数据挖掘技术在医案研究中的应用与讨论

刘兴方，韩学杰*
（中国中医科学院中医临床基础医学研究所，北京 100700）

［摘要］ 探讨数据挖掘技术在医案研究中的应用。通过文献分析整理指出数据挖掘技术在医安研究中的常见目的及方法。医案研究是中医传承的重要手段，数据挖掘技术则是医案研究的重要工具，随着医案数据来源和挖掘目的的变化，挖掘方法也不断增多更新。人机结合才能保证挖掘结果的可靠性，因此对在世名老中医的医案深入挖掘显得尤为迫切。基于数据挖掘技术的名医传承研究平台不断建立，应共享推广，整合优化，促进中医药的学术进步。

［关键词］ 数据挖掘；医案；综述；讨论
［中图分类号］ R287 ［文献标识码］ A ［文章编号］ 1005-9903(2014)09-0247-04
［doi］ 10.13422/j.cnki.syfix.2014090247

Application and Discussion of Data Mining Technology in the Research of in Medical Case

LIU Xing-fang, *HAN Xue-jie**
(*Institute of Basic Research in Clinical Medicine*, *China Academy of Chinese Medical Sciences*, *Beijing* 100700, *China*)

［**Abstract**］ To revievo the application of data mining technology in the research of medical case. Medical case study is an important means of traditional Chinese medicine (TCM). Data mining technology is an important tool in medical case research. Along with the change of medical data sources and mining purposes, mining method is increasing constantly update. Only the man-machine combination can guarantee the reliability of mining results. So the medical case of live and famous specialist of TCM further mining is particularly urgent. Inheritance research platform to constantly build based on data mining technology should be shared, and promotion, integration and optimization should be applied, to promote the TCM academic progress.

［**Key words**］ data mining; medical case; review; discussion

医案是名老中医学术经验传承的重要载体，不仅蕴含了丰富的用药经验，还体现了名老中医对疾病的认知和诊疗策略。通过医案研究汲取名老中医的诊疗经验，是年轻中医大夫提高临床水平的一条捷径。然而中医医案汗牛充栋，百花齐放，体例不一，从信息科学角度来看，中医医案数据是混乱而又复杂的经验数据，这导致了医案研究的困难，亟需新技术和新方法的引进[1]。

在此背景下，运用数据挖掘技术和方法研究中医医案逐渐兴起，且取的了一定的成果。数据挖掘(DM，data mining)是从海量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的，但又是潜在有用的信息和知识的过程，能够进一步提取和挖掘名老中医医案中隐藏的精粹，反映和获取其隐涵的本质的知识，推动中医学术的传承与创新[2]。

1 数据挖掘的一般对象

在中医医案研究中，数据挖掘的对象多种多样，但多以单种疾病的治疗为主线，对单个医家医案[3-4]、

［收稿日期］ 20130603(014)
［基金项目］ 中国中医科学院“名医传承”项目(CM20122002)；第六批自主选题(Z0221)；第七批自主选题(Z0260)
［第一作者］ 刘兴方，硕士，从事中医标准规范共性技术研究，E-mail：liuxingfang@163.com
［通讯作者］ *韩学杰，博士生导师，研究员，从事中医标准规范共性技术研究，Tel：010-64014411-3312，E-mail：xuejiehan@126.com

地域流派医案[5]、专科医案[6]、古今医家医案[7]、特定时期医案[8]等进行研究。名医在其擅长治疗的疾病方面有着丰富宝贵的经验,同一地域或同一流派医家对疾病的认识和诊疗往往有相通之处,而对古今医家治疗某一疾病的医案文献进行对比挖掘研究,能够归纳、总结各历史时期疾病证治的演变规律及医家对疾病的证治规律和特点的总体认识,为疾病的治疗提供新的思路,促进临床水平的提高。

2 数据挖掘的常见目的

用药规律的分析总结是中医医案数据挖掘中最常见也是最重要的目的,通过挖掘一是总结老中医的用药经验,二是发现最佳配伍,常从用药频次[9-10]、治疗剂量[11]、功效治法[12]、药对组方[13]等方面入手,为疗效提高和新药研发[14]打下基础。医案数据挖掘还常用于疾病辨证规律的研究,如证类[15-16]证素[17-18]研究、症状分析[19]、病机研究[20]、病位分析[21]等。治法亦可通过数据挖掘方法来分析研究[22]。

3 数据挖掘的常用方法

中医医案蕴含的信息量大,涉及辨证、用药等多个方面及互相之间的联系,如理-法-方-药-症-证关系等,故应用的数据挖掘方法多种多样,主要有频数分析、关联规则、聚类分析、因子分析等,除根据不同研究目的采用不同的分析方法外,研究中也常综合使用多种方法深入挖掘。如黄利兴[23]以姚荷生医案中的咳嗽医案为对象,从医案的症状与中药出发,进行频数统计、症状因子分析、症状聚类分析、症状-中药关联分析、中药-中药关联分析、中药因子分析、中药聚类分析、症状因子结果-中药因子结果皮尔逊相关分析等处理,挖掘姚荷生医案中咳嗽的证治规律。

3.1 关联规则 是数据挖掘的重要方法之一,可以得到隐含于海量数据中具有潜在价值的有用信息[24]。中医医案总结中常用关联规则挖掘病因、病位、证候、四诊信息及组方配伍,并能为新药的开发提供理论支持[25],相关算法主要有 Apriori 算法[26]、FP-树频集算法[27]等。如李赛等[28]采用 Apriori 算法对 486 例病案中症状 = > 方剂,症状 = > 中药,中药 = > 症状,中药 = > 中药进行关联规则分析以总结聂莉芳治疗慢性肾衰竭患者的经验。李文林等[29]采用基于 FP-tree 的算法对证型-症状、症状-药物、证型-药物之间的关联规则进行了挖掘,发现挖掘出的大部分规则能得到合理的解释并具有一定的实际意义,但也有个别规则在理论是成立的,与实际情况有出入,需要在扩大样本量的基础上进行方法学方面的进一步验证。

3.2 聚类分析 是医案数据挖掘的主要任务之一,常用于分析医案中蕴含的组方配伍及常见证型等信息,在发现新处方方面具有独到之处[30]。它能够作为一个独立的工具获得数据的分布状况,也能够作为其他算法(如分类和定性归纳算法)的预处理步骤。李寿松[31]对 183 份丁书文治疗高血压病病例中 42 味高频中药进行基于功效的聚类分析,结合丁教授的临证经验,得到 5 个聚类方并总结出其常用治法。张菁[32]运用模糊聚类以及关联规则的数据挖掘技术分析和挖掘干祖望治疗耳鼻喉疾病医案中理-法-方-药之间的多重关系,孙明月等则[33]采用双聚类分析法,挖掘魏子孝治疗甲状腺疾病的用药、用量特点。

3.3 因子分析 主要目是用来描述隐藏在一组测量到的变量中的一些更基本的,但又无法直接测量到的隐性变量,因此常应用于证候学研究当中,尤其是对医案中证候要素的提取。金香兰等[34]对 428 例高血压病患者的中医四诊信息进行因子分析,得出 26 个公因子(证候要素),并对患者证候要素分布情况进行了分析。戴霞等[35]则以中医证型为切入点,运用系统聚类与因子分析相结合的方法挖掘近 10 年高血压病中医证型研究文献四诊资料隐含的客观规律,发现高血压病中医辨证可大致分为 5 个单证,并用因子分析提取出了每一证型的主要辨证依据,为高血压病证候规范化研究及建立证候表征体系提供依据和可行的研究思路。

3.4 Logistic 回归分析 主要在流行病学中应用较多,比较常用的情形是探索某疾病的危险因素,在中医医案研究中常用来对与证候相关的病因、症状以及用药等进行挖掘分析。张珊珊等[36]采用非条件 Logistic 多元逐步回归分析方法和判别分析的方法,对 987 例古籍医案和现代医案中治疗原发性高血压肝阳上亢证的方剂和药物进行分析,从而界定原发性高血压的方药表征。傅爽等[37]则采用非条件 Logistic 多元逐步回归方法从 305 例与原发性高血压相关的古今医案中筛选症状或体征变量,最终获得回归方程数学模型以用于原发性高血压肝阳上亢证的诊断。

3.5 粗糙集理论[38] 是处理模糊和不确定性知识的一种较新的数学工具,能够将医案中经验的描述和概况进行高层次的综合分析,从宏观的角度对医案中的临证经验进行规范化研究。它无需提供问题所需处理的数据集合之外的任何先验知识,因此能有效避免专家的主观经验,但对噪声较敏感。如傅

爽等[39]利用粗糙集技术的 Johnson 算法对 776 首治疗高血压病肝阳上亢证的中药处方中出现频率高的 53 种药物进行属性约简,进一步明确了高血压病肝阳上亢证的用药规律。

3.6 人工神经网络 特有的非线性适应性信息处理能力克服了传统人工智能方法对于直觉,如非结构化信息处理方面的缺陷,有较好的抑制噪声干扰的能力。因此在医案挖掘中将两者结合起来,根据粗集方法预处理后的信息结构构成神经网络信息处理系统,不但可以减少信息表达的属性数量,减小神经网络构成系统的复杂性,而且具有较强的容错及抗干扰能力,为处理医案中不确定、不完整信息提供了一条强有力的途径。如温宗良等[40]运用人工神经网络技术,采用成比例的共轭梯度算法从 79 例高血压病案样本中获取规则,建立高血压中医证候诊断模型,并将其运用于新病例的判别,具有较好的诊断、预测能力。

3.7 贝叶斯网络 是目前不确定知识表达和推理领域最有效的理论模型之一,是一种强有力的不确定性推理方法,能在有限的,不完整的,不确定的信息条件下进行学习和推理。如唐启盛等[41]运用贝叶斯网络模型对 611 例抑郁症患者的横断面证候进行数据研究,并结合前期聚类分析研究结果,拟定出抑郁症的 6 个中医证型,认为基于贝叶斯网络研究的中医证型具有一定客观性、科学性,较符合中医理论。

3.8 典型相关分析 利用综合变量对之间的相关关系来反映两组指标之间的整体相关性的多元统计分析方法,其条件是两组变量都是连续变量,其资料都必须服从多元正态分布。如张晶[42]对 743 例情志致病症分类医案中的情志因子与左尺脉象进行典型相关分析,总结出恐、烦、精神萎靡、郁、狂与左尺脉象的相关性,为情志相关脉诊临床实践提供了文献支持。

4 讨论

名老中医经验是提高我国卫生健康保障水平和发展中医药学术的重要支撑,如何提取和挖掘其医案中隐藏的精粹,反映和获取其隐涵的本质的知识,是当前亟需解决的重要研究课题[43]。

数据挖掘技术给医案研究者提供了一条有效捷径,而如何从医案中挖掘出契合临床实际的数据也给数据挖掘技术提出了挑战。医案数据挖掘的结果错杂,务必人机交互,多次反复校正,去粗取精、去伪存真后才能符合临床实际,所得的结果方能真正体现医家的诊疗策略。人机交互的最佳模式莫过于医生亲自对自己医案的数据挖掘结果进行筛选分析,即“以人为主”,因此对在世的名老中医的医案进行数据挖掘显得尤为迫切[44]。

“临床医生经验”就是循证医学三要素之一。名医经验是中医循证医学的重要证据,然而目前医案数据挖掘多针对于单个名医,由于诊治患者的数量有限,尤其对于某些单一疾病,样本量更小,对其经验的总结仅仅停留在个人的体会层面上[45],证据级别较低。若从病种角度对多位名家的医案进行挖掘,综合比较研究后所得的共性结果在制定临床指南时较单一名家的临床经验证据强度更高。而这需要从政府和行业学会角度统筹支持,有必要建立一个面向全国的开放的数据平台,从采集、录入、统计挖掘各流程确保数据质量,并提高数据的横向可比性,最终获得名老中医经验的共性规律[46],为中医临床指南制定提供即反映中医特色,又符合循证医学要求的高质量证据。也可对现有的应用较多的平台软件进行推广,如名老中医经验分析挖掘平台、中医传承辅助系统软件[47]等,并根据实践运用不断更新软件,提高科研成果转化率,避免资源浪费,推动中医药学术的发展。

[参考文献]

[1] 王佑华,陆金根,柳涛,等. 中医医案中的知识发现研究[J]. 中西医结合学报,2007,5(4):368.

[2] 王树鹏,刘书宇. 数据挖掘技术在中医药领域中的应用研究[J]. 中华中医药学刊,2011,29(1):36.

[3] 宋军,路志正. 路志正教授调理脾胃法治疗冠心病的用药规律研究[J]. 世界中西医结合杂志,2011,6(9):801.

[4] 宋军,路志正. 路志正教授调理脾胃治疗冠心病的临床证候学研究[J]. 中华中医药杂志,2010,25(12):2261.

[5] 晏婷婷,吴丽,王旭东. 基于数据挖掘的孟河医家治疗痹证的治法及用药规律研究[J]. 新中医,2012,44(9):98.

[6] 杨朝杰. 当代中医耳鼻喉科名家医案常见病用药规律数据挖掘研究[D]. 广州:广州中医药大学,2012.

[7] 王兵. 基于古今医案数据分析的水气病证治规律研究[D]. 哈尔滨:黑龙江中医药大学,2010.

[8] 张玉辉. 民国时期温病医案证候要素与应证组合规律研究[D]. 北京:中国中医科学院,2008.

[9] 张京春,谢元华,蒋跃绒,等. 陈可冀辨治冠心病医案证法方药的频数分析[J]. 中医杂志,2008,49(10):901.

[10] 程宾. 邓铁涛医案研究及交互式网站的构建[D]. 广州:广州中医药大学,2006.

[11] 金末淑. 基于数据挖掘的仝小林教授应用干姜黄芩黄连人参汤治疗 T2DM 用药规律研究[D]. 北京:北

京中医药大学,2012.
[12] 吴丽丽,周莺,严灿,等. 古代情志病证医案中组方用药规律分析[J]. 安徽中医学院学报,2008,27(1):25.
[13] 欧阳志强,蒋力生,王如意,等. 名中医牙痛医案63例中药配伍及方证对应规律分析[J]. 江西中医学院学报,2007,19(5):88.
[14] 李鑫颉. 基于无监督数据挖掘技术的中风病用药规律及处方发现研究[D]. 石家庄:河北医科大学,2012.
[15] 薛声波,王米渠,邓雪梅,等. 糖尿病医案证候的半定量聚类分析[J]. 辽宁中医杂志,2010,37(1):11.
[16] 陈素玲,付爽,高云,等. 基于粗糙集理论的原发性高血压肝阳上亢证辨证系统的建立[J]. 山东中医药大学学报,2010,34(3):201.
[17] 李敬华,蔡顺利,赵林冰,等. 中医医案的证素与检索研究[J]. 辽宁中医杂志,2012,39(3):439.
[18] 叶放,李国春,沈波,等. 基于周仲瑛教授大样本"瘀热"病案数据挖掘分析研究报告[J]. 中华中医药杂志,2012,27(5):1294.
[19] 王俊文,崔蒙,赵英凯,等. 基于成人哮喘复诊医案建立中医疗效判断模型[J]. 中国中医基础医学杂志,2012,18(3):324.
[20] 李万斌. "瘀血生风"假说检验[D]. 济南:山东中医药大学,2008.
[21] 于凌. 基于对《临证指南医案》的数据挖掘探讨不寐病位的相关问题[J]. 北京中医药大学学报,2012,31(9):682.
[22] 金海浩. 基于粗糙集重要度和因子载荷对"滋水涵木法"古代医案的数据分析[J]. 中国中医急症,2012,21(6):897.
[23] 黄利兴. 基于文本挖掘技术探索姚荷生咳嗽医案的证治规律[D]. 长沙:湖南中医药大学,2010.
[24] 王爱平,王占凤,陶嗣干,等. 数据挖掘中常用关联规则挖掘算法[J]. 计算机技术与发展,2010,(4):105.
[25] 黄颖琦,贾恒,何前松,等. 关联度最强药物配伍的中医止呕类方数据挖掘[J]. 中国实验方剂学杂志,2012,18(14):1.
[26] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]// Proceedings ofthe 1993 ACM SIGMOD Conference. Washington DC:[s. n.],1993:207.
[27] Han Jiawei, Pei Jian, Yin Yiwen, et al, Mining frequent patterns without candidate generation[J]. Data Mining and Knowledge Discovery,2004,8(1):53.
[28] 李赛,聂莉芳,孙红颖. 聂莉芳治疗慢性肾功能衰竭经验的关联规则分析[J]. 中华中医药杂志,2011,26(7):1602.
[29] 李文林,赵国平,陆建峰,等. 关联规则在名医临证经验分析挖掘中的应用[J]. 南京中医药大学学报,2008,24(1):21.
[30] 赵鑫,崔向宁. 基于中医传承辅助系统的治疗慢性心力衰竭方剂组方规律分析[J]. 中国实验方剂学杂志,2012,18(19):8.
[31] 李寿松. 基于数据挖掘的丁书文教授治疗高血压病用药分析及流行病学调查[D]. 济南:山东中医药大学,2012.
[32] 张菁. 基于模糊聚类-关联方法的干祖望耳鼻喉疾病医案分析挖掘研究[D]. 南京:南京中医药大学,2011.
[33] 孙明月,高蕊. 魏子孝治疗甲状腺疾病医案挖掘分析[J]. 中国中医药信息杂志,2012,19(2):25.
[34] 金香兰,张允岭,陈志刚,等. 运用因子分析探讨原发性高血压病证候要素[J]. 北京中医药大学学报,2011,34(2):131.
[35] 戴霞,姜婷,于杰,等. 基于现代文献的高血压病证候多元统计分析[J]. 中西医结合心脑血管病杂志,2009,7(11):1339.
[36] 张珊珊,贺立娟,李运伦. 原发性高血压中医历代医案数据库规范化建设探讨[J]. 山东中医药大学学报,2009,33(2):103.
[37] 傅爽,李运伦. 基于多元统计分析方法的原发性高血压肝阳上亢证判别模型的建立[J]. 山东中医药大学学报,2010,34(1):14.
[38] 张文修,吴伟业,梁吉业. 粗糙集理论及方法[M]. 北京:科学出版社,2001:78.
[39] 傅爽,陈素玲,李运伦. 基于数据挖掘的高血压病肝阳上亢证用药规律分析[J]. 中国中医基础医学杂志,2011,17(1):48.
[40] 温宗良,岳桂华,杨靖,等. 基于共轭梯度算法的 BP 神经网络在高血压证候诊断中的应用[J]. 山东中医药大学学报,2012,36(3):183.
[41] 唐启盛,曲淼,包祖晓,等. 抑郁症中医证候的贝叶斯网络研究[J]. 中医杂志,2008,49(11):1013.
[42] 张晶. 古代情志相关医案中情志因子与脉象的典型相关分析[J]. 四川中医,2011,29(7):26.
[43] 胡镜清,路洁,刘喜明,等. 名老中医经验传承研究内容与方法的思考[J]. 中华中医药杂志,2009,24(10):1346.
[44] 张华,刘保延,田从豁,等. "人机结合、以人为主"的名老中医经验整理研究方法[J]. 中医研究,2007,(2):4.
[45] 唐仕欢,杨洪军. 中医组方用药规律研究进展述评[J]. 中国实验方剂学杂志,2013,19(5):359.
[46] 王映辉,张润顺,周雪忠,等. 名老中医经验共性规律及个性差异比较研究[J]. 世界科学技术——中医药现代化,2009,11(6):793.
[47] 卢朋,李健,唐仕欢,等. 中医传承辅助系统软件开发与应用[J]. 中国实验方剂学杂志,2012,18(9):1.

[责任编辑 邹晓翠]