第29卷第5期
2010年10月

红外与毫米波学报
J. Infrared Millim. Waves

Vol. 29, No. 5
October, 2010

文章编号:1001-9014(2010)05-0397-04

基于小波数据压缩的光谱技术在水质分析中的应用

潘国锋[1], 杨慧中[1], 孔 军[1,2]

(1. 江南大学 通信与控制工程学院,江苏 无锡 214122;2. 中国科学院上海技术物理研究所,上海 200083)

摘要:一定浓度氮磷水样的紫外吸收光谱数据量非常大,采用基于软阈值的小波变换可以对这些光谱数据进行有效压缩.不同浓度氮磷水样的紫外吸收光谱信号之间存在很强的相关性,利用偏最小二乘回归(PLSR)方法对光谱信号的强度和水样中氨氮浓度之间的关系进行回归建模可以降低这种相关性的影响,提高所建模型的拟合精度.实际水样测试数据的建模结果表明,用这种方法所建立的模型,氮磷浓度检测的最大相对误差为8.9%,完全满足检测精度的要求.

关 键 词:小波变换;偏最小二乘回归;吸收光谱;软阈值

中图分类号:TP731 **文献标识码**:A

APPLICATION OF SPECTROSCOPY TECHNIQUE TO WATER QUALITY ANALYSIS BASED ON WAVELET DATA COMPRESSION

PAN Guo-Feng[1], YANG Hui-Zhong[1], KONG Jun[1,2]

(1. School of Communication&Control, Jiangnan University, Wuxi 214122, China;
2. Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China)

Abstract: Data of the UV absorption spectra from the water samples with a certain concentration of the nitrogen and phosphorus is very large. Wavelet transformation (WT), based on the soft-threshold, can be effectively used to compress the spectrum signal of UV absorption spectra. High correlation often exists among the spectral intensity for the water samples with different concentration of the nitrogen and phosphorus. Partial least square regression (PLSR) can be used to decrease such correlation and to build an effective regression model for the sampled water. The detection precision is also increased in this way. Simulation of the tested water samples showed that the maximum relative error in the concentrations of nitrogen and phosphorus was 8.9%. The model can fully meet the requirements of detection precision.

Key words: wavelet transform; partial least square regression; absorption spectroscopy; soft-threshold

引言

随着经济的快速发展和人口的急剧增加,水体的富营养化已成为公害,水体中的氨氮、总磷成为衡量水体是否富营养化的重要指标,也是水质监测的一项重要任务.由于被检测水样成份相对复杂,除含氮和磷的化合物外,还含有大量其它未知物质,此类物质对水样检测会产生干扰.传统的光谱分析方法分析上述水样存在很大难度,研究和探索新的水样光谱分析技术成为必然[1].

多元校正技术通过校正集样品的光谱数据与组成成分数据的量测,采用合适的化学计量学方法建立关联两者的校正模型,通过所建立的校正模型和未知样品的光谱数据推定未知样品的组分,成为光谱信息处理的有效途径.为了最大限度地获取谱图信息,对复杂体系的分析常采用全光谱校正技术对整个谱区的光谱信息和待预测的组成或性质指标进行相关分析,从中优选出少数相关系数高的波长点进行定量分析并获得最佳的分析结果.由于全光谱分析信号的采集通常包括数千个数据点,光谱信息处理面临一个高维数据空间,计算量大,建模效率低.相关系数法、主成分分析(PCA)法、偏最小二乘

收稿日期:2009-09-03,**修回日期**:2010-04-10 **Received date**: 2009-09-03, **revised date**: 2010-04-10
基金项目:国家自然科学基金(60674092);江苏省高技术研究项目(工业)(BG2006010)
作者简介:潘国锋(1972-),男,浙江湖州人,讲师,博士研究生,主要研究方向为过程参数检测及信号处理,E-mail:pgf134@sina.com.

回归分析(PLSR)法是多元分析中常用的降维技术,它们通过分解算法从光谱信息矩阵中提取少数几个相互正交的隐变量,由于这几个隐变量能反映原始光谱信息的最大变差,既可以降维,又能充分利用光谱信息.不过当光谱信息矩阵很大时,提取隐变量的计算量依然很大,不利于快速建立校正模型.

利用小波变换的时—频局部化和能对信号进行多尺度多分辨率细化分析的优良特性,针对建立紫外光谱定量分析模型中面临的高维数据处理对象,利用小波变换压缩技术实现降维处理并建立相应的定量分析预测模型,可以大大提高建模效率.本文采用小波变换作为光谱数据的压缩工具,并直接利用压缩后的数据,通过 PLSR 方法,对水体中的氮磷含量建立分析模型,可以减少存储空间,提高建模和分析的速度.

1 数据压缩算法

1.1 基于多分辨率分析的数据压缩方法

对于水体的吸光度信号 $f(x)$,其连续小波变换定义为:

$$W_f(a,b) = \int_R f(x)\Psi_{(a,b)}(x)\mathrm{d}x = \frac{1}{\sqrt{a}}\int_R f(x)\Psi\left(\frac{x-b}{a}\right)\mathrm{d}x \quad , \tag{1}$$

其中,$\Psi_{(a,b)}$ 为小波函数,参数 a 为水体吸光度信号的不同频率,参数 b 为不同时间处的波长.可以看出,每一个不同的吸光度信号,都有一个小波系数 $W_f(a,b)$ 与其对应.如果小波系数很小,对吸光度信号的表示没有意义.如果将其去除,在重构的信号中将不会丢失有意义的信息.因此,小波变换可用于吸光度信号的压缩.

光谱数据的小波压缩包括 3 个步骤:

①将原始光谱信号进行小波变换,得到小波空间系数;②将系数中足够小的系数删除,得到压缩后的数据;③对压缩后的信号进行重构.这样,在用多元校正方法建立模型时,可以取较大的小波系数组成新的数据矩阵代替原始数据矩阵,大大降低数据量,既能有效消除噪声,又能提高多元校正的速度.同时,采用较少的变量建模,有利于减少模型的随机性并提高预测精度[2~4].

在光谱信号处理中,利用 Mallat 快速小波分解算法,其分解公式为[4]:

$$C_n^j = \sum_{k\in z} h_{k-2n}C_k^{j-1} \quad ,$$

$$D_n^j = \sum_{k\in z} g_{k-2n}C_k^{j-1} \tag{2}$$

其中,$k=1,2,\cdots,N$,N 是输入序列的个数.C_n^j 是被采样的吸光度信号分解后的低频分量,D_n^j 是分解后的高频分量,j 代表分解的层数,即第 j 级分解.h_{n-2k} 和 g_{n-2k} 分别是小波分解后的高通滤波器系数和低通滤波器系数.进行 j 尺度的小波分解所得到小波系数的总数据点数与原始信号相同,由于在小波系数中有大量的系数值很小,忽略这些系数不会带来信息的丢失,因此小波变换是一种高效的数据压缩方法.

Mallat 算法中的小波重构公式为[4]:

$$C_n^{j-1} = \sum_{k\in z} h'_{n-2k}C_k^j + \sum_{k\in z} g'_{n-2k}D_k^j \quad , \tag{3}$$

其中,C_k^j 是小波重构后的低频分量序列,D_k^j 是高频分量序列,并且它们的长度是序列 C_k^{j-1} 和 D_k^{j-1} 的长度的一半.h'_{n-2k} 和 g'_{n-2k} 分别是小波重构后的高通滤波器系数和低通滤波器系数.

在对原始采样数据进行小波分解得到 D_k^j 以后,可以对其选用适当的阈值函数进行阈值处理,以最大程度地恢复原始光谱信息.

1.2 阈值选取及压缩效果评判方法

如何选取阈值是光谱数据压缩的核心.本文采用如下软阈值法确定阈值[5]

$$T = \gamma\sigma\sqrt{2N} \quad , \tag{4}$$

式(4)中,σ 为噪声强度,计算公式为:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}(d^{n-1}-\bar{d})^2} \quad , \tag{5}$$

$\bar{d}$ 为 d^{n-i} 的均值.原始光谱信号序列 A^n 的长度为 2^N,γ 为常数.

为使原始光谱数据压缩后不丢失有用信息,在保证一定压缩比的情况下,要求压缩失真率 P 小于某一设定值 ε,即:

$$P = \frac{\| A_m^n - \overline{A^n} \|_2}{\| \overline{A^n} \|} < \varepsilon \quad , \tag{6}$$

若不能满足这一指标,可调整 γ,修改小波系数阈值 T 重新进行压缩.

2 浓度解析算法

经上述方法压缩的光谱数据组成的序列 A_M^n,是被测水样中各物质吸收光谱相互叠加的结果.多组分吸收光谱的相互干扰会严重影响分析检测精度,简单利用朗伯-比尔定律会产生很大的误差[6].采用

PLSR 方法对待测成分浓度进行解析. 其建模过程主要包括以下几步[7~10].

2.1　数据的标准化处理

标准化处理可以使不同量纲、不同数量级的数据能在一起进行比较,其过程为:

$$\begin{cases} E_{0_i} = x^* = (x_i - \overline{x_i})/S_{x_i} \\ F_0 = y^* = (y - \overline{y})/S_y \end{cases}, \quad (7)$$

式中, E_{0_i}、F_0 分别为光谱数据 x_i、浓度 y 的标准化向量;$\overline{x}$、$\overline{y}$ 分别为 x_i、y 的均值;S_{x_i}、S_y 分别为 x_i、y 的均方差.

2.2　第一主成分 t_1 的提取

从 E_0 中提取第一主成分 t_1,从 F_0 中提取第一主成分 ,

$$\begin{cases} t_1 = E_0 w_1 \\ w_1 = E_0^T F_0 / \| E_0^T F_0 \| \\ u_1 = F_0 C_1 \end{cases}, \quad (8)$$

式中,w_1 为 E_0 的第一个轴,C_1 为 F_0 的第一个轴,$\| C_1 \| = 1$.

2.3　第 h 主成分的提取

以 E_h 代替 E_{h-1},F_h 代替 F_{h-1},用上面的方法可以求出第 h 主成分:

$$\begin{cases} t_h = E_{h-1} w_h & w_h = E_{h-1}^T F_{h-1} / \| E_{h-1}^T F_{h-1} \| \\ E_h = t_h p_h^T + E_{h-1} & F_h = t_h r_h^T + F_{h-1} \\ p_h = E_h^T t_h / \| t_h \|^2 & r_h = F_{h-1}^T t_h / \| t_h \|^2 \end{cases}, \quad (9)$$

2.4　偏最小二乘回归模型的推导

F_0 关于 $t_1 \sim t_h$ 成分的回归方程为:

$$F_0 = r_1 t_1 + r_2 t_2 + \cdots + r_h t_h, \quad (10)$$

$$t_j = E_{j-1} w_j = E_0 w_j^* = \prod_{i=1}^{j-1} (I - w_i p_i^T) w_j . \quad (11)$$

则标准化变量的回归方程为:

$$\begin{cases} \hat{y}^* = \sum_{i=1}^{m} \alpha_i x_i^* \\ \alpha_i = \sum_{j=1}^{h} r_j w_{ij}^* \end{cases}, \quad (12)$$

可以采用"舍 2 交叉验证法"来确定主成分数 m,具体确定过程见文献[8].

3　实验与结果分析

为了验证上述方法对水体中氮磷浓度检测的有效性,用硝酸钠($NaNO_3$)、磷酸钠(Na_3PO_4)、氯化钠(Nacl)、蒸馏水(H_2O)等配制不同组分、不同浓度的氮、磷的混合溶液,将该混合溶液用 U-3000 型紫外分光光度计进行光谱扫描,扫描参数为:扫描步长为 0.1nm,扫描速度为 60 次/分钟. 扫描范围从 200nm 到 400nm. 图 1 是氮、磷浓度分别为 0.2mg/L、0.6mg/L 的混合溶液的扫描谱图. 从谱图可以看出,该混合溶液对波长大于 341nm 的紫外光的吸收度趋于 0. 对所有水样,本文选取波长在 200 ~ 341nm 之间,每隔 1nm 进行一次采样得到光谱数据,对这些数据经过小波压缩后再进行 PLSR 解析. 经过对光谱数据使用不同层数和不同的小波基进行压缩后的数据处理效果的分析比照后发现:选用 DB4 小波基,小波变换的层数为 4 进行小波变换,可以将经过压缩的数据由原来的 142 个波长点压缩到 32 个波长点,数据压缩比达到 4,大大减轻了数据处理的运算量,并且浓度检测的误差也达到最小.

在使用紫外分光光度计进行光谱扫描分析时,可以将一组 $n = 20$ 的已知氮、磷浓度水样的紫外吸收光谱数据集,用作训练样本求得回归系数矩阵,建立数学模型,从而根据待测水样的光谱信息计算氮磷的浓度.

训练样本不同主成分数 h 所对应的 PRESS 和浓度矩阵与不同成分 t_i 之间的相关系数如表 1 所示. 可以看出,当 $h = 6$ 时,预测残差平方和最小,可选择 $h = 6$ 建立 PLSR 回归模型. 同时,各成分与自变量之间的相关系数表明,氮、磷的浓度分布与第一成分 t_1 的相关程度最高,与其它成分相关程度较小.

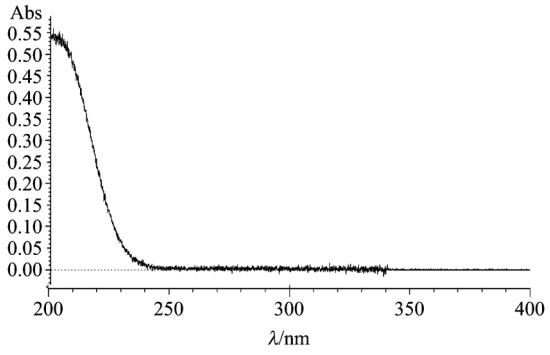

图 1　氮、磷浓度分别为 0.2mg/L、0.6mg/L 混合液的扫描谱图

Fig. 1　Scanning calorimetry of the mixed solution of 0.2mg/L nitrogen and 0.6mg/L phosphorus

表 1　不同成分与自变量的相关系数及 PRESS 表

Table 1　Table of the PRESS and the correlation coefficient among the different components and variables

成分	t_1	t_2	t_3	t_4	t_5	t_6	t_7
氮	0.9467	0.2489	-0.1417	0.0292	-0.0341	-0.0149	-0.0118
磷	0.9654	0.16	0.1687	-0.1129	0.0226	0.0191	-0.0138
PRESS(h)	0.105	0.779	1.3935	0.0347	0.0425	0.0003406	0.1412

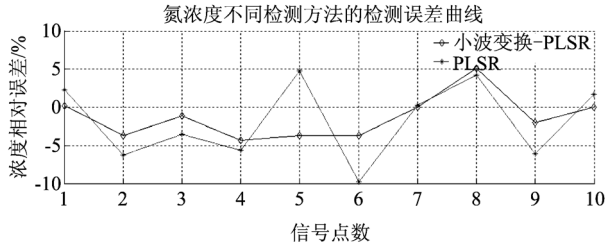

图2 氮浓度检测结果相对误差比较

Fig. 2 Comparison of relative error of the nitrogen's concentration

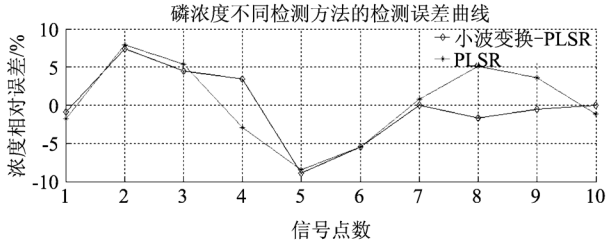

图3 磷浓度检测结果相对误差比较

Fig. 3 Relative error of the phosphorus's concentration

为了比较小波压缩光谱信号的效果,运用上述20组训练样本水样分别进行光谱信号小波压缩并进行PLSR回归和直接进行PLSR回归,得到两个回归模型,再分别对10组不同浓度的氮、磷组成的待测水样进行预测,得到如图2和图3所示的检测结果的相对误差. 从图中可以看出,检测的最大相对误差为8.9%,符合这方面的国家检测标准.

表2列出了不同检测方法的最大相对误差和标准差. 可以看出,与没有经过小波压缩的PLSR方法相比,本文方法所建模型不仅检测结果精度得到提高,参与建模的数据也要少得多. 因此,小波变换-PLSR方法兼具有小波分析的滤波能力和PLSR充分利用浓度矩阵信息的优点,是进行污水中氮磷检测的一种较好的检测方法.

表2 不同检测方法的相对误差和误差的标准差

Table 2 Relative error and the standard error of the different presisions

PLSR最大检测相对误差(%)		小波变换-PLSR最大检测相对误差(%)		PLSR最大检测误差标准差(mg)		小波变换-PLSR最大检测误差的标准差(mg)	
氮	磷	氮	磷	氮	磷	氮	磷
9.52	8.975	5.1	8.9	0.0381	0.0737	0.0239	0.0502

4 结论

通过基于小波变换的偏最小二乘回归方法对水样中的氮磷含量进行了检测. 检测结果表明,该方法具有运算量小、运算精度高的优点,与常用的其它氮磷检测方法相比,提高了检测精度.

REFERENCES

[1] Wu Xi-Ping. Study on Prediction Models of Determination of Chemical Oxygen Demand (COD) in Wastewater by Near-Infrared Spectroscopy Based on Matlab[D]. *Fujian Agriculture University*(吴喜平. 基于Matlab的水质COD近红外光谱预测模型的研究. **福建农业大学**),2007.

[2] YUE Quan-Ming, YU Wei-Yung, BAI Chuan-Jun, *et al*. Novel compression scheme of fault recording data in power systems based on lifting algorithm[J]. *Automation of Electric Power Systems*(乐全明,郁惟墉,柏传军,等. 基于提升算法的电力系统故障录波数据压缩新方案. **电力系统自动化**),2005,**29**(5):74—78.

[3] Wang Aiping, Wang Huinan. 1/f noise eliminating based on wavelet analysis[J]. *Journal of Data Acquisition & Processing*(王爱萍,王惠南. 基于小波分析的1/f噪声降噪. **数据采集与处理**),2006,**21**(2):218—221.

[4] Mallat S. A Theory for multiresolution signal decomposition: the wavelet representation[J]. *IEEE Trans. on Pattern Anal. Mach. Intell*,1989,**11**(7):674—691.

[5] Ren Xiao-Mei, Wang Zhi-Zhong, Hu Xiao. EMG signal decomposition based on wavelet transform and ICA method [J]. *Journal of Data Acquisition & Processing*(任小梅,王志中,胡晓. 应用小波变换和ICA方法的肌电信号分解. **数据采集与处理**),2006,**21**(3):272—276.

[6] LI Qiong-Fei, YANG Zeng-Ling, HAN Lu-Jia. Analysis of cattle and sheep content In pig or poultry meat and bone meal by near infrared reflectance spectroscopy[J]. *Journal of Infrared and Millimeter Waves*(李琼飞,杨增玲,韩鲁佳. 肉骨粉中牛羊源成分含量的近红外漫反射光谱分析[J]. **红外与毫米波学报**),2007(6):414—418.

[7] LI Jun-hui, QIN Xi-yun, ZHANG Wen-juan. Influence of LPLS algorithm parameters on NIR veracity[J]. *Spectroscopy and Spectral Analysis*(李军会,秦西云. 局部偏最小二乘回归建模参数对近红外检测结果的影响研究. **光谱学与光谱分析**),2007,**27**(2):262—264.

[8] CHENG Zhong. Research on Several Key Technologies of Partial Least Squares Regression in Chemistry and Chemical Process Modeling[D]. *Zhejiang University*(成忠. PLSR用于化工建模的几个关键问题的研究. **浙江大学**),2005.

[9] LIU Bo-Ping, QIN Hua-Jun, LUO Xiang. Determination of four contents of feedstuff powder using near infrared spectroscopy by PLS-BP Model[J]. *Spectroscopy and Spectral Analysis*(刘波平,秦华俊,罗香. PLS-BP法近红外光谱同时检测饲料组分的研究. **光谱学与光谱分析**),2007,**27**(10):2007—2009.

[10] LIU Ming-Yang, MENG Yu, REN Yu-Lin, *et al*. Nondestructive quantitative analysis of cofrel medicines by improved partial least squares-NIR spectroscopy[J]. *Spectroscopy and Spectral Analysis*(刘名扬,孟昱,任玉林,等. 改进的偏最小二乘-近红外光谱法非破坏定量分析Cofrel药品. **光谱学与光谱分析**),2007,**27**(6):1098—1101.