_j)]}, \quad (6)$$

其中 $D = 1.7$,区分度 a_j、难度 b_j 和猜测度 c_j 均为已知. 同时模拟生成一个服从 0 到 1 均匀分布的随机数 r,记为 $r \sim U(0,1)$. 当 $r \leqslant P_{ij}(\theta)$,则认为被试正确作答题目 j,得 1 分; 否则得 0 分.

2.4 施测过程

施测过程分为 2 个阶段: 能力粗估阶段,从题库中随机抽取 3 道题让被试作答,根据被试的作答反应,使用 EAP 方法估计,得到被试的能力初值; 精确施测阶段,根据被试的能力初值,分别使用按 a 分层法和按 b 分块 a 分层法选题,再根据被试的作答反应使用 EAP 方法重估被试的能力值,再选题,如此反复,直至满足测验的终止规则. 其中,测验中被试的得分根据其作答反应获取.

2.5 评价指标

本文采用 7 个评价指标[9-10]评价终止规则的优劣: 能力估计准确性(Re)、选题策略稳定性(Se)、项目调用均匀性(De)、人均用题数(Nf)、测验效率(Eff)、卡方统计量(χ^2)、测验重叠率(Rt),其中 Eff 越大越好外,其他均为越小越好.

2.6 实验结果分析

实验 1 题库均按 a 分层,分为 4 层,采用能力值与难度最匹配法选题,测验信息量取 16,最大答题数为 60,每个实验重复 30 次. 所有 CAT 模拟实验均在 Matlab 7.1 下进行.

根据下列 4 个表的实验数据显示,新的终止规则,除了能力估计准确性(Re)和选题策略稳定性(Se) 2 个指标与其他方案相当外,其他指标均远优于其他方案,文献[5-7]方案、文献[8]方案总体效果相当.

表 1　$a \sim U(0.2,2.5)$, $b \sim U(-3,3)$, $c \sim Beta(5,17)$ 实验结果

终止规则	*Re*	*Se*	*De*	*Nf*	*Eff*	χ^2	*Rt*
1:3:5:7	0.197 1	0.237 1	67.220 0	38.130 0	0.451 7	118.520 0	0.155 8
文献[6-7]	0.197 8	0.239 0	42.261 0	27.327 0	0.634 1	65.363 0	0.091 8
文献[8] 方案 1	0.198 9	0.239 6	42.095 0	27.283 0	0.635 0	64.955 0	0.091 3
文献[8] 方案 2	0.196 8	0.237 7	42.282 0	27.304 0	0.634 5	65.484 0	0.091 9
新方案 1	0.201 0	0.244 7	31.029 0	22.232 0	0.780 9	43.310 0	0.064 6
新方案 2	0.199 8	0.242 4	30.990 0	22.219 0	0.780 9	43.226 0	0.064 5

表 2　$a \sim U(0.2,2.5)$, $b \sim N(0,1) \wedge b \in [-3,3]$, $c \sim Beta(5,17)$ 实验结果

终止规则	*Re*	*Se*	*De*	*Nf*	*Eff*	χ^2	*Rt*
1:3:5:7	0.194 8	0.235 4	43.514 0	41.569 0	0.414 7	45.554 0	0.086 2
文献[6-7]	0.196 7	0.238 1	24.756 0	29.295 0	0.590 2	20.923 0	0.049 3
文献[8] 方案 1	0.198 0	0.239 9	24.728 0	29.301 0	0.590 4	20.870 0	0.049 2
文献[8] 方案 2	0.198 2	0.239 3	24.708 0	29.275 0	0.590 5	20.856 0	0.049 2
新方案 1	0.197 9	0.240 2	19.186 0	23.685 0	0.730 6	15.544 0	0.038 3
新方案 2	0.200 0	0.241 3	19.234 0	23.700 0	0.730 7	15.612 0	0.038 4

表 3　$\ln a \sim N(0,1) \wedge a \in [0.2,2.5]$, $b \sim U[-3,3]$, $c \sim Beta(5,17)$ 实验结果

终止规则	*Re*	*Se*	*De*	*Nf*	*Eff*	χ^2	*Rt*
1:3:5:7	0.290 8	0.332 1	120.500 0	60.000 0	0.113 8	242.010 0	0.301 3
文献[6-7]	0.196 6	0.237 4	78.279 0	48.552 0	0.351 6	126.220 0	0.174 0
文献[8] 方案 1	0.197 3	0.236 8	78.407 0	48.535 0	0.351 7	126.670 0	0.174 4
文献[8] 方案 2	0.195 8	0.234 5	78.209 0	48.515 0	0.351 7	126.090 0	0.173 8
新方案 1	0.198 2	0.238 6	58.653 0	37.501 0	0.456 0	91.743 0	0.128 4
新方案 2	0.195 0	0.235 2	58.597 0	37.533 0	0.455 4	91.492 0	0.128 2

表 4　$\ln a \sim N(0,1) \wedge a \in [0.2,2.5]$, $b \sim N(0,1) \wedge b \in [-3,3]$, $c \sim Beta(5,17)$ 实验结果

终止规则	*Re*	*Se*	*De*	*Nf*	*Eff*	χ^2	*Rt*
1:3:5:7	0.302 8	0.344 1	81.415 0	60.000 0	0.104 1	110.480 0	0.169 7
文献[6-7]	0.197 4	0.235 8	44.959 0	50.251 0	0.336 5	40.228 0	0.089 6
文献[8] 方案 1	0.196 3	0.235 1	44.759 0	50.283 0	0.336 2	39.846 0	0.089 2
文献[8] 方案 2	0.197 4	0.235 7	44.872 0	50.259 0	0.336 3	40.068 0	0.089 4
新方案 1	0.195 4	0.235 5	33.647 0	38.923 0	0.436 3	29.090 0	0.067 1
新方案 2	0.198 3	0.237 9	33.701 0	38.928 0	0.436 3	29.177 0	0.067 2

实验 2　题库均按 b 分块按 a 分层,分为 4 层,采用能力值与难度最匹配法选题,测验信息量取 16,最大答题数为 60,每个实验重复 30 次. 所有 CAT 模拟实验均在 Matlab 7.1 下进行.

由表 5 及表 7 知新方案 1 及 2 除了能力估计准确性(*Re*) 和选题策略稳定性(*Se*) 2 个指标与其他方案相当外,其他指标均远优于其他方案. 由表 6 及表 8 知新方案与文献[8] 方案相当,略优于文献[8] 方案. 综上所述,新方案可行,且新方案 1 及 2 表现相当.

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

终止规则	*Re*	*Se*	*De*	*Nf*	*Eff*	χ^2	*Rt*
1:3:5:7	0.197 8	0.237 4	40.827 0	31.157 0	0.553 6	53.501 0	0.083 7
文献[6-7]	0.201 8	0.243 4	26.400 0	23.633 0	0.730 6	29.495 0	0.052 2
文献[8] 方案 1	0.201 0	0.243 5	29.750 0	25.824 0	0.669 5	34.275 0	0.059 2
文献[8] 方案 2	0.199 3	0.242 1	28.945 0	25.263 0	0.685 4	33.168 0	0.057 5
新方案 1	0.199 8	0.244 3	24.106 0	21.174 0	0.820 3	27.447 0	0.047 7
新方案 2	0.200 2	0.243 5	25.138 0	22.240 0	0.778 4	28.416 0	0.049 7

表 6　$a \sim U(0.2,2.5)$, $b \sim N(0,1) \wedge b \in [-3,3]$, $c \sim Beta(5,17)$ 实验结果

终止规则	*Re*	*Se*	*De*	*Nf*	*Eff*	χ^2	*Rt*
1:3:5:7	0.196 2	0.236 3	22.091 0	29.302 0	0.590 9	16.657 0	0.045 0
文献[6-7]	0.199 7	0.241 0	13.671 0	24.274 0	0.712 8	7.702 2	0.031 0
文献[8] 方案 1	0.199 9	0.241 9	13.405 0	23.818 0	0.726 3	7.545 6	0.030 4
文献[8] 方案 2	0.200 5	0.243 7	13.795 0	24.597 0	0.704 1	7.738 0	0.031 4
新方案 1	0.199 3	0.241 0	13.362 0	22.341 0	0.773 3	7.992 6	0.029 4
新方案 2	0.199 5	0.242 1	13.114 0	21.865 0	0.792 6	7.866 1	0.028 8

表 7 $\ln a \sim N(0,1) \wedge a \in [0.2,2.5], b \sim U(-3,3), c \sim Beta(5,17)$ 实验结果

终止规则	*Re*	*Se*	*De*	*Nf*	*Eff*	χ^2	*Rt*
1:3:5:7	0.228 8	0.271 3	99.945 0	56.180 0	0.232 0	177.830 0	0.233 2
文献[6-7]	0.197 9	0.237 9	49.374 0	37.937 0	0.449 6	64.267 0	0.101 3
文献[8] 方案 1	0.197 0	0.237 0	44.579 0	38.607 0	0.440 4	51.480 0	0.089 2
文献[8] 方案 2	0.197 8	0.237 7	47.745 0	37.318 0	0.455 9	61.089 0	0.097 5
新方案 1	0.196 5	0.236 4	39.212 0	33.189 0	0.512 8	46.332 0	0.078 6
新方案 2	0.197 5	0.237 4	38.720 0	35.021 0	0.486 0	42.815 0	0.076 9

表 8 $\ln a \sim N(0,1) \wedge a \in [0.2,2.5], b \sim N(0,1) \wedge b \in [-3,3], c \sim Beta(5,17)$ 实验结果

终止规则	*Re*	*Se*	*De*	*Nf*	*Eff*	χ^2	*Rt*
1:3:5:7	0.197 7	0.236 4	39.456 0	49.149 0	0.342 2	31.676 0	0.079 9
文献[6-7]	0.199 8	0.237 3	26.238 0	43.073 0	0.386 6	15.986 0	0.058 1
文献[8] 方案 1	0.199 1	0.237 3	25.162 0	41.702 0	0.405 5	15.184 0	0.055 9
文献[8] 方案 2	0.197 4	0.236 4	24.056 0	39.695 0	0.427 7	14.581 0	0.053 3
新方案 1	0.198 1	0.238 9	22.424 0	33.446 0	0.507 0	15.036 0	0.047 5
新方案 2	0.198 1	0.238 1	21.248 0	34.425 0	0.493 1	13.116 0	0.046 6

3 小结与展望

本文综述分层化方法在安全性等方面的优越性以后,陈述了几种分层退出方案,提出了在 0-1 评分 3PLM 下的按 a 分层和按 b 分块 a 分层方法的 CAT 中新的分层终止规则. 通过 2 个实验对比,得出新的分层退出方案和已有方案相比,在人均用题数、测验效率、卡方统计量、测验重叠率等方面,都有优势,特别是与 b 分块 a 分层相结合以后,除了 *Re* 和 *Se* 相当外,其他指标都比和 a 分层结合表现更好,对于提高题库的安全性和测验效率方面有更好的表现. 新的分层退出方案如何推广到多级评分模型下,值得探讨. 由于不定长的分层退出规则还处于探索阶段,相关研究还比较薄弱,是否还有更好的分层退出方法,值得研究.

4 参考文献

[1] Chang Huahua, Ying Zhiliang. A-stratified multistage computerized adaptive testing [J]. Applied Psychological Measurement, 1999, 25: 211-222.

[2] Chang Huahua, Qian J, Ying Zhiliang. A-stratified multistage CAT with b-blocking [J]. Applied Psychological Measurement, 2001, 25: 333-341.

[3] 王茜娟,丁树良,谭渊. 按 c-分层不定长 CAT 的研究[J]. 江西师范大学学报: 自然科学版, 2005, 29(3): 227-230.

[4] 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.

[5] 文剑冰,侯杰泰. A-stratified 方法在不定长 CAT 中的应用[R]. 第五届华人社会心理与教育学术研讨会, 2001.

[6] 陈德枝. Samejima 等级反应模型下 CAT 选题策略比较研究[D]. 南昌: 江西师范大学, 2004.

[7] 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.

[8] 朱隆尹,丁树良,王茜娟. 不定长 CAT 区分度分层终止规则研究[J]. 心理学探新, 2008, 28(4): 80-84.

[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] 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.

Exploration of Hierarchical Termination Rules for CAT

HU Shan, DING Shu-liang*, CHENG Yan, XIONG Jian-hua

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

Abstract: The question how to control each hierarchical information boundary in variable length CAT is still empirical exploration, two new termination rules for quit of stratifications have been raised. Simulations show that the new methods have a good effect on improving the safety and efficiency.

Key words: CAT; termination rules; variable length; stratified

(责任编辑: 冉小晓)