

人工神经网络用于对位取代苯酚定量构效关系的研究*

宋新华 陈 茁 俞汝勤**

(湖南大学化学化工系, 长沙 410082)

摘 要

本文利用误差反向传播(BP)的人工神经网络(ANN)模型研究了对位取代苯酚衍生物的生物活性与其结构及物理化学性质参数之间的定量构效关系。优化了ANN模型的参数设计,提出了动态调节网络学习速率的经验规则以改善网络的性能。采用 $f(x)=1/(1+e^{-x})$ 作为网络节点的输入输出转换函数的三层神经网络具有较佳性能,当取隐含节点数为10时,该网络预测26个对位取代苯酚衍生物生物活性的均方误差(mse)为0.036,优于常规构效关系预测结果。

关键词 人工神经网络、定量构效关系、对位取代苯酚

对位取代苯酚衍生物定量构效关系(QSAR)的研究对于药物设计、毒理学及环境监测等方面均有重要意义^[1,2]。QSAR是值得进一步深入研究的化学计量学领域^[3],通常用到的主要是多元统计方法,如多元线性回归、逐步回归、统计模式识别等。探索一些新的计算方法以进一步提高传统方法定量预测化合物生物活性的准确度是一项有意义的研究。近年来引起研究者兴趣的人工神经网络(ANN)^[4,5]是一类模拟人脑功能的信息加工处理系统,其理论基础是神经网络的数学模型。它可用于处理因果关系不甚明确、推理规则不确定的复杂非线性问题,故亦适用于QSAR研究以提高预测准确度。本文以对位取代苯酚衍生物体系为例,详细探讨网络参数对QSAR结果的影响,证实采用 $f(x)=\frac{1}{1+e^{-x}}$ 作为网络节点的输入输出转换函数,取隐含节点数为10的三层网络能用于对位取代苯酚生物活性预测,预报功能优于文献[1]报道的结果。

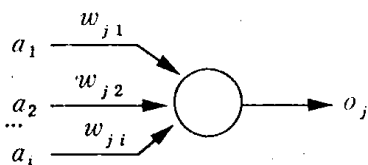
1 基 本 理 论

ANN由多层高度连接的基本处理单元或称节点(图1)组成。节点 j 通过连接权重接收

1992-01-24 收稿, 1992-08-06 收修改稿。

* 国家自然科学基金资助项目。

** 联系人。

图 1 节点 j 示意图

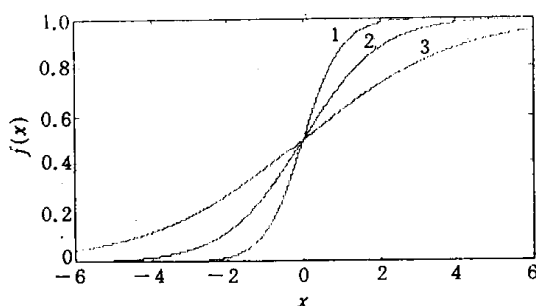
来自其它节点的信息,然后通过输入输出转换函数 $f(x)$ 给出输出信息:

$$o_j = f\left(\sum_i a_i W_{ji} + r_j\right), \quad (1)$$

此处, a_i 表示节点 j 的输入; o_j 表示节点 j 的输出; W_{ji} 表示节点 i 与节点 j 之间的连接权重; r_j 系偏置项.

通常 $f(x)$ 取如下非线性形式:

$$f(x) = \frac{1}{1 + e^{(-x/\theta)}}, \quad (2)$$

图 2 $f(x)$ 形状示意图

1—— $\theta=0.5$, 2—— $\theta=1.0$, 3—— $\theta=2.0$

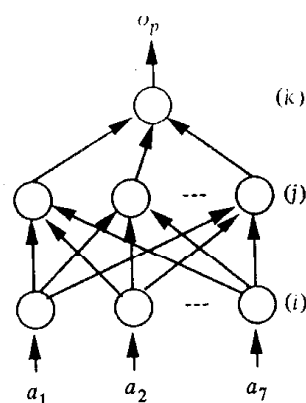


图 3 三层神经网络

(i)——输入层节点; (j)——隐含层节点; (k)——输出层节点; $a_i (i=1, 2, \dots, 7)$ 为网络的输入信号, o_p 为网络的输出信号

该函数的值在 0—1 之间. 其中 θ 系一可调参数, 用以修正该 S 形函数的形状, 即相当于改变其非线性度(图 2). 本文采用三层神经网络(图 3)进行对位取代苯酚的 QSAR 研究. 对位取代苯酚衍生物的结构通式为: $R-\text{C}_6\text{H}_4-\text{OH}$. 所收集到的样本共 26 个, 取自文献[1], R 为取代基. 本文采用的三层网络的输入层代表对位取代苯酚的结构参数及物理化学性质参数, 包括对位取代苯酚在 1-辛醇-水中的分配系数 $\log P_{ow}$, 形成氢键的能力 HB , 电场参数 F , 共振参数 R , 摩尔折射率 MR , 酸离解常数 pKa 以及空间位阻参数 σ_p 共 7 个参数^[1], 即输入层含有 7 个节点, 而输出层代表对位取代苯酚衍生物的生物活性 $\log(1/c50)$ ($c50$ 是对 *Tetrahymena pyriformis* 菌株的 50% 抑制生长浓度)^[2], 故输出层仅含 1 个节点. 依 Wold 的定义, 这是第 3 个水平的 QSAR 研究^[6].

在网络学习过程中, 一系列的输入输出学习模式提供给网络令其学习, 并依据一定的学习规则调节各层节点之间的连接权重以使网络的实际输出的 $\log(1/c50)$ 与期望输出相比较达到一定的精度要求. 根据误差反向传播(BP)算法^[7], 对于学习模式 p , 输出误差可表示为:

$$E_p = \frac{1}{2} (t_p - o_p)^2, \quad (3)$$

其中 t_p 是对于学习模式 p 输出节点(k)输出的 $\log(1/c50)$ 期望值, o_p 是网络实际输出的

$\log(1/c50)$ 值.

依 BP 算法中的 delta 规则, 输出层的 delta 误差项表示为:

$$\delta_p = (t_p - o_p) o_p (1 - o_p), \quad (4)$$

对于隐含层的节点 j , 其 delta 误差项可表示为:

$$\delta_{pj} = o_{pj}(1 - o_{pj}) \delta_p W_{kj}. \quad (5)$$

输出层产生的 delta 误差项通过迭代方式调节各层节点之间的连接权重在网络中反向传播, 直至所有输入输出学习模式中最大输出误差 $E_p(\max)$ 收敛至所要求的精度, 本文中以训练集循环次数表示.

连接权重调节为:

$$\Delta W_{ji}(q) = \varepsilon \delta_{pj} o_{pi} + \alpha \Delta W_{ji}(q-1), \quad (6)$$

此处, ΔW_{ji} 为隐含层中节点 j 与输入层中节点 i 之间连接权重的变化值(类似的公式亦适用于输出节点 k 与隐含节点 j 之间的情况); ε 为学习速率; α 为动量项; o_{pi} 为学习模式 p 在输入节点 j 的输出, 本文中 $o_{pi} = a_{pi}$, 即相当于输入层(仅输入层)中节点的输入输出转换函数取 $g(x) = x$; $q, q-1$ 分别表示当前迭代及前一次迭代.

在网络学习开始时, 各层节点之间的连接权重及偏置项均设置为任意小的非零随机数.

2 结果与讨论

2.1 网络参数的优化

为确定一组最佳网络参数用于训练, 使之能以尽可能少的迭代次数给出尽可能小的输出误差, 实验研究了数据预处理、学习速率、动量项、参数 θ 、隐含节点数等诸因素对网络性能的影响, 分别叙述如下.

2.1.1 数据预处理

对输入输出学习模式中的输入的对位取代苯酚衍生物的结构与物理化学性质参数 x_i 作如下归一化处理:

$$a_i = (x_i - x_{\min} + 0.01) / (x_{\max} - x_{\min} + 0.01), \quad (7)$$

其中 $x_{\max}$ 及 $x_{\min}$ 是对于参数 i 所有学习模式中输入数据的最大值及最小值, $i=1, 2, \dots, 7$.

而根据节点输入输出转换函数 $f(x)$ 的要求, 必须使网络的输出取值在 0—1 之间. 实际训练时, 为了使网络易于收敛, 对学习模式中输出的对位取代苯酚衍生物生物活性 y (即 $\log(1/c50)$) 作如下变换:

$$t_p = (y + 0.40) \times 0.40, \quad (8)$$

以使网络的期望输出值在 0.15—0.85 之间.

经如上预处理得到的 a_i, t_p 输入输出学习模式对构成训练集, 提供网络学习.

2.1.2 学习速率与动量项的影响

公式(6)中学习速率 ε 实际上系目标函数 E_p 在网络所有连接权重空间优化时的逼近步长. 通常并不要求对 ε 求精确解. 唯 ε 若过大, 则网络易陷于局部最优解, 且不易稳定; 而 ε 若过小, 则影响网络收敛速度. 为此, 本文提出如下经验规则以调节 ε 的大小. 当 $E_p(q) \leq E_p(q-1)$ 时, 置 $\varepsilon(q+1) = \varepsilon(q) \times 0.70$, 使学习速率下降; 当 $E_p(q) > E_p(q-1)$ 时, 置 $\varepsilon(q+1) = \varepsilon(q) \times 1.05$, 使学习速率稍增大, 这相当于使网络快速跨过当前不利的搜索空间. 显然, 网络收敛到接近目标时, 随 E_p 下降学习速率亦下降, 此时需保持 ε 不致过小而导致网络收敛过慢, 置最小学习速率 $= 0.10$. 实验证实使用此规则较使用固定学习速率时效果为佳, 主要表现在网络收敛速度加快, 且 QSAR 预测结果亦趋改善. ε 的初值对网络输出结果及迭代次数有一定影响, 因 ε 初值不同常导致网络在训练开始的第一次迭代后状态(即所有连接权重的取值)的不同, 随后的收敛过程也有较大差别. 研究表明, 设置如表 1 所示的 ε 初值, 网络性能较好.

表 1 ε 初值和 α 对网络性能的影响

序号	取代基 $R^{[1]}$	ε 初值	α	训练集循环次数	ANN 预测结果 ^{a)}	实际观测值 ^[1]
1	CONH ₂	0.40	0.50	3369	0.736	0.780
2	CH ₂ CH ₂ OH	0.40	0.50	36	0.910	0.830
3	NHCOCH ₃	0.30	0.50	2	0.849	0.820
4	CH ₂ CN	0.50	0.20	174	0.296	0.380
5	OCH ₃	0.40	0.20	87	0.101	0.140
6	CHO	0.40	0.30	145	0.549	0.270
7	COCH ₃	0.50	0.50	144	0.223	0.090
8	H	0.30	0.50	7244	0.265	0.430
9	COC ₂ H ₅	0.60	0.40	19	0.382	0.050
10	CN	0.50	0.50	154	0.431	0.520
11	F	0.60	0.40	41	0.163	0.020
12	OCH ₂ CH ₃	0.70	0.20	2961	0.157	0.010
13	NO ₂	0.40	0.60	14	0.975	1.420
14	CH ₃	0.80	0.60	553	0.160	0.190
15	Cl	0.50	0.50	354	0.551	0.550
16	CH ₂ CH ₃	0.60	0.30	5166	0.471	0.210
17	Br	0.50	0.60	398	0.627	0.680
18	I	0.40	0.50	1339	0.832	0.850
19	OC ₄ H ₉	0.40	0.60	380	0.662	0.700
20	CH(CH ₃) ₂	0.40	0.50	36	0.596	0.470
21	COC ₆ H ₅	0.40	0.50	524	0.998	1.020
22	C ₃ H ₇	0.50	0.60	354	0.685	0.630
23	C ₆ H ₅	0.50	0.50	1499	1.002	1.380
24	C(CH ₃) ₃	0.50	0.50	591	0.840	0.910
25	OC ₆ H ₅	0.50	0.70	898	1.355	1.360
26	CH ₂ C ₆ H ₅	0.40	0.80	1064	1.271	1.714

ANN 预测均方误差(mse) $= 0.036$

a) 网络节点的输入输出转换函数取 $f(x) = 1/(1 + e^{-x})$, 隐含节点数取 10.

动量项 α 的作用是滤除网络的高频振荡, 并保证网络在较低学习速率下亦具有一定的收敛速度. 本文结合初始学习速率的设置同时段优化网络动量项参数. 实验发现如表 1 所示的 ε 初值

与 α 的组合能使网络性能较好. 采用逐次抽取1个样本校验的方法(leave-one-out)对26个对位取代苯酚衍生物的生物活性进行定量预测, 即当预测某化合物的生物活性时, 其余25个化合物的输入输出学习模式对 (a_i, t_p) 构成训练集供网络学习, 直至收敛到一定精度, 然后用此网络对该化合物进行预测. 表1中亦列出了网络的预测结果及均方误差.

2.1.3 $f(x)$ 函数中参数 θ 的影响

θ 的大小决定 $f(x)$ 的非线性度, 故 θ 影响网络性能(表2). 由表2可知, 当 $\theta=1.0$ 时, 网络性能较佳. 而当 $\theta=2.0$, 即 $f(x)$ 函数的非线性度降低时, 网络性能明显下降, 说明本文研究的对位取代苯酚衍生物体系的生物活性与其结构、物化性质参数之间的关系是高度非线性的. 减小 θ 值似有利于改善网络性能, 但若 θ 过小(如取0.5), 则 $f(x)$ 函数的值在0—1之间有一急剧突变, 将导致网络在给出较高及较低生物活性输出时产生较大偏差, 网络亦不易收敛.

表2 参数 θ 对网络性能的影响

参数 θ	预测均方误差(mse)
0.5	0.044
1.0	0.036
2.0	0.050

2.1.4 隐含节点数的影响

对于不同的研究对象, 依其不同复杂程度及固有特性, 神经网络所需要的最优隐含节点数有较大差别, 目前在选取隐含节点数时尚无成熟理论规则可循. 本文考察了隐含节点数对网络性能的影响, 结果如图4所示. 当取隐含节点数为10时网络性能较佳. 再增加若干隐含节点(如取12, 14等), 网络性能亦可, 但训练需时较长. 而取过多隐含节点数时(如20以上), 则网络性能明显下降. 过多的隐含节点使网络的复杂化程度超过了对位取代苯酚衍生物体系本身固有的复杂程度, 而引进附加随机噪声. 如隐含节点数取值过小, 则网络的复杂程度不足以表征研究体系的固有特性, 亦使预测误差增大.

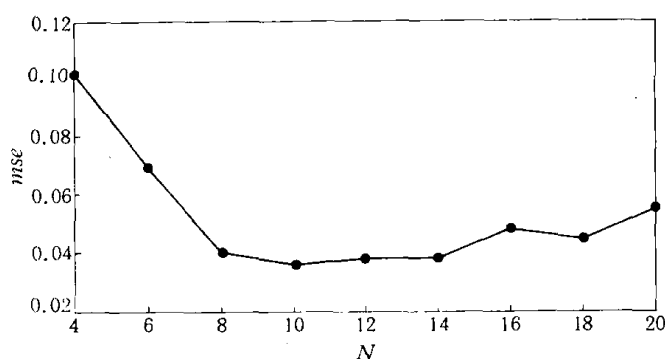


图4 隐含节点数(N)对预测均方误差(mse)的影响

2.1.5 影响网络收敛速度的因素

研究表明, 除初始学习速率及动量项外, 训练集中学习模式的排序亦影响网络收敛速度, 因调节网络各层节点之间连接权重时, 当前迭代(相应于当前学习模式)是以前一次迭代(相应于前一次学习模式)为基础的. 表1中列出的训练集循环次数系基于如下排序规则: 当预

测 No.1 化合物时, No.2, No.3, No.4, ..., No.26 依次排序构成训练集, 而预测 No.2 化合物时, 则以 No.3, No.4, No.5, ..., No.26, No.1 为序构成训练集, 依此循环类推. 实验发现学习模式如此排序时网络收敛速度较快. 本文用的是“NNSS”软件(长沙工学院提供), 在 Sun 3/260 工作站上运行, 训练集循环次数在几次到数千次变化时, 所需时间在几秒到几小时之内变化, 至于其间的变化规律, 主要取决于网络各参数及初始连接权重的设置, 尚无明确的理论可循.

2.1.6 过拟合现象

神经网络在训练过程中, “过拟合”问题是较难掌握的, 因很难决断网络在何时达到“正拟合”. 本文在考察网络性能时发现, 当训练集循环次数超过如表 1 所示的次数时, 网络性能反而下降, 表明网络过度拟合了训练集中学习模式提供的信息, 因而当一未知化合物提供给网络识别时, 可能给出较差的预测. 当然, 当网络训练不足时, 亦不能给出正确输出结果.

2.2 对位取代苯酚衍生物 QSAR 研究结果

Breyer^[1] 等人曾研究了对位取代苯酚衍生物的生物活性与其逆相液相色谱保留指数 K' 这一单一参数之间的回归关系, 提出了如下 QRAR 模型:

$$\log(1/c50) = 0.087 K' - 0.922 \quad (9)$$

文献[1]亦用逐步回归分析对此研究体系提出了 QSAR 模型:

$$\log(1/c50) = 0.668 \log P_{ow} - 0.405 pKa - 0.614R + 2.723 \quad (10)$$

本文则利用三层神经网络模型, 将文献[1]提供的对位取代苯酚衍生物的全部结构及物化性质参数作为输入信息以训练网络, 当取节点输入输出转换函数 $f(x) = \frac{1}{1+e^{-x}}$, 隐含节点数为 10 时, 获得了如表 1 所示的预测结果, 它优于模型(9)及(10)的预测结果(表 3).

表 3 ANN 模型与 QRAR, QSAR 模型比较^{a)}

模型	预测均方误差 (mse)
ANN	0.036
QRAR	0.051
QSAR	0.047

a) 均用 Leave-one-out 方法.

前已述及, 本文研究的对位取代苯酚衍生物的生物活性与其结构及物化性质参数之间的关系是高度非线性的, 因而用多元线性统计方法(如逐步回归分析)较难解决好这一问题. 文献[1]试图用对位取代苯酚衍生物的色谱保留指数来提供综合信息,

当它与结构参数 F (见前文)结合建立对位取代苯酚衍生物的生物活性预测模型时, 确能改善预报准确度. 但本文的研究结果表明, 结构及物化性质参数蕴含了足够的 QSAR 信息, 只是传统的统计方法难以最大限度地提取出这种信息. ANN 模型能利用对位取代苯酚衍生物体系的所有结构及物化性质参数, 尽管这些参数之间存在不同程度的相关关系, 但网络总能充分利用它们各自提供的独特信息, 亦能同时考虑它们之间的相互影响, 通过自适应学习, 对研究体系自动建立起非线性模型(由训练好的网络各层节点之间的连接权重来表示), 因而获得了较好的预测结果.

3 结 论

本文利用误差反向传播的人工神经网络模型对对位取代苯酚衍生物体系的定量构效关系

(QSAR)进行了研究,详细探讨了网络参数对网络性能的影响,提出了动态调节网络学习速率的经验规则,获得较满意结果.神经网络模型是具有自适应、自学习功能的高度非线性系统,因而用于非线性的QSAR体系,具有一定的优越性,是克服传统QSAR研究方法局限性的一种有用的新型研究方法.

参 考 文 献

- [1] Breyer, E. D., Strasters, J. K., Khaledi, M. G., *Anal. Chem.*, 1991, 63(8): 828.
- [2] Schultz, W., Riggin, G. W., Wesley, S. K., *QSAR in Environmental Toxicology-II* (ed. Kaiser, K. L.), Reidel Publishing Co., Dordrecht, The Netherlands, 1987, 333.
- [3] Yu, R.-Q., *Chemometrics and Intelligent Laboratory Systems*, 1992, 14: 23.
- [4] Rumelhart, D. E., McClelland, J. L., *Parallel Distributed processing*, Vols. 1 & 2, MIT Press, Bradford, 1986.
- [5] Zupan, J., Gasteiger, J., *Anal. Chim. Acta*, 1991, 248: 1.
- [6] Wold, S., Albano, C., Dunn III, W. J. *et al.*, in *Chemometrics: Mathematics and Statistics in Chemistry* (ed. Kowalski, B. R.), NATO ASI Series C: Mathematics and Physical Science, V. 138, p.17. D. Reidel Pub. Co., Dordrecht, 1984.
- [7] Pao, Y.-H., *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley Publishing Company, Inc. 1989, 120.