文章编号:1000-5862(2018)02-0139-05

动态加权区间的 CAT 选题策略研究

邱 敏,罗 芬,熊建华,丁树良,甘登文*
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要:在计算机化自适应测验中,由于测验的性质不同,在衡量测验优劣的多个指标中,有的测验侧重于测量精度,有的侧重于测验的公平性,还有的侧重于测验的效率. 指标之间或许有冲突,但希望尽可能多方兼顾. 该文构造了动态加权区间的选题策略以适应测验目的多样性:先构造一个包含最大信息量的区间,该区间的题目集相当于一个"影子题库",再设置一个权值调节影子题库的大小. 区间的使用可以提高题库利用的均匀性,保证题库安全,而权值根据测验关注点进行调整可实现测验目标. 模拟实验显示:新的选题方法效果比较理想.

关键词:加权区间;计算机化自适应测验;选题策略;3 参数逻辑斯蒂克模型
中图分类号:B 841.7 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2018.02.04

0 引言

计算机化自适应测验(Computerized Adaptive Testing,CAT)是以计算机为媒介的量体裁衣式测验,它不同于传统考试的"千人一卷",其目的是为每一个被试构造一个最优测验. 项目反应理论(Item Response Theory,IRT)是实现 CAT 的基础,使得题库中不同项目之间的参数可以相互比较,确保了测验的公平性[1].

在 CAT 的组成部分中[2],选题策略是一个研究热点. 一般来说,CAT 的选题策略要求选出的项目达到以下目标:尽可能准确地估计被试的潜在特质,尽量保护题库的安全性(即控制项目的曝光率),各个项目使用频率尽量均匀,兼顾测验内容均衡,以及合理、高效地利用题库[3]. 选题策略是 CAT 能够实现自适应的关键,其好坏直接关系到 CAT 的质量.

在 IRT 中,Fisher 信息量是一个重要的概念,它有如下性质:(i)项目所能提供信息量与项目区分度的平方成正比;(ii)信息量越大,能力测量方差越小,测量越准确;(iii)由局部独立性假设,测验信息量(也称为累积信息量)可以表示为各个项目提供的信息量之和[4]. 从性质(i)不难看出,若选题策略仅仅考虑准确性,则区分度高的项目会被频繁地选择,而区分度低的项目较少调用(甚至不用),从而导致一些试题过度曝光,被试之间可能共享试题信息而影响能力估计的准确性和考试的公平性[5].

传统的选题策略有随机选题法[6](RAN)和最大 Fisher 信息量选题方法[7](MFI). RAN 在整个题库中,随机选择一个项目给被试作答,直到测验结束;它的优点是项目曝光率均匀,可作为项目使用均匀性的参照标准,但是测量精度较低. MFI 是根据被试当前能力估计值,计算该被试的剩余题库(题库中的项目减去该被试已经作答的项目后,余下的项目集合称为该被试的剩余题库,简称为剩余题库)中所有项目的 Fisher 信息量,选取信息量最大的项目作为被试的下一作答项目. 该方法的优点是能力估计准确性较高,可作为测量精度的参照标准,其缺点在于会频繁选取高区分度的项目,使得某些项目曝光率过高,影响考试安全.

针对 MFI 策略的缺点,程小扬等[8]提出了引入曝光因子的最大信息量选题策略(Modi-MIC 1),在 MFI 的基础上引入了 3 个量:(i)项目 j 的控制曝光因子(exposure-control factor),记为 $ecf(j)$;(ii)$ecf(j)$ 的调节因子 λ_j,调节 $ecf(j)$ 对选择项目的影响(Modi-MIC 1 策略中取 $\lambda_j = 1$,即 λ_j 不施加影响);(iii)区分度 a_j 的幂函数 $a(j,T,k) = a_j^{2(T-k)/(T-1)}$,$T$ 表示分 T 个阶段选题(Modi-MIC 1 策

收稿日期:2017-11-02
基金项目:国家自然科学基金(31360237,31500909)资助项目.
通信作者:甘登文(1956-),男,江西奉新人,教授,主要从事智能教学软件和应用统计的研究. E-mail:gdw8120429@126.com

略中 $T=1$),k 表示当前所处阶段,并且 $1 \leqslant k \leqslant T$. 当 $k=1$ 时,f_j 是项目信息量与 3 个引入量乘积之比,即 $f_j = I_j(\hat{\theta}) / (ecf(j))$. 在 CAT 作答过程中,从剩余题库中选取满足 $\max f_j$ 的项目作为被试的下一个测试项目,引入曝光因子之后,项目的均匀性得到提高,曝光率得到明显改善.

测验由于测量的关注点不同,有的测验偏重于测量精度,有的更关注测验的公平性,还有的侧重测验的效率. 而程小扬方法(Modi-MIC 1)虽然有提高项目均匀性的优势,但是不能兼顾测验的各种需求,故本文在程小扬的基础上提出了动态加权区间方法,可以根据测验当时的要求对区间进行动态调整以达成测量目标,并探讨其在不同题库结构下的表现,做法如下:先构造一个包含最大信息量的区间,落入该区间的项目一般不止一个,此区间的题目集相当于一个"影子题库"[8],再构造一个权值,调节影子题库的大小. 区间的使用可以提高题库利用的均匀性,保证题库安全,而权值根据测验关注点进行调整实现测验目标.

1 动态加权区间选题策略

1.1 3PLM 模型与 Fisher 信息量

本文使用的模型为 3 参数逻辑斯蒂克模型(3PLM)[9],其项目反应函数为

$$P_{(u_j=1\mid\theta)} = c_j + \frac{1-c_j}{1+\exp(-D\cdot a_j(\theta-b_j))},$$

其中 $P_{(u_j=1\mid\theta)}$ 表示能力值为 θ 的被试在项目 j 上正确作答的概率,θ 为被试的能力值,a_j 表示项目 j 的区分度,b_j 表示项目 j 的难度,c_j 表示项目 j 的猜测度,D 为量表因子,一般取常数 1 或者 1.7(在本文中 D 取 1),u_{ij} 是被试 i 在项目 j 上的作答反应,3PLM 的项目信息函数为[10]

$$I_{ij}(\theta) = \frac{D^2 a_j^2 (1-c_j)}{(c_j + e^{Da_j(\theta_i-b_j)})(1+e^{-Da_j(\theta_i-b_j)})^2}.$$

由局部独立性假设,累积信息函数为 $I(\theta) = \sum_{j=1}^{n} I_j(\theta)$,$I_j(\theta)$ 表示能力值为 θ 的被试在项目 j 上的信息量,n 为测验的项目总数.

1.2 能力估计方法与终止规则

估计被试能力水平的常用方法有极大似然估计法(Maximum Likelihood Method,MLE)和贝叶斯方法(Bayesian Method),其中最常用的是贝叶斯期望后验估计(Expected a Posteriori,EAP)和贝叶斯极大后验估计(Maximum a Posteriori,MAP). 本文采用 EAP 能力估计方法,该方法不需要进行迭代运算,计算简单、速度快.

CAT 常用的终止规则大致分为 2 类:定长和不定长. 定长是指在一场测验中所有被试的测验长度是一个统一的预设值[11];不定长则是不固定测验长度,测验累积信息量达到预设值就结束. 定长终止规则操作比较简单,但各个被试的测量精度可能相差比较大. 本文选择测验长度为 35 题的定长测验和累积信息量达到 9 时测验中止的变长测验.

1.3 新的选题策略

本文提出的新策略同 Modi-MIC1 策略一样根据信息量来选题,Modi-MIC1 通过曝光因子提高项目均匀性,保障题库的安全. 新策略则通过放大区间来实现,并且通过设置权值动态调整区间,根据测验关注点实现测验目标. 新的选题策略如下:令 $\varepsilon = 1/\sqrt{I(\theta)}$,其中 $I(\theta) = \sum_{i=1}^{n} I_i(\theta)$ 是累积信息量,而 $I_i(\theta)$ 表示能力值为 θ 的被试在项目 i 上的信息量,ε 为能力的极大似然估计理论的根方差. 随着测验深入,$I(\theta)$ 增大,ε 减少,意味着测验开始区间大,可以使用低区分度项目,随着测验累积信息量增大,区间变小,逐步使用高区分度项目. 具体的选题过程如下:首先根据被试估计的能力值计算剩余题库中单个项目所能提供的最大信息量 $infor_{\max}$,再把能提供项目信息量介于区间 $(infor_{\max}-\varepsilon, infor_{\max})$ 的题目保存下来,最后选择这些题目中被选次数最少的那一题,不断重复前面 3 步,直到测验终止. 引入区间的思想增加了可选题目的数量,提高题库利用的均匀性,从而保障题库的安全性.

由于测验的初始阶段区间较大,造成达到同样测量精度,被试的人均用题数增加较多,再选取一个权值 w,将 ε 改为 $w\varepsilon$,用权值 w 动态调整 ε,即通过调整权值达到调整区间大小的目的. 注意 $w=0$ 时,即为 MFI 策略. 根据测验目的(关注测验准确性还是关注题库使用的均匀性)来调整 w 的值实现测验目标. 这里称新方法为动态加权区间的 MFI 策略(DWI-MFI).

2 CAT 的模拟过程

2.1 被试及题库模拟

题库和被试按照如下分布进行模拟[12]:1)项目参数. 根据项目参数服从不同的分布,可以得到 4

种不同的题库:(i) $\ln a \sim N(0,1)$,$b \sim N(0,1)$,记为题库 1.(ii) $\ln a \sim N(0,1)$,$b \sim U(-3,3)$,记为题库 2.(iii) $a \sim U(0.2,2.5)$,$b \sim N(0,1)$,记为题库 3.(iv) $a \sim U(0.2,2.5)$,$b \sim U(-3,3)$,记为题库 4. 对以上 4 种题库分布中的猜测参数 c 均服从 α 为 5,β 为 17 的 Beta 分布,记为 $c \sim \beta(5,17)$. 2) 能力参数. 产生 1 000 个能力值为 θ 的被试,能力参数 θ 服从标准的正态分布,记 $\theta \sim N(0,1)$.

2.2　施测过程

本文讨论 3PLM 模型下的随机选题策略(RAN)、最大信息量选题策略(MFI)、引入曝光因子的最大信息量选题策略(Modi-MIC 1) 以及本文提出的 DWI-MFI. 施测过程一般分为 2 个阶段: 第 1 阶段从剩余题库中随机抽取 5 题给被试作答,计算得到被试的能力初值; 第 2 阶段再根据本文介绍的选题策略选择项目 j 作答,求得被试在该项目上的得分,运用 EAP 计算能力估计值. 重复以上步骤,直到测验终止. 为消除实验随机误差,每个实验重复 15 次,取 15 次实验结果的平均值.

2.3　评价指标

本文用能力估计准确性、χ^2 检验统计量、人均用题数共 3 个评价指标来评价比较选题策略的优劣,指标值越小表示选题策略越好.(i) 能力估计准确性: $ABS = \sum_{i=1}^{N} |\theta_i - \hat{\theta}_{ij}| / N$;(ii) χ^2 检验统计量: $\chi^2 = \sum_{j=1}^{M} \left((A_j - (\sum_{j=1}^{M} A_j / M) \right)^2 / (\sum_{j=1}^{M} A_j / M))$, 其中 A_j 为第 j 题曝光率. 计算 A_j 的方法为: A_j = 第 j 题被使用的次数 $/N$. M 为题库的总项目数.(iii) 人均用题数[13]: $Nf = \sum_{i=1}^{N} r_i / N$,其中 r_i 为第 i 个被试在模拟中作答的项目数.

2.4　实验 1 DWI-MFI 的权值 w 选取

采用不定长累积信息量达到 9 终止的规则,重复 15 次测验. 收集了新方法在 4 个题库下,权值 w 从 0.1 ~ 1.3 的各项评价指标值,如图 1 所示.

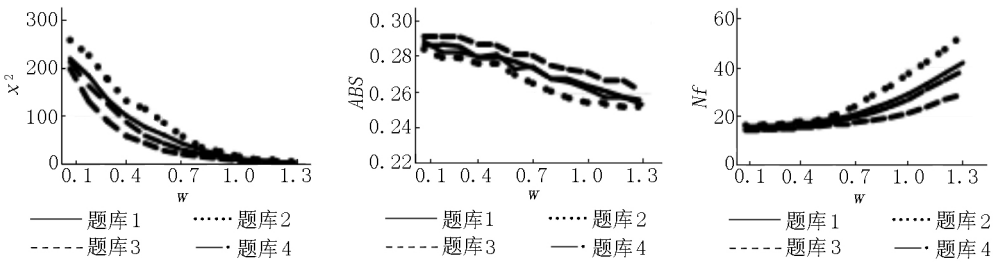

图 1　各项评价指标

图 1 表明,在不定长实验中,权值越大,χ^2 检验统计量越小,表示均匀性越好,从而可以调整题库的安全性; 人均用题数呈上升趋势,由于测验长度增加,因此能力估计精度也随之提高. 可见权值的调整可以调节测验的测量精度和题库使用均匀性. 本实验在题库 3 下的均匀性最好,测验长度最短,即效率较高. 由于测验长度短,精度相对于其他题库要稍差些.

2.5　实验 2 DWI-MFI 与其他选题策略的比较

2.5.1　不定长测验结果分析　实验 1 结果表明,权值增大,题库利用均匀性更好,测验效率有所下降,反之亦然. 本文采用动态权值均衡这 2 个指标的变化. 在不定长测验的条件下,使用如下 w 来权衡安全性和精度: $w = w_{max} - (w_{max} - w_{min}) M / T$,其中 M 为累积信息量的值,T 为要达到的信息量的值(本文中 $T = 9$). w_{max} 表示选取的较大权值,w_{min} 表示选取的较小权值.

例如对于题库 1,保证精度差距不大并且人均用题数尽量少的前提下,在图 1 中寻找均匀性与 RAN 策略更接近的权值作为 w_{max},这里可以取 1.2; 寻找与 MFI 策略更接近的权值作为 w_{min},这里可以取 0.2. 根据公式,运用 $w = 1.2 - M / T$ 在测验进行过程中动态调整权值.

在定长的条件下,$w = w_{max} - (w_{max} - w_{min}) l / L$,其中 l 为测验的当前长度,L 为预设的测验(本文中 $L = 35$). w_{max} 与 w_{min} 选取方法同上.

本文选取了不同的 w_{max} 与 w_{min} 的值进行对比,从而找到最优测验结果. 根据定长和不定长测验,分别利用上述公式找出的权值,再将新策略与其他方法进行比较,先看不定长测验条件下 15 次试验平均的结果,如表 1,各题库权值选取如下: $w = 1 - 2M / (5T)$(题库 1),$w = 0.9 - M / (10T)$(题库 2),$w = 1.1 - M / (2T)$(题库 3),$w = 1 - 3M / (5T)$(题库 4).

表 1 表明在不定长测验中,由于新策略兼顾测验精度和测验安全性,当新方法的均匀性和精度与 Modi-MIC 1 策略等效甚至好于它的时候,人均用题数增加 1 ~ 2 题,并且同实验 1 一样,新方法在题库 3 下达到相同的测量精度时,测验长度最短,题库利

用率最均匀.

表 1 不定长测验的实验结果

选题策略	题库 1			题库 2			题库 3			题库 4		
	χ^2	*ABS*	*Nf*	χ^2	*ABS*	*Nf*	χ^2	*ABS*	*Nf*	χ^2	*ABS*	*Nf*
MFI	259.32	0.28	15.48	287.37	0.28	16.38	249.81	0.27	14.22	257.09	0.29	14.63
RAN	0.89	0.25	75.63	0.91	0.25	93.97	0.96	0.25	52.53	0.97	0.24	65.43
Modi-MIC 1	25.95	0.26	22.02	34.71	0.25	26.02	10.91	0.26	18.99	18.42	0.27	20.28
DWI-MFI	19.74	0.26	25.92	29.52	0.25	30.86	7.07	0.27	21.33	16.20	0.26	23.90

2.5.2 定长测验结果分析 各题库权值选取如下: $w=1.2-3l/(10L)$(题库 1),$w=1.2-l/(5L)$(题库 2),$w=1.2-l/(2L)$(题库 3),$w=1.2-l/(5L)$(题库 4). 实验结果如表 2 所示.

表 2 定长测验的实验结果

选题策略	题库 1		题库 2		题库 3		题库 4	
	χ^2	*ABS*	χ^2	*ABS*	χ^2	*ABS*	χ^2	*ABS*
MFI	291.71	0.19	305.59	0.19	256.78	0.17	275.68	0.18
RAN	1.02	0.34	1.24	0.38	1.09	0.28	1.11	0.33
Modi-MIC 1	41.38	0.21	50.38	0.22	23.42	0.19	35.36	0.21
DWI-MFI	33.58	0.23	56.40	0.23	24.86	0.20	38.80	0.21

表 2 表明,新方法在定长测验时依旧可行,并且在各题库下的表现和不定长结果相似. 本研究引入权值,通过调整权值达到调整区间大小的目的,从而达成测验目标. 当测验注重安全性的时候,主要考虑用 χ^2 统计量调整; 如果测验注重效率,则要从测验精度以及测验长度 2 个方面来调整. 结果见表 3.

表 3 权值动态调整的结果

题库	不定长				定长		
	χ^2	*ABS*	*Nf*	*w*	χ^2	*ABS*	*w*
1	19.74	0.26	25.92	$w=1-2M/(5T)$	40.07	0.22	$w=1.2-2l/(5L)$
	17.35	0.26	28.58	$w=1.2-M/T$	33.58	0.23	$w=1.2-3l/(10L)$
2	29.52	0.25	30.86	$w=0.9-M/(10T)$	56.40	0.23	$w=1.2-3l/(5L)$
	25.51	0.25	33.01	$w=1-2M/(5T)$	45.49	0.24	$w=1.2-l/(2L)$
3	11.10	0.27	19.38	$w=1-M/(2T)$	24.86	0.20	$w=1.2-l/(5L)$
	7.07	0.27	21.33	$w=1.1-M/(2T)$	42.32	0.197 6	$w=1.2-2l/(5L)$
4	16.20	0.26	23.90	$w=1-3M/(5T)$	38.80	0.21	$w=1.2-l/(2L)$
	21.45	0.26	23.36	$w=1-7M/(10T)$	12.01	0.24	$w=1.2-l/(5L)$

表 3 表明,根据本文提出的公式选取的动态的权值可以实现各项指标的改变.

综合以上实验,本方法能根据测验目的需求,通过调整权值进而控制考试的安全性和精度,若需要精度高,可以缩小权值; 若需要安全性高,则可增大权值. 本文还提出了一个权衡安全性和精度的公式,利用公式可以找出一个优良解,使得项目的曝光率大大降低,提高了考试的安全性,相比于 Modi-MIC 1 的测验长度会有所增加,但是能力估计的准确性要好些. 在今后的 CAT 测验中,本方法可以成为一种灵活安全且高效的选题策略.

3 讨论

G. G. Kingsbury 等[14] 提出的 RAND 策略是从一组最优项目中随机选择一个项目给被试作答,J. Revuelta等[15] 提出的 PROG 策略是给项目信息量添加一个随机组件,每次选择二者加权和最大的项目给被试作答. 本文的新方法是对上述 2 个控制项目曝光率策略的改进: 对于 RAND,新方法是从最优项目组中,再选被选次数最少的题作答; 对于 PROG,新方法的 ε 是比随机组件更精确的调节信息量的因子. 因此,动态加权区间策略同样达到了控制项目曝光率保证题库安全性的目的.

本文进行了 2 个实验,一个是验证权值大小对精度和均匀性的影响,另一个是为权衡精度和安全性提出的一个公式的实践. 从实验 1 可以看出,修改权值的大小确实可以动态调整均匀性和精度,使得测验结果灵活化,并且通过对比可以看出,若不加以权值的调整,相比于 Modi-MIC 1,精度和均匀性都较好,但是测验长度更长. 在实验 2 中,通过定长和不定长条件下的实验结果,由本文提出的公式,都能找

到一个比较理想的结果.

当然,本文只是提出了一个权值的概念和一个权衡均匀性和精度的公式,是否还有其他方法让选题策略的效果更佳,并且适用于多级评分模型,是否可以根据测验的关注点自动选择权值? 这些都有待于进一步研究.

4　参考文献

[1] 丁树良,罗芬,涂冬波. 项目反应理论新进展专题研究[M]. 北京: 北京师范大学出版社,2012: 173.

[2] 游晓锋. CAT 中原始题目项目参数的估计[D]. 南昌: 江西师范大学,2008.

[3] 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.

[4] 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.

[5] 胡珊. 基于 GPCM 和平 PLM 的 CAT 研究[D]. 南昌: 江西师范大学,2015.

[6] 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.

[7] 汪文义. 计算机化自适应测验选题策略研究[D]. 南昌: 江西师范大学,2009.

[8] 戴勰,甘登文,丁树良. 结合影子题库的选题策略[J]. 江西师范大学学报: 自然科学版,2013,37(6) : 657-660.

[9] 罗照盛. 项目反应理论基础[M]. 北京: 北京师范大学出版社,2012.

[10] 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: 89-90,150-153.

[11] 漆书青. 计算机化自适应测验的编制和应用[J]. 江西教育科研,1999(2) : 65-67.

[12] 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(3) : 461-467.

[13] 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.

[14] Kingsbury G G,Zara A R. Procedures for selecting items for computerized adaptive tests[J]. Applied Psychological Measurement,1991,26: 412-432.

[15] Revuelta J,Posonda V. A comparison of item exposure control methods in computerized adaptive testing[J]. Journal of Educational Measurement,1998,35: 311-327.

The Study on Item Selection Method of CAT with Dynamic Weighted Interval

QIU Min,LUO Fen,XIONG Jianhua,DING Shuliang,GAN Dengwen*

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

Abstract: Due to the different property of computerized adaptive testing,among a number of indicators to measure the quality of the test,some tests focus on measurement accuracy or fairness and others focus on the efficiency,of course there are conflicts between those indicators,but as far as possible consider more indicators. In the paper,a new item selection strategy with dynamic weighted interval is proposed,which can meet the demand of diversity. Firstly,constructing an interval contains the maximum information,the group of those items are equivalent to a "shadow bank",then a weight is set to adjust the size of shadow bank. The interval can balance the item pool usage and ensure test security,the weight can adjust the indicators according to test focus. Monte Carlo simulation shows that the new method works ideally.

Key words: weighted interval; computerized adaptive testing; item selection method; 3PLM

(责任编辑: 冉小晓)