·理论研究·

# 中医证候研究的现代方法学述评(一)
## ——中医证候数据挖掘技术

龚燕冰[1] 倪 青[2] 王永炎[1]
(1中国中医科学院临床基础医学研究所 北京 100700)
(2中国中医科学院广安门医院内分泌科 北京 100053)

**摘要:目的** 探索中医证候的现代研究方法。**方法** 对近年来的中国中医期刊有关中医证候的数据挖掘技术进行汇总,分析其优势与不足。**结果** 目前用于中医证候研究的数据挖掘方法主要有:关联规则、集对分析、粗糙集理论、聚类分析、人工神经网络、决策树、支持向量机、贝叶斯网络等。**结论** 中医数据具有非线性、模糊性、复杂性、非定量等特征,针对具体的医学数据和不同的挖掘目标往往要将几种方法综合起来应用,以发挥各自的技术优势。

**关键词**: 中医证候;数据挖掘技术;方法学
**中图分类号**: R2-03

## Modern methodology of TCM syndrome study(Ⅰ): Data mining technology of TCM syndrome

GONG Yan-bing[1], NI Qing[2], WANG Yong-yan[1]
(1 Institute of Fundamental Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700)
(2 Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing 100053)

Abstract: Objective To explore the modern research methods for TCM symptomatology. Method The data mining techniques of TCM symptomatology were summarized from different TCM magazines in recent years, and their advantages and disadvantages were analyzed. Result The result showed that the methods for data mining in TCM symptomatology included association rules, set pair analysis, rough set theory, cluster analysis, artificial neural network, decision tree, support vector machine and Bayes network, etc. Conclusion The TCM data have the characteristics of nonlinearity, indistinction, complicacy and unquantification and so on. These methods should be applied integratedly in accordance with the specific TCM data and different mining aims, and their advantages will be given a full play to.

Key words: TCM syndrome; data mining technique; methodology

中医证候信息的多模式特性是它区分于其他领域数据的最显著特征,这种多属性模式并存加大了中医数据挖掘的难度。许多证候信息的表达本身就具有不确定性和模糊性的特点,证候信息所体现出的客观不完整和描述疾病的主观不确切,形成了中医证候信息的复杂性。数据挖掘技术善于从海量数据中发现隐含的有意义的知识,预测未来趋势及行为,做出前瞻性的决策,正是这种优势使得数据挖掘技术在分析中医证候的研究中被广泛地采用并取得了许多有价值的成果。在数据挖掘之前必须对中医证候信息进行清理和过滤,将其变成适合挖掘的形式,以确保数据一致性。数据挖掘的基本步骤[1]包

龚燕冰,女,29岁,在读医学博士生

括数据选择、处理、转换、采掘和解释与评价几个阶段，目前应用于中医证候研究的数据挖掘方法主要有以下几种：

## 1 关联规则

关联规则[2]（association rules）是数据挖掘中最活跃的研究方法之一，最初提出的动机是针对购物篮分析问题，提出的目的是为了发现交易数据库中不同商品之间的联系规则，侧重于确定数据中不同领域之间的关系，找出满足给定条件下的多个域间的依赖关系。关联规则挖掘对象一般是大型数据库（Transational Database），该规则一般表示式为：A1∧A2∧…Am⇒B1∧B2∧…Bn，其中，Ak（k＝1，2，…，m），Bj（j＝1，2，…，n）是数据库中的数据项，有支持度（Support）（A⇒B）＝P（A∪B），置信度（Confidence）（A⇒B）＝P（A∪B），数据项之间的关联，即根据一个事务中某些数据项的出现可以导出另一些数据项在同一事务中的出现。

张氏等[3]收集了中医古籍文献中有名称的中医肾病治疗方剂，并建立相应的数据库，然后应用关联规则挖掘算法对该数据库进行复方配伍规律的研究，以获取能够表达复方配伍的确定性知识和随机性知识。结果发现在治疗肾病时有各药味同时出现的模式，因此在一定程度上该结果更能体现出中药复方配伍的科学内涵。

在中医证候的研究中，由大量的中医症状、舌脉表现组成的数据库相当繁杂，关联规则可以寻找出相关联的各个数据，当某些症状总是同时出现时，我们可以从中找出某种证型规律，甚至是病机规律。当然，其缺陷在于观测症状、证候与实验室指标之间的关联关系时不足以反映临床实际，比如口干这个症状与心电图正常这个指标有关联关系，但并不具有临床意义。

这里归结于预处理数据的关联规则很多，而绝大多数对于中医证候的研究者来说是没用的。为了在建模过程中提高模型在实际应用中的准确性，通常我们用最小支持度、最小置信度和兴趣度来衡量关联规则，只有支持度和置信度分别大于指定的最小值的关联规则，才是符合要求的关联规则模型，需要注意的是最小支持度和置信度的设定是由专家或专家群体设定的，这里又不可避免地加入了研究者的主观性。兴趣度（interesting）＝P（条件和结果）/P（条件）P（结果）。当兴趣度大于 1的时候，这条规则就是比较好的，当兴趣度小于 1的时候，这条规则就是没有很大意义的。兴趣度越大，规则的实际意义就越好。

## 2 集对分析

不确定性是自然界和人类社会中普遍存在的一种客观现象，也是中医药学的主要特点之一，目前不确定性的研究已经逐渐被提高到很重要的位置。集对分析[4]（set pair analysis，SPA）是一种用联系数 a＋bi＋cj统一处理由于模糊、随机、中介和信息不完全所致不确定性的系统理论和方法，很可能会成为处理中医药不确定性的捷径，将很大程度地推动证候规范化和中医药客观化的实现。

集对分析的特点是对客观存在的种种不确定性给予客观承认，并把确定性与不确定性作为一个既确定又不确定的同异反系统进行辨证分析和数学处理。集对分析处理不确定性理论的特色在于：①对不确定性给予“客观承认”；②把确定性与不确定性作为一个系统进行处理和分析；③从系统层次的观点认识不确定性的本质，微观层次上的不确定性不能随便确定；④不确定性以及不确定性系统不能孤立地存在；⑤集对分析中的不确定性理论也适用于对确定性问题的研究。归纳起来，集对分析对不确定性的处理思路可以简要地概括为“客观承认、系统描述、定量刻画、具体分析”。

集对分析理论用于中医数据挖掘的优势是对不确定性采取了与某些不确定性理论不同的处理方法，就是不像以往那样一味地去把不确定性转化为确定性来加以研究，而是把不确定性与确定性作为一个系统来加以研究。借助对这个系统中确定性与不确定性相互依存、相互联系、相互渗透，以及在一定条件下相互转化过程的描述、分析、处理，来研究不确定性在具体条件下的取值规律。

集对分析的不足之处在于对于不确定性的描述只能在系统层次上，而微观层面的认识还不能随便确定，所以适用于对中医宏观问题的研究，而对于相对精细的问题则束手无策了。如果能够和其他可以解决精确问题的方法结合，比如与粗糙集理论协同应用，则有扬长避短之效。

## 3 粗糙集理论

粗糙集理论[5]（rough set），是继模糊数学理论之后的又一种处理不精确和不确定问题的数学方法，是波兰学者 Z. Pawlak在 80年代初提出来的。它是一种研究不完整数据、不确定知识的表达、学习及归纳的数学方法，为研究不精确数据的分析和推理、挖掘数据间的关系、发现潜在的知识提供了行之有效的工具。粗糙集理论认为知识是对对象的分类

能力,对于知识,可以用属性和相应的值来描述。粗糙集理论的出发点是知识的不可识辨关系。这意味着由于缺乏信息,不可能通过已有信息识辨对象,只能将不可识辨族作为知识的一部分来处理。

确定规则就是某种证候诊断时的必要条件,可能规则就是可出现的症状和体征或检查结果,这样就在很大程度上避免了临床医生的主观性和片面性,粗糙集理论认为知识和概念可随知识本身的发展而不断扩展和完善,从而使中医证候诊断更加规范化、科学化。

将粗糙集理论引入到中医学中来,将为实现中医诊断智能化提供一种方法[6]。秦氏等[7]利用粗糙集理论建立中医诊断类风湿的模型。具体方法是以患者一般情况、症状、体征(包括舌象、脉象)、物理检查、实验室检查结果为主要依据,建立信息表(不是将其数据化),继而利用差别矩阵法进行属性简约与病例简约,得到下近似集和上近似集,从而抽取中医诊断的确定规则和可能规则。他还以该方法与模糊数学方法进行了比较,前者的诊断正确率远远高于后者。

粗糙集理论的优势在于它仅利用数据本身提供的信息,能搜索数据的最小集合,能从经验数据中获取易于证实的规则知识,它同时允许使用定性和定量的数据。并且与其他理论相结合,产生了大量的可以扬长避短的科学方法。重要的是,粗糙集理论并不是把中医症状单纯看作一组数据,而是将之看作是一种知识,然后运用复杂的数学原理对之进行知识自挖掘及学习,故而能处理大量非线性的不精确的、模糊的数据。它避免了临床医生的主观性和片面性,使其得出的概率不会因样本量的大小而影响其判断效果。此外,粗糙集理论的知识和概念不是固化的,随着知识的深化,粗糙集理论也可以使中医证候诊断更为准确。

## 4　聚类分析

聚类分析(cluster analysis)方法有两种分析策略,一种是对应分析方法,这种方法的基本思想是尽量保留主要信息,放弃次要信息,所划分出来的类别由于充分利用了数据的信息,而使分类更加合理,但由于这种方法理论的复杂性和计算上有一定的难度,应用并不太广泛。

另一种较常用的方法叫系统聚类,系统聚类可以对指标进行分类,在证候研究中,由于变量间普遍存在多重共线性关系,系统聚类可以把证候变量按相似程度大小进行归类,具有共线性关系的变量经聚类分析后归到一类,从而达到降维的目的,消除共线性对回归分析结果的影响,研究者可根据变量的情况选择具有代表性的指标进行下一步研究。由于聚类分析是对整个样本资料按指标和样品的相似程度进行归类,并不得出结论,故属于探索性分析[8]。

袁氏等[9]对 67个肾虚症状变量的轻、中、重不同程度总积分进行排序,对前 20个症状进行聚类分析,发现这 20个症状的类群基本反映了肾虚证候的几个主要方面。这 20个肾虚症状经过不同角度的聚类分析,其症状群落的结构和关系与中医理论的描述基本一致,为中医诊断学中有关肾虚的症状结构提供了科学的解释。

聚类的方法可以很容易地得出研究者所需的症状群或者数据群,并进行简单的一维解释,聚类技术的根本问题是对两个对象间距离和相异度度量的选择,针对两两对象之间的"相似度"或"相异度"划分不同类别。并不能从多维和多层次角度来全面分析数据并解释数据中真正复杂结构,而中医症状以及症状与证候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,这可能是目前的聚类分析方法所无法解决的。所以聚类的方法在中医证候的研究中,始终是一种辅助的手段。在统计学中,聚类分析和关联规则一样是属于无指导学习(unsupervised learning)的范畴[10]1, 306, 316,意即只能观察特征,而没有结果度量。

## 5　人工神经网络

人工神经网络(artificial neural network)的原理是通过模拟生物的神经网络结构和功能,实现对各种信息的有效处理。它通常包含输入层、输出层以及一个或几个隐含层,它的基本组成单位为神经元。输入层接受外界信号,不对其进行加工和处理,直接将其引入神经网络;隐含层位于网络的输入层和输出层之间,可包括多层,对输入的信息进行处理并将处理后的信息传给输出层(或下一个隐含层);输出层则输出经隐含层处理后的结果。可见,人工神经网络不需要精确的数学模型,而是通过模拟人的联想推理和抽象思维能力,来解决传统自动化技术无法解决的许多复杂的、不确定性的、非线性的自动化问题。

将此方法用于中医证候量化诊断模型已经有了初步的探索[11]:将数据让改进的 BP网络学习和训练,同时用录入的原始数据让改进的 BP网络学习和训练,用抽样检验的方法,采用相同的数据进行证候诊断检验,检验的结果是前者的证候诊断准确率

为 94.47%,后者的证候诊断准确率为 61.1%,前者远高于后者,说明中医证候特征矢量的提取,可以提高证候的诊断准确率。

人工神经网络具有很强的自组织性和容错性,在医学数据挖掘中得到了广泛的应用[12]。因此,基于神经网络的中医证候量化诊断模型研究,有可能为解决中医证候诊断标准研究中症状权值难以明确的问题提供更为科学的方法与途径。

## 6 决策树

决策树(decision tree)根据不同的特征,以树型结构表示分类或决策集合,产生规则和发现规律,其思路是找出最有分辨能力的属性,把数据库划分为多个子集,直到所有子集包含同一类型的数据,最后得到的决策树能对新的例子进行分类。决策树是发现概念描述空间的一种有效方法,也是许多归纳系统常采用的知识表示形式。

刘氏[13]用决策树的方法做出 2型糖尿病中医证素及其下属症状的模型,得出各证素的下属症状及其对该证素的贡献度,得出的糖尿病病性证素主要有:气虚、阴虚、阳虚、热盛、血瘀、痰、湿及湿热等;病位证素主要有:脾、肾、肝。将决策树方法运用于2型糖尿病证素的研究,简化了糖尿病气虚和燥热的现行诊断标准,方便了临床应用。

决策树的主要优点是描述简单、分类速度快,特别适合大规模的数据处理。决策树根据变量的值切分数据,应用"if—then"语句组成一个体系结构来分类数据。该方法的主要优点是比神经网络更快也更易于理解。但是,其主要缺陷是,数据类型是不连续的或者必须归为某类,这样使得连续数据不得不转换成一项数据类型,可能会导致有重要意义的数据点被删除。此外,如果条件比较复杂时,"if—then"语句也会变得复杂[14]。

## 7 支持向量机

支持向量机[10]271,263(Support vector machine SVM)可以扩展到多类问题,本质上是通过求解多个2分类问题。为每个类构造一个分类器,而最终的分类器是最有优势的一个,SVM在许多其他有指导和无指导学习问题中具有广泛的应用。经验表明它在许多实际学习问题中表现得很好。它允许扩大的空间维数非常大,在某些情况下可能无穷大。

支持向量机可以用于分子生物学中基因的分类、蛋白质一级结构的识别和预测蛋白质亚细胞水平的分布等[15]。SVM在医学数据挖掘和医学信息处理中的研究还处于起步阶段。由于它在分类和回归问题上的精确性以及对样本维数不敏感,相信支持向量机在中医的数据挖掘中也会得到更广泛的应用。

支持向量机是基于统计学习理论、针对小样本学习问题的一个理论框架[16],用于数据挖掘的最大优势在于:其计算复杂性与数据的维数不成正比,只和样本的数量有关,SVM对数据库中模式分类的准确率一般要高于神经网络。它的缺点在于对于维数非常大的数据,看上去计算量可能变得让人望而生畏,也许使用充足的基函数数据是可分的,但可能出现过分拟合。

## 8 贝叶斯网络

贝叶斯网络(bayes network)包括网络结构和参数集合两部分。网络结构是个有向无环图,由一个节点集合和一个节点间的有向边集合组成,任意两个节点间最多存在一条有向边,贝叶斯网络能够利用简明的图形方式定性的表示事件之间复杂的因果关系或概率关系,在给定某些先验信息后,还可以根据条件概率表定量地表示这些关系的强度。

王氏[17]等应用贝叶斯网络的相关技术从 474例病例的临床数据中发现血瘀证的关键症状,定量计算这些症状对诊断的贡献度,并建立血瘀证的诊断模型,用贝叶斯网络的方法发现了血瘀证的 7个关键症状,并定量计算其诊断贡献度。基于这些关键症状建立的简单贝叶斯分类器模型对血瘀证诊断的准确率达到 96.6%。结果表明贝叶斯网络技术适合于解决中医定量诊断问题。它可以揭示众多症状间以及症状与证候间的复杂关系,从中发现证候的主要症状和次要症状,并定量确定其诊断价值,有助于确定证候诊断的标准和规范。

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,具有以下优点:可以在更少的数据中学习到更准确的模型,学习 P(A)和 P(B)比联合概率分布 P(AB)需要的数据少;揭示了研究对象或领域的结构性质,有助于深入理解领域问题,丰富对领域对象的认识;网络结构蕴含的条件独立关系有助于认识事件间的先后关系,进行灵敏度分析和推理;网络结构的因果语义使得人们可以学习到事件间的因果关系,从而预测某些行为的可能结果。它的缺点是,由于任意两个节点间最多存在一条有向边,这就决定了两个结点的关系是有方向性的,是有先有后的,是一因一果的,不存在交互的、逆向的相关关系。对于繁杂的中医证候研究,单一的贝叶斯也无能为力了。

总之,在汇总了诸多的数据挖掘方法之后,我们力图寻找其在中医证候研究中的最有优势的一面,然而任何方法都不可能面面俱到。当每一种方法面对多维多阶的中医数据信息都力不从心时,我们深深地体会出中医数据挖掘的难度,最重要的是从另一个侧面体现了由王永炎院士首次提出的中医证候的"内实外虚、动态时空、多维界面"的特征,并遵循"以象为素,以素为候,以候为证,病证结合"的原则,所以中医证候是一个高维性、高阶性和非线性的复杂系统[18]。

面对这样错综复杂的定量与定性结合、主观与客观结合、确定与模糊结合、线性与非线性结合的海量的中医数据,针对具体挖掘目标,往往要将几种方法综合起来应用,以发挥各自的技术优势。如用聚类分析和关联规则等无指导的学习方法做探索性分析,并求助于有指导的学习方法如贝叶斯网络法、支持向量基方法等求得结果;粗糙集理论、人工神经网络、支持向量机等适用于复杂的、不确定性的、非线性的数据,结合起来应用可能会弥补单一方法的不足。

**参考文献**:


[1] LANK E. The Human Factor Long Range Planning[J]. Leveraging Invisible Assets, 1997, 30(3): 406-412.

[2] 王　华,胡学钢. 基于关联规则的数据挖掘在临床上的应用[J]. 安徽大学学报(自然科学版), 2006, 30(2): 21-25.

[3] 张承江,闫朝升,宋立群. 中医肾病治疗信息中关联规则的挖掘算法[J]. 黑龙江大学自然科学学报, 2005, 22(6): 842-845.

[4] 孟庆刚,王连心. 浅谈集对分析在证候规范化研究中的应用[J]. 北京中医药大学学报, 2005, 28(4): 9-14.

[5] 张文修,吴伟志. 粗糙集理论与方法[M]. 北京:科学出版社, 2003: 22-23.

[6] 王相东,殷　鑫. 粗糙集理论与证候规范化研究[J]. 陕西中医学院学报, 2005, 28(2): 70-71.

[7] 秦中广,毛宗源,邓兆志. 粗糙集在中医类风湿证候诊断中的应用[J]. 中国生物医学工程学报, 2001, 20(4): 357-363.

[8] 查青林,林色奇,吕爱平,等. 多元统计分析在中医证候研究中的应用探析[J]. 江西中医学院学报, 2004, 16(6): 79-80.

[9] 袁世宏,王米渠,王天芳. 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.

[10] TREVOR H, ROBERT T, JEROME F. 统计学习基础-数据挖掘、推测与预测[M]. 范　明,柴玉梅,咎红英,等译. 北京:电子工业出版社. 2004.

[11] 李建生,胡金亮,余学庆. 基于神经网络的中医证候量化诊断模型探索. 河南中医学院学报[J], 2005, 20(3): 6-8.

[12] 阎平凡,张民水. 人工神经网络与模拟进化计算[M]. 北京:清华大学出版社, 1999: 421-430.

[13] 刘延华. 糖尿病中医证候量化标准研究[D]. 中国博硕论文库.

[14] 崔　雷. 数据采掘及其在医学研究中的应用[J]. 情报理论与实践, 2001, 24(5): 330-333.

[15] CAI YD, LIU XJ, XU XB, et al. Support vector machines for perdiction of protein subcellular location[J]. Mol Cell Biol Res Commun, 2000, 4(4): 2305.

[16] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32.

[17] 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, 2006, 28(1): 4-7.

[18] 王永炎. 完善中医辨证方法体系的建议. 中医杂志[J], 2004, 45(10): 729-731.


(收稿日期: 2006-05-28)

# 《北京中医药大学学报》医学名词著录格式

医学名词以全国自然科学名词审定委员会公布的《医学名词》(科学出版社出版)为准。无通用译名的名词术语于文内第 1次出现时应注原词或注释。药名以《中华人民共和国药典》(2005年版)或《中国药品通用名称》(卫生部药典委员会, 1997年版)为准。药物名称不用商品名。统计学符号按 GB3358-82《统计学名词及符号》的有关规定书写,如:样本的算术平均数用英文小写 $\bar{x}$;标准差用英文小写 $s$; $t$检验用英文小写 $t$; $F$检验用英文大写 $F$;卡方检验用希腊文小写 $\chi^2$;相关系数用英文小写 $r$;概率用英文大写 $P$($P$值前应给出具体检验值,如 $t$值、$q$值等)。以上符号均用斜体。