·理论研究·

隐结构模型与中医辨证研究(Ⅱ)*
——肾虚数据分析

张连文[1] 袁世宏[2] 陈 弢[1] 王 燚[1]
(1香港科技大学 香港; 2北京中医药大学)

关键词:隐结构模型;中医;辨证;肾虚
中图分类号:R241

我们提出一种为中医辨证建立客观定量标准的研究方法,即隐结构法。在文献[1]中已经介绍了隐结构法的基本步骤:首先对症状在人群中的分布情况做流行病学调查,然后利用隐结构模型对数据进行聚类分析,最后用得到的类指导辨证、建立起辨证标准。为了探讨隐结构法的可行性,我们用多层隐类(HLC)模型对肾虚辨证进行研究,收集和分析肾虚数据,并讨论所得隐变量及隐类的中医学意义。

1 数据收集及分析

在进行流行病学调查之前,需要确定调查所涉及的症状。依据国家证候诊疗标准[2]和中医诊断学教材[3],我们选择了67个肾虚症状变量,包括精神不振程度、腰部酸困程度、夜间尿频程度等。这些变量反映了肾脏藏精主水纳气、主生殖、主骨生髓、司二便、开窍于耳、其华在发、齿为骨之余等各方面的生理功能和病理变化。变量的程度分为无、轻、中、重4级。为保证可操作性、一致性,减少人为误差,调查过程采用了严氏等[4]所制定的肾虚症状轻重程度判别标准。这个标准的建立考虑到了症状表象轻重程度、伴随条件、持续时间、发作频率和诱发条件5个因素。抽样调查的样本空间定为肾虚证常常出现的60岁或以上人群。调查在成都、重庆、山西、内蒙古等地的敬老院、老年大学、福利院、工厂、农村进行,共收集数据2 600例[5]。

一个HLC模型的质量可以用贝叶斯信息标准评分(BIC)[6]来度量。BIC评分一方面要求模型与数据拟合、揭示数据中隐含的规律,另一方面要求模型尽量简单。用HLC模型分析数据就是要寻找BIC分数最高的最优HLC模型[7]。需要逐一考察每一个可能的模型、计算其BIC分、取分数最高者。但因为所有可能的模型数量巨大,计算机无法一一列举,于是需要搜索算法,故使用目前最好的启发式单重爬山法(HSHC)[8]分析肾虚数据。此算法暂时无法处理数据所涉及的全部67个变量,故选择了其中对肾虚辨证最重要的35个变量进行分析。分析过程在1台2.4GHz的奔腾Ⅳ计算机上耗时98.5 h完成,得到的模型记为M_0,它的BIC分数为-73947。爬山法有一个共同的缺点,就是有可能被困在局部最优而找不到全局最优,为了帮助HSHC算法跳出局部最优,我们利用背景知识对M_0进行多方面修改,得到一系列新模型,然后由此出发用HSHC算法进一步自动搜索,最后取BIC分最高的模型作为分析结果。这个模型记为M^*,它的BIC分数是-73860。数据分析过程中,人为因素的作用是,基于HSHC算法所得的中间结果,为它的进一步搜索提供出发点。整个搜索过程考察大量模型,取哪一个作为最后结果是由BIC评分来决定的。这意味着对模型所作的人为修改如果违反BIC原则就会被否决,所以人为因素不影响数据分析结果的客观性。

模型M^*的结构如图1所示,其中Y_0~Y_{34}是来自数据的症状显变量,而X_0~X_{13}是在数据分析过程中HSHC算法引入的隐变量。文献[1]中指出,用隐结构模型分析数据就是对数据进行多维同时聚类。通过对肾虚数据进行分析,得到包含14个隐变量的模型M^*,即找出了14个(相互关联的)隐聚类指标,按这些指标被数据聚成了各式各样的类。这些类称为隐类。例如,变量X_2有3个取值,即按隐指标X_0把数据聚成了5个类;变量X_1有5个取值,即按隐指标X_2同时又把数据聚成另外3个类等等。

张连文,男,副教授

* 香港研究资助局项目(No 622015),国家重点基础研究发展计划(973计划)项目(No 2003CB517101)

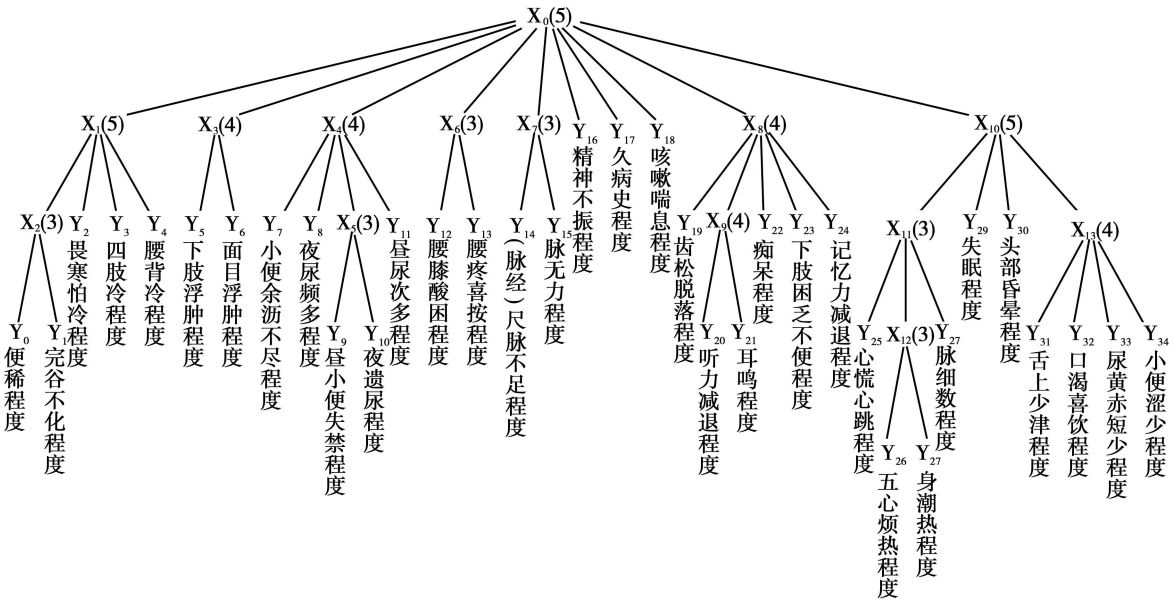

图 1　模型 M^* 结构图

2　隐变量诠释

模型 M^* 结构图中的隐变量 X_1,它直接影响畏寒怕冷、四肢冷、腰背冷这 3个症状的出现和轻重程度,同时又通过隐变量 X_2 间接影响便稀、完谷不化的出现和轻重程度。这与中医理论是否吻合? 隐变量 X_1 和 X_2 的意义是什么? 中医理论认为,肾藏元阳内寓命火,既可温煦脏腑经络、推动激发脏腑功能活动,又能温肢体暖周身。当肾阳不足、失于温煦时则可见畏寒肢冷、腰背怕冷。同时肾阳虚衰火不温脾土,则可见完谷不化、大便溏薄。仔细分析对照模型和中医理论中以上两段话,可以发现,它们讲的完全是同一件事,即:有一个隐因子,它直接影响畏寒怕冷、四肢冷、腰背冷这 3个症状;同时它通过另一隐因子间接影响便稀、完谷不化这 2个症状。所以,这个 M^* 模型局部与中医理论基本吻合。中医理论中与这 2个隐因子对应的分别为肾阳虚失于温煦和火不温脾土,即模型 M^* 中的隐变量 X_1 可诠释为肾阳虚失温煦情况,X_2 可诠释为火不温脾土情况。

模型 M^* 中的隐变量 X_3,它影响下肢浮肿、面目浮肿的出现及其轻重情况。而中医理论认为,肾的阳气亏虚,气化失司,水液失于蒸化则内停泛滥,溢于上则面目浮肿,趋于下而见下肢浮肿。所以 M^* 模型这一局部与中医理论基本吻合,X_3 可诠释为肾阳虚水泛情况。

在 X_4 及其下这个局部,隐变量 X_4 直接影响小便余沥、夜尿频多、昼尿次多的出现和轻重程度,同时又通过隐变量 X_5 间接影响昼小便失禁和夜遗尿的出现及轻重程度。中医理论认为膀胱为洲渎之官,它的气化约束作用控制着小便开合排泄。当膀胱气化失司,约束失职关门不利,则可见小便余沥不尽、夜尿频多、昼尿次多。气化约束失职较重,关门不能,则可使尿失禁锢而见夜遗尿、昼小便失禁。因此就这一局部而言,模型 M^* 与中医理论基本吻合,其中 X_4 可诠释为膀胱失约情况,而 X_5 可诠释为尿失禁锢情况。

中医认为肾精不足的临床表现有:耳鸣耳聋、健忘恍惚、两足痿软、发脱齿摇、神情呆钝等,这些与 X_8 以下的症状变量基本一致,所以 X_8 可诠释为肾精亏虚情况。中医认为(肾)阴虚的临床表现有:失眠、眩晕、口燥咽干、潮热颧红、五心烦热、盗汗、小便短赤、舌红少津少苔、脉细数等,这些与 X_{10}以下的症状变量基本一致,所以 X_{10}可诠释为(肾)阴虚情况。中医认为久病可致肾虚,而肾虚的典型症状有腰膝酸软、精神不振、脉弱,其子证型包括肾阳虚、肾阳虚水泛、肾阴虚、肾精亏虚、膀胱失约等,这些与 X_0 周围的情况基本一致,所以 X_0 可诠释为肾虚情况。

数据分析所涉及的 35个症状变量全部与肾虚有关,因此在模型 M^* 中有一个隐变量与肾虚对应是意料之中的。然而模型 M^* 中有隐变量与肾阳虚失温煦、肾阳虚水泛、膀胱失约、肾精亏虚、(肾)阴虚等证候因子对应这件事却十分有意义。首先,根据中医理论肾阳虚失温煦等证候因子只与部分症状变量直接相关,而不是与全部症状变量直接相关,并

且不同证候因子影响不同症状变量。其次，模型 M^* 中的隐变量是按照统计学原则从数据里面提取出来的，具有客观性。因此，我们的数据分析显示了肾阳虚失温煦等证候因子的客观性。

模型 M^* 与中医理论也有不一致的地方。首先，肾虚的其他子证候如肾气不固、肾不纳气等没有在 M^* 中反映出来。肾气不固没有在 M^* 中反映出来是因为没有收集到生殖、性生活方面的有关数据；而肾不纳气没有在 M^* 中反映出来是由于除了咳嗽喘息以外，肾不纳气的其他症状，如呼吸表浅、呼多吸少、咳嗽无力等，被排除在分析范围以外。另外，在中医理论中，阴虚和精亏都可以导致失眠和头晕，然而在模型 M^* 中失眠和头晕却只与 X_{10}（阴虚）相连。这样的差异源自 HLC模型的一个限制：HLC模型要求隐、显变量之间的关系成树状结构[7]，从而一个显变量只能与一个隐变量相连。放松这个限制是隐结构模型研究下一步要解决的主要问题之一。

中医理论有不完整和模糊的地方，不能完全转化为数学或逻辑命题，因此任何一个数学模型都不可能与之百分之百吻合。中医理论是从几千年不断临床实践中提炼出来的，我们的目标就是要通过现代数据分析，重复这个提炼过程，使中医理论得到发展。上述结果表明，这个目标是可以实现的。

3 隐类诠释

模型 M^* 中隐变量 X_1 被诠释为肾阳虚失于温煦情况，它有 5个可能的取值，即按照肾阳虚失于温煦情况这个隐指标把数据聚成了 5个隐类；隐变量 X_4 被诠释为膀胱失约情况，有 4个可能的取值，即按照膀胱失约情况，把数据聚成了另外 4个隐类等。下面以 X_4 给出的 4个隐类为例，讨论这些隐类的中医学意义。

在模型 M^* 中位于 X_4 下面的症状变量有小便余沥（Y_7）、夜尿频（Y_8）、昼小便失禁（Y_9）、夜遗尿（Y_{10}）和昼尿次多（Y_{11}）。图 2展示这些显变量在 X_4 给出的 4个隐类中的概率分布：共有 4幅直方图，对应 4个隐类；每幅直方图中有 5个柱体，对应 Y_7 等 5个症状变量；一个柱体最多由 4段组成，每一段对应症状变量的一个取值：黑色段的长度代表症状变量取值为“重”的概率，白色段（位于顶端）的长度代表症状变量取值为“无”的概率，而 2个灰色段的长度分别代表症状变量取值为“轻”和“中”的概率。概率直方图展示了隐类的整体性质。另一方面，典型成员表（见表 1）给出各隐类中最典型成员的情况，让我们对隐类的成员有一个具体认识。类 $X_4=s_0$ 的典型成员是‘$Y_7=3$、$Y_8=3$、$Y_9=3$、$Y_{10}=3$、$Y_{11}=3$’，这是因为是它属于类 $X_4=s_0$ 的概率 $P(X_4=s_0 \mid Y_7=3、Y_8=3、Y_9=3、Y_{10}=3、Y_{11}=3)$ 在所有可能成员中是最高的。

表 1 隐变量 X_4 各隐类典型成员比较

隐类	Y_7	Y_8	Y_9	Y_{10}	Y_{11}
s_0	0	0	0	0	0
s_1	1	1	0	0	1
s_2	0	3	0	0	0
s_3	3	3	3	3	3

通过考察图 2可以把握 X_4 给出的 4个隐类的含义。首先，隐变量 X_4 的意义是膀胱失约情况，因此它所给出的 4个类可视为膀胱失约的 4种不同状态。在类 $X_4=s_0$ 中，小便余沥等 5个症状出现的概率很低，其典型成员完全没有这些症状，所以它可以诠释为无膀胱失约。其次，在类 $X_4=s_3$ 中，小便余沥等 5个症状出现的概率很高，它的典型成员 5个症状都有，而且是重度，所以它可以诠释为重度膀胱失约。接下来考虑 $X_4=s_1$ 和 $X_4=s_2$ 这 2个类。在这 2个类中小便余沥等 5个症状出现的概率介乎于类 $X_4=s_0$ 和 $X_4=s_3$ 之间。而就它们两者之间比较而言，各症状在 $X_4=s_1$ 中出现的总概率虽然高于在

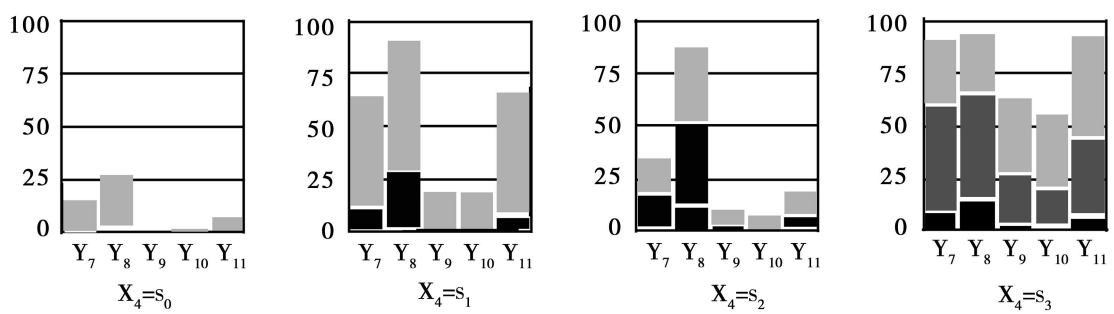

图 2 各症状变量在隐变量 X_4 各隐类中的概率比较

表 2　隐变量中医证候意义诠释分析列表

隐变量	隐变量诠释	s_0隐类	s_1隐类	s_2隐类	s_3隐类	s_4隐类
X_1	肾阳虚失温煦情况	无	轻 1	轻 2	中	重
X_3	肾阳虚水泛情况	无	轻 1	轻 2	重	—
X_4	膀胱失约情况	无	轻 1	轻 2	重	—
X_8	肾精亏虚情况	无	轻 1	轻 2	重	—
X_{10}	(肾)阴虚情况	无	轻	中 1	中 2	重

$X_4=s_2$中出现的总概率,但各症状在 $X_4=s_1$中以中度或重度出现的概率却低于在 $X_4=s_2$中以中度或重度出现的概率。类 $X_4=s_1$的典型成员有 3个轻度症状,即轻度小便余沥、轻度夜尿频、和轻度昼尿次多;而类 $X_4=s_2$的典型成员有 1个重度症状,即重度夜尿频。综合这些情况来看,类 $X_4=s_1$和 $X_4=s_2$都有轻度膀胱失约的意义。据此,我们把他们分别诠释成轻度膀胱失约(1)和轻度膀胱失约(2)。

除 X_4以外,我们还对 X_1、X_3、X_8和 X_{10}所给出的隐类的意义进行了分析[3],发现它们可按表 2诠释。之所以只考虑上述 5个隐变量,是因为它们反映了肾虚证几个主要方面。

中医理论中有一些关于症状出现频率的定性论断。比方说,中医认为若肾气亏虚,膀胱气化失司,失去约束固摄功能,轻则夜尿频多、小便余沥不尽,重则昼小便失禁、夜遗尿。换句话说,小便失禁和夜遗尿出现的频率低于小便余沥等其他 3个症状出现的频率。这与图 2不谋而合。只有在 $X_4=s_3$(重)时,小便失禁(Y_9)和夜遗尿(Y_{10})才有较大概率会出现;即使此时,它们以重度出现的概率还是很小。另外,中医认为夜尿频多(Y_8)是肾虚的典型症状,这与它在 $X_4=s_1$、$X_4=s_2$和 $X_4=s_3$这 3个类中出现的概率都很大的事实相吻合。

中医理论中关于症状出现频率的定性论断有很多,其中与模型 M^*相关的都与 M^*吻合[3]。我们的数据分析为这些论断的正确性提供了客观的统计学支持,并进一步把它们进行了定量化。

4　隐结构法与辨证论治

清代医家叶天士在《临证指南医案》中说:“医道在乎识证、立法、用方,此为三大关键,一有草率,不堪为医也。…… 然三者之中,识证尤为紧要。”隐结构法通过对数据进行多维聚类分析,能够深化我们对证候的认识,帮助我们更好地识证、立方,从而提高辨证论治水平。

用隐结构模型分析数据可以找出隐变量,并且以它们为指标把数据聚为多个隐类。研究结果表明,使用隐结构法能够获得有明显证候意义的隐类,而对这些隐类的研究能对中医证候学的发展起推动作用。具体地说,我们可以:①对每个证候隐类进行研究,根据其症状的统计特性找到针对它们的治疗方案(基本证候论治研究);②在临床辨证论治时,计算出病人属于各隐类的后验概率(识证);③基于这些概率将基本证候的治疗方案进行适当组合,得到对病人的治疗方案(立法、用方)。这样,辨证论治的客观性、定量性、规范性都可以得到提高。其中模型辨证内容,我们会在后文继续讨论,其他是需要将来进一步研究的课题。

参考文献:

[1] 张连文,袁世宏.隐结构模型与中医辨证研究(I)——隐结构法的基本思想以及隐结构分析工具[J].北京中医药大学学报, 2006, 29(6): 365—369.

[2] 国家技术监督局.中华人民共和国国家标准 GB/T16751.2—1997:中医临床诊疗术语证候部分[S].北京:中国标准出版社, 1997.

[3] 朱文锋.中医诊断学[M].上海:上海科学技术出版社, 1995: 160—161.

[4] 严石林,张连文,王米渠,等.肾虚证辨证因子等级评判操作标准的研究[J].成都中医药大学学报, 2001, 24(1): 56—59.

[5] 张连文,袁世宏.隐结构模型与中医辨证[EB/OL].香港科技大学计算机科学及工程系技术报告 HKUST-CS04-12. http://www.cse.ust.hk/~lzhang/tcm/.

[6] SCHWARZ G. Estimating the dimension of a model[J]. Annals of Statistics, 1978, 6(2): 461—464.

[7] ZHANG N L. Hierarchical latent class models for cluster analysis[J]. Journal of Machine Learning Research, 2004, 5(6): 697—723.

[8] ZHANG N L, KOCKA T. Efficient Learning of Hierarchical Latent Class Models[C]. Proc. of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-2004), Florida, 15—17.

(收稿日期: 2008-03-06)