文章编号: 1000-5862(2012)06-0632-04

基于等级反应模型的广义距离判别法

李　娟, 丁树良*, 罗　芬

(江西师范大学计算机信息工程学院, 江西　南昌 330027)

摘要: 受0-1评分广义距离判别法的启发, 给出一种新的多级评分认知诊断方法.蒙特卡洛模拟实验结果显示: 基于等级反应模型的广义距离判别法(GRM-GDD)比属性层级模型(GRM-AHM)有更高的模式判准率, 尽管随着被试作答失误率的提高, 模式归准率均有所下降.

关键词: 认知诊断; 广义距离判别法; 等级反应模型
中图分类号: O 626.4　　文献标志码: A

0　引言

认知诊断能够揭示被试的认知状况, 以便对其不足有针对性地补救, 这对教师开展因材施教具有重要意义.认知诊断模型(cognitive diagnostic model, CDM)有助于诊断被试对每个属性的掌握情况, 其中规则空间模型和属性层次方法是 2 种常见的认知诊断模型.规则空间模型(rule space model, RSM)是 K.K. Tatsuoka 等[1]提出的一种认知诊断方法.RSM 主要包括了 $\boldsymbol{Q}$ 矩阵理论和分类识别 2 个部分, 引入 $\boldsymbol{Q}$ 矩阵理论是为了建立不可观察的属性掌握模式(知识状态)和观察反应模式(observed response pattern, ORP)之间的联系.RSM 的分类识别是根据项目反应理论和多元分析中模式识别的原理构造一个规则空间, 将被试的观察反应模式与理想反应模式都转化为规则空间的点, 理想反应模式对应于规则空间中的点称为纯规则点, 它是分类判别的类中心.通过比较观察反应模式对应于规则空间中的点与纯规则点的马氏距离的大小来对观察反应模式进行判别, 以达到认知诊断的目的.RSM 可以根据已有的项目来构造 $\boldsymbol{Q}$ 矩阵.J.P. Leighton 等[2]认为这样构造的 $\boldsymbol{Q}$ 阵可能没有反映所测属性之间真正的层级关系, 这会影响诊断的精确性, 因而他们在 RSM 的基础上提出了一种新的认知诊断模型: 属性层级方法(attribute hierarchy method, AHM).

AHM 认为认知技能是一个相互联系的加工网络, 通常是非孤立操作的.认知属性被假设是具有某种层次关系, 其中属性定义为正确求解测试项目所要求的基本认知过程或技巧[3].AHM 有利于指导测验的编制和开发, 因为一旦确定了某一领域的属性及属性层级结构, 测验开发者就能够根据属性的层级结构来编写测验项目.

丁树良等[3]在 AHM 改进的 $\boldsymbol{Q}$ 矩阵理论的基础上, 指出 K.K. Tatsuoka 的 $\boldsymbol{Q}$ 矩阵理论仍存在缺陷且该缺陷可能影响判别分类结果, 并对 $\boldsymbol{Q}$ 矩阵理论作了进一步补充和修正, 给出如下定理.

定理 **1**　在 0-1 评分方式下, 如果认知加工的特点是非补偿、连接的, 那么欲使观察反应模式和期望反应模式建立双射当且仅当可达矩阵是认知诊断测验蓝图(测验 $\boldsymbol{Q}$ 矩阵 $\boldsymbol{Q}_t$)的子矩阵.

孙佳楠等[4]在上述定理的基础上, 提出了一种新的认知诊断方法: 广义距离判别法(generalized distance discrimination, GDD).一方面它是遵循 RSM 和 AHM 的认知诊断思路, 定义了一种广义距离(generalized distance between response Patterns, GDRP)来度量观察反应模式和理想反应模式之间的距离, 根据距离最小准则获得对应的期望反应模式; 另一方面 GDD 根据定理 1 设计认知诊断测验, 然后由定理 1 找到和期望反应模式对应的知识状态, 即将被试的知识状态进行归类, 从而实现诊断.

本研究则是受孙佳楠等[4]提出的适用于 0-1 评分模型的广义距离判别法的启发, 推导出了适用多级评分模型的 GDD.本文使用的项目反应模型为 F. Samejima[5]等级反应模型(grade response model,

收稿日期: 2012-05-09
基金项目: 国家自然科学基金(30860084, 31160203, 31100756)资助项目.
作者简介: 李树良(1949-), 男, 江西樟树人, 教授, 博士生导师, 主要人事计算机辅助教学, 应用及教育和心理测量方面的研究.

GRM), 简记为 GRM-GDD.

1　0-1 评分广义距离判别法(GDD)

0-1 评分广义距离判别法的广义距离定义为[4]

$$d(Y_\alpha, X_\beta)\triangleq\sum_{j=1}^{J} d(Y_{\alpha j}, X_{\beta j}), \tag{1}$$

其中

$$d(Y_{\alpha j}, X_{\beta j}) = \left|Y_{\alpha j} - X_{\beta j}\right| P_j(\theta_\alpha)^{Y_{\alpha j}}(1-P_j(\theta_\alpha))^{1-Y_{\alpha j}}, \tag{2}$$

其中 J 为项目个数, $Y_\alpha = (Y_{\alpha 1}, \cdots, Y_{\alpha j})$ 表示被试的观察反应模式; $X_\beta = (X_{\beta 1}, \cdots, X_{\beta j})$ 表示第 β 种理想反应模式; $d(Y_{\alpha j}, X_{\beta j})$ 表示项目 j 上被试 α 的观察反应 $Y_{\alpha j}$ 与第 β 种理想反应 $X_{\beta j}$ 的广义距离; $d(Y_\alpha, X_\beta)$ 表示 Y_α 到 X_β 的广义距离, 即所有项目的广义距离之和; $P_j(\theta_\alpha)$ 表示为具有能力水平 θ_α 的被试 α 正确作答项目 j 的概率, 如用 IRT 模型中的 2PLM 具体化概率函数, 甚至可以是 DINA 模型的项目反应函数.

2　GRM-GDD 模型的分类方法

F. Samejima[5]在 Logistic 模型的框架下, 建立了可用于多级评分的等级反应模型(GRM), 突破了过去项目反应理论只能用于 0-1 评分项目的限制.等级反应模型假设每个项目只有一个区分度值、有多个难度等级值, 而且每个项目在各个等级上的难度值是严格单调递增的.若项目 j 满分为 f_j, 即有 f_j+1个等级, 则 $b_{j1} < b_{j2} < \cdots < b_{j,f_j+1}$. F. Samejima[5]提出分 2 个步骤获得能力为 θ 的被试 α 在某个项目上恰好得某个等级分的概率: (i) 能力为 θ_α 的被试 α 在第 j 个项目上得分不低于 h 分的概率表示为

$$\begin{aligned} P^*_{\alpha jh} &= \{1+\exp[-D_{\alpha j}(\theta_\alpha - b_{jh})]\}^{-1}, \\ P^*_{\alpha j0} &= 1, P^*_{\alpha j, f_j+1} = 0, \end{aligned} \tag{3}$$

其中 D 为量表因子, 一般取 1.7, α_j 为第 j 个项目的区分度. (ii) 能力为 θ_α 的被试 α 在第 j 个项目上恰好得 h 分的概率 $P_{\alpha jh}$ 表示为

$$P_{\alpha jh}(\theta) = P^*_{\alpha jh}(\theta) - P^*_{\alpha j,h+1}(\theta), h = 0,1,\cdots, f_j. \tag{4}$$

由于多级评分 GDD 是 0-1 评分 GDD 的推广, 故多级评分 GDD 判别法的核心仍是衡量被试的观察反应模式(ORP)与每种理想反应模式(IRP)之间的广义距离, 依据距离最小准则对被试的观察反应模式进行归类.对于多级评分 GDD, 设 $\boldsymbol{U}$ 为 0-1 矩阵, 当被试 α 在第 j 个项目上恰好得 h 分时 $U_{\alpha jh}$ 的值为 1, 否则为 0, 称 $\boldsymbol{U}$ 为观察得分矩阵的示性矩阵; $\boldsymbol{V}$ 亦为 0-1 矩阵, 当第 β 个理想反应模式在第 j 个项目上恰好得 h 分时 $V_{\beta jh}$ 的值为 1, 否则为 0, 称 $\boldsymbol{V}$ 为理想得分矩阵的示性矩阵.第 j 题上观察反应 $Y_{\alpha j}$ 与理想反应 $X_{\beta j}$ 的广义距离则是 0 f_j 分(共 f_j+1个等级)加权距离的总和, 即为 $\sum_{h=0}^{f_j}\left|U_{\alpha jh}-V_{\beta jh}\right| P_{\alpha jh}^{U_{\beta jh}}$, 其中 $P_{\alpha jh}$ 为被试 α 在第 j 题上恰好得 h 分的概率, 本文使用 Samejima 等级反应模型中的反应函数具体化 $P_{\alpha jh}$, 因此, 可定义多级评分 GDD(PGDD)的归类函数为

$$d(Y_\alpha, X_\beta)\triangleq\sum_{j=1}^{J} d(Y_{\alpha j}, X_{\beta j}) = \sum_{j=1}^{J}\sum_{h=0}^{f_j}\left|U_{\alpha jh}-V_{\beta jh}\right| P_{\alpha jh}^{U_{\alpha jh}}(\theta_\alpha). \tag{5}$$

3　模拟研究

为了比较 GRM-GDD 和 GRM-AHM 分类方法的分类效果, 进行 Monte Carlo 模拟研究.

3.1　实验设计

本文采用 J.P. Leighton 等[2]提出的 4 种属性层级结构(如图 1, 分别为发散型、线型、合流型和无结构型), 在 4 种被试作答失误率 *slip*(分别为 2%、5%、10%和 15%)下考虑这 4 种属性层级结构的诊断结果.即用 4×4 交叉设计, 共 16 个试验条件, 每个试验条件下取被试 5 000 人, 重复 30 次试验以减少误差.每个试验都对 2 种分类方法(GRM-GDD 和 GRM-AHM, 对 AHM 又分别考虑 A 方法和 B 方法)进行比较, 以考察 GRM-GDD 的表现.

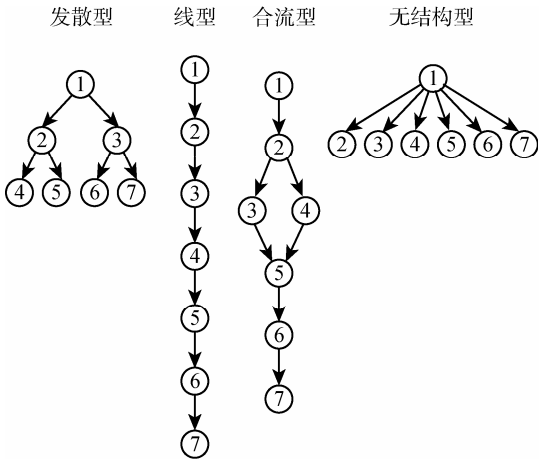

图 1　7 个属性的 4 种层次结构

3.1.1　测验 $\boldsymbol{Q}$ 矩阵的设计　应用杨淑群等[6]提出的

生成简化 $\boldsymbol{Q}$ 矩阵的扩张方法，以及根据丁树良等[7-8]给出的上述定理 1 的结论，运用祝玉芳等[9-10]关于多级评分 AHM 模型确定期望反应模式全集的方法，可得到包含 7 个属性产生的 $\boldsymbol{Q}_r$ 所包含的题数分别是发散型 25 个、合流型 8 个、线型 7 个、无结构型 64 个. 本研究设计了 2 个实验: (1)为了与祝玉芳等[9]实验结果有可比性，本文采用了相似模拟条件下的试验结果相比较，设计的测验 $\boldsymbol{Q}_t$ 矩阵的测验项目个数分别为 25, 8, 7, 64 题.(2)本文认为 7 个属性合流型和线型的测验项目数太少，而无结构型测验项目数太多，势必会影响诊断的效率.因而为了比较在不同的层级结构下 GRM-GDD 的效率，通过重复项目方法增加了合流型和线型的测验项目数，删减了一些无结构型的测验项目数，这 3 种结构的测验项目数为 28 题.

3.1.2 *被试作答矩阵的模拟* 假设每个被试的观察反应模式或者来自理想反应模式，或者是理想反应模式受到随机误差的“污染”.所以模拟这些被试的观察反应模式可以使用如下的方法: 对期望反应模式得分向量加上随机误差，造成 *slip*(这里的 *slip* 是指由于失误或猜测造成与期望项目作答反应不一致的反应)后所得到的反应向量作为被试观察反应模式. 把期望反应模式按总得分从小到大排序，然后使具有这些得分的被试人数满足标准正态分布，产生 5 000 个人进行分配，其中得分相同的期望反应模式平均分配人数.为了产生发生了 *slip* 的观察反应模式，按如下的方法模拟: 如要模拟每个模式的每个项目的得分有 5%的概率发生 *slip* 的情况，采用一个服从开区间(0, 1)上均匀分布 $U(0, 1)$的随机数 r, 如果 $r>0.95$ 且该得分不是满分，则该项目得分增加 1 分，如果 $r<0.05$ 且该项目得分不是 0 则该项目得分减 1 分，否则该项目得分不变，这样就模拟产生一个有 5%率的观察项目反应模式.

3.1.3 *归类判别* 自编程序用 EM 算法估计项目参数和 EAP 估计能力参数，然后分别用 GRM-GDD 和 GRM-AHM 的 A 方法，B 方法把观察反应模式归类到期望反应模式中.

3.1.4 *重复实验* 对同一种实验条件(在某种属性层级结构时发生某种 *slip*)，重复执行 3.1.2 节中的作答反应和 3.1.3 节中判别分类 30 次，30 次的属性模式归准率的平均值作为实验结果.

3.2 评价指标

用理想反应模式对应的知识状态作为真值，然后根据属性模式归类的正确率来比较方法的好坏. 如诊断测验共有 K 个属性(本实验 K=7)且有 N 个被试参加测验，被试 α 对应的理想反应模式为 X_α，而在发生 *slip* 后观察反应模式归类结果为 Z_α，如果 $X_\alpha = Z_\alpha$，令 $h_\alpha = 1$；否则 $h_\alpha = 0$，则属性模式的归准率为 $\sum_{\alpha=1}^{N} h_\alpha / N$.

3.3 实验结果

实验结果如表 1~表 4 所示.

表 1 发散型结构的属性模式归准率

分类方法	*slip*			
	0.02	0.05	0.10	0.15
GRM-GDD	0.999 5	0.999 0	0.983 7	0.912 9
A	0.973 2	0.939 0	0.872 5	0.805 6
B	0.652 3	0.346 5	0.132 1	0.058 2

表 2 合流型结构的属性模式归准率

分类方法	*slip*			
	0.02	0.05	0.10	0.15
GRM-GDD	0.989 0	0.969 1	0.918 8	0.840 4
A	0.928 1	0.903 0	0.852 1	0.832 8
B	0.888 0	0.747 1	0.561 2	0.411 6

表 3 线型结构的属性模式归准率

分类方法	*slip*			
	0.02	0.05	0.10	0.15
GRM-GDD	0.994 6	0.980 1	0.938 5	0.872 0
A	0.953 2	0.914 2	0.895 5	0.835 6
B	0.902 3	0.776 5	0.601 1	0.448 2

表 4 无结构型结构的属性模式归准率

分类方法	*slip*			
	0.02	0.05	0.10	0.15
GRM-GDD	1.000 0	0.999 8	0.999 4	0.998 3
A	0.983 2	0.952 0	0.901 5	0.879 6
B	0.341 3	0.097 5	0.043 1	0.034 2

表 5 **28** 题合流型、线型、无结构型属性模式归准率

层次结构	*slip*			
	0.02	0.05	0.10	0.15
合流型	0.999 8	0.999 6	0.995 4	0.969 0
线型	0.999 9	0.999 6	0.996 4	0.980 4
无结构型	0.999 9	0.996 6	0.959 0	0.837 6

根据表 1~表 4 显示的 4 种结构的属性模式判准率可知，发散型、合流型、线型、无结构型的 GRM-GDD 在同等条件下都比 GRM-AHM 要好，当然随着 *slip* 的增大，GRM-GDD 和 GRM-AHM 方法的 4 种结构类型归准率都有所下降，但是总体来看，同种条件下 GRM-GDD 方法的归类效果更好.由 3.1 节的试验设计，知表 1~表 4 中不同层级对应的测验的长度不同，所以不同层级的结果就不具有可比性. 而表 5 消除了测验长度的影响，所得到的结果具有可比性.由表 5 也可以看出 GRM-GDD 的 3 种不同类

型结构的模式归准率的效果都很好.结合表 1 中发散型结构的属性模式归准率, 它的测验长度为 25, 尽管测验长度小于表 5, 但是结果还是比较好.

4　讨论

本文受到孙佳楠等[9]的 0-1 评分 GDD 的启发, 给出了基于 GRM 的 PGDD.如果某个被试在项目 j 上观察得分与理想得分一致, 则(5)式中相应的距离等于 0, 否则在一个项目上, (5)式中绝对值有两项不等于 0 而等于 1, 权重分别为某个 $P_{\alpha jh}$ 和 1, 这是 PGDD 和 0-1 评分的 GDD 不同之处.

本文仅仅讨论了基于 GRM 的多级评分 GDD, 而基于 GPCM 的多级评分 GDD 的诊断效果如何, 值得研究.另外, 本文只是假设每个属性值都为 1, 即计算期望得分时, 假设掌握项目中一个属性便得到一分的评分方式, 这种属性等权重的假设显然过于简单, 而对于属性权重不等同的情况也有待进一步的研究.

5　参考文献

[1] Tatsuoka K K. Architecture of knowledge structure and cognitive diagnosis: a statistical pattern recognition and classification approach [D]. NJ: Erlbaum, 1995: 327-361.

[2] Leighton J P, Gierl M J, Hunka S M. The attribute hierarchy method for cognitive assessment : a variation on Tatsuoka's rule space app roach [J]. Journal of Educational Measurement, 2004, 41(3): 205-237.

[3] 丁树良, 祝玉芳, 林海菁, 等. Tatsuoka ***Q*** 矩阵理论的修正 [J]. 心理学报, 2009, 41(2): 175-181.

[4] 孙佳楠, 张淑梅, 辛涛, 等. 基于 ***Q***. 矩阵和广义距离的认知诊断方法 [J]. 心理学报, 2011, 43(9): 1095-1102.

[5] Samejima F.Estimation of latent ability using a response pattern of graded scores [EB/OL].[2011-12-16].http: //www. psychometrika.org/journal/online/MN17.pdf.

[6] 杨淑群, 蔡声镇, 丁树良, 等. 求解简化 ***Q*** 矩阵的扩张算法 [J]. 兰州大学学报: 自然科学版, 2008, 44(3): 87-91, 96.

[7] 丁树良, 杨淑群, 汪文义. 可达矩阵在认知诊断测验编制中的重要作用 [J]. 江西师范大学学报: 自然科学版, 2010, 34(5): 490-494.

[8] 丁树良, 汪文义, 杨淑群. 认知诊断测验蓝图的设计 [J]. 心理科学, 2011, 34(2): 258-265.

[9] 祝玉芳, 丁树良. 基于等级反应模型的属性层级方法 [J]. 心理学报, 2009, 41(3): 267- 275.

[10] Ding Shuliang, Luo Fen, Cai Yan, et al.Complement to Tatsuoka's ***Q*** matrix theory [C]. Tokyo : Universal Academy Press INC, 2008: 417-423.

The Generalized Distance Discrimination Based on Graded Response Model

LI Juan, DING Shu-liang*, LUO Fen

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

Abstract: It is motivated by generalized distance discrimination(GDD) for 0-1 scoring proposed by Sun and her colleagues, the GDD for polytomous scoring based on the graded response model(GRM-GDD)is proposed. The results of Monte Carlo simulation study show that the behavior of GRM-GDD is better than that of polytomous attribute hierarchy method based on GRM(GRM-AHM) although the pattern classification ratio decreasing with the slipping of the responses increasing.

Key words: cognitive diagnosis; generalized distance discrimination; graded response model

(责任编辑: 冉小晓)