高光谱无损识别野生和种植黑枸杞

赵 凡，闫昭如，薛建新，徐 兵

山西农业大学工学院，山西 太谷 030801

摘 要 高光谱图像技术在农产品检测及识别方面有广阔的应用前景。野生黑枸杞经济效益显著，经常被种植黑枸杞冒充。提出一种利用高光谱图像对野生黑枸杞无损快速识别的方法。主要内容和结果如下：(1)共采集256份(野生、种植各128份)黑枸杞在900～1 700 nm范围的高光谱反射光谱，每份平均光谱作为此样品的光谱；(2)采用标准正态变换(SNV)对采集的光谱预处理；基于Kennard-Stone法，按照校正集和预测集比例为2∶1对样品划分，用连续投影算法(SPA)对光谱进行降维处理，提取特征波长30个；分别将全光谱和SPA提取的30个特征波长作为模型输入，建立支持向量机(SVM)、极限学习机(ELM)和随机森林(RF)识别模型。(3)结果表明，在识别野生黑枸杞模型中，基于全光谱和SPA建立的SVM，ELM和RF模型校正集识别率均高于98.8%，基于全光谱和SPA建立的SVM，ELM和RF模型预测集识别率均高于97.7%。基于全光谱(FS)建立的三种识别模型略优于基于SPA建立的三种识别模型。但从简化模型方面，SPA提取的特征波常数仅为全光谱的11.8%，大大降低了模型运算量。三种模型中，基于随机森林模型无损识别野生黑枸杞效果最好，均达到100%。研究表明，利用高光谱图像技术结合分类模型可快速识别野生黑枸杞。

关键词 野生黑枸杞；高光谱图像；支持向量机；极限学习机；随机森林算法

中图分类号：O433 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)01-0201-05

引 言

黑枸杞有“花青素之王”美誉。它具有抗衰老、降血脂、防癌等功效[1-2]。野生黑枸杞生长在野外，生长周期长、光照充足、无重金属和农药残留等问题，而且产量远远低于种植黑枸杞，所以野生黑枸杞更加珍贵。野生黑枸杞市场价格远远高于种植黑枸杞，故一些不法商家用种植黑枸杞冒充野生黑枸杞来欺骗消费。识别野生黑枸杞已成为黑枸杞市场急需解决的关键问题。

高光谱图像技术具有无损、高效、非接触等优势[3]。它在农产品检测及识别方面具有非常广泛的应用前景[4]。Liu等[5]对带真菌和瘀伤的草莓进行识别；Dong等[6]利用高光谱图像对不同浓度的猕猴桃膨大果进行识别；鲍一丹等[7]利用光谱图像对国产咖啡豆品种识别。目前尚未见用高光谱图像识别野生黑枸杞的报道。

为研究高光谱图像识别野生黑枸杞，本研究以野生和种植黑枸杞为研究对象，建立支持向量机(support vector machine，SVM)、极限学习机(extreme learning machine，ELM)和随机森林(random forest，RF)识别模型，并采用连续投影法(successive projections algorithm，SPA)提取特征波长，比较全光谱(FS)和连续投影算法对模型精度的影响。

1 实验部分

1.1 材料

实验用野生和种植黑枸杞均由青海千拓贸易有限公司提供，原产地为青海。黑枸杞根据颗粒大小分为特级(0.6 cm以上)、高级(0.5～0.6 cm)、中级((0.4～0.5 cm)三级，选用颗粒在0.4～0.5 cm范围的中级野生和种植黑枸杞作为实验材料。野生黑枸杞如图1(a)所示。为防止实验受到影响，将所有黑枸杞去除果柄和杂质。去除果柄野生黑枸杞如图1(b)所示。每(5±0.1)g黑枸杞作为一份样品。野生和种植黑枸杞样品数分别均为128份，总样品数为256份。

1.2 仪器

高光谱图像系统：GaiaSorter“盖亚”高光谱分选仪北京汉光卓立公司；4个35 W溴钨灯、电控平台；物镜以及计算机等部件。图像光谱范围：900～1 700 nm；光谱分辨率：

收稿日期：2019-11-29，**修订日期**：2020-03-22

基金项目：国家自然科学基金项目(31801632)资助

作者简介：赵 凡，女，1989年生，山西农业大学工学院讲师 e-mail：1140117238@qq.com

3.19 nm；曝光时间：10 ms；物距：20 cm；图像采集速率：7.2 mm·s^{-1}。

图 1(a) 野生黑枸杞

Fig. 1(a) Wild black Goji berries

图 1(b) 去除果柄后的野生黑枸杞

Fig. 1(b) Wild black Goji berries after removing the stalks

1.3 高光谱图像的采集

仪器箱体内存在暗电流、光源分布不均匀，这些因素会使采集到的高光谱图像含有较大噪音，故需对高光谱图像进行黑白校正[8-9]。公式如式(1)

$$R/\% = (R_0 - B)/(W - B) \times 100\% \quad (1)$$

式(1)中：R_0 为反射光谱图像；W 为白板漫反射图像；B 为暗图像；R 为校正后漫反射光谱图像。

利用 ENVI4.8 软件建立掩膜提取高光谱图像。选取第 140 波段处的图像进行阈值分割，当阈值为 0.18 时，能够提取完整的黑枸杞图像，因此设定阈值为 0.18 进行图像提取。将黑枸杞图像区域的平均光谱作为此黑枸杞单个样品的反射光谱。

1.4 光谱处理和样品划分

采用标准正态变换(standardized normal variate，SNV)进行光谱预处理。采用 Kennard-Stone(K-S)法划分样品；K-S 算法已经被证明在选择代表性样品方面的具有很好的效果。采用 SPA 法对光谱降维从而简化模型；SPA 法可以在高光谱庞大复杂的数据中去除冗杂数据、提取特征波长数据[10]。

1.5 建模方法

1.5.1 SVM 模型

SVM 是将向量映射到更高维空间，构建最大间隔的超平面，剪力合适的分隔超平面，使两个与之平行的超平面距离最大化，从而来解决复杂数据的分类和回归问题[11]。

1.5.2 ELM 模型

ELM 是由南洋理工大学黄广斌教授提出的一种有效单隐含层前馈神经网络算法，它学习速度快、泛化能力好[12]。

1.5.3 RF 模型

RF 是一种用多棵树对样品进行训练并预测的分类器。它包含多个决策树算法，具有数据选择随机性。RF 具有实现简单、能处理高维数据、避免过拟合等优势[13]。

2 结果与讨论

2.1 光谱预处理和样本划分

野生和种植黑枸杞样品各 128 份。图 2 所示是野生和种植黑枸杞样品平均反射光谱，共 254 个波段。由图 2 可知，2 条平均反射光谱变化趋势一致，其中，波长 1 000～1 350 nm 范围内，野生黑枸杞光谱反射率明显高于种植黑枸杞；在波长 1 500～1 650 nm 范围内，种植黑枸杞光谱反射率略高于野生黑枸杞。这两条光谱反射曲线都有 2 个明显的波谷，即在波长 1 235 和 1 350～1 650 nm 处均有明显吸收峰。而雷建刚等在近红外对不同产地枸杞优化论文中也提到枸杞在 1 235和 1 535 nm 处均有明显吸收峰[14]。

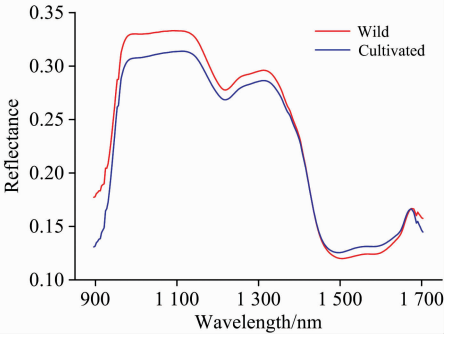

图 2 野生和种植黑枸杞的原始平均光谱

Fig. 2 Average reflectance spectra of wild and cultivated black Goji berries

对所有样品光谱进行 SNV 预处理。对经 SNV 后的光谱样品进行样品划分，按照校正集和预测集样品数为 2∶1 的比例，用 K⨋S 法划分 256 份样品，得到校正集 170 个(野生和种植黑枸杞各 85 份)。预测集 86 个(野生和种植黑枸杞各 43 份)。

2.2 光谱数据降维

设定 SPA 选择最多波长数为 50，用均方根误差确定最佳特征波常数，均方根误差随特征波长数变化曲线如图 3 所示。选取最佳特征波长数为 30。

2.3 建模结果

分别将全光谱 254 个波段、经 SPA 提取的 30 个特征波长作为输入变量，建立 SVM，ELM 和 RF 野生黑枸杞和种植黑枸杞识别模型。图 4—图 6 是三种模型对黑枸杞的识别结果；每个图纵坐标中，1.0 代表野生黑枸杞，2.0 代表种植黑枸杞。

2.3.1 SVM 模型

在 SVM 中，采用 RBF(radial base function)作为核函数，通过留一交叉验证方法(cross validation，CV)寻找最佳

惩罚因子(c)、核函数参数(g)，基于 FS 和 SPA 不同模型确定的 c 和 g 见表 1。SVM 模型对野生黑枸杞和种植黑枸杞识别结果如图 4 所示。

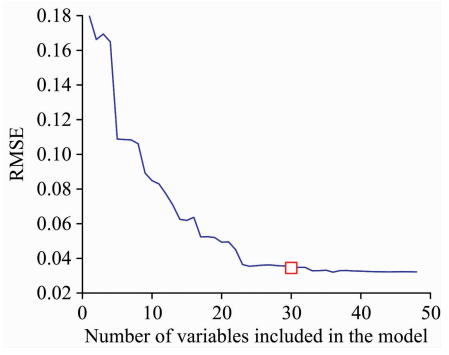

图 3　均方根误差随 SPA 中特征波长数变化曲线

Fig. 3　Changed RMSE with the number of characteristic wavelength in SPA

表 1　SVM 模型参数

Table 1　Parameters of SVM

特征波长选择	c	g
FS	0.003 9	0.108 8
SPA	0.003 906 3	0.003 903 6

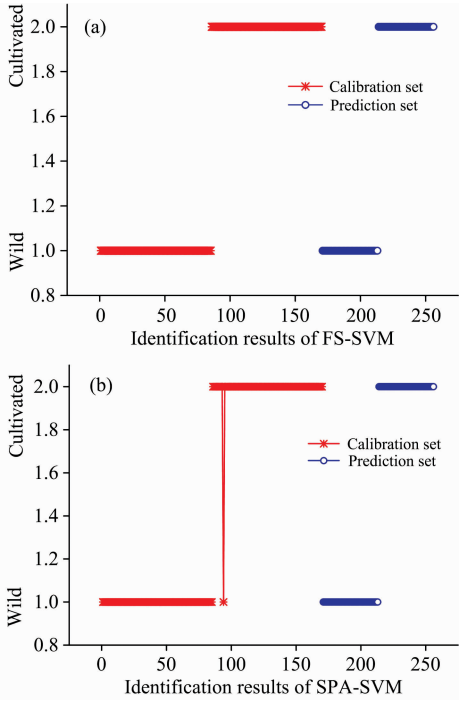

图 4　SVM 黑枸杞识别结果

Fig. 4　Identification results of black Goji berries by SVM

由图 4 可知，FS-SVM 校正集和预测集平均识别率均为 100%。SPA-SVM 校正集中，有 1 份种植黑枸杞识别错误，野生、种植黑枸杞识别率分别为 100%和 98.8%；所以 SPA-SVM 校正集平均识别率为 99.4%。SPA-SVM 预测集平均识别率为 100%。FS-SVM 模型识别率均整体略优于 SPA-SVM 模型。

2.3.2　ELM 模型

在 ELM 模型中，采用“sigmoidal”函数作为激活函数，设置隐含层神经元个数为 1～100，步长为 1，确定 FS 和 SPA

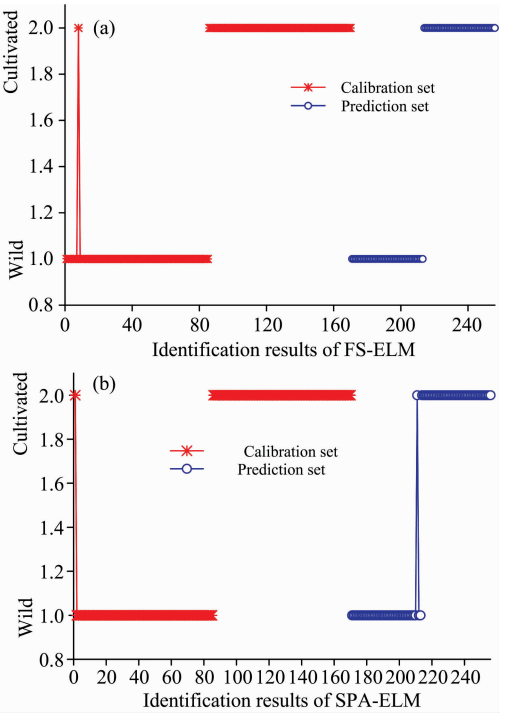

图 5　ELM 模型黑枸杞识别结果

Fig. 5　Identification results of black Goji berries by ELM

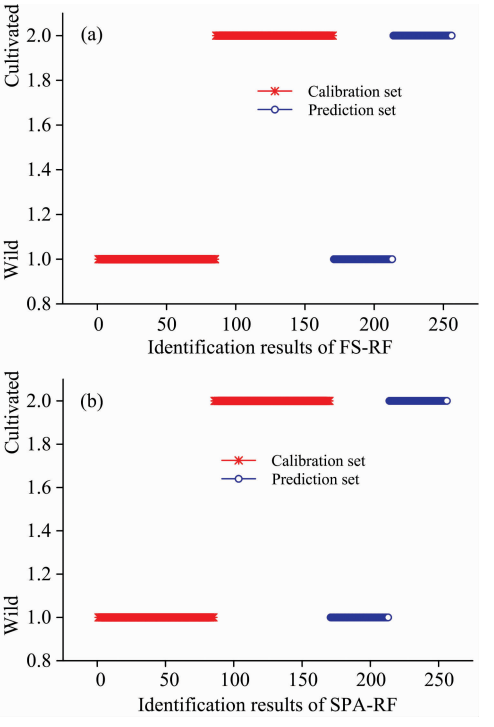

图 6　RF 黑枸杞识别结果

Fig. 6　Identification results of black Goji berries by RF

的隐含层神经元个数为 10 和 7。ELM 模型对野生黑枸杞和种植黑枸杞识别结果如图 5 所示。

由图 5 可知，FS-ELM 校正集中，有 1 份野生黑枸杞识别错误，野生和种植黑枸杞识别率分别为 98.8%和 100%；FS-ELM 校正集平均识别率为 99.4%。FS-ELM 预测集识别率均为 100%。SPA-ELM 校正集中，有 1 份野生黑枸杞识别错误，野生、种植黑枸杞分别为 98.8%和 100%；SPA-ELM 校正集平均识别率均为 99.4%。SPA-ELM 预测集中，有 1 份野生黑枸杞识别错误，野生、种植黑枸杞识别率分别为 97.7%和 100%；SPA-ELM 预测集平均识别率为 98.8%。整体来说，FS-ELM 模型识别率略高于 SPA-ELM 模型。

2.3.3 RF 模型

建立随机森林识别模型，树的数目为 500。RF 模型结果见图 6。由图 6 可知，FS-RF 和 SPA-RF 的校正集和预测集识别率全部达到了 100%。这说明 FS-RF 和 SPA-RF 模型可完全识别野生和种植黑枸杞。

2.4 建模结果比较

FS-SVM 和 FS-RF，SPA-RF 模型对校正集和预测集识别率都达到了 100%。SVM 模型识别率整体优于 ELM 模型，而 RF 模型识别率是三种模型中最高，达到 100%。所以 RF 模型是最优识别模型。

3 结 论

(1)识别野生黑枸杞模型中，基于 FS 和 SPA 建立的 SVM，ELM 和 RF 模型校正集识别率高于 98.8%，基于全光谱和 SPA 建立的 SVM，ELM 和 RF 模型预测集识别率高于 97.7%。

(2)基于 FS 建立的模型识别效果最好，基于 SPA 建立的模型识别效果略低于 FS 建立的模型。但从简化模型方面，SPA 提取的特征波常数仅为 FS 的 11.8%，大大降低了模型运算量。

(3)RF 识别模型最优，野生黑枸杞识别率均达到了 100%。

References

[1] Li Y H, Zou X B, Shen T T, et al. Food Analytical Methods, 2016, 10(4): 1.

[2] Tian Zhihao, Aierken Aizezijiang, Pang Huanhuan, et al. Journal of Liquid Chromatography & Related Technologies, 2016, 39(9): 453.

[3] DONG Jin-lei, GUO Wen-chuan(董金磊，郭文川). Food Science(食品科学), 2015, 36(16): 101.

[4] LU Na, HAN Ping, WANG Ji-hua(卢 娜，韩 平，王纪华). Journal of Food Safety & Quality(食品安全与检测学报), 2017, 8(12): 4594.

[5] Liu Q, Sun K, Peng J, et al. Food Analytical Methods, 2018, 11(5): 1518.

[6] Dong J L, Guo W C, Zhao F, et al. Food Analytical Methods, 2017, 10(2): 477.

[7] BAO Yi-dan, CHEN Na, HE Yong, et al(鲍一丹，陈 纳，何 勇，等). Optical Precision Engineering(光学精密工程), 2015, 23(2): 349.

[8] ElMasry G, Wang N, Vigneault C. Postharvest Biology and Technology, 2009, 52(1): 1.

[9] Menesatti P, Zanella A, D'Andrea S, et al. Food and Bioprocess Technology, 2009, 2(3): 308.

[10] Wu D, Shi H, Wang S J, et al. Analytica Chimica Acta, 2012, 726: 57.

[11] KONG Qing-ming, SU Zhong-bin, SHEN Wei-zheng, et al(孔庆明，苏中滨，沈维政，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2015, 35(5): 1233.

[12] Yuan Peipei, Chen Hong, Zhou Yicong, et al. Neurocomputing, 2015, 167: 528.

[13] Zhang Xiaolong, Lin Xiaoli, Zhao Jiafu, et al. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, 16(3): 774.

[14] LEI Jian-gang, LIU Dun-hua(雷建刚，刘敦华). Food Science(食品科学), 2013, 34(20): 148.

Identification of Wild Black and Cultivated Goji Berries by Hyperspectral Image

ZHAO Fan, YAN Zhao-ru, XUE Jian-xin, XU Bing
College of Engineering, Shanxi Agricultural University, Taigu 030801, China

Abstract Hyperspectral image technology has a broad application in the detection and identification of agricultural products. Wild black Goji berries have remarkable economic benefits, and are often impersonated by growing black Goji berries. A nondestructive and fast identification method for wild black Goji berries using hyperspectral image technology is proposed. Obtained results were as follows:(1) a total of 256 samples of black Goji berries (Wild, Growing, 128 each) in the range of 900～1 700 nm were observed, and each average spectra were used as simple spectra. (2) spectral is preprocessed with standardized normal variate transform (SNV) based on the Kennard-Stone(K-S) method, the calibration set and prediction set samples ratio were observed in 2∶1 (pairs). However, the spectra were found reduced in dimension by the successive projections algorithm method (SPA), and the 30 characteristic wavelengths extracted by the full spectra (FS). Then the 30 characteristic wavelengths and the full spectra are used as model inputs, the support vector machine (SVM), extreme learning machine (ELM), and random forest (RF) recognition models were established. (3) In the identification of wild black Goji berries models, the results showed that the calibration identification rate of SVM, ELM, and RF model with reference to FS and SPA were higher than 98.8%, and the prediction set samples rate of SVM, ELM, and RF model were also higher than 97.7%. The identification model of FS was slightly better than the identification model of SPA. However, the characteristic wave constant extracted by SPA is 11.8% less compared to FS, which eventually reduces the calculated model. RF identification model was reported better compared to SVM, and ELM, RF identification rate is 100%. The study has shown that the use of hyperspectral image technology combined with classification models can quickly identify wild black Goji berries.

Keywords Wild black Goji berry; Hyperspectral image technology; Support vector machine; Extreme learning machine; Random forest

(Received Nov. 29, 2019; accepted Mar. 22, 2020)