

运用回归分析与人工神经网络预测含硫芳香族化合物好氧生物降解速率常数*

张爱茜 韩朔暎 沈 洲 马 静

(南京大学环境科学与工程系, 南京 210093)

摘要 为了研究人工神经网络预测有机化合物生物降解的性能, 运用多元线性回归方法和误差反向传递人工神经网络模型以基团代码作为结构描述符, 分别拟合、预测了一批含硫芳香族化合物的一级好氧生物降解速率常数。结果表明, 由于神经网络自动考虑了基团间的交互作用, 它对生物降解这类复杂问题有极高的求解能力, 预测的均方误差为 0.00102, 远低于线性回归模型的预测误差 0.01591。

关键词 人工神经网络, 预测, 含硫芳香族化合物, 好氧生物降解, 多元线性回归。

Prediction of Aerobic Biodegradation Rate Constants of Sulfur-Containing Aromatic Compounds Using Regression and Neural Network Techniques

Zhang Aiqian Han Shuokui Shen Zhou Ma Jing

(Department of Environmental Science and Engineering, Nanjing University, Nanjing 210093)

Abstract In order to evaluate the ability of artificial neural networks to predict the biodegradability of organic compounds, the codes of function groups were employed as structural descriptors to fit and predict the pseudo-first-order aerobic biodegradation rate constants of a batch of sulfur-containing aromatic compounds using multiple linear regression and error back propagation neural network models, respectively. The results indicated that the neural network has higher ability to solve the complex problems affected by confounding structure factors such as biodegradation process due to its ability to conclude interactions between groups automatically. The mean-square error of the network is 0.00102, which is much lower than that of the linear model 0.01591.

Keywords artificial neural networks, prediction, sulfur-containing aromatic compounds, aerobic microbiodegradation, multiple linear regression.

在有机物化学结构与生物降解性能关系的研究中, 化合物生物降解性的拟合、预测一般是通过生物降解速率常数与化合物理化参数之间的回归模型获得的, 因此如何选取及能否获得所需的理化参数对建模至关重要。在缺少所需的理化参数时, 直接利用化学基团作为结构描述符具有特别的优越性。然而, 化学基团之间存在的复杂相互作用却在一定程度上限制了基团

回归模型的应用。人工神经网络因其具有强大的处理非线性问题的能力, 可以任意逼近最佳刻划样本特性的函数, 能够自动考虑基团间的相互作用, 提供了一条解决此类问题的有效途

* 国家自然科学基金资助项目 (Project of Supported by National Natural Science Foundation of China) 批准号: 29477270

张爱茜: 女, 25 岁, 博士研究生

收稿日期: 1997-06-13

径^[1]. 本文采用简单的基团代码作为输入参数, 运用人工神经网络对一批含硫芳香族化合物的生物降解速率常数进行拟合与预测, 得到比线性回归方法更好的结果.

1 理论与模型实践

1.1 误差反向传递人工神经网络(Back Propagation, BP) 模型

在人工神经网络中, 误差反向传递模型是应用最为广泛的一种^[2,3]. 本研究用 C 语言编制的 BP 模型是由 3 个神经元层次组成, 其主要特点是:

(1) 4 个输入节点代表化合物的 4 个不同取代位置, 输入参数为基团代码. 被研究的 26 个含硫芳香族化合物的结构通式为:

$2-R_1-4-R_2-Ph-R_3-CH_2-CO_2-R_4$
式中, $R_1=NO_2, Cl, H$; $R_2=NO_2, Cl, H$; $R_3=S, SO, SO_2$; $R_4=CH_3, CH(CH_3)_2$. 针对这类化合物的结构特点, 对不同取代位置上的基团采用了简单的编码规则: 苯环上的邻、对位取代基 NO_2, Cl, H , 其代码分别是 1, 2, 3; 含硫基团 S, SO, SO_2 的代码分别是 1, 2, 3; 而酯烷基 $-CH_3$ 取 1, $-CH(CH_3)_2$ 取 2. 在本研究中基团代码只起类别标识作用, 经验证代码取不同值对结果没有影响.

(2) 采用双极性 S 型压缩函数($f(x) = -\frac{1}{2} + \frac{1}{1 + \exp(-x/\theta)}$)作为非线性转换函数^[4], 用以改进网络的训练性能, 其中 θ 为非线性系数. Stornetta 与 Huberman 提出的这种函数使节点的输出范围限定在 $\pm 1/2$ 之间, 网络收敛时间可大为缩短.

(3) 采用 Fahlman 改进的误差函数^[4], 该函数既保留了转换函数 $f(x)$ 一阶导数的作用, 又有助于消除网络在训练过程中出现的麻痹现象. 具体公式为: $\delta_p = [f'(x) + 0.1](t_p - O_p)$, 其中 δ_p, t_p 与 O_p 分别是网络输出层权重调节值, 目标值(即实验值)与网络的实际输出值.

(4) 采用扩展的 delta-bar-delta 算法. 该算法中附加的冲量项可防止网络振荡, 保证训练过程的稳定.

1.2 网络参数设定

在利用神经网络预测被试化合物一级好氧生物降解速率常数 K_d 的负对数值 pK_d 时(K_d 的测定方法另文发表), 为了减少网络的冗余度, 避免随机关联的发生, 本文构建的网络符合样本数大于网络可调参数的原则. 网络结构为 4:3:1, 即输入层 4 个节点, 隐含层 3 个节点, 输出层 1 个节点. 隐含层转换函数非线性调节系数 θ 取 0.8, 输出层取 0.4; 输入参数 C_i 按下式进行标准化:

$$\alpha = (C_i - C_{\min} + 0.01) / (C_{\max} - C_{\min} + 0.01)$$

其中, $C_{\max}$ 及 $C_{\min}$ 为输入参数 C_i 所有学习样本中的最大、最小值, 而 0.01 是为了避免零输入而附加的修正值. 同时 pK_d 还须进行如下变换, 使网络期望输出在 $-0.45 \sim 0.45$ 之间^[5].

$$t_p = (pK_d - 0.44) \times 2.53$$

1.3 选择训练组与预测组

采用聚类分析的方法, 在 26 个样本中合理地选择训练样本和预测样本. 本文基于 10 个理化参数^[6], 用类间平均法对被试化合物进行系统聚类分析(SPSS 统计分析软件包), 表征相似性的统计量为平方欧氏距离, 其系统聚类谱系见图 1. 图 1 中 case(1-26) 代表化合物的编号, 具体参见表 1. 根据聚类结果, 在保证每类至少包含 2 个化合物的情况下, 26 个化合物被分为

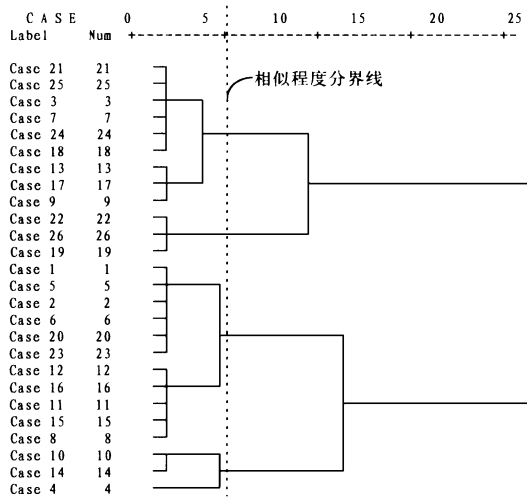


图 1 系统聚类谱图

4类,从每类中再随机挑选1–2个化合物作为预测样本,构成含5个样本的预测组;剩余的21个化合物构成训练组.聚类分析法保证了所取的训练组能覆盖样本提供的结构信息,避免挑选的预测样本在结构特征上过于相似.

2 生物降解速率常数的拟合与预测

分别用多元线性回归法和BP型神经网络对化合物的生物降解速率常数 pK_d 进行拟合与预测.

2.1 多元线性回归(Multiple Linear Regression, MLR)模型

由训练组的21个化合物构建的线性回归方程为: $pK_d = -0.0925(0.0276)R_1 - 0.0687(0.0316)R_2 + 0.0223(0.0229)R_3 + 0.0664(0.0363)R_4 + 0.536(0.112)$

$R^2 = 0.352, S = 0.0807, F = 3.717, N = 21$ 式中,变量 R_1, R_2, R_3, R_4 与神经网络的输入参数一致.由 R^2 可见该模型对生物降解速率常数的拟合能力很差,对5个预测样本的预测结果列于表1,其均方误差(Mean-Square Error, MSE)是0.01591.

表1 MLR与BP模型对被试化合物的拟合与预测结果¹⁾

		拟合与预测结果					实验值
化合物编号	输入代码	SE _{MLR}	O _{pMLR}	SE _{ANN}	O _{pANN}	pK _d	
训	1	3112	0.01020	0.393	0.00001	0.289	0.292
	3	2112	0.00422	0.461	0.00002	0.521	0.526
	4	1311	0.00026	0.279	0.00005	0.270	0.263
	5	1312	0.00533	0.345	0.00002	0.267	0.272
	6	1211	0.00053	0.371	0.00040	0.374	0.394
	7	1212	0.00168	0.438	0.00000	0.398	0.397
	8	1111	0.00032	0.464	0.00017	0.433	0.446
	9	1112	0.00230	0.530	0.00314	0.522	0.578
练	10	3121	0.00160	0.349	0.00012	0.378	0.389
	12	2121	0.02074	0.417	0.00029	0.544	0.561
	13	2122	0.01796	0.484	0.00029	0.601	0.618
	15	1322	0.00078	0.367	0.00017	0.382	0.395
	16	1221	0.00001	0.393	0.00014	0.384	0.396
	17	1222	0.00292	0.460	0.00722	0.491	0.406
	19	1122	0.00449	0.552	0.00002	0.480	0.485
	20	3131	0.00160	0.371	0.00006	0.323	0.331
组	21	2131	0.00593	0.440	0.00001	0.366	0.363
	22	2132	0.00422	0.506	0.00001	0.438	0.441
	23	1331	0.00001	0.323	0.00029	0.337	0.320
	24	1332	0.01613	0.390	0.00490	0.447	0.517
	25	1231	0.00314	0.416	0.00102	0.328	0.360
模型拟合误差(MSE)		0.00497		0.00087			
预 测 组	2	2111	0.00004	0.458	0.00116	0.486	0.452
	11	3122	0.00325	0.479	0.00026	0.438	0.422
	14	1321	0.01102	0.490	0.00122	0.350	0.385
	18	1121	0.01988	0.549	0.00185	0.365	0.408
	26	1232	0.04537	0.608	0.00062	0.420	0.395
模型预测误差(MSE)		0.01591		0.00102			

1) 表中SE_{MLR}与SE_{ANN}分别是回归模型(MLR)和神经网络模型(ANN)的平方误差(Square Error, SE), 计算公式为SE= (pK_d – O_p)²; 模型拟合(预测)误差MSE= 拟合(预测)样本SE之和/拟合(预测)样本数;
O_{pMLR}与O_{pANN}分别是回归模型与网络模型实际输出的pK_d值, 即拟合(或预测)值; pK_d是实验值.

2.2 BP 模型

对构建的 BP 网络训练 13100 次,得到的训练均方误差为 0.00087,预测均方误差是 0.00102,结果见表 1.训练次数对网络预测能力的影响如图 2 所示,训练次数大于 20000 时,人为加大了问题的复杂性,网络噪声变大,预测能力迅速降低.此外,隐含层节点数减小,预测误差也会增大,这是因为隐含节点过少,网络处理非线性问题的能力下降,达不到解决实际问题所需要的复杂程度.

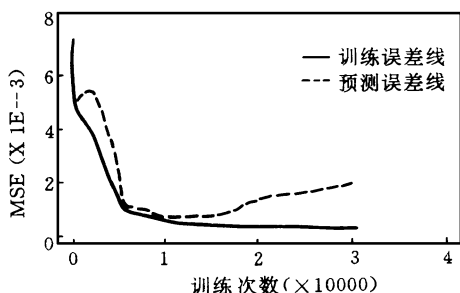


图 2 训练次数对网络预测能力的影响

3 结论

(1) 由于生物降解是一个复杂的生化反应过程,各基团对生物降解的影响并不独立,因此线性模型在预测生物降解速率常数时误差

较大,而人工神经网络能自动考虑基团间的相互影响,显示出解决非线性问题的独特能力.

(2) 利用聚类分析选择预测组,是一个行之有效的办法.它不似逐一交互验证法那么费时,又避免了逐次提取 K 个化合物交互验证法在化合物选取上所遇到的组合膨胀;在较短的时间内,就达到了满意的预测效果.

(3) 在拟合、预测精度允许的范围内,优化网络结构,减少冗余度,既可降低网络振荡,又避免了随机关联,增强了网络外推能力.

参 考 文 献

- 1 Salt D W et al. . The Use of Artificial Neural Networks in QSAR. Pestic. Sci. , 1992, **36**(2): 161
- 2 Gakh A A et al. . Neural Network-Graph Theory Approach to the Prediction of the Physical Properties of Organic Compounds. J. Chem. Inf. Comput. Sci. , 1994, **34**(4): 832
- 3 Doucet J P et al. . Neural Networks and ^{13}C NMR Shift Prediction. J. Chem. Inf. Comput. Sci. , 1993, **33**(3): 320
- 4 周继成等编著. 人工神经网络. 北京: 科学普及出版社, 1992: 52, 53
- 5 Tomoo Aoyama et al. . Neural Networks Applied to Quantitative Structure-Activity Relationship Analysis. J. Med. Chem. , 1990, **33**(9): 2583
- 6 韩朔睽, 李鹏等. 37 个含硫芳族化合物理化参数的相关性分析. 环境化学, 1992, **11**(4): 30

Internet 上的部分环境资源

世界卫生组织 WHO(<http://www.who.ch>). 世界卫生组织致力于人类健康事业, 其中涉及大量环境问题, 事实上 WHO 有自己的环境部门, 对大气污染、水体污染以及有毒有害物质在环境中的迁移转化都有深入的研究. 尤其在污染物对人体危害的发生机理、形式方面的研究具有权威性.

全球生物多样性和生态系统网络(<http://ht.tamu.edu/bene>). 这个网点建立在美国 Texas A&M 大学, 受美国国家环保局的资助, 提供全球有关生态保护、生态环境恢复和生态

管理方面的信息交流.

世界银行(<http://www.worldbank.org>). 随着环保意识的不断提高, 世界银行在进行全球性建设贷款时, 需要对资助项目进行环境影响分析, 因此对于全球性环境问题也有深入的研究.

同样涉及环境问题的国际组织还有设在加拿大的国际持续发展组织(<http://www.iisd.ca>) 和设在丹麦的国际能源与环境合作中心(<http://www.risoe.dk>) 等. (下转第 86 页)

清华大学环境工程系 钱海燕 供稿