引文格式：袁乃全, 熊丽军. 基于 GEE 与集成学习的土地覆盖信息提取方法[J]. 地理空间信息, 2023, 21(5): 76-79. doi:10. 3969/j.issn. 1672-4623.2023.05.018

基于 GEE 与集成学习的土地覆盖信息提取方法

袁乃全[1]，熊丽军[2]
（1. 湖南地科院规划设计有限公司，湖南 长沙 410118；2. 安化县自然资源局，湖南 益阳 413599）

摘　要：高精度的土地利用制图对国土空间规划和社会经济发展具有重要意义。为快速、准确获取大范围土地利用/覆盖信息，以洞庭湖区为研究对象，以 Google Earth Engine（GEE）云平台为支撑，基于 Sentinel-2 多光谱影像和 Sentinel-1 SAR 影像，利用 Stacking 集成学习算法进行土地覆盖类型分类。结果表明，该方法能较好地区分不同土地利用类型，总体精度和 Kappa 系数分别达到了 88.53%和 0.85。研究结果以期为大尺度土地利用制图以及国土资源管理与规划提供科学方法和依据。

关键词：土地利用/覆盖；洞庭湖；GEE；Stacking 集成学习算法
中图分类号：P237　　**文献标志码：**B　　**文章编号：**1672-4623（2023）05-0076-04

Land Coverage Information Extraction Method Based on GEE and Integrated Learning

YUAN Naiquan[1], XIONG Lijun[2]
(1. Hunan Academy of Geosciences Planning and Design Co., Ltd., Changsha 410118, China;
2. Anhua County Natural Resources Bureau, Yiyang 413599, China)

Abstract: Highly accurate land-use mapping is of great importance to the territory spatial planning and socio-economic development. In order to obtain large scale land-use/land coverage(LULC) information quickly and accurately, taking Google Earth Engine(GEE) cloud platform as the research platform, taking Dongting Lake as the research object, based on Sentinel-2 multi-spectral images and Sentinel-1 SAR images, we used Stacking integrated learning algorithm to classify land coverage types. The results show that the method could better distinguish different land-use types, and the overall accuracy and Kappa coefficient reach 88.53% and 0.85, respectively. The results of this paper are intended to provide a scientific method for large-scale land-use mapping and territory resource management and planning.

Key words: LULC, Dongting Lake, GEE, Stacking integrated learning algorithm

由于与生态变化、社会经济问题、水文学、气候变化以及生态系统服务密切相关，土地利用/覆盖（LULC）制图一直是研究人员、政府和国际组织关注的热点[1-2]。然而，利用遥感技术进行大范围的 LULC 制图是具有挑战性的，一方面是由于环境的异质性和地表的光谱复杂性，另一方面是由于硬件和软件的限制[3-4]。鉴于卫星图像数据量的空前增加、卫星图像处理的实质性进展以及遥感新平台（GEE）的出现，确定和提出大尺度遥感图像的解决方案至关重要。

GEE 是一个强大的云计算平台，用于处理卫星图像并在全球范围内提供地理空间数据。它不仅是一个多 PB 存储的超级数据库，还提供先进的处理模型和可视化工具[5-6]。近年来，学者们已将 GEE 用于各种尺度的遥感应用，如湿地划定、土地覆盖信息提取和作物制图[7-8]；LONG X R[9]等基于 GEE 平台的 Landsat TM/ETM+/OLI 数据，利用 LandTrendr 算法对洞庭湖湿地植被近 20 年的变化情况进行了分析[9]；熊元康[10]等以 GEE 云平台为支撑，以 Sentinel-1/2 和 Landsat7/8 为数据源，采用机器学习分类方法提取了广东省土地覆盖信息；毛丽君[11]等基于 GEE 云平台，利用随机森林算法生成了多年的钱江源国家公园土地覆盖数据集；文可[12]等提出了一种基于 GEE 和时序 Sentinel-1/2 遥感数据的水产养殖池识别方法；姜伊兰[13]等提出了一种基于 GEE 与时序 NDVI 差异指数的作物种植区提取方法；付甜梦[14]等基于 GEE 平台的 Landsat TM/OLI 遥感影像，采用随机森林方法提取了 2000—2018 年苏拉威西岛地表覆盖信息。上述研究在利用 GEE 云平台提取土地覆盖信息时偏向于采用其自带的机器学习算法，如支持向量机（SVM）、随机森林（RF）和决策树（DT）等方法；但这些机器学习算法各有优缺点，没

收稿日期：2021-11-09；修回日期：2022-02-22。
项目来源：湖南省自然资源科技计划资助项目（2020-37）。
第一作者简介：袁乃全（1988—），硕士研究生，高级工程师，研究方向为国土空间规划与区域发展、地理空间大数据应用，E-mail: 1426647089@qq.com。

有一种分类方法对所有数据集都表现良好[15]。此外，对于景观结构复杂且异质性程度高的区域，单一分类器受到了更大的挑战。Bagging、Boosting、Stacking等集成学习算法集成了多种机器学习，可提供全局的分类精度，且已有部分学者尝试利用集成学习算法对遥感影像进行分类，并取得了比机器学习算法更高的分类精度[15]。在集成学习算法中，Stacking算法的学习泛化能力较好，依靠学习数据集的特征来促进基于模型性能的输出加权，比简单投票和平均等组合规则更具优势[15]。基于此，本文基于GEE平台，采用Stacking算法提取区域土地覆盖信息，以弥补机器学习泛化能力不足和精度不高的缺陷。

1　研究区概况与数据来源

1.1　研究区概况

环洞庭湖区位于我国长江中游地区的湖南省东北部，行政范围主要包括岳阳市、常德市和益阳市，是我国重要的湿地分布区，也是我国的商品粮基地。研究区总体海拔低于40 m，以平原和低矮丘陵为主；属亚热带季风气候，雨热同期，年均气温为17℃，年均降水量为1 400 mm，最低和最高气温分别出现在1月（4.7℃）和7月（29℃）[16]。由于多年人类活动的影响（城市化和工业化），洞庭湖区的土地覆盖类型发生了巨大变化，景观结构日趋复杂，异质性逐渐增强。因此，准确、快速地获取洞庭湖区土地覆盖信息，能为国土资源管理与规划提供科学的基础数据。

1.2　数据来源与预处理

本文采用的遥感数据主要为2020年Sentinel-2A MSI多光谱影像（云量<5%）和Sentinel-1 SAR影像。Sentinel-2多光谱影像包括10 m空间分辨率的4个光谱波段（蓝光、绿光、红光和近红外波段）、20 m空间分辨率的6个波段（4个红边波段和两个短波红外波段）以及60 m空间分辨率的3个波段（沿海气溶胶、水汽和卷积波段），本文采用10 m和20 m空间分辨率的Sentinel-2影像光谱波段，其中20 m的光谱波段通过重采样使其空间分辨率变为10 m。本文采用的Sentinel-1 SAR影像主要为VV和VH后向散射系数，并将Sentinel-1后向散射系数重采样为10 m，与Sentinel-2影像空间分辨率一致。

实地验证数据主要通过外业调查的形式获取，获取时间为2020年5—8月，收集到的实地验证数据包括林地（32）、耕地（45）、泥沙滩地（26）、苔草（30）、芦苇（38）、建筑用地（35）6类。调查样地规格为10 m×10 m，并用GPS记录样地位置信息与土地覆盖类型。由于天气和其他因素影响，获取的实地验证数据有限，因此本文结合Google Earth高分辨率影像以目视判别的方式增加验证样本集。其他辅助数据包括GF-1号多光谱影像（2 m）和湖南省土地利用现状图，主要用于辅助选取分类过程中的样本训练。

2　研究方法

2.1　分类特征选择与组合

本文用于分类的遥感数据包括Sentinel-2多光谱影像、植被指数（NDVI、$NDVI_{re}$）以及Sentinel-1 SAR影像的后向散射系数。由于各地类的样本数量不同，在计算多类可分性时，权重也不同。为减少数据冗余，本文采用J_{Bh}距离获取用于土地覆盖信息提取的最佳数据组合。该方法基于Bhattacharyya原理，对可能性更高的类别赋予更高的权重[17]，计算公式为：

$$J_{Bh}=\sum_{i=1}^{N}\sum_{j>i}^{N}\sqrt{p(w_i)\times p(w_j)}\times JM^2(i,j) \quad (1)$$

式中，N为地类数量；$p(w_i)$与$p(w_j)$为土地利用类型的先验概率；i、j分别为土地利用类型；JM为不同地类之间的Jeffries-Matusita距离，范围在0～2之间，值越大，地类可分类性越大。

2.2　Stacking算法

本文利用Stacking算法对洞庭湖区土地覆盖信息进行提取。Stacking算法分为两层，第一层为多种机器学习算法，即基分类器，主要用于从训练数据集中提取主要特征，并为第二层的学习算法提供训练数据集[15]，通常为了达到令人满意的预测性能，基分类器应是多样化且性能足够好的；第二层为单一机器学习算法，即元分类器，将基分类器提取的特征作为训练数据进行训练和分类，得到最终的分类结果。以往研究表明，以RF作为元分类器比简单的逻辑回归分类器要好[15]，因此本文选择RF作为Stacking算法的元分类器。基分类器的种类越多，分类精度就越高，本文选取DT、最小距离、贝叶斯、SVM和RF等5种分类方法作为基分类器。

为了避免过度拟合和选择不同机器学习算法的最佳参数，通过k-fold交叉验证和分层抽样的方法将训练样本分为k个大小相同的集合，以$k-1$组为训练集，其余为测试集，再将测试集得出的结果作为元分类器的训练样本，元分类器的分类结果则为最终分类结果（图1）。

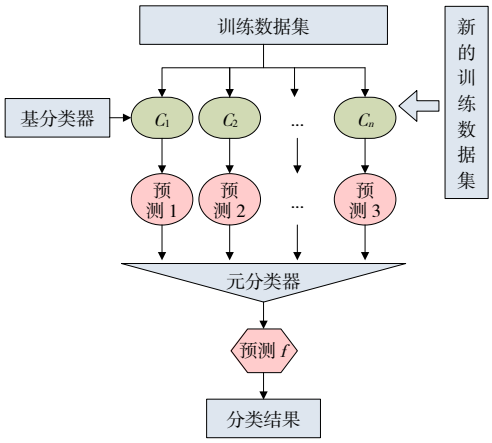


图1　Stacking集成算法

2.3　精度评价

本文基于实地调查数据，利用混淆矩阵对最后的分类结果进行精度验证。评价指标包括总体精度（*OA*）、Kappa系数、用户精度（*UA*）和生产者精度（*PA*）。

3　研究结果与分析

3.1　最佳分类特征组合

为减少数据冗余对分类精度造成的影响，本文利用 J_{Bh} 距离指数对最佳分类特征组合进行了筛选，结果见表1，可以看出，Sentinel-2多光谱影像结合Sentinel-1 SAR影像对于区分不同土地利用类型的可分离性具有较好的结果；但不是分类特征数量越多，J_{Bh} 距离值就越大，因为特征数量多在一定程度上会造成数据冗余，反而会降低土地利用类型间的可分离性。通过多次试验与计算得到，当分类特征组合为Sentinel-2多光谱影像+$NDVI_{re}$+Sentinel-1后向散射系数时，可得到最大的 J_{Bh} 距离（9.457）。

表1　不同分类特征组合下的 J_{Bh} 距离

分类特征组合	J_{Bh} 距离
Sentinel-2多光谱影像	8.237
Sentinel-2多光谱影像+NDVI	8.654
Sentinel-2多光谱影像+$NDVI_{re}$	8.889
Sentinel-2多光谱影像+NDVI+$NDVI_{re}$	8.661
Sentinel-2多光谱影像+NDVI+ Sentinel-1	9.028
Sentinel-2多光谱影像+$NDVI_{re}$+ Sentinel-1	9.457
Sentinel-2多光谱影像+NDVI+$NDVI_{re}$+Sentinel-1	9.364

3.2　土地利用分类结果与分析

本文根据最佳分类特征组合，基于GEE云平台，采用Stacking算法对洞庭湖区的LULC信息进行提取。为验证Stacking算法的优势，本文同时采用RF和SVM两种机器学习算法对研究区的土地利用类型进行分类（图2），可以看出，SVM的分类结果中许多林地和耕地像元被误分为建筑物，RF和Stacking算法的表现比SVM更好。为了更直观地展示各算法的分类细节信息，本文选择了典型的子区域进行详细说明（图3），可以看出，Stacking算法的分类结果具有更清晰的土地覆盖类型边界。比较分类结果和获得的Sentinel-2图像发现，SVM分类器在高度异质性区域不能准确区分多种土地类型，部分泥滩像元被误分为水体和耕地，部分苔草滩地像元被误分为耕地；RF分类器可以较准确地区分高度异质地区的多种土地类型，但在建筑物和泥沙滩地之间仍存在一些误分情况。总的来说，与其他两种算法相比，Stacking算法的分类结果具有更清晰的边界，对水体、泥沙滩地、苔草滩地、芦苇滩地和林地的区分更精准。

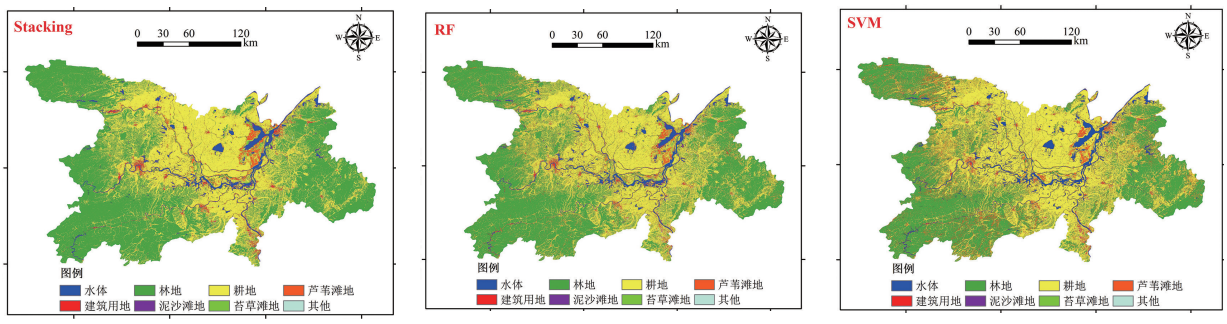


图2　不同分类算法下的洞庭湖区土地利用分类结果

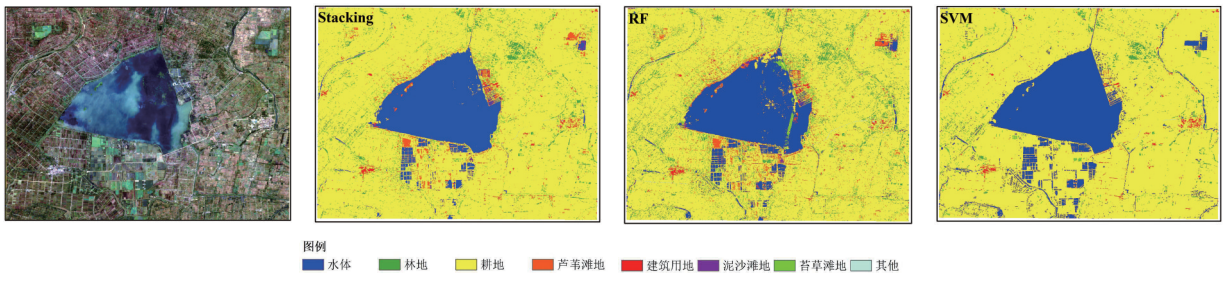


图3　3种分类算法下典型区域的土地利用分类结果

3种算法分类结果的准确性见表2，Stacking算法的*OA*和Kappa系数分别为88.53%和0.85，与SVM和RF相比，其*OA*分别提高了6.17%和4.27%，Kappa系数分别提高了0.10和0.04。虽然SVM和RF的*OA*尚可，但一些土地利用类型的*PA*和*UA*并不理想。Stacking算法提高了大部分土地利用类型的*PA*和*UA*，特别是对耕地、苔草滩地和芦苇滩地的识别精度。在高度异质地区的土地利用分类中，Stacking算法比RF和SVM等算法具有更好的分类性能。

表2　不同分类算法的精度评价结果/%

土地利用类型	SVM		RF		Stacking算法	
	PA	*UA*	*PA*	*UA*	*PA*	*UA*
水体	96.18	95.21	96.31	96.18	97.58	98.56
建筑用地	85.23	84.65	85.23	87.19	90.86	89.42
林地	84.62	85.13	86.59	86.54	89.25	90.32
泥沙滩地	72.73	70.36	74.31	73.92	80.53	83.41
耕地	86.34	85.49	86.53	86.99	87.49	89.05
苔草滩地	83.04	82.50	84.02	84.87	88.63	88.59
芦苇滩地	78.60	80.15	80.25	81.63	87.35	88.71
其他	83.59	82.28	84.13	87.12	88.65	89.58

注：SVM、RF、Stacking算法的*OA*分别为82.36%、84.26%、88.53%，Kappa系数分别为0.75、0.81、0.85。

4　结　语

本文以洞庭湖区为研究对象，基于Sentinel-2多光谱影像和Sentinel-1 SAR影像，通过分析研究区内不同LULC类型的可分离性，选取Sentinel-2多光谱影像、植被指数和Sentinel-1 SAR后向散射系数作为研究区土地覆盖分类的最佳特征组合；分别采用RF、SVM和Stacking算法提取了洞庭湖区的LULC类型信息。

1）多光谱影像结合SAR影像能较好地区分不同的土地利用类型，但不是分类特征数量越多，精度越高，因为特征数量多在一定程度上会造成数据冗余，反而会降低土地利用类型间的可分离性。

2）Stacking算法在土地利用信息提取的过程中具有更好的分类精度。集成学习能够综合不同机器学习的特点，在一定程度上提高分类的准确性和稳定性。

尽管Stacking算法具有很强的泛化能力，但其提取遥感图像特征的能力还需进一步提高。因此，基于深度学习特征提取能力和Stacking算法的泛化能力，将深度学习与Stacking结合起来以提高LULC分类的准确性和精度是后续研究的重点。

参考文献

[1] 孙姣姣，王琦，郑洁，等. 大英县2002-2017年土地利用分类动态分析[J]. 地理空间信息，2022，20(1)：56-60

[2] 熊晓峰，张德州. 长三角地区地表覆盖与城市热岛时空演变分析[J]. 地理空间信息，2021，19(11)：71-74

[3] 孙晓莉，张顺安，刘应芳，等. 基于GIS的土地利用动态变化及其驱动力分析[J]. 地理空间信息，2021，19(11)：101-103

[4] 刘睿，王志勇，高瑞. 时序SAR影像的干旱地区土地利用分类应用[J]. 测绘科学，2021，46(10)：90-97

[5] 秦海超，骆焕成，王笑. 基于GEE和Sentinel-2影像的临沂市土地利用/覆被分类信息提取[J]. 工程勘察，2021，49(8)：69-73

[6] 熊皓丽，周小成，汪小钦，等. 基于GEE云平台的福建省10 m分辨率茶园专题空间分布制图[J]. 地球信息科学学报，2021，23(7)：1 325-1 337

[7] 刘凯，彭力恒，李想，等. 基于Google Earth Engine的红树林年际变化监测研究[J]. 地球信息科学学报，2019，21(5)：731-739

[8] 牟晓莉，李贺，黄翀，等. Google Earth Engine在土地覆被遥感信息提取中的研究进展[J]. 国土资源遥感，2021，33(2)：1-10

[9] LONG X R，LI X Y，LIN H，et al. Mapping the Vegetation Distribution and Dynamics of a Wetland Using Adaptive-stacking and Google Earth Engine Based on Multi-source Remote Sensing Data[J]. International Journal of Applied Earth Observation and Geoinformation，2021(102)：102 453

[10] 熊元康，张鸿辉，梁宇哲，等. 基于Google Earth Engine与多源遥感数据的土地覆盖变化监测[J]. 地理空间信息，2021，19(9)：73-78

[11] 毛丽君，李海涛，薛晓明，等. 基于GEE平台的国家公园土地覆盖变化遥感检测方法构建[J]. 南京林业大学学报(自然科学版)，2021，46(2)：213-220

[12] 文可，姚焕玫，黄以，等. 基于GEE的广西北部湾沿海水产养殖池塘遥感提取[J]. 农业工程学报，2021，37(12)：280-288

[13] 姜伊兰，陈保旺，黄玉芳，等. 基于Google Earth Engine和NDVI时序差异指数的作物种植区提取[J]. 地球信息科学学报，2021，23(5)：938-947

[14] 付甜梦，张丽，陈博伟，等. 基于GEE平台的海岛地表覆盖提取及变化监测：以苏拉威西岛为例[J]. 遥感技术与应用，2021，36(1)：55-64

[15] CAI Y T，LI X Y，ZHANG M，et al. Mapping Wetland Using the Object-based Stacked Generalization Method Based on Multi-temporal Optical and SAR Data[J]. International Journal of Applied Earth Observation and Geoinformation，2020(92)：102 164

[16] 陈磊士，高霞霞，廖玉芳，等. 基于CART决策树的高分二号洞庭湖区湿地分类方法[J]. 测绘通报，2021(6)：12-15

[17] HAO P，WANG L，NIU Z，et al. The Potential of Time Series Merged from Landsat5 TM and HJ-1 CCD for Crop Classification：A Case Study for Bole and Manas Counties in Xinjiang，China[J]. Remote Sensing，2014(6)：7 610-7 631