doi:10.3969/j.issn.1672－0504.2019.05.006

基于随机森林算法的玉米品种高光谱图像鉴别

邵 琦[1,2],陈云浩[3,4],杨淑婷[5],赵逸飞[3],李 京[1*]

(1.北京师范大学遥感科学国家重点实验室,北京 100875;2.北京师范大学地理科学学部,北京 100875;3.环境遥感与数字城市北京市重点实验室,北京师范大学地理科学学部,北京 100875;4.地表过程与资源生态国家重点实验室,北京师范大学地理科学学部,北京 100875;5.宁夏农林科学院农业经济与信息技术研究所,宁夏 银川 750002)

摘要:玉米品种直接影响到玉米的产量和品质,事关农业收入和食品安全,因此,如何准确、高效、无损地鉴别玉米品种具有重要意义。该文基于高光谱成像系统采集3个品种共600粒玉米在533～893.4 nm波段(共146个波段)范围的高光谱图像,对其进行校正和预处理,利用Boruta算法筛选有效波段。在全波段、全波段和纹理信息、有效波段以及有效波段和纹理信息4种特征组合下,利用随机森林算法进行玉米品种识别研究。结果表明:4种特征组合下,随机森林的平均分类准确率达76.25%,Kappa系数均在0.6以上,分类效果均优于传统的偏最小二乘判别分析方法;从4种特征组合的分类结果看,融合纹理信息的随机森林判别模型识别精度显著提升,分类准确率达77.20%,Kappa系数在0.64以上;基于有效波段和纹理信息判别模型的分类准确率达78.30%,Kappa系数为0.675。由此可见,有效波段和纹理信息特征组合下的随机森林算法能充分利用高光谱图像的光谱和纹理信息,准确地鉴别玉米品种,为玉米品种的自动识别提供了一种新方法。

关键词:高光谱图像;玉米;随机森林;偏最小二乘判别分析

中图分类号:TP79;S513 **文献标识码**:A **文章编号**:1672－0504(2019)05－0034－06

0 引言

玉米是我国重要的粮食作物之一,占粮食总产量的30%以上,制种量年均超过10亿t[1]。随着玉米需求多样化,我国玉米品种也逐年增多。在育种过程中,有不法商贩用劣质玉米品种冒充优质玉米品种,导致玉米产量下降,给国家粮食生产和农业安全带来巨大隐患,因此,如何准确、高效、无损地鉴别玉米品种具有重要意义。目前鉴别玉米品种的方法主要有:1)人工检测方法。主要靠肉眼根据玉米种子的大小、形态和颜色等特征做出判断,受人的主观因素影响很大,不仅检测精度低、速度慢,而且劳动强度大,难以形成统一的标准[2]。2)理化检测方法。主要是基于不同品种玉米的生物化学特性进行鉴别,准确率虽高,但只能进行抽样检测,有损且过程复杂、操作繁琐,难以满足市场需求[2]。3)计算机视觉检测方法。主要应用图像模式识别技术,基于玉米的近红外光谱信息和可见光的形态学特征信息建立品种判别模型[3－6],虽具有即时、高效、无损和准确的优点,但提取的特征信息较少,且多是基于外部特征的评价,无法获取表征种子内部成分的特征,影响结果的可靠性[7];另外,随着玉米品种数据增多,特征交叉现象严重,导致模型的可分性变差[8]。4)高光谱技术检测方法。高光谱集光谱和图像于一体,相比于机器视觉和近红外光谱分析技术而言,其能够获取对象的图像信息和光谱信息,可有效表征观测对象的内部结构特征和化学成分特征;此外,高光谱技术以其快速、无损的技术优点有效地解决了农产品检测领域操作繁琐的难题,在农产品无损检测方面的应用日趋广泛[9－12]。

许多学者应用高光谱技术对玉米品种分类展开了研究。例如:冯朝丽等[13]采集11类共528粒玉米的高光谱影像,基于偏最小二乘判别分析(PLSDA)建立分类判别模型,测试集精度达94%;唐金亚等[14]采集了6类共720粒玉米的高光谱影像,利用局部学习算法筛选最优波段,并基于PLSDA建立分类预测模型,分类精度达95.97%;吴翔等[15]采集了4类共384粒玉米的高光谱影像,利用连续投影算法和PLSDA建立了分类判别模型,测试集分类精度达70.8%;魏利峰[2]采集10类共1 000粒玉米的高光谱影像,基于PLSDA分别建立全波段、多波段和特征波段下融合形态和纹理特征的玉米品种判别模

收稿日期:2019－03－10; 修回日期:2019－04－13

基金项目:宁夏农林科学院科技创新引导项目(NKYG-18-01)

作者简介:邵琦(1994－),女,硕士研究生,研究方向为遥感应用。*通讯作者 E－mail:lijing@bnu.edu.cn

型，测试集的分类精度高达 98%；黄敏等[8]采集 9 类共 432 粒玉米的高光谱影像，基于 PLSDA 分别建立全波段、单波段和多波段融合纹理特征的玉米品种判别模型，测试集的分类精度达 93%。

高光谱图像波段繁多，数据量大，很多学者采用主成分分析方法（PCA）对其进行降维处理，虽可有效减少数据量，但 PCA 并非合适的高光谱图像降维方法[16]，会丢失高光谱图像的特性，导致光谱信息损失；此外，多数研究采用偏最小二乘判别分析算法，该方法难以捕捉特征之间的非线性关系，当特征数量增多时，分类效果变差。因此，本文采集 3 类共 600 粒玉米的高光谱图像，基于 Boruta 算法[17]筛选有效波段，在全波段、全波段和纹理信息、有效波段、有效波段和纹理信息 4 种特征组合下，利用随机森林算法建立分类识别模型并进行对比分析，为玉米品种自动识别提供新方法。

1　数据采集与研究方法

1.1　玉米影像采集

本文选取农华 213、银玉 274 和银玉 439 玉米种子作为研究对象，每类取 200 粒，样本总数共 600 粒。这 3 类玉米均于 2018 年在宁夏农林科学研究院的基础试验田生产，主色调均为黄色，消除了种子年龄和种植土壤的影响。在玉米收获和干燥后，密封保存，防止在储存期间吸收水分。高光谱图像采集系统由光源、移动平台和采集单元组成。图像采集单元包括图像光谱仪、CCD 探测器（1 392 * 1 040 的面阵 CCD 相机）和镜头（F/2.4，MCT，C-mount），图像光谱仪的狭缝宽度为 30 μm，光谱范围为 400～1 000 nm，光谱分辨率为 2.8 nm。

1.2　数据预处理

1.2.1　影像校正　数据采集过程中存在信号噪声，为保证数据质量，只保留 533～893.4 nm 波段范围内 146 个波段的数据。由于光源光照强度分布不均，加之暗电流的影响，因此，需对影像进行校正[8]：

$$R=(I_S-I_D)/(I_W-I_D) \quad (1)$$

式中：I_S 为原始高光谱影像；I_D 为黑板标定影像；I_W 为白板标定影像；R 为校正后的高光谱影像。

1.2.2　玉米感兴趣区域（ROI）提取　为提取每个玉米粒的光谱和纹理特征，需从原始图像中提取籽粒的 ROI[18]。本文采用基于距离变换的标记分水岭分割算法[19]提取每个籽粒的 ROI 区域，并将提取的 ROI 放置在 44 * 44 * 146 的空白数组里，背景值为 0。玉米 ROI 的提取过程如图 1 所示。

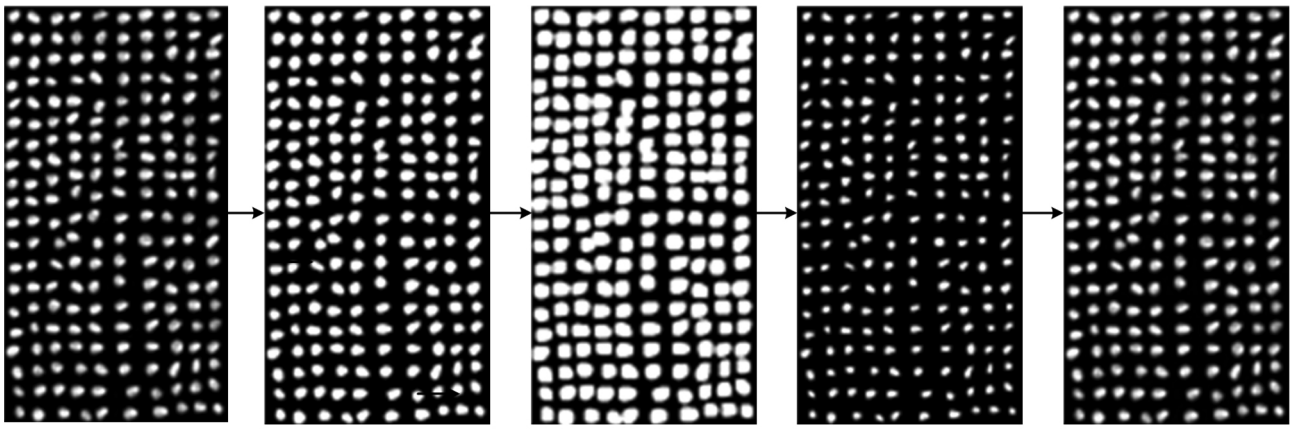


图 1　玉米 ROI 提取过程

Fig. 1　Extraction process of maize ROI

1.3　研究方法

将每类玉米样本按照 7∶3 的比例随机划分为训练集 420 粒（用于建立模型）和测试集 180 粒（用于精度评价）。在全波段、全波段和纹理信息、有效波段、有效波段和纹理信息 4 种特征组合下，采用随机森林算法建立分类识别模型，在训练集上利用五折交叉验证方法确定模型参数，然后在测试集上进行预测并计算分类混淆矩阵[20]（Kappa 系数和准确率）和分类指标[21]（精确率、召回率和 $F1$ 值），查看分类效果。技术路线如图 2 所示。

1.3.1　光谱特征提取

（1）光谱预处理。计算各种玉米 ROI 的平均光谱值，分别作为该品种的光谱特征；由于光谱信息存在噪声，故需进行平滑处理以消除噪声干扰，提高模型预测能力及稳定性。本文采用 Savitzky-Golay（S-G）平滑算法[22]对提取的光谱特征进行校正：将奇数个（本文设定参数为 5）光谱点作为一个窗口，采用多项式算法对窗口内的数据做最小二乘拟合，计算窗

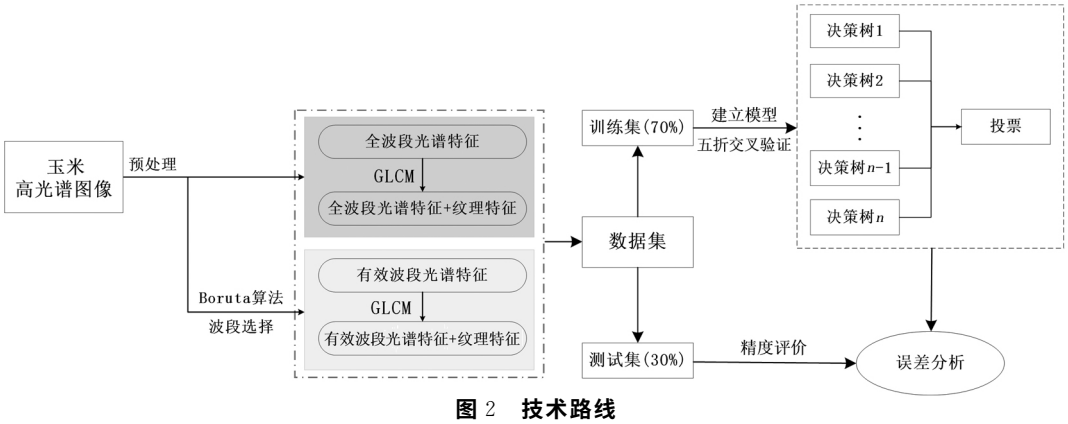


图 2　技术路线
Fig. 2　Flow chart of technology route

口中心点的一阶导数和平滑数据值，移动窗口重复上述过程，对光谱特征做平滑处理[23]。

(2)光谱波段选择。为降低高光谱数据的高维性，本文采用 Boruta 算法[17]从 146 个波段中筛选出有效波段，作为模型分类的特征集。Boruta 算法主要利用 *Z*-score 度量特征的重要性，进而筛选出重要特征[24]，算法过程为：1)为每个特征构建 shadow 特征并混入训练集，同时去除这些特征与类别的关联；2)在混入 shadow 特征的训练集上训练随机森林分类器；3)利用特征的平均损失和标准偏差计算 *Z*-score；4)删除 *Z*-score 比 shadow 特征差的特征；5)当所有特征被确认或者算法达到设置的迭代次数时，算法停止。

1.3.2　纹理特征提取　采用灰度共生矩阵(GLCM)[25]提取玉米粒的纹理特征，通过提取图像的空间相关矩阵描述图像的纹理信息。本文将像素间的距离参数设置为 1，依次选取 0°、45°、90°和 135°分别提取玉米粒的能量值(energy)、对比度值(contrast)、相关性值(correlation)和熵值(entropy)[7] 4 个纹理特征，玉米 ROI 的平均纹理特征值作为该玉米粒的纹理特征。

1.3.3　随机森林算法　随机森林是一种分类与回归技术，也是一种组合式的自学习技术[26]，其通过对大量回归决策子树的汇总提高了模型的预测精度[27]，同时避免了过拟合等问题，具有实现简单、计算开销小的优点，对异常值和噪声容忍度高，非常适用于非线性数据建模，并且可以对变量进行重要性分析[28－30]。

2　结果与分析

利用 Boruta 算法对训练集的 146 个波段进行五折交叉验证，最终筛选出 58 个有效波段。在全波段、全波段和纹理信息、有效波段、有效波段和纹理信息 4 种特征组合下，采用随机森林算法建立分类判别模型，利用五折交叉验证方法确定模型参数，在测试集上进行预测并计算混淆矩阵(表 1－表 4)和分类指标(图 3)。

从玉米的分类混淆矩阵(表 1－表 4)看，随机森林算法的平均分类准确率达 76.25%，Kappa 系数均

表 1　全波段的分类混淆矩阵
Table 1　Classification confusion matrix of full bands

类别	随机森林			PLSDA			总计
	农华 213	银玉 274	银玉 439	农华 213	银玉 274	银玉 439	
农华 213	40	2	18	42	2	16	60
银玉 274	0	57	3	0	60	0	60
银玉 439	16	6	38	22	9	29	60
总计	56	65	59	64	71	45	180
Kappa 系数		0.625			0.592		
准确率		0.750			0.728		

表 2　全波段和纹理信息的分类混淆矩阵
Table 2　Classification confusion matrix of full bands and texture information

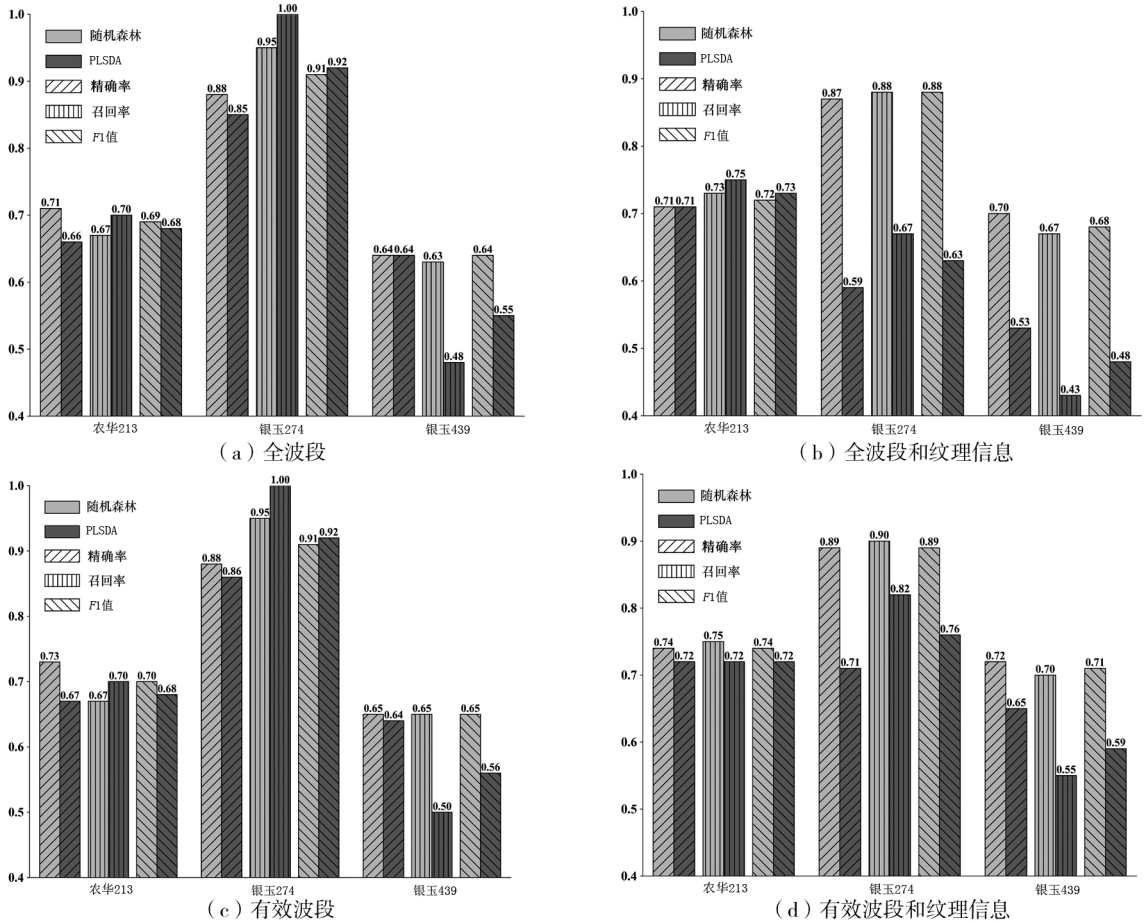
类别	随机森林			PLSDA			总计
	农华 213	银玉 274	银玉 439	农华 213	银玉 274	银玉 439	
农华 213	44	2	14	45	6	9	60
银玉 274	4	53	3	6	40	14	60
银玉 439	14	6	40	12	22	26	60
总计	62	61	57	63	68	49	180
Kappa 系数		0.641			0.425		
准确率		0.761			0.617		

表 3　有效波段的分类混淆矩阵
Table 3　Classification confusion matrix of effective bands

类别	随机森林			PLSDA			总计
	农华 213	银玉 274	银玉 439	农华 213	银玉 274	银玉 439	
农华 213	40	2	18	42	1	17	60
银玉 274	0	57	3	0	60	0	60
银玉 439	15	6	39	21	9	30	60
总计	55	65	60	63	70	47	180
Kappa 系数		0.633			0.600		
准确率		0.756			0.733		

表 4　有效波段和纹理信息的分类混淆矩阵
Table 4　Classification confusion matrix of effective bands and texture information

类别	随机森林			PLSDA			总计
	农华 213	银玉 274	银玉 439	农华 213	银玉 274	银玉 439	
农华 213	45	2	13	43	5	12	60
银玉 274	3	54	3	5	49	6	60
银玉 439	13	5	42	12	15	33	60
总计	61	61	58	60	69	51	180
Kappa 系数		0.675			0.541		
准确率		0.783			0.694		



（a）全波段　（b）全波段和纹理信息
（c）有效波段　（d）有效波段和纹理信息

图 3　模型分类结果比较
Fig. 3　Comparison of model classification results

在 0.6 以上，而 PLSDA 算法的平均分类准确率达 69.30%，Kappa 系数均在 0.6 以下，这表明随机森林的分类结果优于 PLSDA 算法；从玉米分类的精确率、召回率和 $F1$ 值（图 3）看，随机森林算法的分类结果仍优于 PLSDA 算法。

对于表 1—表 4 和图 3 中随机森林算法的分类结果，全波段和纹理信息的分类结果优于全波段，有效波段和纹理信息的分类结果优于有效波段，融合纹理信息的分类平均准确率达 77.20%，Kappa 系数均在 0.64 以上，这表明融合纹理信息的分类模型可有效提高分类精度。有效波段的分类结果优于全波段，有效波段和纹理信息的分类结果优于全波段和

纹理信息，这表明 Boruta 算法可以有效地筛选表征玉米性质的特征波段，从而减少数据量，提高分类精度。

对于表 1－表 4 及图 3 中的 PLSDA 分类结果，有效波段的分类结果优于全波段，全波段的分类结果优于全波段和纹理信息，这表明：当特征数量增多时，PLSDA 算法无法准确捕捉特征之间的非线性关系，导致分类精度下降。

3　结论

本文采集 3 类共 600 粒玉米的高光谱图像，基于 Boruta 算法筛选有效波段，在全波段、全波段和纹理信息、有效波段以及有效波段和纹理信息 4 种特征组合下，利用随机森林算法建立分类判别模型，结果表明：1）4 种特征组合的分类结果中，随机森林算法的平均分类准确率达 76.25%，Kappa 系数均在 0.6 以上，分类效果优于传统的 PLSDA；2）从随机森林在 4 种特征组合的分类结果看，融合纹理信息的随机森林判别模型识别精度显著提升，分类平均准确率达 77.20%，Kappa 系数在 0.64 以上；其中，有效波段和纹理信息特征组合下的随机森林算法判别模型的分类准确率达 78.30%，Kappa 系数达 0.675。

深度卷积神经网络算法在多类别、大样本数据集下的分类效果显著，精度较高，但由于此次采集的玉米品种和样本量较少，无法利用深度学习算法建立判别模型。后续将采集更多的样本数据集，探索深度学习算法在高光谱玉米品种鉴别方面的应用。

参考文献：

[1] 赵久然，王凤格. 玉米品种 DNA 指纹鉴定技术研究与应用[M]. 北京：中国农业科学技术出版社，2009.

[2] 魏利峰. 玉米种子高光谱图像品种检测方法研究[D]. 沈阳：沈阳农业大学，2017.

[3] KIRATIRATANAPRUK K，SINTHUPINYO W. Color and texture for corn seed classification by machine vision[A]. 2011 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS)[C]. 2011. 1－5.

[4] 刘双喜，王盼，张春庆，等. 基于优化 DBSCAN 算法的玉米种子纯度识别[J]. 农业机械学报，2012，43(4)：188－192.

[5] 闫小梅，刘双喜，张春庆，等. 基于颜色特征的玉米种子纯度识别[J]. 农业工程学报，2010，26(S1)：46－50.

[6] 朱启兵，冯朝丽，黄敏，等. 基于图像熵信息的玉米种子纯度高光谱图像识别[J]. 农业工程学报，2012，28(23)：271－276.

[7] 王润博. 基于高光谱图像技术的枸杞品质检测方法研究[D]. 洛阳：河南科技大学，2017.

[8] 黄敏，朱晓，朱启兵，等. 基于高光谱图像的玉米种子特征提取与识别[J]. 光子学报，2012，41(7)：868－873.

[9] 李江波，饶秀勤，应义斌. 农产品外部品质无损检测中高光谱成像技术的应用研究进展[J]. 光谱学与光谱分析，2011，31(8)：2021－2026.

[10] 刘燕德，张光伟. 高光谱成像技术在农产品检测中的应用[J]. 食品与机械，2012(5)：223－226.

[11] 张保华，李江波，樊书祥，等. 高光谱成像技术在果蔬品质与安全无损检测中的原理及应用[J]. 光谱学与光谱分析，2014，34(10)：2743－2751.

[12] 高攀，张初，吕新，等. 近红外高光谱成像的微破损棉种可视化识别[J]. 光谱学与光谱分析，2018，38(6)：1712－1718.

[13] 冯朝丽，朱启兵，朱晓，等. 基于光谱特征的玉米品种高光谱图像识别[J]. 江南大学学报(自然科学版)，2012，11(2)：149－153.

[14] 唐金亚，黄敏，朱启兵. 基于局部学习的玉米种子近红外高光谱图像鉴选[J]. 激光与光电子学进展，2015，52(4)：041102.

[15] 吴翔，张卫正，陆江锋，等. 基于高光谱技术的玉米种子可视化鉴别研究[J]. 光谱学与光谱分析，2016，36(2)：511－514.

[16] CHERIYADAT A，BRUCE L M. Why principal component analysis is not an appropriate feature extraction method for hyperspectral data[A]. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477)[C]. 2003，6：3420－3422.

[17] KURSA M B，RUDNICKI W R. Feature selection with the Boruta package[J]. Journal of Statistical Software，2010，36(11)：1－13.

[18] WANG L，SUN D W，PU H，et al. Application of hyperspectral imaging to discriminate the variety of maize seeds[J]. Food Analytical Methods，2016，9(1)：225－234.

[19] TARABALKA Y，CHANUSSOT J，BENEDIKTSSON J A. Segmentation and classification of hyperspectral images using watershed transformation[J]. Pattern Recognition，2010，43(7)：2367－2379.

[20] COHEN J. A coefficient of agreement for nominal scales[J]. Educational and Psychological Measurement，1960，20(1)：37－46.

[21] 张宁，贾自艳，史忠植. 使用 KNN 算法的文本分类[J]. 计算机工程，2005，31(8)：171－172.

[22] SCHAFER R W. What is a Savitzky-Golay filter[J]. IEEE Signal Processing Magazine，2011，28(4)：111－117.

[23] 侯培国，李宁，常江，等. SG 平滑和 IBPLS 联合优化水中油分析方法的研究[J]. 光谱学与光谱分析，2015，35(6)：1529－1533.

[24] 张凯. 高维小样本数据的互信息特征选择方法研究[D]. 太原：山西大学，2017.

[25] POURREZA A，POURREZA H，ABBASPOUR-FARD M H，et al. Identification of nine Iranian wheat seed varieties by textural analysis with image processing[J]. Computers and Electronics in Agriculture，2012，83：102－108.

[26] BREIMAN L. Random forests[J]. Machine Learning，2001，45(1)：5－32.

[27] 李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报，2013，50(4)：1190－1197.

[28] 方匡南，吴见彬，朱建平，等. 随机森林方法研究综述[J]. 统计与信息论坛，2011，26(3)：32－38.
[29] 张雷，王琳琳，张旭东，等. 随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例[J]. 生态学报，2014，34(3)：650－659.
[30] 邵琦，陈云浩，李京. 基于卫星遥感和气象再分析资料的北京市 $PM_{2.5}$ 浓度反演研究[J]. 地理与地理信息科学，2018，34(3)：32－38.

Identification of Maize Seed Varieties Based on Random Forest and Hyperspectral Technique

SHAO Qi[1,2]，CHEN Yun－hao[3,4]，YANG Shu－ting[5]，ZHAO Yi－fei[3]，LI Jing[1]

(*1. State Key Laboratory of Remote Sensing Science，Beijing Normal University，Beijing 100875；2. Faculty of Geographical Science，Beijing Normal University，Beijing 100875；3. Beijing Key Laboratory of Environmental Remote Sensing and Digital Cities，Faculty of Geographical Science，Beijing Normal University，Beijing 100875；4. State Key Laboratory of Earth Surface Processes and Resource Ecology，Faculty of Geographical Science，Beijing Normal University，Beijing 100875；5. Institute of Agricultural Economics and Information Technology，Ningxia Academy of Agricultural and Forestry Sciences，Yinchuan 750002，China*)

Abstract：Maize varieties directly affect the yield and quality of corn，which is related to agricultural income and food safety. Therefore，it is of great significance to accurately and efficiently identify the varieties of maize seeds. In this paper，a hyperspectral imaging system was used to acquire hyperspectral images of 600 maize seeds from 3 varieties within the wavelength range of 533～893. 4 nm (146 bands). The image was then corrected and preprocessed，and the effective band was screened by the Boruta algorithm. The random forest algorithm was used to identify maize varieties under the combinations of full-band，full-band and texture information，effective band，and effective band and texture information. The results show that using the four combinations，the average classification accuracy by the random forest algorithm is 76. 25%，the Kappa coefficients are above 0. 6，and the classification effect is better than the traditional partial least squares discriminant analysis. According to the classification results，the recognition accuracy of the random forest discriminant model with fusion of texture information is significantly improved，the classification average accuracy is 77. 20%，the Kappa coefficient is above 0. 64，and the classification average accuracy based on the effective band and texture information discriminant model reaches 78. 30%. The Kappa coefficient reaches 0. 675. The research shows that the random forest algorithm under the combination of effective band and texture information features can make full use of hyperspectral spectral features and texture features to accurately identify maize varieties and provide a new method for automatic identification of maize varieties.

Key words：hyperspectral image；maize；random forest；partial least squares discriminant analysis

“农业高光谱遥感”专栏组稿专家简介

本期“农业高光谱遥感”专栏由北京师范大学地理学部陈云浩教授(博士生导师)和中国矿业大学(北京)地球科学与测绘工程学院蒋金豹教授(博士生导师)组稿并担任责编。陈云浩于 2000 年获中国矿业大学(北京)博士学位，2001－2002 年在北京师范大学开展博士后研究工作，入选北京市科技新星计划 A 类、北京市优秀人才培养资助计划、教育部新世纪优秀人才，是北京市优秀博士学位论文指导教师；研究方向为热红外与高光谱遥感，目前主要关注机载与实验室成像高光谱技术研发与应用。蒋金豹于 2007－2008 年在英国诺丁汉大学地理学院深造，2009 年获北京师范大学博士学位并就职于中国矿业大学(北京)；主研方向为高光谱遥感在农业病害、天然气泄漏监测、食品安全检测及矿物识别方面的应用，构建了一系列高光谱遥感识别与检测模型，拓展了高光谱遥感应用研究的广度与深度。