

基于多目标遗传随机森林特征选择的面向对象湿地分类

刘舒¹ 姜琦刚¹ 马玥¹ 肖艳¹ 李远华¹ 崔璨²

(1. 吉林大学地球探测科学与技术学院, 长春 130026; 2. 大连海事大学航海学院, 大连 116000)

摘要: 以多时相 Landsat8 影像和 SRTM DEM 为数据源,对南瓮河流域进行了面向对象湿地分类。为削弱高维特征集对分类精度的影响,提出一种多目标遗传随机森林组合式特征选择算法(MOGARF)进行特征集优化。利用 Relief F 算法对完整特征集进行特征初选,再以基于随机森林的封装式多目标遗传算法进一步提取优化特征集。将所得特征集结合随机森林分类法提取湿地信息。并将结果分别与基于完整特征集和仅采用 Relief F 算法及 Boruta 算法提取的优化特征集的 3 种随机森林分类结果对比。试验结果表明,采用 MOGARF 算法特征选择后,特征维度降低至原来的 10%,且分类精度最高,总体精度为 92.61%,比其他分类方案提高 0.35%~1.94%,Kappa 系数为 0.907 5,袋外误差为 7.77%,比其他分类方案降低 0.91%~1.48%。利用 MOGARF 特征选择的随机森林分类法是湿地分类的有效方法。

关键词: 湿地分类; 多光谱遥感影像; 面向对象; 多目标遗传随机森林算法; 特征选择

中图分类号: TP79 **文献标识码:** A **文章编号:** 1000-1298(2017)01-0119-09

Object-oriented Wetland Classification Based on Hybrid Feature Selection Method Combining with Relief F, Multi-objective Genetic Algorithm and Random Forest

LIU Shu¹ JIANG Qigang¹ MA Yue¹ XIAO Yan¹ LI Yuanhua¹ CUI Can²
(1. College of Geo-exploration Science and Technology, Jilin University, Changchun 130026, China
2. Navigation College, Dalian Maritime University, Dalian 116000, China)

Abstract: Recently, researchers adopted object-oriented method to extract wetland distributions. Multi-temporal and multi-sources of data can facilitate the extraction process but meanwhile it enlarges the amount of features. It needs a large quantity of experiment based on the expert knowledge to determine the optimal feature sets and the threshold values. In order to improve the classification accuracy and relieve the researchers from large amount of work, a filter-wrapper hybrid feature selection method combining relief F, multi-objective genetic algorithm and random forest was proposed, which was a two-step method. In the first step, relief F algorithm was adopted to select features with class separability. In the second step, multi-objective genetic algorithm based on random forest (MOGARF) was built. Four measures such as out-of-bag (OOB) error of random forest algorithm, dimension of the feature space, correlations among features and the variable weight of relief F algorithm were acted as four objectives of MOGA. The probability whether the feature was expressed was determined by the variable importance measures from random forest algorithm. The crowded distance of each feature collection was calculated and the feature collection with the least crowded distance was the optimal feature set. Nanweng river basin was taken as the study site. Object-oriented classification using random forest classifier was conducted based on the optimal feature set. Then the result was compared with three other random forest classification schemes by using the entire feature set or the feature set selected by relief F algorithm or the Boruta algorithm. The classification scheme with MOGARF had the best performance and the feature dimension was reduced to 10% of the entire one. The overall accuracy reached 92.61% which was 0.35%~1.94% higher than those of the other three schemes with Kappa coefficient of 0.930 6. The

收稿日期: 2016-09-02 修回日期: 2016-11-04

基金项目: 东北地区国土资源遥感综合调查项目(85015B01009)

作者简介: 刘舒(1988—),女,博士生,主要从事遥感地学和环境遥感研究,E-mail: liushu8877@126.com

通信作者: 姜琦刚(1964—),男,教授,博士生导师,主要从事 GIS 与遥感地学环境研究,E-mail: jiangqigang@jlu.edu.cn

OOB error of MOGARF was 7.77% which was 0.91% ~ 1.48% lower than those of the other schemes. All these indicated that the MOGARF feature selection method was an effective feature selection method when it was combined with random forest classifier.

Key words: wetland classification; multi-spectral remote sensed imagery; object-oriented; multi-objective genetic and random forest algorithm; feature selection

引言

利用多光谱遥感影像对湿地进行基于植物类型的准确分类,是湿地碳循环过程监测的有效手段^[1]。目前研究者常引入多源数据采取面向对象分类方式,以削弱多光谱影像分类的弊端^[2-4]。但面向对象方法增大了特征维度,在使用常规方法分类时需要先进行特征选择^[4-7]。随机森林(Random forest, RF)被认为是能够直接处理高维数据的高效分类算法^[8-10],广泛应用于多个领域,近年来也被引入到湿地分类研究中,并获得较高精度的结果^[2,5,10-12]。但当特征维度过高时,特征集中次要特征仍会降低分类精度,在分类前也需要优化特征空间^[13]。优化的特征空间应具有较低维度,特征应具有类内聚合性和类间可分性,特征间相关度低,能得到较高的分类精度^[14]。

目前与 RF 算法结合的特征选择算法主要分为过滤式(Filter)和封装式(Wrapper)。也有学者构造 Filter - Wrapper 组合特征选择法,发挥两种算法优势,兼顾多个特征空间评价因素,寻求精度和效率的平衡^[4,15]。

Relief F 算法是一种过滤式多类别特征选择算法,通过计算特征权重,判断特征重要性。

目前基于 RF 模型的面向对象湿地分类研究较

少,将组合式特征选择算法应用于面向对象分类的研究也相对较少^[4]。本文将 Relief F Filter (Relief F)算法和基于随机森林的多目标遗传 Wrapper 算法结合,提出多目标遗传随机森林 Filter - Wrapper 组合式特征选择算法提取优化特征集,并应用于南瓮河流域,实现基于 RF 的面向对象湿地分类。将分类结果与基于完整特征集 (No_FS) 和分别利用 Relief F 算法、Boruta 算法提取的优化特征集的 RF 面向对象湿地分类对比,验证该方法的有效性。

1 研究区数据

1.1 研究区概况

南瓮河流域位于大兴安岭地区东部,地理坐标为北纬 50°56'12" ~ 51°39'40",东经 124°24'54" ~ 126°13'15"(图 1)。研究区内植被丰富,乔木主要有兴安落叶松、柞树等;灌木有兴安杜鹃、丛桦等;草本植物主要为杜香、大叶章等。该区为寒温带大陆性季风气候,寒冷季节较长,植物生长周期约为 110 d。研究区湿地资源丰富,主要包括森林湿地、灌丛湿地、草本湿地、岛状林湿地、湖泊湿地和河流湿地,人工湿地为采矿后产生的积水地带。其内包含我国唯一以寒温带森林湿地生态系统为保护对象的国家级自然保护区^[7],也是我国最大的森林湿地分布区之一^[7,16]。

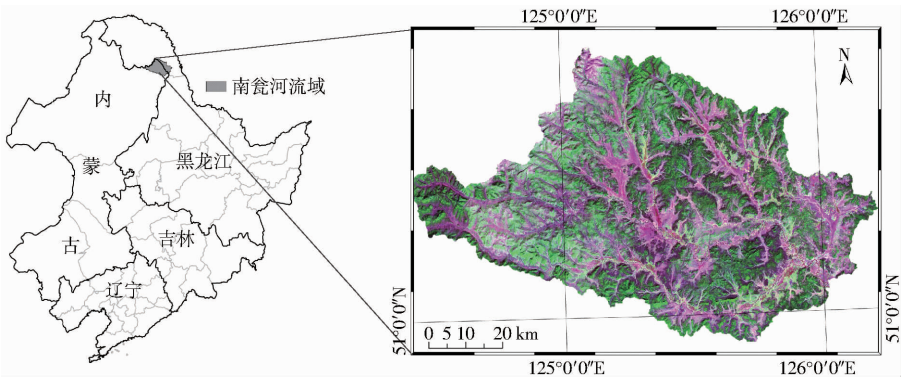


图 1 研究区位置图
Fig.1 Location map of study area

1.2 数据源与预处理

本文以春、夏和秋季 30 m 空间分辨率的 Landsat8 (OLI) 影像和 90 m 空间分辨率的 SRTM DEM 影像为数据源,提取南瓮河流域光谱、形状、纹理和地形特征,并计算指数特征。春季 OLI 影像选

用 LC81200242014098LGN00, 夏季影像选用 LC81200242015197LGN00, 秋季影像选用 LC81200242015293LGN00。夏季影像分布有 1% 的薄云,经大气校正后不影响光谱特征的提取。由于选取的 OLI 影像拍摄时间接近,且南瓮河流域大部

分区域较少人为活动干预,本文假定不同时相影像同一位置像素间没有土地覆被类型转换,不影响像素类别划分过程。

将 3 个时相 Landsat8 影像在 ENVI5.1 中辐射定标和大气校正,SRTM DEM 重采样为 30 m,并统一两种影像投影参数。将每个时相 Landsat8 影像的第 1、2、3、4、5、7 波段,连同高程信息合成具有 19 个波段的影像分类基础数据。利用 Ecognition 多尺度分割模块分割该基础数据,分割时令所有波段权重为 1,形状因子为 0.1,紧凑度因子为 0.3,分割尺度确定为 15。

2 研究方法

2.1 湿地分类体系建立与湿地特点分析

根据《湿地公约》和《东北地区国土遥感综合调

查》项目建立的分类体系,南瓮河流域湿地类型以天然湿地为主,具体分为河流湿地、湖泊湿地和沼泽草甸湿地。区内的人工湿地主要为采矿积水区,由于面积较小,在本研究中不单独列为一类。参照 2000 年和 2007 年对该区湿地的解译结果和相关文献,区内河流湿地主要为永久性河流,湖泊湿地主要为季节性湖泊和坑塘,沼泽草甸湿地主要为草本湿地、森林湿地和灌丛湿地。结合研究区实际情况,建立湿地分类体系如表 1 所示^[7]。

本研究的目的在于提取湿地信息,因此将区内其他土地覆被类型作为背景地类。背景地类参考 2007 年《土地利用现状分类》国家标准划分至一级类,分别为:林地、草地、耕地、交通运输用地。其中耕地以旱地为主,交通运输用地以道路为主,在此简称为道路。林地特指不包括森林湿地的一般有林区域。

表 1 研究区湿地分类体系
Tab.1 Classification system for wetlands in study area

一级	二级	三级	划分依据
天然湿地	河流湿地	永久性河流	常年有水的河流,仅包括河床部分
	湖泊湿地	季节性湖泊	季节性积水的漫滩湖泊或浅滩,以水面为主
		草本湿地	草本植物为主,植被盖度不低于 30%,有泥炭层
	沼泽草甸湿地	灌丛湿地	灌木植物为主,植被盖度不低于 30%,无泥炭累积
		森林湿地	落叶乔木为主,植被郁闭度不低于 0.2,有泥炭层

研究区内的乔木主要为落叶乔木。处于植被繁茂的有叶时期时,林地与森林湿地的植被群落组成和冠层反射率相似,直接利用光学遥感影像分离森林湿地较为困难。但二者地表水文特征有显著差异,在早春或晚秋的落叶期对二者林下水文信息的提取是区分林地和森林湿地的关键。季节性积水的特点也要求森林湿地多处于坡度较小的区域。本研究区存在岛状林湿地,在影像上表现为被草本湿地包围的林地小斑块,易于辨识^[7,10]。

湿地与旱地的水文特征差异还影响上层植被的长势和生长期分布,进行野外验证时发现,与季节性干涸的草本湿地相比,水分充足的草本湿地中植被更晚枯萎。研究区内草本湿地和草地的植被生长周期不同,在春季和秋季的影像上二者区别明显^[7]。草本湿地和森林湿地的植被类型差异较大,可直接通过植被指数进行区分。

灌丛湿地属于森林湿地和草本湿地之间的过渡阶段,虽与草本湿地、森林湿地之间的界限不明显,但它不同于森林湿地和草木湿地,表层土壤无泥炭累积,且植被类型以灌木为主,落叶时期的土壤指数和有叶时期植被指数为灌丛湿地识别的关键。

对于永久性河流和季节性湖泊而言,当处于丰水期时,二者在影像上均表现为深蓝色水面,当枯水

期时,二者区别明显。

2.2 样本提取与精度验证

样本质量直接关系到湿地分类的精度,应选择具有典型性、代表性的纯净对象作为样本。常用的样本选取方式有:简单随机抽样、聚点抽样、等距抽样和分层随机抽样^[17]。由于 RF 算法分类精度受各类别样本数量和空间分布的影响^[9,18],因此采取分层随机抽样的方式,使各类别样本点数量大致与该类别的总面积成比例。为保证比例较小的类别也有一定数量样本,设各类别最小样本数为 100。

依据 Google Earth 高分辨率影像解译研究区的岛状林湿地和沿河的灌丛湿地,并将解译结果作为森林湿地和灌丛湿地典型样区;利用收集到的 2007 年湿地解译数据确定草本湿地、永久性河流和季节性湖泊典型样区;结合部分 2016 年东北地区自然资源解译数据确定森林、草地、耕地和道路样区。在高清影像中岛状林斑块和灌丛湿地特征明显,边界清晰;选用的已有参考解译数据都已经验证过,是样本选取和精度评价的可靠基础数据。

查阅相关统计数据,设定覆被类型面积比例。提取落在各样区内相对纯净的分割对象组成样本空间,按照设定比例在每类地物样本空间中随机抽取样本。其中 2/3 作为训练样本,用于特征选择和 RF

建模,剩余作为测试样本,利用模型将其分类,并与评价 RF 分类精度。各类样本数量如表 2 所示。

样本实际类别对比,计算混淆矩阵和相关精度指标,

表 2 各类型样本数量分布

Tab.2 Numbers of samples for each class

类别	林地	森林湿地	灌丛湿地	草本湿地	草地	耕地	道路	永久性河流	季节性湖泊
训练集	500	80	100	500	200	100	100	100	80
测试集	250	30	50	250	100	50	50	50	40

2.3 对象特征提取

依据不同湿地类型特点建立完整对象特征集,共提取与水体、土壤和植被状态有关的 473 个特征用于湿地分类。主要包括以下类别:

(1)光谱特征:提取各对象 3 个时相 Landsat8 影像 1、2、3、4、5、7 波段的均值 (Mean)、标准差 (Std)、最大像素值 (Max.)、最小像素值 (Min.)、偏斜度 (Skewness)、波段最大差异 (Max. diff)、总体亮度值 (Brightness)^[6]。

(2)指数特征:提取归一化植被指数 NDVI,差值植被指数 DVI,比值植被指数 RVI,土壤调节植被指数 SAVI,优化土壤调节植被指数 OSAVI,修正土壤调节植被指数 MSAVI,垂直植被指数 PVI,增强植被指数 EVI^[19],结构不敏感色素指数 SIPI^[20],修正差值植被指数 RDVI,改进叶绿素溶解绿指数 MCARI2^[21],绿度植被指数 GVI^[22],抗大气植被指数 ARVI^[23],归一化湿度指数 NDMI,穗帽变换绿度指数 TCG,湿度指数 TCW,亮度指数 TCB,归一化建筑指数 NDBI^[24],修正归一化水体指数 RNDWI,改进归一化水体指数 MNDWI,新型水体指数 NWI,增强水体指数 EWI^[25],全球植被水分指数 GVMI^[26],可见光短波红外干旱指数 VSDI,地表水指数 LSWI,水分胁迫指数 MSI,地表水容量指数 SWCI^[27],叶面积指数 LAI^[28],叶绿素含量指数 CVI,表征绿量 VQ^[29]。

(3)纹理特征:提取对象全方位灰度共生矩阵

$$W_m = W_m - \sum_{j=1}^k \frac{|R[m] - H_j[m]|}{\max(m) - \min(m)} \Big/ (nk) + \sum_{c \notin \text{class}(R)} \left[\frac{p(c)}{1 - p(\text{class}(R))} \sum_{j=1}^k \frac{|R[m] - T(c)_j[m]|}{\max(m) - \min(m)} \right] \Big/ (nk)$$

式中 c ——异类样本的样本类别
 $R[m]$ ——个体 R 特征 m 的值
 $H_j[m]$ ——第 j 个最邻近同类样本特征 m 的值
 $p(c)$ ——异类样本类别为 c 的概率
 $\text{class}(R)$ ——个体 R 的类别
 $p(\text{class}(R))$ ——样本类别与 R 相同的概率
 $T(c)_j[m]$ ——第 j 个最邻近 c 类样本特征 m 的值
特征权重越大,说明该特征使得样本的类间距

均值 (GLCM Mean),熵 (GLCM Ent),同质度 (GLCM Homo),标准差 (GLCM Std),非相似性 (GLCM Dissim),对比度 (GLCM Contrast),角二阶矩 (GLCM Ang. 2nd Moment) 和相关性 (GLCM Corr);全方位 (all dir.) 归一化灰度矢量 (GLDV) 的均值 (GLDV Mean),熵 (GLDV Ent),反差 (GLDV Contrast),角二阶矩 (GLDV Ang. 2nd Moment) 来描述对象纹理特征。

(4)地形特征:从 SRTM 影像中提取高程 (Elev)、坡向 (Aspect)、坡度 (Slope) 和山体阴影 (Hillshade)等地形信息。

(5)形状特征:提取对象的形状特征,包括对象的边界长度、长度、宽度、圆度、紧凑度、不对称性、密度、包含像素数、面积、形状指数等^[6]。

2.4 多目标遗传随机森林 (MOGARF) 特征选择方案

2.4.1 Relief F 算法原理

Relief F 算法是目前广泛应用的一种过滤式 (Filter) 多类别特征选择算法,通过计算特征权重判断特征重要性。该算法不受数据类型的限制,不受噪声干扰,实施简单,运算效率高。具体过程为:在某类样本中抽取一个体 R ,分别在同类和异类样本中寻找 k 个最邻近样本,构成同类临近样本集 H 和异类临近样本集 T 。再以 R 与 H 和 T 中样本各特征平均差异的差值定义特征权重 W 。对于任意特征 m ,完成 n 次抽样的特征权重 W_m 计算式为

(1)

离大,类内距离小,对类别识别作用大^[4,30]。

2.4.2 随机森林算法原理

随机森林分类器 (Random forest classifier) 是根据随机抽取的样本信息,通过建立多棵独立决策树分别预测目标类别的无参数分类器。假定生成决策树的棵数为 N ,样本的全部特征数为 M ,模型的每棵决策树都是在所有训练样本中依据 bootstrap 抽样方式提取子训练集后,再从参与建模的全部特征中随机选择指定数目的特征建立的,并利用不属于子训

练集的样本计算该决策树的分类内部误差。为新数据分类时,综合所有决策树的独立预测结果,通过投票方式决定待分类目标的最可能类别,得出综合袋外误差 (Out-of-bag error, OOB)。OOB 误差越小,说明基于 RF 的分类错误率越小,分类精度越高。

OOB 误差除体现分类精度外,也可用于计算特征重要性评分 (VIM)。在每棵决策树得到 OOB 误差 (B_o) 后,对于每一个参与决策树运算的特征变量,保持其他特征取值不变,将该特征变量袋外数据取值随机打乱,重新计算决策树的 OOB 误差 (B_n)。所有决策树两类 OOB 误差的差值和的百分比即为被打乱特征的 VIM。对于任意特征 M_A ,决策树编号为 t ,特征重要性评分 $V(M_A)$ 可以表示为

$$V(M_A) = \frac{1}{N} \sum_{t=1}^N (B_{n_t}^{M_A} - B_{o_t}^{M_A}) \tag{2}$$

式中 $B_{o_t}^{M_A}$ ——任意特征 M_A 值未被打乱时第 t 棵决策树的 OOB 误差
 $B_{n_t}^{M_A}$ ——任意特征 M_A 值被打乱时第 t 棵决策树的 OOB 误差
 V 越大,特征越重要^[8,9,31]。

特征选择过程中,RF 算法主要用于评价各特征组合分类精度和组合内特征的重要程度。获得最优特征集后,还将利用 RF 算法提取研究区湿地信息。

RF 模型基于 R 语言的 Random Forest 语言包建立。运行时需为两个参数赋值:生成决策树的数目 $ntree$ 和节点分裂时输入的特征变量个数 $mtry$ 。软件默认 $mtry$ 为 $\sqrt{M}$, $ntree$ 为 500^[3]。当特征数目较多时,使用默认值得到的模型 OOB 误差可能偏高,因此本文在每次利用 RF 算法分类前,都分别对

两个参数取值进行试验,提取使 OOB 误差最小参数组合并以此建立 RF 模型。具体方法为:首先固定 $ntree$ 为 2 000,对 $mtry$ 取遍 1 ~ M 的所有整数分别建立 RF 模型,计算 OOB 误差,最佳 $mtry$ 取值为使 OOB 误差最小的 $mtry_0$;再令 $mtry = mtry_0$,使 $ntree$ 取遍 1 ~ 2 000 的所有整数,比较相应的 OOB 误差,确定使 OOB 误差最小的最佳 $ntree$ 取值。

2.4.3 MOGARF 特征选择方案构建

做某种决策时常需要同时综合考虑多项限制条件,有时这些约束条件甚至是相互矛盾的,这就需要借助多目标优化思想来解决问题。NSGAI 多目标遗传算法是对传统遗传算法的改进,目前已成为解决多目标优化问题的基准算法之一。NSGAI 的步骤主要有编码、初始化种群、设计目标函数、非支配排序和计算拥挤距离,通过对拥挤距离的评估进行选择、交叉和变异产生新种群^[14]。

本文提出的 MOGARF 算法是基于 NSGAI 的特征选择算法,首先利用 Relief F 算法对参与分类的特征进行初选,剔除类间距离小于类内距离的特征;再根据特征优化的 4 个因素确定算法目标,建立非支配集,计算拥挤距离;在生成初始化种群时,由 RF 算法得到 VIM 基于 Gama 分布产生每个保留特征入选的概率,确保重要的特征入选的几率更大^[13]。

设遗传代数为 G ,种群中个体数为 S ,选择率为 p_s ,交叉率为 p_c ,变异率为 p_m ,由包含全部特征的 No_FS 特征集得到的最佳 RF 建模参数为 $mtry_0$ 和 $ntree_0$,MOGARF 算法具体实现流程如图 2 所示^[13]。

对于第 G 代种群中的个体,直接选取其中拥挤距离最大的个体作为最优解,该个体包含的特征构

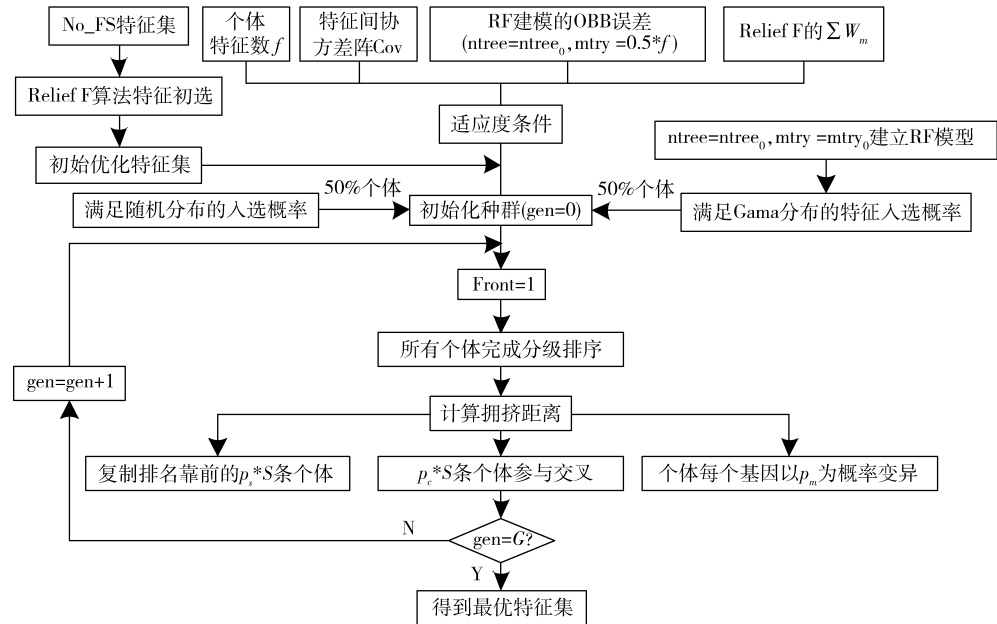


图 2 MOGARF 特征选择流程图

Fig. 2 Flow chart of MOGARF feature selection method

成本试验最优特征集。

MOGARF 选择和湿地 RF 分类过程借助 Matlab 和 R 语言软件平台交互实现。对选出的最优特征集进行 RF 模型参数选择试验,选出最佳参数组合 $ntree_n$ 和 $mtry_n$,并对最优特征集建立 RF 模型,提取南瓮河流域湿地信息。

3 结果与讨论

3.1 MOGARF 模型选取特征集

对 No_FS 特征集,当 $mtry$ 取 94 时 OOB 误差取最小值 13.37%, $ntree$ 超过 1 000 时模型 OOB 误差趋于稳定,故将最佳参数组合确定为 $mtry_0 = 94$, $ntree_0 = 1\,000$ 。此参数组合也作为初始 RF 建模参数提取满足 Gama 分布的特征入选概率。利用 Matlab 和 R 语言软件交互实现 MOGARF 特征选择

过程,得到了具有 45 个特征的最优特征集,如表 3 所示。表中指数特征以“特征类型 季节缩写”的形式命名,其他特征以“特征类型 波段 季节”命名。

在最优特征集中,包含除形状特征外的其余 4 种特征类别,说明想要达到理想的分类效果,需要综合利用多源信息,整合不同类别的特征,共同用于分类过程。入选光谱特征和指数特征数量最多,说明在本研究中光谱特征和指数特征是重要特征类别。纹理特征和地形特征种类较少,且没有形态特征入选,可能由以下原因造成:对于中等分辨率的影像而言,由于存在混合像元,且在影像分割后以对象边界为窗口提取对象的纹理特征,削弱了纹理特征取值的区别,使纹理特征在分类过程中起到作用较小。本文采用统一尺度分割影像,对象的形状特征相似,削弱了形状特征的作用。

表 3 最优特征集
Tab.3 Optimal feature set

特征类别	特征命名	数量
光谱特征	Mean NIR spr,Mean NIR sum、Mean SWIR2 sum、Mean Red aut,Mean SWIR1 aut、Skewness SWIR1 aut,Skewness SWIR2 aut,Skewness SWIR2 spr,Min. pixel Blue spr,Min. pixel Elev、Min. pixel Green sum、Min. pixel SWIR1 sum、Max. pixel Blue sum、Max. pixel Slope	14
纹理特征	GLCM Ang. 2nd moment (all dir.),GLCM Mean Blue spr、GLCM Mean Green sum、GLCM Mean Red spr、GLCM Mean Elev、GLCM Mean Slope、GLDV Ent NIR aut	7
指数特征	MCAR12 aut、EWI spr、LAI spr、SAVI spr、OSAVI sum、DVI spr、DVI aut、EVI spr、EVI sum、EVI aut、ARVI spr、GVI spr、GVI sum、NDBI spr、NDBI aut、CVI spr、CVI aut、MNDWI aut、LSWI spr、NDMI sum、NDMI aut、SWCI spr、SWCI aut	23
地形特征	Mean Elev	1

3.2 MOGARF_RF 分类方案结果与精度分析

令 OOB 误差最小的参数组合为 $mtry_{\text{MOGARF}} = 6$, $ntree = 1\,000$ 。建立 RF 模型对研究区进行分类,湿地分布情况如图 3a 所示。研究区内各类湿地总面积为 1 661.698 km²,占总面积的 30%,多分布于平缓低洼地带。其中森林湿地约为 274.576 km²,主要集中在研究区东南,占湿地总面积的 16.52%;灌丛湿地面积 36.60 km²,多分布于河岸两侧,占湿地总

面积的 0.02%;草本湿地面积 1 340.120 km²,面积最大,遍布整个研究区,占湿地总面积的 80.65%。与图 3b 对比可知,本文得到的湿地面积和分布情况基本符合研究区实际情况。

计算混淆矩阵,获取的精度评定指标如表 4 所示。混淆矩阵中的元素以百分比形式表示。由表 4 可知,MOGARF_RF 分类方案的总体精度为 92.61%,Kappa 系数为 0.907 5,森林湿地、灌丛湿

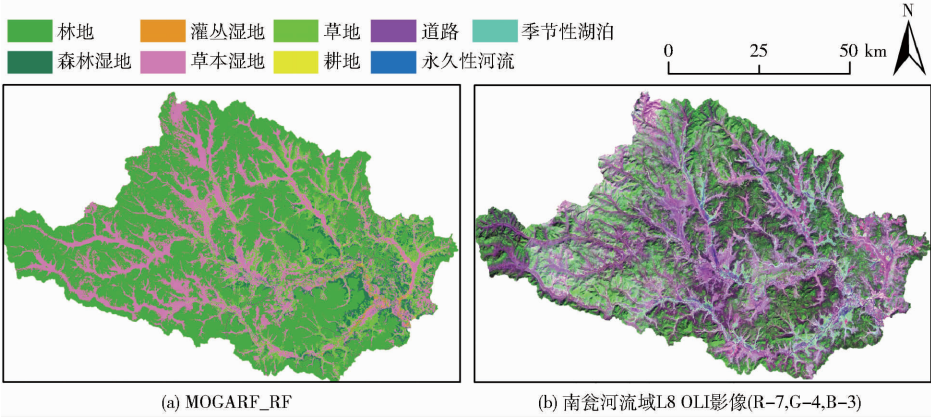


图 3 MOGARF_RF 方案分类结果

Fig.3 Result maps of MOGARF_RF classification strategy

地、草本湿地、永久性湖泊和季节性河流的用户精度分别为 87.5%、82.0%、97.6%、88.0% 和 79.5%。这 5 类湿地的制图精度分别为 94.7%、93.3%、86.3%、86.8% 和 97.0%。该方案分类精度较高，是研究区内湿地信息提取的可行方案。由于样本具有代表性，不同类别湿地间的混淆较少。不同地类的混淆主要发生在草本湿地和其他地类之间，灌丛湿地和永久性河流之间，以及季节性湖泊和永久性河流之间。部分混淆是由地类间光谱特征的相似性造成。季节性湖泊和永久性河流在一定时段内都存在特征相似的水面，草地和草本湿地在一定时段内水文、植被特征相似，在草相对茂盛的区域，即使草

下水文状态不同，也可能由于 Landsat 影像的弱穿透能力而导致误分错分现象。其余混淆主要是由于相邻两种地类间边界较模糊，且存在一定的混合像元造成的。一般情况下，湿地通常从森林湿地经由灌丛湿地和草本湿地过渡至泥炭沼泽湿地。相邻湿地类型常混生分布，之间没有明确的界限，在中等分辨率的影像中常以混合像元形式存在，从而导致湿地类型的误判断。由于岛状林的特殊性，研究区中部分森林湿地与草本湿地直接相连，二者也存在一定量的混合像元^[7]；本研究选取的灌丛湿地样本分布在永久性河流河岸处，灌丛湿地和永久性河流也存在混合像元。

表 4 MOGARF_RF 分类方案混淆矩阵

Tab.4 Confusion matrix of MOGARF_RF classification scheme

	林地	森林湿地	灌丛湿地	草本湿地	草地	耕地	道路	永久性河流	季节性湖泊
林地	1	0	0	0	0	0	0	0	0
森林湿地	0	0.90	0	0.08	0	0	0	0	0.03
灌丛湿地	0	0	0.84	0.04	0	0	0	0.12	0
草本湿地	0	0	0	0.98	0.01	0	0	0	0
草地	0	0.02	0	0.13	0.83	0.02	0	0	0
耕地	0	0	0	0.08	0.02	0.88	0.02	0	0
道路	0.04	0	0	0.20	0.04	0	0.72	0	0
永久性河流	0	0	0.04	0.04	0	0	0	0.92	0
季节性湖泊	0	0	0	0.13	0	0	0.03	0.03	0.82
制图精度/%	99.2	94.7	93.3	86.3	93.3	95.6	92.3	86.8	97.0
用户精度/%	100	87.5	82.0	97.6	83.0	86.0	64.0	88.0	79.5
总体精度/%	92.61								
Kappa 系数	0.9075								

3.3 不同方案分类精度指标比较

为比较 MOGARF 算法的特征选择效果，将其分别与利用 No_FS 特征集、Relief F 优化特征集和 Boruta 优化特征集的 RF 分类对比，令 Relief F 优化特征集和 Boruta 优化特征集的特征数也为 45。RF 模型的 ntree 取 1 000，mtry_{0No_FS} = 5，mtry_{0MOGARF} = 6，mtry_{0Relieff} = 9，mtry_{0Boruta} = 5。由 4 种分类方案的混淆矩阵分析总体精度、Kappa 系数、OOB 误差、漏分和错分误差等指标，量化不同方案的分类精度。各分类方案精度评定指标如图 4 所示。

从图 4a~4c 中可以看出，在 4 种分类方式中，利用 MOGARF 优化特征集，总体精度达到 92.61%，OOB 误差为 7.77%；利用 Boruta 优化特征集的 RF 分类，总体精度达到 90.67%，OOB 误差为 9.25%；而考虑类间可分性的 Relief F 优化特征集分类精度为 92.26%。分类结果的 Kappa 系数都在 0.8 以上，分类结果具有较高可信度。在建立 RF 模型时都采用了每种分类方式各自的最优参数组合，保证所有分类结果都为对应方案中的相对最优解。

比较可知，MOGARF 进行面向对象的 RF 分类精度最高，总体精度比其他分类方案提高 0.35%~1.94%，OOB 误差比其他分类方案降低 0.91%~1.48%。

从图 4d、4e 可以看出，采用 MOGARF 特征选择算法对所有湿地类别的错分误差总体上小于其他方法，错分主要集中在森林湿地、灌丛湿地和季节性湖泊中。原因可能在于这 3 类湿地覆盖面积小，而在选取样本时为保证类别具有足够样本，实际采集的样本数所占比例大于实际地类区内所占比例，从而增大了其他类别个体被误判为这几类的可能性。永久性河流和草本湿地的漏分现象严重，道路与其他类别的错分也由相似原因导致。此外，道路在影像图中较为细小，易与其他类别形成混合像元，在一定程度上影响了分类精度。

4 结束语

从上述的研究过程与结果中可以看出，在采用 RF 算法进行湿地分类前，先利用 MOGARF 特征选择算法提取优化的特征集，能够在一定程度上提高

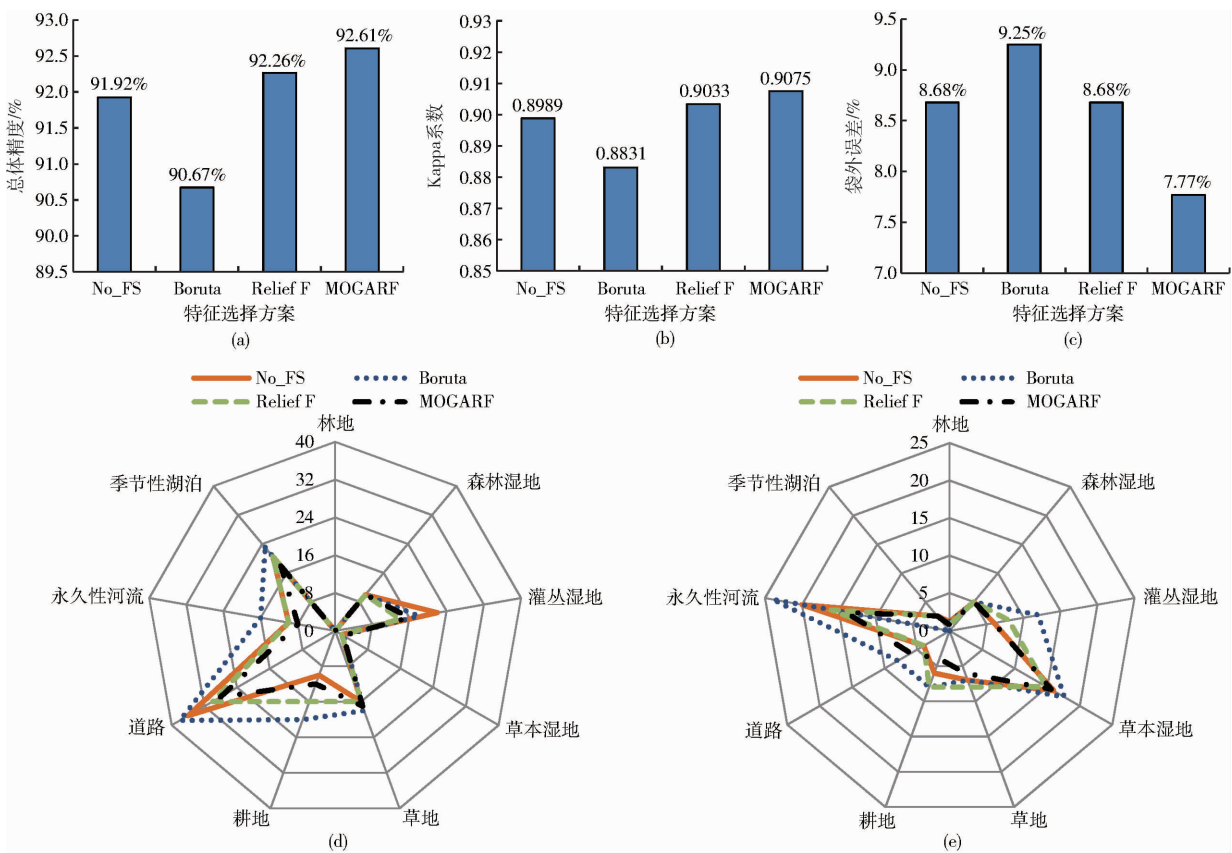


图 4 不同特征选择方案分类误差对比图

Fig. 4 Comparisons of classification errors with different feature selection strategies

RF 分类精度。相比于 Boruta 和 Relief F 特征选择方案,MOGARF 特征选择算法能够独立确定特征集维数,并平衡分类精度、数据冗余度、数据维度等多项特征集评价指标。此外,利用 MOGARF 算法选取最优特征集,能够大幅度降低数据维度,降维后的数

据通过 RF 算法进行分类,在保证高分类精度的前提下,减少研究人员在数据准备过程中的工作量,节约存储空间,提高运算效率。因此,基于 MOGARF 特征选择的面向对象分类方法是湿地分类研究可靠的实施方案,更适用于对湿地主要类别的提取。

参 考 文 献

1 毛德华. 定量评价人类活动对东北地区沼泽湿地植被 NPP 的影响[D]. 长春: 中国科学院东北地理与农业生态研究所, 2014.
MAO Dehua. Quantitative assessment in the impacts of human activities on net primary productivity of wetlands in the northeast China[D]. Changchun: Northeast Institute of Geography and Agro ecology, Chinese Academy of Sciences,2014. (in Chinese)

2 DRONOVA I. Object-based image analysis in wetland research: a review[J]. Remote Sensing, 2015,7(5):6380-6413.

3 NA Xiaodong, ZANG Shuying, WU Changshan, et al. Mapping forested wetlands in the Great Zhan River Basin through integrating optical, radar, and topographical data classification techniques[J]. Environmental Monitoring and Assessment, 2015, 187(11): 1-17.

4 肖艳,姜琦刚,王斌,等. Relief F 和 PSO 混合特征选择的面向对象土地利用分类[J]. 农业工程学报,2016,32(4):211-216.
XIAO Yan, JIANG Qigang, WANG Bin, et al. Object based land-use classification based on hybrid feature selection method of combining Relief F and PSO[J]. Transactions of the CSAE, 2016, 32(4):211-216. (in Chinese)

5 SHRUTHI R B, KERLE N, JETTEN V, et al. Object-based gully system prediction from medium resolution imagery using random forests[J]. Geomorphology, 2014,216: 283-294.

6 谢静. 基于多季相遥感信息的三江平原湿地信息提取[D]. 长春: 中国科学院研究生院东北地理与农业生态研究所,2013.
XIE Jing. Classification of wetlands using object-oriented method and multi-season remote sensing images in Sanjiang Plain[D]. Changchun: Northeast Institute of Geography and Agro ecology, Chinese Academy of Sciences, 2013. (in Chinese)

7 孙俊杰,马大喜,任春颖,等. 基于多时相环境卫星数据的南瓮河流域湿地信息提取方法研究[J]. 湿地科学,2013,11(1): 60-67.
SUN Junjie, MA Daxi, REN Chunying, et al. Method of extraction of wetlands' information in Nanweng River Basin based on multi-temporal environment satellite images[J]. Wetland Science, 2013, 11(1):60-67. (in Chinese)

8 BREIMAN L. Random forest[J]. Machine Learning,2001,45(1):5-32.

9 BELGIU M, DRĂGUȚ L. Random forest in remote sensing: a review of applications and future directions[J]. ISPRS Journal of

- Photogrammetry and Remote Sensing, 2016, 114: 24–31.
- 10 严婷婷, 边红枫, 廖桂项, 等. 森林湿地遥感信息提取方法研究现状[J]. 国土资源遥感, 2014, 26(2): 11–18.
YAN Tingting, BIAN Hongfeng, LIAO Guixiang, et al. Research status of methods for mapping forested wetlands based on remote sensing[J]. Remote Sensing for Land & Resources, 2014, 26(2): 11–18. (in Chinese)
- 11 王书玉, 张羽威, 于振华. 基于随机森林的洪河湿地遥感影像分类研究[J]. 测绘与空间地理信息, 2014, 37(4): 83–85, 93.
WANG Shuyu, ZHANG Yuwei, YU Zhenhua. Classification of Honghe wetland remote sensing image based on random forests[J]. Geomatics & Spatial Information Technology, 2014, 37(4): 83–85, 93. (in Chinese)
- 12 ZHANG Caiyun, XIE Zhixiao. Object-based vegetation mapping in the Kissimmee River Watershed using HyMap data and machine learning techniques[J]. Wetlands, 2013, 33(2): 233–244.
- 13 赵发林, 张涛, 李康. 基于遗传算法的随机森林模型(GARF)用于特征基因筛选[C]//2011 年中国卫生统计学会会议论文集, 2011: 7.
- 14 宋羚. 基于多目标遗传算法和 SVM 的特征选择方法[D]. 武汉: 华中科技大学, 2007.
SONG Ling. A feature selection method based on multi-objective genetic algorithm and support vector machines[D]. Wuhan: Huazhong University of Science and Technology, 2007. (in Chinese)
- 15 裘国永, 王娜, 汪万紫. 基于互信息和遗传算法的两阶段特征选择方法[J]. 计算机应用研究, 2012, 29(8): 2903–2905.
QIU Guoyong, WANG Na, WANG Wanzi. Two-stage feature selection algorithm based on mutual information and genetic algorithm[J]. Application Research of Computers, 2012, 29(8): 2903–2905. (in Chinese)
- 16 刘航宏. 黑龙江南瓮河国家级自然保护区大型真菌多样性研究[D]. 长春: 吉林农业大学, 2013.
LIU Hanghong. Study on diversity of macrofungi in Nanwenghe national nature reserve[D]. Changchun: Jilin Agricultural University, 2013. (in Chinese)
- 17 裴亚军. 滇东南石漠化多尺度遥感监测的精度评价研究[D]. 昆明: 昆明理工大学, 2014.
- 18 MILLARD K, RICHARDSON M. On the importance of training data sample selection in random forest image classification; a case study in Peatland ecosystem mapping[J]. Remote Sensing, 2015, 7(7): 8489–8515.
- 19 刘金锋. 基于多源遥感数据的青海湖流域植被指数研究[D]. 西宁: 青海师范大学, 2014.
LIU Jinfeng. Vegetation indices based on multi-source remote sensing data of the Qinghai Lake basin[D]. Xining: Qinghai Normal University, 2014. (in Chinese)
- 20 朱旭珍. 三种不同方法估算森林叶面积指数的比较研究[D]. 临安: 浙江农林大学, 2014.
ZHU Xuzhen. A comparative study of three different methods for estimating forest leaf area index[D]. Lin'an: Zhejiang A&F University, 2014. (in Chinese)
- 21 李子扬, 钱永刚, 申庆丰, 等. 基于高光谱数据的叶面积指数遥感反演[J]. 红外与激光工程, 2014, 43(3): 944–949.
LI Ziyang, QIAN Yonggang, SHEN Qingfeng, et al. Leaf area index retrieval from remotely sensed hyperspectral data[J]. Infrared and Laser Engineering, 2014, 43(3): 944–949. (in Chinese)
- 22 朱绪超, 袁国富, 易小波, 等. 基于 Landsat 8 OLI 影像的塔里木河下游河岸林叶面积指数反演[J]. 干旱区地理, 2014, 37(6): 1248–1256.
ZHU Xuchao, YUAN Guofu, YI Xiaobo, et al. Leaf area index inversion of riparian forest in the lower basin of Tarim River based on Landsat 8 OLI images[J]. Arid Land Geography, 2014, 37(6): 1248–1256. (in Chinese)
- 23 任安才. 基于 TM 影像的川西北理塘草地生物量与植被指数关系研究[D]. 雅安: 四川农业大学, 2008.
REN Ancai. Grassland biomass on north-western Plateau of Sichuan and vegetation indexes relation using Landsat TM image[D]. Ya'an: Sichuan Agricultural University, 2008. (in Chinese)
- 24 樊辉. 基于 Landsat TM 的城市热岛效应与地表特征参数稳健关系模型[J]. 国土资源遥感, 2008, 19(3): 45–51.
FAN Hui. The robust linear regression model between satellite-derived urban heat island and underlying surface parameters[J]. Remote Sensing for Land & Resources, 2008, 19(3): 45–51. (in Chinese)
- 25 罗崇亮. 基于水体指数的艾比湖湖水面积提取对比研究[J]. 科技创新导报, 2015(24): 34–35.
- 26 李梦云. 基于 VSDI 指数的土壤湿度遥感降尺度研究[D]. 长春: 东北师范大学, 2014.
LI Mengyun. Downscaling of passive microwave soil moisture using visible and shortwave infrared drought index: a case study of the western Jilin province[D]. Changchun: Northeast Normal University, 2014. (in Chinese)
- 27 ZHANG Ning, YANG Hong, QIN Qiming, et al. Evaluation of the visible and shortwave infrared drought index in China[J]. International Journal of Disaster Risk Science, 2013, 4(2): 68–76.
- 28 邢丽玮, 李小娟, 李昂晟, 等. 基于高光谱与多光谱植被指数的洪河沼泽植被叶面积指数估算模型对比研究[J]. 湿地科学, 2013, 11(3): 313–319.
XING Liwei, LI Xiaojuan, LI Angsheng, et al. A comparative study on estimation model for leaf area index of vegetation in marshes in Honghe National Nature Reserve based on hyperspectral and multispectral vegetation indices[J]. Wetland Science, 2013, 11(3): 313–319. (in Chinese)
- 29 康峰峰. 北京西南山地森林绿量遥感反演的研究[D]. 北京: 北京林业大学, 2011.
KANG Fengfeng. Study on remote sensing retrieval of forest vegetation quantity in mountainous area south-western Beijing China[D]. Beijing: Beijing Forestry University, 2011. (in Chinese)
- 30 李晓岚. 基于 Relief 特征选择算法的研究与应用[D]. 大连: 大连理工大学, 2013.
LI Xiaolan. The study and application of feature selection algorithms based on Relief[D]. Dalian: Dalian University of Technology, 2013. (in Chinese)
- 31 马玥, 姜琦刚, 孟治国, 等. 基于随机森林算法的农耕区土地利用分类研究[J/OL]. 农业机械学报, 2016, 47(1): 297–303.
http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?file_no=20160140&flag=1. DOI:10.6041/j.issn.1000-1298.2016.01.040.
MA Yue, JIANG Qigang, MENG Zhiguo, et al. Random forest classification of land use in farming area[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(1): 297–303. (in Chinese)