【文章编号】 1004-1540(2005)03-0182-03

最优小波包变换的化学模式特征选择方法

林 敏，毛谦敏，吕 进，刘辉军
（中国计量学院 计量技术工程学院，浙江 杭州 310018）

【摘 要】 提出了用近红外光谱作为化学模式的原始特征矢量，通过引入 Shannon 信息熵并根据最小熵准则寻求原始特征矢量小波包分解的最优小波树，使分解有最大的规律性，从而实现特征的选择. 结果表明，该方法既能有效地将原始的高维特征空间变为低维的特征空间，又能使样品间的差异性变大，这对化学模式识别和多元校正模型的建立，特别是分析方法灵敏度的提高都具有非常重要的意义.

【关键词】 小波包变换；化学模式 ；近红外光谱；Shannon 熵；Euclidian 距离
【中图分类号】 TP391.4 【文献标识码】 A

A characteristic selection method of chemical patterns based on the optimum wavelet-packet transformation

LIN Min, MAO Qian-min, LU Jin, LIU Hui-jun
(College of Metrological Technology and Engineering, China Jiliang University, HangZhou 310018, China)

Abstract: Based on the minimum entropy criteria and with using near infrared spectrum as the primitive characteristic vector of chemical patterns through the Shannon entropy, the optimal wavelet-tree is achieved by transforming the characteristic vector for maximum regularity so as to realize feature selection. The result shows that it can transform the high-dimension characteristic space into low-dimension characteristic space effectively This method is an important in the chemical patternrecognition and multivariate calibration, especially in improving analysis sensitivity.

Key words: wavelet-packet transform ation; chemical pattern; near infrared spectrum; Shannon entropy; Euclidian distance

在化学研究中，大量地存在着如何根据研究对象的某些可测数据来估计或判断对象的某种性质的问题[1]. 近红外光谱能反映化学样品的组成与结构信息，相同的或近似的样品有着相同或接近的光谱. 因此，对化学样品进行科学的抽象，用量测得到的近红外光谱作为表征它们特征的一组数据，可实现对化学样品的定性和定量分析[2].

随着现代分析仪器技术的发展，通过计算机

【收稿日期】 2005-05-20
【基金项目】 浙江省自然科学基金资助项目(No. 202081)；浙江省留学人员科技活动择优项目.
【作者简介】 林 敏(1962－)，男，浙江台州人，副教授. 主要研究方向为化学计量学方法.

数据采集系统可以获取大量的近红外光谱数据.但原始的近红外光谱存在着变量个数多,数据间相关性大以及不同样品间的相异性小,在多元回归分析及神经网络法中,过多的变量不仅计算量大,而且可能导致所得的数学模型不稳定,使预测结果较差,而当各变量间相关系数较高时,又会导致偶然相关,使预测模型不稳定.因此,样品的原始近红外光谱不适宜直接作为特征变量,必须进行特征的选择[3,4].

在分析化学领域,尤其是光谱分析中,人们关心的主要是图谱中各谱峰的位置、形状和大小特征,即需要了解化学信号的局部特征.近年来小波变换在信号处理和特征提取中得到了广泛应用,小波变换是一种多尺度分解的时频局域变换,可分析包含不同尺度的信号,但用于特征提取,固定的时频分解形式并不是最优的.小波包变换对信号具有任意的多尺度分解形式,小波包库包含了丰富的小波包基,不同的小波包基具有不同的性质,反映不同的信号特性,能提供其它变换所不能提供的信号的重要特征[5,6].本文通过引入 Shannon 信息熵并根据最小熵准则寻求原始光谱特征矢量小波包分解的最优小波树,从而实现特征的选择.该方法既能有效地将原始的高维特征空间变为低维的特征空间,又能使样品间的差异性变大,这对化学模式识别和多元校正模型的建立,特别是分析方法灵敏度的提高都具有非常重要的意义.

1　原理与方法

1.1　化学模式空间

设一个化学样品的 n 个特征量测值分别为 $x_1,x_2,\cdots,x_n$,由于它们来自同一个对象,所以将它们作为一个整体构成一个 n 维特征矢量 X,即 $X=(x_1,x_2,\cdots,x_n)'$.X 是化学样品的一种数学抽象,用其来代表一化学样品,即为该化学样品的模式.当采用近红外光谱分析技术时,化学样品在 n 个波长下的吸收或透过数据就构成 n 维的特征矢量,各种不同取值的 X 的全体就构成了 n 维化学模式空间.一般来说,高维模式空间提供了更多的信息,有可能解决一些低维空间中难以解决的问题.

1.2　小波包变换与特征选择

小波包由 Coifman,Meyer,Quaker,Wickerhauser 提出,小波包分解对近似系数和细节系数都进行分解(图 1).

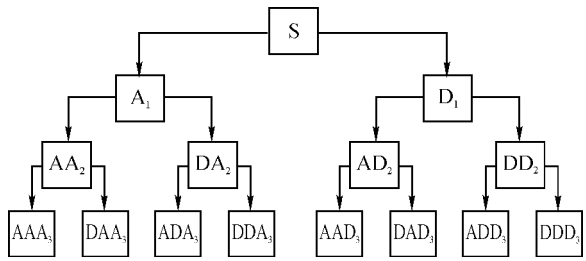

图 1　小波包分解结构

如果 $\{h_k\}_{k\in z}$ 和 $\{g_k\}_{k\in z}$ 是一组共轭镜象滤波器(QMF)即:

$$\sum_{n\in Z}h_{n-2k}h_{n-2l}=\delta_{kl},\ \sum_{n\in Z}h_n=\sqrt{2},\tag{1}$$

$$g_k=(-1)^k h_{l-k}\quad k\in Z.\tag{2}$$

则可定义一系列函数 $\{u_n(t)\}(n=0,1,2,\cdots)$ 满足如下方程:

$$u_{2n}(t)=\sqrt{2}\sum_{k\in Z}h_k u_n(2t-k),\tag{3}$$

$$u_{2n+1}(t)=\sqrt{2}\sum_{k\in Z}g_k u_n(2t-k).\tag{4}$$

当 $n=0$ 时,$u_0(t)$ 即为尺度函数 $\phi(t)$,$u_1(t)$ 即为小波函数 $\Psi(t)$.每一形如 $2^{-j/2}u_n(2^{-j}t-k)$,$j,k\in Z,n\in Z_+$ 的函数称作一个小波包函数,其集合 $\{2^{-j/2}u_n(2^{-j}t-k),j,k\in Z,n\in Z_+\}$ 称为一个小波包库,其中 j 是尺度参数,k 是平移参数,而 n 是频率参数.从小波包库中选择能构成 $L^2(R)$ 空间的一个基函数系称为 $L^2(R)$ 的一个小波包基.对任一固定的 j 值,$\{2^{-j/2}u_n(2^{-j}t-k),k\in Z,n\in Z_+\}$ 均可构成 $L^2(R)$ 的一个正交基.引入小波包变换是为了让信息能量集中,为此需要有一定的衡量准则.本文采用熵最小原则.在信息论中,熵是用来度量信息规律性的概念,熵越小,则信息的规律性越强,所以一般的判别方法是根据系数分解后的系数的熵之和是否大于原系数的熵.假设 s 为信号(原信号或各层小波变换系数),s_i 为该信号在某组正交基上的第 i 项系数,用 E 表示熵,则 E 是一个由每个正交基上的系数的某种变换叠加起来的值,即 $E=\sum_i E(s_i)$.本文采用

Shannon 熵，其定义为：

$$E(s) = -\sum_i s_i^2 \log(s_i^2). \quad (5)$$

根据可分性准则，从小波包库中选择一个对分类最优的小波包基，从与该小波包基对应的小波包系数中，选择一组具有最大可分性的系数作为化学模式的特征矢量.

2 结 果

近红外光谱分析技术是对化合物进行定性和定量分析的重要手段，而一张红外光谱需要相当多的数据才能准确地反映所测样品的结构，谱图简化，不仅可以节省储存空间、加快检索速度，而且能有效地提取特征.

本文选用 80 种玉米样品，在近红外光谱仪上进行光谱扫描，波长范围是 1 100－2 498 nm，间隔为 2 nm，每个样品具有 700 个数据(图 2). 如果直接将原始光谱作为特征矢量，不仅维数高达 700 维，而且数据间存在着极强的相关性，不同样品间的相异性较小，采用 Euclidian 距离来度量样品间的差异，图 3 所示是序号为 1 的样品与其余 79 个样品间的 Euclidian 距离.

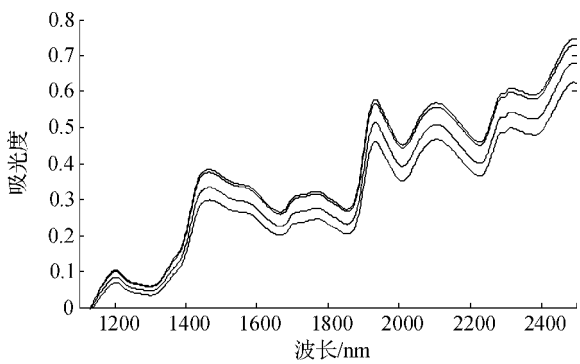

图 2 玉米近红外光谱

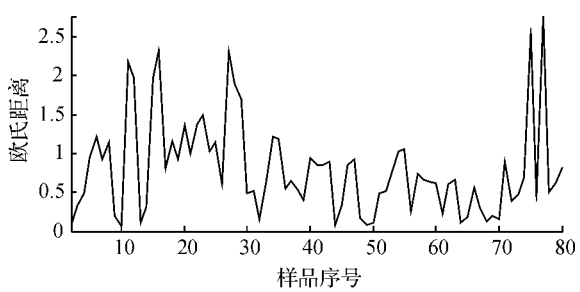

图 3 原始光谱样品间的欧氏距离

选择正交小波系 Symlets 中的 Sym6 对玉米近红外光谱进行分析. 用 Sym6 对玉米近红外光谱数据进行 5 层小波包分解. 引入 Shannon 熵并根据最小熵准则，从根节点开始，当子节点 N_1 和 N_2 的熵之和小于母节点 N 的熵时，一个节点 N 可以被分成两个子节点 N_1 和 N_2；如果反之，N_1 和 N_2 的熵之和大于 N 的熵，则保留节点 N 不作分解，寻找最佳的小波包分解方式即最佳分解树结构，如图 4 是最佳树结构图，左图各分解节点上标注的是熵值，右图是各节点上的能量分布. 可见，玉米的近红外光谱经小波包分解后，各节点上的熵值相差极大，能量主要集中在个别节点上，最佳小波树使分解有规律性，能将近红外光谱的内在规律显现出来. 选择熵值最小、能量最大的节点上的小波系数作为化学模式的特征矢量，其数据个数仅为 32，而所占有的能量却高达 99.94%，相应的样品间的 Euclidian 距离也变大(图 5).

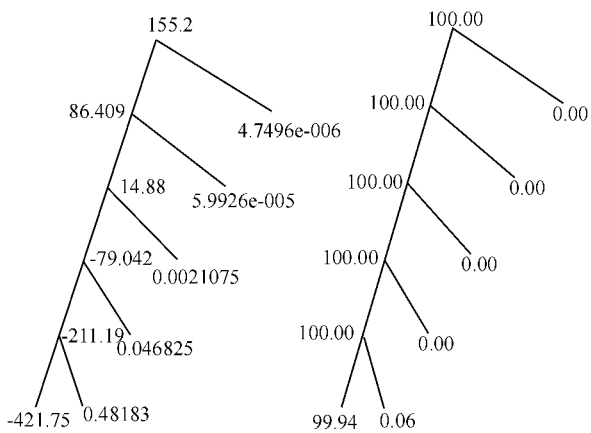

图 4 最佳树结构 左:熵值;右:能量

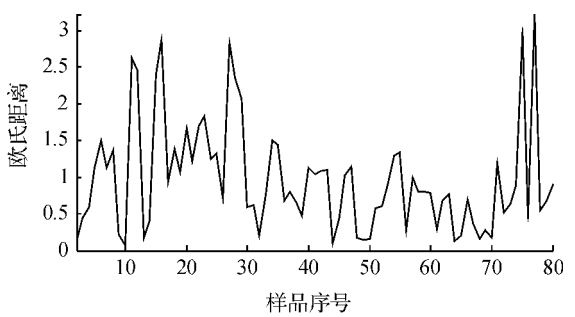

图 5 特征选择后的样品间的欧氏距离

(下转第 187 页)

表 1　检定结果

压力(MPa)	0	0.01	0.02	0.03	0.04	0.05
电压(mV)	−2.682 4	−1.021 3	0.657 8	2.349 4	4.051 0	5.760 1
压力(MPa)	0.06	0.07	0.08	0.09	0.10	
电压(mV)	7.482 7	9.208 3	10.944 9	12.708 4	14.464 2	

全部用 MATLAB 编程计算，可大大减少代数法的计算工作量.

按本文第二节线性方程的最小二乘法计算，得到输入输出关系为：

$y = -2.764\ 4 + 17.151\ 1p$.

标准差为：$\sigma_y = 0.051\ 3$(mV).

按本文第三节二次方程的最小二乘法计算，得到输入输出关系为：

$y = -2.685\ 8 + 16.627\ 1p + 0.524\ 1p^2$，

标准差为：$\sigma'_y = 0.003\ 8$(mV).

比较 σ'_y 与 σ_y，可见 $\sigma_y' \ll \sigma_y$，精度得到很大提高. 而且，利用 MATLAB 程序，计算机可以迅速显示输入压力输出电压的二次曲线图形以及对应的数值. 从这一实例可以明显看出其拟合精度可显著提高.

4　讨　论

以往在处理压力传感器的输入输出关系时，都是用代数法进行的. 该法的计算量大，而且输入输出关系是用直线表示. 随着计算机软件的发展，MATLAB 成为数值计算和矩阵运算中最方便的工具，根据检定时得到的实验数据，计算二次多项式关系，计算过程简单明确，计算结果准确可靠，使用十分方便.

另外，从理论上分析，输入输出关系可进行更高次多项式拟合，笔者对 3 中的例子也作了计算，从直线关系到二次多项式关系精度提高最明显，从二次多项式关系到三次多项式关系，精度提高很小.

【参　考　文　献】

[1] 孙以才，刘玉岭，孟庆皓. 压力传感器的设计制造与应用[M]. 北京：冶金工业出版社，2000.

[2] 李国玉，孙以才，潘国峰，等. 非线性函数规范化多项式拟合精度分析[J]. 传感器世界，2004(3)：30－34.

[3] JJG860－94. 压力传感器(静态)检定规程[S].

[4] 费业泰. 误差理论与数据处理[M]. 北京：机械工业出版社，2004.

[5] 袁慰平，张令敏. 计算方法与实习[M]. 南京：东南大学出版社，1994.

[6] 张　平，吴云洁，王　东，等. MATLAB 基础与应用简明教程[M]. 北京：航空航天大学出版社，2001.

(上接第 184 页)

4　结　论

本文提出的基于最优小波包变换的化学模式特征选择方法，既能有效地将原始的高维特征空间变为低维的特征空间，又能使样品间的差异性变大，这对化学模式识别和多元校正模型的建立，特别是分析方法灵敏度的提高都具有非常重要的意义.

【参　考　文　献】

[1] 汪尔康. 21 世纪的分析化学[M]. 北京：科学出版社，1999.

[2] 陆婉珍，袁洪福，徐广通. 现代近红外光谱分析技术[M]. 北京：中国石化出版社，2000 .

[3] 边肇祺，张学工. 模式识别[M]. 北京：清华大学出版社，2000.

[4] 许　禄. 化学计量学[M]. 北京：科学出版社，2004.

[5] MALLAT S G. A Theory for multiresolution signal decomposition：the wavelet representation. IEEE[J]. Transaction on Pattern Analysis and Machine Intelligence，1989，11(7)：674－693.

[6] H LIANG，I HARTIMO. A feature extraction algorithm based on wavelet packet decompo，sition for heart sound signals[M]. Proc of IEEE－SP intre Symp USA：IEEE，1998：93－96.