数据库技术在中医药领域的应用概述

★ 张媛媛 莫芳芳 张国霞* （天津中医药大学 天津 300193）

摘要：随着计算机技术和数据库技术的不断发展，其对中医药研究的辅助作用也逐步显露出来，尤其是促进了中医药学信息数字化的发展。数据库技术已经应用于中医药领域的诸多方面，如在中医文献、中医证候、中药有效成分、方剂的配伍规律研究等方面都积累可参考和借鉴的经验。

关键词：数据库；中医药；数据挖掘
中图分类号：TP 39 **文献标识码**：A

An overview of data mining used in Chinese medicine research

ZHANG Yuan-yuan, MO Fang-fang, ZHANG Guo-xia
Tianjin University of Traditional Chinese Medicine, Tianjin 300193

Abstract: With the development of computer and database technology, the supporting role of Chinese medical research can been seen, especially for the promote of information technology. Database has been applied in many aspects of Chinese medicine research.
Key words: database; Traditional Chinese Medicine; Data Mining

数据库技术产生于上世纪60年代，经过40多年的发展，已经成为计算机信息处理与应用不可或缺的一项技术支持，运用于数据处理、信息检索和人工智能等许多方面。中医药学作为一门古老的科学，要从海量信息中获得新知识，发现隐藏的关系与规律，以及对未知情况进行预测，已经不能再依靠传统的还原论方法。信息科学是促进中医药学术发展的有力手段，是中医理论进行现代化研究的一个切入点[1]。近年来，数据库应用于中医药研究由开始的单纯数据检索逐步向数据挖掘方向转变，目前已经广泛应用于中医文献整理、中药、证候诊断以及方剂组成配比等各个领域。本文对近年来数据库技术在中医药领域的应用概述如下。

1 中医古籍文献的检索与分析

中医古籍文献浩如烟海，从奠定中医基础的《黄帝内经》，到中医又一辉煌“金元四大家”的著作，再到明清集大成之医家的著作，都是中国人民五千年智慧的丰硕成果，运用数据库技术可挖掘大量尚未被发现和总结的信息。烟建华等[2]将中医基础理论数据库拟分为中医古籍原著数据库、主题词检索数据库与专题文献数据库三部分。周莲菊等[3]建立的《黄帝内经》计算机检索系统对《黄帝内经》的一次文献和二次文献进行了分类整理和加工。为学习、研究、引证提供了现代化检索手段。陈擎文[4]用频数分析和关联原则对是山人医案、吴鞠通医案、叶天士医案、临证指南医案中的中风医案进行分析。找出医案中治疗中风的药物出现频率由高到低依次是：地黄、枸杞子、当归、肉苁蓉、天麻、茯神、甘草、石斛、牛膝。常用的9种对药：肉苁蓉与枸杞子、当归与枸杞子、茯神与地黄、天麻与当归、牛膝与枸杞子、天麻与枸杞子、石斛与牛膝、肉苁蓉与地黄、石斛与枸杞子。以及常用的3种3味对药：石斛、牛膝与枸杞子，肉苁蓉、枸杞子与地黄，天麻、当归与枸杞子。赵凌等[5]对古代针灸治疗偏头疼的109条文献进行挖掘。结果显示偏头痛处方配伍中以合谷配风池出现的频次最高，其次为丝竹空配合谷，丝竹空配风池，丝竹空配合谷、风池，列缺配太渊分列第2~4位。任玉兰等[6]还采集了古代文献中

* 通讯作者

江西中医学院学报 Journal of Jiangxi University of Traditional Chinese Medicine
http://www.ajutcm.com

针灸治疗“痞证”、“痞满”、“胃痞”、“胃脘痛”等证的 393 首针灸处方为数据挖掘的样本，通过频次分析与多层关联规则分析方法寻找腧穴分布规律以及经络选择规律。结果在方中出现的总频次为 900 次，频次和支持度居前 10 位的腧穴依次为足三里、中脘、脾俞、公孙、胃俞、内关、膈俞、水分、太白、内庭，其中胃俞、内关、膈俞、水分；排在第 1 位的腧穴配伍项集为“足三里≥中脘”（其支持度与置信度分别为 5.3% 和 24.7%）；主要选取经脉集中在任脉、膀胱经、胃经和脾经。

上述分析可以看出，数据挖掘对传统中医理论的深层凝练，提供了可行的技术与方法。

2 中药研究

2.1 *药物性质的研究* 数据挖掘对于中药性质的研究主要应用频数分析、Logistic 回归分析、聚类分析、决策树等方法，来验证中药功效、性味、归经描述的科学性以及预测记载缺失的中药的性质。王虹菲等[7]对中药的生物学分类、入药部位、性味、归经、功效和免疫功能作用等进行标引，并建立中药免疫信息数据库。以中药免疫信息数据库为统计源，根据检索的结果统计不同类别的中药的数量和其中具有免疫作用的中药味数，用 Excel 处理数据，并用统计软件 SPSS 进行 Logistic 回归分析，证实了中医药理论的科学性。五加科中药中具有免疫作用的比例最高，为 60%；百合科、桑科、伞形科、毛茛科，均有 30% 以上的药物具有免疫作用。依照免疫作用强弱排列不同药性的顺序为：性平药物最高，为 27.06%；其次性寒药物为 22.19%；凉性药物为 19.23%。温热药物中具有免疫作用的药物比例比较少，而五味中淡味的药物中具有免疫作用的药物比例最高，其次为甘味，以及归心经、肾经和脾经的药物具有免疫作用的比较多。何前锋等[8]应用聚类分析对含有 4 种以上功效的单味药共有 639 种，单味药对应功效数据共有 2822 条进行分析归类，希望把具有相同或者相似功效的单味药归类在一起。周密等[9]在《中华人民共和国药典》中筛选药性与功能记录齐备的 507 味药物建立数据库，其中归肝经药物为 249 味。随机分出 95 味药为检验数据，其余 412 味药为建模数据。采用 C4.5 决策树算法预测有功能、性味记录、无归经记录的 20 味中药是否归肝经。结果得出其中 17 味药的归经，如“天山雪莲”归肝经的概率为 81.8%；沙棘归肝经的概率为 100%；而藏菖蒲不归肝经的概率为 80.8%。众所周知中药是由多种有效组分构成的整体，各有效组分发挥协同作用，相辅相成，其药效表现为综合效应[10]。而传统中药性质与药效的关系尚不清楚，数据挖掘为这一关系的研究提供了一种新方法。

2.2 *药物不良反应的研究* 随着中药研究的深入，人们开始重视药品不良反应的评价。以往文献中的不良反应分析性论文，往往仅是对不良反应病案信息进行简单的归类论述。数据挖掘作为分析可疑药物与不良事件之间关联程度的一种手段其深度挖掘作用已经得到认可，目前主要采用决策树等方法。吴嘉瑞等[11]收集了 95 种医学刊物中的 179 篇关于清开灵注射液不良反应的文章。应用 ACCESS 建立病案数据库，对患者原患疾病、过敏史及不良反应累及器官进行分析。结论清开灵注射剂不良反应类型可能存在规律性，即低年龄段相对更易发生皮肤损害类不良反应，女性患者和高年龄段患者相对更易出现过敏性休克类严重不良反应。作者[12]还收集 422 例双黄连注射剂不良反应病例，构建数据库，应用数据挖掘决策树法进行分析。结果双黄连注射剂不良反应显著特点为：50 岁以上患者群的过敏性休克构成比显著高于 50 岁以下患者群（$P<0.05$），而皮肤损害构成比显著低于 50 岁以下患者群（$P<0.01$）；20～49 岁患者群的过敏性休克构成比显著高于 0～19 岁患者群（$P<0.05$）；18 岁以上患者群的过敏性休克构成比极显著高于 18 岁以下患者群（$P<0.001$）。成年患者群中，使用注射液患者群的过敏性休克构成比显著高于使用粉针剂患者群（$P<0.01$）。决策树方法揭示出药物不良反应与患者年龄、性别、体质、过敏史、原患疾病的内在联系。对指导临床用药、确定不良反应易发人群、减低临床不良反应率有一定的指导作用。同时也对预测不良反应易累计器官与预后有指导意义。

3 方剂组成的研究

方剂具体表现为方-方、方-药、药-药、药-量，以及方药、方证、药症（证）与证症的多维关联，其本质就是一种数据关系，具有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。数据挖掘运用频数分析、Logistical 回归分析、聚类分析、关联原则的方法对药证关系以及类方进行挖掘。

3.1 *药-证关系研究* 方证关系常用回归分析的原始方法进行分析，但其必须符合大样本正态分布，使其应用极大受限。于润桥等[13]运用粗糙集强大的数据分析处理功能，结合相关回归的方法分析方剂中各药物与证侯指标之间的关系，为方证关系分析提供了一种新方法。翁晓红等[14]收集了 120 余部明清时期重要古医籍中关于脾胃湿热证方 591 条，其中治疗外感脾胃湿热证方 144 条，内伤脾胃湿热

江西中医学院学报 Journal of Jiangxi University of Traditional Chinese Medicine http://www.ajutcm.com

证方 210 条,脾胃湿热型黄疸方 93 条,脾胃湿热型泻泄和痢疾方 144 条。使用 ACCESS 建立数据库,采用频数分析和 Logistical 多元逐步回归分析,对药-证关联规则挖掘。结果分析出明清医家治疗脾胃湿热证的核心药物,合为一剂为:茯苓、黄连、橘皮、泽泻、白术、厚朴、猪苓、黄芩、甘草、苍术、半夏、滑石、人参、茵陈蒿、木香。大体由 5 个经方加减而成,即茵陈五苓散、半夏泻心汤、平胃散、香连丸和六一散。印证了明清医家治疗脾胃湿热证主要基于利水渗湿药、健脾益气药、清热燥湿药构建方剂的记载。需要看到的是数据挖掘虽然为揭示药-证关系提供了一种科学的方法,但也有其弊病:即初期数据采集较为繁琐、纳入标准不易确立,若标准选择不正确很可能影响最终结果的科学性与准确性。

3.2 *方剂类方的研究* 这里指的类方是以方类方,即由经典处方(经方与时方)繁衍发展而成的方剂系统,乃历代名医灵活运用古方之精华,亦被奉为临床处方用药之规矩与准绳。李国春等[15]用动态聚类分析研究半夏泻心汤中药量的组合规律。将 189 个相关病案聚成四类:原方用量 3 倍(I 类);只加大黄连、黄芩比例(II 类);加大党参、大枣、甘草用量(III 类);通方减少药量(IV 类)。I、II 两类以和降胃气、清热燥湿为主;III 类以补益脾胃药为主;IV 类极有可能是低年龄患者或成年患者病情较轻者的用量。张昱等[16]运用布尔关联原则的 FP-树频集算法,对所收集的宋以后的 967 首四君子汤类方进行分析挖掘。结果显示人参 + 茯苓、人参 + 白术、白术 + 茯苓、人参 + 白术 + 茯苓的配伍方式支持度最高,是四君子汤类方的核心药物组团;当归 + 人参的支持度仅次于四君子汤中的原有药物。而"人参 + 白术 + 茯苓"的组合在频繁 3 项集中支持度最高,选定其为核心药组。类方中与核心药组配伍支持度前三位的是化湿类、理气类、温里类药。叶亮等[17]运用布尔关联规则的 Apriori 算法对收集的四物汤类方 195 首进行挖掘,其中古方 74 首,今方 121 首。结论与四物汤四味药共同出现频率越高低依次为:延胡、香附、桃仁、红花、甘草(置信度 >0.3)。常出现的高频药对为香附与延胡(置信度 >0.3)。可以看到在研究类方时主要采用频数分析、聚类分析、关联原则来揭示方剂中各个组成之间的关系,确定核心药物组团。

4 中医证候研究

中医的辨证体系是一个非线性的、多维多阶的、可以无限组合的复杂巨系统[18]。从信息处理的角度看,中医辨证的诊断过程可看作是一个映射过程,即通过症状(原象)找出对应的证型集(象)的过程[19],缺乏定量标准,不易规范化。

4.1 *证候标准规范化的研究* 缺乏统一的辨证分型标准,一直是影响中医临床规范及疗效评价体系建立的关键所在。近年来,一些学者运用数据挖掘的手段进行证候标准化研究,以求确立统一的辨证分型标准,进一步提高中医诊疗评价体系的质量。宇文亚等[20]将胃炎实证 5 型、虚证 6 型各自建立单证候库(A 库)、单证候诊断标准数据库(B 库)、虚实关联匹配子数据库(C 库)。虚实主次结构通过症状体征指标程度赋值的比值反应,波动 <2/3 为绝等关联证;波动 >2/3 为主兼虚实关联证。张启明等[21]用 SAS6.12 统计软件的非条件 logistic 多元回归法对"肝气郁结"进行统计。以"肝气郁结"为自变量,筛选发病部位、病因、病例结果、常见症状及常用药物录入。结论指出在"肝气郁结"症状中依照常见度排列为:弦脉、胁痛、胁胀、脉沉、嗳气、月经紫暗、瘕聚、咽如梗物、吐咽不解、善太息、乳房胀痛。按照各症状对判断"肝气郁结"的特异性有强到弱依次为:善太息、乳房胀痛、咽如梗物、吐咽不解、胁胀、月经紫暗、胁痛、脉弦、嗳气、瘕聚、脉沉。刘刚等[22]应用偏微分效用分析方法对中医肝阳上亢证的症状敏感性进行了研究,绘制出了程度影响参考曲线。进而深挖掘得出对肝阳上亢证影响最大的变量是弦脉、烦躁易怒、头痛、口苦而渴、面目烘热、眩晕。王雪峰等[23]在进行了小儿肺炎的中医辨证规范研究的基础上,筛选出 177 项对小儿肺炎病证诊断有临床意义的证候变量。运用包括关联分析、聚类分析、分类分析、序列分析等进行数据挖掘,采用的技术包括粗糙集、模糊集、神经网络、遗传算法等,确立初步小儿肺炎诊疗判定标准。经过验证处理后,我们得到挖掘后的疗效评定标准,与现有诊疗标准数据库进行试测并最终构建小儿肺炎的中医疗效评价指标和方法体系。沈亚诚[24]首先通过聚类分析等方法对原有绝经综合征数据库进行数据清洗,选择具有代表性的指标进行研究。以聚类分析和主成分分析对中医绝经综合征肾阴虚型证和西医症状进行对比,然后分别分别建立中医疗效矩阵和对照的西医疗效矩阵。根据欧氏距离公式,计算出两序列的距离 =1.30,距离比较小,表明两序列比较相似。基于欧氏距离,分别对中医疗效序列、西医疗效序列和生存质量指标序列进行聚类,245 个病例的三类指标序列的聚类结果基本相同。发现中医药疗效分析结果与生存质量分析在有效性方面表现基本一致,表明采用中医证侯评分量表进行疗效评价比

江西中医学院学报 Journal of Jiangxi University of Traditional Chinese Medicine
http://www.ajutcm.com

较合理。综上所述,运用数理统计方法研究中医证候标准,有助于对大量中医辨证信息的简化处理和其中潜在规律的挖掘,从而可能得到更为客观的结论。

4.2 人工神经网络 人工神经网络是一种新型的数据挖掘技术,它是将整体论与还原分析方法有机结合的研究复杂系统的有效方法,能够有效处理复杂系统中杂乱无章的海量数据,并能够在海量数据中寻找模式,寻找规律,归纳隐含在信息单元之间的关联规则。叶进等[25]运用人工神经网络,采用头痛、怕冷、身重、恶风、弦脉、洪脉、浮脉、发热、出汗 9 种症状指标,诊断湿温、暑湿、太阳伤寒、太阳中风四证。人工神经网络经过训练后,用现有的病案检验系统检测,效果良好,且获得了很多难以通过人工获取的规则。吴芸等[26]建立八刚辩证神经网络,验证结果表明中医八纲辨证神经网络能够比较准确地对 43 例病例进行中医八纲证型分型,其符合率达到了 90% 以上。实验证明经过训练的人工神经网络可以较准确地进行辨证,为临床症候诊断提供了一种辅助手段。

5 结语

综上所述,中医药与计算机技术的结合是适应中医药信息学的必然结果,为中医药学发展提供了先进的研究方法与辅助手段。从以往中医数据库运用中可以看出大部分都是简单信息储存、检索功能,近年来向数据挖掘深度发展的趋势初步展露出来,并且已经取得一些成果。此外,要从海量的中医药信息中发现新的规律与知识,只采用单一的数据源难以提取出隐含的大量信息,采用多组数据库融合技术还有待进一步研究开发。

●学术广角●

参考文献

[1]任廷革. 浅析中医与信息科学[J]. 世界科学技术——中药现代化,2000,2(3):42 -44.

[2]烟建华,王彤,任廷革,等. 中医基础理论数据库的学术建构思路[J]. 中医教育,2004,23(4):47 -49.

[3]周莲菊,赵敏,刘绍武,等.《<黄帝内经>计算机检索》系统[J]. 天津中医,2002,19(4):77 -78.

[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,2008,26(10):2254 -2257.

[5]赵凌,任玉兰,余毓如,等. 基于数据挖掘技术分析古代针灸治疗偏头痛的经穴特点[J]. 中国中医基础医学杂志,2008,14(10):774 -776.

[6]任玉兰,赵凌,刘迈兰,等. 基于数据挖掘探析古代针灸治疗功能性消化不良的选穴特点[J]. 辽宁中医药杂志,2009,36(2):259 -262.

[7]王虹菲,启贵,刘辉. 构建中药免疫数据库及对中药免疫规律的研究[J]. 中草药,2007,38(2):297 -301.

[8]何前锋,周雪忠,周忠眉,等. 基于中药功效的聚类分析[J]. 中国中医药信息杂志,2004,11(6):561 -562.

[9]周 密,王耘,乔延江. 利用数据挖掘方法预测中药缺失药性的初步研究[J]. 中国中医药信息杂志,2007,15(6):93 -94.

[10]张培,李江,王耘,等. 贝叶斯网络在中药有效组介五味预侧中的应用[J]. 世界科学技术—中医药现代化,2008,10(5):114 -117.

[11]吴嘉瑞,张冰. 清开灵注射液不良反应文献信息的数据挖掘研究[J]. 药物流行病学杂志,2009,18(2):78 -81.

[12]吴嘉瑞,张冰. 双黄连注射剂不良反应文献的数据挖掘研究[J]. 中国药物警戒,2008,5(3):139 -143.

[13]于润桥,杨琳瑜,王建红,等. 基于粗糙集理论的中医方证系统模型的建立[J]. 辽宁中医药大学学报,2007,9(1):5 -7.

[14]翁晓红,肖林榕,杨雪梅,等. 明清时期脾胃湿热证的用药规律研究[J]. 中医杂志,2007,48(5):440 -442.

[15]李国春,戴慎. 动态聚类分析在中医方剂药量组合规律中的应用[J]. 中国卫生统计,2006,23(1):63 -64.

[16]张昱,陈云慧,王燕,等. 布尔关联规则挖掘四君子汤类方药物配伍规律研究[J]. 江苏中医药,2008,40(7):67 -68.

[17]叶亮,范欣生,卞雅莉,等. 古今治疗痛经的四物汤类方关联规则研究[J]. 南京中医药大学学报,2008,24(2):94 -96.

[18]张志斌,王永炎. 证候名称及分类研究的回顾与假设的提出[J]. 北京中医药大学学报,2003,26(2):1 -4.

[19]施明辉,周昌乐. 一种基于 ANN 的中医辨证不确定性推理模型研究[J]. 计算机工程与应用,2007,43(7):10 -13.

[20]宇文亚,沈舒文,彭宇,等. 胃粘膜异型增生虚实关联证数据库建立方法学探讨[J]. 中华中医药学刊,2008,26(1):80 -81.

[21]张启明,王永炎,张志斌,等. 中医历代医案数据库的建立与统计方法[J]. 山东中医药大学学报,2005,29(4):298 -299.

[22]刘刚,陈国林. 基于数据挖掘的肝阳上亢证的症状敏感性研究[J]. 计算机工程与应用,2008,44(20):238 -241.

[23]王雪峰,董丹,梁茂新,等. 数据挖掘技术在小儿肺炎中医临床疗效评价研究中应用的思路与方法[J]. 中国中西医结合杂志,2007,27(10):949 -951.

[24]沈亚诚,王小云,傅昊阳,等. 基于数据挖掘的绝经综合征中医药临床疗效评价研究[J]. 中国卫生统计. 2008,25(4):387 -390.

[25]叶进,邢传鼎. 基于人工人工神经网络的病症诊断原型系统[J]. 华东大学学报,2003,2(4):43 -47.

[26]吴芸,周昌乐. 中医八纲辨证神经网络构造技术研究[J]. 计算机工程与应用,2008,44(18):243 -248.

(收稿日期:2009-09-16)

欢 迎 投 稿 ! 欢 迎 订 阅 !

江西中医学院学报 Journal of Jiangxi University of Traditional Chinese Medicine
http://www.ajutcm.com