中医脾胃方配伍规律的数据挖掘试验*

□蒋永光** （成都中医药大学 成都 610075）
李 力 （西南交通大学 成都 610031）
李认书 李慧琴 陈 波 （成都中医药大学 成都 610075）

摘 要：目的：探索数据挖掘技术在方剂配伍规律研究中的方法。方法：1. 从《中医大辞典·方剂分册》中筛选出1355首脾胃方；2. 按照数据挖掘技术中对原始数据的预处理要求，进行方剂数据的规范化、结构化和数字化处理；3. 根据方剂的数据特点，选用聚类分析、对应分析和频繁集方法，进行多角度、多层次和量化的分析和处理，并形成了有关技术规则和处理程序。结果：就脾胃方的核心药物、方剂结构、药对药组和“方药证”的对应关联所进行的数据挖掘，结果基本符合中医脾胃方组方用药的一般规律和特点，并发现了一些值得深入研究的特殊配伍现象和模式。结论：本研究表明，数据挖掘与传统的数据处理方法不同，能以线性和非线性方式进行数据解析，适宜对包含大量模糊和非量化数据的中医方剂配伍规律的研究。但由于数据挖掘对数据质量的要求高，因而数据预处理在方剂数据挖掘中工作量大，技术性强，是实现方剂配伍规律的知识发现的关键所在。

关键词：中医方剂 脾胃方 配伍规律 数据挖掘

一、目 的

数据挖掘（Data Mining）是当今最前沿的信息处理技术，主要用于海量数据的分析和研究。目前，数据挖掘方法已开始应用于生物医学领域。中医方剂配伍研究，因数据的庞大和复杂而非人力所能完成。本研究以脾胃方为样本，按照数据挖掘的程序、要求和基本方法对其进行处理，试图探索数据挖掘在方剂配伍规律研究中的方法意义和应用策略。

二、方 法

1. 方剂的选择

基于原始数据的可靠性、准确性和完整性方面的考虑，本研究以《中医大辞典·方剂分册》[1]为方剂来源。因为，该辞典出版20年来，至今是普遍被采用的方剂学工具书；所收方剂来自于主要医籍，具有一定代表性；方剂数量（包括一些方剂的加减方）在万首左右，较为适中；对方剂原

收稿日期：2002-12-03
★ 国家中医药管理局课题（2000-J-P-54）：中药类方配伍规律的计算机量化分析系统，负责人：蒋永光。
★★ 本文联系人：蒋永光，教授，长期从事中医有效方药的筛选与评价研究，Tel: 028-8779876, E-mail: jiangyg@netease. com。

文进行一定程度的处理，详实可靠而条理清楚，易于操作。

根据以下原则，共选入脾胃方1355首。这些方剂出自汉代到明清的不同临床医家。

（1）方名冠有“脾”、“胃”、“脾胃”、“大肠”者。

（2）主治为“脾”、“胃”、“脾胃”、“大肠”及相关证候者。

（3）主治病证虽不属脾胃，但治法是从脾胃入手者。

（4）某些方剂后的加减方，如方与证齐全，作独立方剂处理。

（5）方与证齐全者。有方药而无主治病证的，不能入选。

2. 数据预处理

预处理就是将原始方剂的语言描述性信息，分解、转化为计算机能够处理的数据单元，使之规范、准确和有序，实现数据的正确表达和合理组织，这是数据挖掘的基本条件。

本研究对脾胃方数据预处理主要包括三个方面：

（1）数据的规范化。

方剂原始数据的不规范，主要表现为两个方面：

概念词的语义模糊。如眩晕，眩指眼花，晕为头晕，通常并称眩晕，指头晕较重，有天旋地转之感者。其词义显然涉及眩、晕、眩晕和头昏这几个概念的定义和区别。

概念表达的不规范。再如眩晕，其表达用词甚多，如晕眩、眼晕、眼昏、眩冒、头晕、目眩、头旋、昏眩等等。这些又反映了对同一概念的规范表达问题。

上两方面的问题，在中医方剂中又以后者更突出。因此，我们参考《中国中医药主题词表》[2]、《中医症状鉴别诊断学》[3]、《中医大辞典》[4]等权威工具，采取以下原则进行处理：

一词多义，使之单义化。如眩晕，即指有天旋地转，不能站立特点的头晕，并将之与单纯的眼花、头昏、晕厥、晕车船等相区别。

多词一义，用一词表达。在具体操作中，不随便进行归并，尤其舌脉证候的表达，尽可能细致，如微热、发热、大热、夜热、日晡发热，均作单一概念处理。

组合概念，拆分进行表达。如涉及“清热”的治法很多，有“清热化痰”、“清热燥湿”、“清热凉血”，为减少数据量，将这些概念拆分为：清热、化痰、燥湿、凉血4个概念表达。

（2）数据的结构化

数据结构化的目的是对方剂原始数据进行符合数据挖掘要求的合理细化和组织，以实现知识重点的有序排列和数据间关联结构的形成。这对揭示方剂配伍与应用的规律是极为重要的。

方剂数据具有多层关联结构，如药与药、药与症、药性与炮制、功效与主治、原方与加减方、加减方与变症等等。其中，“证、药、方”是核心，“药”又是核心中的要素。三者之关系是：针对“证”，选用“药”，配制“方”。而“证”又是由若干证候组成的，“药”包含性味、归经、剂量等内容，“方”则存在复杂的组配关系及加减变化。

方剂数据的结构化最后是通过多层的数据仓库来体现的。我们所构建的脾胃方数据仓库是由若干子库搭建而成，方剂主库、药物库、证候库、治法库等，各子库则根据数据情况向下细化到不能分解的原数据。数据库及各种数据之间可以借助代码或链接，进行线性的或非线性的方式的关联分析与计算。

（3）数据的数字化。

数字被认知的精度和深度远远高于其它的信息载体。因此，我们尽可能采用数字来替代那些负载有某种知识内容的文字或符号。在本研究中，数据的数字化主要出于两种目的：

用数字来实现数据量化。除方剂的剂量统一用以克为单位的数字来表示外，中药性味、毒力等也采用数字来描述。中药的寒热温凉四性实际包含了不同谱级，通过对30名有关专家的问卷调查及统计处理，我们将平性药取值为0，偏性之药依次有其相应取值：

大热←热←温←微温←平→

1.2　1　0.7　0.4　0

凉→微寒→寒→大寒

-0.4　-0.7　-1　-1.2

用数字来表达数据结构。用文字或符号描述的数据很难表示出其结构和相互的关系，数字则易于实现这一点。中药、证候、治法等数据都能够采用数字来表达，这不仅有助于相关数据的深

表 1　症状的编码处理

1 全身症状
11 身寒
111 恶风
……
12 发热
121 身热不扬
122 潮热
1221 骨蒸潮热
1222 午后潮热
124 夜热
125 恶热
……

入分析和运算，也是实现数据规范化和标准化的重要措施。我们对此进行了一次尝试，有关脾胃方主治症的数字化处理结果（部分示例）见表1。试验证明，数据的数字化对数据挖掘的结果产生极好的作用。

3. 数据挖掘方法

（1）聚类分析。

聚类分析（cluster analysis）是研究事物分类的一种统计方法，是直接比较样本中各指标（或样品）之间的性质，将性质相近的归为一类，性质差别较大的归在不同类。聚类过程是数据挖掘过程的第一个阶段。它首先把数据区分于不同的类，以便做进一步的分析。

本研究分别采用系统聚类（hierarchical clustering analysis）和模糊聚类（fuzzyclustering analysis）。对脾胃方所含414种药物，从功效（117种）、归经（12个）、药性（9种）和药味（11种）的角度，进行了分类特征分析。聚类的统计值是各统计指标的距离，没有考虑中药传统的分类规范，由计算机自动完成，基本是客观的。如：根据药物性味的聚类，脾胃方所含中药分为37类，其中部分举例见表2。

(2) 频繁集方法。

数据挖掘要实现的一个重要目标就是寻找关联规则（associationrule），而其第一步则是要找到相应的频繁集（frequent item set）。所谓关联规则，是指同一个事件中出现的不同项的相关性，因而能够反映存在于不同项之间的某些规律和模式。

通过频繁集方法，我们分析了药物与症状、药物与病机、症状与病机等不同项间的相关性，从多角度和多层次来认识脾胃方之“方、药、证”之间的关联，得到了饶有趣味的发现，见表3。

（3）对应分析。

对应分析（correspondence analysis）是一种有效的多元分析方法，能够通过列联表的行列变量间的关系的低维图示，在同一个直角坐标系内同时表达变量与样品两者之间的相互关系。在中医方剂中，“方、药、证”之间存在错综交织的对应关系，故寻找其特殊对应关系对配伍规律的发现和有效方药筛选非常为有用。如：

对脾胃方主治症分析发现，有的症状发生频数极高，于是根据方中药物与症状的对应关系，我们从1355首脾胃方中筛选出症状频数大于30的方剂，然后进一步挖掘这些方剂的药物组成与病机、症状的关联规律，见表4、表5。

三、结　果

1. 核心药物

在“药症对应”分析中发现，有7味药的出现频次远远高于其

表 2　根据性味的药物聚类

类别	类名	代表性药物	药物种数
C1	辛温	大蒜、草果、砂仁，等	46
C2	苦寒	大黄、黄连、黄芩，等	35
C3	辛苦酸温	佛手、香橼	2
C4	甘温	龙眼肉、黄芪、扁豆，等	15
……			

表 3　脾胃方的功效频数(按序排列)

功　效	在脾胃方中的应用频数	占全部方剂的比例(%)
燥湿化痰	1103	81.4
补气	924	68.2
活血	730	53.9
祛痰	690	51
温中理气	668	49.3
……		

表 4 脾胃方药物与主要病机的关联举例

药物	病机																					合计
	大肠实热	大肠痰热	大肠血瘀	肝脾不和	肝气犯胃	肝气郁	脾肾阳虚	脾胃不和	脾胃寒湿	脾气虚	脾胃寒实	脾胃实热	脾胃虚寒	食积	外感风寒	外感暑邪	胃气上逆	胃气滞	胃阴虚	血虚	中焦痰湿	
扁豆	0	0	0	1	0	0	8	4	1	22	0	1	1	0	2	2	0	0	0	8	2	52
白蔻	1	0	0	0	0	2	0	12	1	7	2	0	0	0	0	1	3	1	0	0	2	32
白芍	2	0	2	2	3	3	0	1	0	21	9	16	5	1	1	0	0	5	0	4	2	77
白术	0	0	0	3	1	0	9	19	3	83	16	6	20	5	1	2	1	11	1	3	17	201
……																						
合计	125	22	20	24	46	23	88	295	59	1006	401	370	329	132	48	31	98	203	35	68	283	3706

表 5 脾胃方药物与主要症状的关联举例

药物	症状																					合计
	便秘	肠鸣	大便不利	大便滞	恶心	腹痛	腹胀满	口渴	里急后重	痢疾	面色黄	纳差	呕吐	疲乏	吞酸	胃脘痛	胁胀满	泄泻	心烦	心下痞满	饮食不消	
扁豆	0	0	0	0	12	16	15	3	1	11	2	24	28	5	5	3	0	42	4	3	3	181
白蔻	1	3	2	3	4	10	13	3	0	0	2	18	16	1	7	4	3	9	3	19	11	132
白芍	6	11	17	4	8	82	32	7	30	39	4	46	29	19	4	12	21	43	18	13	10	455
白术	6	18	7	20	28	73	82	16	7	40	10	121	123	37	16	18	16	131	18	55	35	877
……																						
合计	341	280	349	242	667	1916	2006	458	463	1076	269	1949	2245	616	367	579	462	1945	527	906	714	19651

表 6 “药物－证候－病机”对应

药物	主症	基本病机
白术	气短、疲乏、纳差、腹痛、心下痞	脾气虚
橘皮	纳差、腹胀、恶心、呕吐、胸闷	脾气虚、胃气逆
茯苓	泄泻、纳差、脘闷、呕吐	脾气虚、中焦痰湿
半夏	呕吐、心下痞、腹胀、苔白腻	中焦痰湿
干姜	泄泻、腹痛、疲乏、畏寒	脾胃虚寒、脾气虚

表 7 “药物－主症－病机－方剂”对应

药物	主症	基本病机	代表方剂
茯苓 白术	泄泻	脾气虚	茯苓汤、四君子汤
白术 干姜	呕吐	脾胃虚寒	理中丸
半夏 干姜	呕吐	脾胃寒湿	半夏干姜散
白格 橘皮	心下痞满	脾胃不和	异功散
橘皮 半夏 茯苓	呕吐 恶心 胸闷	中焦痰湿	二陈汤

它药物：甘草(1579)、陈皮(1139)白术 (877)、人参 (843)、茯苓(823)、厚朴(776)、木香(768)。不难看出，由这几味药可以构成：四君子汤、异功散，和香砂六君汤的主药。

而从另一角度分析能够得到完全一致的结果：1355首脾胃方中，所用中药共412种，按方剂中药物的使用频率，排前7位亦为：甘草(709)、橘皮(485)、白术(404)、人参(395)、茯苓(362)、厚朴(310)、木香(293)。

2. 方剂结构

各种分析结果证明，变化复杂的脾胃方，其基本结构主要是：

以四君子汤为代表的补气健脾方剂是脾胃方最基本的用方。

补气药+理气药配伍的方剂。如异功散、香砂六君子汤等。

补气药+温里药配伍的方剂。如理中丸、保元汤等。

补气药+理气药+化痰药（或化湿药）配伍的方剂。如六君子汤、参苓白术散等。

3. 药对药组

具有特殊的组合关系的药对和药组，为更加深入的配伍规律研究提供了线索。如：白术与茯苓，人参与生姜，茯苓与木香，陈皮与当归，人参、甘草与陈皮，陈皮、半夏与茯苓，等等。

4. “方药证”对应关联

对方剂、药物和证候进行多角度和层次的对应分析，则能发现更多的规律性的现象。见表6，表7。

四、结　论

1. 脾胃方的数据挖掘结果基本符合中医脾胃方组方用药的一般规律和特点，并发现了一些值得深入研究的配伍现象和模式。

2. 数据挖掘在技术线路上与传统数据处理方法不同，能以线性和非线性方式解析数据，尤善处理模糊的和非量化的数据，确实非常适宜于方剂数据分析。

3. 数据挖掘对数据质量的要求很高，因而数据预处理在方剂数据挖掘中工作量大，技术要求高，这成为方剂数据挖掘和知识发现的关键所在。

4. 数据挖掘是包含人工智能、模式识别、模糊数学、数据库、数理统计等多种方法的技术集成。本研究采用方法和样本量尚少，故应加大研究范围和深度。

参考文献

1 中医研究院等. 中医大辞典·方剂分册. 北京:人民卫生出版社，1983.

2 吴兰成. 中国中医药学主题词表. 北京:中医古籍出版社，1996.

3 赵金铎. 中医症状鉴别诊断学. 北京:人民卫生出版社，1984.

4 李经纬等. 中医大辞典. 北京:人民卫生出版社，1995.

（责任编辑：张志华　柳　莎）

应该给“SARS”准确通俗的译名

2002年第4季度以来，一种传染性和致命性很强的烈性传染病逐渐在一些国家和地区流行起来，尤其对中国的经济发展、社会进步造成了较大的影响。最近被称为突如其来的“非典”在媒体出现的频率越来越高，已是家喻户晓。

随着病原体分离鉴定取得进展，国际科技界公认此病是由冠状病毒的一个或多个变种引起的，另外一个缩写“SARS”在媒体上的出现频率逐渐上升，为了让大众了解“非典”与“SARS”的区别，及时与国际接轨，略陈己见于下。

在WHO确认此病病原体之前，“非典”的叫法无可厚非，语音朗朗上口，而且全都明白“非典”是非典型肺炎的简称，在分离鉴定病原体的过程中，又曾有报告称发现引起非典型肺炎的病原体一衣原体。但在WHO确认此病病原体之后，再称“非典”，似欠妥当。在中文语义中，非典型总是次、轻于典型，在一段时间内未引起民众甚至医护人员的警惕，不能说与此术语欠妥无关。更主要的是“非典”已经与目前这种烈性传染病没有任何联系，也体现不了这种疾病的任何属性，纯粹是一个符号、代号，无法成为一个学术名称而永远被人使用。目前新闻媒体所SARS翻译成“非典”或将二者划等号，甚至可能出现在抗击“非典”的纪念物上，都可能把这种误解永久化。

“SARS”是严重急性呼吸系综合症（Severe Acute Respiratory Syndrome）的英文缩写，比较准确地概括和体现了此病的特点，世界通用，应该向中国的大众介绍和宣传，而介绍和宣传的第一项工作就是要给“SARS”一个中文术语。作者建议“飒什疫”，“飒”是SA的谐音，如风之飒然而至，寓意突如其来，又急又重；“什”谐S之音，有探索的意思，加“疫”提示传染性、致命性。国际学术界已有SARS病毒的提法，即把冠状病毒的一个或多个变种直接与SARS挂钩，相应地可称为“飒什病毒”，但不可称为“非典”病毒。以上建议是抛砖引玉，希望医学名词术语方面的专家和机构，及早确定一个恰当的名称。

（胡世林 中国中医研究院中药研究所）

TCM and to improve its diagnostic value in modern clinical medicine.

This article is one of the key-note reports made at the first acadmic salon on traditional Chinese medicine and materia medica held by the journal World Science and Technology in 2000.That conference put its theme on "Scientific characteristics, Modernization and Digitalization of TCM Theories".

Key Words: tongue diagnosis, digital tongue image, tongue image analysis system, image analysis

Experiment on Data Mining in Compatibility Law of Spleen-stomach Prescriptions in TCM

Jiang Yongguang and Li Renshu
(Chengdu University of Traditional Chinese Medicine, Chengdu 610075)
Li Li (Southwest Jiaotong University, Chengdu)
Li Huiqin and Chen Bo
(Chengdu University of Traditional Chinese Medicine, Chengdu 610075)

Exploring the methods of dada-mining technology in the study on the law of the compatibility of TCM prescriptions. Method: 1.Screening out 1355 prescriptions for the treatment of Spleen-stomach diseases from the section of Prescriptions, Dictionary of Traditional Chinese Medicine; 2. standardizing, structuralizing and digitalizing the data of prescriptions according to the requirements of pre-processing raw data by data-mining technology; 3.according to the Characteristics of the prescription data, adopting the cluster analysis, the correspondence analysis and the method of frequent set to process and analyze these data from different angles, at different levels and by the way of quantization, and forming technical rules and processing procedures involved. Result: The result of data-mining in main drugs, prescription structure, pairing drugs and corresponding relations among formula, drugs and syndromes basically accords with the general rules and characteristics of the compatibility of spleen-stomach prescriptions in TCM, and some unusual phenomena and modes of drug compatibility, which are worth being studied deeply, are found. Conclusion: Data-mining is different from traditional data processing, which is able to analyze data by ways of linearity and non-linearity and suitable for the study on the compatibility laws of TCM prescriptions with large puzzling and non-quantization data. Since data-mining requires high-quality data, the pre-processing of data means a lot of work and high technology in the mining of prescription data, but the key to the discovery of knowledge in the compatibility laws of TCM prescriptions as well.

Key Words: TCM prescription, spleen-stomach Prescription, compatibility law, data-mining

Importance in Identification of A List of Diseases for Which Traditional Chinese Medicine Has Superior Effectiveness

Shi Shenghong and Zheng Xiaoyan
(Chengdu Housheng medicine R & D Co.,Ltd, Chengdu 610072)
Wang Zongqin
(Chengdu University of Traditional Chinese Medicine, Chengdu 610075)

This article puts forward the viewpoint that a list of diseases for which traditional Chinese medicine (TCM) has superior effectiveness should be identified by the way of contrast and analogy and with the aid of existing investigation materials, explores the thought of how to identify such a list and emphasizes the importance in the identification of the list. It holds: (1) The extensive effectiveness of given Chinese medicines according to the description of TCM literatures forms contrast with their practical limitation. Therefore, traditional theories of TCM should be re-evaluated and revised by the knowledge acquired in consistent practice and from the height of the development of various disciplines; (2) Traditional Chinese medicine and Western medicine belong to their own different systems of medical science and thus it is natural that they should present their features of the medical science they belong to respectively; (3) Facing the modern medical science that is changing with each passing day, the identification of a list of diseases with superiority and features of TCM in their treatment should be explored; (4) The study on the identification of such a list will contribute to probe into the theme "What problems can TCM actually deal with?" and also an important measure for the guidance of correct use of Chinese medicines in clinic treatment so as to improve the curative effectiveness and the enhancement of the reputation of TCM.

Key Words: TCM, list of diseases with superior effectiveness of TCM in their treatment, modernization of TCM