

贝叶斯网在小样本认知诊断中的应用*

汪玲玲

(沈阳师范大学教育科学学院, 沈阳 110034)

摘要: 认知诊断的一个理想应用场景是在小样本情境的课堂测试中提供学生的诊断信息, 进而促进补救教学; 当前多数的认知诊断模型需较大的样本量进行参数估计, 并且有些参数估计方法存在计算效率问题不能提供及时的诊断反馈。基于贝叶斯网的认知诊断方法可以实现小样本情况下诊断分类, 并且能够提供及时的诊断反馈, 这对于推进认知诊断在实践中的应用提供了可能。研究尝试使用贝叶斯网络方法进行小样本认知诊断分类, 模拟研究表明: 贝叶斯网络方法的诊断分类性能优于同样适用于小样本的海明距离法。

关键词: 认知诊断评估; 贝叶斯网络; 课堂评估; 小样本

中图分类号: B841.2

文献标志码: A

文章编号: 1003-5184(2024)03-0264-09

1 引言

认知诊断评估(Leighton & Gierl, 2007a)是心理测量学和认知心理学的结合, 是新一代测量理论的核心。其将学生个体对知识内容的解答过程分解成多个认知成分或加工过程, 称其为认知属性, 基于这些认知属性, 通过专家构建测验题目与属性之间的映射关系, 即测验 Q 矩阵(Embretson, 1984; Tatsuoka & Tatsuoka, 1997), 然后使用基于概率的认知诊断模型对被试的属性掌握情况进行诊断分类。认知诊断评估的一个理想应用场景是在小样本情境的课堂测试中提供学生的诊断信息, 进而促进补救教学。但目前的认知诊断模型大多都是通过基于大样本的参数估计来实现题目参数与被试知识状态的估计, 当样本量较小时会导致算法不收敛, 或参数估计误差增大, 进而影响对被试分类的结果(Liu & Cheng, 2018), 而且有些参数估计方法的运算效率较低, 例如马尔可夫链-蒙特卡罗算法(Markov chain Monte Carlo algorithm, MCMC; DiBello et al., 2007; Henson, Templin, & Willse, 2009; von Davier, 2005, 2008)即使是对比较简单的模型, 其运行时间也比较长, 无法实现课堂教学评估中及时反馈的需求。所以这些适用于大样本的模型并不能够在课堂评估中提供及时的诊断反馈进而真正指导教学促进个性化学习的实现(Huff & Goodman, 2007)。

另一方面, 也有研究者开发了非参数的认知诊断方法, 例如神经网络方法(Neural network; Gierl,

Zheng, & Cui, 2008; Shu, Henson, & Willse, 2013; Cui, Gierl, & Guo, 2016), 粗糙集方法(Rough set; 唐小娟, 丁树良, 俞宗火, 2016), 支持向量机法(Support vector machine; 汪文义, 丁树良, 宋丽红, 邝铮, 曹慧媛, 2016)。这些非参数方法的实现原理在本质上相似: 都是充分利用测验蓝图 Q 矩阵的信息, 通过知识状态与理想反应模式(Ideal Response Pattern, IRP)之间的映射关系, 使用各种方法实现分类, 可以使用这种映射关系信息训练神经网络或支持向量机等机器学习模型进而实现分类, 也可以使用粗糙集学习这种映射关系之中的规则实现分类。这些方法都是在已知题目所测量的属性模式即 Q 矩阵信息时, 利用知识状态与理想反应模式之间的映射关系就可以实现, 也就是说无需考虑题目的猜测失误参数等信息, 所以这些方法中并不能纳入实际的题目作答反应信息。但现实中即使是考察同一属性模式的题目, 从实际题目作答反应中估计的猜测与失误参数也可能是不同的, 利用这种参数还能衡量实测数据与理论模型的拟合程度。所以在以上这些非参数方法中仅仅依赖理论建构的理想反应模式信息而缺失题目实际作答反应信息势必会降低模型分类准确率。此外还有一些非参数方法如聚类分析方法(Cluster analysis; Chiu, Douglas, & Li, 2009), 海明距离法(Hamming distance; Chiu & Douglas, 2013)会纳入题目的实际作答反应信息, 这两种方法或者对观察反应模式进行聚类, 或者计算观察

* 基金项目: 教育部人文社科青年基金“基于贝叶斯网的认知诊断 Q 矩阵估计与检验方法”(22YJCZH160), 辽宁省属本科高校基本科研业务费专项资金资助。

通信作者: 汪玲玲, E-mail: wlling-007@163.com。

反应模式与理想反应模式之间距离,从而实现诊断分类。但是这时两种方法对被试的分类效果仍然依赖于该测验的理论建构的合理程度,也就是真实作答反应与理想反应模式的接近程度,若理论建构不够精准,那么这些非参数方法的分类效果也会打折扣。

如何在参数类认知诊断模型与非参数认知诊断方法之间寻求一种平衡?既能够克服传统参数认知诊断模型对大样本的依赖,真正推进认知诊断评估在教学实践中的应用,同时又能够克服现有非参数认知诊断方法的局限,基于此研究者提出使用贝叶斯网络进行小样本认知诊断的新思路来尝试解决以上问题。

2 贝叶斯网络及其在认知诊断中的应用的现状

贝叶斯网络(Bayesian Networks, BNs)是结合概率论与图论的概率模型(张连文,郭海鹏,2006)。一个贝叶斯网络由一个结构模型和一组条件概率组成,其中结构模型是有向无环图,图中的节点表示随机变量,有向边表示变量之间的相关或因果依赖关系,有向边所联结的两个节点中,都是从父节点指向子节点。这种依赖关系用网络中每个节点在给定父节点前提下的条件概率来量化。例如,图 1 就是一个用于认知诊断的贝叶斯网络(Levy & Mislevy, 2017),在这个结构中属性和题目变量都是节点,题目与属性之间的依赖关系用有向边来表示,整个贝叶斯网络的结构就包含了传统认知诊断中的 Q 矩阵信息。每一个节点都对应着一个条件概率分布表或者边缘概率分布表,如题目 4 测量了属性 1 和属性 2 两个变量,这时的条件概率分布表如图所示,其中的每个概率值也称为贝叶斯网的参数。贝叶斯网络通过概率推理来实现诊断分类。根据网络结构,多维变量的联合概率分解为一系列条件概率的乘积,贝叶斯网利用变量之间的条件独立关系对联合分布进行分解,减少模型的参数个数,从而使知识的获取与表达得以简化(Neapolitan,2004)。

通常获得贝叶斯网络的参数有两种方法,可以由领域专家指定,这在计算机领域早期的专家系统中经常使用(薛薇,陈欢歌,2012),也可以使用参数估计算法从大规模数据中学习条件概率和边缘概率参数。国外已经有一些研究将贝叶斯网络用于认知诊断分类中(Sinharay, 2006; Sinharay & Almond, 2007; Wu, 2013; Almond, Mislevy, Steinberg, Yan, & Williamson, 2015; Levy & Mislevy, 2017)。相较于传统认知诊断方法,该方法没有明确的函数形式,只有表征属性层级结构和 Q 矩阵的网络结构,以及一组条件概率和边缘概率参数构成。将贝叶斯网络用于

认知诊断进行参数估计的一个难点在于网络中的节点变量并不都是已知的,被试的认知属性是需要通过网络参数估计来求取的,这对应着贝叶斯网络中具有缺失值情况下的参数估计问题。处理具有缺失值情况下的参数估计问题通常有两种算法:MCMC (Sinharay, 2006; Sinharay & Almond, 2007) 算法和 EM (expectation maximization, Korb & Nicholson, 2010) 算法。已有将贝叶斯网络用于认知诊断的研究中,都是通过 MCMC 参数估计方法来解决这一问题的。这一解决方案使得贝叶斯网络方法与传统的具有明确函数形式的参数类认知诊断模型面临相同的困境,即 MCMC 参数估计方法计算效率较低,难以在课堂教学中提供及时的诊断反馈信息。

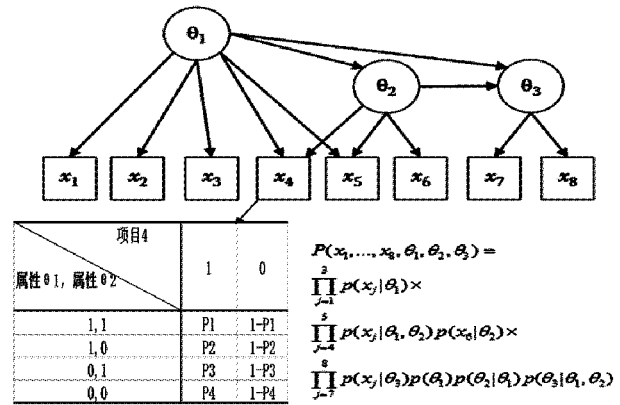


图 1 作为认知诊断模型的贝叶斯网络

此外,国内也有一些研究将贝叶斯网络用于认知诊断中。目前喻晓锋(2009, 2023),王军(2012)使用贝叶斯网络对被试进行分类,但是他们的设计思路是将贝叶斯网作为一种有监督机器学习的分类器,训练数据集中既要包含被试对题目的作答反应,同时也要包含每一个被试对题目作答反应模式背后的知识状态。将被试的知识状态仅用一个多分类的变量来表达,而不是用多个属性变量来表达知识状态,这样设计的缺陷就是当知识状态较多时,被试分类判别将变得很不方便,而且有监督机器学习的分类器在实证研究中的应用受到寻找训练数据集的局限。喻晓锋等人(2011)利用贝叶斯网进行结构学习,从被试的属性掌握模式中挖掘出属性间的层级关系,学习所得到的层级关系可以用来验证由专家得到的层级关系,为认知诊断对被试分类奠定基础。宋丽红等人(2016)使用贝叶斯网络,通过已有认知诊断模型分析测验数据来验证认知模型,也就是属性间的层级关系。这两个研究都是使用贝叶斯网络构建认知诊断中的属性层级关系,还没有涉及到使用贝叶斯网络对被试进行直接分类。也就是说此时

贝叶斯网络与认知诊断计量模型之间的关系是割裂的,还没有整合进一个整体的测量模型中。

3 贝叶斯网络参数估计 IRP-EM 新算法实现

贝叶斯网络的参数估计实际上就是条件概率和边缘概率的求取。如果网络节点中每一个变量的取值都是已知的,那么这种概率的求取问题就变成了极大似然估计(张连文,郭海鹏,2006),也就是通过对网络节点的频次计数即可以实现(Lee, 2011)。下面我们以一个简单的示例来说明如何应用极大似然估计方法进行频次计算,也就是使用完全的没有缺失的数据来训练贝叶斯网络得到网络参数。假设一个测验中有两个题目,项目 1 测量属性 1 和属性 2,项目 2 测量属性 2,这就构建了一个简单的用于认知诊断的贝叶斯网络结构,如图 2 所示。假设我们能够通过专家标定得到某一作答反应下被试的真实属性掌握模式,那么这批数据就可以用来训练贝叶斯网络进而求得参数。为了说明这一问题,我们构造了包含 9 例个案的人工数据,其在图 2 的贝叶斯网络结构中 4 个结点上的取值见表 1(这 9 个数据仅仅作为参数估计方法说明的人工数据,其构成不一定具有现实意义)。基于各案例的实际情况,可以得到各属性的边缘概率以及题目的条件概率,例如,某学生掌握属性 1 的概率为 $4/9$,掌握属性 2 的概率为 $5/9$ 。如下表 2 所示,在属性 1 和属性 2 都掌握情况下,答对项目 1 的概率为 1,两个属性都没有掌握的情况下答对项目 1 的概率为 0,掌握属性 2 而没有掌握属性 1,答对项目 1 的概率为 $2/3$;掌握属性 1 而没有掌握属性 2,答对项目 1 的概率为 $1/2$ 。对项目 2 的作答概率的求取类似项目 1,就不一一赘述。获取了贝叶斯网的参数(条件概率和边缘概率)之后,就可以使用该贝叶斯网预测被试的知识状态,向网络输入题目作答反应,通过概率推理获取表征被试知识状态的几个节点的取值,进而像传统的函数形式认知诊断模型一样估计被试的知识状态。

以上求取条件概率和边缘概率的频次计数过程就是贝叶斯网中极大似然估计的实现(张连文,郭海鹏,2006)。通过极大似然估计来实现贝叶斯网络参数估计的过程本质就是一种有监督学习的过程,这与使用神经网络或者支持向量机(SVMs; Liu & Cheng, 2018; 汪文义等, 2016)用于认知诊断的过程类似。这类问题的关键就是需要有一批已知被试知识状态的作答反应数据来训练各种机器学习模型。受到已有非参数认知诊断方法的启发,理想反应模式与属性掌握模式之间的映射关系包含非常充分的测验信息,可以有效地用来进行诊断分类。

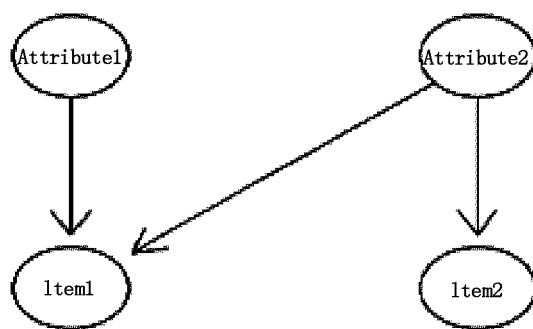


图 2 用于认知诊断的贝叶斯网络

表 1 用于训练该贝叶斯网络的数据

	Attribute1	Attribute2	Item1	Item2
case1	1	1	1	1
case2	1	0	1	0
case3	0	1	1	1
case4	0	0	0	0
case5	1	1	1	1
case6	1	0	0	0
case7	0	1	0	1
case8	0	0	0	0
case9	0	1	1	0

表 2 使用训练数据集通过极大似然方法求取频次计数来估计的条件概率和边缘概率表

P(Attribute1)			
Attribute1 = 1	Attribute1 = 0		
4/9	5/9		
P(Attribute2)			
Attribute1 = 1	Attribute1 = 0		
5/9	4/9		
P(Item1 Attribute1 , Attribute2)			
Attribute1	Attribute2	Item1 = 1	Item1 = 0
1	1	1	0
1	0	1/2	1/2
0	1	2/3	1/3
0	0	0	1
P(Item2 Attribute2)			
Attribute2	Item2 = 1	Item2 = 0	
1	4/5	1/5	
0	0	1	

那么使用理想反应模式及对应的属性掌握模式信息对贝叶斯网络进行训练,通过极大似然方法估计概率参数应该也是可行的。但是这种只依靠理想反应模式与对应的属性掌握模式之间映射关系的方法并不能纳入实测的题目作答反应信息。如果要在贝叶斯网络结构中纳入实测的题目作答反应信息,那么题目的节点值是已知的,题目作答反应所对应

的被试的知识状态信息在网络结构中是未知的,这就涉及到缺失值情况下的参数估计问题。而贝叶斯网络的参数估计方法中除了以上介绍的具有完备数据情况下的极大似然估计方法以及在认知诊断中处理缺失值的MCMC方法,还有EM算法可以处理缺失值问题,但是将贝叶斯网用于认知诊断的研究中,还没有使用EM算法估计参数的研究。因为EM算法在实际应用中会遇到不收敛的问题,尤其是在样本量较少的情况下(Ma & Jiang,2020)。

综上,单纯使用IRP信息的极大似然估计方法没有纳入实测题目作答反应信息,而传统的EM算法在样本量较少时容易出现不收敛问题。将两种算法结合起来是否能够克服二者各自的缺点呢?新的参数估计方法先使用理想反应模式信息通过极大似然估计方法估计贝叶斯网的参数,将这些参数作为先验信息进而使用EM算法继续进行参数估计,从而克服了原有EM算法在样本量较小时不容易收敛的问题。该算法的实现通过贝叶斯网专业计算软件Netica(<http://www.norsys.com>)实现,Netica软件本身可以实现完备数据情况下的极大似然参数估计算法,也可以实现具有缺失数据情况下的EM算法。而在本研究认知诊断的小样本情境中,将两种参数估计方法结合即可实现新的参数估计方法。

通过Netica软件设计贝叶斯网络结构如图3所示,该测验测量5个认知属性,属性之间相互独立,共有30个题目,每个题目与属性之间的映射关系(Q矩阵信息)也通过网络结构表征出来。然后向网络中输入五个独立属性的理想反应模式以及对应的被试知识状态数据,根据独立的属性层级结构,共有32种知识状态,这些理论的知识状态与包含30个题目的Q矩阵信息进行关联共得到 32×30 的题目作答反应矩阵,也就是理想反应模式(IRP),将IRP和被试知识状态信息共同输入所建构的贝叶斯网,此时运用具有完备数据的极大似然估计得到贝叶斯网的初始条件概率参数,这也是BN-IRP算法的实现。用该贝叶斯网处理被试实际作答反应信息进而预测被试知识状态。与其他非参数认知诊断方法类似,此时贝叶斯网的条件概率参数反应的是理想反应模式及对应知识状态所包含的测验理论信息。进一步的,若继续对贝叶斯网的条件概率参数进行校准,需纳入被试对题目的实际作答反应信息,并且这些实际作答反应信息背后的被试真实知识状态是未知的,此时就结合能够在缺失值情况下进行参数估计的EM算法实现,先前使用BN-IRP算法实现的极大似然估计的参数作为EM算法进行参数估计的初始值,进而实现BN-IRP-EM算法。图3

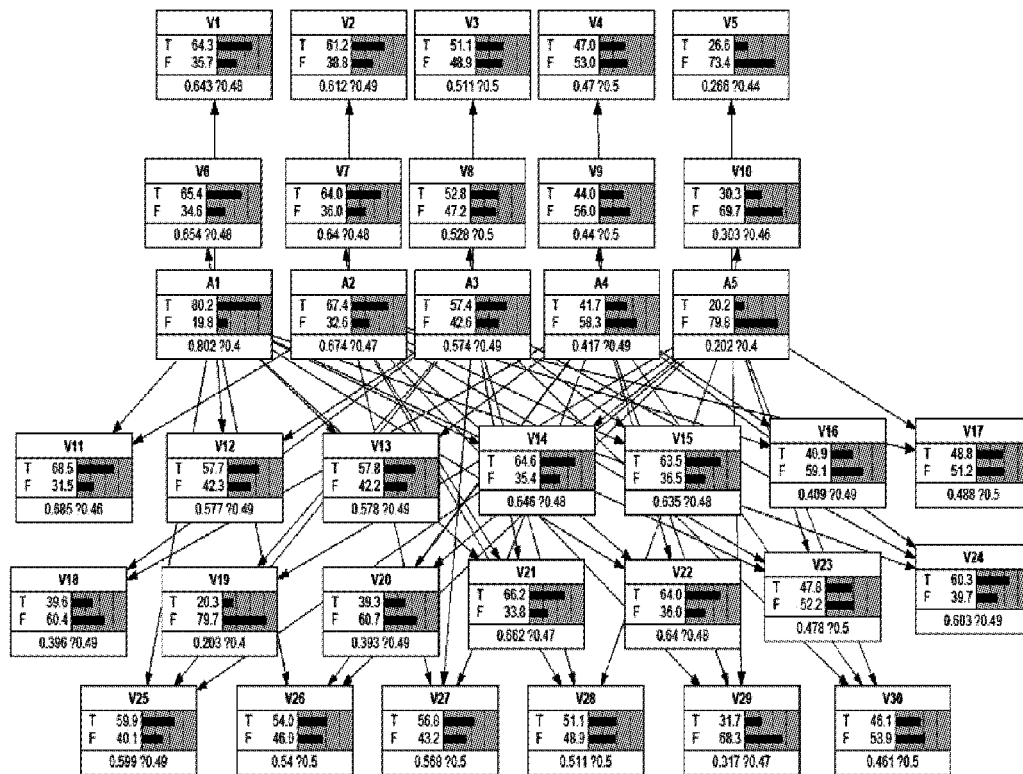


图3 用于认知诊断的贝叶斯网络在Netica中的结构设计

A1	A2	T	F
T	T	99.67	0.33
T	F	19.306	80.694
F	T	67.29	32.71
F	F	9.173	90.827

图4 贝叶斯网中某一个节点下的条件概率参数值

展示了用于认知诊断的贝叶斯网的具体结构,包含Q矩阵信息和5个独立属性,在每个节点变量中,T代表该节点取值为1,F代表该节点取值为0,其中的数值是百分比,即该节点的条件概率参数。

以图4为例说明贝叶斯网中的具体参数信息,根据图3的网络结构,题目V11测量了A1、A2两个认知属性,那么A1、A2的取值组合共有四种,在这四种取值下,题目V11的取值为1(T)或者0(F)的概率分别在图4中呈现出来。依此类推,该网络结构中其他节点的条件概率也以同样的形式存储在条件概率分布表中。当估计参数之后,即可根据这些条件概率参数进行概率推理,进而求出被试的真实知识状态。接下来,本文通过模拟实验研究探查基于贝叶斯网络的新方法(下文简称BN-IRP-EM)进行认知诊断分类的准确性。

4 贝叶斯网在小样本情况下的诊断分类性能探究

4.1 研究方法

Chiu和Douglas(2013)提出了一种可用于小样本认知诊断的方法,即通过计算观察反应模式与理想反应模式之间的海明距离实现。对BN-IRP-EM方法在小样本情况下的诊断分类性能的评价采用与这个研究进行对比的方法。具体的评价指标为平均属性判准确率(Average attribute classification rate, AACR)和模式判准确率(Pattern classification rate, PCR)。

具体设置如下:属性个数为5个,测验长度为30题,Q矩阵如表3所示,题目的猜测和失误参数取自(0,0.3)的均匀分布。被试的作答反应使用DINA模型(Deterministic Input, Noisy Output “AND” gate model; Macready & Dayton, 1977; Junker & Sijtsma, 2001), R-RUM (Reduced - Reparameterized Unified Model, R-RUM, Hartz, 2002) 和ACDM (de la Torre, 2011) 模拟产生。被试的知识状态有两种产生方式:一种是从所有可能的属性掌握模式的均匀分布中随机抽取样本量为30, 50, 100, 200, 500的被试,每种知识状态被抽取的概率都为 $1/2^K$ ($K=5$);另一种是模拟被试的知识状态服从多维正态分布,参考已有研究(Chen, 2017)的模拟方式。模拟被试的知识状态后,根据Q矩阵和参数设置产生被试的作答反应数据。综合以上的模拟条件,共有三个模型(DINA, R-RUM, ACDM) \times 五种样本量水平(30, 50, 100, 200, 500) \times 两种知识状态分布形态(uniform, mvn) 30种模拟情况。每种情况模拟30次,取30次均值为最终结果。

根据已有的Q矩阵构建贝叶斯网络结构,具体的说,每一个题目与所测量的属性之间的映射关系,在网络结构中体现为题目节点与认知属性节点之间的有向边的联结,如图3所示,其中所有题目节点与认知属性节点之间的映射关系都可以通过有向边的联结表达出来,也就构成了整个贝叶斯网的结构。模拟数据产生后,根据Q矩阵,产生测验的理想反应模式,然后将理想反应模式和对应的属性掌握模式作为训练数据训练贝叶斯网,然后根据题目作答反应信息使用EM算法继续估计贝叶斯网的参数,这个过程使用Netica (<http://www.norsys.com>) 软件实现。将贝叶斯网络的诊断分类结果与同等条件下Chiu和Douglas(2013)提出的海明距离方法进行比较,并且为了说明BN-IRP-EM算法的优势,将该方法也与BN-IRP方法进行比较,即仅使用理想反应模式训练网络得到网络参数,而不继续使用EM算法估计参数。

表3 测验Q矩阵设计

	A1	A2	A3	A4	A5		A1	A2	A3	A4	A5
I1	1	0	0	0	0	I16	0	1	0	1	0
I2	0	1	0	0	0	I17	0	1	0	0	1
I3	0	0	1	0	0	I18	0	0	1	1	0
I4	0	0	0	1	0	I19	0	0	1	0	1

续表 3

	A1	A2	A3	A4	A5		A1	A2	A3	A4	A5
I5	0	0	0	0	1	I20	0	0	0	1	1
I6	1	0	0	0	0	I21	1	1	1	0	0
I7	0	1	0	0	0	I22	1	1	0	1	0
I8	0	0	1	0	0	I23	1	1	0	0	1
I9	0	0	0	1	0	I24	1	0	1	1	0
I10	0	0	0	0	1	I25	1	0	1	0	1
I11	1	1	0	0	0	I26	1	0	0	1	1
I12	1	0	1	0	0	I27	0	1	1	1	0
I13	1	0	0	1	0	I28	0	1	1	0	1
I14	1	0	0	0	1	I29	0	1	0	1	1
I15	0	1	1	0	0	I30	0	0	1	1	1

4.2 研究结果

海明距离法 (Hamming Distance, H - D, Chiu & Douglas, 2013) 和基于贝叶斯网的两种方法在不同条件下的判准率如表 4、表 5 所示。从表 4 的平均属性判准率中可以看到, 在 DINA 模型下, 当知识状态服从均匀分布时, BN - IRP 法的 AACR 稍低于 H - D 法, 在不同样本量水平下 AACR 比较稳定, 没有太大变化, 此时 BN - IRP - EM 法的 AACR 高于 H - D 法和 BN - IRP 法, 并且随着样本量增大, AACR 有逐渐增大趋势。这是因为在 BN - IRP - EM 算法中不但能够纳入 IRP 理论建构的信息, 而且能够纳入题目实际作答反应信息。当知识状态服从多元正

态分布时, BN - IRP 法就已经高于 H - D 法, BN - IRP - EM 法更是高于 BN - IRP 法。在 RRUM 模型下, 无论知识状态服从均匀分布, 还是多元正态分布, BN - IRP 法的 AACR 都高于 H - D 法, 而 BN - IRP - EM 法更是高于 BN - IRP 法, 并且平均高出 3 个百分点左右, 随着样本量的增大, 这种高出的幅度越来越大。在 ACDM 模型下, 无论知识状态是服从均匀分布还是多元正态分布, BN - IRP 方法也依然高于 H - D 法, 并且 AACR 所高出的幅度比前两个模型都更多, BN - IRP - EM 方法也同样高于 BN - IRP 方法, 并且高出的幅度超过 3 个百分点。

表 4 三种方法在不同样本量水平下的平均属性判准率 (AACR)

		样本量	30	50	100	200	500
DINA	uni	H - D	0.953	0.954	0.951	0.952	0.952
		BN - IRP	0.952	0.953	0.950	0.951	0.951
		BN - IRP - EM	0.956	0.960	0.962	0.962	0.961
	mvn	H - D	0.949	0.946	0.949	0.950	0.949
		BN - IRP	0.954	0.951	0.953	0.954	0.953
		BN - IRP - EM	0.959	0.956	0.960	0.962	0.963
RRUM	uni	H - D	0.891	0.891	0.892	0.892	0.893
		BN - IRP	0.903	0.904	0.904	0.904	0.914
		BN - IRP - EM	0.936	0.939	0.944	0.957	0.959
	mvn	H - D	0.896	0.900	0.897	0.898	0.902
		BN - IRP	0.914	0.913	0.912	0.912	0.916
		BN - IRP - EM	0.936	0.939	0.951	0.960	0.961
ACDM	uni	H - D	0.841	0.843	0.843	0.844	0.844
		BN - IRP	0.865	0.868	0.860	0.869	0.869
		BN - IRP - EM	0.908	0.917	0.918	0.951	0.955
	mvn	H - D	0.853	0.856	0.854	0.854	0.859
		BN - IRP	0.884	0.883	0.881	0.881	0.885
		BN - IRP - EM	0.913	0.927	0.941	0.945	0.956

注: 表格中 uni 代表均匀分布的被试; mvn 代表多元正态分布的被试;
H - D 代表海明距离判别方法; BN - IRP 和 BN - IRP - EM 代表两种贝叶斯网分类方法, 以下类同。

表 5 三种方法在不同样本量水平下的模式属性判准率 (PCR)

样本量			30	50	100	200	500
DINA	uni	H - D	0.823	0.821	0.816	0.820	0.819
		BN - IRP	0.806	0.806	0.798	0.804	0.803
		BN - IRP - EM	0.824	0.828	0.833	0.838	0.840
	mvn	H - D	0.801	0.794	0.801	0.806	0.804
		BN - IRP	0.813	0.803	0.808	0.813	0.812
		BN - IRP - EM	0.820	0.835	0.837	0.840	0.841
RRUM	uni	H - D	0.555	0.549	0.556	0.554	0.559
		BN - IRP	0.590	0.588	0.593	0.589	0.595
		BN - IRP - EM	0.710	0.734	0.752	0.810	0.813
	mvn	H - D	0.570	0.582	0.576	0.577	0.594
		BN - IRP	0.636	0.634	0.630	0.628	0.644
		BN - IRP - EM	0.707	0.732	0.777	0.813	0.815
ACDM	uni	H - D	0.371	0.374	0.385	0.379	0.381
		BN - IRP	0.443	0.450	0.456	0.451	0.452
		BN - IRP - EM	0.600	0.612	0.640	0.769	0.795
	mvn	H - D	0.419	0.432	0.422	0.422	0.437
		BN - IRP	0.519	0.518	0.513	0.509	0.524
		BN - IRP - EM	0.622	0.671	0.702	0.745	0.794

从表 5 的属性模式判准率结果来看,在 DINA 模型下,当知识状态服从均匀分布时,BN - IRP 法低于 H - D 法的 PCR,但是 BN - IRP - EM 法的 PCR 高于 H - D 法。而当知识状态服从多元正态分布时,无论是 BN - IRP 法还是 BN - IRP - EM 法都高于 H - D 法的 PCR。在 RRUM 模型下,无论是均匀分布还是多元正态分布中,H - D 法的 PCR 为 50% 多,而 BN - IRP - EM 法的 PCR 为 70% 多,在样本量为 200 至 500 时,甚至达到了 80%,增幅达到 20 个百分点左右,而 BN - IRP 法的判准率介于 H - D 法和 BN - IRP - EM 法之间。在 ACDM 模型下,H - D 法的 PCR 更低,在多元正态分布下为 40% 多,在均匀分布下甚至达到了 30% 多,但是使用 BN - IRP 法明显增加了 PCR,平均增加 10 个百分点左右,而最好的结果依然是 BN - IRP - EM 法,比 H - D 法的 PCR 平均增加了 20 多个百分点。

5 小结与讨论

认知诊断的理想应用场景是在课堂测试中提供诊断信息,进而促进补救教学,而课堂评估的数据通常都是小规模,这时使用传统参数类认知诊断方法可能导致参数估计不收敛,或者参数估计误差增大进而对被试分类不准确,或者计算效率较低不能够及时提供诊断反馈。本研究提出使用理想反应模式训练贝叶斯网络并结合 EM 算法估计参数进行诊断分类的方法,考察了贝叶斯网络在小样本情境下的诊断分类表现,将 BN - IRP 法、BN - IRP - EM 法与同

样基于理想反应模式 (IRP) 的海明距离法 (Chiu & Douglas, 2013) 进行比较。理想反应模式和对应的知识状态包含了所开发测验的非常充分的信息。现在已经有很多非参数方法或者机器学习方法都基于这一信息实现认知诊断分类。例如神经网络 (Gierl, Zheng, & Cui, 2008; Shu, Henson, & Willse, 2013; Cui, Gierl, & Guo, 2016), 聚类分析 (Chiu, Douglas, & Li, 2009), 海明距离 (Chiu & Douglas, 2013), 粗糙集 (唐小娟, 丁树良, 俞宗火, 2016), 支持向量机 (汪文义等, 2016) 等。但是这些非参数或机器学习方法所面临的一个共同问题就是无法纳入实测的题目作答反应信息,从而可能降低分类准确率。本研究所提出的 BN - IRP - EM 新方法可以将 IRP 信息和实测的题目作答反应信息结合,所以能够提高模型的分类型准确率,模拟研究结果已经证实了这一结论。从而为小样本认知诊断问题提供了新的解决方案。虽然在本次初始的探索性研究中实现贝叶斯网的 IRP - EM 算法依赖于特定的贝叶斯网专业软件 (Netica),但是在未来若要将该算法用于课堂教学实践中,可以开发封装该算法之后的智能学习系统,实现界面图形化操作,从而快捷简便的对学生的学习情况进行诊断分类,进而根据分类结果,教师进行有针对性的补救教学,智能学习系统可以进行个性化的学习资源推荐。

另外,本研究仅仅初步探讨了贝叶斯网络用于认知诊断的一种新思路。未来可以探查这种方法在

更广泛情境下的诊断分类性能,例如在符合其他认知诊断模型假设的数据中的表现,在Q矩阵界定有误差的情况下的诊断分类性能以及具体识别这种误差的检验能力。为了不失一般性,本研究所考察的情境都是假设属性之间没有层级关系,未来研究可以考察具有属性层级关系的情境中贝叶斯网的表现,贝叶斯网络结构的构建非常灵活,可以将属性之间的层级关系与Q矩阵信息同时纳入一个模型中(如图1所示)。另外本研究仅仅考察了0~1计分的认知诊断情况,未来研究也可以考察贝叶斯网在多级计分中的分类表现。本研究仅仅使用模拟研究来考察贝叶斯网的诊断分类性能。未来研究中可以考虑使用实证研究进一步验证该方法的表现。贝叶斯网的方法为认知诊断评估在课堂教学中的应用提供了可能,未来应有更多的基于贝叶斯网的认知诊断评估促进形成性评价实现个性化教学的实践研究。

参考文献

- 宋丽红,汪文义,戴海琦,丁树良. (2016). 基于贝叶斯网的认知诊断模型构建. *心理科学*, 39(4), 783-789.
- 唐小娟,丁树良,俞宗火. (2016). 粗糙集理论在认知诊断中的应用. *心理科学*, 39(4), 790-795.
- 涂冬波,蔡艳,戴海琦. (2012). 基于DINA模型的Q矩阵修正方法. *心理学报*, 44(4), 558-568.
- 汪文义,丁树良,宋丽红,邝铮,曹慧媛. (2016). 神经网络和支持向量机在认知诊断中的应用. *心理科学*, 39(4), 777-782.
- 王军. (2012). *贝叶斯网络在认知诊断中的应用* (硕士学位论文). 南京师范大学,南京.
- 薛薇,陈欢歌. (2012). *基于Clementine的数据挖掘*. 中国人民大学出版社.
- 喻晓锋. (2009). *贝叶斯网络在认知诊断中的应用* (硕士学位论文). 江西师范大学,南昌.
- 喻晓锋,丁树良,秦春影,陆云娜. (2011). 贝叶斯网在认知诊断属性层级结构确定中的应用. *心理学报*, 43(3), 338-346.
- 喻晓锋,肖遇春,秦春影. (2023). 基于贝叶斯网模型的多级计分诊断测验分类及比较研究. *心理与行为研究*, 21(1), 49-57.
- 张连文,郭海鹏. (2006). *贝叶斯网引论*. 北京:科学出版社.
- Almond, R. G., Mislevy, R. J., et al. (2015). *Bayesian networks in educational assessment*. New York, NY: Springer. doi: 10.1007/978-1-4939-2125-6
- Chiu, C. Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74, 633-665. doi: 10.1007/s11336-009-9125-0
- Chiu, C. - Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, 30, 225-250. doi: 10.1007/s00357-013-9132-9
- Chen, J. S. (2017). A residual-based approach to validate Q Matrix specifications. *Applied Psychological Measurement*, 41, 277-293. doi: 10.1177/0146621616686021
- Cui, Y., Gierl, M., & Guo, Q. (2016). Statistical classification for cognitive diagnostic assessment: An artificial neural network approach. *Educational Psychology*, 36(6), 1065-1082. doi: 10.1080/01443410.2015.1062078
- DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics psychometrics* (Vol. 26, pp. 979-1030). Amsterdam, The Netherlands: Elsevier Science Publishers.
- Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, 49(2), 175-186. doi: 10.1007/BF02294171
- Gierl, M. J., Zheng, Y., & Cui, Y. (2008). Using the Attribute Hierarchy Method to Identify and Interpret Cognitive Skills that Produce Group Differences. *Journal of Educational Measurement*, 45(1), 65-89.
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, IL.
- Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). New York, NY, US: Cambridge University Press.
- Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74, 191-210. doi: 10.1007/s11336-008-9089-5.
- Korb, K., & Nicholson, A. E. (2010). *Bayesian artificial intelligence* (2nd Ed). London: Chapman & Hall.
- Leighton, J. P., & Gierl, M. J. (Eds.). (2007a). *Cognitive diagnostic assessment for education: Theory and practices*. Cambridge: Cambridge University Press.
- Lee, J., & Corter, J. E. (2011). Diagnosis of subtraction bugs using Bayesian networks. *Applied Psychological Measurement*, 35(1), 27-47. doi: 10.1177/0146621610377079
- Ma, W. C., & Jiang, Z. H. (2020). Estimating Cognitive Diagnosis Models in Small Samples: Bayes Modal Estimation and Monotonic Constraints. *Applied Psychological Measurement*, 45

- (2). DOI:10.1177/0146621620977681.
- Neapolitan, R. E. (2004). *Learning Bayesian Networks*. Prentice Hall.
- Shu, Z., Henson, R., & Willse, J. (2013). Using neural network analysis to define methods of DINA model estimation for small sample sizes. *Journal of Classification*, 30, 173 – 194. doi:10.1007/s00357-013-9134-7
- Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics*, 31, 1 – 33.
- Sinharay, S., & Almond, R. G. (2007). Assessing the fit of cognitive diagnostic models; A case study. *Educational and Psychological Measurement*, 67, 239 – 257.
- Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3), 1 – 22. URL <http://www.jstatsoft.org/v35/i03/>.
- Tatsuoka, K. K., & Tatsuoka, M. M. (1997). Computerized cognitive diagnostic adaptive testing: Effect on remedial instruction as empirical validation. *Journal of Educational Measurement*, 34(1), 3 – 20. doi:10.1111/j.1745-3984.1997.tb00504.x
- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service.
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61, 287 – 301. doi:10.1348/000711007X193957
- Wang, C. (2013). Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing With Short Test Length. *Educational and Psychological Measurement*, 73(6), 1017 – 1035. doi:10.1177/0013164413498256
- Wu, H. Y. (2013). *Comparison of general diagnostic models (GDM) and Bayesian networks using a middle school mathematics test* (Unpublished doctoral dissertation). Florida State University, America.

The Application of Bayesian Networks in Cognitive Diagnosis with Small Sample Size

Wang Lingling

(School of Educational Science, Shenyang Normal University, Shenyang 110034)

Abstract: In this study, Bayesian networks (BN) are proposed to conduct cognitive diagnosis in a small sample. The combination of IRP (Ideal Response Pattern) and EM parameter estimating methods can overcome the shortcomings of IRP and EM algorithms respectively, and can realize the BN application in a small sample size. The Monte Carlo simulation study is used to examine the performance of the BN-IRP-EM method in a small sample size, compared with the hamming distance method. In the simulation study, the pattern match ratio and average attribute match ratio are used as criteria to evaluate the classification accuracy of different approaches. To demonstrate the effectiveness of the BN-IRP-EM method, the BN based purely on the IRP method is adopted as the controlling method, another controlling method is the hamming distance (H-D) method. The results are as follows: the classification rate of the BN-IRP method is slightly higher than that of the H-D method which is based on the same IRP information except for some conditions. The classification rate of the BN-IRP-EM method is higher than the BN-IRP method and the H-D method in all circumstances. In the BN-IRP-EM condition, due to the incorporation of the empirical information, the classification rate is gradually increasing with the increase in sample size. These outcomes demonstrated that the BN-IRP-EM method could be used in a small sample size and can promote the application of CDA in classroom assessment.

Key words: cognitive diagnostic assessment; bayesian networks; classroom evaluation; small sample size