

DOI: 10.13288/j.11-2166/r.2016.05.021

综 述

近 5 年中医证候诊断客观化研究述评

徐玮斐, 刘国萍*, 王忆勤, 燕海霞, 郭睿

(上海中医药大学基础医学院, 上海市蔡伦路 1200 号, 201203)

摘要 多学科交叉融合的深入以及多种数据挖掘方法的广泛应用, 为中医证候诊断的客观化研究提供了有益的借鉴。围绕中医证候的兼挟性、不平衡性、动态性、复杂性特点, 梳理近 5 年来分别应用于其中的多标记学习法、不平衡数据集的处理、拓扑结构及前瞻性队列研究思路、复杂网络及从人脑认知等中医证候诊断客观化研究方法, 以为中医证候客观化诊断的深入研究提供参考。发现目前相关研究仍存在数据资料的采集和证候的判定主观性较大、针对临床证候的不平衡性及动态性研究相对偏少、尚缺乏综合可行的研究方法等诸多问题, 建议中医证候诊断的客观化研究应根据研究目的、中医证候特点、数据挖掘方法的使用范围等因素综合考虑。

关键词 证候诊断; 客观化诊断; 多标记学习法; 复杂性

辨证论治是中医认识疾病和解决疾病的过程, 是中医临床医学的灵魂所在^[1]。长期的医疗实践证明, 辨证论治具有其科学性、优越性和必要性^[2]。中医证候基础研究的主要目的在于提高辨证论治水平, 进而提高临床疗效^[3]。近年来有关中医学理论、临床诊疗及实验等研究取得了丰硕成果, 体现出中医学的继承与发展^[4]; 同时非线性科学、人工神经网络、计算数学、模糊数学、信息工程、医学工程、复杂性科学等多种学科交叉融合的发展, 为证候诊断的客观化研究提供了许多可行的方法。本文从中医证候的兼挟性、不平衡性、复杂性等特点切入, 梳理近 5 年来应用于中医证候诊断研究的方法如下。

1 证候的兼挟性及其客观化研究

临床就诊患者的证候分布往往具有兼挟性, 表现为单一证候少见、复合证候形式多见的特征。黄燕等^[5]通过对 4077 例缺血性中风患者中医证候要素变化的研究, 发现临床上该病中医证候组合分布以 2 证和 3 证组合最为常见。王娟等^[6]对 630 例慢性心力衰竭患者的中医证候分布规律进行研究, 发现 3 证组合方式最为常见, 其次依次是 4 证、2 证及 5 证组合, 而单证和 6 证组合的方式相对较少。

贺大强等^[7]对 1000 例慢性萎缩性胃炎患者证候兼挟情况进行研究, 结果显示, 该病的证候以复杂证候群的形式存在, 其中同时有 3 个或 3 个以上证的患者所占最多, 达 98.4%。可见临床疾病的证候分布往往单一证候少见, 两个或两个以上证候兼挟多见。针对临床证候的这种特性, 有些学者将多标记学习的方法引入到中医证候研究之中。

多标记学习概念是由 Schapire 和 Singer 提出的一种基于集成学习的方法, 用于解决文档分类时遇到的歧义性问题^[8], 此后该方法引起了众多学者的关注。现实生活中许多问题都属于多标记分类的问题, 如一个场景可以用不同的标签进行标记、文本的分类中每个文档可能与多个预先定义的主题相符、在生物信息学中每个基因可能与多个功能相关等, 且多标记学习也正被广泛应用于场景分类、文本分类、生物信息学等多个领域^[9-11]。在国内, Zhang 等^[12]首先将国外的多标记学习方法引入到酵母基因的功能分析、自然场景分类和网页自动分类的研究中, 并取得较好的分类效果。

这种一个示例同时属于多个不同类别的情况与中医临床同一个患者具有多种证候兼挟的特点相似, 因此, 也有学者将该方法引用到中医临床诊断数据资料的分析之中。如 Liu 等^[13-14]将多标记学习的方法应用于慢性胃炎及冠心病证候模型的构建, 提高了证候的识别率; 邵欢等^[15]提出一种混

基金项目: 国家自然科学基金(81270050, 30901897)

* 通讯作者: 13564133728@163.com

合优化的特征选择算法用于冠心病问诊数据建模,结果该模型分类效果较当前流行的多种降维算法的分类器有明显优势;任晋涛^[14]提出一种多标记 REAL 学习框架并应用于中医冠心病问诊数据中,其分类判断准确率达到 87.8%。同时上述方法都有别于单一多标记学习,充分考虑了原始数据中特征和标记之间的关联性,更加适合中医临床数据资料的研究。据此可以认为,多标记学习方法为具有兼挟性的中医临床证候的研究提供了一种有效可行的研究思路。

2 证候分布的不平衡及其客观化研究

中医临床中的证候除了兼挟性之外,还存在着某些疾病临床证候分布不平衡的特点,许多常见疾病的证候分布往往存在某些证候较为常见而某些证候偏少的情况。黄艳青^[17]对岭南地区 260 例小儿肺炎患者证候分布进行统计,结果显示,入院当日最常见的为痰热闭肺证占 81.5%,其次为风热闭肺证占 10.0%、痰浊闭肺证占 6.5%。谭倩^[18]通过对 1565 例 2 型糖尿病及其并发症患者的中医证候构成的研究,发现前 3 位依次是气阴两虚证占 48.95%、脾肾亏虚证占 6.20%、肝肾阴虚证占 4.35%。丁成华等^[19]对慢性萎缩性胃炎相关的 122 篇文献整理分析,发现脾胃虚弱证和肝胃不和证的出现频率明显高于其他证型,此二者累积频率达 61.71%,其余证型的出现频率为 38.3%。以上研究均提示,临床上不同疾病的证候分布具有显著的不平衡性,而这种分布不平衡性影响了分类的效果,不适合以准确率评价从不平衡数据集中得到模型分类效果,这也是临床证候客观化研究中亟需解决的问题之一。

具有不平衡类分布的数据集在许多实际应用中都会见到,如信用卡欺诈检测等,尽管不常出现,但稀有类的正确分类比多数类的正确分类更有价值^[20]。对此类不平衡数据集的处理,国外很多研究者都介绍过合适的研究方法^[21-23],而国内许多从事计算机专业的学者也对此类数据集的处理提出研究方法。徐燕等^[24]从特征选择的角度提出了解决不平衡数据集问题的特征选择方法,明显改进了稀有类别文本的分类效果。尤鸣宇等^[25]提出一种解决不平衡问题的特征选择方法,通过提高小类分布在信息熵计算中的权重,提高了小类的分类准确率;此外,该研究团队还提出了一种多类标的非对称偏最小二乘分类算法,并将该算法应用于张云鹏老中医治疗肝病数据集的研究中,结果发现,该算法对平衡各个类别有很好的效果^[26]。

对不平衡数据集的处理已经进行过相关研究,提供了许多可行的思路和方法,但将之运用于中医临床证候不平衡性的研究仍然相对较少,而中医临床不常见的证候也具有不可忽视的诊断价值和意义。所以在临床证候的研究中引入处理不平衡数据集的方法值得进一步推广与借鉴。

3 证候的动态性及其客观化研究

《伤寒论》云“观其脉证,知犯何逆,随证治之。”“证”具有时空性,始终处于动态变化之中^[27],同时证候的演变呈现出渐进性,而且是连续不断的。许多研究者运用统计学方法,通过跟踪一定数量的研究对象,揭示临床上某一疾病在不同阶段的证候变化规律。周扬等^[28]对 86 例肝硬化患者进行 1 至 3 年随访,研究了该病的临床证候演变规律。李四维^[29]分析观察 1333 例冠心病稳定期患者的证候要素变化情况,发现中医证候随着随访时间的延长不断发生着演变。何刚等^[30]对 108 例中风病住院患者进行辨证量化诊断调查,其结果证实,不同阶段证候分布规律及动态变化存在明显差异。

王永炎将证候的动态性概括为“动态时空”特征^[31],中医临床辨证施治的灵活性也正是借此体现的,即随着时间的迁移证候发生了变化,诊断和治疗也随之而变。对于证候动态性研究,孙喜灵等^[32]提出一种基于证候理论内蕴拓扑结构的理论和方法,并应用于脾气虚证候,较好地展现了脾气虚证候动态演化的全貌;另外,该课题组还提出这种拓扑结构可以用于阐明证候动态演化过程中子集合的衍生规律^[33]和用于中医证候结构表征的研究等^[3],这种拓扑结构的理论和方法为掌握中医证候的动态演变规律提供了新思路。此外,牟新等^[34]提出引入前瞻性队列研究思路进行证候学演变规律的研究,可以动态观察同一患病人群在发病后不同时间点的证候特征,对了解证候的形成、演变与转归具有非常重要的价值。但是这种方法实施难度大,耗时较长。

4 证候的复杂性及其客观化研究

复杂性用于描述混沌性的局部与整体之间的非线性关系,中医证候中也存在着这种复杂的非线性关系。中医证候诊断依据的是临床数量庞杂、相互关系的四诊信息,这些信息构成了中医证候的高阶、多维、非线性的特点。证候的高阶性体现在反映证候的各种相关因素之间的复杂关系;证候的高维性体现在证候诊断参考的四诊信息、生化指标和影像学报告等资料的丰富性,证候要素判断所考虑疾病的病因、病位、病性的特点及邪正关系、脏腑、

气血、阴阳等情况的差异性及可供选择的八纲、六经、三焦、卫气营血等辨证方法的多样性^[63]；证候的高维性和高阶性共同构成了中医证候的非线性，这与复杂性的非线性本质相似。随着复杂性科学方法研究的兴起与深入，许多与复杂性科学相关的方法也被引入到中医证候诊断的研究之中。

复杂性科学的主要研究任务是尚处于萌芽阶段的复杂系统理论，而复杂网络构成了复杂系统的基本框架。复杂网络由节点和节点间的连线组成，用节点表示系统中的各个元素，用连线表示元素间的相互作用^[64]。经由复杂网络可以描述现实世界中大量的复杂系统，这为中医证候的复杂性研究提供了一定的方法与思路，同时还可用于解释中医理论的整体性、非线性和动态性的基本特点。史琦等^[65]运用复杂网络的方法挖掘出用于诊断疾病证候的核心四诊信息，并分析出疾病的基本证候类型及其之间的相关关系及关联程度。杨铭等^[66]在信息熵理论及复杂网络方法的基础上提出了复杂系统熵网络方法，可用于分析处理多维混杂交互的中医肿瘤临床相关数据。孙喜灵等^[67]提出了一种基于复杂网络的证候拓扑结构的理论和方法，用于破解证候的非线性结构，并能展现证候复杂多变的全貌。

除了以上针对中医证候复杂性的复杂网络研究方法之外，还有学者从人脑认知角度对中医复杂性进行研究。中医证候的客观化诊断可以通过机器学习模拟人脑认知实现，而更好地实现模拟人脑认知，对于提高客观化辨证的准确率有着重要意义。认知是人脑对客观事物的反映，是人脑意识的主动形成部分。神经学家发现，人的大脑是通过在同一个脉冲反复刺激下改变神经元之间的神经键连接强度来进行学习，人工神经网络就是由一组相互连接的结点和有向链构成的类似人脑结构，通过训练这种模型可解决分类的问题^[68]。目前人工神经网络方法已广泛应用于中医证候诊断的客观化研究之中，并取得了较好的效果。西广成^[69]建立了证候的抽象神经自动机模型，实现了模拟人脑对中医四诊信息的分类，直观展示了证候之间的区别。同时还考虑到现有神经网络都是单独应用，在实际辨证中并不是所有症状都起着同样的作用，同时将所有症状平等训练要花费大量时间，且在样本少的情况下不能得到正确的网络模型，进而提出了实现中医辨证论治过程的多层神经网络模型，大大提高了网络的效率。同样是模拟人脑认知的研究方法，Liu 等^[70]提出运用深度学习算法模拟中医辨证中的大脑深度思维模式，用于处理中医辨证过程中的复

杂性及非线性，并将深度学习方法应用于慢性胃炎诊断模型的构建，与许多方法相比具有明显优势。

5 讨论

随着多学科交叉融合的深入，应用于中医证候研究的方法层出，算法各异，无疑促进了中医证候的客观化研究。针对临床证候具有的不同特点，不同研究者也各有侧重，有的关注于解决临床证候的兼挟性，有的侧重于探索疾病证候的动态演变规律，有的致力于实现高维高阶证候的降维降阶，有的尝试模拟人脑思维实现中医的辨证等，这些都为中医证候诊断研究的客观化提供了许多值得借鉴、有效可行的思路和方法。但是现阶段证候诊断客观化的研究仍然存在诸多不足：其一，数据资料的采集和证候的判定仍然存在主观性较大的问题。临床上四诊信息的采集，特别是问诊信息的采集，可能存在信息遗漏或者患者描述不清的情况，而证候分类往往通过专家辨证得到，故证候分类不易统一且不可避免地具有主观性。其二，中医临床数据的不平衡特征较为显著，但针对临床证候的不平衡性研究仍相对偏少，对相应的研究方法仍有待进一步挖掘。其三，对于中医临床证候动态性的研究也相对较少，而更好地掌握临床证候的动态演变规律对指导治疗疾病有着重要意义。其四，目前方法大都只是针对证候的某类特性进行的研究，存在着“自限性”，更加适合某一类型的数据处理，尚缺乏综合可行的研究方法。

综上，中医证候诊断的客观化研究尚有许多方法可循，但现行方法兼顾不足，对证候诊断的研究仍处于探索性阶段。今后在现有中医学、统计学、系统科学和信息科学等多种学科广泛合作的基础上，中医证候诊断的研究尚需进一步创新。

参考文献

- [1] 邓铁涛. 辨证论治是中医临床医学的灵魂 [J]. 中医药理学, 2002, 20(4): 394-395.
- [2] 朱文锋. 创立统一的辨证方法与体系 [J]. 湖南中医药导报, 2003, 9(1): 7-10.
- [3] 孙喜灵, 郑秋生, 王振华, 等. 关于中医证候动态演化规律的关键科学问题研究 [J]. 世界科学技术-中医药现代化, 2014, 16(9): 2042-2046.
- [4] 张震. 辨证论治新理念与临床应用 [M]. 上海: 上海科学技术出版社, 2014: 2-5.
- [5] 黄燕, 赵敏, 蔡业峰, 等. 4077 例缺血中风患者中医证候要素分布特点及规律 [J]. 中国中医基础医学杂志, 2011, 17(7): 759-761.
- [6] 王娟, 陈婵, 张鹏, 等. 630 例慢性心衰患者中医证候分布规律研究 [J]. 北京中医药大学学报, 2013, 36(8):

- 567-571.
- [7] 贺大强,孙喜灵,王斌胜,等. 慢性萎缩性胃炎中医复杂证候群及其分布规律 [J]. 辽宁中医杂志,2014,41(2):205-207.
- [8] SCHAPIRE RE, SINGER Y. Boostexter: a boosting-based system for text categorization [J]. Machine Learning, 2000,39(2):135-168.
- [9] BOUTELL MR, LUO J, SHEN X, et al. Learning multi-label scene classification [J]. Patt Recog,2004,37(9):1757-1771.
- [10] ELISSEEFF A, WESTON J. A kernel method for multi-labeled classification [C] // Dietferich TG, Becke RS, Ghahramani Z. Advances in Neural Information Processing Systems 14. Cambridge MA: MIT Press,2002:681-687.
- [11] 张敏灵. 一种新型多标记懒惰学习算法 [J]. 计算机研究与发展,2012,49(11):2271-2282.
- [12] ZHANG ML, ZHOU ZH. MI-kNN: a lazy learning approach to multi-label learning [J]. Patt Recog,2007,40(7):2038-2048.
- [13] LIU GP, LI GZ, WANG YQ. Modelling of inquiry diagnosis for coronary heart disease in TCM by using multi-label learning [J]. BMC Complement Altern Med,2010,10:37. doi: 10.1186/1472-6882-10-37.
- [14] LIU GP, YAN JJ, WANG YQ. Application of Multi-label Learning Using the Relevant Feature for Each Label (REAL) algorithm in the diagnosis of chronic gastritis [J]. Evid Based Complement Alternat Med, 2012, 2012:135387. doi: 10.1155/2012/135387.
- [15] 邵欢,李国正,刘国萍,等. 多标记中医问诊数据的症状选择 [J]. 中国科学: 信息科学,2011,41(11):1372-1387.
- [16] 任晋涛. 基于多标记学习的中医问诊系统的研究 [D]. 上海: 华东理工大学,2012.
- [17] 黄艳青. 岭南地区小儿肺炎证候演变规律及其与体质关系的研究 [D]. 广州: 广州中医药大学,2011.
- [18] 谭倩. 1565 例 2 型糖尿病患者中医病证分布规律研究 [D]. 北京: 北京中医药大学,2010.
- [19] 丁成华,李晶晶,方芳,等. 慢性萎缩性胃炎中医病机与证候分布规律研究 [J]. 中华中医药杂志,2011,26(3):582-586.
- [20] PANG-NING T, MICHAEL S, VIPIN K. 数据挖掘导论 [M]. 范明,范宏建,译. 北京: 人民邮电出版社,2011:180.
- [21] KUBAT M, MATWIN S. Addressing the curse of imbalanced training sets: One-sided selection [C] // Proceedings of the 14th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 1997: 179-186.
- [22] JAPKOWICZ N. The class imbalance problem: significance and strategies [C] // Las Vegas. Proceedings of the 2000 International Conference on Artificial Intelligence: special track on inductive learning,2000:111-117.
- [23] DRUMMOND C, HOLTE RC. C4.5, Class imbalance, and Cost sensitivity: Why under-sampling beats over-sampling [C] // Washington DC: In ICML'2004 Workshop on Learning from Imbalanced Data Sets II, 2003:1-8.
- [24] 徐燕,李锦涛,王斌,等. 不均衡数据集上文本分类的特征选择研究 [J]. 计算机研究与发展,2006,43(增刊):58-62.
- [25] 尤鸣宇,陈燕,李国正. 不均衡问题中的特征选择新算法: Im-IG [J]. 山东大学学报(工学版),2010,40(5):123-128.
- [26] YOU MY, ZHAO RW, LI GZ, et al. MAPLSC: a novel multi-class classifier for medical diagnosis [J]. Int J Data Min Bioinform,2011,5(4):383-401.
- [27] 刘保延,王永炎. 证候、证、症的概念及其关系的研究 [J]. 中医杂志,2007,48(4):293-296,298.
- [28] 周扬,胡鑫才,张华,等. 86 例肝硬化随访患者中医证候演变研究 [J]. 世界中医药,2014,9(5):545-548.
- [29] 李四维. 冠心病稳定期患者中医证候演变规律研究 [D]. 北京: 北京中医药大学,2011.
- [30] 何刚,冯子驹,卢静,等. 中风病证候分布及动态变化研究与探讨 [J]. 中国中医急症,2007,16(6):686-687,702.
- [31] 郭蕾,王永炎,张志斌. 关于证候概念的诠释 [J]. 北京中医药大学学报,2003,26(2):5-8.
- [32] 孙喜灵,张晓林,刘琳,等. 中医学证候理论内蕴的拓扑结构研究 [J]. 山东中医药大学学报,2010,34(5):383-388.
- [33] 孙喜灵,姜伟炜,张晓林,等. 中医证候的结构化研究 [J]. 世界中医药,2013,8(2):146-148.
- [34] 牟新,周迪夷,周旦阳,等. 采用前瞻性队列研究在观察中医证候演变规律中的方法学探讨 [J]. 中华中医药杂志,2014,29(2):514-517.
- [35] 李梢. 从维度与阶度探讨中医证候的特征及标准化方法 [J]. 北京中医药大学学报,2003,26(3):1-3.
- [36] 刘建香. 复杂网络及其在国内研究进展的综述 [J]. 系统科学学报,2009,17(4):31-37.
- [37] 史琦,陈建新,赵慧辉,等. 基于复杂网络技术的冠心病患者四诊信息分布模式的研究 [J]. 北京中医药大学学报,2012,35(3):183-188.
- [38] 杨铭,焦丽静,陈佩奇,等. 复杂系统熵网络方法及其在中医肿瘤临床数据挖掘中的应用 [J]. 世界科学技术—中医药现代化,2012,14(2):1376-1383.
- [39] 孙喜灵,姜伟炜,张晓林,等. 中医证候动态演化规律研究与证候判定诊断标准科学内涵的阐释 [J]. 中国医药导报,2012,9(25):127-129.
- [40] 西广成. 复杂系统方法学与中医证候建模 [M]. 北京: 科学出版社,2010:173.
- [41] LIU GP, YAN JJ, WANG YQ. Deep Learning Based Syndrome Diagnosis of Chronic Gastritis [J]. Comput Math Methods Med,2014,938350. doi: 10.1155/2014/938350.

(收稿日期: 2015-07-10; 修回日期: 2015-09-09)

编辑: 崔京艳