

基于拓扑结构的碱金属化合物摩尔磁化率的支持向量回归研究^{*}

蔡从中[†] 庄魏萍 温玉锋 朱星键 裴军芳 肖婷婷

(重庆大学应用物理系,重庆 400044)

(2009 年 1 月 15 日收到,2009 年 4 月 30 日收到修改稿)

基于经典电动力学导出的表征简单离子磁化率的磁性点价 g_i 所构建的分子磁性连接性指数^m F 及 45 种碱金属化合物的摩尔磁化率 χ_m 的实测数据集,利用粒子群寻优的支持向量回归(SVR)方法,建立了基于⁰ F 和¹ F 的碱金属化合物 χ_m 的预测模型,并与基于多元线性回归(MLR)模型的计算结果进行了比较.结果显示,基于 9 次交叉验证的 SVR 模型预测的平均绝对误差、平均相对误差绝对值以及均方根误差均比 MLR 模型小,表明 SVR 模型的回归预测能力优于 MLR.研究表明,磁性连接性指数^m F 是一种合适的分子描述符,SVR 是一种预测碱金属化合物 χ_m 的有效方法.

关键词:碱金属化合物,摩尔磁化率,支持向量回归,预测

PACC:7530C,0270,0650

1. 引 言

物质的磁性与组成它的原子、离子或分子的微观结构有关,研究物质的磁性对于确定原子结构、分子结构具有极其重要的理论意义.物质磁化率反映了磁介质在外磁场中被磁化的信息,人们可以通过测定物质磁化率来计算物质分子中未成对电子数以研究分子中成键状况以及判断配合物的结构类型等.根据物质的不同结构特征可采用不同的方法来测量物质的磁化率,常见磁化率测定方法有 Gouy 磁天平法^[1]、Curie 磁秤法^[2]、Faraday 相对法^[3]、Quincke 法^[4]、Rankine 法^[5]等.除直接测量外,人们还应用蒙特卡罗方法、第一性原理、多元线性回归(multivariate linear regression,简记为 MLR)、神经网络等方法对物质磁化率的模拟计算进行了一系列的探索研究^[6-12].由于早期的加和法在计算分子磁化率时未考虑原子在分子中的价态以及成键的性质等对分子磁化率的影响,其估算结果的准确率往往较低^[13].冯长君等^[13]根据经典电动力学原理,引入与摩尔磁

化率 χ_m 显著相关的表征简单离子磁化率的磁性点价 g_i 结合分子图的邻接矩阵构建了一种新的分子磁性连接性指数^m F ,利用 MLR 方法对碱金属化合物的磁化率进行了拟合研究.结果表明,零阶(⁰ F)和一阶(¹ F)连接性指数与碱金属化合物的磁化率具有良好的二元相关性,其相关系数达 0.957.本文是在文献^[13]的工作基础上,利用支持向量回归(support vector regression,简记为 SVR)方法^[14]并结合粒子群寻优(particle swarm optimization,简记为 PSO)算法,基于零阶和一阶连接性指数,对 45 种碱金属化合物的摩尔磁化率 χ_m 进行了 SVR 建模和预测研究.

2. 原理与方法

2.1. SVR 原理

支持向量机是 Vapnik 及其合作者基于统计学习理论及结构风险最小化原则而提出的一种机器学习方法^[14,15],可用于数据的分类与回归研究,具有很

^{*} 教育部新世纪优秀人才支持计划(批准号:NCET-07-0903)、教育部留学回国人员科研启动基金(批准号:2008101-1)、重庆市自然科学基金(批准号:CSTC2006BB5240)和国家大学生创新性实验计划(批准号:CQUCX-G-2007-016)资助的课题.

[†] E-mail: caiczh@gmail.com

强的学习与泛化能力,已被广泛地应用于实际问题的分类与回归研究^[16-26]。

SVR 的基本思想是针对不满足线性关系的样本集 $(x_1, y_1) \dots (x_m, y_m)$ 通过输入空间到输出空间的非线性映射 Φ 将样本数据 x 映射到高维空间 F , 并在 F 中用下列函数进行线性回归:

$$E(f(x_i) - y_i) = \begin{cases} 0 & (|f(x_i) - y_i| < \epsilon), \\ |f(x_i) - y_i| - \epsilon & (|f(x_i) - y_i| \geq \epsilon), \end{cases} \quad (3)$$

式中 m 为训练样本数, $E(f(x_i) - y_i)$ 是损失函数, C 是惩罚因子, ϵ 为误差。

为控制函数的复杂性,应使线性回归函数尽量平坦,并考虑可能超出精度的回归误差,引入松弛因子 ξ 和 ξ^* ,以处理不满足(3)式的数据点。根据统计学习理论的结构风险最小化准则,SVR 是通过最小化目标函数 $R(w, \xi_i, \xi_i^*)$ 来确定(1)式中的 w 和 b ,

$$R(w, \xi_i, \xi_i^*) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*), \quad (4)$$

且满足

$$\begin{aligned} y_i - w \cdot x_i - b &\leq \epsilon + \xi_i, \\ w \cdot x_i + b - y_i &\leq \epsilon + \xi_i^*, \\ \xi_i &\geq 0, \\ \xi_i^* &\geq 0. \end{aligned}$$

(4)式等号右端第一项是使回归函数更为平坦,泛化能力更好;第二项则为减少误差,惩罚因子 C 是一个常数,且 $C > 0$,用来控制对超出误差 ϵ 的样本的惩罚程度。为求解 w 和 b ,建立拉格朗日方程

$$\begin{aligned} L(w, \xi_i, \xi_i^*) &= \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*) \\ &\quad - \sum_{i=1}^m \alpha_i ((\epsilon + \xi_i) + y_i \\ &\quad + (w \cdot \Phi(x_i)) + b) \\ &\quad - \sum_{i=1}^m \alpha_i^* ((\epsilon + \xi_i^*) + y_i \\ &\quad + (w \cdot \Phi(x_i)) - b) \\ &\quad - \sum_{i=1}^m (\lambda_i \xi_i + \lambda_i^* \xi_i^*). \end{aligned} \quad (5)$$

要使(5)式取得最小值, L 对于参数 w, b, ξ, ξ^* 的偏导数都应等于零,即

$$\begin{aligned} f(x) &= w \cdot \Phi(x) + b, \\ \Phi: R^n &\rightarrow F, w \in F, \end{aligned} \quad (1)$$

式中 b 是阈值, w 是回归系数向量。影响 w 的因素有经验风险的总和及使其在高维空间平坦的 $R(w)$,

$$R(w) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m E(f(x_i) - y_i), \quad (2)$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^m (\alpha_i - \alpha_i^*) \cdot \Phi(x_i) = 0,$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0, \quad (6)$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \lambda_i = 0,$$

$$\frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \lambda_i^* = 0.$$

代入(5)式后可以得到对偶优化问题

$$\begin{aligned} \min \left\{ \frac{1}{2} \sum_{i,j=1}^m (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) \Phi(x_i) \cdot \Phi(x_j) \right. \\ \left. + \sum_{i=1}^m \alpha_i (\epsilon - y_i) + \sum_{i=1}^m \alpha_i^* (\epsilon + y_i) \right\} = Q, \end{aligned} \quad (7)$$

且满足

$$\begin{aligned} \sum_{i=1}^m (\alpha_i - \alpha_i^*) &= 0, \\ \alpha_i &\in [0, C], \\ \alpha_i^* &\in [0, C]. \end{aligned}$$

由此,SVR 回归问题就可以归结为二次规划(7)式,从而得到用训练样本点表示的 w ,即

$$w = \sum_{i=1}^m (\alpha_i - \alpha_i^*) \Phi(x_i), \quad (8)$$

式中 α_i 和 α_i^* 是最小化目标函数 $R(w, \xi_i, \xi_i^*)$ 的解。由此可求得线性回归函数

$$f(x) = \sum_{i=1}^m (\alpha_i - \alpha_i^*) k(x, x_i) + b, \quad (9)$$

式中 $k(x, x_i) = \Phi(x) \cdot \Phi(x_i)$ 为核函数。选择不同形式的核函数就可以生成不同的 SVR 模型。常用的核函数有径向基函数、多项式函数、sigmoid 函数、线性函数等。本文采用径向基函数来建立 SVR 模型。

2.2. SVR 参数的 PSO 算法寻优

PSO 算法是在 1995 年由 Kennedy 和 Eberhart^[27]

模拟鸟群的飞行捕食行为而提出的一种高效多维并行寻优算法. 由于 SVR 模型的泛化性能完全依赖于误差 ϵ 、惩罚因子 C 和核函数参数 γ 这三个参数, 因而进行 ϵ, C, γ 参数寻优对于 SVR 达到最优泛化性能尤其关键. 采用 PSO 算法来寻找 ϵ, C, γ 的最优值, 即采用速度-位置搜索模型来寻找最优参数. 群体中的每个粒子由三维参数向量 (ϵ, C, γ) 组成, 设第 i 个粒子在三维解空间的位置为 $\mathbf{u}_i = (u_{i1}, u_{i2}, u_{i3})^T$, 其速度为 $\mathbf{v}_i = (v_{i1}, v_{i2}, v_{i3})^T$, 当前时刻的个体极值记为 $\mathbf{p}_{i\text{best}}$, 全局极值记为 \mathbf{g}_{best} . 在每次迭代中, 粒子跟踪个体极值、全局极值和自己前一刻的状态来调整当前时刻的位置和速度, 其迭代公式为

$$\mathbf{v}_i(t+1) = \omega \mathbf{v}_i(t) + c_1 \text{rand}(\cdot) \cdot (\mathbf{p}_{i\text{best}} - \mathbf{u}_i(t)) + c_2 \text{rand}(\cdot) \cdot (\mathbf{g}_{\text{best}} - \mathbf{u}_i(t)), \quad (10)$$

$$\mathbf{u}_i(t+1) = \mathbf{u}_i(t) + \mathbf{v}_i(t+1). \quad (11)$$

这里, $\mathbf{v}(t)$ 和 $\mathbf{v}(t+1)$ 分别是粒子在当前时刻和下一时刻的速度; $\mathbf{u}(t)$ 和 $\mathbf{u}(t+1)$ 分别是粒子在当前时刻和下一时刻的位置; $\text{rand}(\cdot)$ 是 $[0, 1]$ 之间的随机数; c_1 和 c_2 是学习因子, 通常取为 2; ω 是权重因子, 为加快收敛速度, 其值随算法迭代的进行而自动调节, 一般定义为

$$\omega = \omega_{\min} + (N_{\max} - N) (\omega_{\max} - \omega_{\min}) / N_{\max} \quad (12)$$

式中, ω_{\max} , ω_{\min} 分别为最大、最小权重因子, 且 ω_{\max} 和 ω_{\min} 的值一般取为 0.9 和 0.4, N 为当前迭代次数, N_{\max} 为总的迭代次数. 为了直接反映 SVR 回归性能, 选用均方根误差 E_{RMS} 作为适应度函数,

$$E_{\text{RMS}} = \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2}, \quad (13)$$

式中, m 是训练样本数, y_i 和 \hat{y}_i 分别是第 i 个训练样本目标量的实测值和预测值.

3. SVR 模型的建立

3.1. 数 据

本文所用数据来源于文献 [13], 该数据集共包含 45 个样本(见表 1), 是冯长君等根据文献 [14] 中列出的 45 种碱金属化合物(包括碱金属卤化物、硫化物、多硫化物、硫酸盐、高氯酸盐及卤酸盐)的摩尔磁化率 χ_m , 在分子图的邻接矩阵基础上引入简单离子的磁性点价 g_i , 构建了零阶和一阶连接性指数. 连接性指数 $^m F$ 的定义及详细计算方法参见文献 [13].

3.2. 建立模型

在利用 SVR 方法对不同碱金属化合物的摩尔磁化率 χ_m 的训练建模过程中, 以 $^0 F$ 和 $^1 F$ 两个参数为输入变量, 以 χ_m 为输出变量进行训练学习. 本文应用 SVR 采用 9 次交叉验证(9-fold cross validation, 简记为 9FCV)对 45 个样本进行建模训练和预测研究.

3.3. 模型预测性能的评价

采用最大相对误差绝对值 $E_{\max \text{AP}}$ 、平均绝对误差 $E_{\text{mean A}}$ 、平均相对误差绝对值 $E_{\text{mean AP}}$ 以及相关系数 r^2 对所建模型的预测性能进行评价. $E_{\max \text{AP}}$, $E_{\text{mean A}}$, $E_{\text{mean AP}}$ 以及 r^2 分别定义为

$$E_{\max \text{AP}} = \max \left| \frac{\hat{y}_j - y_j}{y_j} \right| \quad (1 \leq j \leq n) \quad (14)$$

$$E_{\text{mean A}} = \frac{1}{n} \sum_{j=1}^n |\hat{y}_j - y_j|, \quad (15)$$

$$E_{\text{mean AP}} = \frac{1}{n} \sum_{j=1}^n \left| \frac{\hat{y}_j - y_j}{y_j} \right|, \quad (16)$$

$$r^2 = \frac{\sum_{j=1}^n (\hat{y}_j - \bar{y})^2}{\sum_{j=1}^n (y_j - \bar{y})^2}. \quad (17)$$

这里, n 是检验样本数, y_j 和 \hat{y}_j 分别是第 j 个检验样本的目标值和预测值, \bar{y} 是检验样本的目标平均值.

4. 结果分析与讨论

4.1. SVR-9FCV 预测结果

将 45 个样本随机地等分为 9 组, 各组依次作为测试样本, 其余 8 组作为训练样本, 应用 SVR 进行 9FCV, 所得结果列于表 1.

文献 [13] 曾对表 1 中的 45 个样本数据进行多元线性拟合^[13], 得到 χ_m 与 $^0 F$ 和 $^1 F$ 之间的线性回归方程(以下称为 MLR1)为

$$\chi_m = 54.248 - 36.550 {}^0 F - 30.722 {}^1 F. \quad (18)$$

文献 [13] 还对 45 个样本进行了分段线性拟合, 即将碱金属卤化物(样本 1—样本 16)、碱金属硫化物及多硫化物(样本 17—样本 29)和碱金属含氧酸盐(样本 30—样本 45)分段拟合为如下回归方程^[13](以下合称为 MLR2):

$$\chi_m = 43.039 - 14.700 {}^0 F - 64.604 {}^1 F, \quad (19)$$

$$\chi_m = 57.912 - 24.639 {}^0F - 41.437 {}^1F, \quad (20)$$

$$\chi_m = 46.983 - 15.043 {}^0F - 42.382 {}^1F. \quad (21)$$

为便于与 SVR-9FCV 的预测结果进行直接比较 MLR1 与 MLR2 的拟合结果也列入表 1.

表 1 45 种碱金属化合物的磁性连接性指数与摩尔磁化率的实测值、MLR 和 SVR 计算值及误差

样本编号	化合物	0F	1F	$-\chi_m/10^{-12} \text{ m}^3 \cdot \text{mol}^{-1}$						
				实测值	MLR1 预测值	MLR1 预测误差/%	MLR2 预测值	MLR2 预测误差/%	SVR-9FCV 预测值	SVR-9FCV 预测误差/%
1	NaF	1.391	0.606	16.4	15.2	-7.32	16.6	1.22	16.4	0.00
2	KF	1.707	0.692	23.6	29.4	24.58	26.8	13.56	23.4	-0.84
3	RbF	1.916	0.744	31.9	38.6	21.00	33.2	4.08	35.6	11.60
4	CsF	2.133	0.794	44.5	48.1	8.09	39.6	-11.01	44.0	-1.11
5	NaCl	1.599	0.764	30.2	27.7	-8.28	29.8	-1.32	28.1	-6.94
6	KCl	1.915	0.872	38.8	42.5	9.54	41.4	6.70	39.6	2.06
7	RbCl	2.124	0.937	46.4	52.2	12.50	48.7	4.96	49.5	6.68
8	CsCl	2.341	1.000	56.7	62.0	9.35	56.0	-1.23	61.5	8.47
9	NaBr	1.749	0.859	41.1	36.1	-12.17	38.2	-7.06	35.5	-13.62
10	KBr	2.064	0.981	49.2	51.3	4.27	50.7	3.05	49.7	1.02
11	RbBr	2.274	1.054	56.4	61.2	8.51	58.5	3.72	58.8	4.26
12	CsBr	2.491	1.125	67.2	71.4	6.25	66.3	-1.34	69.1	2.83
13	NaI	2.006	1.003	57.0	49.9	-12.46	51.2	-10.18	48.2	-15.23
14	KI	2.322	1.145	63.8	65.8	3.13	65.1	2.04	62.4	-2.18
15	RbI	2.531	1.230	72.2	76.0	5.26	73.6	1.94	72.3	0.14
16	CsI	2.748	1.313	82.6	86.5	4.72	82.2	-0.48	83.9	1.57
17	Na ₂ S	1.860	0.904	39.0	41.5	6.41	41.4	5.38	40.9	4.87
18	K ₂ S	2.306	1.032	60.0	61.7	2.83	63.1	5.17	58.6	-2.32
19	Rb ₂ S	2.602	1.109	80.0	74.9	-6.38	77.2	-3.50	78.4	-1.99
20	Cs ₂ S	2.910	1.184	104.0	88.5	-14.90	91.8	-11.73	103.5	-0.47
21	Na ₂ S ₂	2.023	1.297	53.0	59.5	12.26	57.9	9.25	59.5	12.26
22	K ₂ S ₂	2.470	1.426	71.0	79.8	12.39	79.6	12.11	70.3	-0.98
23	Rb ₂ S ₂	2.765	1.503	90.0	93.0	3.33	93.7	4.11	81.8	-9.10
24	Na ₂ S ₃	2.148	1.461	68.0	69.1	1.62	67.1	-1.32	68.6	0.88
25	K ₂ S ₃	2.595	1.589	80.0	89.4	11.75	88.8	11.00	82.4	3.00
26	Na ₂ S ₄	2.254	1.586	84.0	76.9	-8.45	74.6	-11.19	83.5	-0.59
27	K ₂ S ₄	2.700	1.714	89.0	97.1	9.10	96.2	8.09	90.9	2.13
28	Na ₂ S ₅	2.347	1.692	99.0	83.5	-15.66	81.0	-18.18	95.1	-3.93
29	K ₂ S ₅	2.793	1.819	98.0	103.7	5.82	102.6	4.69	98.2	0.20
30	Na ₂ SO ₄	2.093	1.165	52.0	58.0	11.54	59.2	13.85	57.1	9.81
31	K ₂ SO ₄	2.540	1.265	67.0	77.5	15.67	79.7	18.96	72.7	8.51
32	Rb ₂ SO ₄	2.836	1.325	88.4	90.1	1.92	93.1	5.32	89.1	0.79
33	Cs ₂ SO ₄	3.143	1.384	116.0	103.1	-11.12	107.0	-7.76	116.5	0.43
34	NaClO ₄	1.632	0.936	37.6	34.2	-9.04	36.3	-3.46	40.6	7.98
35	KClO ₄	1.947	1.007	47.4	47.9	1.05	50.7	6.96	49.5	4.43
36	CsClO ₄	2.374	1.091	69.9	66.0	-5.58	70.0	0.14	65.2	-6.71
37	NaClO ₃	1.611	0.934	34.7	33.3	-4.03	35.3	1.73	36.5	5.19
38	KClO ₃	1.926	1.005	42.8	47.0	9.81	49.8	16.36	47.0	9.81
39	CsClO ₃	2.353	1.089	65.0	65.2	0.31	69.1	6.31	63.9	-1.68
40	NaBrO ₃	1.652	0.988	44.2	36.5	-17.42	37.9	-14.25	44.1	-0.22
41	KBrO ₃	1.967	1.059	52.6	50.2	-4.56	52.3	-0.57	49.9	-5.12
42	CsBrO ₃	2.394	1.143	75.1	68.4	-8.92	71.7	-4.53	66.9	-10.91
43	NaIO ₃	1.695	1.039	53.0	39.6	-25.28	40.5	-23.58	47.6	-10.18
44	KIO ₃	2.010	1.110	63.1	53.3	-15.53	54.9	-13.00	53.2	-15.68
45	CsIO ₃	2.437	1.194	83.1	71.5	-13.96	74.3	-10.59	68.8	-17.42

4.2. 结果分析与讨论

从表 1 可以看出,MLR1,MLR2 和 SVR-9FCV 的回归结果的 $E_{\max AP}$ 依次递减,分别为 25.28%, 23.58% 和 17.42%.

表 2 是 MLR1,MLR2 和 SVR-9FCV 的回归结果的 $E_{\max AP}$, $E_{\text{mean A}}$, $E_{\text{mean AP}}$, E_{RMS} 和 r^2 的统计结果.从表 2 可以看出,MLR2 拟合的 $E_{\max AP}$, $E_{\text{mean A}}$, $E_{\text{mean AP}}$,

E_{RMS} 均比 MLR1 的相应值要小,并且 MLR2 的相关系数 $r^2 = 0.9430$ 也比 MLR1 的 $r^2 = 0.9261$ 要大.这说明分段多元线性回归模型的拟合效果优于整段多元线性回归模型的拟合效果,同时表明 45 个样本数据的 χ_m 与 0F 和 1F 线性关联性不及分段样本的线性关联性强,因此有必要寻求和应用多元非线性回归方法进行拟合.

表 2 不同回归模型下碱金属化合物摩尔磁化率的拟合结果

模型	$E_{\max AP}/\%$	$E_{\text{mean A}}/10^{-12} \text{ m}^3 \cdot \text{mol}^{-1}$	$E_{\text{mean AP}}/\%$	$E_{\text{RMS}}/10^{-12} \text{ m}^3 \cdot \text{mol}^{-1}$	r^2
MLR1 ^[13]	25.28	5.64	9.42	6.82	0.9261
MLR2 ^[13]	23.58	4.62	7.28	6.14	0.9430
SVR-9FCV	17.42	3.04	5.25	4.32	0.9503

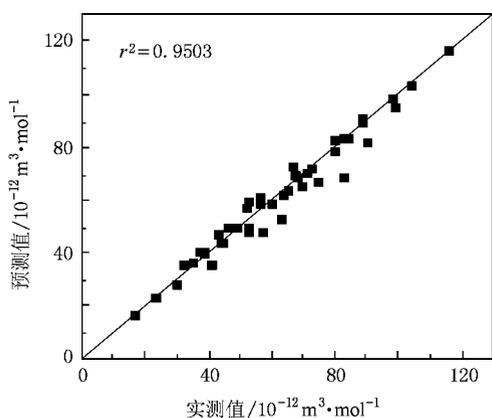


图 1 45 种碱金属化合物摩尔磁化率 χ_m 的 SVR-9FCV 预测值与实测值对比

应用径向基函数的 SVR 方法就是一种基于机器学习的多元非线性回归方法.从表 2 还可以看出,SVR-9FCV 拟合结果的 $E_{\max AP}$, $E_{\text{mean A}}$, $E_{\text{mean AP}}$ 和 E_{RMS} 等误差均比 MLR1 和 MLR2 模型拟合的相应误差要小,SVR 拟合的相关系数 $r^2 = 0.9503$ 也分别比 MLR1 的 $r^2 = 0.9261$ 或 MLR2 的 $r^2 = 0.9430$ 要大.

图 1 给出了 45 种碱金属化合物摩尔磁化率 χ_m 的 SVR-9FCV 预测值与实验测量值的对比.从图 1 可以看出,绝大多数样本的 SVR-9FCV 预测值与实验测量值符合较好.

以上结果表明,SVR 的回归模型的预测能力优于传统的线性回归方法.

5. 结 论

基于分子连接性指数^mF,应用 SVR 方法并结合 PSO 算法,对碱金属化合物的摩尔磁化率 χ_m 进行了预测研究,并与 MLR 方法的拟合结果进行了比较.结果表明(1)零阶和一阶连接性指数与 45 种碱金属化合物的摩尔磁化率 χ_m 具有良好的关联性,可有效地用于预测碱金属化合物的摩尔磁化率 χ_m ; (2)SVR 方法的拟合结果和预测性能明显地优于传统的 MLR 方法.因此,SVR 是一种能够预测碱金属化合物的摩尔磁化率 χ_m 的有效方法.

[1] Gouy L G 1889 *Comput. Rend.* **109** 935

[2] Cusack N 1958 *The Electrical and Magnetic Properties of Solids* (London: Longmans)

[3] Mulay L N, Keys L K 1964 *Anal. Chem.* **36** 2383

[4] Quincke G 1885 *Ann. D. Physik* **260** 347

[5] Bockris J O'M, Parsons D F 1953 *J. Sci. Instrum.* **30** 362

[6] Diaconu M, Puscasu R, Stancu A 2003 *J. Optoelectron. Adv. Mater.* **5** 971

[7] Purss M B J, Cull J F 2005 *Geophysics* **70** L53

[8] Mu L L, Feng C J, He H M 2008 *Ind. Eng. Chem. Res.* **47** 2428

[9] Kaya A, Bain J A 2006 *J. Appl. Phys.* **99** 08B708

[10] Mu L L, Feng C J, He H M 2007 *MATCH-Commun. Math. Comput. Chem.* **58** 591

- [11] Mu L L ,Feng C J ,He H M 2008 *J. Mol. Model.* **14** 109
- [12] Bennet A J ,Xu J M 2003 *Appl. Phys. Lett.* **82** 3304
- [13] Feng C J ,Yang W H 2005 *Chem. Res.* **16** 88 (in Chinese)[冯长君、杨伟华 2005 化学研究 **16** 88]
- [14] Vapnik V 1995 *The Nature of Statistical Learning Theory* (New York : Springer)
- [15] Burges C J C 1998 *Data Min. Knowl. Disc.* **2** 121
- [16] Cai J W ,Hu S S ,Tao H F 2007 *Acta Phys. Sin.* **56** 6820 (in Chinese)[蔡俊伟、胡寿松、陶洪峰 2007 物理学报 **56** 6820]
- [17] Cai C Z ,Han L Y ,Ji Z L ,Chen Y Z 2004 *Proteins* **55** 66
- [18] Cai C Z ,Han L Y ,Ji Z L ,Chen X ,Chen Y Z 2003 *Nucl. Acids Res.* **31** 3692
- [19] Song M H ,Breneman C M ,Bi J B ,Sukumar N ,Bennett K P ,Cramer S ,Tugcu N 2002 *J. Chem. Inform. Comput. Sci.* **42** 1347
- [20] Yuan Z 2005 *BMC Bioinform.* **6** 248
- [21] Cai C Z ,Wang W L ,Chen Y Z 2003 *Int. J. Mod. Phys. C* **14** 575
- [22] Zhang J S ,Hu S S 2008 *Acta Phys. Sin.* **57** 2708 (in Chinese)[张军峰、胡寿松 2008 物理学报 **57** 2708]
- [23] Nguyen M N ,Rajapakse J C 2006 *Proteins* **63** 542
- [24] Cai C Z ,Wang W L ,Sun L Z ,Chen Y Z 2003 *Math. Biosci.* **185** 111
- [25] Camps-Valls G ,Bruzzone L ,Rojo-Alvarez J L ,Melgani F 2006 *IEEE Geosci. Remote. Sens. Lett.* **3** 339
- [26] Lee M ,Jeong C S ,Kim D 2007 *BMC Bioinform.* **8** 471
- [27] Kennedy J ,Eberhart R 1995 *Proc. IEEE Int. Conf. Neural Networks* **4** 1942

Topological research on the molar magnetic susceptibility of alkali metal compounds with support vector regression^{*}

Cai Cong-Zhong[†] Zhuang Wei-Ping Wen Yu-Feng Zhu Xing-Jian Pei Jun-Fang Xiao Ting-Ting

(Department of Applied Physics , Chongqing University , Chongqing 400044 , China)

(Received 15 January 2009 ; revised manuscript received 30 April 2009)

Abstract

According to the experimental dataset on the molar magnetic susceptibility χ_m of 45 alkali metal compounds and the topological descriptor——magnetic connectivity index mF , which is extracted by the magnetic valence g_i of simple ion deduced from classical electrodynamics, support vector regression (SVR) combined with particle swarm optimization for its parameter optimization is proposed to establish a model for predicting the molar magnetic susceptibility of alkali metal compound via 0F and 1F . The performance of SVR model is compared with that of multivariate linear regression (MLR) model. The results show that the mean absolute error, the mean absolute percentage error and the root mean square error for 9-fold cross validation test of SVR models are all smaller than those achieved by MLR models. It is revealed that the generalization ability of SVR model is superior to that of MLR model. This study suggests that magnetic connectivity index is an effective descriptor and the SVR is a powerful approach to the prediction of the molar magnetic susceptibility of alkali metal compounds.

Keywords : alkali metal compounds , molar magnetic susceptibility , support vector regression , prediction

PACC : 7530C , 0270 , 0650

^{*} Project supported by the Program for the New Century Excellent Talents in University of Ministry of Education , China (Grant No. NCET-07-0903) , the Scientific Research Foundation for the Returned Overseas Chinese Scholars of Ministry of Education , China (Grant No. 2008101-1) , the Natural Science Foundation of Chongqing , China (Grant No. CSTC2006BB5240) and the National Innovation Experimental Program for Undergraduate Students in China (Grant No. CQUCX-G-2007-016).

[†] E-mail : caiczh@gmail.com