应用径向基函数网络预测有机化合物的生物活性

戴勇, 卢颖, 张爱茜, 韩朔睽(南京大学环境科学与工程系, 南京 210093, E-mail: ecls@nju.edu.cn)

摘要: 为了构建径向基函数(RBF)神经网络模型, 研究一批取代芳香族化合物结构与其大型蚤生物活性参数 $\log(1/EC_{50})$ 的关系, 利用正交最小平方算法逐一选择非线性高斯函数的中心, 并将归一化参数 σ 和容差 ρ 作为网络的系统参数, 通过广程扫描确定其最佳值. 采用种子聚类分析与简单随机抽样相结合的方法将化合物合理地划分为训练集和预测集. 对构建的 RBF 网络模型的预测质量进行了不确定性分析. 与回归模型相比, RBF 网络具有较好的预测性能, 模型预测的不确定性降低了 30%.

关键词: 径向基函数网络; 高斯函数; 生物活性预测; 不确定性分析
中图分类号: X171.5 文献标识码: A 文章编号: 0250-3301(2000)04-0011-05

Predicting of Bioactivity of Organic Compounds by a Neural Network Based on Radial Basis Functions

Dai Yong, Lu Ying, Zhang Aiqian, Han Shuokui(Dept. of Environmental Science & Engineering, Nanjing University, Nanjing 210093, China E-mail: ecls@nju.edu.cn)

Abstract: A Neural Network Based on Radial Basis Functions(RBF) was applied to study the relationship between structure and bioactivity ($\log(1/EC_{50})$) to *Daphnia magna* for some substituted aromatic compounds. The centers of Gauss functions were selected one by one with the orthogonal least squares algorithm. The optimum values of the two parameters of the system, σ and ρ, were determined by large-scale scanning. The compounds were divided into learning set and predicting set by seeded clustering and random sampling. The predicted quality of the RBF network model was evaluated by an uncertainty analysis method, which shows a 30% decrease of the prediction uncertainty compared with a regression model. This result indicates that the RBF model has a better predicting performance.

Keywords: Radial Basis Function Network; Gauss function; prediction of bioactivity; uncertainly analysis

在有机物结构与活性关系(quantitative structure activity relationship, QSAR)研究中, 多元线性回归分析是常用的建模方法. 回归模型直观, 关系明确, 但难以表达非线性关系. 运用人工神经网络研究非线性问题目前备受关注[1], 其中反向传播(Back Propagation, BP)网络在 QSAR 研究中已获得了成功的应用. 应该指出, BP 网络由于算法问题, 存在训练时间过长, 易陷于局部误差极小的缺点, 对训练量大的应用并不合适. 近年来有关径向基函数网络(Radial Basis Function Network, RBF)的研究[2,3]表明, RBF 网络除保证实现全局误差最小外, 其训练时间之短, 使它具有实时应用的潜力. 本文尝试运用这种网络研究有机物结构与活性的关系.

1 RBF 网络的结构

本文构建的 RBF 网络由输入, 隐含和输出 3 层组成. 同层神经元之间没有连接, 相邻 2 层神经元全互连接. 输入层神经元无计算功能, 其作用是将 n 个输入分布给隐含层中的各神经元. 隐含层神经元通过径向基函数—— 高斯函数对输入产生非线性应答. 输出层神经元对隐含层的输出进行线性加权组合, 产生 1 个标量输出, 实现从 R^n 到 R 的映射.

1.1 高斯函数表达式

基金项目: 国家自然科学基金资助项目(Project Supported by the National Natural Science Foundation of China): 29877012
作者简介: 戴勇(1977~), 男, 硕士研究生
收稿日期: 1999-11-16

隐含层中神经元的高斯函数表达式:

$$R = \exp\left| - \frac{(\boldsymbol{x} - \boldsymbol{c})^2}{2\sigma^2} \right|$$

式中, $\boldsymbol{x}$ 为隐含神经元接收的输入矢量; $\boldsymbol{c}$ 为高斯函数在 n 维空间的中心, 用矢量表示; 其输出值的范围在 0 到 1 之间. 从该式可知, 输入矢量 $\boldsymbol{x}$ 与高斯函数中心 $\boldsymbol{c}$ 的欧氏距离越近, 高斯函数的响应值 R 就越大; 对径向距离相同的输入, 产生相同的响应值 R.

1.2 选取高斯函数的中心 $\boldsymbol{c}$

本文采用正交最小平方(OLS)算法[3], 按前向回归的方式, 每次从训练集的样本中选取一个样本作为高斯函数的中心, 条件是被选中的样本对回归平方和的贡献要大于其它样本. 随着高斯函数中心选取数量的增加, 网络的性能逐渐改善, 一旦网络性能达到要求, 选择下一个高斯函数中心的过程就会自动停止. 实际上, 高斯函数中心选取的数目就是隐含层神经元的个数.

1.3 归一化参数 σ^2 及三维图示

在本文中, 归一化参数 σ^2 是一个系数参数, 影响每个高斯函数响应面的"坡度". 图 1 描绘了在 σ 取值不同的情况下, 中心为原点(0, 0)的高斯函数对 $x \in [-2, 2]$, $y \in [-2, 2]$正方形区域内的输入产生不同的响应 R. 图 1 显示: 随着 σ 值由小变大, 高斯函数响应面从陡峭向平坦变化. 响应面陡峭, 网络对样本与高斯函数中心的区分灵敏. 但是, σ 过小, 网络对噪声太敏感, 易失真; σ 过大, 又会使网络丧失区分和拟合的能力. 因此 RBF 网络需要有合适的 σ 值. 本文采用广程扫描的方法[2], 择出一个适用于隐含层中所有高斯函数的最佳 σ 值.

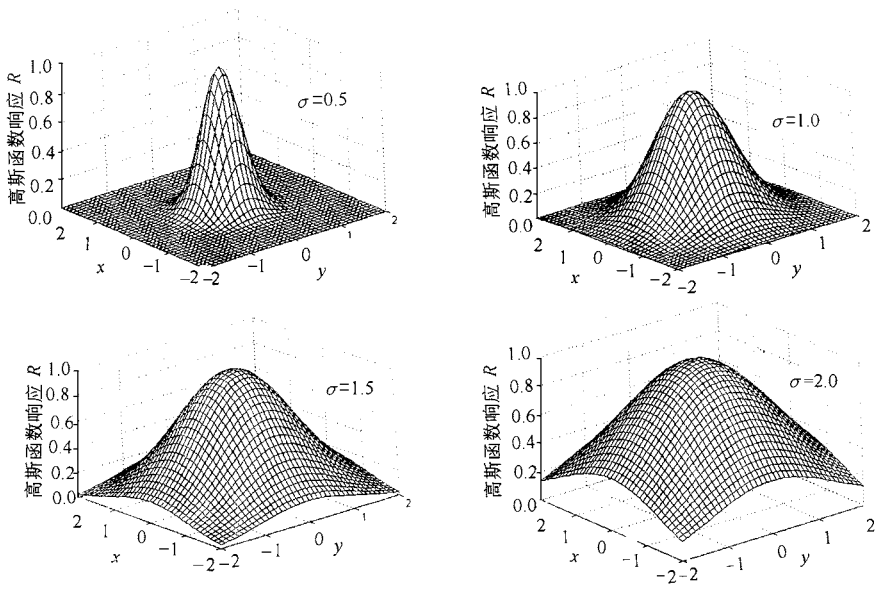

图 1 高斯函数响应面随 σ 变化的三维图

1.4 容差 ρ

容差 ρ 是 OLS 算法中的一个关键参数, 表征网络对训练集的拟合精度. ρ 值的大小也直接影响隐含层神经元的数量, ρ 值越小, 隐含层神经元的数目越多. 对 ρ 的取值应该权衡网络复杂性与网络精度 2 方面来考虑. 较为合适的 ρ 值应取大于并接近于训练集的残差与期望方差的经验比值. 本文也是通过广程扫描, 确定最合适的 ρ 值.

2 建模的数据

利用化合物的理化参数和结构参数作为模

型的输入参数, 以化合物的生物活性参数作为模型的输出参数, 采用回归和 RBF 2 种方法构建可以预测化合物生物活性的 QSAR 模型.

2.1　数据来源

可利用的 7 种输入参数为 46 个取代芳香族化合物的溶解度 $\lg S$, 辛醇/水分配系数 $\lg K_{OW}$, 取代基立体效应常数 E_S, 疏水常数 π_x, 电子效应常数 σ_X, 分子连接性指数 ^1X, 价指数 $^1X^V$; 可利用的模型输出参数是生物活性参数—— 大型蚤 48h 半数抑制浓度的倒数对数值 $\lg(1/EC_{50})$, 均取自文献[4].

2.2　模型输入参数的初筛

将 46 个化合物的 7 种待选输入参数与生物活性参数 $\lg(1/EC_{50})$ 放在一起, 进行偏相关分析, 逐一检查待选输入参数间的偏相关系数值的大小, 弃去相关性过高的参数, 从中挑选出偏相关系数均小于 0.5, 但与 $\lg(1/EC_{50})$ 有较高相关性的 5 种参数即: $\lg S$, E_S, σ_X, ^1X, $^1X^V$ 作为模型的输入参数.

2.3　预测集和训练集的选择

基于偏相关分析挑选的 5 种参数的数值, 采用种子聚类分析方法对化合物进行分类, 经多次随机给点试算, 46 个化合物按欧氏距离最终被较均匀地划分成 5 类, 每类化合物的数目分别为 12, 12, 8, 8 和 6. 随机地从各类中抽取 2 个化合物, 组成了含 10 个化合物的预测集(见表 2), 其余 36 个化合物合并成训练集(见表 1). 本文曾采用平均值, 重心, 中位数, 最远和最近等多种系数聚类方法, 但是分类的结果都很差, 各类化合物的数目之比过于悬殊.

3　建模的方法

3.1　回归模型

用多元回归分析, 在训练集 36 个化合物的 $\lg S$, E_S, σ_X, ^1X, $^1X^V$ 与 $\lg(1/EC_{50})$ 之间建立了如(1)式的线性方程:

$$\lg(1/EC_{50}) = -3.893 + 0.238\lg S + 0.0676E_S + 0.272\sigma_X + 0.710\,^1X + 0.489\,^1X^V \quad (1)$$

$$(R^2 = 0.6826, F = 16.05, SE = 0.4120, N = 36)$$

根据回归系数 t 检验的结果, 5 个参数中有 2 个参数(E_S 和 σ_X)对 $\lg(1/EC_{50})$ 没有显著影响. 采用逐步回归, 剔除不相干参数重建方程:

$$\lg(1/EC_{50}) = -5.162 + 0.269\lg S + 0.812\,^1X + 0.693\,^1X^V \quad (2)$$

$$(R^2 = 0.6756, F = 25.30, SE = 0.4166, N = 36)$$

方程(2)的所有参数对 $\lg(1/EC_{50})$ 均有显著影响. 两方程给出的 $\lg(1/EC_{50})$ 拟合值和预测值见表 1 和表 2. 无论是方程 1 或是方程 2, 复相关系数 R^2 值都不高, 提示 $\lg(1/EC_{50})$ 与参数间可能存在着非线性关系.

3.2　径向基函数网络模型

本文以 MATLAB 神经网络工具箱[5]为工作平台, 完成 RBF 网络的编程. 为了进行比较, 基于方程 1 和方程 2 中的 2 套参数, 构建了 2 套 RBF 网络. RBF-1 网络以 $\lg S$, E_S, σ_X, ^1X, $^1X^V$ 5 种参数作为输入参数, 各层神经元的数目分别为 5: 16: 1, 归一化参数 $\sigma = 2.21$, 容差 $\rho = 0.006$. RBF-2 网络以 $\lg S$, ^1X, $^1X^V$ 3 种参数为输入参数, 各层神经元的数目分别为 3: 16: 1, σ、ρ 分别取值 1.46 和 0.006.

3.3　网络拟合和预测结果

RBF-1 和 RBF-2 网络给出的 $\lg(1/EC_{50})$ 拟合值及预测值分别见表 1 和表 2.

4　模型的预测质量及不确定性分析[6,7]

本文应用基于经验的不确定性分析方法(EBUA), 采用 2 种模型性能指数对所构建的回归模型和 RBF 模型的预测质量进行了比较. 其中模型性能指数 r 是用预测集的 $\lg(1/EC_{50})$ 实验值对其预测值进行回归所得的相关系数表示. r 的最佳值为 1, r 越接近 1, 模型的预测质量越好. 实验值与模型预测值对数之间的"函数距离 d^2"是另一种用来评价模型预测性能的指数, 其定义为:

$$d^2 = \sum(\lg M_i - \lg O_i)^2/n$$

式中, M_i 为实验值, O_i 为模型预测值. d^2 越接近其最佳值 0, 模型的预测值与实验值越相符. 比较表 3 中所列的 r 和 d^2 值, 无论是 5 参数还

是 3 参数模型, RBF 网络模型的预测质量都好 于回归模型.

表 1 回归方程及径向基函数网络对 $\lg(1/EC_{50})$ 的拟合值

训练集			方程(1)		方程(2)		RBF-1		RBF-2	
编号	化合物	实验值	拟合值	误差	拟合值	误差	拟合值	误差	拟合值	误差
1	苯	0.31	0.16	0.15	-0.04	0.35	0.12	0.19	0.45	-0.13
2	1-氯苯	0.59	0.70	-0.11	0.56	0.03	0.48	0.11	0.37	0.22
3	1, 3-二氯苯	0.84	1.23	-0.39	1.12	-0.28	0.83	0.01	0.91	-0.07
4	1, 2, 3-三氯苯	1.54	1.70	-0.16	1.60	-0.06	1.57	-0.03	1.53	0.01
5	1, 2, 4-三氯苯	1.65	1.63	0.02	1.54	0.11	1.71	-0.06	1.64	0.00
6	1, 2, 4, 5-四氯苯	1.87	2.19	-0.31	2.13	-0.26	1.84	0.03	1.88	-0.01
7	六氯苯	2.31	2.93	-0.62	2.91	-0.60	2.32	0.00	2.31	0.00
8	五氯酚	4.50	3.57	0.93	3.40	1.09	4.50	-0.01	4.49	0.00
9	溴苯	0.81	0.89	-0.08	0.84	-0.03	1.06	-0.26	0.77	0.04
10	间-二溴苯	1.35	1.60	-0.24	1.68	-0.33	1.29	0.06	1.35	0.00
11	氟苯	0.67	0.92	-0.25	0.87	-0.20	0.82	-0.15	0.86	-0.19
12	碘苯	1.64	0.98	0.66	1.02	0.62	1.38	0.26	1.55	0.09
13	联苯	1.54	0.86	0.68	0.99	0.54	1.59	-0.05	1.55	-0.01
14	苯胺	0.41	0.59	-0.18	0.57	-0.16	0.58	-0.17	0.35	0.06
15	1, 3-氯苯胺	0.98	1.24	-0.27	1.21	-0.24	1.01	-0.03	1.04	-0.06
16	1, 4-氯苯胺	1.01	1.10	-0.10	1.18	-0.18	1.02	-0.01	1.15	-0.14
17	对-溴苯胺	1.43	1.24	0.18	1.45	-0.02	1.43	-0.01	1.37	0.06
18	2, 6-二氯苯胺	1.49	1.74	-0.25	1.72	-0.23	1.40	0.09	1.46	0.03
19	3, 4-二氯苯胺	1.43	1.69	-0.26	1.69	-0.26	1.43	0.00	1.48	-0.05
20	3, 4-氯氟苯胺	1.57	1.67	-0.10	1.62	-0.05	1.54	0.04	1.39	0.18
21	对-二胺苯	0.93	0.78	0.15	1.01	-0.08	0.84	0.09	0.78	0.15
22	二苯胺	2.05	1.13	0.92	1.34	0.71	2.01	0.04	2.05	0.00
23	硝基苯	0.64	1.03	-0.39	0.95	-0.31	0.81	-0.17	0.80	-0.16
24	邻-氯硝基苯	1.69	1.69	0.00	1.68	0.01	1.58	0.11	1.75	-0.07
25	对-氯硝基苯	1.82	1.61	0.20	1.63	0.19	1.51	0.31	1.65	0.17
26	1, 2-二硝基苯	2.09	1.88	0.20	1.76	0.33	2.21	-0.12	2.29	-0.20
27	1, 3-二硝基苯	2.15	1.89	0.26	1.74	0.41	2.24	-0.09	2.23	-0.08
28	1, 4-二硝基苯	2.41	1.90	0.51	1.73	0.68	2.26	0.14	2.15	0.25
29	间-二硝基甲苯	1.34	1.77	-0.43	1.88	-0.54	1.45	-0.11	1.36	-0.02
30	对-二硝基甲苯	1.36	1.71	-0.35	1.88	-0.52	1.42	-0.06	1.31	0.05
31	邻硝基甲苯	1.19	1.61	-0.43	1.76	-0.57	1.25	-0.06	1.32	-0.13
32	间-硝基甲苯	1.18	1.60	-0.43	1.73	-0.55	1.22	-0.04	1.33	-0.15
33	邻硝基苯胺	1.43	1.31	0.12	1.33	0.10	1.42	0.01	1.44	-0.01
34	间-硝基苯胺	1.50	1.29	0.21	1.32	0.18	1.41	0.09	1.44	0.06
35	对-硝基苯胺	1.53	1.29	0.24	1.31	0.22	1.43	0.10	1.45	0.08
36	对-硝基甲醛	1.31	1.38	-0.07	1.40	-0.09	1.53	-0.22	1.28	0.03

表 2 回归方程及径向基函数网络对 $\lg(1/EC_{50})$ 的预测值

预测集			方程(1)		方程(2)		RBF-1		RBF-2	
编号	化合物	实验值	预测值	误差	预测值	误差	预测值	误差	预测值	误差
1	1, 2-二氯苯	0.83	1.29	-0.46	1.16	-0.33	0.90	-0.07	1.10	-0.27
2	1, 4-二氯苯	0.91	1.34	-0.43	1.27	-0.36	1.02	-0.11	1.19	-0.29
3	1, 2, 3, 4-四氯苯	1.84	2.22	-0.38	2.15	-0.31	1.75	0.09	1.89	-0.05
4	对-二溴苯	1.66	1.44	0.22	1.54	0.12	1.47	0.19	1.35	0.31
5	1, 2-氯苯胺	0.97	1.28	-0.31	1.24	-0.27	1.10	-0.13	1.01	-0.04
6	间-碘苯胺	1.83	1.33	0.50	1.47	0.36	1.44	0.39	1.62	0.21
7	2, 4-二氯苯胺	1.45	1.75	-0.30	1.71	-0.26	1.45	0.00	1.45	0.00
8	间-氯硝基苯	1.77	1.37	0.40	1.64	0.13	1.36	0.41	1.68	0.08
9	二硝基溴苯	2.59	2.30	0.29	2.25	0.34	1.93	0.67	2.18	0.41
10	对-硝基甲苯	1.21	1.61	-0.40	1.74	-0.53	1.37	-0.16	1.32	-0.11

函数距离 d^2 还可用来表征模型预测的不确定性, 计算模型预测的置信区间. 通常模型预测的不确定性用标准偏差 σ 表示. 对于单一数据组, 如果随机变量 $\lg(M_i/O_i)$ 的概率分布为正态分布, 且模型输出不是有偏的, 则可导出标准偏差 $\sigma=(d^2)^{1/2}=d$[6]. 由于 d 与 $\lg(1/EC_{50})$ 的单位一致, 因此直接比较 d 的大小, 就可得出 5 参数和 3 参数的 RBF 网络的模型预测不确定性分别比对应的回归模型降低了 22% 和 30%. 本文还以 M_i/O_i 的 68% 置信区间 $[10^{-\sigma}, 10^{\sigma}]$ 为例, 在表 3 中列出了模型预测的置信区间. 结果表明 RBF 模型预测的 68% 置信区间比回归模型狭窄. 通观 r, d^2, d 和 68% 置信区间, 无论是 RBF 网络还是回归模型, 用 5 个输入参数构建的模型都不如用 3 个参数构建的模型预测质量好, 表明无用的参数影响模型预测的性能.

表 3　模型预测质量比较

模型	5 个输入参数				3 个输入参数			
	r	d^2	d	68% 置信区间	r	d^2	d	68% 置信区间
回归	0.706	0.143	0.378	[0.419, 2.389]	0.828	0.104	0.322	[0.476, 2.099]
RBF	0.929	0.0873	0.295	[0.507, 1.972]	0.948	0.0488	0.221	[0.601, 1.663]

5　讨论

构建 RBF 网络难点在于径向基函数中心的选择. 利用 OLS 法确定的基函数中心是训练集总样本中的一个子集. 这个子集能否表示基函数的最佳中心位置与样本容量的大小很有关系. 若样本容量很大, 最佳中心就有可能包含在其中. 对于小样本, OLS 法若不能在其中选择到最佳中心, 需付出增加隐含神经元的代价, 导致网络结构复杂.

根据具体研究对象, 归一化参数 σ 在本文中被设置为一个系统参数, 它不随每个径向基函数而变化, 减少了待定参数, 有助于模型简化.

参考文献:

1　瞿福平等. 人工神经网络及分子拓扑参数在酚类有机物 QSBR 研究中的应用. 环境科学, 1999, **20**(5): 16~19.

2　Lohninger H. Evaluation of Neural Networks Based on Radial Basis Functions and Their Application to the Prediction of Boiling Points form Structural Parameters. J. Chem. Inf. Comput. Sci., 1993, **33**(5): 736~744.

3　徐秉铮, 张百灵, 韦岗. 神经网络理论与应用. 广州: 华南理工大学出版社, 1994. 304~350.

4　王连生, 刘征涛, 周风帆, 于红霞, 高鸿. 芳烃类有机物结构与活性相关的模式参数研究. 环境科学, 1994, **15**(3): 7~10.

5　楼顺天, 施阳. 基于 MATLAB 的系统分析与设计—— 神经网络. 西安: 西安电子科技大学出版社, 1998. 98~143.

6　Monte L et al. Uncertainty analysis and validation of environmental models: the empirically based uncertainty analysis. Ecological Modelling, 1996, **91**(1~3): 139~152.

7　Dilks D W et al. Development of Bayesian Monte Carlo techniques for water quality model uncertainty. Ecological Modelling, 1992, **62**(1~3): 142~162.