高光谱和集成学习的黑枸杞快速分级方法

卢 伟[1]，蔡苗苗[1]，张 强[2]，李 珊[3]

1. 南京农业大学人工智能学院江苏省现代设施农业技术与装备工程实验室，江苏 南京 210031
2. 青海大学水利电力学院，青海 西宁 810016
3. 同济大学生命科学与技术学院，上海 200092

摘 要 黑枸杞具有较高的营养和医学价值，不同等级的黑枸杞质量不同，价格也具有显著差异，但由于缺乏有效的检测分级手段，造成黑枸杞市场鱼龙混杂、以次充好，影响黑枸杞市场质量监管。为实现黑枸杞快速、无损、高精度分级检测，提出基于高光谱和集成学习的黑枸杞快速无损分级方法。首先，选取诺木洪1级(NMH-grade1)、诺木洪2级(NMH-grade2)、诺木洪3级(NMH-grade3)、诺木洪4级(NMH-grade4)黑枸杞各200颗，在两种放置模式下(果柄朝上、去柄后整体横放)，通过GaiaSorter-Dual宽波段高光谱分选仪得到光谱范围为391.6～2 528.1 nm的光谱图像立方体。掩模处理后结合细胞计数算法实现单颗黑枸杞ROI高光谱信息的自动提取。考虑噪声的影响，截取500～2 400 nm范围内的黑枸杞光谱信息。经过FD(first derivative)，FFT(fast Fourier transform)、HT(hilbert transform)，SG(savitzky golay)，Normalize，SNV(standard normal variate)预处理后，再通过PCA(principal components analysis)，SPA(successie projection algorithm)，CARS(competitive adaptive reweighted sampling)提取特征波长的光谱信息。然后分别建立LIBSVM，LDA(latent dirichlet allocation)，KNN(k-nearest neighbor)，RF(random forest)，NB(naive Bayes)检测模型，其中，果肉-Normalize-SPA-LDA、果肉-FD-CARS-RF和果肉-SNV-CARS-LIBSVM组合方式最优，准确率分别为0.941 7，0.941 7和0.937 5。在预处理方法中，FD，HT，Normalize和SNV效果较好；在降维方法中，SPA和CARS的模型效果较好；在LIBSVM，LDA，KNN，RF和NB所建立的模型中，测试集精度不低于0.9的个数分别为2，7，0，4和1，因此LDA，RF和LIBSVM三个分类器效果最好。为进一步提高黑枸杞的分级精度，以LDA，RF，LIBSVM三个最优分类器为元模型，通过Stacking集成学习建立黑枸杞快速无损分级模型，使用果肉-FD-SPA-Stacking组合，精度可从0.941 7提升到0.983 3，此时共提取17个特征波长，分别为：591.6，609.1，721.6，989.1，1 083.3，1 111.3，1 296.1，1 564.9，1 844.9，1 934.5，1 996.1，2 046.5，2 130.5，2 292.9，2 315.3，2 320.9和2 348.9 nm，其中721.6，1 083.3，1 111.3，2 130.5，2 292.9，2 315.3，2 320.9和2 348.9 nm附近有C—H的倍频峰和吸收峰，721.6，989.1，1 934.5，1 996.1和2 292.9 nm附近有O—H的倍频峰和吸收峰，2 130.5和2 292.9 nm附近有C—O的吸收峰。研究结果表明基于高光谱结合集成学习进行黑枸杞快速无损分级是可行的。

关键词 光谱学；黑枸杞；集成学习；花青素；无损检测

中图分类号：S567；O433 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)07-2196-09

引 言

黑枸杞(*Lycium ruthenicum* Murr.)主要分布在我国内蒙古西部、宁夏、青海和西藏等地区，具有较高的营养和医学价值[1]。黑枸杞因其花青素含量较高，具有抗氧化、抗肿瘤、降低血栓等功能，其医学价值远高于普通的红枸杞。近几年来黑枸杞的价格逐渐上升，市场不断扩大，前景广阔。而不同等级的黑枸杞质量不同，其商业价值也不同，因此实现黑枸杞分级具有重要意义。目前分级方法仍以人工检测和化学方法为主，人工检测根据经验分级，检测精度较低，而化学方法速度较为缓慢，且具有破坏性，因此目前黑枸杞市

收稿日期：2020-05-16，**修订日期**：2020-09-21

基金项目：国家自然科学基金面上项目(32071896，31960487)，江苏省自然科学基金面上项目(BK20181315)，江苏省农业科技自主创新项目[CX(20)3068]，扬州市重点研发计划(现代农业)项目(YZ2018038)资助

作者简介：卢 伟，1978年生，南京农业大学副教授 e-mail：njaurobot@njau.edu.cn

场鱼龙混杂。现在急需一种快速、无损、高精度黑枸杞检测方法。

近红外技术将光谱测量与化学计量学有机结合，在农业种子质量检测等领域取得了广泛的应用[2]。但是该方法因无法获取图片信息，检测精度较低。而高光谱成像技术将光谱和图像结合，可以同时反映样品的外部纹理特征、内部结构以及化学成分，受到许多食品工业研究者的青睐。王磊[3]等利用近红外高光谱图像实现了宁夏枸杞产地的鉴别。于慧春[4]等采用高光谱图像技术对枸杞多糖和总糖含量进行检测，并探寻其最适宜的光谱波段。然而目前对于黑枸杞的研究较少。

为提高检测速度和精度，多元散射校正(multiplicative scatter correction，MSC)[5]、标准正态变量校正(standard normal variate，SNV)[5]等预处理方法被用来去噪；主成分分析(principal components analysis，PCA)[3]、竞争性自适应重加权算法(competitive adaptive reweighted sampling，CARS)[5]，连续投影算法(successie projection algorithm，SPA)[5]等方法被用来数据降维。再通过支持向量机(support vector machine，SVM)[5]、线性判别分析(latent dirichlet allocation，LDA)[3]、K 最近邻(k-nearest neighbor，KNN)[5]、随机森林(random forest，RF)[6]、朴素贝叶斯(naive Bayes，NB)[7]、BP(back propagation)等神经网络等弱分类器建立食品质量检测模型。而基于不同的预处理方法，降维方法和弱分类器建立的模型差异较大，需要进一步研究。为提高弱分类器的检测精度，深度学习[8]和集成学习[9]因其较强的特征学习能力而逐渐被引入食品质量的无损检测领域。但是，深度学习对计算机软硬件配置要求较高且资源消耗较大，计算速度很大程度上受限于计算机的性能。集成学习则通过将多个弱分类器融合成一个强分类器，来增强模型的学习和泛化能力[9]，且在提高模型预测精度的同时，对资源消耗和计算机软硬件的要求并未明显提高。因此本工作拟采用高光谱技术，并基于 Stacking 集成学习[9]实现黑枸杞品质的快速无损检测分级。

1　实验部分

1.1　材料

实验的黑枸杞均为 2019 年产自青海诺木洪地区，按照果实大小分为 4 个等级，从大到小分别为诺木洪 1 级(NMH-grade1)、诺木洪 2 级(NMH-grade2)、诺木洪 3 级(NMH-grade3)、诺木洪 4 级(NMH-grade4)。从每个等级中挑选大小颜色均匀、无明显缺陷、带有果柄的黑枸杞 200 颗，分别采集黑枸杞在两种放置模式下(果柄朝上、去柄后整体横放)的高光谱图像。

1.2　高光谱成像系统

GaiaSorter-Dual 宽波段高光谱分选仪如图 1 所示。仪器主要由均匀光源、宽波段光谱相机、大行程电控移动平台(传送带)、计算机及控制软件等部分构成。均匀光源由 4 个 400 W 溴钨灯和 4 个 800 W 溴钨灯组成。Image-λ“G”系列高光谱相机将分光元件与面阵列相结合，能够同时、快速获取光谱和影像信息。借助移动平台对样品实现线扫描。系统整体采用上下分体设计，使样品以传送带传输的方式实现同步宽波段检测，最终实现连续性测量。采集的光谱范围为 391.6～2 528.1 nm，在 391.6～1 044.1 nm 区间内分辨率为 2.5 nm，在 1 044.1～2 528.1 nm 区间内分辨率为 5.6 nm。

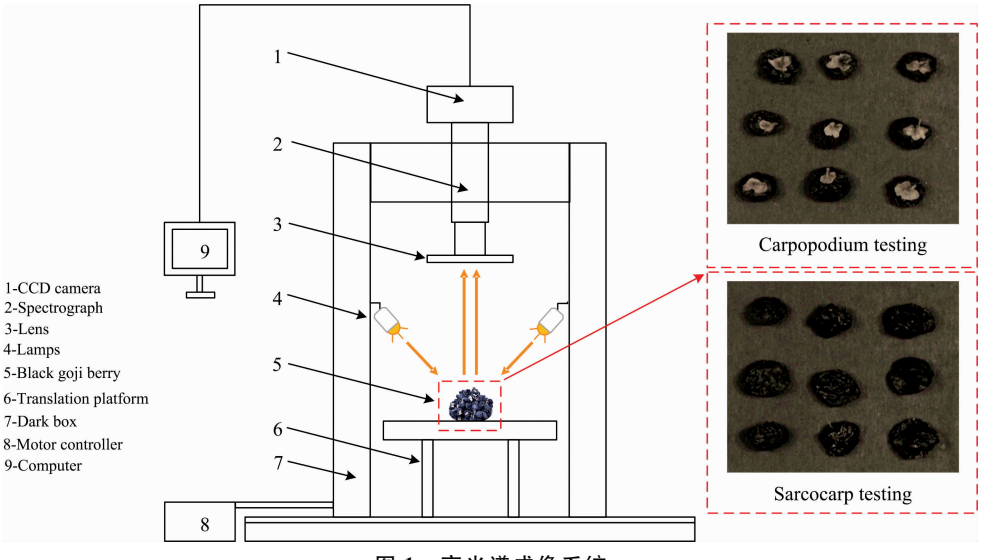

图 1　高光谱成像系统

Fig. 1　Hyperspectral imaging system

1.3　数据的采集与黑白校正

实验开始前，将仪器进行 30 min 的预热以确保实验的精准性。经过多次调整，将相机高度设置为 5 cm，曝光设置为 6 ms，传送带速度设置为 0.36 cm·s^{-1}。同时，对采集到的高光谱图像进行黑白校正以减少因暗电流或者光照不均匀等因素造成的影响[3]，校正公式为

$$R = \frac{I - D}{W - D} \quad (1)$$

其中，I 为原始高光谱图像；D 为黑帧(反射率接近 0)；W 为白帧(反射率接近 1)；R 为经过黑白校正后最终得到的光谱图像。校正工具为系统自带的 SpecVIEW 软件。

1.4 ROI 区域自动提取

通过高光谱成像系统，可得 1 681×1 024×527 的光谱图像立方体。其中共 1 681×1 024 个像素，每个像素点有 527 个光谱信息。为实现单颗黑枸杞果柄和果肉 ROI 区域高光谱信息的自动提取，利用 MATLAB 编写 ROI 提取算法如下(流程见图 2)。

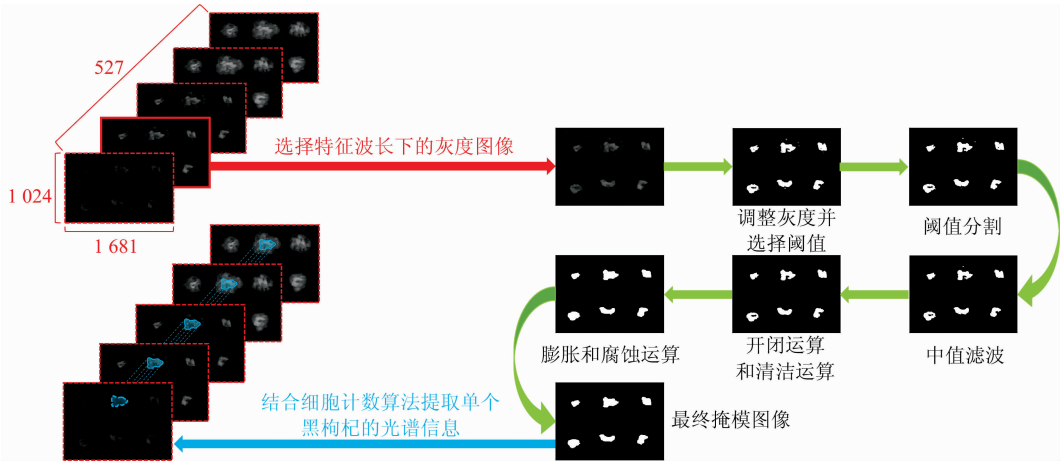

图 2 单颗黑枸杞 ROI 的自动提取

Fig. 2 Automatic ROI extraction in hyperspectral images

Step 1 ROI 掩模

首先，通过手动比较 20 颗随机挑选的黑枸杞的果柄、果肉和背景区域的平均反射信息(如图 3 所示)，选取果柄与果肉反射光差值最大的波长作为提取果柄的特征波长(1 094.5 nm)，再选取果肉与背景反射光差值最大的波长作为提取果肉的特征波长(1 111.3 nm)，分别在两个特征波长下的灰度图中通过阈值分割和中值滤波、开闭运算、清洁运算、膨胀和腐蚀等运算，获取果柄和果肉的 ROI 掩模图像。

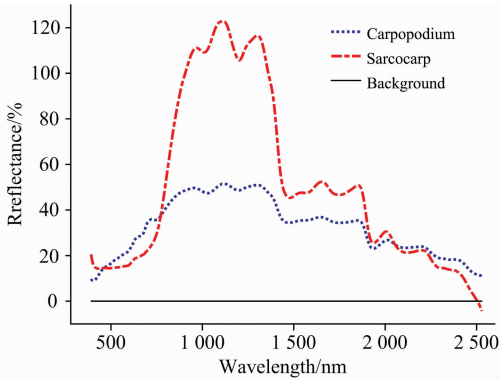

图 3 果柄、果肉和背景平均反射信息

Fig. 3 Average reflection informations of carpopodium, sarcocarp and background

Step 2 ROI 光谱信息的自动提取

通过细胞计数算法，将掩模图像中每颗黑枸杞果柄和果肉的位置提取出来，结合光谱图像立方体，提取对应位置下的光谱信息。最后利用每颗黑枸杞果柄和果肉的平均光谱信息建立模型。

1.5 黑枸杞花青素含量的测定

1.5.1 花青素的提取

采集完样本的高光谱信息后，再采用 pH 示差法测定黑枸杞内花青素的含量 H，具体方法见文献[10]。设置测定波长为 525 nm，温度为 40 ℃。在 pH 1.0 和 pH 4.5 的缓冲液稀释处测定吸光度，其平衡时间分别设定为 30 和 20 min。最后采用 Fuleki T 公式计算花青素含量 H。

1.6 基于弱分类器的建模

为减小仪器、测量环境等因素对模型结果造成的影响，首先进行 smooth 平滑去噪，再分别采用一阶导数(first derivative，FD)、快速傅里叶算法(fast Fourier transform，FFT)低通去噪[5]、希尔伯特变换法(Hilbert transform，HT)[5]、多项式平滑算法(savitzky golay，SG)、正规化(normalize)、标准正态变量校正(standard normal variate，SNV)六种方法进行预处理。其中经参数优选，FFT 算法中低通滤波器的截止频率设置为 0.125，SG 算法中阶数和窗口点数分别设置为 3 和 9，Normalize 算法中采用 0-1 正规化。

为降低计算消耗和提高运算速度，主成分分析(principal components analysis，PCA)、连续投影算法(successie projection algorithm，SPA)、竞争性自适应重加权算法(competitive adaptive reweighted sampling，CARS)三种常用的降维算法被用来提取特征波长[5]。PCA 旨在利用降维的思想，把多变量转化为少数几个综合指标。SPA 是一种变量选择技术，旨在消除变量共线性问题。CARS 将指数衰减函数和自适应重加权采样技术相结合建立 PLS(partial least squares)模型，去掉权重较小的波长点，利用交互验证选出 RMSECV(root mean square error of cross-validation)最低的子集，将其作为最优波长组合。

选用的分类器为 LIBSVM[11]、线性判别分析(latent dirichlet allocation，LDA)[3]、K 最近邻(k-nearest neighbor，KNN)[5]、随机森林(random forest，RF)[6]和朴素贝叶斯(naive Bayes，NB)[7]。LIBSVM 为台湾大学林智仁教授的算法，其利用交叉验证选择最优参数，可实现多分类。通过比较，LIBSVM 中交叉验证次数设为 5，KNN 中相邻数目设为 10，RF 中决策树数目设为 500。在使用分类器时，均将 70%的样本作为训练集，其余 30%的样本为测试集。

首先将采集到的高光谱图像进行 ROI 掩模处理，结合细胞计数算法获取单颗黑枸杞果柄和果肉的平均光谱信息。然后通过 FD，FFT，HT，SG，Normalize 和 SNV 进行预处理，再用 PCA，SPA，CARS 进行特征波长的提取。最后利用 LIBSVM，LDA，KNN，RF 和 NB 分类器建立黑枸杞分级模型。为进一步提高模型精度，通过 Stacking 集成学习将多个弱分类器融合成一个强分类器以提高黑枸杞分级模型的泛化能力[9]，其具体过程如图 4，总算法流程如图 5。

1.7　基于 Stacking 集成学习的建模

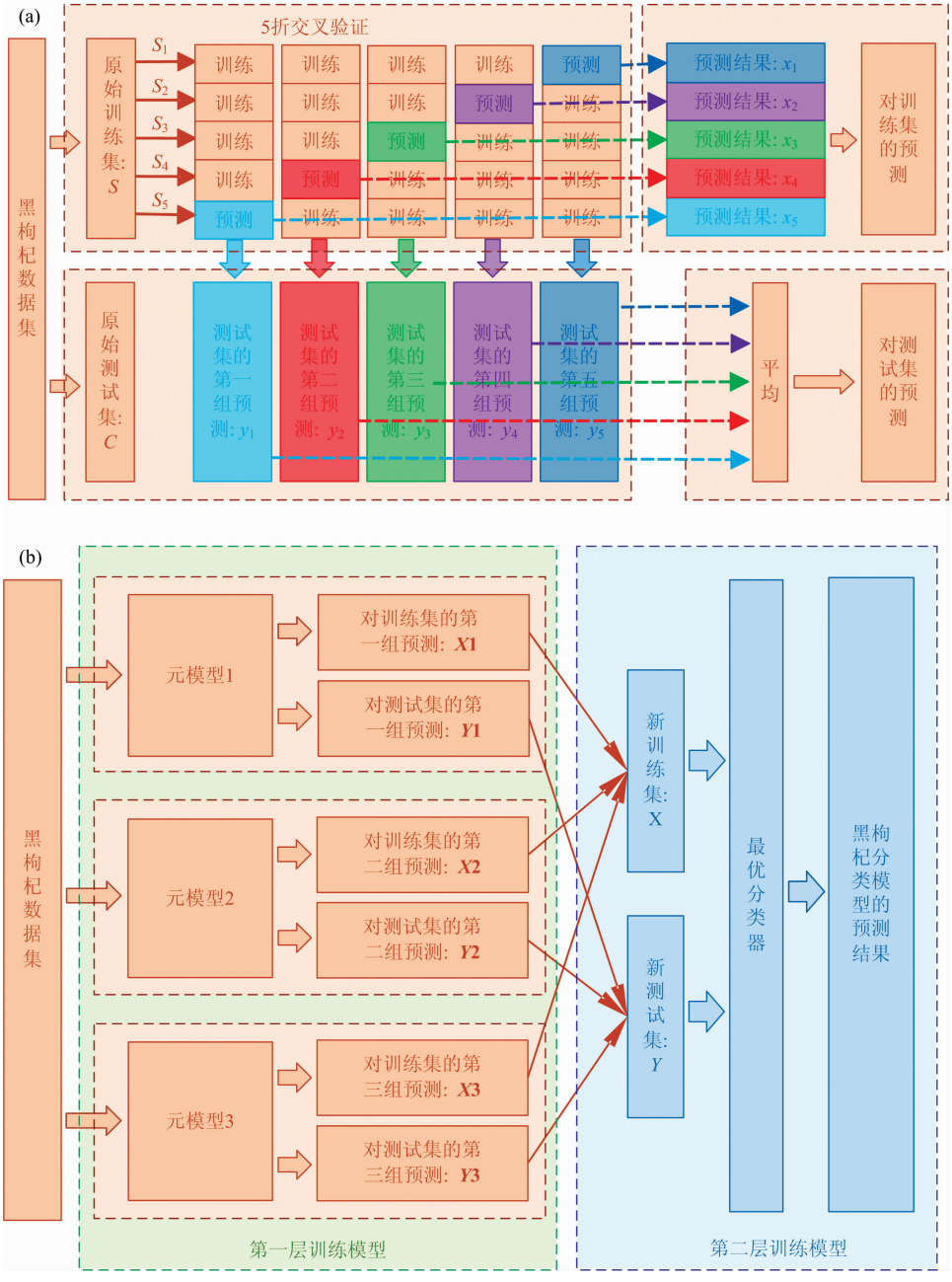

图 4　Stacking 集成学习过程

(a)：元模型的训练和预测模型；(b)：总流程图

Fig. 4　Stacking ensemble learning process

(a)：Training and prediction models for metamodels；(b)：General flowchart

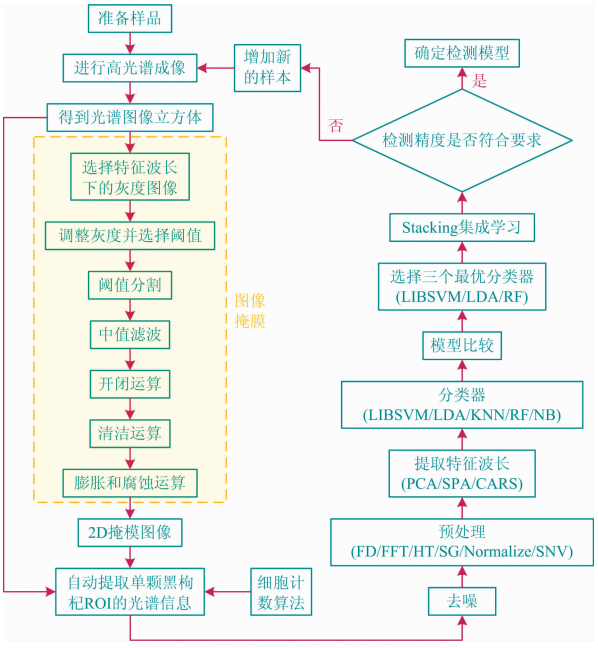

图 5 黑枸杞快速无损分级模型的流程图

Fig. 5 Flow chart of fast and non-destructive grading model of black goji berry

Stacking 集成学习采用两层训练结构。第一层利用不同的分类器构建不同的元模型，将所有元模型的预测结果进行整合，然后将其作为第二层的输入数据，最后对第二层进行训练。

(1)第一层：基于交叉验证思想的元模型的建立

首先选出建模效果最好的三个分类器构建三个元模型。每个元模型中，基于 5 折交叉验证的思想，对训练集和测试集预测。

黑枸杞原始数据集 $\boldsymbol{M}$ 划分为原始训练集 $\boldsymbol{S}$ 和原始测试集 $\boldsymbol{C}$。原始训练集 $\boldsymbol{S}$ 中，每类黑枸杞有 140 颗，共 560 颗；原始测试集 $\boldsymbol{C}$ 中，每类黑枸杞有 60 颗，共 240 颗。基于交叉验证的思想，将黑枸杞原始训练集 $\boldsymbol{S}$ 平均划分为 5 份，记 s_1—s_5。其中，每份有 112 个数据，每份每类黑枸杞有 28 个数据。首先用 s_2—s_5 训练第一个基分类器并且预测 s_1 和 $\boldsymbol{C}$，得到 s_1 的预测值 x_1 和 $\boldsymbol{C}$ 的第一次预测值 y_1；然后用 s_1，s_3—s_5 训练，得到 s_2 的预测值 x_2 和 $\boldsymbol{C}$ 的第二次预测值 y_2。以此类推，最终得到 s_1—s_5 的预测值 x_1—x_5 以及 $\boldsymbol{C}$ 的 5 次预测值 y_1—y_5。将 x_1—x_5 合并得到原始训练集 $\boldsymbol{S}$ 的预测值 $\boldsymbol{X}_1$，将 y_1—y_5 取均值得到原始测试集 $\boldsymbol{C}$ 的预测值 $\boldsymbol{Y}_1$。对剩下两个基分类器进行同样的操作得到 $\boldsymbol{S}$ 的预测结果 $\boldsymbol{X}_2$，$\boldsymbol{X}_3$ 和 $\boldsymbol{C}$ 的预测结果 $\boldsymbol{Y}_2$，$\boldsymbol{Y}_3$。

(2)第二层：利用新的训练集和新的测试集建立黑枸杞分级模型

将第一层得到的结果合并：$\boldsymbol{X}=\{X_1, X_2, X_3\}$，$\boldsymbol{Y}=\{Y_1, Y_2, Y_3\}$。其中，$\boldsymbol{X}$ 为新的训练集，$\boldsymbol{Y}$ 为新的测试集，选择建模效果最好的基分类器进行第二层模型的训练。

2 结果与讨论

2.1 黑枸杞花青素含量的测定结果

pH 示差法测定后的四个等级的黑枸杞花青素含量如表 1 所示。由表 1 可知，诺木洪的分级与花青素含量呈正相关，其中诺木洪 1 级花青素含量达到 30.57 mg·L^{-1}，诺木洪 4 级花青素含量达到 20.37 mg·L^{-1}，二者花青素含量差别显著。

表 1 黑枸杞花青素含量

Table 1 Anthocyanin content of four grades of black goji berry

等级名称	花青素含量 H/(mg·L^{-1})
NMH-grade1	30.57
NMH-grade2	24.32
NMH-grade3	23.44
NMH-grade4	20.37

2.2 光谱信息采集与预处理

测量波长范围为 391.6～2 528.1 nm，共 527 个波长，考虑噪声的影响，截取 500～2 400 nm 作为有用光谱信息，如图 6(a)和(b)所示。

对截取后的原始光谱进行平滑滤波，然后分别通过 FD，FFT，HT，SG，Normalize 和 SNV 方法进行预处理。其中，经过 FD 处理后的光谱曲线及其平均值如图 6(c)，(d)，(e)和(f)所示。FD 的主要思想是对原始光谱求导，进而放大不同样本间的差异。从图 6(c)和(d)可观察出不同黑枸杞之间的光谱差异主要在 580～640，700～1 500 和 1 840～2 100 nm 波段。

2.3 特征波长的提取

在 PCA 中，提取前 30 个成分特征值的贡献率，再选出贡献率大于 1%的主成分作为新坐标系，并将原始数据在新坐标系下解析，得到降维后数据。

在 SPA 中，设置 SPA 可选择波长数量范围为 5～35，步长为 1。

在 CARS 中，对原始光谱采用 CARS 降维，设定蒙特卡罗采样次数为 50，采用 5 折交叉验证法建立 PLS 模型。当某个采样次数所建立的 PLSR 模型达到最小 RMSECV 时，取在该采样次数下的波长作为特征波长。

2.4 基于弱分类器的黑枸杞分级结果

将果柄和果肉的光谱信息经 6 种预处理方法和 3 种特征提取方法去噪和降维后，分别建立 LIBSVM，LDA，KNN，RF 和 NB 模型，训练集和测试集精度如表 2 所示。模型判断标准以测试集精度为主：精度越高，模型效果越好。

由表 2 可见，采用果肉-Normalize-SPA-LDA、果肉-FD-CARS-RF 和果肉-SNV-CARS-LIBSVM 建立的分级模型最优，准确率分别为 0.941 7，0.941 7 和 0.937 5。

果柄光谱信息建立的模型中，有 2 个测试集精度大于等于 0.9，而果肉则有 12 个，表明果肉的光谱信息更有利于分

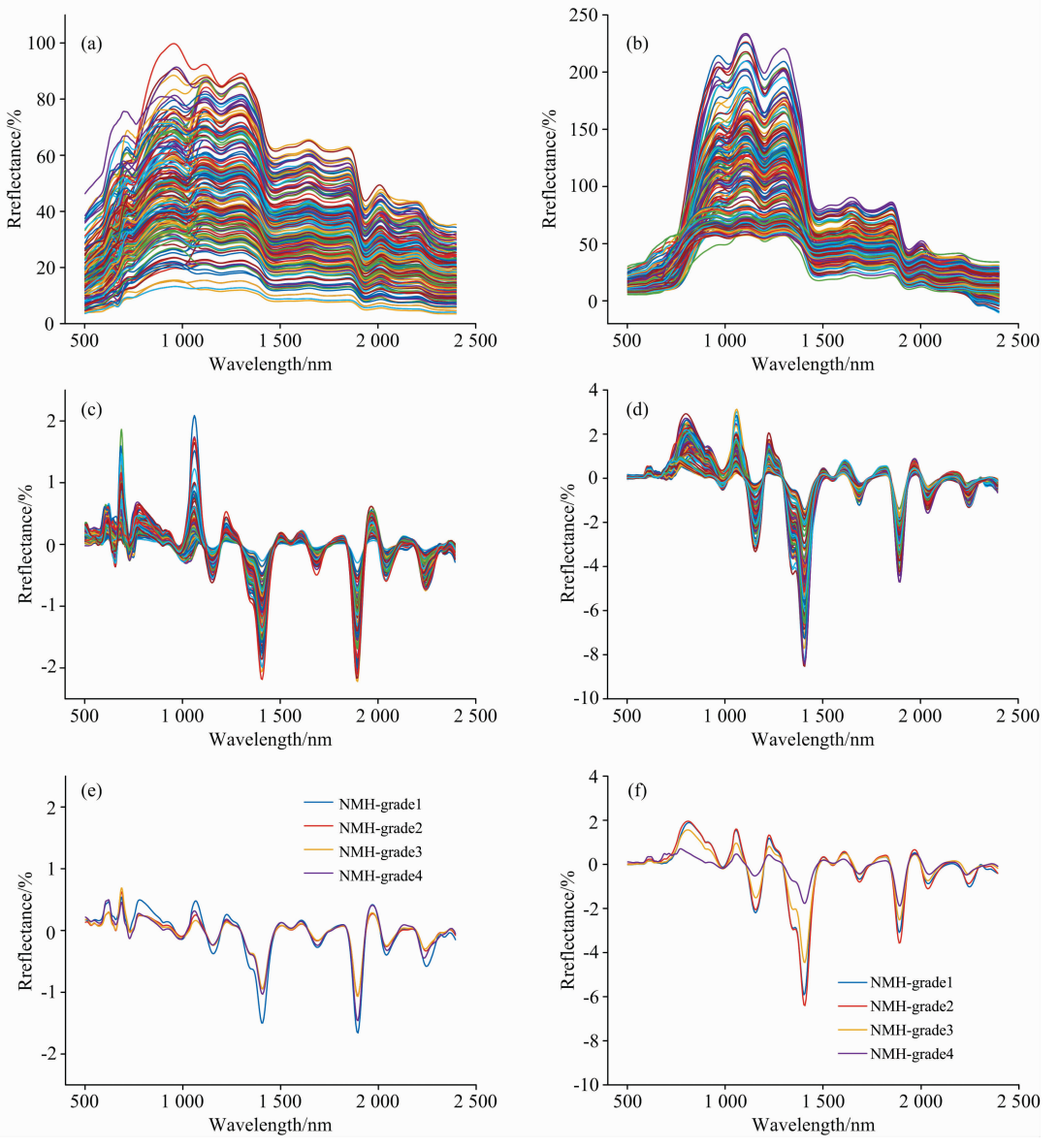

图 6　预处理前后黑枸杞的光谱曲线

(a)：果柄原始光谱曲线；(b)：果肉原始光谱曲线；(c)：FD 处理后果柄光谱曲线；(d)：FD 处理后果肉光谱曲线；(e)：FD 处理后不同等级黑枸杞果柄的平均光谱曲线；(f)：FD 处理后不同等级黑枸杞果肉的平均光谱曲线

Fig. 6　Spectral curves of black goji berries before and after pretreatment

(a)：Raw spectra of carpopodium；(b)：Raw spectra of sarcocarp；(c)：Spectra of carpopodium after FD treatment；(d)：Spectra of sarcocarp after FD treatment；(e)：Average spectra of different grades of black goji berrycarpopodium after FD treatment；(f)：Average spectra of different grades of black goji berrysarcocarp after FD treatment

表 2　基于 PCA 的建模结果

Table 2　PCA-based modeling results

特征提取方法		PCA		SPA		CARS	
分类器	预处理	果柄	果肉	果柄	果肉	0.783 3	**0.933 3**
LIBSVM	FD	0.800 0	0.779 2	0.666 7	0.883 3	0.300 0	0.250 0
	FFT	0.250 0	0.266 7	0.250 0	0.304 2	0.483 3	0.395 8
	HT	0.316 7	0.266 7	0.533 3	0.475 0	0.350 0	0.395 8
	SG	0.300 0	0.333 3	0.516 7	0.333 3	0.483 3	0.787 5
	Normalize	0.366 7	0.508 3	0.483 3	0.725 0	0.783 3	**0.937 5**
	SNV	0.683 3	0.775 0	0.766 7	0.895 8	0.516 7	0.800 0

续表 2

LDA	FD	0.733 3	**0.920 8**	0.650 0	**0.900 0**	0.366 7	0.679 2
	FFT	0.400 0	0.337 5	0.500 0	0.554 2	0.466 7	0.654 2
	HT	0.416 7	0.579 2	0.516 7	0.729 2	0.450 0	0.495 8
	SG	0.633 3	0.687 5	0.833 3	0.820 8	**0.912 5**	**0.933 3**
	Normalize	0.383 3	0.333 3	**0.912 5**	**0.941 7**	0.383 3	0.795 8
	SNV	0.716 7	0.691 7	0.816 7	**0.929 2**	0.700 0	0.812 5
KNN	FD	0.716 7	0.758 3	0.716 7	0.787 5	0.616 7	0.858 3
	FFT	0.366 7	0.383 3	0.516 7	0.595 8	0.600 0	0.775 0
	HT	0.416 7	0.520 8	0.516 7	0.816 7	0.683 3	0.845 8
	SG	0.666 7	0.737 5	0.683 3	0.829 2	0.333 3	0.566 7
	Normalize	0.383 3	0.475 0	0.300 0	0.491 7	0.633 3	0.812 5
	SNV	0.516 7	0.737 5	0.600 0	0.812 5	0.783 3	**0.941 7**
RF	FD	0.666 7	0.829 2	0.766 7	**0.912 5**	0.550 0	0.812 5
	FFT	0.483 3	0.245 8	0.583 3	0.800 0	0.583 3	**0.904 2**
	HT	0.450 0	0.550 0	0.500 0	**0.904 2**	0.683 3	0.854 2
	SG	0.566 7	0.716 7	0.683 3	0.875 0	0.350 0	0.612 5
	Normalize	0.176 7	0.575 0	0.383 3	0.570 8	0.733 3	0.762 5
	SNV	0.566 7	0.854 2	0.616 7	0.783 3	0.750 0	**0.900 0**
NB	FD	0.733 3	0.883 3	0.816 7	0.883 3	0.483 3	0.641 7
	FFT	0.483 3	0.433 3	0.433 3	0.629 2	0.483 3	0.829 2
	HT	0.516 7	0.608 3	0.466 7	0.891 7	0.633 3	0.750 0
	SG	0.683 3	0.691 7	0.566 7	0.750 0	0.250 0	0.366 7
	Normalize	0.233 3	0.658 3	0.333 3	0.354 2	0.666 7	0.800 0
	SNV	0.666 7	0.716 7	0.483 3	0.766 7	0.783 3	**0.933 3**

级。在预处理方法中，FD，HT，Normalize 和 SNV 效果较好，测试集精度不低于 0.9 的模型个数分别为 6，2，4 和 2，其中 FD 效果最好。在降维方法中，SPA 和 CARS 的模型效果较好，测试集精度大于等于 0.9 的模型分别有 6 个和 7 个，而 PCA 由于提取特征波长数量较少，难以反应完整的黑枸杞信息，建模效果欠佳。

在 LIBSVM，LDA，KNN，RF 和 NB 所建立的模型中，准确率不低于 0.9 的个数分别为 2，7，0，4 和 1，因此优选出 LDA，RF 和 LIBSVM 三个分类器用于 Stacking 集成学习。

2.5 基于 Stacking 集成学习的黑枸杞分级结果

选取 LDA，RF 和 LIBSVM 三个分类器作为第一层分类器，建模效果最好的分类器(LDA)作为第二层分类器，结果如表 3 所示。

可见，果肉-FD-SPA-Stacking 的建模效果最佳，可将测试集准确率从原来的 0.941 7 上升到 0.983 3。该模型提取的特征波长有 17 个，分别为(单位 nm)：591.6，609.1，721.6，989.1，1 083.3，1 111.3，1 296.1，1 564.9，1 844.9，1 934.5，1 996.1，2 046.5，2 130.5，2 292.9，2 315.3，2 320.9和 2 348.9。花青素是类黄酮化合物，以 C6-C3-C6 的 C 骨架为基本结构，含有 C—H，O—H，C—O 等化学键[12]。黑枸杞除了含有花青素，还有蛋白质、水分、糖类、脂肪等成分[1]。在提取的特征波长中，721.6 nm 附近有 C—H 五倍频峰、H_2O 四倍频峰以及 O—H 四倍频峰；989.1 nm 附近有 H_2O 三倍频峰；1 083.3 nm 附近有 C—H 第三组组合频峰；1 111.3 nm 附近有 C—H 三级倍频峰；1 934.5 nm 附近有水分 O—H 一级倍频峰；1 996.1 nm 附近有 O—H 一倍频峰；2 130.5 nm 附近有蛋白的 C—H 和 C—O 组合吸收峰；2 292.9 nm 附近有糖类的 O—H 和 C—O 组合吸收峰；2 315.3和 2 320.9 nm 附近有油分的 C—H 键伸缩一级倍频峰；2 348.9 nm 附近有纤维素 C—H 键伸缩振动一级倍频峰。因此，基于高光谱结合集成学习算法建立的黑枸杞无损检测模型可以较好地反应黑枸杞内部的生化参数。

表 3 Stacking 集成学习的建模结果

Table 3 Modeling results of Stacking ensemble learning

特征提取方法	预处理	训练集		测试集	
		果柄	果肉	果柄	果肉
PCA	FD	0.944 6	0.955 4	0.766 7	**0.908 3**
	FFT	0.326 8	0.300 0	0.316 7	0.262 5
	HT	0.598 2	0.337 5	0.383 3	0.312 5
	SG	0.557 1	0.726 8	0.616 7	0.687 5
	Normalize	0.408 9	0.935 7	0.350 0	0.512 5
	SNV	0.930 4	0.935 7	0.683 3	0.762 5
SPA	FD	0.837 5	0.975 0	0.750 0	**0.983 3**
	FFT	0.632 1	0.678 6	0.533 3	0.537 5
	HT	0.553 6	0.846 4	0.516 7	0.729 2
	SG	0.905 4	0.914 3	0.833 3	0.820 8
	Normalize	0.364 3	0.817 9	0.416 7	0.633 3
	SNV	0.955 4	0.950 0	0.816 7	**0.912 5**
CARS	FD	0.980 4	0.894 6	0.500 0	0.800 0
	FFT	0.455 4	0.554 2	0.266 7	0.483 3
	HT	0.728 6	0.782 1	0.450 0	0.604 2
	SG	0.450 0	0.612 5	0.300 0	0.404 2
	Normalize	0.941 1	0.950 0	**0.916 7**	0.804 2
	SNV	0.926 8	0.962 5	0.733 3	**0.916 7**

3　结　论

采用高光谱技术结合 Stacking 集成学习实现黑枸杞快速无损分级。首先，采集黑枸杞在两种放置模式下(果柄朝上、去柄后整体横放)的高光谱图像，通过掩模处理自动提取单颗黑枸杞果柄和果肉的图谱信息。经 FD，FFT，HT，SG，Normalize 和 SNV 进行预处理后，通过 PCA，SPA 和 CARS 获取特征波长下的光谱信息，再比较 LIBSVM，LDA，KNN，RF 和 NB 的分类效果，优选 LDA，RF 和 LIBSVM 三个分类器。结果表明，与果柄相比，黑枸杞果肉信息的建模精度更高，其中，果肉-Normalize-SPA-LDA、果肉-FD-CARS-RF 和果肉-SNV-CARS-LIBSVM 建立的模型最优，精度分别为 0.941 7，0.941 7 和 0.937 5。为进一步提高分类精度，建立 LDA，RF 和 LIBSVM 元模型，构建基于 Stacking 集成学习的黑枸杞快速无损分级模型，在果肉-FD-SPA-Stacking 组合方式时，精度可从 0.941 7 提升到 0.983 3。研究表明，采用高光谱结合 Stacking 集成学习，可进一步提升黑枸杞的分级精度，实现黑枸杞质量的快速、无损、高精度分级。

References

[1] ZHANG Ying, CHEN Hao(张　莹，陈　浩). The Food Industry(食品工业), 2018, 39(3): 312.

[2] He J, Chen L, Chu B, et al. Molecules, 2018, 23(9): 2395.

[3] WANG Lei, QIN Hong, LI Jing, et al(王　磊，覃　鸿，李　静，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(4): 1270.

[4] YU Hui-chun, WANG Run-bo, YIN Yong, et al(于慧春，王润博，殷　勇，等). Journal of Nuclear Agricultural Sciences(核农学报), 2018, 32(3): 523.

[5] Zhu S, Chao M, Zhang J, et al. Sensors, 2019, 19(23): 5225.

[6] Xu Y, Zhang H, Zhang C, et al. Infrared Physics & Technology, 2019, 102: 103034.

[7] Zhang J, Huang Y, Reddy K N, et al. Pest Management Science, 2019, 75(12): 3260.

[8] Ahn D H, Choi J Y, Kim H C, et al. Sensors, 2019, 19(7): 1560.

[9] YUAN Pei-sen, YANG Cheng-lin, SONG Yu-hong, et al(袁培森，杨承林，宋玉红，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2019, 50(11): 144.

[10] GONG Fang-fang, LIAO Bi-fang, DONG Jia-jia(巩芳芳，廖碧芳，董佳佳，等). Journal of Zhengzhou University · Medical Sciences(郑州大学学报 · 医学版), 2019, 54(4): 531.

[11] FAN Xiao-yi, QU Jun-hao, QU Bao-an, et al(范晓易，曲均浩，曲保安，等). Journal of Geodesy and Geodynamics(大地测量与地球动力学), 2019, 39(9): 916.

[12] MA Rong, SHAN Jing, CHEN Bing-bing, et al(马　蓉，单　璟，陈兵兵，等). Guangzhou Chemical Industry(广州化工), 2020, 48(3): 17, 51.

Fast Classification Method of Black Goji Berry (*Lycium Ruthenicum* Murr.) Based on Hyperspectral and Ensemble Learning

LU Wei[1], CAI Miao-miao[1], ZHANG Qiang[2], LI Shan[3]

1. Jiangsu Provincial Laboratory of Modern Facility Agriculture Technology and Equipment Engineering, College of Artificial Intelligence, Nanjing Agricultural University, Nanjing 210031, China
2. School of Water Resources and Hydropower, Qinghai University, Xining 810016, China
3. School of Life Science and Technology, Tongji University, Shanghai 200092, China

Abstract Black goji berry has high nutrition and medical value. Different grades of black goji berry have different quality, and prices are also significantly different. However, due to the lack of effective detection and grading methods, the black goji berry market is chaotic, and the bad become mixed with the good, which affects the black goji berry market's quality supervision. To achieve fast, non-destructive and high-precision classification of black goji berry, this paper proposes a fast non-destructive classification method of black goji berry based on hyperspectral and ensemble learning. First of all, for Nomhong 1st grade (NMH-grade1), Nomhong 2nd grade (NMH-grade2), Nomhong 3rd grade (NMH-grade3), Nomhong 4th grade (NMH-grade4), select 200 for each grade. Then, in two placement modes (carpopodium up and overall horizontal after removing the carpopodium), the spectral image cube with a spectral range of 391.6～2 528.1 nm is acquired using a GaiaSorter-Dual wide-band hyperspectral sorter. Through the mask processing, automatically extract single black goji berry ROI hyperspectral information with cell counting algorithm. The spectral information of black goji berry in the range of 500～2 400 nm is extracted. After FD(First Derivative), FFT(Fast Fourier Transform), HT(Hilbert Transform), SG(Savitzky Golay), Normalize, SNV(Standard Normal Variate) preprocessing, the spectral information of the characteristic wavelength is extracted by PCA(Principal Components Analysis), SPA(Successive Projection Algorithm), CARS(Competitive Adaptive Reweighted Sampling). Then build LIBSVM, LDA(Latent Dirichlet Allocation), KNN(k-Nearest Neighbor), RF(Random Forest), NB (Naive Bayes) detection models. The combination of sarcocarp-Normalize-SPA-LDA, sarcocarp-FD-CARS-RF and sarcocarp-SNV-CARS-LIBSVM is the best, with accuracy rates of 0.941 7, 0.941 7 and 0.937 5, respectively. At the same time, it can be found that in the pretreatment, FD, HT, Normalize, and SNV have better effects. In the dimensionality reduction method, the models of SPA and CARS have better effects. And in the models established by LIBSVM, LDA, KNN, RF, and NB, the number of test set accuracy rates of not less than 0.9 are 2, 7, 0, 4, and 1, respectively, so the three classifiers LDA, RF, and LIBSVM work best. To further improve the classification accuracy of black goji berry, LDA, RF and LIBSVM are used as meta-models to build a fast and non-destructive classification model of black goji berry Stacking ensemble learning. When the sarcocarp-FD-SPA-Stacking is combined, the accuracy can be improved from 0.941 7 to 0.983 3. A total of 17 characteristic wavelengths is extracted, respectively (in nm): 591.6, 609.1, 721.6, 989.1, 1 083.3, 1 111.3, 1 296.1, 1 564.9, 1 844.9, 1 934.5, 1 996.1, 2 046.5, 2 130.5, 2 292.9, 2 315.3, 2 320.9, 2 348.9. Among them, there are C-H frequency doubling peaks and absorption peaks near 721.6, 1 083.3, 1 111.3, 2 130.5, 2 292.9, 2 315.3, 2 320.9, 2 348.9, O—H frequency doubling peaks and absorption peaks near 721.6, 989.1, 1 934.5, 1 996.1, 2 292.9, and C—O absorption peaks near 2 130.5 and 2 292.9. Research has shown that fast and non-destructive classification of black goji berry based on hyperspectral combined with ensemble learning is feasible.

Keywords Spectroscopy; Black goji berry; Ensemble learning; Anthocyanin; Non-destructive testing

(Received May 16, 2020; accepted Sep. 21, 2020)