区间极限学习机在气体 FTIR 光谱浓度反演中的应用研究

陈媛媛[1] 张记龙[1,2,3] 王志斌[1] 赵冬娥[1,2,3] 陈友华[1]

[1]中北大学光电信息与仪器工程技术研究中心，山西 太原 030051
[2]中北大学仪器科学与动态测试教育部重点实验室，山西 太原 030051
[3]中北大学电子测试技术国家重点实验室，山西 太原 030051

摘要 为准确反演气体浓度，节约建模时间，提出了基于区间极限学习机(ELM)定量分析模型的傅里叶变换红外(FTIR)光谱分析技术。该方法基于区间划分思想，将整个光谱范围划分为若干个子区间，利用 ELM 分别建立各个子区间的定量分析模型，并根据各个子区间模型的决定系数大小评价其泛化性能，进而筛选出最具代表性的子区间组合。基于上述方法，对 NO 与 NO_2 气体的红外光谱进行波长筛选，并利用筛选后的特征波长点光谱建立定量分析模型。实验结果表明，NO 气体测试集的决定系数 R^2 为 0.9999，NO_2 气体测试集的决定系数 R^2 为 0.9997。与区间偏最小二乘法相比，利用区间 ELM 方法建模速度更快，模型泛化性能更优。

关键词 傅里叶光学；红外光谱；气体浓度反演；光谱筛选；区间极限学习机

中图分类号 X831 **文献标识码** A **doi**：10.3788/CJL201138.s115006

Research on Concentration Retrieval with Gas FTIR Spectra by Interval Extreme Learning Machine Method

Chen Yuanyuan[1] Zhang Jilong[1,2,3] Wang Zhibin[1] Zhao Donge[1,2,3] Chen Youhua[1]

[1] *Enginering Technology Research Center of Shanxi Province for Opto-Electronic Information and Instrument, North University of China, Taiyuan, Shanxi* 030051, *China*
[2] *Key Laboratory of Instrumentation Science & Dynamic Measurement, Ministry of Education, North University of China, Taiyuan, Shanxi* 030051, *China*
[3] *Science and Technology on Electronic Test & Measurement Laboratory, North University of China, Taiyuan, Shanxi* 030051, *China*

Abstract To retrieve the gas concentration accurately and rapidly, a new quantitative analysis technique based on Fourier transform infrared spectroscopy of interval extreme learning machine (ELM) model is proposed. Based on the idea of interval division, this approach firstly divides the whole spectrum into several subintervals, secondly establishes quantitative analysis model corresponding to each subinterval with ELM method, and finally selects the best subinterval combinations according to the determination coefficient of each model. Based on the above approach, wavelengths are selected in the spectrum of NO and NO_2, and then establishs the quantitative analysis model using the selected spectrum combinations, respectively. The experimental results showed that, the testing set determination coefficient of NO and NO_2 are 0.9999 and 0.9997, respectively. The outcome indicates that, compared with Interval partial least squares method, the proposed Interval ELM method can establish quantitative analysis model more rapidly and accurately.

Key words Fourier optics; infrared spectrum; gas concentration retrieval; spectrum selection; interval extreme learning machine

OCIS codes 070.4790; 300.6300; 280.1120

收稿日期：2011-08-10；**收到修改稿日期**：2011-09-27

基金项目：科技部国际科技合作项目、国家自然科学基金(61040062)、山西省国际科技合作计划(2010081038)和山西省人才引进与开发专项资金资助课题。

作者简介：陈媛媛(1980—)，女，博士研究生，讲师，主要从事红外光谱分析方面的研究。E-mail：chenyy-000@163.com

导师简介：张记龙(1964—)，男，博士，教授，主要从事光电仪器与系统方面的研究。E-mail：zhangjl@nuc.edu.cn(通信联系人)。

1　引　　言

伴随着全球环境的变化，大气污染给人类的生活及动植物的生存带来了严重的影响。作为大气保护基础工作的大气环境监测也越来越显示出其重要性。随着计算机技术的不断发展，大气污染遥测技术也日趋成熟，其中，傅里叶变换红外(FTIR)光谱技术在大气污染物浓度遥测中应用最为广泛[1~3]。朱军等[4,5]利用 FTIR 光谱测量系统，应用非线性最小二乘拟合算法实现了 CO 和 CO_2 气体浓度的定量分析；徐亮等[6]系统地研究了利用 HITRAN 数据库合成校准光谱的方法，魏合理等[7]提出了提高大气吸收光谱测量分辨率的新方法，实现了可以模拟真实环境下的气体光谱。

然而，由于光谱与气体浓度间呈现出非线性关系，且一些光谱波长点间存在多重共线问题，采用传统的如多元线性回归分析、最小二乘法等方法难以准确地描述其数学关系。因此，在前人基础上，本文提出了一种新的定量分析模型——区间极限学习机(ELM)。该方法首先基于区间划分的思想，将整个光谱范围划分为若干个子区间并筛选出最具代表性的子区间组合，然后利用 ELM 建立气体浓度的定量分析模型。

2　ELM 原理

针对传统智能算法[如反向传播(BP)神经网络、支持向量机等]普遍存在的学习速度慢、调节参数多、泛化能力差等缺点和不足，Huang 等[8]于 2004 年提出了 ELM。与 BP 神经网络不同的是，ELM 的基本结构是单隐含层前馈神经网络(SLFN)，如图 1 所示。

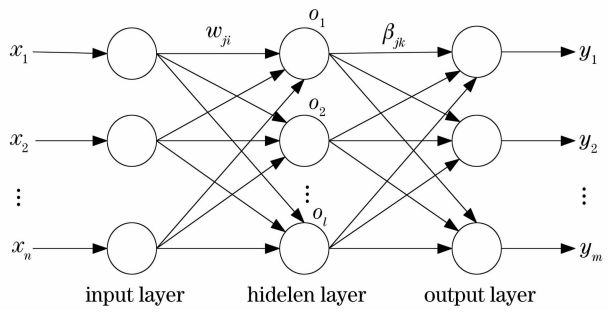

图 1　典型的 SLFN 结构

Fig. 1　Structure of classical SLFN

设具有 Q 个样本的训练集输入矩阵 $\boldsymbol{X}$ 和输出矩阵 $\boldsymbol{Y}$ 分别为

$$\boldsymbol{X}=[x_1,\ x_2,\ \cdots,\ x_Q],\quad \boldsymbol{Y}=[y_1,\ y_2,\ \cdots,\ y_Q].$$

式中 $\boldsymbol{x}_i=[x_{1i},x_{2i},\cdots,x_{mi}]^{\mathrm{T}}$ $(i=1,2,\cdots,Q)$ 和 $\boldsymbol{y}_i=[y_{1i},y_{2i},\cdots,y_{mi}]^{\mathrm{T}}$ $(i=1,2,\cdots,Q)$ 分别表示第 i 个样本的输入向量和输出向量。

设隐含层神经元的激活函数为 $g(x)$，则由图 1 可得，网络的输出 $\boldsymbol{T}$ 为

$$\boldsymbol{T}=[\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_Q]_{m\times Q}=\begin{bmatrix}\sum_{i=1}^{l}\beta_{i1}g(w_ix_j+b_i)\\ \sum_{i=1}^{l}\beta_{i2}g(w_ix_j+b_i)\\ \vdots\\ \sum_{i=1}^{l}\beta_{im}g(w_ix_j+b_i)\end{bmatrix}_{m\times Q},\quad (j=1,2,\cdots,Q) \tag{1}$$

式中 $\boldsymbol{t}_i=[t_{1i},t_{2i},\cdots,t_{mi}]^{\mathrm{T}}$ $(i=1,2,\cdots,Q)$，$\boldsymbol{w}_i=[w_{i1},w_{i2},\cdots,w_{in}]$。(1)式可表示为

$$\boldsymbol{H}\boldsymbol{\beta}=\boldsymbol{T}', \tag{2}$$

式中 $\boldsymbol{T}'$ 为矩阵 $\boldsymbol{T}$ 的转置；$\boldsymbol{H}$ 称为神经网络的隐含层输出矩阵，具体形式为

$$\boldsymbol{H}(w_1,w_2,\cdots,w_l,b_1,b_2,\cdots,b_l,x_1,x_2,\cdots,x_Q)=\begin{bmatrix}g(w_1x_1+b_1),\ g(w_2x_1+b_2),\cdots,g(w_lx_1+b_l),\\ g(w_1x_2+b_1),\ g(w_2x_2+b_2),\cdots,g(w_lx_2+b_l)\\ \vdots\\ g(w_1x_Q+b_1),g(w_2x_Q+b_2),\cdots,g(w_lx_Q+b_l)\end{bmatrix}_{Q\times l}. \tag{3}$$

在前人基础上，Huang 等[8]证明，对于一个任意区间无限可微的激活函数 $g:R\to R$，给定任意小误差 $\varepsilon>0$，则总存在一个含有 K $(K\leqslant Q)$ 个隐含层神经元的 SLFN，在任意赋值 $w_i\in R^n$ 和 $b_i\in R$ 的情况下，有 $\|\boldsymbol{H}_{N\times M}\boldsymbol{\beta}_{M\times m}-\boldsymbol{T}'\|<\varepsilon$。

因此，当激活函数 $g(x)$ 无限可微时，SLFN 的参数并不需要全部进行调整，w 和 b 在训练前可以随机选择，且在训练过程中保持不变。而隐含层与输出层间的连接权值 β 可以通过求解

$$\min_{\beta} \| \boldsymbol{H}\beta - \boldsymbol{T}' \| \tag{4}$$

获得，其解为

$$\hat{\beta} = \boldsymbol{H}^{+}\boldsymbol{T}', \tag{5}$$

式中 $\boldsymbol{H}^{+}$ 为隐含层输出矩阵 $\boldsymbol{H}$ 的 Moore-Penrose 广义逆。

简言之，ELM 在训练之前只需确定隐含层神经元个数及隐含层神经元的激活函数（无限可微），即可计算出 β。具体地，ELM 的学习算法主要有以下几个步骤：

1）确定隐含层神经元个数，随机设定输入层与隐含层间的连接权值 w 和隐含层神经元的偏置 b；

2）选择一个无限可微的函数作为隐含层神经元的激活函数，进而计算隐含层输出矩阵 $\boldsymbol{H}$；

3）计算输出层权值 $\hat{\beta}$：$\hat{\beta}=\boldsymbol{H}^{+}\boldsymbol{T}'$。

3 区间 ELM 波长筛选

相关研究成果表明，波长点之间存在多重共线问题，若利用全部波长点数据建立校正模型，则会引起建模时间长、模型精度低等问题。因此，在建立校正模型之前，有必要对光谱数据进行压缩和筛选。

区间 ELM 的思想起源于丹麦 Norgaard[9] 在 2000 年提出的一种波长筛选方法——区间偏最小二乘法（PLS），其基本思想是：将全部光谱区域划分为若干个子区间，选择具有代表性的子区间建立校正模型。

本文在传统方法的基础上，提出了改进的区间 ELM 波长筛选方法，其主要步骤为：

1）初始化划分的子区间个数范围为 $[N\_\min, N\_\max]$，循环变量初值 $k=N\_\min$；

2）将全部光谱区域划分为 k 个子区间，分别建立 k 个子区间的 ELM 定量分析模型，并计算各个校正模型的测试集均方根误差

$$E_i = \sqrt{\sum_{j=1}^{n}(y_j^i - y_j)^2 \Big/ n}, \quad (i = 1,2,\cdots,k) \tag{6}$$

式中 n 为测试集样本数；y_j^i 为第 j 个测试集样本在第 i 个子区间 ELM 校正模型中的预测值；y_j 为第 j 个测试集样本的真实值；

3）将 E_i 升序排列，并重新记均方根差值的关系为 $E_{\mathrm{RMS\_1}} < E_{\mathrm{RMS\_2}} < \cdots < E_{\mathrm{RMS\_}k}$，以 $E_{\mathrm{RMS\_1}}$ 对应的子区间光谱数据作为训练集的输入，并设循环变量初值 $m=2$；

4）将 $E_{\mathrm{RMS\_}m}$ 对应的子区间光谱数据添加到训练集的输入中，重新建立 ELM 校正模型，并计算测试集的均方根误差 E_{new}，若 $E_{\mathrm{new}} < E_{\mathrm{RMS\_1}}$，保留 $E_{\mathrm{RMS\_}m}$ 对应的子区间光谱数据；否则，删除 $E_{\mathrm{RMS\_}m}$ 对应的子区间光谱数据；

5）$m=m+1$，若 $m<k$，返回步骤 4）；否则，记录此时的子区间组合及对应的均方根误差 E_{best}^{k}；

6）$k=k+1$，若 $k<N\_\max$，返回步骤 2）；否则，算法停止，输出最佳的子区间组合及对应的均方根误差 E_{best}。

4 实验与结果分析

4.1 低分辨率校正光谱计算

在开放光程条件下，采用传统的方法无法采集到用于校准的校正光谱，因此利用 HITRAN Database，通过逐线积分的方法[10]，计算给定环境温度、气压和光程条件下的高分辨率吸收光谱，并且跟实际采用的光谱仪的仪器线型函数做卷积，从而得到相同条件下低分辨率光谱。

利用上述方法，分别计算浓度范围为 $(100\sim10000)\times10^{-6}$，间隔为 50×10^{-6} 的各 199 个 NO 与 NO_2 气体样本的高分辨率光谱，并且与三角切趾函数相卷积得到模拟的低分辨率光谱。其中选择的 NO 气体的待分析光谱区间是 1750～2000 cm^{-1}，波数间隔为 1.929 cm^{-1}，该波段含有 130 个数据点；NO_2 气体的待分析光谱区间是 1550～1660 cm^{-1}，波数间隔为 1.929 cm^{-1}，该波段含有 58 个数据点。

4.2 测试集气体光谱采集

实验使用瑞利 WQF-520 光谱仪，将光谱分辨率设置为 4 cm^{-1}。然后将分别如表 1 所列的 10 个不同浓度的 NO 气体和表 2 所列的 10 个不同浓度的 NO_2 气体充入密闭气室中，测量其透射率。

表 1 NO 气体测试集浓度

Table 1 Volume fraction of NO testing set

Sample	Volume fraction /10^{-6}	Sample	Volume fraction /10^{-6}
1	1500	6	1540
2	2780	7	1160
3	9140	8	5600
4	6780	9	520
5	9940	10	4180

表 2　NO_2 气体测试集浓度

Table 2　Volume fraction of NO_2 testing set

Sample	Volume fraction /10^{-6}	Sample	Volume fraction /10^{-6}
1	5260	6	4520
2	9780	7	3140
3	4240	8	8460
4	6000	9	3060
5	900	10	6600

4.3　实验结果与分析

实验采用近似取整的方法对区间进行划分，例如将整个 NO 光谱区域划分为 20 个子区间，则前 19 个子区间包含的波长点数是相等的，均为[130/20]=6，剩余的波长点归属于最后一个子区间，即第 20 个子区间包含的波长点数为 130−6×19=16。其他情况，以此类推。

4.3.1　NO 气体区间 ELM 定量分析模型

利用区间 ELM 方法对 NO 气体的光谱进行波长筛选，筛选结果表明，当将整个光谱区间划分为 18 个子区间时，均方根误差最小，模型性能最佳。在此情况下，光谱波长筛选的结果如图 2 所示。从图中可以看出，筛选出的最佳建模区间是 12 和 13 号子区间，对应的光谱波数范围为 1898.5～1923.6 cm^{-1}。

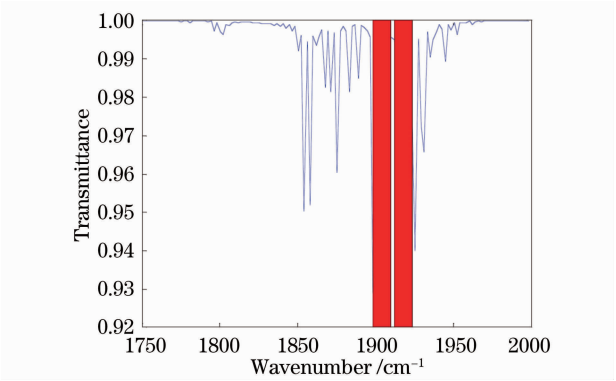

图 2　NO 气体光谱波长筛选结果

Fig. 2　Wavelength selecting results of NO spectrum

利用筛选出的波数范围内的光谱建立模型，预测结果如表 3 所示。从表中可以清晰地看到，与区间 PLS 方法相比，利用区间 ELM 筛选出的光谱建立模型，可以使得建模速度更快、精度更高。

4.3.2　NO_2 气体区间 ELM 定量分析模型

利用区间 ELM 方法对 NO_2 气体的光谱进行波长筛选，筛选结果表明，当将整个光谱区间划分为 15 个子区间时，均方根误差最小，模型性能最佳。在此情况下，光谱波长筛选的结果如图 3 所示。从图中可以看出，筛选出的最佳建模区间是 7，11，12 和 14 号子区间，对应的光谱波数范围为 1584.7～1588.6 cm^{-1}，1607.9～1617.5 cm^{-1}，1625.2～1629.1 cm^{-1}。

利用筛选出的波数范围内的光谱建立模型，预测结果如表 4 所示。从表中可以清晰地看到，利用区间 ELM 筛选出的光谱建立模型，可以使得建模速度更快、精度更高。

表 3　NO 气体定量分析模型预测结果

Table 3　Prediction results of NO quantitative analysis model

Sample	Iinterval ELM /10^{-6}	Interval PLS /10^{-6}
1	1492	1582
2	2789	2705
3	9167	9047
4	6762	6730
5	9904	9886
6	1525	1479
7	1184	1128
8	5580	5704
9	537	651
10	4203	4124

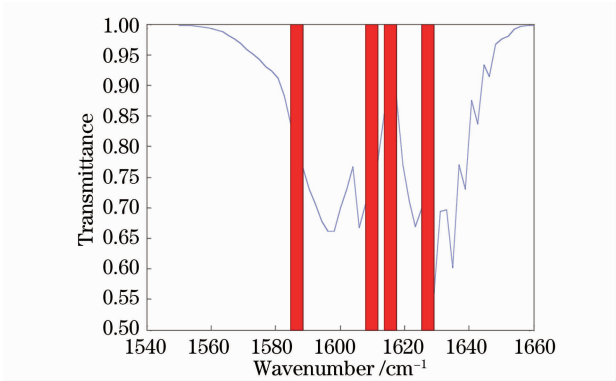

图 3　NO_2 气体光谱波长筛选结果

Fig. 3　Wavelength selecting results of NO_2 spectrum

表 4　NO_2 气体定量分析模型预测结果

Table 4　Prediction results of NO_2 quantitative analysis model

Sample	Iinterval ELM /10^{-6}	Interval PLS /10^{-6}
1	5272	5130
2	9836	9716
3	4215	4351
4	6040	5928
5	868	993
6	4464	4562
7	3120	3047
8	8406	8533
9	3084	3219
10	6697	6488

5 结 论

随着 FTIR 光谱技术的日趋成熟，大气污染物的遥测技术也取得了长足的进步和发展。对于开放光程条件而言，利用 HITRAIN 数据库，模拟合成真实大气环境下的校准光谱建立气体浓度反演模型，已经成为一个公认的理想方法与思路。利用本文提出的区间极限学习机方法，可以方便地在建立模型前，对整个光谱范围进行筛选，从而减少多重共线问题的干扰。同时，利用极限学习机方法，网络的权值和阈值参数可以随机产生，且训练过程无须迭代，一次解析计算便可实现模型的建立。实验结果表明，与区间 PLS 方法相比，区间 ELM 方法可以使得建模速度更快，模型泛化性能更好。

参 考 文 献

1 Zheng Longjiang, Li Peng, Qing Ruifeng *et al*.. Research situation and developing tendency for optical measurement technology of gas density[J]. *Laser & Optoelectronics Progress*, 2008, **45**(8): 24～32
郑龙江，李 鹏，秦瑞峰 等. 气体浓度检测光学技术的研究现状与发展趋势[J]. 激光与光电子学进展，2008，**45**(8)：24～32

2 Lan Tiange, Xiong Wei, Fang Yonghua *et al*.. Research on preprocessing algorithm for infrared spectral signals of biological aerosols[J]. *Acta Optica Sinica*, 2010, **30**(9): 2742～2747
兰天鸽，熊 伟，方勇华 等. 生物气溶胶红外光谱信号预处理算法研究[J]. 光学学报，2010，**30**(9)：2742～2747

3 Liu Zhiming, Liu Wenqing, Gao Minguang *et al*.. Study of the retrieval algorithm of emission gas spatio-temporal distribution of pollution source using the infrared solar occultation flux (SOF) method[J]. *Acta Physica Sinica*, 2010, **59**(8): 5397～5405
刘志明，刘文清，高闽光 等. 基于红外掩日通量法(SOF)污染源排放气体浓度时空分布反演算法研究[J]. 物理学报，2010，**59**(8)：5397～5405

4 Zhu Jun, Liu Wenqing, Liu Jianguo *et al*.. Application of FTIR spectra fitting method in retrieving gas concentrations [J]. *Spectrosc. & Spectral Analy.*, 2005, **25**(10): 1573～1576
朱 军，刘文清，刘建国 等. FTIR 光谱拟合方法在反演气体浓度中的应用[J]. 光谱学与光谱分析，2005，**25**(10)：1573～1576

5 Zhu Jun, Liu Wenqing, Liu Jianguo *et al*.. Quantitative gas analysis using Fourier transform infrared spectroscopy method [J]. *Chin. J. Sci. Instrum.*, 2007, **28**(1): 80～84
朱 军，刘文清，刘建国 等. 傅里叶变换红外光谱学方法用于气体定量分析[J]. 仪器仪表学报，2007，**28**(1)：80～84

6 Xu Liang, Liu Jianguo, Gao Minguang *et al*.. Monitoring of atmospheric NH_3 in urban area using open path FTIR system[J]. *J. Atmos. & Environm. Opt.*, 2007, **2**(1): 60～63
徐 亮，刘建国，高闽光 等. 开放光程傅里叶变换红外光谱系统观测城市空气中的 NH_3[J]. 大气与环境光学学报，2007，**2**(1)：60～63

7 Wei Heli, Wu Chengjiu, Ma Zhijun *et al*.. A new method for improving the measurement spectral resolution of atmospheric absorption apectra[J]. *Acta Optica Sinica*, 2002, **22**(2): 165～169
魏合理，邬承就，马志军 等. 提高大气吸收光谱测量分辨率的新方法[J]. 光学学报，2002，**22**(2)：165～169

8 G.-B. Huang, Q.-Y. Zhu and C.-K. Siew. Extreme learning machine: a new learning scheme of feedforward neural networks [C]. International Joint Conference on Neural Networks, (Budapest, Hungary), 2004

9 L. Norgaard, A. Saudland, J. Wagner *et al*.. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy [J]. *Appl. Spectrosc.*, 2000, **54**(3): 413～419

10 Zhang Hua, Shi Guangyu. A fast and efficient line-by-line calculation method for atmospheric absorption [J]. *Chin. J. Atmos. Sci.*, 2000, **24**(1): 111～121
张 华，石广玉. 一种快速高效的逐线积分大气吸收计算方法[J]. 大气科学，2000，**24**(1)：111～121

栏目编辑：李文喆