最大信息量选题策略的自适应测试系统①

王 鹏, 荆永君, 王海敏

(沈阳师范大学 教育技术学院, 沈阳 110034)

摘 要: 随着心理与教育测量理论以及其与计算机技术相结合的不断研究和发展, 基于计算机的自适应测试成为了一种新型的测试形式.在分析、阐述计算机自适应测试理论和最大信息量选题策略的基础之上, 设计与实现了一个基于最大信息量法选题策略的自适应测试系统. 通过实验对被试者估计能力值、总信息量值、最大信息量值、试题参数等的变化进行分析与总结, 进而证明本系统的有效性. 最后对最大信息量选题策略研究做出总结, 针对其存在试题曝光度较高的问题给出了解决思路.

关键词: 教育测量; 自适应测试; 最大信息量选题策略

Computer Adaptive Test System Based on the Selection Strategy of Item Maximum Information

WANG Peng, JING Yong-Jun, WANG Hai-Min

(School of Education Technology, Shenyang Normal University, Shenyang 110034, China)

Abstract: With the continuous research of the psychological and educational measurement theory combined with computer technology, based on computer adaptive test become a new type of test form. On analysis of the computer adaptive theory and the selection strategy of the item maximum information, the paper designs and implements an adaptive test system. In experiment, the value of student's evaluated capacity, the total information, the maximum information and the parameters of item are analyzed and summarized. The experiment shows that the system is effective.

Key words: psychological and educational measurement theory; computer adaptive test; selection strategy of the item maximum information

随着心理与教育测量理论以及其与计算机技术相结合的不断研究和发展, 自适应测试的选题策略也出现了许多新的方法, 如按 a 分层法[1]、按 b 分层法、按 c 分层法[2]、按选题系数选题法等等, 都是以单维性假设为主的对被试者所测量得到的单一的能力为主要因素, 而忽略了在实际测量中, 被试者在测试上的表现往往受多种因素的影响, 如人格特质、情绪状态、成就动机. 在实际应用中不可能把所有因素都考虑在内, 只有考虑主要因素, 降低其他不可测量因素的影响, 尽量使测试的结果能够更精确、更稳定、更有效, 使测试效果更好[3]. 本文设计、开发了一个基于最大信息量法选题策略的自适应测试系统, 通过实验对每次估计能力值变化、总信息量值变化、最大信息量值变化、试题参数变化进行分析与总结; 通过获得的估计能力值的准确性对最大信息量选题策略做出评价; 并对最大信息量法选题策略存在试题曝光度较高的问题, 给出了解决思路.

1 计算机自适应测试

计算机自适应测试是建立在现代测试理论——项目反应理论基础之上, 题库的建立、参数估计、测试的终止条件、试题的选择都是以项目反应理论为指导进行的.

项目反应理论建立在潜在特质理论之上, 其有两个基本概念: 潜在特质和项目特征曲线. 潜在特质指的是制约被试者在测试上的表现行为的某一种心理品质,

① 基金项目:全国教育科学规划教育部重点课题(GFA111026)
收稿时间:2012-10-31;收到修改稿时间:2012-12-02

也可以解释为一组对被试者在测试上表现行为进行预测或解释的因素. 项目特征曲线是揭示被试者在测试项目上的反应行为与被试者潜在特质之间的关系.

项目反应理论的数学模型有很多, 本文主要应用三个参数 Logistic 模型(3PLM)[4], 其公式为:

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

其中, θ 为被试者能力水平值, a_i 为试题 i 的区分度, b_i 为试题 i 的难度, c_i 为试题 i 的猜测系数, D 为常量且 D=1.7, $P(i)$为答对第 i 道试题的概率.

计算机自适应测试的基本思想是: 计算机先通过一些试探性试题来初步估计学生的能力水平, 再根据选题算法从题库中选择与学生能力相近的试题继续测试, 每测试一题都重新估计学生的能力, 并不断重复这一过程, 直到满足终止条件为止.

计算机自适应测试必须解决的四个问题是题库建设、选题策略、参数估计、终止条件. 本文主要研究选题策略这个问题.

2 最大信息量选题策略

最大信息量选题策略是一种常用的自适应测试选题方法, 其基本思想是根据当前被试者的能力值选择所有试题中信息量最大的试题, 呈现给被试者作答. 基本步骤如下:

(1) 设定被试者初始能力值.

(2) 利用当前的能力值计算题库每道试题信息量.

(3) 选取信息量最大的试题作为下一道试题呈现给被试者.

(4) 根据被试者的作答情况重新估计被试者的能力值.

(5) 返回(2), 直到满足设定的终止条件为止.

终止条件有两种: 不定长度测验和定长度测验. 定长度测验需要设定测试的试题数量, 被试者完成规定数量即停止测试. 不定长度测验需要设定一个终止参数, 即所有被抽取试题的信息量总和的标准差小于某一预先确定的值.

3 系统设计与实现

本系统采用 Java 和 MySQL 数据库实现, 使用 Apache Tomcat 6.0 作为 Web 服务器, 系统包含用户管理、题库管理、考试管理等基本功能, 下文介绍系统实现的几个关键技术.

(1) 设定试题参数

本系统题库中试题的区分度 a、难度 b 和猜测系数 c 的初始值是根据经典测试理论计算得出的[5]. 首先选取一定数量的被试群体进行实验测试, 然后采用极端分组法计算试题的难度和区分度. 即先根据被试的测验总分从高到低排序, 然后在两端分别截取人数比例相等的(一般取 27%比例)高分组和低分组, 分别计算这两组在每一个题目上的通过率, 求其平均值作为该试题的难度指标, 其公式为:

$$P=\frac{P_H+P_L}{2}$$

求其差值作为试题的区分度指标, 其公式为:

$$D=P_H-P_L$$

上式中, P_H 和 P_L 分别代表高分组与低分组在同一个试题上的通过率, 且 $P_H=\frac{R_H}{n}, P_L=\frac{R_L}{n}$. n 代表高(低)分组人数, R_H 和 R_L 分别代表高、低分组答对该题目的人数.

试题区分度 a 的取值范围一般在(-1,1)之间, a 值越大, 试题质量越好. 试题难度 b 的取值范围一般在(0, 1)之间, b 值越大, 试题越难. 本系统选取的试题区分度 a 的取值设定在(0.30, 0.45)之间, 难度 b 的取值设定在(0.35, 0.65)之间, 且平均值在 0.5 左右. 试题猜测系数 c 值越大, 试题越容易被猜对. 本系统试题采用四选一的选择题形式, 所以试题猜测系数 c 的取值设定在(0, 0.25]之间.

(2) 设定被试者初始能力值

一般情况下, 被试者的初始能力值是未知的, 通常的做法是从题库中抽取一个中等难度的试题让被试者作答, 若回答正确则抽取稍难试题; 若回答错误则抽取稍容易试题. 如果经过 m(m 值可由用户自定义)道试题后, 考生全部答对则补测一道最难的试题; 如果全错, 则补测一道最容易的试题. 这样经过 m 或者 m+1 道试题后, 用公式 $\theta_0=\frac{x}{z-x}$(x 是答对试题数, z 是作答试题总数)估算出被试者的初始能力值[5].

(3) 计算试题总信息量

试题总信息量是所有试题信息量的总和, 其计算公式为:

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

式中，m 为试题总数，$I_i(\theta)$ 为每道试题的信息量，其计算公式为:

$$I_i(\theta)=\frac{1.7^2a_i^2(1-c_i)}{\left[c_i+e^{1.7a_i(\theta-b_i)}\right]\left[1+e^{-1.7a_i(\theta-b_i)}\right]^2}$$

式中，a_i、b_i、c_i、θ 与 Logistic 模型中的参数含义一致[6].

(4) 计算被试者能力值

本系统用牛顿—拉夫逊迭代(N-R)方法，根据被试者当前的能力值 θ_k 重新计算被试者答题之后获得的新能力值 θ_{k+1}. 其应用主要公式为:

$$\theta_{k+1}=\theta_k-\frac{g(\theta_k)}{g'(\theta_k)}k=0,1,2,3\cdots$$

$$g(\theta_k)=D\sum_{i=1}^{n}\frac{a_k(u-p_0)(p_0-c_k)}{p_0(1-c_k)}$$

$$g'(\theta_k)=-D\sum_{k=1}^{n}\frac{a_k^2(uc_k-p_0^2)(p_0-c_k)(1-p_k)}{p_0^2(1-c_k)^2}$$

$g(\theta_k)$ 是对 θ 一阶导，$g'(\theta_k)$ 是对 θ 二阶导. 式中 D 为常量且 D=1.7, u 为被试者答题的正误情况, 1 代表答对试题, 0 代表答错试题，p_0 是答对试题的概率[7].

4 实验与结果分析

(1) 实验准备

实验试题采用“大学英语语法词汇(四级)”试题，题型为单选题，试题总数 1000 个，实验对象为 50 名大二学生，每人进行一次测试. 系统在每次实验中分别记录被试者的能力值、总信息量、最大信息量、试题参数等数据，目的是对各项数据变化进行分析与总结，对系统的有效性做出评价.

实验中被试者的初始试题为随机抽取一个中等难度的试题，终止条件选择定长度测验，长度为 30.

(2) 数据分析

通过对记录的 50 名被试者的实验数据分析，我们发现被试者能力值变化、总信息量变化、最大信息量变化都有共同的规律，下面对其中一位被试者 T1 的实验数据做以分析.

① 能力值变化

在完成测试之后，被试者 T_1 能力值变化情况如图 1.

由图 1 可知，在第 23 题之前，被试者能力值变化较大，之后逐渐趋于稳定，最后较快地收敛于某一值. 这表明最大信息量选题策略是一种用时较短，使用试题数量较少的一种选题策略.

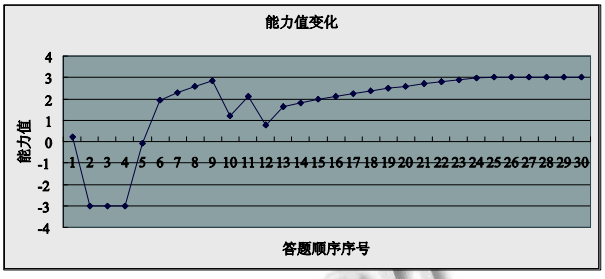

图 1 被试者能力值变化

② 总信息量变化

为被试者 T_1 选取测试试题时根据其当前的能力值，计算获得的总信息量，选取其有效部分，从第 5 题到第 30 题变化情况如图 2.

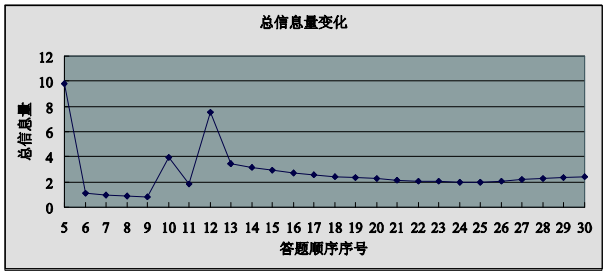

图 2 总信息量变化

由图 2 可知，总信息量的变化在测试初始阶段变化较大，逐渐趋于稳定，这与被试者能力值的变化相同.

③ 最大信息量变化

计算题库中剩余的试题信息量，抽取出具有最大信息量的试题，呈现给被试者，选取其最大信息量值有效部分，从第 5 题到第 30 题变化情况如图 3.

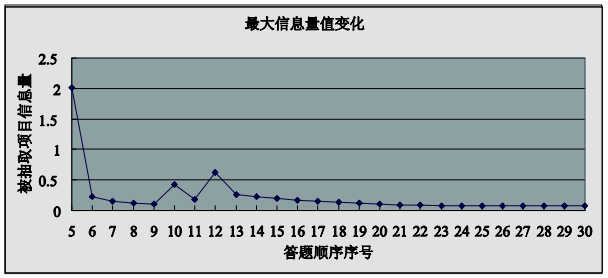

图 3 最大信息量值变化

由图 3 可知，最大信息量与总信息量变化趋势相同，在初始阶段变化较大，最后趋于稳定.

④ 试题参数变化

被试者 T_1 答题过程中，为其选取的测试试题的 a、b、c 参数的变化如图 4.

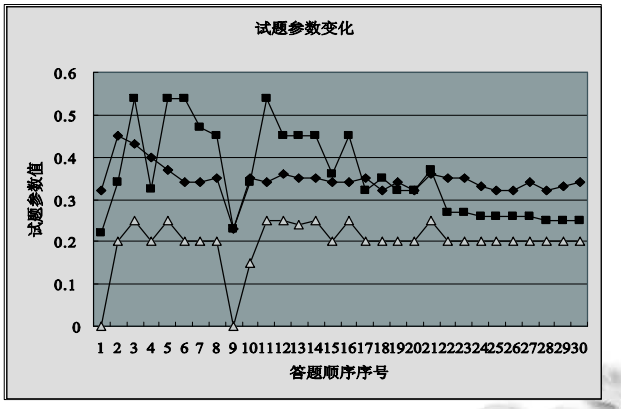

图 4　试题参数变化

由图 4 可知，试题的 a、b、c 值，在测试初始阶段变化较大，逐渐都趋于稳定.

⑤ 试题抽取情况

在实验中题库内试题抽取情况统计结果如下：被抽取 50 次以上的有 170 道试题；被抽取 1 次的有 40 道试题；总共被抽取的试题有 430；从未被抽取的有 570 道试题. ，经常被抽到试题是区分度较大、难度较大，因而试题曝光率较大，而区分度、难度较小的试题很少被使用，试题曝光率较小，这体现了最大信息量选题策略的局限性. 由于本系统的测验长度为 30，可以适当增加测验长度使测验结果更精确，同时也可以增加未被使用试题的曝光率.

(3) 实验小结

在实验中由于初始阶段(尤其是前 5 题)提供给被试者的试题的参数变化较大，被试者的能力值、总信息量和最大信息量在初始阶段变化都较大，最后都逐渐趋于稳定，表明了本系统是有效的.

本系统的性能采用测验效率和测验偏差两个指标作出评价，测验效率越大、测验偏差越小，则系统性能越好. 测验效率公式为[4,5]:

$$E=\frac{\sum_{j=1}^{M}\inf_j}{\sum_{j=1}^{M}L_j}$$

测验偏差公式为:

$$Bias=\frac{1}{M}\sum_{j=1}^{M}\left(\hat{\theta}_j-\theta_j\right)$$

其中，M 为测试总数，$\inf_j$ 为被试者 j 总信息量，L_j 为被试者的测验长度，$\hat{\theta}_j$ 为被试者的能力估计值，θ_j 为被试能力真值.

通过实验数据计算获得本系统的测验效率平均值为 0.25，测验偏差平均值为 1.5869. 由此可以看出本系统的性能指标尚可，基本可以满足实际需求，但仍有很大提升空间.

5　总结

通过以上对最大信息量选题策略的自适应测试研究，可以看出:

(1) 总信息量、最大信息量、试题参数会随能力值的变化而变化，并逐渐收敛于某一数值.

(2) 被试者答对的题数越多，能力值逐渐趋近于 3，答对题数与能力水平值成正比. 反之，能力值趋近于-3.

(3) 试题的区分度、难度、猜测系数影响到试题的曝光率. 区分度相对较大的试题经常被抽到. 在区分度和猜测系数相同的情况下，难度值较大的试题经常被抽到. 这也就是说明，试题区分度和猜测系数相同时，难度值越大，该试题的信息量越大.

针对最大信息量选题策略的曝光率的局限性，可以使用如下解决方法:

(1) 增加题库的容量，新增加试题的参数设置与经常被抽到的试题的参数值相接近.

(2) 在不增加题库容量前提下，适当增加测验长度.

(3) 增加筛选机制，将未被经常抽取的试题筛选掉.

参考文献

1 程小扬,丁树良.子题库题量不平衡的按 a 分层选题策略.江西师范大学学报(自然科学版),2011,35(1):6–9.

2 王茜娟,丁树良,谭渊.按 c-分层不定长 CAT 的研究.江西师范大学学报(自然科学版),2005,29(3):228–230.

3 戴海琦,陈德枝,丁树良,邓太萍.多级评分题计算机自适应测验选题策略比较.心理学报,2006,38(5):778–783.

4 金瑜.心理测量.上海:华东师范大学出版社,2001.

5 漆书青,戴海琦,丁树良.现代教育与心理测量学原理.北京:高等教育出社,2002.

6 Chang HH. A Global Information Approach to Computerized Adaptive Testing. Applied Psychological Measurement, 1996,20(3):213–229.

7 Kingsbury G, Zara A. Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education, 1989,2(4):359–375.