

数据挖掘技术及其在中药分析中的应用

张 铭

贵阳中医学院基础部数微教研室, 贵州省贵阳市市东路 50 号, 550002)

通讯方式: maggiezhangm@163.com

[摘 要] 简要概述了数据挖掘技术的定义、应用范围以及常用的数据挖掘方法, 并着重介绍数据挖掘在中药分析研究领域中的应用。应用数据归约技术、聚类的方法、模糊集理论改进中医药数据的质量, 使得在预处理后的中药方剂数据库中成功挖掘出重要信息, 为研制中药新药提供决策支持。

[关键词] 数据挖掘; 中药分析; 数据归约; 数据预处理

数据库中的知识发现或称为数据挖掘, 旨在从大量原始数据中挖掘出隐含的、有用的、尚未发现的知识(如规则、模型、规律、模式、约束等), 帮助决策者寻找数据间潜在的关联, 找出人工难以发现的因素。中药数据挖掘是中医药数字化工程的重要环节。中医药研究中, 方剂是中药处方的基本数据单位, 它包含方剂的整体性能、性、味、归经、毒性等等信息。每个方剂针对某一种类型的病证, 并以特定的治疗机理(证)对症施治。中药类方含数万个方剂, 用数据挖掘技术提取方剂数据共性, 挖掘出符合中医药理的概念描述具有重要现实意义。

1 数据挖掘技术

1.1 数据挖掘的概念

数据挖掘是一个集统计学、人工智能、模式识别、并行计算、机器学习、数据库等技术于一体的交叉性学科研究领域。数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中, 提取隐含在其中的、人们事先不知道的、有用的信息和知识的过程。通过数据挖掘可以帮助决策者寻找规律, 发现被忽略的要素, 预测趋势, 进行决策^[1]。它涉及对数据库中的大量数据进行抽取、转换、分析以及模型化处理, 从中提取辅助决策的关键性数据。因此, 数据挖掘就是深层次的数据信息分析方法。

1.2 数据挖掘的基本过程

数据挖掘的基本过程包括数据收集、数据整理、数据挖掘、结果评估、分析决策等, 其基本过程与步骤如图 1 所示^[2]。从图 1 可以看出, 数据挖掘的各过程不是一次就能完成的, 其中某些步骤或者全过

程可能需要多次的循环反复, 才有可能达到预期的效果。

1.2.1 数据收集 大量全面丰富的数据是数据挖掘的前提, 没有数据, 数据挖掘也就无从谈起。因此, 数据收集是数据挖掘的首要步骤。在开始数据挖掘之前, 首先要消除噪声或不一致数据, 广泛收集用户的各种信息, 建立数据库与数据表, 为数据挖掘做准备。数据可以来自于现有事务处理系统, 也可以从数据仓库中得到。

1.2.2 数据整理 数据整理是数据挖掘的必要环节。由数据收集阶段得到的数据可能有“污染”, 表现在可能存在自身的不一致性, 或者有缺失数据的存在等, 从数据库中检索与分析任务相关的数据, 确保数据能够真实反映待挖掘的对象, 因此, 数据的整理是必需的。数据挖掘通过汇总或聚集操作将数据变换或统一成适合挖掘的形式, 可单独利用也可综合利用各种数据挖掘方法对数据进行分析, 挖掘用户所需要的各种规则、趋势、类别、模型等。

1.2.3 结果评估 数据挖掘的结果有些是有实际意义的, 而有些是没有实际意义的, 这就需要进行评估。这样对发现的规则、趋势、类别、模型进行评估, 可以保证发现的模式的正确性。分析决策数据挖掘的最终目的是辅助决策。决策者可以根据数据挖掘的结果, 结合实际情况, 调整竞争策略等, 并将挖掘结果以可视化的形式展现在用户面前。

1.3 数据挖掘的方法^[3]

1.3.1 分类 分类在数据挖掘中是一项非常重要的任务。分类的目的是通过统计方法、机器学习方法(包括决策树法和规则归纳法)、神经网络方法等构造一个分类模型, 然后把数据库中的数据映射到给定类别中的某一个中去。



图 1 数据挖掘的基本过程与步骤

1.3.2 聚类 聚类是把一组个体按照相似性归成若干类型,即“物以类聚”。其目的是使同一类别之内的相似性尽可能大,而类别之间的相似性尽可能小。教育知识管理外化过程的下一层次包含了更为强大的搜索工具和文件管理系统,它们对贮存的知识进行分类,并能识别出各信息源之间的相似之处。基于此,可用聚类的方法找出教育信息知识库中各知识结构间隐含的关系或联系。

1.3.3 关联规则 关联规则挖掘就是发现大量数据中项集之间有趣的关联或相关联系。关联规则广泛应用于决策支持系统。关联规则分为布尔关联规则与量化关联规则,这区分于它们所处理值的类型。并根据其涉及的属性维数分为单维关联规则或多维关联规则。这方面比较有影响的算法有 Apriori 算法等。关联规则是教育研究的一种重要分析手段,如分析学生口头表达能力与各学科成绩之间关系。

1.3.4 Web 页挖掘 随着互联网的迅速发展及 Web 的全球普及,使得 Web 上的信息量无比丰富,通过对 Web 的挖掘,可以利用 Web 上的海量数据进行分析,并根据分析结果找出企业管理过程中出现的各种问题,对这些信息进行分析和处理,以便识别、分析、获取、存储、传递信息,并将其转化为知识。

2 数据挖掘在中药学领域中的应用

中药药对的运用一直是凭借着医者经验的积累,其配伍规律一直不为人们所了解。在中医药现代化呼声日渐高涨的今天,不能仅停留于传统的经验及理论上,应该借助数据挖掘等技术方法揭示其科学内涵。目前国内利用数据挖掘技术对药对的配伍规律研究尚处于一个起步阶段。笔者认为,药对本身具有多维性,如药对的性味特点、药对的剂量特点、药对的功效特点,以及药对主治特点等。这些各异的特点,可以将其视为不同的维度加以分析,通过对此类数据进行多维分析来发现药对配伍的规律。中药数据的特点给数据挖掘带来的问题有:①如何从长文本字段中分离出病证、药名及药剂量。②为保证各属性含义唯一,如何处理“同药异名”问题。③中药剂量的数值和度量表示混乱,如何合理处理数值及寻找一个科学的度量单位。④为更好地实施分类,如何尽量减少一个方剂中的药味数而又不影响结果正确性。为解决上述问题,顺利对中医药数据进行分类数据挖掘,笔者参考文献[4]的方法进行预处理。

2.1 方剂规范化

方剂规范化主要完成方剂组成正交分解和消除

冗余的处理。“方剂组成”是一个长文本类型的字段,由一些无规则的数据组成,不利于数据挖掘的进行,因此,须对其进行正交分解,将每一味中药及其剂量分别存放,消除异名,以及清除不重要药名消息。

2.2 剂量规范化

为实现数值型的数据挖掘,进行离散化数值、异名处理、概化归约、药物剂量加权、模糊化处理。离散化数值时,若是离散型数值就直接提取,若是连续数值则取平均值。在异名处理时,中医药专利数据库中的数据都是由手工录入的,由此导致数据存在两个基本问题:①中药异名,同一种中药在不同医学工作者的处方中可能以不同的名称出现;②中药的无序性,药方中各味中药的录入顺序是随机的。为了解决这两个问题,本研究引入了“中药词典”这一映射表。中药词典中收录了现有的所有中药及其在同一种中药名称的所有同义词,并给每一味中药指定了唯一的索引,借用中药词典就可使得存在的两个问题得到顺利解决。概化归约时,方剂组成中存在的多药一来源问题,无疑会加重数据挖掘的负担,因此,本研究采用了数据归约技术,得到数据集的压缩表示,它比源数据集小得多,但仍然保持原数据的完整性,这样在归约的数据集上挖掘将更有效,并能产生相同或几乎相同的分析结果。药物剂量加权时,在数据中,剂量是以质量为衡量的,挖掘结果最终还是要体现在药效上。然而中药的药性比较复杂,四气和五味、归经、升降沉浮及有毒、无毒等方面,均对药效产生影响。为了得出的结果更合理,将每味药的有效成分含量系数与药质量相乘的结果作为药物剂量,对药物剂量进行了修正,更准确地从数值上体现药效。

数据的模糊化时,所要实现的目标是得到如下规则:治疗某种疾病的几味中药及其重量会同时出现。在不同的方剂中同一味中药的含量是不同的,因而药物剂量就形成了一个连续的数值型变量,而某一个含量的出现有其偶然性,不足以说明在此草药含量是对于治疗某种疾病的最佳药剂量,因此,不能使用某一个具体的含量作为分析的数据,而应在某一数据范围内讨论药剂用量,减少偶然性对最终结果的影响。为此借用模糊的概念将具有相近意义的数据划分到同一个模糊区间内。对于每一味草药的用量,都使用此隶属函数来计算隶属度。在分析中药用量时,用“L(低)”、“M(中)”、“H(高)”3个模

糊区域来表示药剂用量。首先,利用 k-means—平均聚类方法将所有用量分为 3 个簇,得到每个簇的中心值:0.0571、0.2047、0.5840。据 3 个中心值本研究给出了关于药剂量的隶属函数,如图 2 所示^[5]。

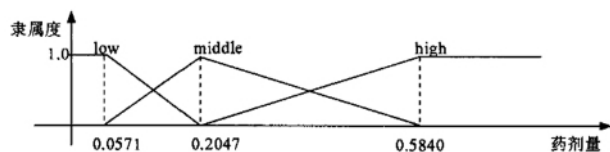


图 2 在本例中使用的隶属函数

3 结语

中药自身的特点和中药配伍的多样性决定了其不同于西药的研发,多种不确定的因素使数据挖掘技术在中药研究过程中的应用变得困难。文章提出了一套数据预处理的方法,结合了数据归约技术、聚

类方法、模糊集理论成功将中药专利数据规范化,为数据挖掘技术应用于中医药研制领域提供参考。

参考文献

- [1] Margaret H Dunham. 数据挖掘教程[M]. 北京:清华大学出版社,2005:3-5.
- [2] 范明,孟小峰. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2003:189-215.
- [3] 袁方,王煜,王丽娟,等. 实用数据挖掘[M]. 北京:电子工业出版社,2004:82-84.
- [4] 朱金伟,鞠时光,辛燕. 基于数据挖掘的中医药数据预处理方法[J]. 计算机工程,2006,32(15):280-282.
- [5] 钱增瑾,辛燕. 中医药数据预处理方法的设计与实现[J]. 计算机工程与设计,2005,12(16):3199-3200.

(收稿日期:2010-12-30;修回日期:2011-02-18)

[编辑:黄 健]

“2011 全国中医妇科临床诊疗新技术高级研修班”通知

为提高我国中医妇科诊疗技术水平,加强学术交流。由中华中医药学会继续教育部、中医杂志社主办,中华中医药学会妇科分会协办,北京东方博联经济信息咨询中心承办的“2011 全国中医妇科临床诊疗新技术高级研修班”,拟于 2011 年 8 月 12 日~14 日在湖南省长沙市举办,届时将邀请我国中医药行业相关领导、中医妇科领域知名专家等列席会议并授课。本次“研修班”主题为“传承名老中医临证经验,挖掘科技创新诊疗成果”,旨在加强全国各地中医妇科领域学术交流与合作,进一步提高中医妇科工作人员理论、临床诊疗水平及科研创新能力。本次“研修班”为 2011 年中华中医药学会继续教育项目,参会代表经考核合格,将颁发继续医学教育 I 类学分 10 分。

1. 拟邀请以下专家授课:尤昭玲(国家首席健康科普专家,中华中医药学会妇科分会主任委员,世界中医药学会联合会妇科分会常务副会长,国家中医药管理局重点学科中医妇科学学术带头人、博士生导师;题目:IVF-ET 中医辅助疗法临床运用);罗颂平(博士生导师,中华中医药学会妇科分会候任主任委员;题目:复发性流产的中医防治);赵新广(湖南中医药大学教授、博士后、研究生导师,中华中医药学会妇科分会常务委员、秘书长;题目:妇科方药安全性研究现状分析及对策思考);杜慧兰(主任医师,博士研究生导师,中华中医药学会妇科分会副主任委员,世界中医药学会联合会妇科专业委员会副主任委员,河北省中医药学会妇科分会主任委员,国际传统与现代生殖医学协会副主席;题目:子宫内膜异位症的中医治疗);谈勇(江苏省中医院主任医师、教授,中华中医药学会妇科分会副主任委员,江苏省中医药学会妇科专业委员会主任委员;题目:妇科名老中医夏桂成教授对排卵障碍性疾病的认识与诊治对策)。

2. 论文征集与交流:为使会议内容更具实用性,方便各位参会代表相互学习和交流,本研修班自即日起将在全国中医妇科领域征集优秀论文,并编辑《优秀论文集》,撰稿内容参见会议主题,凡入选《优秀论文集》的论文将颁发优秀论文证书,符合刊用标准的将优先在《中医杂志》上发表。

3. 投稿要求:论文必须真实可靠,引文准确,2500 字左右;文末附注作者简介:姓名、年龄、性别、职务、职称、地址、电话、传真、电子邮箱等;每篇论文收取工本费 300 元(只接收电子文稿);根据来稿内容和会议安排,将符合要求的论文安排大会或专题分组会上宣读。

4. 学习班费用:1200 元/人(含培训费、资料费、餐费、合影及继

续教育学分证书等),住宿统一安排,费用自理。

联系电话:010-68625036(兼传真)、13522333787,联系人:李阳。

网址:www.jtcm.net.cn,投稿信箱:zyzz1955@163.com。

邮 购 信 息

《中医杂志》光盘合订本 1951~1986 年(4 张光盘)480 元;1987~1994 年(1 张光盘)260 元;1995~2004 年(1 张光盘)280 元。光盘合订本为计算机用数据光盘,收入了同期出版的全部《中医杂志》的内容,可对所有文章进行任意字检索,方便实用。

《〈中医杂志〉“专题笔谈”文萃》收集了本刊“专题笔谈”栏目 1995~2004 年 10 年间关于单味中药单独应用或以单味中药为主配合其他药物治疗各类疾病的临床治疗经验共 900 余篇文献。

“专题笔谈”是《中医杂志》特色栏目之一,汇集了全国各地中医师丰富的临床实践经验,其中不乏对一些疑难病症的独得治法,也包括了一些常用中药的非常规用法,深受广大医务人员和读者的欢迎,对于临床医生临证用药具有较高的参考价值。全书共分为两辑,定价 83 元(免邮寄费)。

联系地址:北京市东直门南小街 16 号中医杂志北京医海林音像书店,邮编:100700,电话:(010)64014411 转 3053,64013162,联系人:崔京艳、邓媛。请在汇款单附言栏注明所购书名及联系电话。

欢迎邮购《中医杂志》

1997 年 2、4、6、8~12 期,每册 3.50 元;1998 年 3~8、10~12 期,每册 4 元;1999 年 4~12 期,每册 4 元,合订本 65 元;2000 年 2~12 期,每册 5.80 元,合订本 85 元;2001 年 2~12 期,每册 6.50 元,合订本 95 元;2002 年 1~12 期,每册 6.80 元,合订本 105 元;2003 年 1~12 期,每册 6.80 元,合订本 105 元;2004 年 1、2、4~12 期,每册 6.80 元,合订本 115 元;2005 年 1~7、9、11、12 期,每册 6.80 元,合订本 125 元;2006 年 1~12 期,每册 7.30 元,合订本 135 元;2007 年 2~12 期,每册 8.80 元;2008 年 2~12 期,每册 9.80 元;2009 年、2010 年每册 9.80 元。2007 年~2010 年合订本每年 140 元。《中医杂志》2006~2009 年增刊每册 40 元。以上免费邮寄,如需挂号每件另加 3 元。

欲购者请汇款至北京东直门内南小街 16 号中医杂志社读者服务部,邮编:100700,电话:(010)64013164。