

文章编号: 1672-8785(2019)07-0047-06

基于机器学习的深圳市坝光湿地园 树种高光谱分类

李 丹¹ 黄钰辉² 孙中宇¹ 张卫强² 甘先华²
王佐霖³ 孙红斌³ 杨 龙^{1*}

- (1. 广东省广州地理研究所地理空间信息技术与应用公共实验室, 广东广州 510070;
2. 广东省林业科学研究院森林培育与保护利用重点实验室, 广东广州 510520;
3. 广东省深圳市野生动物救助中心, 广东深圳 518040)

摘 要: 高光谱遥感数据为树种的精细识别提供了可能。为探索高光谱数据在树种识别中的能力, 本研究基于深圳市坝光古银叶树群落的8种主要树种叶片高光谱数据, 比较了6种光谱预处理方式和2种分类方法对树种分类识别精度的影响, 并基于随机森林算法对不同树种识别的特征波段进行了识别。研究结果表明, 一阶导数预处理方法在分类识别中性能最好, 平均分类精度为76.65%; 随机森林回归方法较支持向量回归算法的性能好, 模型平均分类识别精度为73.07%。从混淆矩阵可以看出, 多毛茜草、银柴、阴香易错分为假萍婆, 鸭脚木与银柴易错分, 银叶树和细叶榕易错分。400 nm、495 nm、615~675 nm、835 nm、915~975 nm、1035~1065 nm、1085~1135 nm、1265~1275 nm、1425~1535 nm、2040 nm、2100~2270 nm、2430 nm附近的光谱数据与8个树种分类识别有密切关系。

关键词: 机器学习; 树种分类; 高光谱; 叶片

中图分类号: S718.49 **文献标志码:** A **DOI:** 10.3969/j.issn.1672-8785.2019.07.006

Tree Species Classification in Baguang Wetland Park in Shenzhen Based on Machine Learning and Hyperspectral Data

LI Dan¹, HUANG Yu-hui², SUN Zhong-yu¹, ZHANG Wei-qiang², GAN Xian-hua²
WANG Zuo-lin³, SUN Hong-bin³, YANG Long^{1*}

- (1. Guangdong Provincial Geospatial Information Technology and Application Public Laboratory,
Guangzhou Institute of Geography, Guangzhou 510070, China;

收稿日期: 2019-06-26

基金项目: 广东省科技计划(2017A020216022、2018B030324002); 广东省科学院创新人才引进资助专项(2017GDASCX-0805); 广东省科学院实施驱动发展能力建设专项(2018GDASCX-0403); 林业科技创新平台运行补助项目(2018-LYPT-DW-069)

作者简介: 李丹(1985-), 女, 河南通许人, 副研究员, 主要从事农业遥感研究。

***通讯作者:** 杨龙(1981-), 男, 研究员, 主要从事恢复生态学研究。E-mail: yanglong@gdas.ac.cn

2. *Guangdong Key Laboratory of Forest Cultivation and Protection and Utilization,*

Guangdong Academy of Forest, Guangzhou 510520, China;

3. *Shenzhen Wildlife Rescue Center, Shenzhen 518040, China)*

Abstract: Hyperspectral remote sensing data provides the possibility for fine identification of tree species. In order to explore the ability of hyperspectral data in tree species identification, this study is based on the leaf hyperspectral data of eight major tree species in the *Heritiera littoralis* community of Baguang, Shenzhen, and compared the performance of six spectral preprocessing methods and two classification methods to classify tree species. Then based on the random forest algorithm, the importance of the each band was evaluated. The results showed that the first derivative preprocessing method had the best performance in classification and identification, and the average classification accuracy was 76.65%. The random forest regression method had better performance than the support vector regression algorithm, and the model average classification recognition accuracy was 73.07%. It can be seen from the confusion matrix that *Aidia pycnantha*, *Aporosa dioica*, *Cinnamomum burmanni* were recognized as *Sterculia lanceolata*. There were the misclassification between *Scheffero octophylla* and *aporosa dioica*. And *Heritiera littoralis* was also misclassified as *Ficus microcarpa*. Spectral data near 400 nm, 495 nm, 615—675 nm, 835 nm, 915—975 nm, 1035—1065 nm, 1085—1135 nm, 1265—1275 nm, 1425—1535 nm, 2040 nm, 2100—2270 nm, and 2430 nm are identified as the spectral features, which are most important for the classification of eight tree species.

Key words: machine learning; tree species classification; hyperspectral; leaf

0 引言

森林树种识别对于理解森林组成和树种分布、保护森林资源具有重要意义^[1]。传统的基于地面调查的树种识别方法在大区域森林资源调查中费时、费力，成本又高。高光谱遥感具有精细的光谱分辨率，能够探测地物细小的光谱差异，为快速、大面积的森林树种识别提供了可能，相关研究已成为国际高光谱植被遥感研究的热点^[2-3]。

国内外学者基于高光谱数据的树种分类研究主要有成像高光谱数据和非成像高光谱数据两类。一类以 Hyperion 高光谱影像为代表，探索用这类高光谱成像卫星遥感数据进行树种分类识别的可行性。另一类是基于叶片尺度的高光谱树种分类机理研究。这些研究表明，高光谱遥感数据能够较好地进行树种分类研究^[4-5]。

高光谱数据的波段间隔小，信息量大，数据质量受到信噪比等因素影响，且波段间的相关性高，这导致地物分类处理时的处理精度和效率下降，因此在分析之前对数据进行预处理、特征降维很有必要^[6]。高光谱数据的预处理形式主要有原始反射数据 (Re-

flectance, R)、一阶导数 (First Derivative, 1st)、二阶导数 (Second Derivative, 2nd)、基线校正 (Baseline)、标准正态变量变换 (Standard Normal Variate, SNV)、多元散射校正 (Multiplicative Scatter Correction, MSC) 等^[7]。特征降维的方法主要有特征提取和特征选择，常用的特征提取方法有主成分分析法 (Principal Component Analysis, PCA) 等。分类方法研究是遥感分类研究的重点内容，支持向量机分类 (Support Vector Machine Classification, SVMC)、随机森林分类 (Random Forest Classification, RFC) 等是最近几年使用较多的分类算法^[8-9]。

当前树种的遥感分类研究较少涉及半红树林树种。本研究以深圳市坝光银叶园的半红树林树种为研究对象，利用 ASD 光谱仪实测了 8 种叶片的可见光近红外光谱数据，分析不同数据预处理方法对树种识别分类精度的影响；并对比了随机森林算法和支持向量机算法对数据进行分类的精度；最后基于随机森林算法进行特征提取，识别了树种分类的特征波段，并分析树种光谱可分性的机理，为影像尺度的遥感分类研究提供参考。

1 研究区域的概况

研究区位于深圳市大鹏半岛最北端的坝光银叶树湿地园, 海拔的变化范围为 0~20 m, 属于亚热带季风气候区, 年均温为 22.1℃, 最高温度为 36.6℃, 最低温度为 1.4℃, 年均湿度为 80%, 年降水量为 1800 mm。

银叶树属于梧桐科银叶树属, 为热带、亚热带海岸红树林植物, 多分布于高潮线附近的海滩内缘以及潮水淹及的滩地, 属于典型的水陆两栖的红树植物^[10-11]。坝光银叶树群落是当前全球发现的树龄最长的天然古银叶树群落。该银叶树群落主要分布在滨海沼泽湿地生境、远离海岸的陆地生境以及邻海陆地生境内海湾滩地的海生环境^[12]。群落内其他建群种还包括阴香、假苹婆、多毛茜草树、海芒果、银柴、水石梓、鸭脚木、细叶榕等阔叶树种。

2 数据处理与建模方法

2.1 高光谱数据的获取

本文中, 叶片高光谱数据使用 ASD FieldSpec Pro@3 野外光谱仪测量。光谱数据的波段范围为 350~2500 nm, 选择了 10 种优势树种作为研究对象(见表 1)。随机选择每个树种的 15 棵成年个体, 用高枝剪摘下一个冠层树枝, 从每个树枝上选择 3 片成熟全展叶作为研究对象。将采集好的叶片保存于样品袋中, 带回实验室进行光谱测量。

利用 ASD 公司生产的 FieldSpec FieldSpec 3 便携式分光辐射光谱仪在室内测定每个叶片

的反射光谱。反射光谱的范围为 350~2500 nm, 光谱间隔为 1 nm。光源采用光谱仪自带的 50 W 卤化灯, 光源的方位角为 60°, 光谱仪的视场角为 25°, 探头垂直向下, 到叶片的距离为 0.10 m。将叶片正面平整置于反射率近似为 0 的黑色橡胶垫上。测定时探头分别正对待测叶片的前、中及后部。每次测量记录 5 个采样光谱, 共测定 15 个采样光谱, 以其平均值作为该叶片的光谱反射率值。测定过程中, 每隔 15 min 进行一次标准白板校正^[9]。

2.2 数据预处理方法

为降低高光谱数据的信息冗余, 提高光谱处理的效率, 对原始光谱数进行 10 波段平均; 同时去除光谱噪声较大的波段, 仅保留 415~2475 nm 光谱范围内共 207 个光谱数据。由于室内光谱测量时, 室内光源照明强度和叶片测量位置等因素对原始反射光谱有一定的影响, 光谱数据预处理时选用 1st、2nd、Baseline、SNV、MSC 数据^[13]。

2.3 数据分类方法

采用随机森林算法和支持向量机法两种分类法比较树种分类的精度。从 8 个树种中每个树种选 15 个样本, 随机选择每个树种中的 10 个样本建模、5 个样本验证, 组成 75 个样本的建模样本集、45 个样本的验证数据集。

3 结果与讨论

3.1 不同数据预处理方式对分类精度的影响

对 6 种不同数据预处理方法进行分类的精度见表 2。从表 2 中可以看出, 一阶导数进行

表 1 坝光古银叶树群落建群种的植物学特征与类别编号

种名	科名	种源	叶型	类别编号
多毛茜草树	茜草科	本地	阔叶	0
海芒果	夹竹桃科	本地	阔叶	1
假苹婆	梧桐科	本地	阔叶	2
细叶榕	桑科	本地	阔叶	3
银柴	大戟科	本地	阔叶	4
银叶树	梧桐科	本地	阔叶	5
阴香	樟科	本地	阔叶	6
鸭脚木	五加科	本地	阔叶	7

分类建模的分类效果最佳,其平均分类精度达到 76.65%。二阶导数类型的数据分类建模的平均精度为 76.07%,两者相差 0.58%。这两种差异的出现主要是由于一阶导数的预处理可以去除同波长无关的漂移,而二阶倒数去除同波长线性相关的漂移,削弱了数据间的线性相关性,同时会引入噪声,最终使得分类精度略低于一阶导数数据类型。基线校正数据类型的分类建模精度为 74.12%,仅次于原始数据的分类精度。相反,由于标准正太变量变换和多元散射校正主要消除颗粒分布不均等产生的散射影响,放大了可见区域的光谱差异,变换不具有压缩原光谱数据的能力,数据形式维度大和信息冗余严重可能是导致分类效果差的主要原因^[14]。

3.2 不同分类方法的精度比较

以上分析了数据预处理方式对分类精度的影响,现将分类建模方式和分类精度进行分析比较,结果见表 2。采用随机森林算法和支持向量机法两种分类建模方式。采用随机森林算法进行分类建模的效果最佳,其平均精度为 73.07%,该算法采用 Bagging 抽样技术,能有效地减少过拟合的风险,具有较好的抗噪声和泛化能力,因此在与支持向量机算法进行比较时,随机森林算法的稳定性好、精度高等特点得到了明显的体现。

通过分析、比较可知,当数据类型为一阶导数、分类算法采用随机森林算法时,分类精度最优。在此基础上,本文将数据集分为训练数据集和验证数据集,分别用来建立分类模型和验证分类精度,采用随机不放回抽样方法进行 10 次建模试验,最终以混淆矩阵的方式将结果进行输出,如图 1 所示。图 1 中,纵向表示实际的分类结果(分为 0~7 共 8 类),横向表示

建模的分类结果(分为 0~7 共 8 类),将以准确率为指标的 8×8 混淆矩阵进行可视化处理,颜色越浅的区域表示其分类精度越高。同时进一步分析试验分类错误的部分。统计 10 次建模试验的分类出错率较高的部分,以错误率为指标构成混淆矩阵,结果如图 2 所示。图 2 中,白色区域表示错误率较高,主要表现在细叶榕与多毛茜草、阴香、银叶树易错分,银柴与阴香易错分,银叶树与鸭脚木存在误分。

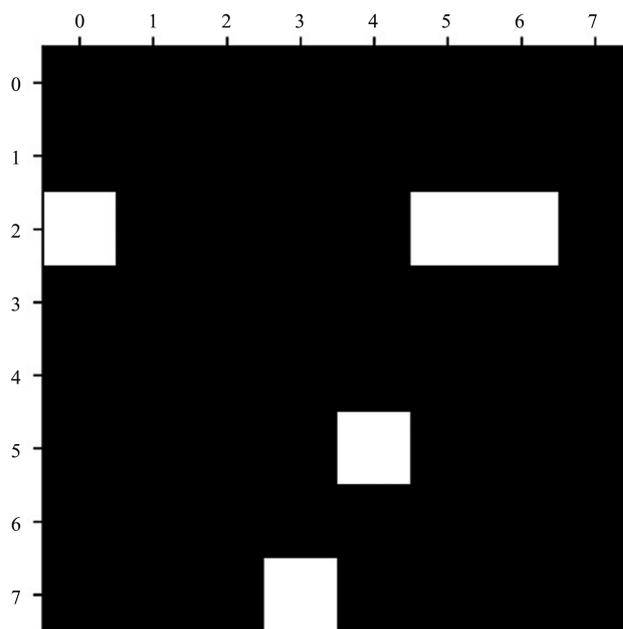


图 1 分类错误率结果的示意图

3.3 树种分类的特征波段

图 2 给出了采用随机森林算法基于原始光谱和一阶导数光谱分类模型的变量重要性图。从图中可以看出,400 nm、495 nm、615~675 nm、835 nm、915~975 nm、1035~1065 nm、1085~1135 nm、1265~1275 nm、1425~1535 nm、2040 nm、2100~2270 nm、2430 nm 附近光谱数据的数值相对较高。在分类中,这些波段的数值对物种分类的贡献大。

表 2 不同数据预处理类型下模型的分类精度

	R	1st	2nd	Baseline	SNV	MSC	平均精度
RFC	74.16%	76.24%	77.13%	72.97%	70.79%	67.13%	73.07%
SVMC	76.18%	77.06%	75%	76.47%	54.41%	56.47%	69.26%
平均精度	75.17%	76.65%	76.07%	74.72%	62.60%	61.80%	/

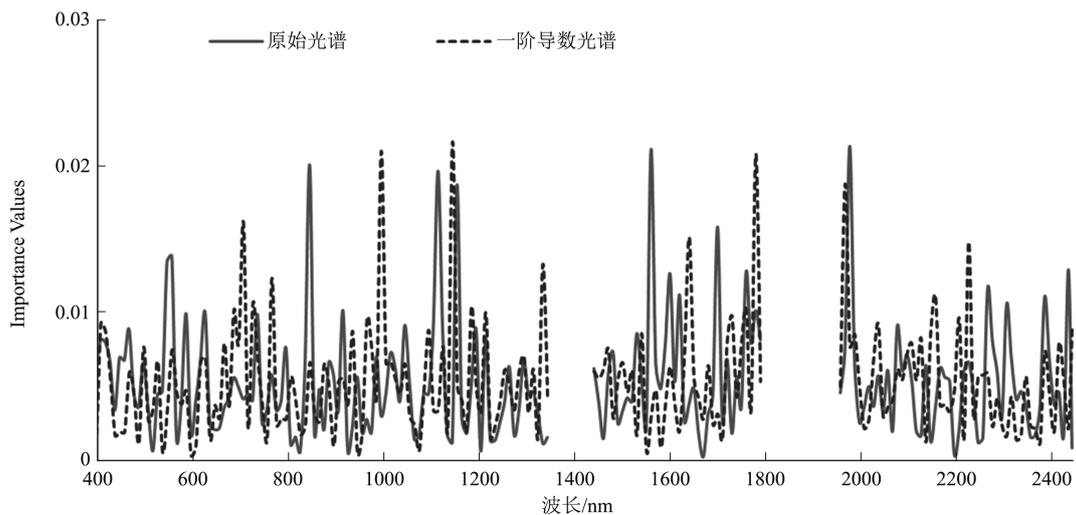


图 2 采用随机森林算法的变量重要性图

400 nm、615~675 nm 的吸收与色素吸收相关^[15-16], 490 nm 附近的吸收与氮素的吸收有密切关系^[17], 835 nm 的吸收与植被生物的化学属性重金属胁迫相关^[13], 915~975 nm 的吸收与水分吸收相关^[16], 1035~1065 nm 位于蛋白质吸收和油脂区间内^[15]; 1085~1135 nm 位于 C-H 伸缩键和 2 阶倍频处, 这附近的光谱与纤维素吸收有关; 1265~1765 nm 在一些研究中被用于反演水分和其他生化参数^[18-20], 1435~1535 nm 位于淀粉、糖类、木质素、纤维素、蛋白质、氮素等的特征吸收区间内^[15], 2040 nm 位于水分吸收区间内^[16]、2100~2270 nm 位于蛋白质、氮素、淀粉等吸收区间内; 而在基于 ASTER 数据进行一些植被分类的研究中, 用 ASTER 数据的 Band 9 (2360 ~ 2430 nm) 进行分类^[21-22]。

4 结论

本研究基于支持向量机分类算法和随机森林分类算法, 分析了深圳市坝光银叶园内重要半红树树种光谱分类的精度, 并识别了用于树种分类的主要光谱特征。进行不同数据预处理时, 一阶导数预处理分类的精度较高。两种分类方法结果显示, 随机森林分类算法的精度较支持向量机算法高, 其平均精度为 73.07%。对树种分类影响较大的光谱特征变量位于 400

nm、495 nm、615~675 nm、835 nm、915~975 nm、1035~1065 nm、1085~1135 nm、1265~1275 nm、1425~1535 nm、2040 nm、2100~2270 nm、2430 nm 附近。这些特征变量的选择多位于木质素、纤维素、淀粉、蛋白质、氮等吸收特征区间内。利用叶片光谱数据对半红树树种进行分类是可行的。进一步的研究将探讨利用冠层遥感数据和遥感影像进行分类识别的可行性。

参考文献

- [1] Sobhan I. Species Discrimination from a Hyperspectral Perspective [D]. Wageningen: Wageningen University, 2007.
- [2] Harrison D, Rivard B, Sánchez-Azofeifa A. Classification of Tree Species Based on Longwave Hyperspectral Data from Leaves, a Case Study for a Tropical Dry Forest [J]. *International Journal of Applied Earth Observation and Geoinformation*, 2018, **66**: 93-105.
- [3] Madonsela S, Cho M A, Mathieu R, et al. Multi-phenology WorldView-2 Imagery Improves Remote Sensing of Savannah Tree Species [J]. *International Journal of Applied Earth Observation and Geoinformation*, 2017, **58**: 65-73.
- [4] George R, Padalia H, Kushwaha S P S. Forest Tree Species Discrimination in Western Himalaya Using EO-1 Hyperion [J]. *International Journal*

- of Applied Earth Observation and Geoinformation*, 2014, **28**: 140–149.
- [5] Fassnacht F E, Latifi H, Stereńczak K, et al. Review of Studies on Tree Species Classification from Remotely Sensed Data [J]. *Remote Sensing of Environment*, 2016, **186**: 64–87.
- [6] 张银涛. 高光谱数据的树种分类波段选择研究 [D]. 杭州: 浙江农林大学, 2014.
- [7] 褚小立, 袁洪福, 陆婉珍. 近红外分析中光谱预处理及波长选择方法进展与应用 [J]. *化学进展*, 2004, **16**(4): 528–542.
- [8] 杨珺雯, 张锦水, 朱秀芳, 等. 随机森林在高光谱遥感数据中降维与分类的应用 [J]. *北京师范大学学报(自然科学版)*, 2015, **51**(s1): 82–88.
- [9] 李丹, 杨龙, 王重洋, 等. 广东省主要人工林树种光谱分类 [J]. *红外*, 2016, **37**(2): 36–41.
- [10] 王伯荪, 李鸣光. 重要值-面积曲线在热带亚热带森林中的应用 [J]. *植物生态学报*, 1986, **10**(3): 161–170.
- [11] 陈晓霞, 李瑜, 茹正忠, 等. 深圳坝光银叶树群落结构与多样性 [J]. *生态学杂志*, 2015, **34**(6): 1487–1498.
- [12] 简曙光, 吴梅, 刘念. 葫芦苏铁遗传多样性的等位酶分析 [J]. *广西植物*, 2005, **25**(6): 566–569.
- [13] 陆婉珍. *现代近红外光谱分析技术(第2版)* [M]. 北京: 中国石化出版社, 2007.
- [14] 史飞飞, 高小红, 杨灵玉, 等. 基于HJ-1A高光谱遥感数据的湟水流域典型农作物分类研究 [J]. *遥感技术与应用*, 2017, **32**(2): 206–217.
- [15] Curran P J. Remote Sensing of Foliar Chemistry [J]. *Remote Sensing of Environment*, 1989, **30**(3): 271–278.
- [16] Thenkabail P S. Optimal Hyperspectral Narrow Bands for Discriminating Agricultural Crops [J]. *Remote Sensing Reviews*, 2002, **20**: 257–291.
- [17] Kalascska M, Arturo Sanchez-Azofeifa G. Hyperspectral Remote Sensing of Tropical and Sub-Tropical Forests [M]. Boca Raton: CRC Press, 2008.
- [18] Sims D A, Gamon J A. Relationships Between Leaf Pigment Content And Spectral Reflectance Across a Wide Range of Species, Leaf Structures And Developmental Stages [J]. *Remote Sensing of Environment*, 2002, **81**: 337–354.
- [19] Sun P, Grignetti A, Liu S, et al. Associated Changes in Physiological Parameters And Spectral Reflectance Indices in Olive (*Olea Europaea* L.) Leaves in Response to Different Levels of Water Stress [J]. *International Journal of Remote Sensing*, 2008, **29**(6): 1725–1743.
- [20] Mahajan G R, Sahoo R N, Pandey R N, et al. Using Hyperspectral Remote Sensing Techniques to Monitor Nitrogen, Phosphorus, Sulphur and Potassium in Wheat (*Triticum Aestivum* L.) [J]. *Precision Agriculture*, 2014, **15**(5): 499–522.
- [21] Breunig, Fábio M. Classification of Soybean Varieties Using Different Techniques: Case Study with Hyperion And Sensor Spectral Resolution Simulations [J]. *Journal of Applied Remote Sensing*, 2011, **5**(1): 053533.
- [22] Prasad S T, Lyon J G. Hyperspectral Remote Sensing of Vegetation [M]. Boca Raton: CRC Press, 2012.