DOI:10.13652/j.issn.1003-5788.2021.06.014

CARS-SVM 预测哈密瓜可溶性固形物含量

Prediction of soluble solids in Hami melon by CARS-SVM

郭 阳[1] 郭俊先[1] 史 勇[1] 李雪莲[1]

GUO Yang[1] *GUO Jun-xian*[1] *SHI Yong*[1] *LI Xue-lian*[1]

刘彦岑[1] 黄 华[2] 李泽平[1]

LIU Yan-cen[1] *HUANG Hua*[2] *LI Ze-ping*[1]

(1. 新疆农业大学机电工程学院,新疆 乌鲁木齐 830052;2. 新疆农业大学数理学院,新疆 乌鲁木齐 830052)

(1. *College of Electrical and Mechanical Engineering, Xinjiang Agricultural University, Urumqi, Xinjiang* 830052, *China*;2. *College of Mathematics and Physics, Xinjiang Agricultural University, Urumqi, Xinjiang* 830052, *China*)

摘要:采用近红外光谱技术结合数据降维的方法,建立了哈密瓜可溶性固形物含量的预测模型:对比多种光谱预处理方法,确定二阶求导用于处理原始光谱;经预处理的光谱数据分别结合特征选择竞争性自适应重加权采样法(CARS)、蒙特卡罗无信息变量消除法(MC-UVE)提取特征波长,以及利用主成分分析进行降维;再使用特征选择和特征提取的光谱数据作为模型的输入变量,建立哈密瓜可溶性固形物含量预测模型。结果显示,CARS+SVM 建立的预测模型最优,模型的校正集相关系数为 0.981 4,预测集相关系数为 0.900 2,模型能够准确预测哈密瓜可溶性固形物含量。

关键词:哈密瓜;CARS;支持向量机;可溶性固形物;无损检测

Abstract: Soluble solid content is one of the important indexes for the internal quality analysis of Hami melon. In this study, the prediction model of soluble solid content of Hami melon was established by using near infrared spectroscopy combined with data dimension reduction method. Compared with a variety of spectral preprocessing methods, the second-order derivative was used to process the original spectrum; the preprocessed spectral data were combined with CARS and MC-UVE to extract the characteristic wavelength, and the principal component analysis was used to reduce the dimension; Finally, the spectral data of feature selection and feature extraction were used as the input variables of support vector machine to establish the prediction model of soluble solid content of Hami melon. The results showed that the prediction model established by CARS + SVM was the best, with the correlation coefficient of the model calibration of 0.981 4, and the correlation coefficient of the prediction set was 0.900 2. This model could be used to accurately predict the soluble solids of Hami melon.

Keywords: Hami melon; CARS; support vector machines; soluble solids; nondestructive testing

哈密瓜是新疆的特色农产品之一,其果肉鲜嫩,爽脆可口,深受广大消费者青睐,其中,可溶性糖含量(SSC)与其口感有很大关系。哈密瓜在生长过程中受田间环境、植株生长形态、植株冠层营养等因素影响,造成可溶性糖积累分布不均匀、含量低。可溶性固形物主要是指可溶性糖类,其是衡量哈密瓜品质好坏的重要指标,同时传统的哈密瓜中可溶性固形物含量的检测方法的准确率高,但需破坏样本。近年来,近红外光谱检测技术因其具有快速、准确及多组分同时检测等特点,已被应用于椰汁品质[1]、鸡蛋新鲜度[2]、肉类品质[3-4]、石榴糖度[5]、梨的可溶性固形物[6-7]、液态奶三聚氰胺[8]等农业生产检测方面。哈密瓜可溶性固形物检测方面,张德虎等[9]采用BiPLS 光谱波长筛选方法提取哈密瓜糖度特征波长,优化后的预测模型校正集和预测集的 RMSE 分别为0.996 1 和1.18;Greensill 等[10]利用 4 种光电二极管陈列近红外光谱仪结合不同光谱预处理方法预测了甜瓜的 SSC;Guthrie[11] 建立了移动窗口偏最小二乘(MWPLS)甜瓜总糖含量的预测模型,其预测集均方根误差和标准偏差分别为 1.1 和 0.04;毕智健等[12]比较了

基金项目:新疆维吾尔自治区教育厅自然科学重点项目(编号:XJEDU2020I009);国家自然科学基金面上项目(编号:61367001)

作者简介:郭阳,男,新疆农业大学在读硕士研究生。

通信作者:李雪莲(1967—),女,新疆农业大学副教授,硕士。E-mail:250585997@qq.com

收稿日期:2021-01-06

哈密瓜样品中可见近红外光谱数据的预处理方法的预测效果；马本学等[13]利用高光谱成像技术比较了偏最小二乘(PLS)、逐步多元线性回归(SMLR)和特征提取(PCR)3 种建模方法对带皮和去皮哈密瓜糖度的检测效果。水果可溶性固形物无损检测中，高升等[14]将光谱信息和图像特征信息进行有机融合，融合后的模型精度较单一的图像与光谱模型都有较大提升，其红提糖度最优的预测模型为最小二乘支持向量机(LS-SVM)，模型的校正集和预测集的相关系数分别为 0.954，0.952；Dong 等[15]研究苹果中可溶性固形物含量无损测定时，分别建立了偏最小二乘回归(PLSR)、LS-SVM、极限学习机(ELM)模型，最优预测模型为 LS-SVM 模型，其模型预测集相关系数为 0.878；杨晓玉等[16]利用特征选择竞争性自适应重加权采样法(CARS)、无信息变量消除法(UVE)、连续投影算法(SPA)对灵武长枣的原始光谱特征波长进行提取，并将提取出的特征波长作为输入变量建立了 PLSR 和 LS-SVM 的灵武长枣维生素 C 含量预测模型，最优模型为无信息变量消除法＋支持向量机(UVE-SVM)模型，其校正集和预测集的决定系数分别为 0.847 1，0.714 9。

综上，非线性模型在水果理化性质无损测定中应用非常广泛，而目前对哈密瓜可溶性固形物建立的定量分析模型多是 PLS、PCR 等线性模型，有关非线性模型下结合特征选择和特征提取对哈密瓜可溶性固形物定量分析模型进行优化的研究尚未见报道；同时哈密瓜成熟采收时，其是否可以采摘主要通过果皮表面颜色的变化以及哈密瓜可溶性固形物含量来进行判断，而可溶性固形物含量是判断哈密瓜是否可以采摘的关键指标。研究拟采用特征波长选择和特征提取 3 种算法对预处理后的光谱数据进行数据降维，同时应用非线性的支持向量机(SVM)、极限学习机(ELM)、最小二乘支持向量机(LS-SVM)算法结合 3 种优化算法建立预测模型，并比较所有模型的优劣，最终选取最优的模型作为哈密瓜可溶性固形物含量的无损检测模型，旨在为哈密瓜可溶性固形物含量的无损检测技术的发展提供依据。

1　材料与方法

1.1　试验地点

选取新疆哈密地区巴里坤县三塘湖镇中湖村为试验地点，该地形呈西高东低之势，气候干燥酷热多风，属典型的大陆性气候。年平均大风日 116.6 d，年日照时间 3 373.4 h，有效积温 3 440 ℃，无霜期 169 d。极端最高气温 40.3 ℃、最低气温－28.5 ℃，年平均气温 8 ℃。年平均降水量 34.4 mm，蒸发量 3 790 mm。试验田位置为东经 1 200.144°，北纬 3 020.626°，土壤的理化性质见表 1。

甜瓜品种：金华蜜 25 号，俗称“新 86”，晚熟品种，生育期 100 d，单瓜重 3.5 kg 左右。

1.2　试验样本与数据采集

待哈密瓜成熟后，从试验田一次性随机采摘 144 个无病虫害和损伤的哈密瓜作为试验样本，标记编号运回实验室。将所有样本在室温下放置 24 h 后开始试验，并尽可能地快速完成试验。光谱数据的采集使用美国海洋光学公司的 maya2000 微型光纤光谱仪，光谱测定范围为 200～1 100 nm，光谱采样间隔 0.2 s。数据采集前，光谱设备预热 30 min，主要工作参数设置包括积分时间 7 200 s，扫描次数 10，平滑点数 3。光谱采集位置选取每个样本赤道中间部位，每隔 120°采集一次，每个样本采集 3 个光谱数据，取平均值作为样本的原始光谱数据。

可溶性固形物含量测定使用 ATAGO PR-101 型数字折光仪(日本爱拓)，测量范围为 0～45 °Brix，测量精度为 0.1 °Brix。切取光谱采集处对应的内缘果肉并榨汁，将汁液滴至折光仪的测量区域，重复测定读数直至 3 次读数稳定，记录为当前样本的可溶性固形物含量。

1.3　原始光谱数据预处理

采集原始光谱数据过程中除了包含样品本身的特征信息外，还掺杂一些影响模型准确性的无用信息，如样品表面散射带来的光程变化所产生的光谱误差、光谱的散射影响、光谱数据中的噪声、以及设备自身造成的干扰。因此，分别用均值中心化(MC)、多元散射校正(MSC)、标准正态变量交化(SNVT)、SavitZky-Golay 卷积平滑法(SG-平滑)、二阶求导、归一化、移动平均平滑(MA)对原始光谱数据进行预处理。

1.4　数据降维

光谱数据具有数据量大、维度高、数据共线性等缺点，且未经过数据降维处理的光谱数据直接作为模型的输入变量，会影响模型的精确度和稳定性，同时大量的数据计算时会影响可溶性固形物含量的无损检测效率，不利于后期在线检测。分别使用特征提取主成分分析(PCA)[17]、特征选择竞争性自适应重加权采样法(CARS)和蒙特卡罗无信息变量消除法(MC-UVE)来实现数据降维。

表 1　大田哈密瓜的土壤理化性质

Table 1　Soil physical and chemical properties of Hami melon in field

土壤类型	pH 值	碱解氮质量分数/(mg·kg^{-1})	有效磷质量分数/(mg·kg^{-1})	速效钾质量分数/(mg·kg^{-1})
沙壤土	6.8	126.42	267.52	41.91

1.5　预测模型与模型评价

支持向量机(SVM)[18]可有效克服神经网络收敛慢、预测能力差的缺点,针对小样本量的预测模型建立具有其独特的优势。SVM 回归预测模型是通过非线性变化转换为某个高维空间的线性问题,并在此空间进行线性求解,实现回归预测模型的建立。极限学习机(ELM)[19]相比于前馈神经网络等在运算过程中不需设定大量的参数,且运算速度更快,只需按照实际情况选择合适的激励函数(AF),在算法运行过程中随机产生网络的输入权值及隐含层单元偏置,且不需要调整,比较容易实现。因此,ELM 具有学习速度快,高强的泛化能力促使模型只有唯一的最优解等特点。最小二乘支持向量机(LS-SVM)[20]是在 SVM 的基础上进行改进的算法,可以完成线性和非线性的多元预测模型的建立,具有降低计算复杂度、提高模型泛化能力、并能使训练集数据在高维特征空间进行学习等优点。

预测模型的评价指标为相关系数(R)和均方根误差(RMSE)。其中校正集均方根误差为 RMSEC、预测集均方根误差为 RMSEP;校正集相关系数为 R_c、预测集相关系数为 R_p,预测模型的相关系数越大表示相关性越高;预测模型的 RMSEP 越小,模型的预测效果越好。

$$R=\frac{\sum_{i=1}^{N}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}},\quad(1)$$

$$S_{RME}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2},\quad(2)$$

式中:

R——相关系数;

S_{RME}——均方根误差;

x_i——样本;

$\bar{x}$、$\bar{y}$——相应变量的平均值;

y_i——实际值;

$\hat{y}$——预测值;

N——样本数。

以上光谱数据处理和定量预测模型的建立均使用 Matlab2017b 软件完成(美国,MathWorks),采用 Matlab2017b 和 OriginPro 8 软件绘图。

2　结果与分析

2.1　样本划分

考虑到光谱理化值共生距离法(SPXY)算法能同时研究光谱特征与样本理化性质的能力,使用该划分法按 3∶1 将原始数据划分为样本校正集和预测集,其测定结果见表 2。

由表 2 可知,哈密瓜可溶性固形物含量的最大值和最小值都被划分到了校正集中,并且划分到预测集的数据值均在校正集区间内,表明利用 SPXY 划分的样本集的分布合理,所建的预测模型也能产生较好的结果。

2.2　光谱预处理

将原始光谱和 7 种预处理后的光谱变量分别结合 PLS 建立预测模型,通过对比多个 PLS 的预测模型的精度,选择最优模型的光谱预处理方法作为哈密瓜光谱变量的预处理方法,建模结果见表 3。

由表 3 可知,最优光谱预处理方法为二阶求导,这是因为利用此种预处理方法处理光谱原始数据可以提高光谱分辨率,减小噪声并提高信噪比,模型的预测精度会有所提高。从所有全波长建模角度来看,相关系数基本为 0.60～0.75,表明全波段作为模型的输入变量建立的预测模型效果不是很理想,说明全波长的变量中存在冗余信息和数据共线等问题,导致模型的准确性不高,故需对全波长的光谱信息进行变量选择和变量提取。

2.3　基于 CARS 数据降维

图 1 为 CARS 算法筛选特征波长变量过程。由图 1 可知,特征波长变量筛选过程中,随着迭代次数的增加,波长变量的总数减少,直至选取最优迭代次数为止。迭

表 2 哈密瓜的可溶性固形物含量

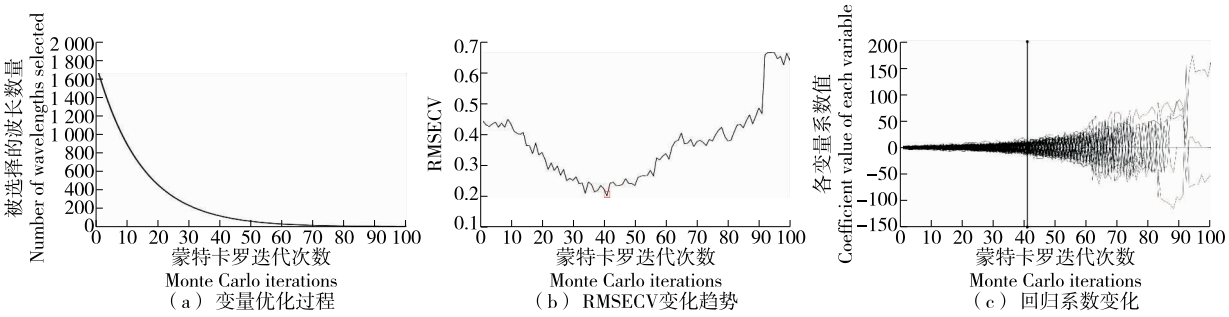
Table 2　Mass fraction of Soluble solids in Hami melon

样本	样本数	可溶性固形物含量/°Brix		
		平均值	最大值	最小值
校正集	108	13.9	15.9	12.5
预测集	36	13.8	15.8	11.9

表 3　不同光谱预处理结合 PLS 哈密瓜可溶性固形物的预测效果

Table 3　Prediction effect of different spectral pretreatment combined with PLS on soluble solids in Hami melon

预处理方法	主成分 PC	R_c	RMSEC	R_p	RMSEP	剩余预测残差 RPD
原始光谱	14	0.816 8	0.351 0	0.721 6	0.415 6	1.898 9
SNVT	14	0.734 0	0.435 8	0.720 0	0.374 3	1.892 1
SG-平滑	14	0.813 7	0.353 9	0.721 1	0.416 0	1.897 7
MSC	13	0.756 4	0.414 2	0.607 3	0.452 3	1.609 4
二阶求导	5	0.852 4	0.317 6	0.751 5	0.348 3	2.017 2
MA	14	0.816 8	0.351 0	0.721 6	0.415 6	1.899 0
归一化	13	0.743 9	0.426 8	0.675 6	0.418 8	1.767 1



（a）变量优化过程　（b）RMSECV变化趋势　（c）回归系数变化

图 1　CARS 筛选光谱变量过程

Figure 1　CARS screening spectral variable process

代次数最优时，RMSECV 越小迭代次数越好，当 RMSECV 为 0.199 7 时，对应的最优迭代次数为 41。因此，确定从原始 1 600 个波长中筛选的特征波长变量为 110 个。

2.4　基于 MC-UVE 数据降维

当 $N=1\ 000$ 时，波长变量的稳定值如图 2 所示。

MC-UVE 算法仅给出了光谱变量的稳定值，未给出最终筛选的光谱变量数作为后续模型建立的输入变量。因此为了剔除多余的变量，减少变量的共线性并提高模型的泛化能力，通过前向变量选择程序选择光谱变量。经 MC-UVE 算法筛选的光谱变量曲线如图 3 所示。由图 3 可知，RMSEP 的最小值为 0.307 9，对应的组数为 13，因此筛选前 13 组作为最佳变量，即共有 130 个特征波长变量。

2.5　基于 PCA 数据降维

主成分分析结果如图4所示，其前15个主成分累计贡献率达 95%以上，可以很好地表征原光谱数据的特征，故使用前 15 个主成分得分值作为模型的输入变量。

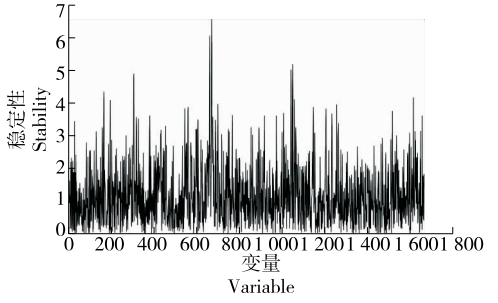


图 2　光谱变量稳定图

Figure 2　Spectral stability diagram

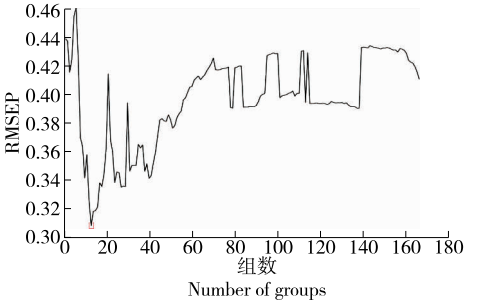


图 3　MC-UVE 筛选光谱变量曲线

Figure 3　MC-UVE filter spectral variable curve

2.6　哈密瓜可溶性固形物含量建模预测分析

3 种数据降维方式结合 SVM、ELM、LS-SVM 的哈密瓜可溶性固形物预测分析结果如表 4～表 6 所示。由表 4～表 6 可知，主成分分析下的建模效果都不是很理想，相关系数仅有 0.79，0.77，0.86，可能是主成分分析只降低了光谱数据的维度，并未减少光谱的变量数；相比而言，特征选择下的数据降维效果优于主成分分析，且二者优于全波长下的 PLS 预测模型。最优的预测模型为二阶求导＋CARS＋SVM，其校正集相关系数为 0.981 4，预测集相关系数为 0.900 2；表明该模型可以准确、快速地预测哈密瓜中可溶性固形物含量。3 种数据降维方法结合 ELM 建立的模型预测精度都不是很理想，可能是因为 ELM 属于神经网络模型的一种，且神经网络都有收敛慢、预测能力差的缺点，故相比于SVM、LS-SVM的建模

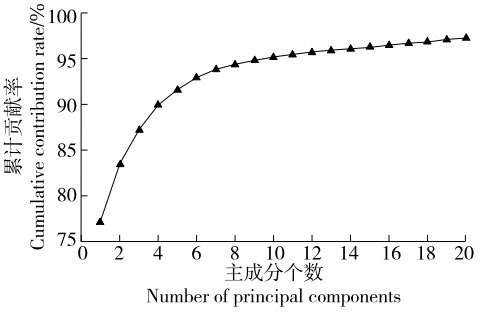


图 4　前 20 个主成分的累计贡献率

Figure 4　Cumulative contribution of the first 20 principal components

表 4　数据降维下结合 SVM 的建模预测效果

Table 4　Modeling and forecasting effect of SVM combined with dimensionality reduction

处理方法	校正集		预测集	
	R_c	RMSEC	R_p	RMSEC
MC-UVE	0.947 6	0.002 4	0.801 2	0.007 0
CARS	0.981 4	0.000 8	0.900 2	0.004 7
PCA	0.792 3	0.009 2	0.676 4	0.011 9

表 5　数据降维下结合 ELM 的建模预测效果

Table 5　Forecasting effect of modeling based on Dimension Reduction and ELM

处理方法	校正集		预测集	
	R_c	RMSEC	R_p	RMSEC
MC-UVE	0.750 9	0.165 1	0.746 4	0.232 3
CARS	0.722 5	0.189 5	0.671 3	0.194 2
PCA	0.775 9	0.144 8	0.736 2	0.182 1

表 6　数据降维下结合 LSSVM 的建模预测效果

Table 6　Modeling and forecasting effect of LSSVM combined with data dimension reduction

处理方法	校正集		预测集	
	R_c	RMSEC	R_p	RMSEC
MC-UVE	0.956 6	0.003 4	0.873 2	0.008 7
CARS	0.906 3	0.006 8	0.878 9	0.008 8
PCA	0.861 9	0.005 7	0.745 1	0.011 0

效果,ELM 的价值是最低的。同时,证明 CARS 算法在定量预测建模中可以对光谱变量中与理化性质相关性高的变量进行准确提取。

3　结论

通过对原始光谱以及经均值中心化、多元散射校正、标准正态变量正交化、SavitZky-Golay 卷积平滑法、二阶求导、归一化、移动平均平滑预处理获得的光谱数据建立全波长的 PLS 预测模型并分析,得出最优的预处理方式为二阶求导法;在二阶求导的基础上再分别使用两种特征选择方法(特征选择竞争性自适应重加权采样法和蒙特卡罗无信息变量消除法)和特征提取主成分分析法对光谱作进一步处理;基于处理后光谱数据分别结合非线性模型支持向量机、极限学习机和最小二乘支持向量机建立定量分析模型。结果表明,最优的预测模型为二阶求导+特征选择竞争性自适应重加权采样法+支持向量机,模型的校正集和预测集相关系数分别为 0.981 4,0.900 2,模型预测效果得到了提升;表明光谱数据与理化性质之间也存在非线性的相关信息,且非线性模型可以准确地预测哈密瓜可溶性固形物含量,实现哈密瓜内部品质的无损检测,同时也为田间便携式哈密瓜是否成熟判别设备的研制提供了新思路。后续应考虑如何将特征选择与特征提取进行有效融合,结合两者的优点实现光谱数据的压缩且保证关键信息不会被丢失,以期建立准确且稳定的定量分析模型。

参考文献

[1] 连媛媛, 熊乾威, 杨木莎, 等. 基于近红外光谱技术快速检测椰汁品质[J]. 食品工业科技, 2019, 40(12): 235-240.

[2] 赵杰文, 毕夏坤, 林颢, 等. 鸡蛋新鲜度的可见—近红外透射光谱快速识别[J]. 激光与光电子学进展, 2013, 50(5): 213-220.

[3] 何鸿举, 王魏, 李波, 等. 近红外高光谱快速无接触评估冷鲜猪肉脂质氧化[J]. 食品与机械, 2020, 36(8): 117-122.

[4] 黄伟, 杨秀娟, 曹志勇, 等. 近红外反射光谱快速检测滇南小耳猪肉中水分、粗脂肪及粗蛋白含量的研究[J]. 中国畜牧杂志, 2015, 51(7): 73-77.

[5] 刘燕德, 张雨, 徐海, 等. 基于近红外光谱检测不同产地石榴的糖度[J]. 激光与光电子学进展, 2020, 57(1): 253-259.

[6] 路敏. 基于近红外光谱的梨的可溶性固形物含量的无损检测[D]. 兰州: 兰州大学, 2019: 11-18.

[7] 孙通, 江水泉. 基于可见/近红外光谱和变量优选的南水梨糖度在线检测[J]. 食品与机械, 2016, 32(3): 69-72.

[8] 程文宇, 管骁, 刘静. 近红外光谱技术检测液态奶中微量三聚氰胺的可行性研究[J]. 食品与机械, 2015, 31(1): 71-74, 81.

[9] 张德虎, 田海清, 武士钥, 等. 河套蜜瓜糖度可见近红外光谱特征波长提取方法研究[J]. 光谱学与光谱分析, 2015, 35(9): 2 505-2 509.

[10] GREENSILL C V, WOLFS P J, SPIEGELMAN C H, et al. Calibration transfer between PDA-Based NIR spectrometers in the NIR assessment of melon soluble solids content[J]. Applied Spectroscopy, 2001, 55(5): 647-653.

[11] GUTHRIE J A. NIR model development and robustness in prediction of melon fruit total soluble solids[J]. Australian Journal of Agricultural Research, 2006, 57(4): 411-418.

[12] 毕智健. 哈密瓜糖度可见近红外光谱在线检测系统设计研究[D]. 石河子: 石河子大学, 2017: 6-10.

[13] 马本学, 肖文东, 祁想想, 等. 基于漫反射高光谱成像技术的哈密瓜糖度无损检测研究[J]. 光谱学与光谱分析, 2012, 32(11): 3 093-3 097.

[14] 高升, 王巧华. 基于高光谱图像信息融合的红提糖度无损检测[J]. 发光学报, 2019, 40(12): 1 575-1 584.

[15] DONG Jin-lei, GUO Wen-chuan, WANG Zhuan-wei, et al. Nondestructive determination of soluble solids content of 'Fuji' apple produced in different areas and bagged with different materials during ripening[J]. Food Analetical Methods, 2016, 9(5): 1 087-1 095.

[16] 杨晓玉, 刘贵珊, 丁佳兴, 等. 灵武长枣 V_C 含量的高光谱快速检测研究[J]. 光谱学与光谱分析, 2019, 39(1): 230-234.

[17] 孟庆龙, 尚静, 黄人帅, 等. 基于主成分回归的苹果可溶性固形物含量预测模型[J]. 保鲜与加工, 2020, 20(5): 185-189.

[18] 王小燕, 王锡昌, 刘源, 等. 基于 SVM 算法的近红外光谱技术在鱼糜水分和蛋白质检测中的应用[J]. 光谱学与光谱分析, 2012, 32(9): 2 418-2 421.

[19] 朱哲燕, 刘飞, 张初, 等. 基于中红外光谱技术的香菇蛋白质含量测定[J]. 光谱学与光谱分析, 2014, 34(7): 1 844-1 848.

[20] 赵杰文, 林颢. 食品、农产品检测中的数据处理与分析方法[M]. 北京: 科学出版社, 2012: 92-97.