·理论研究·

数据挖掘在亚健康状态研究中的应用展望*

王利敏[1] 赵歆[1] 陈家旭[1#] 岳利峰[1] 薛飞飞[2] 赵晖[1]
(1北京中医药大学 北京 100029; 2暨南大学医学院中医系)

摘要:探讨数据挖掘技术在亚健康状态研究中的应用。主要结合以下 3个方面:证候分布、症状构成、症状对证候诊断的贡献率;亚健康状态与证候之间、症状与证候之间、证候与证候之间的关系;亚健康状态及其证候与相关性指标之间的关系,并提出应统一分析宏观资料和微观指标,以便对亚健康状态的证候机理进行深入研究。

关键词:亚健康状态;数据挖掘;微观指标;宏观资料
中图分类号:R2—03

Application prospect of data mining in research of subhealth status*

WANG Li-min[1], ZHAO Xin[1], CHEN Jia-xu[1#], YUE Li-feng[1], XUE Fei-fei[2], ZHAO Hui[1]
(1 School of Preclinical Medicine, Beijing University of Chinese Medicine, Beijing 100029; 2 Chinese Medicine Department, School of Medical, Jinan University)

Abstract: The application of data mining in the research of subhealth status was discussed in the paper from following three points: syndrome distribution, symptom constructure and contribution rate of symptom to syndrome diagnosis; relationship between subhealth status and syndromes; symptoms and syndromes and syndromes and syndromes; relationship between subhealth status and its syndromes and relevant indexes. The author thought that the macroscopic materials and microscopic indexes should be analysed as a whole in order to study deeply the syndrome mechanism of subhealth status.

Key words: subhealth status; data mining; microscopic indexes; macroscopic data

随着社会竞争的加剧、生活节奏的加快、生活方式的改变,人们受到心理压力加大、不良情绪刺激等多种因素的影响,WHO的一项全球性调查表明,真正健康的人仅占 5%,患有疾病的人占 20%,而75%的人处于亚健康状态[1]。但由于对亚健康状态概念认识的模糊,难以和正常状态、疾病状态进行界定,加之其所表现的证候的多样性、复杂性和不确定性,使得亚健康状态的临床研究有着一定难度。

近年来,数据挖掘技术在中医诊断[2]及临床研究中被广泛应用,其目的是从大量、不完全、有噪声、模糊、随机的数据中,提取隐含在其中、人们事先未知但又是潜在有用的信息和知识,并预测未来趋势和行为。理论上认为,只要证候客观存在,那么同一证候的临床表现在大样本人群中必然表现为数据上的内部聚集性,尽管证候的复杂性、多维性使得临床数据表现出各证候之间交叉、重复、多维等复杂关系,但特定证候与其临床表现间必然表现为较强的关联性,因而数据上就会有规律可循,可采用数据挖掘方法进行处理和分析。

目前,数据挖掘技术在亚健康状态研究中的应用,主要是通过临床流行病学调查,获取大量的宏观特征资料,在此基础上,运用传统统计学方法对亚健康状态的人口学特征、亚健康人群分类、证候及症状特征、亚健康状态不同证候特点、影响因素和危险因素、亚健康定量测评等进行研究。而面对亚健康状态研究中的许多问题,如:对大量主

王利敏,女,在读博士生
#通讯作者:陈家旭,男,教授,博士生导师,主要研究方向:中医证候的病理生理基础及中医病证规范化, E-mail: chenjiaxu@hotmail.com
* 国家高技术研究发展计划(863计划)项目(No 2008AA02Z406)
(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

观症状即潜变量的直接测量、利用调查问卷获取真实可靠反映亚健康中医证候学特征和演变规律的资料、寻找亚健康状态相关性指标并与中医证候类型有机结合、亚健康状态研究数据有别于其他临床数据等，运用传统的统计学方法已感到力不从心，需要汇总诸多数据挖掘方法的特点，将几种方法综合起来，扬长避短，寻找出在亚健康状态研究中最有优势的方法。

1 亚健康状态判断、证候分布、症状构成、症状对证候诊断的贡献率

从中医学角度来看，亚健康状态虽无器质性病理改变，但机体已出现阴阳、气血、脏腑、营卫的不平衡状态，因此，亚健康状态有证可辨，有病因病机可究[3]。而明确辨证，明析病因病机，区别各症状在不同证候中的权重，从而确定主证、次症、兼症等问题，则亟待借助恰当的数据挖掘技术，对多中心、大样本的亚健康状态临床流行病学调查资料进行分析和研究。

支持向量机(SVM)分类方法在实际二分类问题的应用中，显示出良好的学习和泛化能力，且在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势，已逐渐应用于证候研究和疾病诊断领域。如：建立脾虚证多证型诊断分类器，应用该分类器对占消化性溃疡约 85%的脾气虚证、肝胃不和证、肝郁脾虚证进行分类，取得了满意的效果[4]；利用 SVM 构建癌症诊断模型，对检测者 ATP、SDH 酶进行探讨，用以辨别正常人与癌症患者[5]。可利用支持向量机，直接从亚健康状态临床流行病学调查数据出发，建立性能优良的分类器模型，以解决亚健康状态的判断问题。

张氏提出隐结构模型，通过临床流行病学调查，系统收集病例，再用电脑对所得数据进行分析，寻找潜在的影响症状出现规律的隐变量，并基于隐变量建立隐结构，用所得的隐结构模型来指导辨证[6]。运用隐类分析以及隐结构模型来揭示亚健康状态的症状(显变量)和证候(隐变量)的关系[7]。

朱氏采用“双层频权剪叉算法”来确定证候的诊断贡献度，提取所研究证型的证候特征，其中，“频权剪叉”是指高频数变量的权值轻、低频数变量的权值重，“双层”是指根据“频权剪叉”原理，对证素所见证候的权值进行分配，将各症状对各证候、证型的贡献度进行分配，形成证候和证素标准化权值[8]。该算法的应用，也为在多变量、非线性基础上，提取亚健康状态中医证型的证候特征提供了一种新的方法。

2 亚健康状态与证候之间、症状与证候之间、证候与证候之间的关系

就可进行中医辨证的亚健康状态而言，其证候与有明确病因病机的疾病状态的同一证候之间可能有着程度的不同，同一证候中症状的轻重程度也可能不同，各个症状对于证候的贡献率也可能存在很大差异。因此，如何通过明确症状对证候的贡献率，来界定亚健康状态或者疾病状态，并为今后的疗效评价提供客观依据是研究人员需要解决的问题。

有学者采用判别分析及 Logistic回归分析，利用标准偏回归系数绝对值的大小来判断各自变量对发病影响的重要性，对慢性乙型肝炎肝郁脾虚证的主要症状进行筛选，观察 2种方法产生的对证型判断具有决定意义的症状异同点，实现对慢性乙型肝炎肝郁脾虚证诊断要点的辨识[9]。该方法可应用于亚健康状态症状与证型之间判别函数的建立。

贝叶斯网络技术可通过数据分析，自动创建贝叶斯网络，发现变量间的因果关系，利用概率定量表示这些因果关系的强度，并进行诊断预测。有研究利用该技术发现了血瘀证的 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%，提示：贝叶斯网络技术适合于解决中医定量诊断问题，可揭示众多症状间以及症状与证候间的复杂关系，且建立的证候诊断模型能够以概率的形式给出诊断结果[10]，亦可应用于亚健康状态的定量诊断研究。

作为决策树算法的一种改进的随机森林，是合并了多个决策树后形成的组合判别模型，能够给出有效的错判率估计、分类器强度、相关性和变量的重要性等指标，据此来进行有效变量的挑选，达到有效降维的目的，目前已广泛应用于基因表达谱研究中。有研究显示：利用随机森林选择的肿瘤特征基因包含更多分类信息，分类准确率更高[11]。因此，可使用随机森林来提取关键症状或指标来进行亚健康状态的判断。

以上为数据挖掘技术在亚健康状态宏观资料分析中的应用，但宏观资料分析仅可阐述亚健康的总体规律，如危险因素、证型分布、症状分布、演变规律等，对于亚健康状态时机体的生物学变化，尚缺乏进一步的探索和深入挖掘。而随着复杂科学在中医证候研究中的应用，其微观分析和宏观综合相结合的研究方法，则更注重揭示客观事物的构成原因及其

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

演变过程，为亚健康状态的综合诊断标准研究提供了技术支持。

3　亚健康状态及其证候与相关性指标之间的关系

各种社会、心理应激因素所致的亚健康状态可导致机体的一系列生物学改变，包括：神经 —内分泌功能、免疫功能、脑功能、体液指标、基因蛋白质表达等方面。故可通过探索神经、内分泌、免疫方面微观指标的变化规律、指标间的关联、与亚健康状态的关系、与证候演变的关系，来研究亚健康状态与中医证候的内在关联以及生物学意义。这些指标包括：免疫球蛋白（IgA、IgG）、β内啡肽（β-EP）、皮质醇（Cor）、睾酮（T）、促肾上腺皮质激素（ACTH）、T细胞亚群（$CD3^+$、$CD4^+$）和去甲肾上腺素（NE）、多巴胺（DA）、5羟色胺（5-HT）及其代谢产物等。因此，需要尝试运用合适的数据挖掘方法来分析亚健康状态及其证候与相关性指标之间的关系。

典型相关分析是研究两因素集团之间相关关系的一种多元统计分析方法。它借助主成分分析降维的思想，分别对两组变量提取主成分，且使从两组变量提取的主成分之间的相关程度达到最大，而从同一组内部提取的各主成分之间不相关，用从两组之间分别提取的主成分的相关性来描述两组变量的整体的线性关系。运用典型相关分析将证候变量看作一组变量，相关性指标也看作一组变量，不必根据患者的证候信息进行辨证，消除了证型判断的主观性对结果的影响，其分析结果将是一定的证候信息组合与一定的相关性指标组合具有相关性[12]。

关联规则可揭示数据之间有意义的关联或者相互联系，且对于处理稀疏和弱相关的数据效率较高。用关联规则分析方法发现亚健康状态与相关性指标之间的复杂关联，通过支持度、可信度及列联系数、作用度指标等对规则的重要性、变量值之间的相关性进行评价，以关联规则作为亚健康状态的初筛工具，为开展亚健康状态与相关性指标间的关系研究提供新的方法学思路。

复杂系统熵聚堆可通过计算每两个四诊变量之间、每个四诊变量与其他变量之间的关联度系数，得出最常见的四诊信息组合，进而分析证候属性，作为证候要素存在的依据，并基于搜集到的信息，运用算法得出既有中医症状、体征，又有微观指标的相关模式。因此，可用作联系宏观临床症状、体征与微观指标的一个桥梁[13]。

叶氏等使用决策树建立慢性乙型肝炎肝胆湿热和肝郁脾虚证候的诊断模型，通过挖掘发现，16项症状体征以及 3项实验室指标能够建立上述证候的诊断模型，准确性达到 70%以上[14]。该方法可作为建立亚健康状态中医证候诊断模型的探索性方法。

但对于亚健康状态相关微观指标的分析，不应仅仅停留在指标本身改变有无统计学意义这个层面，而应从中医整体观的角度，来分析微观层面上的改变与机体整体功能失调之间的关系。此外，亚健康状态在微观层面上涉及多个系统的多种微观指标，这些指标之间又相互影响，因此，需要充分利用多种数据挖掘技术，较好地解决各种混杂因素、交互作用、以及一定程度上的共线性等问题，构建证候与微观指标相互关系的模型，对指标的相互作用关系以及表现出的宏观功能特征进行分析，把宏观资料和微观指标统一起来分析，对亚健康状态的证候机理进行深入研究。

课题组采用 SPSS中 Binary Logistic回归模型之后退剔除式 backward: conditional 探索性的分析一系列生化指标（IgA、IgG、β-EP、Cor、T、ACTH、$CD3^+$）、闪光融合频率值与健康人群和亚健康人群、以及与亚健康人群各证候的关系，获得各证候下所建模型对于该证候的预测概率，准确率最高的为预测心火证的模型，对建模数据总的回判正确率为 100.0%，肝气虚证和胃火证的预测模型正确率次之，均为 98.2%；湿证、心气虚证、肺气虚证和脾气虚证的预测模型正确率在 70% ～90%间，而肝火证和肝郁证的预测模型正确率最低。将亚健康人群的生化指标与宏观证候统一起来进行数据分析，尽管由于样本量较小（亚健康人群 57人，健康人群 33人），会对统计分析结果造成一定的影响，但可以预见，随着样本量的不断扩大，应用合适的数据挖掘技术，将有可能摸索出亚健康状态及其证候与相关性指标之间的关系。

综上，数据挖掘技术强有力的知识发现功能，使我们有理由相信，选择合适的数据挖掘技术，扬长避短，相互结合，将会推动亚健康状态在上述 3个方面的深入研究，探索并确立亚健康状态及其中医证候的诊断标准。

参考文献：

[1] 孙宪民，任平.关于亚健康若干问题的思考[J].中国误诊学杂志，2002，2(8)：1255－1256.

[2] 薛飞飞，陈家旭.数据挖掘在中医诊断学中的应用[J].中医杂志，2009，50(3)：200－202.

（下转第 627页）

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[25] KENEMAN S A, BORDOGNA J, ZEMEL J N. Evaporated films of arsenic trisulfide: physical model of effects of light-exposure and heat cycling[J]. J Appl Phys, 1978, 49(9) : 4663－4673.

[26] TRENTELMAN K, STODULSKI L, PAVLOSKY M. Characterization of pararealgar and other light-induced transformation products from realgar by Roman microspectroscopy[J]. Anal Chem, 1996, 68(10) : 1755－1761.

[27] BINDI L, POPOVA V, BONAZZI P. Uzonite As4S5, from the type-locality: X-ray single-crystal study and lighting experiments[J]. Can Mineral, 2003, 41(6) : 1463－1468.

[28] KYONO A, KIMATA M, HATTA T. Light-induced degradation dynamics in realgar: in situ structural investigation using single-crystal X-ray diffraction study and X-ray photoelectron spectroscopy[J]. Am Mineral, 2005, 90(10) : 1563－1570.

[29] 王金华,叶祖光,梁爱华,等. 安宫牛黄丸中砷、汞在正常和脑缺血模型大鼠体内的吸收与分布研究[J]. 中国中药杂志, 2003, 28(7) : 639－642.

[30] 张伟,余伯阳,寇俊萍,等. 雄黄活性物质的毒效相关性初步研究[J]. 中国天然药物, 2004, 2(2) : 123－125.

[31] 温磊,楼雅卿,江滨,等. 四硫化四砷动物药动学研究[J]. 中国药学杂志, 2006, 41(8) : 619－623.

[32] LU D P, QIU J Y, JIANG B, et al. Tetra-arsenic tetrasulfide for the treatment of acute promyelocytic leukemia[J]. A pilot report, 2002, 99(9) : 3136－3143.

[33] 郝红缨,腾智平,陆道培. 四硫化四砷对急性早幼粒细胞白血病细胞株 NB4的凋亡作用[J]. 中国药理学与毒理学杂志, 2002, 16(1) : 37－40.

[34] 陈思宇,刘陕西,李倍民. 雄黄对急性早幼粒细胞白血病细胞诱导凋亡和促进分化的双重作用[J]. 西安交通大学学报:医学版, 2002, 23(4) : 401－404.

(收稿日期: 2010-03-12)

(上接第 587页)

[3] 赵晖,陈家旭. 亚健康若干问题思考[J]. 山东中医杂志, 2008, 27(9) : 583－584.

[4] 车国海,方思行. 基于支持向量机的脾虚证多证型分类方法[J]. 计算机工程与应用. 2005, 21: 219－221.

[5] 黄英辉,李立奇,罗万春. 支持向量机在临床疾病诊断中的应用[J]. 数学的实践与认识, 2008, 38(23) : 101－103.

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

[7] 王天芳,张连文,赵燕,等. 隐结构模型及其在中医证候研究中的应用[J]. 北京中医药大学学报, 2009, 32(8) : 519－526.

[8] 朱文锋,何军锋,晏峻峰,等. 确定证素辨证权值的“双层频权剪叉”算法[J]. 中西医结合学报学报, 2007, 5(6) : 607－611.

[9] 张晓东,凌其华,聂红明,等. 慢性乙型肝炎肝郁脾虚证的 logistic回归及判别分析[J]. 中国中医药信息杂志, 2009, 16(9) : 23－24.

[10] 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.

[11] 李建更,高志坤. 随机森林:一种重要的肿瘤特征基因选择法[J]. 生物物理学报, 2009, 25(1) : 51－56.

[12] 龚燕冰,倪青,王永炎. 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[J]. 北京中医药大学学报, 2006, 29(12) : 797－801.

[13] 赵慧辉,陈建新,王伟,等. 基于复杂系统熵聚堆算法的不稳定性心绞痛宏观与微观指标相关性研究[J]. 北京生物医学工程, 2008, 27(5) : 462－465.

[14] 李梢,张宁波,李志红,等. 慢性乙型肝炎患者肝胆湿热证和肝郁脾虚证的决策树诊断模型初探[J]. 中国中西医结合杂志, 2009, 29(11) : 993－996.

(收稿日期: 2010-01-16)

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net