第33卷 第3期 贵州工业大学学报(自然科学版) Vol.33 No.3
2004年 6月 JOURNAL OF GUIZHOU UNIVERSITY OF TECHNOLOGY June.2004
(Natural Science Edition)

文章编号:1009-0193(2004)03-0031-04

BP神经网络在氯代羟基苯甲醛QSRR研究中的应用

张军方[1],陈梦瑜[1],罗妮娜[2],陈 桐[1]

(1.贵州工业大学 理化分析中心,贵州 贵阳 550003;2.浙江大学 环境科学系,浙江 杭州 310028)

摘 要:用BP神经网络法对氯代羟基苯甲醛的定量结构-色谱保留值关系(QSRR)进行了研究,采用多元回归分析对此类化合物的QSRR进行了比较研究,并用因子分析法的斜交因子得分图直观地揭示了上述两法在运用中的优劣差异。

关键词:BP神经网络;定量结构-色谱保留值关系;氯代羟基苯甲醛;因子分析法

中图分类号:TP183;Q62 **文献标识码**:B

0 引 言

近年来,人工神经网络法(artificial neural network,简称ANN)作为化学计量学的一种新方法在色谱理论研究中发挥着重要作用。例如,人工神经网络法已在色谱保留值的预测[1]、色谱实验条件的优化、色谱谱图处理[2,3]等方面得到了很好的应用。

有机化合物的定量结构-色谱保留值关系(QSRR)是近年来色谱理论研究的一个热点,它对于预测保留值、选择分离条件以及探索色谱保留机制等具有重要意义[3]。本文运用ANN法中的变步长反向传播(BP)算法,由辛醇-水分配系数($\log K_{ow}$)、官能团位置指数(S_{ox})作为输入向量,以化合物的色谱保留值作为输出向量,对氯代羟基苯甲醛的QSRR进行了研究。同时采用多元回归分析和因子分析法对此类化合物的QSRR进行了比较研究,结果表明,用ANN研究QSRR可以很好地改善多因素拟合时线性程度较低而导致的模型不稳定的问题。

1 方法原理

1.1 辛醇—水分配系数($\log K_{ow}$)

辛醇—水分配系数(Octanol-Water Partition Coefficient)$\log K_{ow}$是一种亲脂性或疏水性健合的量度参数,研究已表明,它与气相色谱保留值具有一定的相关性[4]。并且$\log K_{ow}$与分子的结构直接相关,包括分子的大小,分子的柔韧性、极性,分子之间的氢键等[5],通常化合物的$\log K_{ow}$越大,极性越小。从而它可代替结构参数,预测化合物的色谱行为。本文采用分子片法[6]计算了氯代羟基苯甲醛的辛醇-水分配系数,如下式所示:

$$\log K_{ow} = a + \sum_i b_i B_i + \sum_j c_j C_j \tag{1}$$

其中a是常数,一般取-0.703;b_i代表分子结构中分子片的数目;B_i代表分子片的贡献率;c_j代表分子片需要被修正的数目;C_j代表经过修正后的贡献率。以上参数可由文献[6]获得。

1.2 官能团距离指数(S_{ox})

官能团距离指数是表示化合物分子中的官能团位置(Site)指数,S_{ox}是表示指定的官能团到其它各个原子的距离之和。

$$S_{ox} = \sum d_i \quad (i \text{ 除氢以外的碳、氧、氯等的原子位置序号}) \tag{2}$$

本文参考文献[7],以羟基为指定的官能团。例如:由2-氯-对羟基苯甲醛原子距离定位图(图1),得到$S_{ox}=1+2+2+3+3+3+4+5+6=29$.

收稿日期:2004-03-15

1.3 BP神经网络算法

算法参考文献[8]中的Levenberg－Marguardt BP(LMBP)算法，采用三层(输入层、隐层，输出层)BP网络来描述输入与输出之间的非线性映射。隐层的传递函数为对数S型(sigmoid)函数$f1(x)=1/(1+e^{-x})$，输出层传递函数$f2(x)$为线性激励(purelin)函数。

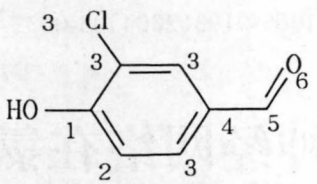

图1 2－氯－对羟基苯甲醛原子距离定位图

1.4 仪器和软件

联想Pentium－MMX 64RAM计算机。神经网络程序在Matlab 5.3上完成，多元回归分析及因子分析在SPSS统计软件上完成。

1.5 数据

氯代羟基苯甲醛的实验气相色谱保留值取自Kornonen的研究[9,10]，实验结果是在非极性柱SE－30柱上，160℃取得的，按前述方法计算的$\log K_{ow}$与S_{ax}和化合物的气相色谱保留值列于表1.运行神经网络程序时以$\log K_{ow}$、S_{ax}作为输入向量，以化合物的色谱保留值作为输出向量。

表1 25种氯代羟基苯甲醛的I、$\log K_{ow}$和S_{ax}值及不同模型下的预测值

序号	化合物	I*	$\log_w$	S_{ax}	*Factor*1	*Factor*2	*ANN*法		*MLR*法	
							$I^{pre}1$	*RE*1(%)	$I^{pre}2$	*RE*2(%)
1	对羟基苯甲醛	1320	1.219	26	－0.6489	－1.9151	1331	0.9	1270	－3.8
2	2－氯－对羟基苯甲醛	1524	1.873	30	0.4289	－1.0239	1536	0.8	1423	－6.6
3	3－氯－对羟基苯甲醛	1291	1.873	29	－0.7299	－0.8461	1302	0.8	1376	6.6
4	2,3－二氯－对羟基苯甲醛	1463	2.527	33	0.1997	0.0803	1460	－0.2	1529	4.5
5	2,5－二氯－对羟基苯甲醛	1449	2.527	33	0.1349	0.0956	1465	1.1	1529	5.5
6	2,6－二氯－对羟基苯甲醛	1753	2.527	34	1.6226	－0.1600	1695	－3.3	1577	－10.1
7	3,5－二氯－对羟基苯甲醛	1465	2.527	32	0.1294	0.0006	1340	－8.6	1482	1.1
8	2,3,5－三氯－对羟基苯甲醛	1632	3.181	36	1.0358	0.9324	1632	0.0	1635	0.2
9	2,3,6－三氯－对羟基苯甲醛	1651	3.181	37	1.2035	0.9891	1621	－1.8	1682	1.9
10	2,3,5,6－四氯－对羟基苯甲醛	1820	3.835	40	2.0396	1.8412	1756	－3.5	1788	－1.7
11	邻羟基苯甲醛	1062	1.219	22	－2.1625	－1.9424	1073	1.0	1080	1.7
12	3－氯－邻羟基苯甲醛	1264	1.873	25	－1.1735	－1.1264	1232	－2.5	1186	－6.2
13	6－氯－邻羟基苯甲醛	1214	1.873	26	－1.3255	－0.9941	1212	－0.2	1233	1.6
14	5－氯－邻羟基苯甲醛	1206	1.873	27	－1.2829	－0.9079	1201	－0.5	1281	6.2
15	3,4－二氯－邻羟基苯甲醛	1434	2.527	29	－0.2532	－0.1979	1388	－3.2	1339	－6.6
16	3,5－二氯－邻羟基苯甲醛	1388	2.527	30	－0.3866	－0.0700	1365	－1.7	1386	－0.1
17	3,6－二氯－邻羟基苯甲醛	1387	2.527	29	－0.4709	－0.1464	1413	1.8	1339	－3.5
18	5,6－二氯－邻羟基苯甲醛	1376	2.527	31	－0.3625	0.0206	1353	－1.6	1434	4.2
19	4,5－二氯－邻羟基苯甲醛	1358	2.527	31	－0.4459	0.0404	1360	0.2	1434	5.6
20	4,6－二氯－邻羟基苯甲醛	1326	2.527	30	－0.6737	－0.0021	1382	4.2	1386	4.6
21	3,4,5－三氯－邻羟基苯甲醛	1577	3.181	34	0.6218	0.8377	1529	－3.1	1540	－2.4
22	3,4,6－三氯－邻羟基苯甲醛	1529	3.181	33	0.3198	0.8128	1562	2.2	1492	－2.4
23	3,5,6－三氯－邻羟基苯甲醛	1522	3.181	34	0.3671	0.8980	1542	1.3	1540	1.2
24	4,5,6－三氯－邻羟基苯甲醛	1510	3.181	35	0.3911	0.9886	1589	5.3	1587	5.1
25	3,4,5,6－四氯－邻羟基苯甲醛	1721	3.835	38	1.4217	1.7947	1718	－0.2	1693	－1.6

* 取自文献[9,10]

2 结果与讨论

2.1 ANN隐含层节点数的优化

隐含层节点数过少，网络不能反映输入节点与输出节点间的复杂函数关系，因而测试误差较大；但隐含层节点数过大时，会将实验误差也作为函数关系的非线性因素引入网络，这样的网络也不能反映输入向量与输出向量间的实质性关系，因而会导致测试误差增大(超拟合)。用1－7个隐含层节点对表1中25个样本组成的训练集进行网络优化，其优化曲线可用测试误差均方根(root mean square error, RMS)对隐含层节点数作图获得，隐含层节点数为3时比较合适。

2.2　最小训练集样本数的计算

训练集中的样本数不能过少,过少则可能使网络的归纳能力变差,进而使测试误差增大。按照 ANN 理论,训练集里的最小样本数可按下式计算:

$$n_{\min} = mn + ny + n + y \tag{3}$$

式中,m、n、y 分别为输入层、隐含层、和输出层节点数。本文 BP 网络结构对应的最小样本数为 $2\times3+3\times1+3+1=13$,因而本文选取的样本数 25 大于最小训练集样本数。

2.3　BP 网络训练结果

设置网络的目标误差为 0.005,学习速率设置为 0.01,训练次数为 1000.训练好的网络中,各节点的连接权值和偏置值见表 2 和表 3.

表 2　输入层到隐含层的连接权值和偏置值

隐含层节点	输入层节点		偏置值
	1	2	
1	0.2635	91.2985	−1.1073
2	−0.0639	0.160	−2.0991
3	0.6779	23.8952	10.1488

表 3　隐含层到输出层的连接权值和偏置值

隐含层节点	1	2	3	偏置偏
权值	−0.8720	105.6657	−2.4700	−8.8032

2.4　BP 网络模型与多元线性回归模型的稳定性比较

用去一法(leave − one − out)检验 BP 网络模型的稳定性。即每次从 25 个样本中随机选出一个样本作为预测集,其余 24 个样本作为训练集进行模型训练,预测结果及其与实验值的相对误差(RE1)见表 1.再对 25 个化合物用 $\log K_{ow}$、S_{ox} 及气相色谱保留值 I 按以下模型进行多元线性回归:

$$I = a\log K_{ow} + bS_{ox} + c \tag{4}$$

得到回归方程及复相关系数 r 如下:

$$I = -56.428\log K_{ow} + 47.548S_{ox} + 102.638 \quad r = 0.928 \tag{5}$$

同样采用去一法得到多元线性回归模型下的 25 个预测值及其与实验值的相对误差($RE2$)(见表 1)。

表 1 中,比较 BP 网络模型的预测值与多元线性回归模型的预测值发现,$RE1>5\%$ 的只有 2 个,而 $RE2>5\%$ 的却达到了 8 个,其中相对误差的绝对值最大达 10.1%,这说明用本文建立的 BP 网络模型预测氯代羟基苯甲醛的气相色谱保留值比多元线性回归方法更可靠,模型更稳定。

2.5　因子分析法的应用

在 SPSS11.0 统计软件中,我们对 25 种氯代羟基苯甲醛化合物的三个参数 $\log K_{ow}$、S_{ox} 及 I 作因子分析。通过 Promax 斜旋转,提取两个因子作为主因子,得到了化合物的斜交因子得分值(见表 1).两个主因子包含了全部信息的 98.612 %,基本上反映全部信息。斜交因子得分值分布如图 2 所示。

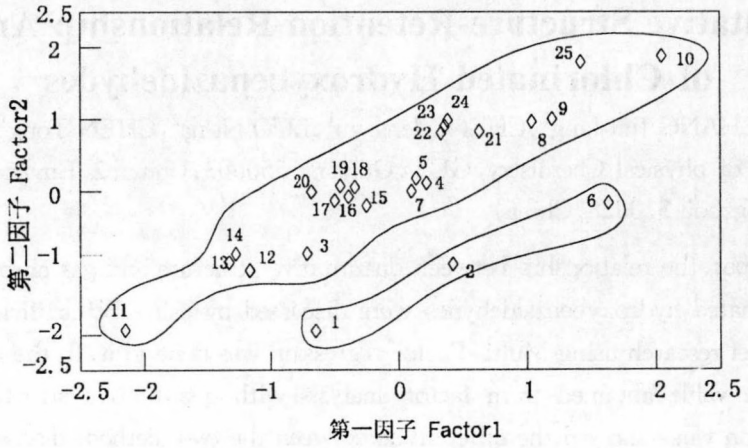

图 2　斜交因子得分图

由图 2 可明显看出,斜交因子得分值将 25 种氯代羟基苯甲醛分为界限清晰的两类,也即是参数

$\log K_{ow}$、S_{ax}对 I 的影响大小分为了重要性不一样的两类。在这样的情况下,运用线性回归模型将使结果不理想,表 1 中有 8 个 $RE2$ 达 5% 以上就说明这一点。然而,我们运用 ANN(BP 网络)可以较好地改善这一问题,在 ANN(BP 网络)得到的观测值的 $RE1>5\%$ 的个数只有 2 个。这也说明 ANN 可以成功地应用于非线性函数映射,对多元非线性函数可以显示其独特的优越性。

3 结 论

本文用 BP 神经网络法对氯代羟基苯甲醛的定量结构 - 色谱保留值关系(QSRR)进行了研究;采用多元回归分析对此类化合物的 QSRR 进行了比较研究,并用因子分析法的斜交得分因子图直观地揭示了上述两法在运用中的优劣差异。

ANN(BP 网络)的运行结果表明,由辛醇 - 水分配系数($\log K_{ow}$)、官能团位置指数(S_{ax})作为输入向量,以化合物的色谱保留值作为输出向量构建的网络稳定性良好。采用多元回归分析对此类化合物的 QSRR 进行的比较研究表明 ANN(BP 网络)可以很好地改善多因素拟合时线性程度较低而导致的模型不稳定的问题。

参考文献:

[1] 史雪岩,陈祥光,傅若农,等.人工神经网络在气相色谱保留值预测上的应用[J].分析科学学报,2000,16(3):196 - 200.
[2] 付大友,何瑾,袁东.人工神经网络在色谱中的应用[J].四川轻化工学院学报,2000,13(4):43 - 47.
[3] 马 波,周在德,李梦龙.人工神经网络及其在色谱中的应用[J].化学研究与应用,2000,12(4):375 - 379.
[4] 王连生,等.有机物定量结构—活性相关[M].北京:中国环境科学出版社,1993.
[5] 张锡辉.高等环境化学与微生物学原理及应用[M].北京:化学工业出版社,2001.
[6] Klopman G, Wang S. A computer automated Structure evaluation (CASE) approach to calculated partition funcions[J]. J Comput Chem, 1991, 12(8): 1025 - 1032.
[7] 杨 林.酚的拓扑指数在色谱分析中的应用[J].化学通报,2000,63(6):50 - 51.
[8] Martin T H, Howard B D, Mark H B. 神经网络设计[M].北京:机械工业出版社,2002.
[9] Kornonen I O O. Gas - liquid chromatographic analyses. XXXIX. ω - chloroethanols on low - polarity (SE - 30) and polar (OV - 351) capillary columns[J]. J Chromatogr, 1985, 324(1): 181 - 191.
[10] Kornonen I O O. Gas - liquid chromatographic analyses. XXXI. Retention increments of isomeric chlorophenols on low - polarity (SE - 30) and polar (FFAP) capillary columns[J]. J Chromatogr, 1984, 315(1): 185 - 200.

Application of BP Artificial Neural Network in Quantitative Structure-Retention Relationship Analysis of Chlorinated Hydroxybenazldehydes

ZHANG Jun-fang[1], CHEN Meng-yu[1], LUO Ni-na[2], CHEN Tong[1]

(1. Analysis Center of physical Chemistry, GUT, Guiyang 550003, China; 2. Environment Science Department, ZJU, Hangzhou 310028, China)

Abstract: In this paper, the relationship between quantitative structure and gas chromatographic retention value of chlorinated hydroxybenzaldehydes were discussed by using BP artificial neural network. And then the parallel research using Multi -Factor regression was done to it. In the end, the diagram of oblique factor score value abtained from factor analysis with quantitative structure and gas chromatographic retention value showed the differentiae between the two methods discussed above.

Key words: BP artificial neural network; QSRR; chlorinated hydroxybenzaldehydes; factor analysis