

文章编号: 0253-2239(2007)11-2054-5

# 黄酒糖度预测的可见-近红外光谱方法研究\*

刘 飞 何 勇 王 莉

(浙江大学生物系统工程与食品科学学院, 杭州 310029)

**摘要:** 提出了用可见-近红外光谱结合不同化学计量学方法预测黄酒糖度的新方法。用 240 个黄酒样本建模, 60 个样本进行预测。通过对光谱数据进行平滑、变量标准化及一阶导数等预处理, 建立并比较了偏最小二乘法, 小波变换与偏最小二乘法相结合, 主成分分析与人工神经网络相结合以及主成分分析与最小二乘支持向量机相结合四种不同建模方法的预测精度, 以相关系数  $r$ 、预测标准差、偏差等为评判标准, 得到黄酒糖度预测的最优模型为最小二乘支持向量机模型。该模型对黄酒糖度预测的相关系数为 0.962、预测标准差为 0.021、偏差为 -0.001, 获得了理想的预测精度。结果表明应用可见-近红外光谱对黄酒糖度进行预测是可行的, 且最小二乘支持向量机模型能得到最优的预测结果。

**关键词:** 医用光学与生物技术; 可见-近红外光谱; 黄酒糖度预测; 最小二乘支持向量机; 人工神经网络; 偏最小二乘法

中图分类号: O657.33; TS262.4 文献标识码: A

## Methods for the Prediction of Sugar Content of Rice Wine Using Visible-Near Infrared Spectroscopy

Liu Fei He Yong Wang Li

(College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310029)

**Abstract:** A new approach of using visible-near infrared spectroscopy combined with different chemometric methods was investigated for the prediction of sugar content of rice wine. 240 wine samples were used for the calibration set, while 60 for the validation set. After some pretreatments of smoothing, standard normal variate (SNV) and first derivative to the spectral data, four different models were developed and the precision were compared, including partial least squares (PLS), combination of wavelet transform and PLS (WT-PLS), combination of principal component analysis and artificial neural network (PCA-ANN), and combination of PCA and least squares-support vector machine (PCA-LS-SVM). According to the standards of correlation coefficient ( $r$ ), the root mean square error of prediction (RMSEP) and bias, a best calibration model of PCA-LS-SVM was achieved for the prediction of sugar content of rice wine. The correlation coefficient  $r=0.962$ , RMSEP=0.021 and Bias=-0.001 by PCA-LS-SVM, and a satisfying prediction precision was achieved. The results indicated that visible-near infrared spectroscopy could be successfully applied for the prediction of sugar content of rice wine and PCA-LS-SVM model could achieve a best prediction results.

**Key words:** medical optics and biotechnology; visible-near infrared spectroscopy; prediction of sugar content of rice wine; least squares-support vector machine; artificial neural network; partial least squares analysis

## 1 引 言

黄酒以其富含的氨基酸、维生素、矿物质等营养物质和独特的滋味著称于世<sup>[1]</sup>, 同时还以其营养保

健功能而闻名<sup>[2]</sup>。黄酒种类繁多, 而其含糖量可以作为黄酒分类的一个标准, 同时含糖量在黄酒发酵过程中也是一个重要的指标, 发酵过程中含糖量的

\* 国家自然科学基金(30671213)、国家十一五科技支撑项目(2006BAD10A04)和教育部高等学校优秀青年教师教学科研奖励计划(02411)资助课题。

作者简介: 刘 飞(1983-), 男, 河北定州人, 博士研究生, 主要从事数字农业和农产品多光谱检测技术等方面的研究。  
E-mail: fliu@zju.edu.cn

导师简介: 何 勇(1963-), 男, 浙江慈溪人, 教授, 博士生导师, 主要从事数字农业和多光谱检测技术等方面的研究。  
E-mail: yhe@zju.edu.cn

收稿日期: 2007-04-09; 收到修改稿日期: 2007-05-08

变化涉及到糖化的速度和发酵的速度,两者之间的平衡关系到黄酒的质量。因此黄酒糖度的快速检测是黄酒品质理化指标分析中的一个重要方面<sup>[3]</sup>。

可见-近红外光谱分析技术,可以充分利用全波段或多波长下的光谱数据进行定性或定量分析,是一种快速、无损、低成本、无污染的分析技术,已广泛应用于农业、石油化工、食品、制药等领域<sup>[4]</sup>,例如在苹果糖度预测<sup>[5~7]</sup>、番茄硬度、可溶性固形物和 pH 值预测<sup>[8]</sup>、杨梅汁、橘子汁可溶性固形物和酸度测定<sup>[9,10]</sup>、黑木耳产地鉴别<sup>[11]</sup>、胆结石中蛋白质的傅里叶变换中红外光谱特性<sup>[12]</sup>和茶饮料可溶性固形物检测<sup>[13]</sup>等方面的应用。在酿酒行业中,例如利用红外光谱技术进行假酒中甲醇的定性分析<sup>[14]</sup>,对葡萄酒产地鉴别<sup>[15]</sup>,利用近红外透射光谱法结合气相色谱建立茅台酒的指纹模型<sup>[16]</sup>,黄酒中品质指标的傅里叶变换近红外光谱检测<sup>[17]</sup>等。但是多为偏最小二乘法或多元线性回归等,建模方法单一,且未能充分考虑光谱数据中存在的非线性信息对预测结果的影响,模型的预测精度和稳健性受到一定限制。

本文应用可见-近红外光谱对黄酒糖度进行预测,并比较不同建模方法的预测精度。比较了偏最小二乘法(Partial least squares, PLS)、小波变换(Wavelet transform, WT)与偏最小二乘法相结合(WT-PLS),主成分分析(Principal component analysis, PCA)与人工神经网络(Artificial neural network, ANN)相结合(PCA-ANN),以及主成分分析与最小二乘支持向量机(Least squares-support vector machine, LS-SVM)相结合(PCA-LS-SVM)四种不同建模方法对预测精度的影响,从而得到应用可见-近红外光谱预测黄酒糖度的最优模型。

## 2 材料与方法

使用美国 ASD (Analytical Spectral Device) 公司的 Handheld FieldSpec 光谱仪,其光谱测定范围 325~1075 nm,采样间隔为 1.5 nm,光谱扫描次数设定为 30 次,每个样本保存 3 条光谱曲线,光谱仪探头视场角为 20°。光源采用 14.5 V 卤素灯。分析软件采用 ASD View Spec Pro, Unscrambler V9.6, Matlab 7.0, 以及 DPS (Data Procession System for Practical Statistics)。

黄酒糖度的测量值通过数字阿贝折射仪 (Model: WAY-2S) 测得,仪器本身带有温度校正,测量单位为布里斯克糖度(°Brix),测量准确度以折射率表示,  $n_D = \pm 0.0002$ 。

购得 6 种不同的绍兴加饭黄酒,包括古越龙山、塔牌、黄中皇、会稽山一年陈、三年陈和五年陈。所有样本都密封保存在恒温  $25 \pm 1$  °C 的实验室内。在样品采集前,先将酒摇匀,每种黄酒制备了 50 个样本。光谱扫描采用透射法,光程为 2 mm,每个样本扫描 30 次,保存 3 条光谱,以其平均光谱作为最终的透射光谱。从全部的 300 个黄酒样本中,每个品种随机选择 40 个共 240 个样本作为定标集,其余 60 个作为预测集。每个样本在光谱扫描完成后,立即进行糖度的测定。

将光谱数据经  $\lg(1/T)$  转换成吸光度值,以 ASCII 码输出。为去除高频随机噪音、基线漂移、光散射、样本不均匀等影响,需对原始光谱数据进行预处理<sup>[18]</sup>。采用 5 点 Savitzky-Golay 平滑法,变量标准化(Standard Normal Variate, SNV)处理和一阶 Savitzky-Golay 导数处理,求导间隔为 5。为消除光谱数据在首端与末端产生的部分噪音,截取 350~1050 nm 波段的数据进行分析。

应用小波变换在 Matlab 中将得到的样本与光谱吸光度值的矩阵进行压缩和消噪,通过计算比较,选择合适的小波基函数和分解尺度,将小波变换后的数据重新进行偏最小二乘法(PLS)回归分析,建立 WT-PLS 模型。

应用主成分分析对光谱数据进行降维。根据提取的主成分的累积贡献率,将主成分作为反向传播神经网络的输入以减少神经网络模型的计算量,采用 Sigmoid 传递函数,通过调节网络层的参量,减少模型的拟合残差,从而建立 PCA-ANN 模型。

LS-SVM 通过非线性映射函数建立回归模型,将输入变量映射到高维特征空间。然后将优化问题改成等式约束条件。利用拉格朗日乘子求解最优化问题,对各个变量求偏微分。根据 Mercer 条件,存在映射函数  $\varphi(\cdot)$  和核函数  $K(x_i, x_j)$  使得

$$\varphi(x_k)^T \varphi(x_l) = K(x_k, x_l), \quad k, l = 1, \dots, N \quad (1)$$

采用径向基函数(RBF)作为核函数,其表达式为

$$K(x_k, x_l) = \exp[-\|x_k - x_l\| / (2\sigma^2)], \quad (2)$$

从而得到 LS-SVM 的函数估计为

$$y(x) = \sum_{k=1}^N \alpha_k K(x, x_k) + b. \quad (3)$$

## 3 结果与讨论

黄酒样本的可见-近红外光谱原始图谱如图 1 所示。每种黄酒随意选取 10 条光谱。图中横坐标为光谱波长,范围为 350~1050 nm,纵坐标为吸光

度。从图 1 中可以看出,三种黄酒的原始图谱的曲线趋势基本相同,并没有太明显的差异。黄酒糖度的测量值如表 1 所示。虽然不同种类的黄酒其糖度值有所不同,但为了便于整体的预测分析,将其作为整体来进行建模分析。

表 1 300 个黄酒样本糖度值

Table 1 Sugar content for 300 rice wine samples

Parameters (Sample No. $n=300$ )	Sugar content / ( $^{\circ}$ Brix)
Mean	16.8
Maximum	17.7
Minimum	15.4
Standard deviation	0.502

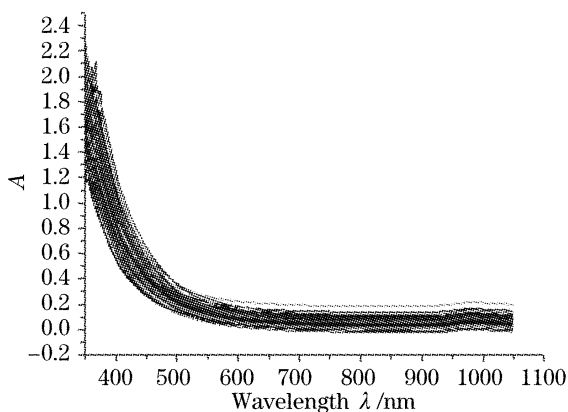


图 1 黄酒样本的可见-近红外光谱原始图谱

Fig. 1 Original visible-near infrared spectra of rice wine samples

将 300 个样本的原始光谱进行预处理后,用随机选择的 240 个样本建立黄酒糖度预测的 PLS 模型,经交互验证判断,确定建立 PLS 模型的最佳主成分数为 12。对剩余的 60 个样本进行预测,预测结果如表 2 所示。用相关系数  $r$ , 预测标准差和偏差等来衡量模型预测的精确性。该 PLS 模型预测结果表明,预测的相关系数  $r=0.938$ , 预测标准差为 0.183, 偏差为 0.051。

利用小波变换对光谱数据进行消噪和压缩。考虑不同的正交小波基函数 (Daubechies) (Db1, Db2, Db3, Db4, Db5) 与分解尺度 (4, 5, 6, 7) 对数据压缩

的影响,以验证集的预测平方和 (PRESS) 为评价标准,选择最优的小波基和分解尺度。PRESS 计算公式如下:

$$P = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (4)$$

式中  $n$  为样本个数;  $y_i$  为第  $i$  个样本实测值;  $\hat{y}_i$  为第  $i$  个样本预测值。

通过 Matlab 编程,实现小波变换对光谱数据的压缩和消噪。通过计算得到最优的  $P$  对应的小波基函数为 Db5, 分解尺度为 5, 将 701 个光谱数据点压缩成 30 个, 然后重新建立 PLS 模型, 即 WT-PLS 模型, 同样采用交互验证来检验, 最佳主成分数为 8。对 60 个未知样本进行预测, 结果如表 2 所示。WT-PLS 预测结果表明, 预测的相关系数  $r=0.942$ ,  $\Delta_{\text{RMSEP}}=0.173$ ,  $\Delta_{\text{Bias}}=0.037$ 。预测精度优于 PLS 模型。

应用主成分分析对光谱数据进行降维, 从原始光谱数据中提取出新的特征变量, 这些特征变量之间是相互独立的, 并且尽可能多的保留与因变量的相关性, 又保留较多的方差, 消除了原有自变量共线性问题。光谱数据可能存在非线性的部分, 采用非线性预测的前馈型反向传播神经网络模型来建立预测模型<sup>[7]</sup>。在主成分分析中, 前 7 个主成分的累积贡献率已达 98.970% (如表 3 所示), 已经能够解释样本的绝大部分重要信息。将前 7 个主成分作为反向传播神经网络的输入值。调节网络参量的设定值, 经过反复实验得到的最优网络结构为: 输入层为 7 个输入单元, 第一隐含层为 5 个隐含单元, 第二隐含层为 3 个隐含单元, 以及 1 个输出层单元的四层反向传播神经网络模型, 其传递函数为 Sigmoid 函数, 允许误差值设定为 0.0001, 学习速度为 0.2, 动态参量为 0.7, 训练次数为 2000, 数据经过标准变换。该模型的预测结果如表 2 所示, 预测的相关系数  $r=0.950$ ,  $\Delta_{\text{RMSEP}}=0.158$ ,  $\Delta_{\text{Bias}}=0.011$ , 预测精度优于 PLS 模型和 WT-PLS 模型。

表 2 不同建模方法的预测结果

Table 2 Prediction results by different calibration models

Parameters	Sample No.	$r$	$\Delta_{\text{RMSEP}}$	$\Delta_{\text{Bias}}$	Slope	Offset
PLS	60	0.938	0.183	0.051	0.896	1.801
WT-PLS	60	0.942	0.173	0.038	0.880	2.044
PCA-ANN	60	0.950	0.158	0.011	0.884	1.960
PCA-LS-SVM	60	0.962	0.021	-0.001	0.842	2.655

表 3 前 7 个主成分及其累积贡献率

Table 3 The first 7 principal components and accumulative reliabilities

Principal components	1	2	3	4	5	6	7
Accumulative reliabilities / %	59.591	88.163	94.033	96.430	97.592	98.369	98.970

采用径向基函数(RBF)作为核函数,该模型主要有两个待定参量:超参量  $\gamma$  和径向基函数核函数参量  $\sigma^2$ ,这两个参量在很大程度上决定了最小二乘支持向量机的学习能力和预测能力。采用带有交互验证(cross-validation)的网格搜索法(grid-search)得到这两个参量的最优组合。寻优过程由粗选和精选两个步骤组成:粗选格点数  $10 \times 10$ ,如图 2“·”所示,搜索步长较大,采用误差等高线确立最优参量范围;精选格点数仍为  $10 \times 10$ ,如图 2 中“×”所示。在粗选基础上,以较小步长更加细致地搜索,以定标集交互验证误差均方根(Root Mean Square Error of Cross-validation, RMSECV)最小为指标,确定最优的  $(\gamma, \sigma^2)$  组合。依据最小二乘支持向量机原理,经验最大取值范围是: $\gamma$  取  $10^{-1} \sim 10^4$ ,  $\sigma^2$  取  $10^{-4} \sim 10^8$  [19]。这是一个遍历过程,初始值的选取对结果没有影响,本文中这两个参量的初始值均取最小值。通过计算得到黄酒糖度建模参量的最优组合为: $\gamma=42.428, \sigma^2=23.369$ 。该模型的预测结果如表 2 和图 3 所示。在图 3 中,横坐标表示黄酒糖度的测量值,纵坐标表示黄酒糖度的预测值,预测的相关系数  $r=0.962, \Delta_{RMSEP}=0.021, \Delta_{Bias}=-0.001$ ,预测精度优于 PLS 模型,WT-PLS 模型和 PCA-ANN 模型。

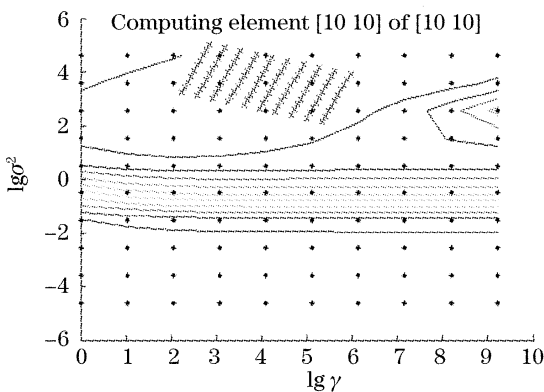


图 2 LS-SVM 模型中  $\gamma$  和  $\sigma^2$  网格寻优过程

Fig. 2 Grid search for the optimal  $\gamma$  and  $\sigma^2$  by LS-SVM model

综合以上四种建模方法的预测结果,可知用 PCA-LS-SVM 模型对黄酒糖度预测效果最好,其次为 PCA-ANN 模型。说明这两种方法能够利用光谱数据中的非线性信息,提高模型的预测精度。而通过小波变换对光谱数据进行压缩和降噪,也提高了 PLS 模型的预测精度。PCA-LS-SVM 模型的预测的精度优于文中相关文献的结果。结果表明,应用可见-近红外光谱结合 PCA-LS-SVM 方法对黄酒糖度进行预测能够获得满意的结果。

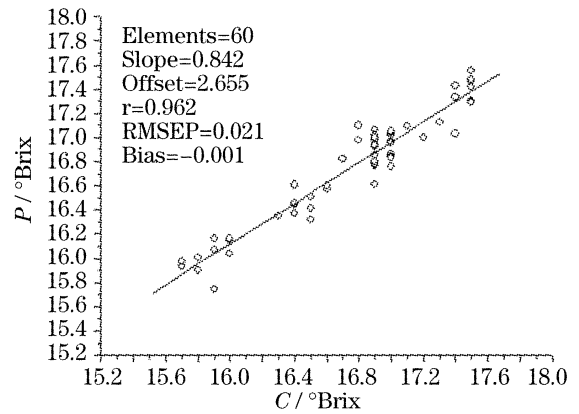


图 3 黄酒糖度的 PCA-LS-SVM 模型预测结果

Fig. 3 Prediction results of sugar content of rice wine by PCA-LS-SVM model

## 4 结 论

应用可见-近红外光谱对黄酒糖度进行了预测,通过四种不同建模方法(PLS、WT-PLS、PCA-ANN、PCA-LS-SVM)的比较,得到了黄酒糖度预测的最优模型,即主成分分析与最小二乘支持向量机相结合(PCA-LS-SVM)的模型。该模型对黄酒糖度进行预测的相关系数  $r=0.962, \Delta_{RMSEP}=0.021, \Delta_{Bias}=-0.001$ ,获得了满意的预测效果。说明应用可见-近红外光谱对黄酒糖度进行预测是可行的,并且通过主成分分析与最小二乘支持向量机相结合的方法建模获得了满意的预测精度。

## 参 考 文 献

- Chen Cheng, Yin Zijian, Xu Su. The history and nutrition value of yellow wine[J]. *Liquor Making*, 2002, **29**(1): 55~56 (in Chinese)  
陈 成,殷子建,徐 速. 浅析黄酒的历史及营养价值[J]. *酿酒*, 2002, **29**(1): 55~56
- Xie Guangfa, Dai Jun, Zhao Guangao *et al.*. Understanding scientifically the health-protect function of rice wine[J]. *China Brewing*, 2004, **1**: 30~31 (in Chinese)  
谢广发,戴 军,赵光鳌 等. 科学认识黄酒的保健养生功能[J]. *中国酿造*, 2004, **1**: 30~31
- Wang Fuyuan. *Modern Techniques for Food Fermentation* [M]. Second Edition, Beijing: China Light Industry Press, 2004. 73~89 (in Chinese)  
王福源. *现代食品发酵技术* [M]. 第 2 版,北京:中国轻工业出版社,2004. 73~89
- Yan Yanlu, Zhao Longlian, Han Donghai *et al.*. *The Foundation and Application of Near-Infrared Spectroscopy Analysis* [M]. Beijing: China Light Industry Press, 2004. 3 (in Chinese)  
严衍禄,赵龙莲,韩东海 等. *近红外光谱分析基础与应用* [M]. 北京:中国轻工业出版社,2004. 3
- Zhao Jiewen, Zhang Haidong, Liu Muhua. Preprocessing methods of near-infrared spectra for simplifying prediction model of sugar content of apples[J]. *Acta Optica Sinica*, 2006, **26**(1): 136~140 (in Chinese)  
赵杰文,张海东,刘木华. 简化苹果糖度预测模型的近红外光谱预处理方法[J]. *光学学报*, 2006, **26**(1): 136~140
- Yande Liu, Yibin Ying, Xiaping Fu *et al.*. Experiments on

- predicting sugar content in apples by FT-NIR Technique[J]. *J. Food Engineering*, 2007, **80**: 986~989
- 7 Bart M. Nicolai, Karen I. Theron, Jeroen Lammertyn. Kernel PLS regression on wavelet transformed NIR spectra for prediction of sugar content of apple [J]. *Chemometrics and Intelligent Laboratory Systems*, 2007, **85**: 243~252
- 8 Yongni Shao, Yong He, Antihus H. Gómez *et al.*. Visible/near infrared spectrometric technique for nondestructive assessment of tomato 'Heatwave' (*Lycopersicon esculentum*) quality characteristics[J]. *J. Food Engineering*, 2007, **81**: 672~678
- 9 Yongni Shao, Yong He. Nondestructive measurement of the internal quality of bayberry juice using Vis/NIR spectroscopy[J]. *J. Food Engineering*, 2007, **79**: 1015~1019
- 10 Haiyan Cen, Yong He, Min Huang. Measurement of soluble solids contents and pH in orange juice using chemometrics and Vis-NIRS[J]. *J. Agricultural and Food Chemistry*, 2006, **54**: 7437~7443
- 11 Shi Youming, Liu Gang, Liu Jianhong *et al.*. Identification of auricularia auricula from different regions by Fourier transform infrared spectroscopy[J]. *Acta Optica Sinica*, 2007, **27**(1): 129~132 (in Chinese)  
时有明,刘刚,刘剑虹等.不同产地黑木耳的傅里叶变换红外光谱鉴别[J].*光学学报*, 2007, **27**(1): 129~132
- 12 Liu Gang, Xing Da, Wang Haimin *et al.*. Study of protein in human gallstones by Fourier transform infrared spectroscopy and surface-enhanced Raman spectroscopy[J]. *Acta Optica Sinica*, 2002, **22**(4): 441~446 (in Chinese)  
刘刚,邢达,王海珉等.胆结石中蛋白质的傅里叶变换红外光谱和表面增强拉曼光谱研究[J].*光学学报*, 2002, **22**(4): 441~446
- 13 Xiaoli Li, Yong He, Changqing Wu *et al.*. Nondestructive measurement and fingerprint analysis of soluble solid content of tea soft drink based on Vis/NIR spectroscopy [J]. *J. Food Engineering*, 2007, **82**: 316~323
- 14 Yao Jie, Yang Qian, Sun Caiyun *et al.*. Qualitative analysis of methanol in the forged wine using infrared spectrophotometer[J]. *Chin. J. Spectroscopy Laboratory*, 2000, **17**(1): 35~37 (in Chinese)  
姚杰,杨倩,孙彩云等.红外光谱法定性分析假酒中的甲醇[J].*光谱实验室*, 2000, **17**(1): 35~37
- 15 L. Liu, D. Cozzolino, W. U. Cynkar *et al.*. Geographic classification of Spanish and Australian tempranillo red wines by visible and near-infrared spectroscopy combined with multivariate analysis[J]. *J. Agricultural and Food Chemistry*, 2006, **54**: 6754~6759
- 16 Wang Li, Wang Diqiang, Wang Hua *et al.*. Combine with NIR transmission spectroscopy and gas chromatography to build the fingerprint model[J]. *Liquor Making*, 2005, **32**(4): 18~20 (in Chinese)  
王莉,汪地强,汪华等.近红外透射光谱法和气相色谱法结合建立茅台酒指纹模型[J].*酿酒*, 2005, **32**(4): 18~20
- 17 Haiyan Yu, Yibin Ying, Xiaping Fu *et al.*. Quality determination of Chinese rice wine based on Fourier transform near infrared spectroscopy[J]. *J. Near Infrared Spectroscopy*, 2006, **14**: 37~44
- 18 Chu Xiaoli, Yuan Hongfu, Lu Wanzhen. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique[J]. *Progress in Chemistry*, 2004, **16**(4): 528~542 (in Chinese)  
褚小立,袁洪福,陆婉珍.近红外分析中光谱预处理及波长选择方法进展与应用[J].*化学进展*, 2004, **16**(4): 528~542
- 19 Zhu Jiayuan, Yang Yun, Zhang Hengxi *et al.*. Data prediction with few observations based on optimized least squares support vector machines[J]. *Acta Aeronautica et Astronautica Sinica*, 2004, **25**(6): 565~568 (in Chinese)  
朱家元,杨云,张恒喜等.基于优化最小二乘支持向量机的小样本预测研究[J].*航空学报*, 2004, **25**(6): 565~568