第 27卷第 1期
2008年 2月

红 外 与 毫 米 波 学 报
J. Infrared Millim. Waves

Vol. 27, No. 1
February, 2008

文章编号: 1001 - 9014(2008)01 - 0051 - 05

应用光谱技术和支持向量机分析方法快速检测啤酒糖度和 pH值

王 莉, 何 勇, 刘 飞, 应霞芳
(浙江大学生物系统工程与食品科学学院,浙江 杭州 310029)

摘要:为实现啤酒糖度和 pH值的快速检测,采用可见/近红外光谱仪器得到 360个啤酒样本的可见/近红外光谱数据.使用主成分分析(PCA)对数据进行降维处理以消除众多信息共存中相互重叠的部分,得到 6个主成分值.将样本数据随机分为定标集和预测集,利用最小二乘支持向量机(LS-SVM)算法在定标集数据基础上建立啤酒糖度和 pH值预测模型,并利用此模型对预测集样本进行预测.根据预测相关系数(*r*)和预测标准偏差(RMSEP)判断预测模型好坏,结果表明该模型对啤酒糖度预测的相关系数 *r*为 0.9829, RMSEP为 0.1506;对啤酒 pH值的预测相关系数 *r*为 0.9563, RMSEP为 0.0494,预测精度明显高于神经网络和 PLS预测,所以利用该模型能够准确的预测啤酒的糖度及 pH值.

关 键 词:啤酒;可见/近红外光谱;最小二乘支持向量机;糖度; pH
中图分类号: TS262.5 **文献标识码**: A

RAPID DETECTION OF SUGAR CONTENT AND pH IN BEER BY USING SPECTROSCOPY TECHNIQUE COMBINED WITH SUPPORT VECTOR MACHINES

WANG Li, HE Yong, LIU Fei, YING Xia-Fang
(College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310029, China)

Abstract: For the rapid detection of sugar content and pH in beer, visible and near infrared (VIS/NIR) spectra of 360 beer samples were collected by using VIS/NIR spectroradiometer. Principal component analysis (PCA) was applied for reducing the dimensionality in order to decrease the overlapped information of the raw spectral data, and 6 principal components (PCs) were selected. The samples were randomly separated into calibration set and validation set, and least squares-support vector machines (LS-SVM) algorithm was used to build calibration model of sugar content and pH in beer, then the model was employed for the prediction of the validation set. Correlation coefficient (*r*) of prediction and root mean square error of prediction (RMSEP) were used as the evaluation standards. The results indicate that the *r* and RMSEP for the prediction of sugar content are 0.9829 and 0.1506, while 0.9563 and 0.0494 for pH, respectively. The precision of prediction was obviously higher than that of neural network and PLS models. Hence, LS-SVM model with high prediction precision can be applied for the determination of sugar content and pH of beer.

Key words: beer; VIS/NIR spectroscopy; least squares-support vector machines; sugar content; pH

引言

可见/近红外光谱技术是一种快速、简便的分析技术,它已被广泛的应用于定性或定量分析食品、药品等的成分、类别,土壤有机质含量等.光谱技术已经在各个领域尤其是在食品科学和农业科学上已经成为一种重要的检测方法[1, 2].

支持向量机(SVM)是一种新型建模方法,它通过结构风险最小化原理来提高泛化能力,较好地解决了小样本、非线性、高维数、局部极小等实际问题,开始成为解决"维数灾"和"过学习"等传统困难的一种有力手段.已经在模式识别、信号处理、函数逼

收稿日期: 2007 - 03 - 21,**修回日期**: 2007 - 12 - 18 **Received date:** 2007 - 03 - 21, **revised date:** 2007 - 12 - 18
基金项目:国家十一五科技支撑计划项目(2006BAD10A04);国家自然科学基金项目(30671213);高等学校优秀青年教师教学科研奖励计划(02411)
作者简介:王 莉(1983-),男,河北保定人,浙江大学生物系统工程与食品科学学院硕士生,主要从事数字农业和多光谱检测技术研究.

近等领域得到了广泛应用[3]. 经典 SVM 模型较为复杂,求解速度较慢. LS-SVM 是对经典 SVM 的一种改进,以求解一组线性方程代替经典 SVM 中复杂的二次优化问题,降低了计算的复杂性,并且加快了计算的速度[4].

用传统的方法测定啤酒的各项指标虽然精确,但是过程繁琐,且不便于进行快速在线检测,已经有学者利用近红外光谱技术对啤酒的一些指标进行检测,但是所用的建模方法都是一些常规的方法,预测精度不甚理想,应用最小二乘支持向量机建立起预测模型,同传统的偏最小二乘(PLS)预测模型和神经网络预测模型相比具有较高的精度.

1 材料与方法

1.1 仪器设备

实验使用美国 ASD(Analytical Spectral Device)公司的 Handheld FieldSpec 光谱仪,其光谱采样间隔(波段宽)1.5nm,测定范围 325～1075nm,扫描次数 30次,探头视场角为 20度. 分析软件为 ASD View Spec Pro、Unscramble V9.6、MATLAB 7 和 LS-SVM 工具包.

啤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,测量值单位为°Brix,测量准确度以折射率表示,$n_D = \pm 0.0002$.

啤酒 pH 值测量使用上海大普仪器有限公司生产的 PHS-4CT 型酸度计,pH 测量范围 0～14,分辨率 0.001,输出电压范围 -1999.9～1999.9mV.

1.2 样品来源及光谱的获取

从超市选得 6 种典型啤酒,百威啤酒、喜力啤酒、青岛啤酒、燕京啤酒、雪花啤酒和西湖啤酒共计 360 个样本,每个品种 60 个样本. 样本随机分成定标集和预测集,定标集设置为 270 个,每个品种 45 个;预测集设置为 90 个,每个品种 15 个. 实验在恒温 25℃下进行,采用透射光谱法,光程 2mm. 光谱仪置于样品池正上方,对每一个样品扫描 30 次,并将其取平均,由此得到一个样本光谱.

1.3 光谱数据预处理

为了去除来自高频随机噪音、基线漂移、样本不均匀、光散射等影响,需要进行光谱预处理来消除噪音. 采用 Moving Average Smoothing 平滑,选用平滑点数为 3,此时能很好滤除各种因素产生的高频噪音;再进行 Standard Normal Variate (SNV)处理[5,6].

1.4 最小二乘支持向量机(LS-SVM)

Suykens J. A. K 提出的最小二乘支持向量机,用最小二乘线性系统代替传统的支持向量机. 最小二乘支持向量机的算法描述如下[4,7,8]:

训练集样本可以设为:

$$D = \{(x_k, y_k) \mid k = 1, 2, \cdots, N\}, x_k \in R^n, y_k \in R \quad . \tag{1}$$

其中 x_k 是输入向量,y_k 是目标值. 在权 w 空间中的函数估计问题可以描述为求解下面问题:

$$\min J(w, e) = \frac{1}{2} w^T + \frac{1}{2} \gamma \sum_{k=1}^{N} e_k^2 \quad . \tag{2}$$

约束条件为:$y_k = w^T \varphi(x) + b + e_k$, $k = 1, \cdots, N$.

其中:$\varphi(g): R^n \rightarrow R^{n_h}$ 是核空间映射函数,权向量 $w \in R^{n_h}$,误差变量 $e_k \in R$, b 是偏差量,γ 是可调超参数. 利用拉格朗日法求解这个优化问题:

$$L(w, b, e, \alpha) = J(w, e) - \sum_{k=1}^{N} \alpha_k \{w^T \varphi(x_k) + b + e_k - y_k\} \quad . \tag{3}$$

其中:α_k, $k = 1, \cdots, N$, 是拉格朗日乘子. 根据优化条件:

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{k=1}^{N} \alpha_k \varphi(x_k) \\ \dfrac{\partial L}{\partial b} = 0 \rightarrow \sum_{k=1}^{N} \alpha_k \varphi(x_k) = 0 \\ \dfrac{\partial L}{\partial e_k} = 0 \rightarrow a_k = \gamma e_k, \; k = 1, \cdots, N \\ \dfrac{\partial L}{\partial \alpha_k} = 0 \rightarrow w^T \varphi(x_k) + b + e_k - y_k \\ \qquad = 0, \; k = 1, \cdots, N \end{cases} \quad . \tag{4}$$

可得:

$$\begin{bmatrix} 0 & I^T \\ I & \Omega + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad . \tag{5}$$

其中:$x = [x_1, \cdots, x_N]$, $y = [y_1, \cdots, y_N]$, $\alpha = [\alpha_1, \cdots, \alpha_N]$, $I = [1, \cdots, 1]$. 核函数 $K(x_k, x_l) = \varphi(x_k)^T \varphi(x_l)$, $k, l = 1, \cdots, N$ 是满足 Mercer 条件的对称函数.

最小二乘支持向量机回归估计为:

$$y(x) = \sum_{k=1}^{N} \alpha_k K(x, x_k) + b \quad . \tag{6}$$

其中 α, b 由式(5)求出,核函数 $K(x_k, x_l)$ 为满足 Mercer 条件的任意对称函数. 常见的核函数有线性核函数、多项式核函数、RBF(Radial Basis Function)核函数、多层感知核函数等. 本文采用了 RBF 核函数:

$$K(x, y) = \exp \frac{-(x - y)^2}{2\sigma^2} \quad . \tag{7}$$

2 实验结果与分析

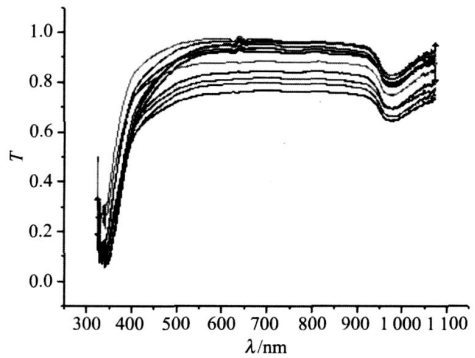

图 1　啤酒样本的可见/近红外光谱原始图谱
Fig 1　Original Vis/NIR spectroscopy for rice beer samples

表 1　主成分的累计可信度
Table 1　The reliabilities of principal components

主成分 Principal component	PC1	PC2	PC3	PC4	PC5
累计可信度 Reliabilities	82.954%	96.265%	98.391%	99.224%	99.402%
主成分 Principal component	PC6	PC7	PC8	PC9	PC10
累计可信度 Reliabilities	99.525%	99.594%	99.662%	99.724%	99.768%

2.1　不同品牌啤酒的可见/近红外光谱图

六种啤酒的可见/近红外光谱如图 1所示(为图形显现清晰,每个品牌的啤酒随意选取 2条光谱曲线). 图 1中横坐标为波长范围 325～1075nm,纵坐标为光谱透射率. 从图 1中可以看出,不同品牌啤酒的光谱曲线交错重叠. 由于实验条件和仪器的限制,光谱数据存在不同程度的平移,还会引入一些噪声,要消除这些因素的影响,我们必须要结合化学计量学方法对所得的光谱数据进行处理,从而建立起啤酒糖度和 pH值检测模型.

2.2　主成分分析

测量得到的每个样本的光谱数据包含 751个透射率值,虽然可以将每个样本的所有透射率值作为输入值进行 LS-SVM建模. 但是这样数据点数量庞大,必然影响运算速度. 为了提高建模速度,减少运算量,通过主成分分析(Principle Component Analysis, PCA)对原始光谱输入变量进行降维压缩. 经主成分分析光谱数据后得到前 10个主成分累计贡献率如表 1所示,前 6个主成分的累计贡献率已经达到了 99.525%,而且这之后,随着主成分数的增加,累计贡献率增加相当缓慢(低于 0.1%),所以选用前 6个主成分作为优化输入特征子集,进行 LS-SVM建模.

2.3　LS-SVM建模及预测

采用 LS-SVM运算时,必须选择合适的核函数,目前还没有形成一个统一的模式来选择核函数. 通过和其他核函数的比较,RBF核函数作为非线性函数能够减少训练过程中计算的复杂性. 采用 RBF作为核函数的 LS-SVM模型主要有两个参数:超参数 γ和 RBF核函数参数 σ^2,这两个参数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.

我们采用定标集 270个样本进行 LS-SVM建模. 对于这两个参数的优化,本研究中采用的是基于交叉验证(Cross-validation)的网格搜索(Grid-search). 这个方法的原理就是把要选择的参数当作一个坐标格子上的点,选择的过程就是遍历空间中的各个方向的参数组合的空间点,寻优过程由粗选和精选两个步骤组成:粗选格点数 10×10,如图 2和图 3中"·"所示,搜索步长较大,采用误差等高线确立最优参数范围;精选格点数仍为 10×10,如图 2和图 3中"×"所示,在粗选基础上,以较小步长更加细致地搜索,在参数优化过程将每组 γ和 σ^2的组合所得到的训练集交叉验证误差均方根(Root Mean Square Error of Cross-validation, RMSECV)最小值为指标,并以此确定最优模型参数. 依据最小二乘支持向量机原理,经验最大取值范围是:γ取 10^{-1}～10^4,σ^2取 10^{-4}～10^{8}[8],因为是一个遍历过程,初始值的选取对结果没有影响,本研究中这两个参数的初始值均取最小值. 图 2是对啤酒糖度建模时的寻优过程,图 3是对啤酒 pH值建模时的寻优过程,这两幅图的横坐标是 γ的对数,纵坐标为 σ^2的对数. 对糖度建模的参数优化结果为:γ = 200.88, σ^2 = 22.331;对 pH值建模的参数优化结果为:γ = 1.639, σ^2 = 1.2787.

预测结果的相关系数和预测标准偏差(Root Mean Square Error for Prediction, RMSEP)被用于评估模型的预测能力. 最终 LS-SVM对啤酒糖度的预测结果如图 4所示,其中相关系数 r = 0.9829, RMSEP = 0.1506;对啤酒 pH值的预测结果如图 5所示. 其中相关系数 r = 0.9563, RMSEP = 0.0494. 表 2列举了偏最小二乘法(PLS)、误差反向传输人工神经网络(BP-ANN)和 LS-SVM三种建模方法预测结果的相关系数 r和预测标准偏差 RMSEP,一般是预测结果相关系数越大,预测标准偏差越小,所建的模型预测能力就越强. 从表 2数据可以发现无论是对啤酒糖度的预测还是对 pH的预测,LS-SVM的预测结果都要优于其它两种建模方法.

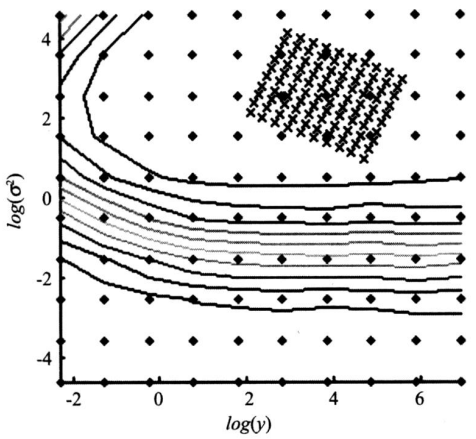

图 2 糖度模型 γ和 σ^2寻优过程

Fig 2 Grid search on γ and σ^2 in sugar content model

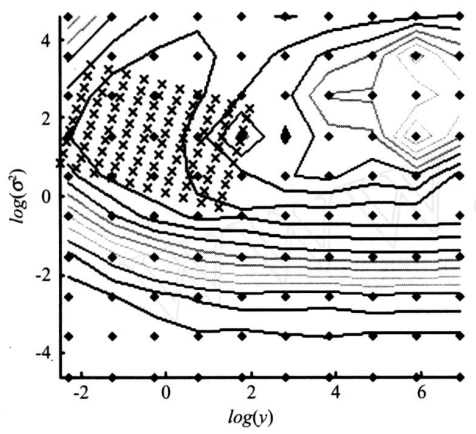

图 3 pH值模型 γ和 σ^2寻优过程

Fig 3 Grid search on γ and σ^2 in pH model

表 2 基于不同模型的预测结果

Table 2 Prediction results for different models

建模方法	糖度预测结果		pH预测结果	
	r	RMSEP	*r*	RMSEP
PLS	0.9165	0.3300	0.8643	0.1264
BP-ANN	0.9539	0.2559	0.8818	0.1645
LS-SVM	0.9829	0.1506	0.9563	0.04943

3 结论

应用可见/近红外光谱技术对啤酒的糖度和 pH 值进行了预测,采用主成分分析和最小二乘支持向量机建立了啤酒糖度和 pH 值预测模型,预测结果的相关系数和均方根误差都达到了较好的效果,同传统的偏最小二乘法和 BP 人工神经网络模型相比具有更高的精度,是对传统啤酒指标测定方法的改进,为开发出更高精度的啤酒糖度、pH 值以及其它一些指标的检测仪器以及实时在线测量提供了依据.对于今后的研究,主要是优化预测模型,使其对

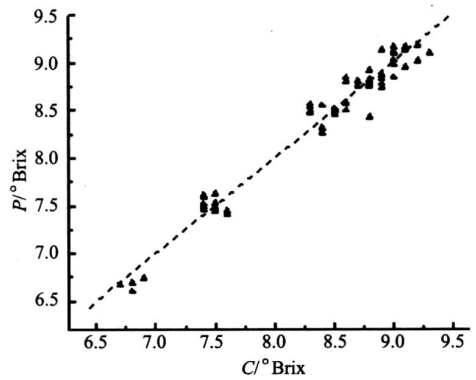

图 4 啤酒糖度预测结果

Fig 4 Prediction results of sugar content in beer

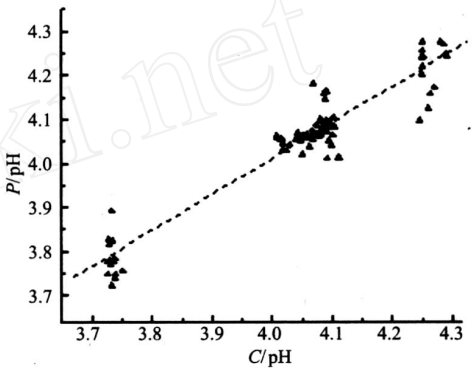

图 5 啤酒 pH值预测结果

Fig 5 Prediction results of pH in beer

糖度、pH以及多类啤酒指标具有更好的预测能力.

REFERENCES

[1] LI Xiao-Li, HU Xing-Yue, HE Yong. New approach of discrimination of varieties of juicy peach by near infrared spectra based on PCA and MDA model[J]. *J. Infrared Millim. Waves* (李晓丽,胡兴越,何勇. 基于主成分和多类判别分析的可见/近红外光谱水蜜桃品种鉴别新方法. 红外与毫米波学报), 2006, **25**(6): 417—420.

[2] HE Yong, LI Xiao-Li. Discrimination of varieties of waxberry using near infrared spectra[J]. *J. Infrared Millim. Waves* (何勇,李晓丽. 近红外光谱杨梅品种鉴别方法的研究. 红外与毫米波学报), 2006, **25**(3): 192—194.

[3] YU Ke, CHENG Yi-Yu. Discriminating the genuineness of Chinese medicines with least squares support vector machines[J]. *Analytical Chemistry* (虞科,程翼宇. 一种基于最小二乘支持向量机算法的近红外光谱判别分析方法. 分析化学), 2006, **34**(4): 561—564.

[4] YU Yan-Fang, GAO Da-Qi. An improved least squares support vector machine and its applications[J]. *Computer Engineering & Science* (余艳芳,高大启. 一种改进的最小二乘支持向量机及其应用. 计算机工程与科学), 2006, **28**(2): 69—71, 85.

[5] CHU Xiao-Li, YUAN Hong-Fu, LU Wan-Zhen. Progress and application of spectral data pretreatment and wavelength

selection methods in NIR analytical technique[J]. *Progress In Chemistry* (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. 化学进展), 2004, **16**(4): 528—539.

[6] SHAO Yong-Ni, HE Yong. Method for predieting acidity of bayberry juice by using Vis/near infrared spectra[J]. *J. Infrared Millim. Waves* (邵咏妮,何勇. 可见/近红外光谱预测极梅法酸度的方法研究. 红外与毫米波学报), 2006, **25**(6): 478-480.

[7] TANG He-Sheng, XUE Song-Tao, CHEN Rong, *et al*. On-line weighted LS-SVM for hysteretic structural system identification[J]. *Eng. Struct*. 2006, **28**(12): 1728—1735.

[8] ZHU Jia-Yuan, YANG-Yun, ZHANG Heng-Xi, *et al*. Data prediction with few observations based on optimized least squares support vector machines[J]. *Acta Aeronautica et Astronautica Sinica* (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. 航空学报), 2004, **25**(6): 565—568.

(上接 50页)

4 结论

本文提出了一种基于 KFKT的红外小目标检测方法.我们利用 KPCA特征避免了直接计算映射函数,给出了 FKT推广为 KFKT的理论推导过程,使 KFKT具有提取图像高阶统计特征的能力.为了检验 KFKT的目标检测性能,我们分别选取了 3种典型背景的红外图像,用 KFKT检测其中的小目标.实验结果表明,KFKT比基于 FKT的二次相关滤波具有更优良的检测性能,这是因为 KFKT具有描述图像高阶统计特征的能力.

REFERENCES

[1] YANG Lei, YANG Jie, ZHENG Zhong-Long. Detecting infrared small target based on adaptive local energy threshold under sea-sky complex backgrounds[J]. *J. Infrared Millim. Waves* (杨磊,杨杰,郑忠龙. 海空复杂背景中基于自适应局部能量阈值的红外小目标检测. 红外与毫米波学报), 2006, **25**(1): 41—45.

[2] Mahalanobis A, Muise R, Stanfill S, *et al*. Design and application of quadratic correlation filters for target detection[J]. *IEEE Trans AES*, 2004, 40(3): 837—850.

[3] Fukunaga K, Koontz W. Representation of random processes using the finite karhunen-loeve expansion[J]. *IEEE Trans. on Information and Control*, 1970, **16**(1): 85—101.

[4] Yang J, Frangi A, Yang J, *et al*. KPCA Plus LDA: A complete kernel fisher discriminant framework for feature extraction and recognition[J]. *IEEE Trans. on PAMI*, 2005, **27**(2): 230—244.

[5] YANG Lei, YANG Jie. Real-time method for detecting multi-small targets in infrared large sight field[J]. *J. Infrared Millim. Waves* (杨磊,杨杰. 一种红外大视场环境下的多小目标适时检测方法. 红外与毫米波学报), 2006, **25**(5): 377—381.

[6] YE Zeng-Jun, WANG Jiang-An, RUAN Yu, *et al*. Detection algorithm of weak infrared point targets under complicated background of sea and sky[J]. *J. Infrared Millim. Waves* (叶增军,王江安,阮玉,等. 海空复杂背景中基于自适应局部能量阈值的红外小目标检测. 红外与毫米波学报), 2000, **19**(2): 121—124.