王小龙, 邓继忠, 黄华盛, 等. 基于高光谱数据的棉田虫害鉴别研究 [J]. 华南农业大学学报, 2019, 40(3): 97-103.
WANG Xiaolong, DENG Jizhong, HUANG Huasheng, et al. Identification of pests in cotton field based on hyperspectral data[J]. Journal of South China Agricultural University, 2019, 40(3): 97-103.

基于高光谱数据的棉田虫害鉴别研究

王小龙[1,2]，邓继忠[1,2]，黄华盛[1,2]，邓宇森[1,2]，蒋统统[1,2]，钟兆基[1,2]，张亚莉[1,2]，文　晟[1,3]
(1 国家精准农业航空施药技术国际联合研究中心，广东 广州 510642； 2 华南农业大学 工程学院，广东 广州 510642；
3 华南农业大学 工程基础教学与训练中心，广东 广州 510642)

摘要:【目的】快速、准确、无损伤地鉴别棉花虫害类别，以便针对性制定植保施药方案。【方法】对棉花叶片高光谱数据进行采集和分析。采用波段范围为 350~2 500 nm 的 FieldSpec®3 便携式光谱分析仪，分别获取受蚜虫和红蜘蛛危害的棉花叶片以及正常棉花叶片的高光谱数据。采用 *K*−近邻和 SVM 算法区分受红蜘蛛和蚜虫侵害的叶片以及正常叶片。为进一步优化虫害识别模型、提高识别精度，利用主成分分析方法 (PCA) 进行特征降维，并利用网格搜索法进行参数寻优。【结果】使用 *K*−近邻算法和 SVM 算法构建了虫害识别模型，2 种模型的识别率分别为 86.08% 和 89.29%；引入 PCA 进行特征降维并使用网格搜索进行参数寻优后，可以提高虫害识别率，*K*−近邻算法和 SVM 算法的识别精度分别达到 88.24% 和 92.16%。【结论】利用高光谱数据可以区分受蚜虫和红蜘蛛侵害以及正常的棉花叶片；结合 PCA 降维和网格搜索法，能够提高识别率且不需要获得具体的特征波段；对于受蚜虫和红蜘蛛侵害以及正常的叶片识别，基于径向基核函数的 SVM 算法优于 *K*−近邻算法。

关键词: 棉花虫害；*K*−近邻；支持向量机；高光谱数据；无损鉴别
中图分类号: S435.62；TP391.4　　**文献标志码:** A　　**文章编号:** 1001-411X(2019)03-0097-07

Identification of pests in cotton field based on hyperspectral data

WANG Xiaolong[1,2], DENG Jizhong[1,2], HUANG Huasheng[1,2], DENG Yusen[1,2], JIANG Tongtong[1,2],
ZHONG Zhaoji[1,2], ZHANG Yali[1,2], WEN Cheng[1,3]
(1 National Center for International Collaboration Research on Precision Agricultural Aviation Pesticides Spraying Technology, Guangzhou 510642, China; 2 College of Engineering, South China Agricultural University, Guangzhou 510642, China; 3 Engineering Fundamental Teaching and Training Center, South China Agricultural University, Guangzhou 510642, China)

Abstract:【Objective】To identify cotton pests quickly and accurately without destruction，and formulate pertinently a plant protection spraying plan.【Method】Hyperspectral data of cotton leaves were collected and analyzed. FieldSpec®3 portable spectrum analyzer with a wavelength range of 350−2 500 nm was used to obtain hyperspectral data of cotton leaves including normal leaves and leaves infected by aphids and red spiders. *K*-nearest neighbor and SVM algorithm were used to distinguish above leaves. In order to further optimize pest identification of the model and improve the recognition accuracy, the principal component analysis method (PCA) was used for feature dimension reduction, and the grid search method was used for parameter optimization.【Result】The models of pest identification were constructed by *K*-nearest neighbor algorithm and

收稿日期:2018–07–25　　**网络首发时间:**2019–04–16 09:12:00
网络首发地址:http://kns.cnki.net/kcms/detail/44.1110.S.20190412.1740.020.html
作者简介:王小龙 (1995—)，男，硕士研究生，E-mail: 931457221@qq.com；通信作者：邓继忠 (1963—)，男，教授，博士，E-mail: jz-deng@scau.edu.cn
基金项目:广东省科技计划项目 (2017A020208046，2017B010117010)；国家重点研发计划项目 (2016YFD0200700)；国家发展和改革委 2014 年北斗卫星导航产业重大应用示范发展专项项目 (20142564)；广东省教育厅重点平台及科研项目 (2015KGJHZ007)；广州市科技计划项目 (201707010047)

SVM algorithm，and recognition rates of two models were 86.08% and 89.29% respectively. Recognition rate increased after introducing PCA for feature dimension reduction and using grid search for parameter optimization. The recognition accuracies of *K*-nearest neighbor algorithm and SVM algorithm reached 88.24% and 92.16% respectively.【Conclusion】Hyperspectral data can be used to distinguish aphid or red spider-infected leaves and normal cotton leaves. Using PCA dimensionality reduction and grid search method, the recognition rate can increase without obtaining specific characteristic bands. For identifying aphid- or red spider-infected leaves and normal leaves, SVM algorithm based on radial basis kernel function is better than *K*-nearest neighbor algorithm.

Key words: cotton pest; *K*−nearest neighbor; support vector machine; hyperspectral data; non-destructive identification

棉花 *Gossypium* spp. 是我国重要的经济作物，在国民经济中占有十分重要的地位[1]。棉花的整个生长过程中会遭受到多种病虫的危害，常年因病虫危害的棉花损失达 15%~20%，严重发生年份损失超过 50%，甚至绝收，通过病虫防控每年可挽回棉花损失 90 万 t 以上[2]。蚜虫 *Aphis gossypii* 和红蜘蛛 *Tetranychus cinnbarinus* 是棉花的主要害虫，其个体小、繁殖快、适应力强，如若不及时防治会严重影响棉花的产量和质量。蚜虫俗称腻虫或蜜虫，是最具破坏性的植食性害虫之一，靠吸食植物幼嫩部位的汁液生存，蚜虫危害后常使叶片、茎尖等部位发生皱缩或卷曲[3]。棉花红蜘蛛，又叫叶螨，是我国各棉区普遍发生危害较重的一类害虫。棉花红蜘蛛主要在棉叶的背面吸食营养汁液，为害初期叶片正面出现较多白点，几天后叶柄处变红，重则落叶垮秆，状如火烧，造成大面积减产或绝收[4]。因此，防治棉花红蜘蛛和蚜虫对棉花生产很关键。

传统的虫害防治是对整个农田按照统一剂量进行均匀的农药喷洒。但是，过度施用农药也造成了环境污染。精准喷施可以有效解决这一问题，而有效识别虫害发生位置能为精准喷施提供决策信息。由于蚜虫和红蜘蛛体积小，且多集中在叶片背部，为数据采集带来困难。光谱技术能够从叶片正面检测出虫害发生所造成的光谱差异[5]，从而有效地解决数据采集的问题，为后续的虫害识别和精准施药奠定基础。光谱分析技术具有速度快、无污染以及不破坏样品等优点，使得该技术成为分析农产品特性的重要手段[5]。国内外学者已经在这个方向开展了多项研究。相关的试验数据表明，光谱技术在多个农情监测的研究上都具有可行性[5-24]。白敬等[5] 利用逐步判别分析法选取了 710、755、950 和 595 nm 的 4 个特征波长并建立判别模型，最终识别率达 98.89%。孙俊等[6] 利用基于径向基内核的支持向量机 (SVM)和十折交叉验证方法建立桑叶农残检测模型，最终识别率 97.78%。黄双萍等[7] 采用卡方–支持向量机分类算法模型建立稻瘟病害程度等级分级模型，最终识别率达 94.72%。Piron 等[8] 利用二次判别的分析方法选取 3 个特征波长作为识别胡萝卜和杂草的特征点，对胡萝卜和杂草的识别率均达 72%。利用光谱技术对棉花虫害的分类研究鲜见报道。本文基于棉花的高光谱数据，采用机器学习的方法对棉花虫害进行自动识别，在预处理阶段使用平均处理消除数据采集的随机噪声，分别以 *K*−近邻算法和基于径向基核函数的 SVM 算法进行分类识别，使用主成分分析和逐步判别分析优化特征向量，采用网格搜索策略优化模型参数，为棉花红蜘蛛和蚜虫虫害的识别提供决策支持。

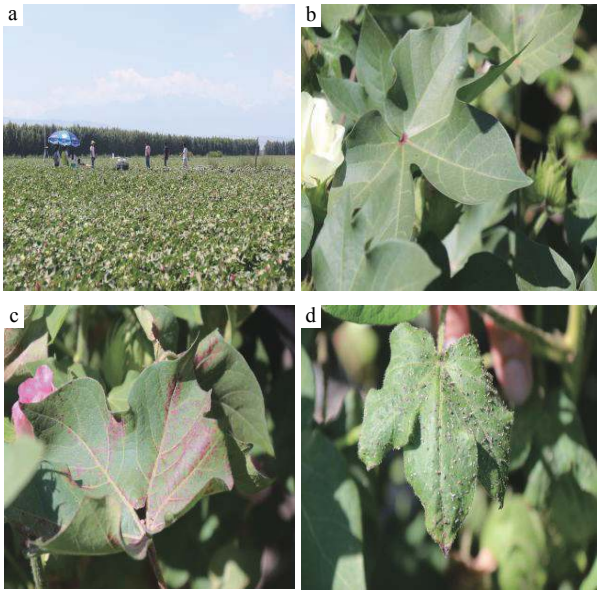
1 材料与方法

1.1 试验样本的制备与采集

红蜘蛛的危害一般从 6 月开始一直持续到 9 月，而蚜虫的爆发时期是 7—8 月。本次试验样本于 2017 年 7 月 19 日 10:00—14:00 从新疆自治区石河子市石总场试验棉田采集，采集的数据涉及红蜘蛛和蚜虫危害的棉花叶片以及正常棉花叶片 (图 1)。其中，红蜘蛛样本数据于石河子市石总场三分场四连试验棉田采集，GPS 坐标为北纬 44°23′、东经 85°55′；蚜虫数据于石河子市石总场六分场二连试验棉田采集，GPS 坐标为北纬 44°36′、东经 85°58′。采集样本当天晴朗无云，田间作业温度达 40 ℃，气候环境利于样本数据的获取。

1.2 高光谱数据的采集和预处理

1.2.1 *数据采集* 试验采用的是美国 ASD 公司的 FieldSpec®3 便携式光谱分析仪，光源采用仪器自带的卤素灯。光谱仪的光谱范围为 350~2 500 nm，采样间隔为 1 nm。本文使用光谱技术采集数据，目的



a：采集地；b：正常叶片；c：红蜘蛛叶片；d：蚜虫叶片

a: Collecting site; b: Normal leaves; c: Leaves with red spiders; d: Leaves with aphids

图 1　数据采集地和棉花叶片

Fig. 1　Data collection sites and cotton leaves

是使用非接触、无损伤的方式进行虫害的胁迫监测；同时，使用机器学习方法来分析数据，目的是利用其黑匣子的优点，在不需要机理分析的前提下就能获得较为满意的结果。

采集高光谱图像前的准备工作：提前 30 min 打开光源进行预热，紧接着通过扫描采集反射率为 100% 的标准白板进行黑白标定，光谱仪设定 1 次采集 5 组数据，即数据样本包含 5 组数据。为了保证测量的是叶片的表征区域，我们测量时让叶片水平朝上，使光谱仪的探头对准叶片中间，以此来保证叶片在光谱仪的覆盖范围之内，同时让仪器与待测叶片保持固定的距离。受红蜘蛛危害的棉田采集样本 161 个，其中，受红蜘蛛虫害的样本 111 个、正常样本 50 个；受蚜虫危害的棉田采集样本 59 个，其中，受蚜虫危害的样本 29 个、正常样本 30 个。

1.2.2　*数据预处理*　由于每个样本包含 5 条光谱曲线，本文对这 5 组数据进行平均处理，从而减少原始光谱数据的内在误差。计算光谱反射率：

$$R_{\text{goal}} = \frac{\text{Rad}_{\text{goal}}}{\text{Rad}_{\text{board}}} \times R_{\text{board}}, \tag{1}$$

式中，R_{goal} 表示通过白板反射率求得的目标光谱反射率，Rad_{goal} 表示通过光谱仪测得的目标物光强值，$\text{Rad}_{\text{board}}$ 表示通过光谱仪测得的白板光强值，R_{board} 表示已知的白板反射率。

经过预处理操作，本次研究的数据集共包含 220 组数据。包含红蜘蛛虫害的棉花数据 111 组，蚜虫虫害的棉花数据 29 组，正常棉花数据 80 组。

1.3　方法

1.3.1　*K–近邻判别模型*　*K*–近邻算法是采用测量不同特征向量之间距离的方法进行分类，其优点是精度高、对异常值不敏感，并且对输入数据无假定；缺点是计算和空间复杂度高[25]。具体工作原理[25]：已知一个样本数据的集合，并且知道样本中每一数据与所属分类的对应关系；在输入新的没有标签的数据后，将新数据的每个特征与已知样本的数据对应特征值进行比较，然后提取已知样本中特征最相似的数据分类标签。一般只选择 K 个最相似的数据，最后选取 K 个最相似数据中出现次数最多的类别，以此作为新样本的分类。本文在进行特征向量对比时，选用的是欧式距离 (d) 公式，具体公式如下：

$$d = \sqrt{\sum_{i,j=1}^{n} (x_i - x_j)^2}, \tag{2}$$

式中，x_i代表待测样本，x_j代表已知样本，n 代表样本数。

1.3.2　*SVM 判别模型*　支持向量机 (SVM) 是一种主要针对二分类任务的模型。简单来说就是找到一个平面 (超平面) 将优化目标最大化分类间隔，此处的间隔是指样本里超平面的距离，而最靠近超平面的样本就称作支持向量[26]。它在解决小样本，非线性及高维识别问题上有独特的优势。SVM 已经被大量用于高光谱数据分析，而且比一些其他的传统机器学习算法呈现出更好的分类性[6]。

原始的支持向量机是线性可分的，基本型如下式：

$$\begin{aligned} &\min \frac{1}{2}\|w\|^2, \\ &\text{s.t.} y_i\left(w^T x_i + b\right) \geqslant 1,\ i = 1,2,...,m, \end{aligned} \tag{3}$$

式中，w 为法向量，决定了超平面的方向；b 为位移项，决定了超平面与原点的距离，s.t.为约束条件。

本文处理的数据是非线性可分的，需要引入核函数将数据映射到高维，利用非线性支持向量机来进行建模处理。非线性支持向量机将原始的支持向量机基本型转换为：

$$\begin{aligned} &\max \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K < x_i, x_j >, \\ &\text{s.t.} \sum_{j=1}^{m} \alpha_i y_i = 0, \\ &0 \leqslant \alpha_i \leqslant C,\ i = 1,2,...,m, \end{aligned} \tag{4}$$

式中，α为拉格朗日因子，x_i和y_i为样本点，$K < x_i, x_j >$为核函数，C为惩罚因子，s.t.为约束条件。

求解后得到：

$$f_x = \sum_{i=0}^{m} \alpha_i y_i K < x_i, x_j > + b, \quad (5)$$

式中，f_x 为分类结果。

SVM 的难点在于模型参数核函数参数 (γ) 和惩罚因子 (C) 的选择。本次 SVM 算法模型的构建采用的是径向基核函数 (RBF)，其公式如下：

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right), \quad (6)$$

式中，x代表所需要输入的样本，y为高斯函数的中心，σ为函数值跌至 0 的速度参数。

径向基核函数是一个采用向量作为自变量的函数，能够基于向量的距离运算输出一个标量，对复杂数据有很好的适用性和稳定性[25]。此次初步构建 SVM 模型时，根据经验设置了参数，并未考虑参数的寻优问题。

1.3.3 *数据降维* 基于 K−近邻和 SVM 的棉花虫害分类算法模型构建的分类结果正确率并不理想。由于数据中存在的噪声和冗余信息会对分类器造成干扰，因此采用数据降维的方法消除这个影响因素。常用的数据降维方法包括线性判别分析 (LDA)、二次判别分析、主成成分分析 (PCA)、逐步判别分析等。本文采用了 PCA 和逐步判别分析进行数据降维，相关方法的降维效果统一采用 K−近邻分类器进行验证。

PCA 目的是进行数据降维，在不丢失主要光谱信息的前提下，选择为数较少的新变量代替原来较多的变量，获得的主成分间是正交的，排除了大量数据信息的重叠性[20]。使用 PCA 获取的不是具体的特征波段，而是降维后的特征向量。PCA 的主要实现过程如下：

1) 获取经过数据预处理后的样本矩阵 $\boldsymbol{X} = [x_1, x_2, ..., x_n]$，$n$ 代表样本数，每一列都代表一个样本；

2) 减去平均值，将矩阵 $\boldsymbol{X}$ 的每一列进行零均值化，再减去这一列的平均值得到矩阵 $\boldsymbol{C_X}$；

3) 计算 $\boldsymbol{C_X}$ 的协方差矩阵；

4) 计算协方差矩阵的特征向量和特征值，将特征值降序排列，选取特征向量；

5) 将样本投影到选取的特征向量上，获取新的数据集作为虫害识别模型的输入样本。

逐步判别分析是将每个变量逐一输入判别函数，对输入的单个变量进行检验，不满足条件的将被剔除，最终保留下来的都是最显著的变量[5]。本文采用的高光谱数据经过逐步判别分析后能够剔除不显著的波段，从而获取显著性强的波段。

1.3.4 *参数寻优* 采用网格搜索策略进行分类器的超参优化，从而提高分类器的预测精度。对于 K−近邻算法，所优化的超参包括近邻数 K；对于 SVM 算法，所优化的超参包括核函数参数 (γ) 和惩罚因子 (C)。具体流程如图 2 所示。

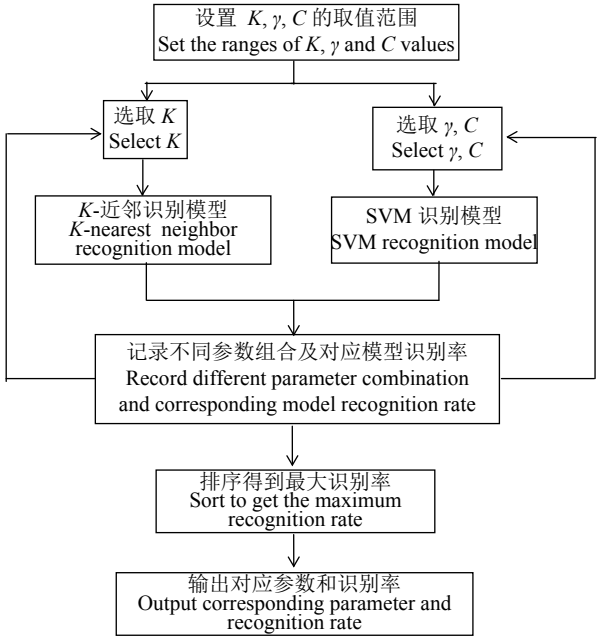


图 2 参数寻优流程图

Fig. 2 The flow chart of parameter optimization

2 结果与分析

2.1 棉花叶片光谱数据信息预处理结果

图 3 为红蜘蛛叶片、蚜虫叶片和正常棉花叶片的光谱数据曲线。由图 3 可以看出，棉花叶片光谱曲线的变化趋势是符合棉花的生理特性变化的：1) 虫害胁迫降低了棉花叶片的叶绿素含量和相对含水量，因此对应的光谱反射率在可见光范围 (350~

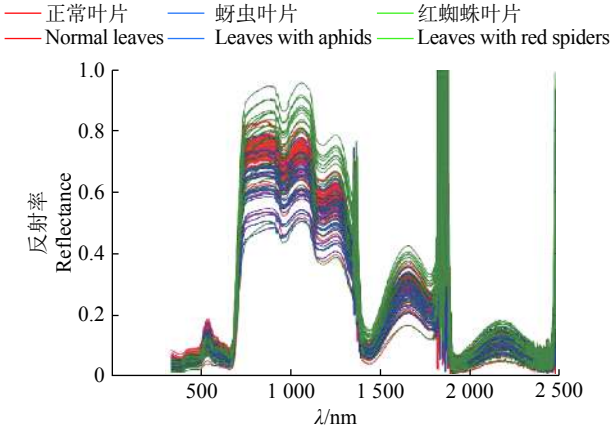


图 3 棉花叶片高光谱曲线

Fig. 3 Hyperspectral curves of cotton leaves

700 nm) 和近红外范围 (700~1 300 nm) 内普遍低于正常叶片[27]。因此这 2 个波段可作为判断棉花受到红蜘蛛危害的识别波段[28]。2) 受虫害胁迫的叶片较正常叶片水分更少[29], 而 1 400~2 500 nm 的波段受水分影响明显[30-31], 因此造成了光谱反射率在这一波段范围的差异。

2.2　2 种判别模型的分类识别结果

2.2.1　*K*−近邻判别模型　利用 *K*−近邻做了三分类的判别: 将有蚜虫虫害的样本标记为 0, 有红蜘蛛虫害的样本标记为 1, 正常棉花样本标记为−1。对于采集到的数据, 随机选择 70% 作为训练集, 剩下的 30% 作为测试集。训练集选取总的样本数为 169 组, 其中, 有蚜虫虫害的样本 19 组、有红蜘蛛虫害的样本 90 组、正常棉花样本 60 组, 测试集样本 51 组 (表 1)。

表 1　样本的选取

Table 1　Sample selection

项目 Item	虫害 Pest	标记 Label	样本数 Sample number
训练集 Training set	正常 Normal	−1	60
	蚜虫 Aphid	0	19
	红蜘蛛 Red spider	1	90
测试集 Test set	正常 Normal	−1	20
	蚜虫 Aphid	0	10
	红蜘蛛 Red spider	1	21

利用 *K*−近邻算法对上述基于特征波长的高光谱数据样本进行分类。经过多次调试发现在 *K*−近邻模型参数近邻数 *K*=12 处得到一个较好结果, 识别率为 86.08%。

2.2.2　SVM 判别模型　原始的 SVM 是一个二分类模型, 本次研究涉及多分类的问题, 因此采取了 2 步二分类的方法。第 1 步, 先将样本判别为正常棉花和虫害棉花; 第 2 步, 对棉花虫害进行分类, 将其分为红蜘蛛危害的棉花和蚜虫危害的棉花。为了与 *K*−近邻算法模型对比, SVM 模型的数据集与 *K*−近邻是一致的。

经过多次尝试, 发现 SVM 模型在核函数参数 (γ) 为 18、惩罚因子 (C) 为 0.000 1 处得到最优的分类效果, 识别率为 89.29%。与 *K*−近邻算法模型 (86.08%) 相比, SVM 算法模型的识别效果更好。

2.3　数据降维结果

2.3.1　主成分分析 (PCA)　PCA 是机器学习中一种常用的数据处理方法, 也是一种降维技术。本次试验获取的数据经过预处理之后维度还是很高, 包含 2 126 个特征, 其中可能包含了很多无效和冗余的波段信息, 运用 PCA 方法可以将这些特征进行降维, 去除无关信息, 找出有效的波段信息。获取的数据保留了 90% 的方差, 最终将特征维度降到了 6 维。降维后的数据输入 *K*−近邻模型, 得到识别率为 88.24%。

2.3.2　逐步判别分析　以各样本的反射率为自变量, 所属类别为因变量, 采用 SPSS 分析软件进行分析, 通过对反射率的筛选, 提取出特征波长。由于样本波长数量过多, 采取每隔 5 nm 选取一个波长的方法进行简化处理, 将处理过的数据输入到 SPSS 软件中。最终提取出 10 个特征波长, 分别为 755、805、920、930、1 125、1 340、1 485、1 889、1 894 和 2 004 nm。对原输入分类模型的数据进行二次处理, 只保留包含上述 10 个波长的数据, 输入 *K*−近邻算法模型进行判别, 最终得到的识别率为 78.43%。

上述结果表明, 利用主成分分析 (PCA) 方法进行数据降维能够有效地提高分类精度, 且 PCA 方法的效果明显优于逐步判别分析, 为此, 本文采用 PCA 方法进行数据降维。

2.4　参数寻优

使用 *K*−近邻算法和 SVM 算法时, 对分类器的参数进行了多次尝试并且选择最优值, 但这种人工调节参数的方式难以获得全局最优值。因此, 采取了网格搜索 (Grid search) 分别对 *K*−近邻算法参数 (K) 和 SVM 算法参数 (γ 和 C) 进行参数的寻优。

由表 2 可以看出, 通过参数优化和数据降维, 2 种分类器的分类精度都有了明显的提高; *K*−近邻和 SVM 算法对蚜虫虫害识别的效果接近, 但两者效果都不理想; 在红蜘蛛虫害和正常棉叶识别上 SVM 算法明显优于 *K*−近邻。从整体的衡量精度来看, SVM 算法性能优于 *K*−近邻算法。综合 PCA 特征降维和网格搜索, 最终在 *K*−近邻算法模型中得到的识别率为 88.24%, SVM 算法模型中得到的识别率为 92.16% (表 2)。

本文使用 *K*−近邻和 SVM 算法对棉花虫害的光谱数据进行分类识别, 采用数据降维和参数寻优的方法优化分类结果。其中 SVM 模型对测试数据的准确率达到了 92.16%, 说明本文采用的方法能够对棉花虫害的数据进行有效识别, 显示了良好的泛化能力。

表 2 2 种算法模型参数寻优后的识别率和混合矩阵精度

Table 2 Hybrid matrix precisions and identification rates of two models after parameter optimization

模型 Model	虫害 Pest	识别率/% Identification rate			混合矩阵精度/% Hybrid matrix precision		
		使用者精度 User's accuracy	生产者精度 Procuder's accuracy	总体精度 Overall accuracy	正常 Normal	蚜虫 Aphid	红蜘蛛 Red spider
网格+PCA+*K*-近邻 Grid+PCA+*K*-nearest neighbor	正常 Normal	95.00	82.61	88.24	95.00	5.00	0.00
	蚜虫 Aphid	60.00	75.00		40.00	60.00	0.00
	红蜘蛛 Red spider	95.23	100.00		0.00	4.76	95.23
网格+PCA+SVM Grid+PCA+SVM	正常 Normal	100.00	87.00	92.16	100.00	0.00	0.00
	蚜虫 Aphid	80.00	100.00		40.00	60.00	0.00
	红蜘蛛 Red spider	85.71	95.23		0.00	0.00	100.00

3 结论

本文采用便携式光谱分析仪，采集了正常棉花、红蜘蛛棉花以及蚜虫棉花的叶片高光谱数据。使用数据降维消除数据间的冗余信息，去除噪声的影响。基于机器学习技术，对棉花叶片虫害进行自动鉴别。

针对高光谱数据中相邻谱带间存在较强的相关性，使用数据降维技术进行噪声滤波，以进一步提高分类精度。本文选取了逐步判别分析和 PCA 方法进行数据降维，并在 *K*−近邻算法上寻求降维效果的对比。逐步判别分析保留的都是最显著的变量，不满足条件的数据被剔除。逐步判别分析虽然能够在很大程度上简化数据，但是数据间的相关性会大大降低，而 PCA 方法在简化数据的基础上会保留数据的相关性。本研究结果也表明，PCA 方法优于逐步判别分析。相比直接进行分类，采用 PCA 方法之后，*K*−近邻和 SVM 分类器的分类精度分别提高了 2.16% 和 2.87%，说明采用 PCA 方法进行数据降维可以有效提高分类的准确率。

经过数据降维之后，采用 *K*−近邻和 SVM 方法对棉花虫害进行分类识别，同时采用网格搜索方法进行参数寻优。结果表明，这 2 种方法在测试数据上的准确率分别达到 88.24% 和 92.16%。SVM 的分类效果优于 *K*−近邻，这可能是因为 SVM 分类器泛化能力好，对未知因素体现了良好的预测结果。从分类结果来看，*K*−近邻和 SVM 分类器对正常棉花和红蜘蛛棉花的分类准确率都很高，均超过了 90%；对蚜虫的识别精度较低，均为 60%。这可能与本试验中蚜虫数据不足有直接关系。另外，数据采集过程中的操作也会对后续的分析过程造成影响。在数据采集中，有些叶片获得的曝光率过大或过小，以及叶脉的存在，都会对采集到的棉花叶片光谱曲线造成干扰，进而影响后续的分类识别。

本文利用便携式光谱分析仪采集棉花叶片高光谱数据，经过数据预处理之后，采用逐步判别和 PCA 方法进行特征降维，最后利用 *K*−近邻算法和 SVM 方法进行识别，同时使用网格搜索进行参数寻优。试验结果表明，基于 PCA 和 SVM 方法的分类效果最好，对测试集的分类精度为 92.16%，说明本文采用的机器学习算法对棉花虫害能够有效鉴别。今后，还需要采集大量数据来进行模型的优化和验证，通过深度学习算法进行特征的自动提取和分析，寻求分类性能的提升。

参考文献：

[1] 雒珺瑜, 张帅, 任相亮, 等. 近十年我国棉花虫害研究进展 [J]. 棉花学报, 2017, 29(增刊): 100-112.

[2] 崔金杰, 陈海燕, 赵新华, 等. 棉花害虫综合防治研究历程与展望 [J]. 棉花学报, 2007, 19(5): 385-390.

[3] 张海娜, 钱玉源, 刘祎, 等. 蚜虫防治研究及在棉花上的应用 [J]. 农学学报, 2015, 5(8): 36-39.

[4] 冯国民. 棉花红蜘蛛的发生与防治 [J]. 北京农业, 2010(25): 41.

[5] 白敬, 徐友, 魏新华, 等. 基于光谱特性分析的冬油菜苗期田间杂草识别 [J]. 农业工程学报, 2013, 20(29): 128-133.

[6] 孙俊, 张梅霞, 毛罕平, 等. 基于高光谱图像桑叶农药残留种类鉴别研究 [J]. 农业机械学报, 2015, 46(6): 251-255.

[7] 黄双萍, 齐龙, 马旭, 等. 基于高光谱成像的水稻稻瘟病害程度分级方法 [J]. 农业工程学报, 2015, 31(1): 212-217.

[8] PIRON A, LEEMANS V, KLEYNEN O, et al. Selection of the most efficient wavelength bands for discriminating weeds from crop[J]. Comput Electron Agric, 2008, 62(2): 689-699.

[9] 田有文, 李天来, 张琳, 等. 高光谱图像技术诊断温室黄瓜病害的方法 [J]. 农业工程学报, 2010, 26(5): 202-205.
[10] 刘波, 方俊永, 刘学, 等. 基于成像光谱技术的作物杂草识别研究 [J]. 光谱学与光谱分析, 2010, 30(7): 1830-1833.
[11] 邓巍, 张录达, 何雄奎, 等. 基于支持向量机的玉米苗期田间杂草光谱识别 [J]. 光谱学与光谱分析, 2009, 29(7): 1906-1910.
[12] 陈树人, 贾移新, 毛罕平, 等. 基于光谱分析技术的作物中杂草识别研究 [J]. 光谱学与光谱分析, 2009, 29(2): 463-466.
[13] 谢传奇, 王佳悦, 冯雷, 等. 应用高光谱图像光谱和纹理特征的番茄早疫病早期检测研究 [J]. 光谱学与光谱分析, 2013, 33(6): 1603-1607.
[14] 薛龙, 黎静, 刘木华, 等. 基于高光谱图像技术的水果表面农药残留检测试验研究 [J]. 光学学报, 2008, 28(12): 2277-2280.
[15] 朱文静, 毛罕平, 周莹, 等. 基于高光谱图像技术的番茄叶片氮素营养诊断 [J]. 江苏大学学报(自然科学版), 2014, 35(4): 290-294.
[16] 洪添胜, 乔军, NGADI M O, et al. 基于高光谱技术的雪花梨品质无损检测 [J]. 农业工程学报, 2007, 23(2): 151-154.
[17] 刘雪梅, 章海亮. 基于 DPLS 和 LS-SVM 的梨品种近红外光谱识别 [J]. 农业机械学报, 2012, 43(9): 160-164.
[18] 吴迪, 黄凌霞, 何勇, 等. 作物和杂草叶片的可见–近红外反射光谱特性 [J]. 光学学报, 2008, 28(8): 1618-1622.
[19] 王立国, 赵亮, 刘丹凤, 等. SVM 在高光谱图像处理中的应用综述 [J]. 哈尔滨工程大学学报, 2018, 39(6): 973-980.
[20] 袁建清, 苏中滨, 贾银江, 等. 基于高光谱成像的寒地水稻叶瘟病与缺氮识别 [J]. 农业工程学报, 2016, 32(13): 155-158.
[21] JIMENEZL O, LANDGREBE D A. Supervised classification in high dimensional space: Gemetrical, statistical, and a symptotical properties of nultivariate data[J]. IEEE Trans Syst Man Cybean C: Appl Rev, 1998, 28(1): 39-54.
[22] 岳学军, 全东平, 洪添胜, 等. 柑橘叶片叶绿素含量高光谱无损检测模型 [J]. 农业工程学报, 2015, 31(1): 294-300.
[23] 孙俊, 金夏明, 毛罕平, 等. 高光谱图像技术在掺假大米检测中的应用 [J]. 农业工程学报, 2014, 30(21): 301-305.
[24] 祝志慧, 刘婷, 马美湖. 基于高光谱信息融合和相关向量机的种蛋无损检测 [J]. 农业工程学报, 2015, 31(15): 285-290.
[25] HARRINGTON P. 机器学习实战 [M]. 李锐，李鹏，曲亚东，等译. 北京: 人民邮电出版社, 2013: 15-110.
[26] 李航. 统计学习方法 [M]. 北京:清华大学出版社，2017: 95-133.
[27] CHEN T, ZENG R, ZHANG L. Detection of stress in cotton (*Gossypium hirsutum* L.) caused by aphids using leaf level hyperspectral measurements[J]. Sensors, 2018, 18(9): 2798.
[28] 牛鲁燕, 郑纪业, 张晓艳, 等. 基于成像高光谱的苹果叶片叶绿素含量估测模型研究 [J]. 江西农业学报, 2018, 30(2): 100-104.
[29] RAVINDER R, GIRIDHAR M. Spectral reflectance from the tomato crop canopy under controlled condition by using spectroradiometer[C]// LANE C, BAEAR S. NCWES. Hyderabad: BS Publications, 2017: 392-397.
[30] 梁守真, 施平, 马万栋, 等. 植被叶片光谱及红边特征与叶片生化组分关系的分析 [J]. 中国生态农业学报, 2010, 18(4): 804-809.
[31] 孙林, 程丽娟. 植被叶片生化组分的光谱响应特征分析 [J]. 光谱学与光谱分析, 2010, 30(11): 3031-3035.

【责任编辑　周志红】