心理学探新 2021,Vol.41,No.5,458-465
PSYCHOLOGICAL EXPLORATION

针对微测评理念的自适应测验技术
——基于 CO-MIRT 诊断模型

孙雄飞[1]*,王永明[2],贾小君[1]
(1.北京学格科技有限公司测评研发部,北京 100031;2.英国伦敦国王学院,精神病学、心理学和神经科学研究所,伦敦)

摘 要:前人研究业已表明 MIRT 模型在自适应测验等诸多领域的测量优势,但面对当前国内教育行业在实践过程中的现状,仍无法有效地解决待测知识点数量、试题数量和测量精度之间的矛盾。对此,本次研究设计了 CO-MIRT 模型,经由前馈层、全连接层的操作以共享试题之间所传递的信息,以及通过控制层、L2 正则化等操作来限制小样本测验下的过拟合,来达到降低估计误差的目的。本次研究采用蒙特卡洛模拟的方式验证了模型效果,并使用数学推演的方式给予理论上的证明。

关键词:微测评;自适应测验;多维能力参数;CO-MIRT 模型
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003-5184(2021)05-0458-08

1 引言

项目反应理论(Item Response Theory,简称为 IRT)是20 世纪 50 年代以后发展起来的第三代教育测量理论,在项目开发、试卷编制、测验等值、题库建设和计算机化自适应测验等诸多方面都发挥了极大的作用。然而,正如 Mislevy 和 Wingersky(1993)所指出的,项目反应理论,作为一种标准测验理论,主要是在宏观层次上来对被试能力进行量化和评价,而在微观层面上却有所忽略。因此,传统的、宏观取向的模型所计算出来的结果,既难以在理论上得到有效的验证,也无法在实践过程中指导工作人员对此给予干预和"补救",存在着较大的局限性。

认知诊断理论的兴起填补了传统的测验理论所造成的空缺。与经典测验理论、概化理论和项目反应理论有所不同,认知诊断理论并非某一个具体的模型,而是包含了若干个目标一致的模型。这类模型总体上而言可以划分为离散型和连续型两个大类,而多维项目反应理论(Multidimensional Item Response Theory,简称为 MIRT 模型)则是连续型认知诊断模型的代表。多维项目反应理论结合了项目反应理论和因素分析的优势,已愈发受到学界和业内的重视,在诸多领域里发挥了不可替代的作用。

随着计算机科学技术的不断革新和国内教育市场的蓬勃发展,"精准教学"的理念也日趋成为业界的共识。要实现对学生的精准教学,关键的一步是要通过细致而准确的测评手段来评估学生在各项知识点、能力等维度上的表现,这也是"纳米级知识点"的理念在教育测评行业逐渐兴起和不断发展的根本原因。

细致的测评意味着要在更微观的粒度上采集学生作答的数据,然而,随着对被试能力测量的不断颗粒化,待测知识点数量也会呈现出倍数增长的态势。在传统的测量模型下,测量一项知识点通常需要 5 到 10 项题目,而如果将 1 项知识点扩展成 5 项知识点,测试题目就会上升至 25 至 50 道题目,否则会造成极大的误差而无法达到精准的目的。从另一方面讲,如果试题数量过高,同样会对测量结果产生负面影响。除了学校组织的统一考试以外,在单次测验中,被试连续回答数十道题目时会承载极大的心理压力,在无监督的情况下甚至会产生拒斥答题的现象,因而同样满足不了精准的要求。在此情况下,如何兼顾测量误差、测量题量和待测知识点数量之间的平衡,是本次研究所需要解决的基本问题。

2 相关研究

2.1 文献综述

除了 EM(Expectation Maximum,期望最大化算法)、MCMC(Markov Chain Monte Carlo Simulation,马尔可夫蒙特卡洛算法)等算法以外,神经网络也是项目反应理论中常见的参数估计方法。台湾地区的学者(郑海东,1999;蔡至煌,2000;黄坤泉,2000)

* 通讯作者:孙雄飞,E-mail:442337729@qq.com。

最早采用神经网络方法对 IRT 试题参数进行了实验。后来,大陆地区的学者(余嘉元,2002;谭云兰,丁树良,辛锐铭,2004)分别提出了使用联结主义模型和 BP 神经网络来对 IRT 项目参数和能力参数进行估计。随后,Bagging 技术也被引入优化广义回归神经网络(generalized regression neutal network, GRNN)的估计过程(余嘉元,2009)。

然而,并非所有的心理与教育测验都是扁平的,相反是隐含着结构性质的关系的(Golay & Lecerf, 2011)。对此,De 和 Song(2009)最早依据 3PL 模型提出了高阶项目反应理论,并在小样本条件下验证了 HO - IRT 较之传统的 IRT 模型的参数估计的优势(Dela Torre & Hong,2010)。目前,HO - IRT 模型被广泛地应用于教育测验中的多个领域,包括计算机自适应测验(Huang, Chen, & Wang, 2012)、题组测验(Huang & Wang,2013),结合混合 IRT 模型来对总体 - 子组测验进行分析(Wang, Kohli, & Henn, 2015)。近年来,一些研究人员尝试将深度学习技术与传统的心理测量模型进行融合,例如:利用 CNN 技术和神经网络诊断模型(Neural Cognitive Diagnosis Model, NCDM)对试题文本信息和能力参数进行抽取、拟合,并以此构建了对试题标签和难度的预测(Wang et al. ,2020);以及使用试题文本信息,并结合 DKT 和 IRT 模型提出了 Deep - IRT 模型(Yeung,2019)。

前人关于神经网络的应用,对本文将深度学习模型与心理、教育测量的结合提供了启示。HO - IRT 模型中对知识点的树状结构的表述和模型,既符合实践工作者对知识图谱的修订规则,也为本文在进一步利用知识点间的先验信息,提供了充足的理论依据。然而在过去研究中,尽管在许多实验条件下证实了神经网络理论对 IRT 参数估计的有效性,但这些实验基本上是围绕着项目参数估计、大样本*条件而设计的,而没有针对被试能力的、小样本**的实验结果。此外,在利用知识点间的结构性信息方面,HO - IRT 模型更多考虑的是父节点与子节点之间的关系,但却假设兄弟节点之间的关系是相互独立的,这一点与许多经验证据是有矛盾的。而如何对传统 MIRT 模型进行优化,使之良好地完成小样本条件下的测样任务,是本次研究的主要目标。

2.2　MIRT

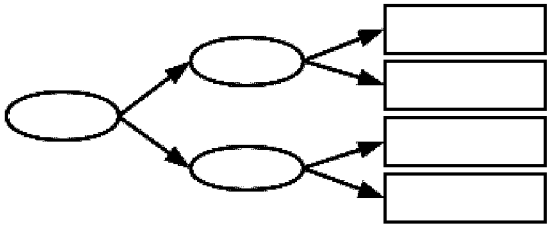
项目反应理论在总体上可划分为三个类别,分别是正态肩型模型、Rasch 模型和 Logistic 模型,其中应用最广泛的是 Logistic 模型。在 Logistic 模型的结构下又可以划分为双参数模型和三参数模型,本次研究主要采用的是双参数模型。

$$p_j(\theta_i) = \frac{1}{1 + \exp^{-\sum_k Da_{jk}(\theta_{jk} - b_{jk})}}$$

其中,D 参数取值为 1.7,a_{jk} 表示第 j 道试题的第 k 个知识点的区分度,b_{jk} 表示第 j 道试题的第 k 个知识点的难度,θ_{jk} 表示第 i 个被试在第 k 个知识点上的能力水平。

2.3　HO - IRT

在心理与教育测量中会经常遇到这样一种情况:待测知识点不是单独而是多维的,而多维知识点之间的关系也不是松散的,而是存在某种结构化的关系。也就是说,多维知识点既有可能是某一个高阶知识点的子节点,也很有可能是某些知识点的兄弟节点,彼此之间存在着潜藏的数量关系。如下图所示,圆圈代表一个知识点,矩形代表一道试题。



HO - IRT 模型是一种为了测量阶层化指标而专门设计的模型,它良好地结合了传统的 IRT 模型对宏观层面上的总体性指标的估算能力,以及 MIRT 模型在微观层面上的计算效率和准确性。HO - IRT 模型总体上而言继承了 MIRT 模型的基本结构,但它将被试能力 θ_{ik} 改造为 $\theta_{ik} = \lambda_k\xi_i + \varepsilon_k$,其中 λ_k 表示 θ_{ik} 的潜在回归系数,ξ_i 表示根节点的 theta 值,ε_k 表示在已知 λ_k 的前提下关于 θ_{ik} 的残差项。

然而,HO - IRT 模型在实践过程中也存在一定的问题。与心理测量相比,针对知识点的测验并不仅仅存在唯一的量表或试卷,而是存在多样性的,这就存在一种潜在的可能,也即在不同条件下,被试总体能力 ξ_i 与单项能力 θ_{ik} 之间的关系,即 λ_k 和 ε_k 两项参数会存在差异,不能采用统一的线性关系予以

*　此处所指的大样本是指在被试能力的参数估计中,通常试题数量会多达几十道。

**　小样本指的是仅仅使用几道题目来对一项知识点进行测量。

表示。如何对 λ_k 和 ε_k 进行处理，尚无相关研究成果。而且，如何在小样本的条件下计算其关系，同样缺乏相关的文献说明。

3 模型构建

基于前人研究成果，本文提出了一种新的项目反应理论模型，即协作式的多维项目反应理论（Cooperative Multiple Item Response Theory，以下简称 CO－MIRT）。简单地说，CO－MIRT 模型一方面继承了 HO－IRT 模型的阶层结构，并在此基础上摒弃了知识点相互独立的假设，补充了对兄弟节点的函数关系的考虑，使之能够协作分享彼此的信息，从而提高估计精度；另一方面，CO－MIRT 模型借鉴了深度学习模型的设计，在前馈过程中使用 tanh 函数对每一层的输出结果进行激活，使其得以处理非线性的节点关系。

3.1 前馈层

比起传统的 MIRT 模型，HO－IRT 模型提出了代表被试整体能力的 ξ_i（见上文）的概念。而本文则在此基础上对它进行了进一步的扩展，提出了具有层级关系的、锥式的能力体系（以下简称层级属性关系）。该体系从纵向、横向两个维度上描述了各项能力维度之间的关系，从纵向来说，该体系描述了子级与父级知识点之间的关系，而横向上则描述了兄弟知识点之间的关系。

本文对于层级属性关系中两类知识点关系的数学关系表示如下，以 children 表示子级知识点的水平，parents 表示父级知识点的水平，w_{p2c} 表示父级与子级知识点之间的权重，b_{p2c} 表示父级与子级知识点之间的偏置项。brothers 表示兄弟知识点的水平，w_{b2b} 表示兄弟知识点之间的权重，b_{b2b} 表示兄弟知识点之间的偏置项。

$$children = f(parents, w_{p2c}, b_{p2c})$$

$$brother = g(brother, w_{b2b}, b_{b2b})$$

基于上述理论假设，本文借鉴并拓展了认知诊断理论中常用的邻接矩阵（又称 A 矩阵）概念，使之适用于层级属性关系中的知识点关系。首先，应邀请学科专家绘制本学科的知识图谱，该图谱绘制的重点集中在父级知识点与子级知识点之间的关系，暂不涉及到兄弟知识点之间的关系。其次，根据知识图谱中的父子关系而绘制了邻接矩阵的图式，由此来控制前馈层中的信息传递。

在绘制父子关系的邻接矩阵的过程中，假设前馈层总共存在 L 层，$L = \{l_1, l_2, \cdots, l_s\}$，$s = 1, 2 \cdots l$，ls 表示 L 层中的第 s 层。其中，第 l_{s-1} 层是第 l_s 层的父节点，第 l_s 层是第 l_{s-1} 层的子节点。第 l_s 层的矩阵尺寸为 m * n，m 是父节点的知识点数量，n 是子节点的知识点数量，若第 m 个父节点与第 n 个子节点之间存在关系，则标记为 1，否则为 0，记为 A_s。w_s 是随机初始化生成的权重矩阵（m * n），用来计算父节点和子节点之间的关系，同时对它进行 softmax 处理*，β_s 是随机初始化生成的偏置项。θ 表示学生的能力值，θ_s 表示当前学生的第 s 层能力向量。

最后，模型对本层的输出结果 θ_s 进行 tanh 处理，并对此后接一个全连接层。关于全连接层的部分会在下一节中进行描述。

具体数学关系如下：

权重：

$$weight_s = \mathrm{softmax}(weight_s)$$

$$weight_s = A_s \bullet weight_s$$

输出：

$$theta_s = theta_{s-1} \circ weight_s + beta_s$$

激活：

$$theta_s = \tanh(theta_s)$$

3.2 全连接层

全连接层是神经网络算法中的常用手段。CO－MIRT 模型假设，属于同一个父节点之下的任意知识点，均与其兄弟节点之间存在相关关系。基于这一假设，模型使用了全连接层对层级属性关系中的每一层知识点，也即兄弟知识点关系，进行一次全连接计算。其中，w_s 表示兄弟知识点之间的关系矩阵，并对其进行 softmax 处理，β_s 表示偏置项，θ_s 表示当前学生的第 ls 层的的能力向量。w_s 和 β_s 的初始化是随机给定的。

具体数学关系如下：

权重计算：

$$weight_s = \mathrm{softmax}(weight_s)$$

能力计算：

$$theta_s = theta_s \circ weight_s + beta_s$$

能力激活（注：k 表示底层知识点，不参与激活）：

$$theta_s = \tanh(theta_s)$$

$$theta_k = theta_k$$

---

* 如果前馈层表示的是根节点与其子节点之间的关系，则不进行 softmax 运算。

除了对前馈层中的每层知识点进行全连接计算以外，模型还针对底层知识点（即本次测评事件中的目标知识点进行一次全连接层的计算，而其输出 θ_k 即是本次测评事件中，所期望获得的被试能力估值。

3.3 控制层

在自适应测验和微测验的双重条件下，对被试能力的估算是一个明显的难题。如果使用 MCMC 算法来进行参数估计，则会由于迭代次数过多而无法满足快速反应的需求（自适应测验通常是根据学生上一道题目的回答结果，来推送下一道题目）。而使用梯度下降法来进行参数估计，又会因为试题数量过少而难以避免极端值的出现。

面对梯度下降法所造成的弊端，相关研究表明 tanh 函数对 θ_s 的估值具有明显的抑制效应。然而，tanh 函数的值域为（-1，1），这实际上会对 theta 的估计值产生人为的限制，比如，若 theta 服从于标准正态分布时，tanh 函数会主动摒弃左右两侧的值域，累计约 31.7%。也就是说，tanh 函数会对成绩较优或较差的学生产生“抛弃”。

为了解决这一难题，CO - MIRT 算法专门设计了 control 层。假设知识图谱中的每一层知识点共有 k_s 个，随机生成一个尺寸为 $1 * k_s$ 的矩阵 $w_{control}$，并对其进行 sigmoid 处理，将其取值映射到（0，1）。此外，令 θ_s 为未经 tanh 处理的估算值，θ'_s 是对 θ_s 进行 tanh 处理的结果。则最终的输出结果如下：

具体数学关系如下：

控制向量，其中 k 表示底层知识点的数量：

$$control = [w_1, w_2, \cdots, w_k]$$

$$control = sigmoid(control)$$

控制计算，其中 k 表示底层知识点的数量：

$$theta_s' = \tanh(theta_s)$$

$$theta_s = theta_s \bullet control + theta_s' \bullet (1 - control)$$

3.4 输出层

输出层是为了将前馈层、全连接层的最后输出，即学生能力的估值 θ_i，与试题参数（包括 Q 矩阵、区分度 a、难度 b，常数 D）相结合而构成 logit，从而构建并完成损失函数的计算。其中，l 表示知识图谱中的底层知识点，j 表示第 j 道题目，q_j 表示第 j 道题目的 Q 矩阵，a_j 表示第 j 道题目的区分度，b_j 表示第 j 道题目的难度，常数 D 取值为 1.7。

具体数学关系如下：

$$\text{logit}_j = q_j^{\circ}[a_j \bullet (theta_i - b_j)]$$

3.5 损失函数和惩罚项

通过上述的神经网络结构，可获得被试在每一道试题中的 logit_j，代入 sigmoid 函数后可计算出被试在该题上的答对概率，并结合被试回答的正误结果，可构建出交叉熵损失函数。

$$Loss = \sum_j y_j \log p_j + (1 - y_j)\log(1 - p_j)$$

同时，为了避免在试题数量过少的条件下，梯度下降法所造成的极端值现象，算法对被试能力向量 θ_i 的结果进行了 L2 正则化处理。

4 实验设计

4.1 实验环境

为了验证 CO - MIRT 模型的有效性，本次实验围绕着测量模型、知识点数量和试题数量共三个维度，结合 Monte Carlo 模拟方法和自适应测验方法进行比较式的研究。

测评模型：参与比较的模型包括传统的双参数 MIRT 模型、HO - IRT 模型、CO - MIRT 模型，总计三个模型；

测评形式：为了验证算法模型最优效果，实验采取自适应测验的方法，也即根据被试在上一道题目中的回答情况来推送下一道测试题目，试题推荐算法采用的是 CAT 中常用的 Fisher 信息量方法；

知识点数量：本次实验将知识点数量划分为 5、10、20、30 道题目共 4 个层次。其中，5、10 项知识点更倾向于国内教育行业，尤其是英语教育行业，每次测验的平均测量知识点数量，例如，英语阅读理解通常会分为细节理解、主旨大意、词义猜测、逻辑推理、七选五，语法填空则包括词性转换、非谓语动词、宾语从句、状语从句、介词等近十项考察点；而 20、30 项知识点则是针对近年来业内提出的纳米级知识点，或是称为细分级知识点的概念。

试题数量：基于微测评的理念，实验对试题数量进行了严格的控制。针对 5 项和 10 项知识点，由于知识点数量相对较少，因此在此基础上乘以 2，分别设置为 10 道和 20 道题目。而对 20 项和 30 项知识点，则统一设置为 30 道题目，以防止试题数量过多。因此，在本次实验中一共存在四个层次，分别是 5 - 10、10 - 20、20 - 30、30 - 30 共四个层次，其中前一个数字表示知识点数量，后一个数字表示试题数量。

4.2 参数设置

被试能力向量：被试能力 $\theta_i = [\theta_{i1}, \theta_{i2}, \cdots, \theta_{ik}]$ 是基于多元正态分布 mvnrnd（μ，$\sum$）而随机生成

的，θ_i 表示第 i 个学生的能力向量，θ_{ik} 表示第 i 个学生的第 k 项能力。μ 是能力均值向量，在本次实验中全部设置为 0。$\sum$ 是能力向量的方差 - 协方差矩阵，在本次实验中，假设 K 项能力之间是相互独立的，因此将其设置为单位矩阵 I。针对不同的知识点，本次实验各生成 1000 个被试能力向量。

属性层级关系：属性层级关系即用于前馈层的、对知识图谱中父子节点关系的描述矩阵。考虑到 HO - IRT 模型只存在两层结构关系，即根结点与其他知识点之间的关系，因此本次实验也只设计了相同的结构关系。例如，实验测量的是五项知识点，则关系矩阵为[1,1,1,1,1]。

Q 矩阵：Q 矩阵指的是待测知识点与测验题目之间的关系，如果第 j 项题目考察第 k 个知识点，则标记 q_{jk} 为 1，否则为 0。在测验类型上，本次实验采取的是项目间 - 多维能力测验的方式，也即每道试题只考查 K 项知识点中的一项。

项目参数：本次实验主要采取的是项目反应理论中的双参数模型，也即一项试题包括了区分度和难度两项参数。其中，难度是从标准正态分布 normrnd (0,1) 中生成的，区分度是从对数正态分布中 lognrnd (0,1) 生成的。此外，常数 D 取 1.7。

被试得分矩阵：在本次实验中，利用上文所述双参数模型计算出学生对该题的答对概率 p，同时从 (0,1) 中随机生成一个数字 r，若 r 小于 p 则标记被试在该题上的回答结果为 1，否则为 0。

4.3 效果评价

尽管从表面上看，MIRT 算法是对 Logistics 分类模型的一种拓展，但对被试能力的参数估计实际上是一个回归问题。因此，采用离均差平方根（Root Mean Square Error，RMSE）作为参数返真性的评价，并采用多个 RMSE 的标准差作为算法稳健性的评价。

参数返真性：

$$RMSE = \sqrt{\frac{\sum_n (\theta' - \theta)^2}{n}}$$

算法稳健性：

$$S = \sqrt{\frac{\sum_n (RMSE - \overline{RMSE})^2}{n-1}}$$

其中，θ 和 θ' 是被试能力的真实值和估计值，n 表示样本容量。

5 研究成果

5.1 结果分析

表 1 展示了三种模型在四个层次的知识点数量和试题数量条件下的参数返真性与算法稳健性的结果。

表 1 三种 MIRT 模型对被试能力估计返真性及稳健性情况

	5 - 10	10 - 20	20 - 30	30 - 30
MIRT	0.587(0.023)	0.702(0.036)	0.784(0.028)	0.868(0.029)
HO - IRT	0.554(0.020)	0.641(0.025)	0.742(0.021)	0.822(0.026)
CO - MIRT	0.507(0.012)	0.556(0.018)	0.643(0.019)	0.724(0.027)

注：括号外的数字是参数返真性的结果，括号内的数字是算法稳健性的结果。

从蒙特卡洛模拟的实验数据中，可以总结如下：

无论是在哪一种知识点数量的条件下，CO - MIRT 模型的参数估计精度均取得了最优效果，HO - IRT 模型次之，而传统的 MIRT 模型则处于末尾；

随着知识点数量的递增，三种模型的估计精度均呈现出下降趋势，且精度偏差* 呈现出上升趋势。按照实验顺序来排序，MIRT 与 CO - MIRT 的精度偏差依次是 0.08、0.146、0.141 和 0.144，HO - IRT 和 CO - MIRT 的精度偏差依次是 0.047、0.085、0.099、0.103；

如果设置测量精度的接受阈值为 0.6，则 CO - MIRT 模型可胜任 10 项知识点以下的测评任务，而 MIRT 和 HO - IRT 模型则只能胜任 5 项知识点以下的测评任务。以国内 K12 英语学科测验为例，这意味着 MIRT 和 HO - IRT 模型将无法胜任完形填空、语法填空等相关测验工作。如果将接受阈值提升至 0.7，则 CO - MIRT 模型可胜任 20 项知识点以下的测评任务，而 MIRT 和 HO - IRT 模型则只能胜任 10 项知识点以下的测评任务，这意味着 MIRT 和 HO - IRT 模型将无法胜任细分知识点（如语法点）的相关

* 精度偏差指的是在同一实验数据的条件下，两种模型的估计精度 rmse 的差值。例如在实验数据 data 的条件下，HO - IRT 模型估计的 rmse 为 0.742，CO - MIRT 模型估计的 rmse 为 0.643，则精度偏差为 0.099。

测验工作；

稳健性指标均处于 0.03 以下，表明三个模型对被试能力的参数估计是比较稳定的。

5.2　数学证明

在项目反应理论的参数估计的问题中，Fisher 信息量是用以估算 θ' 与 θ 之间的误差的重要函数，二者之间存在着如下的数学关系。

$$SE(\theta) = \frac{1}{\sqrt{\sum_j I_j(\theta)}}$$

从上述公式可知，若要减少被试能力 θ 的标准误，一方面可以通过增加试题数量来实现，另一方面则应该提高每一道试题所提供的信息量。而后者除了可以在自适应测验的选题过程中选择更具价值的试题，也可以通过对算法结构的优化来完成，而这正是本次研究所完成的事情。

在 MIRT 的模型中，关于被试能力 θ 的信息函数公式如下(Reckase，2009)：

$$I_j(\theta) = P_j(\theta)Q_j(\theta)(\sum_k Da_{jk}\cos\alpha_{jk})^2$$

其中，$P_j(\theta)$ 是被试在第 j 道题上的答对概率，$Q_j(\theta) = 1 - P_j(\theta)$ 是被试在该题上的答错概率，D 为常数 1.7，a_{jk} 为第 j 道试题第 k 个知识点的区分度，$\alpha_j = [\alpha_{j1},\alpha_{j2},\cdots,\alpha_{jk}]^*$ 是第 j 道试题的方向向量，且 $\sum_k \alpha_{jk} = 1, \alpha_{jk} > 0$，$\cos\alpha_{jk} = \frac{\alpha_{jk}}{\sqrt{\sum_k \alpha_{jk}}}$。

在 CO－MIRT 算法中，由于完成了多次前馈层和前连接层的的计算，因此可将输出层的函数构建如下：

$$p_j(\theta) = \frac{1}{1 + e^{-\sum_k \sum_m Da_{jk}w_{mk}\theta_{mk} - d}}$$

其中，$w_j = [w_{j1},w_{j2},\cdots,w_{jk}]$ 是在神经网络处理之后的权重向量，w_{jk} 是第 k 项知识点上的权重。m 和 k 是对知识点的表示。$d = \sum_k a_{jk}b_{jk}$ 表示当前试题的总体难度。由上述公式可推导出，在前馈层、全连接层的处理之下，测量模型将得到一个全新的方向向量 $\alpha'_j = [\alpha'_{j1},\alpha'_{j2},\cdots,\alpha'_{jk}], \alpha'_{jk} > 0$，而且与传统模型相比是更加稠密的，因为 MIRT 模型方向向量中的权重并非全部大于零，也就是说，并不是所有的知识点都能分享到当前试题所传递出来的信息量，而 ∂_j 则能实现这一效果。因此，可将 CO－MIRT 模型的信息函数定义如下：

$$I'_j(\theta) = P_j(\theta)Q_j(\theta)(\sum_k Da_{jk}\cos\alpha'_{jk})^2$$

经由数学推导(详见附件)可证明 $I'(\theta)$ 大于 $I(\theta)$，说明在不增加测试题量的前提下，CO－MIRT 算法有效地增加了试题所传递出来的信息量，从而减少被试能力 θ 的测量误差，优于传统的 MIRT 模型。

6　总结与展望

结合项目反应理论的研究趋势与国内 K12 教育行业的发展方向，本项研究在汲取和优化多维项目反应理论和高阶项目反应理论的前提下，利用神经网络技术对此完成了进一步的迭代升级。同时，依据 3 项理论模型 *4 种知识点数量进行了蒙特卡洛模拟实验，并通过数学证明的方式证实了 CO－MIRT 在参数估计精度上，较之 MIRT 模型和 HO－IRT 模型的优势。

然而，本次研究仍然存在一些有待改进之处。

首先，尽管实验证明 CO－MIRT 模型在测量精度上的优越性，但是当待测知识点的数量上升到一定程度之后，也难以保证将其控制在可接受范围内。因此，在此基础上，如何进一步地优化模型结构以降低参数估计的误差，是下一阶段的研究重点。

其次，在当前 CO－MIRT 模型结构中也有所不足。第一个不足之处是当待测知识点数量增加到一定程度之后，全连接层的权重将会过分庞大，反而有可能会降低参数估计的精度。第二个不足之处是项目组尚未能在理论上证明 Control 层的优势。而与此相关的工作也是未来研究的一个组成部分。

最后，在本次模拟实验中，研究假定父节点与子节点之间的数学关系是线性的，但有相关研究表明这一假设并非完全成立的。事实上，在研究之初，项目组原本计划将非线性的条件纳入到实验过程中，但最终未能实现。一方面是因为在多项式扩展的情况下，很难将被试能力 θ，$\theta \sim$ normrnd(0,1) 的数量范围控制在[－3,3]之间。另一方面是项目组可以尝试使用 tanh 函数激活来规避这一问题，但担心会有过分贴合模型设计的嫌疑。目前只能依据常理推断，在非线性关系上，CO－MIRT 模型相比 MIRT 和

---

* 因为模型使用的是梯度下降法，也即方向导数与偏导数向量的方向一致。同时，答对概率 P_j 在 θ_k 上的偏导数为 $\frac{\partial P_j\theta}{\partial\theta_k} = P_j\theta Q_j(\theta)$，是恒大于零的。因此，可知方向导数中的元素 α_{jk} 恒大于零。

HO－IRT 模型的优势会更加明显。而这一推断未能在本文中予以展示。如何解决非线性数据生成的这一难题，也是未来工作中的一个重要的方向。

附件：

设 $w = \cos\partial, Info' = P_j(\theta)Q_j(\theta)$

$J(\theta) = I'(\theta) - I(\theta) = Info'(\sum_k Da'_k w'_k)2 - Info'(\sum_k Da_k w_k)^2$

$= D^2 Info'[(\sum_k a'_k w'_k)2 - (\sum_k a_k w_k)2]$

$= D^2 Info'[(\sum_k a'_k w'_k + 2\sum_k a'_k w'_k) - (\sum_k a_k w_k + 2\sum_k a_k w_k)]$

$E[I'(\theta) - I(\theta)] = D^2 Info' E[(\sum_k a'_k w'_k + 2\sum_k a'_k w'_k) - (\sum_k a_k w_k + 2\sum_k a_k w_k)]$

$= D^2 Info'[\sum_k E(a'_k w'_k) + 2\sum_k E(a'_k w'_k) - \sum_k E(a_k w_k) - 2\sum_k E(a_k w_k)]$

$= D^2 Info'[\sum_k E(a'_k)E(w'_k) + 2\sum_k E(a'_k)E(w'_k) - \sum_k E(a_k)E(w_k) - 2\sum_k E(a_k)(w_k)]$

$\because a_k$ 服从于同一分布

$\therefore J(\theta) = D^2 Info'[E(a^2)\sum_k E(w'_k) + 2E(a^2)\sum_k E(w'_k) - E(a^2)\sum_k E(w_k) - 2E(a^2)\sum_k (w_k)]$

$= D^2 Info'[E(a^2)E(\sum_k w'_k) + 2E(a^2)E(\sum_k w'_k) - E(a^2)E(\sum_k w_k) - 2E(a^2)E(\sum_k w_k)]$

$\because \sum_k w'_k = 1, \sum_k w_k = 1$

$\therefore J(\theta) = D^2 Info'[E(a^2) + 2E(a^2)E(\sum_k w'_k) - E(a^2) - 2E(a^2)E(\sum_k w_k)]$

$= D^2 Info'[2E(a^2)E(\sum_k w'_k) - 2E(a^2)E(\sum_k w_k)]$

$= 2D^2 E(a^2) Info' E(\sum_k w'_k - \sum_k w_k)$

令 $\eta = 2D^2 E(a^2) Info'$，则 $J(\theta) = \eta E(\sum_k w'_k - \sum_k w_k)$

因此，欲证 $E[I'(\theta)] > E[I(\theta)]$，则证 $\sum_k w'_k > \sum_k w_k$

$\because \sum_k w'_k = \frac{\sum_k \alpha'_{jk}}{\sqrt{\sum_k \alpha'_{jk}}}, \alpha'_{jk} \geqslant 0$

$\sum_k w'^2_k = \frac{\sum_k \alpha'_{jk} + \sum_k \sum_k \alpha'_{jm}\alpha'_{jn}}{\sum_k \alpha'_{jk}} = 1 + \frac{\sum_k \sum_k \alpha'_{jm}\alpha'_{jn}}{\sum_k \alpha'_{jk}}$

又 $\because$ 当每道题目仅考察一项知识点时，$\sum_k w_k^2 = \frac{\alpha_{j1}}{\alpha_{j1}} = 1$，

$\therefore \sum_k w'_k > \sum_k w_k \Rightarrow E[I'(\theta)] > E[I(\theta)] \Rightarrow I'(\theta)$ 比 $I(\theta)$ 提供更多信息量

参考文献

蔡志煌.(2000).*利用神经网络于题目反应理论参数估计之研究*(硕士学位论文).台南师范学院,台湾.

黄坤泉.(2000).*题目反应理论参数自动化估计之研究*(硕士学位论文).台南师范学院,台湾.

余嘉元.(2002).基于联结主义的连续记分 irt 模型的项目参数和被试能力估计.*心理学报*,(5),80－86.

谭云兰,丁树良,辛锐铭.(2004).基于 irt 模型的 bp 神经网络降维法参数估计及其应用.*江西师范大学学报(自然科学版)*,*28*(6),4.

余嘉元.(2009).基于神经网络集成的 irt 参数估计.*江南大学学报(自然科学版)*,*8*(5),4.

郑海东.(1999).*以类神经网络进行适性测验题目参数估计之研究*(硕士学位论文).台南师范学院,台湾.

De,L.,& Song,H.(2009).Simultaneous estimation of overall and domain abilities:A higher－order irt model approach.*Applied Psychological Measurement*,*33*(8),620－639.

Golay,P.,& Lecerf,T.(2011).Orthogonal higher order structure and confirmatory factor analysis of the french wechsler adult intelligence scale(wais－iii).*Psychological Assessment*,*23*(1),143－152.

Huang,H. Y.,Chen,P. H.,& Wang,W. C.(2012).Computerized adaptive testing using a class of high－order item response theory models.*Applied Psychological Measurement*,*36*(8),689－706.

Huang,H. Y.,& Wang,W. C.(2013).Higher order testlet response models for hierarchical latent traits and testlet－based items.*Educational & Psychological Measurement*,*73*(3),491－511.

Jimmy,D.,& Hong,Y.(2010).Parameter estimation with small

sample size a higher – order irt model approach. *Applied Psychological Measurement*,*34*(4),267 – 285.

Mislevy,R. J. ,& Wingersky,S. M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*,*30*(1),55 – 78.

Reckase,M,D. (2009). *Multidimensional item response theory*. Springer New York.

Wang,F. ,Liu,Q. ,Chen,E. ,Huang,Z. ,& Wang,S. (2020). Neural cognitive diagnosis for intelligent education systems. *Proceedings of the AAAT Conference on Artificial Intelligence*,*34*(4),6153 – 6161.

Wang,C. ,Kohli,N. ,& Henn,L. (2015). A second – order longitudinal model for binary outcomes:Item response theory versus structural equation modeling. *Structural Equation Modeling:A Multidisciplinary Journal*,*23*(3).

Yeung,C. K. (2019). *Deep – irt:Make deep learning based knowledge tracing explainable using item response theory*. arXiv e – prints.

Adaptive Testing Technology for Micro – testing Concept
——Based on CO – MIRT Diagnosis Model

Sun Xiongfei[1],Wang Yongming[2],Jia Xiaojun[1]

(1. Evaluation R&D Department,Beijing Xuege Technology Co. ,Ltd,Beijing 100031;
2. Institute of Psychiatry,Psychology & Neuroscience,King's College London,London)

Abstract:Previous studies have shown the measurement advantages of the Multiple Item Response Theory(MIRT)model in many fields such as adaptive testing. However,in the face of the current status of the domestic education industry in the process of practice,it is still unable to effectively solve the problem contradiction. In this regard,this study designed the Cooperative Multiple Item Response Theory (CO – MIRT) model,through the feedforward layer and the fully connected layer to share the information passed between the test questions,and through the control layer,L2 regularization and other operations to limit the overfitting under the small sample test,and to reduce the estimation error. The Monte Carlo simulation was used to verify the model effect,and the mathematical deduction was used to give theoretical proof in our study.

Key words:micro – evaluation;adaptive testing;multi – dimensional ability parameters;CO – MIRT model