有机化合物水生毒性作用模式的支持向量机分类研究*

Study of Support Vector Machine Classification of Model of Toxicity Action of Organic Compounds

易忠胜[1,2]，刘树深[1,2,3]

YI Zhong-sheng[1,2]，LIU Shu-shen[1,2,3]

（1.桂林工学院材料与化学工程系，广西桂林　541004；2.桂林工学院有色金属及其加工新技术省部共建教育部重点实验室，广西桂林　541004；3.南京大学环境学院，江苏南京　210093）

（1. Department of Material and Chemistry Engineering，Guilin University of Technology，Guilin，Guangxi，541004，China；2. Key Laboratory of Nonferrous Materials Processing Technology of Ministry of Education，Guilin University of Technology，Guilin，Guangxi，541004，China；3. School of Environment，Nanjing University，Nanjing，Jiangsu，210093，China）

摘要：采用正辛醇/水分配系数 $\log K_{ow}$、最低未占有轨道能 E_{LUMO}、最高占有轨道能 E_{HOMO}等参数为描述变量，结合支持向量机算法，选择具有最大(Leave-One-Out，LOO)交互检验识别率，兼顾支持向量样本数和边界支持向量样本数为标准进行支持向量机建模参数优化，建立3组化合物水生毒性作用模式的SVM分类模型，分别对样本数为190、221、88的化合物水生毒性作用模式进行分类研究。结果表明，选用RBF核函数和C-SVC方法，3组参数分别为 $C=512$、$\gamma=2.048$，$C=512$、$\gamma=2.048$，$C=512$、$\gamma=0.512$ 时，建立3个体系的SVM分类模型对全部样本的错误误别个数分别为0、2、0个，训练集模型对全部样本的错误识别个数分别为9，17，7。分类的效果与化合物描述子的选择和数量有关，如果增加合适的分子描述子，其分类结果相应地会得到改善。

关键词：有机化合物　毒性作用　分类　支持向量机

中图法分类号：O6-04　　**文献标识码**：A　　**文章编号**：1005-9164(2006)01-0031-04

Abstract：Taking the octanol-water partition coefficient $\log K_{ow}$，the energy of the lowest unoccupied molecular orbital E_{LUMO}，and the energy of the highest occupied molecular orbitl E_{HOMO}as described variables，and the highest leave-one-out(LOO) rate and the lower number of support vector samples and bounded support vectors as criterion of searching the optimum parameters of SVM，a classification problem about the model of aquatic toxicity action of 3 sets organic compounds(the number of compounds are 190，221 and 88 respectively) has been built by the SVM)(Support Vector Machines) technique. The result shows that the misclassified numbers of 3 sets compounds are 0，2，0 for C-SVC with RBF kernel and 9，17，7 for train-set while the 3 sets of parameters are c = 512 $\gamma=2$，c = 512 $\gamma=0.248$ and c = 512 $\gamma=0.512$ respectively. The qualities of SVM's classification models are relation to the selection and number of descriptors and would be improved after adding the number of descriptors.

Key words：organic compounds，toxicity action，classification，support vector machine

支持向量机(Support Vector Machine，SVM)是Vapnik等[1~3]在统计学习理论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，可推广至多类和回归建模问题。由于SVM具有比神经网络更好的泛化推广能力，能消除神经网络的过拟合现象，能对小样本问题构建稳定可预测的统计分类与回归模型，因而已成为计算智能技术研究及其相关应用领域中新的研究热点[4]。短短几年时间，SVM已在药学、计算生物学、生物信息学、结构-活性相关等研究领域模式识别和回归建模中得到广泛的应用。例如，SVM已用于微阵列基因表达数据分类[5]、蛋白质相互作用模式研究[6]、水生毒性作用模式判别[7]、多环芳烃致癌活性划分[8]、有机物水溶性估算[9]、偶氮染料最大吸收波长的预测[10]等。

收稿日期：2005-06-02

修回日期：2005-08-30

作者简介：易忠胜(1970-)，男，广西资源人，硕士，副教授，主要从事化学计量学方法应用研究。

*广西新世纪十百千人才计划、广西高校百名中青年学科带头人资助项目(桂教人2003年97号文)和桂林工学院青年扶持基金(桂工院科〔2004〕8号文)联合资助。

本文选择样本数为190、221、88的化合物的正辛醇/水分配系数 $\log K_{ow}$、最低未占有轨道能 E_{LUMO}、最高占有轨道能 E_{HOMO} 等参数为描述变量，提出以留一法(Leave-One-Out，LOO 交互检验)最高的交互检验的识别率，兼顾支持向量样本数和边界支持向量样本数为标准进行支持向量机建模参数优化，建立3组化合物水生毒性作用模式的SVM分类模型，并以同样的参数建立训练集模型，取得很好的分类效果。

1 研究方法

1.1 化合物及其毒性作用模式划分

本文分别选取样本数为190、221、88的化合物作为研究的对象。为描述方便，分别称3组化合物为体系1、2、3，这些化合物均采用 $\log K_{ow}$、E_{LUMO}、E_{HOMO} 等参数进行表征，其参数直接取自文献[11～13]，其中体系1共有190个化合物，根据文献[11]的实验结果划分为非极性和极性麻醉(分别用1，2表示)毒性作用模式。这些化合物的结构多样，包括醇、酮、醚、酯、胺等几类化合物，其中属于非极性麻醉毒性作用模式的化合物样本有114个，属于极性麻醉毒性作用模式的化合物样本有76个。体系2中的化合物为苯酚及其衍生物，主要包含烃基、氟、氯、溴、氨基、硝基等多种官能团，共221个，根据文献[12]的实验结果，本文将这些化合物的毒性作用模式划分为麻醉毒性(153个，文献中的第1类)与反应性毒性两种(68个，包括文献中的第2，3，4类)，分别用1，2表示。体系3中的化合物主要为苯酚、苯胺等及其衍生物，共88个，其毒性作用模式划分为麻醉毒性(48个)和反应性毒性(40个)[13]，分别用1，2表示。对3个体系分别均匀抽取大约2/3的样本(分别为130、155、58个样本)作为训练集，余下的约1/3的样本(分别60、66、30个样本)作为检验集，用来检验最优SVM参数建模的稳定性。

1.2 支持向量机的分类方法

水生毒性作用模式通过支持向量机来进行分类研究，而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(Quadratic Programming，QP)问题，它寻求超平面中的最优分类超平面将两类分开，具体的算法过程见参考文献[1～3]，最终得到分类函数为：

$$f(x) = \mathrm{sgn}((w^*)^T\varphi(x) + b) = \mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i^* y_i K(x_i, x) + b^*\right),$$

其中，α^* 为求解二次型规划中的Lagrange乘子 α 的最优解；b^* 为对应的常数。

1.3 描述子的选择

描述子能否表征化合物的毒性作用信息对化合物的毒性作用模式分类有着很大的关系，将直接影响分类的效果[7]。麻醉毒性的疏水性通常以有机物在互为饱和的辛醇和水两相中的分配比例 K_{ow} 来表示，化合物毒性等诸多环境参数与辛醇/水分配系数的对数值($\log K_{ow}$)有较好的相关性，化合物通过疏水性作用进入生物脂质导致生物膜系“溶胀”而丧失功能的能力大小与通过分配作用进入辛醇的能力大小成正相关[14]。极性麻醉的毒性比麻醉毒性稍大，这类有机物往往拥有一些氢键给予基团，氢键成键能力可以区分为受体和给体的成键能力，二者又有包含离子和共价贡献，氢键受体和结体中离子对氢键的贡献分别用 Q^- 和 Q^+ 表示。氢键的共价贡献以作为氢键受体的最高占有轨道(Highest Occupied Molecular Orbital，HOMO)与作为氢键给体的最低未占有轨道(Lowest Unoccupied Molecular Orbital，LUMO)之间能量差的形式给出；作为受体的污染物，其共价贡献与化合物的HOMO和污染对象LUMO之间的能量差相关，而作为给体的污染物共价贡献则与化合物的LUMO和污染对象HOMO之间的能量差相关。因为在同一个研究体系中，污染目标对所有的化合物来说是一样的，所以共价贡献与污染物的前线轨道能相关，因此可以分别用前线轨道能 E_{HOMO} 和 E_{LUMO} 表示氢键受体和氢键给体的特征。

如果分子中的氢形成氢键，那么给予氢键的电荷将优先进入LUMO，这也将得到更低的 E_{LUMO} 值；另一方面，如果分子参与形成外部的氢键，分子中的HOMO电荷优先释放出来给予氢键，通常HOMO能量越高越有利于氢键的形成。一般来说，好的氢键给予就有高的 Q^+ 值和低的 E_{LUMO}；反之，好的氢键受体就有低的 Q^-(大的绝对值)和高的 E_{HOMO} 值[12]。反应性毒性是有机物毒性最大的一类，这类有机物的毒性作用机理种类很多，很难用某种确定的特征描述子建立相关关系[14]。通过以上的分析，本文尝试选择文献中的部分参数 $\log K_{ow}$ 和前线轨道能 E_{LUMO}、E_{HOMO}，作为三组化合物特征的描述子，分别对它们的毒性作用模式用SVM进行分类研究。

1.4 参数优化方法

根据支持向量机原理[1～3]，SVM用于分类时有C-SVC和υ-SVC方法，而常用核函数中，计算速度快的线性函数 $K(x, x_i) = (x^T x_i)$ 本身没有参数，径向基(RBF)函数 $K(x, x_i) = \mathrm{e}^{-\gamma\|x - x_i\|^2}$ 有一个参数γ，因而在进行SVM参数优化时，需要考虑2种SVC算法、SVM本身的参数C/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没有进行大范围讨论或者不讨论的情况，我们对这些参数采用网格的方法进行大范围的搜索，确定参数的范围，根据Lee J H[15]提供的实验结果设置参数，按照网格的方式把各参数组合，用这些参数组合进行SVM计算，选择各参数组合的LOO交互检验识别率最大，支持向量样本数和边界样本数的数量不能太大的参数组合作为最终的建模参数。全部计算采用LIBSVM软件完成[16]。

2 结果与分析

2.1 支持向量机的分类

以3个体系有机物的 $\log K_{ow}$、E_{LUMO}、E_{HOMO}作为自变量，有机物的毒性作用模式作为因变量，分别计算2种核函数在不同参数组合下，C-SVC和υ-SVC的LOO交互检验识别率的计算结果(表1)表明，当使用线性核函数时，2种SVM算法的最优分类效果基本持平，C-SVC对参数不是很敏感，其LOO交互检验识别率随参数 C 的变化不大，ν-SVC对参数 ν 的变化非常敏感；当采用RBF核函数时，C-SVC的LOO交互检验最高识别率达到了100%、99.10%、100%，υ-SVC的LOO交互检验最高识别率只有94.18%，并且2种SVC的参数变化情况类似线性核函数，但是没有前者变化大。因此C-SVC方法相对比较稳定，且核函数采用RBF函数时3个体系的最大LOO交互检验识别率大于99%，所以选择RBF函数为核函数的C-SVC方法作为建模SVM方法。

表1 SVM的参数设置、最大LOO交互检验识别率及其对应的参数

Table 1 The parameters setting of SVM, the largest LOO rate and its parameters in SVM

SVC算法 SVC method	核函数 Kernel function	参数设置 Parameter setting	最大LOO交互检验识别率 The highest LOO rate (%)		
			190个	221个	88个
C-SVC	线性函数 Linear function	$C=1\times2^N$, $N=0,1,\cdots9$	88.89	82.27	82.76
	RBF函数 RBF function	$\gamma=0.00025\times2^N$, $N=0,1,\cdots13$	100.00	99.10	100.00
υ-SVC	线性函数 Linear function	$\upsilon=0.001\times2^N$, $N=0,1,\cdots9$	88.89	84.55	82.76
	RBF函数 RBF function		94.18	90.00	90.80

确定SVM方法和RBF函数为核函数后，从表1可知，只有 γ、C 参数会影响C-SVC分类效果，按照最优参数的选择标准分别得到3个体系最终建模参数为：$C=512,\gamma=2.048$；$C=512,\gamma=2.048$；$C=512,\gamma=0.512$。

根据得到的最优模型参数对全部样本进行训练建模得到的3个体系模型见表2中的 M_{A11}，其支持向量样本数分别为80、57、20，这些模型对全部样本错误识别的个数分别为0、2、0，LOO交互检验错误识别的个数分别为0、2、0。

为了说明建模参数的可靠性而对3个体系分别用训练集样本进行SVM训练建模得到的模型见表2中的 $M_{2/3}$，其支持向量样本数分别为59、50、19，这些模型对全部样本识别错误的样本数分别为9、17、7，其中训练集识别错误的样本数分别为1、0、0，检验集识别错误的样本数分别为8、17、7。

Ren S[11]采用 $\log K_{ow}$、E_{LUMO}、E_{HOMO}、Q^+、Q^-等5个描述子结合判别分析方法对体系1的190个有机化合物的毒性作用模式进行判别，得到8个有机化合物的毒性作用模式判别错误，LOO交互检验识别错误的样本数为8，其中非极性麻醉样本有7个，极性麻醉有1个样本识别错误。我们选择其中的3个描述子：$\log K_{ow}$、E_{LUMO}、E_{HOMO}，结合支持向量机方法，当采用RBF函数作为核函数，C-SVC方法，$C=512,\gamma=2.048$ 时，建立的判别函数对全部样本能够正确识别，LOO交互检验识别率达到100%，当取130个样本作为训练集得到模型，对全部的样本也只有9个样本识别错误，我们的分类效果明显优于文献[11～13，17]。

2.2 描述子与分类效果分析

采用3个描述子分别确定3个体系的SVM分类模型参数，并利用这些模型参数研究单个描述子各自的分类效果。表3表明，E_{LUMO}、E_{HOMO}参数在3个体系中正确分类超过80%，而 $\log K_{ow}$ 对体系1的分类效果超过70%，对体系3的分类效果在68%以上，说明我们选择 $\log K_{ow}$、E_{LUMO} 和 E_{HOMO} 参数能够充分地表达3组化合物的毒性作用模式的信息。

从我们的分类结果可以发现，在增加合适的描述子的情况下，分类效果将会得到适当的改善，这与描述子越多包含有机物毒性作用模式的信息也越多有关，因此我们可以定性地认为分类效果的好坏不但与分类器的好坏有关，而且还与化合物的描述子选择是否合适、描述子的数量有很大的关系。从本文的研究结果来看在相同的条件下，SVM的分类效果明显优于其它分类方法[11～13,17]。

表 2　3 个体系的 SVM 模型及其正确识别样本数

Table 2　The models of SVM in 3 systems and the number of correctly cognized samples respectively

体系 System	M_{A11}	$M_{2/3}$	正确/错误识别样本数 The number of correct /misclassified samples	
			M_{A11}	$M_{2/3}$
1	$f(x)=\mathrm{sign}(\sum_{i=1}^{80}\alpha_i y_i e^{-2.048\|x-x_i\|^2}-0.6288)$	$f(x)=\mathrm{sign}(\sum_{i=1}^{59}\alpha_i y_i e^{-2.048\|x-x_i\|^2}-0.9189)$	190/0	181/9
2	$f(x)=\mathrm{sign}(\sum_{i=1}^{57}\alpha_i y_i e^{-2.048\|x-x_i\|^2}-0.7553)$	$f(x)=\mathrm{sign}(\sum_{i=1}^{50}\alpha_i y_i e^{-2.048\|x-x_i\|^2}-1.2468)$	219/2	204/17
3	$f(x)=\mathrm{sign}(\sum_{i=1}^{20}\alpha_i y_i e^{-0.512\|x-x_i\|^2}+1.2702)$	$f(x)=\mathrm{sign}(\sum_{i=1}^{19}\alpha_i y_i e^{-0.512\|x-x_i\|^2}-0.6454)$	88/0	81/7

表 3　3 个描述子各自在 3 个体系中的分类效果

Table 3　The classified effect of 3 descriptors in 3 systems respectively

体系 System	$\log K_{ow}$(%)	E_{LUMO}(%)	E_{HOMO}(%)
1	71.05(135/190)	82.11(156/190)	84.21(160/190)
2	71.04(157/221)	85.52(189/221)	84.16(186/221)
3	68.18(60/88)	87.50(77/88)	81.82(72/88)

3　结束语

本文采用正辛醇/水分配系数 $\log K_{ow}$、最低未占有轨道能 E_{LUMO}、最高占有轨道能 E_{HOMO} 等参数表征有机物分子，结合支持向量机算法对样本数为 199、221、88 的化合物的毒性作用模式进行分类研究，针对目前 SVM 的最优参数的确定没有合适的理论指导，提出以 LOO 交互检验识别率的大小，并且兼顾支持向量样本数/边界支持向量样本数的多少作为选择最优模型参数的标准，分别建立样本数为 190、221、88 的有机物水生毒性作用模式的 SVM 模型，这些模型对全部样本错误识别的样本数分别 0、2、0，LOO 交互检验的错误识别样本数分别为 0、2、0，取得了非常好的分类效果。

参考文献：

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

[2] VAPNIK V. The nature of statistical learning theory [M]. New York: Springer-Verlag, 1995.

[3] VAPNIK V N. Statistical learning theory[M]. New York: Wiley Publishers Science, 1998.

[4] IVANCIUC O. Support vector machines for cancer diagnosis from the blood concentration of Zn, Ba, Mg, Ca, Cu and Se [J]. Internet Electronic Journal of Molecular Design, 2002, 1(8): 418-427.

[5] BROWN M P S, GRUNDY W N, LIN D, et al. Knowledgebased analysis of microarray gene expression data by using support vector machines [J]. Process Natural Academy Sciences, 2000, 97(1): 163-267.

[6] BOCK J R, GOUGH D A. Predicting protein-protein interactions from primary structure [J]. Bioinformatics, 2001, 17(5): 455-460.

[7] IVANCIUC O. Support vector machine identification of the aquatic toxicity mechanism of organic compounds [J]. Internet Electronic Journal of Molecular Design, 2002, 1(3): 157-172.

[8] IVANCIUC O. Support vector machine classification of the carcinogenic activity of polycyclic aromatic hydrocarbons [J]. Internet Electronic Journal of Molecular Design, 2002, 1(4): 203-218.

[9] LIND P, MALTSEVA T. Support vector machines for the estimation of aqueous solubility[J]. Journal Chemical Information Computer Sciences, 2003, 43: 1855-1859.

[10] 赵　慧，陆文聪，宋海峰. 支持向量回归方法预报偶氮染料最大吸收波长[J]. 化学学报，2004，62(7)：649-656.

[11] REN S. Classifying class I and class II compounds by hydrophobicity and hydrogen bonding Descriptors[J]. Environ Toxicol, 2002, 17: 415-423.

[12] APTULA A O, NETZEVAB T I, VALKOVAC I V, et al. Multivariate Discrimination between Modes of Toxic Action of Phenols[J]. Quantitative Structure-Activity. Relationships, 2002, 21: 12-22.

[13] REN S, SCHULTZ T W. Identifying the mechanism of aquatic toxicity of selected compounds by hydrophobicity and electrophilicity descriptors [J]. Toxicology Letters, 2002, 129: 151-160.

[14] 王连生，韩朔睽. 分子结构、性质与活性[M]. 北京：化学工业出版社，1997.

[15] LEE J H, LIN C J. In Automatic model selection for support vector machines, NIPS workshop on kernel methods, Breckenridge [EB/OL]. 2000. http://www.csie.ntu.edu.tw/~cjlin/papers/modelselect.ps.gz.

[16] CHANG C C, LIN C J. LIBSVM, a library for support vector machines [EB/OL]. 2001-02-06. Http://www.csie.ntu.edu.tw/-cjlin/libsvm.

[17] REN S. Two-step multivariate classification of the mechanisms of toxic action of phenols [J]. QSAR Combinatorial Sciences, 2003, 22: 596-603.

（责任编辑：黎贞崇　邓大玉）