第 22 卷 第 2 期
2007 年 6 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.22 No. 2
Jun. 2007

基于模糊聚类的中药对照指纹图谱研究

王成哲，卢 佩，沈鸿章
（天津科技大学电子信息与自动化学院，天津 300222）

摘 要：针对一般模糊聚类法在样本特征相似时分辨力差，以及现有的建立对照指纹图谱方法在选择样本时容易出现的随机性强，偶然人为错选等弱点，将摄动的模糊数学聚类方法(FCMBP)引用到对照指纹图谱建立的前期样本选择上，提出一种基于模糊模式识别的建立中药对照指纹图谱的新方法. 通过田基黄薄层色谱指纹图谱实验证明，该方法能进一步提高对照指纹图谱的有效性和合理性.

关键词：中药；对照指纹图谱；FCMBP；模糊聚类
中图分类号：TP18 **文献标识码**：A **文章编号**：1672-6510（2007）02-0062-04

Research of Establishing Reference Fingerprint for the Traditional Chinese Medicines Based on Fuzzy Pattern Recognition

WANG Cheng-zhe，LU Pei，SHEN Hong-zhang
（College of Electronic Information and Automation，Tianjin University of Science & Technology，Tianjin 300222，China）

Abstract：The new method of creating traditional Chinese medicine fingerprint based on the fuzzy pattern recognition was put forward by introducing fuzzy clustering method based on perturbation (FCMEP) into sample choosing in the earlier stage of creating reference fingerprint, due to fuzzy cluster method definition poor while sample characteristics similar, creating reference fingerprint method randomness strong when choosing sample, accidental mistake etc. It was proved by Hypercum japonicum fingerprint experiment that this method can improve the validity and rationality of reference fingerprint.

Keywords：traditional Chinese medicine；reference fingerprint； FCMBP；fuzzy clustering

指纹图谱分析通常由指纹图谱获取和指纹图谱计算两部分所组成,前者是采用色谱、光谱或二者联用等分析方法，获取能表征样本化学组成特征的组分来分析图谱或图像；后者是运用计算分析的方法对所获得的图谱进行数据处理，获得专属、宏观、整体的化学特征综合信息，并对样品化学组成的总体波动情况进行估测. 应用指纹图谱监控中药材的质量是否稳定,在国内已逐渐成为一种趋势，在国外也越来越多的受关注. 目前，相似度已被国家药典委员会确定为中药指纹图谱标准中的一项重要评价指标. 其中,建立对照指纹图谱是计算相似度的一个关键环节.

在没有提供对照指纹图谱的情况下，建立对照指纹图谱的方法有两种：（1）选择典型样品的指纹图谱作为对照指纹图谱；（2）选择共有模式作为对照指纹图谱，即通过对一批指纹图谱的研究模拟计算出指纹图谱数据作为对照指纹图谱，目前产生共有模式的算法有平均矢量法和中位数矢量法. 方法（1）对选择者的要求较高，在选择过程中难免会出现随意性. 方法（2）直接用样本计算指纹共有模式也存在超常样本参与合成对照指纹图谱的可能性. 为建立更合理的对照指纹图谱，许多学者也进行了相关的研究，并取得了较好的效果. 马成俊等[1]使用中国药典委员会推荐的“计算机辅助相似度评价软件”进行数据处理，对不同产地金银花药材指纹图谱的相似度进行比较分析,构建金银花药材对照指纹图谱. 赖何季等[2]应用传递闭包法先进行样本选择,再建立中药对照指纹图谱.

然而，这些研究还存在不足：（1）在精确聚类上还有欠缺. 常用的模糊聚类方法求传递闭包的过程经

收稿日期：2006-09-28；**修回日期**：2007-03-20
作者简介：王成哲（1984—），男(回族)，安徽巢湖人，硕士研究生.

过一系列的非恒等变换，因此由模糊等价矩阵得到的聚类是否真实的反映了原始问题的聚类情况，这在理论上还没有得到严格的保证，“失真”问题没有得到解决；（2）对超常样本的处理上还不太完善.大多数方法在建立对照指纹图谱前没有剔除可能影响图谱精确性的超常样本. 本文根据数学上全局最优不具有唯一性，但局部最优有唯一性[3]结论和摄动的解决“失真”问题的思路，提出“收缩”FCMBP 法聚类中药样本，试图寻找一个与原始的模糊相似矩阵按某种“距离”最小的模糊等价矩阵，可在一定程度上解决聚类“失真”的问题；根据设定阈值剔除超常样本，再用平均矢量等方法计算处理后，便可得到中药对照指纹图谱.

1　原理与方法

在色谱法测定指纹图谱的研究中，由于图谱中峰众多，而且每张样本图谱中峰的各项参数并不完全一致，因此把保留时间和峰面积作为数字特征，筛选出能代表图谱特征的一组参数 $x_i=(x_{i1},x_{i2},\cdots,x_{im})$ 作为参与模糊聚类的数据矩阵.

1.1　“收缩”FCMBP 聚类法的基本原理

在“收缩”FCMBP 聚类法分析过程中应用到如下理论：

定义 1　X 为有限论域，模糊关系 R 的模糊矩阵 $R=r(i,j)_{m\times n}$ 当且仅当满足以下条件：（1）自反性 $r_{ij}=1$，（2）对称性 $r_{ij}=r_{ji}$，（3）传递性 $\overset{n}{\underset{k=1}{\vee}}(r_{ik}\wedge r_{kj})\leqslant r_{ij}$，则称 R 是一个模糊等价矩阵.

定义 2　设 R 是论域 X 中的模糊关系，包含 R 的最小模糊传递关系称为 R 的传递闭包，记为 $t(R)$.

定义 3　设 $\lambda_0\in[0,1]$.如果动态聚类时，$\lambda\geqslant\lambda_0$ 所得的类群数目与 $\lambda<\lambda_0$ 所得的类群数目不同，则称此 λ_0 为聚类的分界水平.

由文献[4]的证明可得：

（1）设模糊相似矩阵 $R=(r_{ij})_{n\times n}$，$t(R)=(t_{ij})_{n\times n}$ 是它的传递闭包，$E_k=\{r_{ij}\mid t_{ij}=t_k,r_{ij}\in R\}$，$t_k'$ 为 E_k 中元素的平均值.当 $r_{ij}\in E_k$ 时，令 $r_{ij}'=t_k'$，由此得到等价矩阵 $R'=(r_{ij}')_{n\times n}$，并且 R' 与 R 的距离小于 $t(R)$ 与 R 的距离.

（2）基于相似矩阵的传递闭包，按照“收缩”方法得到矩阵的聚类分界水平小于相对应的传递闭包聚类的分界水平.

1.2　建立对照指纹图谱的方法

1.2.1　数据标准化

由于本实验的所有的数据参数都具有相同的量纲，实验可直接做平移—极差变换：

$$x_{ik}'=x_{ik}-\min_{1\leqslant i\leqslant n}\{x_{ik}\}\Big/\max_{1\leqslant i\leqslant n}\{x_{ik}\}-\min_{1\leqslant i\leqslant n}\{x_{ik}\}\quad(k=1,2,\cdots,m)$$

其中，$(x_{i1},x_{i2},\cdots,x_{im})$ 为参与模糊聚类的数据矩阵. x_{ik}' 是标准化后的变量，且 $x_{ik}'\in\{0,1\}$.

1.2.2　建立模糊相似矩阵

本文引进夹角余弦法，即

$$r_{ij}=\sum_{k=1}^{m}x_{ik}\cdot x_{jk}\Big/\sqrt{\sum_{k=1}^{m}x^2{}_{ik}\cdot\sum_{k=1}^{m}x^2{}_{jk}}$$

作为数量指标来衡量样本间的相似程度.

1.2.3 聚类样本并选择样本

（1）计算 $R=(r_{ij})_{n\times n}$ 的传递闭包 $t(R)=(t_{ij})_{n\times n}$，计算公式为 $R\Rightarrow R^2\Rightarrow R^4\Rightarrow\cdots\Rightarrow R^{2k}=t(R)$，其中 k 为自然数. $R^2=\{r_{ij}{}^{(2)}\}=\overset{n}{\underset{k=1}{\vee}}[r_{ik}\wedge r_{kj}]$，$R^4=\{r_{ij}{}^{(4)}\}=\overset{n}{\underset{k=1}{\vee}}[r_{ik}\wedge r_{kj}]\cdots$，直到 $R^{2k}=R^k$，计算完成. 并标出不同元素值 $t_1,t_2,\cdots,t_m(1\leqslant m\leqslant n)$.

（2）根据 $t(R)$ 中相对应位置的元素取值是否相同，将 $R=(r_{ij})_{n\times n}$ 中的元素分成 m 类并计算每类元素的算术平均值.与 t_k 相应的类为 $E_k=\{r_{ij}\mid t_{ij}=t_k,r_{ij}\in R\}$，其平均值记为 t_k'，$k=1,2,\cdots,m$.

（3）对矩阵 R 中的元素进行变换，即令 $r_{ij}'=t_k'$，当 $r_{ij}\in E_k,k=1,2,\cdots,m$.这样便应用数学方法，由原来的相似矩阵得到了一个新的矩阵 $R'=(r_{ij}')_{n\times n}$，这就是所求的模糊等价矩阵.

（4）将 R' 中不同元素的值按从大到小的顺序依次排列为 $1\geqslant\lambda_1>\lambda_2>\cdots\lambda_t$.可得一系列 λ 的截矩阵，即 $R_\lambda'=(r_{ij}'(\lambda))_{n\times n}$，其中 $r(\lambda)=\begin{cases}1, & r_{ij}'\geqslant\lambda\\ 0, & r_{ij}'<\lambda\end{cases}$.

对于对象 x_i,x_j，若 $r_{ij}'(\lambda)=1$，则在水平上将它们归为一类，从而达到模糊聚类的目的. 根据剔除超常样本的需要，在 λ 值上进行截取，得到所需的分类. 当 λ 从 1 逐步降为 0，类的个数由多变少，逐步归并，就可以形成动态聚类图，此图能简洁明了的表示模糊聚类的结果.

（5）以 λ 所取的值作为阈值，可将样本中超常的部分样本剔除.

1.2.4 计算共有模式

共有模式的计算通常可用两种算法：一种是平均矢量，另一种是中位数矢量. 本实验用平均矢量法计算，其公式为

$$\text{平均矢量}=\sum[(X_{1j},X_{2j},\cdots,X_{ij},\cdots,X_{nj})/n].$$

2　实验与分析

采用田基薄层色谱指纹图谱实验数据，见表1. 计算使用Matlab7.0编程实现. 使用1—10号药材的指纹图谱数据建立对照指纹图谱.

表1　田基黄薄层色谱指纹图谱各指纹峰保留值及峰面积

Tab.1　Retention time and peak area of Hypercum japonicum fingerprints of different resources

药品编号	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
1	134.7	105	4 213.7	1 779.1	4264.2	3 494.1	39.4	109.3	1 109.3	0
2	169.5	233.3	4 050.9	1 556.3	4139.2	3 173.5	21.2	49	913	0
3	45	96.1	5 278.6	665.7	1553.8	5 610.9	0	82.6	431.3	0
4	435	93.3	5 168.8	686.1	1486.7	1 486.7	0	96.1	445.8	0
5	90.1	69.2	1 552.6	298.2	472.3	1 576.8	0	39.2	280.3	0
6	99.9	0	1518	267.2	405.3	1 601.3	0	0	188.8	0
7	90	136.1	1 135.2	804.5	255.3	5 475.3	0	0	90	211.3
8	329.8	0	1 368.1	434.6	953.6	2 228.5	0	0	417.6	0
9	327.9	0	1 314	418.5	932.5	2 146.6	0	0	371.8	0
10	41.5	47.6	1 768	651.9	857.7	2 684.7	176	411	355	0
R_f	0.01	0.02	0.05	0.1	0.13	0.18	0.24	0.26	0.32	0.33
药品编号	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
1	974.7	0	6 478.2	317.1	1 154.1	654.8	541.7	308.5	131.3	0
2	1 339.6	0	8 342.9	399.9	1 204.3	734.8	559	289.2	218.4	0
3	1 123.3	0	1 0037	747.2	1 419.2	542.2	502.9	354.4	243.8	0
4	1 185.9	0	8 780.8	990.9	1 743.2	949.3	1 196.5	1 509.8	534.2	0
5	570.5	0	4 044.1	256.1	6 6.6	394.5	462.8	261.7	71.8	0
6	452.8	0	4 537.6	268.6	788.6	529.2	522.9	287.6	184	0
7	0	312.3	135	5 000	300.6	1 500.2	0	1 361.6	398	158.7
8	1 194.6	0	8 364.1	1 257	1 282.2	571.9	814.2	468.6	361.6	0
9	1 161.9	0	8 162.4	1 221	1 158.9	579.3	787.5	395.2	338.3	0
10	1 080.3	07	6 903.9	0	1 943.7	0	1 130.3	790.5	124.6	0
R_f	0.41	0.44	0.57	0.51	0.55	0.57	0.62	0.7	0.75	0.83

注：P1—P20代表色谱图20个峰的标号，其中1、2号药材产地为广西桂林，3、4号药材产地为广西桂平，5、6号药材产地为广西梧州，8、9号药材为广西百色，10号药材产地未知，7号为加入的超常样本.

由步骤1—3得到所求的模糊等价矩阵 R'.

$$R'=\begin{bmatrix}
1.000\,0 & 0.987\,8 & 0.911\,6 & 0.911\,6 & 0.911\,6 & 0.911\,6 & 0.374\,4 & 0.911\,6 & 0.911\,6 & 0.911\,6\\
0.987\,8 & 1.000\,0 & 0.911\,6 & 0.911\,6 & 0.911\,6 & 0.911\,6 & 0.374\,4 & 0.911\,6 & 0.911\,6 & 0.911\,6\\
0.911\,6 & 0.911\,6 & 1.000\,0 & 0.989\,2 & 0.974\,3 & 0.974\,3 & 0.374\,4 & 0.953\,9 & 0.953\,9 & 0.965\,2\\
0.911\,6 & 0.911\,6 & 0.989\,2 & 1.000\,0 & 0.974\,3 & 0.974\,3 & 0.374\,4 & 0.953\,9 & 0.953\,9 & 0.965\,2\\
0.911\,6 & 0.911\,6 & 0.974\,3 & 0.974\,3 & 1.000\,0 & 0.997\,1 & 0.374\,4 & 0.953\,9 & 0.953\,9 & 0.965\,2\\
0.911\,6 & 0.911\,6 & 0.974\,3 & 0.974\,3 & 0.997\,1 & 1.000\,0 & 0.374\,4 & 0.953\,9 & 0.953\,9 & 0.965\,2\\
0.374\,4 & 0.374\,4 & 0.374\,4 & 0.374\,4 & 0.374\,4 & 0.374\,4 & 1.000\,0 & 0.374\,4 & 0.374\,4 & 0.374\,4\\
0.911\,6 & 0.911\,6 & 0.953\,9 & 0.953\,9 & 0.953\,9 & 0.953\,9 & 0.374\,4 & 1.000\,0 & 0.999\,9 & 0.953\,9\\
0.911\,6 & 0.911\,6 & 0.953\,9 & 0.953\,9 & 0.953\,9 & 0.953\,9 & 0.374\,4 & 0.999\,9 & 1.000\,0 & 0.953\,9\\
0.911\,6 & 0.911\,6 & 0.965\,2 & 0.965\,2 & 0.965\,2 & 0.965\,2 & 0.374\,4 & 0.953\,9 & 0.953\,9 & 1.000\,0
\end{bmatrix}$$

步骤3得到的传递闭包 $t(R)$ 聚类图见图1 ，最终得到的模糊相似矩阵 R' 聚类图见图2. 在图2中选择 λ=0.9，由聚类分析可得7号药材被归为一类，其余药材被归为另一类，所以剔除7号药材，余下药材数据根据平均矢量法计算对照指纹图谱（即共有模式）.

选择所有药材样本与建立的共有模式进行相似度计算，相似度测量用余弦夹角法. 采用本文所述方法，进行模糊聚类对样本进行选择后再使用平均矢量法计算共有模式，然后计算样本与共有模式的相似度，其结果见表2的相似度Ⅱ.（表2相似度Ⅰ为不使用模糊聚类法，直接用平均矢量计算共有模式，然后计算样本与共有模式的相似度）.

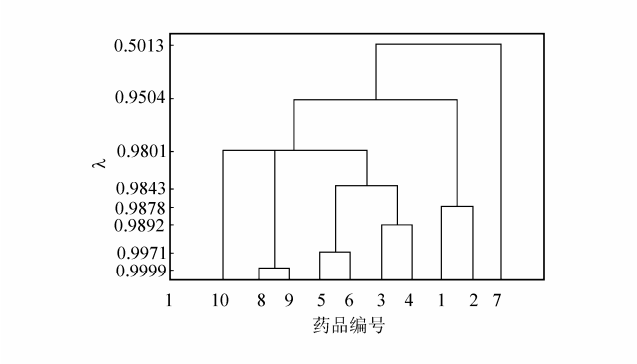

图 1　t (R)聚类分析结果
Fig. 1　t(R)cluster analysis result

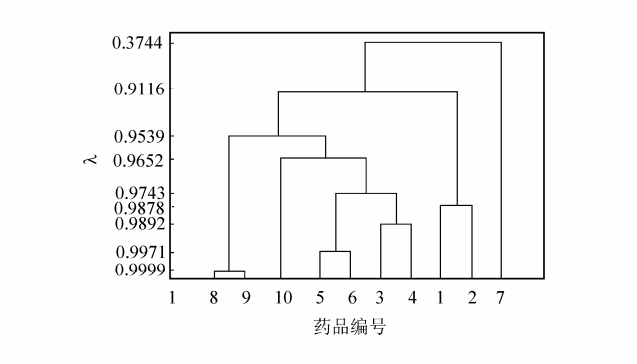

图 2　R' 聚类分析结果
Fig. 2　R'cluster analysis result

表 2　两种方法得出指纹图谱的相似度

Tab. 2　Similarity of the fingerprints with the two proposed methods

药材编号	相似度 Ⅰ	相似度 Ⅱ
1	0.940 2	0.944 0
2	0.967 7	0.973 5
3	0.988 4	0.986 6
4	0.982 2	0.978 8
5	0.993 8	0.994 6
6	0.987 8	0.988 7
7	0.902 0	0.883 0
8	0.964 6	0.966 9
9	0.963 5	0.965 9
10	0.979 0	0.977 0

图2与图1比较可以看出传递闭包法聚类的缺陷：它的聚类分界水平 λ 的集中度比较高，这不利于有效、快速的聚类，相应分类的敏感性也就较弱，这就不能明显的区分类别间的差异.且 $d(R,R')=0.228\,3$，$d(R,t(R))=0.476\,2$，$d(R,R') < d(R,t(R))$，这说明 R' 是相对于 $t(R)$ 失真较小的模糊等价矩阵，尤其是在大量样本聚类时，它的聚类更合理，由它计算选择的样本更准确.

比较表 2 的两个结果，如果按相似度大于 0.9 来评价中药，那么超常样本在方法一中被鉴定为质量过关的产品，在方法二中被鉴定为质量不过关产品.所以方法二的结论能很好的体现实际结果. 同时由表中相似度可以看出同一产地的药材相似度比较接近.

3　结　语

考虑超常样本对建立对照指纹图谱的影响和传递闭包模糊聚类法自身的一些缺陷，结合数学算法，本文对传递闭包模糊聚类法进行改进，提出用“收缩”FCMBP 聚类法先进行样本选择，剔除超常样本，然后再用平均矢量法建立共有模式作为中药对照指纹图谱，该方法能进一步提高对照指纹图谱的合理性，在一定程度上提高了中药指纹图谱相似度计算的有效性. 为没有提供对照指纹图谱的情况下，建立准确可信的对照指纹图谱提供了一种新方法.

参　考　文　献：

［1］ 马成俊，李桂生，任召言，等. 金银花药材对照指纹图谱的建立及质量评价[J]. 时珍国医国药，2006, 17（2）：238—240.

［2］ 赖何季，朱学峰，许建新. 一种建立中药“对照指纹图谱”方法的研究[J]. 天然产物研究与开发，2005，17：80—83.

［3］ 何　清，李洪兴. 模糊聚类中的模糊等价矩阵[J]. 系统工程理论与实践，1999，19（4）：8—11，69.

［4］ 苗　丽. FCMBP 聚类及在学生成绩评估中的应用[D]. 北京：北京师范大学，2004.

［5］ 李　侃，刘玉树. 模糊聚类的自适应算法[J]. 决策与控制，2004，19（5）：595—597.

［6］ 周敬泉，颜春兰，袁　鹏，等. 用模糊聚类研究中药成分特征谱[J]. 控制理论与应用，2004，21（4）：569—573.