文章编号: 1000-5862(2014)05-0441-04

不定长认知诊断计算机化自适应测验终止规则研究

艾国金,甘登文*,丁树良,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: 目前大多数认知诊断计算机化自适应测验采用定长终止规则,该文提出 3 种不定长终止规则. 通过实验对比,新方法较已有方法具有如下优势: i) 能够较好地保障测验精度; ii) 大幅降低了人均测验用题数; iii) 人均测验用时也有所下降; iv) 项目调用均衡性和题库安全性有所增强.

关键词: 认知诊断; 计算机化自适应; 不定长终止规则; Monte Carlo 模拟; 题库安全
中图分类号: B 841.7; TP 301.6　　**文献标志码**: A

0 前言

教育认知诊断由于可以为家长、老师和学生提供诊断信息,为因材施教提供参考和指导,因而备受国内外研究者和应用者的青睐. 认知诊断计算机化自适应测验(cognitive diagnosis computerized adaptive testing,CD_CAT)运用能够体现"因人施教、量体裁衣"的 CD_CAT 选题策略和终止规则,根据被试当前的状态自适应匹配项目进行测验,获得被试对项目的反应,通过反应快速、准确地诊断出被试对测验所涉及属性的掌握情况[1]. 近些年来,国内外对CD_CAT 的研究越来越多,也越来越深入,与传统 CAT 不同的是目前 CD_CAT 中还没有找到类似 Fisher 信息量指标衡量测量误差,因此认知诊断 CAT 通常采用施测起来较为方便的定长 CAT 的形式作为其终止规则或者采用其他指标作为不定长的终止规则.

目前对不定长 CD_CAT 终止规则的研究并不多,如 C. Tatsuoka[2] 建议如果被试的后验概率达到 0.8 以上,测验终止; Cheng Ying[3] 则建议当后验的 SHE 值或邻近 SHE 值的变化足够小时,或邻近 2 次后验 KL 距离足够小时,测验终止; C. L. Hsu 等[4] 通过大量实验提出当最大潜在模式后验概率大于某个预定的值(如 0.7)或当最大潜在模式后验概率大于某个预定的值(如 0.7)且第 2 大潜在模式后验概率小于某个预定值(如 0.1)时,测验终止; 郭磊等[5] 则认为当邻近后验概率之差等于某个足够小的值或属性标准误之差足够小时,测验终止. 以上方法通过模拟实验都获得了较好的效果. 不定长 CD_CAT 至少在用题量方面可能比定长情形要节省一点,本文讨论 CD_CAT 的新的终止规则.

1 不定长 CD_CAT 终止规则

1.1 模型介绍

DINA 模型(deterministic inputs,noisy and gate model)表达式为

$$P(X_{ij}=1 \mid \alpha_i)=(1-s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}, \quad (1)$$

其中 α_i 为被试 i 的知识状态,$\eta_{ij}=\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}$ 描述被试 i 是否掌握项目 j 所考察的所有属性. 若 $\eta_{ij}=1$,说明被试 i 掌握了项目 j 所考察的全部属性; 若 $\eta_{ij}=0$,则说明被试 i 对项目 j 所考察的属性至少有 1 个未掌握,q_{jk} 为项目 j 所考察的属性分量,其值为 0 或 1. 若 $q_{jk}=1$ 说明项目 j 考察了第 k 个属性; 若 $q_{jk}=0$ 则说明项目 j 未考察第 k 个属性.

$s_j=P(X_{ij}=0 \mid \eta_{ij}=1)$ 表示被试 i 在掌握了项目 j 所考察的全部属性的情况下,答错项目 j 的概率,通常称为失误参数,$g_j=P(X_{ij}=1 \mid \eta_{ij}=0)$ 表示被试 i 在未全部掌握项目 j 所考察所有属性的情况下,答对项目 j 的概率,通常称为猜测参数.

1.2 选题策略和知识状态估计方法

对各种不同终止规则本文均采用尚志勇等[6]

收稿日期: 2014-06-21
基金项目: 国家自然科学基金(30860084,31160203,31100756,31360237),国家社会科学基金(12BYY055,13BYY087)和江西省教育厅科技计划(GJJ13208,GJJ13206,GJJ13227,GJJ13208,GJJ13209)资助项目.
通信作者: 甘登文(1956-),男,江西奉新人,教授,主要从事计算机辅助教学和统计应用方面的研究.

提出的按属性模式分层选题策略作为模拟试验的CD_CAT选题策略,利用MAP方法估计被试的知识状态,即将在作答模式 X_i 已知的条件下先计算被试各种可能的知识状态对应的后验概率分布,然后将具有最大后验概率对应的知识状态作为被试知识状态的估计值,公式为

$$\hat{\alpha}_i = \arg\max_{\alpha_c}(P(\alpha_c \mid X_i)); c = 1,2,3,\cdots,2^K, \quad (2)$$

其中 $P(\alpha_c \mid X_i) = \dfrac{P(X_i \mid \alpha_c)P(\alpha_c)}{\sum_{c=1}^{2^K} P(X_i \mid \alpha_c)P(\alpha_c)}$ 为知识状态,为 α_c 的后验概率.

1.3 不定长终止规则

1.3.1 Hsu等方法 当被试属于某个知识状态的最大后验概率 P_{1st} 大于某个预定的值(如0.7)并且第2大后验概率 P_{2nd} 小于某个预定值(如0.1)时,测验终止,并给出了第2大后验概率的计算公式[4]:

$$P_{2nd} = (1 - P_{1st})/(2^K - 1) + ((2^K - 2)(1 - P_{1st})d)/(2^K - 1), 0 \leqslant d \leqslant 1, \quad (3)$$

其中 K 为考察属性个数,通常 d 根据需要取值,Hsu等在模拟实验中 d 取0,0.25,0.5和0.75.

1.3.2 邻近后验概率之差法 邻近后验概率之差法(difference of the adjacent posterior probability method,DAPP)[5] 规定在测试过程中当出现从属于同一个知识状态的前后2次邻近的最大后验概率差的绝对值小于某个预设值时,测验终止.

1.3.3 3种新终止规则 由于被试 i 每做一题,其不同潜在模式的后验概率就会更新一次.因此,若被试 i 做了 t 题,则不同潜在模式的后验概率累积的更新次数更多.对于好的选题策略,t 越大最接近被试 i 真实知识状态的潜在模式后验概率值会越来越大,其他潜在模式的后验概率值则会越来越小.受Hsu等方法2和DAPP法的启发,本文给出几种新的终止规则.

方法1 被试 i 测验 t 题后观察其最大后验概率与第2大后验概率之差,若差值足够大,则说明被试 i 能够较好地区分最大后验概率值对应的知识状态和其他潜在知识状态.最大后验概率与第2大后验概率之差 M 大于某个预设值,计算公式为

$$PPLS_{1st} = \max_{\alpha_c}(P(\alpha_c \mid X_i)), \quad (4)$$

$$PPLS_{2nd} = \max_{\alpha_c \neq \alpha_{PPLS_{1st}}}(P(\alpha_c \mid X_i)), \quad (5)$$

$$M = PPLS_{1st} - PPLS_{2nd}. \quad (6)$$

方法2 若最大后验概率与最小后验概率之差值足够大,则说明被试 i 在作答最大后验概率对应的项目时,其答对的概率非常大.这也说明对被试 i 能够较好地区分最大后验概率值对应的知识状态和其他潜在知识状态.最大后验概率与最小后验概率之差 N 大于某个预设值,计算公式为

$$PPLS_{\min} = \min_{\alpha_c}(P(\alpha_c \mid X_i)), \quad (7)$$

$$N = PPLS_{1st} - PPLS_{\min}. \quad (8)$$

方法3 如果方法1与方法2的差的绝对值,即第2大后验概率与最小后验概率之差的绝对值足够小,说明此时最大后验概率已足够大,按照MAP估计方法也能说明被试 i 能够较好地区分与自己真值接近的知识状态和其他潜在知识状态.方法1与方法2的差的绝对值小于某个预设值 ξ,计算公式为

$$|M - N| < \xi. \quad (9)$$

1.4 评价指标

本文使用模式判准率、人均测验用时、人均测验用题数、单个被试最大用题数和最小用题数、χ^2 统计量和测试重叠率作为考察指标.模式判准率(pattern match ratio,PMR),即被试掌握模式并判准的人数占总人数的百分比,计算公式为:$P_{MR} = N_P/N$,其中 N_P 指被试掌握模式并判对的人数,N 指总人数;$Time$ 为 N 个被试开始测验到结束测验的总耗时,S_{Items} 为 N 个被试总使用题数,人均测验用时:$T = Time/N$,人均测验用题数:$S = S_{Items}/N$,单个被试最大用题数和最小用题数,即被试在不同终止规则下在模拟实验过程中测验需要的最大题数和最小题数;χ^2 统计量是用来反映项目被调用的均匀性,χ^2 指标越小说明整个题库的使用越均匀,计算公式为

$$\chi^2 = \sum_{j=1}^{M}\sum_{t=1}^{J}\left\{\left[A_{jt} - \sum_{j=1}^{M}\sum_{t=1}^{J}(A_{jt}/M)\right]^2 / \sum_{j=1}^{M}\sum_{t=1}^{J}(A_{jt}/M)\right\}, \quad (10)$$

其中 A_{jt} 为第 j 个项目模式下的第 t 个题目的曝光率,计算 A_{jt} 的公式为 $A_{jt} = n_t/N$,n_t 为第 j 个项目模式下的第 t 个题目的使用次数.测试重叠率(Rt)也是用来衡量安全性的指标,计算公式为

$$Rt = 2\sum_{j=1}^{M}\sum_{t=1}^{J}C_{n_t}^2 / \left[(N-1)\sum_{i=1}^{N}L_i\right], \quad (11)$$

其中 L_i 为第 i 个人测试长度.

2 CD_CAT模拟实验

为验证新方法,本文在Window 7系统,内存2 GB的环境下,采用Matlab8.0(R2012b)为工具进行Monte Carlo模拟实验.实验中共考察了6个属性,分为4种结构:线型、收敛型、发散型、无结构型[7],如图1所示,依次为L、C、D、U.

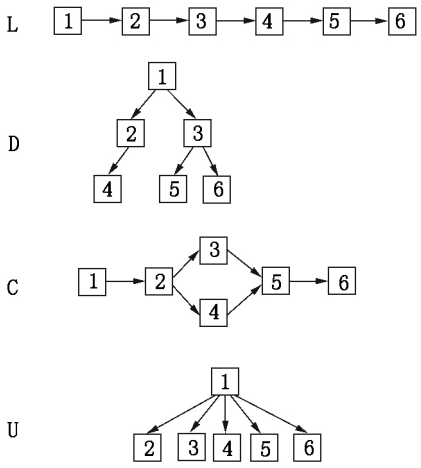

图 1　4 种属性层级结构图

被试人数设为 1 000 人，对于每种类型的属性层级结构，有相应的项目类 q_j（q_j 为潜在 Q 阵的某一列），每个项目类的属性相同但参数不同，每类模式的项目设为 100，项目的失误参数和猜测参数均服从均匀分布 $U(0.05,0.25)$，以此建立题库[8].

实验中将定长 $L=30$、Tatsuoka 提出的方法（以下简称 Tatsuoka 法）、Hsu 方法 2（其中 $P_{1st}>0.95$，$d=0.25$）作为参照终止规则，方法 1 中 $M>0.99$，方法 2 中 $N>0.99$，方法 3 中 $\xi=0.001$. 利用 Monte Carlo 模拟测验并重复 30 次求平均值的方法，得到 4 种结构下不同终止规则的模式判准率如表 1 所示，人均测验用时如表 2 所示，人均测验用题数如表 3 所示，单个被试最大用题数和最小用题数如表 4、表 5 所示，各方法的 χ^2 统计量、测试重叠率如表 6.

表 1　4 种结构下不同终止规则的模式判准率

属性层级结构	定长 $L=30$	Tatsuoka 法	Hsu 方法 2	方法 1	方法 2	方法 3
L	0.999 7	0.872 6	0.989 0	0.999 5	0.999 0	0.990 8
C	0.999 5	0.871 6	0.987 9	0.999 0	0.998 8	0.989 7
D	0.997 7	0.856 5	0.989 6	0.997 0	0.997 1	0.987 1
U	0.987 1	0.843 1	0.987 3	0.989 6	0.992 0	0.987 6

表 2　4 种结构下不同终止规则模拟实验人均测验用时　　单位: s

属性层级结构	定长 $L=30$	Tatsuoka 法	Hsu 方法 2	方法 1	方法 2	方法 3
L	0.002 3	0.000 45	0.001 8	0.001 3	0.001 1	0.001 0
C	0.002 5	0.000 55	0.001 9	0.001 6	0.001 3	0.001 2
D	0.006 0	0.001 80	0.004 9	0.004 4	0.004 0	0.003 8
U	0.029 1	0.011 10	0.025 2	0.024 5	0.023 5	0.022 4

表 3　4 种结构下不同终止规则模拟实验人均测验用题数　　单位: 个

属性层级结构	定长 $L=30$	Tatsuoka 法	Hsu 方法 2	方法 1	方法 2	方法 3
L	30	5.502 1	19.085 3	13.963 9	13.412 4	12.092 8
C	30	5.999 5	18.327 9	15.165 6	14.624 5	13.410 8
D	30	8.520 1	22.438 5	20.231 3	19.604 7	17.977 9
U	30	11.244 7	25.521 4	24.727 5	24.168 7	22.654 1

表 4　4 种结构下不同终止规则模拟实验单个被试最大用题数　　单位: 个

属性层级结构	定长 $L=30$	Tatsuoka 法	Hsu 方法 2	方法 1	方法 2	方法 3
L	30	20	30	30	30	30
C	30	29	30	30	30	30
D	30	30	30	30	30	30
U	30	30	30	30	30	30

表 5　4 种结构下不同终止规则模拟实验单个被试最小用题数　　单位: 个

属性层级结构	定长 $L=30$	Tatsuoka 法	Hsu 方法 2	方法 1	方法 2	方法 3
L	30	3	4	5	5	2
C	30	3	5	5	5	2
D	30	4	5	7	7	2
U	30	6	7	9	8	8

表 6　不同终止规则模拟实验 χ^2 指标和测试重叠率指标

指标	定长 $L=30$	Tatsuoka 法	Hsu 方法 2	方法 1	方法 2	方法 3
χ^2 指标	22.609 7	11.553 7	19.570 8	21.252 5	20.882 1	19.458 2
Rt 指标	0.011 7	0.003 0	0.008 5	0.007 1	0.006 8	0.006 2

从表 1 中可以得出:定长终止规则得到的模式判准率要比不定长终止规则得到的稍好些,但是表现出的优势十分有限;在不定长终止规则中方法 1、方法 2 和方法 3 要比 Tatsuoka 法和 Hsu 方法好,而方法 1 和方法 2 在不同属性层级结构下其模式判准率表现也各有优势. 从表 2、表 3 中可以看出:不定长终止规则的人均测验用时和人均测验用题数表现要优于定长终止规则,方法 3 的表现又优于其他终止规则;从表 4、表 5 中可以看出:不同终止规则在单个被试最大用题数上的表现几乎相当,在单个被试最小用题数上,不定长终止规则要优于定长终止规则. 从表 6 可以看出不定长终止规则 χ^2 指标和 Rt 指标都优于定长终止规则,结合前 5 个指标,在小幅度降低模式判准率的前提下,方法 3 的表现要优于其他终止规则. 考虑到 CD_CAT 要实现"快速、准确、安全"测验这个特点,综合表 1 ~ 表 6 可以得出方法 1、方法 2、方法 3 要优于其他方法.

虽然方法 1、方法 2 和方法 3 在上述 5 个指标上的表现都不错,但在不同指标上的优势却不尽相同. 新方法只讨论了在 DINA 模型下的表现情况,如果改成其他模型新方法[9-11]是否可用. 另外能否开发一个或多个不定长终止规则在上述 7 个指标上的表现都为最佳,这些都有待在未来研究中进一步探索.

3 参考文献

[1] 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.

[2] Tatsuoka C. Data analytic methods for latent partially ordered classification models [J]. Applied Statistics,2002,51(3):337-350.

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

[4] Hsu C L,Wang W C,Chen S Y. Variable-length computerized adaptive testing based on cognitive diagnosis models [J]. Applied Psychological Measurement,2014,4:6-7.

[5] 郭磊,边玉芳. 认知诊断计算机化自适应测验变长终止规则的研究 [C]//心理学与创新能力提升——第十六届全国心理学学术会议论文集,2013.

[6] 尚志勇,丁树良. 认知诊断自适应测验选题策略探新[J]. 江西师范大学学报:自然科学版,2011,35(4):418-421.

[7] Leighton J P,Gierl M,Hunka S M. The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's rule-space approach [J]. Journal of Educational Measurement,2004,41(3):205-236.

[8] 唐小娟,丁树良,毛萌萌,等. 基于属性层级结构的认知诊断测验的组卷 [J]. 心理学探新,2013,33(3):252-259.

[9] 丁树良,罗芬,汪文义. 多级评分认知诊断测验蓝图的设计——独立型和收敛型结构 [J]. 江西师范大学学报:自然科学版,2014,38(2):265-269.

[10] 丁树良,罗芬,汪文义. 多级评分认知诊断测验蓝图的设计——根树型结构 [J]. 江西师范大学学报:自然科学版,2014,38(2):111-118.

[11] 艾国金,甘登文,丁树良. 计算机化自适度认知诊断测验按模式分层选题策略 [J]. 江西师范大学学报:自然科学版,2014,38(3):270-273.

Research on Variable-Length Termination Rules for Computerized Adaptive Testing with Cognitive Diagnosis

AI Guo-jin,GAN Deng-wen*,DING Shu-liang,XIONG Jian-hua

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

Abstract: At present,most computerized adaptive testing with cognitive diagnosis adopt fixed-length termination rule. Three kinds of variable-length termination rules have been presented. Compared with the present methods,the new methods have the following advantages according to simulated experiments: i) guarantee the test accuracy properly; ii) lower the per capita testing items; iii) decrease per capita testing time; iv) enhance the balance of extracting items and security of item pool.

Key words: cognitive diagnosis; computerized adaptive testing; variable-length termination rules; monte carlo simulation; securer item pool

(责任编辑:冉小晓)