贝叶斯网络在中医诊断中的应用研究①

范建平[1]，李常洪[1]，吴美琴[1]，梁嘉骅[1]，侯丽萍[2]
（1. 山西大学管理学院，太原 030006；2. 太原市类风湿病医院，太原 030006）

摘要：数据挖掘技术是分析复杂医疗病例的重要手段，但如何选择合适的指标将中医临床经验科学化以提高医生的诊断和治疗水平、并对其进行有效监控是中医急需解决的问题. 针对这个问题，提出一种医疗指标约简的方法——基于聚类的贝叶斯网络模型，通过分类及进行主特征和类特征的提取来研究中医症候分型、动态演变，为类风湿关节炎病因、病机的研究及临床医生诊断质量控制提供了重要依据.

关键词：临床实验；数据挖掘；聚类；贝叶斯网络
中图分类号：N39 **文献标识码**：A **文章编号**：1007－9807(2008)06－0143－08

0 引 言

现代西医通过定量化的研究使得西医成为世界医学的潮流，与西医诊断定量化的思想不同，中医学的思维方式是经验直观的，因此，如何将中医的抽象思维、非定量、模糊地研究整体问题各方面的关联转变为科学定量的研究是中医发展的必然趋势.

数据挖掘是近年来随着人工智能和数据库技术的交叉融合而兴起的边缘学科，是近十年来使用较多的一种重要方法. 它致力于发现隐含在数据中的关于事物本质和事物发展趋势的知识或规律，并为专家决策提供支持[1].

由于中医有大量病例数据，数据挖掘技术在中医药领域具有良好的应用前景，将其应用于中医诊断，不仅可以从临床诊断数据中辨析症候与症状间的复杂关系，总结归纳中医专家的辨证规律并模拟其诊断推理过程，还可能发现客观有用的知识以丰富专家经验和中医理论[2,3]. 采用统计的数据挖掘从个体的不确定性中归纳出总体的确定性规律，找出某种疾病的病因、病机及动态演变的规律，从统计分析进一步进行理论研究，最后升华到理论层次才能被认为是科学的结论. 将科学结论应用到实践当中去，通过与临床医生的诊断结果不断交互，在使科学规则不断完善总结的基础上，实现对医生的临床诊断进行考核. 但在数据的处理方法上，由于很多处理信息的方法带有太多的主观色彩，目前很少单独采用，须和其他方法结合使用[4].

基于上述原因，本文采用基于聚类的贝叶斯网络来对中医类风湿关节炎的诊断指标进行分型筛选，其中聚类系数的确定以及贝叶斯网络中节点的选择是指标筛选的关键. 因此，首先通过改进的K－均值方法对大量病例进行利于分型的球形聚类，根据聚类结果分析病机及动态演变规律，再加与专家经验结合共同选择进行贝叶斯网络学习的节点来提取类风湿病关节炎中医主特征及其关系，并辨析症候与症状间的复杂关系. 并且采用边收集病历边进行处理的形式，动态地修改病历库、规则库等，同时也遵循人机交互、综合集成的思想[5]. 本文从多方法有机结合的角度建立并扩展了统计模型，并以新型的方法实现对中医诊断质量的控制.

① 收稿日期：2006－12－29；修订日期：2008－04－23.
基金项目：国家自然科学基金资助项目(70501016).
作者简介：范建平(1975—)，男，汉族，山西武乡人，博士生，讲师. Email：fjp@sxu.edu.cn

1 理论与数据预备

1.1 中、西医比较

(1)古代中西方医学虽然同属于经验医学的范畴,中国医学研究的根本思想是“治病救人”,但是中国医学更重视整体、功能和临床的疗效,病因病源的探索局限在整体和宏观上. 而西方医学更重视于局部、物质性和对本原的探索.

(2)西方科学技术发展为西方实验医学发展所造就的技术中介突飞猛进,使西方医学走上了以解剖探索为基础的实验医学道路,同时显微镜的发明,使解剖学进入动物体的微观层次. 从此,西方医学采用实验和定量分析方法,与科学技术相结合大踏步前进. 但由于缺乏临床试验的条件,西方医学由经验医学转向实验医学时并未走临床实验的道路,而采用了类比的动物实验. 中医系统地观、辩证的分析方法、功能性的分析手段是西方医学无法相比的. 但中国医学仍停留在经验医学范畴上探索,虽然也取得了很大的进展,但是始终未进入微观和物质性分析的层次,同时对病因、病机的分析则往往建立在类比分析的基础上,缺乏有目的、有计划的实验验证和定量分析.

(3)对医生的考核方式不同. 西医的定量化研究问题的方法也为它的考核提供了依据,使得西医的中等水平医生占主要比例. 而中医依靠经验的诊断本身就是不可靠的,基于经验的考核也是不可靠的. 因此中医长期缺乏有效的考核依据,形成了中医劣等水平医生占主要比例的状况.

1.2 中医的发展

鉴于以上的原因,本文认为中医发展要从以下几个方面做起.

(1)中国医学研究的五脏六腑不是解剖学意义上的,而是功能意义的. 这就决定了中医经验医学向现代化发展,必须走与西医发展不同的道路.

1)中医现代化首先是进行临床实验,而不是动物实验;

2)以现代数学及信息技术为手段进行结构化、规范化、定量化的整理,形成全面的、统一的、规范科学的理论医学;多学科交叉向现代系统医学发展;

3)中医在中西医结合中更要发挥以“整体性和自发性,研究协调和协同”为核心;

4)证候是中医临床诊断与治疗的关键环节,是中医疗效评价的依据之一,运用临床流行病学和计算机智能为中医“证”的研究提供了重要的方法.

(2) 建立中医的诊断标准.

1)收集患者的发病原因、中医症状特征、中医治疗方法、疗效等完整的病历档案;

2)通过知识挖掘,得出具有统计意义的病因、病机、分类、诊断标准、优化治疗方案、疗效评价;

3)对结果用中医理论进行解释、处理,建立一套以中医理论为基础的病因、病机、证候分型、诊断标准、动态演变、治则治法、治疗方案的理论与实用系统.

(3) 通过最终形成的诊断标准来对医生的诊断进行考核,促使中医水平的整体提高.

对于类风湿关节炎由于缺乏统一规范的诊断标准,对其病理机制看法众说不一. 而使用中医治疗,就要用中医解释病因、病机、分类、定名,最好的办法是临床试验. 本文认为运用“临床试验——统计分析——数据挖掘——理论分析——理论与应用系统重建——诊断质量控制”的方法,是研究类风湿关节炎中医病因、病机、证候分型、诊断标准、动态演变和提高整体中医医疗质量的有效方法,并相信这一方法能广泛地用于中医的各个领域,使中医理论由经验科学迅速上升到现代意义的科学.

为此,参照美国风湿病学会诊断标准(ARA),结合临床研究,选择临床 71 项症状、体征及化验指标,对几年来积累的大量临床资料进行严格筛选,从中筛选出具备较长时期临床观察并具有完整准确记录的 1 512 例确诊病例,同时对这些病例的 71 项指标进行了逻辑数理化处理,作为定量化研究的基础资料.

1.3 数据的集成

由于这些数据是从数据库中取得的,来自不同的数据库或数据表. 因此,会产生数据的不匹配或者数据的冗余,所以,首先要将这些与研究相关的数据库或数据表进行数据集成. 集成后的数据结构如表 1 所示.

表 1　病例样本集成后的数据结构

Table 1 The integrated data structure of cases

样本号	症　状							
	年龄	性别	症状 1	…	症状 e	症状 f	…	症状 p
⋮								
样本 i					x_{ie}	x_{if}		
⋮								
样本 j						x_{jf}		
⋮								
样本 n								

表中，n 是类风湿数据库中所有病例样本的个数，x_{ie} 与 x_{if} 为两个症状属性，$\overline{x_{ie}}$ 与 $\overline{x_{if}}$ 分别是 x_{ie} 与 x_{if} 的平均值，$\sigma_{x_{ie}}$ 与 $\sigma_{x_{if}}$ 分别是 x_{ie} 与 x_{if} 的标准差. 并且通过相关分析检测两个属性的数据是否具有冗余性，两个症状的属性 A 与 B 之间的相关性可以用下式进行度量

$$r_{A,B} = \frac{\sum (x_{ie} - \overline{x_{ie}})(x_{if} - \overline{x_{if}})}{(n-1)\sigma_{x_{ie}}\sigma_{x_{if}}} \tag{1}$$

如果上式的值大于 0，意味 A 的值随 B 的值增加而增加，表明一个症状属性蕴涵另一个的可能性增大；如果结果值等于 0，则表明症状 A 与症状 B 之间是不相关的；如果该值小于 0，则表明一个症状的值会随另一个减少而增加.

1.4　数据的规范化

中医疾病里有很多症状特征，没有明确的量化指标，所以首先要进行量化. 量化后的对象是由区间标度变量，二元变量、标称变量这些混合类型的变量所描述的. 所以最适宜的方法是用一种技术将不同类型的变量组合在单个相异度矩阵中，把所有有意义的变量转换到共同的值域区间[0，100]，有

$$v' = \frac{v - \min}{\max - \min}(new\_\max - new\_\min) + new\_\min \tag{2}$$

式中，$new\_\max = 100$，$new\_\min = 0$.

另外，定义病例样本 i 和 j 之间的相似度 $d(i, j)$ 为

$$d_{i,j} = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}} \tag{3}$$

数据结构的处理如表 1 所示. p 表示症状变量个数，i 和 j 表示样本，f 表示从 1 到 p 的任何一个症状，δ_{ij}^{f} 为指示项，其中，如果对象 i 或对象 j 没有症状变量 f 的度量值，或该症状变量的值均为 0($x_{if} = x_{jf} = 0$)，且症状变量 f 是不对称的二元变量，则指示项 $\delta_{ij}^{f} = 0$；否则，指示项 $\delta_{ij}^{f} = 1$. 症状变量 f 对样本 i 和样本 j 之间相异度的计算方式与其具体类型有关：

(1) 如果 f 是二元变量 若 $x_{if} = x_{jf}$，则 $d_{ij}^{f} = 0$，否则 $d_{ij}^{f} = 1$；

(2) 如果 f 是区间标度变量 则 $d_{ij}^{f} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$，这里的 h 遍取变量 f 的所有非空缺对象.

2　聚类分析模型

聚类分析是指把不同的样本进行分类，它是目前常用的一种统计方法. 聚类对象组成的特点就是在一个类中的对象具有很高的相似性，而与其他类中的对象很不相似. 聚类挖掘技术在挖掘数据规则、类型分析等方面有广泛的应用[6].

由于系统的高度复杂性和指标的模糊性，常常是事先无法确定预分类数. 即使对系统有一些了解，事先给定预分类数，实际上是对算法的一种人工干预，很可能伤害了分类的科学性[7]. 因此，可以根据样本的具体情况预先确定分类数及收敛精度，实现样本的最优划分.

另外，不同的聚类方法适用于探测不同类型的聚类，有的聚类方法划分数据使聚类内的点间最大距离最小化，这样形成的是紧凑的和大体球形的聚类；而有的聚类则使聚类中的每一点与类

中的某一成员尽可能的近，这样形成的是类似于香肠形状的聚类[8]. 如果要对给定的症状判定所属类风湿病的类型，本文研究的对象较适合于采用前一种方法. 这是因为通过聚类分析以寻找这种疾病的子类型时，由于患者可能是在疾病的不同阶段接受检验的，因此，采用后一种方法会使得不同的阶段界限模糊而都聚为一类，不利于各型的分割. 基于此，本文采用一种改进的层次聚类算法，即首先在所有对象中大体为球形的高密度点来确定凝聚点，并计算各个点与凝聚点的互连度来聚类.

步骤 1 求各点的密度 p_i

确定一个正数 d_0，以每个样本点 x_i 为中心，d_0 为半径作 n 维空间的超球，$d(x_i,x_j)$ 为 x_i 到 x_j 的距离，其中

$$d(x_i,x_j) = \sum_{p=1}^{n}(x_i - x_j)^2 \tag{4}$$

若满足 $d(x_i,x_j) \leqslant d_0$ 则认为 x_j 落在超球内，落在超球内的点的总数为 x_i 点的密度 p_i，容易看出，p_i 越大，则以它为凝聚点的资格就越大[9].

步骤 2 确定凝聚点

密度 p_i 由下式得到

$$f^D(x) = \sum_{i=1}^{n} e^{-\frac{d(x_i,x_j)}{2\sigma^2}} \tag{5}$$

依次取密度 p_i 最大的点作为凝聚点. 设凝聚点的集合为 S，已经有 k 个凝聚点 $s_i \in S(i = 1,2,\cdots,k)$，设类间距离参数为 D 与密度参数为 T，若满足 $f_j > T$ 与 $d(s_i,x_j) > D$，则 x_j 为下一个凝聚点.

步骤 3 聚类

k 个凝聚点选出后，设各类为 $C_i, i = 1,2,\cdots,k$，以这 k 个凝聚点为中心进行凝聚. 计算剩余的 $n-k$ 个点 C_j 与这些聚类中心的互连性

$$RI(C_i,C_j) = \frac{|EC_{\{C_i,C_j\}}|}{\frac{1}{2}(|EC_{C_i}| + |EC_{C_j}|)} \tag{6}$$

式中，$EC_{\{C_i,C_j\}}$ 为 C_i 与 C_j 凝聚时被取代的边；EC_{C_i} 是它的最小截断等分线的大小，当 $RI > R$ 时，R 为预先设定的值，就将其聚为一类.

该方法在某些方面有一定的局限性. 因为改进的 k - 均值聚类方法与数据分析都与初始选择的凝聚点有很大关系，分类效果可能受此影响. 为此，文献[10]对该方法进行了调整，使其学习能力更加完善.

在本文中，由于中医症状很多，所以在聚类之前，要对剔出冗余和含有大量空缺值后的 65 项症状指标进行频率统计，目的是再剔除掉那些扰乱聚类含义的指标；在找到频率较高的症状指标后，按照上述改进方法进行聚类. 聚类的指标有 40 项，如表 2 所示.

表 2 聚类时采纳的指标

Table 2 The indexes of clustering

1	2	3	4	5	6	7	8
肿胀	血沉	皮温	白血球	疼痛	恶寒	乏力	RF(+)
9	10	11	12	13	14	15	16
渴不欲饮	夜尿多	关节变形	活动后重	肌瘘着骨	Hb	γ球蛋白	血小板减少
17	18	19	20	21	22	23	24
心慌	手足心热	自汗	数	滑	滑数	沉滑	沉紧
25	26	27	28	29	30	31	32
弦细数	滑细数	舌红	红瘦	白腻	黄腻	薄黄	薄白
33	34	35	36	37	38	39	40
胖大	少苔	镜面	裂纹	晨僵	GTP	沉弦细	红绛

结果显示，在采用聚类时，聚类系数变化率最大的是 2、3、4、6 类，他们的变化率如表 3 所示.

在对聚类的结果进行有效性分析时，采用单因素方差分析法，发现在分为 4 类的情况下，系统聚类结果在关节疼痛、肿胀等 14 项症状中无显著差异外，其他各变量类间均有显著差异，p 值小于 0.05，这与中医的实际经验是相符的. 因此，本文将类风湿关节炎分为 4 型，并将这里的 14 项指标（如下表 4 所示）作为下一步贝叶斯网络学习的变量.

表3　聚类系数变化率

Table 3 The clustering coefficients diversification rates

聚类数	1	2	3	4	5	6	7
系数变化率(%)	21.2	45.6	41.3	44.2	16.7	39.9	15.4

表4　聚类分析后得到的主要症状变量

Table 4 The main symptoms after clustering analysis

1	2	3	4	5	6	7
关节肿胀	血沉	皮温	白血球	关节疼痛	恶寒	舌质红
8	9	10	11	12	13	14
RF(+)	渴不欲饮	晨僵	关节变形	活动后重	乏力	脉滑

3　样本主特征及类特征提取

中医的诊断与西医不同，应该类似于以下的诊断方程(诊断 = 诊断要素 × 贡献率(或称影响因子) + 诊断要素 × 贡献率 + … +)[11]. 诊断要素包括证候、症状、体征、舌脉等，贡献率大的要重视其诊断意义，贡献率小的可以忽略，但哪些证候、症状、体征应重视或忽略呢?

3.1　基本定义

事物往往有许多特征，有些是在事物的发展变化中起着主导作用的，有些特征对事物的发展变化影响并不大，故需将起主导作用的特征提取出来，以识别其类型. 这些起主导作用的特征即为主特征. 如本文谈及的类风湿关节炎，其临床症状若一一罗列出来，可能有上百个，但不可能每个症状都是它的主要症状，所以需要将其主要症状找出来，以区别于其它疾病，此即为类风湿关节炎的主特征.

将事物进行分类后，每一类区别于其它类型的主要特征即为类特征. 类特征是标志着该类型的根本特征，据此可以判定某病例样本是否属于此类.

定义 1　V_{Ei} 为样本总体关于特征 Ei 的特征值.

定义2　设定一个值 KV_{Ei}，当 $V_{Ei} > KV_{Ei}$ 时，称 Ei 为样本总体的主特征.

定义 3　特征集中主特征组成的子集称为主特征子集.

定义 4　记

$$T(a_k,a_j) = \frac{\sum \min[FEi(a_k),FEi(a_j)]}{\frac{FEi(a_k),FEi(a_j)}{n}} \tag{7}$$

$T(a_k,a_j)$ 为样本 a_k 与 a_j 的对于症状特征 Ei 的贴近度，若 Ei 为主特征，则称为主特征集贴近度.

很显然 T 在[0,1]上取值，当 $T(a_k,a_j) = 1$ 时 a_k 与 a_j 完全重合，$T(a_k,a_j) = 0$ 时 a_k 与 a_j 完全不重合.

定义 5　记 $L(a_k,a_j) = 1 - T(a_k,a_j)$，$L(a_k,a_j)$ 称为 a_k,a_j 间的 FUZZY 距离.

3.2　贝叶斯网络模型用于主特征及类特征提取

贝叶斯网络可以进行知识发现，贝叶斯网络学习技术能够通过数据分析自动创建贝叶斯网络. 贝叶斯网络通过从数据中发现变量间的因果关系，并用概率定量表示这些因果关系的强度. 与神经网络及规则库相比，贝叶斯网络更适合于大型数据库的数据挖掘. 但由于贝叶斯网络的计算量会随着节点数目的增加呈指数级增长，因此，节点的数目应尽量得到控制.

本文结合贝叶斯理论，提出节点约简方法，即首先利用删除冗余数据、频率统计与聚类分析将节点数目进行约简，然后通过专家经验对这些症状节点进行验证，再构造贝叶斯网络结构.

(1) 贝叶斯网络模型及贝叶斯分类器的构造

贝叶斯分类器是一类诊断效果很好的简化贝叶斯网络模型，它假设各属性在以类别变量为条件时相互独立. 设置 C_m(m = Ⅰ型，Ⅱ型、Ⅲ型、Ⅳ型)是类风湿关节炎的类别变量，即聚类后分成的类型，每一类型对应一种类别，A_k(k = 1,2,3)则是用于描述类别变量的属性(如疾病的症状或理化检测指标)，即每一种类型所具有的症状特征，结构如图 1 所示. 贝叶斯分类器模型先从病例数据中学习得到每个症状 A 在每种分型下的条件

概率,再根据待诊断病例的症状取值情况,应用贝叶斯定理计算每种类别的后验概率,最后把后验概率最大的类型作为该病例的 4 型判定结果.

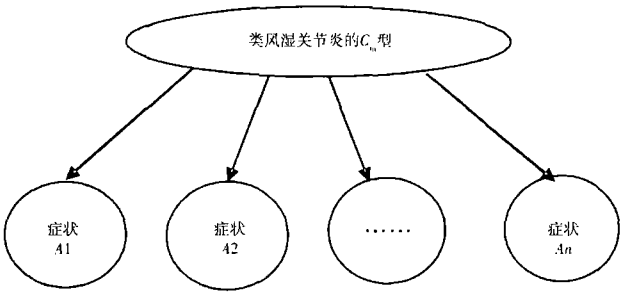

图 1 类风湿关节炎的贝叶斯分类器

Fig. 1 The Bayesian belief classifier of RA

(2) 贝叶斯信念网络的学习

由于病症的发展是不断演变的过程,在每个阶段症状都表现不同. 但是本文通过聚类分析已经将所有病例分为 4 类,分类的有效性经过了数学方法与专家经验的综合判定,即初步形成了贝叶斯网络的结构. 因此,可以采用梯度下降的方法来训练网络学习和形成条件概率表的值[12]. 仍然使用上面的 n 个训练样本$[a_1,a_n]$的集合 N,其中的每个样本为 X_d,$w_{p_1\cdots p_n}$ 是对于区分类风湿关节炎各型有意义的症状的一个条件概率表,这些症状用 $U_i = u_{ik}$ 来表示,类风湿关节炎由 $Y = y_{ij}$ 表示,u_{ik} 与 y_{ij} 取值为 T 与 F. 如表 6 所示.

表 6 类风湿关节炎条件概率表的设计

Table 6 The design of RA conditional probability

$Y = y_{ij}$	$U_i = u_{ik}$			
	症状 $p_1,p_2,\cdots,p_n$			
	$T,F,F,\cdots,F$	$F,T,F,\cdots,F$	…	$F,F,\cdots,T$
	$w_{p_1\cdots p_n}$			
类风湿(T)	0.83	0.69	…	0.32
类风湿(F)	0.17	0.31	…	0.68

目标是最大化 $P_w(N) = \prod_{d=1}^{n} P_w(X_d)$,但为了使问题简单,可以按 $\ln P_w(N)$ 来做,将所筛选症状的各种组合的初始权值都设置为 0.5,并用下面的方式进行迭代计算,在每次迭代中修改这些权,最终收敛到局部最优解.

1) 计算梯度. 对每个 i,j,k,计算

$$\frac{\partial \ln P_w(N)}{\partial w_{ijk}} = \sum \frac{p(Y_i = y_{ij}, U_i = u_{ik} \mid X_d)}{w_{ijk}} \quad (4)$$

2) 沿梯度方向趋近,使权值得到更新,学习率 r 被设置为一个小常数

$$w_{ijk} \leftarrow w_{ijk} + (r)\frac{\partial \ln P_w(N)}{\partial w_{ijk}} \quad (5)$$

3) 重新规格化权值. 由于权值 w_{ijk} 是概率值,必须介于 0 到 1 之间,并且对于所有的 i,k,$\sum_j w_{ijk}$ 必须等于 1. 在权值更新之后,可以对它们重新规格化来保证这一条件.

(3) 结果

在聚类之后,采用基于信息论的贝叶斯网络结构学习算法学习 14 个变量间的因果关系,找到与"是否类风湿关节炎" 直接相关的症状,称之为"主特征",它们分别为"关节疼痛、关节肿胀、关节变形、晨僵、类风湿因子阳性、舌质红、脉滑" 这 7 项症状,因果关系如图 2 所示. 计算出各症状以"是否风湿关节炎" 为条件时的条件概率后,再用信息熵分析法分别计算每个症状对"是否类风湿关节炎" 的信息贡献量,它能够反映该症状对类风湿关节炎证诊断的贡献. 通过计算得到 14 项症状与"是否风湿关节炎" 间的诊断贡献度,如表 7 所示. 可以发现"晨僵、类风湿因子阳性" 的贡献率最大,"关节肿胀","关节变形","白血球","血沉" 的贡献度依次降低.

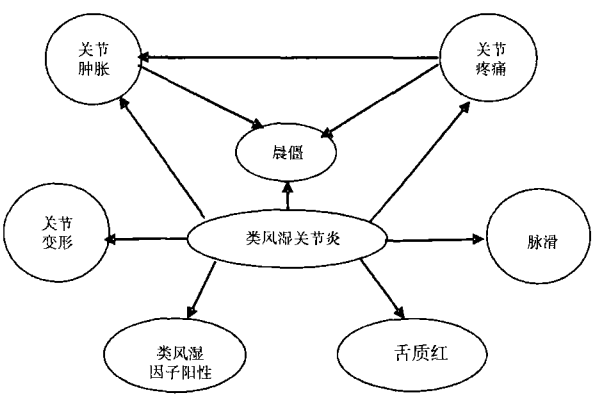

图 2 提取出的 7 个主特征症状的因果关系图

Fig. 2 The consequence of 7 main characters of symptoms

表7　主特征的诊断贡献度

Table 7 The diagnose contribution degree of main character

症状	类风湿因子阳性	晨僵	关节肿胀	关节疼痛	舌质红	脉滑	关节变形
诊断贡献度	0.980	0.826	0.796	0.78	0.679	0.644	0.590
症状	血沉	皮温	白血球	恶寒	渴不欲饮	活动后重	乏力
诊断贡献度	0.486	0.347	0.358	0.273	0.211	0.198	0.187

由上面的分析可以得出类风湿关节炎的主特征为：

1）关节肿胀；

2）晨僵 ≥ 1h；

3）关节疼痛；

4）关节变形；

5）类风湿因子阳性；

6）脉象滑；

7）舌质红.

同理，对分型后的4型分别进行了贝叶斯学习，得到了各类的类特征.

4　结束语

本文对收集到的1 512例类风湿关节炎患者临床资料进行了数据整理，从原始病例中提取出71项指标. 经过数据集成后形成65项指标，再经过频率统计、聚类分析将指标缩减为40项指标，并将类风湿关节炎病情发展的动态过程分为4型. 最后通过贝叶斯网络学习提取出类风湿关节炎的7项主特征及Ⅰ型、Ⅱ型、Ⅲ型、Ⅳ型的类特征. 为中医辨证分型及类风湿关节炎中医诊断标准提供临床依据. 通过对4型的症状变化分析，可以帮助临床医生分析类风湿关节炎的病因、病机及病理演变规律.

进一步的研究工作包括，优化用于分析的指标体系；改进聚类算法；此外，将本文的方法推广到其他各种疾病的效用性也是本文将要继续探讨的内容.

参 考 文 献：

[1]由　松，胡立胜. 中医症状及证候的量化方法探讨[J]. 北京中医药大学学报，2002，25(2)：13—15.
You Song, Hu Li-sheng. Study on the methodology of the quantification of TCM symptoms and syndromes[J]. Journal of Beijing University of Traditional Chinese Medicine, 2002, 25(2): 13—15. (in Chinese)

[2]吴美琴，范建平，梁嘉华，等. 医疗质量考核的新思想[J]. 中华医学科研管理杂志，2006，(6)：217—220.
Wu Mei-qin, Fan Jian-ping, Liang Jia-hua, *et al*. The new idea of the medical quality supervisor and assess[J]. Chinese Journal of Medical Science Research Management, 2006, (6): 217—220. (in Chinese)

[3]梁嘉华，范建平，曹暄伟，等. 管理信息与管理信息系统[M]. 太原：山西科学技术出版社，2001.
Liang Jia-hua, Fan Jian-ping, Cao Xuan-wei, *et al*. Management Information & Management Information System[M]. Taiyuan: Shanxi Science and Technology Publishing House, 2001. (in Chinese)

[4]赵卫东，李旗号，盛昭瀚. 基于案例推理的决策问题求解[J]. 管理科学学报，2000，3(4)：29—36.
Zhao Wei-dong, Li Qi-hao, Sheng Shao-han. Decision problem solution based on case-based reasoning[J]. Journal of Management Sciences in China, 2000, 3(4): 29—36. (in Chinese)

[5]于景元，周晓纪. 从定性到定量综合集成方法的实现和应用[J]. 系统工程理论与实践，2002，22(10)：26—32.
Yu Jing-yuan, Zhou Xiao-ji. The realization and application of meta-synthesis[J]. Systems Engineering—Theory & Practice, 2002, 22(10): 26—32. (in Chinese)

[6]程　岩，卢　涛，黄梯云. 在数据库中挖掘定量关联规则的方法研究[J]. 管理科学学报，2001，4(4)：41—48.
Cheng yan, Lu Tao, Huang Ti-yun. Research on methods of mining quantitative association rules in database[J]. Journal of Management Sciences in China, 2001, 4(4): 41—48. (in Chinese)

[7]诸克军，成金华，郭海湘. 模糊软分类中最佳聚类数的确定[J]. 管理科学学报，2005，8(3)：8—14.
Zhu Ke-jun, Cheng Jin-hua, Guo Hai-xiang. Optimal number of clusters in fuzzy soft clustering[J]. Journal of Management Sciences in China, 2005, 8(3): 8—14. (in Chinese)
[8]Hand D, Mannila H, Smyth P. Principles of Data Mining[M]. Cambridge: Massachusetts Institute of Technology, 2001.
[9]初 颖，刘 鲁，张 巍. 基于聚类挖掘的供应链绩效评价的标杆选择法[J]. 管理科学学报，2004，6(5)：49—53.
Chu Ying, Liu Lu, Zhang Wei. Benchmark selection for supply chain assessment based on clustering mining technology [J]. Journal of Management Sciences in China, 2004, 6(5): 49—53. (in Chinese)
[10]王碧泉，陈祖荫. 模式识别：理论、方法和应用[M]. 北京：地震出版社，1989.
Wang Bi-quan, Chen Zu-yin. Pattern Recognition: Theory, Method and Application[M]. Beijing: Earthquake Press, 1989. (in Chinese)
[11]王学伟，瞿海斌，王 阶. 一种基于数据挖掘的中医定量诊断方法[J]. 北京中医药大学学报，2005，28(1)：4—7.
Wang Xue-wei, Qu Hai-bing, Wang Jie. A quantitative diagnostic method based on data-mining approach in TCM[J]. Journal of Beijing University of Traditional Chinese Medicine, 2005, 28(1): 4—7. (in Chinese)
[12]Han J, Kamber M. Data Mining: Concepts & Techniques[M]. San Francisco: Morgan Kaufmann Publishers, 2001.

Study on application of Bayesian network in Chinese medicine diagnose

FAN Jian-ping[1], *LI Chang-hong*[1], *WU Mei-qin*[1], *LIANG Jia-hua*[1], *HOU Li-ping*[2]
1. School of Management, Shanxi University, Taiyuan 030006, China;
2. Taiyuan Rheumatism Hospital, Taiyuan 030006, China

Abstract: Data mining technology is an important method to analyze the medical cases. It is important to choose the appropriate indices to make the Chinese medicine more scientific and to improve the diagnosis level by supervising the quality more effectively. This paper brings forward a new medical indices reduction method-the Bayesian network based on clustering to solve this problem. This paper studies TCM syndrome type and dynamic evolution by classifying the data to pick-up the main characters and species characters. It provides the important evidence to RA pathogens and sickness mechanism research, and it also plays an important role in supervising diagnosis quality of the clinician.

Key words: clinic experiment; data mining; clustering; Bayesian network