第42卷，第6期
2022年6月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42, No. 6, pp1716-1720
June, 2022

近红外光谱的玉米种子穗腐病特征提取与判别模型研究

孟繁佳[1]，罗 石[1]，吴月峰[1]，孙 红[1]，刘 飞[2]，李民赞[1*]，黄 威[3]，李 穆[3]

1. 中国农业大学现代精细农业系统集成研究教育部重点实验室，北京 100083
2. 浙江大学生物系统工程与食品科学学院，浙江 杭州 310058
3. 吉林省农业科学院玉米研究所，吉林 长春 130033

摘 要 玉米种子穗腐病是危害玉米产量的主要病害之一。利用近红外光谱开展了玉米种子穗腐病判别模型研究。246粒玉米种子由吉林省农业科学院海南育种基地提供，其中96粒玉米种子为穗腐病染病样本，其他150粒玉米种子为同种玉米正常样本。利用MATRIX-Ⅰ型傅里叶近红外光谱仪采集了样本800～2 500 nm范围的近红外光谱信息，并对样本近红外光谱数据利用多元散射校正(MSC)进行预处理。结合玉米内部有机物质的近红外光谱的敏感波段和样本近红外光谱吸收峰挑选了4个优选区间，并采用相关系数法(CA)、连续投影算法(SPA)和竞争性自适应重加权算法(CARS)三种不同原理的特征波长提取算法分别提取了4(1 362，1 760，2 143和2 311 nm)、5(1 227，1 310，1 382，1 450和1 728 nm)和10(1 232，1 233，1 257，1 279，1 313，1 688，1 703，1 705，2 302和2 323 nm)个特征波长。以提取得到的特征波长作为玉米种子穗腐病判别模型输入变量，用0-1(染病-正常)表示样本染病状况作为输出真实值建立支持向量机(SVM)模型，使用网格搜索法结合十折交叉验证法对模型参数进行优化。结果表明，CA-SVM，SPA-SVM和CARS-SVM三种判别模型中训练集和测试集建模准确率均在90%以上。该研究成果为玉米种子病害诊断装置提供了模型基础，且针对优选区间进行特征波长选择的方式也可以为建立其他种子病害判别模型提供参考。

关键词 近红外光谱；玉米种子；穗腐病；特征波长

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)06-1716-05

引 言

玉米作为世界上主要粮食作物，含有丰富的营养物质。2019年，我国的玉米种植面积也已经超过4 100万hm^2，每年的作物产量占我国粮食总产量的20%左右[1]。玉米还是工业乙醇的主要原料和饲料产业的重要原材料。玉米的生产安全直接影响我国的粮食安全。由于我国多样的气候环境，玉米易受到病害侵袭，每年作物病害均对玉米产量造成严重损失。玉米穗腐病是危害玉米产量的主要病害之一，严重时可造成玉米亩产减少30%～40%。初侵染病原由种子传播。带病种子播种后，存在大概率无法出苗问题，且出苗后病原体仍会感染植株进而通过孢子借助风雨传播，对大田种植具有危害性。在播种前检测玉米种子是否被病原体侵染对于防治玉米穗腐病具有重要实践意义。

目前，对作物种子的病害检测方法包括免疫分析法、理化分析法和人工检测。其中，以酶联免疫吸附法和免疫亲和层析法为代表的免疫分析法和以高效液相色谱法和气相色谱法等仪器探测的理化分析法被认为是高精度检测种子病害病原体的手段。但高昂的仪器设备、复杂的测试过程以及破坏性检测方法等因素限制了它们在种子病害检测领域的应用[2-3]。人工检测手段仅仅能够通过视觉嗅觉等手段甄别病害种子，经验依赖性高、效率低且错检率高。

近红外光谱分析作为一种无损检测手段，通过对物质不同光谱波段的吸收度进行定性及定量分析，被广泛运用于农作物品质检测，生长周期检测和类别检测中[4-5]。Chu等基于近红外高光谱建立了玉米种子真菌感染判别的两种分类模型，准确率分别为97.96%和98.94%[6]；Shen利用多元散射校正(multiplicative scatter correction，MSC)和线性判别分析对玉米种子真菌感染水平进行分类，准确率达到了

收稿日期：2021-05-08，**修订日期**：2021-07-13

基金项目：国家重点研发计划项目(2018YFD010100202)，中央高校基本科研业务费(2020TC143)和国家自然科学基金项目(31401294)资助

作者简介：孟繁佳，1983年生，中国农业大学现代精细农业系统集成研究教育部重点实验室高级工程师 e-mail：mengfanjia@126.com

*通讯作者 e-mail：limz@cau.edu.cn

86.7%[7]；Daniel 等基于近红外高光谱对玉米种子表面黄曲霉毒素 B_1 浓度进行检测，结合主成分分析和因子判别分析法建立的判别模型准确率为 96%[8]；Tao 等基于近红外光谱对感染黄曲霉毒素的玉米种子进行分类，建立的偏最小二乘判别分析模型准确率为 96.3%[9]。上述研究基于单一真菌或真菌代谢毒素感染后的玉米种子近红外光谱信息建立了判别模型，但自然感染穗腐病的玉米种子致病病原菌存在 20 多种，与单一真菌感染情况相比，玉米种子内部物质变化情况不具备一致性。

本工作基于近红外光谱技术，通过优选区间和特征波长提取方法对玉米种子穗腐病判别模型进行研究，以期为后续玉米种子病害诊断装置科学模型依据。

1　实验部分

1.1　试验材料

玉米种子样本均由吉林省农业科学院海南育种基地提供。海南育种中心为中国北方玉米种子产业提供种子培育试验田，玉米穗腐病为试验田中的主要病害。玉米籽粒均在 2019 年收获，收获时籽粒成熟度一致。经过育种基地验证，玉米籽粒被划分为染病籽粒与健康籽粒，运输过程中两类籽粒被分开包装，运输至实验室后放置在低温、干燥的环境中保存。

为避免水分影响，选择样本时剔除了干瘪的染病玉米种子。共选用玉米种子 246 粒，其中包含 150 份健康玉米种子及 96 份穗腐病染病玉米种子。

图 1　试验样本

Fig. 1　Test samples

1.2　光谱采集仪器

试验使用德国 BRUKER 公司生产的 MATRIX-Ⅰ型号傅里叶近红外光谱仪进行全谱光谱信息采集，获得玉米籽粒在 4 000～12 000 cm^{-1} 的近红外光谱吸收度信息。在采集光谱时，玉米种子均为胚面向下放置。仪器参数设置如下：分辨率为 8 cm^{-1}，单次采集扫描次数为 32(重复扫描，求平均光谱)。单个玉米种子样本采集 1 037 个近红外光谱数据。使用波长描述光谱信息，波数与波长转换公式为

$$\nu = 10^7/\lambda \tag{1}$$

式(1)中，ν 为波数，单位为 cm^{-1}；λ 为波长，单位为 nm。

1.3　MSC 预处理方法

近红外区域的光谱信号较弱，易受到外界环境干扰，且光谱采集时的背景光谱也会产生细微偏差，所以连续测量同一类种子时产生的光谱数据会产生基线漂移。近红外光谱数据进行分析处理前必须进行预处理，加强光谱数据所包含信息的可靠性。MSC 是一种光谱预处理手段，能够有效降低光谱采集时的散射影响，提高光谱数据信噪比，修正光谱基线漂移的同时对样品对应的光谱吸收信息没有影响[10]。

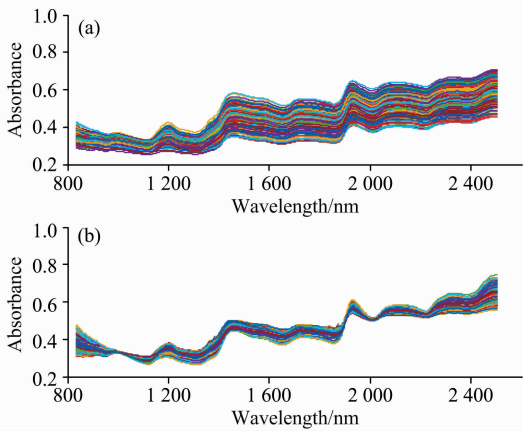

图 2　原始光谱及 MSC 处理

Fig. 2　Original spectra (a) and preprocessed by MSC (b)

1.4　优选光谱区间方法

根据已有研究[11]，玉米种子中水分子—OH 官能团吸收谱带为 920～1 950 nm，蛋白质—NH 官能团吸收谱带为 1 560～1 670 和 2 080～2 220 nm，脂肪 C—H，C—H_2 和 C—H_3 官能团吸收谱带为 2 300～2 350 和 1 680～1 760 nm，碳水化合物—CO 和—OH 官能团吸收谱带为 2 060～2 150 nm。玉米种子受到穗腐病病原菌侵染后，由于病原菌的生理活动导致玉米种子内部脂肪、蛋白质、淀粉等有机物质发生氧化作用和水解反应，物质组成成分区别于正常玉米种子。根据样本玉米种子近红外光谱中的吸收峰，选取四个优选区间 T1(1 204～1 449 nm)、T2(1 560～1 760 nm)、T3 (2 060～2 220 nm)、T4(2 300～2 395 nm)。其中 T1 表征水分子吸收谱带，T2 表征蛋白质及脂肪吸收谱带，T3 表征碳水化合物及蛋白质吸收谱带，T4 表征脂肪吸收谱带。图 3 为经过 MSC 预处理后优选区间的近红外光谱吸收度。

1.5　特征波长选择算法

建立玉米种子穗腐病判别模型时，需要以玉米种子的近红外光谱数据作为输入变量，但以全光谱波段作为输入则模型计算时间过长，且存在信号谱带重叠；进行特征提取可以消除具有共线性关系的原始数据，提高建模稳定性。选择了三种具有代表性的光谱特征提取算法：相关系数法(correlation analysis，CA)[12]、连续投影算法(successive projections algorithm，SPA)[13]和竞争性自适应重加权算法(competitive adaptive reweighted sampling，CARS)[14]，三种算法原理不同，但均可以对原始光谱数据进行波长筛选，提取特征变量作为判别模型的输入。

1.6　SVM 模型

支持向量机(support vector machine，SVM)是一种二分类模型，用于生成样本的特征空间上间隔最大的线性分类器。用样本空间中的少数样本作为支持向量，求出不同类别样本欧几里得距离最大的分离超平面，作为模型输出。SVM 的优势是在样本数量较少的情况下生成模型稳定性高，且可以通过更换 RBF 核函数求解非线性分类问题，适用于本工作

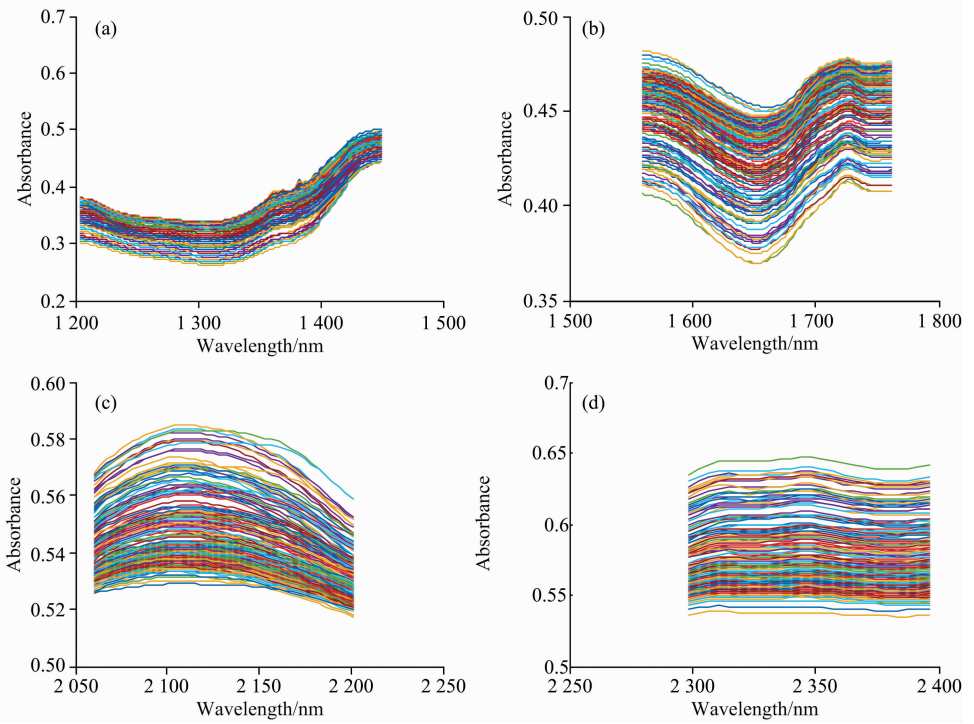

图 3 玉米种子近红外光谱优选区间

Fig. 3 Near infrared spectra of maize seed in 4 optimal wavelength ranges

的二分类判别模型。SVM 的训练集和验证集以 7∶3 的比例随机划分，并将随机种子数设置为 1，保证数据集划分稳定性，训练集及测试集组成如表 1 所示。使用网格搜索法和十折交叉验证法确定模型参数。

表 1 数据集划分结果

Table 1 Data set partitioning results

	训练集	测试集
正常玉米种子	106	44
染病玉米种子	66	30

2 结果与讨论

2.1 CA-SVM 玉米种子穗腐病判别模型

采用 CA 提取与样本真实值相关度最高的波长作为特征波长。用 0-1 表示种子样本状态真实值(染病样本-健康样本，下同)。图 4 为样本近红外光谱区间吸收度与真实值之间的相关系数，从图中可以看出优选区间相关系数均高于 0.6，说明有机物质的变化与样本真实值之间存在紧密联系。

选取 4 个优选区间内相关系数最高的波长作为特征波长，4 个特征波长分别为：1 362，1 760，2 143 和2 311 nm。

SVM 建模时，考虑吸收度与样本真实值非线性相关，选取 RBF 核函数。使用网格搜索法结合十折交叉验证法，对 RBF 核函数中的惩罚因子 C 和 gamma 进行参数优化，C 的取值范围设为 1～50，搜索步长为 0.1，gamma 的取值范围设为 0.1～1，搜索步长为 0.1(下同)。参数优化结果为 C 取值 46.6，gamma 取值 0.8。CA-SVM 模型训练集准确率为 92.44%，测试集准确率为 93.24%。

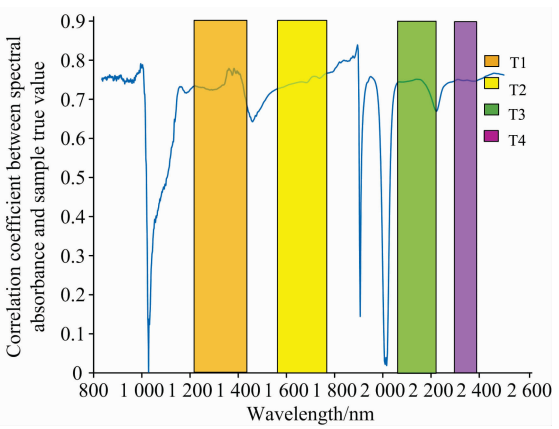

图 4 样本近红外全波段相关系数

Fig. 4 Correlation coefficient from NIR full-band spectral data

2.2 SPA-SVM 玉米种子穗腐病判别模型

采用 SPA 提取变量矢量空间共线性最小的波长组合作为特征波长。SPA 中建模集组成为 100∶60(健康样本∶染病样本，下同)，验证集为 50∶36，最大变量选择数设为 5。如图 5 所示。从优选区间提取了 5 个特征波长，分别为 1 227，1 310，1 382，1 450 和 1 728 nm。

经过参数优化，C 为 45.7，gamma 为 0.8，CA-SVM 模型训练集准确率为 91.86%，测试集准确率为 94.59%。

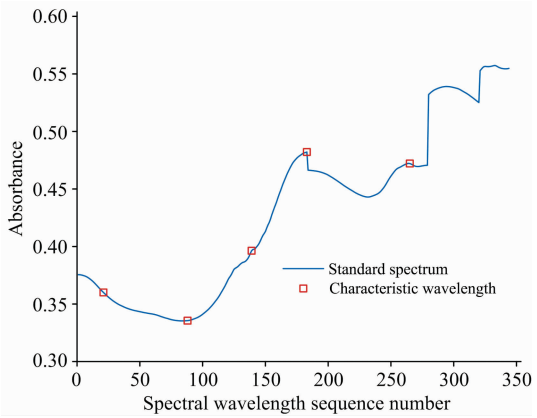

图 5　SPA 特征提取

Fig. 5　SPA feature extraction

2.3　CARS-SVM 玉米种子穗腐病判别模型

CARS 通过自适应重加权技术以 PLS 建模后的回归系数为参考值，寻找最优变量组合。CARS 建模中，设置偏最小二乘回归保留的主成分数最大值为 10，使用十折交叉验证法求均方根误差，并设置自竞争加权算法运行次数为 50 次，输出特征变量筛选结果。通过 CARS 选择的 10 个特征波长为：1 232，1 233，1 257，1 279，1 313，1 688，1 703，1 705，2 302 和 2 323 nm，选择过程如图 6 所示。

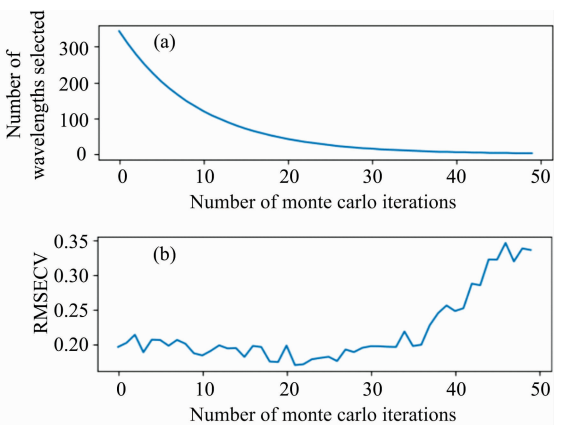

图 6　CARS 特征提取

Fig. 6　CARS feature extraction

经过参数优化，C 值为 48.3，gamma 值为 0.7。CARS-SVM 模型训练集准确率为 90.69%，测试集准确率为 93.24%。

2.4　建模结果对比与分析

由表 2 可知，经过优选区间和特征波长提取，建模所需变量数最多可以减少至原始数据波长数量的 0.38%，三种玉米种子穗腐病判别模型训练集和测试集准确率均高于 90%，且在 SPA-SVM 模型中测试集准确率最高为 94.59%，证明利用近红外光谱可以建立有效玉米种子穗腐病判别模型。同时，三种模型中，训练集准确率低于测试集准确率，这是由于进行样本划分时，选择随机种子数为 1，样本划分情况不理想，与样本数量较少也有一定关系。

表 2　判别模型对比

Table 2　Comparison of discriminant models

特征波长提取区间	判别模型	特征波长数	训练集准确率/%	测试集准确率/%
优选区间	CA-SVM	4	92.44	93.24
	SPA-SVM	5	91.86	94.59
	CARS-SVM	10	90.69	93.24

全光谱波段建立了 SVM 模型，参数优化后，C 取值 42.6，gamma 取值 0.6，训练集准确率为 93.60%，预测集准确率为 97.29%，判别模型准确率得到了一定提升，但该模型输入变量过多，不适宜应用于实际检测。

3　结　论

针对玉米种子穗腐病开展近红外光谱建模研究，结论如下：

(1)分别对经过 MSC 预处理后的原始光谱数据利用 CA，SPA 和 CARS 三种特征波长提取算法进行特征提取，使用提取出的特征波长建立 CA-SVM，SPA-SVM 和 CARS-SVM 判别模型。验证结果表明，三种判别模型中训练集和测试集准确率均在 90%以上；进行特征提取后建立的判别模型可以有效识别玉米种子穗腐病，且输入变量数量最少的仅为原始光谱变量总数的 0.38%。为后期研发玉米种子穗腐病近红外光谱检测装置提供了模型基础。

(2)结合有机物质的敏感波段进而选取优选区间进行建模分析的方法是有效的，可以对其他染病作物的近红外光谱判别模型的建立提供研究思路。

References

[1] WANG Xiao-jun, HE Ya-ping, JIANG He-ping(王晓君，何亚萍，蒋和平). Reform(改革), 2020, (9): 27.

[2] CHEN Xi, ZHANG Ming-zhe, LIN Xiao-jia, et al(陈　曦，张明哲，林晓佳，等). Acta Agriculturae Zhejiangsis(浙江农业学报), 2014, 26(5): 1273.

[3] CHEN Jin-ying, LIU Peng, WANG Xiao-qing, et al(陈晋莹，刘　鹏，王小庆，等). Grain Storage(粮食储藏), 2019, 48(3): 47.

[4] WANG Ya-li, PENG Yan-kun, ZHAO Xin-long, et al(王亚丽，彭彦昆，赵鑫龙，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2020, 51(2): 350.

[5] SUN Hong, LIU Ning, XING Zi-zheng, et al(孙　红，刘　宁，邢子正，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析),

2019, 39(6): 1870.
[6] Chu Xuan, Wang Wei, Ni Xinzhi, et al. Infrared Physics & Technology, 2020, 105: 103242.
[7] Shen Fei, Huang Yi, Jiang Xuesong, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2020, 229: 118012.
[8] Kimuli Daniel, Wang Wei, Lawrence Kurt C, et al. Biosystems Engineering, 2018, 166: 150.
[9] Tao Feifei, Yao Haibo, Hruska Zuzana, et al. Biosystems Engineering, 2020, 200: 415.
[10] LI Shang-ke, LI Pao, DU Guo-rong, et al(李尚科, 李 跑, 杜国荣, 等). Journal of Food Safety & Quality(食品安全质量检测学报), 2019, 10(24): 8024.
[11] CHU Xiao-li, CHEN Pu, LI Jing-yan, et al(褚小立, 陈 瀑, 李敬岩, 等). Journal of Instrumental Analysis(分析测试学报), 2020, 39(10): 1181.
[12] Jiang Weiwei, Lu Changhua, Zhang Yujun, et al. Analytical Methods, 2019, 11(24): 3108.
[13] WANG Xin-zhong, LU Qing, ZHANG Xiao-dong, et al(王新忠, 卢 青, 张晓东, 等). Jiangsu Journal of Agricultural Sciences(江苏农业学报), 2019, 35(5): 1197.
[14] TANG Hai-tao, MENG Xiang-tian, SU Xun-xin, et al(唐海涛, 孟祥添, 苏循新, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2021, 37(2): 105.

Characteristic Extraction Method and Discriminant Model of Ear Rot of Maize Seed Base on NIR Spectra

MENG Fan-jia[1], LUO Shi[1], WU Yue-feng[1], SUN Hong[1], LIU Fei[2], LI Min-zan[1*], HUANG Wei[3], LI Mu[3]
1. Key Laboratory of Modern Precision Agriculture System Integration Research, China Agricultural University, Beijing 100083, China
2. College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China
3. Maize Research Institute, Jilin Academy of Agricultural Sciences, Changchun 130033, China

Abstract Ear rot of corn seeds is one of the main diseases that harm the yield of corn. A discriminant model of ear rot of corn seeds was studied by near-infrared spectroscopy. The study samples were provided by the Hainan Breeding Base of Jilin Academy of Agricultural Sciences. 246 corn seeds were selected as the research objects, 96 of which were infected with ear rot, and the other 150 were normal samples of the same kind of corn. A Matrix-Ⅰ Fourier NIR spectrometer was used to collect the NIR spectra of the samples in the range of 800～2 500 nm, and the NIR spectra were preprocessed by Multiplicative Scatter Correction (MSC). Four optimal regions were selected combined with the sensitive band of NIR spectrum of organic matter in maize and the absorption peak of the NIR spectrum of samples. Correlation analysis (CA), successive projections algorithm, SPA) and Competitive Adaptive Reweighted Sampling (Competitive Adaptive Reweighted Sampling, Cars), 4 (1 362, 1 760, 2 143 and 2 311 nm), 5 (1 227, 1 310, 1 382, 1 450 nm) were extracted by three characteristic wavelength extraction algorithms with different principles, respectively 1 728 nm) and 10 (1 232, 1 233, 1 257, 1 279, 1 313, 1 688, 1 703, 1 705, 2 302 and 2 323 nm). The characteristic wavelengths extracted were used as input variables of the corn seed ear rot identification model. The disease status of samples was represented by 0-1 (infected normal) as the output true value to establish the support vector machine (SVM) model. The model parameters were optimized by the grid search method and the 10-fold cross-validation method. The results show that the modeling accuracy of the training and test set in three discriminant models, CA-SVM, SPA-SVM and CARS-SVM, is above 90%. The research results in this paper provide a model basis for the maize seed disease diagnosis device. The method of selecting characteristic wavelengths for the optimal region can also provide a reference for establishing other seed disease discrimination models.

Keywords Near-infrared spectrum; Corn seeds; Ear rot; Characteristic wavelength

(Received May 8, 2021; accepted Jul. 13, 2021)

* Corresponding author