文章编号:1000-7601(2020)03-0281-08 doi:10.7606/j.issn.1000-7601.2020.03.37

基于时序 Landsat 8 OLI 多特征与随机森林算法的作物精细分类研究

刘 杰[1,2],刘吉凯[3],安晶晶[1,2],章 超[3]

(1.淮河流域气象中心,安徽 合肥 230031;2.安徽省气象台,安徽 合肥 230031;3.安徽科技学院资源与环境学院,安徽 凤阳 233100)

摘 要:利用新疆阿克苏地区温宿县 2014—2015 年生长季的 7 景时序 Landsat 8 OLI 数据,提取光谱特征、纹理特征、植被指数等高维信息,并基于随机森林算法(Random Forest, RF)构建分类模型。分析了 RF 模型中重要参数树个数 k、节点分裂特征个数 m 对分类精度的影响,计算 GINI 系数评估所有特征重要性,探索最佳特征子集,完成模型的参数率定与信息冗余消除,实现了温宿县研究区内的多种作物类型精细分类,并对比分析了随机森林与其他几种机器学习算法的分类性能。结果表明:作物分类的 3 类特征中,重要性排名靠前的分别是影像纹理平均规则程度 Mean、与作物水分含量密切相关的地表水分指数(LSWI)及短波红外波段光谱反射率,对应干旱区作物的 2 个关键时相:生长旺盛期与播种期;随机森林分类精度受分类特征数量的影响。当特征删除量低于总特征数的 30%时,RF 模型的分类精度基本保持不变;当删除量超过 70%时,分类精度下降的幅度加大;随机森林方法相对于决策树、支持向量机、朴素贝叶斯、K-近邻等监督分类算法,无论是分类结果的精度还是分类效率均具有优势。

关键词:随机森林算法;作物分类;时序 Landsat 8 OLI;特征重要性;新疆

中图分类号:S127 **文献标志码**:A

Precise crop classification based on multi-features from time-series Landsat 8 OLI images and Random Forest Algorithm

LIU Jie[1,2], LIU Jikai[3], AN Jingjing[1,2], ZHANG Chao[3]

(1. *Huaihe River Basin Meteorological Center*, *Hefei*, *Anhui* 230031, *China*;
2. *Anhui Meteorological Observatory*, *Hefei*, *Anhui* 230031, *China*;
3. *College of Resource and Environment*, *Anhui Science and Technology University*, *Fengyang*, *Anhui* 233100, *China*)

Abstract: According to the 7-scene landsat 8 OLI data of the 2014-2015 growth season in Wensu County, Aksu region, Xinjiang, we extracted spectral features, texture features, vegetation index, and other high-dimensional information, and constructed a classification model based on the Random Forest (RF) Algorithm. The influence of the number of important parameter trees k and the number of node splitting features m in the RF model on classification accuracy was analyzed, and GINI coefficient was calculated to evaluate the importance of all features to explore the best feature subset. In this study, parameter calibration and information redundancy elimination of the model were completed, and the precise classification of various crop types from the study area of Wensu County was realized. The classification performance of RF and other machine learning algorithms was compared and analyzed. The results showed that among the three characteristics of crop classification, the most important ones were mean of image texture, land surface water index (LSWI) closely related to crop moisture content, and spectral reflectance of short-wave infrared, which were corresponding to two key time phases of crops in arid areas: growth period and sowing period. Moreover, the classification accuracy of RF was affected by the number of classification features. When the deletion amount of "feature" was less than 30% of the total feature number, the classification accuracy of RF model remains unchanged. When the deletion amount was more than 70%, the classification accuracy decline

收稿日期:2019-08-11 **修回日期**:2020-03-02

基金项目:淮河流域气象开放研究基金(HRM201606);安徽科技学院引进人才资助项目(ZHYJ201603)

作者简介:刘杰(1990-),男,江苏盐城人,硕士,研究方向为农业遥感应用。E-mail:maginas_nuist@163.com

通信作者:刘吉凯(1989-),男,安徽蚌埠人,助教,主要从事信息农业研究。E-mail:liujkahstu@163.com

increased. Finally, compared with decision tree, support vector machine, naive Bayes, k-nearest neighbor, and other supervised classification algorithms, Random Forest Algorithm had advantages in both accuracy and efficiency of classification results.

Keywords: Random Forest Algorithm; crops classification; time-series Landsat 8 OLI images; variable importance; Xinjiang

作物类别识别是农业遥感应用的重要方向,是农业精细化管理、农情监测的基础[1-2]。传统的作物信息获取主要以行政部门的地面抽样为主,费时费力,数据获取量少、分布离散,在国家、州、省等区域尺度推广应用的时效性差[3-4]。随着农业遥感技术的深入发展,基于不同分辨率的遥感影像可以快速、无损、实时地获取全球、区域、局部范围内的作物信息,对粮食估产、作物监测、作物生长周期模拟等研究具有重要意义[2-7]。

如何实现作物类别的精细分类已经成为当前农业遥感精细化、智能化发展的一个关键问题[6,8-9]。相对比传统的作物关键生育期遥感影像的信息识别,利用作物生育期的时间序列遥感影像可以最大限度地发现作物与背景干扰地物的差异,增加光谱特征相似的作物之间的分离[2,6-8,10]。充分利用不同作物生育期的物候信息,可以弥补“异物同谱、同物易谱”现象,能够较高精度获取作物信息,其与关键生育期单一时相影像相比识别的精度有较大提高[2,6-8,10]。

作物类别识别研究中,除原始影像的光谱特征被利用最多外,由影像波段经线性或非线性变换而来的植被指数能有效增加作物信息识别的效率,是农业遥感研究中至关重要的特征参数[2,8,10-11]。此外,研究表明作物的纹理特征作为影像空间特征的局部表示,对于作物类别及耕作方式等较为敏感,在作物精细分类中的应用日益深入[11-14]。

在中等分辨率层面(一般其空间分辨率在 10~250m 范围内),作物类别识别常用的方法主要是基于像元的分类方法,如最大似然法、支持向量机和决策树分类方法,它们简单易行、高效快捷,且精度有一定的保障,是目前国家级或区域级农情遥感监测业务化平台中的常见方法[1-5,11-12]。随着大量遥感卫星的发射,已经组建起完整的农业遥感观测系统[2,4-5,15],可用于作物类别识别的分类特征呈现出高维、异源、海量等特点,使得传统分类方法在数据处理效率、多源特征组合、数据深度挖掘等方面日益难以满足业务化运行的需求[2,8-9,13-15]。最近几年机器学习算法迅速发展,在处理多维复杂数据时展现出了更好的精度和效率。例如,神经网络[2,16-17]、决策树[2,7,10-12,15]、随机森林[9,13-15,18]、支持向量机[2,8,15-16]等。其中随机森林(Random Forest, RF)是一种具有优秀性能的集成学习算法,被广泛应用于复杂遥感数据集的分析处理[9,13-15,18-19]。

本文选择多时相 Landsat 8 OLI 数据提取研究区时序光谱特征、纹理特征、植被指数等信息,利用随机森林算法对分类特征进行降维以节约计算资源,获取最优参数与特征子集后对研究区农作物实现精细分类。同时评估基于随机森林多时相多特征类型的分类算法对农田作物的辨别能力,为农业遥感的进一步研究提供依据。

1 数据与方法

1.1 研究区概况

研究区主要位于新疆维吾尔自治区阿克苏地区温宿县西南部(见图 1),属典型的大陆性气候。区域内土地肥沃、水源丰富、光照充足、无霜期长,适宜各类农作物生长,是国家重要的商品粮、商品棉基地。

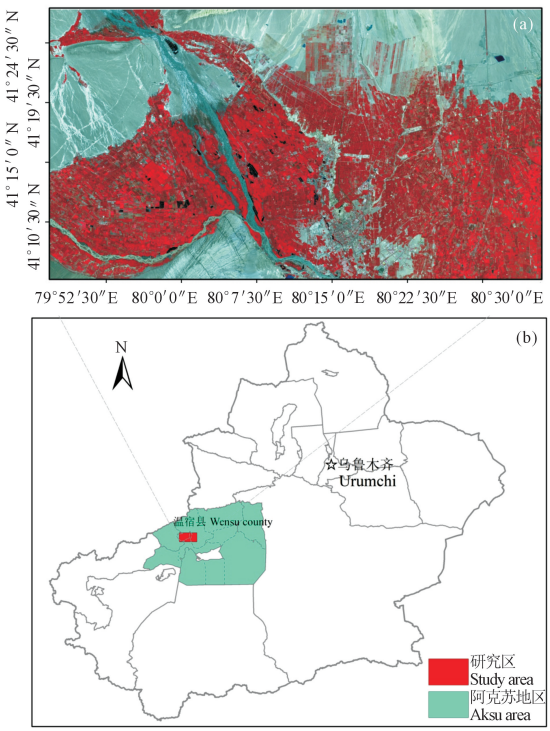


图 1 研究区 2015 年 8 月 14 日 Landsat 8 OLI 标准假彩色合成卫星影像(a)及其在新疆阿克苏地区的地理位置(b)

Fig.1 The Landsat 8 OLI false color image of study area on 14 August 2015 (a) and the location of study area in Aksu, Xinjiang (b)

通过对县域实地调查,研究区内农作物类型多样,种植复杂,物候期高度重叠。主要种植的农作物有水稻、棉花、春玉米和冬小麦等,主要的林果种类为枣树、核桃、苹果、香梨、葡萄等。

1.2　数据源与预处理

研究区属大陆干旱半干旱性气候区,在农作物生长周期内(4—11 月)的光学影像主要受沙尘影响。本文经筛选后共获取 2014—2015 年生长季的 7 景无云、无沙尘影响的 Landsat 8 业务化陆地成像仪(operational land imager, OLI)数据,成像时间分别为 2015 年 3 月 23 日、4 月 24 日、5 月 26 日、7 月 13 日、8 月 14 日、9 月 15 日和 10 月 17 日,格式为 L1T,多光谱空间分辨率为 30 m,下载至 USGS[20]。对所获取的多时相多光谱数据使用 ENVI5.5 软件分别进行了辐射定标、FLAASH 大气校正、几何配准(双线性内插,几何误差小于 0.5 个像元)处理,投影选择为 UTM(44N)/WGS-84。

结合研究区野外实测数据、Google Earth 高分辨率影像目视解译结果,将研究区分为棉花、小麦、玉米、水稻、香梨、核桃、苹果、葡萄、枣树、林地、草地、水域、沙地、戈壁和建筑共 15 个类别。野外实测主要是采用手持 GPS 仪获取农作物的解译样本标志,人工目视解译主要在高分 Google Earth 影像上根据实测解译标志勾绘研究区内的主要地物类别,利用 ENVI5.5 软件将矢量化结果转换至 OLI 影像,共选取了 667754 个像元点,按 7∶3 随机分为训练样本和验证样本。

1.3　分类特征处理

植被指数(vegetable indices, VIs),作为地表植被特征的重要表征参数,在植被长势、生物量、结构信息等应用中具有重要意义[2,8,10-11]。本文在波段反射率基础上提取 16 种常用于农作物信息识别的植被指数[21-23]。

纹理特征能够弥补基于像元光谱分类的不足,可以突出作物细节信息,是作物分类识别的常用特征之一[12-14]。本文对波段影像进行主成分变换,利用变换后的第一主成分(principal component analysis 1, PCA1)替代原始影像基于灰度共生矩阵(gray level co-occurrence matrix, GLCM)的方法进行纹理特征提取。

综上所述,本文根据研究区内主要农作物的波段反射率、植被指数特征和纹理特征共 217 个特征(见表 1),利用随机森林方法根据特征重要性选取最佳分类特征子集,并实现最优特征集支持下的农作物精细分类识别。

表 1　参与分类的所有特征

Table 1　Features used in the classifications

特征类别 Features type	特征参数 Features detail	个数 Number
波段反射率 Band reflectance	B1~B7 共 7 个波段地表反射率 Bands reflectance from OLI image (B1~B7)	7×7
植被指数* Vegetable indices	DVI、RVI、NDVI、GNDVI、EVI、TVI、GCI、GVI、NDWI、LSWI、MNDWI、MSAVI、NDSVI、NDTI、SAVI、STI	16×7
纹理特征 Textural features	均值(Mean)、方差(VAR)、同质性(HOM)、对比度(CON)、非相似性(DIS)、熵(ENT)、二阶矩(SM)、相关性(COR)	8×7

注:* DVI:差值植被指数;RVI:比值植被指数;NDVI:归一化差值植被指数;GNDVI:绿色归一化差值植被指数;EVI:增强型植被指数;TVI:三角植被指数;GCI:绿色叶绿素指数;GVI:绿色植被指数;LSWI:陆表水体指数;NDWI:归一化差值水体指数;MNDWI:修正的归一化差值水体指数;SAVI:土壤调节植被指数;MSAVI:修正的土壤调节植被指数;NDSVI:归一化差分凋零植被指数;NDTI:归一化差分耕作指数;STI:土壤耕作指数。

Note:* DVI: Difference vegetation index; RVI: Ratio vegetation index; NDVI: Normalized difference vegetation index; GNDVI: Green normalized difference vegetation index; EVI: Enhance vegetation index; TVI: Triangular vegetation index; GCI: Green chlorophyll index; GVI: Green vegetation index; LSWI: Land surface water index; NDWI: Normalized difference water index; MNDWI: Modified normalized difference water index; SAVI: Soil adjusted vegetation index; MSAVI: Modified soil adjusted vegetation index; NDSVI: Normalized difference senescent vegetation index; NDTI: Normalized difference tillage index; STI: Soil tillage index.

1.4　随机森林算法

2001 年,美国科学家 Breiman 提出了一种称为随机森林(RF)的新型分类算法[19]。它由多棵 CART 决策树分类器构成,能够高效处理多维特征的数据集,并具有准确性高、模型稳定等优点[9,13,18]。RF 通过 k 次 Bootstrap 随机有放回抽样,每次随机抽取约 2/3 的原始数据建立单棵决策树,形成 k 棵树组成的随机森林。在每棵树节点分裂时再从 M 维的特征向量中随机选择 m($m \leqslant M$)个参与,最终通过所有树的统计投票,决定最可能的分类结果[18-19]。

1.4.1　随机森林的关键参数　在构建随机森林时,树的个数 k 和节点分裂特征个数 m 是影响模型精度与运行效率的最重要的两个参数[13,18-19]。一般来讲,随着决策树个数 k 的增加,模型泛化误差有效降低,但计算效率下降;节点分裂特征个数 m 决定单棵决策树分类能力,并影响树之间的相关性。本文使用 Python Scikit-learn 库实现随机森林的构建,以另外的约 1/3 未被抽中的袋外数据(out-of-bag)

计算 OOB 误差(oob_error)和验证数据计算的误差(test_error)作为评价依据,综合考虑模型效率和精度,选择最优参数 k 和 m 获取分类结果[18-19,21]。

1.4.2 随机森林的特征重要性 在随机森林中,特征的有效增加能提高分类精度,但高维度的特征互相之间可能具有相似性,继而对模型分类能力贡献少,并影响计算效率。因此,筛选各特征变量对模型的影响非常重要。

基尼系数(GINI)通常可作为衡量输入特征对模型贡献大小的评价标准。其主要思想:假设随机森林中有 n 棵树,对于第 i 棵树,样本中某一特征变量 X_j 在节点 m 分裂前后的 GINI 系数的变化值,即可作为在节点 m 上特征变量的重要性。再对第 i 棵树上特征变量 X_j 参与的所有节点集合 M 累加计算出在第 i 棵树上该特征的重要性。最后计算 n 棵树平均 GINI 系数变化值,即特征 X_j 在随机森林所有树中节点分裂不纯度的平均改变量($VIM_j^{(Gini)}$)作为重要性评分。对样本中所有特征变量来说,基于 GINI 系数的归一化重要性评分能直观量化各个特征对模型的贡献大小,值越高,特征重要性越高[18-19,24]。

本文以归一化重要性评分作为指标,客观评价各个分类特征的重要程度,并在试验中逐步减少输入特征维度,在保证模型分类性能和效率的基础上探索最好的特征子集,达到降维目的。

2 结果与分析

2.1 随机森林关键参数的确定

在提取了研究区 217 个特征后,构建不同 k 和 m 参数下的随机森林模型,利用 oob_error 和 test_error 作为评价判断标准,测试关键参数 k(值的范围 1~1 000)和 m(值的范围 1~30)对模型的影响。如图 2 所示,随着 k 值的增加,模型精度均有所提高,特别是少于 100 棵树,还未形成“森林”时精度提升明显。但在数目超过 100 棵后,oob_error 和 test_error 两种误差均缓慢收敛并趋于稳定。以 $m=16$ 的模型为例,当 k 从 1 增加到 100,oob_error 从 68.85%下降到 6.67%,test_error 从 15.72%下降到 6.35%;k 从 100 增大到 1 000,oob_error 仅从 6.67%下降到 6.18%,test_error 仅从 6.35%下降到 6.21%。因此,本文认为树的数量能有效提高随机森林分类精度,但在超过 100 以后,模型对树的增加变得不那么敏感,分类模型趋于稳定。

节点随机分裂的特征数 m 的增加也可以有效降低模型误差,但是相比参数 k 影响较小。m 在小于 5 时,模型精度提升相对明显;在 m 大于 15 以后模型精度提升幅度很小,特别是当树的数量大于 100 后,m 超过 10 时模型就已基本稳定。

为平衡模型的稳定、精度与效率,需要选取适当的参数 k 和 m。本文选取了 $k=200$、$m=10$ 参数下的模型用于进一步的最优特征子集筛选研究。

2.2 最优特征子集的选取

在获取 217 个特征变量重要性评分的基础上,选择随机森林分类模型的最优特征子集,该子集可使模型分类精度的降低最小,以实现降维的目的。图 3 展示了随着信息量最少特征的逐渐去除,分类的 Kappa 系数与总体精度随之变化的关系。当所有特征参与建模时,分类的(删减 0%)Kappa 系数为 0.926、总体精度为 0.935。在删减特征变量不超过 30%时,随机森林模型的分类能力基本维持不变,Kappa

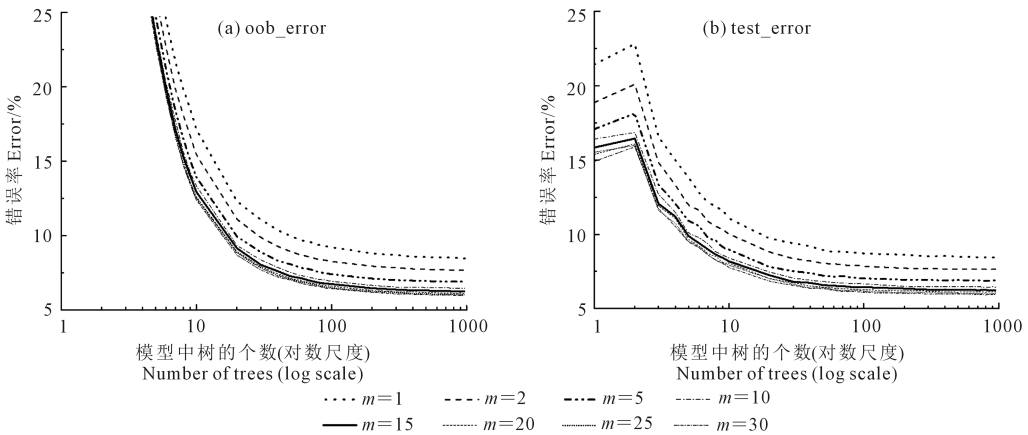


图 2 树的个数 k、节点随机分裂特征数 m 与 oob_error(a)、test_error(b)的关系

Fig.2 Relationship between the accuracy of out of bag dataset (a) or test dataset (b) and the number of trees (k) and number of random split variables (m)

系数在 0.925 左右,总体精度在 0.935 左右。当删减在 30%至 70%区间内,随着特征数的减少,模型的分类精度缓慢下滑。在删减超过 70%以后,分类精度下降的幅度迅速加快,特别是在删减超过 90%后模型的 Kappa 系数和总体精度均近直线快速下降。当仅保留归一化重要性评分最高的 10 个特征时,模型的分类能力仍然令人满意,Kappa 系数和总体精度分别为 0.842 和 0.861。经对比分析,本文选取了 161 个特征(约删减 26%)作为随机森林模型的最优特征子集,该子集分类的总体精度为 0.935,Kappa 系数为 0.926。

2.3　分类结果的精度评价

2.3.1　分类结果的混淆矩阵分析　经 2.2 节的分析与对比,本文选取由 161 个特征构成的最优特征子集利用随机森林实现分类,分类结果如图 4(e)。分类结果的精度评价采用混淆矩阵的方式,利用未参与建模的 30%,200326 个像元建立精度评价矩阵表(见表 2)。对研究区内的 9 种作物,分类精度最高

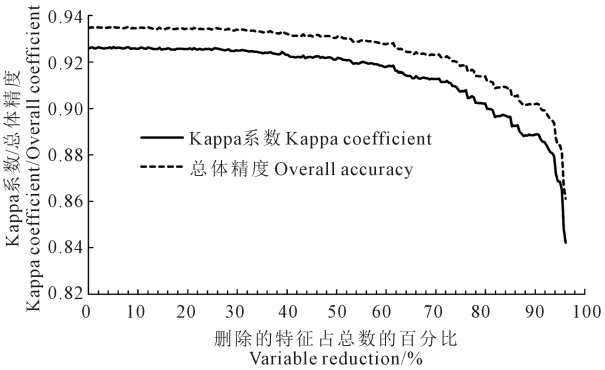


图 3　随机分类模型精度随特征删减比的变化关系

Fig.3　The effect of variable reduction on classification accuracies

表 2　随机森林分类模型混淆矩阵

Table 2　Confusion matrix of random forest classification model

类别 Type	小麦 Wheat	玉米 Corn	水稻 Rice	香梨 Pear	核桃 Walnut	苹果 Apple	葡萄 Grape	枣树 Jujube	棉花 Cotton	林地 Forest	草地 Grass	建筑 Building	戈壁 Gobi	沙地 Sand	水域 Water	总计 Total	用户精度 UA
小麦 Wheat	2703	0	5	0	1104	4	0	90	158	1	3	19	0	4	0	4091	0.661
玉米 Corn	0	117	0	0	0	0	0	1	97	0	1	1	0	0	0	217	0.539
水稻 Rice	1	0	17981	0	18	0	0	0	1243	1	0	13	0	0	29	19286	0.932
香梨 Pear	0	0	0	1753	339	457	0	64	7	0	2	34	0	0	0	2656	0.660
核桃 Walnut	39	0	3	19	14477	593	0	441	386	0	301	78	0	0	1	16338	0.886
苹果 Apple	9	0	0	60	706	5702	0	56	31	0	88	16	0	0	0	6668	0.855
葡萄 Grape	0	0	0	18	100	10	34	23	1	0	10	7	0	0	0	203	0.168
枣树 Jujube	26	0	1	11	1112	180	0	9648	302	1	426	71	0	2	0	11780	0.819
棉花 Cotton	65	0	2432	1	365	12	0	169	29432	2	31	78	1	4	32	32624	0.902
林地 Forest	0	0	0	4	75	17	0	9	79	1715	6	33	0	0	12	1950	0.880
草地 Grass	0	0	0	0	80	2	0	140	4	0	33470	154	10	1	0	33861	0.989
建筑 Building	6	0	3	8	42	18	0	6	59	2	144	17776	5	2	7	18078	0.983
戈壁 Gobi	0	0	0	0	0	0	0	0	0	0	0	4	34329	4	0	34337	1
沙地 Sand	3	0	0	0	0	0	0	1	6	0	1	22	4	5570	0	5607	0.993
水域 Water	0	0	1	0	0	0	0	0	3	5	0	28	0	0	12593	12630	0.997
合计 Total	2852	117	20426	1874	18418	6995	34	10648	31808	1727	34483	18334	34349	5587	12674	200326	
生产精度 PA	0.948	1	0.880	0.935	0.786	0.815	1	0.906	0.925	0.993	0.971	0.970	0.999	0.997	0.994		0.935

注 Note:UA: User accuracy; PA: Producer accuracy.

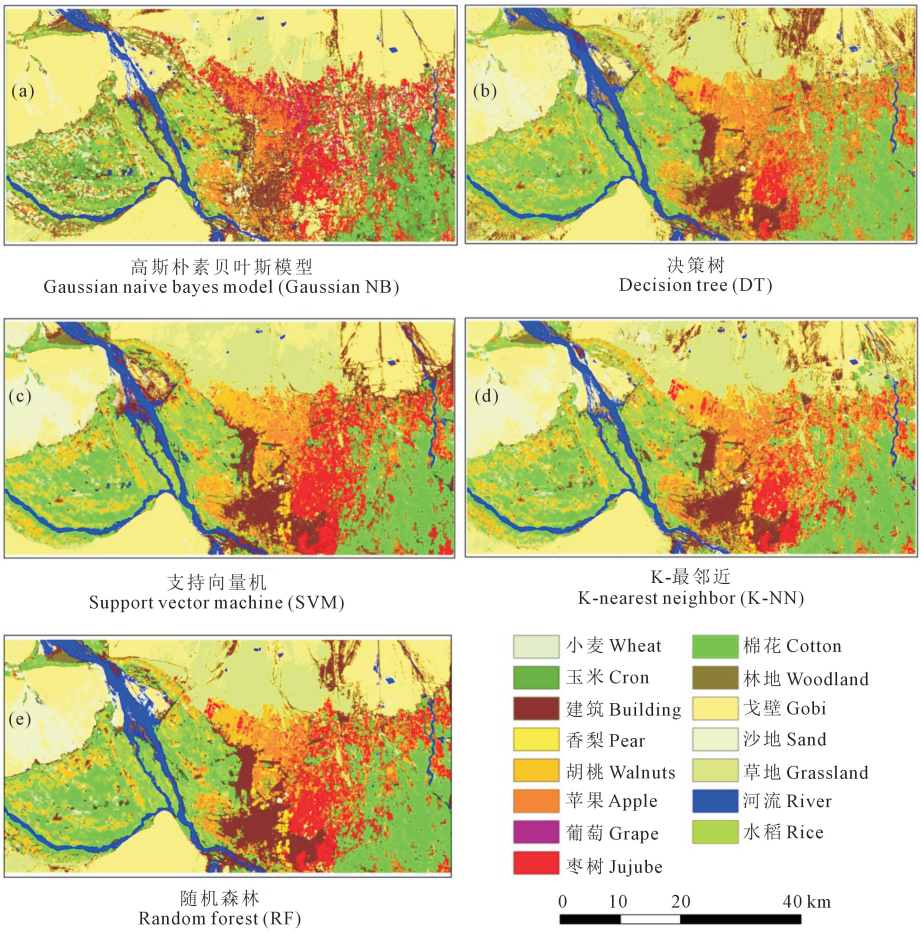


图 4 5 种监督分类算法的分类结果图

Fig.4 Classification results produced by Gaussian NB (a) DT (b) SVM (c) K-NN (d) and RF (e)

表 3 不同监督分类方法的精度对比

Table 3 Accuracy assessment of the different supervised classification models

分类方法 Classifiers	总体精度/% Overall accuracy	Kappa 系数 Kappa coefficient
朴素贝叶斯 Gaussian NB	0.655	0.616
决策树 DT	0.851	0.832
支持向量机 SVM	0.864	0.845
K-最邻近 K-NN	0.913	0.901
随机森林 RF	0.935	0.926

的是棉花,用户精度和生产精度分别为 0.902 和 0.925,其次为水稻(用户精度和生产精度分别为 0.932和 0.88,下同)、枣树(0.819 和 0.906)、苹果(0.855和 0.815)与核桃(0.886 和 0.786)。香梨、小麦、玉米和葡萄的用户精度都低于 0.7。由表 2 可知,核桃和苹果类别易被错分为香梨,这是由于三者同属果树类别,具有高度重叠的物候期。进一步分析发现(图 5):参与分类的特征集中在 3 月到 8 月间,此生长期内的小麦与林果物候期重叠,是造成小麦用户精度低的主要原因。研究区内玉米、葡萄种植分散、地块较小,将高分辨率的影像上目视解译的结果叠置在 30 m 分辨率影像上时,样本数量(分别为 117 个和 34 个)明显小于其他作物类别,且远低于分类的特征数,由此计算而来的生产精度和用户精度不具有代表性[18],本文认为对其精度分析没有意义。但为了研究的客观性,其精度值仍被列在混淆矩阵表中,由此说明 30 m 空间分辨率数据对研究区小地块少量样本类别分类的局限性,亦说明随机森林方法对样本数量具有一定的依赖性,但该研究在本文中并未深入说明。其他非作物类别的用户精度和生产精度均较高,除林地外,均超过了 0.9。

2.3.2 随机森林方法与其他监督分类方法的比较分析 除了随机森林方法外,本文利用 Python Scikit-learn 模块实现了朴素贝叶斯高斯模型(Gaussian NB)、支持向量机(Support Vector Machine, SVM)、K-最邻近算法(K-Nearest Neighbor, K-NN)和决策树(Decision Tree, DT)4 种常用的分类算法在研究区的地物分类,分类结果见

图 4(a~d),精度评价见表 3。对 5 种常见监督分类方法的对比分析可知,随机森林分类模型的效果明显优于其他算法,其次是 K-NN 分类模型,其总体精度和 Kappa 系数也均超过了 0.9,朴素贝叶斯高斯模型分类能力较差,Kappa 系数仅为 0.616。

对比分析 5 种分类算法的分类结果,朴素贝叶斯高斯模型分类结果中存在明显错分漏分现象,其中棉花被错分为玉米,草地错分为戈壁。在决策树和支持向量机分类结果中,棉花与枣树混淆严重;在 K-NN 分类结果中存在棉花与苹果、枣树的混淆。但相比于其余 4 种监督分类模型,随机森林在特征选取后保持了较好的分类能力,在提取作物信息过程中,精度与效率均表现最好。因此,本文认为随机森林分类算法在作物的遥感提取中有很好的可用性。

3　讨　论

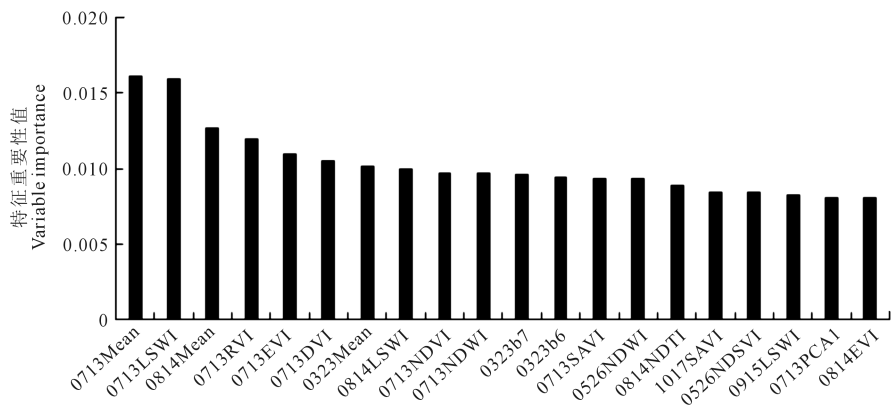
3.1　分类特征的重要性分析

为了评估模型(k=200、m=10)217 个特征的重要性,利用归一化重要性评分量化各个特征对模型的贡献大小。图 5 排列出了模型中重要性排名前 20 的特征,可知 7 月 13 日、8 月 14 日的特征重要性均较高,其次是 3 月 23 日。特别是 7 月 13 日影像的特征明显高于其他时相,高排名的特征数量也最多,这是因为在盛夏(7—8 月)时,作物生长处于旺盛阶段,植被信息较为明显,有利于作物信息的识别。3 月 23 日影像上的主要绿色作物为小麦和果树,其他农作物尚未播种,易于区别。

重要性排名靠前的特征主要是纹理特征的均值 0713Mean 和 0814Mean、植被指数的 0713LSWI、0713RVI、0713EVI、0713DVI、0814LSWI 和 0713NDVI 等。其中 Mean 表示的是纹理规则的平均值,是作物在卫星遥感影像上的形态特性反映,与作物种类及其生长状态相关。LSWI 指数表示了作物体内的水分含量,可知不同作物体内水分差异明显,成为类别识别的重要依据。其他排名靠前的植被指数(RVI、EVI、DVI 和 NDVI 等)均与红波段和近红外波段密切相关,是农业信息识别中常用的波段或植被指数。由植被指数可知尽管不同农作物的生育期重叠,但结合其生长所需水分和生长状态的不同可以实现有效的识别。排名靠前的波段反射率为 0323b7 和 0323b6,这两个波段均为短波红外波段,对水分信息敏感,可用于对绿色作物或水分含量差异的类别识别。

3.2　分类时相的重要性分析

遥感时相的选择是光学遥感农业应用的关键环节[2,6-8,10]。即使是处于干旱半干旱气候类型的研究区也未必每个生长季均能获取完整的时间序列影像数据,对其他气候区由于云、雾、雨、沙等因素的存在,仅能获取作物生长季的关键时期数据,因此对不同时相的重要性分析将对于指导影像的选择具有实用价值,图 6 列出了参与分类的不同时相特征的重要性值。由图 6 可知 7 月 13 日、8 月 14 日和 3 月 23 日的特征重要性均较高,与上文 3.1 节分析相同。其中 7—8 月为研究区作物生长旺盛季节,3 月为作物播种(栽培)或抽枝发芽季节,两个时期信息差异显著。由此可知,对研究区所代表的干旱区气候类型的作物识别,可选择两个关键时相:生长旺盛期与播种期,这一分析结果与赵良斌[25]与曹卫彬[26]等的研究结果一致。



注:研究中将所有特征按照影像时相加名称标记,如 3 月 23 日第 4 波段光谱反射率标记为 0323b4;7 月 13 日归一化水体指数(LSWI)标记为 0713LSWI,其他类同。

Note: Features are named according to the acquisition time of images and feature name. For example, the spectral reflectance of the 4th band from the image on March 23 is named 0323b4; The land surface water index (LSWI) from the image on July 13 is named 0713LSWI. Other features are named similarly.

图 5　特征重要性统计

Fig.5　Variable importance of the three feature sets

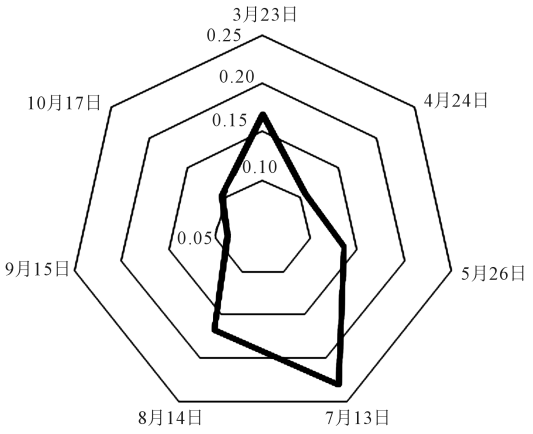


图 6 不同时相的重要性值

Fig.6 Spider charts representing the temporal importance

3.3 随机森林分类精度分析

随机森林方法作为机器学习领域的研究热点被广泛应用于地物信息分类识别中。黄双燕等[9]基于机器学习方法,采用时间序列 Sentinel 2A 遥感数据提取典型干旱区的农作物分类信息,探讨了不同分类特征组合对随机森林分类精度的影响,结果表明:随机森林分类器以有效集成光谱和植被指数等多维向量的优势,将其应用于干旱区典型农作物分类上的精度均在 89% 以上,总体精度最高可达 94.02%。同样是典型的干旱区作物分类,本文运用随机森林算法虽然得到了令人满意的分类结果(总体精度 93.5%,Kappa 系数 0.926),但也发现对不同的作物,RF 分类结果的生产与用户准确度之间存在差异。本文研究结果与黄双燕等[9]研究结果有所出入,主要原因在于,黄双燕等的研究对象仅有棉花、春小麦和冬小麦三类作物,所选研究区内作物类型单一,地块规整,与本文地块破碎、作物多样的研究区差异巨大。岳俊等[12]运用多种监督分类方法,结合光谱与纹理特征对南疆盆地 4 种主栽果树(核桃、枣树、香梨和苹果)进行遥感识别,结果表明枣树的分类精度远高于其他 3 种果树,而香梨、苹果、核桃光谱和纹理特征差异较小,分类精度较低。虽然岳俊等[12]在分类方法选择上没有利用随机森林方法,但对于不同林果类别分类精度的结论与本文一致(枣树分类精度最高,苹果、核桃和香梨三者易混淆)。苏腾飞等[27]基于多种植被指数时间序列和机器学习算法研究了内蒙古五原县的作物遥感分类,结果表明对于随机森林而言,EVI、、NDVI 和 NDSVI 等组合具有最佳分类精度,与本文特征重要性排名靠前的植被指数一致。

综合以上分析,可知随机森林法对农作物的精细分类具有高精度、泛化能力强、高维特征处理等优势,但对于不同类别样本量的不平衡性(如本文葡萄和玉米的过小样本)的适应性差,如何选择最佳分类样本数仍需深入研究,以确定随机森林方法的适用性。

4 结 论

本研究探究了随机森林算法在干旱地区作物遥感分类的适用性,利用多时相的时间序列 Landsat 8 OLI 遥感数据提取多种分类特征(波段反射率、植被指数和纹理特征),探寻了高维特征支持下的随机森林作物精细分类,并分析参与分类的特征重要性,以期为分类最佳时相数据的选择、最佳分类特征集的选取等关键问题提供参考。主要结论如下:(1)随机森林算法通过 GINI 系数可以实现分类特征的重要性评价。在作物分类中,表示影像纹理平均规则程度的特征 Mean、对作物含水量十分敏感的地表水分指数 LSWI 及短波红外光谱反射率均有较高贡献度。(2)最佳分类时相的选择可以依据分类特征重要性确定。对研究区所代表的干旱区气候类型的作物识别而言,可选择两个关键时相:生长旺盛期与播种期。(3)随机森林分类精度受分类特征数量的影响。按照重要性评分值从低到高的顺序删除部分特征,当删除数量低于总特征数的 30% 时,RF 模型的分类精度基本保持不变;当删除量超过 70%时,分类精度下降的幅度加大。(4)随机森林方法相对于决策树、支持向量机、朴素贝叶斯、K-近邻等监督分类算法,无论是分类结果的精度上,还是分类效率上均具有优势。

本研究的不足之处有:(1)对于参与分类的原始特征选取缺乏目的性,导致选择很多特征,如多个植被指数间存在信息冗余,可能会限制随机森林方法的分类敏感性。(2)研究区内作物类别的样本选择没有考虑到影像的分辨能力,导致对葡萄和玉米的分类结果不可靠,因此需要进一步研究随机森林算法对样本数量的敏感性。(3)对最佳时相的选择仅研究了单一时相的分类重要性,缺乏不同时相的组合研究。

参 考 文 献:

[1] 吴炳方.中国农情遥感速报系统[J].遥感学报,2004,8(6):481-497.
[2] 胡琼,吴文斌,宋茜,等.农作物种植结构遥感提取研究进展[J].中国农业科学,2015,48(10):1900-1914.
[3] 史舟,梁宗正,杨媛媛,等. 农业遥感研究现状与展望[J].农业机械学报,2015,46(2):247-260.
[4] 陈仲新,任建强,唐华俊,等. 农业遥感研究应用进展与展望[J]. 遥感学报,2016,20(5):748-767.

(下转第 298 页)