文章编号:1000-5862(2015)01-0069-04

基于平均数形式的选题策略比较

李　佳,丁树良,方剑英
(江西师范大学计算机信息工程学院,江西 南昌　330022)

摘要:在 3PLM 模型下,将改进的最大优先级指标(MMPI)方法和各类平均数形式相结合得到的 4 种新选题策略在提高测验精度、控制项目曝光均匀性、降低平均违规次数、提高题库利用率等方面均表现更好. 经定长测验和不定长测验的蒙特卡洛模拟,MMPI 下算术平方根平均数形式的选题策略表现最优.

关键词:选题策略;改进的最大优先级指标方法;平均数形式
中图分类号:B 841.7;TP 301.6　**文献标志码**:A　**DOI**:10.16357/j.cnki.issn1000-5862.2015.01.13

0　问题的提出

计算机化自适应测验(CAT)是一种个性化的测验,它能用较少项目精确地测量出被试能力,并且每个被试的测验项目都不一样但又是最适合被试自己的题目,真正实现了对被试而言的"因人施测,量体裁衣". 目前,CAT 已广泛地应用于美国研究生入学考试、美国医生护士资格考试和中国汉语水平考试等. 在 CAT 中,最能体现其智能化的是选题策略,它直接影响到测量被试能力的准确性和题库的安全性. F. M. Lord[1] 提出基于最大信息量准则的选题策略(MIC),曾被广泛地应用于 CAT 中,但随着测验各方面指标要求的不断提高,MIC 方法存在着题库安全隐患和题库中题目的大量浪费等现象. 为了改善这种情况,文献[2-3]提出了 a-分层法和 a-分层、b-分块的选题策略,可以较好地控制项目曝光率,提高题库的安全性,但它们均没有考虑到内容平衡的要求. 为了解决这个问题,文献[4]提出按内容分层(c-STR),S. Buyske 提出基于线性规划的离差加权模型[5],程莹等[6]提出最大优先级指标(MPI),潘奕娆等[7-8]对 MPI 方法进行改进,提出改进的最大优先级指标(MMPI)方法. 但是 MPI 和 MMPI 方法在满足内容平衡的同时降低了能力测量的准确性. 为了更好地平衡这相互冲突的 2 个方面,研究发现 MMPI 方法中的优先级指标是由信息量和约束条件 2 个部分构成的,所以把它们分开来和平均数形式相结合,得到 4 种新的选题策略,在测验初期充分满足内容约束等非计量学指标的要求,而在测验中后期逐渐体现信息量的作用,不断提高测验精度. 将这 4 种新方法和随机化最大信息量选题方法以及 MMPI 选题方法一起在定长测验和不定长测验下分别进行模拟比较. 本研究共设计了 6 种评价指标分别从能力估计准确性、卡方检验统计量、被试平均违规次数、项目从未曝光率、项目过低曝光率和不定长测验时平均测验长度等指标对以上 6 种选题策略进行综合评价比较.

0.1　随机化最大信息量方法(改进的 MIC)简介

G. G. Kingsbury[9] 提出的随机化最大信息量选题方法较容易实现,从当前 5 个信息量最大的项目中随机地挑选出一个项目给被试作答.

0.2　改进的最大优先级指标方法(MMPI)简介

程莹等提出的最大优先级指标方法(MPI)是一种 2 阶段选题策略,其思想是建立一个综合各类约束条件的变量,在选题过程中先满足约束条件个数多的项目,并且充分考虑每类约束条件的上下限. 预先定义约束条件的上下界和约束关联矩阵,其中约束条件为 $l_k \leqslant \mu_k \leqslant u_k$,其中 $l_k, u_k (k = 1, 2, \cdots, K)$ 分别为第 k 个约束的下界和上界,K 为总的内容约束个数,约束关联矩阵 $\boldsymbol{C} = (c_{jk})_{J\times K}$,$J$ 为项目数,$c_{jk} = 1$ 表示项目 j 受到条件 k 的约束. 同时 MPI 也将曝光控制作为一个约束条件 $f_k = (r - n/N)/r$,其中 r 取 0.2,表示项目的最大曝光率控制在 0.2 以下,n 表示当前此项目被使用过的次数,N 表示被试总人数.

收稿日期:2014-11-16
基金项目:国家自然科学基金(30860084,31160203,31100756,31360237,11401271);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.
作者简介:李　佳(1979-),女,江西南昌人,讲师,主要从事计算机辅助教学及教育和心理测量方面的研究.

具体选题策略如下: 项目 j 的最大优先级指标为 $PI_j = I_j^t \prod_{k=1}^{K} (w_k f_k)^{c_{jk}}$,其中 I_j^t 为项目 j 在当前能力 $\hat{\theta}$ 上的信息量,w_k 为约束条件 k 所对应的权重,$f_k = (X_k - x_k)/X_k$ 表示缺额比例,其中在第 1 阶段 X_k 为约束下界数 l_k,第 2 阶段 X_k 为约束上界数 u_k,x_k 为已达到约束 k 的个数.

潘奕娆等[7-8] 发现 MPI 方法中 $w_k f_k$ 的值有可能小于 1,所以对它进行了修正,得到修正的 MPI 方法(MMPI). 项目 j 的优先级指标修正为 $MMPI_j = I_j^t \prod_{k=1}^{K} [g_k(w_k f_k + 1)]^{c_{jk}}$,其中 g_k 用来判断约束条件 k 是否达到上下界,若第 1 阶段约束条件 k 达到下界时 $g_k = 0$,否则 $g_k = 1$; 若第 2 阶段约束条件 k 达到上界时 $g_k = 0$,否则 $g_k = 1$.

0.3 平均数形式

平均数主要包括算术平方根平均数 $(\lambda a^2 + (1-\lambda) b^2)^{1/2}$,算术平均数 $\lambda a + (1-\lambda) b$,几何平均数 $a^{\lambda} b^{1-\lambda}$ 和调和平均数 $1/(\lambda/a + (1-\lambda)/b)$ 共 4 种,考虑在 MMPI 方法中项目优先级指标主要是由信息量和约束条件这 2 个部分组成,尝试用信息量代替平均数形式中的参数 a,用约束条件代替平均数形式中的参数 b,得到 4 种新的项目优先级指标为:

(i) 项目算术平方根平均数优先级指标: $RAM_j = (\lambda (I_j^t)^2 + (1-\lambda) (\prod_{k=1}^{K} (g_k(w_k f_k + 1))^{c_{jk}})^2)^{1/2}$;

(ii) 项目算术平均数优先级指标: $AM_j = \lambda I_j^t + (1-\lambda) \prod_{k=1}^{K} (g_k(w_k f_k + 1))^{c_{jk}}$;

(iii) 项目几何平均数优先级指标: $GM_j = (I_j^t)^{\lambda} \cdot (\prod_{k=1}^{K} (g_k(w_k f_k + 1))^{c_{jk}})^{1-\lambda}$;

(iv) 项目调和平均数优先级指标: $HM_j = 1/(\lambda/I_j^t + (1-\lambda)/\prod_{k=1}^{K} (g_k(w_k f_k + 1))^{c_{jk}})$,

其中 λ 是一个从小到大动态变化的变量,在测验初期先满足内容平衡等非计量学约束,而在测验中后期尽量满足测验信息量的要求,充分发挥信息量的优势,以提高测验精度.

0.4 参与比较的 6 种选题策略

将随机化最大信息量选题方法(改进的 MIC)和 MMPI 方法作为参照,加上 4 种新的选题策略,共 6 种选题策略进行比较研究. 这 6 种选题策略分别为:

(i) 改进的 MIC: $\underset{j \in T}{\text{rand}}\{I_j^t\}$,$T$ 为当前 5 个信息量最大项目构成的集合;

(ii) MMPI: $\max_{j \in R_a}(I_j^t \prod_{k=1}^{K} (g_k(w_k f_k + 1))^{c_{jk}})$;

(iii) MMPI 下算术平方根平均数形式: $\max_{j \in R_a}((\lambda (I_j^t)^2 + (1-\lambda) (\prod_{k=1}^{K} (g_k(w_k f_k + 1))^{c_{jk}})^2)^{1/2}$;

(iv) MMPI 下算术平均数形式: $\max_{j \in R_a}(\lambda I_j^t + (1-\lambda) \prod_{k=1}^{K} (g_k(w_k f_k + 1))^{c_{jk}})$;

(v) MMPI 下几何平均数形式: $\max_{j \in R_a}((I_j^t)^{\lambda} \cdot (\prod_{k=1}^{K} (g_k(w_k f_k + 1))^{c_{jk}})^{1-\lambda})$;

(vi) MMPI 下调和平均数形式: $\max_{j \in R_a}\left(1/(\lambda/I_j^t + (1-\lambda)/\prod_{k=1}^{K} (g_k(w_k f_k + 1))^{c_{jk}})\right)$.

定长测验时: $\lambda = \frac{L(j)}{test\_length}$,$L(j)$ 为当前测验长度,$test\_length$ 表示测验长度; 不定长测验时: $\lambda = \frac{inf(j)}{Infor}$,$inf(j)$ 表示当前已经实施 j 个项目的测验信息量,$Infor$ 表示测验信息总量,λ 在测验进行过程中逐渐从 0 增大变化到 1.

1 模拟实验

1.1 被试及其题库的模拟

本文中所有试验均采用 Monte Carlo 模拟. 在实验过程中模拟生成 5 000 个被试,且被试能力真值服从标准正态分布. 在实验过程中模拟生成题库共 642 个项目且满足条件 $\ln a \sim N(0,1)$,$b \sim N(0,1)$,$c \sim Beta(5,17)$,且 $0.2 < a < 2.5$,$-3.5 < b < 3.5$,$abs(a-b) < 4$,$c < 0.4$,内容约束同文献[10]. 题库的项目数据见表 1.

表 1 题库的项目数据

项目数据	区分度 a	难度 b	猜测度 c
平均值	1.005 10	0.024 13	0.227 26
标准差	0.601 12	0.979 74	0.080 20

1.2 模拟 CAT 的施测过程

本测验为 3PLM 模型下的 0-1 评分测验. 设被试的能力初值为 0,然后根据不同的选题策略采用 EAP 方法对能力进行估计. 分定长和不定长 2 种测验: (i) 定长测验,设定测验长度为 24; (ii) 不定长测验,所有选题策略的测验在被试累积信息量达到 9 时结束.

1.3　评价指标

本文用能力估计的准确性(ABS)、卡方检验统计量(χ^2)、平均违规次数($\bar{V}$)、项目过低曝光率(UE)、项目从未曝光率(NE)和不定长测验下平均测验长度(AL)等6个评价指标来评价比较选题策略的优劣. 这6个指标都是越小越好.

(i) 能力估计准确性(ABS): $ABS = \sum_{i=1}^{N} |\theta_i - \hat{\theta}_i| / N$,其中$N$为被试总人数,$\theta_i$为第$i$个被试的能力真值,$\hat{\theta}_i$为第$i$个被试的能力估计值. 该指标反映了被试能力真值与其能力估计值的平均绝对偏差,指标值越小说明能力估计越准确.

(ii) 卡方检验统计量(χ^2): $\chi^2 = \sum_{j=1}^{M} \left(\left(A_j - \left(\sum_{j=1}^{M} A_j / M \right) \right)^2 / \left(\sum_{j=1}^{M} A_j / M \right) \right)$,其中$M$为题库中项目数,$A_j$为第$j$题的曝光率,即$A_j$ = 第j题的使用次数$/N$. 该指标用来衡量题库中项目曝光的均匀性,指标值越小说明题库中项目曝光越均匀,测验安全性越高.

(iii) 被试平均违规次数($\bar{V}$): $\bar{V} = \sum_{i=1}^{N} V_i / N$,其中$V_i$表示被试$i$在测验中,违背约束条件的总数,这项指标体现了满足内容平衡的条件约束的程度.

(iv) 项目从未曝光率(UE): $UE = \sum_{i=1}^{M} UE_i / M$,其中$UE_i$表示题库中曝光率等于0的项目数,这项指标体现了题库的利用率.

(v) 项目过低曝光率(NE)[11]: $NE = \sum_{i=1}^{M} NE_i / M$,其中$NE_i$表示题库中曝光率小于0.02的项目数,这项指标能更清楚地体现题库的利用率.

(vi) 测验平均长度(AL): 不定长测验中每个被试的测量精度类似,所以早达到测验精度的被试所需测验长度更短,而晚达到测验精度的被试所需测验长度就更长,这项指标体现了测验效率.

1.4　实验结果及其分析

实验为定长测验时,结果见表2; 实验为不定长时,结果见表3.

表2　定长测验($L = 24$)6种选题策略的表现

选题策略	ABS	χ^2	$\bar{V}$	UE	NE
改进的MIC	0.133 010	124.370	49.911 000	0.764 85	0.823 99
MMPI	0.204 810	70.037	0.000 718	0.543 32	0.719 63
MMPI下算术平方根平均数形式	0.189 750	16.009	0.000 574	0.331 78	0.542 99
MMPI下算术平均数形式	0.193 840	26.643	0.009 460	0.361 37	0.552 34
MMPI下几何平均数形式	0.193 710	44.227	0.014 520	0.407 23	0.552 34
MMPI下调和平均数形式	0.198 820	46.214	0.015 220	0.400 88	0.550 78

表3　不定长测验6种选题策略的表现

选题策略	ABS	χ^2	$\bar{V}$	UE	NE	AL
改进的MIC	0.099 03	108.327 0	29.829 600	0.742 49	0.894 08	15.561 0
MMPI	0.144 05	44.645 0	0.007 718	0.518 38	0.778 50	16.864 2
MMPI下算术平方根平均数形式	0.104 28	8.796 3	0.000 694	0.357 26	0.501 89	21.616 2
MMPI下算术平均数形式	0.112 02	9.257 6	0.001 504	0.362 58	0.506 45	18.730 4
MMPI下几何平均数形式	0.119 83	10.902 0	0.001 671	0.398 91	0.536 14	20.232 8
MMPI下调和平均数形式	0.123 46	11.763 0	0.001 717	0.409 25	0.533 02	19.563 8

当测验为定长时,从表2的结果可以看出,改进的MIC方法的测验精度较高,但是后3个指标表现均不够好,这是因为此方法是按最大信息量方法选题而不理会内容约束的要求. MMPI下各平均数形式的测验精度比MMPI要好,这是因为新方法在测验后期对信息量更为关注,所以测验精度会更高; 在MMPI下各平均数形式被试平均违规次数比MMPI要好,是因为新方法在测验初期特别强调了先满足内容平衡的约束; 而新方法在题库利用率方面优于MMPI的原因是新方法对项目优先级指标做了改进,动态地调整了信息量和约束条件的比例; 熟知调和平均数不超过几何平均数、几何平均数不超过算术平均数、算术平均数不超过算术平方根平均数,所以在4种新方法中MMPI下算术平方根平均数表现更好的理由是算术平方根平均数形式把信息量和约束条件的差异拉得比较大,在实现上更能体现信息量和满足约束条件的优点.

当测验为不定长时,从表3的结果可以看出,实验结果和定长测验类似,但测验精度更高一些,测验平均长度都短于定长测验的测验长度,这也说明了

不定长测验更有利于提高测验效率,所以在 CAT 中采用不定长测验是有道理的.

2 讨论

试验结果表明,MMPI 下各平均数形式的选题策略不但可以提高测验精度,还可以提高题库使用率,更好地满足约束条件. 本文引用了项目过低曝光率这个评价指标,这个指标当然越低越好,然而文献[11]中项目过低曝光率高达 0.8 以上,由于目前对于这个指标的研究还较少,所以它到底多大才算合适,目前还没有统一的认识. 对于试验而言,项目过低曝光率和项目从未曝光率这 2 项指标显示题库当中仍然有一半左右的题目用得很少,并且 30% ~ 40% 题目从未用到,这就造成了题库的极大浪费,如果 MMPI 下各平均数选题策略和程小扬提出的引入曝光因子[12]、李萍提出的自动控制区分度[13]相结合,是否能够既满足非统计计量约束,又提高题库利用率,这是一个有趣的问题. 另外,本实验只考虑了 0-1 评分下 3PLM 模型,对 GRM 和 GPCM 模型值得进一步研究.

3 参考文献

[1] Lord F M. Application of item response theory to practical testing problems [M]. Hillsdale NJ: Erlbaum Associates, 1980.

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

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

[4] Cheng Ying, Chang Huahua. The maximum priority index method for severely constrained item selection in computerized adaptive testing [J]. British Journal of Mathematical and Statistical Psychology, 2009, 62(2): 369-383.

[5] Buyske S. Optimal design in educational testing [J]. Applied Optimal Designs, 2005, 46(2): 1-19.

[6] Cheng Ying, Chang Huahua, Douglas Jeffery, et al. Constraint-weighted *a*-stratification for computerized adaptive testing with nonstatistical constraints [J]. Educational and Psychological Measurement, 2009, 69(1): 35-49.

[7] 潘奕娆,丁树良,尚志勇. 改进的最大优先级指标方法[J]. 江西师范大学学报: 自然科学版, 2011, 35(2): 213-215.

[8] 潘奕娆. 改进的最大优先指标及在计算机化自适应诊断测验中的应用 [D]. 南昌: 江西师范大学, 2011.

[9] Kingsbury G G, Zara A R. Procedures for selecting items for computerized adaptive test [J]. Applied Measure in Education, 1991(2): 359-375.

[10] Cheng Ying. Computerized adaptive testing: new developments and applications [D]. Urbana-Champaign: University of Illinois, 2008.

[11] Cheng Ying, Jeffrey M P, Can Shao. *a*-stratified computerized adaptive testing in the presence of calibration [J]. Educational and Psychological Measurement, 2014, 21(1): 1-24.

[12] 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.

[13] 李萍,甘登文,丁树良,等. 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.

The Comparison of Item Selection Strategies Based on Average Type

LI Jia, DING Shuliang, FANG Jianying

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

Abstract: A new item selection strategies named MMPI with four kinds average types on 3PLM has been put forward. The results of Monte Carlo simulations of fixed length tests and variable length tests show that the MMPI with four kinds average type approaches are more ideal in the performance of improvement the test precision, control the exposure uniformity, reducing the average number of violations and improve the bank utilization. Especially, the MMPI with root square of arithmetic average type is the best one.

Key words: item selection strategy; improved maximum priority index; average type

(责任编辑: 冉小晓)