doi:10.3969/j.issn.1006－4931.2021.10.001

中药性味归经及药用功能的数据挖掘研究*

张好霞[1],侯 钰[1],杨建明[1],陈 浩[2],李天泉[2△]

(1. 重庆康洲医药大数据开发应用研究院,重庆 404100; 2. 重庆康洲大数据有限公司,重庆 404100)

摘要:目的 通过中药的性、味、归经预测其药用功能。**方法** 采用样本分析法,调取药智网数据库的中药基本信息数据表(共收录中药1 690 种)数据,进行预处理、清洗、变换,并参考 Apriori 算法,结合关联数据技术,分析中药的性、味、归经、药用功能之间的联系。**结果** 中药的药性以平性最多,温性、寒性次之,药味的苦味与药性的热性关联性极强,主要涉及祛风湿药,与中医常识"苦味属寒凉"不一致。平性、辛味药易入肾经,且多用于祛风湿;温性、咸味药易入肾,用于补虚;凉性、淡味药可清热;热性、苦味药可祛风湿等。**结论** 该研究可为大数据技术下中药有价值信息的探索和性能预测研究提供借鉴,但仍有待进一步论证。

关键词:数据挖掘;药性;药味;归经;药用功能;关联规则;Apriori 算法

中图分类号:R932;R285.1 **文献标志码:**A **文章编号:**1006－4931(2021)10－0001－04

Data－Mining of the Nature, Flavour, Channel Tropism and Medicinal Function of Traditional Chinese Medicinal Herbs

ZHANG Haoxia[1], HOU Yu[1], YANG Jianming[1], CHEN Hao[2], LI Tianquan[2]

(1. Chongqing Kangzhou Pharmaceutical Big Data Development and Application Research Institute, Chongqing, China 404100; 2. Chongqing Kangzhou Big Data Co., Ltd., Chongqing, China 404100)

Abstract: Objective To predict the medicinal function of traditional Chinese medicinal herbs through their nature, flavour and channel tropism. **Methods** By using the sample analysis method, the basic information data of traditional Chinese medicinal herbs (including 1 690 kinds of traditional Chinese medicinal herbs) in the Yaozhi database were retrieved, pre－processed, cleaned and transformed, and the relationship among the nature, flavour, channel tropism and medicinal function of traditional Chinese medicinal herbs was analyzed by referring to Apriori algorithm and linked data technology. **Results** Most traditional Chinese medicinal herbs have mild nature, followed by warm nature and cold nature. The bitter flavour has a strong correlation with the hot nature of traditional Chinese medicinal herbs (mainly related to herbs for removing wind－dampness), which is inconsistent with the common sense of "bitter flavour has correlation with cold nature". Traditional Chinese medicinal herbs with mild nature and acrid flavour are easy to enter the kidney channel and are mostly used to remove wind－dampness, those with warm nature and salty flavour are easy to enter the kidney channel and are used to tonify deficiency, those with cool nature and light flavour can clear heat, those with hot nature and bitter flavour can remove wind－dampness. **Conclusion** This study can provide a reference for the exploration of valuable information and performance prediction of traditional Chinese medicinal herbs through big data technology, but it still needs further demonstration.

Key words: Data－mining; nature; flavour; channel tropism; medicinal function; association rules; Apriori algorithm

在现代科学技术推动中医药理研究快速发展的背景下,药性、药味理论的研究已逐渐成为国内学者关注的热点,促进了中药性、味、归经、功能间关系数据探究方法的创新[1－4]。目前,数据挖掘已成为中医药研究的重要技术,主要是用于发现数据中的隐含规律和潜在的信息及知识等。关联规则挖掘则是数据挖掘技术中较活跃的技术,一般用来发现研究对象的关联性和相互依存性。其中 Apriori 算法是关联规则挖掘中较典型的算法[5],也是挖掘布尔关联规则频繁项集最有影响的算法,已被广泛用于中药数据挖掘的研究中[6－8]。关联规则的数据挖掘方法多用于药物属性间的关系挖掘,郭小磊[8]、于红艳等[9]报道了部分四性－五味间的紧密关系,以及寒－解毒、寒－清热、温－止痛、寒－凉血、微寒－清热、寒－消肿、凉－清热等的关联关系;尚尔鑫等[10]报道了部分中药属性四性－归经、归经－归经的关联性。本研究中在现有理论研究的基础上,运用数据挖掘方法进行样本分析、处理,并参考 Apriori 算法,结合关联数据技术[10－11],分析中药的性、味、归经及药用功能间的联系,探讨中药性能和传统经验之间的区别和联系,为中药性能研究提供科学、有效的论证方式。现报道如下。

1 数据处理

1.1 数据来源

本研究中涉及的数据来源于药智网数据库的中药基本信息数据表,表中信息主要来源于《中国药典(一

*基金项目:重庆市基础研究与前沿探索项目(重庆市自然科学基金)[cstc2018jcyjAX0047]。

第一作者:张好霞,女,大学本科,研究方向为医药数据挖掘,(电子信箱)1311256345@qq.com。

△通信作者:李天泉,男,硕士,高级工程师,研究方向为生物化学工程,(电子信箱)ltq007@163.com。

部)》及全国各省市自治区药材标准信息,且详细记录了中药材加工、规范、归类等操作,信息权威,可直接利用度高(数据加工规范等过程有内部标准,统一性强)。同时,中医药数据挖掘过程中需要针对性强、数据规范、数据量相对完备的专业数据库,而药智网数据库中的中药基本信息数据表收录了1 690种常见中药,记录了其名称、药性、药味、归经、功效、炮制等基本信息,现抽取其中性、味、归经、功能作为研究对象进行数据挖掘。

1.2 数据预处理

研究中需涉及中药药性、功效等信息,由于其记录不统一、各种信息的名称不规范等原因,造成数据显示的内在规律会呈过度分散状态,使得现有数据库不能直接按研究目的进行数据分析或数据挖掘,所以很有必要在数据分析前对数据进行预处理,数据预处理一般包括数据清洗、数据变换、数据集成等步骤[12]。

数据清洗:主要包括初步清洗和深度清洗两部分。初步清洗主要是指除去数据当中冗余、无关的标点符号、空格等部分,这类错误常由人工录入造成,规律性差,需人工逐一查找、规范并加以改善。深度清洗主要是指对中药的性、味、归经、功能进行清洗,去掉"有小毒"等无关成分;将属性"微寒"变换为"寒";将归经的名词简化,如"大肠经"简化为"大","脾经"简化为"脾","膀胱经"简化为"膀"等。数据清洗过程需提高数据的准确性、完整性和简化性,尽可能地提高数据的质量。

数据变换:是指对某些信息描述中属性值数目不唯一且有一定多样性的数据进行转换处理,将其简化。原数据中每味中药的某个属性可能有多个值,且数目不相同,如漏芦的介绍为"味苦性寒,归胃经,有清热解毒之功效,可消痈,下乳,舒筋通脉";薤白的介绍为"味辛、苦,性温,归心、肺、胃、大肠经,有通阳散结、行气导滞的功效"。分析发现,漏芦药味属性有1种、归经有1种,而薤白药味属性有2种、归经有4种,而数据库中的数据无明显规律,且组合种类繁多,不易于算法识别。故极有必要将原始数据分割、展开,以变换为简单的数据形式。变换后保留2种主要药味、1种主要药性、3种主要归经、3种主要功能,按一一对应原则分别将其展开,共得$2\times1\times3\times3=18$项。经过处理后的数据形式简洁明了。数据处理的难点还在于药物功效的分类。一般药物有主要功效和次要功效,但原有数据的功效多而杂。若仅将功效进行分割和初步清洗,所得功效将超过600种,需将其合并带入Apriori算法中,然后将功能分类后的数据带入算法。此种数据处理方式可降低项集绝对支持度,对于含有项较多的项集,其相对支持度基本无改变,对于含有项较少的项集,其相对支持度和绝对支持度均有改变。故对于最后的结果,项数越多的频繁项集越可靠。在进行算法分析时,可将功效设为目标,设置其他3项为输入项。

数据集成:主要是将来源、格式、特点性质不同的数据在不同的系统定义数据元素,并将这些数据元素在结构化的模式上有效协调存在的差异,同时保持一致的数据视图,最终使得数据共享更加便捷化。

2 模型算法原理及使用工具

中医理论中对"药对"在七情、性、味、归经、功效等多个角度的组成方法及形式均有涉及,但均仅对部分药对进行了阐述,其间还存在部分信息交叉。虽合乎中医药临床的实际操作,但与药对组成结构的现代理论研究之间还存在一定的距离[5]。从药对组成药物的属性入手,利用基于关联规则的数据挖掘方法,可以探究中药性、味、归经、功能属性间的关系,并可描述每两个属性间的关联性强弱。

Apriori算法是关联规则算法中最常用的算法。关联规则的原理是,在数据集中,若大量记录具有特征属性A的同时,也频繁出现特征属性B,则称特征属性A和B构成模式,这些模式可以用关联规则来观察和分析,从而表现A和B之间的关联性[4]。关联规则的质量一般由规则的支持度(support)和置信度(confidence)来度量,它们分别反映所发现规则的有用性和确定性[11]。规则XY在数据库D中的支持度是交易集中同时包含X和Y的事务数与所有事务数之比,记为$\mathrm{support}(XY)=\mathrm{support}(X\cup Y)$,可简化表示为$P(A\cup B)$。支持度描述了X和Y项集在所有事务集D中同时出现的概率。规则X和Y在事务集中的置信度是指同时包含X和Y项集的事务数与包含X项集的事务数之比,它用来衡量关联规则的可信程度,记为$\mathrm{confidence}(XY)=\mathrm{support}(X\cup Y)\mathrm{support}(X)$,可简化表示为$P(B|A)$,即在事务集D中出现项集A的同时,也出现项集B的概率[13-14]。本研究中采用MySQL和Excel作为数据处理工具,R语言、SPSS作为算法挖掘工具。

采用样本分析法,随机抽取一定数量的数据样本,设置合理的参数,将不同样本数据带入算法模型运行,直到得出最理想的结果,则该模型为成功的预测模型。

3 模型调试与预测结果分析

3.1 样本1

随机抽取700种中药,根据性、味、归经进行展开,得到最终数据1 767条。数据以csv格式储存,并带入SPSS运行算法。设置性、味、归经、功能作为关联规则算法模型项集的属性,设置算法模型最低支持度为0.5%,置信度为60%,调用R关联分析规则包中的算法进行调试,详见图1。

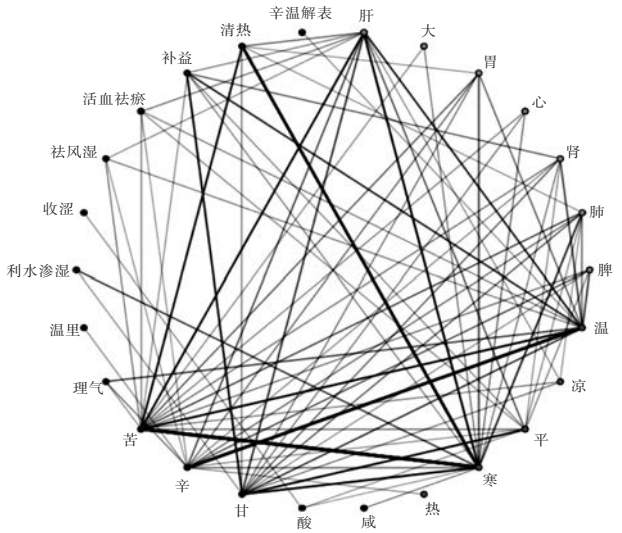


图 1　700 种中药性、味、归经、功效网络图

Fig. 1　Network diagram of the nature, flavour, channel tropism and medicinal fundion of 700 kinds of traditional Chinese medicinal herbs

通过网络节点，可直观显示各属性间的关联程度，其中每个原点各表示 1 种属性，其间的直线表示原点间有关联，直线越粗，关联程度越强[15-16]。由图 1 可知，苦－寒、辛－温、寒－清热间关联性很强，其次，苦－清热、甘－补益、苦－肝间的关联性较强，表明补益功能与温、甘、肾，利水渗透功能与寒，理气功能与温有关联。

利用模型对功能的预测结果可得关联规则频繁项集(见表 1)，分析可知，设置支持度为 0.01%，置信度为 55%时，预测结果最多为 148 项，其中预测准确数目为 0。因此，样本量为 1 767 建立的模型无法得到理想的效果，需进一步扩大样本量对模型进行调试。

3.2　样本 2

抽取全部 1 690 种中药，总数据量为 4 637 条，以 8∶2 的比例选取其中 80% 为训练集，20% 为预测集。设置中药性、味、归经、功能为关联规则算法模型项集的属性，将算法模型的最低支持度设置为 0.2%，置信度设置为 60%，先调用 R 关联分析规则包中的算法模型，将训练集数据带入算法进行调试，详见图 2。可知，关联强度较强的是清热－苦、清热－寒、温－辛、肝－苦。同时，补虚与温、肝、肾间有一定关联，祛风湿与温、苦、辛、平、肝有一定关联，行(理)气与温有轻微关联，利水渗透与寒的关联较弱。

表 1　样本 1 关联规则频繁项集

Tab. 1　Frequent item sets of association rules for sample 1

前项	后项	支持度(%)	置信度(%)	提升
热，甘	温里	0.623	100.000	28.030
咸，肾，温	补益	0.510	88.889	6.568
淡，寒	利水渗湿	0.906	81.250	11.660
肾，甘，温	补益	1.472	73.077	5.400
小，寒	利水渗湿	1.019	66.670	9.572
热，苦	祛风湿	0.680	66.667	9.344
咸，温	补益	1.472	65.385	4.831
热，心	温里	0.623	63.636	17.830
凉，肾	补益	0.623	63.636	4.702
包	活血化瘀	0.453	62.500	8.057
热，胃	温里	0.453	62.500	17.520
膀，甘，寒	利水渗湿	0.453	62.500	8.974
小，甘	利水渗湿	0.566	60.000	8.615
淡，平	利水渗湿	0.566	60.000	8.615

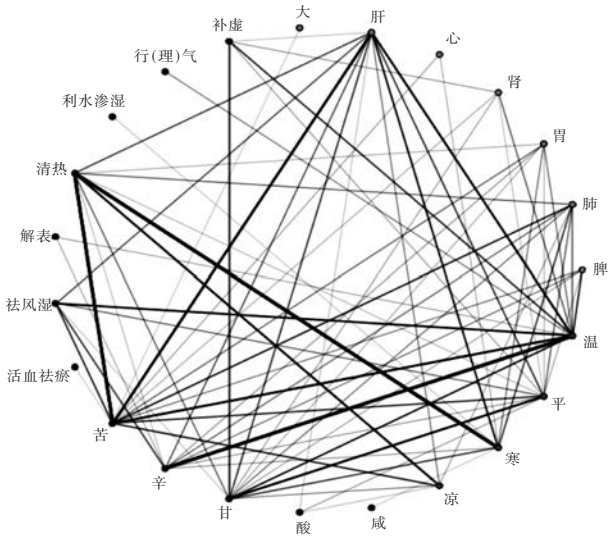


图 2　1 690 种中药性、味、归经、功效网络图

Fig. 2　Network diagram of the nature, flavour, channel tropism and medicinal fundion of 1 690 kinds of traditional Chinese medicinal herbs

利用模型的预测功能预测训练集，对比结果得出准确率，预测准确率作为模型评估选择依据。通过改变支持度和置信度来实现对模型的调试，直至出现最大准确率和预测的最大数量，该模型即为最优模型。预测结果见表 2，基于准确率和预测的数量，考虑采用支持度为 0.2%、置信度为 65% 的模型。将预测集数据带入模型，对数据进行运算，所得频繁项集见表 3。

4　讨论

中药的药性以平性最多，温性、寒性次之。关联分析

表 2　模型预测结果

Tab. 2　Prediction results of model

置信度(%)	支持度(%)	预测准确数	准确率(%)	置信度(%)	支持度(%)	预测准确数	准确率(%)
60	1.0	114	64.04	60	0.03	468	73.88
60	0.8	147	67.57	50	0.1	1 600	54.66
60	0.6	147	67.57	55	0.1	633	65.51
60	0.4	241	70.83	65	0.6	71	75.08
60	0.2	303	70.86	65	0.4	129	75.56
60	0.1	420	71.12	65	0.2	170	73.26
60	0.05	452	71.46	65	0.1	244	72.45

表3 样本2关联规则频繁项集

Tab. 3 Frequent item sets of association rules for sample Ⅱ

前项	后项	支持度(%)	置信度(%)	提升
平,辛,肾	祛风湿	0.237	90.909	5.978
咸,肾,温	补虚	0.582	85.185	7.128
淡,凉	清热	0.755	82.857	3.704
胆,凉	清热	0.237	81.818	3.658
热,苦	祛风湿	0.431	80.000	5.261
热,甘	温里	0.302	78.751	39.593
辛,温	行(理)气	0.561	69.231	10.730

结果显示,药效清热与药味苦,药效清热与药性寒,药性温与药味辛,药入肝经与药味苦有很强的关联;药效补虚与药性温及药入肝、肾经,药效祛风湿与药性温、甘有较强的关联,验证了中医中的"辛味和甘味属温,苦味属寒凉"的理论知识。

根据关联规则分析,置信度超过80%的关联节点有平、辛、肾-祛风湿,温、咸、肾-补虚,凉、淡-清热,淡、胆-清热,热、苦-祛风湿。这表明药性为平、药味为辛的易入肾经,且多用于祛风湿;药性为温、药味为咸的易入肾,用于补虚;凉性、清淡的可清热;药性为热、药味为苦的可祛风湿等。上述分析结果有效地验证了中医古籍理论。

苦味药在传统中医理论中药性寒凉,本研究中发现,药味苦与药性热有很强的关联性,且这些药物主要为祛风湿药,这与中医常识"苦味属寒凉"有明显不同,证明了中医药中味苦药材并非全部属寒凉药性。这一结论有待中医药专业人员进一步研究和论证,并期待发现有创新、全新的中医药应用。

在大数据技术日益成熟的时代,中医药数据不断聚合汇总,形成完整海量的大数据中医药平台,将挖掘模型应用于这些具有极高价值的数据中,研究者不但可挖掘出更多、更有价值的信息,而且对于具有缺陷的中药信息记录书籍或资料,合适的数据挖掘模型也可发挥预测作用加以补全,从而提高中药学资料的完整性。

当下,数据挖掘技术突飞猛进,并在诸多领域发挥作用。中医药领域有极其庞大的数据资源,且中药资料或书籍里的数据之间也存在各种关联,这种关联恰好是数据价值的核心所在。使用关联规则挖掘能将分散的中药数据关联起来,发掘性、味、归经与功能的关联性和数据之间的依存性。该方法仅需分析人员利用自己的经验及知识结构对数据进行一定规范,并带入模型分析即可得出结果,可大幅降低人为主观因素的影响。同时,该分析方法使用线将点与点进行连接,并用线的粗细程度来显示点与点间的关联强度,线条较粗的可用于验证经典书籍或资料中理论的准确性,线条较细的可用于探究目前研究较浅或未知的归经、功能,可能会更便利、有效地获得新的信息。并且,此方法可直观、生动、形象地展现性、味、归经、功能间的关联程度。

利用大数据挖掘方法研究中医药学,建立模型,规避研究人员本身的经验或知识结构固化的限制,可挖掘出现有中医药学之外的潜在信息,大幅提高研究思路的创新性和多样性,从而更快、更优地挖掘出分散的传统中药数据背后潜藏的价值,推动创新中医药发展,传承中医精神。

参考文献

[1] 杨谨成,王会凌,费小非. 大数据在临床药学的相关应用与展望[J]. 中国药业,2017,26(19):1-4.
[2] 王 建,张廷模,曾 南,等. 中药药性理论的研究与思考[J]. 辽宁中医杂志,2008,35(2):212-213.
[3] 杨 欣,李亚辉,李 专,等. 现代的科学技术与方法在《中药药理学》药性研究中的应用[J]. 教育教学论坛,2019(48):214-215.
[4] 王振国,张 冰,曾英姿,等. 中医药理论的近代嬗变及其影响——以本草诠释方法为视角[J]. 山东中医杂志,2019,38(1):1-8.
[5] 牛丽敏. Apriori算法分析与改进综述[J]. 桂林电子科技大学学报,2007,27(1):27-30.
[6] 郑 斌. Apriori算法改进及其在资产管理中的应用[J]. 重庆科技学院学报(自然科学版),2019,21(5):54-56.
[7] 洪卫兰,陈桂强,王 持,等. 基于Apriori算法的自动化药房盒装药品储位优化管理[J]. 中国药业,2020,29(22):14-17.
[8] 郭小磊. 数据挖掘技术在中医方剂科学问题研究中的应用探究[J]. 网络安全技术与应用,2019(10):123-124.
[9] 于红艳,许成刚. 关联挖掘技术在中药药性及其他属性间关系的应用研究[J]. 中国方剂学杂志,2013,19(14):343-346.
[10] 尚尔鑫,叶 亮,范欣生,等. 基于改进关联规则算法的中药药对药味间性味归经功效属性关系的发现研究[J]. 世界科学技术—中医药现代化,2010,12(3):377-381.
[11] 路立峰,解 玲,郑 丹,等.《中国药典》成方制剂的数据挖掘及其问题分析[J]. 中医药临床杂志,2019,31(11):2107-2111.
[12] 张尧庭,谢邦昌,朱世武. 数据采掘入门及应用[M]. 北京:中国统计出版社,2001:1-108.
[13] 戴稳胜. 关联规则挖掘概述[J]. 统计与信息论坛,2009,17(55):5-7.
[14] 李 虹,蔡之华. 关联规则在医疗数据分析中的应用[J]. 微机发展,2003,15(6):94-97.
[15] 廖开际. 数据仓库与数据挖掘[M]. 北京:北京大学出版社,2008:1-251.
[16] 帅文昊,杨 颖,张珊珊,等. 基于数据分析挖掘中医治疗失眠的用药规律[J]. 湖南中医杂志,2021,37(1):117-119.

(收稿日期:2020-11-20)