第28卷，第8期
2008年8月

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

Vol. 28, No. 8, pp1846-1850
August, 2008

DWT-iPLS在烃源岩漫反射光谱数据处理中的应用

宋 宁[1,3]，徐晓轩[1,3]，武中臣[2]，张存洲[1,3]，王 斌[1]

1. 南开大学泰达应用物理学院光子学中心，天津 300457
2. 山东大学威海分校空间科学与应用物理系，山东 威海 264209
3. 天津市南开大学弱光非线性光子学材料先进技术及制备教育部重点实验室，天津 300457

摘 要 红外漫反射光谱技术被广泛地应用于粉末样品的定性、定量测量中。但是由于粉末样品自身的特点，颗粒度、密度、表面粗糙程度等几何参数的影响，使得漫反射光谱数据的信噪比很低、背景干扰很大。因此需要一种有效的方法对漫反射光谱数据进行预处理来提高信噪比，消除背景干扰。文章采用了离散小波变换对红外漫反射光谱进行了预处理，有效地消除了光谱中的高频噪声和低频背景的干扰，并结合iPLS(间隔偏最小二乘回归)方法进行线性回归分析，建立了用于复杂样品体系组分分析的建模方法(DWT-iPLS)。并以烃源岩的红外漫反射光谱为例，将DWT-iPLS应用于数据的预处理及建立数学模型，其结果与未用DWT预处理的方法相比较，准确度有明显的提高，证明了此方法是一种快速有效的定量分析的建模方法。

关键词 漫反射；相关系数；离散小波变换；间隔偏最小二乘回归；均方根误差

中图分类号：O657.3 **文献标识码**：A **文章编号**：1000-0593(2008)08-1846-05

引 言

红外光谱技术由于其快速、无破坏性、无污染等优点，被广泛地应用于定性、定量的测量中，并且可以实现在线的实时检测和监控[1]。在红外光谱技术中，漫反射光谱技术越来越多地被人们应用于定性、定量分析[2-4]。但是，对于复杂体系的红外漫反射光谱，由于样品的颗粒度、密度、表面粗糙程度等几何参数的影响，使得漫反射光谱数据的信噪比很低、背景干扰很大[5-7]，因此需要一种有效的方法对漫反射数据进行预处理来提高信噪比，消除背景干扰。离散小波变换(DWT)[8-10]具有时-频局部化特征，是一种新型的信号处理工具，它被广泛地应用于光谱信号的平滑滤噪、数据压缩、基线校正与背景扣除[11,12]等，可有效地去除漫反射光谱的低频背景和高频噪声，提高信噪比。

间隔偏最小二乘回归方法(interval partial least-squares regression，iPLS[13])是对红外光谱数据定量分析的建模方法，它是一种波长筛选方法[14]，该法主要用于筛选偏最小二乘法建模的波长区域。它将全波段光谱数据均分为n(n为自然数)等分，然后在每个波段上用PLS模型建立一个线性回归模型，各个波段的回归模型以RMSE值作为评判准确度的标准。这种方法将波段进行了分隔，有效地选择出了建模的最佳波段，避免了“过拟合”和“欠拟合”的现象，减少了误差，提高了模型的准确度。

本文以烃源岩的红外漫反射光谱为例，将DWT-iPLS应用于数据的预处理及建立数学模型，将结果与未用DWT预处理的方法相比较，准确度提高了很多。

1 基本理论

小波是为满足一定条件的函数通过平移和伸缩而产生的一个函数族[15]，即：

$$\Psi_{a,b}(t)=\frac{1}{\sqrt{|a|}}\Psi\left(\frac{t-b}{a}\right),\ (a,b\in R,\ a\neq 0) \quad (1)$$

式中：a是尺度参数，b是平移参数，分别用于控制伸缩和位置；$\Psi(t)$为小波基。

小波变换定义为某函数$f(t)\in R^2$在小波上的投影，即$f(t)$和$\Psi_{a,b}(t)$的内积，

$$Wf(a,b)=\langle f(t),\Psi_{a,b}(t)\rangle=\frac{1}{\sqrt{|a|}}\int_{-\infty}^{+\infty}f(t)\Psi_{a,b}(t)\mathrm{d}t \quad (2)$$

对于离散小波变换[16,17]，将变量a和b做如下取值：

收稿日期：2007-05-06，**修订日期**：2007-08-09

基金项目：国家教育部“振兴计划”项目(A01504)资助

作者简介：宋 宁，1978年生，南开大学泰达应用物理学院博士研究生 e-mail：sning@mail.nankai.edu.cn

$$\begin{cases} a = a_0^m \\ b = na_0^m b_0 \end{cases} \quad (m,n \in Z,\ a_0 \neq 0) \tag{3}$$

(1)和(2)式的离散形式为：

$$\Psi_{m,n}(t) = a_0^{-\frac{m}{2}} \Psi(a^{-\frac{m}{2}} t - nb_0) \tag{4}$$

$$Wf(m,n) = \langle f(t), \Psi_{m,n}(t) \rangle = a_0^{-\frac{m}{2}} \int_{-\infty}^{+\infty} f(t) \Psi(a_0^{-m} t - nb_0) \mathrm{d}t \tag{5}$$

离散小波变换得到的小波系数 $Wf(m,n)$ 表示函数 $f(t)$ 中用小波函数 $\Psi_{m,n}(t)$ 所表示的分量。小波变换就是将任意函数或信号表示为小波函数的线性组合，即

$$f(t) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} Wf(m,n) \Psi_{m,n}(t) \tag{6}$$

在离散小波变换的计算方法中，应用最广的是 Mallat 提出的多尺度信号分解(MRSD)算法[18-20]。其计算过程可以用图 1 来表示。

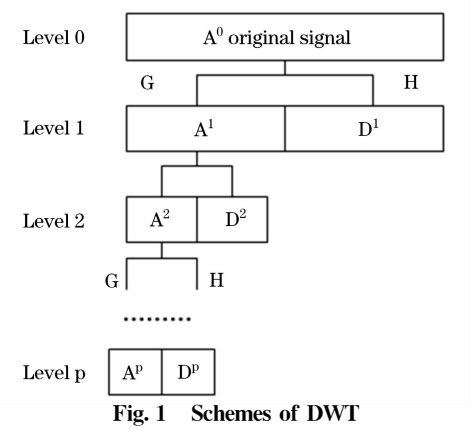

Fig. 1　Schemes of DWT

A: denotes approximation;
D: denotes detail coefficients

某信号经过小波分解后将从原来的空间投影到小波空间，我们得到一系列的小波系数。在这些系数中，有一部分特别小，代表着高频噪声，将这些小的系数扣除不会丢失有意义的信息。还有一部分是低频部分，因为背景和基线往往是信号中频率最低的组成成分，扣除这些低频部分可以实现背景信号的扣除或基线的校正。DWT 对光谱数据的预处理由下面 3 个步骤完成：(1)将原始信号进行小波变换，得到小波系数；(2)将小波系数中代表高频噪声和低频背景的系数删除，得到压缩后的数据；(3)对压缩的信号进行重构，得到扣除了背景干扰和高频噪声的光谱数据。经过小波变换处理的数据，大大降低了数据量，即有效地消除了噪声和背景干扰，又能提高多元校正的速度，同时，经过处理的数据，减少了变量，减少了模型的随机性从而提高了预测精确度。经过 DWT 预处理的光谱数据再用 iPLS 方法分析建模(关于 iPLS 的理论详见文献[13])，即可得到较为理想的线性回归模型。

2　实验部分

2.1　样品和实验仪器与软件

2.1.1　样品

生烃潜含量不同的 22 种烃源岩；仪器：WQF-510 傅里叶变换红外光谱仪(北京瑞利分析仪器公司)，波数范围是 7 800～400 cm^{-1}，分辨率是 0.5 cm^{-1}，分束器为 KBr 基片，探测器是 DTGS，光源是高强度空气冷却红外光源，漫反射附件。

2.1.2　软件

Matlab 的 WaveLab 工具箱和 iPLS 工具箱。

2.2　测定方法

取烃源岩样品，放入研钵中研磨，过 160 目的筛子，将滤出的样品放入样品杯中进行 IR 光谱测量，测试范围为 1 400～4 000 cm^{-1}，分辨率为 4 cm^{-1}，每隔 4 个样品重新校正背景。图 2 为样品的光谱图，可以看出谱图中具有一定的程度的噪声干扰并且不同样品之间又有较严重的基线漂移。

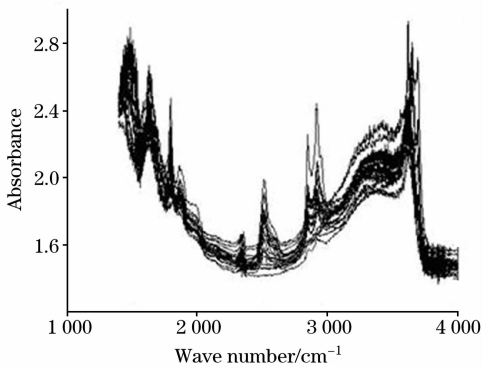

Fig. 2　The infrared diffuse reflection spectrums of 22 samples

2.3　数据处理

用 matlab 的 WaveLab 工具箱和 iPLS 工具箱及结合自编的一些程序，首先用 iPLS 方法对经过平滑消噪处理的谱图进行建模，然后再采用 DWT-iPLS 方法对谱图进行建模，比较可得，DWT-iPLS 方法预测均方根误差(*RMSE*)要远远小于经过平滑处理的预测模型，并且提高了信噪比，消除了漫反射带来的背景干扰。

3　结果与讨论

3.1　DWT 对光谱数据的预处理

小波基是小波变换的核心，不同的小波基在处理同一信号时会表现出不同的效果。在建模过程中以 RMSE 的值为评价标准，取对应于最小 RMSE 值为最佳小波基，RMSE 值计算如下：

$$\mathrm{RMSE} = \sqrt{\frac{\sum (y_{\mathrm{pred}} - y_{\mathrm{ref}})^2}{N}} \tag{7}$$

式中 N 为样品数，y_{pred} 为样品的预测值，y_{ref} 为样品的实测值。

在建模之前我们分别采用正交小波基 Daubechies(db)，symmlet(sym)及 Coiflet(coif)等小波基对数据进行预处理，最终我们选择了 *RMSE* 值最小的 coif2 为最佳小波基。小波变换的分解尺度对数据的压缩效果也有较大的影响，通过采用不同的分解尺度的 *RMSE* 值比较，我们发现当分解尺度

为 8 时，能够有效地去除背景和噪声的干扰，所以我们选择分解尺度为 8。图 3 为分解尺度为 8 的经 coif2 小波变换得到的离散逼近和离散细节，分别将 1 和 8 置为零再重构就得到了扣除低频背景和高频噪声的信号，如图 4。

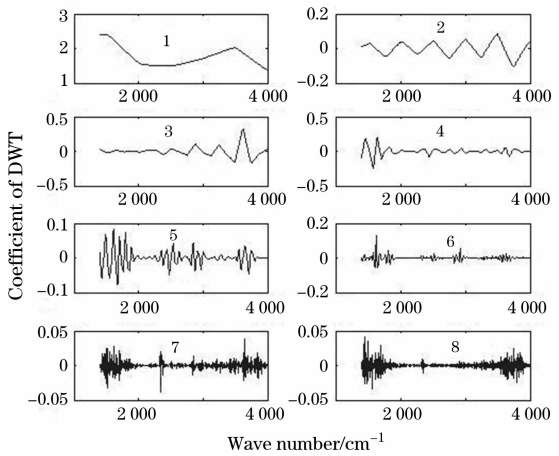

Fig. 3 The approximation and detail coefficient after coif2 transform

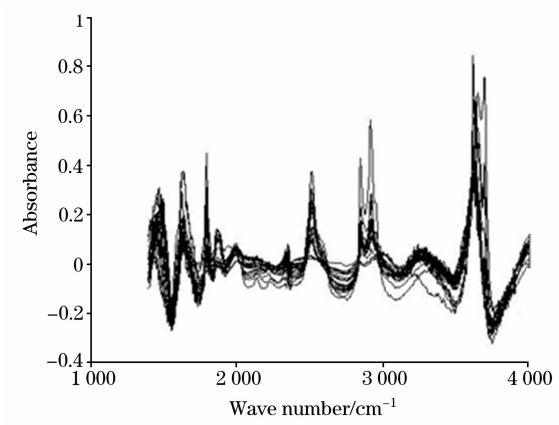

Fig. 4 The infrared diffuse reflection spectrums after DWT

Table 1 The influence of intervals on correlation and RMSE

间隔数	相关性(r)	RMSE
10	0.958 7	1.677 4
20	0.966 8	1.525 4
30	0.959 5	1.616 7
40	0.964 0	1.539 5
50	0.921 8	2.2301
60	0.968 6	1.419 0
70	0.944 1	1.917 2
80	0.931 2	2.085 1
90	0.943 7	1.895 6
100	0.939 4	1.964 6

3.2 DWT-iPLS 建模效果的评价

经过 DWT 预处理的数据，应用 iPLS 工具箱建模。为了得到理想的模型，在建模过程中应选择合适的间隔数。我们将整个谱图从取 10 个间隔到取 100 个间隔，将他们的 *RMSE* 值相比较得出 60 段为最合理的间隔数，见表 1。

将谱图分成 60 个间隔，对每个间隔段内的数据进行 PLS 线性回归分析，得出第 25 段间隔的线性相关性最高(r=0.968 6)，RMSE 值最小(RMSE=1.419 0)。此段的波数为 2 935.13～2 892.7 cm^{-1}，此段为—CH_2 的红外吸收特征峰，其峰值与烃源岩中的总生烃潜含量有一定的线性关系。见图 5。

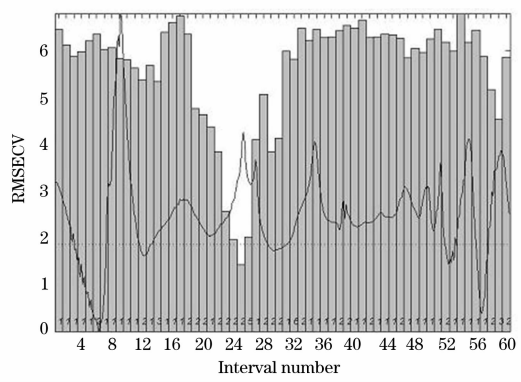

Fig. 5 The PLS regression analysis in 60 intervals

选取最佳的第 25 段(波数为 2 935.13～2 892.7 cm^{-1})建立预测模型，相关系数 r 为 0.968 6，RMSE 值为 1.419 0。将 DWT-iPLS 模型(图 6)与经过平滑处理的 iPLS 模型(图 7)

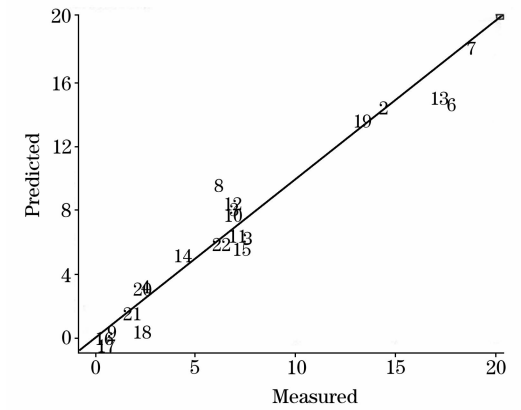

Fig. 6 DWT-iPLS linear regression model

r=0.968 6; RMSECV=1.419 0; Bias=0.003 4

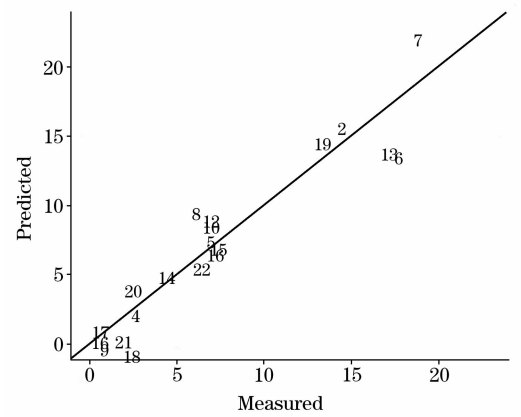

Fig. 7 iPLS linear regression model

r=0.950 3; RMSECV=1.868 8; Bias=0.095 3

相比较可以得出，经过离散小波变换扣除背景和高频噪声的光谱数据，在线性相关性和 RMSE 上都有很大的提高。

4　结　论

将 DWT 用于光谱数据的预处理，消除了背景和高频噪声的干扰，再与 iPLS 技术相结合，建立了用于复杂无机样品的微量有机含量的分析模型方法。与传统的光谱数据预处理方法和建模方法相比，DWT-iPLS 方法在预测的准确度方面有一定优势。离散小波变换快速有效的消除了光谱中的噪声和背景干扰，进一步的提高了 iPLS 方法的预测准确度。因此，DWT-iPLS 方法可作为一种新型的建模方法，在实际复杂样品体系的光谱分析中发挥作用。

参　考　文　献

[1] LU Wan-zhen, YUAN Hong-fu, XU Guang-tong, et al(陆婉珍，袁洪福，徐广通，等). The Modern Analysis Technique of Near Infrared Spectrum(现代近红外光谱的分析技术). Beijing: China Petrochemical Press(北京：中国石化出版社), 2000. 1.

[2] QU Hai-bin, LIU Quan, CHENG Yi-yu(瞿海斌，刘　全，程翼宇). Chinese Journal of Analytical Chemistry(分析化学), 2004, 32(4): 477.

[3] Nagarajan R, Gupta D, Varma S P. Infrared Physics & Technology, 2002, 43(6): 377.

[4] Andres J M, Bona M T. Analytical Chimical Acta, 2005, 535: 123.

[5] Gustav Kortum. Reflectance Spectroscopy Principles, Methods, Applications. Berlin; Heidelberg; New York: Spinger-Verlag, 1969. 58.

[6] ZHAO Li-li, ZHAO Long-lian, LI Jun-hui, et al(赵丽丽，赵龙莲，李军会，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2004, 24(1): 41.

[7] QU Hai-bin, YANG Hai-lei, CHENG Yi-yu(瞿海斌，杨海雷，程翼宇). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(1): 60.

[8] Jetter K , Depczynski U, Molt K, et al. Anal. Chim. Acta, 2000, 420: 169.

[9] SHAO Xue-guang, PANG Chun-yan, SUN Li(邵学广，庞春艳，孙　莉). Progress in Chemistry(化学进展). 2000, 12(3): 233.

[10] Chen Da, Shao Xueguang, Hu Bin, et al. Analytica Chimica Acta, 2004, 511: 37.

[11] Johan Trygg, Svante Wold. Chemometrics Intelligent Laboratory Systems, 1998, 42(1): 209.

[12] YING Yi-bin, LIU Yan-de, FU Xia-ping(应义斌，刘燕德，傅霞萍). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(1): 63.

[13] Norgaard L, Saudland A, Wagner J. Appl. Spectrosc. , 2000, 54(3): 413.

[14] CHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立，袁洪福，陆婉珍). Progress in Chemistry(化学进展), 2004, 16(4): 528.

[15] XU Lu, SHAO Xue-guang(许　禄，邵学广). Methods of Chemometrics(化学计量学方法). Beijing: Science Press(北京：科学出版社), 2004. 202.

[16] Tan Hu-wei, Brown Steven D. Chemometrics, 2002, 16: 228.

[17] Barclay V J, Bonner R F. Analytical Chemistry, 1997, 69: 78.

[18] Walczak B, Massart D L. Chemometrics and Intelligent Laboratory Systems, 1997, 38: 39.

[19] Alexander K L , Chau F, Gao J. Chemometrics and Intelligent Laboratory Systems, 1998, 43: 69.

[20] Walczak B, Massart D L. Chemometrics Intelligent Laboratory Systems 1997, 36: 81.

DWT-iPLS Applied in the Infrared Diffuse Reflection Spectrum of Hydrocarbon Source Rocks

SONG Ning[1, 3], XU Xiao-xuan[1, 3], WU Zhong-chen[2], ZHANG Cun-zhou[1, 3], WANG Bin[1]

1. The TEDA Applied Physics School, Nankai University, Tianjin 300457, China
2. Department of Space Science and Applied Physics, Shandong University at Weihai, Weihai 264209, China
3. The Key Laboratory of Advanced Technique and Fabrication for Weak-Light Nonlinear Photonics Materials, Ministry of Education, Nankai University, Tianjin 300457, China

Abstract Infrared spectroscopy is useful to monitor the quality of products on-line, or to quality multivariate properties simultaneously. The IR spectrometer satisfies the requirements of users who want to have quantitative product information in real-time because the instrument provides the information promptly and easily. However, Samples that are measured using diffuse reflectance often exhibit significant differences in the spectra due to the non-homogeneous distribution of the particles. In fact, multiple spectral measurements of the same sample can look completely different. In many cases, the scattering can be an overpowering contributor to the spectrum, sometimes accounting for most of the variance in the data. Although the degree of scattering is dependent on the wavelength of the light that is used and the particle size and refractive index of the sample, the scattering is not uniform throughout the spectrum. Typically, this appears as a baseline shift, tilt and sometimes curvature, where the degree of influence is more pronounced at the longer-wavelength end of the spectrum. The diffuse reflection spectrum is unsatisfactory and the calibration may provide unsatisfactory prediction results. So we must use some methods to remove the effects of the scattering for multivariate calibration of IR spectral signals. Discrete wavelet transform (DWT) is a good method to remove the effects of the scattering for multivariate calibration of IR spectral signals. By using DWT on individual signals as a preprocessing method in regression modeling on IR spectra, good compression is achieved with almost no loss of information, the low-frequency varying background and the high-frequency noise be removed simultaneously. In this report, we use the iPLS method to establish the calibration models of hydrocarbon source rocks. iPLS is a new regression method and the authors can get better results by using DTW- iPLS.

Keywords Diffuse reflectance; Correlation; DWT; iPLS; RMSE

(Received May 6, 2007; accepted Aug. 9, 2007)