第26卷, 第5期
2006年5月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 26, No. 5, pp908-912
May, 2006

有机环境污染物紫外光谱检索的神经网络方法

王凤霞, 张卓勇*, 王亚敏

首都师范大学化学系, 资源环境与地理信息系统北京市重点实验室, 北京 100037

摘 要 详细讨论了网络优化参数、模拟的测量过程中噪声及杂质对网络收敛性能及预测误差的影响。为加速网络收敛, 提高紫外光谱检索的正确率, 采用了导数光谱对反向传播的人工神经网络(BP-ANN)进行训练和检索, 该方法对检索光谱中噪声、杂质, 尤其是斜坡背景的允许程度明显提高。文章还将ANN方法与普通的相关系数法的识别结果进行了比较。结果表明, 优化参数下的人工神经网络的库检索法在抗噪、容杂等方面都明显地优于普通的相关系数法, 是一种很有效的紫外库检索方法。

主题词 人工神经网络; 有机环境污染物; 紫外光谱; 库检索
中图分类号: O657.3 文献标识码: A 文章编号: 1000-0593(2006)05-0908-05

引 言

由于有机污染物的紫外光谱间存在严重的重叠, 加之光谱测量过程中噪声和杂质等的影响, 使得紫外光谱的库检索非常困难, 如何发展一种快速、方便、准确的紫外谱库检索方法是一个需要研究的问题。目前, 对有机污染物紫外光谱的检索通常采用相关系数法, 即在一个已建的紫外谱库中进行搜索, 通过计算检索光谱与库光谱之间的相关系数, 根据它们的相似性对未知化合物作出识别。但相关系数法受背景吸收、噪声、基线漂移、杂质含量、检测信号的非线性和仪器差异等因素影响较大[1], 而且这些因素的影响程度又无法测量, 因此对最后的检测结果影响较大, 甚至有时可能导致检索的错误。

反向传播的人工神经网络(BP-ANN), 对输入信号和输出信号间的联系能够进行学习和记忆, 具有较强的输入输出映射能力、泛化能力、容错能力及较好的鲁棒性。实验证明, 人工神经网络方法用于谱库检索能消除或减轻上述因素的影响, 取得良好的识别效果[2-6]。本文将BP-ANN用于有机污染物紫外谱库检索, 对该网络的优化参数作了详细的讨论, 针对实际检测中存在的噪声、基线漂移、斜坡背景、含杂等不同情况进行了模拟, 并将该方法与传统方法的识别结果进行了比较。结果表明, 优化参数下的人工神经网络的库检索方法不仅方便、快速, 而且具有更强的抗噪和容杂能力。

1 实验部分

1.1 仪器和试剂

756MC型UV-Vis分光光度计; 所有的计算均在有256兆内存, 奔腾Ⅳ1.8 G处理器的个人电脑上完成。本研究采用反向传播算法的前馈神经网络(BP-ANN)。采用的化合物均为美国国家环保局和中国国家环保局规定的有机环境污染物, 其中包括苯、二甲苯、硝基苯、二硝基甲苯、氯苯、氯酚、硝基酚等27种化合物。所用化合物均为分析纯级以上。每个化合物均以乙醇为溶剂, 配制成稀溶液进行测定。

1.2 数据的准备和预处理

输入数据的选择和处理: 选取212~400 nm紫外光谱区作为鉴别有机污染物的波长范围, 从212 nm开始每隔2 nm记录一个点, 共得95个点的吸光度值。为加速网络收敛, 提高其处理速度, 将上述数据利用小波函数压缩成48个点, 并将压缩后的光谱数据进行归一化, 然后输入网络, 因此输入层神经元的个数为48。

输出数据的选择: 为了突出样本特征, 拉大不同光谱之间的特征差距, 从而利于分类。本实验采用向量一维子空间的27元素作为网络的期望输出, 输出阈值设为0.8和0.2, 对于谱库中的某化合物, 如果输出层对应节点的值大于0.8, 其余节点的值都小于0.2判为识别正确; 所有节点的输出值均小于0.2判为识别错误; 其余所有的情况判为不确定, 因此输出层神经元的个数为27(谱库中共有27个光谱)。

1.3 噪声与杂质的模拟

白噪声即先生成一个随机函数, 然后乘以适当的噪声水

收稿日期: 2005-01-16, 修订日期: 2005-04-26
基金项目: 首都师范大学资源环境与地理信息系统北京市重点实验室开放基金项目(2004211-03)资助
作者简介: 王凤霞, 女, 1979年生, 首都师范大学化学系硕士研究生 * 通讯联系人

平加到训练和检测光谱中, 白噪声的平均值为 0; 偏置噪声即噪声的平均值等于偏置水平, 是在白噪声的基础上生成的, 本实验的偏置水平为 0.08。斜坡噪声由一个线性函数($y = a \times x$)生成, 然后加到待检索的光谱中。杂质即在原光谱的适当位置加上一定杂质水平的正态分布的小峰。本研究中分别在测量范围内的左中右三个位置加入三个谱峰作为杂质加到待检索的光谱中。

2　结果与讨论

2.1　网络参数的优化

2.1.1　隐含层神经元个数

在本实验所用的 BP-ANN 是通常采用的三层结构, 网络输入值是原光谱经小波函数压缩成的 48 个吸光度值, 因此输入层神经元的个数为 48; 输出层每个神经元对应一个光谱, 因此输出层神经元的个数为 27; 因而在该神经网络中, 输入、输出层神经元的个数都已经确定, 隐含层神经元的个数将决定该网络的结构和优劣。对于隐含层神经元的个数与输入、输出层神经元的个数的关系已有许多讨论[7], 但隐含层神经元的个数不仅与输入输出层神经元的个数有关, 还受具体样本噪声的大小及样本中蕴含规律的复杂程度的影响[8, 9], 为寻找本实验最佳隐含层神经元的个数, 在参照理论计算的前提下, 通过实验的方法来确定。图 1 表示了预测误差随隐含层神经元个数的变化情况。从图 1 可以看出本实验最佳的隐含层神经元个数为 35。

2.1.2　学习速率

学习速率是影响神经网络收敛和性能的重要参数, 学习速率选择太大会出现算法不收敛, 选择太小训练过程又太长, 为确定本实验最优的学习速率, 讨论了预测误差随学习速率变化的情况见图 2。可以看出本实验最佳的学习速率为 0.2。

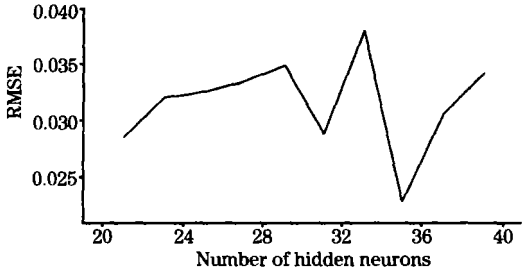

Fig 1　Effect of hidden neurons on network

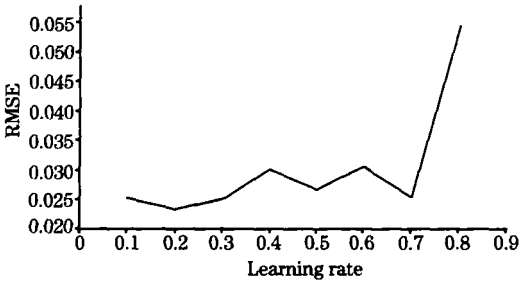

Fig 2　Effect of learning rate on network

2.1.3　动量因子

动量因子是影响神经网络性能的另一个重要参数, 动量因子太小网络处理速度太慢, 而且预测误差大, 动量因子太大网络又极易出现振荡陷入局部最小, 为了寻找本实验最优的动量因子, 作出了预测误差随动量因子的变化图(见图3)。从图 3 可以看出本实验最优的动量因子为 0.8。

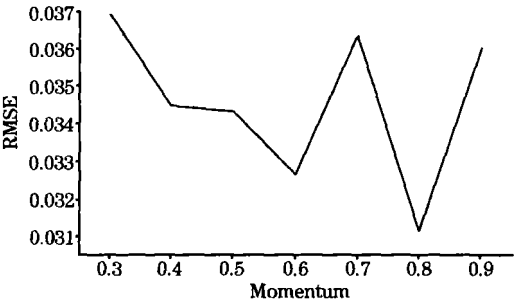

Fig 3　Effect of momentum on network

2.1.4　学习次数

学习次数是网络优化过程中必须调节的一个重要参数, 学习次数太少网络无法达到收敛; 学习次数太多, 网络就会出现"过拟合"问题, 不仅降低了其泛化能力, 增大预测误差, 而且还会延长训练时间。图 4 显示了预测误差随学习次数的变化情况。从图 4 可以看出随着学习次数的增大, 预测误差先变小, 而后又变大, 中间出现了最小值, 因此最小误差对应的学习次数即为本研究的最佳的学习次数。因此本实验最佳的学习次数为12 000。

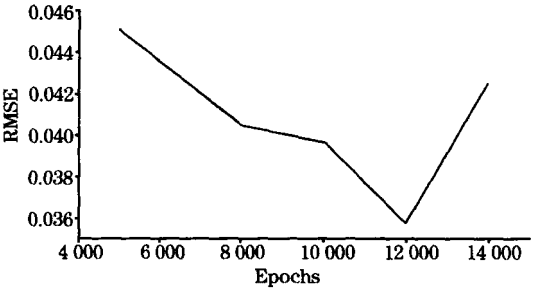

Fig 4　Effect of epochs on network

由以上对参数的讨论可知本组最优的实验参数: 动量因子 0.8, 学习速率 0.2, 隐含层神经元的个数 35, 学习次数 12 000。

2.1.5　优化参数的整体选择

本实验采用的神经网络的优化参数包括隐含层神经元的个数, 动量因子, 学习速率及学习次数, 为了考虑它们之间的相互作用对网络的影响, 本实验根据均匀设计表 $U17(U17^{16})$[10] 在 16 个水平上安排四个参数运行网络, 进一步验证上述参数是否为最优。表 1 给出了参数详细的组合情况。图 5 显示了网络运行参数对预测误差的影响。

由图 5 可知均匀设计法得到的本组最优参数即为水平实验第 5 号: 动量因子 0.7, 学习速率 0.8, 隐含层神经元的个数 15, 学习次数 16 000, 且从图 5 可以看出在该组优化参数下网络的预测误差稍小于在前面优化参数下网络的预测误差

Table 1 U17 (1716) experiment of library search of UV spectra by neural network

Experiment number	Network parameters			
	Momentum	Learning rate	Number of hidden neurons	Epochs
1	0.3	0.2	39	16 000
2	0.4	0.3	33	12 000
3	0.5	0.5	27	8 000
4	0.6	0.6	21	4 000
5	0.7	0.8	15	16 000
6	0.8	0.1	43	12 000
7	0.9	0.2	37	8 000
8	0.3	0.4	31	4 000
9	0.4	0.5	25	18 000
10	0.5	0.7	19	14 000
11	0.6	0.8	13	10 000
12	0.7	0.1	41	6 000
13	0.8	0.3	35	18 000
14	0.9	0.4	29	14 000
15	0.7	0.6	23	10 000
16	0.8	0.7	17	6 000

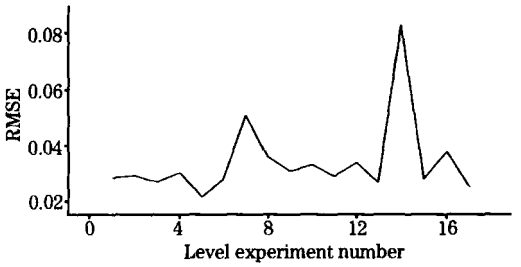

Fig 5 Effect of parameters on network

Note: The 17th point represents optimization parameter obtained with foregoing discussion

(为直观地显示出在两组优化参数下网络的预测误差的相对大小, 将在前面优化参数下对应的预测误差表示在图 5 的第 17 个点)。但是在进一步分析中发现如果学习速率大于 0.4, 对含高水平噪声的光谱预测效果较差, 且网络不稳定易陷入局部最小; 如果学习次数大于 12 000, 随着学习次数增加, 训练误差逐渐变小, 预测误差反而增大, 而且训练时间明显变长。

综合各方面可知, 本实验选用的最优参数: 动量因子 0.8, 学习速率 0.2, 隐含层神经元的个数 35, 学习次数 12 000。

2.2 用常规光谱作训练

采用 27 个有机环境污染物的常规紫外光谱对 BP-ANN 作训练, 对含不同噪声水平的光谱作识别, 所得的结果如图 6所示。从图 6 可知用纯光谱作训练, 对含低水平噪声如 12% 以下的光谱识别效果较好均为 100%; 随着检索光谱中噪声的增加, 识别率越来越低; 而在训练光谱中加入适当水平的噪声如 4%~8% 时, 会提高噪声光谱的检出水平和检索正确率; 但如果训练光谱中的噪声水平过大, 反而导致检索正确率降低, 甚至检索错误。

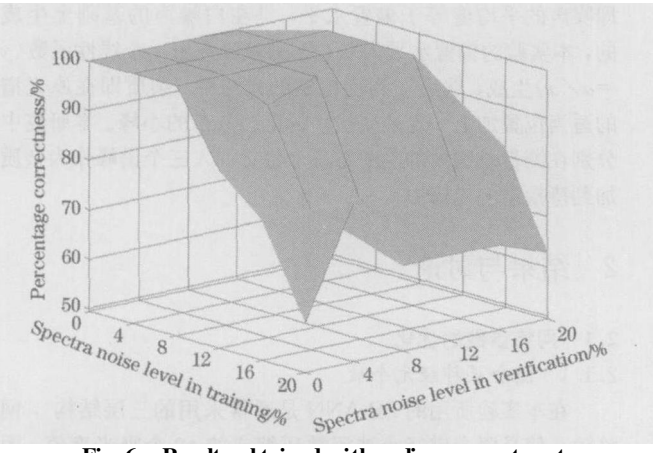

Fig 6 Results obtained with ordinary spectra at different noise levels (with white noise)

2.3 用常规混合光谱作训练

用含有随机函数生成的三个具有同一白噪声水平的光谱再加上原来未加噪声的光谱(共 108 个光谱) 对 BP-ANN 作训练, 然后用于含白噪声光谱的检索, 所得的结果见图 7。用含有 0.08 偏置噪声的光谱加上原来未加噪声的光谱(共 108) 对 BP-ANN 作训练, 用于含偏置噪声的光谱的识别, 所得的结果见图 8。从图 7、图 8 可以看出, 用含噪声的混合光谱作训练, 可提高网络的抗噪能力, 使网络对含高水平噪声光谱有好的识别效果。特别是以含适当噪声水平的混合光谱作训练(如 8%), 网络对噪声水平在 20% 以下的光谱识别率都能达 100%。因此用含噪声的混合光谱训练的 BP-ANN, 可以对那些受噪声、基线漂移等因素响的光谱也能取得好的识别效果。

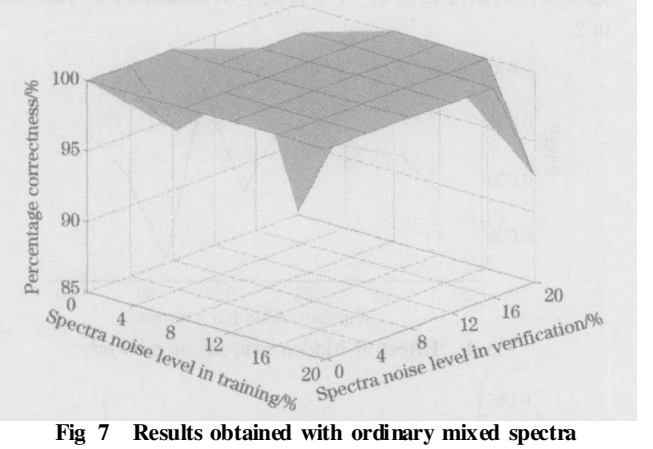

Fig 7 Results obtained with ordinary mixed spectra at different noise levels (with white noise)

2.4 用导数光谱作训练

本实验用有机污染物的导数光谱对 BP-ANN 作训练, 用于含杂质光谱、斜坡背景光谱的检索。由于对光谱求导过程中噪声信号也得到了放大, 为了让网络得到有效的信息, 更快地收敛, 在采用导数光谱的 BP-ANN 作检索预测时, 先

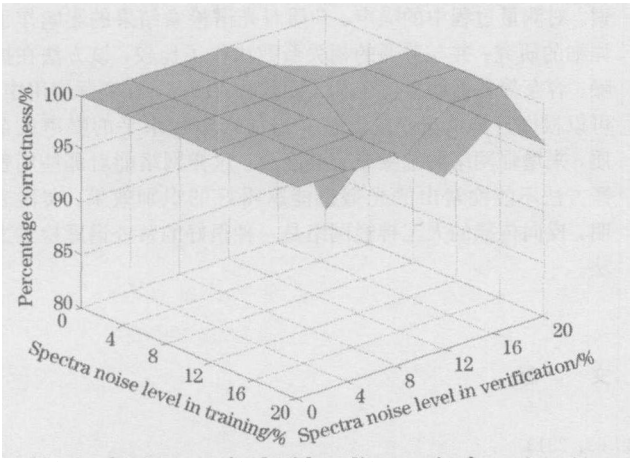

Fig 8　Results obtained with ordinary mixed spectra at different noise levels (with 0.08 bias noise)

对数据进行了平滑滤波、压缩、归一化等信号前处理，再用于含杂质和斜坡背景的峰的检索。表 2 给出了分别用常规光谱和导数光谱作训练的 BP-ANN 对含杂质、斜坡背景光谱的检索结果。从表 2 可以看出，在优化参数下，采用导数光谱训练的 BP-ANN，对含杂质光谱、斜坡背景光谱检索的正确率都有所提高，尤其是对含斜坡背景光谱的识别效果有明显改善；并且该网络的识别效果也明显优于相关的报道[5]。实验表明，优化参数条件下采用导数光谱训练的 BP-ANN 对检索光谱中斜坡背景的允许程度明显提高，可很好地用于含斜坡背景光谱的检索。

2.5　与相关系数法比较

相关系数法是传统的谱库检索方法，是通过比较待测光谱与谱库中所有光谱的相似系数实现对未知物的识别。在本研究中分类阈值设定为 0.997 5，用相关系数法分别对含噪声(白噪声、偏置)、杂质的光谱作识别，并将该方法与神经

Table 2　Identification results obtained with ordinary spectra and with derivative spectra

Spectra with impurity												
Impure levels/%	1		5		10		15		20		25	
	A	B	A	B	A	B	A	B	A	B	A	B
Correctness/%	100	100	100	100	100	100	100	100	96.3	100	92.6	96.3
Unknown/%	0	0	0	0	0	0	0	0	3.7	0	7.4	3.7
Wrong/%	0	0	0	0	0	0	0	0	0	0	0	0
Spectra with slop background												
Slop of slop background(a)	0.005		0.01		0.025		0.05					
	A	B	A	B	A	B	A	B				
Correctness/%	66.7	96.3	25.9	96.3	–	96.3	–	92.6				
Unknown/%	18.5	3.7	7.4	3.7	–	3.7	–	3.7				
Wrong/%	14.8	0	66.7	0	–	0	–	3.7				

Note: A: represents results obtained with ordinary spectra at different noise and impurity levels; B: represents results obtained with derivative spectra at different noise and impurity levels/ represents results obtained with network without convergence; a: represents slop of slop background

Table 3　Identification results obtained with ANN and with correlation coefficient method

Spectra with white noise										
Noise levels/%	0		4		8		12		16	
A	B	A	B	A	B	A	B	A	B	
Correctness/%	100	100	100	96.3	100	25.9	96.3	0	92.6	0
Unknown/%	0	0	0	3.7	0	74.1	3.7	100	7.4	100
Wrong/%	0	0	0	0	0	0	0	0	0	0
Spectra with 0.08 bias noise										
Noise levels/%	A	B	A	B	A	B	A	B	A	B
Correctness/%	100	100	100	96.3	100	25.9	96.3	0	92.6	0
Unknown/%	0	0	0	3.7	0	74.1	3.7	100	3.7	100
Wrong/%	0	0	0	0	0	0	0	0	3.7	0
Spectra with impurity										
Impurity levels/%	0		1		5		10		15	
	A	B	A	B	A	B	A	B	A	B
Correctness/%	100	100	100	100	100	85.2	100	22.2	100	0
Unknown/%	0	0	0	0	0	14.8	0	77.8	0	100
Wrong/%	0	0	0	0	0	0	0	0	0	0

Note: A: represents identification results obtained with BP-ANN;
B: represents identification results obtained with correlation coefficient method

网络方法检索的结果进行了比较, 结果见表 3。从表 3 可以看出, 相关系数法仅对纯光谱或含噪声、杂质水平较低的光谱有较好的识别效果, 对那些含杂质、噪声水平高的光谱识别较差; 神经网络库检索方法在抗噪、容杂等方面都明显优于普通的相关系数库检索方法。

3 结 论

成功地将人工神经网络应用于有机污染物的紫外谱库检索。对测量过程中的噪声、杂质对光谱检索结果的影响作了详细的研究; 并与普通的相关系数法作了比较, 该方法在抗噪、容杂等方面明显优于相关系数法。因此, 在实际应用中, 可以根据具体情况, 在训练光谱加入适当水平的噪声或杂质, 来增强网络的抗噪或容杂能力, 这样网络能对那些用普通方法不能检索出的光谱也能取得好的识别效果。实验表明, 反向传播的人工神经网络是一种很好的紫外谱库检索方法。

参 考 文 献

[1] Gemperline P J, Long J R, Gregorious V G. Anal. Chem., 1991, 63: 2313.

[2] Mittermayr C R, Drouen A C J, Otto M, et al. Anal. Chim. Acta, 1994, 294: 227.

[3] Bruchmann A, Gotze H J, et al. Chemom. Intell. Lab. Syst., 1993, 18: 59.

[4] Benjathapanun N, Boyle W J O, Grattan K T V. Measurement, 1998, 24: 1.

[5] ZHANG Zhuo-yong, LIU Si-dong, DING Bao-jun, et al(张卓勇, 刘思东, 丁保军, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 1998, 18(6): 680.

[6] LIU Si-dong, CUI Xiu-jun, ZHANG Zhuo-yong, et al(刘思东, 崔秀君, 张卓勇, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2003, 23(1): 119.

[7] HAN Li-qun(韩力群). Artificial Neural Network Theory, Design and Application(人工神经网络理论, 设计及应用). Beijing: Chemical Industry Press(北京: 化学工业出版社), 2001. 55.

[8] Xu L, Ball J W, Dixon S L, et al. Environ. Sci. Biochem., 1994, 13: 841.

[9] Andrea T A, Kalayeh H J. Med. Chem., 1991, 34: 2824.

[10] XU Lu, SHAO Xue-guang(许 禄, 邵学广). Methods of Chemometrics(化学计量学方法· 第2版). Beijing: Science Press(北京: 科学出版社), 2004. 564.

Library Search of UV Spectra of Organic Environmental Pollutants Based on Neural Network

WANG Feng-xia, ZHANG Zhuo-yong*, WANG Ya-min

Department of Chemistry, Resource Environment and GIS Key Lab of Beijing, Capital Normal University, Beijing 100037, China

Abstract The effects of optimization of network parameters, noise, and impurity on the network were investigated detailedly. To speed up the convergence of the network and enhance the resolution of the library search of UV spectra, the derivative spectra for BP-ANN library search was proposed. The method has a higher tolerance to noise and impurity levels than using ordinary UV spectra, especially to slop background levels. Finally, the resolutions of library search of UV spectra with ANN with optimized parameters were compared with conventional correlation coefficient method. Results showed that the ANN is superior to conventional correlation coefficient method and is an effective method for library search of UV spectra.

Keywords Artificial neural network; Environmental pollutant; Ultraviolet spectra; Library search

(Received Jan. 16, 2005; accepted Apr. 26, 2005)

* Corresponding author