基于小波变换和支持向量机的光谱多组分分析*

熊宇虹 温志渝 陈 刚 黄 俭 徐 溢

(重庆大学光电工程学院,重庆 400044)

摘 要 以符合朗伯—比尔定律的光谱信号为研究对象,在运用小波变换对光谱信号进行去除噪声处理的基础上,建立了基于支持向量机的多组分分析模型,最后采用计算机模拟的方式对该方法进行了举例说明. 实例表明,该方法能较好地解决非线性、小样本条件下的多组分分析问题.

关键词 光谱分析;小波变换;支持向量机

中图分类号 TP39 **文献标识码** A

0 引言

光谱信号的多组分分析技术是光谱分析技术中的重要组成部分,常用的分析方法有多元线性回归、K 矩阵、卡尔曼滤波、主成分回归、神经网络等. 实际应用时,在仪器噪声干扰、多组分体系交互影响较大、非线性关系明显且已知标准样本少的情况下,上述方法的分析效果均不大理想,因而研究减少噪声干扰、解决非线性、小样本条件下的多组分分析问题也就成为当务之急了.

支持向量机是 Vapnik 等人根据统计学习理论提出的一种建立在结构风险最小化原则的基础上,专门研究小样本情况下统计估计和预测的问题,探索在现有有限样本的情况下如何得到最优解的通用学习方法,体现了兼顾经验风险和置信范围的一种折中的思想,能较好地解决小样本、非线性、高维数等实际问题[1,2]. 本文以符合朗伯—比尔定律的光谱为研究对象,探讨了支持向量机方法在光谱信号的多组分分析中的应用,在简要介绍运用小波变换对光谱信号进行去除噪声处理的基本方法和支持向量机基本原理的基础上,建立了基于支持向量机的多组分分析模型,最后采用计算机模拟的方式对该方法进行了相应的实例分析.

1 基于小波分析的去噪处理

光谱信号包含一定的仪器噪声,这些噪声不可避免地对分析结果产生着影响,因而对分析信号进行去噪预处理是必须的. 近年来,小波分析的方法被广泛应用在信号处理方面. 去噪处理的小波分析方法的基本思想是:利用小波函数,通过小波变换,把信号变换到小波域,在小波域对变换后的信号进行滤波,再通过小波逆变换得到重构的信号[3]. 对光谱信号而言,噪声通常是高频噪声,利用小波分解和重构来消除噪声的最简单方法是:将 Mallat 算法得到的分解信号中的高频系数直接置零,然后对信号进行重构,这样就可以得到去除了高频噪声的信号. 图 1 显示了小波方法去除噪声后的效果,为了便于对比,原始信号和滤波后的信号彼此错开了.

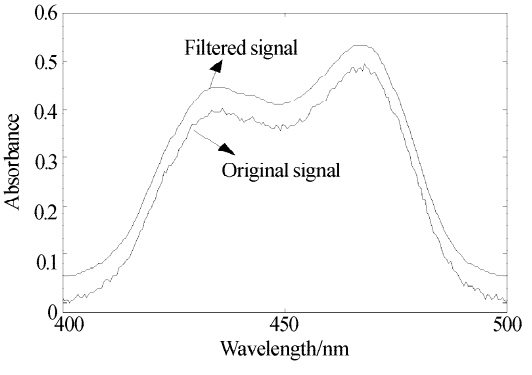

图 1 基于小波变换的去噪

Fig. 1 Denoise with wavelet transform

2 基于支持向量机的回归算法

支持向量机是从线性可分情况下的最优分类面发展而来的,基本思想可用两类线性可分问题进行说明. 图 2 中,○和□分别代表两类样本;H 为最优

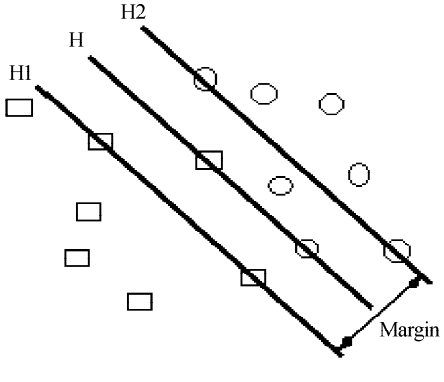

图 2 两类线性可分问题

Fig. 2 Linearity separated two classes problem

* 国家自然科学基金重点项目(69476023)、科技部国际合作项目(2004DFA00600)、国家 863 计划(2004AA4040,2004AA404023)和重庆市科委(CSTC,2005CF2002)资助项目

Tel:023-65104131 Email:xyh341@126.com

收稿日期:2004-08-23

分类线，不仅能将两类样本无错误地分开而且使分类间隔最大；H1、H2 分别为过各类中离分类线最近的样本且平行于分类线的直线，它们之间的距离叫做分类间隔．H1，H2 上的训练样本点就称作支持向量，相应地确定最优分类线的函数被称为支持向量机[4]．

支持向量机的方法也可用于回归分析，其运用思路和在模式识别中类似．首先考虑用线性回归函数 $f(x)=\langle w\cdot x\rangle+b$ 拟合数据 (x_i,y_i)，$x_i\in R^n$，$y_i\in R$，$i=1,\cdots l$ 的问题，并假设所有训练数据都可以在精度 ε 下无误差地用线性函数拟合，即

$$\begin{cases} y_i-w\cdot x_i-b\leqslant\varepsilon, \\ w\cdot x_i+b-y_i\leqslant\varepsilon, \end{cases} \quad i=1,\cdots l \tag{1}$$

考虑到允许拟合误差的情况，引入松弛因子 $\xi_i\geqslant 0$ 和 $\xi_i^*\geqslant 0$，则式(1)变成

$$\begin{cases} y_i-w\cdot x_i-b\leqslant\varepsilon+\xi_i, \\ w\cdot x_i+b-y_i\leqslant\varepsilon+\xi_i^*, \end{cases} \quad i=1,\cdots l \tag{2}$$

优化目标变成最小化 $\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)$，常数 $C>0$ 控制对超出误差 ε 的样本的惩罚程度．采用优化的方法可以得到其对偶问题．在式(3)的条件下，对 Lagrange 因子 a_i,a_i^* 最大化，式(4)为所示的目标函数．

$$\sum_{i=1}^{l}(a_i-a_i^*)=0,0\leqslant a_i,a_i^*\leqslant C,i=1,\cdots l \tag{3}$$

$$W(a_i,a_i^*)=-\varepsilon\sum_{i=1}^{l}(a_i^*+a_i)+\sum_{i=1}^{l}y_i(a_i^*-a_i)-\frac{1}{2}\sum_{i,j=1}^{l}(a_i^*-a_i)(a_j^*-a_j)\langle x_i\cdot x_j\rangle \tag{4}$$

得回归函数为

$$f(x)=\langle w\cdot x\rangle+b=\sum_{i=1}^{l}(a_i^*-a_i)\langle x_i\cdot x\rangle+b^* \tag{5}$$

这里 a_i,a_i^* 也将只有小部分不为 0，它们对应的样本就是支持向量，一般是在函数变化比较剧烈的位置上的样本，而且这里也是只涉及内积运算，只要用核函数 $K(x_i,x_j)$ 替代式(4)、(5)中的形如 $\langle x_i\cdot x_j\rangle$ 的内积运算就可以实现非线性函数拟合，其中，常用的核函数有多项式核函数、径向基函数(RBF)核函数、Sigmoid 核函数等[5]．

3　多组分分析模型的建立

设有 n 个由 h 种组分组成的样本，其浓度矩阵为 C，在 m 个分析通道测得相应的吸光度矩阵为 Y，如式(6)，多组分分析的目的就是根据已知样本的浓度和吸光度矩阵，建立相关的模型，在利用该模型去预测未知试样各组分的浓度．

$$C=\begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,n} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{h,1} & c_{h,2} & \cdots & c_{h,n} \end{bmatrix},Y=\begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,n} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ y_{m,1} & y_{m,2} & \cdots & y_{m,n} \end{bmatrix} \tag{6}$$

式中 $k=1,\cdots\cdots,m;i=1,\cdots,n;f=1,\cdots,h;y_{k,i}$ 是第 i 个样品在第 k 个分析通道处测量吸光度；$c_{f,i}$ 表示第 i 个样品中第 f 个组分的浓度．

由式(6)可以看出，多组分分析是一个多维输入多维输出的问题；而基于支持向量机的回归分析算法只能对多维输入，单维输出的情况进行处理．因此，要想利用它来进行多组分分析就存在一个如何将其推广到多维输入多维输出的问题．本文采用的方法是建立多个支持向量机，每个支持向量机对应一种组分，如有 h 种组分就相应地建立 h 个支持向量机，然后分别以已知标准样本校正集的吸光度矩阵 Y 为输入集合，以校正集的浓度矩阵 C 为目标集合对各个支持向量机进行训练，最后以未知试样的吸光度为输入，利用训练好的各个支持向量机就可得出各组分的浓度．其基本结构如图 3.

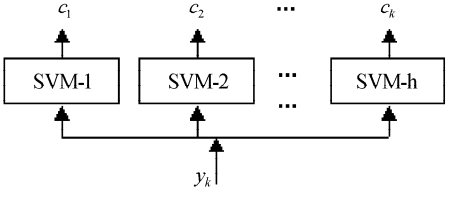

图 3　多组分分析的基本框架

Fig. 3　Basic framework of multicomponent analysis

4　实例分析

为了便于讨论，采用计算机模拟的方式产生三组分(A，B，C)光谱体系数据进行实例分析如图 4．对不同组分设定不同的峰高、峰宽和峰的位置，利用高斯函数得出三组分的单位光谱作为各组分的吸光度系数，进而运用均匀设计的方法设计 12 个不同浓度组成的量测样品，6个作为校正样本集如图5，6

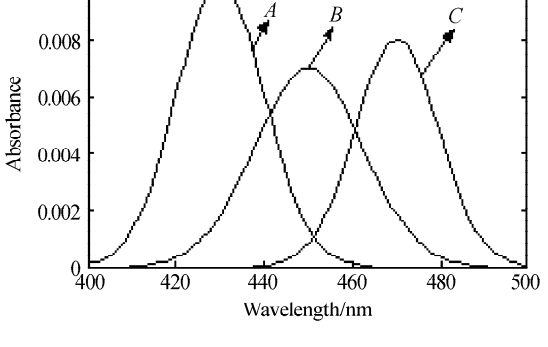

图 4　单组分光谱(A、B、C)

Fig. 4　Single component spectrum(A，B，C)

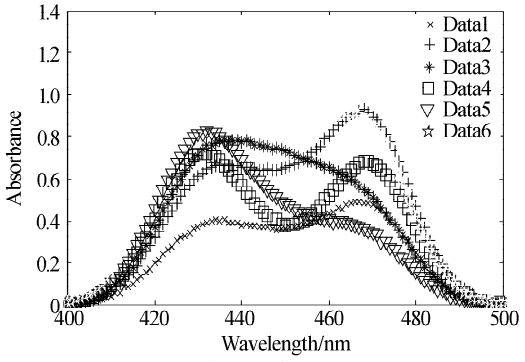

图 5 校正集样本光谱

Fig. 5 Spectrums of calibration samples

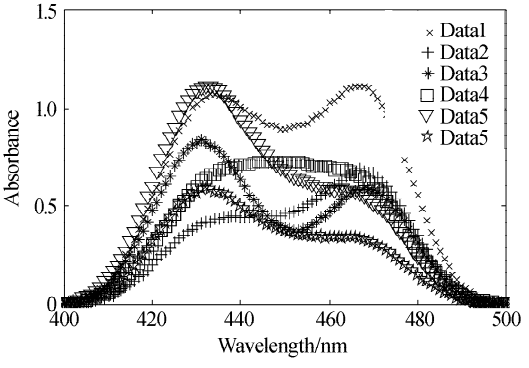

图 6 检验集样本光谱

Fig. 6 Spectrums of test samples samples

个作为检验样本集如图 6,浓度大小如表 1. 利用多组分加和性原理,加上高斯白噪声并乘以相应的干扰系数就得出了各个样品的吸光度曲线,以此作为模拟分析的原始数据.

表 1 样本浓度关系表

No		1	2	3	4	5	6
Calibration samples	A	30	40	50	60	70	80
	B	40	60	80	30	50	70
	C	50	80	40	70	30	60
No		1	2	3	4	5	6
Test samples	A	60	30	70	40	80	50
	B	70	50	30	80	60	40
	C	80	70	60	50	40	30

$$E1(f)=\sum_{i=1}^{6}\frac{100(\hat{c}_{f,i}-c_{f,i})}{6c_{f,i}} \tag{7}$$

$$E2=(E1(A)+E1(B)+E1(C))/3 \tag{8}$$

根据上述条件,分别建立了 3 个支持向量机,选取不同的核函数用校正样本集对其进行了学习训练. 经过比较分析发现,采用一阶多项式核函数,参数 $\varepsilon=1$、$C=\infty$时结果较为理想. 为了评估支持向量机方法(SVM)在多组分分析中的有效性,采用多元线性回归(MLR)、K 矩阵法和人工神经网络(ANN)等方法对上述数据分别进行了处理与分析,$\hat{c}_{f,i}$是对 $c_{f,i}$的估计,并以单组分平均误差百分比 $E1$ 和总的平均误差百分比 $E2$ 作为评价指标,对上述几种方法的计算结果进行了分类列表. 其中表 2 为对校正样本集的拟合误差表;表 3 为对检验样本集的预测误差表,采用的神经网络结构是经过了优化设计的.

表 2 对校正样本集的拟合误差

No		$E1$	$E2$
MLR	A	13.4031	13.6647
	B	14.0758	
	C	13.5151	
K	A	10.4263	10.9832
	B	11.1987	
	C	11.3246	
ANN	A	0	0
	B	0	
	C	0	
SVM	A	2.0298	1.9057
	B	1.6999	
	C	19875	

表 3 对检验正样本集的预测误差

No		$E1$	$E2$
MLR	A	13.5861	13.7232
	B	14.1433	
	C	13.4403	
K	A	10.5055	10.8062
	B	10.9554	
	C	10.9576	
ANN	A	7.0599	6.7622
	B	6.3596	
	C	6.8670	
SVM	A	3.8011	4.4532
	B	3.6019	
	C	5.9568	

从表 2 中的数据可以得出,在拟合性能方面,支持向量机方法优于多元线性回归和 K 矩阵方法,不如神经网络;从表 3 中的数据可以得出,在预测性能方面,支持向量机方法优于多元线性回归和 K 矩阵方法和神经网络. 由于神经网络在网络结构设计和初值选取上都需要许多经验和技巧,而支持向量机方法的复杂度与样本的维数无关,无须太多技巧,同时又具有较好的拟合预测效果,因而在光谱的多组分分析方面,支持向量机方法不失为一种较好的方法.

4 结论

选择合适的多组分分析模型是决定分析结果准确与否的关键所在. 本文探讨了支持向量机方法在光谱信号的多组分分析中的应用,在运用小波变换对光谱信号进行去除噪声处理的基础上,建立了基于支持向量机的多组分分析模型,最后采用计算机模拟的方式对该方法进行了实例分析. 结果表明,在

噪声干扰大、非线性明显、小样本条件下,该方法与多元线性回归、K 矩阵和人工神经网络等传统方法相比具有拟合性较好、预报误差小的优点.

参考文献

1 李凌均. 支持向量机在机械故障诊断中的应用研究. 计算机工程与应用,2002,**49**(19):19～21
Li L J. *Computer Engineering and Application*,2002,**49**(19):19～21

2 丁亚平,陈念贻. 导数光谱-支撑向量回归同时测定 NO_3^--NO_2^-. 计算机与应用化学,2002,**35**(11):752～754
DingY P,Chen N Y. *Computers and Applied Chemistry*,2002,**35**(11):752～754

3 胡昌华. 基于 MATLAB 的系统分析与设计—小波分析. 西安:西安电子科技大学出版社,1999. 1～19
Hu C H. System Analyse and design based on MATLAB—wavelet analysis. Xi′ an: Xi′ an Electron Science Technology University Publishing Company, 1999. 1～19

4 王景雷. 支持向量机在地下水位预报中的应用研究. 水利学报,2003,**30**(5):122～127
Wang J L. *Journal of Hydraulic Engineering*,2003,**30**(5):122～127

5 张学工. 关于统计学习理论与支持向量机. 自动化学报,2000,**42**(1): 32～42
Zhang X G. *Acta Automatica Sinica*,2000,**42**(1): 32～42

Spectral Multicomponent Analysis Based on Wavelet Transform and Support Vector Machine

Xiong Yuhong, Wen Zhiyu, Chen Gang, Huang Jian, Xu Yi

College of Optoelectronic Engineering, Chongqing University, Chongqing 400044

Received date: 2004-08-23

Abstract This paper took spectral signal according with Lambert-beer law as object, and introduced basic method of denoise with wavelet transform, and researched and established model of spectral multicomponent analysis based on support vector machine. Then computer simulation method gave an example to explain in the end, the example indicated that the method based on support vector machine can preferably solve the question of nonlinearity, small-sample in the spectral multicomponent analysis.

Keywords Spectral analysis; Wavelet transform; Support vector machine

Xiong Yuhong was born in 1971. Now he is candidate of the Ph. D. degree in optoelectronic engineering at ChongqingUniversity. He mainly studies signal process, model and simulation based on computer.