基于遗传-支持向量机和遗传-径向基神经网络的有机物正辛醇-水分配系数 QSPR 研究

齐珺，牛军峰*，王丽莉
（北京师范大学环境学院水环境模拟国家重点实验室，北京 100875）

摘要：基于遗传算法（GA）的因子筛选和支持向量机（SVM）的非线性回归，提出了 1 种改进的有机物定量结构-性质相关（QSPR）建模方法——遗传-支持向量机（GA-SVM），并将其用于 38 种食品工业常用有机物正辛醇-水分配系数（K_{ow}）的 QSPR 建模．结果显示，QSPR 模型选取了分子量、Hansen 极性、沸点、含氧率和含氢率 5 种参数；模型的预测值与实测值间的误差平方和（SSE）、均方差（RMSE）和决定系数（R^2）分别为 0.048、0.036 和 0.999，表明模型具有较强的预测能力；同时，交叉验证的结果（SSE = 0.295，RMSE = 0.089，R^2 = 0.995）也表明，模型具有良好的稳健性，因此，GA-SVM 算法适用于对有机物正辛醇-水分配系数的 QSPR 建模．此外，将基于 GA-SVM 的 QSPR 模型分别与基于遗传-径向基神经网络（GA-RBFNN）和基于线性算法的模型进行了比较，结果表明，应用 GA-SVM 建立的 QSPR 模型无论从稳健性还是预测能力上都优于应用其它 2 种算法建立的模型，因此，GA-SVM 算法比 GA-RBFNN 和线性算法更适合于对有机物正辛醇-水分配系数进行 QSPR 建模．

关键词：定量结构-性质相关（QSPR）；正辛醇-水分配系数（K_{ow}）；遗传算法（GA）；支持向量机（SVM）
中图分类号：X131 **文献标识码**：A **文章编号**：0250-3301（2008）01-0212-07

Research on QSPR for *n*-Octanol-Water Partition Coefficients of Organic Compounds Based on Genetic Algorithms-Support Vector Machine and Genetic Algorithms-Radial Basis Function Neural Networks

QI Jun, NIU Jun-feng, WANG Li-li
(State Key Laboratory of Water Environment Simulation, School of Environment, Beijing Normal University, Beijing 100875, China)

Abstract: A modified method to develop quantitative structure-property relationship (QSPR) models of organic compounds was proposed based on genetic algorithm (GA) and support vector machine (SVM) (GA-SVM). GA was used to perform the variable selection, and SVM was used to construct QSPR models. GA-SVM was applied to develop the QSPR models for *n*-octanol-water partition coefficients (K_{ow}) of 38 typical organic compounds in food industry. 5 descriptors (molecular weights, Hansen polarity, boiling point, percent oxygen and percent hydrogen) were selected in the QSPR model. The coefficient of multiple determination (R^2), the sum of squares due to error (SSE) and the root mean squared error (RMSE) values between the measured values and predicted values of the model developed by GA-SVM are 0.999, 0.048 and 0.036, respectively, indicating good predictive capability for lgK_{ow} values of these organic compounds. Based on leave-one-out cross validation, the QSPR model constructed by GA-SVM showed good robustness (SSE = 0.295, RMSE = 0.089, R^2 = 0.995). Moreover, the models developed by GA-SVM were compared with the models constructed by genetic algorithm-radial basis function neural network (GA-RBFNN) and linear method. The models constructed by GA-SVM show the optimal predictive capability and robustness in the comparison, which illustrates GA-SVM is the optimal method for developing QSPR models for lgK_{ow} values of these organic compounds.

Key words: quantitative structure-property relationship (QSPR); *n*-octanol-water partition coefficients (K_{ow}); genetic algorithms (GA); support vector machine (SVM)

食物与包装材料间多种相互作用会引起食物的污染．源自包装材料的挥发物、添加剂、单体和低聚物可以迁移到食物中，聚合物也会吸附食物中挥发化合物，这些都会影响食物质量，直接威胁人类健康[1~3]．研究表明，有机物在食物与包装材料间的迁移与有机物的正辛醇-水分配系数（K_{ow}）密切相关[4~7]，这使得 K_{ow} 成为评价有机物环境行为的 1 个重要理化参数．然而，传统测定 K_{ow} 的摇瓶法、慢搅拌法和产生柱法虽然简单易行，但测定化合物费时，对脂溶性的化合物测定误差较大；采用反相液相色谱法间接测定 K_{ow} 的方法虽然快速、准确，但是要求参考物的 K_{ow} 测定值必须准确，否则可能造成所测定化合物的 K_{ow} 值误差较大[8]．因此建立有机物 K_{ow}

收稿日期：2007-02-06；**修订日期**：2007-04-13
基金项目：国家重点基础研究发展规划（973）项目（2003CB415204）
作者简介：齐珺（1979～），男，博士研究生，主要研究方向为有机污染物的环境行为及模拟预测，E-mail：qijun@mail.bnu.edu.cn
* 通讯联系人，E-mail：junfengn@bnu.edu.cn

的定量结构-性质相关(QSPR)模型显得尤为重要.

应用最广的 QSPR 模型是线性模型,简单方便,但在某些问题上预测能力稍差[9,10];随着智能算法的普遍应用,神经网络等智能算法被用于建立非线性 QSPR 模型[11,12];近年来,偏最小二乘法回归(PLS)[13~15]、支持向量机(SVM)[16,17]等新算法在 QSPR 建模中得到了广泛应用;目前,应用遗传-偏最小二乘法(GA-PLS)、决策树-支持向量机(DT-SVM)等组合方法建模成为 QSPR 建模中新的热点之一[18,19].遗传算法(GA)以概率选择为主要手段,全局搜索能力强,被广泛地应用于参数优化及因子筛选[10,20,21].同时,SVM 在小样本、非线性、高维回归问题方面表现突出[16,17].已有报道表明,以往的研究多集中在利用 GA 优化 SVM 的参数,而将基于 GA 因子筛选的 SVM 方法用于有机物 K_{ow} 的 QSPR 建模方面的研究鲜见.因此,本研究以组成参数和容易计算得到且具有明确物理-化学意义的量子化学参数[8,22,23]为自变量,分别采用遗传-支持向量机(GA-SVM)和遗传-径向基神经网络(GA-RBFNN)算法,建立有机物 $\lg K_{ow}$ 的 QSPR 模型.

1　材料与方法

1.1　方法原理

1.1.1　GA-SVM 方法原理

SVM 是 Vapnik 等[24]根据统计学理论提出的一种新的学习方法.假设样本数据集为(x_1, y_1),(x_2, y_2),……,(x_m, y_m);其中,x_k 为第 k 个样本的自变量因子集合(n 维向量),y_k 为第 k 个样本的因变量,即实测值,$k = 1, 2, \cdots, m$,m 为样本总量.SVM 回归的原理就是利用回归超平面 $y = (w \cdot x) + b$ 最佳拟合空间中的样本.采用 Vapnik 不敏感损失函数,使该平面在给定精度 ε 下放宽限制而允许有一定误差存在,引入稀疏变量和惩罚参数,并利用 Lagrang 函数,将原拟合问题转化为相应的对偶问题.并通过引入核函数 K,实现变换空间内积运算,求解得到决策函数:

$$f(x) = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)K(x \cdot x_i) + b^* \qquad (1)$$

式中,α_i 和 α_i^* 为 Lagrange 乘子,依据 karush-kuhn-tucker(KKT)的二次规划条件限制,只有少数样本的系数 α_i 为非 0 值,对应的数据点被称为支持向量,它们决定该超平面的函数[25].

GA 是 1 种借鉴生物界自然选择和进化机制发展起来的高度并行、随机、自适应搜索算法,它克服了传统优化方法容易陷入局部极值的特点,是 1 种全局优化算法.本研究中 GA 采用二进制 0、1 编码,如图 1 所示,自变量 X 是 1 个 $m \times n$ 维矩阵,每个个体含有 n 个二进制代码,分别与 $S_1 \sim S_n$ 分子结构参数对应.染色体为"1"代表其对应的结构参数被选中,"0"表示未被选中.图 1 所示,由第 i 个个体的染色体结构可以确定其对应的自变量 X_i.

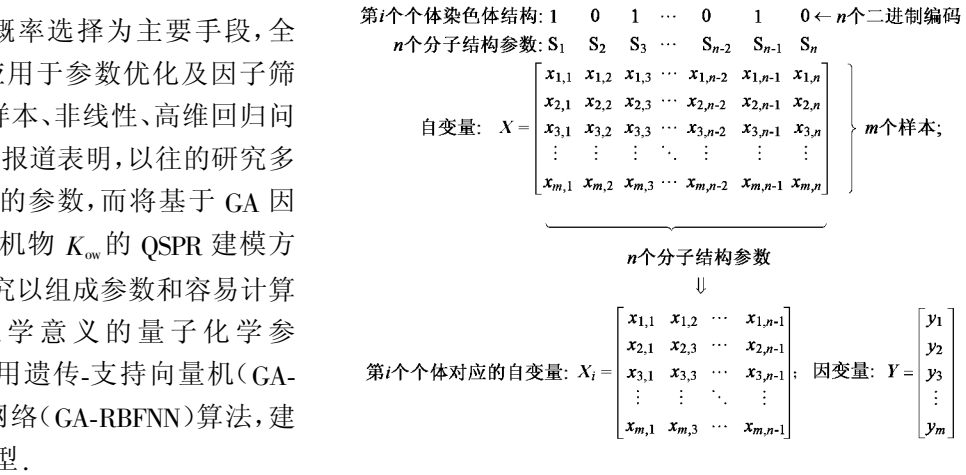

图 1　个体染色体结构示意

Fig.1　Chromosome structure schematic diagram of an individual

在 SVM 建模过程中,选取 $\boldsymbol{X}_i$ 和 $\boldsymbol{Y}$ 中的部分样本作为训练集,分别用矩阵 $\boldsymbol{X}_{i-\text{train}}$,和 $\boldsymbol{Y}_{\text{train}}$ 表示;其余样本作为验证集,分别用矩阵 $\boldsymbol{X}_{i-\text{verify}}$ 和 $\boldsymbol{Y}_{\text{verify}}$ 表示.首先,利用训练集建立 QSPR 模型,再将验证集 $\boldsymbol{X}_{i-\text{verify}}$ 代入模型中,计算模型的预测值,并根据公式(2),求出预测值与实测值间的误差,作为 GA 中第 i 个个体的适应度值.

$$f(y, \hat{y}) = \sum_{i=1}^{l} | y - \hat{y} | \qquad (2)$$

式中,$f(y, \hat{y})$为适应度函数,y 为实测值,$\hat{y}$ 为模型预测值,l 是预测样本数.

图 2 为 GA-SVM 计算流程,其中 Gen 表示当前的遗传代数,MaxGen 表示最大遗传代数.遗传 MaxGen 代后,获得最优个体,将其对应的自变量 X_{best} 通过 SVM 建立 QSPR 模型.

1.1.2　GA-RBFNN 方法原理

径向基函数神经网络(RBFNN)是一种前馈的局部式神经网络,它包括输入层、输出层和 1 个隐含层.隐含层节点由像高斯函数那样的辐射状作用函数构成,因此,从输入层空间到隐含层空间的变换是非线性的,而从隐含层空间到输出层空间的变换通

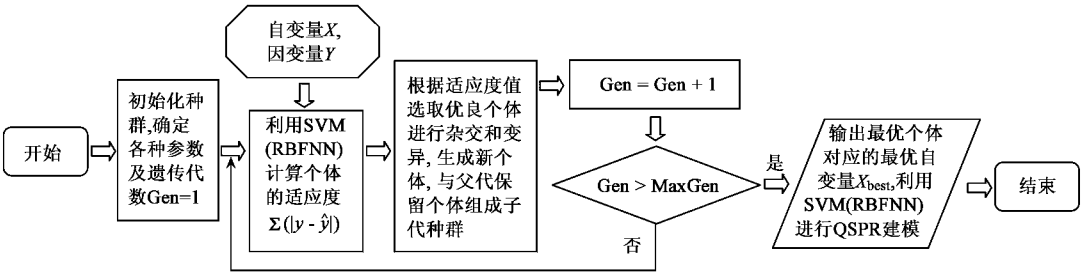

图 2 GA-SVM 算法流程

Fig.2 Flow diagram of GA-SVM

常是线性的,如图 3 所示.与 BP 神经网络相比,RBFNN 不仅学习速度快,而且避免了出现局部极小问题,过拟合现象降低,其推广能力显著增强[26,27].

本研究采用的 GA-RBFNN 方法是将 GA 的因子筛选与 RBFNN 的非线性回归相结合.

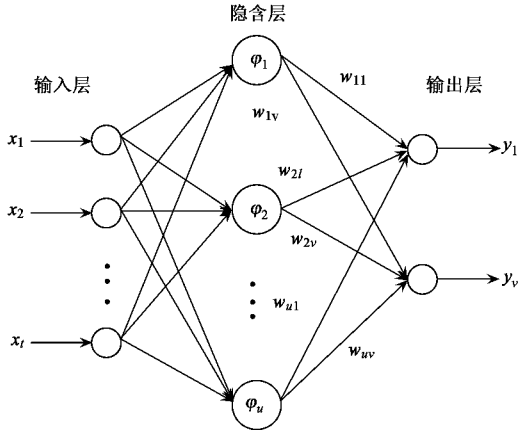

图 3 RBFNN 的拓扑结构

Fig.3 Topological structure of RBF neural network

1.2 研究对象和参数的选取

选取食品工业中常用的 38 种有机化合物作为研究对象,包含芳香烃、醇类、羧酸类、醛类、酮类和酯类等,它们 25℃下的 $\lg K_{ow}$ 值及组成参数和量子化学参数取自文献[1],如表 1 所示.

1.3 统计分析

本研究中 SVM 采用 Matlab LibSVM 工具箱[28],选取 Epsilon-SVR 模式,利用公式(3)所示的径向基(RBF)函数作为核函数.

$$K(x, x_i) = \exp\left(-\frac{\| x - x_i \|}{2\sigma^2}\right) \quad (3)$$

选取误差概率为 0.1,C 值为 100,损失函数 ε 系数为 0.001.RBFNN 采用 Matlab 的 Neural Network 工具箱中的广义回归网络(GRNN)[26],其中输入层节点数由 GA 筛选的因子个数决定,输出层节点数为 1,隐含层节点数由程序确定.GA 采用 Matlab 的 Genetic Algorithm 工具箱,采用二进制编码,更新系数为 0.9;杂交系数为 0.5;变异概率为 0.3;最大遗传代数为 100 代;种群个体数为 50 个.

为了验证和比较模型,对预测结果分别进行误差分析和相关性分析,计算预测值与实测值间的决定系数(R^2)、误差平方和(SSE)和均方差(RMSE)如公式(4)和(5)所示.

$$\mathrm{SSE} = \sum_{i=1}^{l} (y_i - \hat{y}_i)^2 \quad (4)$$

$$\mathrm{RMSE} = \sqrt{\frac{\mathrm{SSE}}{l-1}} = \sqrt{\frac{\sum_{i=1}^{l} (y_i - \hat{y}_i)^2}{l-1}} \quad (5)$$

式中,l 是预测样本数.

2 结果与讨论

2.1 GA-SVM 模型

以表 1 中 38 种有机物的 $\lg K_{ow}$ 值作为因变量,以 7 种组成参数和量子化学参数作为自变量,利用 GA-SVM 建立 QSPR 模型.通过 GA 进行变量筛选得到了包含分子量、Hansen 极性、沸点、含氧率和含氢率 5 个参数的最优个体,并建立了最优 QSPR 模型,模型对这 38 种有机物 $\lg K_{ow}$ 值的预测值列于表 2 中.计算得到 GA-SVM 建立的 QSPR 模型的 R^2、SSE 和 RMSE 分别为 0.999、0.048 和 0.036(表 3),结果表明预测值与实测值间相关关系显著,预测能力较强.

2.2 GA-RBFNN 模型

以表 1 中 38 种有机化合物的 $\lg K_{ow}$ 值作为因变量,以 7 种组成参数和量子化学参数作为自变量,利用 GA-RBFNN 建立 QSPR 模型.通过 GA 进行变量筛选得到了包含所有 7 种参数的最优个体,并建立了

表 1　38 种有机化合物的正辛醇-水分配系数(K_{ow})和参数[1]

Table 1　K_{ow} values and descriptors to 38 organic compounds

序号	有机化合物	分子量	Hansen 极性 /delta·MPa$^{-1/2}$	分子总能量 /a.u.	沸点 /℃	结合能 /a.u.	含氧率	含氢率	lgK_{ow} 实测值
1	甲醇	32.04	12.28	-28.56	54.02	-1.76	49.93	12.58	-0.77
2	乙醇	46.06	8.80	-37.25	78.89	-3.01	34.74	13.12	-0.31
3	1-丙醇	60.09	6.87	-45.94	97.12	-4.26	26.62	13.41	0.25
4	1-丁醇	74.12	5.65	-54.63	116.92	-5.50	21.58	13.59	0.88
5	1-戊醇	88.11	4.49	-63.32	138.63	-6.75	18.15	13.72	1.56
6	1-己醇	102.17	3.91	-72.01	160.64	-8.00	15.65	13.81	2.03
7	1-庚醇	116.20	3.59	-80.70	182.29	-9.24	13.76	13.87	2.38
8	1-辛醇	130.23	3.23	-89.39	160.64	-10.49	12.28	13.93	2.88
9	2-甲基-3-戊醇	102.17	6.05	-72.00	135.83	-7.99	13.81	70.53	2.03
10	甲醛	30.02	18.76	-26.83	-19.50	-1.30	53.28	6.71	0.35
11	丙醛	58.08	6.34	-44.22	47.43	-3.81	27.54	10.41	0.59
12	丁醛	72.10	5.27	-52.91	77.41	-5.06	22.18	11.18	0.88
13	戊醛	86.13	4.45	-61.60	103.06	-6.31	18.57	11.70	1.31
14	己醛	100.16	3.86	-70.29	128.00	-7.55	15.97	12.07	1.78
15	庚醛	114.18	3.40	-78.98	151.93	-8.80	12.47	12.35	2.42
16	丙酮	58.08	10.39	-44.24	56.21	-3.83	27.54	10.41	-0.24
17	2-丁酮	72.10	9.01	-52.93	73.11	-5.08	22.18	11.18	0.29
18	2-戊酮	86.13	7.73	-61.62	102.72	-6.32	18.57	11.70	0.91
19	2-己酮	100.16	6.77	-70.31	128.08	-7.57	15.97	12.07	1.38
20	2-庚酮	114.18	5.96	-78.99	152.43	-8.81	14.01	12.35	1.98
21	2-辛酮	128.21	5.33	-87.68	175.67	-10.06	12.47	12.57	2.37
22	乙酸	60.05	7.91	-54.01	104.90	-2.96	53.28	6.71	-0.17
23	乙酸甲酯	74.07	7.24	-62.68	35.82	-4.19	43.19	8.16	0.18
24	乙酸乙酯	88.10	5.35	-71.38	55.97	-5.44	31.33	9.15	0.73
25	乙酸异丙酯	102.13	4.65	-80.07	73.18	-6.69	36.31	9.86	1.21
26	癸酸	172.26	5.25	-123.52	261.16	-12.93	18.57	11.70	3.91
27	肉桂酸甲酯	162.18	9.12	-115.32	222.63	-11.12	19.72	6.21	2.79
28	柠檬烯	136.23	0.97	-83.47	181.85	-11.60	37.46	11.83	4.20
29	丁内酯	86.09	16.59	-69.94	204.56	-5.29	37.16	7.02	-0.64
30	异戊醇	88.14	4.55	-63.31	135.38	-6.74	18.15	13.72	1.42
31	苯甲酸甲酯	136.15	9.77	-99.68	176.04	-9.08	23.50	5.92	2.10
32	乙酸异戊酯	130.18	3.27	-97.44	130.08	-9.18	24.57	10.83	2.13
33	乙醛	44.05	8.00	-35.53	19.66	-2.57	36.31	9.15	0.43
34	2-甲基丙醛	72.10	5.28	-52.91	62.81	-5.06	22.18	11.18	1.48
35	双乙酰	86.09	13.42	-69.66	141.74	-5.00	37.16	7.02	-1.50
36	醋酸丁酯	116.15	3.63	-88.76	110.33	-7.94	27.54	10.41	1.00
37	香草醛	152.14	9.90	-118.12	230.80	-9.44	31.54	5.29	1.23
38	苯甲酸乙酯	150.17	8.60	-108.37	190.96	-10.33	21.30	6.71	1.84

最优 QSPR 模型,模型的预测值列于表 2 中.计算得到 GA-RBFNN 建立的 QSPR 模型的 R^2、SSE 和 RMSE 分别为 0.997、0.194 和 0.072(表 3),结果表明预测值与实测值间相关关系显著,预测能力较强.

2.3　模型交叉验证

本研究应用留一法(leave-one-out)交叉验证来检验和比较 QSPR 模型的稳健性.分别利用 GA-SVM 和 GA-RBFNN 算法对每一种有机物的 lgK_{ow} 值进行 QSPR 建模预测,交叉验证的预测结果列于表 2 中,并绘制预测值与实测值的相关图,如图 4 所示.

图 4(a)是由 GA-SVM 建立模型的 lgK_{ow} 预测值与实测值的相关图,其中显示预测值与实测值间的相关性显著,误差较小,说明由 GA-SVM 建立的 QSPR 模型稳健性较好.

图 4(b)是由 GA-RBFNN 建立模型的 lgK_{ow} 预测值与实测值的相关图,其中除了点 1 和 2 外其它点的相关性较好.点 1 和 2 分别对应表 1 中 28 号柠檬烯和 35 号双乙酰,这 2 种有机化合物的 lgK_{ow} 值(4.20 和 -1.50)分别是 38 种有机化合物中最大和最小的,这说明由 GA-RBFNN 建立的 QSPR 模型对训练数据集区间内的有机化合物的 lgK_{ow} 值预测比较准确,而对区间外有机化合物的 lgK_{ow} 值预测较

表 2 QSPR 模型 $\lg K_{ow}$ 预测结果比较

Table 2 Comparison of predicted values of $\lg K_{ow}$ by QSPR models

序号	$\lg K_{ow}$ 实测值	全部数据建模						留一法交叉验证			
		线性模型[1]		GA-SVM		GA-RBFNN		GA-SVM		GA-RBFNN	
		预测值	残差	预测值	残差	预测值	残差	预测值	残差	预测值	残差
1	-0.77	0.46	-1.23	-0.77	0.00	-0.77	0.00	-0.76	-0.01	-1.01	0.24
2	-0.31	0.23	-0.54	-0.29	-0.02	-0.30	-0.01	-0.31	0.00	-0.07	-0.24
3	0.25	0.22	0.03	0.22	0.03	0.30	-0.05	0.25	0.00	0.28	-0.03
4	0.88	0.72	0.16	0.88	0.00	0.94	-0.06	0.88	0.00	0.87	0.01
5	1.56	1.31	0.25	1.50	0.06	1.50	0.06	1.55	0.01	1.56	0.00
6	2.03	1.88	0.15	2.00	0.03	2.01	0.02	2.03	0.00	2.03	0.00
7	2.38	2.45	-0.07	2.38	0.00	2.38	0.00	2.38	0.00	2.38	0.00
8	2.88	3.39	-0.51	2.88	0.00	2.78	0.10	2.87	0.01	2.69	0.19
9	2.03	1.82	0.21	2.09	-0.06	2.03	0.00	2.03	0.00	2.05	-0.02
10	0.35	0.28	0.07	0.35	0.00	0.35	0.00	0.36	-0.01	0.34	0.01
11	0.59	0.75	-0.16	0.60	-0.01	0.60	-0.01	0.59	0.00	0.59	0.00
12	0.88	1.01	-0.13	0.87	0.01	1.11	-0.23	0.88	0.00	0.88	0.00
13	1.31	1.42	-0.11	1.28	0.03	1.33	-0.01	1.31	0.00	1.31	0.00
14	1.78	1.87	-0.09	1.76	0.02	1.74	0.04	1.78	0.00	1.79	-0.01
15	2.42	2.38	0.04	2.41	0.01	2.34	0.08	2.42	0.00	2.29	0.13
16	-0.24	0.29	-0.53	-0.25	0.01	-0.23	-0.01	-0.24	0.00	-0.17	-0.07
17	0.29	0.69	-0.4	0.31	-0.02	0.31	-0.02	0.29	0.00	0.30	-0.01
18	0.91	1.09	-0.18	0.88	0.03	0.93	-0.02	0.91	0.00	0.92	-0.01
19	1.38	1.60	-0.22	1.46	-0.08	1.42	-0.04	1.32	0.06	1.40	-0.02
20	1.98	2.13	-0.15	1.97	0.01	2.03	-0.05	1.98	0.00	1.98	0.00
21	2.37	2.71	-0.34	2.38	-0.01	2.41	-0.04	2.37	0.00	2.37	0.00
22	-0.17	-0.79	0.62	-0.16	-0.01	-0.17	0.00	-0.18	0.01	-0.16	-0.01
23	0.18	0.02	0.16	0.21	-0.03	0.18	0.00	0.18	0.00	0.19	-0.01
24	0.73	0.59	0.14	0.69	0.04	0.75	-0.02	0.74	-0.01	0.77	-0.04
25	1.21	1.20	0.01	1.18	0.03	1.20	0.01	1.18	0.03	1.19	0.02
26	3.91	4.04	-0.13	3.91	0.00	3.91	0.00	3.89	0.02	3.89	0.02
27	2.79	2.26	0.53	2.79	0.00	2.72	0.07	2.62	0.17	2.78	0.01
28	4.20	4.67	-0.47	4.20	0.00	4.20	0.00	4.23	-0.03	2.88	1.32
29	-0.64	-1.26	0.62	-0.62	-0.02	-0.64	0.00	-0.64	0.00	-0.65	0.01
30	1.42	1.30	0.12	1.48	-0.06	1.49	-0.07	1.42	0.00	1.42	0.00
31	2.10	1.54	0.56	2.09	0.01	2.09	0.01	2.13	-0.03	2.10	0.00
32	2.13	2.49	-0.36	2.10	0.03	2.07	0.06	2.15	-0.02	2.14	-0.01
33	0.43	0.68	-0.25	0.43	0.00	0.43	0.00	0.43	0.00	0.44	-0.01
34	1.48	1.12	0.36	1.35	0.13	1.19	0.29	1.10	0.38	1.47	0.01
35	-1.50	-1.05	-0.45	-1.47	-0.03	-1.50	0.00	-1.49	-0.01	-0.64	-0.86
36	1.00	1.79	-0.79	1.02	-0.02	1.06	-0.06	0.99	0.01	0.99	0.01
37	1.23	1.52	-0.29	1.23	0.00	1.23	0.00	1.54	-0.31	0.92	0.31
38	1.84	2.29	-0.45	1.87	-0.03	1.92	-0.08	1.98	-0.14	1.81	0.03

表 3 分别由 GA-SVM、GA-RBFNN 和线性算法建立的 3 个 QSPR 模型的参数选取与误差

Table 3 Selected descriptors and errors of 3 QSPR models constructed by GA-SVM, GA-RBFNN and linear method respectively

建模方法	选取参数							误差		
	分子量	Hansen 极性	分子总能量	沸点	结合能	含氧率	含氢率	R^2	SSE	RMSE
GA-SVM	1	1	0	1	0	1	1	0.999	0.048	0.036
GA-RBFNN	1	1	1	1	1	1	1	0.997	0.194	0.072
线性模型[1]	1	1	1	1	1	1	1	0.904	6.050	0.404

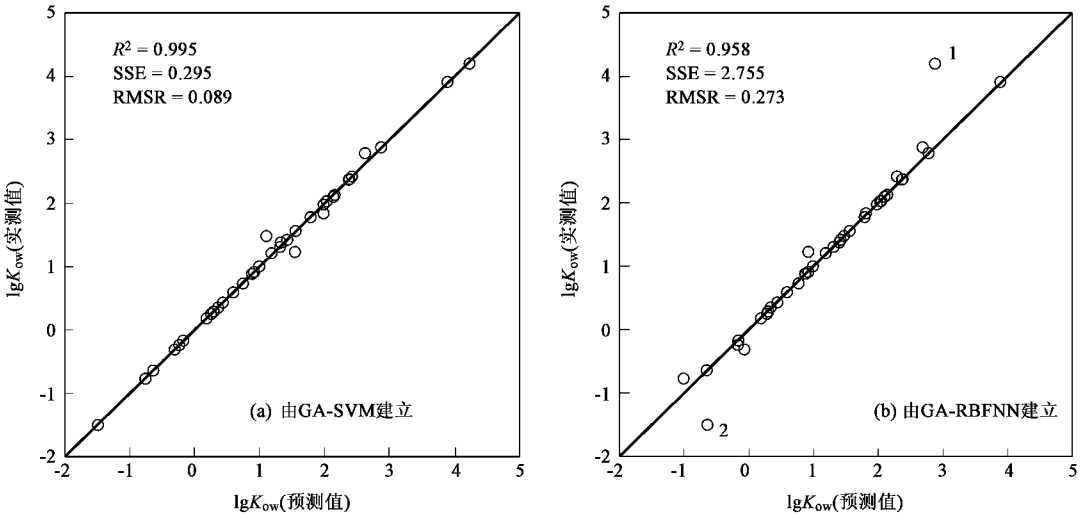

图 4　留一法交叉验证 lgK_{ow} 预测值与实测值的相关图

Fig.4　Plots of predicted values vs. measured values of lgK_{ow} by GA-SVM and GA-RBFNN respectively in leave-one-out cross validation

差.验证结果表明由 GA-RBFNN 建立的模型稳健性较差.

2.4　模型比较

本研究将 GA-SVM 和 GA-RBFNN 算法的计算结果与文献[1]中线性算法的计算结果进行了比较,如图 5 所示.其中,图 5(a)是使用全部数据建立的 3 个 QSPR 模型的误差比较,其中显示,GA-SVM 和 GA-RBFNN 模型的决定系数相差不大,都高于线性模型;同时,GA-SVM 模型的 SSE 和 RMSE 值是 3 个模型中最小的,因此,GA-SVM 模型的预测能力优于其它 2 种模型.图5(b)是交叉验证结果的误差比较.GA-SVM 模型的 SSE 和 RMSE 分别约为 GA-RBFNN 模型的 1/10 和 1/3;而决定系数(0.995)大于 GA-RBFNN 模型(0.958),表明 GA-SVM 模型的稳健性优于 GA-RBFNN 模型.

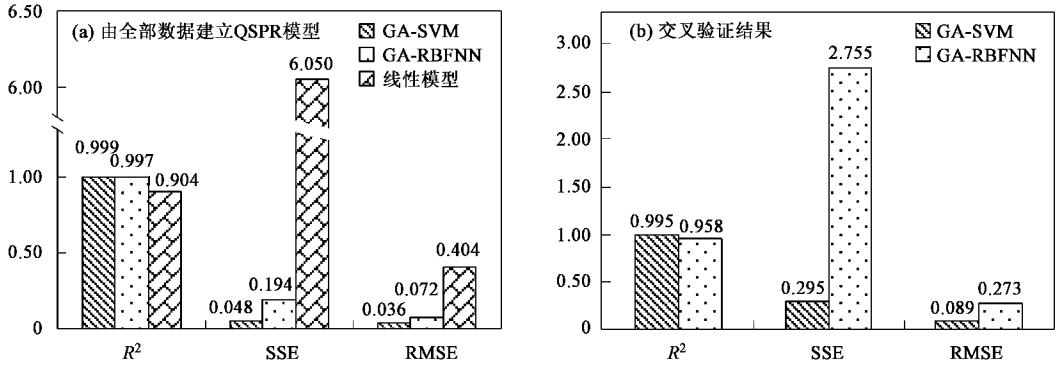

图 5　GA-SVM 和 GA-RBFNN 建立模型与文献[1]中线性模型的误差比较

Fig.5　Comparison plots for the errors among QSPR models developed by GA-SVM, GA-RBFNN and linear method proposed in literature[1] respectively

于 GA-RBFNN 模型.

3　结论

本研究将 GA-SVM 算法用于 38 种有机物正辛醇-水分配系数的 QSPR 建模,误差分析结果表明模型具有较强的预测能力;同时,交叉验证的结果也表明了模型具有良好的稳健性,GA-SVM 算法适用于对有机物正辛醇-水分配系数的 QSPR 建模.此外,通过 GA-SVM 模型预测结果与 GA-RBFNN 和线性模型预测结果的比较表明,在预测能力和稳健性方面,应用 GA-SVM 建立的 QSPR 模型都优于 GA-RBFNN 和线性算法建立的模型.因此,GA-SVM 算法较 GA-RBFNN 和线性算法更适合于对所选取有机物的正辛醇-水分配系数进行 QSPR 建模.

参考文献:

[1]　Tehrany E A, Fournier F, Desobry S. Simple method to

calculate octanol-water partition coefficient of organic compounds [J]. J Food Eng, 2004, **64**: 315-320.

[2] Poças M D F, Hogg T. Exposure assessment of chemicals from packaging materials in foods: a review[J]. Trends Food Sci Tech, 2007, **18**(4): 219-230.

[3] Triantafyllou V I, Akrida-Demertzi k, Demertzis p G. A study on the migration of organic pollutants from recycled paperboard packaging materials to solid food matrices[J]. Food Chem, 2007, **101**(4): 1759-1768.

[4] Durjava M K, Ter Laak T L, Hermens J L, *et al*. Distribution of PAHs and PCBs to dissolved organic matter: High distribution coefficients with consequences for environmental fate modeling[J]. Chemosphere, 2007, **67**(5): 990-997.

[5] 周霞,余刚,黄俊,等. 北京东南郊化工区土壤和植物中氯苯类有机物的残留及分布特征[J]. 环境科学, 2007, **28**(2): 249-254.

[6] Barbour J P, Smith J A, Chiou C T. Sorption of aromatic organic pollutants to grasses from water[J]. Environ Sci Technol, 2005, **39**(21): 8369-8373.

[7] Turner A, Williamson I. On the relationship between D_{ow} and K_{ow} in natural waters[J]. Environ Sci Technol, 2005, **39**(22): 8719-8727.

[8] 隆兴兴,牛军峰,史姝琼. 邻苯二甲酸酯类化合物正辛醇-水分配系数的 QSPR 研究[J]. 环境科学, 2006, **27**(11): 2318-2322.

[9] 王斌,赵劲松,郁亚娟,等. 取代联苯的定量结构活性相关及联合毒性研究[J]. 环境科学, 2004, **25**(3): 89-93.

[10] Gramatica P, Giani E, Papa E. Statistical external validation and consensus modeling: A QSPR case study for K_{oc} prediction[J]. J Mol Graphics Modell, 2007, **25**(6):755-766.

[11] 孙伟,曾光明,魏万之,等. 氯代芳香族化合物结构-电化学还原电位定量关系的贝叶斯规整化 BP 神经网络模型[J]. 环境科学, 2005, **26**(2): 21-27.

[12] 李剑,陈德钊,吴晓华,等. 优化的径向基-循环子空间网络为药物定量构效关系建模[J]. 分析化学, 2005, **33**(6): 767-771.

[13] 黄宏,杨红,樊伟,等. 分子全息 QSAR 方法预测苯衍生物对蝌蚪的急性毒性[J]. 环境科学, 2005, **26**(3): 25-28.

[14] Niu J F, Huang L P, Chen J W, *et al*. Quantitative structure-property relationships on photolysis of PCDD/Fs adsorbed to spruce (*Picea abies* (L.) Karst.) needle surfaces under sunlight irradiation [J]. Chemosphere, 2005, **58**(7): 917-924.

[15] Chen J W, Xue X Y, Schramm K-W, *et al*. Quantitative structure-property relationships for octanol-air partition coefficients of polychlorinated naphthalenes, chlorobenzenes and *p*, *p*′-DDT[J]. Comput Biol Chem, 2003, **27**(3): 165-171.

[16] 李剑,陈德钊,成忠,等. 构建支持向量机-偏最小二乘法为药物构效关系建模[J]. 分析化学, 2006, **34**(2): 263-266.

[17] Liu H X, Yao X J, Zhang R S, *et al*. The accurate QSPR models to predict the bioconcentration factors of nonionic organic compounds based on the heuristic method and support vector machine[J]. Chemosphere, 2006, **63**(5): 722-733.

[18] Lima P D C, Golbraikh A, Oloff S, *et al*. Combinatorial QSAR modeling of P-glycoprotein substrates[J]. J Chem Inf Model, 2006, **46**(3): 1245-1254.

[19] Gieleciak R, Polanski J J. Modeling robust QSAR. 2. iterative variable elimination schemes for CoMSA: application for modeling benzoic acid pK_a values[J]. Chem Inf Model, 2007, **47**(2): 547-556.

[20] Nicolotti O, Carotti A. QSAR and QSPR Studies of a Highly Structured Physicochemical Domain[J]. J Chem Inf Model, 2006, **46**(1): 264-276.

[21] Cho S J, Hermsmeier M A. Genetic Algorithm Guided Selection: Variable Selection and Subset Selection[J]. J Chem Inf Comput Sci, 2002, **42**(4): 927-936.

[22] 牛军峰,余刚,韩文亚. 应用遗传算法建立云杉针叶表面 PCDD/Fs 光解半衰期的预测模型[J]. 环境科学, 2005, **26**(2): 28-33.

[23] Papa E, Dearden J C, Gramatica P. Linear QSAR regression models for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors [J]. Chemosphere, 2007, **62**(2): 351-358.

[24] Vapnik V N. The nature of statistical learning theory[M]. New York: Springer, 1995.

[25] Nello C, John S T. 支持向量机导论(An Introduction to Support Vector Machines and Other Kernel-based Learning Methods)[M]. 北京:电子工业出版社, 2005. 98-107.

[26] Demuth H, Beale M, Hagan M. Neural network toolbox (version 5) user's guide[M]. The MathWorks, Inc, 2006.

[27] 杨晓华,杨志峰,沈珍瑶,等. 区域水资源开发利用程度评价的 RBF 网络模型[J]. 环境科学, 2004, **25**(增刊): 31-34.

[28] Chang C C, Lin C J. LIBSVM: a library for support vector machines [EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html, 2007-01-05.