第26卷, 第4期
2006年4月

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

Vol. 26, No. 4, pp617-619
April, 2006

近红外光谱用于植物样品中水溶性氯离子含量的测定

吴荣晖, 邵学广*

中国科学技术大学化学系, 安徽 合肥 230026

摘 要 基于离散小波变换(DWT)和最小二乘支持向量回归(LSSVR)方法, 建立了近红外光谱测定植物样品中水溶性氯离子的回归校正模型。以烟草样品中水溶性氯离子含量的测定为研究对象, 首先采用DWT对近红外光谱进行数据压缩和背景扣除, 再用LSSVR建立氯离子的校正模型。结果表明, 与偏最小二乘回归(PLSR)和传统的LSSVR方法相比, 作者所建模型的预测准确性具有一定优势。

主题词 离散小波变换; 最小二乘支持向量回归; 近红外光谱; 水溶性氯离子
中图分类号: O657.3 文献标识码: A 文章编号: 1000-0593(2006)04-0617-03

引 言

支持向量机(SVM)是近年发展起来的一种基于统计学习理论的新型学习机器, 已在模式识别、时序分析以及多元校正等领域得到初步应用[1-4]。与传统的主成分回归(PCR)、偏最小二乘回归(PLSR)等线性建模方法相比, 支持向量回归(SVR)方法由于能够解决非线性问题而具有显著优势。虽然人工神经网络(ANN)也常被用于建立非线性模型, SVR方法更能控制过拟合现象, 并具有全局最优及更好的泛化能力等优点[5]。最小二乘支持向量回归(LSSVR)是在SVR基础上发展起来的一种改进技术, 在保持SVR优点的同时大大提高了计算速度[6-8]。

小波变换(WT)在分析化学信号处理中已得到广泛应用, 可有效地进行分析化学信号的平滑滤噪、数据压缩及背景扣除[9-11]。本文基于离散小波变换(DWT)和LSSVR建立了一种近红外光谱(NIR)数据的建模方法, 先利用小波变换扣除光谱中的噪声及背景干扰, 再用LSSVR进行建模, 并应用于植物样品中水溶性氯离子含量的预测。结果表明, 该方法的预测准确性优于PLSR和传统的LSSVR。

1 原理与算法

小波变换的原理、算法及应用已有很多文献报道[12-14]。小波变换的实质是将信号分解成不同频率部分, 因此可有效的实现分析化学信号的平滑滤噪、数据压缩及背景扣除, 提高分析结果的准确性。

支持向量回归(SVR)的基本思想是: 采用非线性变换$\Phi(\cdot)$将给定的样本(x_i, y_i), $i = 1, \cdots, n$, $x_i \in R^d$, $y_i \in R$从原空间映射到高维特征空间, 并在此空间构造最优线性回归函数

$$\hat{y} = \omega \cdot \Phi(x) + b \qquad (1)$$

根据统计学习理论的结构风险最小化原则寻找ω和b。它等价于最小化$\frac{1}{2}\|\omega\|^2 + cR_{emp}$, 其中$\|\omega\|^2$控制模型的复杂度, c为正则化参数, 控制对超出误差范围样本的惩罚程度, R_{emp}为误差控制函数。标准的SVR采用误差ξ作为优化目标的损失函数, 并通过对偶转换和拉格朗日乘子法将问题进一步转化为一个二次规划问题, 得到最终的回归模型

$$\hat{y} = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b \qquad (2)$$

其中α为拉格朗日乘子, $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$为核函数。

LSSVR采用二次误差ξ^2作为优化目标的损失函数, 使优化问题转化为线性方程组的求解问题, 即

$$\begin{bmatrix} 0 & I_n^{\mathrm{T}} \\ I_n & K + C^{-1}I \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \qquad (3)$$

式中a为拉格朗日乘子, I为$[n \times n]$的单位矩阵, I_n为$[n \times 1]$的单位向量, C为常数, K为$K(x_i, x_j)$的矩阵。因此, LSSVR避免了SVR的复杂计算, 具有比SVR更快的计算速度。

2 实验部分

实验使用Bruker Verctor 22/N傅里叶近红外光谱仪对

收稿日期: 2005-03-08, 修订日期: 2005-06-18
基金项目: 国家自然科学基金(20325517)资助项目
作者简介: 吴荣晖, 女, 1975年生, 中国科学技术大学化学系在读硕士研究生 * 通讯联系人

309 个烟叶样品进行测定，以 4 cm^{-1} 间隔记录 4 000~ 9 200 cm^{-1} 之间的光谱测量。烟叶样品中水溶性氯离子按照标准方法(YC/T162) 使用 Bran+ Lubbe 自动分析仪(AutoAnalyzer III) 测定。将 309 个样品的数据随机分为两部分，206 个样品的光谱数据用于校正集，103 个样品的光谱数据用作预测集。

3 结果与讨论

3.1 离散小波变换对光谱的数据压缩和背景扣除

小波变换是一种有效的数据压缩和背景扣除工具[9-11]。在本工作中，首先采用小波变换对 NIR 信号进行分解，按照数据压缩的原理保留小波系数中较大的系数，实现 NIR 的数据压缩。然后再将小波系数中的最低频系数扣除，达到扣除背景干扰的目的。分解尺度和小波基分别以预测值均方误差(RMSEP) 为评价标准进行优化。本文采用了 sym8 小波，分解尺度为 9。

图 1 中(a)，(b) 分别为校正集 NIR 光谱的测量光谱和通过保留 60 个高频系数重构所得到的光谱。可以看出数据压缩和背景扣除后，NIR 光谱的有用信息几乎没有丢失，而背景得到了很好的扣除。

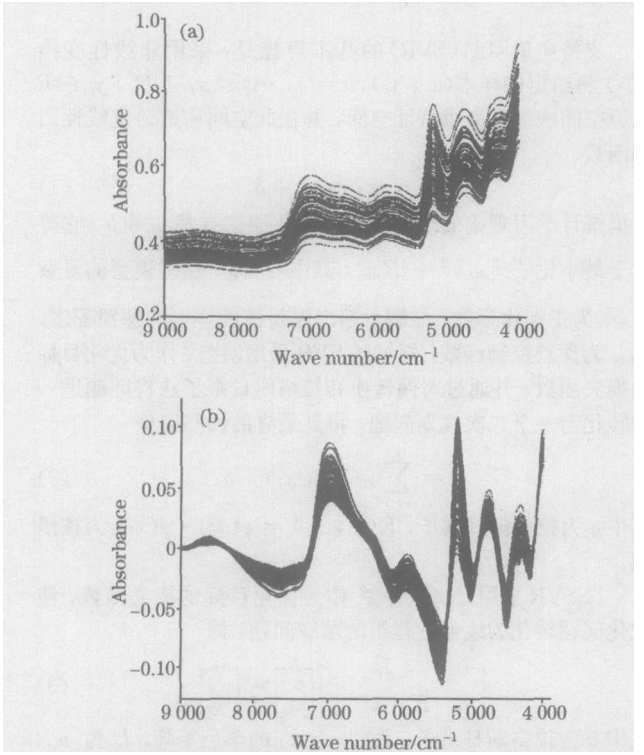

Fig 1 Comparison of the measured (a) and reconstructed (b) spectra of the calibration data set

3.2 预测结果

图 2 显示了氯离子的测量浓度与 DWT-LSSVR 预测浓度之间的关系，a，b 和 r 分别为由最小二乘回归所得到的截距、斜率和相关系数。由图 2 可知，回归直线的斜率为 0.97，与 1 非常接近，相关系数为 0.99。表明预测结果准确且具有良好的线性关系。因此，本方法可以准确的预测出复杂植物样品中氯离子的含量。

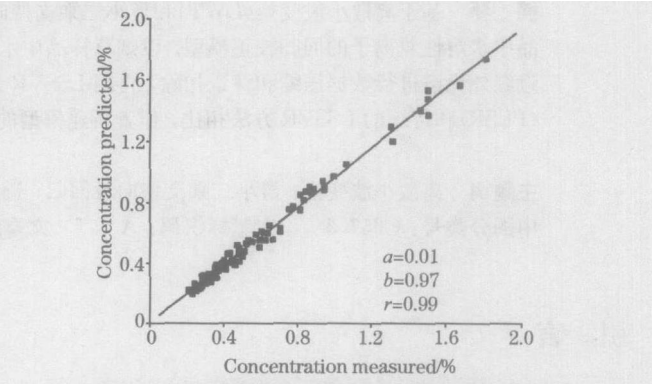

Fig 2 The relationship between the measured and predicted concentrations

同时对 DWT-LSSVR，LSSVR 和 PLSR 三种方法进行了比较，计算结果见表 1。其中，DWT-LSSVR，LSSVR 方法选择了线性核函数，参数 C 分别设置为 200 和 100，而 PLSR 的因子数则根据以往的经验将其设定为 10。

Table 1 Comparison of the prediction results by PLSR, LSSVR and DWT-LSSVR

方法	预测值的均方根误差	相关系数
RPLSR	0.114	0.924
LSSVR	0.069	0.973
DWT-LSSVR	0.059	0.982

可以看出，采用 LSSVR 算法所得到的预测结果明显优于 PLSR 算法，这可能是由于 LSSVR 算法引入核函数能更好的解决了 NIR 光谱中存在的非线性问题。而 DWT-LSSVR 算法，由于引入小波变换，消除了 NIR 中的冗余信息和严重的背景干扰，从而使得预测结果得到进一步提高。

参 考 文 献

[1] Vapnik V. The Nature of Statistical Learning Theory. New York: Springer Verlag, 1998.

[2] LIANG Lu-hong, AI Hai-zhou, XIAO Xi-pan, et al(梁路宏, 艾海舟, 肖习攀, 等). Chinese Journal of Computer(计算机学报), 2002, 25(1): 22.

[3] Thissen U, van Brakel R, de Weijer A P, et al. Chemom. Intell. Lab. Syst., 2003, 69(1-2): 35.

[4] QU Hai-bin, LIU Xiao-xuan, CHENG Yi-yu(瞿海斌, 刘晓宣, 程翼宇). Chemical Journal of Chinese Universities(高等学校化学学报), 2004, 25(1): 39.

[5] CHEN Nian-yi, LU Wen-cong, et al(陈念贻, 陆文聪, 等). Computers and Applied Chemistry(计算机与应用化学), 2004, 21(6): 886.

[6] Suykens J A K, De Brabanter J, Lukas L, et al. Neurocomputing, 2002, 48: 85.

[7] Thissen U, Ustun B, Melssen W J, et al. Anal. Chem., 2004, 76(11): 3099.

[8] Chauchard F, Cogdill R, Roussel S, et al. Chemom. Intell. Lab. Syst., 2004, 71(2): 141.

[9] SHAO Xue-guang, GU Hua, CAI Wen-sheng, PAN Zhong-xiao(邵学广, 顾 华, 蔡文生, 潘忠孝). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 1999, 19(2): 139.

[10] TIAN Gao-you, YUAN Hong-fu, LIU Hui-ying, LU Wan-zhen(田高友, 袁洪福, 刘慧颖, 陆婉珍). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2003, 23(6): 1111.

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

[12] MA Xiao-guo, ZHANG Zhan-xia(马晓国, 张展霞). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2000, 20(4): 507.

[13] Shao X G, Leung A K M, Chau F T. Accounts Chem. Res., 2003, 36(4): 276.

[14] Chau F T, Liang Y Z, Gao J B, et al. Chemometrics: From Basic to Wavelet Transform. New Jersey: A John Wiley & Sons, Hoboken, 2004.

Application of Near-Infrared Spectra in the Determination of Water Soluble Chloride Ion in Plant Samples

WU Rong-hui, SHAO Xue-guang*

Department of Chemistry, University of Science and Technology of China, Hefei 230026, China

Abstract A new method was proposed to extract relevant information from near-infrared(NIR) spectra for multivariate calibration of water soluble chloride ion in complex plant samples. The method is a combination of discrete wavelet transform (DWT) and least squares support vector regression (LSSVR). After data compression and background removal in NIR spectra by DWT, the LSSVR approach was used to build NIR spectra regression models on the retained wavelet coefficients. Compared with partial least square regression (PLSR) and LSSVR, the proposed method is superior both in calculation speed and prediction accuracy.

Keywords Discrete wavelet transform; Least squares support vector regression; Near-infrared spectra; Water soluble chloride ion

(Received Mar. 8, 2005; accepted Jun. 18, 2005)

* Corresponding author