

邓小玲, 曾国亮, 朱梓豪, 等. 基于无人机高光谱遥感的柑橘患病植株分类与特征波段提取 [J]. 华南农业大学学报, 2020, 41(6): 100-108.
DENG Xiaoling, ZENG Guoliang, ZHU Zihao, et al. Classification and feature band extraction of diseased citrus plants based on UAV hyperspectral remote sensing[J]. Journal of South China Agricultural University, 2020, 41(6): 100-108.

基于无人机高光谱遥感的柑橘患病植株 分类与特征波段提取

邓小玲^{1,2}, 曾国亮^{1,2}, 朱梓豪², 黄梓效², 杨佳诚^{1,2},
童泽京², 殷献博^{1,2}, 王天伟^{1,2}, 兰玉彬^{1,2}
(1 华南农业大学 电子工程学院/人工智能学院, 广东 广州 510642;
2 国家精准农业航空施药技术国际联合中心, 广东 广州 510642)

摘要:【目的】结合传统与现代农业病虫害监测的优缺点,探索通过无人机高光谱遥感技术检测出患病的柑橘植株、通过人工田间调查方式判断其患病种类及患病程度的病虫害监测方法。【方法】使用无人机获取原始高光谱图像,经过光谱预处理和特征工程后,采用连续投影算法提取对柑橘患病植株分类贡献值最大的特征波长组合,基于全波段使用 BP 神经网络和 XgBoost 算法、基于特征波段使用逻辑回归和支持向量机算法,建立分类模型。【结果】基于全波段的 BP 神经网络和 XgBoost 算法的 ROC 曲线下面积 (Area under curve, AUC) 分别为 0.883 0 和 0.912 0, 分类准确率均超过 95%; 提取出 698 和 762 nm 的特征波长组合,基于特征波长使用逻辑回归和支持向量机算法建立的分类模型召回率分别达到了 93.00% 和 96.00%。【结论】基于特征波长建模在患病样本分类中表现出很高的准确率,证明了特征波长组合的有效性。本研究结果可为柑橘种植园的病虫害监测提供一定的数据和理论支撑。

关键词: 柑橘病虫害; 无人机; 高光谱遥感; XgBoost; 连续投影算法
中图分类号: TP79; S436.66 **文献标志码:** A **文章编号:** 1001-411X(2020)06-0100-09

Classification and feature band extraction of diseased citrus plants based on UAV hyperspectral remote sensing

DENG Xiaoling^{1,2}, ZENG Guoliang^{1,2}, ZHU Zihao², HUANG Zixiao², YANG Jiacheng^{1,2},
TONG Zejing², YIN Xianbo^{1,2}, WANG Tianwei^{1,2}, LAN Yubin^{1,2}
(1 College of Electronic Engineering/College of Artificial Intelligence, South China Agricultural University, Guangzhou 510642, China; 2 National Center for International Collaboration Research on Precision Agricultural Aviation Pesticides Spraying Technology, Guangzhou 510642, China)

Abstract: 【Objective】Combined with the advantages and disadvantages of traditional and modern agricultural pest monitoring, the method of monitoring pest and disease were discussed, which detected the diseased citrus plants by UAV hyperspectral remote sensing technology and judged the disease species and disease degree by artificial field investigation. 【Method】The original hyperspectral images were obtained by UAV. After spectral preprocessing and feature engineering, continuous projection algorithm was used to extract the feature wavelength combination which contributed the most to the classification of citrus diseased plants.

收稿日期:2020-06-19 网络首发时间:2020-07-30 09:34:24
网络首发地址:<https://kns.cnki.net/kcms/detail/44.1110.S.20200729.1743.002.html>
作者简介: 邓小玲 (1978—), 女, 副教授, 博士, E-mail: dengxl@scau.edu.cn; 通信作者: 兰玉彬 (1961—), 男, 教授, 博士, E-mail: ylan@scau.edu.cn
基金项目: 国家自然科学基金 (61675003); 广东省重点领域研发计划 (2019B020214003); 广东高校重点领域 (人工智能) 专项 (2019KZDZX1012); 广东大学生科技创新培育专项 (pdjh2019b007)

Finally, the BP neural network and XgBoost algorithm were used based on the full band, and the logistic regression and support vector machine algorithm were used to establish the classification model based on the characteristic band. 【Result】 The AUC scores of BP neural network and XgBoost were 0.8830 and 0.9120 respectively, and the accuracy rates of both methods were over 95%. The feature wavelength combination of 698 and 762 nm was extracted. Based on this characteristic band, the recall rates of logistic regression and support vector machine algorithm were 93.00% and 96.00% respectively. 【Conclusion】 The model based on characteristic band shows high accuracy in the classification of disease samples, which proves the effectiveness of characteristic wavelength combination. This result can provide some data and theoretical support for monitoring diseases and pests in citrus plantations.

Key words: citrus disease and pest; UAV; hyperspectral remote sensing; XgBoost; continuous projection algorithm

农作物病虫害是制约农业生产发展的主要因素之一,对农作物产量与品质造成极大损失。目前柑橘 *Citrus reticulata* Blanco 产业的病虫害发展趋势逐渐恶化,其中,柑橘黄龙病 (Citrus Huanglongbing, HLB) 最为严重,患黄龙病的植株会导致果实的产量和品质下降,最后导致树体死亡。美国、中国、巴西等柑橘生产大国都受到柑橘黄龙病的严重威胁^[1]。到目前为止,柑橘黄龙病仍然没有有效的治疗方法,一旦发现,只能连根拔除,以免其他健康植株被传染。因此,如何快速、精准地找到已患黄龙病的柑橘植株成为当今柑橘病虫害防治的首要任务^[2]。传统的柑橘园区病害监测与防治主要是依靠农户田间走访调查取样,不仅耗时、耗力、效率低下,而且因人的主观性而容易遗漏。中国未来的农业产业趋向于高度区域化、智能化、一体化及精准化,依靠人工田间调查的方式显然已经不能满足现代农业的需求^[3]。近几年快速发展的精准农业航空理念非常符合未来农业的发展方向,该理念的基本思想是通过空中和地面遥感采集农田信息,制作处方图,根据处方图实现精准施药。其中低空遥感正是快速获取柑橘园区病虫害信息的重要解决方案之一^[4]。与卫星遥感相比,低空遥感具有成本低、快速、灵活性高、实时获取数据等优点,能够在较大范围内快速获取空间地表信息,在农作物病虫害监测领域具有不可比拟的优势^[5]。高光谱检测技术已经在许多领域被应用,与可见光、多光谱图像相比,高光谱图像中蕴含着大量丰富的光谱信息,患病植株叶片细胞中的水分、色素、氮元素等发生变化会引起反射光谱的变化,通过分析反射光谱,可以建立农作物病虫害的判别模型。在柑橘黄龙病的高光谱遥感研究上,佛罗里达大学的 Li 等^[6]是最早开展无人机遥感柑橘黄龙病的研究团队之一,该团队使用地面高光谱仪建立了光谱库,并对无人机高光谱图像

进行波谱匹配识别分类,但由于获取高光谱的拍摄条件和采集设备的差异,该方法识别准确率较低;Kumar 等^[7]为了实现柑橘黄龙病患病区域的检测,利用多种植被指数特征和红边参数对无人机机载高光谱图像和多光谱图像进行分析,准确率也不理想,表明基于光谱库的方法对柑橘黄龙病检测存在一定的局限性。兰玉彬等^[8]通过无人机获取柑橘果园的低空高光谱影像,对患黄龙病和健康的植株反射光谱进行一阶微分和对数处理,并采用支持向量和 K 邻近进行建模和分类,分类的准确率达到 94.7%,表明传统机器学习方法能够实现无人机高光谱遥感检测柑橘黄龙病。Lan 等^[9]采用 SVM、集成学习等几种机器学习算法,对经过特征压缩与挖掘后的健康柑橘和黄龙病感染柑橘的多光谱样本进行建模,结果表明集成学习和神经网络算法具有很强的鲁棒性和分类效果。

通过机载高光谱相机把每一棵患柑橘黄龙病的植株准确检测并定位出来并非易事,原因在于柑橘种植不仅仅面临黄龙病威胁,还存在着许多其他病虫害,如缺素、黄斑病、黄脉病等,这些病虫害的症状与黄龙病极为相似,在低空遥感条件下很难准确区分^[10]。患病植株与健康植株差别较为明显,本研究提出通过无人机高光谱遥感技术检测出健康的柑橘植株,其余患病的植株通过人工田间调查方式判断其患病种类及患病程度。该方式是传统病虫害监测与现代遥感技术的结合,可以减少柑橘病虫害检测的时间和人力成本,同时确保病虫害识别的准确性。

1 材料与方法

1.1 试验地概况

本研究在广东省柑橘黄龙病绿色防控与新栽培模式研发示范基地进行,研究区域位于广东省惠

州市博罗县杨村镇 (北纬 23°29'57.81"~23°29'59.31", 东经 114°28'8.39"~114°28'12.26", 海拔 40 m), 属于亚热带季风气候, 年平均气温 21 ℃、降水量 1 932.7 mm、平均日照 2 023 h、无霜期长达 345 d, 土地肥沃, 四季如春, 为柑橘等果树的种植提供了优越的气候条件。

试验区域是患病虫害、缺素等症状的高发区, 柑橘品种为砂糖橘, 种植砂糖橘 9 行, 植株间行距 4 m, 每行有 30 余棵植株, 列距 2 m, 共有 334 棵植株。柑橘植株分为健康和患病 (多数患黄龙病、缺素症状, 少数患红蜘蛛病害等)2 类。园区的果树由华南农业大学农学院柑橘研究团队长期追踪检测, 植株健康与否均由该团队专家鉴定确认。

1.2 高光谱数据采集及预处理

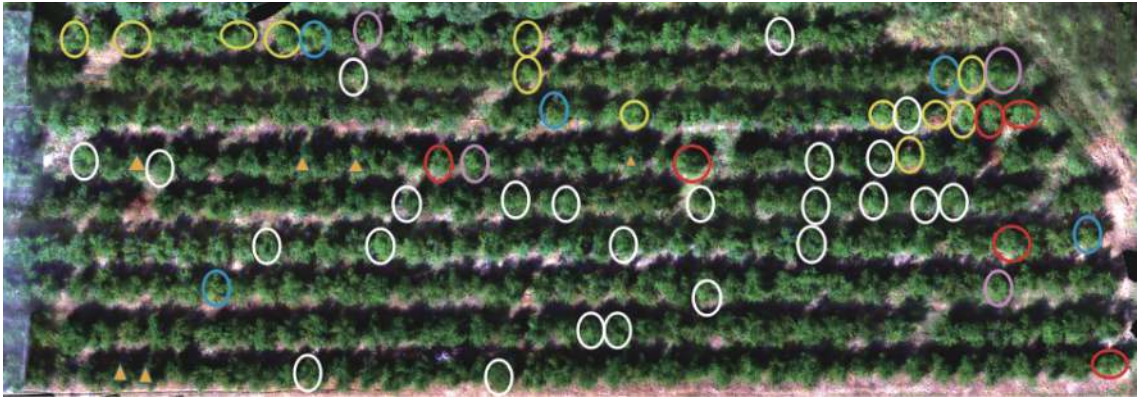
1.2.1 无人机高光谱数据采集 本研究使用的无人机平台是 DJI MATRICE 600 PRO 六旋翼无人机搭载 DJI RONIN-MX 云台。该无人机配备 6 块锂电池 (电池型号为 TB47S), 机臂可快速折叠, 起落架可自由收放, 具有便携灵活安全的优点, 维护成本低, 对起飞降落环境无特殊要求, 可适应复杂的地形环境, 空载质量为 9.5 kg, 最大起飞质量可达 15.5 kg。可实现最多 32 min 的悬停时间。满足本试验对大面积农田的数据采集要求。DJI RONIN-MX 云台是专业级的三轴通用云台系统, 适配多种型号的相机, 可以完美适配搭载在 MATRICE 600 PRO 飞行器上进行专业航拍。

本研究中使用 S185 机载高速成像仪采集高光谱数据, 该成像光谱仪由德国 Cubert 公司生产。S185 采用画幅式高光谱成像技术, 能够以快照式的速度进行所有光谱通道同步成像, 在毫秒级时间内获得范围内的高精度高光谱图像。S185 采用了小型化设计, 质量仅 500 g, 其所有光谱通道同步成像

的特点非常适合机载移动方式快速测量, 配套软件功能丰富, 可以用于批量光谱输出、求取植被指数、图像分类等。S185 具有非扫描式、全画幅、快速成像的特点, 光谱波长范围在 450~950 nm 之间, 采样间隔为 4 nm, 通道数为 125。

执行飞行任务之前, 先在地面进行 S185 高光谱相机的标准白板辐射校正和黑板校正, 在地面基站设定无人机飞行高度 50 m、重叠率 75%、飞行速度 2 m/s。柑橘植株冠层的高光谱数据采集要求晴天无云或少云、无风或风速较小的环境。本研究数据采集时间为 2017 年 12 月 9 日 11:00—13:00, 这个时间段是高光谱数据采集的最佳时间。在飞行试验中, 由遥控器控制无人机升高至指定高度, 然后按照预设飞行路线执行自动飞行任务。高光谱图像数据同时储存在单机版计算机中。

1.2.2 高光谱图像预处理 用 S185 相机采集到的数据由高光谱立体影像和全色图像组成。预处理前, 先删除起飞、降落和飞机转弯时拍摄的照片以及其他与研究区无关的照片, 以减少不必要的数据处理, 加快整体试验进度。然后, 用 Cube-Pilot 软件把高光谱立体影像转换成 jpg 格式, 在 PhotoScan 中拼接高光谱影像, 与拼接数码影像的过程基本一样。全色图像没有 pos 信息, 先拼接全色影像; 在进行照片对齐、建立密集点云、生成网格、生成纹理等操作后, 全色图像的拼接完成; 更改路径, 在全色图像的基础上替换数据, 最后导出全景正射影像。高光谱图像拼接完成后, 使用 ENVI 软件进行高光谱影像柑橘植株冠层的平均光谱提取。华南农业大学农学院柑橘研究团队已在研究区标记出完全健康植株、患黄龙病植株、缺素植株和其他患病植株, 所有植株分为 2 类: 有标记的植株为患病植株 (黄龙病、缺素); 没有标记的植株为完全健康植株 (图 1)。



粉、红、蓝、黄、白圆圈标记分别代表 1、2、3、4 级和患病未定级的黄龙病植株, 三角形标记为缺素植株; 没有标记的植株为完全健康植株
The pink, red, blue, yellow and white circle markers represented the plants with Huanglongbing disease of grades 1, 2, 3, 4 and indefinite respectively, and the triangular markers represented the plants lacking in nutrients; The plants without markers were the complete healthy plants.

图 1 试验区域及样本标注
Fig. 1 Experiment area and sample marking

使用感兴趣区域 (Region of interest, ROI) 提取工具选择 ROI 并提取其平均光谱。对所选取的柑橘植株冠层随机绘制 10 个 ROI, 每个 ROI 为 3×3 的像素矩阵, 通过 ENVI 解译每个 ROI 的光谱曲线, 把一个 ROI 当做一个样本点, 通过 ENVI 均化处理每个 ROI 的光谱值, 把得到的每个 ROI 的平均光谱作为在该样本点的光谱, 得到各样本点的光谱数据。导出所有光谱数据, 按照健康植株与患病植株归类。最后得到 3 400 条光谱数据。

1.3 特征工程

1.3.1 异常数据剔除 数据是科学研究中的最基础的资源, 数据质量的优劣直接影响以此作为依据的试验研究正确性和科学性。在实际应用中, 异常数据一般是因为各种客观因素或人为过失产生的, 异常数据的出现会极大程度地降低数据集的质量, 导致在这些异常数据中学习到的特征, 以至于最后的分类效果不理想。本研究中异常数据出现的原因有: 1) 高光谱图像数据的采集需要稳定的光照条件, 若采集数据过程中出现云层, 太阳光照强度变弱, 导致光线在植株冠层上的反射率变低, 从而出现异常数据; 2) 无人机在飞行过程中, 因环境 (风)、转向、变速等因素, 使得相机在拍摄过程中出现抖动, 相机拍摄角度发生改变, 导致出现异常光谱数据; 3) 由于提取平均光谱的过程需要手动操作, 选取的区域可能是图像拼接时的接合处或者是与柑橘植株较为相似的杂草等, 导致光谱数据异常。

本研究使用 OneClassSVM 算法剔除异常数据。OneClassSVM 是一个单样本检测算法, 常被用于异常数据的检测, 属于支持向量机家族的算法, 它与传统的分类回归支持向量机的区别是采用了无监督学习的方法^[11]。原本线性不可分的数据, 通过核函数将数据映射至更高维空间里, 变成线性可分, 且在数据点间找一个分割超平面使支持向量间隔最大^[12]。经过数据清洗过程, 共剔除异常样本 419 个, 最后剩余有效的健康植株高光谱样本数据 2 474 个、患病样本数据 507 个。

1.3.2 过采样处理 在训练二分类模型时, 经常会遇到正负样本比例不均衡问题。例如, 癌症病人诊断样本, 正负 (患癌与健康) 样本比例达到 1:99, 若分类器把所有样本都判断为健康样本, 模型的正确率达到 99%。然而, 模型的目的是找到患癌的样本, 这显然没有达到我们期望的效果。如果采用原来的不均衡样本进行训练, 算法都会存在类似的问题。过采样是针对数据不均衡问题的最佳解决办法

之一, 随机过采样是从少数类样本中又放回 (随机重复) 抽取样本以扩增少数类样本数量从而均衡正负样本比例, 该方法对少数类样本进行了多次重复取样, 虽然扩大了少数类数据规模, 但会造成过拟合。因此需要采用一些方法生成新的样本。SMOTE (Synthetic minority oversampling technique) 算法^[13]合成了少数类过采样技术, 其基本原理是对少数类集中的每个样本 x_i , 从 x_i 在少数类集中的 K 近邻中选一个样本 $\hat{x}_i$, 然后在 x_i 和 $\hat{x}_i$ 的连线上随机选取一个点作为合成的新样本, 其公式为:

$$x_{\text{new}} = x_i + \text{rand}(0, 1) \times |\hat{x}_i - x_i|, \tag{1}$$

式中, x_{new} 表示新合成的样本, x_i 表示随机选取的样本, $\hat{x}_i$ 表示 x_i 的近邻样本, $\text{rand}(0, 1)$ 表示在 (0, 1) 之间随机选取随机数。

采用 SMOTE 算法对测试集样本进行 1:1 过采样, 最后得出正负样本各 1 974 个, 其中, 生成负类样本 1 567 个。

1.3.3 特征波长提取 高光谱数据的数据量大、波段信息丰富, 可以解决很多在多光谱数据中不能解决的分​​类问题, 但波段间的相关性过大且具有大量冗余信息, 这为实际应用带来海量数据存储和处理的难题。本次试验的光谱数据有 125 个波段, 采用全波段建立判别模型, 每个波段均作为一个特征影响模型的训练结果, 但是, 并不是所有波长都对柑橘植株病虫害敏感, 需要针对柑橘植株病虫害按照一定的判定标准选择出一个最优的波段组合, 该波段组合需要保留原始高光谱数据的主要特征。因此, 需要对光谱数据进行初步压缩, 剔除不敏感的、冗余的波段, 使数据维度降低, 以减少运算的复杂度并降低分类模型的复杂度^[14]。同时为后续在实际应用中定制柑橘植株病虫害专属波段的多光谱相机提供理论依据。

连续投影算法 (Successive projections algorithm, SPA) 是一种新兴的波段提取方法, 能够消除大量波长变量之间的共线影响、减少模型复杂度, 可以消除光谱矩阵间的冗余信息, 被广泛应用于光谱计量学、医学成像、信号处理等领域^[15]。SPA 是一种前向迭代选择方法, 其基本原理是任选一个波长计算它在未选入的波长上的投影, 将投影向量最大的波长加入波长链中, 循环 N 次, 每个被选入的波长都与前一个被选入波长的线性关系最小^[16]。连续投影算法在数据分析软件 Matlab 中实现, 波段选取的评价指标为均方根误差 (Root mean square error, RMSE), RMSE 是观测值与真值偏差的平方和与观测次数比值的平方根, 用于衡量观测值与真

值之间的偏差。

1.4 模型的建立与评价

本研究基于全波段数据使用 BP 神经网络和

XgBoost 算法^[17], 以及特征波段的逻辑回归 (LR) 和支持向量机 (SVM) 算法, 分别建立柑橘健康植株分类模型, 并进行比较分析, 图 2 为特征工程技术路线图。

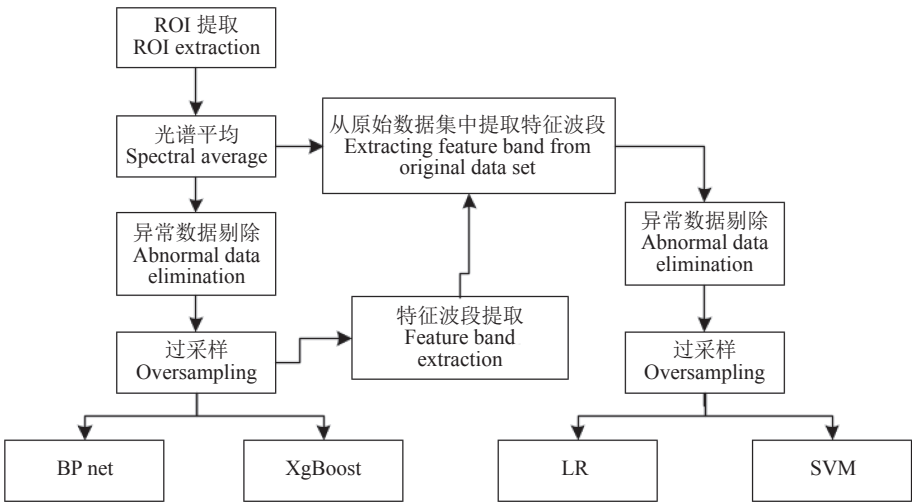


图 2 特征工程技术路线图
Fig. 2 The technology roadmap of feature engineering

BP 神经网络是一种多层前向神经网络, 全称为反向传播网络。其基本过程分为权值参数正向传播以及误差反向传播。每次更新连接权仅针对一个训练样本, 这样的算法需要进行多次迭代, 在训练集较大时, 标准 BP 神经网络会获得较好的解。BP 神经网络具有优秀的表达能力, 但亦有很多显而易见的缺点, 如容易陷入局部最小点、收敛速度慢、网络结构不易确定。XgBoost 算法对梯度提升决策树 (Gradient boosting decision tree, GBDT) 进行了改进, 是 GBDT 的工程实现。它的闪光点在于支持并行化, 在树成型最耗时的阶段, 对某个节点内选择最佳分裂点, 候选分裂点计算增益时用多线程并行, 使训练速度大大加快。在全波段高光谱数据这种分布稠密的数据集上, XgBoost 算法往往能得出非常优秀的泛化效果和表达效果、自动发现特征之间的高阶关系, 而且不需要对数据进行特殊的预处理 (标准化、归一化等), 具有较好的鲁棒性和解释性。

LR 是机器学习领域最常用也是最基础的二分类算法, 能通过未知类别对象的属性特征序列得到对象所处的类别。LR 在数据集线性分布时具有很优秀的分类效果。SVM 是机器学习中用来解决二分类问题的监督学习算法, 对于高维、非线性的数据问题有良好的分类能力。引入核函数能够避免高维变换带来的计算复杂性, 常用的核函数包括线性核函数 (Linear kernel function)、径向基核函数 (RBF

kernel function) 和多项式核函数 (Polynomial kernel function)。

针对不同的问题需要使用不同的模型评价标准, 一般有 Accuracy、Precision、Recall、F1、ROC (Receiver operating characteristic) 曲线和 AUC (Area under curve) 等几种, 前 4 种分别由混淆矩阵计算得出。具体公式如下:

$$ACC = \frac{(TN + TP)}{(TP + TN + FP + FN)}, \tag{2}$$

$$Precision = \frac{(TP)}{(TP + FP)}, \tag{3}$$

$$Recall = \frac{(TP)}{(TP + FN)}, \tag{4}$$

$$\frac{2}{F1} = \frac{1}{Precision} + \frac{1}{Recall}, \tag{5}$$

式中, ACC (Accuracy) 表示分类准确度, TP (True positive) 表示被模型预测为正的正样本, TN (True negative) 表示被模型预测为负的负样本, FP (False positive) 表示被模型预测为正的负样本, FN (False negative) 表示被模型预测为负的正样本。

ROC 起源于二战中敌机检测的雷达信号分析技术, 计算出 2 个重要量的值作为横、纵坐标作图, 得到了 ROC 曲线。在机器学习中, ROC 曲线的纵坐标是真正率 (True positive rate, TPR), 横坐标是假正率 (False positive rate, FPR)。AUC 被定义为 ROC 曲线下的面积, 即对 ROC 曲线求积分, 其值介

于 0.5~1.0 之间, AUC 的值越大表示该模型分类效果越好。计算公式如下:

$$TPR = \frac{TP}{TP + FN},$$

(6)

$$FPR = \frac{FP}{TN + FP},$$

(7)

$$AUC = \frac{1 + TPR - FPR}{2}.$$

(8)

由于测试集数据没有经过过采样算法处理, 测试集仍是不均衡数据集。在不均衡数据集分类算法的评估中, ACC 这一评估标准无法准确地反映少数类的分类效果, 将不再适用于本试验。而在本研究应用中的目标是要尽可能找出所有的患病样本, 宁可误判, 但不能漏判。召回率反映模型的容错性, 因此在本研究中召回率评分是最最重要的一个指标。

1.5 数据集设置

本研究基于全波段数据使用 BP 神经网络和

XgBoost 算法, 基于特征波段使用逻辑回归和支持向量机算法建立健康植株与患病植株的分类模型。为了方便比较模型的效果, 需要统一测试集数据。训练集与测试集的设置如下:

本数据集共有 2 981 个数据, 其中, 正类 (健康样本) 2 474 个、负类 (患病样本) 507 个, 属于不均衡样本集。第 1 步, 按一定比例随机抽选出一部分样本作为测试集 (正类样本 500 个、负类样本 100 个), 其余数据为训练集 (正类样本 1 974 个、负类样本 407 个); 第 2 步, 将训练集分为多数类子集 (正类) 和少数类子集 (负类), 然后通过过采样算法增加少数类子集的样本数据量, 过采样后的少数类子集和多数类子集组成新的数据集, 此时数据集为均衡数据集, 不均衡数据样本处理流程见图 3。最终, 训练集样本共有 4 948 个, 其中, 健康样本与患病样本各 2 474 个; 测试集样本共 600 个, 其中, 健康样本 500 个、患病样本 100 个。

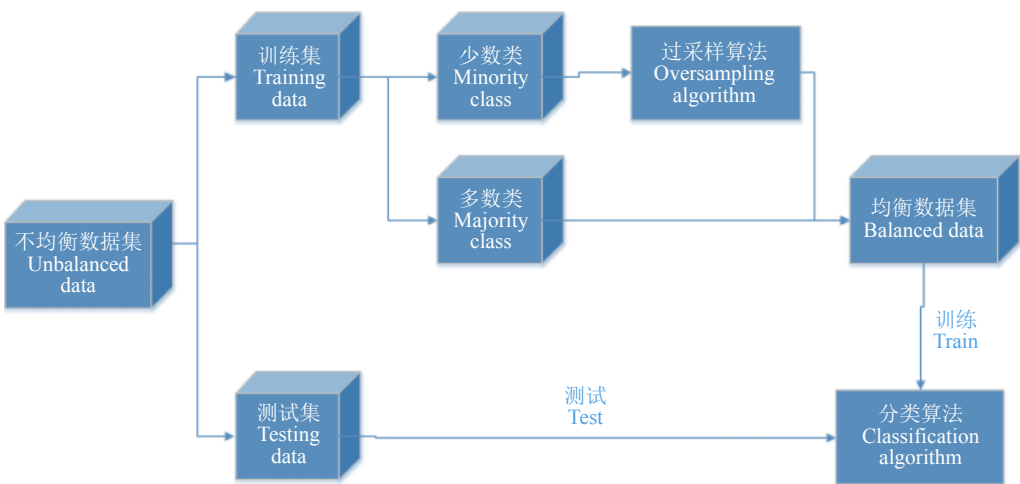


图 3 不均衡数据样本处理流程

Fig. 3 The flow diagram of processing unbalanced data samples

2 结果与分析

2.1 特征波长数的最优结果

图 4 为连续投影算法的均方根误差曲线 (RMSE)。由图 4 可见, 特征波长数为 4 时 RMSE 最小, 因特征波长提取的目的是为了更好地实现健康植株与患病植株的分类, 因此应该选取尽量少的特征波长^[18]。特征波长数为 2 时的 RMSE 为 0.518 05, 与特征波长数为 4 时的 RMSE 相差较小, 因此选择 2 个特征波长作为最优结果。选出来的特征波段是本数据集的第 61、78 个特征波段, 即 698 和 762 nm 是连续投影算法选出来的对健康植株分类结果贡献值最大的特征波长 (图 5)。

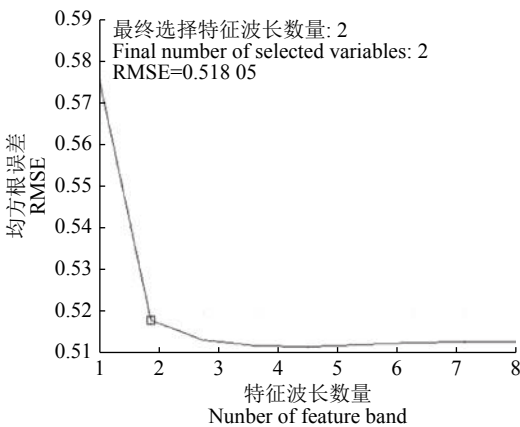


图 4 均方根误差 (RMSE) 曲线图

Fig. 4 The curve of root mean square error(RMSE)

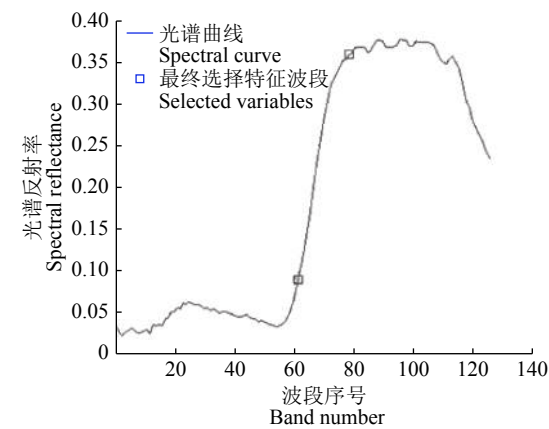


图 5 特征波段示意图
Fig. 5 Schematic diagram of feature band

表 1 BP 神经网络结构
Table 1 The structure of BP neural network

层(类型) Layer (Type)	输入节点数 Number of input node	输出节点数 Number of ouput node	激活函数 Activation function
输入层(全连接) Input layer(Full connect)	125	32	ReLU
隐含层1(全连接) Hidden layer 1(Full connect)	32	32	ReLU
隐含层2(全连接) Hidden layer 1(Full connect)	32	16	ReLU
输出层(全连接) Output layer(Full connect)	16	2	Sigmoid

2.2.2 XgBoost 由于 XgBoost 模型的参数较多，因此优化模型参数步骤是必要的，合适的参数能使模型泛化能力最大化。其中，num_round(树的数量) 和 eta(学习率) 是最终需要的参数，对模型效果影响较大。一般 XgBoost 参数调优步骤需要先通过网格搜索先找出 num_round 和 eta 的最佳取值，再调整其他参数。XgBoost 模型在经过参数优化后，得出的最优模型参数为：num_round=40, eta=0.06, max_depth=7, min_child_weight=3, subsample=0.6, alpha=0.125。模型在测试集上的分类结果见表 2。如表 2 所示，测试集 600 个样本中，使用 BP 神经网络和 XgBoost 模型正确分类的样本数达到 500 个以上，但是对患病样本的分类准确率偏低。

表 2 BP 神经网络和 XgBoost 测试结果 (混淆矩阵)
Table 2 The test results of BP neural network and XgBoost (Confusion matrix)

模型 Model	项目 Item	预测为患病 Predicted illness	预测为健康 Predicted healthy	总计 Total
BP神经网络 BP neural network	真实为患病 True illness	78	22	100
	真实为健康 True healthy	7	493	500
	总计 Total	85	515	600
XgBoost	真实为患病 True illness	85	15	100
	真实为健康 True healthy	13	487	500
	总计 Total	98	502	600

2.2 基于全波段的柑橘患病植株分类结果

2.2.1 BP 神经网络 本次试验使用 4 层网络结构，包括 1 层输入层、2 层隐含层和 1 层输出层，神经网络结构见表 1。训练时设置损失函数，使用交叉熵代价函数 (Cross-entropy cost function)，令 epoch=50, batch_size=100，即每批参加训练的样本大小为 100，一次 epoch 的意思是把所有训练样本都在神经网络中进行正向传播和反向传播。由表 1 可见，模型的输入层为波段数 (输入节点数)125 的光谱曲线向量，前 3 层均使用 ReLu 激活函数，ReLU 函数具有非饱和性，可以很好地解决梯度消失问题；模型输出层的输出节点数为 2，把样本分为健康与患病两类，使用 Sigmoid 激活函数。

2.3 基于特征波段的分类模型及效果

将高光谱数据中的冗余波长数据剔除，只留下特征波长对应的数据。由于对原数据集进行了特征波长提取操作，因此，需要对特征波长数据集重新进行异常数据剔除和重采样操作。上述操作后，最终得到训练样本共 4 016 个 (健康与患病样本各 2 008 个)，测试样本共 600 个 (健康样本 500 个、患病样本 100 个)。预处理同“1.3”，将训练集输入 LR 和 SVM 中，得出分类模型，使用测试集数据验证模型准确度，结果见表 3。由表 3 可见，测试集在特征波长下 LR 与 SVM 模型分类结果对患病样本的分类正确率较高，但对于健康样本的分类效果不佳。

表 3 特征波长下 LR 和 SVM 分类结果 (混淆矩阵)					
Table 3 The results of LR and SVM test bases of feature bands(Confusion matrix)					
模型 Model	项目 Item	预测为患病 Predicted illness	预测为健康 Predicted healthy	总计 Total	
LR	真实为患病 True illness	93	7	100	
	真实为健康 True healthy	44	456	500	
	总计 Total	137	463	600	
SVM	真实为患病 True illness	96	4	100	
	真实为健康 True healthy	61	439	500	
	总计 Total	157	443	600	

2.4 模型效果评估对比

统计 BP 神经网络、XgBoost、LR 和 SVM 4 个模型在测试集数据中的分类结果, 并计算出评价指标, 得分结果见表 4。由表 4 可见, BP 神经网络模型 AUC 得分为 0.883 0; 100 个患病样本中有 78 个样本预测正确, 剩余 22 个样本预测错误, 召回率 (Recall) 只有 78.00%, 效果不理想; 而 500 个健康样本中预测正确的有 493 个, 错判率很低。XgBoost 模型的 AUC 得分为 0.912 0, 且在我们最关注的指标——召回率中, XgBoost 模型有更好的效果, 其召回率为 85.00%。基于全波段的 BP 神经网络和 XgBoost 模型分类准确率 (ACC) 均超过 95%。本文的研究目的是为了尽可能找出全部的患病样本, 要求召回率尽可能高, 因此, 基于全波段的 BP 神经网络和 XgBoost 模型均未达到研究要求。

表 4 4 种模型评估得分				
Table 4 The evaluation scores of four models				
评价指标 Evaluating indicator	全波段 Full-wave band		特征波段 Feature band	
	BP Net	XgBoost	LR	SVM
ACC	0.951 7	0.953 3	0.915 0	0.891 6
Recall	0.780 0	0.850 0	0.930 0	0.960 0
Precision	0.917 6	0.867 3	0.678 8	0.611 5
F1-score	0.843 8	0.858 5	0.784 9	0.747 1
AUC	0.883 0	0.912 0	0.921 0	0.919 0

使用基于特征波长的 LR 和 SVM 分类模型对患病样本的召回率分别为 93.00% 和 96.00%, 分类准确率分别为 91.50% 和 89.16%, 均比 XgBoost 的分类准确率 (85.00%) 高, 基本达到本研究的要求; 但在健康样本分类上, 准确率相对低了一点。LR 和 SVM 分类模型的 AUC 值分别为 0.921 0 和 0.919 0, 均高于全波段数据建模的 BP 神经网络和 XgBoost 模型的 AUC 值, 究其原因, 全波段数据信息量较大, 其中很多波段信息皆为冗余波长, 不仅不能提

升模型效果, 甚至对分类结果有干扰作用^[19]; 而基于特征波段的建模方法只需使用简单的逻辑回归模型就能达到很不错的效果。本研究通过连续投影算法筛选出的特征波长 698 和 762 nm 对柑橘患病植株分类敏感且有效。

3 结论

本文以柑橘植株为对象, 在广东省惠州市博罗县杨村镇柑橘黄龙病绿色防控与新栽培模式研发示范基地, 利用多旋翼无人机 (大疆 M600 PRO) 作为遥感平台搭载高光谱成像仪 (S185), 采集柑橘植株冠层的高光谱遥感图像, 通过 ENVI 软件提取平均光谱作为原始数据, 先后经过数据清洗、过采样等数据预处理操作, 使得原来离散、不均衡的高光谱数据变得规范, 便于建模分析。采用连续投影算法提取特征波长, 基于全波段的 BP 神经网络算法、基于全波段的 XgBoost 算法、基于特征波段的逻辑回归算法和支持向量机算法建立了健康柑橘植株与患病 (患黄龙病、缺素) 植株的分类模型, 4 种模型均得到不错的分类效果, 基于全波段使用 BP 神经网络和 XgBoost 算法建立的分类模型 AUC 值分别是 0.883 0 和 0.912 0, 分类准确率均超过 95%; 基于特征波段使用逻辑回归和支持向量机算法建立分类模型, 召回率 (患病样本的分类准确率) 分别达到了 93.00% 和 96.00%, 证明了特征波长组合的有效性。本研究结果可为柑橘种植园的病虫害监测和精准防治提供一定的数据和理论支撑。

参考文献:

[1] DENG X, LAN Y, HONG T, et al. Citrus greening detection using visible spectrum imaging and C-SVC[J]. *Comput Electron Agr*, 2016, 130: 177-183.

[2] 陈波, 姚林建. 光谱检测技术在柑橘黄龙病诊断中的研究进展[J]. *赣南师范大学学报*, 2018, 39(6): 69-72.

[3] 兰玉彬, 邓小玲, 曾国亮. 无人机农业遥感在农作物病虫害诊断应用研究进展[J]. *智慧农业*, 2019, 1(2): 1-19.

[4] 纪景纯, 赵原, 邹晓娟, 等. 无人机遥感在农田信息监测中的应用进展[J]. 土壤学报, 2019, 56(4): 1-13.

[5] 黄文江, 张竞成, 师越, 等. 作物病虫害遥感监测与预测研究进展[J]. 南京信息工程大学学报(自然科学版), 2018, 10(1): 30-43.

[6] LI X, LEE W S, LI M, et al. Spectral difference analysis and airborne imaging classification for citrus greening infected trees[J]. *Comput Electron Agr*, 2012, 83: 32-46.

[7] KUMAR A, LEE W S, EHSANI R J, et al. Citrus greening disease detection using aerial hyperspectral and multispectral imaging techniques[J]. *J Appl Remote Sens*, 2012, 6(1): 63542.

[8] 兰玉彬, 朱梓豪, 邓小玲, 等. 基于无人机高光谱遥感的柑橘黄龙病植株的监测与分类[J]. *农业工程学报*, 2019, 35(3): 92-100.

[9] LAN Y, HUANG Z, DENG X. Comparison of machine learning methods for citrus greening detection on UAV multispectral images[J]. *Comput Electron Agr*, 2020, 171: 10524.

[10] 李修华, 李民赞, LEE W S, et al. 柑桔黄龙病的可见-近红外光谱特征[J]. *光谱学与光谱分析*, 2014, 34(6): 1553-1559.

[11] 尚方信, 郭浩, 李钢, 等. 基于 One-class SVM 的噪声图像分割方法[J]. *计算机应用*, 2019, 39(3): 874-881.

[12] TAX D M J, DUIN R P W. Support vector data description[J]. *Mach Learn*, 2004(54): 45-66.

[13] CHAWLA N V, BOWYER K W, HALL L O. SMOTE: Synthetic minority over-sampling technique[J]. *J Artif Intell Res*, 2002(16): 321-357.

[14] 成忠, 张立庆, 刘赫扬, 等. 连续投影算法及其在小麦近红外光谱波长选择中的应用[J]. *光谱学与光谱分析*, 2010, 30(4): 949-952.

[15] 高洪智, 卢启鹏, 丁海泉, 等. 基于连续投影算法的土壤总氮近红外特征波长的选取[J]. *光谱学与光谱分析*, 2009, 29(11): 2951-2954.

[16] 吴迪, 金春华, 何勇. 基于连续投影算法的光谱主成分组合优化方法研究[J]. *光谱学与光谱分析*, 2009, 29(10): 2734-2737.

[17] CHEN T, GUESTRIN C. XgBoost: A scalable tree boosting system[C]//The 22nd ACM SIGKDD International Conference. ACM, 2016: 785-794.

[18] 王术波, 韩宇, 陈建, 等. 基于深度学习的无人机遥感生态灌区杂草分类[J]. *排灌机械工程学报*, 2018, 36(11): 1137-1141.

[19] 高林, 杨贵军, 于海洋, 等. 基于无人机高光谱遥感的冬小麦叶面积指数反演[J]. *农业工程学报*, 2016, 32(22): 113-120.

【责任编辑 周志红】