第14卷 第4期 环 境 科 学 学 报 Vol. 14, No. 4
1 9 9 4年12月 ACTA SCIENTIAE CIRCUMSTANTIAE December, 1994

判别符合率递增法辅助酚类化合物构效关系研究*

孙立贤 徐 芬** 宋新华 俞汝勤
(湖南大学化学化工系,长沙 410082)

摘 要

计算了酚类化合物的分子连接性指数($^1X^v$)及其在正辛醇和水之间的分配系数 $\log P$,运用新近提出的数学方法——判别符合率递增法,对酚类化合物的构效关系进行了研究,所得结果优于逐步判别法及穷举法。

关键词:化学计量学;逐步判别法;穷举法;判别符合率递增法。

1 前 言

定量构效关系研究是化学计量学的一个重要分支. 本文试图研究环境毒物——酚类化合物的构效关系,此项工作对于环境保护具有十分重要的意义,这是因为通过实验来测定酚类等环境毒物的毒性类别是十分昂贵和耗时的. 关于定量构效关系的研究方法有多种,例如:SIMCA[1]、聚类分析[2]、线性判别分析[3]、逐步多元线性回归分析[4]等. 建立在 Wilks 统计量u($u = E/W$, E为组内离差阵,W为总离差阵)为基础的逐步判别分析已在许多领域中得到了广泛的应用,但实际上在 Wilks 统计量 u 与错分数之间存在很大的差别,况且人们往往对错分数更感兴趣. 因此,谭良提出了判别符合率递增法[5],此法已成功地预测了工业厂房的抗震性能. 本文运用判别符合率递增法来研究酚类化合物的结构活性关系,结果表明此法优于逐步判别法及穷举法.

2 理 论

2.1 判别符合率递增法

设含m个变量的样本 X_i^α 取自样本容量为 n_α 的第 α 类总体 G_α($t = 1, 2, \cdots, n_\alpha$; $\alpha = 1, 2, \cdots, k$),则, 第 α 类的均值向量为:

$$\bar{x}^\alpha = \sum_{i=1}^{n_\alpha} x_i^\alpha / n_\alpha \qquad (1)$$

总体均值向量为:

$$\bar{x} = \sum_{\alpha=1}^{k} \sum_{i=1}^{n_\alpha} x_i^\alpha / n_\alpha \qquad (2)$$

组内协方差矩阵为:

$$E = \sum_{\alpha=1}^{k} \sum_{i=1}^{n_\alpha} (x_i^\alpha - \bar{x}^\alpha)^{\mathrm{T}} (x_i^\alpha - \bar{x}^\alpha) \qquad (3)$$

式中,T 记为矩阵的转置. 当总体服从正态分布时,对任意样品 x,根据贝叶斯判别

第一作者简介:男,30岁,博士,副教授,现在湖南师范大学化学系工作
* 国家自然科学基金资助课题
** 湖南师范大学化学系

分析方法可建立如下的判别式：

$$y^{\alpha} = \ln q^{\alpha} - 0.5(n-k)\bar{x}^{\alpha}E^{-1}(\bar{x}^{\alpha})^{T} + (n-k)xE^{-1}(\bar{x}^{\alpha})^{T} \quad (4)$$

式中，q^{α} 是 α 类的先验概率：$q^{\alpha} = n_{\alpha}/n$

如 $y^{j} > y^{\alpha}(\alpha = 1, 2, \cdots, k)$，则将样品 x 分配到第 j 类中去.

$$n = n_1 + n_2 + \cdots + n_k \quad (5)$$

令

$$P_i = s/n \quad (6)$$

式中，i 是变量数；s 是用 DCRIM 法判别所得的类与实际类相符的样本数.

为了弥补逐步判别分析法有时不能找到最佳变量组合的不足，本文运用判别符合率递增法来选择变量，然后根据式(4)对化合物毒性进行预测. 具体步骤如下：

第一步：选取研究中所考虑的所有 m 个变量，用式(4)计算每一类的判别函数.

第二步：由式(5)计算所有 n 个样品的判别符合率 P_m.

第三步：从 m 个变量中每次取 $m-1$ 个变量，根据组合原理，有 $C_m^{m-1} = m$ 个组合. 由式(4)和式(5)对每一类分别组建判别函数，并计算相应的判别符合率 P_{m-1}，假设在所有的 P_{m-1} 中数值最大的为 $P_{m-1}(\max)$，若 $P_{m-1}(\max) < P_m$，则上述的 m 个变量的组合是最佳的停止计算；否则继续进行第四步.

第四步：从对应于 $P_{m-1}(\max)$ 的 $m-1$ 个变量中每次取 $m-2$ 个变量，此时有 $m-1$ 个组合，余下的工作与第三步相似，……，直至进行到第 $k+1$ 步的最大判别符合率 $P_{m-k}(\max)$ 较上一步的最大判别符合率 $P_{m-k+1}(\max)$ 下降为止，则对应于 $P_{m-k+1}(\max)$ 的变量组合为所求.

第五步：通过式(4)运用最佳变量组合建立判别函数，然后对未知毒性的化合物进行分类.

2.2 逐步判别分析法

其方法原理参见文献[6].

3 结果与讨论

3.1 酚类化合物的特征参数($^{t}X^{v}$ 和 $\log P$)

按文献[7]所建议的方法计算代表结构特征的参数：分子连接性指数 $^{t}X^{v}$ 和分配系数 $\log P$，结果见表1. 表1中酚类化合物的毒性类别，系参照文献[8]提供的酚类化合物的半致死量(LD_{50})毒性数据确定，若酚类化合物的 $LD_{50} < 500mg/kg$，则其毒性类别属1类；否则，属2类.

3.2 变量选择和结果比较

用于分类的重要变量及其组合分别用逐步判别分析法和判别符合率递增法进行选择，结果见表2. 从表2可以看出，用上述两种方法所得的变量组合结果是不同的. 用判别符合率递增法所得的判别符合率是90.32%，而用逐步判别法所得的判别符合率只有83.87%，因此判别符合率递增法优于逐步判别分析法，由式(4)可以建立每一类的判别函数. 当然，如果运用穷举法，我们也能获得最佳变量组合，但此法是极其耗时的，对于8个变量的情况，根据穷举法的原理，需计算 $2^8 - 1$(255)次才能获得最佳结果，当使用判别符合率递增法时仅需计算43次就能达到目的.

表 1　酚类化合物的结构参数

Table 1　The structure parameters of phenolic compounds

化　合　物	结　构　参　数								实际毒性类别
	$^0x^v$	$^1x^v$	$^2x^v$	$^3x^v$	$^4x^v$	$^5x^v$	$^6x^v$	log$P$	
苯酚	3.834	2.134	1.336	0.756	0.428	0.242	0.029	1.46	2
邻甲酚	4.757	2.551	1.786	1.115	0.563	0.317	0.080	1.95	1
间甲酚	4.757	2.545	1.839	1.002	0.628	0.257	0.074	1.96	1
2,4-二特丁基-间-甲酚	11.602	7.899	6.684	2.959	2.101	1.069	0.978	5.92	2
2,6-二特丁基-对-甲酚	11.602	5.878	3.709	2.752	1.881	1.407	0.653	5.90	2
2,4-二甲基酚	5.679	2.962	2.290	1.352	0.805	0.427	0.159	2.43	2
3,4-二甲基酚	5.679	2.962	2.269	1.490	0.726	0.431	0.136	2.44	2
3,5-二甲基酚	5.679	2.956	2.346	1.206	0.869	0.390	0.178	2.46	2
香芹酚	7.257	3.905	2.623	1.888	1.038	0.709	0.279	3.48	2
邻-苯基苯酚	7.143	4.212	2.717	1.893	1.287	0.809	0.332	3.20	2
b-萘酚	5.989	3.539	2.387	1.605	1.050	0.773	0.309	2.50	2
己基一间-苯二酚	8.662	5.246	4.063	3.042	1.506	0.880	0.393	2.54	2
苯醌	4.126	2.230	1.471	0.824	0.440	0.273	0.045	0.68	1
连苯三酚	4.574	2.451	1.645	0.993	0.524	0.268	0.074	0.22	2
邻-氯苯酚	4.960	2.653	1.896	1.336	0.600	0.338	0.092	2.15	2
a-(2,4-二氯苯氧基)丙酸	8.981	4.675	3.503	2.125	1.315	0.773	0.379	3.38	2
r-(2,4-二氯苯氧基)丁酸	9.525	7.575	3.751	2.164	1.473	0.811	0.422	3.44	2
间-氯苯酚	4.960	2.647	1.957	1.065	0.635	0.326	0.118	2.50	2
2,4-二氯苯酚	6.087	3.165	2.517	1.497	0.936	0.477	0.198	2.08	2
2,4,5-三氯苯酚	7.213	3.684	2.751	2.127	1.112	0.727	0.241	3.12	2
2,3,4,6-四氯苯酚	8.340	4.209	3.532	2.802	1.571	0.733	0.359	4.81	1
五氯酚	6.897	4.733	3.983	3.710	1.855	0.856	0.417	5.85	1
对-氯一间甲苯酚	5.883	3.064	2.514	1.605	0.799	0.463	0.146	2.89	2
2,4-二氯-3,5-二甲酚	7.932	4.005	3.291	2.792	1.271	0.686	0.276	4.08	2
2,4,6-三溴酚	9.554	4.855	4.376	2.670	2.826	0.742	0.426	4.57	1
对甲酚	4.757	2.545	1.836	1.034	0.785	0.280	0.080	1.94	1
间-苯二酚	4.204	2.269	1.520	0.830	0.493	0.253	0.059	0.80	1
双-对氯苯氧基甲烷	10.55	5.832	4.268	2.437	1.502	1.050	0.468	5.56	2
2,2-二羟基-5,5′-二氯苯基甲烷	10.47	5.834	4.677	2.978	2.203	1.337	0.734	4.98	2
2-甲基-4-氯苯氧基丁酸	9.320	5.114	3.732	2.186	1.432	0.782	0.378	3.10	2
抑草蓬	10.11	5.194	4.012	2.752	1.408	1.035	0.469	4.48	2
对-氯苯氧基乙酸*	6.984	3.697	2.522	1.451	0.814	0.547	0.240	2.19	2
2,4,6-三氯苯酚*	7.213	3.690	2.995	2.101	1.095	0.661	0.252	3.77	2
五溴酚*	13.368	6.684	5.934	7.348	3.549	1.837	0.872	6.91	2
对特丁基酚*	7.257	3.795	3.300	1.733	0.913	0.521	0.309	3.44	2
2,3-二甲基酚*	5.679	2.968	2.219	1.576	0.736	0.389	0.139	2.450	2

* 为预测集样本

3.3　酚类化合物毒性的分类预测

为了检验方法的有效性，本文运用表 2 建立的函数对一些酚类化合物的毒性进行分类预测，由表 3 可以看出，预测结果与实际毒性类别一致，表明这些方法是切实可行的.

4. 结 论

综上所述，判别符合率递增法在构效关系研究中是很有用的新方法，它们弥补了逐步判别分析法和穷举法的不足.

表 2 不同方法的结果比较

Table 2 Comparison of results of different methods

名 称	逐步判别分析法	判别符合率递增法
判别符合率，%	83.87	90.32
错分化合物	（邻甲酚 间甲酚 对甲酚 间-苯二酚 苯醌）	（邻甲酚 间-苯二酚 对甲酚）
所选变量组合	（$^1x^v$, $^4x^v$, $^5x^v$, $\log P$）	（$^1x^v$, $^2x^v$, $^4x^v$, $^5x^v$, $^6x^v$, $\log P$）
判别方程	$y^1 = -7.6077 + 3.9843^6x^v - 2.0058^4x^v - 19.2672^5x^v + 0.2423\log P$	$y^1 = -14.3762 + 2.0632^1x^v + 11.8383^2x^v - 1.8016^4x^v + 21.0189^5x^v - 93.1137^6x^v - 0.1845\log P$
	$y^2 = -9.8162 + 5.0410^6x^v - 7.2251^4x^v - 12.75585x^v - 0.4929\log P$	$y^2 = -16.007 + 2.2998^1x^v + 13.8778^2x^v - 7.5396^4x^v + 32.6668^5x^v - 98.2118^6x^v - 1.1053\log P$

表 3 酚类化合物的毒性预测

Table 3 Prediotion of toxicity of phenolic compounds

化 合 物	逐步判别分析法	方 法	
		判别符合率递增法	实际毒性类别
对氯苯氧基乙酸	2	2	2
2,4,6-三氯苯酚	2	2	2
五溴酚	2	2	2
对特丁基酚	2	2	2
2,3-二甲基酚	2	2	2

参 考 文 献

1 Wold S, Sjostrom M. In: Kowalski B, (ed.) Chemometrics: Theory and application. Washington, D C: American Chemical Society, 1977, 243: 243

2 Richard G Lanson, Peter C Jurs. J Chem Inf Comput Sci, 1990, 30: 137

3 Lachenbruch P A. Discrimination analysis. New York: Hafner, 1975

4 Rowberg K L, Hopfinger A J. Chemomctrics and Intelligent Laboratory System, 1990, 8: 183

5 谭良. 数学的实践与认识，1990，2：47

6 中国科学院计算中心概率统计组. 概率统计计算. 北京：科学出版社，1979：197—206

7 王尔华. 定量药物设计. 北京：人民卫生出版社，1983：291—314

8 董华模. 化学物的毒性及其环境保护参数手册. 北京：人民卫生出版社，1988：347—376

1993-05-11收到

DISCRIMINATION COINCIDENCE RATE INCREASING METHOD AIDED STUDIES OF STRUCTURE-ACTIVITY RELATIONSHIP OF PHENOLIC COMPOUNDS

Sun Lixian　Xu Fen　Shong Xinhua　Yu Ruqin

(Department of Chemistry and Chemical Engineering, Hunan University, Changsha 410082)

ABSTRACT

Molecular connectivity index ($^{1}x^{v}$) and Log$P$ (the logarithm of partition coefficient between n-octanol and water) of phenolic compounds are calculated. A new mathematical method-discrimination coincidence rate increasing method (DCRIM) has been used to investigate the structure—activity relationship (SAR) of phenolic compounds. The results obtained with the method are compared favourably with those by using conventional stepwise discrimination method, in addition to this, DCRIM is also simpler and faster than the method of exhaustion.

Keywords: chemometrics; stepwise discrimination method; discrimnation coincidence rate increasing method; method of exhaustion.