

• 思路与方法 •

中医证候的认知思路及其数据挖掘方法

袁世宏¹ 王天芳² 张连文³

(1. 香港大学专业进修学院; 2. 北京中医药大学基础医学院; 3. 香港科技大学计算机科学系)

基金项目: 国家重点基础研究发展计划 (“973”计划)资助项目 (2003CB517106, 2011CB505105);

香港研究资助局项目 (622408)

通讯作者: tianfangwang2000@yahoo.com.cn

摘要】从认知角度而言,中医证候问题的实质是“类”与“分类”,证候是对疾病人群类特征的刻画,辨证是基于该特征对人群的分类。但中医传统对证候“类”的界定是基于人脑的主观经验的积累和总结,要建立证候“类特征”的客观标准,就应该用无监督的聚类学习方法,来构建中医证候的数学模型,为客观定量地辨识证候类别提供帮助。

关键词】数据挖掘;无监督学习;聚类分析;证候类别

要寻找适合中医证候研究的数理统计或数据挖掘方法,就必须清楚中医证候的研究究竟是个什么性质的问题,只有清楚这个问题才能够寻找真正解决中医证候研究问题所需要的数学思想及其适宜的方法。

1 认知分类与证候研究

1.1 划分、分类与认知

对所感知的事物或现象进行划分是人类乃至于动物的本能,它是把一个对象的集合简单地分割成不同的组。要达到这个目的,就必须在划分时有所依据,而且依据的特征能够有效地进行划分,形成不同的类别。由此产生了分类的概念。

分类是划分的特殊形式,是以对象的本质属性或显著特征为根据所作的划分。它以划分为基础,又和一般划分有所不同。基于分类的这种特性,使得分类成为人类认知一切事物的基础。从人类知识的积累和发展来看,按照事物的相似性和特征性对他们进行分类似乎长期以来就是人类获取知识过程中重要的一步,更成为人类所有智能活动的基础^[1]。人类一切认知的初始都是基于对所感知和认知对象的分门别类,确定概念以获得认识。因此,分类是一种思考事物的策略和方法,而不是局限于对事物自身的研究,它体现了人类的思维方式,而非简单的一种技术^[2]。也正是如此,分类也就成为一切科学研究的基本策略。

科学的过程,哪怕在它的纯粹描述性阶段,也是以这样的方式在发生和进行着。我们能够有效地

掌握和支配某个领域,也就意味着它的全部概念(思想)被分门别类。通过分类显示了事物或现象间的某些关系,而借助对这些关系的探索使得人类产生对领域知识理解的加深和改进,这也正是科学知识深化发展的基本线索,在科学发展的各个领域,或各门类中均可看到分类对学科发展所发挥的巨大作用。

1.2 分类与证候

毋庸置疑,分类作为一种研究事物的策略和基本方法,是人类认识生命、健康和疾病的基本方法。

早在人类认识疾病的最初期,就已经开始依据疾病外部的表现特征对疾病人群进行分类,如甲骨文记载曾将疾病人群分成疾首、疾齿、疾腹等不同的群体类别。在《素问·上古天真论》中将古人能否把握肾气、遵守养生之道的不同而分成真人、至人、圣人和贤人四类人群。

五行学说就是以“五”数的形式将各种现象分门别类,成为一个对世间万象进行归类的静态分类体系,并以“五”为纲延伸至各个“目”,最终形成一个庞大的理论框架,更成为对自然界万物进行分门别类的形态学分类体系,并在此基础上升华出一个解释世间万象相生相克的动力学功能系统^[3]。五行的意义不仅是一个分类方式,更上升为一个哲学思想体系^[4]。这种分类体系也成为中医进行疾病证候分类的基本方法。比如《灵枢·阴阳二十五人》按照五行理论,结合五色、五音,把人分为木、火、土、金、水五类,又根据角、徵、宫、商、羽的阴阳多少和左右上下将每一类再分成五种,就体质特征将整个人群分成

二十五种类型的群体。《素问·宣明五气篇》更进一步借助于“五”，通过论述五气所病、五脏化液、五邪所乱、五脏所恶、五病所发、五脏所主、五劳所伤等，对病证进行分类。而在《素问·痹论》更直接提出五脏痹和五体痹的病症分类概念。《伤寒论》实际上就是张仲景根据对当时外感热病的认识而对所有外感热患者群的分类，六经辨证就是六类患有外感热病的人群特征辨别。

中医临床辨证过程其实质就是依据患者的症状、体征的群体特征，运用中医理论和哲学范畴进行概括和解释，并进行划分类型。通过长期的大量的医学实践，在中医理论指导下，形成了对不同疾患者群不同层次的症状体征的群体特征认识，进而形成了中医对疾病认识的特有概念——证候。

1.3 证候类特征界限的缺陷

据上可见，中医辨证过程就是在掌握四诊资料基础上对疾患者群的分类过程。在某一个疾病的样本空间中，不同的人群具有的群体症状、体征特征不同，而同一群体则具有相似的群体特征。这个特征及其所反映的领域意义就是中医的证候。

要分类就必须有清晰（量化的）客观的分割“边界”。但在漫长的中医临床实践的发展过程中，历代医家用人脑来构建证候结构，界定证候分割的“边界”，使其不可避免地夹杂着医家个人的主观性、经验性。一个医生临床经验的多少和中医理论认识的深浅，影响了对分类边界的把握程度。这是中医证候分类所有在的客观性问题。其次，由于人脑的认知程度和古代科学水平的局限性，中医对证候的认识具有模糊性和不确定性。这是中医证候分类中存在的量化问题。缺乏客观化和量化成为中医证候分类“边界”最为现代科学所诟病之处。

要解决这一问题，必须基于与中医证候认知完全一致的分类思想来选择方法和建立模型，并借助于数据挖掘方法来解决证候分类“边界”的客观化和量化的缺陷。

2 证候的客观化与无监督学习

2.1 无监督学习的概念及其实现

利用机器学习的原理，可以从流行病学调查数据中进行学习，学习症状数据中隐含的特征，建立证候模型，得到证候类的定量界定意义。但机器学习的具体算法很多，根据有无先验知识的指导，分为有监督学习和无监督学习。在有监督学习，机器不仅收到序列输入值 $X_1, X_2, X_3, \dots, X_n$ （数据），而且也给了

一组想要的输出值 $Y_1, Y_2, Y_3, \dots, Y_m$ ，此时机器学习的目标是让计算机去学习已经创建好的分类系统，这种学习高度依赖于事先确定的分类系统给出的信息。这显然不符合证候诊断标准客观化研究的初衷，因为中医证候的最大问题就是缺乏公认的“金标准”，而证候的客观化研究就是要寻找“金标准”。

无监督学习就是不告诉计算机怎么做，而是让它（计算机）自己去学习怎样做一些事情，去发现数据体中的本质结构、关系或隶属。此时机器只收到输入的数据 $X_1, X_2, X_3, \dots, X_n$ ，它既没有得到监督的靶输出值，也没有从学习环境受到任何奖赏。其目的是建立输入信息的模型，用来做决策和预测。无监督学习最简单的传统例子就是聚类 and 降维^[5]。

一般来讲，数据的真实分布是未知的，但可以学习这个分布的一个模型。几乎所有的无监督学习的工作都可以看作是学习数据中的一个概率模型。而且，数据模型越好，可以压缩和交流新数据的效果就越好，即使没有给予机器监督和奖赏，机器也能合理地评估一个模型。利用此模型，可以对来自患者群体的症状数据进行分类，以便作出病证的诊断标准，便于将来进行数据之间的交流和数据压缩^[6]。

2.2 客观性问题的解决与无监督学习

解决中医证候的客观性问题，是目前中医现代研究的一个重要课题，也是中医发展所面临的最大困境。中医证候客观性不足的根源，从认识论角度来看，辨证理论依据的资料来源于患者的主观感受和或医生对这些资料的主观感觉，缺少客观的度量标准；从获得方式来看，这些理论大多是个体经验的积累和升华，不同的实践经验、专业知识、认识水平对证的辨识也不尽统一，其客观性必然存在问题；即使从目前中医证候标准的建立程序来看，各种证候诊断和评价标准，也均来源于历代文献描述和专家组甚或专家个人的经验，虽不乏深厚的实践积累，但毕竟带有一定程度的主观偏倚，缺乏现代科学研究方法、技术和数理统计学的支持。这启发人们借助于现代的数理统计学的思想和方法，进行证候标准的客观化和规范化的研究，但就目前多数学者所使用的方法和技术看，与中医证候、症状的数据还有不尽合适之处，因为这些方法从根源讲并不是针对中医问题的特殊性所创立，它们背后的数学思想与证候的研究也有不尽合适之处，况且很多方法在研究证候问题时，要求先验知识的加入，也降低了其结论的客观性和真实性。

要想解决中医证候分类“边界”的客观性问题,就要求在对中医证候症状数据进行机器学习时,首先要判别该数据学习方法是具有监督学习或无监督学习。在中医证候(要素)研究中,二者的主要区别就是:在研究伊始是否存在先验的“金标准”的辨证。由于中医证候认识形成过程的经验性和主观性,使得中医证候诊断的标准存在主观性缺乏客观性,存在不确定性缺乏规范性,而且更为重要的是缺乏量化的标准。因此,基于这种标准的学习使得中医证候研究虽多但质量不可靠,结论很难推广。要解决这个问题就不能依赖以前的“金标准”,不能用有监督的学习,应该用没有先验的“金标准”指导的无监督学习。

3 聚类分析与基于模型的聚类

既然假设证候的获得就是对疾患者群的类特征的把握,要保证分类的客观性就要选用无监督学习的方法,而无监督学习最重要的技术就是聚类。因此,要想客观地对疾患者群进行分类,就要采取良好的聚类技术。

3.1 聚类的概念及其分类

聚类分析(Cluster Analysis, CA)就是分门别类,把一大组观察数据分成更小的组,以至于在每一个组内的观察项目相对类似,而不同组间的观察项目则相对不相似。这种分门别类是自动地搜寻一组数据,并自然地将相似对象分进同一组的过程,因此被看作一种无监督学习方法。

聚类分析的具体技术有很多种。如果根据分配对象到类的方式的不同,聚类算法可分成软聚类(soft cluster)和硬聚类(hard cluster)。在硬聚类,每一个对象分到唯一的一个类;软聚类则允许类成员依据程度分在多个类,即一个对象以某种类成员值可以被分在多个类。实际上,大多数软聚类模型也假定一个对象属于唯一的一个类,不过它与硬聚类的区别在于对象归属的类别存在不确定性。这种软、硬聚类的说法又称为:基于距离或相似性的聚类(distance based)和基于模型的聚类(model based)。基于相似度的聚类需要相似度函数来度量对象之间何等相似。最著名的相似性的度量是基于距离,比如欧几里得距离(Euclidean distance)和曼哈顿距离(Manhattan distance)。由于该方法的聚类是依赖于相似性的度量,使得它不可能应用于某些类型的数据,例如名义性的数据(疾病类别),它们是典型的硬聚类。基于模型的聚类是假设数据对象归

属于某个隐模型,这个模型是从数据中学习而产生的,一个模型对应于一个特别的类。一旦模型被确定,基于该模型来聚类就非常容易。由于基于模型的方法可基于复杂的概率模型提供各种类的定义并分配对象,这种方法本质上支持软聚类^[7]。

目前实际应用中最常用的聚类分析主要是基于相似性度量的硬聚类。但基于模型的聚类方法却越来越受到学者的关注并获得巨大的发展。其原因:一是该方法对数据要求的灵活性,二是群体中的类别不是相互排斥的,而且资料内涵是多维性。这种聚类算法的一个关键特性是把聚类问题看作一个最优化的过程,该过程是在聚类结果的整个空间中寻找最大或最小化的,并基于全局或局部定义的一个特定聚类标准函数^[8]。

3.2 传统聚类分析的局限性

传统的聚类分析虽然较其他统计方法,诸如回归分析和判别分析等具有自然的无监督特性,但就目前应用现状而言,仍然具有一些局限性。

首先,目前的大多聚类分析只是居于相似性或距离的度量,针对两两对象之间的相似性进行简单的一维解释,并不能从多维和多层次角度来全曲分析数据并解释数据中真实复杂结构,而中医症状以及症状与证之间的关联性是非常复杂的,具有多维和多层次的复杂联系,这可能是目前传统聚类分析方法所无法解决的,值得欣慰的是,很多学者提出了改进的方法^[9],而且我们也对此作了多年不懈的努力,做了有益的探索性工作^[10]。

其次,应用聚类分析过程中存在的主观性问题。具体表现:一是相似性度量方法和度量值选择的主观性问题。在聚类过程中有很多聚类标准可供选用,对相同的数据,选用的标准不同产生的分类也不同。即便使用相同的距离尺度、聚类参数不同,其结果也大不相同。在实际中必须尝试性地选择使用不同的聚类标准对数据进行反复多次的试验,以求得到所研究学科可能的结果。而要给予学科经验相近似的感受,必须主观地选择合适的相似性度量方法和适应的量值。二是类数多少确定的主观性问题。聚类分析无有效的算法对最后得出的类数是否最优给出明确的回答,尽管相似性的融合层次树图可以有助于决定最优类数,但在何处“切割”此“树”以得到理想的类数是不确定的,类数的确定必须由研究者依据领域知识去经验性地主观确定。虽然有研究者也提出一些非主观的选择聚类个数的方法,但是这些

选择方法仍然都有一定的局限性,通常只适合于具体的情形^[1]。

再次,分层凝集聚类分析中的单通道问题。一旦一个对象被分到某一类,就不能再撤回或取消,即对于早期阶段被“错误”分组的对象不提供重新分组的机会,从而可能产生一些中医理论无法解释的聚类现象,或实际情况无法解释的类群内容。这实际上是传统聚类分析的刚性分割问题,类的成员可是绝对的,在一个给定的条件下,一个样本属于一个和唯一的一个类,类与类之间的对象则尽可能地毫不相干。但中医证的表现实际的人群样本空间中是交织在一起,相互兼夹错杂的,聚类分析这种“刚性”的“楚河汉界”在实际临床的群体中很难找到和定义。要本着中医证候理论的特色建立症状与证之间的关系,要符合中医理论又不脱离实际的诠释,就必须注意聚类分析的这一局限性^[2]。另外,聚类分析的结果只能解释产生结果的样本,不能给与样本以外的任何对象以同样的解释,而且聚类的多结果性和不稳定性也是一个非常重要问题。总之,尽管聚类分析自身缺乏较为深入和广泛的统计学推理来支持,并存在这样和那样的问题,但作为一种探索性的方法,仍然可给予我们以经验性的提示,以便于采取更有利的方法对数据做进一步分析。

3.3 基于模型聚类的概念及其优势

正是由于上述传统聚类方法应用时所产生的这一系列问题,所以当它作为一种探索性工具建立了正式的分类方法后,马上会发现,聚类算法的结果不是他们分析的结局,并出现了一系列新的疑问:结果的类有意义吗?它反映了真实自然的数据结构,抑或仅仅是一种人工方法的选择?这种聚类分析的作用与随意的分类相比有没有真正的优势?针对这些问题,有学者提出用含有统计推断目的的概率聚类模型的方法来解决聚类和分类中存在的上述问题,并为分类方法和策略提供坚实的理论基础^[3]。实际上,机器学习的技巧或艺术就是发展一个模型,使此模型能适用于所要分析的数据,并具有某些想要的特征。

基于模型的聚类方法,首先假设产生数据样本的整个样本空间存在一个统计模型,或者更准确地讲,所研究的数据对象归属于某个隐藏模型(概率分布)。然后通过对所研究数据的学习去寻找并产生一个复杂的概率模型,而每一个模型对应于一个特别的类结构,进而利用由此模型所获得的似然度(或者后验概率)作为最优化的标准,判断该模型与真实数

据的拟合度;最后基于此模型确定聚类的结构、类数及类的特征等知识^[4-16]。

基于模型聚类方法的优势非常明显。主要有:①一旦模型被确定,基于该模型来聚类就非常容易,从模型中不仅可以得到类特征描述的定性认识,而且还可以获得类数目等大量的定量性认识,这对中医证候的量化研究非常具有现实意义。②该方法既可用于连续变量也可用于离散变量,这种进步主要是由于它允许我们根据他们的形状和结构而不是基于数据点之间的近似性来确定类^[7]。针对中医证候研究中症状变量离散性的取值特点,用统计概率来对其进行描述的方法更为合适。③用统计模型的一个优势就是聚类标准的选择较少武断性,因为该方法通常假设样本集中的k个类服从某k个概率分布,而样本数据属于某个类也通过概率来表达,这类概率算法要用到概率统计中的Bayesian定理和极大似然估计法,可以称之为基于概率的统训方法。④该方法具有严格的统计检测,可以对聚类结果进行检测,判断模型的优劣,最终确定类数和类的特征^[8]。在国外基于模型的聚类分析已经获得最广泛的应用,不过在国内尤其在中医领域的研究才刚刚开始,这些优势必将有利于该方法在中医领域的推广和应用。

4 小结

从认知科学角度分析中医证候的发生和概念,中医的辨证过程就是对疾患者群进行分类划分的过程,证候的认识就是通过对类别特征的把握来确定分类的界限或标准。但传统中医辨证理论的分类界限有主观性,而现存的很多统计学方法和数据挖掘技术也大多属于有监督学习,传统的聚类分析方法虽属无监督,但其对处理中医证候的问题有很多局限性。为此我们提出用基于模型的聚类方法,它既属于无监督学习,又有传统聚类分析所没有的优势。

参考文献

- [1] Estes WK. Classification and Cognition [M]. Oxford: Oxford University Press Inc 1994: 4.
- [2] Arabie P, Hubert LJ, Desoete G. Clustering and Classification [M]. Singapore: World Scientific Publishing Co Pte Ltd. 1996: 1-3.
- [3] 刘宗迪. 五行说考源 [J]. 哲学研究, 2004, (4): 35-41.
- [4] 刘晓琳. “五行”理论的哲学层次 [J]. 云梦学刊, 2005, 26(6): 56-59.
- [5] Boussiquet C, Luxburg U V, Ratsch G. Advanced Lectures on Machine Learning [M]. Germany: Springer Verlag Berlin Heidelberg, 2004: 73.

试论腰椎间盘突出症中医外治疗法中的 辨证论治和整体观

王力平 黄承军

柳州市中医院骨科, 广西壮族自治区柳州市城中区解放北路 32号, 545001)

基金项目: 国家中医药管理局中医药科学技术研究专项课题 (2009ZL37)

通讯作者: bone90@163.com, 13977299876

摘 要 从临床实践出发, 对腰椎间盘突出症的“椎间盘病因说”进行反思, 从证实质的角度分析“非特异性病变”与脊柱整体功能变化之间的关系, 认为外治法所辨之证是脊柱整体功能紊乱状态在某个外在局部的具体表现, 进而提出建立基于微视局部, 并与宏观整体相结合的脊柱整体辨证论治体系。

关键词 腰椎间盘突出症; 中医外治疗法; 整体观

腰椎间盘突出症是引起腰腿痛的最常见原因, 患者寻求医学治疗的目的是为了消除疼痛和恢复腰椎功能, 然而腰椎间盘突出症的病理机制相当复杂, 涉及多个致痛部位, 疼痛表现差异很大, 治疗方法也多种多样。中医学有关腰椎间盘突出症较早的记载见于《史记·扁鹊仓公列传》: “天雨, 黄氏诸倩见建家京下方石, 即弄之。建亦欲效之, 效之不能起, ……腰脊痛, 不可俯仰, 又不得小溲。”《素问·气交变大论》云: “岁火不及, 寒乃大行, 腰背相引而痛, 甚则屈

不能伸, 髀骀如别。”将本病病机归于肾虚。《医林绳墨》则明确指出: “大抵腰痛之证, 因于劳损而肾虚者甚多。”经历代医家的不断发展, 其证型分化渐变繁复, 有统计基本证型和复合证型达 30 种之多^[1]。然而对于腰椎间盘突出症最重要的治疗手段——外治疗法, 尤其是非药物疗法应如何辨证论治, 古今文献对此似未提及, 本文试浅论之。

1 辨证论治和整体观的现代诠释

中医临床辨证论治, 是对疾病过程中某一阶段

[6] Pysard S. M., Ivan B., Miroslav K. Machine Learning and Data Mining Methods and Application [M]. Chichester: John Wiley & Sons Ltd., 1998: 12—99.

[7] Smyth P. Clustering Sequences with Hidden Markov Models [M]. Cambridge MIT Press, 1997: 648—654.

[8] David J. Bartholomew, Steele F., et al. Analysis of Multivariate Social Science Data [M]. 2nd ed. Boca Raton: Taylor & Francis Group LLC, 2008: 271—284.

[9] Benmali H., Cleleax G., Adrian E., et al. Inference in Model based Cluster Analysis [J]. Statistics and Computing, 1997, (7): 1—10.

[10] Zhang NL, Yuan S, Chen T, et al. Statistical validation of traditional chinese medicine theories [J]. A Item Complement Med, 2008, 14 (5): 583—587.

[11] 王珏, 周志华, 周傲英. 机器学习及其应用 [M]. 北京: 清华大学出版社, 2006: 80.

[12] 袁世宏, 王米渠, 王天芳, 等. 聚类分析对肾虚症状的探索性研究 [J]. 北京中医药大学学报, 2006, 29 (4): 254—257.

[13] Banfield JD, Raftery AE. Model based Gaussian and non Gaussian Clustering [J]. Biometrics, 1993, (49):

803—821.

[14] Blines JA. A gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models [R]. Berkeley: International Computer Science Institute, 1998: 3—8.

[15] Rose K. Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems [J]. P. IEEE, 1998, 86 (11): 2210—2239.

[16] Vander Ark LA, Croon MA, Sijtsma K. New Developments in Categorical Data Analysis for the Social and Behavioral Sciences [M]. Mahwah: Lawrence Erlbaum Associates Inc, 2005: 41—75.

[17] Marina Meil A, David Heckerman. An Experimental Comparison of Model Based Clustering Methods [J]. Machine Learning, 2001, (42): 9—29.

[18] Vonev A, Clifford C. Latent Variables Analysis Applications for Developmental Research [M]. Thousand Oaks: SAGE Publications Inc, 1994: 245—272.

收稿日期: 2010—08—02 修回日期: 2010—11—23

编辑: 黄健