文章编号： 1672-8785(2013)12-0030-04

一种新的遗传神经网络组合方法在近红外光谱分析中的应用

孔 笋 祁彬彬 沈 阳 尤国平 张小康

(中海油田服务股份有限公司油田技术研究院，河北 燕郊 065201)

摘 要：近红外光谱分析技术中波长点的选择和建模方法的选取对建立预测分析模型至关重要。在传统相关系数法的基础上，提出了一种基于遗传算法的相关系数阈值优化方法。该方法以决定系数最大为优化目标，寻找最佳阈值。校正时用径向基神经网络来建立定标模型，选取中心时采用正交最小平方法。采用新方法预测了汽油中的碳酸二甲酯含量，把预测结果与偏最小二乘法的实验结果进行了对比。结果表明，新方法的预测精度更高，决定系数可达到 0.9993 。

关键词：近红外光谱；径向基神经网络；遗传算法； OLS 算法；偏最小二乘法

中图分类号： O657 **文献标识码**： A **DOI**: 10.3969/j.issn.1672-8785.2013.12.006

Application of a New Genetic Neural Network Combination Algorithm in Near Infrared Spectroscopy

KONG Sun, QI Bin-bin, SHEN Yang, YOU Guo-ping, ZHANG Xiao-kang

(Oilfield Technology Research Institute, China Oilfield Services Limited, Yanjiao 065201, China)

Abstract: The selection of wavelength points and modeling methods is very important for the establishment of predictive analysis model in near infrared spectroscopy. On the basis of the traditional correlation coefficient method, a correlation coefficient threshold optimization method based on the genetic algorithm is proposed. In the method, the determination coefficient is maximized so as to find the optimal threshold. A radial basis neural network is used to establish a calibration model and an orthogonal least square method is used to select the center for modeling. The new method is used to predict the dimethy carbonate content in gasoline and its result is compared with the experimental result of the partial least square method. The result shows that the new method has a better prediction accuracy. Its determination coefficient is up to 0.9993.

Key words: near infrared spectroscopy; radial basis function; genetic algorithm; orthogonal least squares; partial least squares

0 引言

在近红外光谱分析所使用的化学计量学方法中，波长选取方法和校正方法是核心，定性分析和定量分析都是在此基础上进行的。方法选择不当对后期的模型预测效果有很大影响，因此特征波长点的选择方法以及建模方法显得尤为重要。

相关系数法是多元校正分析中应用最广泛的一种波长选取方法。通常是根据相关的化学知识给定一个阈值 [1,2]，选择的阈值往往不是最优的。由于阈值的大小跟预测结果的好坏不是

收稿日期：2013–10–24

基金项目：“十二五”国家科技重大专项课题 (2011ZX05020-003)

作者简介：孔笋 (1982-)，女，山西绛县人，工程师，主要从事近红外光谱在地层流体分析中的应用研究。

E-mail: ks66_lulu@126.com

一种简单的线性关系，因而很难根据经验选择阈值。本文利用遗传算法 [3] 中的相关系数法建立一种客观的评价标准。根据这个标准选择波长点，可以减小主观随意性带来的误差。

传统的校正方法是利用偏最小二乘法进行建模的。这种方法虽然简单，但都是基于线性回归这一假设，所研究的光谱体系需要符合线性加和性。在实际工作中，硬件设备引起的噪声以及环境都会导致吸收光谱发生“偏离比尔定律”的现象。另外，非吸收光、散射、非平行光和化学因素等，也会使吸光度与浓度间的关系非线性化。所以非线性校正方法更能准确描述吸光度与浓度所确立的目标系统 [4] 。径向基神经网络作为一种非线性校正方法，学习速度快，无局部极小值问题，能以任意精度逼近任意连续函数 [5] 。考虑到径向基神经网络优良的非线性函数逼近能力，本文通过将径向基神经网络与基于遗传算法的相关系数法相结合建立一种新的方法，并将这一新方法用于预测汽油样品中的碳酸二甲酯 (Dimethyl carbonate, DMC) 浓度。

1 基于遗传算法的波长点选取

对于典型的相关系数法，相关系数的绝对值越大，包含的波长信息就越多。建立模型时，通常会给定一个阈值 β ，然后选取相关系数大于 β 的波长点参与模型建立。然而，如果 β 值太大，选取的特征点减少，就会影响拟合效果。若 β 值太小，一些贡献小的波段被选作为特征点，会引入不相关或非线性的变量。为了能够选取适当的 β 值，使得最后的预测结果更加准确，可以采用后面介绍的模型来作为遗传算法的目标函数。

拟合的准确程度用由预测值和实际值求得的决定系数 R^2 来判定。 R^2 越接近 1 ，说明预测效果越好。由于 R^2 与 β 之间存在的是一种非线性关系，可记为 $R^2 = f(\beta)$ 。若要找到最佳 β 值，使 R^2 能够尽可能接近 1 ，则需要求解下面的优化模型：

$$maxf(\beta)$$

$$s.t.\beta \in (0,1) \quad (1)$$

由于该函数存在多个局部极大值，采用具有全局搜索能力的遗传算法来求解该模型可以提高搜索精度。

遗传算法优化阈值的基本步骤为： (1) 编码的实现：将十进制的阈值 β 转换成二进制编码形式，变量 β 的区间为 (0,1) ，精度设定为小数点后 4 位； (2) 种群的初始化：根据经验选择一定的初始种群数量； (3) 适应度函数的选择：目标是使预测模型的决定系数尽可能地接近于 1 ； (4) 遗传进化：通过选择、单点交叉和变异操作，不断修正阈值变量的二进制串； (5) 解码：将满足最终条件的二进制串转换为真实值，从而求得最佳阈值。

2 径向基神经网络建模

2.1 RBFNN

径向基神经网络 (Radial Basis Function Neural Network, RBFNN) 是主要由输入层、隐含层和输出层构成的三层前向网络，其中隐含层采用径向基函数作为激励函数。

RBFNN 的输出和输入的函数关系为

$$y_i(X) = \sum_{j=1}^{M} \omega_{ij}\phi_j(X) + b_j, i = 1,2,\cdots,n \quad (2)$$

式中， $X = (x_1, x_2, \cdots, x_n)^T \in R^n$ 为输入向量， y_i 为第 i 个输出单元的输出值， M 为中心的个数， ω_{ij} 为第 j 个隐神经元到第 i 个输出单元的权值， b_j 为偏置值， $\phi(\cdot)$ 为 RBF 层非线性传递函数。常用的 RBF 层传输函数有薄板样条函数、多二次函数、逆多二次函数和高斯函数。

RBF 网络分为有导师学习和无导师学习两部分，隐含层和输入层之间每一个 RBF 单元的中心 c 和半径 σ 采用的是无导师聚类方法训练，输出层和隐含层之间的权值 ω 采用有导师方法训练。设计 RBF 网络的关键因素是径向基函数中心的选取，如果中心选取不当，会导致构造出来的 RBFNN 的性能不好。本文采用了正交最小平方 (Orthogonal Least Squares, OLS) 算法 [6,7] ，其正交性可以避免由于中心选取太近而产生的近似线性相关。该方法在计算时间和最终效果上均优于随机选取法和 k 均值聚类法。

2.2 模型训练

OLS 算法是从输入样本中随机选取若干样本作为中心矢量，通过正交化回归矩阵 P 的各分量，选择每一次循环过程中误差压缩比最大的回归算子，直到相对二次误差小于选定的容差，最终确定回归算子数，求得权值的最小二乘解。

中心选取步骤如下：

(1) 预选隐层节点数 M， M 一般等于样本个数；

(2) 从输入样本中预选一组中心矢量 c_i；

(3) 结合前面所选取的中心矢量，计算出 $x_i(i=1,2,\cdots,N)$ 的回归矩阵 P；

(4) 采用 OLS 算法选择回归算子 W。每一个回归算子都对应一个误差压缩比，计算公式如下：

$$err_i = \frac{g_i^2 w_i^T w_i}{y^T y} \tag{3}$$

式中， w_i 为回归算子 W 中的第 i 个矢量， y 为期望输出。

k 表示 W 的列数。当 k=1 时，分别令 P 的各列 p_k 为 W 的第一列，计算误差压缩比，然后选择误差压缩比最大的回归算子为 W 的第一列；k≥ 2 时，依次选择回归矩阵 P 中余下的 N-k 列作为 W 的第 k 列，然后用斯密特正交化方法将其与前 k 列向量正交。

$$\left.\begin{aligned} w_i &= pi \\ \alpha_{ik} &= \frac{w_i^T p_k}{w_i^T w_i}, 1 \le i \le k \\ w_k &= p_k - \sum_{i=1}^{k-1} \alpha_{ik} w_i \end{aligned}\right\} k = 2, \cdots, M \tag{4}$$

经过若干次的循环，当 $1 - \sum |err|_i$ 小于设定的容差时，回归算子选择完毕，求得权重的最小二乘解。

3 实例分析

为验证新算法的有效性，本文以预测未知汽油样品中 DMC 的浓度为例 [8]，比较了遗传径向基神经网络与偏最小二乘法两者的预测效果。

3.1 波段选择

在实验中，收集了 61 个汽油样品，波长范围在 1000.709 nm~1999.191 nm 之间，共有 537 个数据点。 DMC 的特征峰值在 1700 nm~1800 nm 之间。

为了提高神经网络模型的分辨能力，需要找出能够代表 DMC 特征的波段作为输入，并且要减少无关信息对网络的干扰。本文采用相关系数法找出特征波段。首先对校正集光谱阵中每个波长对应的吸光度向量与浓度阵中待测组分浓度向量进行相关性计算，得到相关系数图，如图 1 所示。由图可知， 1000 nm~1500 nm 和 1700 nm~1800 nm 波段的线性相关性比较好。但如果直接选取这两个波段，仍然会引入一些不相关的信息量。

下面以决定系数接近 1 为目标，利用遗传算法来搜索最佳阈值。遗传算法的参数设定为：初始种群大小为 100，最大遗传代数为 200，交叉率为 0.4，变异率为 0.08。

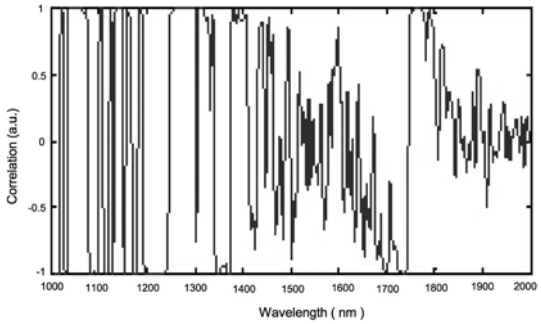

图 1 相关系数图

3.2 径向基网络实现

从 61 个汽油样本中随机选取 12 个作为验证集，其余 49 个样本作为校正样本，用于建立神经网络模型。

采用 MATLAB 语言编写程序， RBF 网络参数如下：输入层的节点数为波长点的个数，输出层的节点数为 1 个，隐含层的最大节点数为 100 个，神经元的扩展速度为 3。

为了找到最优模型，将相关系数的阈值作为一个调整参数，观察其在 [0.1, 0.9] 区间变化时所对应的决定系数 R^2 的变化规律，如图 2 所示。

从图中可以看出，当相关系数在 0.2 到 0.8 之间时，决定系数 R^2 最接近 1，预测结果较好。 R^2 曲线的最高点对应相关系数在 0.2~0.3 之间的位置。通过用遗传算法进行全局搜索得出，当相关系数为 0.248 时， R^2 达到最大值。

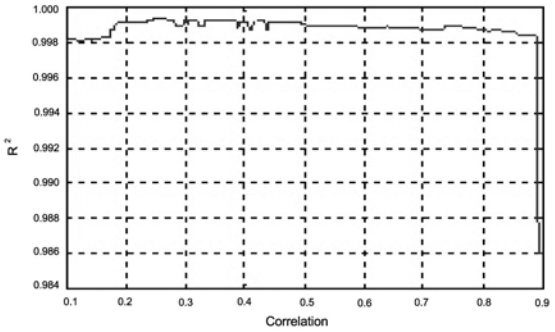

图 2　相关系数与决定系数的关系图

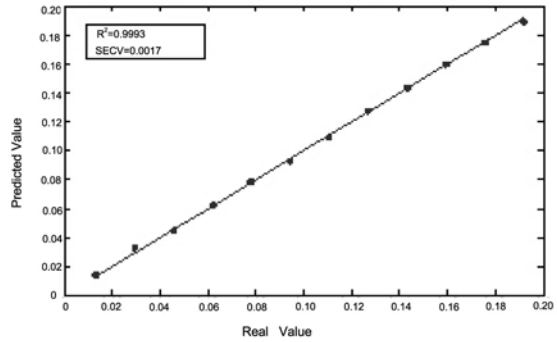

图 3　预测值与实际值

以实测的 DMC 浓度为横坐标，以预测值为纵坐标，选取相关系数大于 0.248 的波段进行仿真建模，预测结果如图 3 所示。从图中可以看出，预测值和实际值几乎相等。新算法具有很好的预测效果。

3.3　结果对比

在同一实验条件下，本文也采用了偏最小二乘方法预测汽油中的 DMC 浓度，并对比了两种方法的预测结果。

在采用偏最小二乘方法建模的过程中，本文还采用遗传算法找到了一个最佳相关系数值 0.355，相应地计算出了最佳决定系数 0.9975。为了能够尽可能地在相同条件下对比两种方法，本文对两种方法的最佳建模效果进行了比较，比较结果见表 1。与偏最小二乘法相比，用 RBF 神经网络计算得到的预测值更加接近真实值，最佳决定系数为 0.9993。

4　结论

本文对传统的相关系数法进行了改进，并用遗传算法优化了相关系数的阈值。通过用 OLS 径向基神经网络建立预测模型，该算法不仅能对波长变量进行优化选择，还能解决光谱中存在的非线性因素问题。为了验证算法的有效性，将该算法用于预测汽油中的 DMC 浓度。在仿真实验中，同时采用偏最小二乘法和 OLS 径向基神经网络法建立定标模型。通过对比发现，用遗传径向基神经网络建立的预测模型具有更快、更准确的预测效果，预测模型的决定系数 R^2 可以达到最佳值。

表 1 两种方法预测结果对比表

样本号	真实值	PLS 预测值	RBF 预测值
(1)	0.0130	0.0154	0.0143
(2)	0.0293	0.0308	0.0333
(3)	0.0455	0.0425	0.0449
(4)	0.0618	0.0559	0.0625
(5)	0.0780	0.0775	0.0791
(6)	0.0943	0.0927	0.0932
(7)	0.1105	0.1098	0.1094
(8)	0.1268	0.1287	0.1280
(9)	0.1430	0.1483	0.1438
(10)	0.1593	0.1603	0.1603
(11)	0.1755	0.1794	0.1756
(12)	0.1918	0.1903	0.1893

参考文献

[1] 王艳斌，袁洪福，陆婉珍．近红外分析方法测定润滑油基础油粘度指数 [J]. **润滑油**,2001,**6**(16):53–56.

[2] 王芳,陈达,邵学广.小波变换和偏最小二乘法在烟草常规成分预测中的应用 [J]. **烟草科技**,2004,**3**(5):23–26.

[3] Conn A R, Gould N I M, Toint P L. A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds [J].*Journal on Numerical Analysis*,1991,**28**(2): 545–572.

[4] 褚小立．**化学计量学方法与分子光谱分析技术**[M]. 北京：化学工业出版社，2011.

[5] 魏海坤．**神经网络结构设计的理论与方法**[M]. 北京：国防工业出版社，2005.

[6] Chen S, Cowan C F N, Grant P M. Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks[J].*IEEE Transactions on Neural Networks*,1991,**2**(2):302–309.

[7] 刘文菊，郭景．RBF 神经网络中心选取 OLS 算法的研究 [J]. **天津工业大学学报**,2002,**21**(2):71–73.

[8] 陆婉珍，龙义成，黎洁，等．碳酸二甲酯作为汽油添加剂的评价 [J].**石油学报 (石油加工)**,1997,**13**(3):40–45.