DOI:10.13652/j.spjx.1003.5788.2022.80560 [文章编号] 1003-5788(2023)10-0123-07

多毛刺小样本高光谱数据下鹰嘴蜜桃含水率的预估

Prediction of moisture content of hummus peach based on multi-burr hyperspectral data

高艾迪[1] 乔奉璋[1] 朱文轩[1] 钟小品[1] 邓元龙[1,2]

GAO Aidi[1] *QIAO Fengzhang*[1] *ZHU Wenxuan*[1] *ZHONG Xiaopin*[1] *DENG Yuanlong*[1,2]

(1. 深圳大学机电与控制工程学院,广东 深圳 510086;2. 深圳技师学院,广东 深圳 518116)

(1. *Shenzhen University*, *College of Mechatronics & Control Engineering*, *Shenzhen*, *Guangdong* 510086, *China*; 2. *Shenzhen Institute of Technology*, *Shenzhen*, *Guangdong* 518116, *China*)

摘要:目的:提出并解决鹰嘴蜜桃高光谱测量数据多毛刺和小样本问题。**方法:**基于高光谱成像技术,使用图像处理方法识别高光谱图像中鹰嘴蜜桃所在区域,计算该区域内的光谱图像从而得到平均光谱反射率数据,形成高光谱曲线图像。对于存在抖动和毛刺的高光谱图像数据,比较多项式平滑算法(SG)、多元散射矫正算法(MSC)、标准正态变量算法(SNV)、一阶导数算子(D1)、二阶导数算子(D2)等数据预处理方法对模型预测精度的影响;针对数据维度高且样本量少的特点,使用主成分分析算法(PCA)对数据进行降维,再对降维后的数据应用马氏距离测度方法(MD)进行异常值剔除;最终利用 Kennard-Stone 算法(KS)划分出训练集和测试集,并选取小样本场景下表现较好的偏最小二乘回归(PLSR)模型对鹰嘴蜜桃的含水率进行估计和分析。**结果:**SG-PCA-MD-KS-PLSR 模型在高光谱曲线存在抖动和毛刺情况时对鹰嘴蜜桃含水率估计的效果最好,训练集下决定系数(R^2)达到 0.928,均方根误差(RMSE)为 0.008 4,测试集下 R^2 达到 0.926,RMSE 为 0.009 2。在进一步对鹰嘴蜜桃以含水率为指标进行分级试验时,该模型的预测结果可以较好地对鹰嘴蜜桃含水状况进行分级,训练集下分级正确率为 0.956,测试集下分级正确率为 0.923。**结论:**利用高光谱成像技术建立 SG-PCA-MD-KS-PLSR 模型,在高光谱样本数较小且存在毛刺的情况下,仍能对鹰嘴蜜桃含水率进行无损估计。

关键词:高光谱;鹰嘴蜜桃;含水率;无损检测

Abstract: Objective: To propose a new solution to overcome the two challenges of data with spikes and small sample sizes in nectarine hyperspectral measurement. **Methods:** Based on hyperspectral imaging technology, image processing methods were used to identify the area of nectarines in the hyperspectral image, and the spectral reflectance data of the area was calculated to form a hyperspectral curve image. For hyperspectral image data with spikes and noise, compared the effects of several data preprocessing methods, including polynomial smoothing algorithm (SG), multivariate scatter correction algorithm (MSC), standard normal variate algorithm (SNV), first-order derivative operator (D1), and second-order derivative operator (D2) on model prediction accuracy. To address the high-dimensional and small sample size characteristics of the data, the principal component analysis algorithm (PCA) was used for dimensionality reduction, followed by outlier removal using the Mahalanobis distance measure method (MD). Finally, the Kennard-Stone algorithm (KS) was used to divide the data into training and testing sets, and the partial least squares regression (PLSR) model, which performed well in the small sample scenario, was selected for estimation and analysis of nectarine water content. **Results:** The SG-PCA-MD-KS-PLSR model performed best for estimating nectarine water content when there were spikes and noise in the hyperspectral curve. The coefficient of determination (R^2) was 0.928, and the root mean square error (RMSE) was 0.008 4 on the training set. The R^2 was 0.926, and the RMSE was 0.009 2 on the testing set. In further experiments grading nectarines based on their water content, the model's predictions showed good performance. The accuracy rate of

基金项目:国家自然科学基金面上项目(编号:62171288);广东省乡村振兴战略专项资金(农村特派员)(编号:163-2019-XMZC-0009-03-0059)

作者简介:高艾迪,男,深圳大学在读本科生。

通信作者:邓元龙(1971—),男,深圳技师学院教授,博士。E-mail:dengyl@szu.edu.cn

收稿日期:2022-07-15 **改回日期:**2023-07-11

grading was 0.956 for the training set and 0.923 for the testing set. **Conclusion**: By using hyperspectral imaging technology and establishing the SG-PCA-MD-KS-PLSR model, non-destructive estimation of nectarine water content and grading of nectarine water content can be achieved in scenarios with small hyperspectral sample sizes and noise.

Keywords: hyperspectral; chick peach; moisture content; nondestructive detection

随着中国经济的快速发展，消费者对高品质水果的需求日益增加，将高精度果品质量评价技术应用于高端果品分选的需求日益增强。检测水果品质指标并以此为依据进行水果的分级具有现实意义，它能够为生产、收购、流通、零售等各个环节提供重要的数据支持并进一步提高农业生产现代化水平，但目前缺少有效手段对果实的品质指标进行高准确率的无损鉴定。含水率是衡量水果品质的指标之一。传统的含水率测量方法为烘干法，该方法的检测精度高，但是烘干后的果实无法正常销售，因此这种方法无法对生产过程中每个果实进行检测。同时整个烘干和测量流程均需要人工参与且耗时较长，无法基于此方法进行自动化水果分级。

近年来，高光谱成像技术因其检测速度快，可以进行无损检测等特点，在果实无损检测领域表现出强大潜力[1-3]。廉孟茹等[4]采用附加散射校正算法（MSC）对玉米的高光谱图像数据进行预处理，使用竞争性自适应加权算法（CARS）提取特征波长，最终建立偏最小二乘回归模型（PLSR）对玉米含水率进行估计。杨佳等[5]采集干燥胡萝卜片的高光谱图像，结合多元数据统计分析和化学计量学方法，构建了支持向量机模型（SVM）对干燥胡萝卜片的水分进行预测。Cogdill 等[6]在 750～1 090 nm 范围内采集了玉米的高光谱图像，使用遗传算法（GA）选择特征波段，建立了偏最小二乘回归模型（PLSR）估计玉米种子水分。吴静珠等[7]采集了玉米波长范围为 968.05～2 575.05 nm 的高光谱图像，使用主成分分析算法（PCA）提取特征波段，经多元散射校正算法（MSC）预处理数据后，建立了基于加权策略的改进随机森林模型（RF）对种子水分进行估计。然而高光谱数据测量容易受到各种隐含干扰，如环境光干扰、设备本身干扰以及实际农业生产和运输过程造成鹰嘴蜜桃表面的不同状态（潮湿、沾染污物等）导致的干扰，这些影响因素在便携式高光谱成像仪成像中更为突出。同时高光谱图像有数据维度高的特点，但由于实际成本制约以及实验室处理能力的限制，可供试验的样本量通常不大，甚至有些情况下样本数小于数据维数，这与现有的高光谱农业应用有所不同，故将其定义为小样本高光谱数据问题。

鹰嘴蜜桃是广东省河源市连平县特产，尾部长相似鹰嘴，果面绿色有茸毛，肉质爽脆多汁味甜，为中国“国家地理标志产品”，曾荣获“岭南十大佳果”殊荣。研究拟通过试验寻找一种解决方法，尽量减少高光谱数据测量的干扰，并能在样本数量较小的条件下对鹰嘴蜜桃含水率能有较好的预测结果，最终实现以含水率为标准的分级，以期为不同含水率等级的鹰嘴蜜桃使用不同加工方法和销售策略提供依据。

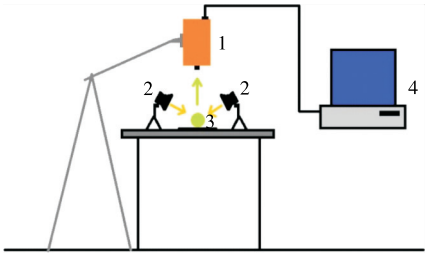
1　材料与方法

1.1　仪器与设备

高光谱成像系统：由高光谱成像仪（GaiaField 便携式地物高光谱成像仪，Schneider Kreuznach Xenoplan 1.4/23-0902 镜头，四川双利合谱公司）、SpecView 软件（高光谱成像仪配套软件，四川双利合谱公司）、光源（4 个 100 W 光纤卤素灯）、高光谱成像仪支架和计算机组成（图 1），实验室自行搭建；

烘干机：WFM 干果机型，爱仕达股份有限公司；

电子秤：CX-I2000 型，500 g/0.01 g，东莞市南城长协电子制品厂。



1. 高光谱成像仪　2. 光源　3. 样品　4. 计算机

图 1　高光谱成像系统示意图

Figure 1　Diagram of a hyperspectral imaging system

1.2　试验材料

鹰嘴蜜桃：2021 年 7 月采摘，河源连平县桃花缘生态农业有限公司。

1.3　试验方法

1.3.1　贮藏与采样　为保证样品新鲜程度，全部样品分 125，142，214，120 个 4 批次采摘并送达。根据批次数量不等分别花费 4，5，6，3 d 完成检测（先采集高光谱数据并测量样品新鲜切片湿重，待新鲜切片烘干 24 h 后测量样品切片干重，视样品送达的时间点到采集完高光谱数据的时间点为完成检测所花费的时间），送达当天即开始检测，无法当天完成检测的样品均放入冰箱冷藏室保存。冷藏样品于每日开始检测前从冰箱中取出提前放置，待其表面凝结的水珠自然蒸发消失后方可开始检测。

1.3.2　含水率测定　用削皮刀在鹰嘴蜜桃样品的前后左右 4 面各削下一片重约 1.20～2.00 g 的薄片，4 片一起称重并记录。将 4 片叠在一起放入烘干机，70 ℃干燥 24 h

后称重并记录,按式(1)计算样品含水率。

$$a=\frac{W_b-W_a}{W_b}\times 100\%, \quad (1)$$

式中:

a——含水率,%;

W_a——样品薄片烘干后的重量,g;

W_b——新鲜样品薄片原始重量,g。

1.3.3　高光谱检测　高光谱成像仪曝光设置选择SpecView软件程序中的"自动曝光"功能,高光谱成像仪电机扫描速度设置选择SpecView软件程序中的"速度匹配"功能,拍摄分辨率设置为696×620,波段数设置为256,白背景数据由拍摄标准反射白板(国家计量院提供)获得,黑背景数据由盖上高光谱成像仪的镜头盖后拍摄获得。拍摄时每次拍摄一个鹰嘴蜜桃样品,4个100 W光纤卤素灯平均分布于样品四周。

1.4　数据处理

1.4.1　反射率校正　在SpecView软件中对原始高光谱图像数据进行反射率校正:将原始高光谱图像数据、白背景数据、黑背景数据按式(2)计算得到校正后的高光谱图像数据。

$$R_{ci}=\frac{\mathrm{Sample}_{ci}-\mathrm{dark}_{ci}}{\mathrm{White}_{ci}-\mathrm{dark}_{ci}}, \quad (2)$$

式中:

Sample_{ci}——原始高光谱图像数据;

dark_{ci}——黑背景数据;

White_{ci}——白背景数据。

1.4.2　图像ROI区域提取　采集的鹰嘴蜜桃高光谱图像波长范围为386.20~1 008.50 nm,共有256个波段,由于波长为438.10 nm(第24个波段)的图像边界最明显,故选用第24个波段的图像进行图像分割。

为了得到可用于建模的高光谱数据,以通过上述流程分割得到的图像区域作为模板,挑选出各样本光谱图像中相应模版区域的像素值,按式(3)计算每段光谱中鹰嘴蜜桃所在区域内的平均光谱值。

$$R_{\mathrm{mean}(\lambda)}=\frac{1}{N}\sum_{i=1}^{N}f(\lambda,i), \quad (3)$$

式中:

N——感兴趣区域的总像素数;

$f(\lambda,i)$——像素i对应波长下的像素值。

1.4.3　PCA主成分分析数据降维　高光谱数据维数为256,但样本数量较小(仅处理了130个鹰嘴蜜桃高光谱数据样本),然而马氏距离计算的样本量需大于特征维度,故无法直接计算样本间的马氏距离,所以需要在计算马氏距离前先利用主成分分析算法(PCA)对光谱数据进行降维处理。主成分分析算法(PCA)会对所有特征进行中心化后求出协方差矩阵的特征值和对应的特征向量,最后将原始的高维特征投影到特征向量上,得到信息量最大(方差最大)的低维特征向量。因此主成分分析算法(PCA)能够在压缩数据维度的同时保留较多的原数据点。

1.4.4　MD马氏距离异常值剔除　为了保证模型预测精度,需要进行数据清洗操作。马氏距离计算公式如式(4)。

$$d_M(x)=\sqrt{(x-\mu)^T\sum^{-1}(x-\mu)}, \quad (4)$$

式中:

x——样本数据光谱矩阵,$x=(x_1,x_2,\cdots,x_p)^T$;

μ——样本数据光谱均值,$\mu=(\mu_1,\mu_2,\cdots,\mu_p)^T$;

$\sum^{-1}$——协方差的逆矩阵。

样本均为鹰嘴蜜桃,因此不同光谱曲线的数值应大致相同,可以用计算马氏距离的方式[8]来对样本的相似性进行定量度量,并衡量测量出的光谱曲线与真实光谱曲线的偏差。这种相似性距离越小,说明样本测量时的偏差越小,样本质量越好,当一个样本与其他样本之间的相似性距离大于某个阈值时,该样本将被视为需要剔除的异常样本。

1.4.5　原始光谱数据的预处理　由于高光谱图像在成像时易受到环境因素的影响,使光谱曲线上出现噪声、失真、毛刺等问题,而大部分去噪手段都会造成原数据有效信息的损失,为了在减少光谱曲线噪声的同时,尽量维持原数据的有效信息,分别使用多项式平滑算法(SG)[9]、多元散射矫正算法(MSC)[10]、标准正态变量算法(SNV)[11]、一阶导数算子(D1)和二阶导数算子(D2)5种方法对鹰嘴蜜桃的高光谱曲线图进行预处理。

1.4.6　训练数据的划分　由于训练样本量有限,为了增加数据集样本间的差异性和代表性,提升模型稳定性,使用Kennard-Stone算法[12](KS)进行数据集划分。KS算法能在最大程度上保证训练集中样本按照欧式距离空间分布均匀。

1.5　建立回归模型

采用偏最小二乘回归(PLSR)模型,设原始光谱特征数据$\boldsymbol{X}_0$为$N\times m$维矩阵,由于此次目标数据只有含水率一个维度,故原始目标数据$\boldsymbol{Y}_0$为$N\times 1$维矩阵。首先对$\boldsymbol{X}_0$和$\boldsymbol{Y}_0$两原始矩阵进行标准化,得到标准化后的特征矩阵$\boldsymbol{X}$、$\boldsymbol{Y}$。设特征矩阵$\boldsymbol{X}$、$\boldsymbol{Y}$的第一组主成分轴向量分别为$\boldsymbol{w}_1$和$\boldsymbol{c}_1$,则特征矩阵$\boldsymbol{X}$、$\boldsymbol{Y}$的第一对主成分为$t_1=\boldsymbol{X}\times\boldsymbol{w}_1$,$u_1=\boldsymbol{Y}\times\boldsymbol{c}_1$。

根据假设,PLSR的求解目标:

$$\begin{aligned}&\mathrm{Maximize}:<\boldsymbol{X}\boldsymbol{w}_1,\boldsymbol{Y}\boldsymbol{c}_1>\\&\mathrm{Subject\ to}:\|\boldsymbol{w}_1\|=1\\&\|\boldsymbol{c}_1\|=1。\end{aligned} \quad (5)$$

在此引入拉格朗日乘子计算方法,计算矩阵$\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}$与$\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{Y}$的最大特征值对应的单位特征向量

得出 $\boldsymbol{w}_1$ 与 $\boldsymbol{c}_1$，由其与 $\boldsymbol{X}$、$\boldsymbol{Y}$ 的关系计算出第一对相关的主成分：

$$\begin{cases} t_1=\boldsymbol{X}\boldsymbol{w}_1 \\ u_1=\boldsymbol{Y}\boldsymbol{c}_1 \end{cases}。\tag{6}$$

建立 $\boldsymbol{X}$、$\boldsymbol{Y}$ 分别对第一对主成分 t_1、u_1 进行回归模型：

$$\begin{cases} \boldsymbol{X}=t_1p_1^T+\boldsymbol{E} \\ \boldsymbol{Y}=u_1q_1^T+\boldsymbol{G} \end{cases},\tag{7}$$

式中：

$\boldsymbol{E}$、$\boldsymbol{G}$——残差矩阵（$\boldsymbol{X}$、$\boldsymbol{Y}$ 中第一对主成分 t_1、u_1 无法解释的部分）。

式(7)仍无法直接建立 $\boldsymbol{X}$、$\boldsymbol{Y}$ 之间的关系，所以在此利用 t_1、u_1 之间的相关性将 $\boldsymbol{Y}$ 改为对 $\boldsymbol{X}$ 的主成分 t_1 进行回归，回归模型为：

$$\boldsymbol{Y}=t_1r_1^T+\boldsymbol{F}。\tag{8}$$

联立式(4)、式(5)和式(7)，可用最小二乘计算出 p_1、q_1、r_1：

$$\begin{cases} p_1=\dfrac{\boldsymbol{X}^Tt_1}{\|t_1\|^2} \\ q_1=\dfrac{\boldsymbol{Y}^Tu_1}{\|u_1\|^2} \\ r_1=\dfrac{\boldsymbol{Y}^Tt_1}{\|t_1\|^2} \end{cases}。\tag{9}$$

联立式(8)、式(9)以推导出 $\boldsymbol{w}_1$，p_1 之间的关系：

$$\boldsymbol{w}_1^Tp_1=\boldsymbol{w}_1^T\ \frac{\boldsymbol{X}^Tt_1}{\|t_1\|^2}=\frac{t_1^Tt_1}{\|t_1\|^2}=1,\tag{10}$$

式中：

$\boldsymbol{w}_1$——$\boldsymbol{X}$ 投影出 t_1 的方向向量。

将前一对主成分不能解释的部分，残差矩阵 $\boldsymbol{E}$、$\boldsymbol{F}$ 作为新的 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$，按上述方法不断重复进行计算，残差矩阵中的数值会随重复的计算不断减小，当残差矩阵 $\boldsymbol{F}$ 小于某个设定好的阈值时，计算停止。设最终计算后共有 k 组主成分，最终可将原始 $\boldsymbol{X}$、$\boldsymbol{Y}$ 表示为：

$$\begin{cases} \boldsymbol{X}=t_1p_1^T+t_2p_2^T+\cdots+t_kp_k^T+\boldsymbol{E} \\ \boldsymbol{Y}=t_1r_1^T+t_2r_2^T+\cdots+t_kp_k^T+\boldsymbol{F} \end{cases}。\tag{11}$$

利用 $w_{ij}^T=1(i=j)$，$w_i^Tt_j=0(i\neq j)$ 两个约束条件可以将式(11)变换为：

$$\begin{cases} \boldsymbol{X}=\boldsymbol{T}P^T+\boldsymbol{E} \\ \boldsymbol{Y}=\boldsymbol{T}R^T+\boldsymbol{F} \end{cases}。\tag{12}$$

对式(12)进行变换可得：

$$\boldsymbol{Y}=\boldsymbol{X}WR^T+\boldsymbol{F}=\boldsymbol{X}A+\boldsymbol{F}。\tag{13}$$

1.6　模型评价指标

回归模型的回归效果一般使用决定系数(R^2)和均方根误差(RMSE)进行衡量，其计算公式：

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)}{\sum_{i=1}^{n}(y_i-\bar{y}_i)}\in(-\infty,1],\tag{14}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}\in[0,+\infty),\tag{15}$$

式中：

$\hat{y}_i$——真实值；

$\bar{y}_i$——真实值的平均值。

决定系数 R^2 越大，表示回归模型估计的数值越接近真实值；均方根误差 RMSE 越小，回归模型估计的数值越接近真实值。

2　结果与分析

2.1　高光谱数据处理结果

2.1.1　图像 ROI 区域提取结果　使用上述图像 ROI 区域提取流程对原始图像进行分割，该套流程可以准确识别出鹰嘴蜜桃的完整轮廓，从而提取出图像的 ROI 区域，分割后的图像如图 2 所示。

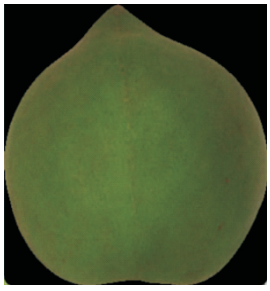


图 2　进行图像分割后的感兴趣区域图像

Figure 2　Region of interest image after image segmentation

2.1.2　波段选择与分析　鹰嘴蜜桃的原始光谱曲线如图 3 所示，每条曲线均为一个样本的光谱曲线，光谱曲线上每个点所对应的数值为该波长下的光谱反射率。

由图 3 可知，不同样本光谱曲线总体趋势相近，在可见光范围 386.20～759.50 nm(光谱曲线范围 0～158)光谱平均反射率较低，短波近红外范围 762.00～1 008.50 nm(光谱曲线范围159～255)光谱平均反射率

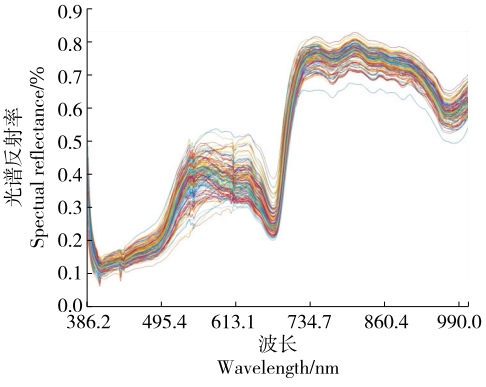


图 3　鹰嘴蜜桃原始光谱曲线图

Figure 3　The original spectral curve of peach

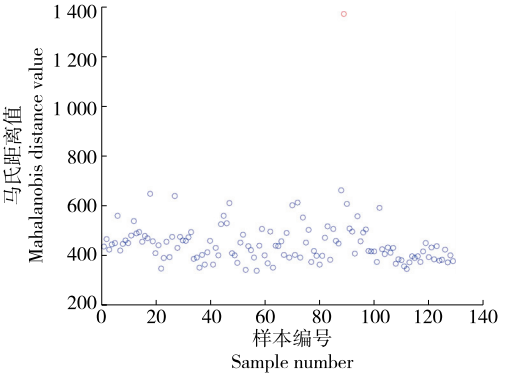
高。680.70 nm(光谱曲线位置 126)处光谱曲线有一处吸收峰,是因为鹰嘴蜜桃表皮中的叶绿素吸收导致的,979.50 nm(光谱曲线位置 244)处光谱曲线亦有一处吸收峰,其与鹰嘴蜜桃中的水分和糖分化合物有关。高光谱曲线有较大的基线偏移现象并伴随微小抖动,并在 435.80,544.40,603.50 nm(光谱曲线位置 23,70,95)附近存在较为严重的毛刺,是由于高光谱成像系统测量时环境因素或设备本身缺陷对高光谱曲线造成的影响,是正常高光谱曲线不应存在的现象。

经后期试验得知 386.20～406.40 nm 信噪较低,故去除,最终结果只保留 406.40～1 008.50 nm 的波长。

2.1.3　MD 马氏距离异常值剔除　由于样本数量较小,仅采集处理了 130 个鹰嘴蜜桃高光谱数据样本,光谱维度数为 256 维,高于样本总量,马氏距离的计算样本量需大于特征维度,故无法直接计算样本间的马氏距离。因此,先利用 PCA 算法对光谱数据进行降维处理,再求得降维后的马氏距离,以此衡量样本质量,剔除异常值。由图 4 可知,红色即为异常样本,需要被剔除。

2.1.4　高光谱图像预处理　分别采用多项式平滑算法(SG)、多元散射矫正算法(MSC)、标准正态变量算法(SNV)、一阶导数算子(D1)和二阶导数算子(D2)5 种方法对光谱进行预处理,处理后的光谱曲线如图 5 所示。

由图 5(d)和图 5(e)可知,光谱在某些特定区域的导数值相对较大,与原始光谱曲线的结论相符。由于一阶导数(D1)和二阶导数(D2)对数据的灵敏度更高,因此对



红色为异常样本

图 4　MD 马氏距离异常值剔除图

Figure 4　Mahalanobis distance outlier removal chart

抖动和毛刺缺陷敏感,同时原始光谱数据本身存在抖动和毛刺缺陷,故在此数据集下,使用一阶导数和二阶导数对数据预处理可能并不合适。

2.1.5　不同数据预处理后的 PLSR 估计结果　基于预处理后的高光谱数据,建立以鹰嘴蜜桃含水率为标签的 PLSR 回归模型,并使用常用的回归模型评价指标,对不同预处理方式下的 PLSR 回归模型性能进行衡量,结果见表 1。

由表 1 可知,一阶导数 D1 和二阶导数 D2 受抖动和毛刺影响较大,虽然这两种算子对光谱信息更为敏感,但由于异常数据的干扰,用这两种方法进行预处理的拟合效果不佳。MSC多元散射矫正算法作为最常用的高光谱

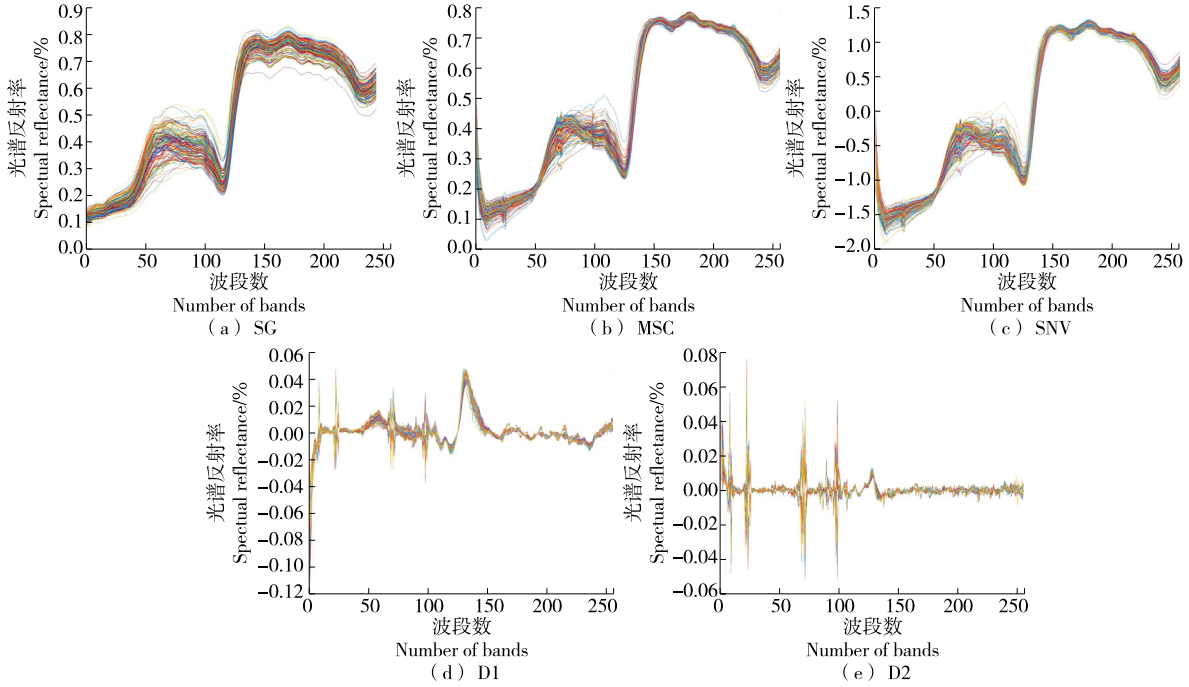


图 5　经不同方法预处理后的光谱曲线图

Figure 5　Spectral curves after different preprocessing methods

表 1　不同数据预处理后的 PSLR 估计结果†

Table 1　Comparison of PLSR estimation results under different preprocessing methods

预处理方法	R_c^2	R_p^2	$RMSE_c$	$RMSE_p$	CMP
RAW	0.911	0.866	0.009 8	0.009 7	17
SG	0.928	0.926	0.008 4	0.009 2	23
MSC	0.942	0.912	0.007 2	0.011 1	23
SNV	0.931	0.863	0.008 8	0.007 9	20
D1	0.994	0.872	0.002 6	0.010 7	26
D2	0.975	0.904	0.004 7	0.011 2	17

† R_c^2为训练集的决定系数；R_p^2为测试集的决定系数；$RMSE_c$为训练集下的均方根误差；$RMSE_p$为测试集下的均方根误差；RAW 为未进行预处理。

预处理方法，其虽然改善了原始数据的共线性，但其拟合程度仍略低于 SG 的。数据经 SG 平滑处理后的光谱图像拟合程度较好，在同时衡量训练集和测试集预测准确性时，SG 在训练集和测试集中预测的准确性均较高，误差较小。因此，在原始高光谱曲线存在多毛刺的情况下，使用多项式平滑算法预处理原始高光谱曲线能够提升模型的回归效果。

2.1.6　PLSR 主成分数的选择　由图 6 可知，当 PLSR 超参数 COMPONENT 为 23 时，回归效果最好，此时训练集下决定系数 R^2 为 0.928，均方根误差 RMSE 为 0.008 4，测试集下 R^2 为 0.926，RMSE 为 0.009 2，对样本拟合程度较好。

2.2　高光谱测定结果

由于在测量样本高光谱图像以及含水量时存在人为误差与系统误差，在综合两类误差后将相对误差＞2%的样本认定为预测错误，将相对误差＜2%的样本认定为预测正确。由图 7 可知，相对误差＞2%的有 2 个样本，相对误差＜2%的有 24 个样本，该模型样本预测正确率为 92.31%，预测正确率较高。

2.3　含水率结果

根据含水率对鹰嘴蜜桃进行分级，统计了全部520 个

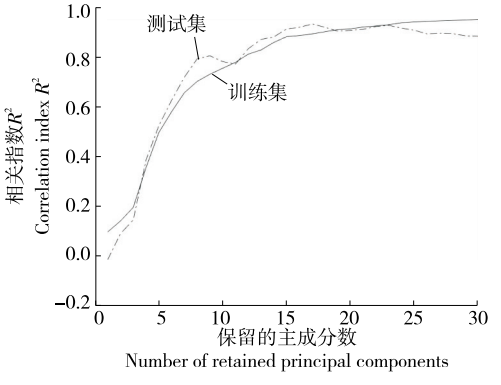


图 6　PLSR 超参数学习曲线

Figure 6　PLSR hyperparameter learning curve

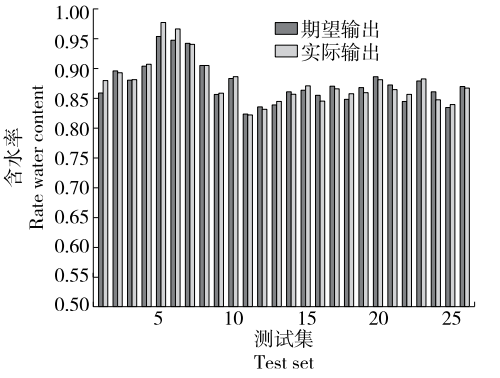


图 7　测试集样本含水率预测值与真实值对比

Figure 7　Chart of predicted values versus true values of the test set

鹰嘴蜜桃含水率，并计算出鹰嘴蜜桃含水率平均值为 0.866 7，方差为 5.31e－04，分布状态近似于正态分布（图 8）。将鹰嘴蜜桃含水率从低到高进行排序，并均衡地划分为 4 级：含水率＜85.29%为含水率低级，含水率 85.29%～86.76%为含水率中低级，含水率 86.76%～88.12%为含水率中高级，含水率＞88.12%为含水率高级。分别使用训练好的模型对训练集和测试集进行预测，并利用上述分级规则对预测含水率进行分级，结果如图 9 所示。由图 9 可知，训练集、测试集的分级正确率分别为 95.56%，92.31%，说明经试验方式处理训练后的数据在鹰嘴蜜桃含水率预测及分级上有较好表现。

3　结论

针对鹰嘴蜜桃含水率无损检测任务的特点与难点，对比了多种高光谱的数据预处理方法，最终确定在多毛刺小样本高光谱数据情况下，SG-PCA-MD-KS-PLSR 为最佳模型，并通过试验发现在超参数 COMPONENT 为 23 时取得最好的回归效果，该模型在训练集中的决定系数为 0.928，均方根误差为 0.008 4，测试集中的决定系数

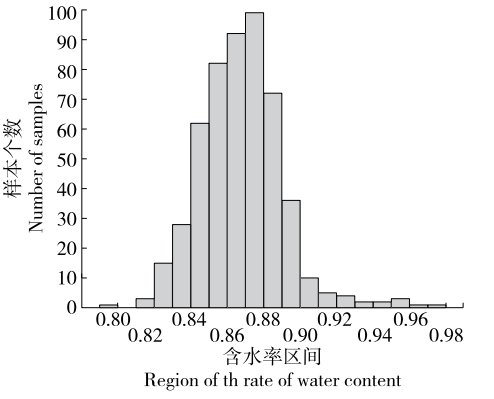
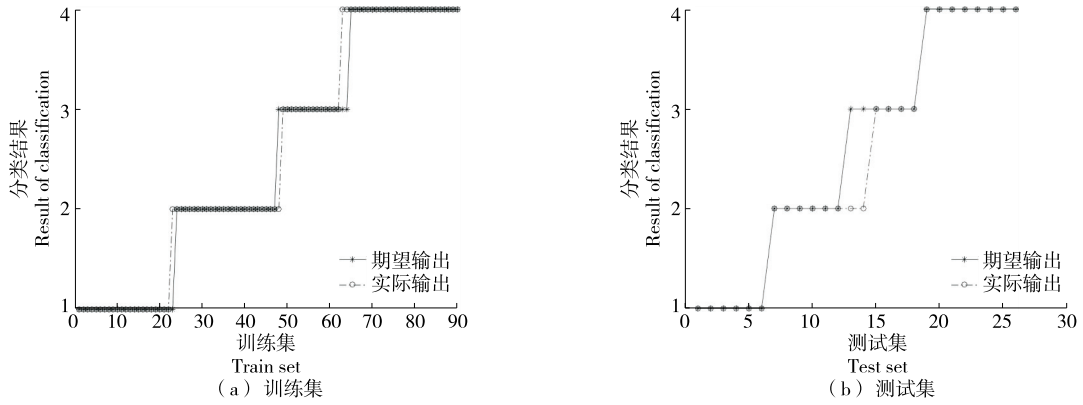


图 8　全部样品的含水率分布直方图

Figure 8　Histogram of water content distribution for all samples



（a）训练集　（b）测试集

图 9　SG-PCA-MD-KS-PLSR 模型依据含水率分级的预测与期望结果

Figure 9　Predicted and expected results of SG-PCA-MD-KS-PLSR model graded according to water content

为 0.926,均方根误差为 0.009 2。在验证试验中,以 2% 的相对误差率衡量含水率估计的正确性,则测试样本含水率的估计正确率为 92.31%。在含水率分级试验中,模型回归的含水率估计值可以较好地对鹰嘴蜜桃进行分级,训练集、测试集的分级正确率分别为 95.56%,92.31%。综上,SG-PCA-MD-KS-PLSR 处理能有效估计出鹰嘴蜜桃含水率,实现对其含水率的无损检测。后续可研究鹰嘴蜜桃其他的品质指标如硬度、甜度等的最佳检测模型。

参考文献

[1] 吴龙国, 何建国, 贺晓光, 等. 高光谱图像技术在水果无损检测中的研究进展[J]. 激光与红外, 2013, 43(9): 990-996.
WU L G, HE J G, HE X G, et al. Research progress of hyperspectral imaging technology in non-destructive detection of fruit[J]. Laser & Infrared, 2013, 43(9): 990-996.

[2] 杨昆程, 孙梅, 陈兴海. 水果成熟度的高光谱成像无损检测研究[J]. 食品科学技术学报, 2015, 33(4): 63-67.
YANG K C, SUN M, CHEN X H. Nondestructive inspect of fruit maturity with hyperspectral imaging technology[J]. Journal of Food Science and Technology, 2015, 33(4): 63-67.

[3] 刘燕德, 邓清. 高光谱成像技术在水果无损检测中的应用[J]. 农机化研究, 2015, 37(7): 227-231, 235.
LIU Y D, DENG Q. Research progress of hyperspectral imaging technology in non-destructive detection of fruit [J]. Journal of Agricultural Mechanization Research, 2015, 37(7): 227-231, 235.

[4] 廉孟茹, 张淑娟, 任锐, 等. 基于高光谱技术的鲜食水果玉米含水率无损检测[J]. 食品与机械, 2021, 37(9): 127-132.
LIAN M R, ZHANG S J, REN R, et al. Nondestructive detection of moisture content in fresh fruit corn based on hyperspectral technology[J]. Food & Machinery, 2021, 37(9): 127-132.

[5] 杨佳, 刘强, 赵楠, 等. 基于高光谱成像的干燥胡萝卜片水分及类胡萝卜素含量无损检测和可视化分析[J]. 食品科学, 2020, 41(12): 285-291.
YANG J, LIU Q, ZHAO N, et al. Hyperspectral imaging for Non-destructive determination and visualization of moisture and carotenoid contents in carrot slices during drying[J]. Food Science, 2020, 41(12): 285-291.

[6] COGDILL R P, HURBURGH C R, RIPPKE G R, et al. Single-kernel maize analysis by near-infrared hyperspectral imaging [J]. Transactions of the ASAE, 2004, 47(1): 311.

[7] 吴静珠, 张乐, 李江波, 等. 基于高光谱与集成学习的单粒玉米种子水分检测模型[J]. 农业机械学报, 2022, 53(5): 302-308.
WU J Z, ZHANG L, LI J B, et al. Detection model of moisture content of single maize seed based on hyperspectral image and ensemble learning [J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(5): 302-308.

[8] 于一凡, 潘军, 邢立新, 等. 基于马氏距离的遥感图像高温目标识别方法研究[J]. 遥感信息, 2013, 28(5): 90-94.
YU Y F, PAN J, XING L X, et al. Identification of high temperature Targets in remote sensing imagery based on mahalanobis distance [J]. Remote Sensing Information, 2013, 28(5): 90-94.

[9] SAVITZKY A, GOLAY M J E. Smoothing and differentiation of data by simplified least squares procedures[J]. Analytical Chemistry, 1964, 36(8): 1 627-1 639.

[10] 郭冰青. 反射光谱及高光谱成像结合多元分析方法快速检测肉类品质[D]. 合肥: 安徽大学, 2021: 9.
GUO B Q. Rapid detection of meat quality using reflectance spectroscopy and hyperspectral imaging with multivariate analysis methods[D]. Hefei: Anhui University, 2021: 9.

[11] 马帅帅, 于慧春, 殷勇, 等. 黄瓜水分和硬度高光谱特征波长选择与预测模型构建[J]. 食品与机械, 2021, 37(2): 145-151.
MA S S, YU H C, YIN Y, et al. Selection of hyperspectral characteristic wavelength and construction of prediction model for cucumber hardness and moisture[J]. Food & Machinery, 2021, 37 (2): 145-151.

[12] KANDPAL L M, LOHUMI S, KIM M S, et al. Near-infrared hyperspectral imaging system coupled with multivariate methods to predict viability and vigor in muskmelon seeds[J]. Sensors and Actuators B: Chemical, 2016, 229: 534-544.