计算机自适应考试(CAT)系统题库的设计与实现

Design and Realization of the Item Pool System for CAT

刘丽平 (中国科学院研究生院 北京 100049)
(北京市电子信息学校 100011)
王文杰 (中国科学院研究生院 北京 100049)
郭世宁 (中国航空结算中心系统服务部 100028)

摘要:计算机自适应考试(CAT)系统以其试题的难度、数量自适应考生个性化需求而日益得到教育界的重视。本文基于项目反应理论(IRT),设计了一个适应于CAT系统的题库,该题库采用三参数Logistic模型(3PLM)拟合项目特征曲线,运用联合极大似然估计法(JML)对项目参数进行估计,再通过评价函数对题库参数进行控制和优化,使得题库能够更好地适合不同考生的特点。

关键词:计算机自适应考试 项目反应原理 Logistic模型 联合极大似然估计法

1 引言

考试作为衡量个人某种能力和水平的工具,是学习型社会中确保教育质量的重要衡量标准,并深入到社会的各个方面。计算机自适应考试(Computerized Adaptive Testing,简称CAT)随着计算机技术的发展和个性化学习的需求而得到越来越多的关注。CAT的基本思想是:计算机先通过一些试探性试题来初步估计考生的水平,再根据选题算法从题库中选择与考生能力相近的题目继续施测,每施测一题都重新估计考生的能力,并不断重复这一过程。随着考生所做题目的增多,计算机对其能力的估计精度也越来越高,最后其估计值将收敛与一点,该点就是考生能力较精确的估计值,具体施测过程见图1[1]。

这种自适应性的考试方式始终围绕着考生的能力进行,考生所做的试题都是系统根据考生的能力从题库中自动选择的,题目针对性强,更加突出了考生的主体地位和个性化需求,大大提高了考试的效度和信度,并且降低了考试所用的题目数量从而提高了考试的效率。

建立优质的题库是CAT编制中最基础也是工作量最大的工程,它不仅提供考试的题目,而且还提供必要的试题参数,确保自适应优质试卷的生成。CAT系统的题库是建立在项目反应理论(Item Response Theory,简称IRT)基础上的。以IRT为理论基础的CAT在世界各国已经引起了广泛的关注,并逐步在社会各个方面得到推广:在教育界,有美国的研究生入学考试GRE和GMAT;在职业资格认证方面,有全美护士国家委员会资格考试NNCLT;在企业界,有Novell公司的认证考试。在国内,对IRT的研究与应用也逐渐得到了教育部门的普遍关注。

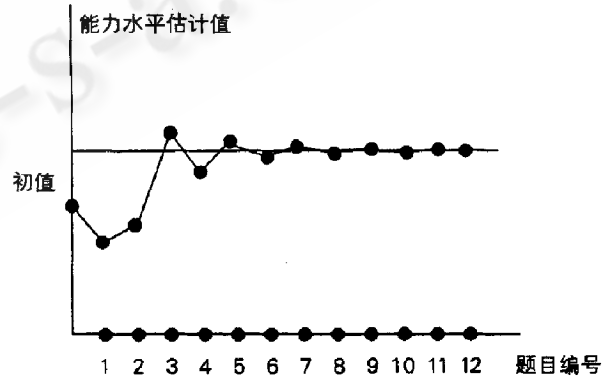

图1 自适应考试施测过程

本文介绍了IRT的基本概念,并着重描述了基于IRT的CAT系统题库设计和实现方案。

2 IRT

IRT 也称项目特征曲线理论或潜在特质理论，它是依据一定的数学模型，用项目特征参数估计潜在特质的一种测量理论。该理论中最重要的两个基本概念是“潜在特质”和“项目特征曲线”。潜在特质是指人具有的相对稳定、支配其对相应的测验做出反应，并使反应表现出一致性的内在特征，一般用表示。决定某一行为的潜在特质往往不是一种，所有潜在特质的集合称为潜在特质空间。大多数考试都是为了考察单一特性而设计的，所以本题库中只考虑单维特质空间的情况。

IRT 研究的主要内容就是被试在测验试题上的反应行为与测验所测的被试潜在特质之间的关系，即项目特征曲线。项目特征曲线是以潜在特质（用表示）为横坐标，以正确反应或肯定反应的概率（用 P( ) 表示）为纵坐标，以此反映项目基本特征的一条曲线见图 2。这条曲线用以下 3 个参数来描述：

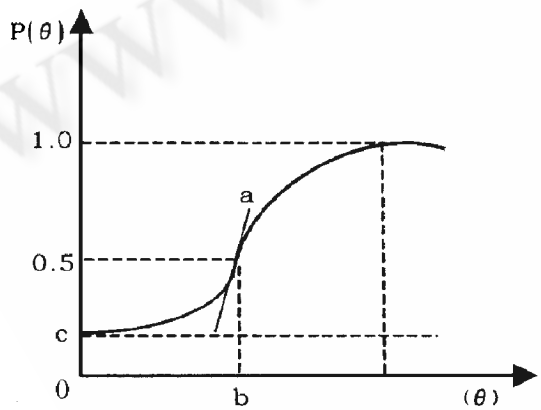

图 2　项目特征曲线

(1) 项目区分度 a。即曲线拐点处的斜率。斜率越大就越陡峭，考生的能力水平 θ 稍有不同，答对题目的概率就有很大变化，即题目的区分能力也就越强。

(2) 项目难度 b。即项目答对概率 P(θ) =0.5 所对应的特质参数 θ 值。

(3) 项目猜测参数 c。即特征曲线的截距，其值越大，越容易猜对本题。

IRT 认为个体的潜在特质与测量该特质的项目反应之间存在一定的函数关系，P(θ) 是随着 θ 的增大而增大，当 θ 大到一定程度以后，P(θ) 就趋向于 1。IRT 就是要研究 P(θ) 与 θ 之间的这种函数关系，并用一定的数学模型来反映这种关系，以此作为系统设计的基础。

3 基于 IRT 的题库设计方案

基于 IRT 的 CAT 题库的建立主要有以下几个部分：(1) 选择模型。首先应选择适合的 IRT 模型，本题库采用的是三参数 logistic 模型 (3PLM)。(2) 题目的开发。开发应按题库命题的规格标准进行，应注重不同知识内容与能力层次、不同难度和不同题型的结合，对开发的试题应组织审查，确保题目的质量。(3) 题目参数的确定。主要是对题目 IRT 各参数值的确定，本题库是经试测后统计分析得到的。(4) 题库的动态维护。包括题目的增加、删除，以及题库参数的动态控制和优化。具体设计方案如下：

3.1 模型选取

项目反应模型是一种数学模型，它是给项目特征曲线配上函数解析式——项目特征函数，用于很好的拟合项目特征曲线。IRT 是依据一定的项目反应模型，用项目特征参数估计潜在特质的一种测量理论。因此，项目反应模型在 IRT 中具有极其重要的意义，是题库建设的首要问题。

目前常用的模型有很多种，如正态卵形模型、Logistic 模型，近年来运用曲线回归的方法又提出了一些新模型如反正切模型、余弦模型等，本题库选用的是单维三参数 Logistic 模型 (3PLM)，其具体表达式为：

$$P(\theta) = c + (1-c)\frac{e^{1.7a(\theta-b)}}{1+e^{1.7a(\theta-b)}} \quad (1)$$

选择该模型的主要依据是：① Logistic 模型在理论和实践上都得到了充分的验证，较为成熟可靠，采用该模型能大大降低系统开发的风险；②该模型计算比较简便；③选择三参数可以提供更多的题目信息，对所测量的特质的估计容易达到较准确的水平。

3.2 参数估计

在题库建设初期试题参数都是未知的，所以需要对其进行估计。本题库中采用联合极大似然估计法[2] (Joint Maximum Likelihood，简称 JML) 对试题参数进行估计。由于本题库采用了单维三参数 Logistic 模型 (3PLM)，所以可设有被试 N 人，试题 M 个，则需要估计试题参数 3M 个 (a，b，c 各 M 个)。假设试题都是 0、1 记分，即答对记 1 分，答错记 0 分，则 N 人回答 M 题的全部结果可用 N 行 M 列的反应矩阵 U 表示：

$$U=(u_{ij})\ N*M=\begin{bmatrix} u_{11} & u_{12} & \cdots u_{1M} \\ u_{21} & u_{22} & \cdots u_{2M} \\ \cdots & \cdots & \cdots \cdots \\ u_{N1} & u_{N2} & \cdots u_{NM} \end{bmatrix} \quad (2)$$

设 P_{ij} 为第 j 个被试答对第 i 题的概率（i =1,2,3,…,M；j =1,2,3,…,N），又令 $Q_{ij}=1-P_{ij}$，即 Q_{ij} 为第 j 个被试答错第 i 题的概率，于是，可以得到第 j 个被试在第 i 题上的分布函数：

$$P_{ij}^{u_{ij}}Q_{ij}^{1-u_{ij}},u_{ij}=0,1 \quad (3)$$

当 $u_{ij}=1$ 时，表示 P_{ij}，当 $u_{ij}=0$ 时，表示 Q_{ij}。

在各个被试作答是相互独立的和同一被试对各题的作答是相互独立的两个假设下，可以得到反应矩阵 $U=(u_{ij})\ N*M$ 的似然函数：

$$L(\theta_1,\cdots,\theta_N;a_1,b_1,c_1,\cdots,a_m,b_m,c_m|u_{11},\cdots,u_{Nm})=\prod_{i=1}^{N}\prod_{j=1}^{m}P_{ij}^{u_{ij}}Q_{ij}^{1-u_{ij}} \quad (4)$$

为了进行参数估计，可先求式(4)的对数似然函数：

$$\ln L=\ln=\prod_{i=1}^{N}\prod_{j=1}^{m}P_{ij}^{u_{ij}}Q_{ij}^{1-u_{ij}}=\sum\sum[U_{ij}\ln P_{ij}+(1-U_{ij}\ln Q_{ij})] \quad (5)$$

将对数似然函数分别对各 M 个 a,b,c 参数求偏导并令其为 0，可得如下方程组：

$$\begin{bmatrix} \sum[(u_{ij}-P_{ij})/P_{ij}\cdot Q_{ij}](P_{ij}/a_i)=0 \\ \sum[(u_{ij}-P_{ij})/P_{ij}\cdot Q_{ij}](P_{ij}/b_i)=0 \\ \sum[(u_{ij}-P_{ij})/P_{ij}\cdot Q_{ij}](P_{ij}/c_i)=0 \end{bmatrix} \quad (6)$$

其中（i =1,2,3,... M；j =1,2,3,…N）。对于非线性方程式(6)，采用牛顿－拉普逊(N－R)迭代法进行计算，则可得到 M 个试题的参数 a,b,c 的值。

3.3 题库建设

(1) 试题开发。试题开发包括制定编题计划和编制试题两个阶段，编题计划要考虑题量、题型、试题分布、目标要求等方面，采用编题三向细目表反应。编制试题采用了组织专家命题、相关资料中选题等方式。

(2) 试测、参数估计。试题编制好后就要对其质量进行审核，质量分析包括定性分析与定量分析两个方面。定性分析包括检查试题是否符合编题计划，测量内容是否有价值，试题表述是否通俗易懂，是否会产生歧义等。定量分析就是根据三参数 Logistic 模型用联合极大似然估计法估计试题的区分度 a、难度 b、猜测参数 c。对于区分度低或猜测参数大的试题应予以淘汰。另外还要检查项目难度的分布是否合理，如不合理要加以调整，以确保在测量各种特质水平的被试时都有足够的试题。

(3) 试题组织方式。本题库将试题正文、评分标准、答案、试题难度、区分度、猜测参数、测量目标层次、背景资料、相关知识点等信息作为题目属性保存下来。题库内部大量试题可以根据内容、认知目标层次、难度、区分度等分区，本题库根据区分度不同分区存储，这样会更利于 CAT 施测时选题的实施。

3.4 题库的动态维护

(1) 题库的扩充。由于 IRT 理论具有参数不变性等优点，所以基于 IRT 的题库的扩充变的更为容易。CAT 需要非常准确的题目参数，而这些参数会随着题库中题目数量的变化而变化，需要做出及时的调整。在一个实际运行的 CAT 系统中，由于某些题目（如区分度高的题目）经常被使用，从控制题目曝光率的角度，需要定期地从题库中删除题目或暂时屏蔽一些题目，从而需要扩充题库。扩充题库并不是简单的增加一些新题目，而是要考虑如何将新的题目与题库中其它题目的参数等值到同一个量表系统上[3]，即找到新旧题目间的线性关系。本题库安排一些连接题目与新题目相混合进行试测，就可以将新题目的参数值与旧题目的参数值统一到一张量表中来。

(2) 题库参数的控制和优化。题库中题目的新增、删除或屏蔽都会影响题库中参数的分布。为了在参数分布不合理时及时调整题库的参数，可设计信息统计模块，用来评价题目参数和知识点参数的分布。当题目参数和知识点参数的分布不合理时给出警告信息。具体评价函数[4]如下：

$$P=\int_{-\infty}^{+\infty}|f(x)-g(x)|dx \quad (7)$$

其中 f(x)是最佳分布密度函数，g(x)是当前实际分布密度函数。根据以下推导：

$$0\leqslant\int_{-\infty}^{+\infty}|f(x)-g(x)|dx\leqslant\int_{-\infty}^{+\infty}(|f(x)+g(x)|)dx=\int_{-\infty}^{+\infty}f(x)dx+\int_{-\infty}^{+\infty}g(x)dx=1+1=2$$

可知：$0\leqslant p\leqslant 2$。当 p 越小分布越接近最佳分布，具体评价指标见表 1。

（下转第 16 页）

(上接第 12 页)

表 1　评价指标表

评价值 p 的范围	[0,0.2]	(0.2,0.6]	(0.6,1.0]	(1.0,1.6]	(1.6,2.0]
结论	很好	合理	可以	需要调整	必须调整

每当 p 值达到需要调整的范围之内时，就给出警告信息，此时可通过向题库中新增或屏蔽题目的方式调整，达到调整题目参数和知识点分布的目的。

4　结束语

将 IRT 应用于题库的建设，关键在于项目反应模型的选取，基于模型参数估计方法的确定，并最终给出题库建设方案。在模型的选择上，本题库采用了三参 Logistic 模型。虽然目前也有人提出了其它一些数学模型，但有些已过时，有些还处于试验阶段。而 Logistic 模型在理论和实践上都得到了较为充分的验证，采用它能大大降低题库开发的风险，也有利于与其它系统的集成，发挥资源整合的优势。参数估计上，本题库应用了联合极大似然估计法。本题库建设方案主要应用于客观题方面，并在美国英语语言学院电子信息教学部考试系统建设中正在被使用，如何做好主观性试题参数的估计问题，还有待于进一步研究。

参考文献

1　陈乙雄、符云清、祝伟华，项目反应原理在远程教育中的应用研究[J]，计算机科学，2005，32(4)：228－230。

2　顾海根，人员测评[M]，合肥 中国科学技术大学出版社，2005，236－240。

3　邵晨辉，基于 WEB 的自适应汉语测试模型的研究[D]，上海交通大学，2001。

4　田怀凤、袁琰、王立 等，机助自适应考试(CAT)系统题库的仿真研究[J]，计算机仿真，2005，22(7)：246－248。