

DOI: 10.13288/j.11-2166/r.2016.10.020

综 述

数据挖掘方法在名老中医用药规律研究中的应用

沈毅^{*}, 傅萍², 孔丽娅³

(1. 浙江中医药大学第三临床医学院, 浙江省杭州市滨江区滨文路 548 号, 310053; 2. 杭州市中医院; 3. 浙江中医药大学基础医学院)

摘要 近年来数据挖掘方法在中医药研究领域的应用日益广泛, 特别是在名老中医用药规律研究中发挥着重要作用。分别对关联规则分析、聚类分析、因子分析、无尺度网络的数据挖掘方法在名老中医用药规律研究中的应用进行总结, 认为以上方法是全面深入挖掘名老中医临床经验及用药特点的有力工具, 应根据研究目的选择最科学有效的数据挖掘方法, 并将最终提炼的成果在实践中反复验证, 方能提高研究结果的可靠性和科学性; 同时将相关数据挖掘方法进行集成并形成相应的软件, 是数据挖掘方法在名老中医经验总结领域应用的重要方向。

关键词 数据挖掘; 名老中医经验; 关联规则分析; 聚类分析; 因子分析; 无尺度网络分析

名老中医具有丰富的临床诊疗经验, 是中医学的带头人。名老中医临床经验、用药规律和学术思想的总结与传承不仅能丰富中医学的理论体系, 还能推动中医学的进步和理论创新, 也是培养新一代名中医的重要途径之一。名老中医组方用药规律研究是中医药总结和发展的核心内容之一。中医处方信息数据错综复杂, 方剂配伍包含了病-证-方-药量的多维关联, 其本质就存在一种数据关系, 具有模糊性、复杂性和非线性动态变化的特性, 符合数据处理的要求和特征。

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、模糊的、随机的实际应用中提取隐含的、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]。采用数据挖掘技术对名老中医用药规律进行研究, 人机交互, 反复多次, 可总结归纳名老中医用药特点, 指导临床实践并提高临床疗效; 同时以药物为连接点进一步探讨疾病的治则治法、病因病机、方药配伍机制, 还可提炼出临证经验中蕴藏的新理论、新方

法、新知识, 实现名医经验的有效总结与传承。本文通过对数据挖掘方法在名老中医用药规律研究中的应用进行评述, 希冀为数据挖掘方法在名老中医用药规律及经验传承领域的进一步深入应用提供参考。

1 关联规则分析

关联规则分析 (association rules) 是从大量的数据中发现项集之间有意义的关联^[2], 常用的算法包括 Apriori 算法、基于划分的算法、FP-树频集算法。关联规则是形如 $X \rightarrow Y$ 的蕴涵式, 支持度和置信度是关联规则的两个重要概念, 其中支持度表达的是 X 和 Y 这两个项集在所有事务中同时出现的概率, 置信度描述了在出现项集 X 的事务集中项集 Y 也同时出现的概率, 支持度反映规则的普遍性, 而置信度反映规则的可靠性^[3]。黄玉华等^[4]运用关联规则分析方法对柴松岩治疗多囊卵巢综合征用药规律进行研究, 总结出核心药物 10 味, 新药对 7 对, 其中补虚药和活血化瘀药使用均数最高, 与柴松岩认为多囊卵巢综合征病机以脾肾不足、痰湿阻滞 (胞络) 为主, 临床治疗以益肾健脾、养血通利为主要法则的学术思想相印证。徐建龙等^[5]以临床收集的聂莉芳治疗的 IgA 肾病患者 251 例为研究对象, 采用 Aprior 算法对其药物-药

基金项目: 国家中医药管理局全国名老中医药专家传承工作室建设项目 (国中医药人教发 [2014] 20 号); 浙江省名老中医专家传承工作室建设项目 (GZS2012023); 2015 年浙江省中医药科技计划项目 (2015ZA162)

* 通讯作者: 1127649268@qq.com

物、处方-症状、症状-药物进行关联规则分析,结果显示,最常用的前 5 味药依次为太子参、金银花、黄芪、生地黄、芡实,六药关联与七药关联出现了六味地黄丸的经典配伍,并在此基础上加太子参而成气阴双补之剂,主要体现了聂莉芳益气养阴固精、疏风清热凉血的治疗大法。赵宇明等^[8]经关联分析得出在王庆国临床用药处方中,具有特征性的药物组合与《伤寒论》经方及刘渡舟常用方相切合,尤以小柴胡汤、柴胡桂枝干姜汤、泻心汤中药组使用为主,在体现刘渡舟“重视和解少阳枢机”学术思想的同时在处方用药上有新的突破。吴嘉瑞等^[9]收集杨博良治疗月经病处方,采用关联规则 Aprior 算法、复杂系统熵聚类无监督数据挖掘方法,结果发现,杨博良治疗月经病常用药对包括泽泻配黄芩、川芎配青皮、当归配黄芩等,常用药物以凉血利湿、活血调经者为主。张北华等^[10]收集整理 108 位中医药专家治疗慢性胃炎的医案,采用关联规则分析药物配伍规律,结果显示,中医药治疗慢性胃炎以健脾、疏肝、理气、清热、消食药最常用,常用处方包括六君子汤、半夏泻心汤、芍药甘草汤、四逆散、左金丸等。

关联规则分析是中医临床资料研究中运用最广泛、最成熟的数据挖掘方法,可见于病-证-症-药之间关联的研究。关联规则分析的优点是非常简单,其分析结果常以类似产生式规则的形式表达,相对比较清晰直观,因而更容易被解释和理解。但其在用药规律的挖掘中也存在一些问题,如高置信度、低支持度的关联规则发现效率低,不能确定“相互依赖”的规则,找到的强规则并不一定是有趣的、甚至是错误的。针对以上问题有学者提出双向关联规则挖掘算法用于用药规律研究,同时将分析结果与临床经验互相印证,能得到有意义的药组及药对^[11]。

2 聚类分析

聚类分析 (cluster analysis) 又叫集群分析,是按“物以类聚”原则研究事物分类的一种多元统计分析方法^[12]。聚类在事先不知道有几类的情况下分析数据间的相似程度,根据类间不同、类内相似的原则把数据归类。聚类分析算法包括划分法、层次法、密度法、网格法、模型法等。王祖龙等^[13]采用聚类分析方法对褚玉霞治疗的 368 例先兆流产病例用药规律进行研究,结果发现,使用频率 >10% 的药物共 27 味,通过聚类分析将核心药物分为补肾、益气健脾、养血、清热、止血、理

气、和胃、宁心安神、润肠通便药 9 大类,并在此基础上遵循“补肾健脾、养阴清热”的原则,将党参、白术、白芍、炙甘草、黄芩、砂仁、续断、杜仲、枸杞子、菟丝子、阿胶、墨旱莲、仙鹤草构成新方“双保煎剂”。闫禹竹等^[12]应用熵聚类方法对程为平治疗癫痫的方剂进行组方规律分析,发现一些新药对或药物组合,并分析这些药对及组合具有重要临床意义。许金海等^[13]应用分层样本聚类方法分析施杞治疗腰椎间盘突出症用药特点,结果显示,益气活血药、补肾助阳药、祛风解表药、利水渗湿药、化瘀通络药及健脾理气和胃药为施杞治疗腰椎间盘突出症的常用配伍药类,聚类所得四类药物组分别以圣愈汤、身痛逐瘀汤、独活寄生汤、苓桂术甘汤为基本方药,该结果客观反映了施杞临床用药规律。田野等^[14]采用聚类分析挖掘和贝叶斯网络分析战丽彬治疗恶性肿瘤的用药特点,得出 5 个高频药物聚类组,组方以补虚药为主、清热解毒药为辅,配伍理气、活血化瘀、消食、安神之品等,体现了恶性肿瘤“病中常虚”的病机特点及扶正抗癌的治疗思路。

聚类分析过程中没有一个已知类别归属的学习样本集,是一种无监督学习,因此,可以避免传统分类方法的人为主观性。对于名老中医用药规律研究而言,它可以从药物客观属性(四性、五味、归经、功效)出发,将药物分成不同的类别,并获得具有临床治疗意义的基础方。但其在实际应用中也存在一定问题,如聚类分析不具有线性的计算复杂度,难以适用于样本量庞大的数据库;无法利用数据类别归属的先验知识,其结果的测试和验证较为复杂;预先输入参数例如希望产生类的数目及不同的聚类算法对聚类分析结果的影响较大。聚类算法的选择取决于数据的类型、聚类的目的及应用,可以对同样的数据尝试多种算法,以发现数据可能揭示的结果。针对不同分析结果,需征求名老中医反馈意见,将客观的数据结果与主观的思想结合,最终得到具有临床指导意义的基础方。

3 因子分析

因子分析 (factor analysis) 是从分析多个原始指标的相关关系入手,找到支配这种相关关系的有限的不可观测的潜在变量,即公因子,并用这些潜在变量来解释原始指标之间相关性的多元统计分析方法^[15]。谢辉辉等^[16]以汪受传门诊治疗小儿厌食的 208 份病案为基础,通过对处方中 28 味核心药物进行相关分析、聚类分析及因子分析,3 种方法

得出的结果基本一致,揭示了汪受传治疗厌食以运脾消食为核心,配伍行气化湿之品,针对不同病机辨证选用清热润下和补气固表药的用药规律。陈亮等^[17]运用因子分析探索丁震临床治疗慢性胃炎的用药规律,结果共提取药物组合公因子 9 个,能够涵盖全部变量信息的 74.9%,9 个公因子均以调畅肝、脾、胃的气机为主要特点,同时注重化湿、健脾、消积、祛瘀、止痛等不同的方面,与丁震临床治疗本病从肝论治、调畅脾胃气机为主要法则的学术思想一致。郭闫葵等^[18]从浦家祚治疗眩晕的临证用药入手,对治疗眩晕病的 34 种高频药物进行因子分析,归纳出 10 个常见药物组合,用药以滋补肝肾为主,配合平肝熄风、活血化瘀、健脾利水、行气化痰等药物,提示眩晕病机主要为虚、瘀、痰、风 4 个方面。

因子分析可以消除人为干扰,从大量处方数据中概括出治疗疾病的药物群组,进而推断疾病病机特点、辨证分型及治则治法。其优点包括以下两点:第一,它不是对原有变量进行取舍,而是根据原始变量的信息进行重新组合,找出影响变量的共同因子,化简数据;第二,它通过旋转使得因子变量更具有可解释性,命名清晰性高。因子分析建立在数据的正态假设的基础上,因此,应用因子分析进行名老中医用药规律研究时应注意样本量要足够大,而且各变量之间具有相关性。同时为保证结果的真实性和稳定性,应当做一些基于多中心、大样本、前瞻性基础上的用药调查,必要时可以进行重复性检验。

4 无尺度网络

无尺度网络(scale-free networks)是最为经典的复杂网络模型之一,是指网络的节点度分布满足幂律分布^[19]。研究表明,中医理论指导下的复方配伍过程具有无尺度复杂网络现象^[20],无尺度网络分析为名老中医复方配伍、药物相互作用等的研究提供了依据。高剑虹等^[21]应用无尺度网络模型分析方法探讨方和谦治疗肝郁脾虚证药物配伍规律,结果发现,位于网络中心的药物即核心药物为党参、炒白术、茯苓、当归、香附、柴胡、炒白芍、炙甘草、紫苏梗、大枣、薄荷、生姜,加减应用频次较高的药物种类有补肾滋阴药、理气和胃药、清热散结药、滋养胃阴药。郭旸等^[22]基于无尺度网络方法对戴希文治疗慢性肾脏病用药经验进行研究,得出处方配伍网络及网络药物关联频度信息,总结出核心方药 15 味,随证加减 18 组,结果

体现出戴希文诊治慢性肾脏病扶正祛邪的治疗总则及扶正不壅补的用药特点。

无尺度网络以处方为基础,以处方中药物为节点,以药物间配伍关系为边,边的权重表示两药物间配伍的频度,建立药物配伍网络图,分析得出核心药物组成及处方配伍规律。无尺度网络的优势在于可以提供图示化、直观化的药物配伍关联图,其结果更具可读性。但同时也应当注意到无尺度网络在名老中医用药规律研究中可能存在的问题,例如在实际研究中样本太少或取样偏差会导致对网络结构的错误认识等。

5 展望

如上所述,关联规则、聚类分析、因子分析、无尺度网络数据挖掘方法在名老中医用药规律的研究中发挥着重要作用,是全面深入挖掘名老中医的临床经验、用药特点及学术思想的有力工具。在数据挖掘过程中应遵循“人机结合、以人为本”的原则,在数据挖掘方法的选择上应根据研究目的的不同选择最科学有效的挖掘方法,综合考虑各种方法的优点、缺点和适用范围,取长补短、联合应用,比较分析结果,将最终提炼的成果在实践中反复验证,从而提高研究结果的可靠性、科学性。

数据挖掘是一个多学科交叉研究领域,其融合了数据库、统计学、人工智能、信息检索、高性能计算以及数据可视化等最新技术,其具体方法需要有计算机学、信息学、统计学等人员参与才能熟练应用,而不易被中医药临床医师和学术继承人所掌握,这给数据挖掘方法在名老中医经验总结中的应用带来困难和挑战。因此,将相关数据挖掘方法进行集成并形成相应的软件,是数据挖掘方法在名老中医经验总结领域应用的重要方向^[23]。

数据挖掘是基于整理而重于发扬、基于经验但超越经验的复杂信息研究平台,可满足名老中医从个体化诊疗数据中发现隐性知识的需要,为名老中医组方用药经验的研究提供了有效工具,具有广阔的应用前景和实用价值,值得进一步深入研究。

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,2004:1.
- [2] 曾孝文.关联规则数据挖掘方法的研究[J].计算机与现代化,2006(9):90-92,96.
- [3] 苏新宁,杨建林,江念南,等.数据仓库和数据挖掘[M].北京:清华大学出版社,2006:116-150.
- [4] 黄玉华,柴松岩,张德政,等.基于关联规则的柴松岩治疗多囊卵巢综合征用药规律研究[J].中国中医药信

- 息杂志, 2012, 19(5): 23-26.
- 〔5〕 徐建龙, 聂莉芳. 基于关联规则的聂莉芳教授治疗 IgA 肾病用药规律分析 〔J〕. 中国中医药信息杂志, 2010, 17(4): 96-98.
- 〔6〕 赵宇明, 尤海燕, 刘哲, 等. 基于数据挖掘的王庆国教授对刘渡舟教授用药传承规律的研究 〔J〕. 北京中医药大学学报, 2012, 35(5): 293-296.
- 〔7〕 吴嘉瑞, 赵梦迪, 郭位先, 等. 基于数据挖掘的孟河名医杨博良治疗月经病用药规律研究 〔J〕. 中华中医药杂志, 2014, 29(10): 3086-3089.
- 〔8〕 张北华, 高蕊, 李振华, 等. 中医药治疗慢性胃炎的专家经验数据挖掘分析 〔J〕. 中医杂志, 2015, 56(8): 704-708.
- 〔9〕 曾令明, 唐常杰, 阴小雄, 等. 双向关联规则挖掘及其相关性分析 〔J〕. 计算机工程与设计, 2005, 26(10): 2585-2587.
- 〔10〕 何清波, 苏炳华, 钱元. 医学统计学及其软件包 〔M〕. 上海: 上海科学技术文献出版社, 2003: 285.
- 〔11〕 王祖龙, 李晖, 孙红. 褚玉霞治疗先兆流产用药经验分析 〔J〕. 中国中医基础医学杂志, 2011, 17(10): 1108-1109.
- 〔12〕 闫禹竹, 程为平, 毕述明, 等. 基于数据挖掘方法分析程为平教授治疗癫痫的用药经验 〔J〕. 中医药学报, 2014, 24(3): 42-44.
- 〔13〕 许金海, 郭学群, 叶洁, 等. 基于聚类分析的施杞教授治疗腰椎间盘突出症的用药规律研究 〔J〕. 中国中医骨伤科杂志, 2013, 21(3): 20-24.
- 〔14〕 田野, 李舒. 战丽彬治疗恶性肿瘤用药规律的数据挖掘分析 〔J〕. 中医杂志, 2014, 55(8): 657-660.
- 〔15〕 孙振球, 徐勇勇. 医学统计学 〔M〕. 北京: 人民卫生出版社, 2002: 26.
- 〔16〕 谢辉辉, 袁丹, 祁波, 等. 汪受传教授治疗厌食用药的相关分析和因子分析 〔J〕. 世界科学技术-中医药现代化, 2014, 16(10): 2087-2092.
- 〔17〕 陈亮, 张寅, 杜娜, 等. 运用因子分析探讨丁霞教授治疗慢性胃炎的用药规律 〔J〕. 中华中医药杂志, 2014, 29(8): 2489-2492.
- 〔18〕 郭闫葵, 郭松霞, 高琛, 等. 浦家祚教授辨治眩晕用药规律研究 〔J〕. 中国中医药科技, 2014, 21(1): 14-15.
- 〔19〕 BARABASI AL, ALBERT R, JEONG H. Mean-field theory for scale-free random networks 〔J〕. Physical A, 1999, 272: 173-187.
- 〔20〕 周雪忠, 刘保延, 王映辉, 等. 复方药物配伍的复杂网络方法研究 〔J〕. 中国中医药信息杂志, 2008, 15(1): 98-100.
- 〔21〕 高剑虹, 李文泉, 范春琦, 等. 方和谦治疗肝郁脾虚证中药配伍规律研究 〔J〕. 北京中医药, 2013, 32(2): 95-98.
- 〔22〕 郭旸, 饶向荣, 张润顺. 基于无尺度网络方法总结戴希文治疗慢性肾脏病经验 〔J〕. 中国中医药信息杂志, 2011, 18(7): 26-28.
- 〔23〕 吴嘉瑞, 唐仕欢, 郭位先, 等. 基于数据挖掘的名老中医经验传承研究述评 〔J〕. 中国中药杂志, 2014, 39(4): 614-617.

(收稿日期: 2015 - 05 - 12; 修回日期: 2015 - 09 - 14)

编辑: 崔京艳