麦类作物学报 2025,45(10):1363－1371
Journal of Triticeae Crops doi: 10.7606/j.issn.1009-1041.2025.10.08

网络出版时间:2025-05-12
网络出版地址:https://link.cnki.net/urlid/61.1359.S.20250508.1512.028

小麦籽粒类胡萝卜素含量近红外快速检测

刘洁琼[1,2],罗 斌[1,2],张 晗[2],康 凯[2],陈 泉[2],邱朝阳[2]

(1.新疆农业大学机电工程学院,新疆乌鲁木齐 830052;2.北京市农林科学院智能装备技术研究中心,北京 100097)

摘 要:小麦籽粒类胡萝卜素含量关系到小麦制品的颜色外观和商品价值,是小麦育种过程中的重要指标之一。目前检测小麦籽粒类胡萝卜素含量的方法主要有紫外分光光度法、薄层色谱法、高效液相色谱法等化学方法,成本高且耗时耗力。为实现小麦籽粒类胡萝卜素含量的快速无损预测,将近红外光谱技术与化学计量技术相结合,利用留出法(hold-out method,HOM)、K 折交叉验证(K-fold cross-validation,KFCV)和时间序列划分(time series split,TSS)3 种样本集划分方法,Savitzky-Golay 平滑(Savitzky-Golay smoothing,SG)、多元散射校正(multivariate scatter correction,MSC)、标准正态变量变换(standard normal variable transformation,SNV)和趋势校正(trend correction,TC)4 种光谱预处理方法,方差阈值特征选择(variance threshold feature selection,VTFS)、SelectKBest 特征选择(SelectKBest feature selection,SKB)、递归特征消除(recursive feature elimination,RFE)分别与主成分分析(PCA)算法融合的 3 种特征选择方法,建立偏最小二乘回归(partial least squares regression,PLSR)、支持向量机回归(support vector machine regression,SVR)、梯度提升回归(gradient boosting regression,GBR)三种模型,比较分析了不同模型预测精度。结果表明,样本集最佳划分方法为留出法,光谱最佳预处理方法为 SG 预处理,最佳特征选择方法为 PCA-SKB,最优模型为 PCA-SKB-GBR,校正集和预测集决定系数 R^2 分别为 0.99 和 0.89,均方根误差 RMSE 分别为 0.03 和 0.34 $\mu g \cdot g^{-1}$,剩余预测偏差 RPD 为 3.01。因此,基于留出法划分样本集、SG 光谱预处理和 PCA-SKB 特征选择方法,建立 PCA-SKB-GBR 模型,可实现小麦籽粒类胡萝卜素含量快速高效预测。

关键词:小麦;类胡萝卜素;近红外光谱;特征选择;估测模型

中图分类号:S512.1;S330 **文献标识码**:A **文章编号**:1009-1041(2025)10-1363-09

Rapid Detection of Carotenoid Content in Wheat Grain Based on Feature Selection by Near Infrared Spectoroscopy

LIU Jieqiong[1,2], LUO Bin[1,2], ZHANG Han[2], KANG Kai[2], CHEN Quan[2], QIU Zhaoyang[2]

(1. College of Mechanical and Electrical Engineering, Xinjiang Agricultural University, Urumqi, Xinjiang 830052, China; 2. Intelligent Equipment Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China)

Abstract: The carotenoid content in wheat kernels serves as a crucial determinant of color appearance and commercial value, representing an important indicator in wheat breeding. While conventional detection methods (ultraviolet spectrophotometry, thin-layer chromatography, and high-performance liquid chromatography) demonstrate accuracy, their operational costs and time requirements prove prohibitive for large-scale applications. This study provided a rapid non-destructive prediction method through the integration of near-infrared spectroscopy(NIRS) with chemometric techniques. We systematically evaluated:(1) three dataset partitioning methods hold-out method(HOM), K-fold cross-validation(KFCV), time series split(TSS); (2) four spectral preprocessing techniques [Savitzky-Golay smoothing(SG), multiplicative scatter correction(MSC), standard normal variate transformation

收稿日期:2024-10-15 修回日期:2024-12-15
基金项目:国家"十四五"重点研发计划项目(2023YFD2000405)
第一作者 E-mail:1549303724@qq.com(刘洁琼)
通讯作者 E-mail:luob@nercita.org.cn(罗 斌)

(SNV), trend correction(TC)], three feature selection methods were obtained by fusing variance threshold feature selection(VTFS), SelectKBest(SKB), and recursive feature elimination(RFE) with principal component analysis (PCA), respectively. Subsequent modeling employed partial least squares regression(PLSR), support vector machine regression(SVR), and gradient boosting regression(GBR) were used to build different prediction models. The results showed that HOM achieved optimal dataset partitioning, SG preprocessing provided superior spectral enhancement, and PCA-SKB feature fusion delivered maximum information retention. The optimized was PCA-SKB-GBR. In the validation set and prediction set, R^2 were 0.99 and 0.89 with RMSE of 0.04 and 0.34 $\mu g \cdot g^{-1}$, respectively. The residual prediction deviation(RPD) reached 3.01. Therefore, based on the hold-out method of sample set, SG spectral pretreatment and PCA-SKB feature selection method, the PCA-SKB-GBR model can be used to predict the carotenoid content of wheat grain quickly and efficiently.
Keywords: Wheat; Carotenoids; Near-infrared spectroscopy; Feature selection; Estimation model

小麦不仅是全世界种植范围最广的粮食作物,也是中国最重要的农作物之一[1]。小麦籽粒中除了蛋白质、水分、碳水化合物、脂肪等主要营养成分外,还含有类胡萝卜素、矿物质、纤维素等物质,这些物质是人体所需要的重要营养[2]。小麦籽粒中类胡萝卜素可以为食物提供亮黄色,对小麦制品的颜色外观和商品价值十分重要,还可以对视觉和一些癌症治疗产生积极作用[3]。因此,开展类胡萝卜素含量的研究对于小麦品质育种有重要意义。目前,小麦籽粒类胡萝卜素含量的检测大多采用紫外分光光度法、薄层色谱法、高效液相色谱法等传统的化学检测方法[4],虽然这些方法的准确度较高,但样本制备和操作过程对人员素质要求较高,难以实现大样本、快速无损测定,不利于小麦材料的高效评价。

利用近红外光谱建立预测模型,可以实现小麦籽粒蛋白质、淀粉、水分、碳水化合物等含量的快速预测[5,6],从而表现出较强的应用潜力。此外,部分研究人员基于近红外光谱开展了小麦籽粒、面粉中含量较少且重要的化学物质含量预测研究。如,Zhao 等[7,8]利用两种新型非破坏性检测方法预测小麦籽粒中玉米赤霉烯酮的污染程度,探索了两种特征区间选择方法对模型预测精度的影响;Shi 等[9]利用近红外光谱预测小麦面粉中非法添加剂的含量,主要是通过不同波段选择算法对关键波长进行筛选。近红外光谱技术也被用于预测作物类胡萝卜素含量。Damián 等[10]认为,利用近红外反射光谱技术可以预测西葫芦果皮和果肉中总类胡萝卜素、叶黄素和 β-胡萝卜素含量。Nicola 等[11]和 Bonierbale 等[12]通过近红外光谱技术实现了玉米籽粒和马铃薯中的类胡萝卜素含量评估。然而,目前有关利用近红外光谱技术预测小麦中类胡萝卜素含量的研究尚未见报道。

本研究基于近红外光谱技术,通过比较留出法(hold-out method, HOM)、K 折交叉验证(K-fold cross-validation, KFCV)和时间序列划分(time series split, TSS)来选择适合该样本集的最优划分方法。将方差阈值特征选择(variance threshold feature selection, VTFS)、SelectKBest 特征选择(SelectKBest feature selection, SKB)、递归特征消除(recursive feature elimination, RFE)特征选择算法分别与主成分分析(PCA)算法相融合,结合 Savitzky-Golay 平滑(Savitzky-Golay smoothing, SG)、多元散射校正(multivariate scatter correction, MSC)、标准正态变量变换(standard normal variable transformation, SNV)、趋势校正(trend correction, TC)四种预处理方法,建立偏最小二乘回归(partial least squares regression, PLSR)、支持向量机回归(support vector machine regression, SVR)、梯度提升回归(gradient boosting regression, GBR)三种机器学习模型,比较分析不同特征选择方法在机器学习模型中对小麦籽粒类胡萝卜素含量的预测效果,确定最优模型,以期为小麦籽粒类胡萝卜素快速无损检测提供方法参考。

1 材料与方法

1.1 试验材料

试验以小麦品种济麦 22 和济麦 23 的种子作为研究对象,每个品种选取 1 kg 种子,去除颜色与外观有瑕疵的种子样品,每 50 粒种子作为一组样品,每个品种共获得 75 组样品。测定样品近红

外光谱信息,并将对应的小麦种子样品进行编号保存,便于后续进行理化值测定。

1.2　试验方法

1.2.1　近红外光谱的采集

采用自主研发的型号为 BIO-NIRDLP-HEM 手持式近红外光谱仪,采集小麦种子样本的近红外光谱数据。该仪器采用的分光系统为 DLP 分光系统,该系统可以通过控制一个镜列中的像素数量来改变到达探测器的光强度,并通过控制镜列的宽度来改变系统的分辨率[13]。图 1 为 DLP 分光系统示意图。DLP 技术利用数字微镜器件和单点探测器取代了传统的线性阵列探测器,相较于传统的近红外光谱仪具有检测速度较快、检测精度高、运行稳定等优点。该手持式近红外光谱仪的波长采集范围为 900～1 700 nm,共 228 个波长点[14]。测量参数为吸光度,每组样品扫描 10 次,共获得近红外光谱曲线 1 500 条,计算每组样品的平均光谱,最后将计算获得的平均光谱作为该样本的分析光谱[15]。利用 Python 3.9 版本的软件分析处理数据并进行后续建模分析。

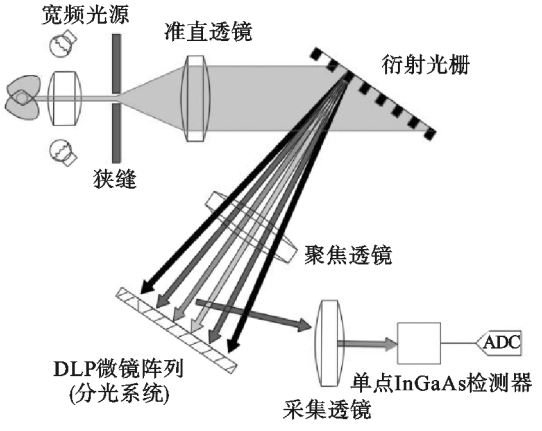

图 1　DLP 分光系统示意图

Fig. 1　Schematic diagram of DLP spectroscopic system

1.2.2　类胡萝卜素含量的测定

小麦样本近红外光谱数据采集结束后,使用规格型号为 TissueLyserⅡ的高效样品破碎仪,设置工作时间为 30 s,将小麦籽粒加工为全麦粉备用。然后使用规格型号为 XPR36 的高精度种子重量分析仪器称取 1.00 g 全麦粉,置于 15 mL 离心管中,加入 10 mL 95%乙醇溶液,避光保存。为了达到高效提取类胡萝卜素的目的,将样品进行振荡后静置,采用胶头滴管吸取上清液。以 95%乙醇为空白,利用型号为 UV-3600 的紫外可见分光光度计测定吸光度,每份样品设置 3 个重复。类胡萝卜素含量的计算[16]:$X=C\times V\div m$,其中 X 为类胡萝卜素含量($\mu g\cdot g^{-1}$),C 为吸光度对应浓度($\mu g\cdot mL^{-1}$),V 为类胡萝卜素溶液体积(mL),m 为全麦粉的质量(g)。

1.3　数据分析与模型构建

1.3.1　样本集划分

不同的样本集划分方法会导致模型预测效果产生差异[17,18]。为了提高数据质量,剔除光谱采集过程中存在异常的光谱曲线,挑选出 120 组样本,采用 HOM、KFCV 和 TSS 三种样本集划分方法将其按照 7∶3 比例划分为校正集和预测集,比较不同划分方法在三种机器学习模型中的效果。其中,HOM 算法将样本集随机分成校正集和预测集,具有实现简单、计算效率高等优点;KFCV 算法通过多次划分样本集并进行模型训练和评估,可以更准确地估计模型的性能,减少过拟合的风险[19];TSS 算法按照时间顺序将数据划分为校正集和预测集,可以有效提高模型的预测精度和泛化能力[20]。

1.3.2　光谱预处理

近红外光谱数据容易受到外部条件、测量环境以及仪器本身的噪声干扰,预处理可以有效地去除光谱中的干扰信息[21]。通过采用 SG、MSC、SNV 和 TC 四种预处理方法对原始光谱进行预处理。其中,SG 预处理可以消除随机噪声对光谱产生的影响;MSC 预处理可以增强光谱吸收信息,提高信噪比;SNV 预处理可以消除散射和基线偏移,突出光谱特征[22];TC 预处理可以使光谱数据更平稳,突出局部特征[23]。

1.3.3　特征选择

特征选择是数据处理过程中至关重要的一步,选择适合的特征选择方法,可以剔除原始光谱中的无效信息,进一步提高模型预测精度。采用了 VTFS 特征选择、SKB 特征选择、RFE 特征选择算法分别与 PCA 算法相融合的三种特征选择方法对光谱数据进行特征筛选,通过比较不同特征选择算法在模型中的预测效果,确定小麦类胡萝卜素最佳的特征选择方法。其中,VTFS 特征选择算法通过设定变量阈值,计算特征的统计量与阈值比较来筛选特征[24];SKB 特征选择算法可以依据特征与目标变量间的统计显著性,衡量特征对目标的贡献程度,筛选出关键特征;RFE 特征选择算法基于模型性能迭代剔除不重要特征,从而排除冗余特征[25]。

1.3.4 模型构建

采用 PLSR、SVR、GBR 三种机器学习模型对所采集的小麦近红外光谱数据进行建模分析。其中，PLSR 模型是一种高效处理数据的方法，可以同时考虑化学值和光谱曲线之间的基本关系，具有降维、减少过拟合等优点[26]。SVR 模型是一种基于统计学习理论和结构风险最小化的机器学习方法，具有较好的泛化能力，适合小样本、高维、非线性数据集的回归问题[27]。GBR 模型是一种强大的集成学习方法，具有高预测精度和灵活性，可以很好地处理非线性、非平滑关系的问题[28]。

1.3.5 模型评价

采用决定系数(R^2)和均方根误差(RMSE)评价模型建模和预测精度[29]。剩余预测偏差(RPD)也可用于评价模型对小麦类胡萝卜素含量的预测能力。当 RPD 值小于 1.5 时，模型预测结果较差，准确性较低；当 RPD 值为 1.5～2 时，预测结果一般，模型需要进一步优化；当 RPD 值为 2～3 时，预测结果较好，模型具有一定的准确性；当 RPD 值大于 3 时，预测结果非常好，模型具有很高的准确性和可靠性，可以实现高精度的预测[30]。

2 结果与分析

2.1 小麦籽粒近红外光谱曲线特征及预处理效果

从图 2 可以看出，小麦的原始光谱曲线分布在 900～1 700 nm 范围间，虽然小麦样品的平均光谱曲线存在一定的差异，但整体变化趋势大致相同，相似度较高且重叠现象明显。光谱曲线具有明显的吸收峰，而且与一些化学成分含量之间存在相关性[31,32]。波长区间 960～980、1 400～1 420 nm 内含有 O－H 基团，1 150～1 220、1 360～1 390 nm 含有 C－H 基团，1 490～1 540 nm 含有 N－H 基团，类胡萝卜素是具有共轭双键长链结构的化合物，分子中的 C－H、O－H 等基团在近红外光区域有特定的吸收。因此，近红外光谱技术可以被用来分析与这些基团有直接或间接关系的相关成分，小麦样品的光谱数据也可以用于类胡萝卜素含量预测模型构建。图 3～图 6 为四种方法预处理后的光谱曲线。与原始光谱曲线相比，SG 预处理后曲线整体表现较平滑；MSC 预处理后曲线特征峰变得更加一致；SNV 预处理后曲线峰形更加明显；TC 预处理后曲线的光谱信号更加突出。总之，四种预处理方法各有特点，均可用于处理数据。

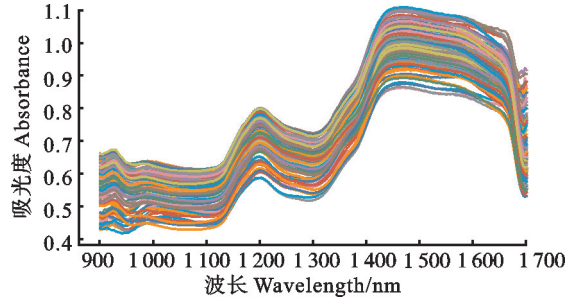

图 2 平均近红外光谱曲线

Fig. 2 Average near-infrared spectral curve

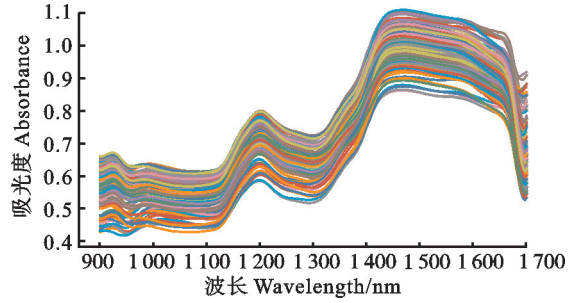

图 3 SG 预处理光谱曲线

Fig. 3 SG pre-processing spectral curve

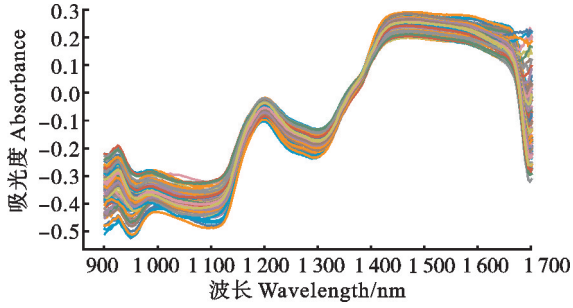

图 4 MSC 预处理光谱曲线

Fig. 4 MSC pre-processing spectral curve

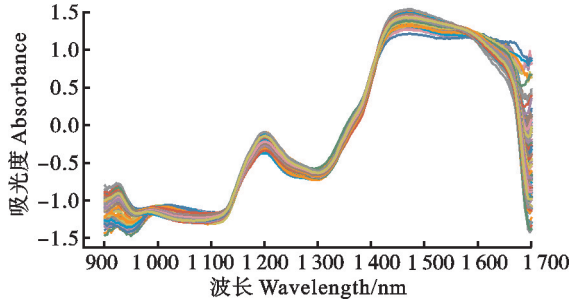

图 5 SNV 预处理光谱曲线

Fig. 5 SNV pre-processing spectral curve

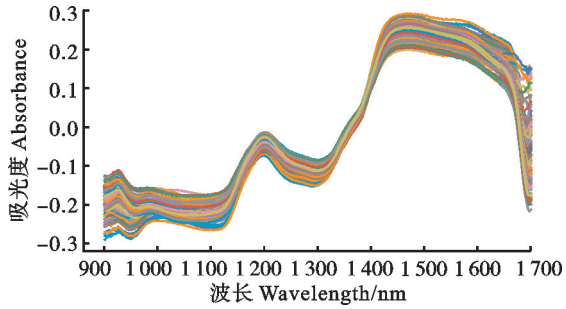

图 6　TC 预处理光谱曲线

Fig. 6　TC pre-processing spectral curve

2.2　原始光谱数据下不同样本集划分方法的比较

使用 HOM、KFCV 和 TSS 三种样本集划分方法将原始光谱数据(raw spectrum data,RSD)划分为校正集和预测集,分别建立 PLSR、SVR、GBR 三种机器学习模型,对小麦类胡萝卜素含量进行预测并比较模型预测精度。从模型的预测结果(表 1)看,HOM 方法在三个模型中的表现明显优于 KFCV 和 TSS 两种方法。在 SVR 模型中,HOM 方法预测集 R^2 为 0.62,预测效果高于其他两种模型。在 GBR 模型中,TSS 方法的校正集 R^2 可达到 0.98,但是在预测集中表现较差;HOM 方法校正集 R^2 为 0.90,预测集 R^2 明显高于其余两种样本集划分方法。因此,选择 HOM 方法划分样本集,构建小麦类胡萝卜素含量的预测模型。

2.3　不同预处理下全波长模型的精度比较

利用 HOM 方法按照 7∶3 的比例将样本集划分为校正集和预测集。选择 SG、MSC、SNV、TC 四种预处理方法对原始光谱数据进行处理,比较原始光谱数据和不同预处理后光谱数据在 PLSR 模型中的预测效果,选择最佳预处理方法。从使用不同预处理光谱数据建立的 PLSR 模型结果来看(表 2),使用 SG 预处理后的近红外光谱数据建立模型后得到的预测效果优于原始光谱数据和 MSC、SNV、TC 三种预处理光谱数据。使用 SG 预处理数据建立模型后校正集和预测集 R^2 可分别为 0.66 和 0.59,RMSE 值分别为 0.56 和 0.72 $\mu g \cdot g^{-1}$。相比原始光谱数据建模效果,预测集 R^2 提升了 0.06。相比于使用 MSC、SNV、TC 四种预处理数据建模后,SG 预处理后预测集 R^2 分别提升了 0.39、0.14 和 0.15。因此,可采用 SG 预处理后的光谱数据建模,对小麦类胡萝卜素含量进行预测。

表 1　不同样本集划分方法下模型的精度比较

Table 1　Comparison of model accuracy under different sample set partition methods

样本集划分方法 Sample set partitioning method	模型 Model	校正集 Validation set		预测集 Prediction set	
		R^2	RMSE/($\mu g \cdot g^{-1}$)	R^2	RMSE/($\mu g \cdot g^{-1}$)
HOM	PLSR	0.67	0.60	0.53	0.66
	SVR	0.65	0.59	0.62	0.65
	GBR	0.90	0.31	0.50	0.73
KFCV	PLSR	0.47	0.76	0.36	0.80
	SVR	0.58	0.66	0.26	0.78
	GBR	0.95	0.20	0.31	0.74
TSS	PLSR	0.58	0.56	0.33	0.78
	SVR	0.64	0.68	0.28	0.82
	GBR	0.98	0.12	0.36	0.77

表 2　不同预处理方法的 PLSR 预测模型

Table 2　PLSR prediction models with different pre-processing methods

预处理方法 Pre-processing method	校正集 Validation set		预测集 Prediction set	
	R^2	RMSE/($\mu g \cdot g^{-1}$)	R^2	RMSE/($\mu g \cdot g^{-1}$)
RSD	0.67	0.60	0.53	0.66
SG	0.66	0.56	0.59	0.72
MSC	0.54	0.70	0.20	0.86
SNV	0.45	0.71	0.45	0.71
TC	0.54	0.68	0.44	0.78

2.4 特征选择方法对建模效果的影响

将 VTFS、RFE、SKB 特征选择算法分别与 PCA 算法融合后对光谱数据进行特征波段选择，并建立 PLSR、SVR、GBR 模型。结果（表 3）表明，使用三种特征选择算法建立 PLSR 模型后，PCA-VTFS 特征选择后的数据预测效果最佳，校正集和预测集 R^2 分别为 0.71 和 0.63，RMSE 值分别为 0.53 和 0.69 μg·g^{-1}，RPD 为 0.75，相比于使用 PCA-REF、PCA-SKB 两种特征选择方法，PCA-VTFS 特征选择后预测集 R^2 分别提升了 0.27 和 0.06。使用三种特征选择算法建立 SVR 模型后，PCA-SKB 特征选择后的数据预测效果最佳，校正集和预测集 R^2 分别为 0.68 和 0.67，RMSE 值分别为 0.55 和 0.64 μg·g^{-1}，RPD 为 1.74；相比于使用 PCA-VTFS、PCA-REF 特征选择方法，PCA-SKB 特征选择后预测集 R^2 分别提升了 0.01 和 0.46。使用三种特征选择算法建立 GBR 模型后，PCA-SKB 特征选择后的数据预测效果最佳，校正集和预测集 R^2 分别为 0.99 和 0.89，RMSE 值分别为 0.03 和 0.34 μg·g^{-1}，RPD 为 3.01；相比于使用 PCA-VTFS、PCA-REF 两种特征选择方法，PCA-SKB 特征选择后预测集 R^2 分别提升了 0.11 和 0.16。因此，利用 PCA-VTFS、PCA-REF、PCA-SKB 三种特征选择方法分别建立 PLSR、SVR、GBR 模型后，PCA-SKB 特征选择在 SVR 和 GBR 模型中预测精度提升效果优于其他两种特征选择方法。相比 VTFS 特征选择和 RFE 特征选择后得到的位于 990、1 520、1 600 nm 附近光谱特征，SKB 特征选择后得到的 1 168、1 230、1 364、1 432 nm 附近光谱特征能够反映 C—H、O—H 等基团，对类胡萝卜素含量预测模型具有更为关键的意义。因此，SKB 特征选择方法结合 PCA 算法可以更好地保留光谱关键信息，提高模型的预测精度和稳定性。

表 3 特征选择方法下模型的精度比较

Table 3 Comparison of model accuracy under feature selection methods

特征选择方法 Feature selection method	模型 Model	校正集 Validation set		预测集 Prediction set		RPD
		R^2	RMSE/(μg·g^{-1})	R^2	RMSE/(μg·g^{-1})	
PCA-VTFS	PLSR	0.71	0.53	0.63	0.69	0.75
	SVR	0.84	0.37	0.66	0.69	1.72
	GBR	0.82	0.40	0.78	0.53	2.14
PCA-RFE	PLSR	0.55	0.70	0.36	0.79	0.83
	SVR	0.78	0.49	0.21	0.88	1.12
	GBR	0.99	0.08	0.73	0.53	1.92
PCA-SKB	PLSR	0.72	0.56	0.57	0.63	0.80
	SVR	0.68	0.55	0.67	0.64	1.74
	GBR	0.99	0.03	0.89	0.34	3.01

2.5 最优模型的建立

综合以上结果，采用 HOM 样本集划分方法，SG 预处理后的数据经过 PCA-SKB 特征选择后提取的光谱变量建立的 PLSR、SVR、GBR 三种模型中，GBR 模型预测精度最好，因而将 HOM-SG-PCA-SKB-GBR 作为样本划分、预处理、特征选择和建模算法的最优组合，用于小麦类胡萝卜素含量预测模型构建。从图 7 可以看出，所建立的 HOM-SG-PCA-SKB-GBR 模型，预测值和实测值线性拟合斜率接近 1，说明模型精度较高，可以用于小麦类胡萝卜素含量的预测。

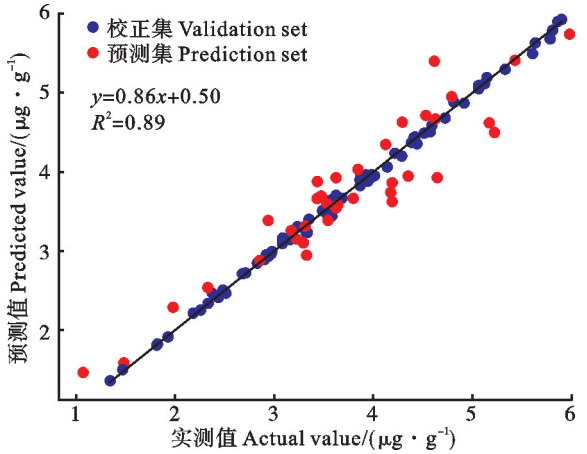

图 7 基于 HOM-SG-PCA-SKB-GBR 的小麦类胡萝卜素含量近红外预测模型散点图

Fig. 7 Scatter plot of near-infrared prediction model for carotenoid content in wheat based on HOM-SG-PCA-SKB-GBR

3 讨论

小麦在人类饮食中占有主要地位，小麦籽粒中类胡萝卜素含量与其品质密切相关。研究表明，

小麦籽粒中类胡萝卜素含量与全麦粉的黄度相关[33]。在测定植物类胡萝卜素方面,李文爽等[34]以 HPLC 测定方法为基础,建立了 UPLC 测定技术,能够更加准确地测定类胡萝卜素含量。但是 HPLC 和 UPLC 都属于化学计量方法,具有一定的局限性,难以实现类胡萝卜素含量的快速无损检测。

目前机器学习模型在农业领域得到了广泛应用。张北举等[35]建立了 Modified PLS 模型,可以快速准确地检测高粱籽粒中直链淀粉、支链淀粉的含量。吕都等[36]建立支持向量机判别模型,可以快速判别小麦霉菌污染。金秀等[37]对 9 种算法进行优化,然后利用随机森林、提升树和梯度提升树三种算法进行模型组合和二次优化,提升了土壤速效磷的预测精度。本研究使用 PLSR、SVR 和 GBR 三种模型对小麦籽粒中类胡萝卜素含量进行预测,结果表明,GBR 模型的预测效果明显优于 PLSR 模型和 SVR 模型,因为 GBR 模型具有灵活性和准确性,可以处理非线性的问题且易于调优。

特征选择对模型的预测效果十分重要。Badrouchi 等[38]利用 7 种不同的特征选择和 5 种模型组合来预测肾移植后的存活率,其中包括 SFM、RFE、SKB 等特征选择方法的比较。郝磊晓等[39]对牛奶中钠钾镁含量进行预测时,采用了多种特征选择算法来提高模型预测精度。项颂阳等[40]为了快速从大量高光谱图像中提取识别能力较好的特征,建立了 RelieF-RFE 特征选择算法。为了有效提取光谱数据中的有效信息,本研究采用 PCA-VTFS、PCA-SKB、PCA-RFE 三种特征选择算法对数据进行处理,结果表明,PCA-SKB 算法在 GBR 模型中对小麦类胡萝卜素含量的预测效果最佳。究其原因,在 PCA-SKB-GBR 模型中,PCA 算法对 SKB 特征选择后获得的相关特征会进一步进行筛选,从而降低了模型的复杂度,提升了模型精度;而 VTFS 特征选择和 RFE 特征选择算法会忽略特征之间可能存在的相关性,从而丧失部分重要的非线性关系特征,进而降低模型精度。此外,SKB 特征选择算法相对上述几种算法而言,可以降低过拟合风险且适应性较强,因此 PCA 算法与 SKB 算法融合后对光谱特征的筛选效果更好。

综上所述,基于近红外技术对小麦类胡萝卜素的预测研究中,该样本集的最优预测模型为 PCA-SKB-GBR。本研究针对小麦品种济麦 22 和济麦 23 进行类胡萝卜素含量的预测试验,SG 预处理后的数据经过 PCA-SKB 特征选择后用于预测类胡萝卜素含量可以获得较为满意的结果,对于其他样品,则需要进一步进行研究和验证。

参考文献:

[1]LI S,LUO J Y,ZHOU X L,*et al*. Identification of characteristic proteins of wheat varieties used to commercially produce dried noodles by electrophoresis and proteomics analysis [J]. *Journal of Food Composition and Analysis*, 2021, 96: 103685.

[2]聂 森,马劭瑾,彭彦昆,等. 主要粮食品质快速光学检测技术与装备研究进展[J]. 农业机械学报,2022,53(11):2.
NIE S,MA S J,PENG Y K,*et al*. Research progress of rapid optical detection technology and equipment for grain quality [J]. *Transactions of the Chinese Society for Agricultural Machinery*,2022,53(11):2.

[3]赵 欣,梁克红,朱 宏,等.类胡萝卜素含量对谷物蒸煮品质的影响[J]. 中国食品学报,2022,22(9):398.
ZHAO X,LIANG K H,ZHU H,*et al*. Effects of carotenoid content on cooking quality of grain [J]. *Journal of Chinese Institute of Food Science and Technology*,2022,22(9):398.

[4]孙延芳,王成社,杨进荣,等. 硬粒小麦类胡萝卜素含量的定量分析[J]. 西北农林科技大学学报(自然科学版),2007,35(5):103.
SUN Y F,WANG C S,YANG J R,*et al*. Study on quantitative analysis of total carotenoids content in *Triticum durum* [J]. *Journal of Northwest A & F University*(*Natural Science Edition*),2007,35(5):103.

[5]ZHANG J,GUO Z,REN Z S,*et al*. Rapid determination of protein,starch and moisture content in wheat flour by near-infrared hyperspectral imaging [J]. *Journal of Food Composition and Analysis*,2023,117:105134.

[6]KAMBOJ U,GUHA P,MISHRA S. Comparison of PLSR, MLR,SVM regression methods for determination of crude protein and carbohydrate content in stored wheat using near infrared spectroscopy [J]. *Materials Today*: *Proceedings*, 2022,48(P3):576.

[7]ZHAO Y Q,DENG J H,CHEN Q S,*et al*. Near-infrared spectroscopy based on colorimetric sensor array coupled with convolutional neural network detecting zearalenone in wheat [J]. *Food Chemistry*:*X*,2024,22:101322.

[8]ZHAO Y Q,ZHU C Y,JIANG H. Quantitative detection of Zearalenone in wheat using intervals selection coupled to near-infrared spectroscopy [J]. *Infrared Physics & Technology*, 2024,136:105004.

[9]SHI S J,FENG J H,MA Y Y,*et al*. Rapid determination of two illegal additives in wheat flour by near-infrared spectroscopy and different key wavelength selection algorithms [J]. *LWT*,2023,189:115437.

[10]DAMIÁN M V, RAFAEL F, MARIA TERESA B D, *et al*. Application of near-infrared reflectance spectroscopy for predicting carotenoid content in summer squash fruit [J]. *Computers and Electronics in Agriculture*, 2014, 108: 78.

[11]BERARDO N, BRENNA O V, AMATO A, *et al*. Carotenoids concentration among maize genotypes measured by near infrared reflectance spectroscopy (NIRS) [J]. *Innovative Food Science & Emerging Technologies*, 2004, 5(3): 397.

[12]BONIERBALE M, GRÜNEBERG W, AMOROS W, *et al*. Total and individual carotenoid profiles in *Solanum phureja* cultivated potatoes: II. Development and application of near-infrared reflectance spectroscopy (NIRS) calibrations for germplasm characterization [J]. *Journal of Food Composition and Analysis*, 2009, 22(6): 515.

[13]褚小立, 李亚辉. 近红外光谱实战宝典[M]. 北京: 化学工业出版社, 2023.
CHU X L, LI Y H. Near-infrared spectroscopy [M]. Beijing: Chemical Industry Press, 2023.

[14]刘燕德, 黎丽莎, 李 斌, 等. 多品种苹果可溶性固形物近红外无损检测通用模型研究[J]. 华中农业大学学报(自然科学版), 2022, 41(2): 238.
LIU Y D, LI L S, LI B, *et al*. General near-infrared model of soluble solids content in multi-variety apples [J]. *Journal of Huazhong Agricultural University* (*Natural Science Edition*), 2022, 41(2): 238.

[15]康明月, 罗 斌, 周亚男, 等. 基于近红外光谱技术结合改进的 CS-BPNN 樱桃番茄 SSC 和 Vc 含量检测[J]. 现代食品科技, 2023, 39(8): 288.
KANG M Y, LUO B, ZHOU Y N, *et al*. The detection of SSC and Vc content in cherry tomatoes based near infrared spectroscopy combined with improved CS-BPNN [J]. *Modern Food Science and Technology*, 2023, 39(8): 288.

[16]吴媛媛, 周 健, 包晓婷, 等. 基因型和环境对小麦类胡萝卜素含量及其品质性状的影响[J]. 麦类作物学报, 2015, 35(9): 1258.
WU Y Y, ZHOU J, BAO X T, *et al*. Effect of genotypes and environments to carotenoid contents and some quality traits of wheat varieties [J]. *Journal of Triticeae Crops*, 2015, 35(9): 1258.

[17]詹雪艳, 赵 娜, 林兆洲, 等. 校正集选择方法对于积雪草总苷中积雪草苷 NIR 定量模型的影响[J]. 光谱学与光谱分析, 2014, 34(12): 3270.
ZHAN X Y, ZHAO N, LIN Z Z, *et al*. Effect of algorithms for calibration set selection on quantitatively determining asiaticoside content in *Centella* total glucosides by near infrared spectroscopy [J]. *Spectroscopy and Spectral Analysis*, 2014, 34(12): 3270.

[18]陈奕云, 齐天赐, 黄颖菁, 等. 土壤有机质含量可见-近红外光谱反演模型校正集优选方法[J]. 农业工程学报, 2017, 33(6): 109.
CHEN Y Y, QI T C, HUANG Y J, *et al*. Optimization method of calibration dataset for VIS-NIR spectral inversion model of soil organic matter content [J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2017, 33(6): 109.

[19]OUF S, ELSEDDAWY A I B. A proposed paradigm for intelligent heart disease prediction system using data mining techniques [J]. *Journal of Southwest Jiaotong University*, 2021, 56(4): 229.

[20]RATNASARI A P, SUSETYO B, NOTODIPUTRO K A. Comparison of double random forest and long short-term memory methods for analyzing economic indicator data [J]. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 2023, 17(2): 762.

[21]刘文政, 周雪健, 平凤娇, 等. 基于可见-近红外光谱的鲜食葡萄成熟品质关键指标检测[J]. 农业机械学报, 2024, 55(2): 375.
LIU W Z, ZHOU X J, PING F J, *et al*. Detection of key indicators of ripening quality in table grapes based on visible-near-infrared spectroscopy [J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2024, 55(2): 375.

[22]张朱珊莹, 顾瀚文, 谢凯文, 等. 基于近红外光谱的预处理及组合方法[J]. 激光与光电子学进展, 2021, 58(16): 473.
ZHANG Z, GU H W, XIE K W, *et al*. Pretreatment and combined method based on near infrared spectroscopy [J]. *Laser & Optoelectronics Progress*, 2021, 58(16): 473.

[23]朱思聪, 高西娅, 张朱珊莹, 等. 红外光谱数据集划分比例及预处理方法研究[J]. 分析化学, 2022, 50(9): 1416.
ZHU S C, GAO X Y, ZHANG Z, *et al*. Partitioning proportion and pretreatment method of infrared spectral dataset [J]. *Chinese Journal of Analytical Chemistry*, 2022, 50(9): 1416.

[24]WANG W H, LU L X, WEI W. A novel supervised filter feature selection method based on Gaussian probability density for fault diagnosis of permanent magnet DC motors [J]. *Sensors*, 2022, 22(19): 7121.

[25]LIU R, TAN F, WANG Y X, *et al*. Machine learning identification of saline-alkali-tolerant japonica rice varieties based on raman spectroscopy and Python visual analysis [J]. *Agriculture*, 2022, 12(7): 1048.

[26]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): 388.
ZHAO J, SHEN M S, PU Y G, *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): 388.

[27]YU S, HUAN K W, LIU X X. Application of quantitative non-destructive determination of protein in wheat based on pretreatment combined with parallel convolutional neural network [J]. *Infrared Physics and Technology*, 2023, 135: 104958.

[28]祝元丽,冯向阳,闫庆武,等. 基于 GBDT 的望奎县农田土壤有机碳主控因子研究[J]. 中国环境科学,2024,44(3):1407.
ZHU Y L,FENG X Y,YAN Q W,*et al*. Spatial distribution and main controlling factors of soil organic carbon under cultivated land based on GBDT model in black soil region of northeast China [J]. *China Environmental Science*,2024,44(3):1407.

[29]YANG Z Y,CHENG Z,SU P Y,*et al*. A model for the detection of β-glucan content in oat grain based on near infrared spectroscopy [J]. *Journal of Food Composition and Analysis*,2024,129:106105.

[30]ZHAO X X,SONG Y T,ZHANG Y P,*et al*. Predictions of milk fatty acid contents by mid-infrared spectroscopy in Chinese holstein cows [J]. *Molecules*,2023,28(2):666.

[31]郭文川,朱德宽,张 乾,等. 基于近红外光谱的掺伪油茶籽油检测[J]. 农业机械学报,2020,51(9):352.
GUO W C,ZHU D K,ZHANG Q,*et al*. Detection on adulterated oil-tea camellia seed oil based on near-infrared spectroscopy [J]. *Transactions of the Chinese Society for Agricultural Machinery*,2020,51(9):352.

[32]刘倩男,黄 伟,丁云花,等. 青花菜中硫代葡萄糖苷 RAA 和 GBC 的近红外光谱快速测定[J]. 中国农业科学,2020,53(21):4500.
LIU Q N,HUANG W,DING Y H,*et al*. Rapid determination of RAA and GBC in broccoli by near infrared spectroscopy [J]. *Scientia Agricultura Sinica*,2020,53(21):4500.

[33]郑文寅,汪 帆,司红起,等. 普通小麦籽粒 LOX、PPO 活性和类胡萝卜素含量变异及对全麦粉色泽的影响[J]. 中国农业科学,2013,46(6):1088.
ZHENG W Y,WANG F,SI H Q,*et al*. Variations of LOX and PPO activities and carotenoid content as well as their influence on whole flour color in common wheat [J]. *Scientia Agricultura Sinica*,2013,46(6):1088.

[34]李文爽,夏先春,何中虎. 普通小麦类胡萝卜素组分的超高效液相色谱分离方法[J]. 作物学报,2016,42(5):711.
LI W S,XIA X C,HE Z H. Establishment of ultra performance liquid chromatography(UPLC) protocol for analyzing carotenoids in common wheat [J]. *Acta Agronomica Sinica*,2016,42(5):711.

[35]张北举,陈松树,李魁印,等. 基于近红外光谱的高粱籽粒直链淀粉、支链淀粉含量检测模型的构建与应用[J]. 中国农业科学,2022,55(1):33.
ZHANG B J,CHEN S S,LI K Y,*et al*. Construction and application of detection model for amylose and amylopectin content in *Sorghum* grains based on near infrared spectroscopy [J]. *Scientia Agricultura Sinica*,2022,55(1):33.

[36]吕 都,唐健波,赵绪婷,等. 小麦霉菌污染支持向量机判别模型的建立[J]. 食品研究与开发,2021,42(18):140.
LÜ D,TANG J B,ZHAO X T,*et al*. Rapid identification of mold contamination in wheat using support vector machine classification [J]. *Food Research and Development*,2021,42(18):140.

[37]金 秀,朱先志,李绍稳,等. 基于梯度提升树的土壤速效磷高光谱回归预测方法[J]. 激光与光电子学进展,2019,56(13):141.
JIN X,ZHU X Z,LI S W,*et al*. Predicting soil available phosphorus by hyperspectral regression method based on gradient boosting decision tree [J]. *Laser & Optoelectronics Progress*,2019,56(13):141.

[38]BADROUCHI S,AHMED A,MONGI BACHA M,*et al*. A machine learning framework for predicting long-term graft survival after kidney transplantation [J]. *Expert Systems with Applications*,2021,182:115235.

[39]郝磊晓,褚 楚,温佩佩,等. 基于中红外光谱的中国荷斯坦牛牛奶中钠钾镁含量预测模型的建立[J]. 中国农业科学,2024,57(14):2865.
HAO L X,CHU C,WEN P P,*et al*. Establishment of prediction models for sodium,potassium and magnesium content in milk of Chinese holstein cows based on mid-infrared spectroscopy [J]. *Scientia Agricultura Sinica*,2024,57(14):2865.

[40]项颂阳,许章华,张艺伟,等. 高光谱图像分类的 ReliefF-RFE 特征选择算法构建与应用[J]. 光谱学与光谱分析,2022,42(10):3284.
XIANG S Y,XU Z H,ZHANG Y W,*et al*. Construction and application of ReliefF-RFE feature selection algorithm for hyperspectral image classification [J]. *Spectroscopy and Spectral Analysis*,2022,42(10):3284.