基于复杂系统熵聚类方法的中药新药处方发现研究思路*

□唐仕欢 （中国中医科学院中药研究所 北京 100700）
陈建新 （北京中医药大学 北京 100029）
（中国科学院自动化研究所 北京 100190）
杨洪军** （中国中医科学院中药研究所 北京 100700）
王 波** （中国人民解放军总医院第二附属医院 北京 100091）

摘 要：本文对中药新药处方发现模式及其问题进行了分析，在此基础上，提出中药新药处方发现的新工具-复杂系统熵聚类方法，并详细阐述应用此方法进行中药新药处方发现的研究思路。主要包括：建立病证方剂数据库；应用复杂系统熵聚类方法快速筛选出中药核心组合；对筛选出的中药核心组合进行专家判断，从而确立用于中药新药研发的候选处方。此方法的应用将为中药新药处方快速发现提供新的思路。

关键词：中药新药 处方发现 熵 复杂系统熵聚类方法

中药复方是中医临床用药的主要形式，也是中药新药研发的主要来源。新药研发的过程中，候选处方的确立是中药新药创制的首要步骤，因此，处方发现是中药新药开发的源头。古今临床医家记载了数以万计的疗效确切的方剂，如何从中确定研发对象，成为中药新药研究的热点和难点。为此，引进新技术和新方法，从浩如烟海的数据中快速挖掘出针对病证具有疗效确切的中药核心组合是解决这一问题的关键。

一、中药新药处方发现模式及其问题分析

对于中药新药创制而言，中药新药必须具备自身的特点，不同于天然药物寻找单一有效化合物为目标的创制模式。从中药新药候选处方的现状来看，主要来源于经典方剂、名医名家的经验方剂、确有成效的中成药的二次开发和根据实验室现代药理研究结果组成的方剂。随着社会需求的迅速发展，从古今医家文献记载的中药复方中筛选新药研发对象成为中药新药研究的热点。几千年来，中医药领域的无数临床实践与理论研究积累了海量的古籍文献以及当前的研究文献，面对如此浩大的文献数据，盲目从中筛选中药新药研发对象，不仅命中率大大降低，而且浪费大量的人力、物力和财力，延长中药新药研发的周期。面对处方筛选存在的问题，有学者提出利用经典古方、名医名方作为中药

收稿日期：2008-12-28
修回日期：2009-02-09
* 国家自然科学基金项目（30873465）：抗痨中药筛选新模式的研究，负责人：王波。
** 联系人：杨洪军，研究员，硕士生导师，主要研究方向：中药新药处方发现和组效关系研究，E-mail：hongjun0420@vip.sina.com；王波，主治医师，博士，主要研究方向：中药新药的临床和实验研究，E-mail：waterflowergxh@sohu.com。

新药研发的主要来源[1]。不可否认,经方、名方是中医药学家临床实践的经验总结和智慧结晶,是中医药学的重要宝贵资料。然而,医家名方仍然出于有限的个人经验,加上中医各家学派精彩纷呈,临床处方各具风格,方剂疗效没有确切的统计数据,存在一定的局限性。

因此,能否有效利用古今文献资料,既可继承名医、名家、名方的精髓,又可确保临床疗效的准确性,集中医名家之大成,融各家精华于一方,是中药新药研发普遍关注、期待解决的问题。随着"非相关文献知识发现方法"、"熵方法"等人工智能数据挖掘技术的迅速发展,为从文献中发现组方规律,进行处方发现提供了有力工具。为此,以数据挖掘和分析计算所确定的组方为基础,再进行有针对性的处方筛选将大大提高命中率,而且能避免个人经验的局限,突破了以往完全依赖实验评价的模式。这种从建立病证方剂数据库的角度创立的中药新药处方发现的新模式,将更快速、更准确、更为可行地进行中药核心组合的筛选,以缩短新药研究的周期。

二、中药新药处方发现的新工具——复杂系统熵聚类方法[2]

数据挖掘是从大量的数据中,抽取出潜在的、有价值的知识的过程,也称为数据库中的知识发现,融合了数据库、人工智能、机器学习、统计学、知识工程、面向对象方法、信息检索、高性能计算以及数据可视化等最新技术的研究成果,是一个多学科交叉研究领域。相关技术方法包括:遗传算法、粗集方法、决策树、人工神经网络、模糊逻辑、规则归纳、聚类分析、模式识别、频繁集方法、最近邻技术、可视化技术等。

中医药数据挖掘的对象是中医药领域中积累的海量数据,这些数据的属性既有离散型的,又有连续型和混合型的,挖掘过程需要人机交互、多次反复,在中医药专业背景知识引导下,针对具体问题,选择合适的数据挖掘方法。复杂系统熵聚类方法是一种非监督的模式发现算法,它能自组织地从海量的数据中提取出信息量最大的组合,同时,此方法特别适用于高度离散性类型的数据。相比于经典的统计方法,它有以下几个优点。

1. 不需要数据的一致性,对于各类数据都适合。特别针对具有随机性,模糊性,非平衡性,非遍历性,多维性特点的中医药数据。

2. 它客观地反映数据的情况,聚出来的组合内元素的相关都特别大,是最优的组合,这些组合为新药发现中候选处方的筛选奠定了基础。

3. 相关是不对称的,为定义贡献度奠定了基础。

4. 算法收敛速度快,对于处理大量的数据有优势。

中医药文献数据中的方剂数据也具有离散型的,又有连续型和混合型的特点,因此,复杂系统熵聚类方法非常适合作为中药新药的处方发现,此方法具有两方面的显著优势:一方面,不仅可以定性、还可以定量挖掘出药物之间、病-证-症-药之间的相关性;另一方面,不仅可以挖掘出名医名家经验的核心组合,还可以挖掘出隐藏于方剂配伍之中的而没有被临床医家所重视的核心组合。

三、研究思路

历代中医经过长期临床实践,积累了丰富的防治经验和大量有效方剂,以图书文献和在世的名医为载体得以保存。以复杂系统熵聚类方法为数据挖掘工具,对名老中医医案、验方、古代方剂、民族药等数据进行挖掘,从数据中寻找组方规律,确定组方,在此基础上,结合专家经验判断,再进行有针对性的处方筛选,将大大提高命中率。这种新药处方发现方法,突破了既往完全依赖具于实验评价和个人经验的局限,创新了现代中药处方发现方法。应用复杂系统熵聚类方法进行中药新药处方发现研究,主要分为以下几个步骤:(1)病证方剂数据库的设计和建立;(2)复杂系统熵聚类方法的数学建模和数据筛选;(3)专家判断和候选处方的确立。由于在中药新药处方发现研究中,必须依托一个具体的疾病进行,为此,基于我们的工作基础,以下以感冒方剂为例加以说明。

1. 数据库设计和建立

对历代医家的宝贵经验处方进行数据挖掘,首先必须建立适合应用复杂系统熵聚类方法进行挖掘的相应病证方剂数据库。在建立数据库的过程中,数据来源的准确性对后期处方发现的结果至关重要,是确保中药新药处方筛选的基础,因此,精选名医、名家、名著的记载处方为首选。数据库中的信息应包括病名、证型、方剂药物组成、药物剂量、方剂来源等,并以数据格式保存。例如,应用复杂系统熵聚类

方法对感冒方剂进行新药处方发现研究，建立感冒方剂数据库，应收集、整理古今医家治疗感冒的相关方剂，主要有：古代记载治疗感冒（包括外感风寒、风热、暑湿等表证）的相关方剂，包括《肘后备急方》、《太平惠民和剂局方》、《普济方》等；当代中医名家治疗感冒的医案、医话中的方剂，包括《施今墨临床经验集》、《孔伯华医集》、《中国百年百名中医临床家·朱良春》、《中国百年百名中医临床家·邓铁涛》、《中国百年百名中医临床家·任继学》等；医学核心期刊发表治疗感冒疗效确切的相关方剂。将总结、整理的方剂以 Visual FoxPro 为开发平台，编制中文操作界面，建立感冒方剂数据库，使得所有信息中涉及的病名、证型、症状及用药等都成为取值为 0 或 1 的二值变量（dichotomous variable），从而为数学建模和数据筛选奠定基础。

2. 数学建模与数据筛选

复杂系统熵聚类方法是一种非监督的学习算法，能快速地从数据库中提取出药物的核心组合及病-证-症-药最相关的组合。采用复杂系统熵聚类方法作为新药处方发现的数学方法，对数据库中数据进行建模和数据筛选。其实现如下：

为了叙述，下面先介绍几个定义：

对于一个复杂系统，可以表示为矢量

$$s=(X_1,X_2,\Lambda,X_i,\Lambda,X_p)^T \qquad (1)$$

其中，$X_i=(X_{ia})(i=1,2,\Lambda,p,a=1,2,\Lambda,q)$ 是描述系统特征的变量。令 $C_i(i=1,2,\Lambda,p)$ 为 X_i 分类的集合，C_i 的第 a 个元素 $C_{ia}=a$，则有 $C_i=\{1,2,\Lambda,a,\Lambda,k\}$，$k\leqslant q$，并令 $n_a\left(\sum_{a=1}^{k}n_a=q\right)$，为事件 X_i 属于 C_i 第 a 类的数量，则变量 X_i 的熵定义为

$$H(X_i)=-\sum_{a=1}^{k}n_a/q\log n_a/q \qquad (2)$$

X_i 和 X_j 的联合熵定义为

$$H(X_i,X_j)=-\sum_{a}\sum_{b}n_{ab}/q\log n_{ab}/q \qquad (3)$$

其中 n_{ab} 表示事件 X_i 属于 C_i 的第 a 类同时 X_j 属于 C_j 的第 b 类的数量。

式（2）、（3）可分别表示成

$$H(X_i)=\log q-\frac{1}{q}\sum_{a=1}^{k}n_a\log n_a \qquad (2)'$$

$$H(X_i,X_j)=\log q-\frac{1}{q}\sum_{a}\sum_{b}n_{ab}\log n_{ab} \qquad (3)'$$

有了上述熵的定义，下面给出基于互信息的关联度的定义。

定义 1. 假设 $X_i\cap X_j=\phi$，则称熵

$$\mu(X_i,X_j)=H(X_i)+H(X_j)-H(X_i,X_j) \qquad (4)$$

为 X_i 和 X_j 之间的关联度。

定义 2. 假设 $X_i\cap X_j=\phi$，则称

$$\mu(X_i,X_j)=\frac{H(X_i)+H(X_j)-H(X_i,X_j)}{H(X_i)} \qquad (5)$$

为 X_i 和 X_j 之间的关联度系数。

通过计算每两个药对之间关联度系数，作为文献中两两药物之间的关联性度量。但药物的相关有两种情况：同时出现和不出现（正相关）；不能同时出现（负相关）。由于关联度系数只能是非负值，无法区分变量间正相关与负相关的差别，所以必须改进关联度系数以使得正相关和负相关能够分开。

从基于 Shannon 熵的关联度 $\mu(X_i,X_j)=H(X_i)+H(X_j)-H(X_i,X_j)$ 的定义可以看出，两负相关的症状，同时出现的概率为 0，重新定义改进的关联度系数为，如果两个信息同时为阳性的频率，记作 $P_0(i,j)$，大于某个数值 δ，那么就保持原来的定义式不变，如果小于 δ，则分子增加惩罚项 $H(X_i,X_j)$，将关联度系数定义式改为：

$$\mu'(X_i,X_j)=\frac{H(X_i)+H(X_j)-2H(X_i,X_j)}{H(X_j)}。$$

于是，改进的动态关联度系数就可以写成：

$$\Delta\mu'(X_i,X_j)=\begin{cases}\dfrac{H(X_i)+H(X_j)-H(X_i,X_j)}{H(X_j)} & P_0(i,j)\geqslant\delta\\ \dfrac{H(X_i)+H(X_j)-2H(X_i,X_j)}{H(X_j)} & P_0(i,j)<\delta\end{cases}$$

这样，两个负相关的药物之间的关联度系数就变小了，甚至可能变负。

通过上述的建模过程和数学运算，计算每一个药对变量与其它变量之间的关联度系数，对于每一个变量，根据与其它变量关联度系数的大小关系，将系数最大的前个变量称为该变量的“亲密变量”。记为全部的变量数。如果两个变量互为“亲密变量”，那么这两个变量是正相关的。如果三个变量之间任意两个变量都是正相关的，那么这三个变量就聚成一堆，依次类推，直到算法收敛，即不能再往堆里加任何元素。据此，在 Matlab7.0 平台上编写程序，进行运算处理，得出相关较大的一系列药物核心组合，作为专家

判断,确立候选处方的依据。

3. 专家判断和候选处方确立

根据复杂系统熵聚类方法对数据库中的数据进行运算和筛选的结果,聘请相应病证的有关专家,结合他们的临床经验,对数学计算出的一系列药物的核心组合进行综合判断,即数学计算与专家判断相结合,确立用于中药新药研发的候选处方。例如,对于感冒方剂而言,可以选取筛选结果中治疗感冒中药相关性最大的核心组合,包括3味药物、4味药物、5味药物的不同组合,聘请有经验的临床专家,对上述数学计算出的核心组合进行判断和评价,最终确立用于抗感冒新药研发的候选处方。应用复杂系统熵聚类方法对感冒方剂进行处方发现研究,其初步的研究结果见表1。

总之,以古今专病方剂文献数据为基础,建立相应病证方剂数据库,通过复杂系统熵聚类方法,筛选出治疗该病的中药核心组合,采用数学算法与专家经验判断相结合的方法,确立中药新药研发的候选处方,将为开发新药奠定基础,同时,对建立"疗效确切、方法便捷"的中药新药创制新模式和新方法具有重要意义。

表1 基于复杂系统熵聚类方法的抗感冒新药处方发现研究结果

编号	候选处方				
1	板蓝根	薄荷	金银花	连翘	石膏
2	薄荷	金银花	荆芥	连翘	牛蒡子
3	薄荷	淡竹叶	金银花	连翘	
4	陈皮	桔梗	前胡	紫苏叶	
5	荆芥	桔梗	连翘	牛蒡子	
6	白术	防风	黄芪		
7	板蓝根	黄芩	石膏		
8	半夏	陈皮	茯苓		
9	薄荷	菊花	连翘		
10	柴胡	黄芩	石膏		
11	陈皮	茯苓	桔梗		
12	淡竹叶	桔梗	连翘		
13	防风	黄芩	荆芥		

参考文献

1 张铁军. 中药新药研究的思路方法和实践. 中草药,2007,28(1):1~6.

2 陈建新. 中医证候的复杂系统建模及其与疾病的相关性研究. 中国科学院研究生院,博士论文, 2008.

Designing New TCM Prescriptions Based on Complex System Entropy Cluster

Tang Shihuan

(Institute of Chinese Materia Medica, China Academy of TCM, Beijing, 100700, China)

Chen Jianxin

(Beijing University of Traditional Chinese Medicine, Beijing, 100029, China;Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China)

Yang Hongjun

(Institute of Chinese Materia Medica, China Academy of TCM, Beijing, 100700, China)

Wang Bo

(PLA General Hospital, Beijing, 100091, China)

This paper analyzes the New TCM Prescription Design Model (NTCMPDM) and the associated problems. Based on the analysis, we present a new technique, or complex system entropy cluster, for designing new TCM prescriptions, along with some research strategies for the application. The strategies are mainly made in three parts. The first part is to build a database for the formula and syndromes in the context of a disease. The next is to screen out the core herbal combinations though complex system entropy cluster. Finally, the qualified herbal combinations are evaluated and screened by TCM experts to be a candidate for new drugs. The strategies presented here constitute a basis for quick screening the desired candidates for new TCM drugs.

Keywords: new drug of Chinese medicine; prescription discovery; entropy; complex system entropy cluster

(责任编辑:王 瑀,责任译审:邹春申)