引文格式：杨柳,罗福正. 基于XGBoost算法的丘陵地区土地覆盖类型提取[J]. 地理空间信息, 2025, 23(9): 96-99. doi:10. 3969/j.issn. 1672-4623.2025.09.021

基于XGBoost算法的丘陵地区土地覆盖类型提取

杨　柳[1,2]，罗福正[1,2]
（1. 青海省地理空间和自然资源大数据中心，青海 西宁 810000；2. 青海省地理空间信息技术与应用重点实验室，青海 西宁 810000）

摘　要：为准确、高效地提取大范围丘陵地区土地覆盖类型，以安徽省宣城市为研究区，基于Sentinel-1 SAR影像资料和Sentinel-2多光谱影像构建多时相光谱特征、指数特征和极化特征；以Gini系数为阈值进行特征优选，采用XGBoost算法构建模型并提取了典型复杂丘陵城市的地物覆盖类型。结果表明，春夏两季地物间的差异性明显增大，NDVI、近红外波段的重要性较高，可分离性也有所提高，Gini系数均达到0.750以上；XGBoost算法的分类精度高于RF、SVM算法，总体精度和Kappa系数分别为84.39%和0.82。通过研究最优特征组合，在多时相特征支持下提取了宣城市土地覆盖类型，可为复杂地物和地貌特征覆盖区的土地利用类型研究提供科学依据。

关键词：丘陵地区；Sentinel-2；土地覆盖类型提取；XGBoost算法
中图分类号：P237　　**文献标志码：**B　　**文章编号：**1672-4623（2025）09-0096-04

Land Coverage Type Extraction in Hilly Areas Based on XGBoost Algorithm

YANG Liu[1,2], LUO Fuzheng[1,2]
(1. Qinghai Geospatial and Natural Resources Big Data Center, Xining 810000, China;
2. Qinghai Provincial Key Laboratory of Geospatial Information Technology and Application, Xining 810000, China)

Abstract: To accurately and efficiently extract land coverage types in hilly areas, taking Xuancheng City, Anhui Province as the research area, based on Sentinel-1 SAR images and Sentinel-2 multi-spectral images, we constructed multi-temporal spectral features, index features and polarization features. Taking the Gini coefficient as a threshold, we performed feature selection. We used XGBoost algorithm to build models, and extracted land coverage types in a typical complex hilly city. The results show that the differences between land coverage types increase significantly in spring and summer, with the NDVI index and near-infrared band being of high importance, and the separability also improves, with Gini coefficients all reaching above 0.750. The classification accuracy of XGBoost algorithm is higher than that of RF and SVM algorithms, with an overall accuracy and Kappa coefficient of 84.39% and 0.82, respectively. By studying the optimal feature combinations and leveraging multi-temporal features, the extraction of land coverage types in Xuancheng City was achieved, which could provide scientific evidence for land-use types research in areas with complex land coverage and topographical features.

Key words: hilly area, Sentinel-2, land coverage type extraction, XGBoost algorithm

土地利用与覆盖信息提取对于丘陵城市规划建设具有举足轻重的指导意义[1-2]。遥感技术因其广域覆盖度、短重访周期、连续观测等优势，在土地利用分类和变化监测领域得到了广泛应用[3]。越来越丰富的遥感数据为土地利用分类的识别研究提供了坚实的数据基础[4]，尤其是雷达影像的问世，有效规避了传统光学遥感图像受云雨天气影响易产生大量无效像元的弊端[5]。将光学影像与雷达影像相结合开展复杂地形下的地物识别分类正成为热点研究领域。近年来，大量学者[6-12]开展了丘陵地区地物覆盖提取与定量遥感研究工作，大多基于光学影像采用单一机器学习算法构建识别模型，虽具有结构简单、易理解、实用性强等优势，但在处理复杂地形条件时仍存在高维特征处理运算速度慢、在结构复杂区域鲁棒性差等缺点，且单一光学影像也易受云雨天气影响从而降低识别精度。极度梯度提升树（XGBoost）算法作为一种集成机器学习算法，在遥感地物识别领域具有更快的训练和预测速度、更高的精度和泛化能力、强大的特征选择和重要性解释功能、灵活的参数调整以及易于集成的特性[13]。因此，本文将光学影像与雷达影像相结合，采用XGBoost算法开展丘陵地区复杂地形和地貌特征下的土地覆盖类型提取，以提高识别精度，弥补传统机器学习识别算法在丘陵遥感识别领域的不足。

项目来源：青海省自然科学基金面上项目（2020-ZJ-927）。
第一作者简介：杨柳（1989—），本科，工程师，从事地图制图测绘类工作，E-mail：paperYangliu@163.com。

1　研究区概况与数据来源

本文选取的研究区为安徽省宣城市，位于安徽省东南部，皖南山区与沿江平原结合地带；地貌复杂多样，南部山地、丘陵和盆谷交错，中部丘陵、岗冲起伏，北部除一部分丘陵外，绝大部分为广袤的平原和星罗棋布的河湖港汊，境内河流属长江和钱塘江流域，河湖众多；属亚热带季风气候，常年平均气温为16.0℃，四季分明、气候温和、雨量充沛；丰富的地貌特征和适宜的气候特点衍生出以针叶林、阔叶林等多种森林植被以及草地、耕地为主的地物覆盖特征[14]。随着城镇化和工业化进程的加快，宣城市土地利用类型正在迅速变化，空间异质性逐渐增强，因此快速、准确提取复杂地形下的土地覆盖特征，对于城市空间规划和资源管理利用至关重要。

本文采用的数据为2020年3月—2021年2月覆盖宣城市的Sentinel-1卫星C波段雷达数据和Sentinel-2卫星多光谱成像仪监测数据产品，均从GEE平台调用和处理。为有效提升Sentinel-2影像质量、减少云污染影响，对Sentinel-2卫星的13个波段进行几何校正、辐射定标和大气校正，并通过云概率数据集QA60波段进行去云[15]，再按照春（3—5月）、夏（6—8月）、秋（9—11月）、冬（12月—次年2月）的顺序进中值合成，生成具有季节代表性的高质量无云影像集。对同期Sentinel-1卫星数据进行热噪声去除、辐射校准和地形校正等预处理，并采用Refined Lee滤波最大化降低噪声影响[16]，逐像素点进行VV、VH后向散射系数的季节性中位数计算，以消除观测角度带来的像素噪声。为保证研究样本的代表性和准确性，于2020年3月15—18日、2020年10月10—13日开展两次连续外业调查工作，利用GPS记录仪记录样本经纬度信息，并结合Google Earth高分辨率影像进行目视解译。经验证，得到有效水体样本75处、阔叶林样本83处、针叶林样本62处、灌木样本65处、人工表面样本55处、农业用地样本78处、草地样本50处、裸地样本58处。采用分层随机抽样的方式，按照5∶1的比例划分训练样本和验证样本。

2　研究方法

2.1　特征空间构建与选择

利用Sentinel-2预处理中值合成光学影像，提取8个光谱波段并计算构造归一化植被指数（NDVI）、归一化建筑指数（NDBI）和归一化水体指数（NDWI），同时提取对应时序的Sentinel-1中值合成后的VV、VH后向散射系数作为初选特征。由于各地类的样本数量存在差异，加上季节中值合成后特征维度成倍数增加，极易造成特征信息冗余，为提高分类模型实际效率，尽可能维持高维特征分类模型精度，本文采用以Gini系数为衡量标准的随机森林（RF）模型对特征进行优选，并通过对比特征在决策树上的平均贡献率来评估特征重要性，优选出贡献率大于均值的特征用于分类。

2.2　XGBoost算法

本文利用Gini系数对特征进行重要性排序，再采用XGBoost算法建立宣城市地物覆盖类型提取模型。作为一种集成机器学习算法，XGBoost由梯度提升树（GBT）算法改进而来，具有预测能力高、鲁棒性强以及支持并行运算等优势[17]。与GBT算法采用一阶导数实现梯度下降不同，XGBoost算法采用的二阶泰勒算法具有更强的适应能力，每棵新树都是在原树的基础上进行改进，目的是最小化损失函数。本文中损失函数采用对数损失方案[18]。梯度提升的基本流程是先计算当前模型的预测误差，再利用该误差来调整模型参数，使每次迭代都能降低误差，提高模型的预测能力。为了验证效果，本文同时采用支持向量机（SVM）[19]和RF算法[20]构建分类模型，其中SVM算法的核函数采用径向基核函数。各算法模型构建的最优参数确定方案均采用格点搜索方法。

2.3　精度评价

本文以实地调查数据为基础，结合卫星目视解译结果，对模型分类结果进行混淆矩阵计算，从而完成模型精度评价工作。评价指标包括总体精度（*OA*）、用户精度（*UA*）、生产者精度（*PA*）和Kappa系数。

3　研究结果与分析

3.1　特征优选分析

为尽可能避免高维特征信息冗余导致的计算资源浪费和模型精度降低，本文基于GEE云计算平台，计算各地类样本的Gini系数并进行归一化处理；再根据各地类样本Gini值大小来评价样本集合的区间纯度，并计算得到42个特征的Gini系数均值为0.384；最后以均值为阈值，初步优选出16个贡献率大于均值的特征（表1），这些特征的Gini系数均在0.565以上，对于不同地类均体现出较好的线性可分离性，其中NDVI、近红外波段（B8）的重要性排序明显靠前，尤其是在春季和夏季。由于植被信息凸显，地物间的差异性明显增大，可分离性也有所提高，春季NDVI、夏季

NDVI以及春季B8的Gini系数均达到0.750以上。

表1 不同分类特征的Gini系数

序号	特征	Gini系数	序号	特征	Gini系数
1	NDVI_1	0.762	9	NDWI_2	0.622
2	NDVI_2	0.755	10	NDBI_2	0.621
3	B8_1	0.752	11	B2_2	0.616
4	B4_3	0.733	12	NDWI_1	0.611
5	NDBI_3	0.681	13	NDVI_4	0.601
6	NDVI_3	0.675	14	B8_2	0.582
7	B8_4	0.655	15	NDBI_3	0.571
8	VV_3	0.632	16	VV_2	0.565

注：1、2、3、4分别代表春、夏、秋、冬季。

为减少分类计算过程中的数据冗余，本文利用RF算法按照Gini系数由高到低的顺序依次向模型中增加特征量，观察地物分类*OA*和Kappa系数在特征量增加后的具体变化。结果表明（图1），随着特征数的增加，前期*OA*和Kappa系数呈快速增长态势，当特征数为8时，RF模型的*OA*达到最大（91.25%），Kappa系数为0.85；随后*OA*基本趋于稳定并在某些特征变量增加后出现异常降低的情况。因此，本文采用前8个特征变量作为优选特征集合，参与地物分类模型的下一步构建。

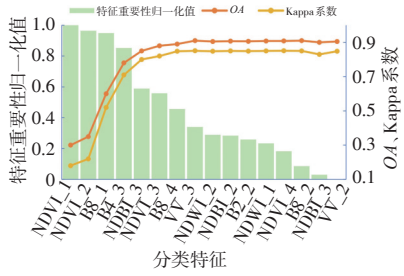


图1 分类特征优化分析

3.2 不同模型土地分类结果分析

本文以最优特征分类集合为输入，基于GEE平台，采用XGBoost算法提取宣城市的地表覆盖类型。为对比常用机器学习算法与XGBoost算法的性能差异，本文对最优特征集合同时采用RF算法和SVM算法建立识别模型（图2），可以看出，3种算法在复杂地形中均具有良好的识别准确率，在水体、建筑等与植被差异较大的地物分类中呈现出较好的识别效果，但RF算法在北部农业种植区中存在农业用地像元被误分为草地的现象，在南部林业用地区域内过拟合现象严重，大面积阔叶林被误分为针叶林或灌木林，而SVM和XGBoost算法的识别结果在整体农林业资源结构布局上更合理。为更鲜明地对比和展示不同算法的分类细节差异，本文选取中部典型复杂地形和地物覆盖区域进行详细验证说明（图3），可以看出，相较于其他两种算法，XGBoost算法的分类识别结果地物间的界限分明，SVM算法对于裸地、草地和针叶林的提取效果较RF算法均有了明显提升，但在处理高异质性的区域尤其是“异物同谱”现象严重的植被覆盖区时存在错分现象，对于零散分布于阔叶林中的针叶林和灌木，识别效果较差。总体来看，与其他两种算法相比，XGBoost算法不仅实现了土地利用分布结构的准确提取，在分类细节上也体现出较强的鲁棒性，具备更清晰明确的边界特征，对不同植被类型的提取更准确。

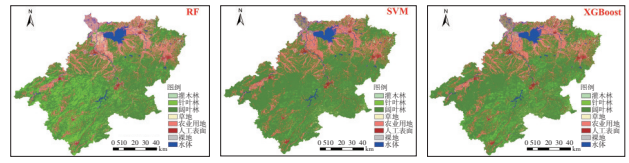


图2 3种分类算法下丘陵区土地覆盖提取结果

审图号:GS(2024)0650

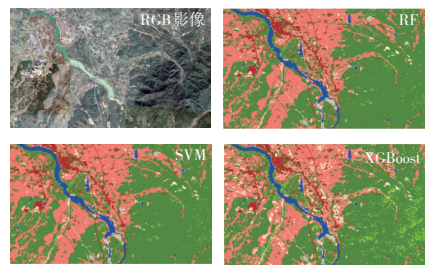


图3 典型复杂地形下3种算法的提取结果

注：卫星影像由Sentinel-2卫星红绿蓝通道合成。

基于RF、SVM和XGBoost算法的分类模型，各地类的分类精度见表2，XGBoost算法的*OA*为84.39%，比RF和SVM算法分别提高了9.01%和3.42%，Kappa系数为0.82，明显高于其他两种算法；RF和SVM算法的*OA*不错，但在灌木林、针叶林和阔叶林的区分以及草地的*UA*和*PA*表现上与XGBoost算法有较大差距。在复杂地貌和地物覆盖情况下，XGBoost算法对于零散目标特征地物具有更好的分类识别性能，体现

表2 3种分类模型精度评价表/%

地物类别	RF算法		SVM算法		XGBoost算法	
	UA	*PA*	*UA*	*PA*	*UA*	*PA*
灌木林	70.61	72.33	75.82	77.35	80.24	81.38
针叶林	73.32	72.25	78.55	76.93	83.38	82.72
阔叶林	75.02	70.35	80.38	79.65	86.12	84.32
草地	78.28	76.72	80.32	82.38	85.32	86.35
农业用地	84.33	82.97	85.38	86.24	88.28	87.35
人工表面	83.33	84.82	87.32	86.45	90.25	92.72
裸地	72.35	73.38	76.38	77.53	78.32	79.25
水体	93.35	91.82	94.32	92.28	95.67	94.29

注：RF、SVM、XGBoost算法的*OA*分别为75.35%、78.97%、84.39%，Kappa系数分别为0.70、0.75、0.82。

出明显的算法优势。

4　结　语

本文以宣城市为研究区，基于Sentinel-1 SAR影像和Sentinel-2多光谱影像，通过GEE云平台以季节中值合成的方式构建光谱特征、指数特征和极化特征，并计算各特征的Gini系数进行特征优选；选取地物覆盖分类模型构建所需的最优特征组合，采用XGBoost算法构建模型并识别提取了典型复杂丘陵城市的地物覆盖类型。

1）对植被信息较敏感的NDVI、B8在特征中的重要性排序明显靠前，尤其是在春季和夏季，可分离性也有所提高，春季NDVI、夏季NDVI和春季B8的Gini系数均达到0.750以上。

2）XGBoost算法的分类精度高于RF、SVM算法，*OA*和Kappa系数分别达到84.39%和0.82，不仅能准确提取地类的分布结构，而且在分类细节上也体现出较强的鲁棒性，具备更清晰明确的边界特征，能更精细地提取不同植被类型。

XGBoost算法因引入正则化目标函数和分裂点搜索近似算法，在训练速度、准确性和处理复杂数据能力上都表现出显著优势，但应用于其他复杂地物覆盖区域时，识别精度仍有一定的提升空间。因此，如何利用XGBoost算法进行模型自适应性参数设置以及与深度学习算法的结合应用是未来继续研究的重点方向。

参考文献

[1] 王周龙,冯学智,刘晓枚,等. 秦淮河丘陵地区土地利用遥感信息提取及制图[J]. 遥感学报,2003,7(2):131-135

[2] 李志国,张过师,刘毅,等. 湖北省主要城市园林绿地土壤养分评价[J]. 应用生态学报,2013,24(8):2 159-2 165

[3] 贾坤,李强子,田亦陈,等. 遥感影像分类方法研究进展[J]. 光谱学与光谱分析,2011,31(10):2 618-2 623

[4] 史泽鹏,马友华,王玉佳,等. 遥感影像土地利用/覆盖分类方法研究进展[J]. 中国农学通报,2012,28(12):273-278

[5] Dubeau P, King D J, Unbushe D G, et al. Mapping the Dabus Wetlands, Ethiopia, Using Random Forest Classification of Landsat, PALSAR and Topographic Data[J]. Remote Sensing,2017,9(10):1 056

[6] 赖格英,杨星卫. 南方丘陵地区水稻种植面积遥感信息提取的试验[J]. 应用气象学报,2000,11(1):47-56

[7] 谭磊,赵书河,罗云霄,等. 基于对象特征的山东省丘陵地区多时相遥感土地覆被自动分类[J]. 生态学报,2014,34(24):7 251-7 260

[8] 易凤佳,李仁东,常变蓉,等. 面向对象的丘陵区水田遥感识别方法[J]. 农业工程学报,2015,31(11):186-193

[9] 秦泉,王冰,李峰,等. 面向对象的GF-1卫星影像苹果树种植面积遥感提取研究:以山地丘陵地区的栖霞市为例[J]. 沙漠与绿洲气象,2020,14(2):129-136

[10] Wang H. Impacts of Topography on the Land Cover Classification in the Qilian Mountains, Northwest China[J]. Canadian Journal of Remote Sensing,2020,46(3):344-359

[11] 张亚新,吴志勇,何海. 苏南丘陵区土地利用遥感分类方法适用性研究[J]. 湖北农业科学,2021,60(5):138-143

[12] 宋明明,郑文龙,卞国栋,等. 基于多时相Landsat8数据的南方丘陵区典型地物信息提取:以秦淮河流域为例[J]. 科学技术与工程,2016,16(18):209-216

[13] Zamani J M, Cao C, Ni X, et al. $PM_{2.5}$ Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data[J]. Atmosphere,2019,10(7):373

[14] 殷梦杰,赵萍,朱翠翠,等. 基于Sentinel-2卫星数据的南陵—宣城矿集区矿化蚀变信息提取[J]. 合肥工业大学学报(自然科学版),2023,46(9):1 254-1 262

[15] 梁顺林,白瑞,陈晓娜,等. 2019年中国陆表定量遥感发展综述[J]. 遥感学报,2020,24(6):618-671

[16] 郑少兰,刘龙威. 单时相Sentinel-1A卫星SAR数据水稻信息提取[J]. 地理空间信息,2020,18(4):61-64

[17] Chen T, Guestrin C. Xgboost: A Scalable Tree Boosting System[C]//Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016:785-794

[18] 王植,方锦雄,李安翼. 多尺度卷积神经网络的遥感影像多标签分类[J]. 测绘与空间地理信息,2020,43(9):7-10

[19] 王振武,孙佳骏,于忠义,等. 基于支持向量机的遥感图像分类研究综述[J]. 计算机科学,2016,43(9):11-17

[20] 李敏,刘国栋,谭凌. 基于随机森林的土地利用分类与景观格局分析[J]. 地理空间信息,2022,20(2):51-56