

高光谱成像鉴别玉米品种早期抗倒性

张天亮, 张东兴, 崔涛, 杨丽*, 解春季, 杜兆辉, 钟翔君

中国农业大学工学院, 农业部土壤-机器-植物系统技术重点实验室, 北京 100083

摘要 针对传统玉米品种抗倒性鉴别方法费时费力、时效滞后的问题, 采用高光谱成像数据结合机器学习方法对9叶期的玉米品种抗倒性进行鉴别, 并给出适于进行玉米品种抗倒性鉴别的种植密度和建模方法。试验设置了5 000, 7 000和9 000株·亩⁻¹3个种植密度和6个典型的抗倒/不抗倒玉米品种, 采集9叶期玉米顶叶的高光谱图像, 使用目标区域分割的方式自动进行光谱图像反射率校正和目标光谱曲线提取。对采集的样本数据使用Kennard Stone算法划分样本训练集和测试集, 用主成分分析法(PCA)和连续投影算法(SPA)提取光谱特征, 建立了基于高斯核函数的支持向量机(SVM)模型并进行参数训练和优化。通过对不同种植密度下各特征提取方法的效果和各模型训练效果及其预测结果的对比, 找到进行玉米抗倒性鉴别的最佳种植密度和建模方法。试验结果表明: 在各种种植密度下PCA方法对光谱特征的降维效果最为显著, 而SPA算法选择的特征波长分布比较均匀、抗倒性分类特征比较明显; 种植密度的增加对于玉米品种抗倒性的鉴别是有益的, 在种植密度为7 000株·亩⁻¹时, 使用SPA-SVM方法建立的模型训练效果和预测结果最佳, 此时模型对训练集数据的10折交叉验证正确率为97.40%, 对测试集数据的预测正确率为98.33%。

关键词 玉米抗倒; 高光谱成像; 主成分分析; 连续投影; 支持向量机

中图分类号: S126 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2022)04-1229-06

引言

倒伏是影响玉米产量和田间机械化收获的关键因素, 严重的倒伏情况会造成玉米大幅减产甚至绝收, 因此对玉米品种的抗倒性进行评价和筛选具有十分重要的意义。传统的玉米品种抗倒性鉴别方法主要是通过对玉米灌浆期、穗期等的株高叶形, 茎秆力学性质, 茎秆物质积累等的统计分析, 结合茎秆的显微结构和玉米基因型对玉米品种的抗倒性进行评价和鉴定。例如吴琼等^[1]通过对不同种植密度下玉米品种在灌浆期的株高穗高、茎秆基部穿刺强度、可溶性糖含量等10项指标的统计分析, 提出了4个互相独立的综合指标来鉴定玉米品种的耐密植抗倒伏能力。Zhang等^[2]研究了两个不同玉米品种在第二和第三茎节间的维管束面积、直径等解剖特征, 证明了维管束的微观表型是评估玉米茎秆机械性的重要指标, 并开发了相应的表型检测软件。田再民等^[3]研究了种植密度对玉米品种抗倒伏性的影响, 发现随着种植密度的增

加, 玉米的穗高系数逐渐升高、茎秆穿刺强度逐渐下降, 玉米品种抗倒性也逐渐降低。

传统玉米品种抗倒性鉴定方法普遍是在穗期采集数据进行分析, 具有人力时间成本高、抗倒性评价滞后等缺点。如果能在玉米生长早期对品种抗倒性进行快速鉴别, 将对玉米品种抗倒性的筛选产生重大意义。高光谱成像技术具有高通量、图谱合一等优势, 可以对植物的内外表型进行检测, 在对玉米表型检测和识别方面也有非常广泛的应用前景。Qin等^[4]使用机载高光谱数据对玉米冠层吸收的光合有效辐射进行估算, 以此来监测玉米的生长和估算产量。Feng等^[5]使用高光谱成像技术结合化学计量学方法对不同品种的玉米种子活力进行评估, 实现了对不同加速老化条件下的玉米种子的老化程度的准确评估。本研究将使用高光谱成像对玉米生长早期的品种抗倒性进行研究, 通过对比不同抗倒性玉米品种在不同种植密度条件下的高光谱数据差异, 筛选出适合进行玉米品种抗倒性鉴定的种植密度和特征波段并建立模型进行抗倒性评价, 为玉米品种抗倒性鉴别提供了一种新的方法。

收稿日期: 2021-04-09, 修订日期: 2021-06-24

基金项目: 国家现代玉米产业技术体系建设项目(CARS-02), 国家重点研发计划项目(2016YFD0300302)资助

作者简介: 张天亮, 1993年生, 中国农业大学工学院博士研究生 e-mail: tianliangzn@163.com

* 通讯作者 e-mail: yangli@cau.edu.cn

1 实验部分

1.1 材料

选择 3 个玉米抗倒伏品种登海 605(DH605)、登海 618(DH618)、京单 28(JD28) 和 3 个不抗倒伏品种浚单 20(XD20)、隆平 208(LP208)、纪元 1 号(JY1), 于 2019 年夏季在河北省沧州市吴桥县中国农业大学吴桥试验站进行种植。在同一地块内设置了 5 000, 7 000 和 9 000 株·亩⁻¹ 3 个密度小区, 每个小区种植 9 行, 行距 0.66 m, 行长 7 m。每个品种在每个密度下重复种植 2 次, 共种植 36 个小区。在玉米生长过程中未喷施植物生长调节剂类的药物, 其他田间管理措施按当地的大田管理方式实行。

1.2 高光谱图像采集

在玉米生长至第 9 片顶叶完全展开时, 于田间采集玉米的第 9 片叶。在小区内挑选长势接近的植株, 自叶片中间向两端取 1/3 叶片长度的样品, 用自封袋密封放入装有冰袋的保鲜盒内, 带回试验站用于拍摄叶片图像。为了确保试验的样本容量且能在短时间内完成样品拍摄, 经计算拍摄时间确定在每个小区取样 21 株, 并以每 3 个叶片为一组进行拍摄。

表 1 各密度下玉米样本数据集划分情况
Table 1 The division of maize sample data sets

项目	5 000 株·亩 ⁻¹			7 000 株·亩 ⁻¹			9 000 株·亩 ⁻¹		
	抗倒样本	不抗倒样本	总计	抗倒样本	不抗倒样本	总计	抗倒样本	不抗倒样本	总计
训练集	95	91	186	96	96	192	94	95	189
测试集	30	30	60	30	30	60	30	30	60
总计	125	121	246	126	126	252	124	125	249

1.4 特征提取与建模方法

分别使用主成分分析(principal component analysis, PCA)和连续投影方法(successive projections algorithm, SPA)对预处理后的训练集数据进行特征提取, 以选择含有最少冗余信息和最小共线性的特征变量组合。其中主成分方法对线性相关的原特征变量进行了线性变换以得到新的线性不相关的主成分, 并使用主成分贡献率来筛选特征变量^[6-7]。而连续投影方法只是筛选出冗余性少、共线性小的特征变量, 不对原特征变量进行操作, 并根据迭代特征向量与待选变量个数回归模型的均方根误差来确定候选特征子集^[8]。

使用高斯核函数(Gaussian kernel)的支持向量机(support vector machines, SVM)方法对筛选出的特征进行建模, 对模型中的惩罚参数 C 和核参数 γ 使用网格搜索(grid search)法进行模型优化^[9]。其中训练集的模型建立使用 10 折交叉验证法进行模型评价, 以增强训练模型的泛化能力、避免模型过拟合, 测试集验证使用分类正确率和受试者工作特征曲线(receiver operating characteristic curve, ROC)进行模型分类效果评价。

2 结果与讨论

2.1 原始光谱图像分析

试验用的是美国 SOC(Surface Optics Corporation)公司的 SOC710VP 型号高光谱成像光谱仪, 光谱波长范围 374~1 038 nm, 光谱分辨率 4.68 nm, 图像分辨率 696×520, 可以在 128 个波段上进行成像, 使用灰度标准板进行反射率校正。在光学暗箱内进行拍摄, 使用 4 个 100 W 的卤素钨灯光源照明, 光谱相机布置在距叶片正上方 75 cm 处, 使叶片样本和标准板一起进行拍摄。

1.3 样品处理

试验在每一个种植密度下分别采集了 252 个样本, 并提取整片叶的平均光谱进行分析, 以确保不丢失任何光谱信息。对每个品种的光谱曲线进行观察, 剔除明显偏离样本中心的光谱曲线。然后对每个品种单独使用 Kennard Stone 算法, 按训练集和测试集 3:1 的比例划分样本, 再将各品种划分的数据集组合成训练集和测试集数据, 以确保数据集内各品种的样本分布均匀, 具体数据集划分结果如表 1 所示。然后对每条样本数据使用标准正态变换(standardized normal variate, SNV)方法进行预处理, 再对每个密度下的训练集数据的特征进行归一化处理, 最后对每个测试集数据的特征使用相对应的训练集数据参数进行归一化处理, 使得训练集数据和测试集数据满足同分布的条件。

采集的光谱图像首先要进行反射率校正。根据反射率校正公式(1), 要得到校正后图像的反射率 R , 首先要得到标准板区域的反射强度 I_{std} , 计算标准板反射强度 I_{std} 与反射率 R_{std} 的比值得出总光照强度, 再用整个光谱图像的反射强度 I 除以总光照强度即得到光谱图像的反射率, 其中标准板的反射率 R_{std} 是已知的。

$$R = \frac{I}{I_{std}/R_{std}} \quad (1)$$

经过分析叶片、标准板和图像背景在 128 个波段上的光谱反射强度曲线, 选取 779 和 648 nm 处的波段图像, 分别设置阈值范围为 [500, 1 800] 及 [1 000, 2 000], 对图像进行二值化处理并取图像的交集, 可以得到图像的标准板区域。然后使用 k-means 聚类方法将该区域聚类成阴影区和正常反射区^[10-11], 聚类效果如图 1(b)所示。最后提取出标准板的正常反射区在各波段上的平均反射强度得到 I_{std} , 再根据式(1)即可求得光谱图像的反射率校正数据。

对校正后的反射率图像还要提取出每个叶片的平均光谱反射率曲线。同样使用阈值分割的方法, 选择 779 nm 处的反射率图像, 设置反射率分割阈值为 0.25, 二值化处理后的叶片分割效果如图 1(c)所示。然后依次计算每个叶片区域在所有波段上的反射率平均值, 即得到各叶片的平均光谱反射率曲线。计算整片叶的光谱数据, 可以确保不丢失任何有用

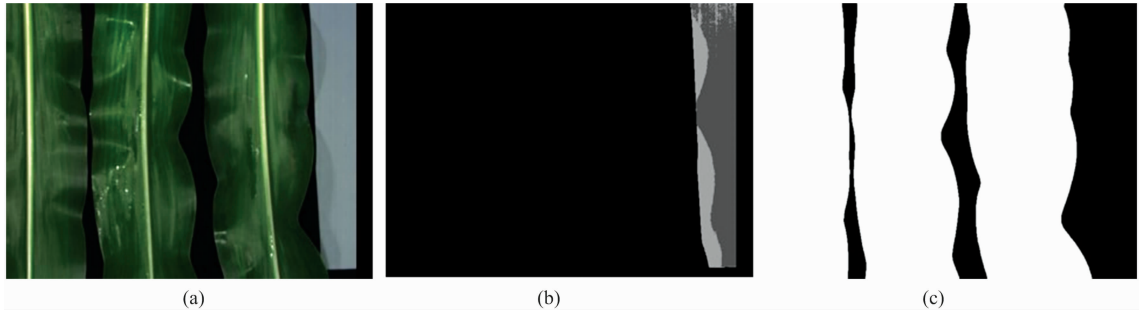


图 1 光谱图像校正与光谱提取方法

(a): 叶片彩图; (b): 标准板聚类分割; (c): 叶片分割效果

Fig. 1 Spectral image correction and spectral extraction method

(a): Color image of leaves; (b): k-means cluster result; (c): Leaves segmentation result

的信息,同时使用平均光谱反射率可以降低不同类别样本之间可能存在的相似像素处的光谱信息对最终模型分类效果产生的影响。

经过反射率校正后,分别提取各密度下的抗倒品种登海 605 和不抗倒品种隆平 208 的所有样本的反射率光谱曲线,对各组样本曲线求反射率平均值后进行绘图比较,如图 2 所示。从图中可以看出各玉米品种的光谱曲线变化趋势基本一致,但在具体的波段范围内也存在明显的光谱差异。品种间进行比较,可以看出在 750~980 nm 波长范围内,登海 605 品种的光谱反射率在各密度下的光谱反射率都高于隆平 208 的最大反射率。在品种内部进行比较,可以看出登海 605 品种在 760~940 nm 波长范围内存在不同密度下的光谱反射率差异,7 000 株·亩⁻¹密度下的光谱反射率最高,而 9 000 株·亩⁻¹密度下的光谱反射率最低;隆平 208 品种在 960~1 000 nm 波长范围内也存在和登海 605 一样的密度光谱差异,而在 760~940 nm 波长范围内则是 9 000 株·亩⁻¹密度下的反射率最低、5 000 和 7 000 株·亩⁻¹密度下的反射率光谱十分接近。综上比较可以看出,不仅不同抗倒性品种的光谱曲线间存在差异,同一品种不同密度下的光谱反射率曲线也存在差异,为玉米品种抗倒性的鉴别提供了条件。

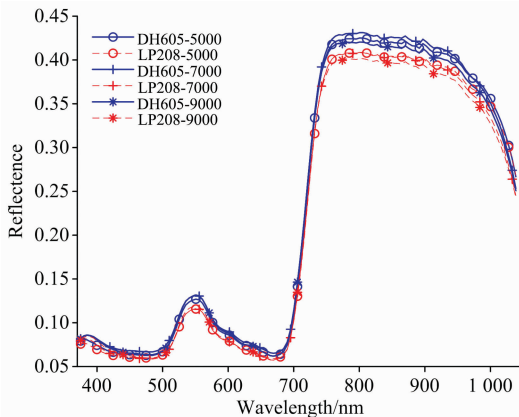


图 2 抗倒品种和不抗倒品种的平均光谱曲线

Fig. 2 Average spectral curves of lodging-resistant varieties and lodging varieties

2.2 光谱数据特征提取

对所有样品的反射率光谱数据进行 SNV 处理,再对训练集数据的各特征进行中心化处理^[12],然后使用主成分分析和连续投影算法进行特征降维与特征选择。主成分分析法在实际使用中只需要保留方差贡献率最大的前几个主成分,就可以包含原始数据中的主要信息。以累计方差贡献率达到 99%为限,决定待选主成分数量,各密度下主成分的贡献率如表 2 所示。可以看出各密度下的前 2 个主成分贡献率最大,前 3 个主成分的累计贡献率都已经超过 90%,说明主成分分析法对光谱特征的降维效果显著,同时各密度下的光谱特征都存在光谱共线性问题。根据表 2 选择三种密度下参与模型训练的主成分数分别为前 8, 10 和 8 个主成分。连续投影算法将各波长特征向量投影到其他特征波长上,以投影向量最大的波长作为待选的特征波长,并根据迭代特征向量与待选变量个数回归模型的均方根误差来确定候选特征数量^[13]。通过选择最小的均方根误差值确定各密度下的特征波长数量分别为 37, 28 和 14 个,各密度下的特征波长分布情况如图 3 所示。从图中可以看到,各密度下在 750~1 000 nm 范围内被选中的特征波长分布最密集,说明此波段范围是

表 2 各密度下主成分的贡献率

Table 2 Contribution rate of principal component at each density

主成分	5 000 株·亩 ⁻¹		7 000 株·亩 ⁻¹		9 000 株·亩 ⁻¹	
	贡献率 /%	累计贡献率 /%	贡献率 /%	累计贡献率 /%	贡献率 /%	累计贡献率 /%
PC1	49.28	49.28	45.59	45.59	48.31	48.31
PC2	34.30	83.58	32.10	77.69	36.06	84.37
PC3	8.77	92.35	12.42	90.11	7.58	91.95
PC4	3.07	95.42	3.16	93.27	3.40	95.35
PC5	1.92	97.34	2.52	95.79	1.76	97.11
PC6	1.13	98.47	1.15	96.94	0.87	97.78
PC7	0.50	98.97	1.09	98.03	0.61	98.59
PC8	0.21	99.18	0.56	98.59	0.41	99.00
PC9	0.19	99.37	0.39	98.98	0.24	99.24
PC10	0.17	99.54	0.25	99.23	0.18	99.42

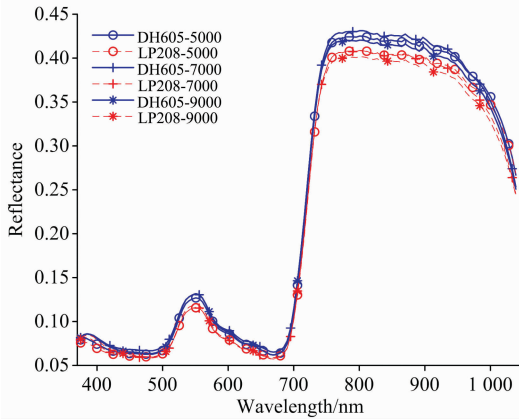


图 3 用 SPA 算法在各密度下选择的光谱特征
Fig. 3 Spectral characteristics selected by the SPA algorithm at various densities

鉴别玉米品种抗倒性的特征波段, 与对原始光谱图像进行分析时得出的结论保持一致。同时用 SPA 算法选择的特征波长在整个光谱曲线范围上都有分布, 而使用 PCA 方法选择的特征都集中在前几个主成分里。

2.3 模型结果与分析

2.3.1 模型训练与优化

将各密度下训练集中被选中的主成分特征与 SPA 算法选出的波段特征分别代入 SVM 模型, 用网格搜索法对模型的惩罚参数 C 和核参数 γ 进行优化选择, 用 10 折交叉验证法对每次的模型训练结果进行评价, 选择模型效果最好的一组参数作为最终的模型参数进行建模预测。其中惩罚参数 C 设定的搜索范围是: $2^{-5}, 2^{-3}, \dots, 2^{29}$, 核参数 γ 设定的搜索范围是: $2^{-27}, 2^{-25}, \dots, 2^{13}$ 。被选中的模型参数及对应的训练集交叉验证正确率见表 3, 各密度下的参数优化效果如图 4 所示。从模型的优化效果可以看出, 参数的选择对于模型分类正确率的影响十分显著。

表 3 各密度下被选中的模型参数

Table 3 The selected model parameters at each density

项目	5 000 株 · 亩 ⁻¹		7 000 株 · 亩 ⁻¹		9 000 株 · 亩 ⁻¹	
	PCA-SVM	SPA-SVM	PCA-SVM	SPA-SVM	PCA-SVM	SPA-SVM
惩罚参数 C	2^9	2^{17}	2^{11}	2^{19}	2^3	2^{11}
核参数 γ	2^{-11}	2^{-13}	2^{-11}	2^{-15}	2^{-7}	2^{-7}
交叉验证正确率/%	79.03	100	83.33	97.40	76.19	98.94
测试集分类正确率/%	78.33	83.33	83.33	98.33	71.67	95

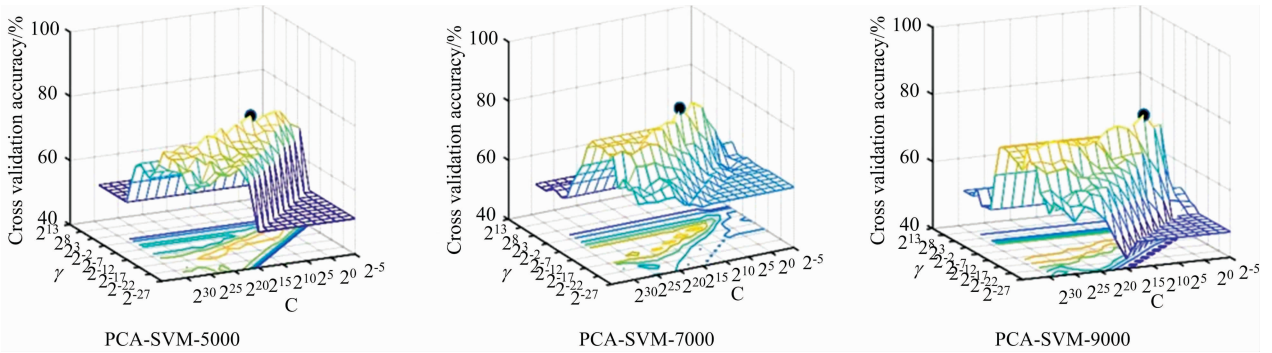


图 4(a) 各密度下 PCA-SVM 模型的参数优化效果

Fig. 4(a) Parameter optimization effects of PCA-SVM model at various densities

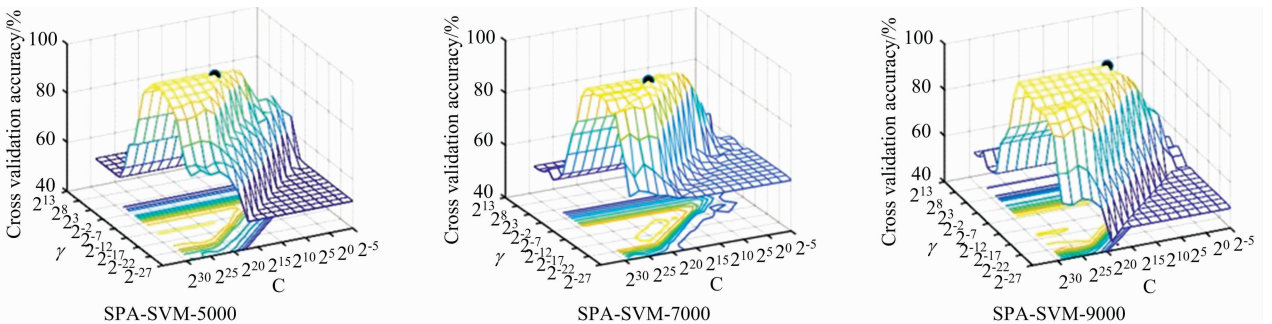


图 4(b) 各密度下 SPA-SVM 模型的参数优化效果

Fig. 4(b) Parameter optimization effects of SPA-SVM model at various densities

2.3.2 模型结果对比

将各密度下 PCA-SVM 模型和 SPA-SVM 模型的最佳参数分别代入训练集数据建立预测模型, 对测试集数据进行预测, 各模型对测试集的预测结果见表 3 所示。其中在 7 000 株·亩⁻¹密度下 PCA-SVM 模型和 SPA-SVM 模型的预测效果均达到最好, 预测正确率分别为 83.33% 和 98.33%; SPA-SVM 模型在各密度下的预测效果均优于 PCA-SVM 模型, 与训练集各模型的训练效果保持一致; SPA-SVM 模型在 5 000 株·亩⁻¹密度下训练集的分类正确率达到 100%, 而在测试集上的分类正确率为 83.33%, 预测正确率明显降低, 而其他模型训练集和测试集的分类正确率差异不大, 说

明 SPA-SVM 模型在 5 000 株·亩⁻¹密度下训练时应该存在过拟合问题。绘制各模型预测效果的 ROC 曲线, 如图 5 所示。在各密度下 SPA-SVM 模型的 ROC 曲线均“包裹”住了 PCA-SVM 模型的 ROC 曲线, 说明 SPA-SVM 模型的预测性能要优于 PCA-SVM 模型^[14-15]; 其中在 7 000 株·亩⁻¹密度下, SPA-SVM 模型的 ROC 曲线下面积为 1, 模型分类正确率可以达到 100%。从各密度下模型的训练效果和预测的结果可以看出, 7 000 株·亩⁻¹密度下模型的训练和预测效果已经达到最佳, 是预测玉米品种抗倒性比较推荐的种植密度, 同时 SPA-SVM 模型则是推荐的预测模型。

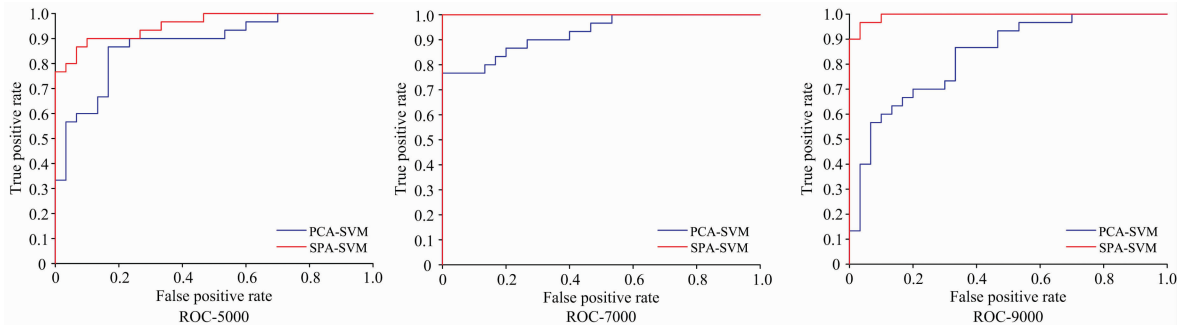


图 5 各密度下模型预测效果的 ROC 曲线

Fig. 5 ROC curve of model prediction effect at each density

3 结 论

采用高光谱成像技术对 3 个种植密度下 6 个玉米品种的抗倒性进行鉴别, 实现了对光谱图像反射率的自动校正和光谱提取, 对比了主成分分析法和连续投影算法的特征提取效果以及对分类模型训练和预测效果的影响, 得出了玉米抗倒性鉴别时最佳的种植密度及分类模型。主要结论如下:

(1) 主成分分析法提取光谱特征具有明显的降维效果, 只用前几个主成分就能代表光谱数据的大部分特征信息, 但用 PCA 方法提取特征建立的模型对玉米抗倒性的分类效果

一般; 使用连续投影算法选择的光谱特征在整个光谱范围都有分布, 且选中的特征波段在抗倒性区分明显的 750~1 000 nm 范围内分布最为集中。SPA 算法对于玉米品种抗倒性特征的选择要优于 PCA 方法。

(2) SPA-SVM 模型在各密度下的模型训练效果和预测结果均优于 PCA-SVM 模型, 其在 7 000 株·亩⁻¹密度下的测试集分类正确率达到了 98.33%, 是推荐的建模方法。适当的增加种植密度对于玉米品种的抗倒性分类是有益的, 7 000 株·亩⁻¹密度下各模型的训练效果和预测结果已经达到最佳, 是适于进行玉米抗倒性鉴别的种植密度。

References

- [1] WU Qiong, YANG Ke-jun, ZHANG Yi-fei, et al(吴琼, 杨克军, 张翼飞, 等). Journal of Maize Sciences(玉米科学), 2017, 25(4): 40.
- [2] Zhang Ying, Du Jianjun, Wang Jinglu, et al. Biological Research, 2018, 51(1): 40.
- [3] TIAN Zai-min, HUANG Zhi-hong, CHEN Jian-xin, et al(田再民, 黄智鸿, 陈建新, 等). Journal of Maize Sciences(玉米科学), 2016, 24(5): 83.
- [4] Qin Haiming, Wang Cheng, Zhao Kaiguang, et al. PLOS ONE, 2018, 13(5): e197510.
- [5] Feng L, Zhu S, Zhang C, et al. Molecules, 2018, 23(12): 3078.
- [6] LI Xiao-bin, WANG Yu-shun, FU Li-hong, et al(李晓斌, 王玉顺, 付丽红, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2016, 36(12): 179.
- [7] Chen Y, Zhao X, Jia X. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015, 8(6): 2381.
- [8] ZHAO Fan, YAN Zhao-ru, XUE Jian-xin, et al(赵凡, 闫昭如, 薛建新, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2021, 41(1): 201.
- [9] Chang C, Lin C. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1.
- [10] XIE Wen-yong, CHAI Qin-qin, GAN Yong-hui, et al(谢文涌, 柴琴琴, 甘勇辉, 等). Transactions of the Chinese Society of Agricultural

- Engineering(农业工程学报), 2020, 36(14): 203.
- [11] Arthur D, Vassilvitskii S. K-Means++: The Advantages of Careful Seeding, Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007. 1027.
- [12] Munera S, Amigo J M, Aleixos N, et al. Food Control, 2018, 86: 1.
- [13] Araújo M C U, Saldanha T C B, Galvão R K H, et al. Chemometrics and Intelligent Laboratory Systems, 2001, 57(2): 65.
- [14] Wu Y, López S, Zhang B, et al. Journal of Real-Time Image Processing, 2019, 16(1): 99.
- [15] Wang L, Chang C, Lee L, et al. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(9): 4887.

Identification of Early Lodging Resistance of Maize by Hyperspectral Imaging Technology

ZHANG Tian-liang, ZHANG Dong-xing, CUI Tao, YANG Li*, XIE Chun-ji, DU Zhao-hui, ZHONG Xiang-jun
Key Laboratory of Soil-Machine-Plant System Technology of Ministry of Agriculture, College of Engineering, China Agricultural University, Beijing 100083, China

Abstract Given the time-consuming, labor-intensive and time-lagging problems of traditional methods for identifying lodging resistance of maize, this study uses hyperspectral imaging data combined with machine learning methods to identify lodging resistance of maize at the 9-leaf stage. It gives the recommended planting density and modeling methods. The experiment set up 3 planting densities of 5 000, 7 000 and 9 000 plants \cdot μm^{-2} and 6 typical lodging resistant/non-lodging resistant varieties. Hyperspectral images of corn top leaves at the 9-leaf stage were collected, reflectance correction and target spectral curve extraction were automatically performed by segmentation of the target area. For the collected sample data, the Kennard Stone algorithm is used to divide the sample training set and the test set, the principal component analysis (PCA) and the successive projections algorithm (SPA) are used to extract the spectral features. A support vector machines (SVM) model of Gaussian kernel function is established, with the performing of parameter training and optimization. By comparing the effect of each feature extraction method and the training effect of each model, and its prediction results under different planting densities, the planting density and modeling method recommended for the identification of maize lodging resistance were found. The test results show that the PCA method has the most significant dimensionality reduction effect on the spectral features at various planting densities. At the same time, the characteristic wavelength distribution selected by the SPA algorithm is relatively uniform, and the lodging resistance classification characteristics are obvious. The increase of planting density is beneficial to identifying the lodging resistance of maize. When the planting density is 7 000 plants \cdot μm^{-2} , the training effect and prediction results of the model established by the SPA-SVM method are the best. The 10-fold cross-validation accuracy of the model on the training set data is 97.40%. The prediction accuracy rate of the set data is 98.33%.

Keywords Maize lodging resistance; Hyperspectral imaging; PCA; SPA; SVM

(Received Apr. 9, 2021; accepted Jun. 24, 2021)

* Corresponding author