基于关联规则挖掘的方剂配伍规律初步研究

马金刚,胡志帅,曹慧*,来建梅
(山东中医药大学理工学院,济南 250355)

[摘要] **目的**:构建结构化的方剂信息数据库,挖掘方剂中药物之间的关联规则,进一步研究方剂的配伍规律。**方法**:运用基于正则表达式的信息抽取方法,以 vs 2008 编程实现自动抽取,并以 SQL Sever 2005 进行关联规则挖掘。**结果**:生成了结构化的方剂信息数据库,并初步抽取方剂信息 81 305 条;在最小支持度为 10%,最小置信度为 70% 的条件下,得到高频药物 9 味,生成规则 4 条。**结论**:基于正则表达式的信息抽取方法使全面大规模研究方剂配伍规律成为可能,挖掘结果与中医理论和临床经验总体相符,为进一步研究提供也依据。

[关键词] 关联规则; 方剂配伍; 信息抽取
[中图分类号] R285 [文献标识码] A [文章编号] 1005-9903(2013)07-0351-03
[doi] 10.11653/zgsyfjxzz2013070351

Preliminary Study of Prescription Compatibility Law based on Association Rule Mining

MA Jin-gang, *HU Zhi-shuai*, *CAO Hui**, *LAI Jian-mei*
(*Institute of Science and Technology*, *Shandong University of Traditional Chinese Medicine*, *Ji'nan* 250355, *China*)

[**Abstract**] **Objective**: This paper aims to build a prescription information database, mining the rules of drugs in the prescription, and study the compatibility law. **Method**: Using information extraction based on regular expressions and vs 2008 programming to realize automatic extraction, and realized association rules mining in SQL Sever 2005. **Result**: Generated a structured prescription information database, and initial extracted prescription information 81 305. When the support degree is 10% and the confidence level is 70%, 9 kinds of high frequency drus and 4 rules were geted. **Conclution**: The method of information extraction based on regular expressions make it be possible that make a full-scale study of prescription compatibility law. The results of data mining and the theory of traditional Chinese medicine and clinical experience were matching overall, which lay a foundation for further study.

[**Key words**] association rules; prescription compatibility; information extraction

目前运用计算机相关技术对方剂配伍规律进行研究,是一种新的研究方法和研究方向,其中应用最多的计算机技术是关联规则挖掘。关联规则属于数据挖掘的范畴,是数据挖掘中的一项重要技术,它是目前应用于方剂配伍规律研究的一种最经典的方法[1]。运用数据挖掘技术研究方剂配伍规律,旨在中医理论指导下,通过相关数学方法与计算机技术的结合,从大量的、不完全的、有噪声的、模糊的、随机的数据中挖掘出其中已经存在的、隐含的、未被发现的但又潜在有用的、最终可被理解的知识的过程[2]。本文以主治咳嗽的方剂为例,尝试运用关联规则挖掘对方剂配伍规律进行初步研究。

1 一般资料

1.1 实验工具 本实验所用的编程和数据处理工

[收稿日期] 20120509(003)
[基金项目] 济南市科技局自主创新计划项目(200906007)
[第一作者] 马金刚,硕士,讲师,从事智能信息处理与软件工程研究,Tel:13012989908,E-mail:ma_jingang@126.com
[通讯作者] *曹慧,硕士,教授,从事生物医学信息处理与分析、医学虚拟现实研究,Tel:0531-89628102,E-mail:caohui63@163.com

具为 Visual Studio 2008 专业版,Microsoft SQL Sever 2005 企业版;参考书为彭怀仁所著的《中医方剂大辞典》[3]。

1.2 关联规则算法相关知识

1.2.1 Apriori 算法基本思想 关联规则挖掘包括两个主要问题,即发现频繁项集和生成关联规则,其中发现所有的频繁项集是生成关联规则的基础[4],这一过程运用到的最经典的算法,即是 Apriori 算法。Apriori 算法的基本思想是,基于频繁项集的先验知识,使用逐层搜索的迭代方法 k-项集搜索(k + 1)-项集。首先要找出频繁 1-项集的集合 L1,并以 L1 找出频繁 2-项集的集合 L2,再以 L2 找到 L3,以此类推,直至不能找到频繁 k-项集为止[5]。

1.2.2 Apriori 算法相关参数 支持度:表示同时包含药物 A 和药物 B 占总的百分比,即数据库中 A,B 两味药同时出现在同一首方中的比例。反应了二者关联的可行性。置信度:表示包含药物 A 和药物 B 占只包含药物 A 的百分比,即 A,B 两药同时出现占 A 出现总频数的比例。反映了二者关联的可靠性。概率:规则 A→B 的概率用项集{A,B}的支持度除以{A}的支持度,即数据挖掘中的置信度。重要性:又叫做兴趣度或增益,其表示方法为,规则 A→B 除以 notA→B 的概率。重要性表示药物 A,B 的关联程度。

2 研究方法

2.1 方剂信息数据库的构建 信息抽取的来源分为两个大的方面,即中医药书籍和中医药网站。纸质书籍信息的抽取,主要通过扫描设备扫描后生成电子文档,并借助 OCR 技术识别页面文字。中医药网站信息的抽取,主要是通过编程解析网页结构,然后书写抽取规则,进行自动抽取。根据方剂网站的内容编排及其具有名称、功效、主治、出处等信息的统一结构,本文通过书写正则表达式制定对应的抽取规则,如方剂名称对应的正则表达式为:

"(?:[^>]|(?! >\d + <) >) + >(\d +)(?:[^<]|(?! <strong) <) + 名称"。

方剂配伍规律的数据挖掘研究需要在大量的方剂数据的基础上进行[6]。本文以 Visual Studio 2008 专业版为开发工具,通过上述基于正则表达式的信息抽取方法,初步获取结构化的方剂信息 81 305 条,并存储到本地数据库即方剂信息数据库中,为数据挖掘的进行提供了充分的数据准备。

2.2 挖掘结构的构建 在原始的方剂信息数据库中,方剂中药物组成是作为表的一个字段结构并列出现的,而要挖掘方剂中药物之间的关联规则,这样显然是不符合要求的,需将原有的表拆分成药物组成表(zuchengtable)和方剂主治表(fangjitable)。由于每首方剂中组成的中药达数味到十数味,这样大量的数据,手工操作是不现实的,本文以 Visual Studio 2008 专业版为开发工具,通过编写相关程序实现。其中药物组成拆分的部分代码如下:

```
private void getzhongyao(int id)
        {
            fangjiinformation fj = new fangjiinformation ();
            fj = fangjiinformationManager.Getfangjiinformation(id);
            string zuchengtable5 = fj.zuchengtable;
            Regex reg = new Regex(@"([^)] + .", RegexOptions.Compiled | RegexOptions.IgnoreCase);
            string zuchengtable1 = reg.Replace(zuchengtable5, "");
            string[] zuchengtable2 = zuchengtable1.Split(new char [] { ',', '、'}, StringSplitOptions.RemoveEmptyEntries);
            for (int i = 0; i < zuchengtable2.Length; i++)
            {
                Regex reg1 = new Regex(@"\d. *", RegexOptions.Compiled | RegexOptions.IgnoreCase);
                string zhongyao = reg1.Replace(zuchengtable2[i],"");
                Zuchengtable com = new Zuchengtable();
                com.fjid = id;
                com.zhongyao = zhongyao;
                ZuchengtableManager.Add(com);
            }
        }
```

生成的药物组成表(zuchengtable)字段包括序号(id)、中药序号(zcid)和中药名称(zhongyao),方剂主治表(fangjitable)字段包括序号(id)、方剂序号(fjid)、方剂名称(name)和方剂主治(zhuzhi)。其中两表的 zcid 和 fjid 均是沿用原始方剂信息表中的方剂编号,故以其作为逻辑主键。本文以 Microsoft

SQL Sever 2005 企业版为数据挖掘工具进行关联规则挖掘,其中所用的算法即为 Apriori 算法。挖掘结构的构建,是通过以药物组成表(zuchengtable)作为源(主键)表,以方剂主治表(fangjitable)作为目标(外键)表,两表嵌套来实现的。嵌套表挖掘结构和逻辑关系如图 1 所示。

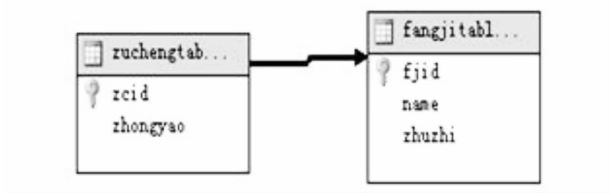

图 1　嵌套表结构和表间逻辑关系

3　研究结果

本文以主治咳嗽的方剂为例,进行关联规则挖掘。在最小支持度为 10%,最小置信度为 70% 下条件下,频繁项集见表 1,关联规则见图 2,关系网络见图 3。

表 1　药物频繁项集

支持度/次	项集	支持度/次	项集
1 324	甘草	543	贝母
926	杏仁	422	前胡
763	人参	414	陈皮
736	桔梗	412	五味子
626	半夏		

图 2　以主治咳嗽的方剂关联规则

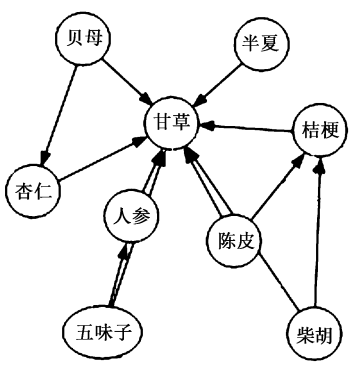

图 3　高频药物依赖关系网

4　讨论

关联规则研究方药的配伍规律,可以在不同层次发现药物的组合使用情况,发现其治疗思想,并可指导临床用药[7]。

本研究前期通过基于正则表达式的信息抽取,以《中医方剂大辞典》作为参考依据,生成了结构化的方剂信息数据库;这种通过编程实现的智能处理方式,为大规模研究方剂配伍规律提供了良好的技术支持,使得进一步全面研究方剂配伍规律成为可能。后期运用 SQL Sever 2005 技术进行关联规则挖掘,在设定的支持度和置信度下,得到高频药物 9 味,生成关联规则 5 条。这一结果从数据挖掘角度来说是成立的,且与中医理论和临床经验总体相符。如高频药物甘草,典型的具有清热解毒、祛痰止咳的功效;高频药物苦参,具有镇咳、平喘的作用。如生成的规则,前胡,枳壳 - > 桔梗,其置信度为 74.7%,重要性为 62.7%,具有较高的可信性;3 味药单独使用,均具有宣肺、祛痰、止咳等功效,参阅《中医方剂大辞典》,可以发现,在实际的临床应用中,此 3 味也经常配伍使用,分别用以治疗各类咳嗽症状,因此,实验结果在实际的方剂配伍中也具有一定的参考价值。然而,与人们印象中的中医理论和临床经验相比,某些结果仍存在着一定的差别。关联规则挖掘的意义或许就在于此,能够揭示一些隐藏的、未被发掘的潜在规律,但也有可能是因为噪声、数据的不完整等原因导致的部分结果失真,需要进一步的验证和改进。

[参考文献]

[1] 宫俊,董俊龙,梁茂新,等. 基于关联规则的广义药对最适合病症的挖掘方法[J]. 东北大学学报:自然科学版,2011,32(8):1097.

[2] Venkatadri M, Lokanatha C R. A review on data mining from past to the future [J]. Inter J Computer Applications, 2011,15(7):19.

[3] 彭怀仁. 中医方剂大辞典[M]. 北京:人民教育出版社,2005.

[4] Cai C H,Fu A,Cheng C H,et al. Mining association rules with weighted items[R]. IEEE DEAS. [s. l.]:[s. n.], 2002:67.

[5] 杨烨,邢斌,高成勉,等. 基于数据挖掘的二陈汤类方关联分析[J]. 中国中医药信息杂志,2009,16(11):89.

[6] 尚尔鑫,范欣生,段金廒,等. 基于三维图形化数据挖掘方法的四物汤类方配伍规律研究[J]. 中国实验方剂学杂志,2011,17(1):217.

[7] 吴荣,刘晛,王阶,等. 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.

[责任编辑　邹晓翠]