

引用格式:何云,黄翀,李贺,等.基于Sentinel-2A影像特征优选的随机森林土地覆盖分类[J].资源科学,2019,41(5):992-1001.  
[He Y, Huang C, Li H, et al. Land-cover classification of random forest based on Sentinel-2A image feature optimization[J].  
Resources Science, 2019, 41(5): 992-1001.] DOI: 10.18402/resci.2019.05.15

# 基于Sentinel-2A影像特征优选的随机森林土地覆盖分类

何云<sup>1,2</sup>, 黄翀<sup>1</sup>, 李贺<sup>1</sup>, 刘庆生<sup>1</sup>, 刘高焕<sup>1</sup>, 周振超<sup>3</sup>, 张晨晨<sup>1,2</sup>

(1. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室, 北京 100101;  
2. 中国科学院大学, 北京 100049; 3. 吉林大学 地球探测科学与技术学院, 长春 130026)

**摘要:**中南半岛地处热带、亚热带地区,由于水热条件适宜,植被生长旺盛,土地利用强度高,地表覆盖类型的光谱特征时空变异复杂,使用传统的基于光谱特征的遥感分类精度难以保证。Sentinel-2A卫星遥感数据具有较丰富的光谱波段和较高的空间分辨率,为土地覆盖遥感分类提供了多维特征空间。但多维特征参与分类容易造成信息冗余,从而导致分类速度和精度降低。因此,如何充分利用Sentinel-2A数据丰富的光谱和空间信息,并通过高维特征空间降维进行特征优选对于提高分类精度具有重要意义。本文以中南半岛典型地区土地覆盖分类为例,利用Sentinel-2A多波段光谱特征,归一化植被指数(NDVI)、比值植被指数(RVI)、差值植被指数(DVI)、归一化水体指数(NDWI)等指数特征以及对比度、相关性、能量、均值、熵等纹理特征,在随机森林模型框架下,采用平均不纯度减少方法对不同特征在土地覆盖分类中的重要程度进行识别;利用袋外(OOB)误差方法,对重要特征组合进行了优选;利用优选特征进行随机森林土地覆盖分类,并与原始随机森林分类结果进行对比。结果表明:Sentinel-2A影像的光谱特征和纹理特征在土地覆盖分类中具有较为重要的作用,光谱特征中短波红外、可见光、植被红边波段重要性较大,纹理特征中均值、能量法重要性较高。选择重要性排名前9位的特征参与分类时,OOB精度达到最高;继续增加特征会使模型复杂度过高,容易发生过拟合而使得分类精度不增反降。通过特征优选高效利用了Sentinel-2A丰富的光谱和纹理信息,其总体分类精度达87.53%,Kappa系数达0.8461,优于原始随机森林方法,一定程度上提高了热带亚热带地区复杂土地覆盖分类精度。

**关键词:** Sentinel-2A; 特征优选; 随机森林; 土地覆盖分类; 袋外(OOB)误差方法; 中南半岛; 泰国穆河流域

DOI: 10.18402/resci.2019.05.15

## 1 引言

遥感数据以其成本低、效率高的优势已成为土地覆盖分类的主要数据源<sup>[1]</sup>,其中,以Landsat中等分辨率光学遥感数据为代表、基于影像光谱特征差异的土地覆盖分类得到广泛的应用<sup>[2]</sup>。然而,在热带、亚热带的中南半岛地区,一年内大部分时期水热条件充足,植被生长旺盛,自然植被和栽培作物光谱特征差异不明显,分类误差较大<sup>[3]</sup>。与此同时,作物种植制度灵活,一年两熟、一年三熟多有发

生。由于作物没有固定的生长季,不同作物在不同的时间和空间上光谱特征变异复杂<sup>[4]</sup>。因此,对于中南半岛地区复杂土地覆盖类型来说,仅依赖有限光谱特征的遥感分类精度难以保证。2015年6月欧空局(ESA)成功发射全球环境与安全监测的第2颗卫星Sentinel-2A,该卫星具有覆盖宽、时空分辨率高、光谱特征丰富等优势,为土地覆盖分类提供了新的数据支持<sup>[5]</sup>。Immitzer等<sup>[6]</sup>利用Sentinel-2A影像的光谱特征,使用随机森林进行农作物和树种分

收稿日期:2018-08-16 修订日期:2019-02-26

基金项目:国家自然科学基金国际(地区)合作与交流项目(41661144030);国家自然科学基金项目(41471335)。

作者简介:何云,女,湖南益阳人,硕士生,主要从事生态遥感研究。E-mail: cug\_heyun@163.com

通讯作者:黄翀, E-mail: huangch@lreis.ac.cn

2019年5月

类,评估了 Sentinel-2A 不同光谱特征在作物提取和树种分类中的重要性,并发现 Sentinel-2A 红边波段在植被分类中发挥着重要的作用;时丽娜等<sup>[7]</sup>利用 Sentinel-2A 影像的光谱特征、冰崩的专题属性和数字高程模型,采用基于规则和最近邻分析的面向对象分类方法,对西藏阿里冰崩范围进行了提取;Ru-joiu-Mare 等<sup>[8]</sup>利用 Sentinel-2A 影像的光谱特征,使用最大似然法和支持向量机分别对土地异质性较低和较高区域进行土地覆盖分类,取得了不错的效果。但是,在利用 Sentinel-2A 数据丰富的光谱和纹理特征的同时,由于不同特征间存在相关性,所有特征参与分类将致使信息冗余,从而导致分类精度降低、分类速度下降。因此,如何在高维特征空间中选择最优特征集合参与分类,在降低模型复杂度的同时保证分类精度显得尤为重要。随机森林作为一种集成学习方法,具有高效、准确度高等特点,在中高分辨率影像分类中不仅能保证较高的精度也能保证较快的速度,且具有特征选择的能力<sup>[9-11]</sup>。

中南半岛位于南海周边,与中国陆海相连,是连通“一带一路”的重要桥梁和纽带。本文以位于

中南半岛的泰国穆河流域典型地区土地覆盖分类为例,在随机森林模型框架下,对 Sentinel-2A 数据不同波段、指数、纹理特征在土地覆盖分类中的重要性进行评估,对最优特征集合进行优选,在此基础上,探讨了基于特征优选的随机森林分类方法的优势和适用性,为热带、亚热带地区复杂土地覆盖遥感分类提供技术支持。

## 2 数据来源与研究方法

### 2.1 研究区域概况

本文选取中南半岛穆河流域的东南部区域(图1)来开展案例研究。穆河流域位于 107°00'E—111°51'E, 14°02'N—16°29'N, 包括泰国的 Nakhon Ratchasima(呵叻府)、Buri Ram(武里南)、Surin(素林)、SiSa Ket(四色菊)、Ubon Ratchathani(乌汶府)等十府,流域面积约 8.2 万 km<sup>2</sup>。穆河流域属于湿润亚热带季风气候,干湿季明显,常年气温不低于 18℃,平均年降水量约 1300~1500 mm。穆河流域南部山区分布有大范围的林地,水系发达,土地覆盖类型主要包括林地、旱地、水田、水体、城建用地、裸地。穆河流域以农业为主要产业,由于水热条件适宜,其

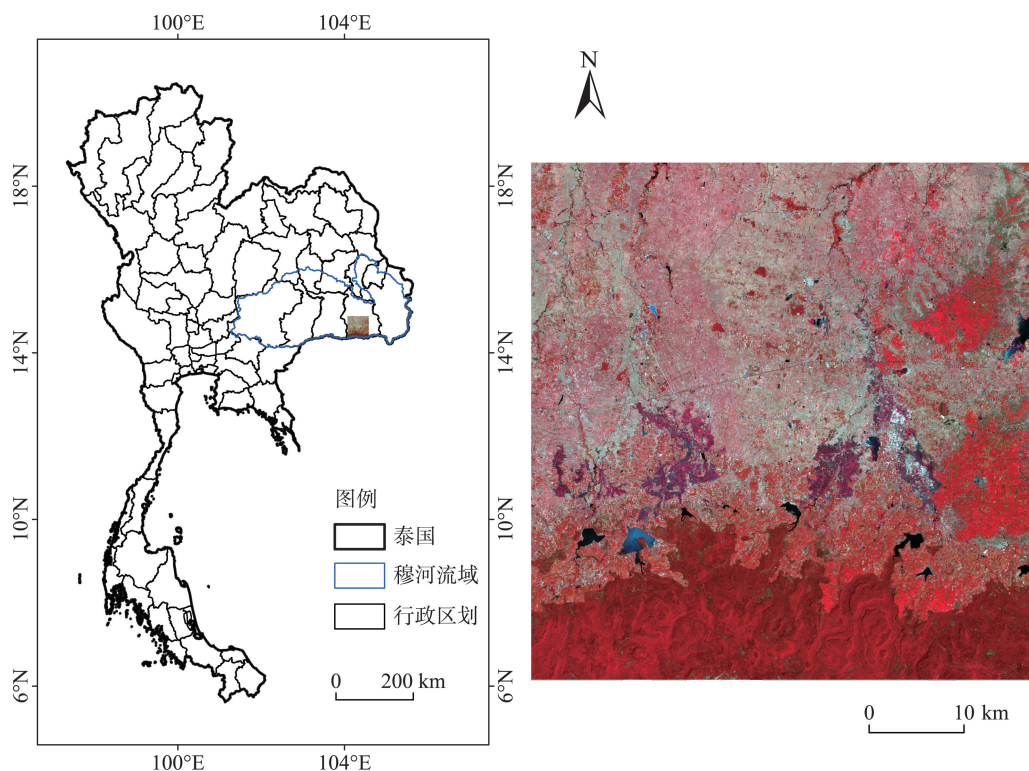


图1 穆河流域地理位置

Figure 1 Location of the Mun River Basin

农业种植方式灵活,受季节限制较小,在有灌溉措施条件下可达到一年多熟。根据研究区土地覆盖情况,同时考虑到 Sentinel-2A 影像的分辨率,将研究区的土地覆盖类型在一级类上分为耕地、林地、城镇与建设用地、水体以及裸地,进一步对该区域的主要覆盖类型耕地细分为水田和旱地。

## 2.2 数据来源与预处理

Sentinel-2 是全球环境与安全监测的第 2 颗卫星,卫星携带了一枚多光谱成像仪,具有 13 个光谱波段,包括 10 m 分辨率的 3 个可见光波段和 1 个近红外波段,20 m 分辨率的 3 个红边波段、1 个近红外波段和 2 个短波红外波段,以及 60 m 分辨率的海岸、水汽和卷积云波段。Sentinel-2A 卫星幅宽达 290 km,时间分辨率达 10 天<sup>[5]</sup>。研究所用的影像是从美国地质调查局(USGS)网站(<https://glovis.usgs.gov/>)下载。选择研究区无云且质量良好(2017 年 2 月 13 日)的一幅 Sentinel-2A 影像,该数据为经过正射校正和亚像元级几何精校正后的 L1C 大气表观反射率产品,因此只需进行大气校正。本文使用欧洲航空局 ESA 提供的 Sen2cor 插件对 Sentinel-2A 影像进行大气校正,使用 SNAP 软件将 20 m 分辨率处的 6 个波段使用最近邻法重采样到 10 m 分辨率,结合 4 个原始 10 m 分辨率的波段共得到 10 个 10 m 分辨率的波段,并转换成 Envi 标准格式,在 Envi 里进行裁剪。本文验证数据为 2015 年穆河流域土地覆盖数据,来源于泰国土地利用规划局和土地开发部。

## 2.3 分类流程

为充分利用 Sentinel-2A 数据提供的丰富的光谱信息和空间信息,首先对影像的植被指数及纹理特征进行了提取。为解决高维海量数据信息冗余问题,利用随机森林方法,对多个特征的重要性进行评估,进一步通过特征优选,筛选出对土地覆盖分类起关键作用的特征组合进行分类,并与特征优选前的分类结果对比,以识别在该区域的土地覆盖分类中最优特征组合,提高分类精度。本文的技术流程如图 2 所示。

### 2.3.1 特征提取

植被指数和水体指数在反映土地覆盖类型时比单一波段更为稳定,结合植被指数和水体指数进

行分类可在一定程度上改善影像分类的精度<sup>[12]</sup>,因此本文选择具有代表性的 3 种植被指数和 1 种水体指数,分别为归一化植被指数(NDVI)、比值植被指数(RVI)、差值植被指数(DVI)以及归一化水体指数(NDWI),其中 NDVI(式 1)为遥感影像分类中应用最广泛的植被指数,是植被生长状态和植被分布密度的最佳指数因子<sup>[13]</sup>,RVI(式 2)适应于高密度植被覆盖区域植被监测,对绿色植被与非绿色植被(如裸地和水体)具有良好的区分度<sup>[14]</sup>;DVI(式 3)对土壤背景变化敏感,适用于植被覆盖率较低或植被发育早中期的植被监测<sup>[15]</sup>,NDWI(式 4)广泛应用于提取影像中的水体信息<sup>[16]</sup>。

$$NDVI = (b_{nir} - b_r) / (b_{nir} + b_r) \quad (1)$$

$$RVI = b_{nir} / b_r \quad (2)$$

$$DVI = b_{nir} - b_r \quad (3)$$

$$NDWI = (b_g - b_{nir}) / (b_g + b_{nir}) \quad (4)$$

式中: $b_{nir}$ 表示近红外波段, $b_r$ 表示红波段, $b_g$ 表示绿波段。

纹理特征能反映丰富的地物信息,在中高分辨率影像分类中已被证实能提高影像分类的精度<sup>[17,18]</sup>,

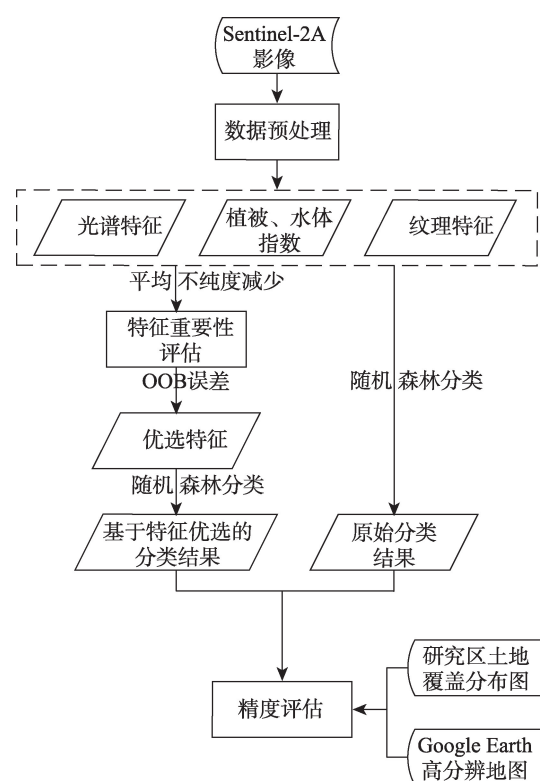


图 2 研究技术路线图

Figure 2 Technical roadmap of the study



2019年5月

本文使用灰度共生矩阵方法提取影像的纹理特征。首先对原始影像进行主成分分析,其中第1主成分累积贡献率达到72.36%,第2主成分累积贡献率达到95.58%,因此取第1、第2个主成分来提取纹理特征。为较好地反映影像的粗纹理和细纹理,通过多次实验对比分析,设置滑动窗口大小为5,步长为1,利用灰度共生矩阵提取了每个主成分的均值(Mean)、对比度(Contrast)、熵(Entropy)、能量法(Energy)、相关性(Correlation)共5个参数,共得到10个纹理特征。

因此本文共选取了10个光谱特征、4个指数特征、10个纹理特征共24个特征。

### 2.3.2 特征重要性评估与优选

本文特征重要性评估和特征选择在Python平台下编程实现。①根据影像本身特征和2015年研究区土地覆盖图选取多边形训练样本(图3),并转换成TIFF格式,共选取了21166个训练像素点。其中包含4327个林地像素点,2266个旱地像素点,5234个水田像素点,7845个水体像素点,422个城建用地像素点,1072个裸地像素点。②利用栅格空间数据转换库(Geospatial Data Abstraction Library, GDAL)读取训练样本数据,通过调用Scikit-learn库中的随机森林分类器建立随机森林模型,并使用平均不纯度减少的方法计算模型中各特征的重要性。③使用OOB误差进行模型评估以确定模型最优特征数量。对得到的特征重要性结果排序,按重要性从高到低的顺序依次选择特征,第1次选择重

要性列首位的特征,第2次选择重要性列前2位的特征,依此类推,得到24个特征组合生成的随机森林模型,通过计算不同特征组合模型的OOB误差,综合考虑模型精度和计算复杂度来确定最优特征个数。

### 2.3.3 随机森林分类

随机森林由美国科学家Leo Breiman于2001年提出<sup>[19]</sup>,结合了Bagging集成学习理论与随机子空间方法,是以决策树为基本分类器的一种集成学习方法。随机森林是一种非参数分类与回归方法,不需要先验知识,易于使用;以决策树为基础分类器,能保证良好的精度;基于Bagging集成学习理论,能容忍一定的噪声和异常值,能并行化处理高维海量数据,是一种高效的机器学习算法。

随机森林构建的基本过程为:①首先通过bootstrap的方式从原始训练样本中有放回地随机抽取样本,假设原始训练样本集共有 $N$ 个样本,每个样本具有 $M$ 个特征,每次有放回地从中抽取 $N$ 个样本,那么某个样本未被抽中的概率为 $(1-1/N)^N$ ,当 $N$ 很大时,这个值趋向于 $1/e \approx 1/3$ ,即抽取时大概有1/3的原始样本未被抽取到,这部分样本称为袋外(Out of Bag, OOB)数据,可使用这部分样本来估计误差,称为袋外(OOB)误差。②对 $N$ 个样本进行训练得到一个决策树模型,在决策树的每个结点处随机选取 $m(m < M)$ 个特征,使用信息熵、信息增益或者基尼指数来选择特征进行结点分裂;由于随机森林是一种集成学习方法,不容易出现过拟合现象,因此在构

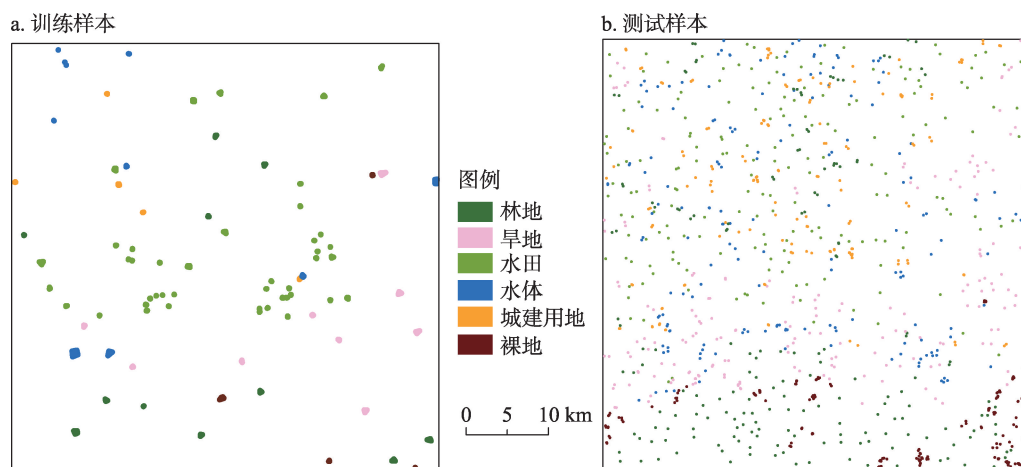


图3 训练及测试样本分布图

Figure 3 Distribution of the training and testing samples



建决策树时不需要进行剪枝<sup>[20,21]</sup>。③重复①、②,通过 $k$ 次样本抽取和样本训练得到 $k$ 个决策树模型。④最后采用集成学习理论将 $k$ 个决策树进行线性组合,其中每个决策树占相等的权重。当输入一个待分类样本时,分类结果由每个决策树的结果通过多数表决的方式投票决定。

随机森林中决策树构建的一个关键步骤是结点分裂时的特征选择。理想状态下,由最优的分裂特征得到的每个子节点是纯的,即每个子节点中的样本属于同一类,可以利用基尼指数来度量样本集合的不纯度(不确定性程度),基尼指数表示在集合中一个随机样本被分错的概率,基尼指数越小表示随机选中的样本被分错的概率越小,集合的纯度越高;反之,集合越不纯。集合 $D$ 的基尼指数的定义如下:

$$Gini(D) = \sum_{b=1}^B p_b(1-p_b) = 1 - \sum_{b=1}^B p_b^2 \quad (5)$$

式中: $B$ 为训练样本中样本种类数, $p_b$ 表示集合 $D$ 中随机选中的样本属于类别 $b$ 的概率, $(1-p_b)$ 表示样本被分错的概率。如果样本集合 $D$ 根据特征 $A$ 是否取某一可能值 $a$ 被划分为 $D_1$ 和 $D_2$ 两个部分,则在特征 $A$ 的条件下,集合 $D$ 的基尼指数为:

$$Gini(D, A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \quad (6)$$

式中: $|D|$ 表示集合 $D$ 中的样本数, $|D_1|$ 表示集合 $D_1$ 中的样本数, $|D_2|$ 表示集合 $D_2$ 中的样本数。可以看出,在随机森林中,若通过某特征划分后平均基尼指数减少的程度越大,即通过该特征划分后集合变纯的程度越大,则可以认为该特征的分类能力越强,在模型中的重要性越大,反之亦然。这种特征重要性评估方法称为平均不纯度减少,平均不纯度减少的定义为:

$$\Delta Gini = \frac{\sum_{n=1}^K [Gini_n(D) - Gini_n(D, A)]}{K} \quad (7)$$

式中: $K$ 为随机森林中决策树的个数, $Gini_n(D)$ 表示第 $n$ 棵决策树 $Gini(D)$ 划分前集合 $D$ 的基尼指数, $Gini_n(D, A)$ 表示第 $n$ 棵决策树 $Gini(D, A)$ 通过特征 $A$ 划分后集合 $D$ 的基尼指数。

本文在Python平台下,利用GDAL库读取训练样本数据,通过调用scikit-learn库中的随机森林分类器,分别利用全部24个原始特征和优选特征建立

随机森林模型,对Sentinel-2A影像进行分类,并对两者的分类结果进行对比分析。

### 2.3.4 分类精度比较

为比较2种模型的优劣,对2种方法得到的土地覆盖分类结果进行精度评估。参考2015年研究区土地覆盖分类图随机选取测试样本,并使得选取的测试样本尽量均匀地分布在研究区。由于所选影像时相为2018年,而研究区作物种植制度灵活,为进一步保证测试样本的正确性,将选取的测试样本导入Google Earth,参考Google Earth高分辨率地图剔除错误测试样本(图3),本文共选取了962个测试样本,其中包括178个林地样本、180个旱地样本、188个水田样本、172个水体样本、156个城镇与建设用地样本、176个裸地样本。

## 3 结果分析

### 3.1 特征重要性评估

利用平均不纯度减少的方法计算模型中各特征重要性结果如表1所示。从表1中可以看出,重要性排前9的特征中,Sentinel-2A影像原始光谱特征占6个,纹理特征占2个,植被指数占1个,说明Sentinel-2A影像光谱特征在土地覆盖分类中发挥着重要的作用,Sentinel-2A影像纹理特征对土地覆盖分类具有较大的贡献度;而Sentinel-2A植被、水体指数在土地覆盖分类中重要性则相对较低。

在Sentinel-2A影像10个光谱特征中,可见光中的蓝、绿、红波段、短波红外的2个波段重要性高,达到14.51%、8.48%、8.35%、8.75%、7.43%,分别列第

表1 特征重要性评估结果

Table 1 Feature importance assessment result

特征	重要性/%	特征	重要性/%
Blue	14.51	NIR_2	2.84
SWIR_1	8.75	RVI	2.04
Green	8.48	Energy_2	2.01
Mean_1	8.41	NIR_1	1.93
Red	8.35	NDWI	1.57
SWIR_2	7.43	NDVI	0.99
Red edge_1	6.58	Contrast_2	0.84
DVI	6.46	Contrast_1	0.54
Mean_2	5.61	Entropy_1	0.33
Red edge_3	5.44	Entropy_2	0.22
Red edge_2	3.51	Correlation_2	0.13
Energy_1	2.90	Correlation_1	0.13

2019年5月

1、第3、第5、第2、第6;植被红边的3个波段重要性较高,达到6.58%、3.51%、5.44%,分别列第7、第11、第10;近红外的2个波段重要性相对较高,但重要性均在1%以上;说明 Sentinel-2A 中可见光、短波红外波段对土地覆盖分类贡献度较大,植被红边波段在土地覆盖分类中具有一定的作用,而近红外波段对土地覆盖分类的贡献度相对较小。在4个指数特征中, $DVI$ 重要性最高,达到6.46%, $RVI$ 、 $NDWI$ 、 $NDVI$ 重要性相对较低,考虑到所选影像日期为2017年2月13日,为旱季,研究区水田和旱地覆盖面积大,而此时水稻和旱地作物大多处于种植早期或未种植,植被覆盖率较低, $DVI$ 对土壤背景变化敏感,适用于植被覆盖率较低或植被发育早中期的植被监测,因此 $DVI$ 的重要性较高。在10个纹理特征中,Sentinel-2A 影像纹理特征中均值的重要性均较高,达到8.41%和5.61%,分别列第4和第9;Sentinel-2A 影像纹理特征中能量法具有一定的重要性,均在2%以上;Sentinel-2A 影像纹理特征中对比度、熵、相关性的重要性较低;可以看出 Sentinel-2A 影像纹理特征中的均值、能量法对土地覆盖分类的贡献度较大,而对对比度、熵、相关性在土地覆盖分类中发挥的作用较小。

### 3.2 最优特征组合

对24个不同特征组合模型的OOB误差分析如图4所示。从图中可以看出,特征个数从1增加到9,OOB精度有逐渐增加的趋势,在特征个数为9

时,OOB精度达到最高值0.9614;特征个数从9增加到24,OOB精度有微小的降低趋势,但变化不大;特征个数过少,会导致分类精度不高;特征个数过多,模型复杂度过高,使得运行时间过长,并且容易发生拟合而使分类精度不增反降。因此,本文选择特征重要性列前9的特征作为基于特征优选的随机森林模型的输入特征。

### 3.3 基于特征优选的随机森林与原始随机森林分类结果的比较

分别利用原始随机森林模型和基于特征优选的随机森林模型得到的土地覆盖分类结果如图5所示。从目视效果来看,水田、旱地、林地、水体分类图斑较为规整:因旱地常与水田镶嵌分布,相似的光谱特征导致旱地易与水田混淆,且难以与林地区分;而植被指数、纹理特征的引入较好地地区分了旱地、水田和林地,在一定程度上缓解了“椒盐现象”;

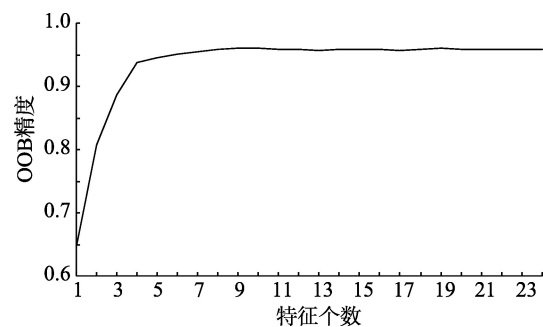


图4 不同特征组合模型 OOB 精度

Figure 4 Out-of-bag (OOB) accuracy of different feature combination models

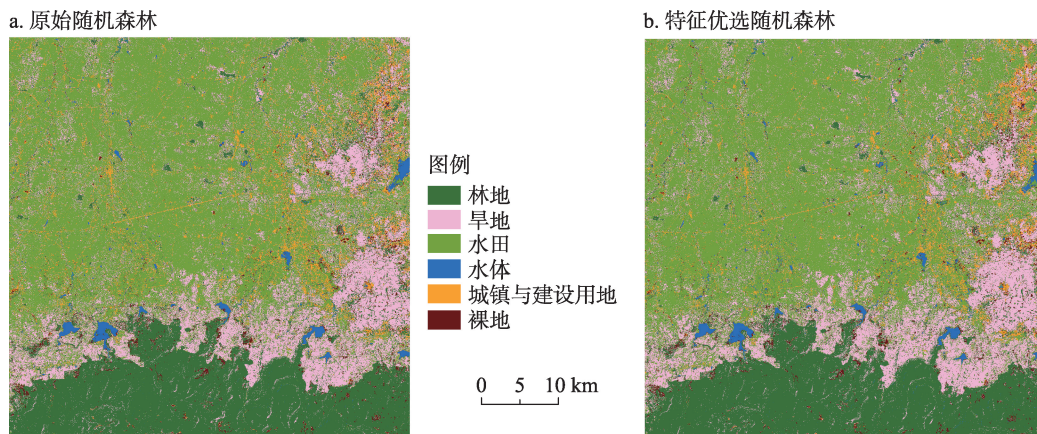


图5 基于2种方法生成的土地覆盖分类图

Figure 5 Land-cover classification map based on two methods

水体与其他地物差异较大,且内部异质性小,因此分类图斑规整。城镇与建设用地分类图斑相对破碎,城镇常与园地、草地等城市绿化带镶嵌分布,内部结构相对多样,造成了分类结果的破碎。

2种方法得到的各土地覆盖类型的制图精度和用户精度如表2所示。从表2中可以看出,基于特征优选随机森林分类结果中,林地、水体、城镇与建设用地的制图精度和用户精度均在85%以上,旱地、水田、裸地的制图精度和用户精度均在80%左右。汇总其混淆矩阵(表3),可以看出旱地易与水田和林地混淆,所选影像位于旱季,水田、旱地部分未种植,光谱间的相似性使得未种植作物的旱地易与未种植水稻的水田混淆,已种植作物的旱地易与林地混淆;水田易与旱地以及城镇与建设用地混淆,研究区水田大多为未种植水稻的裸露地块,因此易与城镇与建设用地混淆;研究区南部裸地与林地、旱地交错分布,导致裸地易与林地、旱地混淆,也易与城镇与建设用地混淆。

从模型精度和分类精度对比可以看出,基于特征优选的随机森林方法分类结果较好(表2)。与基于原始随机森林方法分类结果相比,基于特征优选的随机森林方法分类结果中,林地、旱地、水田、城镇与建设用地的用户精度和制图精度均有所提高,其中用户精度分别提高了1.01%、6.25%、4.56%、1.15%,制图精度分别提高了3.94%、3.33%、4.25%、7.69%。水体的用户精度没有发生变化,制图精度分别下降了0.58%;裸地的用户精度提高了6.79%,制图精度分别下降了2.27%。可以看出,基于特征优选的随机森林方法在对林地、旱地、水田、城镇与建设用地分类时分类精度优于原始随机森林方法,

表2 2种方法各类的制图精度和用户精度

Table 2 Producer's accuracy and User's accuracy for the two methods

类别	原始随机森林		基于特征优选随机森林	
	制图精度/%	用户精度/%	制图精度/%	用户精度/%
林地	88.76	84.04	92.70	85.05
旱地	81.11	81.11	84.44	87.36
水田	82.98	74.29	87.23	78.85
水体	90.70	100.00	90.12	100.00
城镇与建设用地	79.49	84.93	87.18	86.08
裸地	81.82	87.80	79.55	94.59

表3 基于特征优选随机森林方法的土地覆盖分类结果混淆矩阵

Table 3 Confusion matrix of land-cover classification result based on feature-optimized random forest

	林地	旱地	水田	水体	城镇与建设用地	裸地	总计
林地	165	10	2		1		178
旱地	10	152	14		2	2	180
水田	2	10	164		12		188
水体	5		8	155	2	2	172
城镇与建设用地			20		136		156
裸地	12	2			4	70	88
总计	194	174	208	155	157	74	962

在对水体、裸地分类时精度与原始随机森林方法差别不大。基于特征优选的随机森林方法保留了在土地覆盖分类中重要的光谱特征、植被、水体指数、纹理特征,去除了冗余成分,能有效地区分自然植被和栽培作物,对旱地作物和水田作物的区分也具有优势。通过计算得到基于原始随机森林 Sentinel-2A 影像分类的总体精度为84.41%,Kappa系数为0.8110,而基于特征优选的随机森林 Sentinel-2A 影像分类的总体精度为87.53%,Kappa系数为0.8487。可以看出,基于特征优选的随机森林方法进行土地覆盖分类的总体精度高于原始随机森林方法,说明具有丰富光谱特征和空间纹理特征的 Sentinel-2A 影像在热带亚热带地区土地覆盖分类中具有较好的适用性。

4 讨论与结论

4.1 讨论

随着遥感技术的不断进步,遥感数据获取能力越来越强,未来的遥感数据源必将在时间、空间以及光谱分辨率上得到极大提高。土地覆盖遥感分类也将更多地采用多特征进行分类。然而,在多个特征之间存在一定的相关性,所有特征参与分类将致使信息冗余,从而导致分类精度降低、分类速度下降,因此选择最优特征集合参与分类,在未来高分辨率遥感影像土地覆盖分类中具有重要意义。常用的降维或特征优选方法包括主成分分析<sup>[22,23]</sup>、平均精度下降<sup>[6]</sup>、ReliefF 算法<sup>[24]</sup>等,不同的方法适用



2019年5月

于不同的数据集和场景,尚未形成统一的定论。本文基于随机森林模型中的平均不纯度减少方法进行特征重要性评估和确定最优特征。对于一个决策树森林来说,平均不纯度减少方法对于处理 Sentinel-2A 这种较高分辨率的遥感多维特征空间时简单高效。此外,在本文优选的 Sentinel-2A 影像 9 个重要特征可不同程度地提高对地物的区分能力。光谱特征中的可见光、短波红外波段发挥了重要的作用,而 Sentinel-2A 的植被红边波段也表现出较高价值,这与 Immitzer 等<sup>[6]</sup>得到的结果一致。纹理特征具有较高的重要性,研究区水田、旱地分布广泛,具有较为明显的纹理特征,因此能在一定程度上加大地物间的区分度。

本文蓝光波段的重要性偏高,可能受到其他关联特征的影响,这与 Immitzer 等<sup>[6]</sup>得到的结果一致,即如果多个特征存在关联,某个特征被选定后,会引起其他相关特征的重要性降低。通常植被指数和水体指数在湿地类型提取中重要性较高<sup>[26]</sup>,而在本文中可能由于所选影像处于旱季, Sentinel-2A 植被指数和水体指数差异性较小,因而重要性较低。

## 4.2 结论

本文以 Sentinel-2A 影像为数据源,提取了影像的光谱特征、植被、水体指数和纹理特征,使用平均不纯度减少和 OOB 误差评估了不同特征在泰国土地覆盖分类中的重要性,在此基础上进行特征选择并建立基于特征优选的随机森林模型,对泰国典型地区进行土地覆盖分类。通过与特征优选前的分类结果对比,分析了基于特征优选的随机森林分类方法在 Sentinel-2A 土地覆盖分类中的适用性。得到以下结论:

(1) Sentinel-2A 影像光谱特征中,可见光波段、短波红外波段在土地覆盖分类中的重要性较高,植被红边波段则具有一定的重要性;纹理特征中的均值、能量法在土地覆盖分类中重要性较高; Sentinel-2A 影像水体、植被指数中 DVI 的重要性相对较高,其他指数则重要性较低。

(2) 在全部 24 个特征中,选择重要性前 9 位特征参与分类时, OOB 精度达到最高值 0.9614;此后随着特征个数的增加, OOB 精度有微小的降低趋

势,特征个数过多,模型复杂度过高,容易发生过拟合而使得分类精度不增反降。

(3) 基于特征优选的随机森林方法在 Sentinel-2A 影像穆河流域土地覆盖分类中具有较高的分类精度,总体分类精度达到 80% 以上;基于特征优选的随机森林方法能有效地区分自然植被和栽培作物,对旱地作物和水田作物的区分也具有优势,与原始随机森林方法相比,总体分类精度提高了 2.91%, kappa 系数提高 0.0351,而且降低了计算复杂度,提高了分类速度,在 Sentinel-2A 影像土地覆盖分类中具有较好的适用性。

由于研究区地处热带季风区域,多云多雨、种植制度灵活使得同物异谱或同谱异物现象严重,有些土地覆盖类型之间错分度较高。进一步研究,可结合 Sentinel-2A 影像的高时间分辨率优势,利用地物光谱的时间变化特征来区分单景影像特定时刻光谱相似的地物,以进一步提高土地覆盖分类的精度。

## 参考文献(References):

- [1] 张景,姚凤梅,徐永明,等. 基于 MODIS 的土地覆盖遥感分类特征的评价与比较[J]. 地理科学, 2010, 30(2): 248-253. [Zhang J, Yao F M, Xu Y M, et al. Comparison and evaluation of classification features in land cover based on remote sensing[J]. Scientia Geographica Sinica, 2010, 30(2): 248-253.]
- [2] 宋军伟,张友静,李鑫川,等. 基于 GF-1 与 Landsat-8 影像的土地覆盖分类比较[J]. 地理科学进展, 2016, 35(2): 255-263. [Song J W, Zhang Y J, Li X C, et al. Comparison between GF-1 and Landsat-8 images in land cover classification[J]. Progress in Geography, 2016, 35(2): 255-263.]
- [3] Guan X, Huang C, Liu G, et al. Mapping rice cropping systems in Vietnam using an NDVI-based time-series similarity measurement based on DTW distance[J]. Remote Sensing, 2016, doi: 10.3390/rs8010019.
- [4] 管续栋,黄翀,刘高焕,等. 基于 DTW 距离的时序相似性方法提取水稻遥感信息: 以泰国为例[J]. 资源科学, 2014, 36(2): 267-272. [Guan X D, Huang C, Liu G H, et al. Extraction of paddy rice area using a DTW distance based similarity measure: Taking Thailand as an example[J]. Resources Science, 2014, 36(2): 267-272.]
- [5] Lefebvre A, Sannier C, Corpetti T. Monitoring urban areas with Sentinel-2A Data: Application to the update of the Copernicus high resolution layer imperviousness degree[J]. Remote Sensing,

- 2016, DOI: 10.3390/rs8070606.
- [6] Immitzer M, Vuolo F, Atzberger C. First experience with Sentinel-2 Data for crop and tree species classifications in central Europe [J]. *Remote Sensing*, 2016, DOI: 10.3390/rs8030166.
- [7] 时丽娜, 藉雪峰, 王星. 利用 Sentinel-2A 数据的西藏阿里冰崩范围提取[J]. *内蒙古煤炭经济*, 2017, (2): 157-160. [Shi L N, Xue X F, Wang X. Extraction of the ice depression range of Tibet Ali using Sentinel-2A data[J]. *Inner Mongolia Coal Economy*, 2017, (2): 157-160.]
- [8] Rujoiu-Mare M, Olariu B, Mihai B, et al. Land cover classification in Romanian Carpathians and Sub Carpathians using multi-date Sentinel-2 remote sensing imagery[J]. *European Journal of Remote Sensing*, 2017, 50(1): 496-508.
- [9] Hayes M M, Miller S N, Murphy M A. High-resolution land cover classification using Random Forest[J]. *Remote Sensing Letters*, 2014, 5(2): 112-121.
- [10] 姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. *吉林大学学报(工学版)*, 2014, 44(1): 137-141. [Yao D J, Yang J, Zhan X J. Feature selection algorithm based on random forest[J]. *Journal of Jilin University(Engineering Science)*, 2014, 44(1): 137-141.]
- [11] Pal M. Random forest classifier for remote sensing classification [J]. *International Journal of Remote Sensing*, 2005, 26(1): 217-222.
- [12] 张晓羽, 李凤日, 甄贞, 等. 基于随机森林模型的陆地卫星-8 遥感影像森林植被分类[J]. *东北林业大学学报*, 2016, 44(6): 53-57. [Zhang X Y, Li F R, Zhen Z, et al. Forest vegetation classification of landsat8 remote sensing image based on random forests model[J]. *Journal of Northeast Forestry University*, 2016, 44(6): 53-57.]
- [13] 宫攀. 基于 MODIS 数据关键物候特征参数的东北地区植被覆盖分类[J]. *资源科学*, 2010, 32(6): 1154-1160. [Gong P. Vegetation classification based on phenology indices derived from MODIS Data in Northeastern China[J]. *Resources Science*, 2010, 32(6): 1154-1160.]
- [14] 张喆, 丁建丽, 李鑫, 等. TVDI 用于干旱区农业旱情监测的适宜性[J]. *中国沙漠*, 2015, 35(1): 220-227. [Zhang Z, Ding J L, Li X, et al. Suitability of TVDI used to monitor agricultural drought in arid area [J]. *Journal of Desert Research*, 2015, 35(1): 220-227.]
- [15] 周晓双, 姚霞, 田永超, 等. 基于高光谱的稻麦叶面积指数监测研究[C]. 北京: 2014 年中国作物学会学术年会论文集, 2014. [Zhou X S, Yao X, Tian Y C, et al. Monitoring of Rice Leaf Area Index Based on Hyperspectral[C]. Beijing: China Crop Society Academic Annual Meeting, 2014.]
- [16] 陈文倩, 丁建丽, 李艳华, 等. 基于国产 GF-1 遥感影像的水体提取方法[J]. *资源科学*, 2015, 37(6): 1166-1172. [Chen W Q, Ding J L, Li Y H, et al. Extraction of water information based on China-made GF-1 remote sense image [J]. *Resources Science*, 2015, 37(6): 1166-1172.]
- [17] 李智峰, 朱谷昌, 董泰锋. 基于灰度共生矩阵的图像纹理特征地物分类应用[J]. *地质与勘探*, 2011, 47(3): 456-461. [Li Z F, Zhu G C, Dong T F. Application of GLCM-based texture features to remote sensing image classification [J]. *Geology and Prospecting*, 2011, 47(3): 456-461.]
- [18] 胡玉福, 邓良基, 匡先辉, 等. 基于纹理特征的高分辨率遥感图像土地利用分类研究[J]. *地理与地理信息科学*, 2011, 27(5): 42-45. [Hu Y F, Deng L J, Kuang X H, et al. Study on land use classification of high resolution remote sensing image based on texture feature [J]. *Geography and Geo-Information Science*, 2011, 27(5): 42-45.]
- [19] Breiman L. Random forests, machine learning 45[J]. *Journal of Clinical Microbiology*, 2001, 2: 199-228.
- [20] Gislason P O, Benediktsson J A, Sveinsson J R. Random forests for land cover classification[J]. *Pattern Recognition Letters*, 2006, 27(4): 294-300.
- [21] Fan H. Land-cover mapping in the Nujiang Grand Canyon: Integrating spectral, textural, and topographic data in a random forest classifier[J]. *International Journal of Remote Sensing*, 2013, 34(21): 7545-7567.
- [22] 张洪敏, 张艳芳, 田茂, 等. 基于主成分分析的生态变化遥感监测: 以宝鸡市城区为例[J]. *国土资源遥感*, 2018(1): 203-209. [Zhang H M, Zhang Y F, Tian M, et al. Remote sensing monitoring of ecological change based on principal component analysis: A case study of Baoji City[J]. *Remote sensing of land and resources*, 2018(1): 203-209.]
- [23] 田野, 赵春晖, 季亚新. 主成分分析在高光谱遥感图像降维中的应用[J]. *哈尔滨师范大学自然科学学报*, 2007, 23(5): 58-60. [Tian Y, Zhao C H, Ji Y X. Application of principal component analysis in dimensionality reduction of hyperspectral remote sensing images[J]. *Journal of Natural Science of Harbin Normal University*, 2007, 23(5): 58-60.]
- [24] 李冰, 卢小平, 李新社, 等. 特征优选的 GF-2 影像湿地地表覆盖要素提取[J]. *测绘与空间地理信息*, 2018(9): 49-52. [Li B, Lu X P, Li X S, et al. Surface coverage factor extraction of Gf-2 image of wetland with optimized features[J]. *Mapping and Spatial Geographic Information*, 2018(9): 49-52.]
- [25] Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot. Variable selection using Random Forests[J]. *Pattern Recognition Letters*, 2010, 31(14): 2225-2236.
- [26] 张磊, 宫兆宁, 王启为, 等. Sentinel-2 影像多特征优选的黄河三角洲湿地信息提取[J]. *遥感学报*, 2018, 23(2): 313-326. [Zhang L, Gong Z N, Wang Q W, et al. Sentinel-2 image multi-feature optimization for information extraction of Yellow River delta wetland [J]. *Journal of Remote Sensing*, 2018, 23(2): 313-326.]

# Land-cover classification of random forest based on Sentinel-2A image feature optimization

HE Yun<sup>1,2</sup>, HUANG Chong<sup>1</sup>, LI He<sup>1</sup>, LIU Qingsheng<sup>1</sup>, LIU Gaohuan<sup>1</sup>,  
ZHOU Zhenchao<sup>3</sup>, ZHANG Chenchen<sup>1,2</sup>

(1.State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China;

3. College of Geo-Exploration Science and Technology, Jilin University, Changchun 130026, China)

**Abstract:** Due to the suitable hydrothermal conditions, vigorous vegetation growth, high land use intensity and complex spatiotemporal variation of spectral characteristics of surface cover types, it is difficult to guarantee the accuracy of remote sensing classification using traditional spectral characteristics in tropical and subtropical regions. Multi-spectral, high spatial resolution Sentinel-2A imageries provide a new source of data for land-cover classification. In order to improve the speed and accuracy of land-cover classification using Sentinel-2A images, we propose a classification method with feature-optimized random forests. In this study, we took the Mun River Basin of Indo-China Peninsula as our research area and made full use of the rich spectral characteristics, normalized vegetation index (NDVI), ratio vegetation index (RVI), difference vegetation index (DVI), normalized water body index (NDWI), and texture features including contrast, correlation, energy, mean, and entropy, of Sentinel-2A images for the analyses. We used the average impurity reduction method in random forests to evaluate the importance of different spectral features, indices, and texture features. Combining the out-of-bag (OOB) error to select features, the results of land-cover classification with feature-optimized random forests were obtained. They show that the spectral features and texture features of Sentinel-2A images play an important role in our classification compared with the original random forest land-cover classification results. The short-wave infrared, visible, and vegetation red-edge bands are of greater importance in spectral features, and the mean and energy are of high importance in texture features. The accuracy of OOB is the highest when the top 9 important features are selected. Sentinel-2A images have good adaptability in tropical and subtropical region land-cover classification. It can effectively improve the accuracy of land-cover classification in tropical and subtropical regions. The accuracy of our classification method reaches 87.53%, and the Kappa coefficient reaches 0.8461, better than the original random forest method. The random forest method based on feature optimization not only has a fast classification speed, but also can guarantee high classification accuracy under the condition that the sample is representative, especially suitable for the land-cover classification of medium and high spatial resolution images of Sentinel-2A.

**Key words:** Sentinel-2A; feature optimization; random forest; land-cover classification; out-of-bag (OOB) method; Indo-China Peninsula; Mun River Basin of Thailand