doi:10.6041/j.issn.1000-1298.2024.S1.036

基于距离校正和数据融合的苹果可溶性固形物含量预测模型优化

李　阳[1,2]　彭彦昆[1,2]　李永玉[1,2]

（1. 中国农业大学工学院，北京 100083；2. 国家农产品加工技术装备研发分中心，北京 100083）

摘要： 采用可见/近红外漫反射光谱技术对苹果可溶性固形物含量(Soluble solids content, SSC)检测时，光谱采集探头到苹果表面的距离变化是随机和不可控的，造成检测精度降低。并且采用特征波长筛选算法优化预测模型时，忽略了被舍弃光谱数据中所包含的与成分含量相关信息，造成光谱信息丢失。针对以上问题，通过探究检测距离对漫反射光谱的影响规律，提出一种距离校正方法(Distance correction, DC)，并采用数据融合方法将特征波长和非特征波长数据中的有效信息相结合，以提高苹果SSC预测模型的预测性能。为了验证所提出方法的有效性，分别采用多元散射校正(Multiple scattering correction, MSC)、标准正态变换(Standard normal variate transform, SNV)和DC算法对苹果光谱预处理后，建立苹果SSC的偏最小二乘回归(Partial least squares regression, PLSR)模型。结果表明，DC能更加有效提升PLSR模型的预测性能。为了减少模型数据量、消除光谱中共线性和无效信息，在DC预处理光谱的基础上，采用竞争性自适应加权采样算法(Competitive adaptive reweighted sampling, CARS)、自举软收缩(Bootstrapping soft shrinkage, BOSS)和区间变量迭代空间收缩法(Interval variable iterative space shrinkage approach, iVISSA)对光谱数据进行特征波长筛选。建模结果表明，DC-CARS-PLSR模型具有较好预测结果，并且大幅减少了光谱数据量。为了充分利用特征波长和非特征波长数据中与苹果SSC相关的信息，将特征和非特征波长PLSR模型的潜变量得分相融合，建立融合PLSR预测模型。结果表明，所提出的数据融合方法能够进一步提高模型预测性能。其中CARS算法的特征波长和非特征波长数据融合建模结果具有最佳预测性能，校正集相关系数R_c、校正集均方根误差(Root mean square error of calibration, RMSEC)、预测集相关系数R_p、预测集均方根误差(Root mean square error of prediction, RMSEP)和相对分析误差(Relative percentage difference, RPD)分别为0.981、0.297%、0.957、0.469%和3.424。

关键词： 苹果可溶性固形物含量；可见/近红外漫反射光谱；距离校正；特征波长筛选；数据融合

中图分类号： O657.33；TS255.7　　**文献标识码：** A　　**文章编号：** 1000-1298(2024)S1-0336-10

Optimization of Apple Soluble Solids Content Prediction Models Based on Distance Correction and Data Fusion

LI Yang[1,2]　PENG Yankun[1,2]　LI Yongyu[1,2]

(1. *College of Engineering*, *China Agricultural University*, *Beijing* 100083, *China*

2. *National R&D Center for Agro-processing Equipment*, *Beijing* 100083, *China*)

Abstract: When using visible/near-infrared diffuse reflectance spectroscopy for the detection of soluble solids content (SSC) in apples, the distance from the spectral acquisition probe to the sample surface varies randomly and uncontrollably, resulting in a reduction of detection accuracy. Moreover, when using characteristic wavelengths to establish the prediction models, the contribution of non-characteristic wavelengths to the prediction results is often neglected, resulting in the loss of spectral information. Therefore, a distance correction (DC) method was proposed by exploring the law of the influence of detection distance on diffuse reflectance spectra and establishing prediction models for apple SSC by combining the modeling method of fusion of characteristic wavelength and non-characteristic wavelength data. The results showed that DC could more effectively improve the prediction performance of the PLSR model; the use of the competitive adaptive reweighted sampling (CARS) algorithm for characteristic wavelength screening based on DC preprocessing could effectively simplify the model and improve the model prediction performance; and the fusion

收稿日期：2024-08-13　修回日期：2024-09-18

基金项目： 国家重点研发计划项目（2021YFD1600101-06）和中国农业大学 2115 人才工程项目

作者简介： 李阳（1991—），男，博士生，主要从事农产品无损检测技术与装备研究，E-mail: lya6699332@163.com

通信作者： 彭彦昆（1960—），男，教授，博士生导师，主要从事农产品无损检测技术与装备研究，E-mail: ypeng@cau.edu.cn

modeling results of characteristic and non-characteristic wavelength data of the CARS algorithm had the best prediction performance, with the correlation coefficient of calibration (R_c), root mean square error of calibration (RMSEC), the correlation coefficient of prediction (R_p), root mean square error of prediction (RMSEP) and relative percentage difference (RPD) of 0.981, 0.297%, 0.957, 0.469% and 3.424, respectively.

Key words: apple soluble solids content; visible/near-infrared diffuse reflectance spectroscopy; distance correction; characteristic wavelength screening; data fusion

0　引言

苹果是常见的水果之一，因其丰富的营养价值深受消费者的喜爱[1]。随着生活水平的提高，消费者对于苹果内部品质的关注度越来越高[2]。可溶性固形物含量（Soluble solids content, SSC）与苹果的风味密切相关，是主要的内部品质属性[3-4]。通过SSC检测对苹果进行品质分级，以满足不同消费者的需求是未来的发展趋势之一[5]。采用传统理化试验的方法对苹果SSC检测，存在效率低、破坏性大和成本高等问题。可见/近红外光谱检测技术具有准确、快速、非破坏性等优点，在近十几年间发展十分迅速[6-7]。

采用可见/近红外光谱技术对农产品品质检测时，由于样本物理属性差异和操作误差导致探头到样本表面的距离变化是随机和不可控的，造成漫反射光谱出现差异，这些差异会降低预测模型的检测精度[8]。因此，需要探究消除检测距离变化对光谱影响的方法，从而提高模型预测性能[9]。

相关学者研究了检测距离对光谱的影响[10-14]。常见的光谱预处理算法，如多元散射校正（Multiple scattering correction, MSC）等可以消除物理属性差异导致的光谱变化，也可在一定程度上消除检测距离对光谱影响。以上研究表明，检测距离变化会影响获取光谱的质量，保证样本与探头检测距离的一致性，是保证检测精度关键。但是在实际检测中，探头和样本之间的距离一致性很难得到保障[15-16]。因此，有必要探究检测距离变化对漫反射光谱的影响规律，从而消除检测距离造成的光谱差异，提高农产品品质检测精度。

以往对光谱数据进行特征波长筛选时，都是采用不同筛选策略挑选与目标成分含量相关性相对较高的波长数据作为特征波长[17]。然而一些相关性相对较弱的非特征波长数据会被舍弃。这些舍弃的波长数据中包含的与成分含量相关信息往往会被忽略，导致光谱信息的损失。探究利用被舍弃波长数据中与成分含量相关信息的方法，可能是进一步提高模型预测性能的关键。数据融合技术可以充分利用不同数据之间的补充信息，从而开发更准确和鲁棒的预测模型[18]。因此，本研究采用数据融合方法将特征波长数据和非特征波长数据中与苹果SSC相关的信息提取融合，以探究该方法对于提高模型预测性能的有效性。

1　材料和方法

1.1　实验材料

本研究以栖霞红富士为研究对象，通过构建苹果SSC预测模型，验证所提出方法的有效性。这批苹果于2022年8月购买于中国农业大学东校区附近美廉美超市，选取无机械损伤和外部缺陷的苹果共计120个，运往中国农业大学工学院无损检测实验室。将苹果样本表面擦拭干净，编号后，放置在4℃下保存。光谱采集前，将苹果放置在室温（20 ℃）下24 h，以减小温度变化对光谱采集的影响。采用随机分组法，按照3∶1的比例将样本分成校正集和预测集来构建苹果SSC预测模型。

1.2　光谱采集装置组成

本研究通过自行搭建的漫反射光谱采集系统，采集漫反射光谱信息。该系统包括：采集探头（FCR-COL型，Avantes公司，荷兰）、光谱仪（USB2000+型，Ocean Optics公司，美国）、高度调节平台（带有刻度标尺）、4个35 W卤钨灯杯（ESS MR 11 35 W型，Philips公司，德国）、计算机、电源和暗箱，如图1所示。

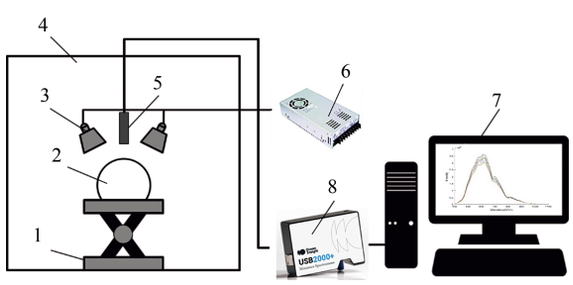


图1　漫反射光谱采集系统示意图

Fig. 1　Schematic of diffuse reflectance spectral acquisition system

1. 高度调节平台　2. 样本　3. 卤钨灯杯　4. 暗箱　5. 探头　6. 电源　7. 计算机　8. 光谱仪

1.3　光谱采集方式

1.3.1　不同距离漫反射光谱采集

为了在理想状况下探究距离变化对漫反射光谱的影响规律，采用直径为80 mm的聚四氟乙烯参考球作为研究对象，在所搭建的采集系统上采集漫反射光谱信息。聚四氟乙烯具有低吸收和与植物材料相似的光散射特性，可以看作成分含量相同、均匀的物质，通常被用作农产品光谱分析中的参考材料。将参考球放置在高度调节平台上，保持探头位置不变，通过旋钮调节平台，

使参考球上表面到探头的距离为3~22 mm，每间隔1 mm采集一次光谱信息。

1.3.2 苹果漫反射光谱采集

为了验证所提出的距离校正方法的有效性，将苹果沿果梗-花萼水平放置在高度调节平台上，保持高度调节平台和探头位置固定，采集苹果赤道区域的漫反射光谱构建苹果SSC预测模型。将光谱仪通过USB数据线连接计算机，使用海洋光学光谱采集软件Ocean View采集光谱信息。软件采集参数如下：积分时间1 ms，启用暗噪声校正，平均次数3。计算吸光度后进行后续的苹果SSC建模分析。采集光谱后使用Matlab软件进行数据分析。

1.4 SSC测定

SSC测定依据农业行业标准NY/T 2637－2014的规定，利用折光仪（PAL-BX/AC5型，ATAGO Co., Ltd.，日本）结合破坏性方法测定苹果的SSC。折光仪的SSC测量范围为0~60.0%，分辨率为0.1%，精度为±0.2%。采集光谱后，将光谱采集区域的果肉取出约30 g，使用手持式榨汁器榨出果汁放入烧杯，搅拌均匀后，使用胶头滴管将果汁滴入折光仪采集位置测定苹果SSC。每个样本采集3次取平均值，作为该样本的SSC。

1.5 距离校正方法

为了减少距离对漫反射光谱的影响，首先采集不同距离的漫反射光谱，通过函数拟合的方法，建立距离与光谱强度之间的方程。然后根据所建立方程，推导出将不同距离光谱校正为参考距离光谱的校正公式。最后将所推导出的校正公式应用于苹果光谱的校正中，验证所提出的DC方法的有效性。

1.6 光谱预处理算法

原始光谱存在大量的无关信息和噪声，会干扰苹果SSC预测模型精度。MSC和SNV算法已经被证实可以在一定程度上消除样本物理性质差异所造成的光谱变化，提高预测模型的预测能力[19]。本研究采用DC、SNV和MSC算法对光谱预处理后，建立苹果SSC的PLSR模型，根据建模效果分析预处理算法的优劣。

1.7 特征波长筛选方法

由于全光谱包含许多无关和共线性信息，影响预测模型的性能。因此，在最佳预处理光谱的基础上采用特征波长提取算法，挑选光谱中与苹果SSC信息密切相关的波长点，可以减少光谱变量数，提高模型预测性能。本研究采用CARS、BOSS和iVISSA算法挑选特征波长，减少光谱数据量。

CARS是与PLSR中的回归系数结合，用来筛选光谱中的波长变量的算法。先从样本的校正集中随机抽取一部份进行PLSR建模，重复随机建模多次，采用指数衰减函数，去除回归系数绝对值权重较小的波长点[20]。经过多次建模，筛选出回归系数绝对值权重较大的波长点，用产生的新变量的子集再进行PLSR建模分析，进行交叉验证，选取交叉验证均方误差（Root mean square of the standard error of cross validation，RMSECV）最小值的子集，这个子集就是选取的最佳波长变量组合。在本研究中，CARS参数设置如下：潜变量（Latent variables，LVs）最大数量设置为15；五折交叉验证；预处理方法为“center”；采样次数为100次。

BOSS算法[21]提供了一种从PLSR模型的回归系数中提取信息的思路。根据权重应用加权自举抽样（Weighted bootstrap sampling，WBS）生成子模型，并使用模型种群分析（Model population analysis，MPA）分析子模型以更新变量的权重。通过基于WBS的变量权重的逐步更新，变量空间软收缩。该算法迭代运行，直到变量数达到1。优化过程遵循软收缩规则，与硬收缩策略相比，它为信息较少的变量分配较小的权重，并且不会直接消除不重要的变量。它可以降低在优化过程中消除重要变量的风险。在本研究中，BOSS参数设置如下：LVs最大数量设置为15；五折交叉验证；预处理方法为“center”；采样次数为1000次。

iVISSA是一种波长区间选择的迭代方法[22]。在该方法中，在全局搜索过程中通过VISSA策略搜索信息性单个波长的位置和组合。每个区间的宽度在局部搜索过程中进行了优化。最后，通过交替执行全局搜索和局部搜索程序，对信息波长间隔的位置、宽度和组合进行智能优化。在本研究中，iVISSA参数设置如下：LVs最大数量设置为15；五折交叉验证；预处理方法为“center”；采样数为500次。

1.8 数据融合方法

首先采用不同的特征波长筛选算法对光谱数据进行特征波长筛选，建立特征波长和非特征波长的PLSR模型，并探究非特征波长数据对苹果SSC的预测能力。然后将两类模型的潜变量得分（Latent variables scores，LVs scores）串联起来，构建融合PLSR预测模型。其中模型的LVs数量都是通过RMSECV最小为准则确定。为了简化描述方式，本研究将CARS特征波长和非特征波长融合方法称为FM1，将BOSS特征波长和非特征波长融合方法称为FM2，iVISSA特征波长和非特征波长融合方法称为FM3。数据融合过程如图2所示。

1.9 建模及评价方法

采用PLSR方法[23]建立苹果SSC预测模型，并通过蒙特卡洛交叉验证结果选择PLSR模型的LVs数量。建立模型后，根据校正集相关系数R_c、预测集相关系数R_p、校正集均方误差（Root mean square error of calibration，RMSEC）、预测集均方误差（Root mean square error of prediction，RMSEP）和相对分析误差（Relative percentage difference，RPD）来评价模型的预测性能[24]。

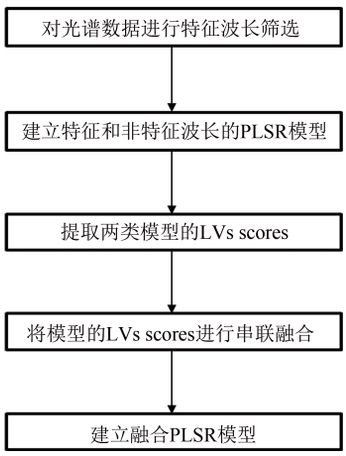


图2　数据融合方法

Fig. 2　Data fusion methods

2　结果与讨论

2.1　距离校正方法探究

2.1.1　距离变化对漫反射光谱的影响规律

采用1.3.1节中光谱采集方法，采集距离为3~22 mm的20条经过暗噪声校正的参考球漫反射光谱，如图3a所示。从图3a可以看出，随着检测距离的增加，光谱整体强度呈上升趋势。这是由于随着检测距离的增加，光源照射的位置逐渐上移，导致检测探头捕获的光子数逐渐增加，造成光谱总体强度上升。绘制600、700、800、900 nm光谱强度I随距离i变化图，如图3b所示。从图3b可以看出，不同波长的光谱强度I随着距离i的增加呈逐渐递增的规律。假设I与i的关系满足

$$I_{i,\lambda} = a_\lambda e^{b_\lambda i} \tag{1}$$

式中　$I_{i,\lambda}$——波长λ距离为i的光谱强度

a_λ——与波长和样本相关的常数，可以看作波长λ的初始距离的光强

b_λ——与波长和样本相关的常数，可以看作波长λ的光强变化系数

使用Matlab的Curve Fitting Tool工具对不同波长I和i按照式(1)进行拟合，拟合结果如图4所示。

从图4可以看出，每个波长的拟合相关系数接近于1，且均方误差较小，表明式(1)拟合效果较好，可以较为可靠的反映I随i变化规律。

2.1.2　距离校正公式推导

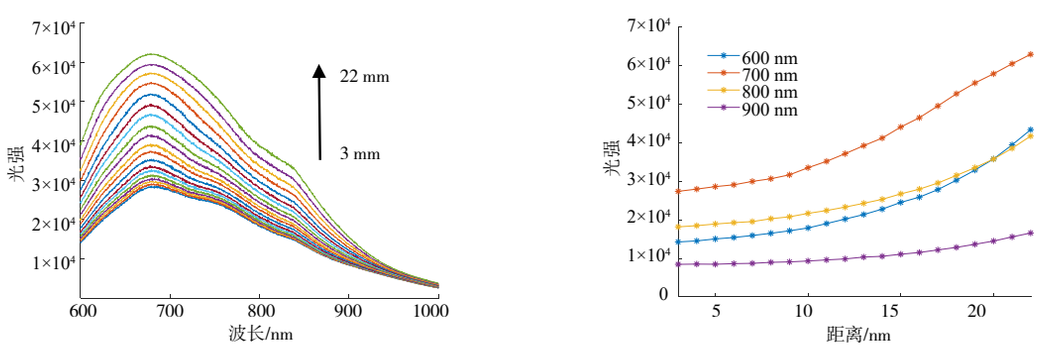
假设$I_{s,\lambda}$为距离为s、波长为λ的待校正光谱，$I_{k,\lambda}$为距离为k、波长为λ的参考光谱，则式(1)可变换为

$$\frac{I_{s,\lambda}}{I_{k,\lambda}} = e^{b_\lambda (s-k)} \tag{2}$$

如果$I_{s,\lambda}$、$I_{k,\lambda}$、s和k已知，则b_λ计算公式为

$$b_\lambda = \frac{\ln \dfrac{I_{s,\lambda}}{I_{k,\lambda}}}{s-k} \tag{3}$$

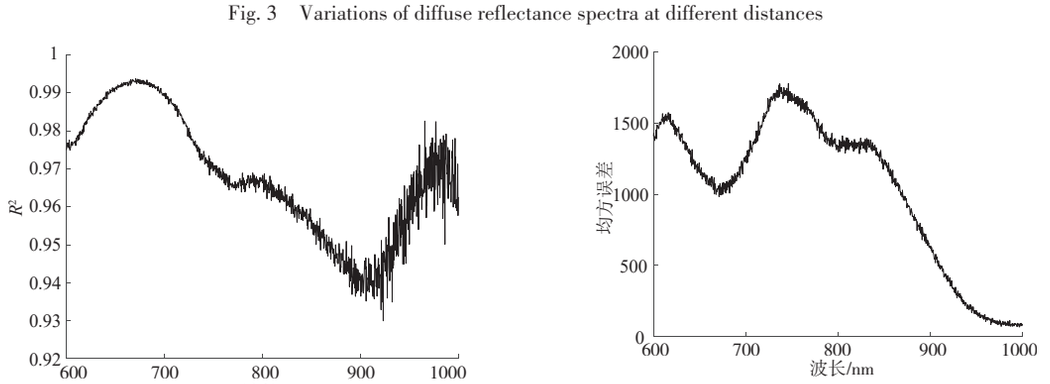
当参考光谱已知，则适用于所有样本的b_λ^*计算公



(a) 不同距离漫反射光谱　　(b) 600、700、800、900 nm光谱强度I随距离i变化

图3　不同距离漫反射光谱变化曲线

Fig. 3　Variations of diffuse reflectance spectra at different distances



(a) 不同波长I与i拟合相关系数　　(b) 不同波长I与i拟合均方误差

图4　不同波长I和i拟合相关系数和均方误差

Fig. 4　Correlation coefficients and mean square errors for different wavelengths I and i fits

式为

$$b_\lambda^* = \frac{\sum_{a=1}^{n} b_{a,\lambda}}{n} \tag{4}$$

式中 $b_{a,\lambda}$——样本a的b_λ

n——样本总数

综上所述，距离校正公式为

$$I_{k,\lambda}^* = \frac{I_{s,\lambda}}{e^{b_\lambda^*(s-k)}} \tag{5}$$

式中 $I_{k,\lambda}^*$——距离校正后光谱强度

使用式(5)可以将不同距离光谱校正为参考距离光谱，从而实现距离校正。在校正过程中，获取探头到样本表面的距离再进行光谱校正会降低在线检测效率。为了简化校正过程，本研究将与距离相关性较高的波长处光谱强度作为距离参数，代入公式进行距离校正。进行光谱校正时，需要确保待校正样本和参考的距离参数为同一波长下的光谱数据。为了寻求与距离相关性较高的波长，本研究依次将每个波长处的光强数据代入公式进行校正，然后采用蒙特卡洛交叉验证法构建校正后光谱与目标成分含量的预测模型，根据交叉验证均方误差最小作为准则，选取与距离相关性较高的波长处光谱强度。

2.2 苹果光谱分析及分组结果统计

采用1.3.2节中光谱采集方法，采集苹果的漫反射光谱用于验证所提出DC方法的校正效果。采集到光谱后，首先截取650~1 000 nm范围内的数据信息以消除光谱两端噪声，然后将原始强度光谱转化为吸光度(图5)。

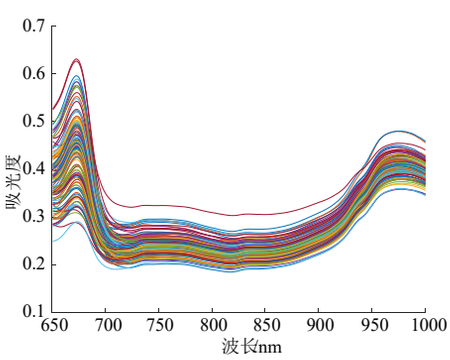


图5 苹果光谱

Fig. 5 Apple spectrum

从图5可以看出，明显的几个峰集中在675、750、840、960 nm附近。其中，675 nm处的吸收峰与苹果皮中的叶绿素和花青素有关[25]。750 nm和960 nm处的峰与含水率有关，分别对应了水的O—H键3倍和2倍频的吸收，840 nm处的峰与苹果SSC中的含C—H键化合物有关[26]。所有苹果样本的光谱曲线趋势相似，但数值有所不同，说明苹果内部物质种类基本相同，但是化合物含量有所差异。光谱差异为建立预测模型提供了基础支撑[27]。

采用1.4节中苹果SSC测定方法，测定120个苹果的SSC数据。统计结果如表1所示。从表1可以看出，校正集和预测集样本的SSC分布较为相似，校正集包含预测集的SSC范围，且样本集的SSC覆盖范围较大。因此，校正集和预测集划分是合理的，有利于构建更为稳健的预测模型。

表1 样本划分及SSC统计

Tab. 1 Sample division and SSC statistics

样本集	样本数量	最大值/%	最小值/%	均值/%	标准差/%
全部	120	14.900	7.400	11.426	1.561
校正集	90	14.900	7.400	11.439	1.554
预测集	30	14.900	7.700	11.387	1.606

2.3 距离校正效果对比分析

为了验证DC方法的校正效果，分别采用MSC、SNV和DC算法对光谱预处理后，建立苹果SSC预测模型。采用DC算法校正时，首先将经过暗噪声校正的强度光谱按照2.1节中方法进行校正，然后将校正后的光谱转化为吸光度后建模分析。选取样本的平均光谱作为DC算法的参考光谱。

建立苹果SSC的PLSR模型时，合理的LVs数量是构建稳定预测模型的关键，LVs数量过少会导致模型预测能力差，LVs数量过多会导致模型过拟合，RMSECV最小值对应的LVs数量通常被认为是最佳LVs数量。本研究采用蒙特卡洛交叉验证方法计算RMSECV随LVs数量变化情况，并根据RMSECV最小原则选取LVs数量。不同预处理算法的苹果SSC建模结果，如表2所示。

表2 不同预处理算法的苹果SSC建模结果

Tab. 2 Apple SSC modeling results with different preprocessing algorithms

预处理方法	LVs数量	校正集		预测集		RPD
		R_c	RMSEC/%	R_p	RMSEP/%	
无	13	0.950	0.486	0.928	0.596	2.695
MSC	13	0.946	0.502	0.933	0.575	2.793
SNV	13	0.939	0.533	0.899	0.698	2.301
DC	14	0.966	0.399	0.943	0.540	2.974

通过对建模结果对比分析可知，MSC和DC预处理后，模型的预测性能有所提升。其中MSC预处理后，使得RPD由原始光谱的2.695提升为2.793。DC预处理后，使得RPD由原始光谱的2.695提升为2.974。SNV预处理后，模型的预测性能有所下降。所有模型中，DC预处理光谱的建模效果提升最为显著。

MSC和SNV预处理算法是光谱分析中常用的预处理算法，它们都是通过消除光谱中的加法和乘法效应，来消除由样本物理属性差异造成的散射效应导致的光谱偏移现象。因此，以上两种算法也能够在一定程度上消除检测距离变化造成的光谱偏移，也是本研究采用以上两种算法进行对比分析的原因。虽然MSC预处理算法也能够提高模型预测性能，但是与DC算法相比，提升效果相对较弱。这可能是检测距离变化是影响苹果光

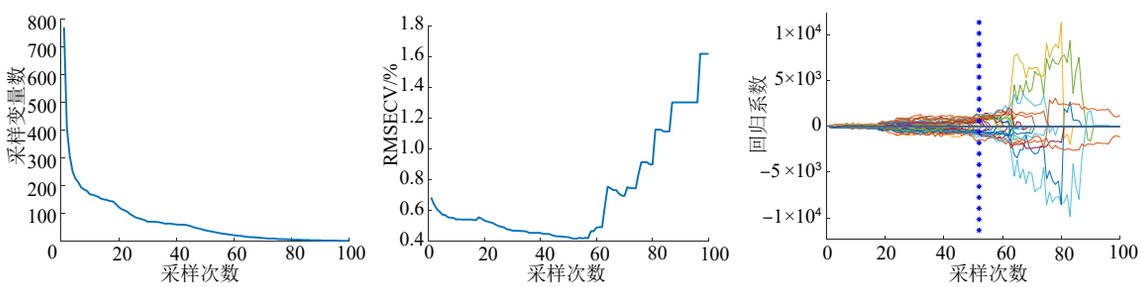
谱质量的主要原因之一，DC算法能够更有针对性地消除检测距离变化对苹果光谱的影响，从而有效提高模型的预测性能。SNV预处理算法导致模型预测性能下降，主要原因可能是这种方法虽然能够消除光谱中的加法和乘法效应，但是也使得SSC与光谱的相关性有所降低。以往研究并没有探究检测距离变化对苹果等类球形果蔬漫反射光谱的影响规律，因此研究结果并不适用于苹果光谱的校正。本研究结果对于提高苹果等类球形果蔬品质检测模型预测性能具有一定的参考价值。

2.4　基于特征波长的苹果SSC模型建立

基于以上的建模结果分析，已经证明了DC算法对苹果光谱的校正效果。为了消除光谱中的共线信息和噪声，简化模型并提高模型预测性能，本研究在DC预处理光谱的基础上，采用CARS、BOSS和iVISSA算法筛选出与苹果SSC密切相关的波长点，并建立苹果SSC预测模型。CARS、BOSS和iVISSA算法都是基于线性建模方法筛选变量，由于其原理不同，不一定与非线性建模方法相关[28]。因此，本研究只构建了基于特征波长的PLSR模型，并对建模结果进行对比分析。

2.4.1　基于CARS的特征波长筛选

在DC预处理光谱的基础上，采用CARS进行特征波长筛选。CARS特征波长筛选过程，如图6所示。由于CARS算法本身就是随机采样法，具有随机性和不确定性，所以经过100次CARS运算选取波长点数少，RMSECV相对较小的特征波长子集作为所筛选出的特征波长。图6a~6c表示在筛选结果最好的一次，CARS运算过程中随着采样次数的增加，变量数、RMSECV和每个变量回归系数变化情况。由图6a可知，在指数衰减函数作用下，变量个数随着采样次数增加由快至慢递减。由图6b可知，随采样次数增加，RMSECV呈先递减后递增变化，当采样次数为52次时，达到最小值，表明在第1~51次采样运算中，光谱中与苹果SSC大量无关信息被去除。52次采样后RMSECV开始递增，表明一些关键信息被剔除导致模型性能变差。图6c中“*”线标示出最小RMSECV所对应采样次数。



(a)　(b)　(c)

图6　CARS特征波长选取过程

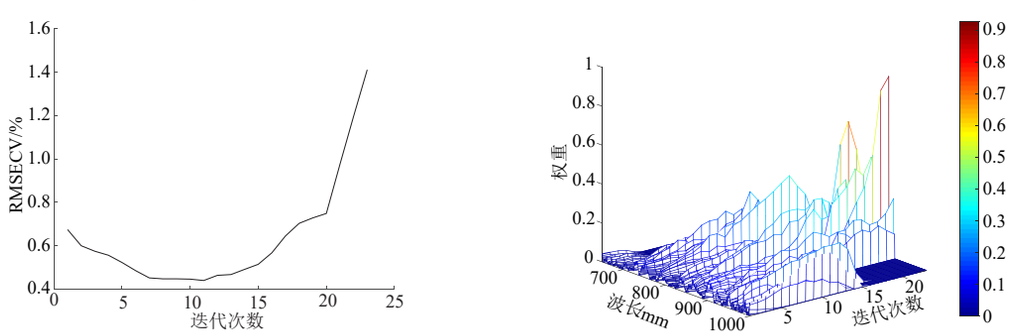
Fig. 6　CARS characteristic wavelength selection process

2.4.2　基于BOSS的特征波长筛选

在DC预处理光谱的基础上，采用BOSS算法进行特征波长筛选。为了避免算法的随机性对变量选择结果的影响，重复进行50次筛选过程，将50次独立运行的最佳结果作为BOSS算法筛选结果。在BOSS方法中，子模型是根据变量的权重生成的。变量的权重是根据子模型的回归系数得到的。每个子模型对应于变量的随机组合，其中变量的权重越大，参与的概率就越大。图7a为RMSECV随迭代次数变化情况。从图中看出，在第11次迭代时达到了最小RMSECV，此时构建子模型的变量即为所筛选出的特征波长。图7b显示了波长的采样权重随迭代次数的变化情况，权重越大被选中的概率就越大。迭代刚开始时，每个波长被选择的概率基本相同，随着迭代的进行重要变量的权重越来越大。

2.4.3　基于iVISSA的特征波长筛选

在DC预处理光谱的基础上，采用iVISSA选取特征



(a)　(b)

图7　BOSS特征波长选取过程

Fig. 7　BOSS characteristic wavelength selection process

波长。图 8 为 iVISSA 特征波长筛选时，RMSECV 随迭代数量增加时的变化情况。从图中可以看出 RMSECV 随着迭代数量的增加先急剧减小，随后趋于平稳状态，说明随着迭代数量的增加，模型的预测能力逐渐提升，当迭代数量增加到一定程度时，模型预测性能趋于稳定。在迭代过程中，波长的采样权重值随迭代次数的变化而变化，算法设置每个波长权重的初始值为 0.5，权重在 0 ~1 之间变化。经过 70 次迭代，采样权重值基本稳定，共提取 355 个特征波长。

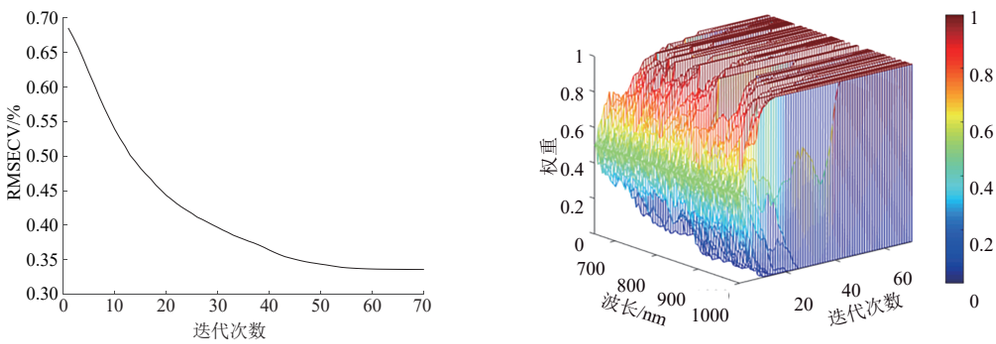


图 8 iVISSA 特征波长选取过程

Fig. 8 iVISSA characteristic wavelength selection process

2.4.4 特征波长筛选结果

特征波长筛选结果如图 9 所示。从图中可以看出，CARS 和 BOSS 筛选特征波长数量较为接近，且波长点重合度较高。iVISSA 算法筛选的特征波长数量较多，基本都包括了 CARS 和 BOSS 算法所筛选出的特征波长。其中，750 nm 波长附近与 O—H 键伸缩的第三泛音有关[29]，760 nm 波长附近与 O—H 键第三泛音和 C—H 键第四泛音有关[30]，800、842、899、920 nm 波长附近与 C—H 键第三泛音有关[31]，960 nm 附近的波长为 O—H 键的第二泛音有关。以上波长均与苹果中的水和 SSC（O—H 和 C—H）有密切关系，说明筛选出的特征波长能够反映苹果 SSC 变化情况。

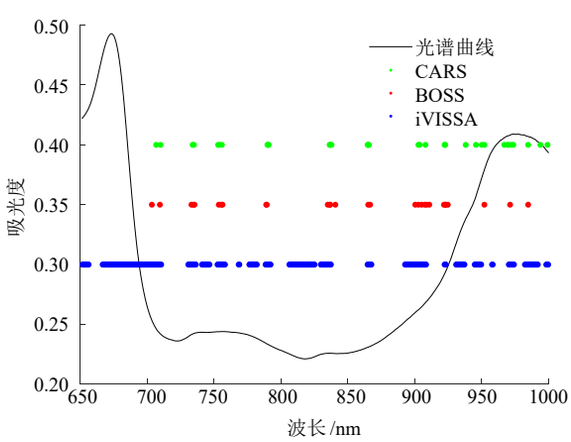


图 9 不同特征波长筛选算法筛选结果

Fig. 9 Screening results of different characteristic wavelength screening algorithms

2.4.5 基于特征波长的 PLSR 模型建立

采用筛选的特征波长建立苹果 SSC 预测模型，结果如表 3 所示。从表 3 可看出，相较于 DC 预处理的全光谱建模结果，CARS、BOSS 和 iVISSA 特征波长筛选算法能够提升模型预测性能，说明特征波长筛选可以有效消除原始光谱中的无关和共线性信息，提高苹果 SSC 预测模型的预测性能。CARS、BOSS 和 iVISSA 算法所选的特征波长数量分别为 36、32 和 355，在全光谱 771 个波长的基础上分别减少 95.331%、95.850% 和 53.956% 的数据量，大幅简化了模型。CARS 算法筛选的特征波长建模结果最好，RPD 为 3.338。BOSS 算法筛选出的特征波长数最少，模型的预测性能仅次于 CARS 算法。iVISSA 算法所筛选出的特征波长数量最多，建模结果相对最差，可能是保留了部分无关信息和噪声所导致。总的来说，DC-CARS-PLSR 在保证模型预测精度的同时，具有较少的特征波长数量。说明该模型可以有效消除检测距离对光谱的影响，提高苹果 SSC 的预测精度。

表 3 基于特征波长的苹果 SSC 建模结果

Tab. 3 Results of apple SSC modeling based on characteristic wavelengths

波长筛选算法	波长数	LVs 数量	校正集		预测集		RPD
			R_c	RMSEC/%	R_p	RMSEP/%	
CARS	36	15	0.980	0.307	0.956	0.474	3.388
BOSS	32	12	0.971	0.372	0.953	0.486	3.305
iVISSA	355	15	0.982	0.294	0.942	0.531	3.024

DC-CARS-PLSR 预测结果散点图如图 10 所示。从图 10 可以看出，DC-CARS-PLSR 预测结果散点图中的散点集中在理想回归线附近（1∶1 线），说明 DC-CARS-PLSR 模型可以准确预测苹果 SSC。

2.5 非特征波长建模结果

为了探究非特征波长是否还包含与 SSC 相关的信息，建立非特征波长的苹果 SSC 的 PLSR 预测模型，结果如表 4 所示。从表 4 可以看出，非特征波长预测模型的性能相比于特征波长有所下降，但是非特征波长预测模型对苹果 SSC 也具有一定的预测能力。说明非特征波

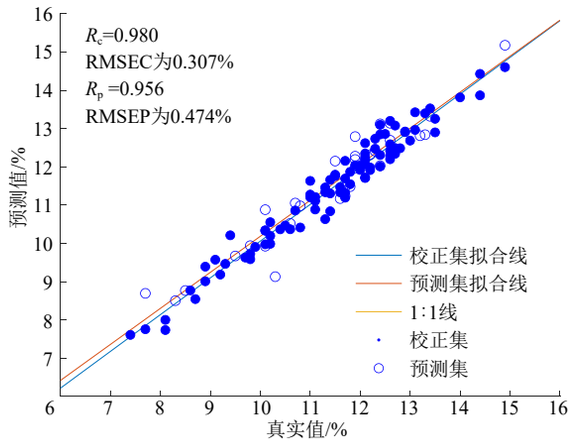


图 10　DC-CARS-PLSR 预测结果散点图

Fig. 10　Scatterplot of DC-CARS-PLSR prediction results

表 4　基于非特征波长数据的苹果 SSC 建模结果

Tab. 4　Apple SSC modeling results based on non-characteristic wavelength data

数据类型	波长数	LVs 数量	校正集		预测集		RPD
			R_c	RMSEC/%	R_p	RMSEP/%	
非 CARS	735	15	0.964	0.412	0.942	0.549	2.925
非 BOSS	739	15	0.964	0.412	0.941	0.552	2.909
非 iVISSA	416	13	0.945	0.504	0.927	0.596	2.695

长中还包含与苹果 SSC 相关的信息。以往采用特征波长建模的方法，并没有充分利用光谱中与苹果 SSC 相关的信息。因此，采用特征波长和非特征波长数据融合的方法，可能会充分利用非特征波长模型对预测结果的贡献，从而进一步提高模型预测性能。

2.6　特征和非特征波长数据融合建模结果

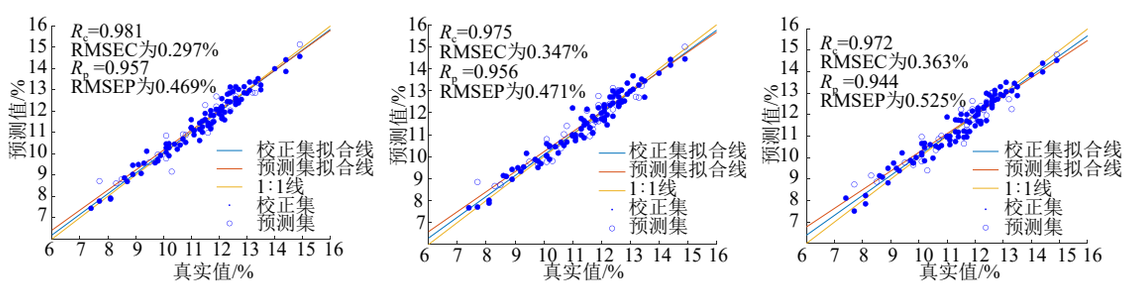
采用 1.8 节中所提出的数据融合方法建立苹果 SSC 的预测模型。建模结果如表 5 所示。

表 5　特征和非特征波长数据融合建模结果

Tab. 5　Fusion modeling results for characteristic and non-characteristic wavelength data

融合方法	变量数	LVs 数量	校正集		预测集		RPD
			R_c	RMSEC/%	R_p	RMSEP/%	
FM1	30	8	0.981	0.297	0.957	0.469	3.424
FM2	27	5	0.975	0.347	0.956	0.471	3.410
FM3	28	1	0.972	0.363	0.944	0.525	3.059

从表 5 可以看出，所提出的数据融合方法均能够进一步提升模型的预测性能。其中，FM1 建模结果最佳，使模型的 RPD 由 3.388 提升到 3.424。FM2 使模型的 RPD 由 3.305 提升到 3.410。FM3 使模型的 RPD 由 3.024 提升到 3.059。建模结果证实了所提出的数据融合方法对于提高模型预测性能的潜力。图 11 为融合模型的苹果 SSC 真实值和预测值的散点图。从图中可以看出，融合模型的真实值和预测值之间差距较小，说明融合模型对苹果 SSC 的预测性能较好。



(a) FM1　　(b) FM2　　(c) FM3

图 11　融合模型的真实值和预测值散点图

Fig. 11　Scatterplot of true and predicted values of fusion models

3　结论

(1)提出了一种检测距离对光谱影响的校正方法，以校正探头到苹果表面距离的变化对漫反射光谱的影响，提高苹果 SSC 检测精度。为了验证 DC 算法的校正效果，在 MSC、SNV 和 DC 预处理光谱的基础上，采用 PLSR 方法构建苹果 SSC 的预测模型。结果表明，DC-PLSR 模型对苹果 SSC 的预测效果最好，R_c、RMSEC、R_p、RMSEP 和 RPD 分别为 0.966、0.399%、0.943、0.540% 和 2.974。

(2)为了减少模型数据量、消除光谱中共线性和无效信息，在 DC 预处理的基础上，采用 CARS、BOSS 和 iVISSA 特征波长筛选算法对光谱进行特征波长筛选，并建立苹果 SSC 的 PLSR 模型。结果表明，DC-CARS-PLSR 模型的预测结果最好，R_c、RMSEC、R_p、RMSEP 和 RPD 分别为 0.980、0.307%、0.956、0.474% 和 3.388，并且减少了 95.331% 的光谱数据量；通过将光谱校正后进行特征波长筛选，可以在简化模型的基础上进一步提高模型预测性能。

(3)探究了特征和非特征波长数据融合对模型的提升效果。结果表明，相比于未融合模型，所提出的数据融合方法能够进一步提高模型预测性能。其中 FM1 融合模型预测效果最好，R_c、RMSEC、R_p、RMSEP 和 RPD 分别为 0.981、0.297%、0.957、0.469% 和 3.424。

参 考 文 献

[1] GIOVANELLI G, SINELLI N, BEGHI R, et al. NIR spectroscopy for the optimization of postharvest apple management[J]. Postharvest Biology and Technology, 2014, 87: 13 - 20.

[2] HARKER F R, GUNSON F A, JAEGER S R. The case for fruit quality: an interpretive review of consumer attitudes, and preferences for apples[J]. Postharvest Biology and Technology, 2003, 28(3): 333 - 347.

[3] LU R F. Multispectral imaging for predicting firmness and soluble solids content of apple fruit[J]. Postharvest Biology and Technology, 2004, 31(2): 147 - 157.

[4] LI J L, SUN D W, CHENG J H. Recent advances in nondestructive analytical techniques for determining the total soluble solids in fruits: a review[J]. Comprehensive Reviews in Food Science and Food Safety , 2016, 15(5): 897 - 911.

[5] HU W H, SUN D W, PU H B, et al. Recent developments in methods and techniques for rapid monitoring of sugar metabolism in fruits [J]. Comprehensive Reviews in Food Science and Food Safety , 2016, 15(6): 1067 - 1079.

[6] CORTES V, BLASCO J, ALEIXOS N, et al. Monitoring strategies for quality control of agricultural products using visible and near-infrared spectroscopy: a review[J]. Trends in Food Science & Technology, 2019, 85: 138 - 148.

[7] IBANEZ G, CEBOLLA-CORNEJO J, MARTI R, et al. Non-destructive determination of taste-related compounds in tomato using NIR spectra[J]. Journal of Food Engineering, 2019, 263: 237 - 242.

[8] TIAN S J, XU H R. Nondestructive methods for the quality assessment of fruits and vegetables considering their physical and biological variability[J]. Food Engineering Reviews, 2022, 14(3): 380 - 407.

[9] ZHANG B H, DAI D J, HUANG J C, et al. Influence of physical and biological variability and solution methods in fruit and vegetable quality nondestructive inspection by using imaging and near-infrared spectroscopy techniques: a review[J]. Critical Reviews in Food Science and Nutrition, 2018, 58(12): 2099 - 2118.

[10] LIU Y D, YING Y B, FU X P, et al. Experiments on predicting sugar content in apples by FT-NIR technique[J]. Journal of Food Engineering, 2007, 80(3): 986 - 989.

[11] PAHLAWAN M F R, WATI R K, MASITHOH R E. Development of a low-cost modular VIS/NIR spectroscopy for predicting soluble solid content of banana[J]. IOP Conference Series: Earth and Environmental Science, 2021, 644(1): 012047.

[12] DIXIT Y, CASADO-GAVALDA M P, CAMA-MONCUNILL R, et al. Multipoint NIR spectrometry and collimated light for predicting the composition of meat samples with high standoff distances[J]. Journal of Food Engineering, 2016, 175: 58 - 64.

[13] ZHU S, CHEN H, WANG M, et al. Plastic solid waste identification system based on near infrared spectroscopy in combination with support vector machine[J]. Advanced Industrial and Engineering Polymer Research, 2019, 2(2): 77 - 81.

[14] CORTES V, CUBERO S, BLASCO J, et al. In-line application of visible and near-infrared diffuse reflectance spectroscopy to identify apple varieties[J]. Food and Bioprocess Technology, 2019, 12(6): 1021 - 1030.

[15] RODIONOVA O Y, BALYKLOVA K S, TITOVA A V, et al. The influence of fiber-probe accessories application on the results of near-infrared (NIR) measurements[J]. Applied Spectroscopy, 2013, 67(12): 1401 - 1407.

[16] HONG F W, CHIA K S. A review on recent near infrared spectroscopic measurement setups and their challenges[J]. Measurement, 2021, 171: 108732.

[17] 赵娟,沈懋生,浦育歌,等. 基于近红外光谱与多品质指标的苹果出库评价模型研究[J]. 农业机械学报, 2023, 54(2): 386 - 395.
ZHAO Juan, SHEN Maosheng, PU Yuge, et al. Out-of-warehouse evaluation and prediction model of apple based on near-infrared spectroscopy combined with multiple quality indexes[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(2): 386 - 395. (in Chinese)

[18] 刘翠玲,秦冬,凌彩金,等. 基于高光谱图谱融合技术的英德红茶等级快速无损判别[J]. 农业机械学报, 2023, 54(3): 402 - 410.
LIU Cuiling, QIN Dong, LING Caijin, et al. Fast nondestructive discrimination of Yingde black tea grade based on fusion of image spectral features of hyperspectral technique[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(3): 402 - 410. (in Chinese)

[19] TIAN S J, ZHANG J H, ZHANG Z X, et al. Effective modification through transmission Vis/NIR spectra affected by fruit size to improve the prediction of moldy apple core[J]. Infrared Physics & Technology, 2019, 100: 117 - 124.

[20] LIU Y H, WANG Q Q, GAO X W, et al. Total phenolic content prediction in Flos Lonicerae using hyperspectral imaging combined with wavelengths selection methods[J]. Journal of Food Process Engineering, 2019, 42(6): e13224.

[21] DENG B C, YUN Y H, CAO D S, et al. A bootstrapping soft shrinkage approach for variable selection in chemical modeling[J]. Analytica Chimica Acta, 2016, 908: 63 - 74.

[22] DENG B C, YUN Y H, MA P, et al. A new method for wavelength interval selection that intelligently optimizes the locations, widths and combinations of the intervals[J]. Analyst, 2015, 140(6): 1876 - 1885.

[23] CHENG J H, SUN D W. Partial least squares regression (PLSR) applied to Nir and Hsi spectral data modeling to predict chemical properties of fish muscle[J]. Food Engineering Reviews, 2017, 9(1): 36 - 49.

[24] FENG L, ZHANG M, ADHIKARI B, et al. Nondestructive detection of postharvest quality of cherry tomatoes using a portable NIR spectrometer and chemometric algorithms[J]. Food Analytical Methods, 2019, 12(4): 914 - 925.

[25] MERZLYAK M N, SOLOVCHENKO A E, GITELSON A A. Reflectance spectral features and non-destructive estimation of chlorophyll, carotenoid and anthocyanin content in apple fruit[J]. Postharvest Biology and Technology, 2003, 27(2): 197 - 211.

[26] GUO Z, WANG M, AGYEKUM A A, et al. Quantitative detection of apple watercore and soluble solids content by near infrared transmittance spectroscopy[J]. Journal of Food Engineering, 2020, 279: 109955.

[27] ZHANG D Y, XU Y F, HUANG W Q, et al. Nondestructive measurement of soluble solids content in apple using near infrared hyperspectral imaging coupled with wavelength selection algorithm[J]. Infrared Physics & Technology, 2019, 98: 297 - 304.

[28] YUN Y H, LI H D, DENG B C, et al. An overview of variable selection methods in multivariate analysis of near-infrared spectra[J]. TrAC Trends in Analytical Chemistry, 2019, 113: 102 - 115.

[29] LIU D, SUN D W, ZENG X A. Recent advances in wavelength selection techniques for hyperspectral image processing in the food industry[J]. Food and Bioprocess Technology, 2014, 7(2): 307 - 323.

[30] JAMSHIDI M, HAMDAMI N, DOKHANI S, et al. Single- and multi-objective optimization of low fat icecream formulation, based on genetic algorithms[J]. Journal of Agricultural Science and Technology, 2012, 14(6): 1285 - 1296.

[31] FU X P, YING Y B. Food safety evaluation based on near infrared spectroscopy and imaging: a review[J]. Critical Reviews in Food Science and Nutrition, 2016, 56(11): 1913 - 1924.

(上接第324页)

[21] 张阳, 李海森, 马礼, 等. 基于ORB-SLAM2算法的水下机器人实时定位研究[J]. 测绘通报, 2019(12): 1 - 7.
ZHANG Yang, LI Haiseng, MA Li, et al. Research on real-time positioning of underwater robot based on ORB-SLAM2 algorithm[J]. Bulletin of Surveying and Mapping, 2019(12): 1 - 7. (in Chinese)

[22] 刘庆运,杨华阳,刘涛,等.基于激光雷达与深度相机融合的SLAM算法[J]. 农业机械学报, 2023, 54(11): 29 - 38.
LIU Qingyun, YANG Huayang, LIU Tao, et al. SLAM algorithm based on fusion of LiDAR and depth camera[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(11): 29 - 38. (in Chinese)

[23] 邓晨, 李宏伟, 张斌, 等. 基于深度学习的语义SLAM关键帧图像处理[J]. 测绘学报, 2021, 50(11): 1605 - 1616.
DENG Chen, LI Hongwei, ZHANG Bin, et al. Research on key frame image processing of semantic SLAM based on deep learnig[J]. Acta Geodaetica et Cartographica Sinica, 2021, 50(11): 1605 - 1616. (in Chinese)

[24] 赵薇, 王峰, 马星宇, 等. 基于动态区域剔除与稠密地图构建的视觉SLAM算法[J/OL]. 兵工学报, 1 - 11[2024-08-31]. http://kns.cnki.net/kcms/detail/11.2176.tj.20240726.1235.0002.html.
ZHAO Wei, WANG Feng, MA Xingyu, et al. Visual SLAM algorithm based on dynamic regionculling and dense map construction[J/OL]. Acta Armamentarii, 1 - 11[2024-08-31]. http://kns.cnki.net/kcms/detail/11.2176.tj.20240726.1235.0002.html. (in Chinese)

[25] 张晨阳, 杨健. 一种自适应点线特征和IMU耦合的视觉SLAM方法[J/OL]. 武汉大学学报(信息科学版), 1 - 19[2024-08-31]. http://doi.org/10.13203/j.whugis20230347.html.
ZHANG Chenyang, YANG Jian. A visual SLAM method coupled with adaptive point-line features and IMU[J/OL]. Geomatics and Information Science of Wuhan University, 1 - 19[2024-08-31]. http://doi.org/10.13203/j.whugis20230347.html. (in Chinese)

[26] 齐咏生, 宋继鹏, 刘利强, 等. 基于点线特征融合的延迟边缘化视觉惯性SLAM方法[J]. 农业机械学报, 2024, 55(12): 373 - 382.
QI Yongsheng, SONG Jipeng, LIU Liqiang, et al. Delayed marginalized visual inertial SLAM method based on point-line feature fusion[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(12): 373 - 382. (in Chinese)