Available online at www.sciencedirect.com
ScienceDirect
本文经编委遴选,英文版将通过 ScienceDirect 全球发行。

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*

□尚尔鑫** 叶 亮 范欣生 唐于平 段金廒**
（南京中医药大学江苏省方剂研究重点实验室 南京 210046）

摘 要:中药药对是联系方剂和中药的重要环节。虽然传统中医理论对药对中药味的组成方式有多角度的论述,但相比论著丰富的方剂理论还不够系统。本文分别采用标准关联规则发现 Apriori 算法以及改进多数据库计算方法，对从历代药对文献中收集整理得到的 625 个药对 347 味药中包括性味、归经、功效等共 49 个属性形成的数据库进行挖掘研究,并对两种方法得到的结果进行比较。结果表明,本文提出的新方法更适用于分析药物间的关联规则,集中度更好,更易于发现有价值的关联规则。

关键词:中药药对 组成结构 关联规则 Apriori 数据挖掘
doi: 10.3969/j.issn.1674-3849.2010.03.013

中医方剂是中医药治疗疾病的主要形式，在从单味药应用到多味药组成方剂的过程中，药对随之形成,并不断充实和发展。“药对”是临床上常用且相对固定的中药配伍形式,是方剂最小的组方单位,由两味中药组成,与单味药之间有密不可分的关系,却不是简单的一加一的关系。药对的组成是以中医药基本理论为指导,着重结合中药本身的性能和功用,选择性的将两味中药进行组合配对而成。在中医理论中药对的组成结构有七情和合、性味配伍、功效配伍以及阴阳、刚柔、引经等多角度的论述。本研究小组基于多年方药研究的认识和体会，提出了从单味药-药对-基本方-类方体系,由简单到复杂、由浅入深、逐层递进的研究思路和方法,认识到药对从药味到方剂形成发展过程中所具有的重要性和蕴含的科学内涵[1~2]。

针对药对在中药及方剂研究中的重要地位,本文对药对中两味药物之间的关联关系进行了数据挖掘研究，分别采用了标准的关联规则 Apriori 算法和改进的数据库间关联规则发现算法以及概率检验,试图在药味的性味、归经,以及功效等属性间发现有意义的规则，为分析药对内部的组成关系提供数据支持。

收稿日期: 2009-12-24
修回日期: 2010-05-08
* 江苏省高校自然科学研究计划重大基础研究项目(06KJA36022):四物汤及类方治疗妇科血瘀证的配伍规律与物质基础研究,负责人:段金廒;国家自然科学基金项目(30902015):方剂图形化技术导向下的方剂效应成分群发现方法的构建,负责人:尚尔鑫;江苏省中医药局科研项目（HZ07048）：中药性味归经配伍规律的可视化方法研究，负责人：尚尔鑫；江苏省中医药局康缘中医药科技创新基金项目(HZ0801KY):特征性关联方药数据库与知识挖掘,负责人:范欣生。
** 联系人:尚尔鑫,讲师,博士,主要研究方向:方剂信息数据挖掘研究,E-mail: shex@njutcm.edu.cn;段金廒,本刊编委,教授,主要研究方向:方药功效物质基础与配伍规律研究、中药资源及资源化学研究与开发,Tel:025-85811116,E-mail:dja@njutcm.edu.cn。

PDF 文件使用 "pdfFactory Pro" 试用版本创建 www.fineprint.cn

一、数据挖掘方法建立

Apriori 算法是应用非常广泛的一种关联规则挖掘方法，其核心是基于频繁集理论的一种递推方法，最早用于顾客交易数据库中项集间的关联规则。该算法的优点是原理简洁、程序简单易懂，但运算量较大，需要多遍扫描数据库。此后，针对 Apriori 算法的不足之处出现了多种改良算法，大大减少扫描数据库的次数，提高运行效率。

但各种改进算法的目的几乎都是为了提高运行效率，计算对象仍然是单一数据库。在本文的研究中，药对是由两种相互独立的药味组成的，每味药物具有其各自的属性，即药对需要两个独立的数据库才能准确描述。若应用 Apriori 算法，需要预先将两个数据库合并，会造成部分信息损失。针对这种情况，本文提出了改进的关联规则发现算法，以用于在两个数据库间发现有意义的属性关联。

1. 标准关联规则方法

Rakesh Agrawal 等[3] 采用标准 Apriori 算法进行挖掘。为了方便本文中的其他研究，支持度同时给出规则 A→B 出现的次数及在全部数据库中的比率两个数据。置信度同时计算 A→B 及 B→A 两个数据。

2. 数据库间关联规则方法

标准 Apriori 算法只能在一个数据库上进行关联规则挖掘。药对由两味药物组成，每味药物都有其各自的属性，应用 Apriori 算法时必须将两味药物的属性合并形成一个数据库再进行挖掘。这种合并可能对挖掘结果造成影响。本文对标准 Apriori 算法进行了改进，采用分步挖掘的策略，使其能在两个数据库间进行关联规则挖掘，避免出现单个数据库中的关联。此外，为避免药对中两个药物排列顺序造成的影响，还需要对挖掘结果进行合并，例如，肝→脾和脾→肝两个规则都表示肝经药物与脾经药物搭配，应将二者合并。

3. 自然出现概率和调整支持率

若某一个规则有特定意义，则构成该规则的两个属性应具有特定的联系或禁忌，即这两个属性联系更加紧密或几乎没有联系。由于各个属性在药物中的分布并不均匀，例如，在本文 347 味药物中，清热药有 68 味，而收涩药仅有 14 味，直接从挖掘结果的支持度很难发现属性间特有的联系或禁忌。某一规则的支持度高，既可能由于这两个属性间具有特定的联系，也可能是这两个属性的药物数较多，而药物数量较多自然该规则出现的支持度就高。

为了避免上述两种可能，尽可能的发现属性间特有的联系或禁忌，本文引入了属性、规则的自然出现概率和调整支持率两个概念。属性 A 的自然出现概率 $P_n(A)$定义为全部药物中具有该属性药物的出现频率，即：$P_n(A)=n_A/N$，其中 n_A：具有属性 A 的药物数，N：全部药物数。

药对中的两味药物可以看作相互独立的随机过程，则规则 A→B 的自然出现概率 $P_n(AB)$可定义为两个属性出现概率的乘积，即：$P_n(AB)=P_n(A)\times P_n(B)$。

规则的自然出现概率相当于支持率的“基础”水平或正常水平，反映了当两个属性间不存在任何特异性的联系时所具有的支持率。因此，能够反映某一规则特定联系的调整支持率 P_{adjust} 定义为规则实际支持率与其自然出现概率之差：$P_{adjust}=P_{support}-P_n$。

若某一规则的调整支持率为零，表明该规则两个属性间不存在任何特定的联系或禁忌，若不为零，正值表明这两个属性具有特定的关联，负值表明两个属性相互禁忌，不能互相搭配。

4. 调整支持率的统计检验

调整支持率是否为零代表了两个属性间的特定联系。检验调整支持率是否为零，也就是检验某一规则的支持率是否和其自然出现概率相一致。将自然出现概率看作全部药对的总体率，任一药对均有存在或不存在该规则两种可能，其支持度符合以自然出现概率为总体率、支持度为抽样数的二项分布。基于二项分布的概率累积函数，可计算出每一规则调整支持率为零的概率 P_{val}。选择置信水平 $\alpha=0.05$，则 $P_{val}<0.05$ 表明该规则两个属性间具有特定的联系，具有研究价值。

5. 工具软件

本文药物属性数据来自于本实验室自建的中药方剂数据库，以 Microsoft SQL Server 2000 为后台服务软件，可提供单机版检索及 Web 访问服务。数据挖掘程序全部自行编写，运行于 Matlab R2009a (MathWorks)。

二、数据挖掘结果

1. 药对数据来源

《神农本草经》提出了对中药“七情和合”的认识，东汉《伤寒杂病论》中已有许多应用药对治病的

PDF 文件使用 "pdfFactory Pro" 试用版本创建 www.fineprint.cn

经验，不少沿用于今，明代《本草纲目》中进一步丰富了药对的应用，到近代秦伯末在《谦斋医学讲稿》[4]对药对进行了总结，列出了 81 对两味中药的配伍。吕景山编著的《施今墨对药临床经验集》[5]，收编对药 277 对，分 24 类，论述其组成、功用[6]。陈维华等著《药对论》[7]，列举药对 400 余对，按功效特点进行阐述。

基于多部药对著作，本研究去除重复、异名、罕见药物、多于两味药物等情况的药对，共收集 625 个有效药对，其中组成药物 347 味。根据本实验室开发的中药方剂数据库确定每味药物的性、味、归经以及功效类别等属性，形成属性数据库，以备挖掘研究。

性包括寒、热、温、凉、平 5 个属性值，味包括辛、甘、酸、苦、咸 5 个值，归经为常见的十二归经，功效类别包括清热药、解表药、化痰止咳平喘药等 27 类。全部共 49 个属性值，分别以逻辑值 1 表示药物存在该属性，0 不存在。每味药物的属性以一个 49 个元素逻辑向量来表示全部属性。

药对由两味药物组成，将两味药物的属性合并，使每个药对同样获得一个 49 元素的属性逻辑向量。例如，麻黄桂枝药对，麻黄性温，味辛苦，归肺、大肠、膀胱、心包经，功效为解表、利水药；桂枝性温，味辛甘，归肺、肝、肾、膀胱、心经，功效为解表药，则麻黄桂枝药对的属性值为性温，味辛甘苦，归心、肝、肺、肾、心包、大肠、膀胱经，功效类别为解表、利水。

2. 药对组成药物频率分析

本研究采用 Apriori 算法计算全部 625 个药对中组成药物的出现频次和频率。在 625 个药对中，共出现 347 味药物，出现频次最高的前 15 味药物见表 1。

黄连药对最多，共 32 个，占 5.1%。可见各药对在药物分布上几乎不存在特定的药物优势，各药物分布比较分散，集中趋势不明显，且绝大部分药物出现频次都较低。

3. 药对组成药物属性频率分析

对全部药对形成的 625 行 49 个属性的数据库。采用标准 Apriori 算法对其进行关联规则发现，出现频次最高前 10 位属性分布见表 2。

归经属性出现较多，在前 10 位属性中归经就有 6 个。五味中甘、苦出现较多，排在第 4、7 位，辛味排在第 8 位。四性中只出现一个温，排在第 9 位。功效属性全部未出现在前 10 位，这可能与功效类别较多（27 个）有关，使得药对在功效上分布相对比较分散。功效中最多的是清热药，排在第 13 位，占 36.48%。可见，药对中在归经属性上分布较为集中，特别是前 3 位的肝、肺、脾、经都在 70%以上。各属性支持度都较高，均大于 50%，这可能是由于中药通常都同时具有多个属性，使各属性支持度都较高，变化趋势不明显。

4. 药对组成药物属性间关联规则分析

在单属性频率分析的基础上，进一步利用 Apriori 算法进行了属性间的关联规则挖掘，结果见表 3。

各规则中肝经、脾经药物搭配出现最多，占 66.24%。肝、脾、肺、胃 4 个归经间相互搭配较多，在前 10 位规则中占据 6 位，这直接与表 2 中这几个归经频率最高有关。除归经间关系外，甘、辛、苦味与脾、肺、肝经搭配出现也比较多。各规则的双置信度都在 80%左右，其中肝、脾搭配的置信度均在 80%左

表 1 药对中出现频次最高的前 15 味药物

序号	药物	支持度	支持率(%)
1	黄连	32	5.10
2	人参	24	3.80
3	半夏	22	3.50
4	大黄	20	3.20
5	麻黄	20	3.20
6	白术	19	3.00
7	当归	18	2.90
8	石膏	18	2.90
9	附子	17	2.70
10	陈皮	17	2.70
11	甘草	16	2.60
12	白芍	16	2.60
13	香附	16	2.60
14	生地黄	15	2.40
15	黄芪	15	2.40

表 2 药对中出现频次最高的前 10 位属性分布

序号	属性	支持度	支持率(%)
1	肝	523	83.68
2	肺	497	79.52
3	脾	493	78.88
4	苦	435	69.60
5	胃	428	68.48
6	心	425	68.00
7	甘	420	67.20
8	辛	399	63.84
9	温	358	57.28
10	肾	340	54.40

PDF 文件使用 "pdfFactory Pro" 试用版本创建 www.fineprint.cn

右，说明二者存在着较为紧密的搭配关系，即在全部含脾经药物药对中，超过 80%的药对含有肝经药物，反之亦然。单个置信度最高出现在辛味与肝经的搭配上，达到 88.97%，但置信度 2 为 67.88%，在前 10 位规则中几乎最低。这说明几乎全部含辛味药物的药对中都存在归肝经药物，但在全部含肝经药物的药对中辛味药物仅占 74%。可见，虽然肝、脾、肺、胃经属性间搭配较多，但搭配的专一度不高，比较广泛的和各属性药物搭配，而相对来说，甘、辛、苦等属性与肝、脾、肺经属性间的专一度较高。总体上各规则虽然支持率较高，但分布分散，支持率下降缓慢，属性间专一性不够高，发现有意义的药对结构关系存在一定的困难。

5. 药对组成药物属性间关联规则分析

研究采用改进的分步式 Apriori 算法分析药对中各属性间的关系。药对中两个药物各形成一个属性数据库，直接分析两个数据库间各属性的关联规则。为消除两个药物排列顺序带来的影响，需对结果进行合并，例如，将肝→脾和脾→肝两条规则进行合并，只保留其中一条，重新计算支持度和支持率。数据库间属性关联，结果见表 4。

与表 3 类似，表 4 出现较多的仍然是肝、脾、肺、胃等归经，肝、肺归经药物搭配排在首位。甘、辛味药物及心经药物未出现在前 10 位规则中。表 3 中最高的肝、脾组合支持度为 66.24%，在表 4 中下降为 58.72%，说明表 3 中的规则可能并不一定是在药对中两个药物间产生的。例如，若某一药物同时归肝、脾经，则不论该药物与其它什么药物搭配，都对肝、脾搭配规则产生贡献。这种情况对我们研究药对中药物间的结构关系会产生一定影响。

此外，表 4 中出现了肝经、肝经药物组合，即药对由两个同归肝经的药物组成，支持率为 47.36%，置信度比较接近，分别为 71.5%和 73.09%。该规则对研究药对的结构关系是有意义的，却无法通过标准 Apriori 算法得出。前 10 位规则的支持率从 59.04%下降到 47.36%，约下降了 12%，相对于表 3 中 8%的下降速度，该结果的离散程度更好，规则更加集中，便于发现规律。可见，本文采用的数据库间分步式关联规则挖掘方法更适用于药对组成结构的挖掘研究。

各规则的自然出现概率各不相同，其调整支持率也不相同。表 5 是按照表 4 结果调整支持率降序排列后的结果。

重新排序后，表 4 排在首位的肝、肺组合下降到了第 8 位，而肝、肝组合则由第 10 位上升到了第 2 位。可见，表 4 中结果未排除自然出现概率的影响，出现频率高的属性在规则中同样出现较多，未能反映出特异性的属性关联。在表 5 中，调整支持率排除了属性自然出现概率的影响，使各规则的关联体现在同一基础水平上，能更好地反映出属性间的特异关联。

表 3　药对属性间出现频率前 10 位关联规则

序号	属性 1	属性 2	支持度	支持率(%)	置信度 1(%)	置信度 2(%)
1	肝	脾	414	66.24	79.16	83.98
2	肝	肺	409	65.44	78.20	82.29
3	脾	肺	394	63.04	79.92	79.28
4	苦	肝	360	57.60	82.76	68.83
5	辛	肝	355	56.80	88.97	67.88
6	肺	胃	354	56.64	71.23	82.71
7	心	肝	352	56.32	82.82	67.30
8	脾	胃	347	55.68	70.59	81.31
9	肝	胃	345	55.20	65.97	80.61
10	甘	肺	343	54.88	81.67	69.01

表 4　药对中药物间出现频率前 10 位属性关联规则

序号	属性 1	属性 2	支持度	支持率(%)	自然概率(%)	调整支持率(%)	Pval	置信度 1(%)	置信度 2(%)
1	肝	肺	369	59.04	36.42	22.62	0	65.54	66.13
2	肝	脾	367	58.72	37.57	21.15	0	68.86	67.46
3	脾	肺	332	53.12	30.61	22.51	0	65.87	62.88
4	苦	肝	319	51.04	28.32	22.72	0	59.74	63.04
5	肝	胃	310	49.60	29.09	20.51	0	57.73	56.88
6	肝	甘	307	49.12	29.29	19.83	0	58.03	56.23
7	苦	肺	304	48.64	23.07	25.57	0	58.80	60.20
8	胃	肺	301	48.16	23.70	24.46	0	62.06	61.93
9	苦	脾	297	47.52	23.81	23.71	0	57.67	57.34
10	肝	肝	296	47.36	44.70	2.66	0.08439	71.50	73.09

PDF 文件使用 "pdfFactory Pro" 试用版本创建 www.fineprint.cn

三、讨 论

中医理论中对药对的组成方式在七情、性味、归经、功效等多个角度都有所论述，但各种方式都只对部分药对做出了论述，其间还有部分交叉。这种方式符合中医药临床实践，但与药对组成结构的现代研究之间还存在距离。基于关联规则的数据挖掘方法从药对组成药物的属性入手，可以建立对药对结构的现代认识与传统理论之间的联系。

在关联规则挖掘中，最常用的就是 Apriori 算法。该算法可对事务性数据库进行挖掘，得到相互关联的属性，并可计算支持度和置信度，以得到较为可靠的结果。Apriori 算法需要多遍扫描数据库，效率较低，已有多种改进算法报道，主要集中在改进算法效率上，减少扫描数据库的遍数。但各种算法都是在一个数据库上进行挖掘[8]。

本文中研究药对中药物间的结构关系，适合应用关联规则算法挖掘，并同时运用了标准 Apriori 算法和改进的分步式挖掘算法进行挖掘。在应用 Apriori 算法时，必须将两个药物属性数据库进行合并，而分步式算法则可以直接在两个数据库中发现规则。从结果对比看（表 3 和表 4），Apriori 算法结果普遍支持度较高，集中趋势不明显，在前 10 个关联中支持度仅下降了 4%。改进算法结果在前 10 个关联中支持度下降了 12%，集中趋势较为明显。由于 Apriori 算法要求合并数据库，合并后难以分辨某一关联是发生在两个药物之间还是来自于同时具有两个属性的一个药物。因此，Apriori 算法结果较为分散，集中度不高。此外，Apriori 算法不可能发现同一属性关联，例如，药对中两个肝经药物的肝-肝关联，合并后只表现为同一个肝经属性，自然无法发现。本文的改进算法所发现的规则全部来自于药对中两个药物之间的属性关联，可以发现相同的属性关联，而且排出了药物自身属性间的关联规则的影响，结果的集中度更好，便于发现有价值的内在关联。

在 Apriori 算法的基础上，本文提出了改进的分步式计算方法，非常适用于药对中药物间关联规则的发现，也可用于不同方剂间关联规则的发现。但该方法仍存在着一些不足，结果的集中度仍有待提高，支持度较高的规则数量较多，分析有一定的困难，缺乏概率检验的能力。本实验室正在进行进一步的研究，力争解决以上问题。

表 5 药对中药物间属性关联调整支持率前 10 位的规则

序号	属性 1	属性 2	支持度	支持率（%）	自然概率（%）	调整支持率（%）	Pval	置信度 1（%）	置信度 2（%）
1	苦	肺	304	48.64	23.07	25.57	0	58.80	60.20
2	胃	肺	301	48.16	23.70	24.46	0	62.06	61.93
3	苦	脾	297	47.52	23.81	23.71	0	57.67	57.34
4	肺	心	284	45.44	21.82	23.62	0	58.44	56.24
5	脾	心	286	45.76	22.51	23.25	0	60.21	55.75
6	苦	辛	254	40.64	17.46	23.18	0	52.70	57.86
7	苦	肝	319	51.04	28.32	22.72	0	59.74	63.04
8	肝	肺	369	59.04	36.42	22.62	0	65.54	66.13
9	脾	肺	332	53.12	30.61	22.51	0	65.87	62.88
10	苦	胃	254	40.64	18.43	22.21	0	55.95	58.93

参考文献

1 段金廒，宿树兰，唐于平，等.中药药对配伍组合的现代认识.南京中医药大学学报，2009，25(5)：330~333.

2 唐于平，段金廒，郭盛，等.药对量效关系研究的认识与思考.南京中医药大学学报，2009，25(1)：21~24.

3 Rakesh Agrawal, Tomasz Imielinski, Arun Swami. Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD Conference, 1993.

4 秦伯未.谦斋医学讲稿.上海：上海科学技术出版社，1964：3.

5 吕景山编.施今墨对药临床经验集.太原：山西人民出版社，1982：1.

6 陈维华，徐国龙，张明淮，等.药对论.合肥：安徽科学技术出版社，1984：11.

7 胥庆华.中药药对大全.北京：中国医药出版社，1996.

8 Rakesh Agrawal, Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. Proceedings of the 20th VLDB Conference, 1994.

Discovery of Association Rules between TCM Properties in Crude Drug Pairs by Mining between Datasets and Probability Test

Shang Erxin, Ye Liang, Fan Xinsheng, Tang Yuping, Duan Jinao

PDF 文件使用 "pdfFactory Pro" 试用版本创建 www.fineprint.cn

(Jiangsu Key Laboratory of TCM Formulae Research, Nanjing University of Chinese Medicine, Nanjing 210046, China)

Abstract: The TCM Crude Drug Pair is an important link node between TCM drugs and formulae. In the traditional TCM theory, there are many descriptions in multi-aspects. However, they seem still scattered compared with formulae theories. In this work, the association rule mining method was applied to reveal the associations between the structures of the crude drug pairs. The standard and enhanced Apriori algorithms were applied and run on the crude drug pair database including 625 drug pairs, 347 drugs and 49 properties. The results from the two algorithms were compared and analyzed. The new enhanced Apriori algorithm was found to be more suitable for the association rule discovery between drugs in the drug pair or formulae.

Keywords: Crude Drug Pairs, Integrated Structure, Association Rule, Apriori, Data Mining

(责任编辑:李沙沙,责任译审:张立崴)

PDF 文件使用 "pdfFactory Pro" 试用版本创建 www.fineprint.cn