约束隐树分析：一种将中医辨证标准定量化的无监督学习方法

张连文[1]，许玉龙[2*]，陈周荣[1]，吕雅丽[2]，陈锦华[3*]

（1. 香港科技大学计算机科学与工程学系 香港；2. 河南中医药大学信息技术学院 河南郑州 450046;

3. 香港大学内科学系 香港）

摘要：中医辨证标准常以主-次症形式定性给出，有必要进行定量化，以便提高其临床指导和参考作用。定量化过程可以使用有监督学习方法，也可以使用无监督学习方法，前者要求在收集数据时对个案所属证型进行判断，因此受主观因素影响较大，而后者没有这样的要求，因此可以减少主观因素的影响。现有无监督学习方法包括隐类分析和隐树分析，这些方法未考虑主证、次症的语义（主症是对证候辨识最重要甚至必要的症状，而次症是对证候辨识有帮助但并非必要的症状），从而难以得到满足这些语义的定量规则。本文提出将主证、次症的语义作为约束条件加入隐类分析和隐树分析，开发出一种将辨证标准定量化的无监督学习新方法，这种方法既减少主观因素对分类规则定量化过程的影响，又能得到反映主、次症特点的规则。

1. 引言

中医辨证标准有国家、学会颁发的[1,2,3]，也有科研团队制定的[4]，一种常见的形式是定性地给出证候对应的主症和次症[5,6]。例如，下面是香港大学一个研究二型糖尿病的课题组根据德尔菲法专家共识确定的胃肠湿热证辨证标准：

主症: 口臭，大便秘结或便溏不爽，舌苔黄，舌苔腻，舌色红，脉滑；

次症: 胃脘痞胀，腹痞胀。

本文探讨如何将这样主-次症形式的辨证规则定量化，以便提高其在临床和科研中的指导和参考作用。确定证候的主症和次症也是一个重要问题，但不是本文的关注点。

* 文通信作者：许玉龙 （flyxyl@126.com）, 陈锦华(chriskwc@hku.hk). 本项研究得到香港研究资助局项目（No: 16202118）、医疗卫生研究基金(No: 14151731)、国家自然科学基金项目(No: 81703946、81973791、61902113)的资助。

辨证是基于症状（包括体征）对患者进行分类，诸如逻辑回归、支持向量机的有监督学习方法经常被用来建立定量分类规则[7]。这些方法基于有标签数据，要求在收集数据时对个案所属证型进行判断，因此受主观因素影响较大。为减少主观因素的影响，本文探索无监督学习方法，这些方法不要求在收集数据时对个案所属证型进行判断。

隐类分析 (Latent Class Analysis) 是一种基于无标签数据对患者进行分类的方法，在西医研究中已经被广泛运用[8]。关于隐类分析的基本原理，文献[9]说："临床诊断方法，往往是由有经验的临床专家通过观察许多不同患者，总结出关键症状和体征出现的规律，而慢慢开始形成的。隐类分析以相对严谨的统计学方式模拟这个过程"。隐类分析有一个不切实际的假设条件，叫局部独立假设(local independence assumption) [8]，其内容是：在一个给定的证候类中，不同症状之间相互独立，即是说不同症状之间相互关联的唯一原因是该证候的出现。为了放宽这个假设，使得模型与数据更好地拟合，文献[10, 11]对隐类分析进行改进和推广，提出了隐树分析 (Latent Tree Analysis)。

隐类分析和隐树分析都是纯粹根据症状分布规律对患者进行分类，其结果往往不能反映主症、次症的特点。中医对主症、次症的认识不仅仅来源于症状出现的规律，还包括对这些规律的理论提升，以及在临床实践中得到的反馈。因此，有必要将对主症、次症的定性认识与症状分布数据结合起来。本文探讨如何将主症、次症的语义作为约束条件加入隐类分析和隐树分析，使主症和次症的差别在辨证规则中得到体现。在本文之前，已有学者提出约束隐类分析(Constrained Latent Class Analysis)[12]的思想，其目的是保证患者类的特性具有单调性，即疾病程度越严重，症状出现概率越高。本文提出的约束条件是保证主症的作用大于次症，其数学形式和模型优化方法与以往的约束条件均不相同。

2. 主症-次症的语义与约束条件

中医诊断学及不同机构颁发的临床指引[1,2,13-17]对主症、次症给出了一致的定义：**主症**是证候辨识最核心的症状，是证候本质的表现，对定性证候具有必要性，并且对证候的辨识起主要作用；而**次症**是证候常见的伴随症状，对证候辨识有帮助，但出现机会差异较大，其必要性不如主症，具有从属关系。下面讨论如何将上述定义转化为约束条件，在数据分析过程中使用，以保证主症的重要性以及它与次症的区别得到恰当体现。

用 Z 记一个证候变量，它有两个取值，$Z=1$ 代表证候出现，$Z=0$代表证候不出现。用 x 记一个症状，它有两个取值，$x=1$ 代表症状出现，$x=0$ 代表症状不出现。$P(x=$

$1|Z=1)$ 和$P(x=1|Z=0)$ 分别是症状 x 在有证候 Z 和无证候 Z 时出现的概率，分别简记为 p 和 q。

症状 x 在无证候 Z 时出现的概率 $q=P(x=1|Z=0)$ 有可能是 0，也可能不是 0。如果第二种情况发生，就意味着除证候 Z 以外，还有其它因素能导致症状 x 出现，而$q=P(x=1|Z=0)$ 正是其它因素导致症状 x 出现的概率。

假设这些其它因素独立于 Z，它们的存在和作用与证候Z 是否出现无关[†]。它们不但在证候Z不出现时起作用，在证候 Z 出现时也起作用。这样，症状 x 在有证候 Z 时出现的概率 $p=P(x=1|Z=1)$ 不但反映证候 Z对症状 x 的影响，还反映其它因素对症状 x 的影响。

用 p_z 记症状 x 纯粹因为证候 Z 而出现的概率。在证候 Z 出现时，有 p_z 概率导致症状 x 出现，也有 $1-p_z$ 概率不导致症状 x 出现。在第二种情况下，其它因素可能导致症状 x 出现，其概率为 q。 因此，症状 x 出现的总概率是 $p=p_z+(1-p_z)q$ 。从而：

$$p_z=\frac{p-q}{1-q}=\frac{P(x=1|Z=1)-P(x=1|Z=0)}{1-P(x=1|Z=0)}。$$

本文把“症状 x 是证候 Z的主症”诠释为证候 Z 对症状 x 的影响大，即症状 x 纯粹因为证候 Z 而出现的概率 p_z 大。于是，引入如下**主症约束条件**：

$$\delta_1 \le \frac{P(x=1|Z=1)-P(x=1|Z=0)}{1-P(x=1|Z=0)}.$$

其中，参数 δ_1是主症纯粹因为证候而出现的概率的下界，可以理解为证候对主症的**最低影响程度**。

另一方面，我们**把**“症状 x 是证候 Z的次症”诠释为证候 Z 对症状 x 的影响相对小，即症状 x 纯粹因为证候 Z 而出现的概率 p_z 相对小。于是，引入如下**次症约束条件**：

$$0 \le \frac{P(x=1|Z=1)-P(x=1|Z=0)}{1-P(x=1|Z=0)} \le \delta_2.$$

其中参数 δ_2是次症纯粹因为证候而出现的概率的上界，可以理解为证候对次症的**最高影响程度**。另外注意，不等式的左边部分意味着 $P(x=1|Z=1)\ge P(x=1|Z=0)$ ，表示次症对证候的影响不是负面的。

主症与次症是相对而言的，于是要求 $\delta_2 \le \delta_1$。至于约束参数δ_1和 δ_2 的具体取值如何确定，将在后文中第 4 节讨论。

[†] 这称为因果机制独立(causal independence) 假设[18：第 2.4.3 节].

3. 约束隐类分析

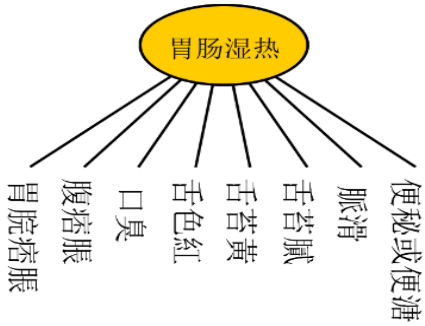

图 1. 一个关于二型糖尿病中胃肠湿热证的隐类模型

在讨论约束隐类分析之前，先简单介绍一下隐类分析。图 1 给出一个例子，它由一个证候隐变量和多个症状显变量组成。一般而言，用Z 记一个隐类模型中的证候变量，用 $x_1, \ldots, x_n$ 记其中的症状变量。型参数包括概率分布 $P(Z)$ 和 $P(x_j|Z)$ $(j = 1, \ldots, n)$。把所有模型参数组成的向量记为 θ，其所有可能取值的集合记为 Θ，称为参数空间。 给定一组症状数据$D = \{D_1, \ldots, D_m\}$，模型参数的对数似然函数是：

$$l(\theta|D) = \log P(D|\theta) = \sum_{i=1}^{m} \log P(D_i|\theta).$$

隐类分析的目的是计算参数的最大似然估计：

$$\theta^* = \arg\max_{\theta \in \Theta} l(\theta|D).$$

由于有隐变量，最大似然估计一般用 EM 算法[18：第 7.7 节]来求解。

约束隐类分析 [12]就是对隐类模型的参数加上约束条件，并在约束条件下进行参数估计。关于症状变量 x_j 的参数有两个，$\theta_{j11} = P(x_j = 1|z = 1)$ 和 $\theta_{j10} = P(x_j = 1|z = 0)$。如果 x_j 是主症， 这两个参数需要满足主症约束条件；如果 x_j 是次症， 这两个参数需要满足次症约束条件。这些约束条件在参数空间中定义了一个子空间，记为 $\widehat{\Theta}$。约束隐类分析的目的是计算在约束参数子空间范围内参数的最大似然估计：

$$\hat{\theta}^* = \arg\max_{\theta \in \widehat{\Theta}} l(\theta|D).$$

为了估计 $\hat{\theta}^*$，本文采用块坐标梯度上升（block coordinate gradient ascent）算法[19]，它从 θ 的随机初始值出发进行迭代，每次迭代逐个考虑症状变量 x_j，更新相关参数 $\theta_{j11} =$

$P(x_j = 1|z = 1)$ 和 $\theta_{j10} = P\left(x_j = 1\middle|z = 0\right)$，最后更新参数$\theta_{z1} = P(z = 1)$。 在迭代过程中，对数似然函数$l(\theta|D)$会逐渐上升，当它停止上升或迭代次数超过一个预先设定的阈值，迭代停止。

参数的初始值采用拒绝采样 (rejection sampling)方法随机选择，即重复从均匀分布中对参数 θ_{j11}、 θ_{j10} $(j = 1, \dots, n)$和 θ_{z1} 进行采样，直到满足所有约束条件为止。

参数的取值在迭代过程中不断得到更新和改进。对于每一个症状变量 x_j，为了更新概率分布 $P\left(x_j\middle|Z\right)$ 参数 θ_{j11}和 θ_{j10}，首先计算梯度向量 $(g_{j1}, g_{j0}) = (\frac{\partial l}{\partial \theta_{j11}}, \frac{\partial l}{\partial \theta_{j10}})$，其中的偏导用下列公式来计算：

$$\frac{\partial l}{\partial \theta_{j1b}} = \sum_{i=1}^{m} \frac{\partial \log P(D_i|\theta)}{\partial \theta_{j1b}} ,$$

$$\frac{\partial \log P(D_i|\theta)}{\partial \theta_{j1b}} = \begin{cases} \dfrac{P(Z = b|D_i, \theta)}{\theta_{j1b}} & \text{如果在} D_i \text{中 } x_j = 1 \\ -\dfrac{P(Z = b|D_i, \theta)}{1 - \theta_{j1b}} & \text{如果在} D_i \text{中 } x_j = 0 \end{cases}$$

其中，$b = 1$ *或* 0 ‡。参数 θ_{j11}和 θ_{j10}用如下公式更新：

$$\theta_{j11} \leftarrow \theta_{j11} + \lambda_j \alpha\, g_{j1};\ \theta_{j10} \leftarrow \theta_{j10} + \lambda_j \alpha\, g_{j0},$$

这里，α 是整体学习率，λ_j 是 $[0,1]$区间中使得各种约束条件得到满足的最大值。约束条件包括 $0 \le \theta_{j11} \le 1, 0 \le \theta_{j10} \le 1$。如果 x_j 是主症，还包括主症约束条件; 如果 x_j 是次症，还包括次症约束条件。为了确定 λ_j， 先令$\lambda_j = 1$，然后把它逐渐降低，直到所有相关约束满足为止。注意，当 $\lambda_j = 0$ 时所有相关约束是满足的，所以确定 λ_j的过程一定会成功。

至于症状变量概率分布中 $P(Z)$ 的参数 θ_{z1}，它的更新公式是：

$$\theta_{z1} \leftarrow \theta_{z1} + \lambda_z \alpha\, g_{z1},$$

其中λ_z 是 $[0,1]$区间中使得条件 $0 \le \theta_{z1} \le 1$得到满足的最大值，而

---

‡ 公式证明的基本思路：如果在数据 D_i 中$x_j = 1$，$\frac{\partial \log P(D_i|\theta)}{\partial \theta_{j1b}} = \frac{1}{P(D_i|\theta)} \frac{\partial P(D_i|\theta)}{\partial \theta_{j1b}} = \frac{1}{P(D_i|\theta)} \frac{P(D_i, Z=b|\theta)}{\theta_{j1b}} = \frac{P(Z=b|D_i,\theta)}{\theta_{j1b}}$。 如果在数据 D_i 中$x_j = 0$, 设 $\theta_{j0b} = 1 - \theta_{j1b}$。 有 $\frac{\partial \log P(D_i|\theta)}{\partial \theta_{j1b}} = -\frac{\partial \log P(D_i|\theta)}{\partial \theta_{j0b}} = -\frac{P(Z = b|D_i, \theta)}{\theta_{j0b}} = -\frac{P(Z = b|D_i, \theta)}{1 - \theta_{j1b}}$。稍后在考虑 $\frac{\partial \log P(D_i|\theta)}{\partial \theta_{z1}}$ 时， 需要注意$P(D_i|\theta) = P(z = 0, D_i|\theta) + P(z = 1, D_i|\theta)$，从而偏导由两项组成。 本文其它公式的证明类似。

$$g_{z1}=\frac{\partial l}{\partial\,\theta_{z1}}=\sum_{i=1}^{m}\frac{\partial\,\log P(D_i|\theta)}{\partial\,\theta_{z1}}\ ,$$

$$\frac{\partial\,\log P(D_i|\theta)}{\partial\,\theta_{z1}}=\frac{P(Z=1|D_i,\theta)}{\theta_{z1}}-\frac{P(Z=0|D_i,\theta)}{1-\theta_{z1}}\ 。$$

为避免陷入局部最优解，算法从多个初始值出发，取结果中似然度最大者作为最终结果。这样，共有三个算法参数：学习率、迭代次数和初始值个数。在分析具体数据时，需要运行算法多次，若发现结果不同，需要酌情改变算法参数。一般而言，初始值个数越多，结果越稳定。降低学习率也有利于结果稳定，但迭代次数需要相应提高。初始值个数和迭代次数的增加会相应增加运算时间。

4. 约束参数δ_1和δ_2的确定

约束隐类分析是在约束条件下估计隐类模型的参数值。在模型参数确定后，可以使用文献[11]所述的方法获得定量辨证规则。这个过程受约束参数δ_1、δ_2的影响，约束参数的取值不同，隐类模型的参数以及最终的定量辨证规则也不同。

以第 1 节提及的关于二型糖尿病的研究为例，数据来源于香港大学一个研究二型糖尿病的课题组[20]。针对胃肠湿热证，使用不同的δ_1、δ_2取值，对数据进行约束隐类分析，得到如表 1 所示的结果。其中，模型的 BIC 评分是对模型与数据拟合程度的一种度量。另外，辨证规则是用打分形式给出的，阈值做了规范化，都是 10。

直观地讲，约束条件越强，模型对数据的拟合程度就越低，BIC 评分就会下降。这一点在表 1 得到充分反映。在固定δ_2情况下，δ_1越大，约束越强，BIC 评分相应相应下降；在固定δ_1的情况下，δ_2越大，约束越弱，模型的 BIC 评分相应上升。

表 1. 使用不同约束参数值对二型糖尿病中胃肠湿热证进行约束隐类分析的结果

δ_1	δ_2	模型的 BIC 分	证候类大小	辨证规则
0.4	0.1	-5538.37	0.19	主症：口臭(1.6)，大便秘结或便溏不爽(1.5)，舌苔黄(4.5)，舌苔腻(3.0)，舌色红(1.4)，脉滑(1.4)。 次症：胃脘痞胀(0.8)，腹痞胀(0.6)。
0.4	0.2	-5536.83	0.17	主症：口臭(1.4)，大便秘结或便溏不爽(1.4)，舌苔黄(4.7)，舌苔腻(3.3)，舌色红(1.2)，脉滑(1.2)。 次症：胃脘痞胀(0.6)，腹痞胀(0.6)。
0.5	0.1	-5550.76	0.11	主症：口臭(1.6)，大便秘结或便溏不爽(1.4)，舌苔黄(3.7)，舌苔腻(2.7)，舌色红(1.2)，脉滑(1.4)。 次症：胃脘痞胀(0.4)，腹痞胀(0.5)。

0.5	0.2	-5547.91	0.12	主症：口臭(1.9)，大便秘结或便溏不爽(1.7)，舌苔黄(2.9)，舌苔腻(2.7)，舌色红(1.5)，脉滑(1.5)。 次症：胃脘痞胀(0.8)，腹痞胀(1.1)。
0.5	0.3	-5547.19	0.14	主症：口臭(2.0)，大便秘结或便溏不爽(1.9)，舌苔黄(2.8)，舌苔腻(2.2)，舌色红(1.6)，脉滑(1.6)。 次症：胃脘痞胀(1.4)，腹痞胀(1.6)。
0.6	0.1	-5561.09	0.08	主症：口臭(2.5)，大便秘结或便溏不爽(1.8)，舌苔黄(2.2)，舌苔腻(2.1)，舌色红(1.5)，脉滑(1.6)。 次症：胃脘痞胀(0.4)，腹痞胀(0.5)。
0.6	0.2	-5555.17	0.08	主症：口臭(2.3)，大便秘结或便溏不爽(1.6)，舌苔黄(2.5)，舌苔腻(2.0)，舌色红(1.5)，脉滑(1.5)。 次症：胃脘痞胀(0.7)，腹痞胀(0.9)。
0.6	0.3	-5553.98	0.07	主症：口臭(2.7)，大便秘结或便溏不爽(1.5)，舌苔黄(2.4)，舌苔腻(1.8)，舌色红(1.3)，脉滑(1.3)。 次症：胃脘痞胀(0.9)，腹痞胀(1.2)。
0.6	0.4	-5552.83	0.09	主症：口臭(2.1)，大便秘结或便溏不爽(2.0)，舌苔黄(2.0)，舌苔腻(1.8)，舌色红(1.7)，脉滑(1.7)。 次症：胃脘痞胀(1.4)，腹痞胀(1.6)。

本文的核心思想是通过整合两种不同来源的信息，确定证候的定量辨证规则。一种信息来源是“流调”数据，在统计学意义下是定量的；另一种是中医对主症、次症的认识，是定性的。为了整合这两种信息，需要将主症、次症的语义定量化。为此，先使用约束参数δ_1和δ_2的不同取值进行约束隐类分析，然后通过考察不同取值的结果，用德尔菲法来确定约束参数的取值。定性辨证标准本身是根据德菲尔法专家共识厘定的，约束隐类分析要求再次使用德尔菲法确定δ_1和δ_2的取值。

就二型糖尿病中的胃肠湿热证而言，基于表 1 中的结果，把δ_1和δ_2的值分别定为 0.5 和 0.1。做出这个选择的理由如下：当δ_1值为 0.4 时，舌苔黄及舌苔腻得分过高，过于支配辨证。当δ_1值高于 0.5 时，辨证需要所有主症同时出现，与辨证理念不吻合。在以往的流行病学调查中，胃肠湿热的流行率约为 11.1%[20]，这与δ_1等于 0.5 时的证候类大小基本吻合。基于上述原因，把δ_1的值定为 0.5。而当δ_1值为 0.5 时，如果δ_2取值为 0.2 或 0.3 时，次症得分接近主症，未能体现两者的区别；当δ_2值为 0.1 时，主次症关系以及证候类大小都与实际情况最吻合， 因此把δ_2的取值定为 0.1。

5. 隐类分析与约束隐类分析的比较

隐类分析和约束隐类分析的出发点都是无标签症状数据，其结果也都是隐类模型，所不同的是约束隐类分析加入了先验约束条件。相比之下，隐类分析得到的模型与数据的拟合程度更高，而约束隐类分析得到的模型则更能反映主症、次症的特点。

以二型糖尿病中的胃肠湿热证为例，隐类分析得到模型的 BIC 分是 -5490。它把患者分为两类，其中有肠湿热证患者占 40%，相应的辨证规则如下：

主症：口臭(0.3)，大便秘结或便溏不爽(-0.1)，舌苔黄(18.5)，舌苔腻(1.2)，舌色红(0.5)，脉滑(0.0)；

次症：胃脘痞胀(0.0)，腹痞胀(-0.1)。

这个辨证规则明显不合理。同是主症，舌苔黄的分值很高，超过阈值（10），但舌色紅、口臭、脉滑、大便秘结或便溏不爽的分值却很低，甚至为 0 和负数。次症胃脘痞胀、腹痞胀的分值也分别为 0 和负数，表示它们对证候没有影响和有负面影响。这些违背了主症和次症的特点。

约束隐类分析得到的模型的 BIC 分是-5550，低于隐类分析得到的模型，表示它与数据的拟合程度相对较低。约束隐类分析也把患者分为两类，其中有肠湿热证患者占 11%，相应的辨证规则如下:

主症：口臭(1.6)，大便秘结或便溏不爽(1.4)，舌苔黄(3.7)，舌苔腻(2.7)，舌色红(1.2)，脉滑(1.4)；

次症：胃脘痞胀(0.4)，腹痞胀(0.5)。

这个规则较好地反映了主症、次症的特点，与隐类分析得到的规则相比较，显得更为合理。

为什么隐类分析得不到理想的结果而加上约束后情况有所改善呢？这个问题值得深入探讨。初步回答是：症状的出现有多种原因，隐类分析试图仅仅基于症状出现情况，确定症状与其中一种原因的定量关系，这无疑是十分困难的。约束隐类分析在分析过程中加入关于目标原因（证候）与症状关系的定性认识，并且允许通过考察结果来确定约束参数的取值，这样使得分析过程“有的放矢”，从而能够找到症状与目标原因的定量关系。

6. 约束隐树分析

在如图 1 所示的隐类模型中，所有症状都与证候直接相连，这意味着给定证候隐变量，症状变量之间相互独立，或者说在每个隐类中，症状变量之间相互独立，因此称为局部独

立假设。在实际运用中，这个假设往往不成立，从而导致估计偏差[8]。为了放宽局部独立假设，使得模型能与数据更好地拟合，文献[21,11]提出使用如图 2 所示的三层模型，在症状变量与证候隐变量之间引入一些中间隐变量。这样的模型是一种特殊的隐树模型[10]，中间层的隐变量代表证候的不同侧面，因此也称为综合聚类模型[21,22]。一般的隐树模型不一定是三层的，隐类模型也是隐树模型的一个特例，所以隐树模型是隐类模型的推广。

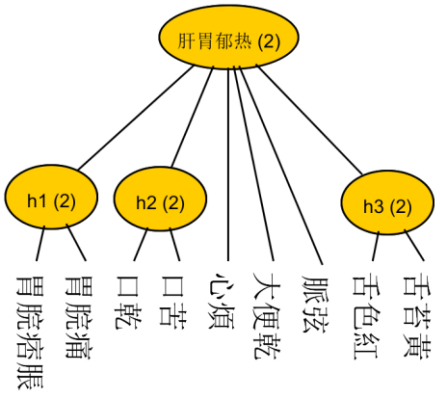

图 2. 一个关于二型糖尿病中肝胃郁热的隐树模型

下面讨论如何在约束条件下估计三层聚树模型的参数。如果一个症状变量 x_j 与证候变量 Z 直接相连，相关模型参数的初始化和更新用第 3 节所描述的方法。证候变量 Z 的概率分布 $P(Z)$ 初始化和更新也一样。其它症状变量按中间层隐变量分组，每组所涉及的所有参数作为一个整体处理。

设 h 是一个中间层隐变量，与其相连的症状变量是 $x_1, \ldots, x_r$，相关的模型参数有 $\theta_{h11} = P(h = 1|Z = 1)$，$\theta_{h10} = P(h = 1|Z = 0)$，$\theta_{l11} = P(x_l = 1|h = 1)$ 和 $\theta_{l10} = P(x_l = 1|h = 0)$ $(l = 1, \ldots, r)$，总共有 $2 + 2r$ 个参数。基于这些参数，可以计算：

$$P(x_l = 1|Z = 1) = P(h = 1|Z = 1)P(x_l = 1|h = 1) + P(h = 0|Z = 1)P(x_l = 1|h = 0),$$

$$P(x_l = 1|Z = 0) = P(h = 1|Z = 0)P(x_l = 1|h = 1) + P(h = 0|Z = 0)P(x_l = 1|h = 0)。$$

如果 x_l 是主症，$P(x_l = 1|Z = 1)$ 和 $P(x_l = 1|Z = 0)$ 需要满足主症约束条件；如果 x_l 是次症，它们需要满次症约束条件。这样的约束共有 r 个。另外，每个参数必须在 0、1 之间。这组参数的初始化也采用拒绝采样方法。参数更新所需的偏导如下：

$$\frac{\partial \log P(D_i|\theta)}{\partial \theta_{l1b}} = \begin{cases} \dfrac{P(h=b|D_i,\theta)}{\theta_{l1b}} & \text{如果在}D_i\text{中 } x_j = 1 \\ -\dfrac{P(h=b|D_i,\theta)}{1-\theta_{l1b}} & \text{如果在}D_i\text{中 } x_j = 0 \end{cases}$$

$$\frac{\partial \log P(D_i|\theta)}{\partial \theta_{h1b}} = \frac{P(h=1,Z=b|D_i,\theta)}{\theta_{h1b}} - \frac{P(h=0,Z=b|D_i,\theta)}{1-\theta_{h1b}},$$

其中，$b = 1$ *或* 0。

基于文献[20]提及的数据，使用不同的δ_1、δ_2取值，对图 2 中的模型进行约束参数估计，并基于结果利用文献[11]所述的方法获得辨证规则，结果如表 2 所示。通过观察辨证规则的特点，使用德尔菲法，基于与第 4 节类似的考虑，把δ_1和δ_2的值分别定为 0.5 和 0.2。

表 2. 使用不同约束参数值对二型糖尿病中肝胃郁热进行约束隐树分析的结果

δ_1	δ_2	模型的 BIC 分	证候类大小	辨证规则
0.4	0.1	-5702.01	0.11	主症：胃脘痛(3.4)，心烦(2.3)，口苦(2.8)，口干(2.8)，舌色红(1.3)，舌苔黄(1.3)，脉弦(1.3)。 次症：胃脘痞胀(0.5)，大便干(0.7)。
0.4	0.2	-5688.2	0.13	主症：胃脘痛(4.0)，心烦(2.4)，口苦(2.8)，口干(2.4)，舌色红(1.3)，舌苔黄(1.5)，脉弦(1.4)。 次症：胃脘痞胀(1.1)，大便干(1.4)。
0.5	0.1	-5722.62	0.05	主症：胃脘痛(3.1)，心烦(2.3)，口苦(2.2)，口干(2.0)，舌色红(1.1)，舌苔黄(1.2)，脉弦(1.2)。 次症：胃脘痞胀(0.4)，大便干(0.6)。
0.5	0.2	-5706.99	0.08	主症：胃脘痛(3.1)，心烦(1.6)，口苦(2.4)，口干(2.9)，舌色红(1.3)，舌苔黄(1.3)，脉弦(1.4)。 次症：胃脘痞胀(0.7)，大便干(0.8)。
0.5	0.3	-5682.2	0.06	主症：胃脘痛(2.5)，心烦(1.7)，口苦(2.1)，口干(3.2)，舌色红(1.0)，舌苔黄(1.4)，脉弦(1.1)。 次症：胃脘痞胀(0.7)，大便干(0.9)。
0.6	0.1	-5743.96	0.04	主症：胃脘痛(2.0)，心烦(1.8)，口苦(2.0)，口干(2.8)，舌色红(1.2)，舌苔黄(1.1)，脉弦(1.2)。 次症：胃脘痞胀(0.3)，大便干(0.5)
0.6	0.2	-5707.43	0.04	主症：胃脘痛(2.6)，心烦(1.7)，口苦(1.9)，口干(1.7)，舌色红(1.5)，舌苔黄(1.3)，脉弦(1.8)。 次症：胃脘痞胀(0.6)，大便干(0.8)。
0.6	0.3	-5685.17	0.05	主症：胃脘痛(2.4)，心烦(1.6)，口苦(2.0)，口干(2.8)，舌色红(1.3)，舌苔黄(1.2)，脉弦(1.3)。 次症：胃脘痞胀(0.7)，大便干(0.9)。
0.6	0.4	-5683.74	0.04	主症：胃脘痛(2.3)，心烦(1.9)，口苦(1.8)，口干(2.3)，舌色红(1.5)，舌苔黄(1.1)，脉弦(1.1)。

				次症：胃脘痞胀(0.8)，大便干(1.0)。

7. 隐树分析与约束隐树分析的比较

与隐类分析的情况类似，在隐树分析中加上约束会降低模型与数据的拟合程度，但是能使模型更好地反映主、次证的特点。例如，肝胃郁热证的主、次症分别是：

主症：胃脘灼热、胃脘痛、心烦、口苦、口干、舌色紅、舌苔黄、脉弦。

次症：胁胀闷、胃脘痞胀、大便干。

由于数据中症状“胃脘灼热”和“胁胀闷”出现少于 5 次，所以在数据分析中把它们忽略。隐树分析得到的模型的 BIC 评分是 -5639。它把患者分为两类，其中有肝胃郁热患者占 27%，相应的辨证规则如下：

主症：胃脘痛(4.7)、心烦(2.8)、口苦(4.1)、口干(4.3)、舌色红(0.3)、舌苔黄(0.4)、脉弦(-0.4)；

次症：胃脘痞胀(4.8)、大便干(2.2)。

这个辨证规则明显不合理。胃脘痞胀是次症，但其分值却最高，而主症舌色紅、舌苔黄、脉弦的分值却很低，甚至为负数。

约束隐树分析得到的模型的 BIC 评分是-5707，低于隐树分析得到的模型，表示它与数据的拟合程度相对较低。约束隐树分析也把患者分为两类，其中有肝胃郁热证患者占 8%，相应的辨证规则如下:

主症：胃脘痛(3.1)，心烦(1.6)，口苦(2.4)，口干(2.9)，舌色红(1.3)，舌苔黄(1.3)，脉弦(1.4)；

次症：胃脘痞胀(0.7)，大便干(0.8)。

这个规则较好地反映了主症、次症的特点，比隐树分析得到的规则更为合理。由此可见，约束隐树分析能够切当地把主、次症语义与数据结合，得出定量辨证规则。

最后需要说明的是，在实践中应该同时使用束隐类分析和约束隐树分析，取模型 BIC 评分大者作为组合结果[22]。由于隐类模型是隐树模型的特例，这种方法称为广义的约束隐树分析。对于胃肠湿热证它得到的结果是一个隐类模型（图 1）。对于肝胃郁热证，它得到的是一个三层隐树模型（图 2），这个模型的 BIC 评分是-5707，高于相应的隐类模

型的 BIC 评分-5747。另外，图 2 比相应的隐类模型也更符合中医理论，中层的隐变量 h1、h2 和 h3 分别反映胃肠湿热证在胃脘、口腔和舌像上的表现。

8. 结束语

建立证候分类定量规则的常用方法是诸如逻辑回归和支持向量机等的有监督学习方法，这些方法的出发点是有标签数据，要求在数据收集过程中，不但考察患者的各种症状，而且还要对患者所属证候进行判断。其分析结果是对数据收集过程中做出的证候判断的总结，因此受主观因素影响较大。无监督学习方法基于症状分布规律对患者进行聚类，然后再建立定量辨证规则。它的出发点是无标签数据，不要求在数据收集过程中对患者所属证候进行判断，因此受主观因素影响程度较低。但是，无监督方法的结果往往与中医对证候的定性认识不吻合。症状的出现有多种原因，试图仅仅基于症状出现情况，确定症状与各种原因之间的定量关系，是不可能的。

本文提出一种新方法，即约束隐树分析，它将中医对证候的定性认识作为约束条件加入无监督数据分析过程[§]。相对于有监督方法，它的优点是不要求在数据收集过程中对患者所属证候进行判断，因此受主观因素影响程度相对较低。与无监督方法相比，它的优点是其结果与学界对辨证规则的定性认识吻合。约束隐树分析可以用来将主-次症形式的定性辨证标准定量化，使其在临床实践和学术科研中得到更好地运用。

参考文献

1. 国家中医药管理局批准发布. 中医病证诊断疗效标准[M]. 北京: 中国医药科技出版社, 2012: 93-94.
2. 中华中医药学会肺系病专业委员会. 支气管哮喘中医证候诊断标准（2016 版）[J], 中医杂志, 2016, 57(22): 1978-1980.
3. 中华中医药学会肝胆病分会. 病毒性肝炎中医辨证标准(2017 年版)[J], 中西医结合肝病杂志, 2017, 27(3):附 I.
4. 中国中医科学院. 中医循证临床实践指南-中医内科[M]. 北京: 中国医药科技出版社, 2012: 95-118

[§]就机器学习而言，本文提出的方法可以称为约束无监督学习，它位于有监督学习与无监督学习之间，又有别于半监督学习。

5. 李建生, 王至婉, 李素云, 王海峰, 晁恩祥, 王永炎. 慢性呼吸衰竭的中医证候诊断标准研制[C], 中华中医药学会肺系病分会成立大会暨第十五次全国中医肺系病学术交流大会, 2012:337-339.
6. 李建生, 王至婉, 李素云, 王明航, & 王海峰[J]. 慢性呼吸衰竭中医证候诊断的研究. 北京中医药大学学报, 2011, 34(11), 780-785.
7. Arji, G., Safdari, R., Rezaeizadeh, H., Abbassian, A., Mokhtaran, M., & Ayati, M. H.[J]. A systematic literature review and classification of knowledge discovery in traditional medicine. Computer methods and programs in biomedicine, 2019, 168: 39-57.
8. Van Smeden, M., Naaktgeboren, C. A., Reitsma, J. B., Moons, K. G., & de Groot, J. A.[J] Latent class models in diagnostic studies when there is no reference standard—a systematic review. American journal of epidemiology, 2013, 179(4): 423-431.
9. Li, Y., Aggen, S., Shi, S., Gao, J., Tao, M., Zhang, K., ... & Li, K.[J] Subtypes of major depression: latent class analysis in depressed Han Chinese women. Psychological medicine, 2014, 44(15): 3275-3288.
10. Zhang, N. L., Yuan, S., Chen, T., & Wang, Y.[J] Latent tree models and diagnosis in traditional Chinese medicine. Artificial intelligence in medicine, 2008, 42(3), 229-245.
11. Zhang, N. L., Fu, C., Liu, T. F., Chen, B. X., Poon, K. M., Chen, P. X., & Zhang, Y. L.[J] A data-driven method for syndrome type identification and classification in traditional Chinese medicine. Journal of integrative medicine, 2017,15(2): 110-123.
12. Laudy, O., Boom, J., & Hoijtink, H.[J] Bayesian computational methods for inequality constrained latent class analysis. New developments in categorical data analysis for the social and behavioral sciences, 2005,63-82.
13. 李晓斌, 王文萍, 李晓, et al. 中医证候主症、次症地位在中药临床试验中的体现[J]. 中华中医药学刊, 2014(1):30-32.
14. 潘定举. 主症不同赋分间中医证候疗效评价结果差异的分析及预测[J]. 中国新药与临床杂志, 2011(4):309-312.
15. 邢玉瑞. 中医辨证思维之主症分析[J]. 陕西中医学院院报, 2010, 33(1): 1-2.
16. 成肇智, 刘雁云.中医主症证治学简述[C]. 中华中医药学会中医诊断学分会第十次学术研讨会论文集, 2009 年.
17. 朱文锋. 论主症可为病名[J]. 陕西中医. 1989, 10(6): 258-259.
18. 张连文, 郭海鹏[M]. 贝叶斯网引论. 科学出版社, 2006.

19. Wright, S. J.[J] Coordinate descent algorithms. Mathematical Programming, 2015,151(1):3-34.
20. CHAN K., WONG V., TANG S. [C] Pragmatic Trial for Chinese Medicine: The Conceptual Design and Implementation of a Semi-individualized Trial for Diabetic Kidney Disease. Paper presented at the 7th Annual Meeting of the Good Practice in Traditional Chinese Medicine Research Association, Daegu, Korea, 2019, 7.
21. 张连文, 许朝霞, 王忆勤, 刘腾飞, 刘国萍, 李福凤, ... & 郭睿[J]. 隐结构分析与西医疾病的辨证分型 (Ⅱ): 综合聚类. 世界科学技术: 中医药现代化, 2012, 14(2), 1422-1427.
22. 许玉龙, 吴秀艳, 李延龙, 王天芳, 张连文, & 薛晓琳[J]. 基于隐结构分析建立中医证候分型规则的三种方法. 世界科学技术-中医药现代化, 2019, 22 (1):101-107.

Constrained Latent Tree Analysis: An Unsupervised Method for Turning Qualitative TCM Diagnosis Standards into Quantitative Rules

Nevin L. Zhang[1] Yulong Xu[2] Zhourong Chen[1] Yali Lv[2] Kam Wa Chan[3]

（1. Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, 2. School of Information Technology, Henan University of Traditional Chinese Medicine, 3. Department of Medicine, The University of Hong Kong.)

Abstract

Abundant TCM diagnosis standards have been published. They are mostly qualitative in nature and often in the form of major-minor symptoms. They are not as useful as quantitative classification rules. In this paper, we propose a method to turn qualitative standards in the form major-minor symptoms into quantitative classification rules.

The method is built upon a previous method, known as latent tree analysis (LTA), that first clusters patients into groups solely based on symptom occurrence data, and then establish quantitative classification rules. Symptom occurrence can be caused by multiple latent factors. It is difficult, if possible at all, to separate the impacts of one latent factor from those of another. Consequently, the results of LTA are often not consistent with the qualitative standards.

We propose to use the semantics of major and minor symptoms to constrain LTA so that the resulting classification rules are consistent with the qualitative standards. Symptom data and qualitative standards originate from different sources. Our method can be viewed as a method to combine the two sources of information. As such, it can improve the results based on either of the sources.

Corresponding authors: Yulong Xu (flyxyl@126.com), Kam Wa Chan (chriskwc@hku.hk).