[文章编号] 1002 - 0624(2003)08 - 0019 - 03

松花江有毒有机物定量构效关系研究

刘永懋,刘　巍

(松辽流域水资源保护局,吉林 长春 130021)

[摘　要] 本文采用定量构效关系的研究方法,对松花江 45 种有毒有机物的毒性进行了预测与理论研究。文中主要对人工神经网络模型及分子连接性指数进行了研究,并对人工神经网络模型在预测松花江有毒有机物中的应用进行了阐述。

[关键词] 有毒有机物;定量构效;人工神经网络;松花江

[中图分类号] X522　　[文献标识码] B

以分子拓朴学研究为基础,对有机物分子结构与生物活性进行定量研究,采用拓朴指数——分子连接指数作为有机分子结构的描述符,同时,引进邻接矩阵和距离矩阵,使分子连接性指数的计算实现了软件化。

所谓定量构效关系,简称 QSAR,就是定量地描述、研究有机物的结构与活性之间的相互关系。对环境化学来说,它是一个科学有效的研究方法。

目前,生物毒性预测多数采用多元线性回归分析方法,建立构效关系方程来实现。但是,由于这种构效关系的复杂性和非线性,在很多情况下其误差较大,且对样本的选取有较高要求,人工神经网络技术以其可逼近任意非线性映射,并具有高度的容错等优点,在有机化学品生物活性预测研究中具有广阔的应用前景。

1　人工神经网络模型的研究

人工神经网络(简称 ANN)是最近发展起来的十分热门的交叉学科,它涉及生物、电子、计算机、数学和物理等学科领域内的知识,有着非常广泛的应用前景,这门学科的发展对目前和未来的科学技术的发展将有重要的影响。

1.1　基本原理与技术方法

BP 法神经网络为多层网络结构,分别由一个输入层、一个输出层、若干个隐含层构成,每一层有若干个神经元(节点),相邻层间的各神经元通过权重相连,同层内神经元无连接,运行算法可分为前向计算和反向误差传播 2 个过程,它们交替运行直至达到误差要求,其基本算法如下:

(1)网络结构的选择,本文采用 3 层网络结构;

(2)权值、阈值的初始化及样本输入;

(3)前向计算,计算各节点的输入 I_j 和输出 O_j:

$$I_j = \sum_i - W_{ij}O_I + \theta_j \qquad (1)$$

$$O_j = f(I_j) = (1 + e^{-I_j})^{-I} \qquad (2)$$

其中 $f(I_j)$为 S 函数,W_{ij} 为连接相值,θ_j 为阈值;

(4)计算误差,包括输出层节点误差 e_k 和样本平均误差 Ep:

$$e_k = O_k - Y_k\ , \qquad (3)$$

$$Ep = [\sum_p \sum_n e_k^2/(p \cdot n)]^{1/2}\ ; \qquad (4)$$

(5)反向误差传播,当 Ep 大于设定值时,调整修正各连接权值 W_{ij} 和阈值 θ_j:

$$W_{ij}(t) = W_{ij}(t-1) + \eta\delta_j O_i + \alpha\triangle W_{ij}(t-1) \qquad (5)$$

$$\theta_j(t) = \theta_j(t-1) + \eta\delta_j + \alpha\triangle \theta_j(t-1) \qquad (6)$$

式中 η 为学习步长;α 为动量因子 δ_j 为节点误差:

$$\delta_j = \begin{cases} O_j(1-O_j)(O_j - Y_j) & (\text{当}j\text{为输出节点时}) \\ O_j(1-O_j) & (\text{当}j\text{为隐含节点时}). \end{cases} \qquad (7)$$

1.2　在有机污染物生物活性预测中的应用

将上述神经网络系统用于松花江部分有机污染物生物活性的预测中,其结果列入表 1 中,对比分析的结果表明:本系统的预测精度明显优于多元回归分析(LR)的预测结果,见表 1。

采用"批样本"训练法,以测试样本误差的增量 Δe 为控制条件,有效地解决了网络训练中常见的过拟合和局部极小等问题;通过采取步长动态变化、激活函数与归一化公式的改进以及权值初始化范围增大等具体措施,实现了网络的快速收敛,提高了网络的预测精度。

2　分子连接性指数的研究

2.1　分子连接性指数的概述

分子连接性指数(简称 MCI)法,目前已建立了与溶解度、土壤吸附系统、分子表面积、分配系统、毒性、生物富集、因子与气相保留指数等关系式,在水环境方面也具有较好的应用前景。

[收稿日期] 2003 - 02 - 17

[作者简介] 刘永懋(1939—),男,吉林省德惠市人,教授级高级工程师,从事水资源学和水污染防治学研究。

表 1 部分有机污染物的生物活性预测结果

化合物名称	AT1		实验值* log(EC_{50})	LR		ANN	
	V_0	V_4		预测值	误差	预测值	误差
1,2,3—三氯苯	1.968	0.917	4.53	4.67	0.14	4.47	-0.06
2—丁醇	0.939	0.000	1.90	2.05	0.15	2.10	0.20
氯苯	1.480	0.224	3.86	3.58	-0.28	3.56	-0.30
2,5—二氯甲苯	1.930	0.879	4.38	4.57	0.19	4.51	0.13
对氯苯甲醛	1.801	0.738	4.15	4.26	0.11	4.21	0.06
2,4—二氯苯胺	1.865	0.863	4.09	4.38	0.29	4.34	0.25
苯胺	1.377	0.170	3.28	3.30	0.02	3.26	-0.0
2,4,5—三氯甲苯	2.174	1.347	4.86	5.04	0.18	4.85	-0.01
对氯甲苯	1.686	0.430	3.88	4.09	0.21	4.04	0.16
邻甲苯酚	1.557	0.360	3.75	3.74	-0.01	3.66	-0.09
2,4—二氯苯酚	1.839	0.946	4.45	4.25	-0.20	4.23	-0.22
1,2,4—三氯苯	1.968	0.673	4.50	4.81	0.31	4.69	0.19

2.1.1 分子连接性指数法的特点

(1) 以有机分子结构为基础建立的参数,不是实验值或经验值,因而能够客观全面地反映分子的结构;

(2) 计算参数的方法容易掌握,特别是将该拓扑指数实现计算机化后,有一定化学知识和数学知识者都能够掌握和应用;

(3) 方法具有较强的灵活性,能处理含杂原子、不饱和键、环及芳香类化合物等特殊分子,任何分子只要知道它的分子结构就可以计算出它的 MCI;

(4) 方法能借鉴量子力学的某些指数、扑法和分子轨道法的结合是建立 QSAR 的一条重要途径。

2.1.2 分子连接性指数法的表示方法

(1)结构图:采用拓扑结构来表示分子结构,即用点来表示碳或其它杂原子,用短线来表示连接键,即所谓隐氢图或拓扑图。

$$^0X = \sum_{i=1}^{n} (\delta_i)^{-\frac{1}{2}} \tag{8}$$

$$^1X = \sum (\delta_i\delta_j)^{-\frac{1}{2}} \tag{9}$$

$$^nX = \sum (\delta_i\delta_j\delta_k \wedge \delta_{n-1}\delta_n)^{-\frac{1}{2}} \tag{10}$$

其中:i, j, k 为原子编号;δ 为点价;X 为 MCI 指数。

(2)分解: 由拓扑图分解成各级子图,包括各级路径、簇、树(路径/簇)和环(链)。

(3) 计算:算出各非碳原子的点价,然后按式(8)~(10)计算出各阶指数。

2.1.3 分子连接性指数的计算

MCI 法在 QSAR 的研究、应用等方面得到了较好的成果,虽然其本身的思想及算法并不十分复杂,但对计算机来说却非常困难,直链还好,而象簇、环、树等结构就麻烦了,特别是当分子较大原子个数较多时,不仅识别判断非常困难而且计算量相当大。如何让计算机认识成千上万而结构又各不相同的分子,是算法实现的关键,无论是直链、簇、单环,还是树、多环及杂原子的分子结构,都能够识别计算。

(1)分子结构的输入:为了让计算机能够认识分子结构,需要对分子结构数学化,即用矩阵来表示分子结构,点价计算式为:$\delta = 4 - h_i$,可通过邻接矩阵来实现。

邻接矩阵:根据分子结构画出分子的拓扑结构图,凡是 2 个原子间有键相连(直接)即为 1,否则为 0。因其为对称矩阵,实际操作只需输入矩阵的一半,依此可确定含有 n 个原子的分子的 $n \times n$ 阶邻接矩阵。恰好有:

$$\delta = \sum_{j=1}^{n} a_{ij}$$

其中,a_{ij}—邻近矩阵元素。

(2) 编码规则:确定各原子的编号顺序:即①先编环原子,若为单环且起始位置任意,只要按照顺时针的次序就可以了,若为多环或为簇、树结构,则起始原为环上的簇原子,顺时针编号,一个环编完后再编另一个环(按顺时针方向),直至所有的环都编完为止;②再编簇原子,以族原子为中心再编与其距离为 1 的其它原子;③然后编树原子即与距离大于 1 的原子,则按顺时针方向编完一条支链后再编另一条;④最后是直链,由左到右、由上到下。

2.1.4 计算步骤

以异戊烷为例,大致计算过程为:(1)画出分子的拓扑结构,按照编码原则对分子结构进行编号,然后输入邻矩阵和距离矩阵;(2)计算点价,邻接矩阵中相应的行各元素的和即为点价,即$\delta_i = \sum_{i=1}^{n} a_{ij}$, n, a 为分子中的原子个数;(3)计算 MCI 是按照结构的不同分别计算,最后根据式(8)~(10)计算各阶指数。

通过对分子连接性指数的研究,并通过编码、修正等等手段,使其更具有一般性。同时,通过引进邻结矩阵和距离矩阵,使分子连接性指数的计算软件化,摆脱了人工计算的繁琐,提高了时效和精确度。化学与计算机、数学的结合,为环境科学的研究开拓了更广泛的空间,是有极其重要意义的。MCI 法在 QSAR 的研究、应用等方面得到了较好的成果。

3 模型的应用

模型建立起来,要用来预测松花江有毒有机物的生物毒性,为此,本文以 EC50 毒性指标,MCI 为关联因子对松花江中 45 种有毒有机物进行了预测,模型算出分子连接性指数,然后,用逐步回归法和 BP 法模型进行预测。分子连接性指数数量之多,很难关联且其中有许多次要的,甚至无用的指数,而且,其中一些指数之间还可能存在相关性,这一切都会影响到如何关联及关联的效果。如:含有一个三阶族的 6 原子(不算氢原子)的异已烷,它的 MCI 就多达 8 个,更复杂、原子个数更多的分子的 MCI 就更多。为此,采用逐步回归来进行主因子提取,同时,又可与 ANN 进行比较。

3.1 逐步回归法

逐步回归方法是根据各个自变量的重要性大小,每步选一个重要变量进入回归方程。第一步是在所有可供选择的变量中选出一个变量,是它组成的一元回归方程比其他的变量

有更大的回归平方和或更小的剩余平方和。第二步是在未选的变量中选一个变量，使它与已选的那个变量组成的二元回归方程比其他的变量与已选量组成的二元回归方程有更大的回归平方和。如此下去，直到无变量可选和可剔除为止。

3.2　主因子选取算法

(1)系数初始相关阵

设有 n 个自变量 X_j，应变量 y，k 个观测点。首先作出 $(n+1)\times(n+1)$ 的规格化的系数初始相关阵，即：

$$\begin{bmatrix} r_{00} & r_{01} & \cdots\cdots & r_{0,n-1} & r_{0,y} \\ r_{10} & r_{11} & \cdots\cdots & r_{1,n-1} & r_{1,y} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{n-1,0} & r_{n-1,1} & \cdots\cdots & r_{n-1,n-1} & r_{n-1,y} \\ r_{y0} & r_{y1} & \cdots\cdots & r_{y,n-1} & r_{yy} \end{bmatrix}$$

阵中各元素为：

$$r_{ij}=\frac{d_{ij}}{dd_i}=\frac{\sum_{t=0}^{k-1}(X-\bar{X})(X-\bar{X})}{\sqrt{\sum_{t=0}^{k-1}(x_{ij}-\bar{x}_j)^2}\sqrt{\sum_{t=0}^{k-1}(x_{ij}-\bar{x}_j)^2}} \qquad (11)$$

式中 $\bar{x}=\sum_{t=0}^{k-1}X_{ti}/k$，$ij=0,1,\cdots\cdots,n-1,n$。

(2)计算偏回归平方和

$V_i=r_{iy}r_{yi}/r_{ii}$，$i=0,1,\cdots\cdots,n-1$。

(3)因子筛选

选入：从所有 $V_i>0$ 的 V_i 中选出 $V_{\max}=\max|V_i|$，则其对应的因子为 $X_{\max}$，后检验它的显著性：若

$$\varphi V_{\min}r_{yy}<F_2$$

因子应选入，并对系数相关阵 R 进行该因子的消元变换，转第二步。

(4)结束

上述过程一直进行到无因子可选可剔为止，计算结束。算法框图见图 1。

3.3　生物毒性预测计算实例

(1) 生物毒性线性回归法预测

对 45 种有毒有机物的发光菌的 EC50 值与 12 种 MCI 拓扑指数进行逐步回归，得到如下方程：

$$-\log(EC50)=3.074X_0-0.595X_4-0.836 \qquad (12)$$

对全部 45 种有毒有机物进行了回归计算，结果略，尽管这 45 种物质的结构差异很大，但回归效果说明 MCI 可以在较宽的范围内定量描述 QSAR，而逐步回归所选取的主因子是成功的，确实代表了分子结构中与生物毒性相关的那部分。

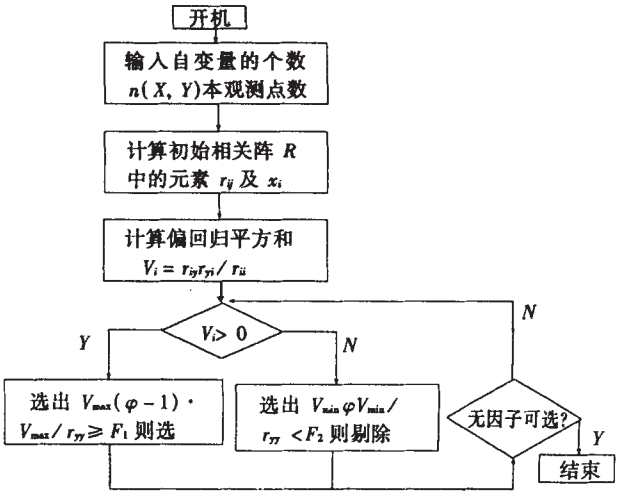

图 1 算法框图

(2)生物毒性人工神经网络法预测

根据自变量、因变量的个数确定网络的输入、输出节点数，采用三输入节点，$N_i=2$；输出节点，$N_0=1$。然后利用式(12)，算出隐节数：$N_h=3$。

传统 QSAR 关联的方法是多元线性回归，然而实际上结构与活性的关系是非线性的，这样不可避免会产生一定的方法误差，而神经网络 BP 法可以任意逼近非线性映射，为 QSAR 的研究提供了另一种研究手段，为证实它的可行性，特对 45 种有机化合物进行了计算。计算结果略。

比较逐步回归与 BP 法，其最大误差、平均误差、均方差分别为：逐步回归，$E_{\max}=0.78$，$E=0.174$，$s=0.150\ 11$；BP 网络，$E_{\max}=0.77$，$E=0.140$，$s=0.122\ 41$。

由此可见，用神经网络法进行预测明显好于线性回归(逐步回归)，说明神经网络用于 QSAR 的回归分析是可行的，具有较高的准确性。

通过对 45 种不同分子结构的有毒有机物进行预测，可以看到人工神经网络用于 QSAR 的研究是成功的，它具有比传统的线性回归方法更大的优越性；通过采用误差增量控制训练，有效地解决了网络过拟合和局部极小问题。

QSAR 应用于松花江中存在的 45 种有毒有机物的毒性预测中，取得了与测定值完全一致的好结果。可见，其技术方法在河流水域中具有普遍地应用意义。

(上接第 18 页)

均小于 3.5%。

4　结　论

根据模型试验资料统计分析，表明引进的量水短管下卧后，其量水性能受到了不同程度的影响，总结其要点有：圆弧型进口短管，因其量水误差已超过 5%，而且 μ_k 值随卧深与卧长的变化摆动明显，水头损失大，所以不宜使用；文丘利短管量水，下卧后量水精度不受影响，量水误差小于 5%，水头损失略有增加，但仍可满足灌区量水要求，其 μ_k 值按式(5)计算，测流比 $Q_{\max}/Q_{\min}$ 控制在 4 以内为好。

Study on artificial nerve network method of surrounding rock classification for hydraulic tunnel

YE Bao – min, LI Xing – wen, ZHANG Jian – hua

[**Abstract**] The hydraulic tunnel surrounding rock classification, being a nonlinearity and uncertainty question, is resolved by artificial nerve network principle, a reliable way for hydraulic tunnel surrounding rock classification is provided by means of setting up mechanics index of hydraulic tunnel surrounding rock and BP determinant model of environment factor.

[**Key words**] artificial nerve network; hydraulic tunnel; surrounding rock; classification

Water sail machine and water sail power station

WANG Ji – tang

[**Abstract**] The paper quantitatively explains the basic law of energy conversion for water sail machine by the mathematical method, introduces the basic technical features of first water sail power station at home and abroad. The water sail power station breaks through the basic technical features of the traditional hydraulic machine and the water power station, simplifies the structure of water power station, and reduces the cost of kW about 50%.

[**Key words**] water sail; water sail power station; energy conversion

Test of affect moisture motion in soil boy by vacuum negative pressure

ZHAO Bin, DONG Zheng – chuan, TAN Ying

[**Abstract**] The paper discusses the influence of vacuum negative pressure to the moisture motion in soil boy and the conditions of speed drainage by means of model test. In view of this test, Darcy law is applied to the unsaturated and the internal water motion, the unsaturated seepage flow law is discussed.

[**Key words**] vacuum negative pressure; speed drainage; Darcy low; moisture motion

Study on quantitative structure – activity relationship of toxic organic compounds in Songhuajiang river

LIU Yong – mao, LIU Wei

[**Abstract**] The paper forecasts and studies the toxicity of 45 noxious organic compounds in Songhuajiang river by the study method of quantitative structure – activity relationship. The paper studies the artificial nerve network model and the molecule linking index and explains the application of the network model in the forecast of noxious organic compounds in Songhuajiang river.

[**Key words**] noxious organic compounds; quantitative structure – activity relationship; artificial nerve network; Songhuajiang river