数据挖掘技术在中医药领域中的应用概况

高毅超[1],王 凡[1],郭 晶[2]
(1. 天津市南开区中医医院,天津,300102;
2. 天津市南开区鼓楼街社区卫生服务中心,天津,300102)

[关键词] 数据挖掘技术;中医药研究;综述,学术性
[中图分类号]R311.13 [文献标识码]A DOI:10.16808/j.cnki.issn1003-7705.2019.07.075

运用中医理论诊治疾病是一个复杂的过程,无论是诊病辨证,还是遣方用药,都包含了大量的信息。中药方剂充分展现了中医辨证施治的特点,在药物的配伍、加减,方证对应等方面也同样包含很多信息。通过对这些信息的研究,利用合适的方法,可以发现问题,总结经验。在计算机、互联网技术、云计算、移动终端、数据存储方式等高速发展下,大数据时代应运而生。它为我们的生活、工作、娱乐等多个方面带来了前所未有的变化,医学领域也不例外。近几年数据挖掘在中医药研究领域运用越来越广泛。

数据挖掘,又称数据库知识发现,是从大量的、不完全的、有噪声的、迷糊的、随机的实际应用数据中提取隐含的、人们事先未知的但又潜在有用的并且最终可理解的信息和知识的过程[1]。目前,数据挖掘涉及的方法很多,主要包括统计方法、机器学习方法、神经网络方法和数据库方法[2]。中医药研究领域引入数据挖掘技术,既解决了实际问题,又推动中医药客观化进程的发展。数据挖掘融合了计算机智能和数理统计等方法,在研究大量数据的基础上,模拟临床诊疗过程,分析症、证、药等之间的关系。在中医药研究领域的数据挖掘方法主要有:关联分析、聚类分析、因子分析、回归分析、贝叶斯网络、决策树、支持向量机、粗集理论、人工神经网络、数据库方法等。现将主要方法及应用综述如下。

1 统计方法

1.1 关联规则 关联规则是从大量的数据中挖掘发现项集之间有意义的关联、相关联系、因果结构以及项集的频繁模式等,并寻找给定数据集中项之间的有趣联系的算法。关联规则可以研究庞大数据库中数据集中项之间存在的隐藏关系,进而总结分析其特征和规律。然而,关联规则也有不足之处,比如计算量增长非常严重,使正确的数据容易被忽视。温丹婷[3]运用关联规则分析名老中医李丽芸治疗不孕症的用药规律,总结李丽芸治疗不孕症的学术思想。结果说明数据挖掘研究中心运用关联规则分析方法能够比较客观地呈现名老中医对病的诊治特色与经验,总结出其诊病的辨证思维模式并发现一些新知识。王昆阳[4]运用关联规则对焦虑抑郁共病的中医证候进行分析。结果说明焦虑抑郁共病的常见中医症状是担忧、心烦易怒、紧张、神疲乏力、善食易饥、入睡困难、多梦、兴趣索然等。进一步基于关联规则下和中医学理论相结合,提取出证候要素,即肝气郁、肝火旺、心气虚、心血虚心阴虚心阳虚、心火亢盛、脾气虚、脾阳虚、肾阴虚、肾阳虚、胆气虚、痰湿、血瘀。郑丹文等[5]整理72则当代名老中医治疗流行性感冒的医案,分析中药和病因、中药和证候、中药和症状的关联规律。有效地总结凝练了名老中医临证的用药规律,使理论与临床实际相联系,能有效指导临床。田瑾[6]对郭蓉娟教授治疗失眠症的3084首方剂进行关联规则分析,总结出其治疗失眠的常用药对21对,涉及药物12味。杨青[7]通过数据挖掘分析15年紧张型头痛的用药特点和配伍规律,关联规则分析后结果显示紧张型头痛的治疗以风药为主,临床上风药与风药,风药与白芍、活血化瘀药等常相须为用,说明川芎是治疗紧张型头痛的要药。

1.2 聚类分析 聚类是指将数据按照某些共同特征分到不同的类或者簇的过程。聚类可以观察每类或者一簇数据的特征,集中对特定的聚簇集合作进一步地分析。聚类分析的分类过程可以在没有先验知识的数据资料自动分类,是一种探索性的分析。任毅等[8]对当代名老中医治疗冠心病病案进行数据挖掘,采用频数分析和聚类分析后,发现当代名老中医治疗冠心病常用药对16个,3味药物组合体7个,多味药物组合体5个。孙庆亮[9]通过总结特发性肺纤维化和结缔组织病相关性间质性肺疾病患者的四诊信息,应用聚类分析出常见的中医症状和证型。特发性肺纤维化有4种证型:气阳两虚、痰瘀内结,痰瘀内结、兼有气虚,气阴两虚、痰热内蕴,气虚血瘀、痰湿内阻。结缔组织病相关性间质性肺疾病有3种证型:气阴两虚、痰热内蕴,气虚血瘀、痰浊内阻,痰热内蕴。丁冠福[10]对240例腹泻型肠易激综合征患者进行临床调查,将采集的四诊信息建立数据库。所

第一作者:高毅超,男,医师,2014级博士研究生,研究方向:临床流行病学与统计学方法及原理在中医内科学中的应用
通讯作者:王凡,女,主任医师,硕士研究生导师,研究方向:中医内科学,E-mail:wangfan@sina.cn

调查的患者常见症状、舌象和脉象经聚类分析后得出 4 个主要证型，即肝郁脾虚证、脾胃气虚证、脾肾阳虚证和脾胃湿热证。鉴于肝郁脾虚证、脾胃气虚证所占比例最多，故本病主要病机为肝郁和脾虚，治疗以肝脾为重。牟新等[11]把糖尿病肾病患者根据实验室检查结果分层调查，运用聚类分析研究出现频数较高的症状，得到 5 个证候类型。提示糖尿病肾病为本虚标实之证，以阳虚更为明显，逐渐发展为阴阳两虚，瘀血证贯穿病程始终。

1.3　因子分析　因子分析是通过研究多个变量间相关系数矩阵的关系，根据变量内部相关性将变量分组，使得同组内的变量之间相关性较高，而不同组间的变量相关性较低。每个分组后的变量都用一个可描述该组特点的公因子表示。因子分析中的因子通常是隐藏在一组测量到的变量中的，又无法直接测量到的隐性变量中。因此，因子分析属于多元统计中处理降维的一种统计方法，其目的就是要减少变量的个数，用少数因子代表多个原始变量。张琳婷[12]整理了来自全国六大地区共 1218 份《冠心病病证候要素、证候特征及证候病机演变规律专家问卷》，通过因子分析得出 9 个公因子，并将其对应为为某一证候要素。例如：公因子有偶有胸闷、经常胸闷，活动较多即感心悸，归纳为胸闷心悸因子，证候要素为气虚，与心相关。基于因子分析，使症状与证素相联系，揭示了辨证特点，为冠心病的二级预防提供了指导作用。魏艺等[13]应用因子分析研究老年原发高血压病患者的中医证素分布特点，结果显示证素分布情况为：肝肾同病 > 脾肾同病 > 心 > 心肝肾同病 > 肾 > 肝 > 心肝同病；病性证素分布情况为：痰浊 + 气滞 + 血瘀 + 阳亢 > 阳虚 + 血瘀 > 水停 + 痰浊 > 阴虚 + 痰浊 + 湿困 > 阴虚 + 气虚 > 血瘀 + 气虚 > 痰浊 + 血瘀 > 气滞 + 血瘀 > 痰浊 + 血瘀 + 气滞。王阶等[14]在文献研究和专家咨询的基础上制定临床流行病学调查表，对不稳定型心绞痛患者的症状和体征进行因子分析，得出 5 个公因子代表 5 种证候，分别为气虚血瘀、痰瘀互阻、阳虚寒凝、心肾阴虚和心脾两虚。根据各公因子所包含的变量及其权重大小，结合专家意见，可以初步建立不稳定型心绞痛的中医证候的诊断标准。气虚血瘀证：主症为气短，次症为自汗、倦怠乏力；舌有瘀斑瘀点，脉细弱。痰瘀互阻证：主症为胸痛；次症为咳嗽、痰多；舌苔厚腻，舌有瘀斑瘀点。阳虚寒凝证：主症为心悸；次症为畏寒肢冷、脘腹腰冷、下肢水肿。心肾阴虚证：主症为胸闷、心悸；次症为腰膝酸软、盗汗；舌红少苔。心脾两虚证：主症为气短；次症为嗳气、腹胀、便溏、自汗、脘痞。

1.4　回归分析　Logistic 回归分析方法是为自变量建立回归模型，估计参数的一种方法，它可以消除变量间的多重共线性。刘艳等[15]运用 Logistic 回归分析方法总结溃疡性结肠炎常见证候特异性的主要及次要症状指标，为该病的辨证治疗提供了参考。张瑞等[16]对 1627 例高校新生的体质类别进行判定，并采用多因素 Logistic 回归分析对体质影响因素和体制类别进行研究。此外，辛海等[17]通过调查高血压病患者的中医九型体质量表和 SF－36 生命质量量表数据，建立二者之间的多元回归分析方程。结果说明调整高血压患者的中医体制可以提高生命质量。

1.5　判别分析　判别分析，又称“分辨法”，在分类已确定的前提下，将研究对象的各种特征值进行判别，进而确定类型归属的一种多变量统计分析方法。杨勇等[18]将功能性便秘患者的症状与肝脾不调、肺脾气虚、肝肾阴虚、脾肾阳虚4 个证型的四诊信息进行判定分析，得出区分上述 4 个证型有显著贡献的 15 个变量，提高了功能性便秘的辨证分型的准确性。李琦等[19]对比围绝经期综合征肾虚型患者和健康人的 TT3、FT3、FT4、FSH、CORT 等实验室指标的质变，应用逐步判定分析法得出该证型与检验指标的相关性。

2　机器学习方法

2.1　贝叶斯网络　贝叶斯网络是由网络结构和概率集合组成，用简明的图形来表现变量与概率之间的关系的一种数据分析方法。贝叶斯网络存在两部分定义。其一为网络拓扑结构，表现为有向无环图，每个节点代表一个随机变量或命题。用弧线连接有直接关系的变量或者命题，表示一个概率依赖。吴宏进等[20]收集围绝经期综合征门诊患者的四诊信息，采用贝叶斯网络算法、最近邻算法、支持向量机算法 3 种数据挖掘分类算法将其进行分析，比较、评价不同分类算法的所需时间、分类准确性、覆盖率及 margin 曲线等，并得出贝叶斯网络算法优于其他 2 种算法的结论。张霆等[21]以肺癌患者症状之间的关联性及关联强度为基础，利用贝叶斯网络概括出了肺癌的证候要素。其中病机要素有痰湿、气虚、阴虚、血虚、内热等 9 个，病位要素有心、肝、脾、肺等 5 个。基于贝叶斯网络还得出了病机要素的主要症状和次要症状。

2.2　决策树　决策树是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。采用自顶向下的递归方式，在决策树的内部节点进行属性值的比较并根据不同的属性值判断从根节点向下的分支，在决策树的叶节点得到结论。张丽娜等[22]将体检人员的健康状况（包括健康、可疑亚健康、亚健康）和影响因素经单因素方差分析筛选出具有统计学意义的因素，将这些因素进行决策树分析归纳出模型的诊断规则，并验证后模型识别正确率和体质归类。余学杰等[23]整理四诊记录和专家对患者证候的辨证，通过决策树分析，得出证名和证候的

规律。

2.3 支持向量机 支持向量机是一种学习机器,以建立在统计学习理论和结构风险最小原则的基础上,根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折中,以求获得最好的推广能力。许明东等[24]用支持向量机学习算法,以高血压的常见 21 个症状、舌苔及舌体、脉象等为输入,高血压证型为输出,建立模型,并用高血压样本进行训练。经实际样本测试后,总体准确率较高,证实了其可行性。

2.4 粗集理论 粗集理论是一种采用新方法来处理模糊性和不精确性知识的数学工具,是处理模糊空间的一种数学方法。粗集理论的特点是:具有严格的数学定义,并且粗集信息处理不需要附加任何先决条件。粗集理论可以研究不完整数据、不精确知识表达、学习、归纳的方法。在中医研究领域,粗集理论则是有利分析的工具。它可以从中医诊疗过程的辨证、用药等进行量化,从中抽取出确定与可能规则。阎红灿等[25]通过基于粗集理论的 ROSETTA 系统处理中医喘证医案中的症状信息,并归纳用药配伍规律。经研究发现四诊信息的组合与相应药物组合有对应规律;药物组合与病因、病位、证候对应,但相同病因、病位和证候可以使用多种药物。喘症常用中药配伍为半夏、茯苓、甘草、橘皮、苦杏仁、麻黄、紫苏子等。

3 人工神经网络

人工神经网络兴起于 20 世纪 80 年代,模拟生物神经系统的原理而构建。它是一种能实现非线性映射功能的新型智能信息处理系统,是信息科学、数学、生物、医学等多学科交叉的边缘学科,具有自学习、自组织、并行分布式处理、良好的容错性等特点;能高速寻找优化解;具有联想存储功能;在处理噪声和漂移方面比传统的统计方法要好[26]。人工神经网络可以完成分类、聚类、特征挖掘等多种数据挖掘任务。李玉森等[27]运用人工神经网络技术,建立脾气虚弱证和肺脾气虚证 HIV/AIDS 患者的人工神经网络模型,将样本按 3∶1比例进行训练和测试,结果证明该模型的诊断准确率可达 80% 以上。覃裕旺等[28]应用改进的共轭梯度学习算法,以高血压症状为输入,高血压危险分层为输出,用样本进行训练和测试。结果表明人工神经网络网络方法可以准确将高血压患者进行危险分层。

4 数据库方法

王映辉等[29]建立了名老中医数据仓库,针对不同主题,通过设置病例数据的范围,利用多种数据挖掘方法,实现对不同名老中医共性及个性经验的知识发现。通过研究比较名老中医经验的异同,有利于更加深入地认识中医个体化医学的核心内涵,发现其共性规律,明确个性经验产生的原因,为今后开展验证性研究以及形成更加科学的中医经验理论提供了基础。并且为后学者学习和汲取多位名老中医经验提供了示范,对促进中医药的传承和发展具有一定意义。肖永华等[30]收集 128 例吕仁和教授诊治的糖尿病患者的医案,医案数据经过一致化处理之后存储于“中医医案数据库”中,对糖尿病类型、分期、并发症、病因、病位等出现的概率及其关系进行统计分析。结果说明吕教授临床治疗的糖尿病患者以出现糖尿病并发症和伴发病的 2 型糖尿病患者为绝对主体;并发症以糖尿病肾病出现的概率最高,其次为糖尿病神经病变;“肥美之所发”的饮食因素和“怒则气上逆”的情绪因素是重要的中医病因;肾、肝为主要病位。结果表明建立“中医医案数据库”并通过数据挖掘技术可以全面总结专家的学术思想,并且有利于名老中医学术思想的传承。

5 总 结

综上所述,数据挖掘方法在中医相关研究中应用广泛,特别是对中医临证经验的总结以及名老中医学术思想的传承取得了卓越的成就。数据挖掘方法为中医特色各个信息单元之间内在隐含关系的挖掘、规律的总结、问题的发现等提供了技术和方法学上的支持。然而,数据挖掘作为一种新兴前沿技术,仍然有很多方面需要完善,如数据挖掘得到的结果具有不确定性、数据挖掘过程中受人为主观影响较大等,相信在日后的实践与完善中一定能够得到校正和改良。

参考文献

[1] 张云涛,龚玲. 数据挖掘原理与技术[M]. 北京:电子工业出版社,2004.

[2] 李雄飞,李军. 数据挖掘与知识发现[M]. 北京:高等教育出版社,2003:10.

[3] 温丹婷. 基于关联规则对李丽芸教授辨治不孕症的用药规律研究[D]. 广州:广州中医药大学,2015.

[4] 王昆阳. 基于关联规则的焦虑抑郁共病中医证候规律研究[D]. 北京:北京中医药大学,2017.

[5] 郑丹文,刘擎,金晓阳,等. 当代名老中医治疗流行性感冒医案 72 则的中药配伍及方证规律关联分析[J]. 时珍国医国药,2013,24(7):1767 - 1769.

[6] 田瑾. 基于复杂网络及关联规则的失眠用药中医临床数据挖掘研[D]. 北京:北京中医药大学,2015.

[7] 杨青. 基于关联规则的近十五年紧张型头痛中医用药规律研究[D]. 济南:山东中医药大学,2016.

[8] 任毅,陈志强,张敏州,等. 当代名老中医治疗冠心病用药规律的聚类分析[J]. 中国中西医结合杂志,2016,36(4):411 - 414.

[9] 孙庆亮. 基于聚类分析对常见间质性肺疾病中医证型特点的研究[D]. 北京:北京中医药大学,2016.

[10] 丁冠福. 基于聚类分析的腹泻型肠易激综合征中医证候特

征研究[D]. 广州:广州中医药大学,2012.
[11] 牟新,庄爱文,马国玲,等. 237 例临床期糖尿病肾病患者中医证候聚类分析[J]. 中华中医药学刊,2016,34(2):332 – 335.
[12] 张琳婷. 基于因子分析的冠心病发病早期中医证候研究[D]. 沈阳:辽宁中医药大学,2013.
[13] 魏艺,曹雪滨,胡元会,等. 基于因子分析对老老年原发性高血压病患者中医证素分析[J]. 中华中医药学刊,2015,30(10):3474 – 3477.
[14] 王阶,何庆勇,李海霞,等. 815 例不稳定型心绞痛中医证候的因子分析[J]. 中西医结合学报,2008,6(8):788 – 792.
[15] 刘艳,李毅,刘力,等. 基于 Logistic 回归分析的溃疡性结肠炎中医症状组合规律研究[J]. 中医药导报,2017,23(12):52 – 56.
[16] 张瑞,刘岷,闫国立,等. 1627 例高校新生中医体质影响因素 Logistic 回归分析[J]. 中医杂志,2015,56(21):1858 – 1861.
[17] 辛海,金玫,沈蔷,等. 663 例高血压病中医体质与生命质量相关规律的多元回归分析[J]. 中国中医基础医学杂志,2011,17(7):798 – 799.
[18] 杨勇,丁曙晴,杨光,等. 功能性便秘中医证候的判别分析[J]. 广州中医药大学学报,2015,32(2):189 – 193.
[19] 李琦,周佩云,李浩,等. 更年期综合征中医肾虚证患者实验室指标判别分析[J]. 中国中西医结合杂志,2013,33(8):1064 – 1068.
[20] 吴宏进,许家佗,张志枫,等. 基于数据挖掘的围绝经期综合征中医证候分类算法分析[J]. 中国中医药信息杂志,2016,23(1):39 – 42.
[21] 张霆,陈波,徐涛,等. 基于贝叶斯网络的肺癌证候研究[J]. 中国中医药科技,2014,21(6):599 – 600,603.
[22] 张丽娜,刘声,陈素平,等. 基于决策树的亚健康状态判定及其与中医体质分类相关性研究[J]. 中华中医药学刊,2012,30(10):2185 – 2187.
[23] 余学杰,李书珍,李晓燕,等. 基于决策树提取中医专家辨证规律初探[J]. 辽宁中医杂志,2015,42(1):19 – 24.
[24] 许明东,马晓聪,温宗良,等. 支持向量机在高血压病中医证候诊断中的应用[J]. 中华中医药杂志,2017,32(6):2497 – 2500.
[25] 阎红灿,李丽红,马会霞,等. 基于粗集理论的中医喘证临床医案关联规则分析[J]. 辽宁中医杂志,2012,3(7):1218 – 1220.
[26] 刘旺华,洪净,李花,等. 人工神经网络在中医诊断信息化中的应用[J]. 湖南中医药大学学报,2017,37(7):809 – 812.
[27] 李玉森,施学忠,杨永利,等. 人工神经网络在 HTV/AIDS 患者主要虚证诊断中的应用[J]. 中华中医药杂志,2012,27(5):1269 – 1271.
[28] 覃裕旺,张爱珍,岳桂华,等. 基于 BP 神经网络的高血压中医证候与危险分层关系研究[J]. 中国中医基础医学杂志,2013,19(4):464 – 466.
[29] 王映辉,张润顺,周雪忠. 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,2009,11(6):793 – 799.
[30] 肖永华,王世东,李靖,等. 吕仁和教授辨治糖尿病医案数据挖掘分析[J]. 北京中医药大学学报,2009,16(3):1 – 4.

(收稿日期:2018 – 11 – 14)

(上接第 175 页)

[6] 朱丽冰,王济,郑燕飞,等. 超重和肥胖人群的中医体质分布特点及相关的影响因素分析[C]//中华中医药学会第十三次中医体质学术年会,2015:24 – 25.
[7] 魏宏,沈涛. “阳化气,阴成形”理论对肥胖症的指导[J]. 湖南中医杂志,2016,32(9):135 – 136.
[8] 张星辉. 浅析肥胖与郁证的关系[J]. 江西中医药,2015,46(6):9 – 11.
[9] 杨玲玲,倪诚,李英帅,等. 王琦治疗肥胖经验[J]. 中医杂志,2013,54(21):1811 – 1813.
[10] 冯博,徐云生. 徐云生从脾虚论治单纯型肥胖经验[J]. 河北中医,2014,36(5):646 – 648.
[11] 金昕,陈思,徐杰,等. 单纯性肥胖就诊患者的中医证素特征分析[J]. 中华中医药杂志,2016,31(7):2774 – 2778.
[12] 林志燕,田怀平,李方,等. 舒肝祛脂胶囊治疗成人单纯性肥胖的临床疗效及安全性评价(英文)[J]. 中国药学:英文版,2017,26(12):890 – 894.
[13] 张生玉. 临床中医针灸推拿学[M]. 西安:西安交通大学出版社,2014:377.
[14] 兰晓,周国平,杨路,等. 基于经络检测指导针刺循经取穴治疗单纯性肥胖的临床研究[J]. 中华中医药学刊,2017,35(4):977 – 980.
[15] 陈勇. 推拿手法治疗脾虚湿阻型单纯性肥胖症的临床观察[D]. 北京:北京中医药大学,2013.
[16] 阎博华,彭趣思,魏启华,等. 经穴推拿对单纯性肥胖患者体质量、体质量指数、腰围及臀围的影响:随机对照研究[J]. 世界针灸杂志,2014,24(1):6 – 9.
[17] 卓越. 运腹通经推拿法治疗单纯性肥胖症(脾失健运证)及对脂肪代谢影响的机制研究[D]. 长春:长春中医药大学,2014.
[18] 叶伊琳,王洁萍,秦勤,等. 穴位埋线治疗肥胖症的禁忌症及治疗后反应处理[J]. 中医药临床杂志,2017,29(9):1446 – 1448.
[19] 梁银利. 中医辨证联合穴位埋线治疗单纯性肥胖症的临床疗效[J]. 临床医学研究与实践,2017(2):134 – 135.
[20] 于天狐. 双活疗法对单纯性肥胖症患者 BMI 和血脂含量的影响[J]. 河北医学,2016,22(8):1349 – 1350.
[21] Tseng CC,Tseng A,Tseng J,et al. Effect of Laser Acupuncture on Anthropometric Measurements and Appetite Sensations in Obese Subjects[J]. Evidence – based Complementary and Alternative Medicine,2016,2016(1):1 – 8.

(收稿日期:2018 – 11 – 07)