优化星座聚类法和有机氟农药构效关系识别*1)

陈曦　王丽君　胡上序
（浙江大学化工系）
林学圃　吴军
（浙江化工研究院）

THE OPTIMIZED CLUSTERING METHOD OF THE CONSTELLATION GRAPH AND THE STRUCTURE-ACTIVITY RELATIONSHIP OF ORGANIC FLUORINATED PESTICIDES

Chen Xi　Wang Lijun　Hu Shangxu
(*Zhejiang University*)
Lin Xuepu　Wu Jun
(*Zhejiang Research Institute of Chemical Engineering*)

Abstract

Since the study of organic fluorinated pesticides has been a noticeable field in the seaking of efficient pesticides, the research of their structure-activity relationships becomes more and more important. If the structure features of the chemicals are considered as pattern parameters, and the activities of pesticides are discretized into classes, then the quantitative structure-activity relationship (QSAR) problem can be treated by cluster analysis. Hence, the clustering method of constellation graph is employed to identify the QSAR of fluorinated organic pesticides. In order to obtain better clustering results, the weights of star-tracks are optimized by the Lagrange operator method and good results are achieved.

§1. 引　言

有机氟农药是近十余年发展起来的一类新型农药，在杀菌、杀虫、除草等方面都有显著效果．因此，对有机氟农药的定量构效关系 (QSAR) 的研究具有重要意义．

* 1996 年 4 月 29 日收到．
1) 国家自然科学基金资助项目．

QSAR 研究的实质 [1] 是找出化合物分子结构与药效活性间的定量统计关系，从而发现结构参数对化合物活性影响的规律，为合成新药物指引方向.

可用以表达农药结构的物化参数有许多种类，例如疏水性、电特性、立体性、环境分支性等，但它们与药效的关系在理论上尚无定量描述. 将化合物的结构特征作为信息模式空间的坐标变量，把药物的活性离散地用类别表示，从而把 QSAR 转换为一种聚类问题. 考虑到星座图 (Constellation Graph) 法在多元聚类分析中的简捷和直观性，本文对星座路径进行优化处理，然后将它应用于药物的构效关系识别.

§2. 星座图的建造

2.1. 数据变换. 使用星座图法 [2,3] 时，首先要将模式数据作线性变换，使变换后的数值在 $[0,\pi]$ 内. 设有 n 个模式，每个模式由 p 个分量构成，用矩阵 X 表示全部模式数据为

$$X=\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix},$$

并令

$$\xi_{ij}=\frac{x_{ij}-x_{\min,j}}{R_j}\pi, \tag{1}$$

其中

$$x_{\min,j}=\min_{1\le i\le n} x_{ij},\quad x_{\max,j}=\max_{1\le i\le n} x_{ij},\ R_j=x_{\max,j}-x_{\min,j},$$

则变换后的数据 $[\xi_{ij}]$ 均落在 $[0,\pi]$ 内.

2.2. 星的路径和位置. 取一组路径权值 $\{\omega_j\}$，使满足

$$\omega_j\ge 0,\quad j=1,2,\cdots,p,\quad \sum_{j=1}^{p}\omega_j=1.$$

一般都取等权，即

$$\omega_1=\omega_2=\cdots=\omega_p=1/p.$$

画一半径为 1 的半圆及底边直径，每个模式对应半圆内的一个点，称为星. 设有模式 X_1，首先以 0 为圆心，ω_1 为半径画一上半圆，在圆周上对应弧度为 ξ_{11} 的点是 0_1；然后再以 0_1 为圆心，ω_2 为半径画一上半圆，在圆周上对应弧度为 ξ_{12} 的点是 0_2；依此类推，直到 0_p 为止，如图 1 所示. 即 0_p 为与 X_1 对应的星座的位置，而以 $(00_10_2\cdots 0_p)$ 表示该星从原点出发的路径. 由以上可得出与任一模式 X_α 对应的星座位置坐标为

$$\Big(\sum_{j=1}^{p}\omega_j\cos\xi_{\alpha j},\ \sum_{j=1}^{p}\omega_j\sin\xi_{\alpha j}\Big). \tag{2}$$

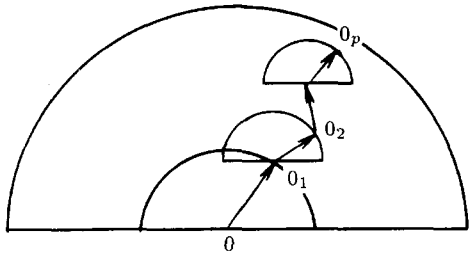

图 1.　星座图示意

§3. 路径的优化

为使同类模式在星座图中的位置距离尽量靠近，不同类模式的距离尽量拉开，本文对路径权值的确定改进如下：设待识别的数据要求分为 m 类，对于两类问题，$m=2$. 如果任何第 g 类中有 n_g 个模式，每个模式都有一个 p 维向量，即

$$X^{(g)}=\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n_g 1} & \cdots & x_{n_g p} \end{bmatrix},\quad g=1,\cdots,m,\quad X=\begin{bmatrix} X^{(1)} \\ \vdots \\ X^{(m)} \end{bmatrix}.$$

令

$$\xi_{kj}^{(g)}=\pi\frac{x_{kj}^{(g)}-x_{\min,j}}{x_{\max,j}-x_{\min,j}},\quad k=1,\cdots,n_g,\quad j=1,\cdots,p,\quad g=1,\cdots,m,$$

$$x_{\min,j}=\min_{1\le i\le n}x_{ij}^{(g)},\quad x_{\max,j}=\max_{1\le i\le n}x_{ij}^{(g)}. \tag{3}$$

为使同类模式在星座图中的位置距离尽量靠近，不同类模式间的距离尽量拉开，定义目标函数

$$Q=\sum_{g=1}^{m}\sum_{k=1}^{n_g}\left[\left(x^{(g)}-\sum_{j=1}^{p}\omega_j\cos\xi_{kj}^{(g)}\right)^2+\left(y^{(g)}-\sum_{j=1}^{p}\omega_j\sin\xi_{kj}^{(g)}\right)^2\right],$$

$$\sum_{j=1}^{p}\omega_j=1,\quad \omega_j\ge 0,\quad j=1,\cdots,p,$$

其中

$$x^{(g)}=\sum_{j=1}^{p}\omega_j\overline{\cos\xi_j^{(g)}},\quad y^{(g)}=\sum_{j=1}^{p}\omega_j\overline{\sin\xi_j^{(g)}},\quad g=1,\cdots,m,$$

$$\overline{\cos\xi_j^{(g)}}=\frac{1}{n_g}\sum_{k=1}^{n_g}\cos\xi_{kj}^{(g)},\quad \overline{\sin\xi_j^{(g)}}=\frac{1}{n_g}\sum_{k=1}^{n_g}\sin\xi_{kj}^{(g)}.$$

为了找出能使目标函数 Q 为最小的一组权值，对上述二次规划模型，引入拉格朗日乘子 λ, 得到

$$L = Q + \lambda\Big(\sum_{j=1}^{p}\omega_j - 1\Big). \tag{4}$$

根据极值原理

$$\begin{cases} \partial L/\partial\omega_j = 0, \\ \partial L/\partial\lambda = 0, \end{cases}$$

可得到

$$\begin{cases} \partial\Big[Q + \lambda\Big(\sum_{j=1}^{p}\omega_j - 1\Big)\Big]\Big/\partial\omega_j = 0, \\ \sum_{j=1}^{p}\omega_j - 1 = 0. \end{cases} \tag{5}$$

因为

$$L = \sum_{g=1}^{m}\sum_{k=1}^{n_g}\Big\{\Big[\sum_{j=1}^{p}\omega_j(\overline{\cos\xi_j^{(g)}} - \cos\xi_{kj}^{(g)})\Big]^2 + \Big[\sum_{j=1}^{p}\omega_j(\overline{\sin\xi_j^{(g)}} - \sin\xi_{kj}^{(g)})\Big]^2\Big\} + \lambda\Big(\sum_{j=1}^{p}\omega_j - 1\Big),$$

所以

$$\frac{\partial L}{\partial\omega_j} = \sum_{g=1}^{m}\sum_{k=1}^{n_g}\left\{\begin{array}{l} 2\Big[\sum_{h=1}^{p}\omega_h(\overline{\cos\xi_h^{(g)}} - \cos\xi_{kh}^{(g)})\Big](\overline{\cos\xi_j^{(g)}} - \cos\xi_{kj}^{(g)}) \\ +2\Big[\sum_{h=1}^{p}\omega_h(\overline{\sin\xi_h^{(g)}} - \sin\xi_{kh}^{(g)})\Big](\overline{\sin\xi_j^{(g)}} - \sin\xi_{kj}^{(g)}) \end{array}\right\} + \lambda = 0.$$

将 $j = 1, 2, \cdots, p$ 代入，并令

$$s_{jh} = \sum_{g=1}^{m}\sum_{k=1}^{n_g}\left[\begin{array}{l} (\overline{\cos\xi_j^{(g)}} - \cos\xi_{kj}^{(g)})(\overline{\cos\xi_h^{(g)}} - \cos\xi_{kh}^{(g)}) \\ +(\overline{\sin\xi_j^{(g)}} - \sin\xi_{kj}^{(g)})(\overline{\sin\xi_h^{(g)}} - \sin\xi_{kh}^{(g)}) \end{array}\right],$$

得到

$$\begin{cases} \omega_1 s_{11} + \omega_2 s_{12} + \cdots + \omega_p s_{1p} + \lambda = 0, \\ \omega_1 s_{21} + \omega_2 s_{22} + \cdots + \omega_p s_{2p} + \lambda = 0, \\ \quad\vdots \quad\quad \vdots \quad\quad \vdots\ \vdots \\ \omega_1 s_{p1} + \omega_2 s_{p2} + \cdots + \omega_p s_{pp} + \lambda = 0. \end{cases} \tag{6}$$

令

$$\omega = (\omega_1, \omega_2, \cdots, \omega_p)^T, \quad S = (s_{jh})_{p\times p},$$

得

$$S\omega = -\lambda I_p, \tag{7}$$

其中 I_p 为 p 维单位列向量.

又由

$$\omega_1 + \omega_2 + \cdots + \omega_p = 1$$

得到

$$I_p^T \omega = 1. \tag{8}$$

根据式 (7) 得

$$\omega = -S^{-1}\lambda I_p = -\lambda S^{-1} I_p. \tag{9}$$

代入式 (8) 得 $-I_p^T \lambda S^{-1} I_p = 1$, 即

$$\lambda = \frac{-1}{I_p^T S^{-1} I_p}. \tag{10}$$

代入式 (9) 得

$$\omega = \frac{S^{-1} I_p}{I_p^T S^{-1} I_p}. \tag{11}$$

由式 (11) 可以得到经过优化后的权值，由此得到的星座图具有同类模式的距离接近，不同类模式的距离拉开的特点.

§4. 特征参数的组织

本文用 17 个含氟农药试样，并选取在研究药物构效关系中最常用的四个参数：1) 疏水参数； 2) 电性参数； 3) 立体性参数； 4) 环境参数，作为特征参数. 各试样经实验确定的活性类别和特征参数值列于表 1.

这样，就把构效关系问题转变为由以上各参数表示的模式与活性类别之间的模式聚类问题.

在经过优化后的星座图中，设 “好” 类的样本中各模式之间的欧氏距离的平均值为 $d1$, “差” 类的样本中各模式之间的欧氏距离的平均值为 $d2$, 两类样本中心的欧氏距离为 $d12$, 取 $D = (d1 + d2)/d12$ 作为总样本聚类好坏的判别标准. 显然， D 越小，聚类结果越好.

表 1. 试样的活性类别和特征参数数据

		模式数据			
试样编号	活性类别	1: 疏水参数	2: 电性参数	3: 立体参数	4: 环境参数
X1	好	3.36	64.35	−1.39	21.93
X2	好	2.93	60.16	−1.08	21.29
X3	好	2.57	76.98	−1.81	20.15
X4	好	2.65	58.32	−1.76	22.56
X5	好	3.62	79.32	−1.70	20.03
X6	好	3.92	70.00	−1.46	21.28
X7	好	4.26	67.14	−0.13	21.80
X8	差	3.99	80.63	−1.33	20.15
X9	差	4.63	67.62	−1.35	22.02
X10	差	4.39	81.71	−2.39	20.35
X11	差	2.84	86.74	−0.76	18.58
X12	差	6.46	86.62	−0.74	20.03
X13	差	5.90	101.34	−2.28	18.81
X14	差	4.74	90.62	−1.84	18.42
X15	差	3.21	63.97	−1.83	21.93
X16	差	4.27	73.27	−0.56	20.65
X17	差	4.18	84.97	−1.77	19.53

首先，分别取四个参数之一作为模式参数，即输入数据为一维的情况，可以得到四个不同的星座图，表 2 列出了聚类结果，可以看出无论哪一个都无法将两种类别有效的分开，尤以第三个参数为差，说明任何单个参数都无法反映试样的构效关系：

表 2. 一维参数星座图结果

参数选择	1	2	3	4
D 的数值	1.5125	1.3454	9.9967	1.7559

然后，考虑输入参数为两维，即分别将以上四个参数两两组合作为输入模式数据，得到六个不同的星座图，结果见表 3.

表 3. 二维参数星座图结果

参数选择	1～2	2～3	3～4	1～4	1～3	2～4
D 的数值	1.1739	2.3125	2.9803	6.7037	2.8414	4.9471

其中，第四个参数 (环境参数) 与其它参数的组合性很差，组合后的结果还不如将它单独作为特征参数时的结果. 另外，在表 2 中我们知道立体参数的识别性很差，而将它与其它参数组合后，效果相对它本身而言有所改善，但比起和它组合的单个参数的聚类结果而言，反而有所降低. 这说明了立体参数无论单独作为特征参数，还是与

其它参数组合，都不具备好的结果．因此，这个参数在本问题研究中应该被剔除．最后，可以从表 3 中看到，由疏水参数和电性参数组合的星座图，可以很好地将不同活性的试样分开，其组合结果好于各自的单个结果，并且从以上分析中我们可以预测到：选择疏水参数和电性参数作为特征参数应当是最佳的，其它两个参数的介入将会降低聚类结果．

为了验证上述预测，我们依次计算输入模式为三维和四维时的情况，可以分别得到四个和一个星座图，结果列于表 4．由此可知，各分类效果均不如由疏水参数和电性参数组合时的结果，这也有力地证明了前述的预测．

表 4．多维参数星座图结果

参数选择	1～2～3	2～3～4	1～3～4	1～2～4	1～2～3～4
D 的数值	1.8431	8.0538	17.638	3.429	4.7577

由此可以判断，对此问题，疏水参数和电性参数是最有效的特征参数组合．在随后的训练和验证过程中，将只选它们作为模式参数．

§5．训练与验证

由上知道，对于不同的模式集合，我们可以得到经过优化后的不同的权值．在全部试样中取出一定数量的试样作为训练样本，其余的作为验证样本．当训练样本确定后，由式 (11) 可以得到对应于该样本的经优化的星座路径权值．最后，根据这些权值，用验证样本检验分类效果．

§6．结果与讨论

首先，我们比较权值是否优化对聚类结果的影响．选择疏水参数和电性参数作为特征参数，对于无权值优化的星座图法，取 $\omega_1 = \omega_2 = 1/2$；而优化星座图法按式 (11) 计算路径权值．两种方法的结果比较如表 5 所示．因此可以验证，权值优化星座图法比普通星座图法具有更好的聚类效果．

表 5．方法结果比较

	权值优化方法	普通方法
聚类结果	$D = 1.1739$	$D = 1.2708$

然后，利用权值优化星座图法进行识别分类．将 1—5 号试样构成活性好的一类的训练样本，8—15 为差活性类的训练样本．除第 15 号外，其它所有试样可以很好地按活性分开．然后，分别用活性好的 6 和 7 号试样和活性差的 16 和 17 号试样作为两类的验证样本进行验证，结果非常理想，它们都落在正确的区域内 (图 2)．

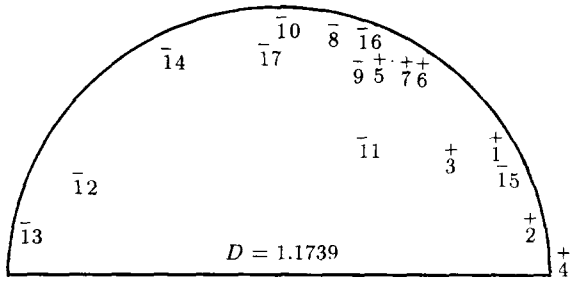

图 2. 由疏水参数和电性参数组合的星座图

因此认为，权值优化星座图法不但可以较好地解决复杂模式分类问题，而且能有效地组织和选择特征参数，最好地表达对象的特性，是研究含氟农药构效关系的一种有效工具.

参 考 文 献

[1] 徐景达译，药物结构与活性的关系，人民卫生出版社， 1987.
[2] 王玺等著，计算机与应用化学， **10**:4 (1993).
[3] 方开泰著，实用多元统计分析，华东师范大学出版社， 1989.