

· 数据挖掘与中医传承 ·

基于数据挖掘的名老中医经验传承研究述评

吴嘉瑞¹, 唐仕欢², 郭位先¹, 张晓滕¹, 张冰^{1*}

(1. 北京中医药大学, 北京 100102; 2. 中国中医科学院 中药研究所, 北京 100700)

〔摘要〕 数据挖掘又称数据库知识发现,是从海量数据中揭示出隐含的、先前未知的并有潜在价值的信息的非平凡过程。近年来,数据挖掘在中医药研究领域的应用日益广泛,特别是在名老中医经验传承领域发挥着重要作用。该文对关联规则、贝叶斯网络、神经网络、决策树、复杂系统熵方法等数据挖掘方法在名老中医经验传承研究中的应用进行总结,并对常用研究方法的优势和不足进行剖析,指出将数据挖掘方法进行集成而开发相应的软件,是其重要的研究方向。

〔关键词〕 数据挖掘;名老中医;经验传承

名老中医是中医药学术发展的杰出代表,是联系传统和实现中医发展的灵魂,名老中医经验的总结和传承不仅能丰富中医学的理论体系,还能对中医药学术进步产生巨大的推动作用,也是培养造就新一代名中医的重要途径之一。因此,总结名老中医的临床经验、用药规律和学术思想,对中医药薪火相传具有重要的理论意义和应用价值,也是中医药创新发展的源泉。数据挖掘(data mining, DM)又称数据库知识发现(knowledge discover in database, KDD),是指从数据库的大量数据中揭示出隐含的、未知的并有潜在价值的信息的非平凡过程。采用数据挖掘技术对名老中医学术思想和临证经验进行研究,可以全面解析其中的规律,分析名老中医个体化诊疗信息特征,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。自从基于信息和数据技术的名老中医临床诊疗经验研究思路提出以来,利用多种数据挖掘技术对病案进行分析的研究报道与日俱增,在名老中医经验传承领域发挥着重要作用^[1-3]。本文中,笔者结合多年从事数据挖掘研究的经验与体会,并借鉴国内同行的相关研究成果,对数据挖掘在名老中医经验传承研究中的应用进行论述,希冀为数据挖掘方法在名老中医经验传承领域的进一步深入应用提供参考。

1 名老中医经验传承研究中数据挖掘方法的应用

1.1 关联规则 关联规则(association rules)是从大量的数

据中挖掘发现项集之间有意义的关联,并寻找给定的数据集中项之间的有趣联系的一种算法。常用的关联规则算法包括 Apriori 算法、FP-树频集算法等^[4-6]。吴荣等采用关联规则算法对名老中医治疗冠心病的用药规律进行数据挖掘。结果显示,治疗冠心病处方中最常用的中药是丹参、瓜蒌、川芎,最常用的方剂是瓜蒌薤白类方、活血通脉剂及生脉散^[7]。刘晓怡等采用关联规则方法对李佃贵教授辨治慢性萎缩性胃炎医案进行数据挖掘,得出李佃贵教授诊治慢性萎缩性胃炎最常用的药物是茵陈、黄连^[8]。田琳等采用关联规则算法对名老中医诊疗眩晕病辨证思维模式进行数据挖掘,得出眩晕病的证候多为风阳上扰、肝肾阴虚、痰浊上蒙、肝肾阴虚、气血两虚等,病因病机多为虚、风、痰、瘀等,发病与肝、肾、脾三脏关系密切^[9-10]。笔者采用关联规则方法对国医大师颜正华教授治疗胃脘痛用药规律进行数据挖掘,得出颜正华教授在治疗胃脘痛过程中常用药物为陈皮、佛手、香附、白芍、煅瓦楞子、赤芍、当归、丹参等,使用频率前3位的药对是佛手、陈皮,陈皮、香附,佛手、香附^[11-12]。作为在名老中医处方规律研究中使用最广泛的数据挖掘算法,关联规则具有明显的优点,如它可以产生清晰有用的结果,支持间接数据挖掘,可以处理变长的数据等。但是,关联规则也有其不足,如计算量增长相当严重,难以决定正确的数据,容易忽略稀有的数据等。

1.2 贝叶斯网络的应用

贝叶斯网络(Bayesian network)是一种以贝叶斯公式为基础的概率网络,是一个有向无环图,由代表变量结点及连接这些结点的有向边构成。贝叶斯网络可以将具体问题中复杂的变量关系在一个网络结构表示,通过网络模型反映问题领域中变量的依赖关系^[13]。吴荣等利用贝叶斯网络技术对名老中医治疗冠心病辨证规律进行数据挖掘,提取出名老中医诊疗冠心病、心绞痛的8个证候要素^[14]。须义贞等采用贝叶斯网络和方剂智能分析软件对沈仲理教授诊疗子宫肌瘤医案进行数据挖掘,得出沈仲理教

〔收稿日期〕 2013-10-29

〔基金项目〕 国家科技支撑计划项目(2007BAI10B01);北京市自然科学基金项目(7112075);北京市中医药科技发展基金项目(JJ-2010-70);北京中医药大学科研创新团队项目(2011-CXTD-14);北京中医药大学“重点学科”开放课题(2013-ZDXKKF-19)

〔通信作者〕 *张冰,教授,博士,博士生导师, Tel: (010) 84738606, E-mail: zhangbing@263.net

〔作者简介〕 吴嘉瑞,副教授,博士,硕士生导师, Tel: (010) 84738662, E-mail: exogamy@163.com



授诊子宫肌瘤善用活血化瘀、清热软坚法,常用治疗药物包括三棱、莪术、丹参、半枝莲、蚤休等^[15]。贝叶斯网络具有分类、聚类、因果分析等功能,面对大规模数据有其独特的优势,具体包括:①学习机制高效灵活,可发现潜在的有用的模式或者联系,实现对数据实例的分类、聚类、预测;②语义清晰,易于理解和接受,具有良好的预测能力;③有效避免对数据的过度拟合;④贝叶斯网络可将先验知识、样本、主观、客观有机结合起来,更加全面反映数据对象内在的联系和本质^[16-17]。

1.3 神经网络的应用 人工神经网络(artificial neural network, ANN)是由大量处理单位(即神经元, neurons)广泛连接而成的网络,是人工建立的以有项图为拓扑结构的动态系统,通过对连续或断续的输入作状态响应而进行信息处理^[18]。陈肇文利用人工神经网络建立名老中医中医方剂系统和中医方证医学诊疗系统,可在线根据患者症状自动进行处方^[19]。白云静等利用人工神经网络技术开展中医证候非线性建模研究,在充分辨识证候表征信息的基础上,建立证候与症状之间的非线性映射函数^[20]。陆萍等基于面诊的证素辨证建立面诊-证素辨证神经网络,对64个病例进行证型辨别,结果表明,神经网络模型构建的中医面诊系统能根据样本值对网络构造和训练,证型辨证结果准确率比较高,是可行的^[21-23]。作为最常用的数据挖掘方法之一,神经网络具有很强的非线性拟合能力,可映射任意复杂的非线性关系,而且学习规则简单,便于计算机实现,具有很强的鲁棒性、记忆能力、非线性映射能力以及强大的自学习能力。

1.4 决策树算法的应用 决策树(decision tree)模型是一种自上而下的预测模型,其基本运算原理属于分类规则,也就是说每个决策或者事件都能引出2个或者多个事件,继而产生不同的后果,这种决策分支的图像就像一棵树的枝干,称其为决策树。决策树以样本的根节点开始,如果样本在同一个类,则该结点为树叶,算法选择最有分类能力的属性作为决策树的当前结点,根据信息增益进行分裂,直到给定节点的所有样本属于同一类或没有剩余属性可供划分为止^[24]。翟海斌等采用决策树方法对血瘀证的诊断规则进行归纳,得到血瘀证决策分类模型,结果表明决策树提取的诊断规则与传统的医学统计方法相比,更为简单、方便,为从病例中自动归纳诊断规则提供了一种新的方法^[25]。谢雁鸣等从临床流行病学的角度用决策树等方法对原发性骨质疏松症的中医基本证候进行研究,得出原发性骨质疏松症的阴虚和阳虚诊断模型^[26]。徐蕾等用信息熵的决策树方法对慢性胃炎的中医辨证进行研究,构建以信息熵减少为特征的决策树分类模型^[27]。

1.5 复杂系统熵方法的应用 2004年,有学者发表了《“熵”在中医方证研究中的运用》一文,在总结熵理论的基础上,首次提出运用熵理论进行中医方证相关研究的设想,并探讨了证、熵、方结合的原理与前景^[28]。通过实践表

明^[29-30],该方法不仅可以应用证候与方剂之间的相关性研究,还可以研究证候与证候之间、证候与症状之间、方剂中药物与药物之间、配伍中成分与成分之间的关系。其后,有学者提出了复杂系统熵聚类的算法,并在疾病、证候、方剂的相关性研究中得到了应用^[31]。复杂系统熵聚类方法是一种非监督的模式发现算法,它能自组织地从海量的数据中提取出信息量最大的组合,同时,此方法特别适用于高度离散性类型的数据。相比于经典的统计方法,它有以下几个优点。①不需要数据的一致性,对于各类数据都适合。特别针对具有随机性、模糊性、非平衡性、非遍历性、多维性特点的中医药数据。②它客观地反映数据的情况,聚出来的组合内元素的相关都特别大,是最优的组合,这些组合为新药发现中候选处方的筛选奠定了基础。③相关是不对称的,为定义贡献度奠定了基础。④算法收敛速度快,对于处理大量的数据有优势。此方法具有两方面的显著优势:一方面,不仅可以定性、还可以定量挖掘出药物之间、病-证-症-药之间的相关性;另一方面,不仅可以挖掘出名医名家经验的核心组合,还可以挖掘出隐藏于方剂配伍之中的而没有被临床医家所重视的核心组合。笔者采用该方法对国医大师颜正华教授治疗胃痛、失眠等用药规律进行挖掘,得出颜正华教授在治疗胃痛、失眠的潜在药物组合和新方,为深入研究和药物开发奠定了基础^[11-12]。

2 数据挖掘在名老中医经验传承中的应用展望

如上所述,关联规则、贝叶斯网络、神经网络、决策树等数据挖掘方法在名老中医经验传承研究中发挥着重要作用,是深入挖掘、继承名老中医的学术思想和临床经验的有力工具。然而,每种数据挖掘方法都有其局限性和不足,均有其适应范围,且对数据有一定的要求。中医药数据挖掘的对象是中医药领域中积累的海量数据,这些数据的属性既有离散型的,又有连续型和混合型的特点,挖掘过程需要人机交互、多次反复,在中医药专业背景知识引导下,针对具体问题,选择合适的数据挖掘方法。同时,数据挖掘是从大量的数据中,抽取潜在的、有价值的知识的过程,融合了数据库、人工智能、机器学习、统计学、知识工程、面向对象方法、信息检索、高性能计算以及数据可视化等最新技术的研究成果,是一个多学科交叉研究领域。因此,数据挖掘具体方法需要有计算机、统计学等多学科交叉人员才能熟练应用,不易被中医药临床人员和学术继承人等所掌握,这给数据挖掘方法在名老中医经验传承中的应用带来困难和挑战。名老中医经验的传承最终的目的是老专家临床经验和学术思想的总结和传播,数据挖掘方法的应用是有力的辅助手段,如何将相关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传承领域应用的重要方向。

中医药的传承经过了口传心授、纸质文献、电子文献的过程,但是,尚不能有效满足传承的需要。中医传承包括丰富的内容,有多种传承模式,无论何种模式,收集整理临床医

案是必不可缺的,因此,基于临床数据的循证传承是核心模式。中医药信息非标准化与个性化的特点,是中医的一大特色,但同时也是中医药传承信息化工作面临的重要瓶颈。数据的零散性与非标准性,从根本上制约了中医药传承信息化工作的效率与质量。为此,中国中医科学院中药研究所与中国科学院自动化研究所联合开发了中医传承辅助系统(traditional Chinese medicine inheritance support system)软件^[32]。该软件在全面客观地对中医药传承模式及方法特点进行分析的基础上,采用基于个性化需求的自助式服务平台的构建思路进行系统架构,以人工智能、数据挖掘、网络科学等学科的方法和技术为支撑,围绕临床诊疗和中医传承工作中的继承、发展、传播和创新4个方面的核心需求,分别构建面向临床数据的中医诊疗信息采集模块、面向中医药本体的知识管理和服务模块以及面向传承创新的隐性知识挖掘模块等几大功能模块,有效地解决了上述问题,从而保证了系统功能需求的有效实现。该软件是计算机科学和中医药学的有机结合,它以智能信息处理方面的研究成果为支撑,遵循基于临床数据的循证传承理念,提出并镶嵌了复杂系统熵聚类的挖掘方法,围绕名老中医学术思想总结和经验传承、中药新药处方筛选与处方发现等方面的信息处理进行了积极的探索^[33-34]。实现了“数据的录入-管理-查询-分析-网络可视化展示”等功能的中医传承辅助平台软件的开发与应用,在中医传承工作的具有里程碑意义,是中医传承的有力辅助工具,值得进一步推广应用。

【致谢】该论文撰写得到中国中医科学院中药研究所杨洪军研究员的指导。

【参考文献】

- [1] 王映辉,姜在阳,刘保证,等. 基于信息和数据挖掘技术的名老中医临床诊疗经验研究思路[J]. 世界科学技术——中医药现代化,2005,7(1):98.
- [2] 周春祥. 名老中医经验总结与传承过程中的问题与思考[J]. 江苏中医药,2004,25(12):1.
- [3] 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):278.
- [4] 曾孝文. 关联规则数据挖掘方法的研究[J]. 计算机与现代化,2006,9:91.
- [5] 吴嘉瑞,张冰. 试论数据挖掘方法在药品不良反应评价领域的应用[J]. 中药新药与临床药理,2007,18(6):485.
- [6] 曾孝文. 关联规则数据挖掘与发展趋势研究[J]. 电脑知识与技术,2005,35:4.
- [7] 吴荣,王阶,周雪忠. 基于关联规则的名老中医冠心病用药规律研究[J]. 中国中药杂志,2007,32(17):1786.
- [8] 刘晓怡,唐晓亮. 基于关联规则方法的李佃贵教授辨治慢性萎缩性胃炎医案分析[J]. 中国外资,2012,264:258.
- [9] 田琳,同英杰,朱建贵,等. 整理挖掘名老中医诊疗眩晕病辨证思维模式的思路与探讨[J]. 中国中医基础医学杂志,2006,12(8):618.
- [10] 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.
- [11] 吴嘉瑞,张冰,杨冰. 基于关联规则和复杂系统熵聚类的颜正华教授治疗胃脘痛用药规律研究[J]. 中国实验方剂学杂志,2012,18(20):1.
- [12] 吴嘉瑞,张冰,杨冰,等. 基于关联规则和复杂系统熵聚类的颜正华教授治疗失眠用药规律研究[J]. 中国实验方剂学杂志,2012,18(24):1.
- [13] Muirhead R J, Pu R D. A Bayesian classification of heart rate variability data[J]. Physica A,2004(336):503.
- [14] 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.
- [15] 须义贞,姚静,赵莉. 沈仲理治疗子宫肌瘤诊治规律的研究[J]. 陕西中医,2010,31(2):201.
- [16] 李艳美,张卓奎. 基于贝叶斯网络的数据挖掘方法[J]. 计算机仿真,2008,25(2):87.
- [17] 刘晓洁. 基于PCA的贝叶斯网络分类器研究[J]. 电子设计工程,2009,9:86.
- [18] 岳桂华. 人工神经网络及其在中医研究中的应用进展[J]. 大众科技,2012,14(156):165.
- [19] 陈肇文. 基于人工神经网络的中医证治模型探析[J]. 中华中医药学刊,2009,27(7):1517.
- [20] 白云静,申洪波,孟庆刚,等. 基于人工神经网络的中医证候非线性建模研究[J]. 中国中药信息杂志,2007,14(7):3.
- [21] 陆萍,林坤辉,周昌乐. 基于神经网络的中医面诊证素辨证的研究[J]. 计算机应用研究,2008,25(9):2655.
- [22] 施明辉,周昌乐. 人工神经网络在中医诊断中的应用现状与趋势[J]. 中国中医药信息杂志,2007,14(1):2.
- [23] 岳沛平. BP神经网络识别在中医脉象信号辨识系统中的运用[J]. 江苏中医药,2005,26(11):4.
- [24] 吴嘉瑞,张冰. 试论数据挖掘决策树方法在药物警戒研究中的应用[J]. 中国药物警戒,2012,9(1):29.
- [25] 瞿海斌,毛利锋,王阶. 基于决策树的血瘀证诊断规则自动归纳方法[J]. 中国生物医学工程学报,2005,24(6):709.
- [26] 谢雁鸣,朱芸茵,葛继荣,等. 基于临床流行病学调查的原发性骨质疏松症中医基本证候研究[J]. 世界科学技术——中医药现代化,2007,9(2):38.
- [27] 贺宪民,孟虹,王忆勤,等. 基于熵的决策树理论及其在中医证型研究中的应用[J]. 数理统计与管理,2004,23(5):57.
- [28] 杨洪军,赵亚丽,刘艳桥,等. “熵”在中医方证研究中的运用[J]. 中国中医基础医学杂志,2004,10(9):16.
- [29] 赵亚丽. 一类复杂系统的熵方法研究[D]. 北京:中国科学院研究生院,2005.
- [30] 孙占全. 复杂系统建模——一类智能计算模型及其应用研究[D]. 北京:中国科学院研究生院,2007.
- [31] 陈建新. 中医证候的复杂系统建模及其与疾病的相关性研究[D]. 北京:中国科学院研究生院,2008.
- [32] 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.
- [33] 唐仕欢,陈建新,杨洪军,等. 基于复杂系统熵聚堆方法的中

药新药处方发现研究思路[J]. 世界科学技术——中医药现代化, 2009, 11(2): 225.

[34] 申丹, 唐仕欢, 卢朋, 等. 含人参中成药用药规律分析[J]. 中国中药杂志, 2013, 38(12): 2029.

Comment on applications of data mining used in studies of heritage of experiences of national medical masters

WU Jia-rui¹, TANG Shi-huan², GUO Wei-xian¹, ZHANG Xiao-meng¹, ZHANG Bing^{1*}

(1. *Beijing University of Chinese Medicine, Beijing 100102, China;*

2. *Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China*)

[Abstract] Data mining, as known as knowledge discovery in databases, is a non-trivial process of revealing the implied, previously unknown and potentially useful information from the massive data. In recently years, the applications of data mining in the field of pharmaceutical research of traditional Chinese medicine have widespread. Especially in the field of the heritage of experiences of national medical masters, data mining plays an important role. In this study, we would expound of the use of methods of data mining in the heritage of experiences of national medical masters, and analyze their advantages and disadvantages, such as association rules, Bayesian networks, neural networks, and decision trees.

[Key words] data mining; national medical masters; heritage

doi:10.4268/cjcm20140411

[责任编辑 陈玲]