·综述·

数据挖掘在中医药研究中的应用分析

董雪燕 祁烁 姜苗 韩丹阳 董石 崔玉

【摘要】 数据挖掘是从大量的、不完整的、复杂繁琐的数据中发现隐藏的、有价值的知识的过程。利用数据挖掘技术从中医累积的数据中探索和发现其中蕴含的规律,有利于中医学的继承、探索与发展。本文通过检索、总结中医学研究中数据挖掘技术相关文献,发现此类技术主要应用于中药方剂配伍规律、症状分析、辨证规律分析三方面;数据挖掘技术主要包括频数统计、聚类分析、因子分析、对应分析、关联规则、分类模型、贝叶斯网络、异常点分析等。

【关键词】 中医; 数据挖掘; 配伍规律

【中图分类号】 R2-03 【文献标识码】 A doi:10.3969/j.issn.1674-1749.2017.03.034

Application of data mining in the research of traditional Chinese medicine

DONG Xueyan, QI Shuo, JIANG Miao, et al. Dongzhimen hospital, Beijing University of Chinese Mcdicine, Beijing 100700, China

Corresponding author: QI Shuo, E-mail: qishuo1202@163.com

【Abstract】 Data mining is a process of discovering hidden and valuable knowledge from a large number of incomplete and complicated data. Using data mining techniques to explore and discover the law in the cumulative data of TCM, which is conducive to the inheritance, exploration and development of traditional Chinese medicine. The article summarizes the literature of using data mining techniques for traditional Chinese medicine research, or found the technology is mainly used in the law of compatibility of Chinese medical research, symptoms, syndrome differentiation regularity analysis; Data mining technology mainly include the frequency statistics, cluster analysis, factor analysis and correspondence analysis, association rules, classification model, Bayesian networks, abnormal point analysis, etc.

【Key word】 Traditional Chinese medicine; Data mining; Law of concerted application compatibility regularity

数据挖掘[1](Knowledge Discovery in Data-base, KDD)是指从大量的、不完整的、有噪音的数据中发现隐藏的、有价值的知识的过程,是当今大数据时代最前沿的数据处理技术之一。其主要表现形式为:规则、概念、规律及模式等。目前,数据挖掘方法逐渐应用于临床医学研究领域。

中医学具有系统性、整体性、复杂性、不确定性等特点。中医临床积累的信息颇多,数据类型及相互关系错综复杂,数据中隐藏大量有价值的信息。但由于此类数据的庞大和复杂,对有效信息及其相关性的分析与探索单纯依靠人力很难完成。而从大量数据中探索和发现其中蕴含的潜在规律与价值,正是数据挖掘技术的优势所在。本文通过检索、总结中医学研究中数据挖掘技术相关文献,发现此类技术主要应用于中药方剂配伍规律、症状分析、辨证规律分析三方面;数据挖掘技术主要包括频数统计、聚类分析、因子分析、对应分析、关联规则、分类模型、贝叶斯网络、异常点分析等,现简述于下。

1 配伍规律分析

在中药方剂配伍规律的研究领域,数据挖掘主要体现为应用频率统计法总结复方中的主要药物及其功效、药性,从而归纳复方的治则;运用聚类分

基金项目:2016 年北京中医药大学优秀青年骨干教师专项计划(2016-JYB-QNJSZX019)

作者单位:100700 北京中医药大学东直门医院血液肿瘤科[董雪燕(硕士研究生)、祁烁、姜苗、韩丹阳(硕士研究生)、董石(硕士研究生)、崔玉(硕士研究生)]

作者简介:董雪燕(1990-),女,2014 级在读硕士研究生。研究方向:中西医结合治疗血液肿瘤病。E-mail:675140504@qq.com

通信作者:祁烁(1985-),硕士,主治医师。研究方向:中西医结合治疗血液肿瘤病。E-mail:qishuo1202@163.com

析、因子分析法对复方及药物进行分类、归纳；通过药物的药效，透析病机病理，探讨证治规律；通过关联规则挖掘复方中的药对和药组，进而发现药物的配伍规律，总结治则、治法。

刘创[2]基于万方数据库、中国学术期刊全文数据库(2000-2014 年)统计关于肺纤维化中医证型的相关文献中治疗特发性肺纤维化(idiopathic pulmonary fibrosis，IPF)的中药复方，利用数据挖掘中的聚类分析和关联分析等方法，分别对复方和单药做聚类分析，总结 IPF 的中医证治规律、每类复方及单药的功效，进而归纳复方的治则；同时使用关联规则挖掘复方中的药对，提出将药对用于优化复方的思路。

数据挖掘技术也在中药复方研究中大放异彩。张天嵩等[3]基于中国生物医学文献数据库中肺纤维化的中药复方数据，对药物做频数统计，提取 36 种主要药物；进而对主要药物进行聚类分析，共分为补益药、活血药、化痰药等 6 类；最后经关联规则分析发现药对 19 条、药组 25 条，主要为益气药与活血药的配伍组合。通过对有趣的药组关联规则判读发现，大多数医家喜用益气药黄芪配伍丹参、当归、川芎等活血通络药；而药对关联规则中，益气药黄芪与党参配伍，益气药黄芪分别与丹参、当归、川芎等活血药配伍，而活血药丹参、当归、川芎两两配伍，说明益气活血通络法是众多医家治疗肺纤维化的共识。陈擎文[4]采用频数统计和关联规则方法分析古代中医古籍有关中风的医案，找出古籍背后隐藏的信息，从而分析出治疗中风的特有数据规则与规律；提出了一个萃取古代各家名医治疗各种病症经验的方法与模式，并经由中风病症的验证，证明本模式的可行性。研究结果显示：发掘出古代名医治疗中风最常用的 7 种中药，9 个药对，3 个药组。因此，本研究证明经由古代中医医案的数据挖掘的确可以有效萃取古代名医的治疗经验，其探勘后的知识不但可行，而且具有显著的临床应用价值。焦秋粉[5]基于张艳萍教授门诊中特发性肺纤维化的 113 条病例数据，使用频数统计方法分析了复方的主要药物，并使用均值法计算了不同药物的剂量，结果显示剂量会根据具体病情轻重缓急变化；使用 Apriori 算法对药物做关联规则分析，通过对药对进行分析发现：在用药规律方面，其多以破血、益气、养阴、通络、化痰、软坚散结之品配伍；最后，通过对用药做异常点分析发现：异常点药物多是针对合并症的药物。

数据挖掘技术的优势在古籍医案的研究中更为显著。古籍医案历史悠久，言语晦涩，后世医家继承时多加用自己的观点，很难客观、真实、全面地的反应作者当时的辨证思路。而数据挖掘技术则可以抛开局限的辨证思维，对整个古籍医案进行分析。马君[6]收集清末以前中医古籍文献中治疗肺痿的方药资料，对方药进行频数分析、因子分析、聚类分析，探讨证治规律。通过频数分析法对治疗肺痿各类药物的使用频率进行比较发现：治疗肺痿的主要药类为补益、清热、止咳等 7 类，并统计了主要药类中的常用药物。通过对方药的因子分析得到 7 个方药潜在因子，分别反映了肺痿的不同病机变化，结果提示燥热伤肺、气阴两虚、痰热蕴结为肺痿的主要病理机制。以药物作为变量聚类，得到 5 类配伍关系密切的药物组成的聚类方。结合药物频数统计和因子分析发现：古代肺痿的治疗原则以补益为主，益气养阴、清热化痰、止咳平喘为主要治法，这与现代肺纤维化治疗有相同之处。蒋永光等[7]从《中医大辞典·方剂分册》中筛选出 1355 首脾胃方，选用聚类分析、对应分析和频繁集方法，分析了核心药物、方剂结构、药对药组。结合聚类分析和核心药物分析结果发现：以四君子汤为代表的补气健脾方剂是脾胃方最基本的用方；复方主要结构有补气药配伍理气药、补气药配伍温里药、补气药配伍理气药及化痰药(或化湿药)。使用关联规则方法对药对、药组进行分析，为配伍规律研究提供了线索，如白术与茯苓、人参与生姜的配伍等。

古今医集医案数量之多，所用处方、中药数量之繁，仅仅依靠人力很难进行全面总结。再者，单纯的人力并不能发现多种药物之间潜在的规律，而数据挖掘恰好弥补了这一不足。中医历来有“医者意也”的说法。中医药知识体系中大量的隐性知识，无法通过文字或语言进行表达，这就是很多疗效显著的名医却很难把医术传承下来的原因。在中医药领域引入数据挖掘技术就是为了从大量中医病案中提取隐性知识，为理论研究和临床实施提供科学依据[8]。

2 临床症状分析

在症状分析方面，通过对症状、舌象、脉象的频率统计，发现相应疾病的主要症状和伴随症状；对症状、舌象、脉象建立贝叶斯网络，构建症状、舌象、脉象间

的因果关系网络,总结概括病机病位等证候要素。

焦秋粉[5]基于张艳萍教授门诊中特发性肺纤维化的 113 条病例数据,对病例中的临床症状做频率统计分析发现,发病人群常见症状以活动后喘息、气短,咳嗽、咳痰,胸闷、疲倦乏力、进行性呼吸困难为主,伴或不伴有口唇紫绀、杵状指及听诊闻及爆裂音等;对舌象、脉象做类似的频率统计分析,舌黯红、舌下脉络瘀滞、脉涩为主要舌象和脉象,表明血瘀证在 IPF 患者中所占比例较大。曲淼等[9]基于 611 例抑郁症患者的资料,对 86 项症状建立贝叶斯网络,发现其中的 39 个症状、舌象、脉象之间有较强的关联性,再由专家组概括出病机、病位等证候要素。病机要素包括:精亏、气虚、血虚、阴虚、阳虚、气郁(滞)、血瘀、痰湿、火热;病位要素包括:心、肾、肝、脾。庞博[10]使用基于贝叶斯的分类分析、基于支持向量的 SMO 分类分析、关联规则分析、聚类分析法分别对施今墨、祝谌予、吕仁和、赵进喜 4 位医家处方中药量、药性、药味、归经、方剂功效、症状、证素、证候进行分析,得到许多相互关联的症状,如口渴多饮与疲乏、淡白舌与紫舌、弦脉与细脉,分别提示了阴虚与气虚、气虚与血瘀、气滞与血虚之间的关系,即常说的气阴两虚证候、气虚血瘀证候和肝郁脾虚证候,进一步总结了四位医家的临床经验。陈为[11]采用数据挖掘的方法,以慢性肾炎的文献资料和临床患者为研究对象,归纳常见临床表现,探讨慢性肾炎肾阳虚证的症状规律和特点,提炼慢性肾炎肾阳虚证的主症、次症和一般症状,探索建立慢性肾炎肾阳虚证的诊断标准。结果得出历代医论中慢性肾炎肾阳虚证的常见症状有 16 个;进行相关性分析发现水肿和腰膝酸痛之间存在正相关性,畏寒肢冷和腰膝酸痛之间存在正相关性,气喘和咳嗽、咯痰之间存在正相关性,小便不利和气喘之间存在正相关性。李园白[12]采用数据挖掘的方法对中医妇科常见病(崩漏、闭经、不孕、痛经)进行研究,把中药药物与患者出现的共同症状中次数比较多的组合筛选出来,得出症状-中药之间的配伍关系。此项数据挖掘中共包含“症状-药物”100 组,例如“经血量少-当归”。这部分结果非常容易被接受而使用,且与中药的功效主治理论相吻合。

中医临床症状看似繁琐而散乱,但是舌脉、症状之间却有内在规律。中医临床是在整体观念指导下的辨证论治,不能管中窥豹,否则,很有可能陷入“头痛医头,脚痛医脚”的境地。但是大量的医案、医集单纯人力很难完整分析出医者的用药规律,不能囊括所有的临床症状,而数据挖掘技术则可以更全面地、系统地分析各种看似毫无关联的症状之间、症状与药物之间的潜在规律。

3 辨证规律分析

在辨证规律方面,数据挖掘方面的研究主要集中在:通过频率统计、关联规则探讨证候、证素、证型及其组合间的分布特点和规律;建立证候要素间的贝叶斯网络,获得各证候要素间的关系,结合中医理论提取证型;运用关联规则,对应分析症状与处方、辨证与处方、症状与辨证之间的关系;通过决策树等分类算法建立由症状到证型的分类模型。在目前中医证候标准化并未取得令人信服结果的背景下,数种挖掘方法的综合运用,更好地反映了中医证候、症状及证型间的关系[13]。

滑振等[14]通过万方数据库、CNKI 检索 2000-2014 年肺纤维化现代文献,运用数据挖掘方法探讨肺纤维化证素、证型及其组合间的分布特点和规律。通过主要病位的频率统计,发现肺纤维化病位主要在肺、肾,涉及脾、肝。统计单证素、双证素、三证素、四证素的频率,对证素做深入分析后得出:在具体临证时当侧重“补肺肾之气”和“补肺肾之阴”,兼以“活血化瘀祛痰”。在证素分析的基础上,使用关联规则挖掘证素组合,进一步说明肺纤维化病机“虚”和“瘀”的特性,也反映出中医“久病多虚多瘀”的理论特征。最终综合分析得出:肺纤维化的证素以气虚、血瘀、阴虚最为常见,其证型以多证素相互组合为主。肖光磊[15]将关联规则的数据挖掘方法应用于中医临床诊断发现:不仅可以从临床诊断数据中辨析症状与处方之间、辨证与处方之间、症状与辨证之间的关联关系,总结归纳名老中医的辨证规律并模拟其诊断、推理过程,还可以发现客观有用的新知识,以进一步促进专家经验的传承及其理论的完善。研究者收集了某位名老中医慢性胃炎诊治医案数据库,采用 FP-Growth 算法进行了挖掘,分别挖掘了症状与处方之间、辨证与处方之间、症状与辨证之间的关联规则,获取了该名老中医在中医诊疗中的经验。吴荣等[16]从证候要素应证组合规律两方面入手,建立冠心病心绞痛名老中医诊疗数据库,通过频率统计的方法进行研究探索常见证候要素应证组合特征,进一步将其分为实证、虚证和虚实夹杂证;运用贝叶斯网络构建证候要素与症状间的因果关系网络,发现与证候相关联

的症状,并以条件概率 0.5 为界,判定证候要素的主要症状和次要症状。江丽杰等[17]以 379 例缺血性中风病临床数据为基础,在 5 个不同时点(0 天、3 天、14 天、28 天、3 个月)采集《中风病辨证诊断标准》中"风、火、痰、瘀、气虚、阴虚阳亢"6 个证候要素的评分和《美国国立卫生院卒中量表》(NIHSS)评分,运用贝叶斯网络分析其相关性。结果表明:NIHSS 评分与中医证候要素评分之间存在相关性,其相关程度随时间呈动态变化。

4 讨论

4.1 数据挖掘在中医药研究中的现状

一般而言,数据挖掘主要应用在上述三个方面。其他如中药复方开发、中药产业化、国际化进程也都需要数据挖掘技术。目前常用的数据挖掘方法主要包括:多元线性回归分析、Logistic 回归分析、判别分析、聚类分析、因子分析、关联规则、粗糙集理论、决策树、人工神经网络、贝叶斯网络、信息熵等[18]。数据挖掘从应用角度可分为描述、预测、评估;从所用算法角度可分为预测类、非预测类、数据降维。常用的线性回归、非线性回归、决策树、贝叶斯网络、Logistic 回归属于预测类模型;聚类分析、关联分析则属于非预测类模型;因子分析、主成分分析属于数据降维。在具体的方法选择上,中药方剂配伍规律的研究常用聚类分析、因子分析、相关性研究等;对中医的症状分析以及辨证规律分析上常选用贝叶斯网络、异常点分析、决策树、Logistic 回归等。面对庞大而复杂的中医药数据,单一的数据挖掘方法难以揭示中医药的全貌,联合应用多种方法可以起到取长补短的效果,能够更深刻地反映中医证候的本质。

目前,无论在临床治疗方面还是科学研究方面,数据挖掘工作仍处于初级阶段,虽然有很多挖掘结果与临床吻合性较高,但仍有部分结论与临床实践有明显的出入,需要数据挖掘和中医学相关领域的专家深入研究与探索,不断改进数据挖掘的技术[19]。

4.2 大数据助力中医临床研究

继承名老中医学术思想是中医药发展的迫切需要,要在采用合理的科研方法上多下功夫。而数据挖掘在中医药信息化领域的重要地位日益突显,在名老中医经验传承方面尤为突出[20]。但是直至目前,对名老中医经验的数据挖掘仍存在许多问题:第一、信息采集的结构化、规范化问题。第二、门诊病历系统不利于病例保存,且各家医院门诊系统不统一,使科研仍局限在少数三级甲等大型医院,不能人人参与到其中。第三、数据挖掘技术的发展水平与中医药相结合的深度还未能改变中医药数据难处理的大格局。第四、人才问题。数据挖掘专业与中医药专业交流少,行业信息不能互通。随着计算机技术的快速发展和广泛应用,现代中医学研究者正在探索新的名老中医经验继承模式,通过数据挖掘得出的名家经验必定会得到越来越多专家认可。在国家的大力支持下,信息化助力中医药事业发展,为中医电子病历的发展提供广阔空间,也有利于利用中医电子病历实现名老中医临床诊疗经验数据采集、挖掘[21]。众所周之,名老中医治疗经验主要以隐性知识的形式存在,而对于挖掘出的隐形知识如何判读是一个更重要的问题,这也是进行数据挖掘的目的,单纯的计算机算法得出的信息不能称之为中医理论,只有在把握中医特有理论和思维规律的前提下,保证研究方向合理的同时借鉴数据挖掘所得到的结果,最终提炼出大量中医处方背后蕴藏的新理论、新方法、新知识。

只有把每一位患者的有效信息都记录在大数据仓库中,人人贡献数据,未来就可以凭借海量的病人信息,借鉴各派名家经验,模拟处方,探索人工智能,人人分享成果。这项艰巨的任务,需要几代人不懈的努力。设想如果全国所有的中医学者共享大数据,不论是从业几十年的老专家,还是刚刚毕业的青年医生,人人都可以挖掘出自己渴望的中医精华。

数据挖掘作为一个在海量数据中获取知识的有力工具,必将对带动中医药学术水平的提高、拓展中医药的生存空间产生巨大的启迪和促进作用[22]。大数据时代无疑是中医药发展的机遇,中医药研究者当抓住机遇,顺势而生,卓然而立,应世而壮[23]。

参 考 文 献

[1] Helma C. Data mining and knowledge discovery in predictive toxicology[J]. SAR QSAR Environ Res, 2004, 15(5-6): 367-83.

[2] 刘创. 基于数据挖掘和导师经验的特发性肺纤维化中医证治研究[D]. 沈阳:辽宁中医药大学,2014.

[3] 张天嵩,张素,李秀娟,等. 治疗肺纤维化中药复方用药规律的数据挖掘[J]. 中国中医药信息杂志,2011,18(2):31-34.

[4] 陈擎文. 数据挖掘技术在古代名中医中风医案之应用研究

[J]. 中华中医药学刊,2008,(10):2254-2257.
[5] 焦秋粉. 张燕萍教授诊治特发性肺纤维化经验的数据挖掘[D]. 北京:中国中医科学院,2014.
[6] 马君. 基于文献整理的肺痿方药证治规律探析[J]. 山东中医药大学学报,2012,(6):474-476.
[7] 蒋永光,李力,李认书,等. 中医脾胃方配伍规律的数据挖掘试验[J]. 世界科学技术,2003,(3):33-37,78.
[8] 王琦,金玉琴,周金海,等. 基于 WPF 的中医电子病历系统设计与实现[J]. 医学信息学杂志,2015,36(12):26-30.
[9] 曲淼,唐启盛,包祖晓,等. 贝叶斯网络模型在中医证候研究中的应用[J]. 中华中医药学刊,2008,26(7):1497-1498.
[10] 庞博. 施今墨学派名老中医诊治糖尿病学术思想与经验传承研究[D]. 北京:北京中医药大学,2012.
[11] 陈为. 基于数据挖掘的慢性肾炎肾阳虚证候诊断标准研究[D]. 成都:成都中医药大学,2011.
[12] 李园白. 中医妇科常见病医案数据挖掘方法研究[D]. 中国中医科学院,2006.
[13] 贾运滨,魏江磊. 数据挖掘技术在中医证候研究中的应用述评[J]. 中国中医急症,2010,19(7):1184-1186.
[14] 滑振,吕晓东,庞立健,等. 基于现代文献的肺纤维化中医证候及证素特征数据挖掘[J]. 世界中西医结合杂志,2015,(9):1188-1191.
[15] 肖光磊. 名老中医经验传承中的数据挖掘技术研究[D]. 南京:南京理工大学,2008.
[16] 吴荣,聂晓燕,王阶,等. 基于贝叶斯网络的名老中医治疗冠心病辨证规律研究[J]. 中国中医药信息杂志,2010,17(5):98-99.
[17] 江丽杰,胡镜清,易丹辉,等. 缺血性中风病中医证候要素动态变化与 NIHSS 评分变化相关性的贝叶斯网络分析[J]. 世界中医药,2013,8(6):613-617.
[18] 黄粤,高颖,马斌. 中医证候研究常用数据挖掘方法述评[J]. 中医药学报,2010,38(3):6-10.
[19] 王倩,生慧,金卫. 中医药领域数据挖掘技术的研究与应用概况[J]. 湖南中医杂志,2015,31(3):186-188.
[20] 沈春锋,王彩华,陆炜青. 名老中医学术思想的传承[J]. 实用中医内科杂志,2015,(12):184-185.
[21] 刘鸿燕,胡红濮,张越. 基于电子病历结构化的名老中医经验数据挖掘研究[J]. 医学信息学杂志,2015,36(12):13-18.
[22] 姚美村,袁月梅,艾路,等. 数据挖掘及其在中医药现代化研究中的应用[J]. 北京中医药大学学报,2002,25(5):20-23.
[23] 孟庆云. 顺势而生 卓然而立 应世而生——当代中医药发展的大数据、网络化机遇[J]. 中国中医基础医学杂志,2016,(1):1.

(收稿日期: 2016-01-06)
(本文编辑: 董历华)