基于支撑向量机方法的有机化合物的生成 Gibbs 自由能的预测

王　冰　刘焕香　姚小军　任月英　胡之德*
（兰州大学化学化工学院　兰州 730000）

摘　要　用支撑向量机研究了 607 种有机化合物的结构和他们的 Gibbs 自由能的关系，建立了相应的 QSPR 模型。表示分子结构的描述符是从 CODESSA 软件中计算得到的，通过前向逐步回归分析选择其中的 13 个描述符，多元线性回归（MLR）和支撑向量机（SVM）被分别用来构造线性和非线性模型预测有机化合物的 Gibbs 自由能。支撑向量机方法得到的模型对整个数据集、训练集、测试集的平均绝对偏差分别为31.098 9、30.569 5 和 35.924 6 kJ/mol，预测结果令人满意。

关键词　支撑向量机，多元线形回归，吉布斯自由能
中图分类号：O657.3；TP389.1　　**文献标识码**：A　　**文章编号**：1000-0518(2006)05-0552-05

分子的定量结构-性质关系（QSPR）分析可以从结构参数定量预测物质的性质。自从 1980 年以来，人工智能技术已经被用在 QSPR 分析中，这类机器学习法虽然可以提供更高的精确度，但也可能导致数据集的过拟合及重现性问题，很大程度是由于网络的初始化和终止标准的变化造成的，分类信息的缺乏也是导致结果重现性差的一个原因。遗传算法也会产生类似问题。由于上述原因，QSPR 分析急需引入更精确、更新的信息化技术。SVM 是一个新的机器学习算法，由于它卓越的泛化性能已引起了广泛注意。QSPR 应用中的另一个主要问题是化学结构的数字表示（或称为分子描述），它会直接影响建模性能和结果的精确性。在有机化合物的 QSPR 研究中，可以采用结构描述符，拓扑描述符，数字编码，量子化学描述符等多种结果表示方法。由 Katritzky 小组发明的 CODESSA 软件[1] 可以计算组成结构、拓扑、电子、量子化学描述符，并且已成功用于各种 QSPR 研究中。

有机化合物的 Gibbs 自由能在化学工程中是一个非常重要的物理-化学参数。在一定的温度和压强下，它是判断化学反应平衡的重要标准。本文用 SVM 计算了 298 K 时以 CODESSA 软件计算的描述符作为输入结构参数的有机化合物的吉布斯自由能，预测结果对于训练集和测试集均令人满意。

1.1　实验部分

1.1　数据集

用于本研究的 607 种从 C1 到 C6 的有机化合物的结构数据和 298 K 下的吉布斯自由能的测定值取自 Yaws 编辑的物理化学性质手册。化合物类型包括烃类、醇类、醛类、酮类、酸类、酯类等。数据集被分为 547 种化合物的训练集和 60 种化合物的测试集，分别用于选择 SVM 参数和 SVM 预测能力的评估。

1.2　分子描述符的产生

用 ISISDRAW 程序[2] 画出分子的三维结构，最后的分子构型在 HYPERCHEM 程序[3] 中用半经验 PM3 方法得到。所有的计算都在 Hartree Fock 限制下实现并且不考虑构象影响。用 Polak-Ribiere 算法优化分子结构到均方误差梯度平方根为 0.001。将计算出的几何结构输入 CODESSA 软件计算分子的结构、拓扑、电子、量子化学描述符。

1.3　支撑向量机回归原理

SVM 是由 Vapnik 首先提出的一类指导学习算法，已成功用于蛋白质折叠识别[4]、蛋白质结构分类预测[5]、蛋白质分裂点鉴别[6] 以及 QSAR 和药物数据分析[7] 中。

2005-06-17 收稿，2005-12-20 修回
通讯联系人：胡之德，男，1931 年生，教授，博士生导师；E-mail：huzd@lzu.edu.cn；研究方向：分析化学

SVM 的基本思想是将数据 X 通过一个非线性映射 Φ 映射到更高维特征空间 F 中,并在这个空间中作线性回归。其拟合函数有如下形式:

$$y = \sum_{i=1}^{l} w_i \Phi_i(x) + b \tag{1}$$

式中 $\{\Phi_i(x)\}_{i=1}^{l}$ 为输入的特征值。$\{w_i\}_{i=1}^{l}$ 和 b 为系数,它们由最小化风险估算函数(2)估算得到

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i) + \frac{1}{2}\| w \|^2 \tag{2}$$

式中,

$$L_\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon & |d - y| \geqslant \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

此处的 ε 是一个指定的参数。

在式(2)中,$C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i)$ 被称作经验风险,它由通过 ε-不敏感丢失函数 $L_\varepsilon(d, y)$ 获得,它表示不惩罚低于 ε 的错误,加号后面的$\frac{1}{2}\| w \|^2$ 项是函数绝对性的测量,C 是常量,用来决定在训练错误和模型平滑度之间的平衡。引进稀疏变量‘ ξ ’导致式(2)有以下限制函数:

$$\text{Max} \quad R(w, \xi^*) = \frac{1}{2}\| w \|^2 + C^* \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{4}$$

$$\text{s. t.} \quad \begin{aligned} & w\Phi(x_i) + b - d_i \leqslant \varepsilon + \xi_i, \\ & d_i - w\Phi(x_i) - b_i \leqslant \varepsilon + \xi_i, \\ & \xi, \xi^* \geqslant 0 \end{aligned} \tag{5}$$

拟合函数(1)成为:

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{l} (\alpha^* - \alpha_i) K(x, x_i) + b \tag{6}$$

α_i, α_i^* 为 Lagrange 乘子,满足 $\alpha_i \cdot \alpha_i^* = 0, \alpha_i^* \geqslant 0$; $i = 1, \cdots, l$ 通过最大化限制函数(4)的二相性获得以下等式:

$$\Phi(\alpha_i, \alpha_i^*) = \sum_{i=1}^{l} d_i(\alpha_i - \alpha_i^*) - \varepsilon\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(\alpha_i, \alpha_j) \tag{7}$$

它有以下限制:

$$\begin{cases} 0 \leqslant \alpha_i \leqslant C, \ i = 1, \cdots, l, \\ 0 \leqslant \alpha_i^* \leqslant C, \ i = 1, \cdots, l \\ \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \end{cases}$$

依据 Karush-Kuhn-Tucker(KKT)的二次规划条件限制,只有少数样本的系数($\alpha_i - \alpha_i^*$)为非 0 值,和他们对应的数据点被称为支撑向量。

在等式(6)中,$K(x_i, x_j)$ 是核函数,2 个向量 x_i 和 x_j 的内积和特征空间 $\Phi(x_i)$,$\Phi(x_j)$ 相对应,即 $K(x_i, x_j) = <\Phi(x_i) \cdot \Phi(x_j)>$。使用核函数的优点在于事实上它可以处理任意维特征空间而不用精确计算映射 $\Phi(x)$。任何函数只要满足 Mercer 的条件都可以用作核函数,在支撑向量回归中,最常用的是高斯核函数 $K(x, y) = \exp(-(x - y)^2/\delta^2)$。

2.4　SVM 的实现和计算环境

所有用来实现 SVM 的计算程序都写入 R 文件,所有文件用 R1.6.2 编译器编译,并在 Pentium 4 的机器上运行这些文件。

2　结果与讨论

2.1　MLR 的结果

经过 CODESSA 软件计算,选出了 72 个描述符,在 R 中,以计算过的分子描述符为基础用向前逐步回归建立线性 QSPR 模型,最好的线性模型包含了 13 个分子描述符,其中有 7 个为组成描述符,6 个为

拓扑描述符，见表 1。该模型对训练集的 Gibbs 自由能平均绝对偏差(MAE)为 45.15 kJ/mol 和相关系数(R)为 0.970。对于测试集其 MAE 为 37.59 kJ/mol。图 1 给出了 MLR 方法模型的预测结果。

表 1 描述符、系数、标准偏差、线性模型的 T 检验值

Table 1 Descriptors, coefficient(R), standard deviation(SD) and linear model's T values(T)

Chemical meaning	Descriptors	R	SD	T
Intercept	(Constant)	15.875 1	24.895 3	0.637 7
Number of C atoms	NCA	−116.626 2	6.696 2	−17.416 7
Number of O atoms	NOA	−161.311 9	3.800 9	−42.440 3
Number of F atoms	NFA	−149.936 5	4.996 4	−30.008 9
Number of double bonds	NDB	152.282 0	6.512 5	23.383 0
Number of triple bonds	NTB	320.368 5	12.370 3	25.898 2
Number of aromatic bonds	NAB	62.018 7	3.078 5	20.146 0
Number of rings	NR	216.392 9	12.040 2	17.972 5
Randic index(order 2)	RI2	−91.593 1	7.171 8	−12.771 3
Kier & Hall index(order 0)	KHI0	85.273 4	7.006 8	12.170 2
Kier & Hall index(order 1)	KHI1	−98.608 5	7.433 5	−13.265 4
Average structural information Content (order 0)	ASIC0	890.137 4	113.283 9	7.857 6
Complementary information Content(order 0)	CIC0	11.298 7	0.581 7	19.422 7
Average bonding information Content(order 0)	ABIC0	−780.558 8	83.237 7	−9.377 5
R = 0.970, F = 644.1, MAE = 45.15				

2.2 SVM 的结果

2.2.1 SVM 参数的优化　对回归问题来说，和其他统计方法一样，SVM 的性能依靠容量参数 C、γ 和 ε-不敏感丢失函数，核函数类型以及它的相关参数的组合。C 是控制最大化边界和最小化误差平衡的规则化参数，C 太小将对训练集产生欠拟合，如果 C 太大则会出现过拟合。但是文献[8]认为 C 对预测结果的影响很小，为了让学习过程稳定，可首先选定一个较大的 C 值(例如 C = 1 000)。通常情况下对 ε 来说最优值的选择是依赖于数据中的噪音，但它是不可知的，对 ε 的最优参数的选择可获得足够多的噪音知识，从而对模型的支撑向量数目可以有一个实际的考虑。因此，选择一个合适的 ε 是非常重要的。

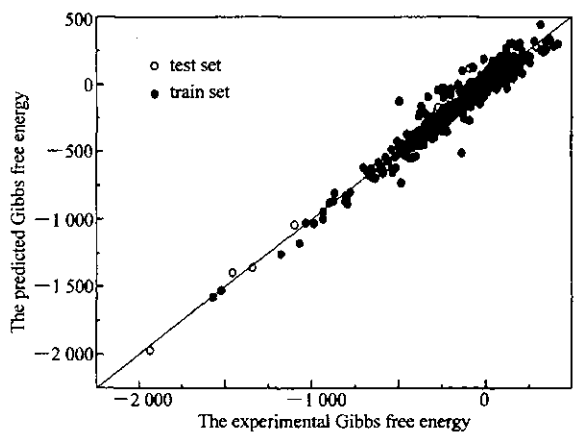

图 1 预测-试验的吉布斯能量(MLR)

Fig. 1 MLR of prediction-test

最常用的核函数类型是 Gaussian 核函数，在 R 中 Gaussian 核函数的形式为：

$$K(u,v) = \exp(\gamma^* \mid u - v \mid^2) \tag{8}$$

式中，u，v 为 2 个独立变量，γ 为一个常数，它控制 Gaussian 核函数的振幅，因此也控制了 SVM 的泛化能力，所以必须优化 γ 并找到其最优解。为了得到上述参数的最优组合，对整个训练集进行留一法交互检验。留一法步骤包括从训练集中取出 1 个样本，以剩余样本为基础构造决定函数，再测试所取出的样本，如此循环获得训练模型。并用平方相关系数来评价训练模型的行为。平方相关系数定义如下：

$$R^2 = \frac{\sum (y_i - \bar{y}_i)(\hat{y}_i - \bar{\hat{y}}_i)}{\sum (y_i - \bar{y}_i)^2 \sum (\hat{y}_i - \bar{\hat{y}}_i)^2} \tag{9}$$

式中，y_i，$\hat{y}_i$ 分别为化合物 Gibbs 自由能的实验值和预测值。

选择参数的详细过程及每个参数对支撑向量及概括泛化性能的影响见表 2、图 2、图 3、图 4。从图 2 及表 2 可以看出，γ 对支撑向量数及均方相关系数有较大的影响，其最优值为 0.003。为了寻找最优的 ε，用不同 ε 值的 SVM 模型的训练值对实验值的均方相关系数 R^2 与 ε 作图(如图 3)，得 ε 的最优值为

0.14。规则化参数 C 对均方相关系数 R^2 的影响见图 4。从图 4 可以看出，模型的行为随参数 C 的增加首先增加，然后减小，最后变化不敏感，这和参考文献[5]描述一致，其最优值为 2 000。

表 2　SVM 参数的优化

Table 2　The optimization of SVM's parameters

No.	C	ε	γ	Support Vectors	Squared correlation coefficient
1	1 000	0.20	0.000 7	137	0.939 7
2	1 000	0.20	0.001	130	0.943 3
3	1 000	0.20	0.003	121	0.946 5
4	1 000	0.20	0.000 5	117	0.943 7
5	1 000	0.20	0.000 7	123	0.938 0
6	1 000	0.20	0.000 9	126	0.934 1
7	1 000	0.08	0.003	268	0.946 8
8	1 000	0.10	0.003	228	0.948 0
9	1 000	0.12	0.003	194	0.948 5
10	1 000	0.14	0.003	173	0.948 5
11	1 000	0.16	0.003	152	0.948 0
12	1 000	0.18	0.003	140	0.947 2
13	100	0.14	0.003	186	0.941 0
14	500			174	0.946 5
15	2 000			171	0.949 8
16	3 000			166	0.949 1
17	4 000			166	0.949 1
18	5 000			170	0.948 6

支撑向量数目是评价 SVM 模型的另一个重要指标。一般来说，当均方相关系数相同时，支撑向量数目越少，模型越好。因为支撑向量数目越少，训练时间越短，这一点对大多数化学信息学问题非常重要。例如本文试验 9、10 所得到的模型的均方相关系数相同，但是我们选定 $\gamma = 0.14$ 为最优值，就是由于其支撑向量数较少。

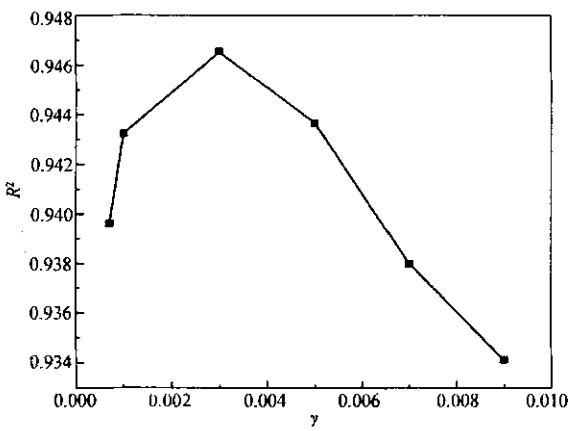

图 2　γ 值对均方相关系数的影响

Fig. 2　The effect of γ value on the squared correlation coefficient

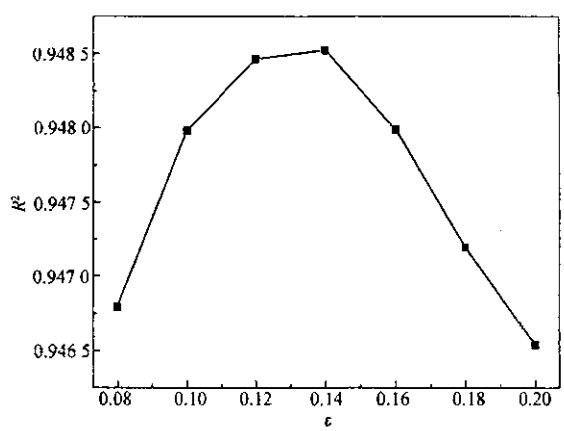

图 3　ε 值对均方相关系数的影响

Fig. 3　The effect of ε on the squared correlation coefficient

2.2.2　SVM 的预测结果　上述结果表明，γ、ε、C 的值分别固定在 0.003、0.14 和 2 000 时，SVM 的支撑向量数为 171 个（仅是训练样本的一部分），最优的 SVM 预测结果如图 5 所示。整个数据集、训练集和测试集 Gibbs 自由能对实验值的平均绝对误差分别为 31.098 9、30.569 5 和 35.924 6 kJ/mol，训练集的相关系数为 0.983 2，比 MLR 模型的相关性（$R = 0.970$）好，表明本文由 SVM 得到的模型可以更准确地预测这些化合物的结构-性质关系。虽然所处理化合物的结构差别很大但结果令人满意，证明了以分子结构的描述符为基础，用 SVM 方法对系列有机化合物 Gibbs 自由能的测试值进行拟合，得到的非线性

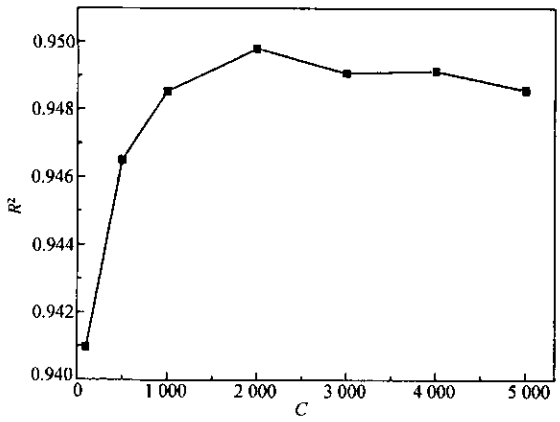

图4 C 值对均方相关系数的影响

Fig. 4 The effect of C value on the squared correlation coefficient

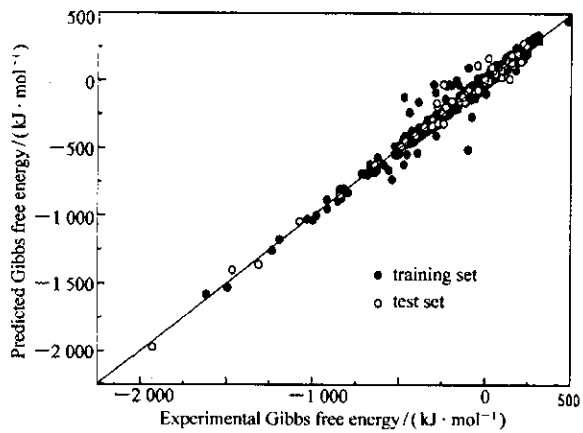

图5 预测和实验吉布斯自由能的相关性

Fig. 5 Relationship between prediction and experimental Gibbs free energy

模型具有很好的预测能力。

参 考 文 献

1 Katritzky A R, Lobanov V S, Karelson M. CODESSA Version 2.0 Reference Manual[M], 1995 ~ 1997

2 ISIS Draw2.3 MDL Information Systems Inc, 1990

3 HyperChem, Release 4.0 for Windows, Hypercube, Inc, 1995

4 Ding C H Q, Dubchak I. *Bioinformatics*[J], 2001, **17**: 349

5 Karchin R, Karplus K, Haussler D. *Bioinformatics*[J], 2002, **18**: 147

6 Cai Y D, Liu X J, Xu X B, *et al*. *J Comput Chem*[J], 2002, **23**: 267

7 Czerminski R, Yasri A, Hartsough D. *Quant Struct-Act Relat*[J], 2001, **20**: 227

8 Wang W J, Xu Z B, Lu W Z, *et al*. *Neurocomputing*[J], 2003, **55**: 643

Prediction of Gibbs Free Energy of Formation of Organic Compounds Based on Support Vector Machines

WANG Bing, LIU Huang-Xiang, YAO Xiao-Jun, REN Yue-Ying, HU Zhi-De*

(*College of Chemical and Engineering, Lanzhou University, Lanzhou* 730000)

Abstract The support vector machine (SVM), as a novel type of learning machine, was used to develop a QSPR model of Gibbs free energies of 607 organic compounds. The descriptors calculated by CODESSA were used to represent the molecular structures. Thirteen of those descriptors were selected by forward stepwise regression and were used developing models to predict Gibbs free energy of the formation of an organic compound. Multiple linear regression (MLR) and SVM were utilized to construct linear and non-linear models of the organic compound. The optimal QSPR model based on the support vector machine was obtained. The mean-absolute error (MAE) of Gibbs free energy of formation was 31.098 9 kJ/mol for the whole set, 30.569 5 kJ/mol for the training set, and 35.924 6 kJ/mol for the test set respectively. The prediction results are more satisfactory than those of MLR.

Keywords support vector machine, linear discriminant analysis, Gibbs free energy