

doi: 10.6046/zrzyg.2022272

引用格式: 陈健,李虎,刘玉锋,等. 基于 Sentinel-2 数据多特征优选的农作物遥感识别研究[J]. 自然资源遥感,2023,35(4): 292-300. (Chen J, Li H, Liu Y F, et al. Crops identification based on Sentinel-2 data with multi-feature optimization[J]. Remote Sensing for Natural Resources, 2023, 35(4): 292-300.)

基于 Sentinel-2 数据多特征优选的农作物遥感识别研究

陈健^{1,3}, 李虎^{1,3}, 刘玉锋², 常竹^{1,3}, 韩伟杰^{1,3}, 刘赛赛²

(1. 安徽师范大学地理与旅游学院, 芜湖 241003; 2. 滁州学院计算机与信息工程学院, 滁州 239000; 3. 资源环境与地理信息工程安徽省工程技术研究中心, 芜湖 241003)

摘要: 本研究以滁州市全椒县为研究区, 借助 GEE 平台, 基于 Sentinel-2 卫星数据构建光谱特征、传统植被指数特征、红边植被指数特征、纹理特征等 90 个特征, 选用基于随机森林的递归特征消除算法 (random forest - recursive feature elimination, RF_RFE)、基于 Relief 拓展的 Relief F 算法、基于相似性的特征优选算法 (correlation - based feature selection, CFS) 结合随机森林分类器对农作物种植类型进行识别, 探究不同的特征优选算法在农作物遥感识别中的效果优劣。在此基础上, 进一步分析最佳特征优选算法在不同机器学习分类方法中的分类效果。研究表明: ①光谱特征在农作物识别中最为重要, 其次是红边指数特征, 纹理特征影响较小; ②基于 RF_RFE 特征优选方法的遥感识别结果精度最好, 总体精度为 92%, Kappa 系数为 0.89; ③在 RF_RFE 特征优选方法下, 随机森林 (random forest, RF) 的 Kappa 系数比支持向量机分类 (support vector machine, SVM) 和最小距离分类 (minimum distance classification, MDC) 分别高 0.01 和 0.41, 说明基于多特征的 RF_RFE 特征优选方法结合 RF 算法可以有效提高农作物遥感识别精度和效率。

关键词: Google Earth Engine; Sentinel-2; 农作物识别; 特征优选; 随机森林

中图法分类号: TP 79 **文献标志码:** A **文章编号:** 2097-034X(2023)04-0292-09

0 引言

农作物种植面积是农情监测和农业种植结构调整的重要基础数据。传统农作物种植面积统计通常采用抽样和逐级上报的方式进行, 不仅费时费力, 还受到一定的主观因素影响^[1]。卫星遥感技术能够实现大面积、长时间的对地观测, 可以在短时间内客观、准确地获取农作物的分布信息, 是目前农业遥感研究的一个热点^[2-3]。

由于存在“同物异谱、同谱异物”的现象, 单一时相的遥感数据往往难以准确反映农作物的分布情况^[4]。学者们选择从多时相数据源入手, 根据作物在遥感影像上的表征随时间变化的特性实现作物分

类, 如王德军等^[5]、杨欢等^[6]通过构建能够反映作物物候信息的时间序列植被指数, 有效识别出了不同农作物。遥感分类过程中, 如果将所有时相的特征都参与计算容易弱化分类器性能, 造成维度灾难和信息冗余, 从而降低精度和效率。因此, 如何从多时相、多特征组合中优选出最佳特征是解决农作物分类问题的关键所在。

近年来, 特征优选在农作物遥感分类中得到了不同程度的应用, 为更好地识别出农作物提供了可能。RF_RFE 由于能够量化每个变量的相对重要性, 通常被用于高维的特征优选。梁继等^[7]采用 RF_RFE 算法分析了不同特征对农作物识别的重要程度, 筛选出有利于农作物识别的特征, 提高了农作物分类的精度和效率。周小成等^[8]使用 RF_RFE 算

收稿日期: 2022-06-26; 修订日期: 2023-02-16

基金项目: 高分辨率对地观测系统科技重大专项 (编号: 76-Y50G14-0038-22/23)、安徽省科技重大专项“现代农业遥感监测系统构建与产业化应用” (编号: 202003A06020002)、滁州市科技计划项目“智慧农业遥感信息服务关键技术研究” (编号: 2021ZD013)、滁州市科技计划项目“滁州市农业气象干旱灾害遥感监测与预警” (编号: 22021ZD015)、安徽省重点研究与开发计划项目“国产陆地卫星星群图像智能处理关键技术及应用” (编号: 2021003)、安徽高等学校自然科学研究重点项目“基于国产高分卫星遥感数据的农作物种植结构信息提取关键技术研究” (编号: KJ2021A1063) 及安徽省特支计划 (编号: 2019) 共同资助。

第一作者: 陈健 (1998-), 男, 硕士研究生, 主要研究方向为现代农业遥感应用。Email: 599776160@qq.com。

通信作者: 刘玉锋 (1981-), 男, 副教授, 主要研究方向为资源遥感监测领域的应用研究。Email: Liuyufeng@chzu.edu.cn。

法优选出有利于林地信息提取的特征,实现了基于无人机数据对竹林、针叶林和阔叶林的分类识别。Relief F 是经典的过滤式特征选择算法,在特征优选中具有运算速度快、泛化能力强等特点。刘家福等^[9]基于 Landsat8 数据研究提取滨海湿地信息时,发现基于 Relief F 算法结合随机森林(random forest,RF)分类器具有高精度、高效率的优势。刘莹等^[10]通过 Relief F 算法对 Landsat8 数据的 71 个特征进行特征优选,然后结合 SVM 分类器有效的识别出城市的不透水面覆盖面积。张东彦等^[11]在对安徽北部平原的大豆区进行提取时,使用 Relief F 算法提取特征,并对比了 RF、反向传播神经网络(back-propagation,BP)、支持向量机(support vector machine,SVM)等 3 种分类器的分类效果,结果表明 RF 分类器结合 Relief F 算法的识别结果更加接近真实情况。基于相似性的特征优选(correlation-based feature selection,CFS)算法是用来综合评价特征与分类结果之间的相关性和特征之间冗余度的方法。张文博等^[12]对比了 CFS 算法和 Relief F 算法提取旱区植被的分类效果,结果表明 CFS 算法能够在保证较少特征的同时具备较高的分类精度。

综上所述,在基于遥感数据多特征优选的分类

识别研究中,分类特征的选择固然重要,但还存在盲目性和不稳定性,对不同的地物识别效果差异也较大,不同的特征优选方法适用于何种遥感数据和分类场景尚未形成统一的定论。本研究以滁州市全椒县为研究区,借助 GEE 平台,基于 Sentinel - 2 卫星数据构建多时相多维遥感特征,选用 RF_RFE,Relief F,CFS 这 3 种特征优选算法,探究不同的特征优选算法在农作物分类中的效果优劣,并确定可用于农作物种植类型识别的最佳特征优选方法;在此基础上,通过与其他分类方法比较,进一步探究最佳特征优选算法在不同分类器中的识别农作物种植结构的性能差异。

1 研究区概况与数据源

1.1 研究区概况

全椒县地处安徽省滁州市的最南端,E117°48'~118°24'、N31°50'~32°14'之间。位于江淮分水岭和滁河之间,为亚热带季风气候,年平均气温范围为 11.4~16.6℃。北部为山区,海拔最高 396 m,南部为平原带,东部为主城区,上半年以种植小麦、油菜为主,下半年以种植水稻为主,地理位置及样本点分布如图 1 所示。

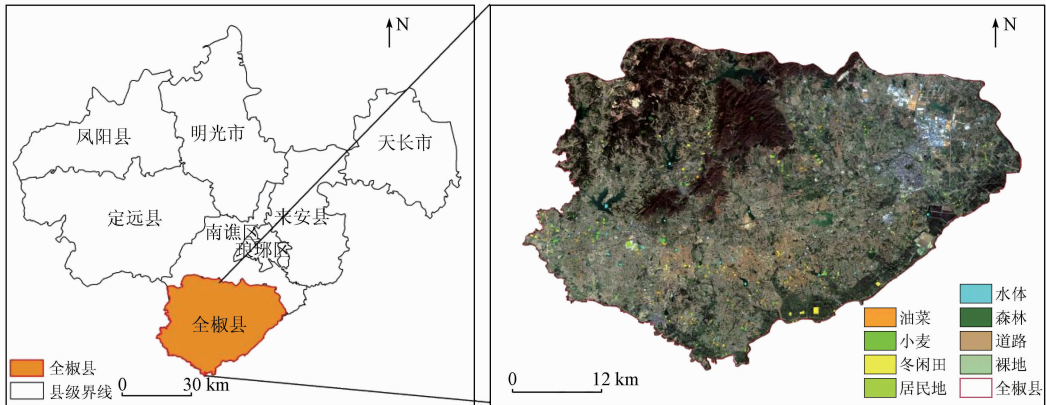


图 1 研究区地理位置及其样本分布

Fig.1 Geographical location of the study area and its sample distribution

1.2 数据源与预处理

1.2.1 Sentinel - 2 影像数据

相较于 GF - 2 卫星数据和 Landsat 卫星数据, Sentinel - 2 卫星数据具有重访周期短、分辨率高的综合优势,且具备对农作物分类有重要影响的红边

特征。依据全椒县农作物的物候特征(表 1)以及相关的前期研究,选择农作物的返青期 2022 年 2 月 25 日、拔节期 2022 年 3 月 12 日、孕穗期 2022 年 4 月 21 日共 3 景 Sentinel - 2 多光谱数据作为数据源。在 GEE 中选择的 Sentinel - 2 数据是经过辐射定标

表 1 全椒县午季农作物生长周期

Tab.1 The growth cycle of mid-season crops in Quanjiao County

时间	10 月			11 月			12 月			来年 1 月			2 月			3 月			4 月			5 月			6 月		
	上	中	下	上	中	下	上	中	下	上	中	下	上	中	下	上	中	下	上	中	下	上	中	下	上	中	下
小麦	播种			出苗			分蘖			越冬			返青			拔节			孕穗			成熟					
油菜	播种			移栽			越冬						现蕾			初花			中花			成熟					
冬闲田	水稻成熟												空闲									水稻播种			移栽		

和几何校正的 Level-1C 产品,其中包含 13 个光谱波段,本文剔除掉气溶胶、水蒸气以及卷云波段这 3 个无关波段,并利用 QA60 波段进行去云掩模操作,除去卷云和厚云的影响,最终得到 3 个时相的全椒县 Sentinel-2 无云影像。

1.2.2 样本数据

本文根据全椒县实地调查情况,将研究区土地覆盖类型分为居民地、道路、裸地、冬闲田、小麦、油菜、水体和森林 8 类,并于 2022 年 2 月 24 日—2022 年 2 月 26 日对研究区开展地面调查,获取各种地物样本以及农作物的类型、种植结构、地理位置并拍照记录。共获得 512 个样本数据,其中居民地 68 个、道路 23 个、裸地 27 个、冬闲田 76 个、小麦 105 个、油菜 148 个、水体 26 个、林地 39 个,具体的样本分布情况如图 1 所示。本文将样本数据按照 7:3 的比例随机分为训练样本和验证样本。

2 研究方法

基于 GEE 平台提取全椒县农作物种植类型的具体流程如图 2 所示。首先通过 GEE 平台获取并处理覆盖全椒县的 Sentinel-2 数据,从中提取出光谱特征、纹理特征、植被指数特征。然后采用 RF_RFE,Relief F,CFS 对多时相、多维遥感特征进行优选,分析不同特征在农作物识别中的重要性程度,并采用混淆矩阵方法评估分类精度确定最佳的特征优选算法。在此基础上,对比 RF,SVM、最小距离分类(minimum distance classification,MDC)3 种分类方法,探究最佳特征优选算法在不同分类器中的性能差异,采用全国第二次土地调查数据中的耕地范围对分类结果做掩模处理,并将非农作物类型合并在一起,只分析农作物的识别结果。

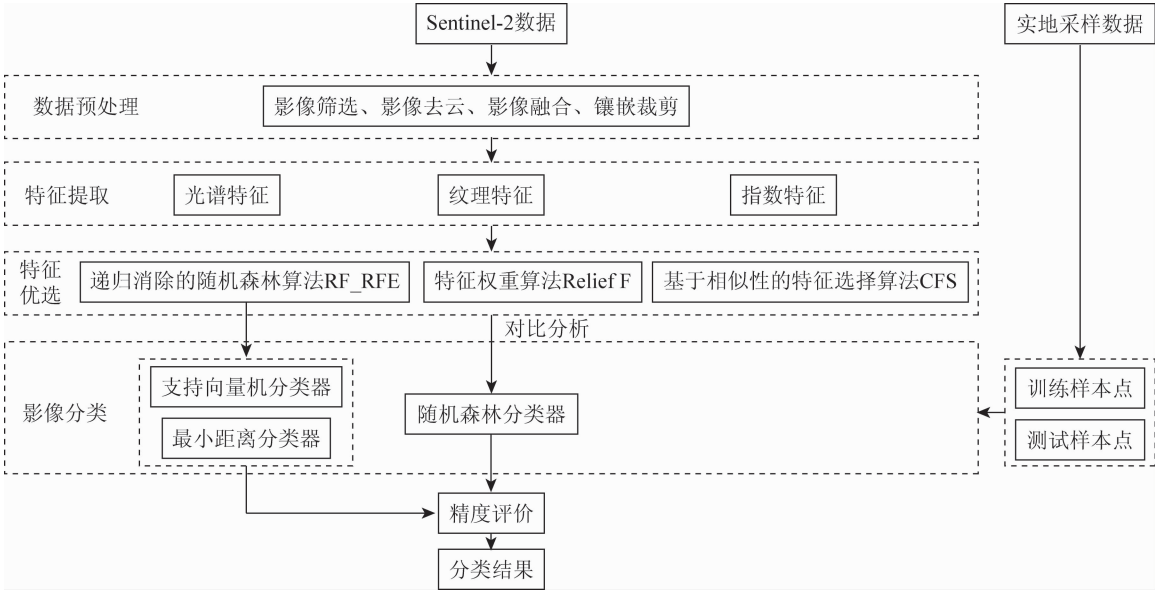


图 2 研究区农作物信息提取的技术流程图

Fig. 2 Technical flow chart of crop information extraction in the study area

2.1 特征集构建

为了分析不同的特征对农作物遥感识别的影响,提高农作物的识别精度,本文选择光谱特征、植

被指数特征、纹理特征等共计 90 个特征构成数据集,详见表 2。表中,光谱特征直接选择每个时相影像的 10 个原始波段,3 个时相的影像共计 30 个光

表 2 农作物遥感识别特征集

Tab. 2 Crop remote sensing identification feature set

特征变量	指数	特征公式及说明	参考文献	特征数目
光谱特征	B *	B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12		30
传统植被指数	归一化植被指数 (NDVI)	$NDVI = (B8 - B4) / (B8 + B4)$	Ni 等 ^[13]	18
	陆地表面水分指数 (LSWI)	$LSWI = (B8 - B11) / (B8 + B11)$	Ni 等 ^[13]	
	增强型植被指数 (EVI)	$EVI = 2.5 (B8 - B4) / (B8 + 6B4 - 7.5B2 + 1)$	Ni 等 ^[13]	
	土壤调整植被指数 (SAVI)	$SAVI = (B7 - B3) / (B7 + B3 - 0.5) * (1 + 0.5)$	王李娟等 ^[14]	
	比值植被指数 (RVI)	$RVI = B7 / B3$	王李娟等 ^[14]	
	归一化差异耕作指数 (NDTI)	$NDTI = (B11 - B12) / (B11 + B12)$	熊皓丽等 ^[15]	

(续表)

特征变量	指数	特征公式及说明	参考文献	特征数目
红边指数特征	归一化植被指数红边 1 (NDVI _{re1})	$NDVI_{re1} = (B8A - B5) / (B8A + B5)$	张磊等 ^[16]	18
	归一化植被指数红边 2 (NDVI _{re2})	$NDVI_{re2} = (B8A - B6) / (B8A + B6)$	张磊等 ^[16]	
	归一化植被指数红边 3 (NDVI _{re3})	$NDVI_{re3} = (B8A - B7) / (B8A + B7)$	张磊等 ^[16]	
	新型倒红边叶绿素指数 (IRECI)	$IRECI = (B7 - B4) / (B5 / B6)$	熊皓丽等 ^[15]	
	红边叶绿素指数 (Cire)	$Cire = B7 / B5 - 1$	王李娟等 ^[14]	
	地面叶绿素指数 (MTCI)	$MTCI = (B6 - B5) / (B5 - B4)$	熊皓丽等 ^[15]	
纹理特征	Contrast	对比度	张磊等 ^[16] 、熊皓丽等 ^[15]	24
	Asm	角二阶矩		
	Corr	相关性		
	Idm	逆差距		
	Ent	熵		
	Var	方差		
	Dvar	差方差		
	Diss	不相似性		
总计				90

谱特征。植被指数中,红边波段是哨兵数据特有的,且红边指数特征对植被更加敏感。本文不仅考虑常规植被指数,而且加入了 6 个与农作物生长过程有关的红边指数,分 3 个时相共计 18 个无红边植被特征和 18 个红边植被指数特征。

纹理特征^[17]的构建选择灰度共生矩阵 GLCM 来提取,GEE 平台上提供了基于 GLCM 快速计算纹理特征的函数,可以计算出 18 种不同的纹理特征,考虑到若将所有特征都用于分类必定会导致冗余。因此根据前人研究,通过对原始影像的 10 个波段进行主成分分析,提取主成分的第一成分来计算最常见的纹理指标用于区分不同地物的空间结构差异^[18],同样分 3 个时相共计 24 个纹理特征。

2.2 特征优选方法

2.2.1 基于 RF_RFE 算法的特征优选

基于 RF_RFE 算法的特征优选方法如下:首先,对于随机森林中的每一个决策树而言,使用相应的袋外数据来计算袋外误差,记作 err_{OOB1} ;然后,随机地对袋外数据 OOB 所有样本的特征 X 加入噪声干扰,再计算其袋外数据误差,记作 err_{OOB2} ;最后,计算特征重要性 X ,计算公式如下:

$$X = \frac{1}{N} \sum_{i=1}^N (err_{OOB2} - err_{OOB1}) \quad , \quad (1)$$

式中 N 为树的个数,式中 X 的值越高,则说明特征 X 越重要^[8]。把需要的特征子集初始化为整个特征子集,每次剔除掉重要性分数最低的数据,直到获得最后的特征集。

2.2.2 基于 Relief F 算法的特征优选

Relief F 算法是根据样本邻近原则,赋予不同特征的权重。具体步骤如下:①选择特征样本,从样本集中随机选择一个样本 X ,从不同类中找到一个最近邻的样本 Y ,从同类的样本中找到一个最近邻

的样本 Z 。②计算特征权重,在某个特征的条件下,比较 X 与 Y 之间的距离和 X 与 Z 之间的距离大小;如果不同类之间的距离较大,说明该特征容易区分,增加特征的权重;反之,则减少特征的权重。③重复上面的步骤,求取各个特征权重的平均值,特征权重越大代表该类的区分能力较强^[19]。

2.2.3 基于 CFS 算法的特征优选

基于相似性的特征选择是过滤式选择的一种的方法,该方法的核心是采用启发的方式评估特征子集的价值,该方法假设思想是:好的特征子集包含与类高度相关的特征,但特征之间彼此不相关。启发方程式为:

$$Merit_{tk} = \frac{kZ_{cf}}{\sqrt{k + k(k - 1) Z_{ff}}} \quad , \quad (2)$$

式中: $Merit_{tk}$ 为包含 k 个特征子集 t 的启发值; Z_{cf} 为特征与类之间的相关性的平均值; Z_{ff} 为特征与特征之间的相关性; Z 为相关系数。启发值越大代表这个特征的分类效果越好^[12]。

2.3 随机森林分类

随机森林分类是集成分类的一个子类,它依靠决策树投票选择来决定最后的分类结果,将若干个弱分类器的分类结果进行投票选择,从而组成强分类器。其具体操作步骤包括以下几个方面:首先,在原始的样本中,随机且有放回地抽取 N (约为总体样本集的 2/3) 个训练样本组成训练样本集,剩余的 1/3 样本作为袋外数据进行内部交叉验证;然后,根据抽取的样本集分别建立 N 棵决策树组成的随机森林,每个决策树随机抽取 M 个特征,采用基尼系数最小的原则进行节点分裂;最后,将生成的多颗决策树组成随机森林分类器,采用投票的方式决定新样本的类别^[20]。

2.4 精度评价方法

根据地面实测样本数据,采用混淆矩阵来评估不同模型的分类效果。使用制图精度来反映分类中的漏分误差、使用用户精度来反映分类中的错分误差。通过 Kappa 系数来评估不同分类模型的优劣,其计算公式如下:

$$Kappa = \frac{N \sum_{i=1}^m x_{ii} - \sum_{i=1}^m x_{i+} x_{j+}}{N^2 - \sum_{i=1}^m x_{i+} x_{j+}}, \quad (3)$$

式中: N 为像元的总数; m 为类别数; x_{ii} 为混淆矩阵对角线上的像元个数; x_{i+} 和 x_{j+} 分别为第 i 行和第 j 列的像元总数。

3 结果与分析

3.1 特征重要性分析

本文共选择了 90 个特征参与全椒县的农作物遥感提取,通过 GEE 平台以及分类样本的特征值,结合 RF_RFE,Relief F,CFS 这 3 种特征优选算法分别计算出相应的特征重要性。为了避免传统阈值方法的人为主观性影响,本文参考前人已有的研究^[18],根据特征重要性大小对不同特征进行降序排列,并从中选出前 50 个特征组成特征集进行实验。每次从特征集中删除一个特征重要性排在最后的特征,并将保留下来的特征子集用于农作物分类和计

算分类精度。重复以上过程,通过逐次迭代计算,从而最终确定特征优选结果变量数目与分类精度之间的关联关系。为了方便统一比较,这里都选择 RF 作为特征优选后特征子集的分类器,保留下来的特征子集需要确保分类精度较高但特征个数较少,不同特征子集的特征个数对应的 Kappa 系数见图 3。

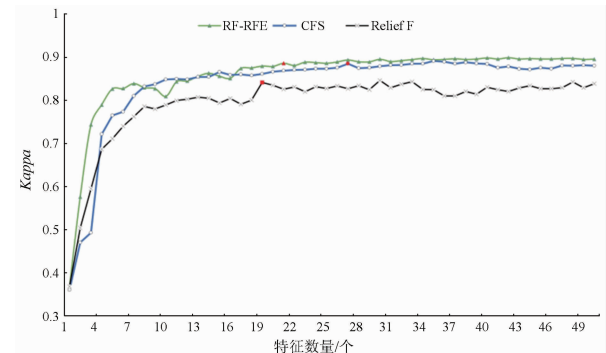


图3 不同特征数目下的 Kappa 系数
Fig.3 The Kappa coefficient for different numbers of features

由图 3 可知,随着特征参入分类的特征个数的增加,分类精度开始随着特征数的增加而逐渐增大,当达到一定值的时候,Kappa 系数出现小范围的波动。最终的 RF_RFE 算法、CFS 算法、Relief F 算法优选特征的 Kappa 系数分别维持在 0.89、0.88 和 0.83 左右波动。图中用红色标记了特征优选的最佳精度的位置,3 种特征优选算法的结果其对应的特征见表 3。

表 3 3 种优选结果的特征分布
Tab.3 Characteristic distribution of 3 optimization results

特征	2022/02/25	2022/03/12	2022/04/21	特征个数
Relief F	LSWI0225 , NDTI0225 , pc1_dvar0225, pc1_diss0225, pc1_contrast0225	NDTI0312 , NDVIre30312, LSWI031, pc1_dvar0312, pc1_diss0312, pc1_contrast0312	B110421 , NDVIre20421, NDVI0421, NDTI0421 , LSWI0421 , pc1_diss0421, pc1_dvar0421, pc1_cotrast0421	19
CFS	B30225, EVI0225, LSWI0225 , pc1_var0225, pc1_corr0225, MTCI0225	B30312 , B40312, B50312 , NDVIre30312, MTCI0312, EVI0312 , NDVIre20312, pc1_corr0312, pc1_contrast0312 , pc1_var0312	B50421 , B8A0421 , B110421 , LSWI0421, MTCI0421, B40421, NDTI0421 , EVI0421, NDVIre20421 , pc1_corr0421, NDVIre30421	27
RF_RFE	B20225, B110225, LSWI0225 , NDTI0225	B20312, B30312 , B50312 , B60312, B110312, pc1_idm0312, NDTI0312 , EVI0312	B20421, B30421, B50421 , B60421, B8A0421 , B70421, B110421 , B120421, NDVIre20421	21

从表 3 中可以看出(表中特征名以特征加时间命名,如 B20225 代表 2022 年 2 月 25 日影像的 B2 波段,同一个特征在不同的特征优选算法里面出现至少 2 次的用斜体显示),3 种特征优选算法在 4 月份的特征数量最多,是农作物提取的最佳时相,其原因是 4 月份是小麦和油菜的孕穗期和中花期,两者之间的形态和光谱反射都会产生较大的差异,易于辨识区分。其次是 3 月份,此时的小麦和油菜正处于拔节期和初花期,此阶段小麦和荒地之间有一定

的差异。2 月份的小麦和油菜正处于拔节期和现蕾期,小麦和油菜刚有成长,处于苗期,而树林处于长势茂盛期,易于区分,此时间段容易产生“同谱异物”的现象,对分类结果存在一定的干扰。表 3 中有 13 个特征至少被 2 种优选算法同时优选出来,表明这 13 个特征在分类中起到重要作用,有利于农作物的识别。这些重要的特征在 RF_RFE 算法、Relief F 算法、CFS 算法优选的特征集中分别占比 47.61%、42.10% 和 40.74%。因此,相比于 Relief F 和 CFS,

RF_RFE 在 Kappa 系数和优选出的特征稳定性上均有利于农作物的地物分类。

对于 RF_RFE 算法,优选出来的特征集的特征重要性得分如图 4 所示。在优选的特征中,原始光谱特征占比最多,21 个特征中包含 15 个原始光谱特征,其中的短波红外 B11 和 B12 共出现 4 次,表明短波红外的加入能够在一定程度上提高农作物的分类精度。其次,B5,B6,B7 相关的红边特征在农作物分类中也起到了重要的作用。此外,纹理特征中的逆差距有利于农作物的分类,展现出较好的分类效果。

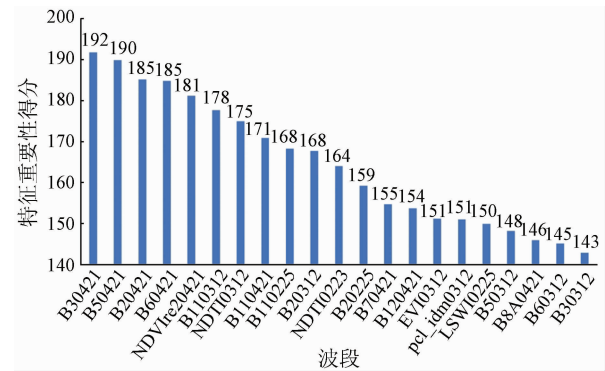


图 4 特征名称及其对应的重要性得分

Fig.4 The importance score for feature names and their correspondings

3.2 不同特征优选方法下 RF 分类精度对比

为了比较 3 种不同的特征优选方法在农作物分类中的效果,本文以 RF 分类器为例进行实验研究,表 4 展示了 3 种不同特征优选方法在 RF 下的分类结果,表中 3 种分类结果的 Kappa 系数均高于 0.83。RF_RFE 的分类精度最高,在特征数目为 21 时,总体精度为 92%,Kappa 系数为 0.89。其余的特征优选算法 Kappa 系数略低于 RF_RFE,从相同的 RF 分类器下,对比不同的特征优选方法来看,Relief F 在特征变量 19 个时,总体精度和 Kappa 系数分别为 0.83 和 0.88,相比于 RF_RFE 算法,特征维度减少了 2 个,但是总体精度和 Kappa 系数分别降低了 4% 和 5%。CFS 相比于 RF_RFE,在 Kappa 系数略微降低 0.01,同时在特征上多使用了 6 个特征参加计算。从用户精度和生产者精度来看,小麦均高于油菜,说明小麦的可分离性要优于油菜。在 RF_RFE 算法中,小麦和油菜的精度均达到了最大值,其中,小麦的生产者精度和用户精度分别为 96.2% 和 93.2%,油菜的生产者精度跟用户精度分别为 83.5% 和 88.8%。为了能够更加清楚地知道分类结果,本文将其与 2022 年 3 月 2 日的高分一号卫星多光谱与全色的融合影像进行对比(表 5)。

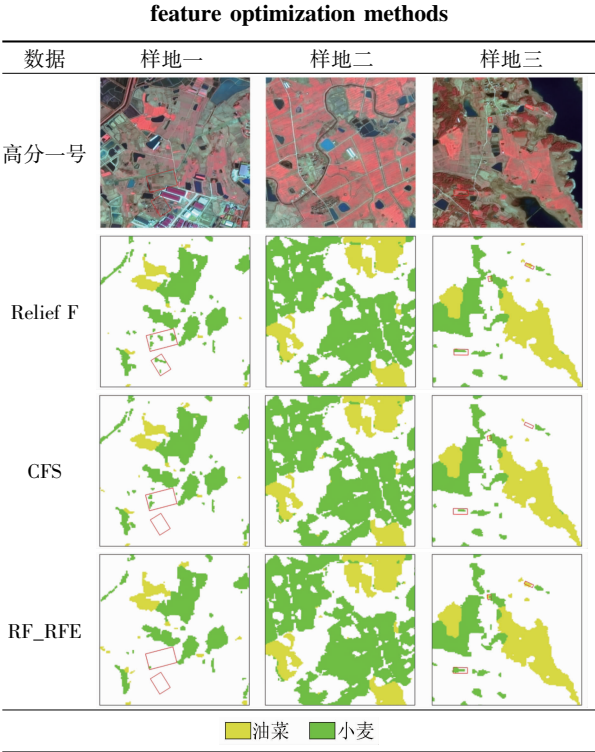
表 4 基于不同特征优选方法和随机森林的地物分类精度

Tab.4 Accuracy assessment of different feature optimization methods

不同的特征 优选方法	RF_RFE		CFS		Relief F	
	PA/%	UA/%	PA/%	UA/%	PA/%	UA/%
油菜	83.5	88.8	83.0	88.7	80.0	82.9
小麦	96.2	93.2	95.0	92.5	94.1	91.9
OA/%	91.7		91.5		87.71	
Kappa	0.89		0.88		0.83	

表 5 不同特征优选方法的局部结果图

Tab.5 Local result plots of different feature optimization methods



由表 5 可知,从高分一号融合影像上看,油菜和小麦在影像上呈现不一样的色调。从 RF 的 3 种分类方法的局部结果图中看,样地一表示破碎地块的农作物分类结果,从中可以看出 Relief F 出现错分,误分的情况较为严重,部分居民地旁边的冬闲田被错误识别为小麦,CFS 和 RF 出现这种现象的程度较轻。样地二表示地块较完整的农作物分类结果,3 种分类结果中有部分将田埂、道路错分为农作物的现象,导致这种现象的原因是 10 m 分辨率的 Sentinel - 2 数据存在混合像元现象。但从整体上来看 3 个分类结果都较为准确的识别出农作物,局部的差异较小。样地三表示林地附近的农作物分类结果,其中 Relief F 和 CFS 都出现漏分现象,对图中红色标记内的农作物,未能准确的识别处理,而 RF 能够较完整的识别农作物。综上所述,将 RF_RFE 作为最佳的优选特征,主要有以下 2 个原因: ①从分类的精度来看,RF_RFE 算法在不同的特征优选结果

中各项分类指标都最高,且特征个数也偏少;②从分类效果来看,分类结果中出现较少的错分或漏分的程度最轻。

3.3 不同机器学习分类对农作物识别精度的影响

基于 RF_RFE 特征优选算法,对不同分类方法的农作物识别精度进行比较,结果见图 5。与 SVM (图 5(a))和 RF(图 5(b))相比,MDC(图 5(c))的分类结果存在明显的差异,MDC 在分类过程中存在将冬闲田错分和误分为农作物的状况,并没有准确

的将农作物提取出来。而 SVM 与 RF 的分类结果较为接近,SVM 和 RF 的农作物主要分布在东北角的耕地区,西南角的耕地区农作物较少,这与西南角耕地区实施高标准农田建设有关。从分类器的机理来看,这种现象可能是由于本实验特征数量较多,MDC 处理大量的特征变量时,出现负载情况导致分类效果较差,而 SVM 和 RF 能够很好地使用小样本和特征进行分类。

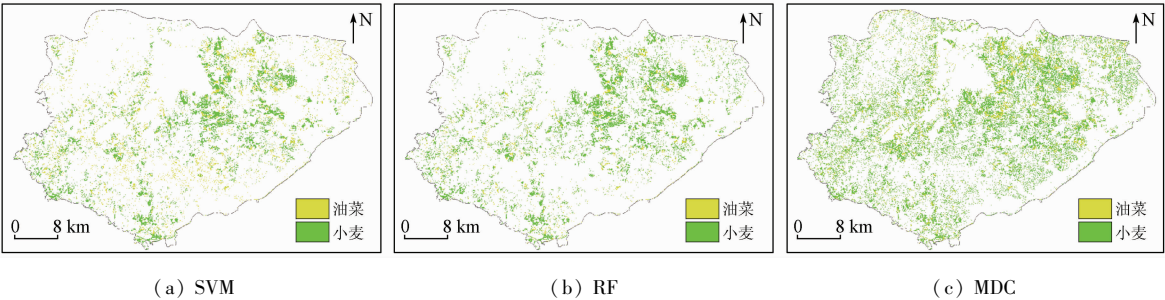


图 5 RF_RFE 下基于不同机器学习方法的研究区农作物空间分布

Fig.5 Spatial distribution of crops in the study area based on different machine learning methods under RF_RFE

对比不同分类方法的分类精度(表 6)可知,RF 的分类精度表现较为优越,总体精度比 SVM 和 MDC 分别高 0.7 个百分点和 30.5 百分点;Kappa 系数比 SVM 和 MDC 分别高 0.01 和 0.41,表明 SVM 和 RF 更适用于研究区的地物分类。在 RF 分类器下,小麦和油菜的用户精度和生产者精度均高于 83.5%。小麦的生产者精度和用户精度和油菜的用户精度均达到最大值 96.2% ,93.2% 和 88.8% ;在油菜的生产者精度略比 SVM 低 1.2 百分点,从数值上来看,除了 RF 在油菜的生产者精度小于 SVM 以外,其余的分类精度均大于 SVM。综上所述可知,3 种机器学习分类方法中,RF 在农作物信息提取的分类结果上与 SVM 分类结果相近,在总体精度上略高于 SVM,而 MDC 分类算法不适合高维度、大数据量的分类,对比可知,RF 分类算法可以有效的识别出农作物。

表 6 RF_RFE 特征优选下基于不同机器学习分类的农作物分类精度

Tab.6 RF_RFE characteristics are preferably based on different crop classification accuracy of machine learning algorithms

机器学习方法	RF		SVM		MDC	
评价指标	PA/%	UA/%	PA/%	UA/%	PA/%	UA/%
油菜	83.5	88.8	84.7	83.0	80.3	63.5
小麦	96.2	93.2	91.1	91.3	70.8	65.4
OA/%	91.7		91.0		61.2	
Kappa	0.89		0.88		0.48	

4 结论与讨论

4.1 结论

1)根据 RF_RFE 的特征重要性得分可知,不同的特征影响着农作物的识别精度,各类特征的得分值从高到低排列依次是光谱特征、红边特征、传统植被指数特征、纹理特征。其中短波红外波段 B11 和 B12,红边波段 B5 和 B6 对农作物的识别具有重要的作用。

2)对比 RF 分类器下的不同的特征优选方法的分类精度。基于 RF_RFE 的特征优选算法的分类精度最高,总体精度为 92% ,Kappa 系数为 0.88 ,且将特征维度从 90 维降低到 21 维。

3)基于相同的 RF_RFE 特征优选的条件下,RF 的 Kappa 系数比 SVM,MDC 分别高 0.01 和 0.41 ,在分类结果中,RF 相比于其他的分类,存在较少的错分和漏分现象。因此,RF 结合 RF_RFE 算法是适用于农作物种植结构信息提取的方法。

4.2 讨论

本文基于从 Sentinel - 2 卫星数据中提取的多时相、多维度遥感特征,采取不同的特征优选算法结合 RF 分类器实现对全椒县的农作物种植结构信息提取。选择有关农作物生理状况和形态结构的特征构造出多维特征,使用 RF_RFE,Relief F 和 CFS 计算出不同特征的重要性,依次消除特征重要性最小的特征,避免了采用传统的阈值方法在判定最佳维

度时存在的主观性。B3, B5, B8A, B11, NDVI, ND-TI, LSWI, EVI 和 pc1_contrast 在至少 2 种特征优选方法中出现过,表明红边、短波红外以及纹理特征在农作物种植结构识别中具有显著优势,这与文献[7,10,14]的研究结论一致。在特征子集相同的条件下,通过对比不同的机器学习分类方法,证明了 RF 结合 RF_RFE 算法在农作物种植结构的信息提取中的有效性。对于接下来的研究中,能否基于多时相多特征的特征优选算法,结合高分辨率遥感影像并推广应用到其他地区需要进一步探究。

参考文献 (References):

[1] 黄健熙,武思杰,刘兴权,等. 基于遥感信息与作物模型集合卡尔曼滤波同化的区域冬小麦产量预测[J]. 农业工程学报, 2012, 28(4): 142-148.

Huang J X, Wu S J, Liu X Q, et al. Regional winter wheat yield forecasting based on assimilation of remote sensing data and crop growth model with Ensemble Kalman method[J]. Transactions of the Chinese Society of Agricultural Engineering, 2012, 28(4): 142-148.

[2] 王利民,刘 佳,季富华. 中国农业遥感技术应用现状及发展趋势[J]. 中国农学通报, 2021, 37(25): 138-143.

Wang L M, Liu J, Ji F H. Application status and development trend of agriculture remote sensing technology application in China[J]. Chinese Agricultural Science Bulletin, 2021, 37(25): 138-143.

[3] 贾 坤,李强子. 农作物遥感分类特征变量选择研究现状与展望[J]. 资源科学, 2013, 35(12): 2507-2516.

Jia K, Li Q Z. Review of features selection in crop classification using remote sensing data[J]. Resources Science, 2013, 35(12): 2507-2516.

[4] 程 彬. 基于支持向量机的乾安县土地利用遥感分类研究[J]. 长春师范大学学报, 2017, 36(12): 86-88.

Cheng B. Land use information extraction based on support vector machine using multitemporal remote sensing in Qian'an County[J]. Journal of Changchun Normal University, 2017, 36(12): 86-88.

[5] 王德军,姜琦刚,李远华,等. 基于 Sentinel-2A/B 时序数据与随机森林算法的农耕地土地利用分类[J]. 国土资源遥感, 2020, 32(4): 236-243. doi:10.6046/gtzyyg. 2020. 04. 29.

Wang D J, Jiang Q G, Li Y H, et, al. Land use classification of farming areas based on time series Sentinel-2A/B data and random forest algorithm[J]. Remote Sensing of Land and Resources, 2020, 32(4): 236-243. doi:10.6046/gtzyyg. 2020. 04. 29.

[6] 杨 欢,邓 帆,张佳华,等. 基于 MODIS EVI 的江汉平原油菜和冬小麦种植信息提取研究[J]. 国土资源遥感, 2020, 32(3): 208-215. doi:10.6046/gtzyyg. 2020. 03. 27.

Yang H, Deng F, Zhang J H, et, al. A study of information extraction of rape and winter wheat planting in Jianghan Plain based on MODIS EVI[J]. Remote Sensing of Land and Resources, 2020, 32(3): 208-215. doi:10.6046/gtzyyg. 2020. 03. 27.

[7] 梁 继,郑镇炜,夏诗婷,等. 高分六号红边特征的农作物识别与评估[J]. 遥感学报, 2020, 24(10): 1168-1179.

Liang J, Zheng Z W, Xia S T, et al. Crop recognition and evaluation

using red edge features of GF-6 satellite[J]. Journal of Remote Sensing, 2020, 24(10): 1168-1179.

[8] 周小成,郑 磊,黄洪宇. 基于多特征优选的无人机可见光遥感林分类型分类[J]. 林业科学, 2021, 57(6): 24-36.

Zhou X C, Zheng L, Huang H Y. Classification of forest stand based on multi-feature optimization of UAV visible light remote sensing[J]. Forestry Science, 2021, 57(6): 24-36.

[9] 刘家福,李林峰,任春颖,等. 基于特征优选的随机森林模型的黄河口滨海湿地信息提取研究[J]. 湿地科学, 2018, 16(2): 97-105.

Liu J F, Li L F, Ren C Y, et al. Information extraction of coastal wetlands in Yellow River Estuary by optimal feature-based random forest model[J]. Journal of Wetland Science, 2018, 16(2): 97-105.

[10] 刘 莹,孟庆岩,王永吉,等. 基于特征优选与支持向量机的不透水面覆盖度估算方法[J]. 地理与地理信息科学, 2018, 34(1): 24-31, 3.

Liu Y, Meng Q Y, Wang Y J, et, al. A method for estimating impervious surface percentage based on feature optimization and SVM[J]. Geography and Geoinformatics Science, 2018, 34(1): 24-31, 3.

[11] 张东彦,杨玉莹,黄林生,等. 结合 Sentinel-2 影像和特征优选模型提取大豆种植区[J]. 农业工程学报, 2021, 37(9): 110-119.

Zhang D Y, Yang Y Y, Huang L S, et, al. Extraction of soybean planting areas combining Sentinel-2 images and optimized feature model[J]. Transactions of the Chinese Society of Agricultural Engineering, 2021, 37(9): 110-119.

[12] 张文博,孔金玲,杨园园,等. 面向对象的旱区植被遥感精细分类研究[J]. 测绘科学, 2021, 46(1): 136-140, 183.

Zhang W B, Kong J L, Yang Y Y, et al. Fine vegetation classification of remote sensing in arid areas based on object-oriented method[J]. Science of Surveying and Mapping, 2021, 46(1): 136-140, 183.

[13] Ni R, Tian J, Li X, et al. An enhanced pixel-based phenological feature for accurate paddy rice mapping with Sentinel-2 imagery in Google Earth Engine[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 178, 282-296.

[14] 王李娟,孔钰如,杨小冬,等. 基于特征优选随机森林算法的农耕地土地利用分类[J]. 农业工程学报, 2020, 36(4): 244-250.

Wang L J, Kong Y R, Yang X D, et, al. Classification of land use in farming areas based on feature optimization random forest algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering, 2020, 36(4): 244-250.

[15] 熊皓丽,周小成,汪小钦,等. 基于 GEE 云平台的福建省 10 m 茶园专题空间分布制图[J]. 地球信息科学学报, 2021, 23(7): 1325-1337.

Xiong H L, Zhou X C, Wang X Q, et al. Mapping the spatial distribution of tea plantations with 10 m resolution in Fujian Province using google earth engine[J]. Journal of Geoinformatics Science, 2021, 23(7): 1325-1337.

[16] 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, 2019, 23(2): 313-326.

Zhang L, Gong Z N, Wang Q W, et al. Wetland mapping of Yellow River Delta wetlands based on multi – feature optimization of Sentinel – 2 images[J]. Journal of Remote Sensing, 2019, 23(2) : 313 – 326.

[17] Haralick R M, Shanmugam K, Dinstein I H. Textural features for image classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1973, 3(6) : 610 – 621.

[18] 王庚泽, 靳海亮, 顾晓鹤, 等. 基于改进分离阈值特征优选的秋季作物遥感分类[J]. 农业机械学报, 2021, 52(2) : 199 – 210.

Wang G Z, Jin H L, Gu X H, et al. Remote sensing classification of autumn crops based on hybrid feature selection model combining with relief F and improved separability and thresholds[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(2) : 199 – 210.

[19] 高国龙, 杜华强, 韩 凝, 等. 基于特征优选的面向对象毛竹林分布信息提取[J]. 林业科学, 2016, 52(9) : 77 – 85.

Gao G L, Du H Q, Han N, et al. Mapping of moso bamboo forest using object – based approach based on the optimal features[J]. Forestry Sciences, 2016, 52(9) : 77 – 85.

[20] 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.

Crops identification based on Sentinel – 2 data with multi – feature optimization

CHEN Jian^{1,3}, LI Hu^{1,3}, LIU Yufeng², CHANG Zhu^{1,3}, HAN Weijie^{1,3}, LIU Saisai²

(1. College of Geography and Tourism, Anhui Normal University, Wuhu 241003, China; 2. College of Computer and Information Engineering, Chuzhou University, Chuzhou 239000, China; 3. Engineering Technology Research Center of Resources Environment and GIS, Wuhu 241003, China)

Abstract: Focusing on Quanjiao County in Chuzhou City, this study determined 90 features, including spectral, traditional vegetation index, red – edge vegetation index, and texture features, from Sentinel – 2 satellite data on the GEE platform. This study examined the effects of diverse feature optimization algorithms combined with a random forest classifier on identifying crop planting types in the study area. These algorithms included the random forest – recursive feature elimination (RF_RFE) algorithm, the Relief F algorithm based on Relief expansion, and the correlation – based feature selection (CFS) algorithm. On this basis, this study further analyzed the classification effects of the optimal feature optimization algorithm in various machine learning classification approaches. The study demonstrates that: ① Spectral features proved to be the most crucial for crop identification, followed by red – edge index features, and texture features manifested minimal effects; ② RF_RFE – based remote sensing identification results exhibited the highest accuracy, with overall accuracy of 92% and a Kappa coefficient of 0.89; ③ Under the RF_RFE feature optimization method, the RF’s Kappa coefficient was 0.01 and 0.41 higher than that of the support vector machine (SVM) and the minimum distance classification (MDC), respectively. This indicates that the RF_RFE feature optimization method based on multiple features, combined with the RF algorithm, can effectively enhance the accuracy and efficiency of remote sensing identification of crops.

Keywords: Google Earth Engine; Sentinel – 2; crop identification; feature optimization; random forest

(责任编辑: 李 瑜)