doi:10.3772/j.issn.1002-0470.2014.01.015

基于多参数距离融合聚类原理建立中药标准指纹图谱的研究①

崔建新②* ** 崔建凤*** 洪文学* 高海波③* ****

(*燕山大学电气工程学院 秦皇岛 066004)
(**河北省测试计量技术及仪器重点实验室 秦皇岛 066004)
(***秦皇岛职业技术学院旅游系 秦皇岛 066004)
(****中国中医科学院针灸研究所 北京 100700)

摘 要 分析了中药标准指纹图谱的传统构建方法,针对传统方法建立的中药标准指纹图谱在中药质量鉴别及评价上的局限性,进行了标准指纹图谱建立方法研究,提出了基于多参数距离融合聚类原理建立中药标准指纹图谱的方法。该方法通过计算中药多维多息图谱数据的各维特征聚类规则并融合在一起获得中药标准指纹图谱,实现了多种图谱信息的融合。采用黄芩数据进行了实验,实验结果显示依据聚类原理构建的中药标准指纹图谱更利于中药的分类。该方法是一种可行的中药标准指纹图谱构建方法。

关键词 中药指纹图谱,聚类分析,信息融合,可视化

0 引 言

中药现代化是目前国内外医药界研究的一大热点[1]。中药指纹图谱技术是实现中药现代化的有力工具[2]。任何药物,只有达到一定质量标准才能产生一定疗效。药物质量的优劣直接关系到疾病的治疗及患者的健康和生命安全[3-5]。而中药由于其化学成分的多样性、复杂性以及有效成分的不均衡性,使得中药分类和质量研究的任务极其艰巨。中药指纹图谱可以用来鉴别中药的真伪、控制中药质量及评价其安全性和有效性。中药指纹图谱的解析与处理通常是指,借助于计算机,分析其化学成分测定值及相关的药理作用,快速、准确地寻找出内在规律,作为中药材质量控制的指标[6,7]。

中药指纹图谱的标准图谱,又称为共有模式,是指能够表征某种中药产品化学组成特征的一个图谱。在中药的鉴别与评价以及中药指纹图谱数据处理中,标准指纹图谱可以作为参照图谱。将待测试的指纹图谱与标准指纹图谱进行对比匹配,可以得到测试样本的质量评价结果。对于中药材,可以把经过鉴定的药材样本作为对照品,构建该品种药材的标准指纹图谱。也可以将道地产区的样品或严格按 GAP(中药材生产质量管理规范认证管理办法)要求生产的药材样品作为标准品进行对照。中药指纹图谱标准图谱的建立,对于中药指纹图谱数据处理准确度的提高具有重要的指导作用,有利于形成对中药材或中成药的客观准确的鉴定或评价。而中药的鉴别与评价最终都归结为分类问题,所以本文基于分类目的,提出以多参数距离融合聚类规则作为评价与鉴别标准,也就是标准指纹图谱不再是传统意义上的有形的特征指纹图谱或有实际意义的特征库,而是一种以分类为目的的聚类规则。由于以分类为目的,以分类结果准确率高的聚类规则来定义标准指纹图谱,所以用这种方法构建的标准指纹图谱更利于指纹图谱的分类,同时在基因研究方面更有广泛应用[8-10]。

1 中药标准指纹图谱传统构建方法

常用标准指纹图谱构建方法主要有三种:

(1)典型指纹图谱选择法,即通过对一组或一

① 中国博士后科学基金(2012M510722),中国中医科学研究院项目(ZZ12001),燕山大学博士基金(B692)和秦皇岛市科学技术研究与发展计划(201001A119)资助项目。
② 女,1977 年生,博士;研究方向:模式识别,中药指纹图谱;E-mail:ydcuijianxin@yahoo.com.cn
③ 通讯作者,E-mail:hhghb@ysu.edu.cn
(收稿日期:2013-05-27)

系列样品的指纹图谱进行研究，并结合待测样品的性状以及其它理化鉴别方法，从样品中选择一个具有代表性或者是有典型意义的指纹图谱来作为对照指纹图谱，构建出该品种的标准指纹图谱。当各个样品的指纹图谱特征相近时，这种方法值得推荐。但是由于选择的典型指纹图谱只包含单个样品的特征，所以选择过程难以避免随意性。而且，在各个样品的指纹特征差异比较大的情况下，典型指纹图谱的选择比较困难。

（2）共有模式生成法。共有模式指的是每个指纹图谱都包含的广义谱峰，它可能是一个真正的谱峰，也可能是一段区域，即包含有几个谱峰。通过对一批指纹图谱进行研究，模拟出用于对照的指纹图谱即标准指纹图谱或生成对照指纹图谱数据。这种方法的优点是综合了一批样品的指纹图谱信息，虽然信息较丰富，但仍然会受样品批次的影响，很难保证结果的准确性。

（3）特征指纹图谱库汇集法。这种方法采用数字化指纹图谱的原理和方法，对 10 个批次以上的样品进行指纹图谱处理，首先生成特征峰的指纹图谱集，然后再根据基准样品、共有峰、n 强峰的数目以及 n 强峰出现频次要求和共有率要求等的设定，来对数据进行处理，这样就得到了 10 个批次以上样品的指纹图谱库，也就是标准指纹图谱。这种方法的优点是综合了多个批次样品的指纹图谱信息，信息丰富。但是，由于该方法是对图谱表层信息的综合，并没有深入挖掘利于鉴别和评价的信息，另外，各种要求的设定也没有一个标准，仍然由人的主观因素决定，所以该方法最终同样保证不了相似度计算的准确性。

2 基于多参数距离融合聚类原理的中药标准指纹图谱的构建

2.1 标准指纹图谱的建立原理

标准指纹图谱的建立以聚类分析为基础，为了得到最优解，在聚类分析之前，首先考察各个样本的多维信息数据之间的相关性。一般情况下，如果两个样本的数据相关性比较弱，则说明这两维数据比单纯任一维数据所包含的信息会更加丰富，因而更能比较全面地表征样本的特异性。如果相关性较强，则可根据实际情况进行数据的约简即降维处理。然后对样本的各个特征分别采用不同的相似度方法计算样本的相似度，如各种距离及相关系数计算方法，并依据相似度对样本的多维特征进行聚类。最后，训练分类器，并选择鉴别效果较好的距离计算方法，得到多个距离分类规则，将多维距离分类规则融合在一起就组成了标准指纹图谱。

图 1 为标准指纹图谱生成示意图。$A,B,\cdots$为多维多息图谱数据表征的样本。$A_{ti},A_{si};B_{ti},B_{si};\cdots$分别表示样本的保留时间、相对保留时间、峰面积和归一化峰面积，m 表示样本的变量个数。t,S 分别表示保留时间和峰面积的聚类规则，它们是各个矩阵的相应行聚类的规则，如 A_{ti} 与 B_{ti} 的聚类，然后得到矩阵 $\boldsymbol{H}$ 为总的聚类规则，即为我们的标准指纹图谱。

$$\begin{bmatrix} A_{t1} & A_{t2} & \cdots & A_{tm} \\ A_{s1} & A_{s2} & \cdots & A_{sm} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

$$\begin{bmatrix} B_{t1} & B_{t2} & \cdots & B_{tm} \\ B_{s1} & B_{s2} & \cdots & B_{sm} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

$$\vdots$$

$$\xrightarrow{\text{聚类}} \begin{bmatrix} t \\ s \\ \vdots \end{bmatrix} = \boldsymbol{H}$$

图 1 标准指纹图谱生成示意图

2.2 标准指纹图谱的建立步骤

中药标准指纹图谱建立的基本步骤如下：

（1）特征选择。首先对指纹图谱数据进行特征选择，由于选择的特征将作为聚类判别的依据，所以这些特征应该最大限度地区分不同的类别样本，同时又能够容易确定相同的类别样本。这与特征空间中样本特征的显著性程度是有很大关系的。

（2）测试指标的确定。由于聚类过程所依据的相似度的判别是需要有度量的，所以应确定相应的度量标准，即选取合适的距离或相关系数的范围，这是一个必要的步骤。

（3）准则函数的确定。准则函数或目标函数的选择将会对聚类结果有很大的影响，不同的数据集可能蕴涵不同的类别形式，相应地就会有不同的准则函数。

（4）聚类算法的确定。基于“距离”的聚类算法有很多种，由于不同数据集有不同的特征形式，且同一数据集的不同特征也可能存在很大差别，所以应针对不同数据集甚至是数据集中的不同特征确定有效的算法。

（5）结果的验证。验证聚类结果的正确性以及聚类方法的有效性。这是非常关键的一步。

3 实验结果及分析

3.1 实验数据

本实验采用文献[11]中的表3-6和表3-9所示的黄芩数据,包括不同产地的不同种类黄芩样品23个。黄芩是一种常用的中药材,可清热燥湿,泻火清毒。但除了中国药典中记载的正品黄芩(Scutellaria Baicalensis Georgl)外,市场上有其它的伪品黄芩也在不同程度地使用,包括甘肃黄芩、滇黄芩、粘毛黄芩及丽江黄芩。甘肃黄芩、滇黄芩、粘毛黄芩及丽江黄芩与正品黄芩既有相似又有不同,实验的目的是鉴别黄芩的质量及区分黄芩的种类。本实验数据由气相色谱(GC)、薄层色谱(TLC)、纸色谱(PC)、紫外光谱(UV)4项技术联用共得到8个特征值。

3.2 实验过程

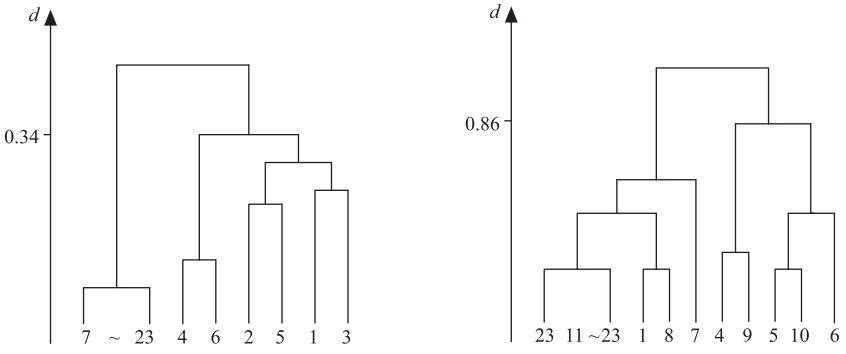
本实验采用层次聚类法,其基本原理是开始每个样本自成一类,然后依次将最相似的两类合并,并计算新类与其他类之间的距离或其它相似性测度,直到所有的样本都归为一类,这一过程可以用谱系聚类图来描述。该方法的聚类原则是由样本间的距离以及类间距离的定义决定的,所以不同的距离定义方法将产生不同的层次聚类计算方法。

本文中类与类之间的距离采用重心法进行计算。距离采用欧式距离,计算每个特征的各个样本间距离值,可以得到23×23的距离矩阵,8个特征得到8个距离矩阵。由于矩阵是对称矩阵,所以只需写出下三角矩阵。由于篇幅关系,这里就不给出。每个特征随机选取矩阵表中的20个样本进行训练,并且保证在训练样本中包含所有黄芩种类,计算对已知类别的各个样本定义边界距离值,对于8个特征值可得到8个距离范围即聚类规则,将其融合在一起即最终的标准指纹图谱。

3.3 实验结果及分析讨论

由于特征7和特征8的距离矩阵数据值比较特殊,首先进行聚类分析,且无须绘出聚类图。对于特征值7取阈值$d_7=0.41$,可分为两大类$G_{71}\{4\sim10\}$和$G_{72}=\{1\sim3,11\sim23\}$。因此,特征值7可将正品黄芩与除甘肃外的其它黄芩区别开来。对于特征值8取阈值$d_8=1$,可分为两大类$G_{81}=\{4,5,9,10\}$和$G_{82}=\{1\sim3,6\sim8,11\sim23\}$。特征值8显然对黄芩类别的区分能力不如特征值7,且比较混乱。但说明了甘肃黄芩和粘毛黄芩与正品黄芩的相似性。

直观上,特征4和特征5在距离矩阵上有一定的相似性。所以,接下来分析特征4和特征5。首先,绘出聚类图如图2所示。对于特征4取阈值$d_4=0.34$,可分为$G_{41}=\{7\sim23\}$、$G_{42}=\{4,6\}$和$G_{43}=\{1\sim3,5\}$。可见,特征值4可将甘肃黄芩和滇黄芩与正品黄芩区分开来。所以,采用特征值7和4共同作用可鉴别甘肃黄芩、滇黄芩与正品黄芩。同时说明了粘毛黄芩和丽江黄芩与正品黄芩的相似性。



(a)特征4聚类图　　(b)特征5聚类图

图2　特征4和特征5聚类图

对于特征5取阈值$d_5=0.86$,可分为$G_{51}=\{1\sim3,7,8,11\sim23\}$和$G_{52}=\{4\sim6,9,10\}$。这时,$G_{52}=\{4\sim6,9,10\}$中为滇黄芩和丽江黄芩,从而利用特征7、4和5可鉴别丽江黄芩。同时,说明了甘肃黄芩、粘毛黄芩与正品黄芩的相似性,以及滇黄芩与丽江黄芩的相似性。

继续分析特征2和特征3。首先绘出特征2聚类图如图3所示。对于特征2,取阈值$d_2=0.45$时,可分为$G_{21}=\{1\sim8\}$、$G_{22}=\{11\sim15,18,19,23\}$、$G_{23}=\{10,16,17,20\sim22\}$和$G_{24}=\{9\}$。所以,特征2与特征7、4可用于鉴别粘毛黄芩。同时也说明了丽江黄芩与正品黄芩的相似性。

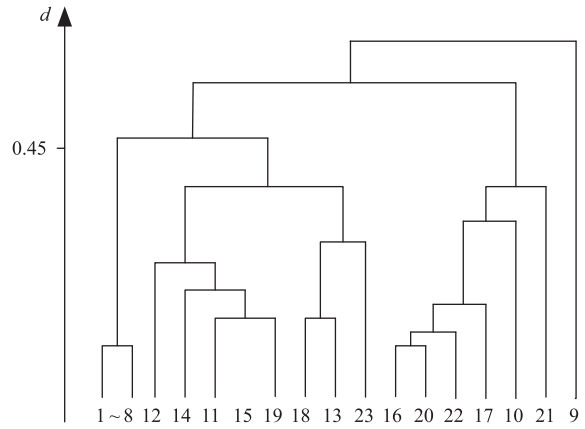


图 3 特征 2 聚类图

绘出特征 3 聚类图如图 4 所示。当特征 3 取阈值 $d_3 = 1$, 可分为 $G_{31} = \{4 \sim 6, ,9 \sim 13\}$ 和 $G_{32} = \{1 \sim 3,7,8,14 \sim 23\}$。所以,特征 3 、7 可鉴别粘毛黄芩,特征 3 、2 可鉴别丽江黄芩,同时也说明了甘肃黄芩、粘毛黄芩与正品黄芩的相似性。

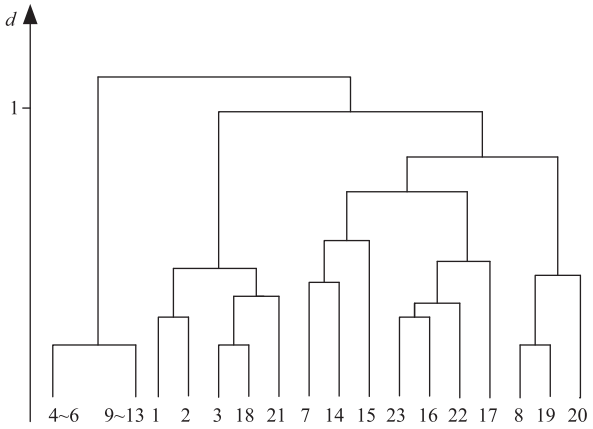


图 4 特征 3 聚类图

接着分析特征 1 和特征 6,首先绘出特征值 1 聚类图如图 5 所示。对于特征值 1 取阈值 $d_1 = 0.31$, 可分为 $G_{11} = \{6,11,18,19,23\}$、$G_{12} = \{1,2,5,9,10,12,13,20 \sim 22\}$、$G_{13} = \{4\}$ 和 $G_{14} = \{3,7,8,14 \sim 17\}$。虽然总体类别比较散乱,但丽江黄芩与粘毛黄芩聚类效果好,可与前面特征相结合进行鉴别。

特征值 6 的聚类图如图 6 所示。对于特征值 6 取阈值 $d_6 = 0.13$, 可分为 $G_{61} = \{4,7\}$、$G_{62} = \{5,6,9\}$、$G_{63} = \{10\}$ 和 $G_{64} = \{1 \sim 3,8,11 \sim 23\}$。特征 6 的类别分布更加混乱,但说明了甘肃黄芩与正品黄芩的相似性,同时也可结合前面特征鉴别甘肃黄芩。

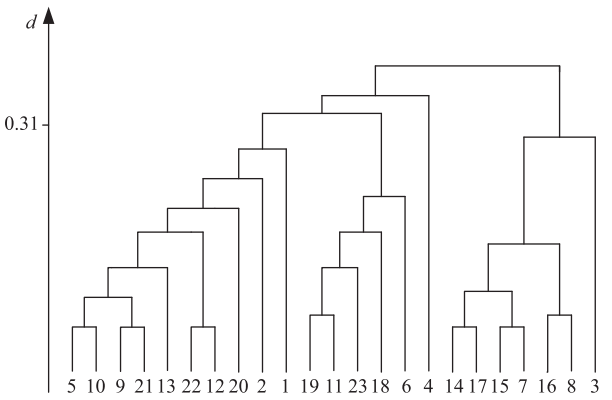


图 5 特征 1 聚类图

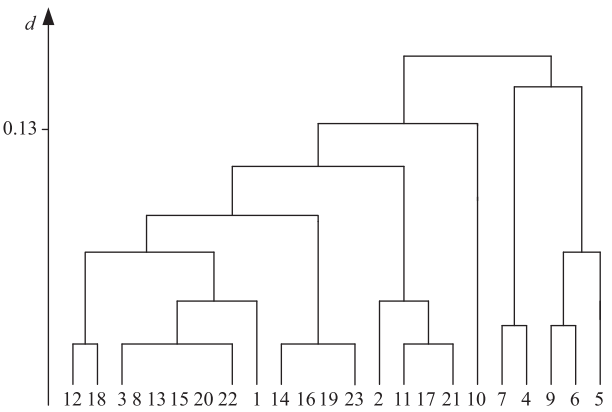


图 6 特征 6 聚类图

由前面的分析可以得到黄芩的标准指纹图谱,即为

$$\boldsymbol{H} = [d_7, d_4, d_5, d_2, d_3, d_8, d_1, d_6] = [0.41, 0.34, 0.86, 0.45, 1, 1, 0.31, 0.13] \quad (1)$$

式中特征值的顺序是按特征的分类能力由高到低排列的。同时,从以上分析可知,甘肃黄芩和粘毛黄芩与正品黄芩的相似性更大,而滇黄芩和丽江黄芩与正品黄芩的相似性较小。这与文献[11]中的结论一致。

依据上述分析,在保证训练样本中包含所有黄芩种类的前提下,随机选取矩阵表中的 20 个样本进行训练,其余作为测试,分类结果以测试集的 10 次交叉检验误差率作为评价指标。选择的分类器分别是线性分类器、k 近邻分类器以及以标准概率密度为基准的贝叶斯分类器。实验结果误差率分别为 5.13% 、3.87% 和 1.98%。由此可见,依据聚类规则建立标准指纹图谱的方法不失为一种可行的好方法。

此外,本实验采用的数据是气相色谱(GC)、薄

层色谱(TLC)、纸色谱(PC)、紫外光谱(UV)四种技术联用测得的中药指纹图谱数据,由于中药的复杂性,目前各种单一的测定方法往往很难得到比较完善的指纹图谱,因而采用多种技术联用来进行测量可获得较全面的中药指纹图谱。而且,依据聚类规则构建标准指纹图谱,实现了多种图谱信息的融合。

4 结 论

中药现代化是中药发展的必由之路。中药指纹图谱技术借助于现代分析手段,以标准指纹图谱作为参照图谱,控制中药质量,鉴别中药真伪,评价其安全性和有效性。但标准指纹图谱的传统建立方法存在诸多缺点,影响了结果的准确性。本文提出的标准指纹图谱建立方法以多参数距离融合聚类规则作为评价与鉴别标准,是一种聚类规则。由于它以分类为目的,以分类结果准确率高的聚类规则来定义标准指纹图谱,所以它所构建的标准指纹图谱一定更利于指纹图谱的分类。实验证明,该方法是个有意义的研究方向。同时,指纹图谱技术可广泛应用于基因研究领域,该方法的研究成果应用于这一领域的研究将会深入进行。

参考文献

[1] 邹纯才,鄢海燕. 中药指纹图谱及其数字化. 安徽:科学技术出版社,2008

[2] 罗国安,梁琼麟,王义明. 中药指纹图谱-质量评价、质量控制与新药开发. 北京:化学工业出版社,2009

[3] 高燕萍,周月芳,胡春湘. 易混品种的药材鉴别比较. 中华现代中医药杂志,2005,3(10):932-933

[4] 张铁军,姜顺善. 决明子的原植物研究. 中草药. 1993,24(1):40-41

[5] 马利飞,唐伯灵,李红等. 决明子及其伪品刺田菁种子的鉴别. 中药材,1993,16(10):20-21

[6] 罗国安,王义明,曹进. 多维多息特征谱及其应用. 中成药,2000,22(6):395-397

[7] 谢培山. 中药质量控制模式的发展趋势. 中药新药与临床药理,2001,12(3):188-191

[8] 张志永,张劲松,巩学千等. 抗 SMV 栽培大豆种质资源的 SCAR 标记指纹图谱分析. 高技术通讯,1998,10(1):49-53

[9] 白史且,高荣,沈翼等. 假俭草遗传多样性的 AFLP 指纹分析. 高技术通讯,2002,10:45-49

[10] 张勇,邓科君,张韬等. 水稻基因组 MSAP 指纹图谱构建及 DNA 甲基化修饰位点分离与鉴定. 高技术通讯,2009,19(9):83-990

[11] 张福良. 聚类分析与中药质量研究. 北京:人民卫生出版社,1994

Research on construction of standard TCM fingerprints based on multi-parameter distance clustering theories

Cui Jianxin[*,**], Cui Jianfeng[***], Hong Wenxue[*], Gao Haibo[*,****]

([*] School of Electric and Electronic Engineering, Yanshan University, Qinhuangdao 066004)

([**] Measurement Technology and Instrumentation Key Lab of Hebei Province, Qinhuangdao 066004)

([***] The Department of Tourism, Qinhuangdao Institute of Technology, Qinhuangdao 066004)

([****] Institute of Acupuncture and Moxibustion China Acadimy of Chinese Medical Sciences, Beijing 100700)

Abstract

The traditional methods for building standard TCM(traditional Chinese medicine) fingerprints were analyzed, and aiming at the limitations of the TCM fingerprints constructed by traditional methods in TCM's quality identification and evaluation, a novel method for construction of standard TCM fingerprints based on multi-parameter distance clusting theories was advanced. The new method builds standard TCM fingerprints by computing each feature clustering rule of multi-dimensional print data of TCM and integrating them, to realize the fusion of multi-print information. The total culstering rule is the standard TCM fingerprint. The experiments were conducted by using Scutellaria data. The experimental results show that the standard TCM fingerprints constructed based on cluster theories are better than the traditional standard TCM fingerprints in TCM classification. It is a feasible method for construction of standard TCM fingerprints.

Key words: TCM fingerprint, cluster analysis, information fusion, visualization