第29卷，第4期
2009年4月

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

Vol. 29, No. 4, pp931-934
April, 2009

基于最小二乘支持向量机的番茄汁糖酸度分析研究

黄 康，汪辉君，徐惠荣*，王剑平，应义斌

浙江大学生物系统工程与食品科学学院，浙江 杭州 310029

摘 要 近红外光谱应用于农产品内部品质无损检测的方法引起人们的广泛关注，在分析过程中建立一个稳定可靠的模型用于处理非线性数据集是十分重要的，也是有一定难度的。目前常用的偏最小二乘(PLS)、主成分回归(PCR)以及逐步多元线性回归(SMLR)等方法还不能解决这类问题。文章提出了将基于统计学原理的最小二乘支持向量机(LS-SVM)回归方法用于番茄汁的近红外(NIR)光谱分析，预测番茄汁品质(糖度和有效酸度)。运用LS-SVM方法以67个番茄汁样本建模，采用高斯径向基函数(RBF)为核函数，对33个样本进行糖酸度预测，糖度的相关系数为0.990 25，均方根标准预测误差为0.0056° Brix；有效酸度的相关系数为0.967 5，均方根标准预测误差为0.024 5。结果表明，LS-SVM方法要优于PLS和PCR建模方法，是一种快速、准确的近红外光谱分析方法。

关键词 近红外光谱；最小二乘支持向量机；番茄汁；糖度；有效酸度

中图分类号：S132 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2009)04-0931-04

引 言

番茄是人们十分爱吃的果类蔬菜。用番茄榨汁得到的番茄汁，不仅具有鲜美的颜色和适口的味道，还含有维生素、多种氨基酸和矿物质，营养价值非常高。对于保护血管，防止高血压和心脏疾病，也有一定的作用。目前，市场上出现了越来越多的番茄汁饮品，其产品品质受到人们的关注。糖度和酸度是衡量番茄汁品质的主要理化指标，对于番茄汁的色泽、口味和营养都有重要的影响。

偏最小二乘(PLS)、逐步多元线性回归(SMLR)、主成分回归(PCR)是近红外定量分析的主要建模方法[1]。但是当样品数量较少或变量(波长)较多，模型非线性，就会出现“过拟合”和“欠拟合”问题，产生较大的误差。近年来新提出的“支持向量机”算法能限制过拟合和欠拟合，而且因其采用核函数算法，能有效处理非线性数据集[2]。

支持向量机(support vector machines，SVM)是Vapnik等提出的一种用于分类与非线性回归的方法，属于神经网络和非线性建模的范畴[3]。由于SVM在实际应用中表现出传统学习方法没有的优越性，从理论上讲得到的是全局最优点，避免了神经网络方法中的局部极小值问题，因此受到了越来越多的关注[2, 4]。张录达等以中药大黄样品作为实验材料，通过SVM近红外光谱法建立大黄样品真伪识别模型，识别准确率达到96.77%[5]。张录达等还采用四种不同核函数方法对小麦样品蛋白质含量与小麦样品近红外光谱进行SVM回归建模，预测结果的相关系数均在0.97以上，平均绝对误差小于0.32[6]。白鹏等提出了一种基于支持向量机的混合气体红外光谱组分浓度和种类分析的新方法。混合气体组分浓度实验结果的最大平均绝对误差为0.132%，混合气体组分种类识别的准确率大于94%。解决了传统的光谱分析中光谱特征谱线重叠、光谱数据的维数大、定性和定量分析无法使用同一方法等问题[7]。最小二乘支持向量机(least square support vector machines，LS-SVM)是经典SVM的一种改进，以求解一组线性方程代替经典SVM中较复杂的二次优化问题，降低了计算复杂性，加快了求解速度[8]。姚肖刚等人将LS-SVM应用于汽油辛烷值近红外光谱分析，计算量显著减少，特别适合于在线应用，预测精度显著优于多元线性回归(MLR)以及偏最小二乘(PLS)等方法[9]。Chauchard等把LS-SVM用于非线性近红外光谱分析，应用于构造预测葡萄酸度的轻便式传感器[10]。Alessandra等在奶粉掺杂物(淀粉、乳清和蔗糖)含量的定量分析中将LS-SVM与近红外光谱相结合，建立的预测模型精确度比用偏最小二乘回归方法更好[11]。

本文运用LS-SVM建模方法对鲜榨番茄汁的NIR光谱

收稿日期：2007-10-16，**修订日期**：2008-01-22

基金项目：国家自然科学基金项目(60778024)和国家科技支撑计划项目(2006BAD10A04)资助

作者简介：黄 康，1985年生，浙江大学生物系统工程与食品科学学院博士研究生 e-mail：huangkang23@hotmail.com

* 通讯联系人 e-mail：hrxu@zju.edu.cn

数据进行建模预测处理，对 LS-SVM 建模方法得出的预测结果进行了讨论。并比较了 LS-SVM，PLS 和 PCR 方法对鲜榨番茄汁品质预测的结果。

1 材料和方法

1.1 材料

试验用的番茄样品均为直接从超市购买的普通番茄，把表皮清理擦净后，用市售榨汁机榨出番茄汁，然后过滤，再将过滤所得溶液离心，取清液作为样本，依次作好标记。将样本放进冰箱，冷藏一天后取出，在 25 ℃的条件下放置 7 h。所有试验均在 25 ℃左右室温下进行。总共 100 个试验样品，分为校正组、预测组两组，数量分别为 67 和 33 个。

1.2 仪器及工作参数

番茄汁的近红外透射光谱采用 Nexus 傅里叶变换近红外光谱仪及其相应的透射附件进行采集。光源为 50 W 石英卤素灯，检测器选用 InGaAs，液体样品池为 1 mm 光程的石英比色皿。试验采用 OMNIC6.1(尼高力仪器公司，美国)软件进行光谱参数设置和光谱数据的采集和存储。光谱采集参数为：测量波长范围 12 500～4 000 cm^{-1}，扫描次数 64，仪器分辨率 4 cm^{-1}，动镜速度 0.949 4 $cm \cdot s^{-1}$。采集样品光谱前以空白比色皿为背景进行背景光谱采集。

样品的糖度测定具体参照食品卫生检验方法理化部分总则(GB/T5009.1－2003)。用 PR-10 型手持糖量计(日本 ATAGO 公司)测量糖度。样品的有效酸度测定参照卫生检验方法理化总则(GB/T5009.1－2003)及食品中总酸的测定方法(GB/T12456－90)。用上海精密科学仪器有限公司生产的 SJ-4A 型实验室 pH 计进行测定。

1.3 光谱处理和数据分析

Nexus FTIR 光谱仪所获得的光谱为原始吸收光谱，利用 OMINIC6.1 对原始吸收光谱进行处理，提取光谱的有效信息。常用的光谱数据预处理方法有平滑和微分两种方法。平滑方法可去除光谱信号中高频噪声的干扰，较多的平滑点数可以使信噪比提高，但同时也会导致信号失真。微分方法可消除基线和背景的干扰，提高分析精度，但是原始光谱经微分后，噪声增大。本研究采用平滑处理。

数据分析使用 Nicolet 公司智能定量分析软件 TQ Analyst v 6.2.1 软件，提取光谱的主成分信息，并结合 Matlab6.1。通过 LS-SVM 建立校正模型，最后用预测样本通过回归分析来评价校正模型的精度。

2 结果与分析

2.1 番茄汁的近红外透射光谱

光谱采集在相同的实验条件下进行，得到原始光谱图如图 1 所示。

由于原光谱峰值出现在 7 300～4 200 cm^{-1}之间，且在此范围内噪声比较小。因此本实验中用来建模的数据为此波段的吸收值。

2.2 糖酸度测量结果

样品糖度(SC)测量结果为：最大值 5.5° Brix，最小值 4.0° Brix，平均值 4.67° Brix，标准偏差 0.23° Brix。样品有效酸度(VA)测量结果为：最大值 pH 4.80，最小值 pH 4.33，平均值 pH 4.54，标准偏差 pH 0.09。表 1 为校正集与验证集样品的糖酸度分布情况。

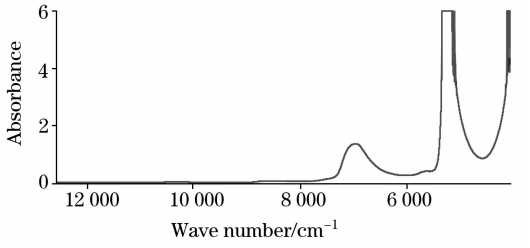

Fig. 1 NIR spectra of tomato juice samples

Table 1 Statistic data of SC and VA in samples

Samples indexes	Calibration	Prediction	Calibration	Prediction
	SC/(°Brix)		VA/pH	
No. of samples	67	33	67	33
Mean	4.66	4.69	4.53	4.54
Maximum	5.5	4.9	4.80	4.75
Minimum	4.0	4.3	4.33	4.38
Standard deviation	0.246 9	0.159 4	0.087 1	0.084 8

2.3 模型的建立与预测分析

利用样品的原始光谱 7 300～4 200 cm^{-1}的 1 612 个数据点进行主成分分析。提取 10 个主成分，以 67 个建模样品的 10 个主成分信息作为 LS-SVM 学习的输入因子，核函数采用径向基函数(RBF)[12]，通过 MATLAB 语言设计 LS-SVM 建模分析程序。所建立的模型预测集与校正集的相关系数(r)和均方根误差(校正集均方根误差 RMSEC，预测集均方根误差 RMSEP)如表 2 所示。糖度的预测精度优于有效酸度，这可能是由于在试验过程中，果汁容易发生化学反应，影响了其 pH 值，使得其预测准确度下降[13]。

利用已建立的 LS-SVM 模型对 33 个已知糖度的样品进行预测，番茄汁糖度预测值和真实值的对应关系如图 2 所示，其中样本集的相关系数(r)为 0.990 25，均方根标准预测误差(RMSEP)为 0.005 6 °Brix。

Table 2 Calibration and validation results of LS-SVM method

组分	校正集(67 个样品)		预测集(33 个样品)	
	r	RMSEC	r	RMSEP
SC	0.946 94	0.0247	0.990 25	0.005 6
VA	0.915 26	0.036 5	0.967 50	0.024 5

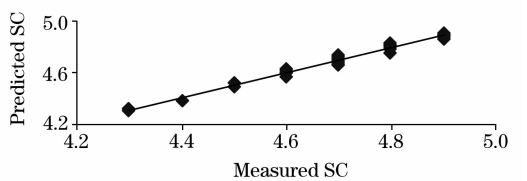

Fig. 2 Relationship between the measured and predicted sugar contents by LS-SVM

利用 LS-SVM 方法建立的模型对 33 个已知酸度的样品进行预测，番茄汁糖度预测值和真实值的对应关系如图 3 所示，其中样本集的相关系数(r)为 0.967 50，均方根标准预测误差(RMSEP)为 pH 0.024 5。由此可知，LS-SVM 所建立的模型能较好地预测番茄汁的糖度和有效酸度。

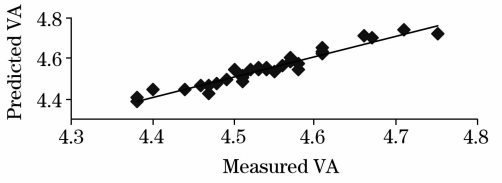

Fig. 3　Relationship between the measured VA and predicted VA by LS-SVM

为了比较不同建模方法的特点和优劣，采用偏最小二乘(PLS)和主成分回归(PCR)二种建模方法所得的结果与先前所得结果进行对比，由表 3 可以看出，对于糖度和有效酸度二种组分，LS-SVM 方法的 RMSEP 均小于 PLS 和 PCR 方法，相关系数也有一定程度的提高。这说明 LS-SVM 方法的预测准确度高，优于 PLS 和 PCR 方法。

在理化指标与光谱信息的变化关系呈非线性时，传统的 PLS 和 PCR 方法就很难达到预测目标，这是因为这些方法的处理过程是依靠找出光谱变量中的线性相关性来实现的[6, 14]。在这一方面，SVM 方法凭借其解决非线性关系的能力以及其广泛的适应能力，可以较好地处理非线性数据集。

在实际应用过程中，建模方式的选择一般倾向于建模方法尽可能简单[15]。LS-SVM 方法在建模精确度和处理非线性问题时比 PLS 和 PCR 方法更可靠；在处理速度上，PLS 方法比 LS-SVM 方法更有优势。所以在选择时应该根据具体问题的不同特点决定。

Table 3　Predicted results by LS-SVM, PLS and PCR

Component	LS-SVM		PLS		PCR	
	RMSEP	r	r	RMSEP	RMSEP	r
SC	0.005 6	0.990 25	0.028 7	0.983 82	0.057 0	0.939 31
VA	0.024 5	0.967 50	0.032 0	0.952 73	0.029 5	0.941 28

3　结　论

本文通过应用最小二乘支持向量机(LS-SVM)建模方法对番茄汁糖酸度分析，并且对比偏最小二乘(PLS)和主成分回归(PCR)建模方法，可以看出 LS-SVM 方法得到的预测结果要优于 PLS 和 PCR 法。运用 LS-SVM 方法，样本糖度的相关系数(r)为 0.990 25，标准预测误差(RMSEP)为 0.005 6；样本有效酸度的相关系数(r)为 0.967 5，标准预测误差(RMSEP)为 0.024 5。结果表明，基于最小二乘支持向量机的建模方法较偏最小二乘和主成分回归建模方法更具优势，能减少或消除“过拟合”及“欠拟合”现象所造成的误差，并且可以更好地解决非线性数据集的建模问题，在实际复杂的 NIR 分析中将发挥更重要的作用。

参　考　文　献

[1] FU Xia-ping, YING Yi-bin, LU Hui-shan, et al(傅霞萍，应义斌，陆辉山，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2007, 27(5): 911.

[2] CHEN Nian-yi, LU Wen-cong, YE Chen-zhou, et al(陈念贻，陆文聪，叶辰洲，等). Computers and Applied Chemistry(计算机与化学应用), 2002, 19(6): 691.

[3] Vapnik V N. The Nature of Statistical Learning, 2nd. ed. New York: Springer, 2000.

[4] Belousov A I, Verzakov S A, Von Frese J. Chemometrics and Intelligent Laboratory Systems, 2002, 64: 15.

[5] ZHANG Lu-da, SU Shi-guang, WANG Lai-sheng, et al(张录达，苏时光，王来生，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(1): 33.

[6] ZHANG Lu-da, JIN Ze-chen, SHEN Xiao-nan, et al(张录达，金泽宸，沈晓南，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(9): 1400.

[7] BAI Peng, XIE Wen-jun, LIU Jun-hua(白　鹏，谢文俊，刘君华). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2007, 27(7): 1323.

[8] Suykens J A K, Gestel T V, de Brabanter J, et al. Least Squares Support Vector Machines. Singapore: World Scientific, 2002.

[9] YAO Xiao-gang, DAI Lian-kui(姚肖刚，戴连奎). Proceeding of the 5th World Congress on Intelligent Control and Automation, June 15, 2004, Hangzhou, China(第五届全球智能控制与自动化大会 2004 年 6 月 15 日，中国杭州).

[10] Chauchard F, Cogdill R, Roussel S, et al. Chemometrics and Intelligent Laboratory Systems, 2004, 71: 141.

[11] Alessandra B, Marco F F, Cesar M, et al. Analytica Chimica Acta, 2006, 579: 25.

[12] GUO Hui, LIU He-ping, WANG Ling(郭　辉，刘贺平，王　玲). Journal of System Simulation(系统仿真学报), 2006, 18(7): 2033.

[13] Curda L, Kukacková O. Journal of Food Engineering, 2004, 61(4): 557.

[14] Bülent U. A Comparison of Support Vector Machines and Partial Least Squares Regression on Spectral data, 2003.

[15] Thissen U, Pepers M, Ustun B, et al. Chemometrics and Intelligent Laboratory Systems, 2004, 73: 169.

NIR Spectroscopy Based on Least Square Support Vector Machines for Quality Prediction of Tomato Juice

HUANG Kang, WANG Hui-jun, XU Hui-rong*, WANG Jian-ping, YING Yi-bin
College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310029, China

Abstract The application of least square support vector machines (LS-SVM) regression method based on statistics study theory to the analysis with near infrared (NIR) spectra of tomato juice was introduced in the present paper. In this method, LS-SVM was used for establishing model of spectral analysis, and was applied to predict the sugar contents (SC) and available acid (VA) in tomato juice samples. NIR transmission spectra of tomato juice were measured in the spectral range of 800-2 500 nm using InGaAs detector. The radial basis function (RBF) was adopted as a kernel function of LS-SVM. Sixty seven tomato juice samples were used as calibration set, and thirty three samples were used as validation set. The results of the method for sugar contents (SC) and available acid (VA) prediction were: a high correlation coefficient of 0.990 3 and 0.967 5, and a low root mean square error of prediction (RMSEP) of 0.005 6° Brix and 0.024 5, respectively. And compared to PLS and PCR methods, the performance of the LS-SVM method was better. The results indicated that it was possible to built statistic models to quantify some common components in tomato juice using near-infrared (NIR) spectroscopy and least square support vector machines (LS-SVM) regression method as a nonlinear multivariate calibration procedure, and LS-SVM could be a rapid and accurate method for juice components determination based on NIR spectra.

Keywords NIR spectroscopy; LS-SVM; Tomato juice; Sugar contents; Available acid

(Received Oct. 16, 2007; accepted Jan. 22, 2008)

* Corresponding author