

基于最小二乘支持向量机的白酒酒醅成分定量分析

熊雅婷, 李宗朋, 王 健*, 冯斯雯, 李子文, 尹建军, 宋全厚
(中国食品发酵工业研究院, 北京 100015)

摘 要: 利用近红外光谱技术实现对白酒发酵过程中酒醅主要成分的质量控制, 并进行模型优化, 提高性能。采用偏最小二乘法提取的潜在变量作为最小二乘支持向量机的输入变量, 先后建立了白酒酒醅中酒精度、淀粉、水分、酸度的近红外定量模型, 并与经无信息变量消除法波段筛选后建立的偏最小二乘模型结果进行比较。结果表明: 与偏最小二乘模型相比, 4 个指标的最小二乘支持向量机定量模型的相关系数 (R^2)、预测均方根误差以及相对分析误差 3 个评价参数均有更优表现; 对未知样品进行预测时, 最小二乘支持向量机模型的预测准确度明显高于偏最小二乘模型。说明最小二乘支持向量机模型的准确度、稳定性及预测性能均优于偏最小二乘法模型, 为白酒酒醅的品质分析方法研究提供了新的思路。

关键词: 白酒酒醅; 最小二乘支持向量机; 潜在变量; 偏最小二乘法; 波段筛选

Quantitative Analysis of Chemical Compositions of Fermented Grains of Chinese Liquor Based on Least Squares Support Vector Machine (LS-SVM)

XIONG Yating, LI Zongpeng, WANG Jian*, FENG Siwen, LI Ziwen, YIN Jianjun, SONG Quanhou
(China National Research Institute of Food and Fermentation Industries, Beijing 100015, China)

Abstract: Near infrared spectroscopy was used to predict the main chemical ingredients of fermented grains of Chinese liquor by modeling. The established models were optimized for improved prediction performance. Latent variables (LVs) were extracted by partial least squares (PLS) and used as the input variables of least squares support vector machine (LS-SVM) for the establishment of NIR quantitative models to predict the alcohol, starch, moisture contents and acidity of fermented grains. Furthermore, a comparison with the PLS models built with waveband selection using uninformative variable elimination (UVE) was carried out. The results showed that compared with the PLS models, quantitative correlation coefficients (R^2), root mean square errors of prediction (RMSEP), and relative percent differences (RPD) of alcohol, starch, moisture and acidity showed better performances in the LS-SVM models, respectively. The accuracy of the LS-SVM models in predicting unknown samples was significantly higher than that of the PLS models. In summary, the accuracy, stability and prediction performance of the LS-SVM models were better than those of the PLS ones. This study can provide a new way for quantitative analysis of fermented grains of Chinese liquor.

Key words: fermented grains of Chinese liquor; least squares support vector machines (LS-SVM); latent variables (LVs); partial least squares (PLS); waveband selection

DOI:10.7506/spkx1002-6630-201612029

中图分类号: TS261.7

文献标志码: A

文章编号: 1002-6630 (2016) 12-0163-06

引文格式:

熊雅婷, 李宗朋, 王健, 等. 基于最小二乘支持向量机的白酒酒醅成分定量分析[J]. 食品科学, 2016, 37(12): 163-168.

DOI:10.7506/spkx1002-6630-201612029. <http://www.spkx.net.cn>

XIONG Yating, LI Zongpeng, WANG Jian, et al. Quantitative analysis of chemical compositions of fermented grains of chinese liquor based on least squares support vector machine (LS-SVM)[J]. Food Science, 2016, 37(12): 163-168. (in Chinese with English abstract) DOI:10.7506/spkx1002-6630-201612029. <http://www.spkx.net.cn>

收稿日期: 2015-08-10

作者简介: 熊雅婷 (1990—), 女, 硕士研究生, 研究方向为食品无损检测。E-mail: xiongyating1130@163.com

*通信作者: 王健 (1973—), 男, 高级工程师, 博士, 研究方向为食品无损检测。E-mail: onlykissjohn@hotmail.com

酒醅发酵是白酒酿造过程的重要环节,酒醅在窖池环境中充当着物质循环、能量流动、信息传递的“三流运转”规律的载体^[1],酒醅中的淀粉、水分、总酸以及酒精等主要成分,是影响白酒品质和风格的主要物质^[2],因此,酒醅发酵直接影响白酒酿造的产量与质量。目前,常规的检测方法普遍操作复杂,分析时间长、消耗试剂,并且存在污染环境的风险,发酵过程质量变化也难以及时准确把握^[3]。

近年来,近红外光谱检测技术凭借操作简单、分析迅速、无损检测、重现性好且环保无污染等优势^[4],逐渐引入到中国传统的酿酒行业中,在白酒品质检测及组分定量分析方面发挥了重要的作用^[5]。彭帮柱^[6]、陈妍^[7]等分别利用人工神经网络、最小二乘法回归等方法,实现了白酒总酸、总酯、乙醇、乙酸乙酯等关键指标含量的快速检测。杨国强等^[8]通过近红外透射光谱分析技术对汾阳王酒进行了真假鉴别,准确率可达100%。但是,目前多数研究以成品白酒为对象,关于近红外光谱技术在白酒酿造酒醅发酵过程控制中的应用研究较为少见。

综合浓香型、白干型和芝麻香型等多种香型的白酒酒醅,分别采用偏最小二乘(partial least squares, PLS)法及最小二乘支持向量机(least squares-support vector machine, LS-SVM)两种算法观察白酒酒醅中酒精度、淀粉含量、水分、酸度等多个指标的近红外检测模型性能改善情况。建立白酒酒醅主要成分快速检测模型,并结合波段筛选方法,提高模型运算速度、稳定性、准确性以及模型预测能力,为白酒发酵过程质量控制提供参考。

1 材料与方法

1.1 材料

选用浓香型(290个)、白干型(275个)和芝麻香型(260个)酒醅样品共825个,由某白酒企业提供。

1.2 仪器与设备

傅里叶变换近红外光谱仪 瑞士步琪有限公司;光谱仪光源为卤钨灯,检测器为温控InGaAs,配有固体测量池。光谱范围为10 000~4 000 cm^{-1} ,分辨率为8 cm^{-1} ,扫描次数为32次;利用配套软件NIRWare Operator采集样品的近红外光谱信息,采用UnscramblerX10.3光谱分析软件(挪威CAMO公司)进行光谱预处理、PLS计算及潜在变量(latent variable, LV)提取,无信息变量消除(uninformative variables elimination, UVE)法、LS-SVM等程序均在MATLAB环境下运行。

1.3 方法

1.3.1 基础数据采集

为保证实验可靠性及模型准确性,对酒醅样品进行3次采样,采用漫反射方式扫描采集酒醅近红外光谱,

并对光谱进行平均。对获得的酒醅样品光谱进行预处理,剔除掉个别的异常光谱,最终剩余实验用光谱数量为816个。酒醅水分实测值测定:采用GB 5009.3—2010《食品中水分的测定》中的直接干燥法;淀粉测定:采用GB/T 5009.9—2008《食品中淀粉的测定》中的酶水解法;酸度测定:采用GB/T 5517—2010《粮油检验:粮食及制品酸度测定》。

1.3.2 样本集划分

本实验采用Kennard-Stone(K-S)法来进行样本集划分,其原理为:选择马氏距离最远的2个样本加入建模集,计算剩余的每个样本到建模集中每个已选样本的距离,找出最大和最小距离值样本,加入建模集,重复操作,直至建模集样本数目满足要求为止^[9]。本实验最终选择500个酒醅样本作为校正集,250个样品作为验证集,其余66个酒醅样品不参与建模分析,作为预测集样品,用于预测模型效果。其统计信息如表1所示。

表1 校正集与验证集数据统计结果
Table 1 Statistical results for calibration set and validation set

建模集	样品数	指标	最大值/%	最小值/%	平均值/%	标准差/%
校正集	500	酒精度	7.07	2.14	4.45	1.06
		淀粉	14.52	6.16	11.02	1.88
		水分	66.20	56.50	61.11	2.10
		酸度	5.60	1.50	3.77	0.92
验证集	250	酒精度	6.78	2.23	4.46	0.99
		淀粉	14.07	6.19	10.76	1.87
		水分	65.90	57.40	61.26	1.79
		酸度	5.60	1.70	3.78	0.83

1.3.3 UVE法

UVE是一种基于回归系数稳定性分析的变量选择方法,可以有效消除冗余信息变量,减少模型运算量,提高模型适用性^[10]。

1.3.4 PLS模型及LV

根据UVE优选的波段,以样品光谱吸光度作为模型输入,分别对酒醅的酒精度、淀粉、水分、酸度4个指标建立PLS模型。作为对比,同时基于全光谱数据作为模型输入,建立全光谱PLS模型。

在建立PLS模型时,选取最优LV个数提高有效信息率。类似于PCA中的主成分,LV的贡献率第一个最大,依次减小,若选择的建模用LV数目过少,则不能全面体现光谱特性,模型精度较差,预测能力降低;反之,若LV个数过多则会引入噪声,降低模型性能^[11]。

1.3.5 LS-SVM校正模型

LS-SVM^[12-15]是基于SVM扩展的一种新兴的非线性定量校正方法,把SVM中的不等式约束转化为等式约束,降低了计算的复杂性,从而提高了建模效率,可以极大改善近红外光谱定性、定量模型的预测能力,已广泛应用于模式识别、函数逼近、人脸检测和数据挖掘等领域。

采用PLS降维提取的潜在变量作为LS-SVM模型的输入变量,建立酒醑酒精度、淀粉、水分、酸度4个指标模型。

1.3.6 模型评价标准

模型的评价指标主要有决定系数 R^2 、预测集均方根误差(root mean square error of prediction, RMSEP)和相对分析误差(relative percent difference, RPD)。通常, R^2 越接近1,则模型相关性越好,预测效果好;RMSEP越小,表明模型预测精度越高,误差越小。RPD越大,模型分辨能力越强,准确度越高,当 $RPD > 3$ 时,认为模型效果良好^[16-18]。

2 结果与分析

2.1 光谱波段优选

UVE波段选择变量稳定性分析结果如图1所示。竖直实线左边是波长变量,右边为引入的系统噪音变量。虚线表示变量稳定性的上下阈值,处于两阈值之间的变量可认为是无关的信息量,超出阈值的部分为有用信息波长变量^[19]。

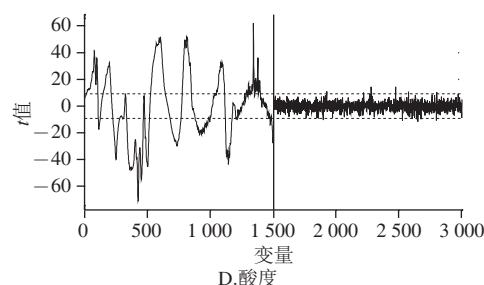
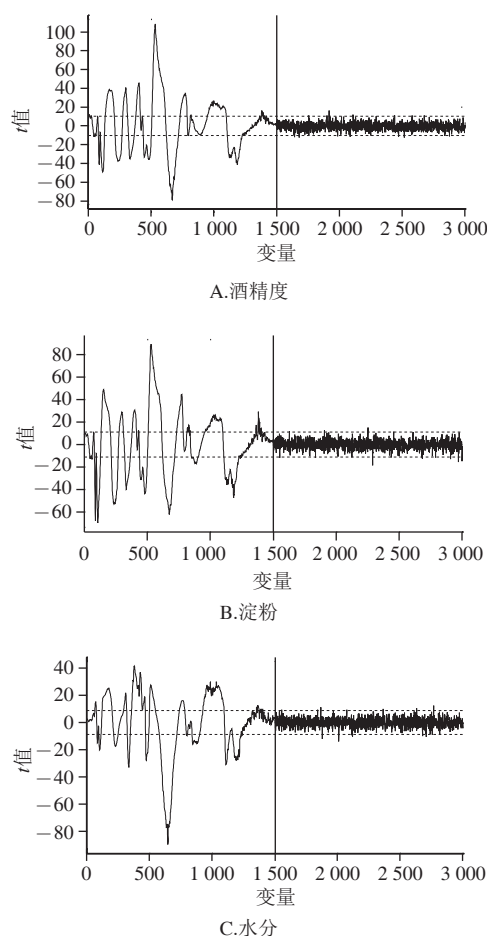


图1 UVE变量稳定性分析结果

Fig. 1 Results of UVE variable stability analysis

酒精度、淀粉、水分、酸度4个指标经UVE筛选后分别得到635、928、1 028、835个波长点数,相较于原始光谱1 501个波长点,变量数显著减少,并且保留了各指标的特征官能团倍频及组合频的所在波段位置,如9 960、6 897、6 711、5 155 cm^{-1} 等,有效地简化了模型,并为后续建模准确性提供了保障。

2.2 PLS法建立定标模型

在UVE波段筛选结果的基础上,分别对酒醑酒精度、淀粉、水分、酸度4个指标进行PLS建模,并以全光谱的PLS模型作为对照,结果如表2所示。经UVE筛选波段能有效提高模型性能, R^2 、RPD均有提高, RMSEP明显减小,模型的稳定性及准确度都有显著改善,说明UVE波段选择方法能够在减少建模运算用变量数、简化模型的同时,优化模型性能,提取有效信息。

表2 白酒酒醑4个指标PLS建模结果
Table 2 Result of PLS models for the four properties

指标	建模方法	波段点数	R^2	RMSEP/%	RPD
酒精度	ALL-PLS	1 501	0.940	0.260	3.807
	UVE-PLS	635	0.947	0.245	4.041
淀粉	ALL-PLS	1 501	0.926	0.498	3.591
	UVE-PLS	928	0.951	0.415	3.755
水分	ALL-PLS	1 501	0.895	0.648	2.762
	UVE-PLS	1 028	0.905	0.635	2.818
酸度	ALL-PLS	1 501	0.931	0.241	3.441
	UVE-PLS	835	0.951	0.214	3.878

注: ALL-PLS.全光谱变量的PLS建模结果; UVE-PLS.经UVE筛选后变量的PLS建模结果。

2.3 LV个数选择

建立PLS模型时,确定合理的LV个数对提高模型准确度和稳定性十分重要^[20],通过观察RMSEP值随LV个数的变化情况可以优选出最佳LV个数。如图2所示,随着LV个数的增加, RMSEP呈下降趋势,当达到最低点后平缓上升并趋于稳定。计算潜在变量数是对数据压缩优化的过程,为了提高模型光谱数据的有效信息率,选择RMSEP最低时对应的LV个数为最佳潜在变量数^[21],此时模型预测效果达到最佳,能充分实现光谱数据“少而精”的目的。因此,酒醑酒精度、淀粉、水分、酸度的LV个数分别选择为11、9、9、10个。

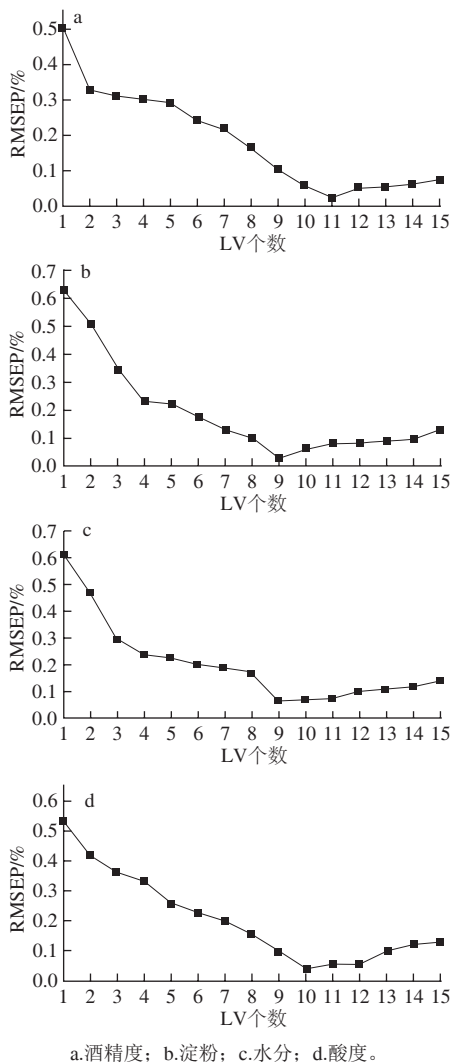


图2 潜在变量与RMSEP的关系

Fig. 2 Relationship between latent variables and RMSEP

此外，为证实所选LV个数数据可靠性，观察4个指标不同LV个数的累计贡献率。如表3所示，当酒精度、淀粉、水分、酸度的LV个数分别选择为11、9、9、10个时，累计贡献率均达到99.99%以上，几乎可以表达原始光谱的全部有用信息。

表3 4种指标潜在变量累计贡献率

Table 3 The explained variance of the first 11 LVs for the four properties by the best PLS models

LV个数	酒精度	淀粉	水分	酸度
1	72.72	85.18	76.67	85.27
2	95.90	96.14	94.00	93.18
3	98.40	98.36	95.54	96.11
4	99.38	99.48	99.02	99.00
5	99.84	99.85	99.59	99.21
6	99.91	99.88	99.91	99.89
7	99.93	99.90	99.94	99.95
8	99.95	99.95	99.95	99.96
9	99.96	99.99	99.99	99.97
10	99.96	99.99	99.99	99.99
11	99.99	99.99	99.99	99.99

2.4 基于LV建立LS-SVM定标模型

将PLS降维得到的LV分别作为酒精度、淀粉、水分、酸度LS-SVM模型的输入。本实验选取RBF核函数作为LS-SVM建模的核函数，以此减少训练过程的计算复杂性。此外，建立LS-SVM模型仍需两个重要调节参数： γ 和 σ^2 ，这两个参数对模型的学习能力和预测能力具有很大的决定性^[22]。其中， γ 是正规参数，取决于训练误差最小化和估计函数平滑的权衡，对提高模型的泛化性起重要作用。 σ^2 是核函数参数，控制模型回归误差，反映模型的灵敏度^[23-25]。目前，对于 γ 和 σ^2 参数的选择没有一定的模式，本实验采用10倍交叉验证的方法分析以确定优化参数。建模结果如表4所示。

表4 白酒酒醅4个指标LS-SVM建模结果

Table 4 Results of LS-SVM models for the four properties

指标	LVs	γ	σ^2	R^2	RMSEP/%	RPD
酒精度	11	193.83	231.32	0.970	0.186	5.322
淀粉	9	62.48	143.98	0.974	0.301	6.213
水分	9	67.16	159.43	0.948	0.464	3.850
酸度	10	109.12	157.38	0.976	0.145	5.724

由表4可见，LS-SVM模型效果较好，4个指标的RPD值均大于3，说明模型分析的分辨能力较高，同时 R^2 与RMSEP均有较好表现，说明LS-SVM模型的稳健型、准确度均达到较高水平。并且，以LV作为LS-SVM模型的输入，不仅保留了准确的光谱有效信息，而且显著地降低模型复杂度，提高了模型的运算速度与性能。

2.5 UVE-PLS模型与LS-SVM模型建模效果比较

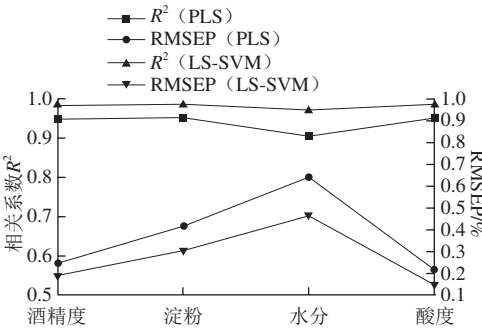


Fig. 3 Comparison of performance parameters between PLS models and LS-SVM models

对比PLS模型与LS-SVM模型效果，结果如图3所示，无论是 R^2 还是RMSEP，LS-SVM模型都取得了最优效果，模型的精度与性能都有较大改善。

2.6 UVE-PLS模型与LS-SVM模型预测效果比较

为比较模型的预测性能，分别将测试用66个未参与建模的酒醅样品带入已建立LS-SVM模型和PLS模型中进行验证，计算预测结果与理化检测值的相对偏差并绘制箱线图。由图4可知，与PLS模型相比，LS-SVM模型的

预测结果相对偏差极值差异更小,中位数接近0值,且样品相对偏差分布更为集中,说明LS-SVM模型预测结果与理化检验测量值更为接近,直观地表明了LS-SVM模型在实际检测应用中具有更强的预测性能。

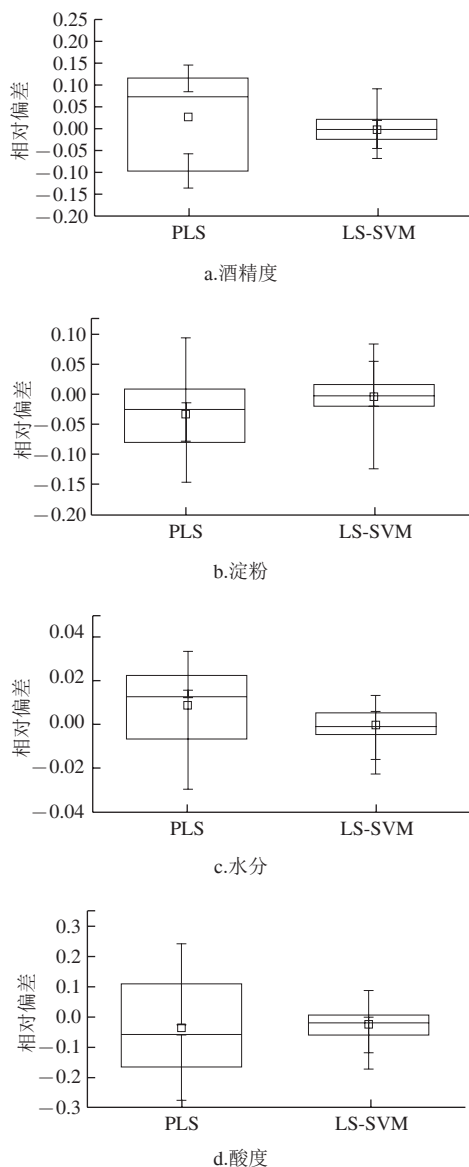


图4 PLS模型与LS-SVM模型预测结果对比图

Fig. 4 Comparison of predicted results obtained with PLS models and LS-SVM models

由于白酒酒醅本身是一种成分复杂的不均匀物质体系,体系中各组分的相互作用、噪声及基线漂移等因素会对光谱测量造成影响,因此需要更为稳健准确的多元校正方法来实现其主要成分的测量。LS-SVM作为一种基于SVM算法简化改良的非线性定量校正方法,与线性的PLS算法相比较,具有更强的对复杂样品的处理能力,可以使潜在变量与浓度之间的相关性增强^[9]。

此外,在建模时,通过潜在变量将PLS和LS-SVM有机结合了起来,在线性模型基础上补偿了酒醅体系的非

线性,因此,基于LV的LS-SVM建模方法更适用于白酒酒醅等复杂的非线性体系的准确分析,能够更明显提高预测精确度,降低误差率,简化模型复杂度。

3 结论

基于近红外光谱技术,对白酒酒醅的酒精度、淀粉、水分和酸度4个指标的定量分析进行了深入研究。利用UVE算法优选了有效特征波段,并通过潜在变量将PLS与LS-SVM两种算法结合了起来,建立了更为简化、高效的LS-SVM模型。

经比较发现:LS-SVM所建模型的酒精度、淀粉、水分和酸度的相关系数 R^2 、RMSEP以及RPD均表现良好,相比PLS模型有突出的优化效果。经待测样品进行验证,LS-SVM模型也具有更强的预测能力,预测结果更为准确。说明基于LV的LS-SVM建模方法能够明显提高预测精确度,降低误差率,简化模型复杂度并适用于白酒酒醅的实际生产应用。

考虑到白酒酒醅的复杂性以及LS-SVM算法优越的非线性校正特性,证明了在实际生产应用中,LS-SVM算法具有良好的可行性,且模型检测性能稳定,可用于白酒酿造过程酒醅的主要成分的检测。为白酒酒醅及相似的复杂体系物质的品质分析方法研究提供了新的思路,同时,为近红外光谱技术在白酒生产过程的应用提供了技术借鉴。

参考文献:

- [1] 黄治国,侯海波,罗惠波,等.浓香型白酒酒醅发酵过程中淀粉和还原糖的变化规律研究[J].中国酿造,2012,31(7):107-110. DOI:10.3969/j.issn.0254-5071.2012.07.028.
- [2] 郝建国,任晶婧.近红外光谱测定酒醅[J].酿酒科技,2011(5):106-107.
- [3] 周杨,刘杰,王纪元.中国酿酒行业中近红外光谱技术的应用进展[J].食品安全质量检测学报,2014(4):1100-1104.
- [4] 邵春甫,李长文,王珊,等.红外光谱技术在中国酿酒行业中的应用研究进展[J].中国酿造,2013,32(4):15-19. DOI:10.3969/j.issn.0254-5071.2013.04.004.
- [5] 魏赫楠,谭红,杨昌彪,等.近红外光谱技术在白酒行业的应用现状及展望[J].酿酒科技,2013(11):85-87.
- [6] 彭帮柱,龙明华,岳田利,等.傅立叶变换近红外光谱法检测白酒总酸和总酯[J].农业工程学报,2006(12):216-219. DOI:10.3321/j.issn.1002-6819.2006.12.045.
- [7] 陈妍,胡慧,汪凤祖.近红外光谱法快速分析白酒中的关键指标[J].酿酒科技,2010(11):90-94.
- [8] 杨国强,张淑娟,赵艳茹.基于近红外透射光谱的汾阳王酒快速鉴别[J].农业机械学报,2013(增刊1):189-193. DOI:10.6041/j.issn.1000-1298.2013.S1.034.
- [9] 褚小立.化学计量学方法与分子光谱分析技术[M].北京:化学工业出版社,2011:61-83.
- [10] 李倩倩.无信息变量消除法在三种谱学方法中的定量分析研究[D].北京:中国农业大学,2014.
- [11] 王莉,何勇,刘飞,等.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[J].红外与毫米波学报,2008,27(1):51-55. DOI:10.3321/j.issn.1001-9014.2008.01.012.

- [12] 张德虎, 田海清, 刘超, 等. 可见近红外光谱检测河套蜜瓜糖度和硬度研究: 基于LS-SVM[J]. 农机化研究, 2014(2): 10-14.
- [13] 安欣, 徐硕, 张录达, 等. 多因变量LS-SVM回归算法及其在近红外光谱定量分析中的应用[J]. 光谱学与光谱分析, 2009, 29(1): 127-130. DOI:10.3964/j.issn.1000-0593(2009)01-0127-04.
- [14] 孙俊, 毛罕平, 羊一清, 等. 基于GA-LS-SVM的水稻叶片含氮率预测[J]. 江苏大学学报(自然科学版), 2010, 31(1): 6-10. DOI:10.3969/j.issn.1671-7775.2010.01.002.
- [15] 刘燕德, 周延睿. 基于GA-LS-SVM的苹果糖度近红外光谱检测[J]. 西北农林科技大学学报(自然科学版), 2013, 41(7): 229-234.
- [16] 高珏, 王从庆. 基于LS-SVM的苹果近红外光谱回归模型的研究[J]. 计算机测量与控制, 2011, 19(1): 176-191.
- [17] ZHENG Hong, LU Hongfei. A least-squares support vector machine (LS-SVM) based on fractal analysis and CIELab parameters for the detection of browning degree on mango (*Mangifera indica* L.)[J]. Computers and Electronics in Agriculture, 2012, 83: 47-51. DOI:10.1016/j.compag.2012.01.012.
- [18] GHAEDI M, GHAEDI A M, HOSSAINPOUR M, et al. Least square-support vector (LS-SVM) method for modeling of methylene blue dye adsorption using copper oxide loaded on activated carbon: kinetic and isotherm study[J]. Journal of Industrial and Engineering Chemistry, 2014, 20(4): 1641-1649. DOI:10.1016/j.jiec.2013.08.011.
- [19] 周林峰. 黄酒主要品质的近红外光谱检测模型建立与转移的研究[D]. 杭州: 中国计量学院, 2013: 1-82.
- [20] 杨晓雷. LSSVM优化方法的研究[D]. 南宁: 广西大学, 2012.
- [21] 王义峰, 张中卫. LSSVM在酒类近红外光谱检测中的应用[J]. 信息技术, 2009(11): 90-92. DOI:10.3969/j.issn.1009-2552.2009.11.026.
- [22] NIAZI A, SHARIFI S, AMJADI E. Least-squares support vector machines for simultaneous voltammetric determination of lead and tin: a comparison between LS-SVM and PLS in voltammetric data[J]. Journal of Electroanalytical Chemistry, 2008, 623(1): 86-92. DOI:10.1016/j.jelechem.2008.06.021.
- [23] 孙鑫. 基于LS-SVM建立发酵过程动态模型及其参数优化[D]. 北京: 北京工业大学, 2013.
- [24] SHAHLAEI M, FASSIHI A, SAGHAIE L. Application of PC-ANN and PC-LS-SVM in QSAR of CCR1 antagonist compounds: a comparative study[J]. European Journal of Medicinal Chemistry, 2010, 45(4): 1572-1582. DOI:10.1016/j.ejmech.2009.12.066.
- [25] SUN Tong, LIN Hongjian, XU Huirong, et al. Effect of fruit moving speed on predicting soluble solids content of 'Cuiguan' pears (*Pomaceae pyrifolia* Nakai cv. Cuiguan) using PLS and LS-SVM regression[J]. Postharvest Biology and Technology, 2009, 51(1): 86-90.