文章编号：1006－544X（2005）02－0211－06

MEDV描述子对多氯代二苯并二恶英的QSRR支持向量机建模

易忠胜[1]，刘树深[1,2]，

（1. 桂林工学院 材料与化学工程系，广西 桂林 541004；2. 南京大学 环境学院，南京 210093）

摘 要：采用分子电性距离矢量（MEDV）描述子表征多氯代二苯并二恶英（PCDDs），结合支持向量机（SVM）算法，对不同固定相下PCDDs的气相色谱保留行为值建立定量模型，模型的相关系数（R）均大于0.99，留一法交互检验的相关系数（q）也都大于0.99. 从有实验数据的异构体中均匀挑选2/3作为训练集，余下的1/3作为检验集进行了建模，所得模型的相关系数也都大于0.99，并对没有实验值的异构体进行了预测.

关键词：多氯代二苯并二恶英（PCDDs）；定量结构－色谱保留关系（QSRR）；气相色谱保留行为；分子电性距离矢量（MEDV）；支持向量机（SVM）

中图分类号：O6－04；X132 **文献标识码：**A①

二恶英类化合物因为其高毒性、持久污染性，近年来备受人们关注，它实际上是指一类氯代含氧三环芳烃类化合物，主要是PCDDs、多氯代二苯呋喃类（PCDFs）、多氯联二苯（PCBs）以及溴代物和其它混合卤代物等物质的总称. PCDDs没有进行商业生产，也没有直接的用途，它的主要来源有：生产1,2,4,5-四氯苯酚的副产物；存在于一些除草剂、杀菌剂、杀虫剂内；露天燃烧城市工业废物[1]. PCDDs和其它二恶英类化合物的共同性质是有较高的热稳定性，一般在加热到800℃才能分解. 在自然环境中的自然降解很慢，半衰期约为9年. 这种持久性有机污染物近年来引起了各国政府和科学界的极大关注[2－4]，初步研究结果表明，二恶英不仅对人类有致癌性，还能降低人体免疫能力，具有内分泌干扰作用. 因此了解和研究PCDDs的各种理化性质非常必要，而PCDDs共有75个异构体，其中有些异构体到目前还没有分离出来，因此它们的理化性质也就没有办法通过实验方法获得，但可以通过QSAR/QSPR（Quantitative Structure－Activity/Property Relationship）方法进行预测. 气相色谱Kováts保留指数（RI）和相对保留时间（RRT）通常用来区分不同的异构体. 化合物在色谱柱上的保留行为与化合物和固定相之间的相互作用有关，当固定相一定时，这种相互作用的程度大小直接与化合物的拓扑、几何和电性特征等相关，不同的化合物在相同的色谱柱上通常表现出不同的特征保留行为，通过这些特点可以达到区分和分离不同化合物的目的.

SVM是Vapnik等人在统计学习理论[5－7]基础上提出的一种确定两类问题最优分类超平面的有效算法，并能推广至多类和回归建模问题[8]. 本文利用分子电性距离矢量（MEDV）[9]表征PCDDs异构体的分子结构，对PCDDs在不同固定相（DB－5、SE－54、OV－1701、SP2100）上的气相色谱保留行为值用SVM算法建立定量色谱保留关系模型进行了研究，由于目前SVM算法中选择最优参数没有统一的标准，笔者选择在QSAR/QSPR研

① **收稿日期：**2004－09－13

基金项目：广西自然科学基金资助项目（桂科自0236063）. 广西高校百名青年科学带头人资助计划项目（桂教人[2003]97）

作者简介：易忠胜（1970－），男，硕士，高级实验师，分析化学专业.

究中使用的交互检验（Cross Validation）的方法，通过大量的对比实验，以交互检验相关系数 q^2 最大并且兼顾支持向量样本数和边界支持向量（Bounded Support Vector）样本数为标准来选择最优的 SVM 建模参数，建立的模型预测效果满意，其预测或估计结果优于多元线性回归.

1 支持向量机回归原理

多元线性回归算法利用有限的已知样本，通过求解 $\boldsymbol{Y}=\boldsymbol{XB}$ 中的回归系数 $\boldsymbol{B}$ 而建立模型，从该算法的求解原理中很清楚地看到它将有限样本数据中的误差拟合进数学模型中[10]，它并没有对数据误差作任何处理. 针对这一缺点，SVR 通过引入损失函数（通常采用“ ε 不敏感函数”），用函数 $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$ 拟合目标值 y，在一定约束条件下，以 $\|\boldsymbol{w}\|$ 取极小的标准来选取数学模型的唯一解，并求解出最佳回归方程. 这一求解策略使过拟合受到限制，显著提高了数学模型的预测能力. 支持向量机的最终求解问题归结为一个有约束的二次型规划（QP，Quadratic Programming）问题（具体的算法过程参考相应的文献［5－7］），最终得到回归方程

$$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)K(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x})+b^*.$$

其中 α_i,α_i^* 为拉格朗日乘子，$\alpha_i\times\alpha_i^*=0$，$\alpha,\alpha^*$ 只有小部分不为 0，它们对应的样本就构成支持向量，一般是在函数变化比较剧烈的位置上的样本. 支持向量回归方法根据实际操作时采用参数的不同可以分为两种方法：ε－SVR 和 ν－SVR[11].

2 结果与讨论

2.1 PCDDs 气相色谱保留行为和分子描述子

75 个 PCDDs 异构体在 4 种不同固定相（DB－5、SE－54、OV－1701、SP－2100）上的气相色谱保留行为（相对保留时间，*RRT*；Kováts 保留指数，*RI*）的数据取自文献［4］（表 1），其中有实验值的样本分别为：固定相为 DB－5 时有 41 个，固定相为 SP－2100 时有 39 个；固定相为 SE－54 时只有 15 个；固定相为 OV－1701 时也只有 15 个. 将有实验值的样本分作训练集和检验集，选取检验集时，先按保留行为值排序，然后均匀抽取大约 1/3，表 1 中固定相的保留行为数据列中带＊号的样本即为检验集，检验集样本数分别为 12，11，5，5；没有实验值的样本作为预测集，各预测集的样本数分别为 26，28，60，60.

根据文献［9］中 MEDV 对原子属性的定义，PCDDs 分子中只有第 2（～C～）、3（≈C～）、10（—O—）和 13（Cl—）等 4 种类型，这 4 种类型两两相互作用构成的描述子序号为：14，15，22，25，26，33，36，82，85，91，它们分别对应于 13 种原子类型中 2－2，2－3，2－10，2－13，3－3，3－10，3－13，10－10，10－13，13－13 的两两相互作用对，也就是 PCDDs 的 MEDV 描述子共有 10 个.

2.2 计算方法

根据支持向量机回归的原理，SVM 用于回归的方法有 2 种，即 *ε-SVR* 和 *ν-SVR*，而常用核函数有 4 类，即：(1)线性函数，$K(\boldsymbol{x},\boldsymbol{x}_i)=(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}_i)$；(2)多项式核函数，$K(\boldsymbol{x},\boldsymbol{x}_i)=((\gamma\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}_i)+c)^q$（当 $q=1$ 及 $c=0$ 时即为线性核函数）；(3)径向基(RBF)函数，$K(\boldsymbol{x},\boldsymbol{x}_i)=e^{-\gamma\|\boldsymbol{w},\boldsymbol{x}_i\|}$；(4)Sigmoid 函数，$K(\boldsymbol{x},\boldsymbol{x}_i)=\tanh(\gamma(\boldsymbol{x}^{\mathrm{T}},\boldsymbol{x}_i)+c)$. 计算过程中发现采用多项式、Sigmoid 函数作为核函数时，计算速度很慢（以后不再对这 2 个函数进行讨论）. 其中线性函数本身没有参数，RBF 函数有一个参数 γ，因此进行 SVM 参数优化时，需要考虑 2 种 SVR 算法以及 SVM 本身的参数 C、ε、ν 和核函数参数的影响. 针对目前还没有统一的 SVM 最优参数选择方法，并且大多数文献对这些参数选择没有进行讨论的情况，对这些参数采用网格的方法进行大范围的搜索. 首先确定参数的范围，根据 Lee[12] 提供的实验结果，一般可以采用 2 的指数增长方式增加参数，并且设置的点一般有 10 个就可以满足要求了，本文的参数设置如表 2 所示. 然后按照网格的方式把各参数组合，利用这些参数组合进行 SVM 计算. 最后以各参数组合的 LOO(Leave-One-Out)交互检验为参数选择的标准，也就是选择 LOO 交互检验相关系数 q^2 最大，并且还要兼顾支持向量样本数和边界样本数的数量不能太多的参数组合作为最终的建模参数. 所有的计算采用 LIBSVM 软件完成[11].

2.3 气相色谱保留指数模型

按照表 2 进行参数设置，然后对 2 种 SVR 方法分别采用线性、RBF 核函数，以固定相为 DB－5 的气相色谱保留指数为因变量，10 个 MEDV 参

表 1　75 个 PCDDs 在 4 种不同固定相的气相色谱保留行为值

Table 1　Retention values of 75 PCDDs on gas chromatographic with 4 stationary phases

No	Isomers	DB－5		SP－2100B		SE－54		OV－1701	
		RI^{Obs}	RI^{Cal}	RRT^{Obs}	RRT^{Cal}	RRT^{Obs}	RRT^{Cal}	RRT^{Obs}	RRT^{Cal}
1	1-Cl		819	0.293	0.294		1.290		1.095
2	2-Cl		907	0.299 *	0.298		1.286		1.085
3	1,2-diCl		1170		0.123		1.251		1.115
4	1,3-diCl		1262		0.146		1.223		1.090
5	1,4-diCl		1172		0.118		1.240		1.097
6	1,6-diCl		1178		0.119		1.237		1.094
7	1,7-diCl		1274		0.146		1.224		1.083
8	1,8-diCl		1271		0.146		1.224		1.083
9	1,9-diCl		1173		0.119		1.239		1.095
10	2,3-diCl	1993	1993	0.433	0.434		1.226		1.102
11	2,7-diCl	1985 *	1985	0.424	0.425		1.227		1.075
12	2,8-diCl	1985	1985		0.179		1.227		1.076
13	1,2,3-trCl		1495		0.243		1.251		1.169
14	1,2,4-trCl	2152	2154	0.600	0.598		1.226		1.145
15	1,2,6-trCl		1527		0.235		1.208		1.127
16	1,2,7-trCl		1626		0.285		1.182		1.112
17	1,2,8-trCl		1622		0.284		1.184		1.114
18	1,2,9-trCl		1517		0.234		1.212		1.130
19	1,3,6-trCl		1620		0.274		1.171		1.101
20	1,3,7-trCl		1719		0.330		1.163		1.091
21	1,3,8-trCl		1718		0.330		1.163		1.092
22	1,3,9-trCl		1617		0.273		1.172		1.102
23	1,4,6-trCl		1520		0.223		1.200		1.111
24	1,4,7-trCl		1626		0.270		1.172		1.096
25	1,7,8-trCl		1666		0.312		1.171		1.113
26	2,3,7-trCl		1761	0.651	0.650		1.174		1.105
27	1,2,3,4-teCl	2379 *	2379	0.980	0.984		1.331		1.304
28	1,2,3,6-teCl	2378	2386	0.975 *	0.992		1.242		1.195
29	1,2,3,7-teCl	2382	2386	0.985	0.987		1.213		1.174
30	1,2,3,8-teCl	2382	2386	0.985	0.988		1.215		1.175
31	1,2,3,9-teCl	2392 *	2386	1.010	0.992		1.245		1.198
32	1,2,4,6-teCl	2346	2343	0.910	0.913		1.212		1.173
33	1,2,4,7-teCl	2340	2343	0.897	0.898		1.185		1.151
34	1,2,4,8-teCl	2340	2343	0.897	0.899		1.186		1.152
35	1,2,4,9-teCl	2346 *	2343	0.910	0.914		1.214		1.175
36	1,2,6,7-teCl	2408	2416	1.040 *	1.043		1.218		1.174
37	1,2,6,8-teCl	2349	2358	0.918 *	0.928		1.175		1.145
38	1,2,6,9-teCl	2378	2378	0.972	0.968		1.201		1.159
39	1,2,7,8-teCl	2400	2396	1.030	1.022		1.180		1.157
40	1,2,7,9-teCl	2364 *	2358	0.951	0.928		1.177		1.146
41	1,2,8,9-teCl	2428 *	2416	1.090	1.045		1.226		1.180
42	1,3,6,8-teCl	2290	2290	0.813	0.816	1.075	1.075	1.052	1.056
43	1,3,6,9-teCl	2315	2320	0.852 *	0.852		1.164		1.129
44	1,3,7,8-teCl	2340 *	2338	0.905 *	0.911		1.152		1.135
45	1,3,7,9-teCl	2304 *	2300	0.833	0.829	1.082 *	1.082	1.063 *	1.066
46	1,4,6,9-teCl	2341	2339	0.896	0.900		1.189		1.143
47	1,4,7,8-teCl	2353	2359	0.928	0.932		1.161		1.139
48	2,3,7,8-teCl	2386	2375	1.000 *	1.003	1.125	1.124	1.106	1.110
49	1,2,3,4,6-peCl		1952		0.847		1.357		1.360
50	1,2,3,4,7-peCl	2573	2573	1.540	1.537		1.328		1.311
51	1,2,3,6,7-peCl	2604	2606		1.146		1.292		1.259
52	1,2,3,6,8-peCl		2285		1.290	1.215	1.224	1.189	1.193
53	1,2,3,6,9-peCl		2174		1.064		1.253		1.244
54	1,2,3,7,8-peCl	2587 *	2587	1.630 *	1.624	1.253	1.254	1.229	1.212
55	1,2,3,7,9-peCl		2280		1.284	1.225 *	1.224	1.203 *	1.194
56	1,2,3,8,9-peCl	2623	2607		1.137		1.296		1.265
57	1,2,4,6,7-peCl		2174		1.051		1.245		1.238
58	1,2,4,6,8-peCl	2501	2506		1.176	1.192 *	1.192	1.170 *	1.174
59	1,2,4,6,9-peCl		2160		0.979		1.213		1.223
60	1,2,4,7,8-peCl		2331	1.460	1.452	1.220	1.219	1.196	1.192
61	1,2,4,7,9-peCl	2501	2506		1.175	1.192	1.192	1.170	1.174
62	1,2,4,8,9-peCl		2166		1.047		1.248		1.241
63	1,2,3,4,6,7-heCl	2812	2796		1.940		1.489		1.455
64	1,2,3,4,6,8-heCl	2742	2739		2.149		1.390		1.389
65	1,2,3,4,6,9-heCl		2266		1.851		1.453		1.443
66	1,2,3,4,7,8-heCl	2781 *	2778	2.540 *	2.550	1.411	1.411	1.370	1.375
67	1,2,3,6,7,8-heCl	2788	2801	2.650	2.674	1.409	1.397	1.363	1.359
68	1,2,3,6,7,9-heCl		2481	2.420	2.427	1.337	1.336	1.338	1.340
69	1,2,3,6,8,9-heCl		2476		2.392	1.337 *	1.338	1.338 *	1.341
70	1,2,3,7,8,9-heCl		2488	2.760	2.675	1.432 *	1.399	1.395 *	1.362
71	1,2,4,6,7,9-heCl	2713 *	2713	2.220	2.211		1.281		1.323
72	1,2,4,6,8,9-heCl	2713	2713		2.190		1.282		1.325
73	1,2,3,4,6,7,8-hpCl	2994 *	2994	4.180 *	4.164	1.659	1.658	1.588	1.584
74	1,2,3,4,6,7,9-hpCl	2949	2949	3.780	3.956		1.630		1.572
75	1,2,3,4,6,7,8,9-ocCl	3196	3188	6.760	6.741		1.506		1.800

数作为自变量,进行 SVR 的 LOO 计算. 计算结果表明,ε-SVR 采用线性和 RBF 核函数时,最大的 q^2 分别为 0.997 5 和 0.988 9;而相应的 ν-SVR 的 q^2 分别为 0.992 8 和 0.988 1. 按照笔者定义的 SVM 建模参数选择的标准,线性核函数比 RBF 核函数的回归效果更好,ε-SVR 比 ν-SVR 的回归效果稍好. 通过计算,选择 ε-SVR 方法,线性核函数,q^2 最大为 0.997 5,对应的 SVM 参数为 $C=512$,$\varepsilon=0.256$ 来建模. 最终得到如下的模型,也就是 SVM 回归方程

$$RI = \sum_{i=1}^{35} (\alpha_i - \alpha_i^*)\boldsymbol{x}_i\boldsymbol{x} - 436.431,$$

其中模型的支持向量样本数为 35.

得到模型后用计算得到的拉格朗日系数、支持向量样本的描述子和待预测样本的描述子代入上述回归方程,就可得到样本的保留指数预测值. 模型对 41 个样本和预测集的计算结果见表 1 的 RI^{Cal} 列. 实验值与计算值之间的相关系数为 0.999 7,标准偏差为 6.098 5. 对预测集的预测结果见表 1 的 RI^{Cal} 列,计算值与实验值的相关关系见图 1.

同样地,对其它 3 种固定相的色谱保留时间进行了支持向量机回归建模,连同固定相为 DB-5 的色谱保留指数的支持向量机模型汇总在表 3 中,将各固定相有实验值的样本均匀抽出约 1/3 作为检验集,余下 2/3 的作为训练集,用上面得到的参数重新建模,分别对训练集、检验集和预测集进行了计算,其中检验集的实验值和计算值的相关系数分别为 0.999 5,0.995 5,0.998 0,0.999 9. 固定相为 DB-5 时的训练集和检验集的实验值和计算值之间的相关关系图 2,其它固定相的保留时间计算值

表 2 SVM 的参数设置和最大 LOO 交互检验识别率以及对应的参数

Table 2 Parameters setting of SVM and the largest q^2 parameters

SVC 算法	核函数	参数设置	最大 q^2			
			DB-5	OV-1701	SE-5	SP-2100
ε-SVR	线性函数	$C = 1 \times 2^N, N = 0,1,\cdots 9$	0.9975	0.8915	0.8907	0.9367
	RBF 函数	$\gamma = 0.00025 \times 2^N, N = 0,1,\cdots 13$	0.9889	0.9169	0.9101	0.9866
ν-SVR	线性函数	$\upsilon = 0.001 \times 2^N, N = 0,1,\cdots 9$	0.9928	0.9774	0.9788	0.9647
	RBF 函数	$\varepsilon = 0.001 \times 2^N, N = 0,1,\cdots 9$	0.9881	0.9906	0.9942	0.9936

表 3 4 种固定相的 PCDDs 气相色谱保留行为的 QSPR 支持向量机模型

Table 3 Models of SVM for retention values of 75 PCDDs on gas chromatographic with 4 stationary phases

固定相	色谱保留行为	模 型	相关系数 r	q^2
DB-5	*RI*	$f(\boldsymbol{x}) = \sum_{i=1}^{35} (\alpha_i - \alpha_i^*)\boldsymbol{x}_i\boldsymbol{x} - 436.431$	0.999 7	0.997 5
OV-1701	*RRT*	$f(\boldsymbol{x}) = \sum_{i=1}^{12} ((\alpha_i - \alpha_j^*)\exp(-\gamma^* \parallel \boldsymbol{x}_i,\boldsymbol{x} \parallel^2)) - 1.384\ 22$	0.999 5	0.990 6
SE-54	*RRT*	$f(\boldsymbol{x}) = \sum_{i=1}^{6} ((\alpha_i - \alpha_j^*)\exp(-\gamma^* \parallel \boldsymbol{x}_i,\boldsymbol{x} \parallel^2)) - 0.924\ 15$	0.998 4	0.994 2
SP-2100	*RRT*	$f(\boldsymbol{x}) = \sum_{i=1}^{32} ((\alpha_j - \alpha_j^*)\exp(-\gamma^* \parallel \boldsymbol{x}_i,\boldsymbol{x} \parallel^2)) - 6.526\ 3$	0.999 8	0.993 6

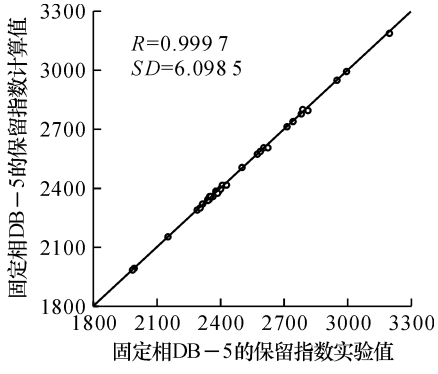

图 1 41 个训练样本的气相色谱保留指数的实验值与 SVR 预测值的相关图

Fig. 1 Object value versus calculated value for 41 train set for retention values of 75 PCDDs on gas chromatographic with stationary phases DB-5

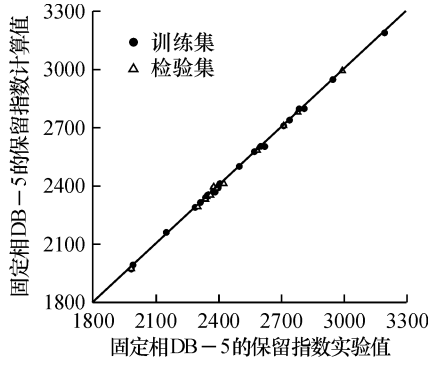

图 2 29 个训练样本保留指数和检验集的实验值与 SVR 预测值相关图

Fig. 2 Object value versus calculated value for 29 train set for retention values of 75 PCDDs on gas chromatographic with stationary phases DB-5

与实验值的相关图限于篇幅不再列出.

2.4　影响 SVM 建模的因素

支持向量机建模的影响因素除了算法本身（参数的选择、核函数选择等）以外，还存在一些算法以外的影响因素. 首先是支持向量样本的多少直接与模型性能有关. 为此，对不同参数的 SVM 模型的支持向量样本数进行了统计. 从表 3 的模型可以看到，所建立的模型中，支持向量样本的数量很大，与训练样本数很接近；固定相为 DB－5 时，按照表 2 用全部的参数组合建立模型，发现这些模型中的支持向量样本最小的是 37，只出现了 1 次，大部分是 40；而边界支持向量样本数也是很大，最小的是 28. 笔者曾经查阅了很多相关的应用文献，但大多没有提及或者讨论支持向量样本数量的问题. Liu[13] 提到了支持向量样本数，其训练样本数为 45，而支持向量样本数同样也很高，达到了 43 个，但是也没有作进一步的研究. 根据支持向量机的原理，一般来说，支持向量样本的数量越多，越容易产生过拟合现象. 但是笔者用训练集建模，再对外部检验集（没有参与建模）进行预测时的结果看，并没有出现过拟合的现象，检验集的实验值与预测值之间的相关系数达到了 0.999 5，与用全部已知样本建模的预测值也非常接近. 支持向量样本数量与训练集样本数量、特征变量的数量之间达到什么样的关系，才不至于产生过拟合现象这个问题还是非常值得进一步研究的.

其次是描述子的选择. 描述子的选择也直接影响 SVM 建模的效果，如果所选择的描述子不能表达所描述分子的信息，建立的模型肯定是失败的，这不是算法所能解决的问题. 根据 MEDV 对原子属性的定义，PCDDs 分子中只有第 2（～C～），3（≈C～），10（—O—），13（Cl—）等 4 种类型，这样 PCDDs 的 MEDV 描述子共有 10 个，这些描述子描述了基于 MEDV 的全部信息，笔者采用基于预测的变量选择和建模的最佳子集回归算法（VSMP）进行建模时发现，MEDV 描述子与本文所研究的 4 种固定相的气相色谱保留行为值具有良好的线性相关性，并且 4 种固定相的气相色谱保留行为值与取代氯和苯环上的碳以及它们之间的相互作用构成的描述子密切相关，特别是共轭体系中的碳原子与取代氯之间的相互作用构成的描述子的影响非常明显，而其它的描述子的作用相对没有那么明显（另文发表）. 笔者采用了全部 10 个描述子进行建模，包含了基于 MEDV 描述 PCDDs 的全部信息，非常有利于 SVM 模型的建立，从表 3 中可以看到所建立的模型非常理想.

再次就是描述子的数量对模型的建立也有很大的关系，描述子越多，表达的分子信息自然越多，而这些描述子各自所包含的信息量是不一样的，就目前的研究现状，可以提取上百的描述子，但是建模过程中，不能一味追求描述子数量，而把一些几乎与活性/性质无关的描述子用来建模，这样反而不利. 因此选择合适数量的描述子也是非常重要的，本文并没有详细讨论描述子数量与建立最佳模型的关系，但初步的计算结果表明，当描述子增加的情况下，所建立模型普遍较好. 今后要通过不同数量的描述子组合选取最佳的 SVM 模型.

3　结　论

采用 SVM 对 PCDDs 4 种固定相的气相色谱保留行为与 MEDV 分子描述子进行了回归研究，以 q^2 最大且兼顾支持向量样本数和边界支持向量样本数作为选择最优 SVM 建模参数，得到了 MEDV 描述子和气相色谱保留行为之间具有很高预测能力的 QSPR 模型，即 q^2 都大于 0.99. 分别用全部由已知实验值的样本建立的模型和 2/3 已知实验值的样本建立的模型对所有的样本进行了预测，预测值与实验值的相关系数都大于 0.99.

参考文献

[1] 王连生. 环境化学进展[M]. 北京：化学工业出版社，1995.

[2] 岳瑞生.《关于就某些持久性有机污染物采取国际行动的斯德哥尔摩公约》及其谈判背景[J]. 世界环境，2001(1)：24－28.

[3] Rayne S. Development of a multiple－class high－resolution gas chromatographic relative retention time model for halogenated environmental contaminants[J]. Anal. Chem. 2003, 75：1049－1057.

[4] D Needham M, Adams K C, C Jurs P. Quantitative structure-retention relationship studies of polychlorinated dibenzodioxins on gas chromatographic stationary phases of varying polarity[J]. Anal. Chim. Acta. 1992, 258：183－198.

[5] Vapnik V. Estimation of dependencies based on empirical data [M]. New York：Springer，1982.

[6] Vapnik V. The nature of statistical learning theory[M]. New York: Springer – Verlag,1995.

[7] Vapnik V. Statistical Learning Theory[M]. New York: Wiley publishers science,1998.

[8] 易忠胜,刘红艳,刘树深. 基于支持向景机的聚氯乙烯耐有机溶剂性能分类[J]. 桂林工学院学报,2004,(4):474 – 479.

[9] 刘树深. 药物分子电性距离矢量及其应用[D]. 重庆:重庆大学, 2001.

[10] 刘树深,易忠胜. 基础化学计量学[M]. 北京:科学出版社, 1999.

[11] Chang C C, Lin C J. Training nu-support vector classifiers: Theory and algorithms[J]. Neural Computation, 2001, 13 (9): 2119 – 2147. (Implementation available in http://www. csie. ntu. edu. tw/ ~ cjlin/libsvm)

[12] Lee J H, Lin C J. Automatic model selection for support vector machines[DB/OL]. http://www. csie. ntu. edu. tw/ ~ cjlin/papers/guide/guide. pdf,2000.

[13] Liu H X, Zhang R S, Yao X J, *et al*. QSAR study of ethyl 2-[(3-Methyl-2,5-dioxo(3-pyrrolinyl)) amino]-4- (trifluoromethyl) pyrimidine-5-carboxylate: An inhibitor of AP-1 and NF – KB mediated gene expression based on support vector machines[J]. J. Chem. Inf. Comput. Sci. ,2003, 43:1288 – 1296.

[14] Liu S S, Liu H L, Yin C S, *et al*. VSMP: A novel variable selection and modeling method based on the prediction[J]. J Chem. Inf. Comput. Sci. ,2003, 43:964 – 969.

MEDV for Quantitative Structure – Retention Relationship of Polychlorinated Dibenzodioxins on Gas Chromatographic Stationary Phases Based on Support Vector Machine

YI Zhong-sheng[1], LIU Shu-shen[1,2]

(1. *Department of Materials and Chemistry Engineering*, *Guilin University of Technology*, *Guilin* 541004, *China*; 2. *Institute of Environment*, *Nanjing University*, *Nanjing* 210093, *China*)

Abstract: The molecular electronegativity distance vector based on 13 atomic types (MEDV) is employed to describe PCDDs isomers structure, and develops the quantitative structure – retention relationships (QSRR) between MEDV and the retention behavior of polychlorinated dibenzodioxins (PCDDs) on gas chromatographic for 4 stationary phases. Using support vector machine (SVM), 4 support vector regression models of experiment value as train-set are developed, with correlation coefficient (R) of 4 models beyond 0.99, and the correlation coefficient left cross validation (q) is also beyond 0.99. 2/3 samples are selected uniformly from experiment value as a new train-set, the rest samples as test-set and the unknown samples as prediction-set built new models, the correlation coefficient (R) of 4 models will also be beyond 0.99. With the established models, the retention behaves of PCDDs not from the modeling samples are predicted.

Key words: polychlorinated dibenzodioxins (PCDDs); quantitative structure – retention relationships(QSRR); retention behave on gas chromatographic; molecular electronegativity distance vector(MEDV); support vector machine (SVM)