文章编号:1671-6833(2004)03-0026-03

神经网络与结构编码法预测直馏汽油色谱保留指数

刘 伟[1], 刘 赞[2], 王玲玲[2]

(1. 郑州大学生物工程系,河南 郑州 450052; 2. 河南省环境监测中心站,河南 郑州 450004)

摘 要:对直馏汽油中的单体烃的分子结构进行了数字编码,并采用误差反向传播神经网络算法构造了直馏汽油中单体烃的气相色谱保留指数与其分子结构的非线性相关模型,神经网络结构为 3 层,隐含层节点为 7 个,有 15 个输入,对应单体烃的 15 位数字编码,1 个输出,对应气相色谱保留指数.预测结果表明,由误差反传算法所得的相关系数和标准偏差均优于多元线性回归方法.

关键词:神经网络; 结构编码; 气相色谱保留指数

中图分类号:O 652.9 **文献标识码**:A

0 引言

神经网络(NN)是一类摸拟生物大脑来处理非线性问题的网络系统.反向传播(BP)学习算法是一种前馈多层神经网络模型[1].BP 模型在化学中已获得一定应用[2].气相色谱保留指数是进行化合物定性分析的一项重要参数,可用神经网络进行预测研究[3].有关汽油中化合物气相色谱保留指数的神经网络应用和研究还很缺乏.本文采用BP 算法构造了直馏汽油中单体烃的气相色谱保留指数与其分子结构的非线性相关模型,并对其气相色谱保留指数进行了有效预测.

1 原理与方法

1.1 BP 学习算法

BP 模型是由输入层、隐含层(0~k 层)、输出层组成的通过神经元(Neuron)或节点高度连接而成的多层网络.图 1 给出了 3 层BP 模型的示意结构.每一节点j 通过连接权重w_{ji} 与上一层节点i 相连接,其输入值i_j 为前一层节点输出值o_i 的加权和,并设置一偏置项 θ,则

$$i_j=\sum_i(o_i w_{ji})+\theta \tag{1}$$

将输入值i_j 经过适当函数变换后得到节点j 的输出值o_j.通常利用Sigmoid 函数:

$$f(u)=1/[1+\exp(-u)] \tag{2}$$

BP 算法的误差函数可表示如下:

$$E=\frac{1}{2n}\sum_n\sum_p(t_p-o_p)^2 \tag{3}$$

式中:n,t_p 分别为训练集样本数和目标输出值.

采用Sigmoid 函数,则输出层节点p 的delta 误差项表示为

$$\delta_p=(t_p-o_p)o_p(1-o_p) \tag{4}$$

对于隐含层的节点h,其delta 误差项表示为

$$\delta_h=o_h(1-o_h)\sum_p w_{ph}\delta_p \tag{5}$$

网络在训练过程中新权重w_{ji}^n 按下列学习规则调整如下:

$$w_{ji}^n=w_{ji}+\eta\delta_j o_i+\alpha(\Delta w_{ji}) \tag{6}$$

式中:η, α 分别为学习速率和动量因子.

开始训练时,网络权重及偏置项设置为随机数.在学习过程中,网络不断调整权重直至输出误差E 达到最小或小于某一设定值.

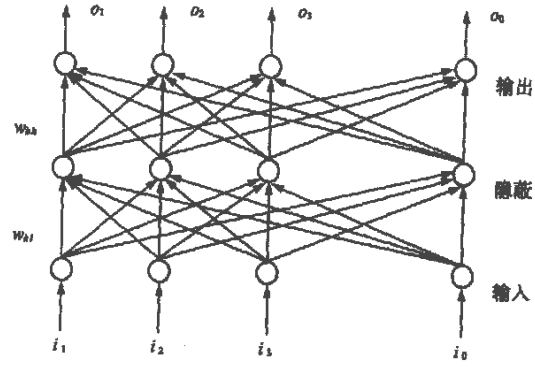

图 1 3 层BP 模型的结构

Fig. 1 Structure of three-layer BP model

收稿日期:2004-04-01;**修订日期**:2004-06-19

基金项目:河南省重点科技攻关项目(0223031800)

作者简介:刘 伟(1967-),男,河南省柘城县人,郑州大学副教授,博士,主要从事化学与生物信息学研究.

1.2　结构编码

分子结构的表述方法之一是用数字编码来实现的. 本研究根据 Driss Cherqaoui 等所采用的数字编码并加以扩展后来表述直馏汽油中的单体烃的分子结构[4]. 每一单体烃由 1 个 15 位的数字编码来表示. 图 2 给出了几种具有代表性的单体烃的分子结构的编码示例.

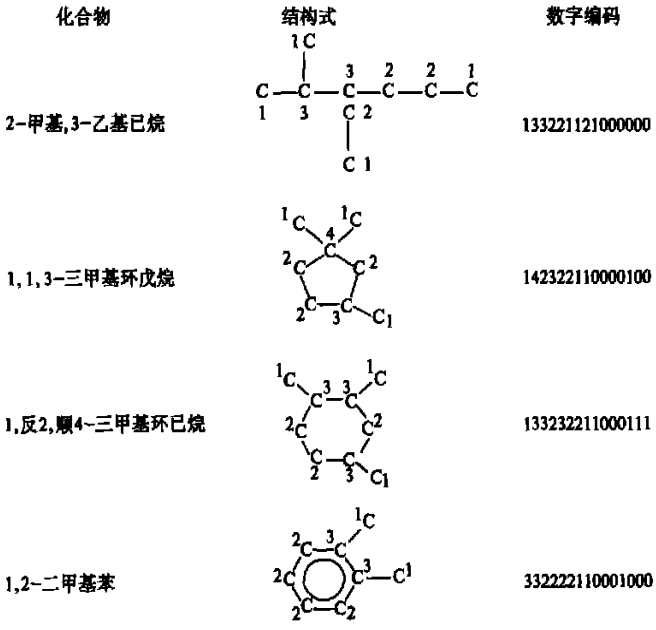

图 2　分子结构式和结构编码示意

Fig .2　Scheme of molecular structure and coding

编码中的前 11 位数字分别表示分子中每 1 个碳原子所联结的碳原子个数,碳原子个数不足时补为 0. 编码中第 12 位数字表示是否含有苯环,有为 1,无为 0. 编码中第 13 位数字表示是否含有环烷烃结构,有为 1,无为 0. 编码中第 14,15 位数字分别表示环烷烃是否含有顺式或反式结构,有为 1,无为 0.

2　结果与讨论

2.1 网络参数的选择

网络为三层结构,含有 1 个隐含层,输入层节点为 15,输出层节点为 1. 隐含层节点在 3～20 之间对网络影响不大,本文选取 7 个节点;学习速率 η 较大时,网络处于振荡状态,选取值为 0.1 ;动量因子 α 小则网络收敛速度慢,选取值为 0.925;训练次数为 100 时,网络已达到稳定状态,选取值为 1 800.

2.2　数据预处理

对输入的结构编码的预处理如下:

$$a_i = 0.1 + cx_i, \begin{cases} i < 12, c = 0.2 \\ i = 12, c = 1.9 \\ i = 13, c = 1.4 \\ i > 13, c = 0.9 \end{cases} \quad (7)$$

式中:x_i 为结构编码中第 i 位的数值. 在预处理中,当 $i > 12$ 时,取值在 0.4～1.5 之间对网络影响不大,但当 $i = 12, c < 1.7$ 时,网络对含苯单体烃的预测能力较差.

对输出的气相色谱保留指数的预处理为

$$b = (y - 200) / 1\,000 \quad (8)$$

式中:y 为单体烃的气相色谱保留指数.

2.3　计算和预测

本文按上述方式对直馏汽油中单体烃[3]进行结构编码和预处理,并对其单体烃的气相色谱保留指数进行了计算和预测. 除正十二烷烃外,采用了文献中全部已知结构的单体烃数据,共 150 个. 从表 1 的结果可看到,将全部数据作训练集时的网络模型,其相关系数和标准偏差均优于多元线性回归 (MLR) . 每次从数据中按类似 1, 11, …, 141 的排列方式选取 15 个数据组成预测集,其余为训练集,对全体数据进行了预测,预测结果良好,其相关系数和标准偏差分别为 0.993 4 和 16.51. 全部预测结果的统计分析见表 1.

表 1　神经网络和多元线性回归的相关系数及标准偏差

Tab .1　Correlation coefficients and standard deviation obtained by neural network and multi-linear regression

方法	r	s
NN	0.997 0	11.19
MLR	0.990 2	20.16

3 结论

通过对直馏汽油中的单体烃的分子结构进行了数字编码,并采用误差反向传播神经网络算法,对直馏汽油中单体烃的气相色谱保留指数进行了预测.预测结果表明,神经网络算法优于多元线性回归方法.

参考文献:

[1] RUMELHART D E, MCCLELLAND J L. Parallel Distributed Processing: Vol. 1[M]. Cambridge: MIT Press, 1986. 321~327.

[2] ZUPAN J, GASTEIGER J. Neural networks: A new method for solving chemical problem or just a passing phase ?[J] Anal Chim Acta, 1991(248): 1~30.

[3] YAN Aixia, JIAO Guimei, HU Zhide, et al. Use of artificial neural networks to predict the gas chromatographic retention index data of alkylbenzenes on carbowax—20M[J]. Computers & Chemistry, 2000, 24(2): 171~179.

[4] CHERQAOUI D, VILLEMIN D. Use of a neural network to determine the boiling point of alkanes[J]. J Chem Soc Faraday Trans, 1994, 90(1): 97~102.

[5] 武 杰,陆婉珍,程序升温毛细管气相色谱分析180℃以前直馏汽油中单体烃[J].分析化学, 1984, 12(7): 572~578.

Prediction of Gas Chromatography Retention Indices of Straight Run Gasoline by Using Neural Networks and Structure Coding

LIU Wei[1], LIU Zan[2], WANG Ling-ling[2]

(1. Department of Biological Engineering, Zhengzhou University, Zhengzhou 450052, China; 2. Environment Monitoring Center Station of Henan Province, Zhengzhou 450004, China)

Abstract: The molecular structures of the hydrocarbons of straight run gasoline are numerically coded. The nonlinear models of relationships between the chromatography retention indices of the hydrocarbons and their molecular structures are constructed by using error back-propagation neural network algorithm and their chromatography retention indices are predicted. The three-layer BPN which contains only one hidden layer, comprising fifteen input nodes, one output nodes and seven hidden nodes is employed. The molecular structures and the chromatography retention indices are used as input and output, respectively. The results show that the correlation coefficient and the standard derivation obtained by means of error back-propagation algorithm are better than those obtained by using multi-linear regression.

Key words: neural network; structure coding; chromatography retention indices