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

孙伟[1] ,曾光明[1] ,魏万之[2] ,黄国和[1]
(1. 湖南大学环境科学与工程系 ,长沙 410082 ; 2. 湖南大学化学化工学院 ,长沙 410082)

摘要 :将贝叶斯规整化误差反向传播神经网络(BRBPNN) 应用于环境领域的 QSPR 模型 .采用 ChemOffice2004 内置的 MOPAC 2000 计算了 6 种量子化学参数(分子最高占据能 E_{HOMO} 、分子最低占据能 E_{LUMO} 、分子生成热 HF 、分子偶极矩 DIP 、分子的电子能量 EE 和分子的核核排斥能 CCR) 以及氯原子数(Cl) 和分子量(MW) ,建立了 87 种氯代芳香族化合物结构与电化学还原电位定量关系的 BRBPNN 模型 .最优网络模型结构为 6-20-1 ,其电化学还原电位的拟合及预测能力明显优于逐步线性回归模型 ,其训练集和预测集的相关系数平方和均方根误差(MSE) 分别达到 0.999 和 0.000105 ,0.965 和 0.00159 .最优模型输入节点到隐含层权重平方和的分布规律揭示出各种描述符对还原电位的影响大小依次为 : $E_{LUMO} > E_{HOMO} > HF > CCR > EE > DIP$.由散点图揭示出影响为正有 EE ;影响为负有 E_{LUMO} , HF , DIP ;影响无明显正负性的有 E_{HOMO} , CCR .结果表明 ,贝叶斯规整化大大方便了网络规整化参数选择 ,保证了网络的优良概括能力和稳健性 .本研究对氯代芳香族化合物采用电化学处理的适用性以及分析相应电化学降解机理提供了依据 .

关键词 :氯代芳香族化合物 ; QSPR ; 还原电位 ;贝叶斯规整化神经网络 ;权重平方和
中图分类号 : X13 文献标识码 : A 文章编号 :0250-3301(2005)02-0021-07

Bayesian Regularized BP Neural Network Model for Quantitative Relationship Between the Electrochemical Reduction Potential and Molecular Structures of Chlorinated Aromatic Compounds

SUN Wei[1] , ZENG Guang-ming[1] , WEI Wan-zhi[2] , HUANG Guo-he[1]
(1 . Department of Environmental Science & Engineering , Hunan University , Changsha 410082 , China ; 2 . School of Chemistry and Chemical Engineering , Hunan University , Changsha 410082 , China)

Abstract : Bayesian regularized BP neural network (BRBPNN) technique was applied in QSPR model in environmental field . The BRBPNN model for quantitative relationship between the electrochemical reduction potential (ERP) and chemical structures of 87 chlorinated aromatic compounds was established . The structure descriptor pool is consisted of Cl number(Cl) , molecular weight (MW) and 6 quantum chemistry parameters which are calculated by MOPAC2000 built in ChemOffice2004 , including energy of the highest occupied molecular orbital (E_{HOMO}) , energy of the lowest occupied molecular orbital (E_{LUMO}) , heat of formation(HF) , dipole(DIP) , electronic energy(EE) , core-core repulsion(CCR) . The achieved optimal network structure was 6-20-1 , which possessed stronger fitting and prediction capacity than that of the stepwise linear regression and with the correlation coefficients square and the mean square error for the training set and the test set as 0.999 and 0.000105 , 0.965 and 0.00159 respectively . The sum of square weights between each input neuron and the hidden layer of BRBPNN(6-20-1) indicate the effect of descriptor on the electric potential declining in the order of $E_{LUMO} > E_{HOMO} > HF > CCR > EE > DIP$. The scatter diagrams show that the EE descriptors had positive effect on ERP , and E_{LUMO} , HF , DIP had negative effects , and E_{HOMO} and CCR showed ambiguous effects . Results show that Bayesian regularized BP neural network is of automated regularization parameter selection capability and thus may ensure the excellent generation ability and robustness . This study threw more light on the applicability of electrochemical treatment for the chlorinated aromatic compounds and the analysis on electrochemical reduction mechanism .

Key words : chlorinated aromatic compounds ; QSPR ; electric potential ; Bayesian regularized neural model ; sum of square weights

氯代芳香烃及其衍生物是广泛应用的化工原料 ,但一般都难降解且具有毒性 ,给传统生物处理带来了困难 .电化学技术由于其设备相对简单 ,固液气态适应能力和主要试剂是“电子”这一清洁试剂而引起了人们极大关注[1 ,2] .电化学技术可以消除污染物毒性而并不使它们完全矿化 ,从而可以作为生物处理的一种前端技术[3] . Farwell 等运用改进伏安法考察了多卤代芳香族化合物的多级还原过程及其机理[4 ,5] ,而且电化学技术用于含氯代芳香烃的污水

收稿日期 :2004-04-23 ;修订日期 :2004-08-10
基金项目 : 国家自然科学基金资助项目(20077006 , 50179011 , 70171055) ;国家杰出青年科学基金项目(50225926) ;教育部高等学校优秀青年教师教学科研奖励计划资助项目 ;国家高技术研究发展计划(863) 项目(2001AA644020)
作者简介 :孙伟(1978 ~) ,男 ,硕士研究生 .

处理的研究已有报道[1,3].但是逐一考察电化学法处理含有氯代芳香烃及其衍生物废水的适用性,需要消耗大量物力财力,QSPR 研究可以较好的解决这一问题[6].魏东斌等[7]采用逐步线性回归,利用 3 种量子化学描述符建立了氯代芳香族化合物结构与电化学还原电位之间的 QSPR 模型,取得较好的预测结果.然而,结构与活性之间大多情况下并非线性关系,而是高度复杂的非线性关系,这实际上已经超出了经典统计学的能力,因此反映非线性关系的各种神经网络已经广泛地成功应用在 QSPR 和 QSAR 的研究当中[8].

反向传播神经网络(BPNN)是最经常用于结构活性预测的神经网络,但是存在无法保证全局最优和过拟合等缺点.全局最优可以简单的通过多次运行获得满意的可行结果;影响网络概括能力的过拟合通常可以通过规整化(Regularization)来较好避免.规整化参数选择可以通过贝叶斯定理自动选取,构成所谓的贝叶斯规整化 BP 神经网络模型(BRBPNNs)[9~11].BRBPNNs 吸取了传统 BPNN 收敛快速和贝叶斯统计充分利用先验信息的优点,具有稳健、数据拟合良好且概括能力强等特点,新近在药学和毒理学的 QSAR 研究中得以成功应用[12,13],但在环境领域 QSPR 研究中鲜见应用报道.因此,本文采用贝叶斯规整化 BP 神经网络,基于 Chem Office2004 内置 MOPAC2000 计算了 6 种量子化学参数,同时考虑氯原子数和分子量,对 87 种氯代芳香烃及其衍生物的还原电位建立了 QSPR 模型[7],并且尝试进行了网络信息的挖掘.

1 材料与方法

1.1 贝叶斯规整化的 BP 神经网络模型(BRBPNNs)理论

BP 神经网络的拟合与泛化能力的矛盾一直是网络训练的困难所在.采用规整化技术可以较好的提高网络的泛化能力,这样网络训练的目标函数 F 可以表示为:

$$F = \alpha E_W + \beta E_D \tag{1}$$

其中,E_W 是网络权重的平方和,E_D 是网络响应和目标值的残差平方和,α 和 β 是目标函数参数(规整化参数).α 和 β 的相对大小决定了网络训练的侧重点是输出残差的减少还是网络体积的减小.实施规整化技术的主要困难在于选择恰当的目标函数参数,贝叶斯统计恰好可以用来优化规整化参数.将 α 和 β 视为随机变量,那么根据贝叶斯定理则有:

$$P(\alpha,\beta|D,M) = \frac{P(D|\alpha,\beta,M)\,P(\alpha,\beta|M)}{P(D|M)} \tag{2}$$

其中,D 代表训练数据集,M 代表使用的网络模型,w 代表网络权重(3).根据贝叶斯假设可以认为 α 和 β 先验分布服从均匀分布,那么欲使式(2)中 α 和 β 的后验分布概率最大,只需使似然函数 $P(D|\alpha,\beta,M)$ 最大.将残差和权重视为随机变量,又根据贝叶斯定理有:

$$P(w|D,\alpha,\beta,M) = \frac{P(D|w,\beta,M)\,P(w|\alpha,M)}{P(D|\alpha,\beta,M)} \tag{3}$$

如果假设残差和权重的分布均为高斯分布,则有:

$$P(D|w,\beta,M) = \frac{\exp(-\beta E_D)}{Z_D(\beta)} \quad 和$$

$$P(w|\alpha,M) = \frac{\exp(-\alpha E_W)}{Z_W(\alpha)} \tag{4}$$

为保证 $P(D|\alpha,\beta,M)$ 在式(3)中为规整化因子,则有

$$P(w|D,\alpha,\beta,M) = \frac{\exp[-F(w)]}{Z_F(\alpha,\beta)} \tag{5}$$

将式(4)和式(5)代入式(3)则得到:

$$P(D|\alpha,\beta,M) = \frac{Z_F(\alpha,\beta)}{Z_W(\alpha)\,Z_D(\beta)} \tag{6}$$

其中,$Z_W(\alpha) = (\pi/\alpha)^{N/2}$,$Z_D(\beta) = (\pi/\beta)^{n/2}$,$Z_F(\alpha,\beta) = (2\pi)^{N/2}\det^{-1/2}(H)\exp[-F(w^{MP})]$,$H = \beta\nabla^2 E_D + \alpha\nabla^2 E_W$ 是目标函数 F 的 Hessian 矩阵.将式(6)展开取对数再分别对 α 和 β 求导并令其为 0,解得 $P(\alpha,\beta|D,M)$ 最大和权重后验密度达到最小点(MP)时对应的 α 和 β 表达:

$$\alpha^{MP} = \frac{\gamma}{2E_W(w^{MP})},\ \beta^{MP} = \frac{n-\gamma}{2E_D(w^{MP})},$$

$$\gamma = N - \alpha^{MP}\mathrm{trace}^{-1}(H^{MP}) \tag{7}$$

其中,n 是样本集个数,N 是网络参数总数,γ 表示网络中对误差函数减少起作用较大的有效参数的个数.贝叶斯规整化的 BP 神经网络的训练是个迭代过程,初始设定 α 和 β 后,采用 Levenberg-Marquardt 算法求得 $F(w)$ 的极小值,按式(7)对 α 和 β 进行更新,获得后验分布最大意义上的最优值,然后再对新 $F(w)$ 求极小值,反复迭代直至收敛.BRBPNNs 的具体实现细节可参考文献[10,14].

1.2 数据集和处理方法

所有研究的氯代芳香烃及其衍生物的还原电位取自文献[7],包括萘、联苯、苯和苯酚等 4 类共 87 种.尽管在理论上 BRBPNN 只需要训练集就可以获得概括能力最大的网络结构并且可以避免常规交叉

检验的繁琐运算,但一般仍采用预测集进行检验.因此数据集分为 2 个子基:74 个化合物的训练集和 13 个化合物的预测集(数据集 15 %).预测集是由 4 类化合物中随机挑选而成.

用 ChemOffice 2004 内置的 MOPAC 2000 计算的量子化学参数包括:分子最高占据能 E_{HOMO}、分子最低占据能 E_{LUMO}、分子生成热 HF、分子偶极矩 DIP、分子的电子能量 EE 和分子的核核排斥能 CCR. MOPAC 计算关键字为"EF GNORM = 0.000 MMOK GEO-OK PM3 PRECISE LET".为使对应的量子化学参数具有可比性,实际计算中所有分子几何构型都反复计算直至控制在 GNORM(0.005 以内.另外,分子量和氯原子个数也作为候选描述符.所有受试化合物的 8 种描述符及还原电位列在表 1.候选描述符选取采用逐步线性回归和试算法.所有数据采用式(8)标准化变换,控制在[- 1,1]之间:

$$P_{标准} = 2 \times (P - P_{min})/(P_{max} - P_{min}) - 1 \quad (8)$$

前向逐步线性回归和其他统计分析是在 SPSS10.5 中进行. BRBPNNs 模型是在 MATLAB6.5中神经网络工具箱中编制调试通过.

表 1　受试化合物名称、描述符以及电化学还原电位

Table 1　Names, descriptors and electrochemical reduction potentials of the tested chemicals

化合物	分子描述符								还原电位[2]	
	最高占据能[3]	最低占据能[3]	Cl	生成热	电子能量	核核排斥能	偶极矩	分子量	实验值	BRBPNN
萘	- 8.836	- 0.408	0	169.701	- 6 640.176	5 332.869	0.000	128.173	- 2.197	- 2.198
1	- 8.804	- 0.620	1	144.495	- 8 018.957	6 410.270	0.905	162.618	- 1.940	- 1.942
2[1]	- 8.897	- 0.605	1	141.357	- 7 927.915	6 319.195	1.066	162.618	- 1.975	- 1.993
1,2	- 8.816	- 0.779	2	120.824	- 9 466.378	7 556.326	1.453	197.063	- 1.726	- 1.736
1,3	- 8.878	- 0.803	2	117.317	- 9 400.432	7 490.343	1.145	197.063	- 1.752	- 1.777
1,4	- 8.770	- 0.827	2	120.074	- 9 479.967	7 569.907	0.371	197.063	- 1.751	- 1.747
1,5	- 8.804	- 0.816	2	119.973	- 9 473.893	7 563.832	0.000	197.063	- 1.765	- 1.776
1,6	- 8.889	- 0.802	2	116.717	- 9 379.608	7 469.513	0.829	197.063	- 1.802	- 1.802
1,7	- 8.861	- 0.799	2	116.758	- 9 399.714	7 489.619	1.480	197.063	- 1.793	- 1.747
1,8[1]	- 8.719	- 0.809	2	131.043	- 9 551.786	7 641.839	1.590	197.063	- 1.704	- 1.652
2,3	- 8.969	- 0.757	2	118.488	- 9 375.104	7 465.028	1.602	197.063	- 1.769	- 1.791
2,6	- 8.910	- 0.791	2	113.522	- 9 276.419	7 366.291	0.000	197.063	- 1.844	- 1.834
2,7	- 9.004	- 0.764	2	113.571	- 9 280.706	7 370.579	0.908	197.063	- 1.838	- 1.828
1,2,3	- 8.888	- 0.922	3	98.404	- 11 007.431	8 796.027	1.675	231.509	- 1.554	- 1.577
1,2,4	- 8.796	- 0.971	3	97.287	- 11 021.141	8 809.725	1.100	231.509	- 1.565	- 1.568
1,2,5[1]	- 8.835	- 0.961	3	96.707	- 10 994.327	8 782.905	0.758	231.509	- 1.581	- 1.600
1,2,6	- 8.869	- 0.951	3	93.399	- 10 887.720	8 676.265	0.614	231.509	- 1.620	- 1.621
1,2,7	- 8.901	- 0.943	3	96.444	- 10 912.030	8 700.575	1.483	231.509	- 1.613	- 1.591
1,2,8	- 8.740	- 0.954	3	107.658	- 11 092.664	8 881.357	1.908	231.509	- 1.512	- 1.516
1,3,5	- 8.865	- 0.982	3	93.325	- 10 948.586	8 737.130	0.966	231.509	- 1.578	- 1.593
1,3,6	- 8.982	- 0.969	3	90.024	- 10 826.181	8 614.691	0.183	231.509	- 1.634	- 1.657
1,3,7	- 8.900	- 0.972	3	90.005	- 10 841.951	8 630.460	0.795	231.509	- 1.635	- 1.615
1,3,8	- 8.810	- 0.977	3	104.391	- 11 005.535	8 794.193	1.358	231.509	- 1.541	- 1.546
1,4,5	- 8.715	- 1.002	3	107.754	- 11 089.804	8 878.497	0.725	231.509	- 1.540	- 1.532
1,4,6	- 8.842	- 0.991	3	92.828	- 10 933.846	8 722.384	0.665	231.509	- 1.618	- 1.591
1,6,7[1]	- 8.934	- 0.939	3	94.263	- 10 919.697	8 708.250	1.567	231.509	- 1.599	- 1.591
2,3,6	- 9.006	- 0.929	3	91.052	- 10 788.599	8 577.119	0.779	231.509	- 1.657	- 1.651
1,2,3,4	- 8.816	- 1.082	4	78.673	- 12 722.121	10 209.394	1.523	265.954	- 1.393	- 1.383
1,2,3,5	- 8.887	- 1.092	4	74.657	- 12 628.506	10 115.737	1.323	265.954	- 1.411	- 1.417
1,2,3,7	- 8.939	- 1.080	4	71.325	- 12 513.795	10 000.992	1.176	265.954	- 1.445	- 1.450
1,2,4,6	- 8.846	- 1.128	4	70.334	- 12 535.702	10 022.888	0.247	265.954	- 1.445	- 1.439
1,3,5,7[1]	- 8.899	- 1.139	4	67.028	- 12 484.023	9 971.175	0.000	265.954	- 1.444	- 1.481
1,3,5,8	- 8.790	- 1.155	4	81.553	- 12 637.581	10 124.883	0.685	265.954	- 1.373	- 1.386
1,3,6,7	- 8.985	- 1.100	4	67.855	- 12 426.963	9 914.123	0.556	265.954	- 1.490	- 1.486
1,4,5,8	- 8.652	- 1.180	4	98.873	- 12 782.287	10 269.769	0.000	265.954	- 1.345	- 1.344
1,4,6,7	- 8.895	- 1.118	4	70.689	- 12 546.912	10 034.102	1.096	265.954	- 1.431	- 1.430
1,2,3,5,7,	- 8.937	- 1.239	5	48.555	- 14 228.831	11 414.672	0.494	300.399	- 1.340	- 1.336
octa	- 8.769	- 1.583	8	18.912	- 20 189.254	16 471.430	0.000	403.734	- 0.940	- 0.941
联苯	- 8.918	- 0.361	0	198.493	- 8 700.176	7 125.222	0.000	154.211	- 2.410	- 2.407
2	- 9.166	- 0.222	1	180.032	- 10 298.979	8 422.714	0.820	188.656	- 2.097	- 2.100
3[1]	- 9.030	- 0.524	1	170.456	- 10 108.643	8 232.278	1.038	188.656	- 2.108	- 2.155
4	- 8.867	- 0.561	1	170.140	- 10 061.051	8 184.684	1.062	188.656	- 2.056	- 2.063
2,3	- 9.244	- 0.352	2	156.879	- 11 864.309	9 686.684	1.339	223.101	- 1.956	- 1.957
2,4	- 9.141	- 0.433	2	152.847	- 11 754.644	9 576.977	1.020	223.101	- 1.983	- 1.972
2,5[1]	- 9.122	- 0.402	2	152.755	- 11 788.974	9 611.306	0.321	223.101	- 1.942	- 2.010

续表 1

化合物	分子描述符								还原电位[2]	
	最高占据能[3]	最低占据能[3]	Cl	生成热	电子能量	核核排斥能	偶极矩	分子量	实验值	BRBPNN
2,6	-9.341	-0.147	2	159.634	-11 992.261	9 814.665	0.498	223.101	-2.107	-2.103
3,4	-8.921	-0.692	2	147.207	-11 627.799	9 450.074	1.567	223.101	-1.871	-1.858
3,5	-9.160	-0.677	2	143.602	-11 610.622	9 432.859	1.209	223.101	-1.897	-1.894
2,3,4	-9.188	-0.537	3	134.101	-13 478.483	10 999.503	1.535	257.546	-1.852	-1.849
2,3,5	-9.173	-0.525	3	130.496	-13 447.912	10 968.894	0.954	257.546	-1.783	-1.808
2,3,6[1]	-9.170	-0.387	3	136.900	-13 639.853	11 160.902	0.538	257.546	-1.937	-1.925
2,4,5	-9.090	-0.585	3	130.401	-13 402.860	10 923.842	1.043	257.546	-1.837	-1.838
2,4,6	-9.454	-0.382	3	133.413	-13 541.985	11 062.998	0.464	257.546	-1.966	-1.968
3,4,5	-8.990	-0.818	3	124.721	-13 287.966	10 808.888	1.728	257.546	-1.696	-1.701
2,3,4,5	-9.151	-0.680	4	111.929	-15 220.287	12 439.958	1.394	291.991	-1.679	-1.678
2,3,4,6	-9.222	-0.551	4	114.873	-15 347.910	12 567.611	0.863	291.991	-1.784	-1.772
2,3,5,6	-9.119	-0.577	4	114.854	-15 381.068	12 600.769	0.542	291.991	-1.787	-1.790
2,3,4,5,6[1]	-9.179	-0.676	5	96.495	-17 247.495	14 165.886	1.019	326.436	-1.566	-1.633
2,2'	-9.322	-0.070	2	158.695	-12 008.602	9 830.996	1.030	223.101	-2.126	-2.127
3,3'	-9.129	-0.675	2	142.750	-11 572.980	9 395.209	0.000	223.101	-2.030	-2.032
4,4'	-8.846	-0.741	2	142.082	-11 470.871	9 293.093	0.000	223.101	-2.000	-1.995
2,4'	-9.116	-0.405	2	151.817	-11 738.902	9 561.225	1.497	223.101	-2.042	-2.041
2,2',6,6'	-9.350	-0.209	4	117.900	-15 786.874	13 006.607	0.000	291.991	-2.123	-2.121
2,2',5,5'	-9.246	-0.384	4	104.683	-15 227.078	12 446.673	0.065	291.991	-1.900	-1.903
3,3',4,4'	-8.957	-0.970	4	96.838	-14 769.943	11 989.457	0.000	291.991	-1.764	-1.762
3,3',5,5'	-9.364	-0.948	4	90.060	-14 771.369	11 990.813	0.000	291.991	-1.720	-1.717
2,2',4',5,5'	-9.245	-0.556	5	82.547	-16 974.829	13 893.075	0.702	326.436	-1.771	-1.771
2,2',4,4',5,5'[1]	-9.268	-0.649	6	60.508	-18 771.616	15 388.515	0.041	360.882	-1.764	-1.712
2,2',4,4',6,6'	-9.567	-0.534	6	65.949	-19 250.670	15 867.625	0.000	360.882	-1.908	-1.909
deca	-9.279	-0.843	10	-6.695	-27 831.062	23 242.787	0.000	498.662	-1.406	-1.406
2,酚	-9.041	0.028	1	-113.261	-5 551.084	4 153.027	1.918	128.558	-2.500	-2.499
2,4	-9.007	-0.184	2	-140.267	-6 773.920	5 074.464	1.841	163.003	-2.350	-2.351
2,4,5	-8.986	-0.429	3	-162.956	-8 156.530	6 155.720	1.331	197.448	-2.200	-2.200
2,3,4,6[1]	-9.077	-0.625	4	-183.964	-9 707.799	7 405.651	0.683	231.893	-1.950	-1.969
2,3,4,5,6	-9.136	-0.789	5	-202.472	-11 343.159	8 739.700	1.104	266.338	-1.700	-1.699
苯	-9.751	0.396	1	97.846	-3 171.560	2 368.735	0.000	78.113	-2.440	-2.441
1,2	-9.295	-0.168	2	46.269	-5 418.492	4 012.894	1.351	147.004	-2.220	-2.219
1,3	-9.421	-0.191	2	42.337	-5 353.315	3 947.676	0.880	147.004	-2.200	-2.200
1,4	-9.235	-0.243	2	41.990	-5 341.558	3 935.916	0.000	147.004	-2.200	-2.200
1,2,3[1]	-9.379	-0.333	3	23.604	-6 714.628	5 007.676	1.350	181.449	-1.960	-1.986
1,2,4	-9.241	-0.435	3	19.755	-6 637.759	4 930.767	0.666	181.449	-2.000	-2.001
1,3,5	-9.588	-0.381	3	16.328	-6 584.448	4 877.420	0.000	181.449	-1.990	-1.990
1,2,3,4	-9.281	-0.556	4	1.312	-8 092.412	6 084.110	0.995	215.894	-1.760	-1.764
1,2,3,5	-9.298	-0.592	4	-2.088	-8 027.411	6 019.073	0.462	215.894	-1.790	-1.782
1,2,4,5[1]	-9.190	-0.635	4	-2.135	-8 015.556	6 007.217	0.000	215.894	-1.810	-1.804
penta	-9.251	-0.737	5	-20.349	-9 563.672	7 254.026	0.432	250.339	-1.570	-1.575
hexa	-9.309	-0.832	6	-38.482	-11 193.383	8 582.430	0.000	284.784	-1.320	-1.321

1) 预测集 2) 还原电位的实验值取自文献[7] 3) 个别化合物的最高和最低占据能与文献[7]有差异,可能是分子构型优化程度不同所致.

2 结果与讨论

2.1 特征选取

最终由标准化训练集数据得到的氯代芳香烃及其衍生物的还原电位多元线性回归模型含有 5 种描述符.描述符的回归系数和其统计学检验结果详见表 2.从表 2 可以看出,所有系数为 0 的 t 检验的显著性概率都小于 0.05,因此可认为这些系数都有显著性意义.同时各个系数的 VIF 值都不是很大,表明此线性模型可以接受.然而,逐步线性回归模型对训练集和预测集的预测效果并不令人满意,相关系数平方(R^2)和均方误差(MSE)分别只有 0.956 和 0.00378 以及 0.882 和 0.00459,列在表 3.

2.2 BRBPNN 的结构确定

理论上,一个隐含层传递函数为 sigmoid 函数和输出层为线性传递函数的 3 层网络可以拟合任意连续或有限间断的函数[14],这里分别选取 Matlab 中的 tangsig 和 pureline 函数.网络输入层节点由特征选取的描述符数目确定,网络的输出层节点只有还原电位一个.传统 BPNN 隐含层的数目只能通过反复试验考察网络效果的方法大致确定,而 BRBPNN 则可以自动寻找后验分布最大意义上的

表 2　标准化变换后的训练集的逐步线性回归模型系数

Table 2　Coefficients of stepwise linear regression model based on standardized training data

入选描述符	非标准化		标准化系数	t	Sig.	B 的 95 % 置信区间		共线性分析	
	系数 B	标准误差				下界	上界	容限	VIF
常数	- 0.100	0.022		- 4.520	0.000	- 0.144	- 0.056		
E_{LUMO}	- 0.734	0.034	- 0.670	- 21.514	0.000	- 0.802	- 0.666	0.661	1.513
Cl	0.993	0.084	0.830	11.842	0.000	0.825	1.160	0.130	7.670
HF	0.367	0.046	0.361	8.061	0.000	0.276	0.458	0.319	3.132
CCR	- 0.499	0.072	- 0.394	- 6.930	0.000	- 0.642	- 0.355	0.198	5.039
DIP	0.093	0.017	0.152	5.412	0.000	0.059	0.127	0.815	1.227

表 3　不同代表性模型的结果比较

Table 3　Results comparison among the representative models

模型	描述符组合[1)]	标准化训练集(74)		训练集(74)		预测集(13)		网络总节点数	网络有效节点数
		E_D	E_W	R^2	MSE	R^2	MSE		
逐步线性回归	2 3 4 6 7			0.956	0.00378	0.882	0.00459		
BRBPNN(5-8-1)	2 3 4 6 7	0.0412	32.960	0.996	0.000339	0.931	0.00270	57	36.406
BRBPNN(5-13-1)	2 3 4 6 7	0.0379	31.678	0.996	0.000312	0.940	0.00238	92	38.084
BRBPNN(5-19-1)	2 3 4 6 7	0.0365	33.503	0.997	0.000300	0.927	0.00326	134	37.470
BRBPNN(6-8-1)	1 2 4 5 6 7	0.0263	30.965	0.998	0.000216	0.910	0.00362	65	43.799
BRBPNN(6-10-1)	1 2 4 5 6 7	0.0179	34.045	0.998	0.000147	0.965	0.00154	81	49.194
BRBPNN(6-20-1)	1 2 4 5 6 7	0.0128	40.942	0.999	0.000105	0.965	0.00159	161	51.446
BRBPNN(8-8-1)	1 2 3 4 5 6 7 8	0.0405	15.304	0.996	0.000333	0.915	0.00342	81	37.739
BRBPNN(8-19-1)	1 2 3 4 5 6 7 8	0.0127	28.665	0.999	0.000105	0.969	0.00119	191	48.416

1)描述符的代表数字:1. E_{HOMO}　2. E_{LUMO}　3. Cl　4. HF　5. EE　6. CCR　7. DIP　8. MW

最优值[9~11].为了展示隐含层数量的确定方法,图 1 总结了不同描述符组合下的 BRBPNN,平行训练 20 次每次最大终止代数 2000 代获得的最优网络的状态随着隐含层节点数 S 增加的变化情况.从图 1 可以看出,当 $S > 10$ 以后,3 种描述符组合 BRBPNN 的训练集和预测集 MSE 以及网络有效节点个数都基本趋于稳定.因此,$S > 10$ 后,各种描述符组合下的 BRBPNN 都应该能够给出很好的计算结果,但并不一定是最优结果.

2.3　BRBPNN 与线性模型的预测结果比较

不同描述符组合和不同隐含层节点数下的代表性 BRBPNN 与逐步线性回归模型的结果比较列在表 3.从表 3 可以看出,无论是对训练集还是预测集,体现描述符与还原电位之间非线性关系的各种 BRBPNNs 的相关系数平方和 MSE 都要大大好于假定呈线性关系的逐步线性回归模型,尤其是对于用于逐步线性回归模型相同描述符组合构建的 BRBPNN(5-S-1)要比逐步线性回归模型在拟合和预测能力上好出很多,这都充分表明了描述符和还原电位之间的高度非线性关系.然而,在计算获得的模型当中,通过试算法获得 6 种量子化学描述符组合的 BRBPNN(6-20-1) 结果明显优于 BRBPNN(5-S-1),而且(6-S-1)系列的稳健性最好(见图 1),这

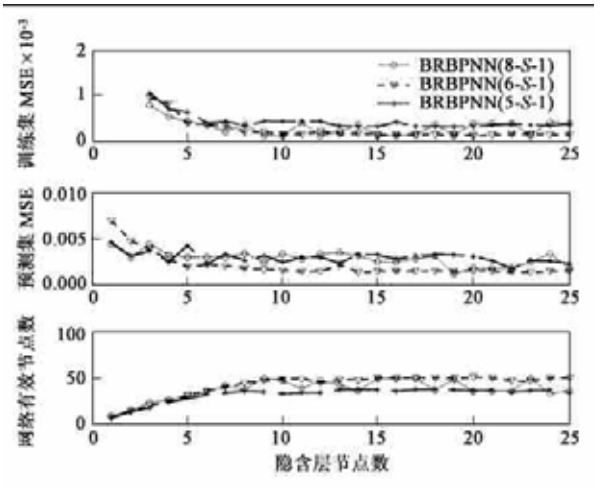

图 1　平行训练 20 次获得的最优 BRBPNNs 随隐含层节点数的变化

Fig.1　Changes of optimal BRBPNNs along with the hidden layer neuron number by parallel training for 20 times

预示着 6 种描述符对于氯代芳香烃及其衍生物的还原电位都可能存在着非线性影响,同时也说明直接简单地通过将逐步线性回归模型获得的描述符组合作为网络输入组合而获得的神经网络模型是无法保证效果的.另外,BRBPNN(8-19-1)的结果相比 BRBPNN(6-20-1)在预测集上略好一些,但(8-S-1)系列的稳健性较差(见图 1),表明前者输入层所包含的氯原子数和分子量对于受试化合物的还原电位

的影响可能较小,且对结果存在干扰.因此,本文选取 BRBPNN(6-20-1)为最优模型.

为了更直观地对比最优 BRBPNN(6-20-1)与逐步线性回归模型的效果,图 2 分别列出了 2 种模型的训练集和预测集的实验值和计算值的对比.可以看到,BRBPNN(6-20-1)的计算值和实验值十分接近,基本上全都聚集在等值线上,显示了其优良的拟合逼近能力和避免过拟合的本领,大大优于逐步线性回归模型的结果.最终获得的 BRBPNN(6-20-1)的还原电位预测值列在表 1.

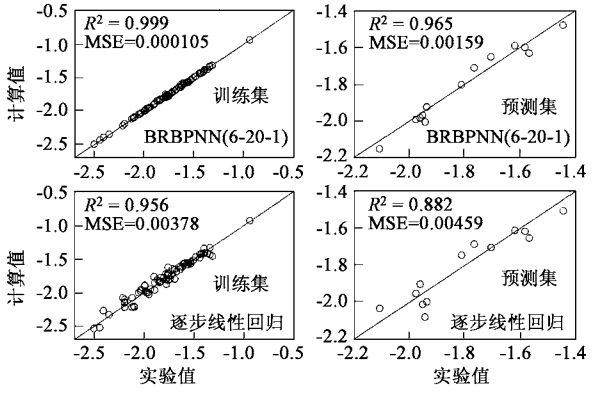

图 2 BRBPNN(6-20-1)和逐步线性回归的结果比较

Fig.2 Comparison results between BRBPN(6-20-1) and stepwise linear regression

2.4 BRBPNNs 的稳健性

通常,初始化权重的随机取值给传统 BPNN 的稳定性带来很大麻烦,从图 1 却可以看出,不同描述符组合和不同隐含层条件下的最优 BRBPNNs 在 $S>10$ 后随着隐含层数的增加显示了较好的稳健性.比较而言,BRBPNN(6-S-1)稳健性最好,训练集和预测集 MSE 都稳定保持在较小的范围之内,有效节点数也基本稳定.BRBPNN(8-S-1)稳健性相对最差,这可能是低效或无效描述符对网络规整化训练的干扰所引起的.

2.5 描述符对还原电位的影响及量子化学解释

如何从人工神经网络结构获得有效信息一直是网络研究的难点之一.如果效果较优模型的最终训练获得的网络结构信息一定程度上反映了描述符和还原电位之间的复杂非线性关系,那么挖掘最优状态下表现神经网络结构信息的网络权重和偏置信息应该能够考察出这些关系.文献[15]采用网络信息流分析的方法是一种可行方法,但略显复杂.由于这里输出层只有还原电位一个输出节点,那么各种描述符对还原电位的影响就直接表现在描述符对网络的影响.而描述符组成的输入层直接作用于隐含层,转而通过隐含层作用于输出层,因此这里尝试通过考察描述符对隐含层的影响作为描述符对还原电位的影响.

输入节点到隐含层所有节点的权重平方和一定程度上表现了该输入节点(描述符)对隐含层的影响.表 4 列出了获得的最优 BRBPNN 最终训练状态下给出的信息.从表 4 可以看出,BRBPNN(6-20-1)的权重平方和揭示的 6 种量子化学参数对还原电位的影响大小依次为:$E_{LUMO} > E_{HOMO} > HF > CCR > EE > DIP$.

为了确定描述符对还原电位的影响的正负性,在图 3 中作出每种描述符和还原电位实验值的散点图.可以看出,还原电位大致随 EE、MW、Cl 增大而增大,随 E_{LUMO}、HF 和 DIP 减小而减小.这些影响是和量子化学理论基本吻合:化合物的分子最低占据能越低,对应空轨道越容易接受外来电子,化合物越容易被还原,还原电位越高;氯原子是吸电子取代基,氯原子增多将会导致苯环上电子云密度降低,使得化合物容易接受电子被还原;分子生成热的增加意味着分子稳定性的增加,化合物不易被还原,还原电位降低;偶极矩表示的是外加电场下能量的一阶导数,反映的是分子电荷分布的不对称性,偶极矩增加,还原电位减小;反映分子总体能量的电子能量越高,分子越不稳定容易失去电子,还原电位增大.

表 4 输入节点到隐含层的权重平方和及百分比例

Table 4 Sum of square weights from input neurons to hidden layer and their percentage

BRBPNN	E_{HOMO}	E_{LUMO}	Cl	HF	EE	CCR	DIP	MW
(6-20-1)	5.4299	7.0512		4.6975	3.2248	3.2807	2.1624	
百分比/%	21.0	27.3		18.2	12.5	12.7	8.4	

另外,图 3 显示的还原电位分别和 E_{HOMO}及 CCR 之间基本没有明显正负影响关系,而化合物的分子最高占据能越高,分子轨道中成键电子的离域性增强,有利于分子还原过程中过渡态形成,还原电位应该越低;分子的核核排斥能对还原电位的影响不直接.这 2 个影响的正负性还需要探索其他方法来更好地揭示.

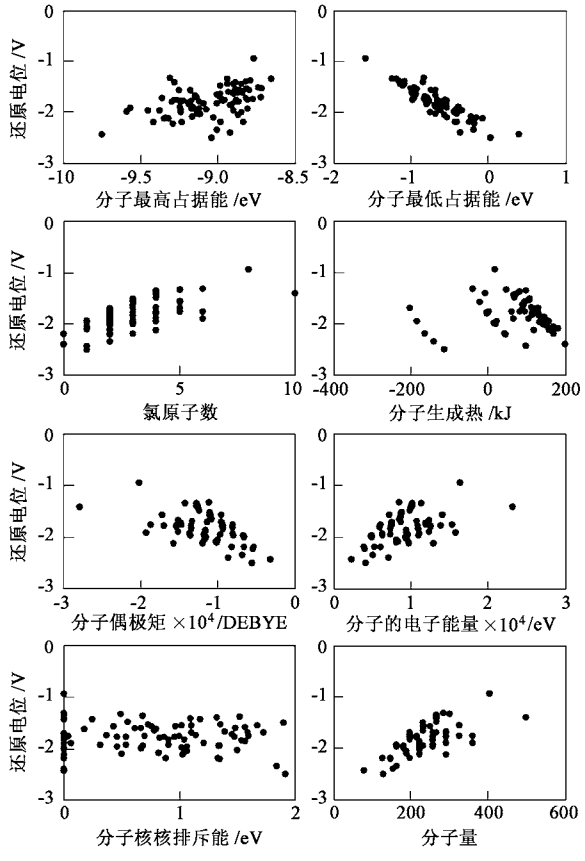

图 3　描述符对还原电位的影响

Fig.3　Descriptors effects on electric potential

3　结论

(1) 本文基于贝叶斯规整化神经网络,采用 Chem Office2004 内置的 MOPAC2000 计算的 6 种量子化学参数以及氯原子数和分子量,对 87 种氯代芳香族化合物结构与电化学还原特性关系建立了 QSPR 模型,取得了满意的预测效果.最优模型结构为 6-20-1,训练集和预测集的相关系数平方和均方根误差分别达到 0.999 和 0.000105, 0.965 和 0.00159.这一模型有益于氯代芳香族化合物采用电化学处理的适用性分析.

(2) 本文提出分析描述符对还原电位影响的简单有效定量方法:即对最优模型 BRBPNN (6-20-1),计算描述符-隐含层权重平方和,从而揭示各种描述符对还原电位的影响大小依次为: $E_{LUMO} > E_{HOMO} > HF > CCR > EE > DIP$.由散点图揭示的影响为正的有 EE;为负的有 E_{LUMO}, HF, DIP;无明显正负性的有 E_{HOMO}和 CCR.这一简便方法对揭示神经网络内在信息有一定借鉴意义并为分析相应的电化学降解机理提供依据.

(3) 描述符和还原电位之间呈现高度非线性关系,各种 BRBPNN 模型都要大大好于假定呈线性关系的逐步线性回归模型.在 BRBPNN 中,(6-*S*-1) 的稳健性和预测结果综合最优,说明低效或无效描述符对网络规整化训练有一定干扰;而且直接将逐步线性回归模型获得的描述符组合作为神经网络模型的输入组合是无法保证效果的.

(4) 贝叶斯规整化自动获取后验分布意义上的最大值,大大方便了神经网络训练时规整化参数的选择,非常好地避免了网络过拟合的发生,并且具有优良稳健性.贝叶斯规整化神经网络对于环境领域的其他 QSAR 和 QSPR 模型的建立具有应用前景.

参考文献:

[1] Rodgers J D, Jedral W, Bunce N J. Electrochemical Oxidation of Chlorinated Phenols[J]. Environ. Sci. Technol., 1999, **33**(9): 1453~1457.

[2] 陶映初,陶举洲. 环境电化学[M]. 北京:化学工业出版社, 2003.1~9.

[3] Panizza Marco, Bocca Cristina, Cerisola Giacomo. Electrochemical treatment of wastewater containing polyaromatic organic pollutants[J]. Water Research, 2000, **34**(9):2601~2605.

[4] Farwell S O, Beland F A, Geer R D. Reduction pathways of organohalogen compounds: part I. Chlorinated benzenes[J]. Electroanalytical Chemistry and Interfacial Electrochemistry, 1975, **61**: 303~313.

[5] Farwell S O, *et al*. Interrupted 2 sweep voltammetry for the identification of polychlorinated biphenyls and naphthalenes[J]. Analytical Chemistry, 1975b, **47**: 895~903.

[6] 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. 45~59.

[7] 魏东斌, 胡洪营, 平本兴正, 藤江幸一. 氯代芳香族化合物电化学还原特性的测定及 QSPR 研究[J]. 环境科学, 2003, **24**(2): 19~22.

[8] Niculescu, Stefan P. Artificial neural networks and genetic algorithms in QSAR[J]. Journal of Molecular Structure: THEOCHEM, 2003, **22**(1-2): 71~83.

[9] MacKay D J C. Bayesian interpolation[J]. Neural Computation, 1992, **4**(3): 415~447.

[10] Foresee F D, M T Hagan. Gauss-Newton Approximation to Bayesian Learning[A]. In: Proceedings of the 1997 International Joint Conference on Neural Networks[C]. Houston: 1997. 1930~1935.

[11] MacKay D J C. A Practical Bayesian Framework for Backprop Networks[J]. Neural Computation, 1992, **4**(3): 448~472.

[12] Burden F R, Winkler D Q. Robust QSAR Models Using Bayesian Regularized Neural Networks[J]. J Med Chem, 1999, **42**(16): 3183~3187.

[13] Burden F R, Winkler D A. A Quantitative Structure-Activity Relationships Model for the Acute Toxicity of Substituted Benzenes to *Tetrahymena pyriformis* Using Bayesian-Regularized Neural Networks[J]. Chem. Res. Toxicol., 2000, **13**(6): 436~440.

[14] The Math Works, Inc. http://www.mathworks.com/access/helpdesk/help/pdf_doc/nnet/nnet.pdf [OL]. Natick, MA, USA.

[15] 高大文, 王鹏, 郑彤, 彭永臻. 多氯酚定量构效关系人工神经网络信息流分析[J].中国环境科学, 2002, **22**(6):561~564.