

doi: 10. 6046/zrzyyg. 2022280
引用格式: 杜晓川, 姜德波, 徐林刚, 等. 基于 GF-2 影像和随机森林算法的花岗伟晶岩提取[J]. 自然资源遥感, 2023, 35(4): 53-60. (Du X C, Lou D B, Xu L G, et al. Extracting granite pegmatite information based on GF-2 images and the random forest algorithm[J]. Remote Sensing for Natural Resources, 2023, 35(4): 53-60.)

基于 GF-2 影像和随机森林算法的花岗伟晶岩提取

杜晓川^{1,2}, 姜德波², 徐林刚¹, 范莹琳^{1,2}, 张琳¹, 李婉悦^{1,2}
(1. 中国地质大学(北京)地球科学与资源学院, 北京 100083; 2. 中国地质科学院矿产资源研究所, 自然资源部成矿作用与资源评价重点实验室, 北京 100037)

摘要: 花岗伟晶岩是花岗伟晶岩型锂矿的重要载体及找矿标志, 青海省天峻县扎卡东南部一带有较好的锂矿找矿潜力, 但该地区具有海拔高、切割深等特点, 使得实地地表调查难度较大, 因此采用随机森林算法对研究区花岗伟晶岩进行遥感提取, 以 GF-2 高空间分辨率遥感影像为数据源提取研究区域各类型地物的光谱特征、纹理特征、指数特征、地形特征、边缘特征及文中新引入的限制对比度自适应直方图均衡化(contrast limited adaptive hitogram equalization, CLAHE)特征等 25 个特征变量构建特征子集, 对子集中的特征变量进行特征重要性评估, 依据特征重要性得分进行特征优选, 确定提取花岗伟晶岩的最优特征组合, 最终选取 16 个特征变量进行随机森林分类, 对分类结果进行精度评价。研究表明: ①CLAHE 特征变量有利于突出地物间的色调差异, 有助于分类精度的提升, 其总体精度上升 2.7 个百分点, Kappa 系数上升 0.035; ②基于 GF-2 影像和随机森林算法的分类结果的总体精度达到了 93.1%, Kappa 系数达到 0.902, 花岗伟晶岩用户精度达到 94.24%, 生产者精度达到 98.00%, 证实方法的有效性, 同时也为该地区进一步工作提供真实可靠的资料。

关键词: 花岗伟晶岩; 随机森林算法; CLAHE; 特征重要性; 青海省

中图法分类号: TP 79 **文献标志码:** A **文章编号:** 2097-034X(2023)04-0053-08

0 引言

随着近年来新能源、新兴材料的发展, 锂等关键金属越来越受到重视^[1], 金属锂由于其特殊的物理化学性质, 被广泛应用到诸如电池、玻璃、陶瓷等各个工业领域, 具有极高的战略价值, 其不仅是高科技产业必需材料, 也逐渐成为日常生活中广泛使用的常规材料。当前我国国内锂业发展高度依赖矿物型锂矿特别是花岗伟晶岩型锂矿的供应^[2], 此类型锂矿床通常位于大型花岗岩侵入体边缘, 受地质构造控制明显, 通常以伟晶岩脉集群的形式产出, 总体呈带状分布, 单条伟晶岩脉形态多为脉状或板状^[3], 因此基于高分辨率影像的遥感技术是寻找和发现伟晶岩脉的有效方法^[4]。

遥感领域相关技术不断发展成熟, 使得遥感技术在地质信息提取的研究中得到了广泛应用, 有许多利用遥感技术进行含锂伟晶岩探测的实例。如代

晶晶等^[5-6]在川西甲基卡地区进行了锂辉石等伟晶岩稀有金属矿床相关矿石矿物的波谱测试研究和伟晶岩遥感地质填图; Cardoso - Fernandes 等^[7]利用多源遥感影像数据, 采用 RGB 彩色合成、波段比值分析、主成分分析等方法识别含锂伟晶岩的相关异常, 完成蚀变填图; 姚佛军等^[4]通过增强 ASTER 遥感图像中的差异来增强地质体的岩性差异, 并突出含锂伟晶岩的遥感特征; 王海宇^[8]以 WorldView-3 为数据源, 并基于全卷积网络(fully convolutional networks, FCN)网络和 U-net 网络 2 种深度学习语义分割模型识别伟晶岩脉信息。前人通过遥感影像数据在含锂伟晶岩提取方面开展了大量的理论研究和实践, 但大多数研究者侧重于使用低空间分辨率影像数据, 提取方法也较为传统, 尽管已有基于高空间分辨率影像数据与深度学习网络模型提取伟晶岩岩脉的先例, 但该方法耗时较长, 效率较低, 且对样本数量有一定要求。因此本文在众多机器学习算法中选用随机森林算法, 其算法具有如下优点: ①数据

收稿日期: 2022-07-05; 修订日期: 2022-12-25
基金项目: 国家重点研发计划项目“西部伟晶岩型粘土型锂等稀有金属成矿规律与潜力评价”(编号: 2021YFC2901905)和基本科研业务费项目“中国铅锌银矿床成矿规律研究”(编号: KK2202)共同资助。
第一作者: 杜晓川(1995-), 男, 硕士研究生, 资源与环境专业。Email: 1169954017@qq.com。
通信作者: 姜德波(1979-), 男, 教授级高级工程师, 主要从事矿产资源评价研究。Email: llddbb_e@126.com。

兼容性强,数据集无须规范化,且可以直接处理高维数据;②抗噪声能力强,受异常值影响较小;③鲁棒性强,不容易发生过拟合现象;④计算效率高、速度快、泛化能力强^[9-11],且该方法结合高空间分辨率影像数据在提取复杂地物的效果较好。

据此,以青海省海西蒙古族藏族自治州天峻县扎卡东南部区域为研究区,提取预处理后的 GF-2 高空间分辨率影像数据中各类型地物的光谱特征、纹理特征、指数特征、地形特征及边缘特征等 25 个特征变量构建特征子集,对子集中的特征变量进行特征重要性分析,依据特征重要性得分进行特征优选,确定最优的提取花岗伟晶岩的特征组合,并进行随机森林分类,对分类结果进行精度评定,为后续该

地区的含锂花岗伟晶岩找矿勘查工作提供依据。

1 研究区概况与数据源

1.1 研究区概况

研究区位于青海省海西蒙古族藏族自治州天峻县扎卡东南部,地理坐标为 E99°27'16"~99°30'32", N36°53'36"~36°54'43"(图 1)。天峻县地处青海省东北部,青海湖西侧、祁连山南麓,柴达木盆地东部。境内高山纵横,山脉呈东南西北走向,地形以山地为主,高山、中低山、山谷和山间盆地相间分布。平均海拔在 4 000 m 以上,地处高原寒带气候,年平均气温 -1.5℃,年降水量 360 mm,干旱少雨,多风沙。

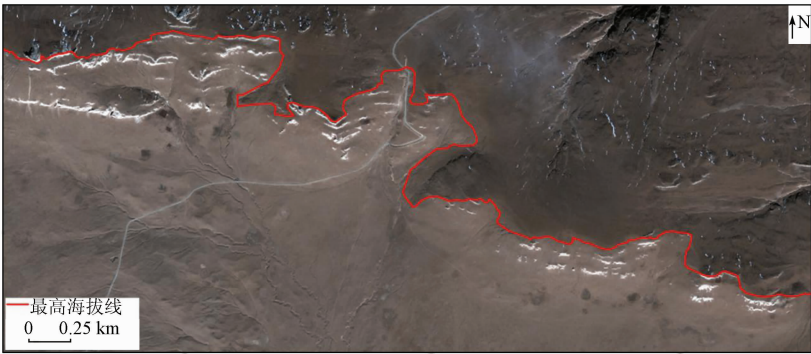


图 1 研究区遥感影像
Fig.1 Remote sensing image in the study area

1.2 数据源及预处理

GF-2 卫星于 2014 年 8 月 19 日发射,是我国自主研制的首颗空间分辨率优于 1 m 的民用光学遥感卫星。搭载有 2 台高分辨率的 1 m 全色与 4 m 多光谱相机,传感器的波段参数如表 1 所示。对大气校正后的多光谱影像以及定标后的全色影像做正射校正处理及融合处理,获得空间分辨率为 1 m 的彩色合成影像,再通过研究区域的矢量范围对影像数据进行裁剪,得到研究区遥感影像数据。

上受构造控制作用明显,研究区内断裂沿 NW-SE 向分布,因此花岗伟晶岩大多沿 NW-SE 向展布,呈条状、脉状产出;在影像色调上,其多呈灰白色或略带土黄色;由于研究区内海拔较高、风沙较多、风化能力较强,且花岗伟晶岩晶体较大,具有较强的抗风化性,因而在高度上要突出于周围地物,并且其表面纹理粗糙,边界较明显。研究区内其他地物类型主要分为道路(图 2(b))、干涸河流(图 2(c))、冰雪(图 2(d))、阳坡裸地(图 2(e))、阴坡裸地(图 2(f))5 类,其中阳坡裸地与阴坡裸地以研究区内最高海拔线为界线(图 1),海拔沿界线南北两侧逐渐降低,加上卫星拍摄角度影响,使得二者地物的色调差别较大。阳坡裸地坡向朝南,色调以土黄色、棕色为主;阴坡裸地坡向朝北,色调以棕黑色为主;干涸河流的形状与边界较为明显,宽度较窄;冰雪主要分布在雪线较低的阴坡裸地上,色调以亮白色为主,形状为长条状,展布大多为南北向;道路色调为墨青色、边界明显。基于不同地物的特点,本文选择提取地物的光谱特征、纹理特征、指数特征、地形特征、边缘特征作为随机森林分类所需的特征变量。光谱特征包括 GF-2 高空间分辨率影像数据 4 个

表 1 GF-2 全色多光谱相机波段参数
Tab.1 Band parameters of GF-2 panchromatic multi-spectral camera

波段	波段范围/ μm	空间分辨率/m	幅宽/km	重访周期/d
Pan	0.45~0.90	1		
Blue	0.45~0.52			
Green	0.52~0.59	4	45	5
Red	0.63~0.69			
NIR	0.77~0.89			

2 研究方法

2.1 特征提取及样本选择

花岗伟晶岩(图 2(a))在分布位置及产出形态



(a) 花岗伟晶岩样本 (b) 道路样本 (c) 干涸河流样本 (d) 冰雪样本 (e) 阳坡裸地样本 (f) 阴坡裸地样本

图 2 影像块样本

Fig. 2 Image samples

波段的灰度特征(B,G,R,NIR)、影像灰度平均值特征(Img Mean)、影像灰度标准差特征(Img Std),以及对影像原始 4 个波段进行限制对比度自适应直方图均衡化的 CLAHE 特征(CLAHE B, CLAHE G, CLAHE R, CLAHE NIR)。

纹理特征是一种全局特征,对噪声具有较强抵御能力,能刻画出图像区域所对应地物表面的性质,具有旋转不变的特性^[12],且已有研究表明纹理特征有助于改善地表信息提取的精度^[13-16]。本文在提取纹理特征前先对影像数据进行主成分分析变换,该方法通过应用正交变换减少相关数据的冗余,减少各组分间的相关性,尽可能将有用信息集中到新的组分中,且各新组分间互不相关,达到信息增强、降维及减少计算量的目的^[17]。其中影像数据的第一主成分(PC1)包含所有波段中 98.5% 的特征信息,因此将 PC1 作为影像的光谱特征的同时,仅对 PC1 采用灰度共生矩阵方法提取纹理特征,选取均值(Mean)、方差(Variance)、均质性(Homogeneity)、对比度(Contrast)、相异性(Dissimilarity)、熵(Entropy)、相关性(Correlation)、角二阶矩(angular second moment,ASM)等 8 种彼此相关性弱的纹理信息特征统计量作为纹理特征。

指数特征是通过一定公式的波段运算所获取的灰度影像。指数特征可以消除地形差异影响,突出地物特征^[18]。基于阴坡裸地存在阴影的特点,本文提取用于突出阴影区域与明亮区域对比的阴影指数(shadow index,SI)特征,其计算公式为:

$$SI = (B + G + R + NIR)/4, \quad (1)$$

式中 B,G,R,NIR 分别为蓝光、绿光、红光及近红外波段的反射率。基于研究区内裸地范围较大,本文提取突出裸地区域亮度的裸地区域指数(bare area index,BAI)特征,其计算公式为:

$$BAI = (B - NIR)/(B + NIR)。 \quad (2)$$

地形特征是描述地势变化的特征,由于研究区内地势起伏较大,且地物与地形因素有一定关联,因此有必要提取坡度(Slope)、坡向(Aspect)、高程等

地形特征。本文从美国地质调查局获取研究区 30 m 空间分辨率的 ASTER GDEM 数字高程模型(digital elevation model,DEM)数据,并分别提取坡度、坡向、DEM 地形特征。

边缘特征是影像中特性(如像素灰度、纹理等)分布不连续处,影像周围特性有阶跃变化的像素集合,其中研究区内花岗伟晶岩、干涸河流、冰雪、道路的边缘特征较明显,边缘灰度梯度变化较大。Canny 边缘检测能较精确估算出图像边缘的强度、梯度方向,具有定位准确、单边响应和信噪比高等优势^[19]。将从 GF-2 高空间分辨率影像中提取出的光谱特征、纹理特征、指数特征、地形特征与边缘特征等 25 个特征变量构成初始特征空间。

各类型地物选取 25 个影像样本构成训练样本集,图 2 为 32 像素×32 像素的影像样本块展示。为提高分类检测精确度,选取原则依据样本要尽可能包含各个类别地物的各种不同类型特征,且单个样本尽量仅包含单个类别地物。根据随机森林实验需求,将样本集按照 2:8 的比例随机划分为测试集和训练集。

2.2 随机森林算法及参数组合优化

随机森林算法最早是由 Breiman^[20]提出的一种机器学习算法,本文依据 Gini 系数最小原则构建分类和决策树(classification and regression tree, CART),并基于多棵决策树对样本进行训练,根据训练得到的模型,对未知待测样本类别进行预测的一种监督学习分类算法^[21]。

本文使用 Python 语言编程调用 sklearn 库进行随机森林模型训练,采用随机搜索交叉验证(RandomizedSearchCV)与网格搜索交叉验证(GridSearchCV)^[22]的双重方式进行参数优化。单一的 GridSearchCV 是对每一组排列组合的参数进行遍历,最后选取出最好的一组参数,但由于参数排列组合数量过多,逐个遍历面临相当耗时与维度灾难等问题,因此在进行 GridSearchCV 前先通过 RandomizedSearchCV(设置随机搭配超参数组合次数(n_iter)为 200 次,交叉验证折数(cv)为 5 次)对取值范围内的

参数进行有限次的随机排列组合,并对其进行遍历,从中选取出一组较优参数组合,再通过进行 10 折的 GridSearchCV 遍历选取最优参数组合。本文最优参数组合设置如表 2 所示。

表 2 最优参数组合		
Tab.2 Optimal parameter combination		
参数名	取值范围	最优值
n_estimators	1,2,3,...,1 000	150
max_features	Auto,Sqrt,Log2	Auto
max_depth	1,2,3,...,200	50
min_samples_split	2,3,...,100	4
min_samples_leaf	1,2,3,...,60	1

2.3 特征重要性评估与特征优选

特征集中包含着不同属性、不同尺度的多样化的众多特征变量,过多的特征变量会造成数据的冗

余、分类时间过长、以及分类精度下降等问题^[23],因此需要通过特征选择筛选出对模型结果影响较大的特征变量构成新的特征子集。特征选择主要是通过一定的准则对特征重要性排序,或者在分类精度保证的情况下选择一组数量最小,且最具有代表性的特征子集^[24]。本文先通过调用 sklearn 库中基于 Gini 系数原理进行特征重要性计算的 feature_importances 参数对每个特征进行重要性评估,对得到的特征重要性得分按降序排列(图 3),然后采用顺序前向特征选择法^[25],在每一次迭代中选择当前特征集中特征重要性得分最高的特征变量进行模型训练,并获取相应模型的测试集精度,依次增加特征变量数量直至全部纳入,从中选取测试集精度最高、特征变量数量相对最少的特征组合。

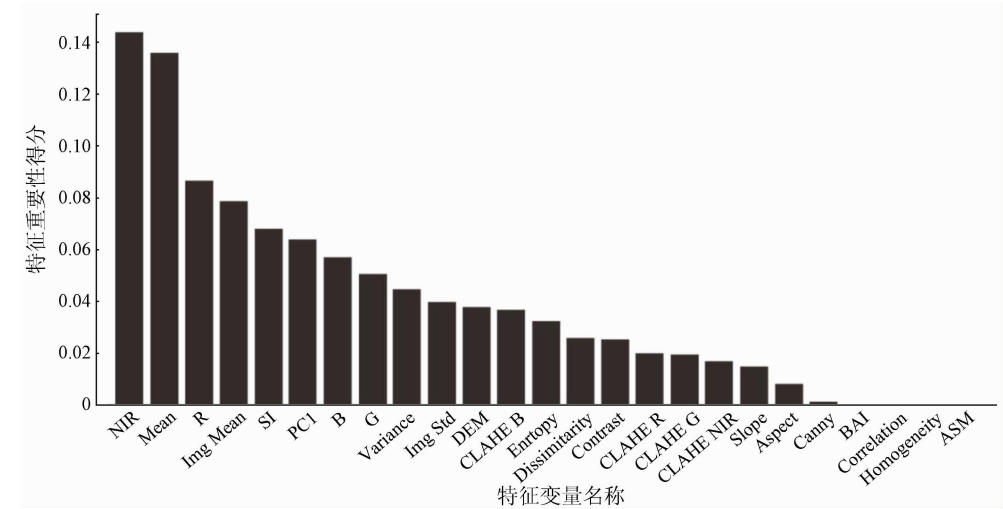


图 3 特征重要性排序
Fig.3 Feature importance ranking

2.4 形态学运算

基于最佳参数组合与最优特征变量组合的随机森林模型对研究区各地物类型进行分类,对分类结果进行先膨胀后腐蚀的形态学闭运算操作,闭运算在填充同类型地物间的细微裂隙,使地物间更连接且更平滑的同时,还可以消除孤立噪声点的影响。

3 实验结果与精度分析

3.1 特征优选评价

分别对特征空间中的 25 个特征变量通过顺序向前特征选择法逐一训练后的结果进行评估,生成特征变量个数与测试集精度关系图(图 4),由图 4 可知:①由首个特征变量逐个增加至 16 个特征变量参与模型训练期间,测试集精度整体呈现迅速上升至平稳上升趋势,由最初的 84.01% 上升至 96.24%;②特征变量数量在 17~25 个期间,由于特征变量数

量过多,产生数据冗余、训练耗时增加等问题,导致分类器出现一定的过拟合现象,造成测试集精度有些许下降,由 96.24% 下降至 95.94%,由此可知

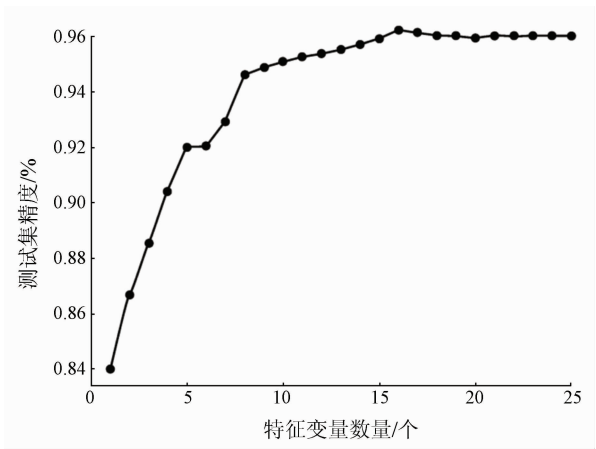


图 4 特征变量个数与测试集精度关系
Fig.4 Relationship between the number of features variables and test set accuracy

并不是特征变量数量越多测试集精度就越高。根据测试集精度结果,选取前 16 个特征变量组合作为最优特征集,并基于该特征集进行随机森林分类。

3.2 特征变量分析

遥感影像中不同波段对于分类不同地物具有重要意义^[24],波段间反射率值的差异有助于区分不同类型地物,通过计算不同特征变量在不同类型地物的反射率值,并对反射率值进行归一化处理,得到各类型地物波谱曲线(图 5)。光谱特征(图 5(a))中的 NIR 波段的地物波谱曲线反射率值差异较大,可以很好地区分各类型地物,并且 NIR 波段的特征重

要性得分最高;纹理特征(图 5(b))中 Mean 波段的各类型地物反射率值有所差异,具有区分地物的能力,花岗伟晶岩的纹理特征在大多数波段反射率值都较高,原因为风化剥蚀等作用使得花岗伟晶岩岩石表面产生较多纹理;指数特征(图 5(c))中阴坡裸地的阴影指数反射率值较低,与其他类型地物有明显区分;地形特征(图 5(d))中,由于研究区内高程起伏较大,高程范围由 3 671~4 029 m,高差相差 358 m,使得 DEM 波段中不同类型地物反射率值差异较大。

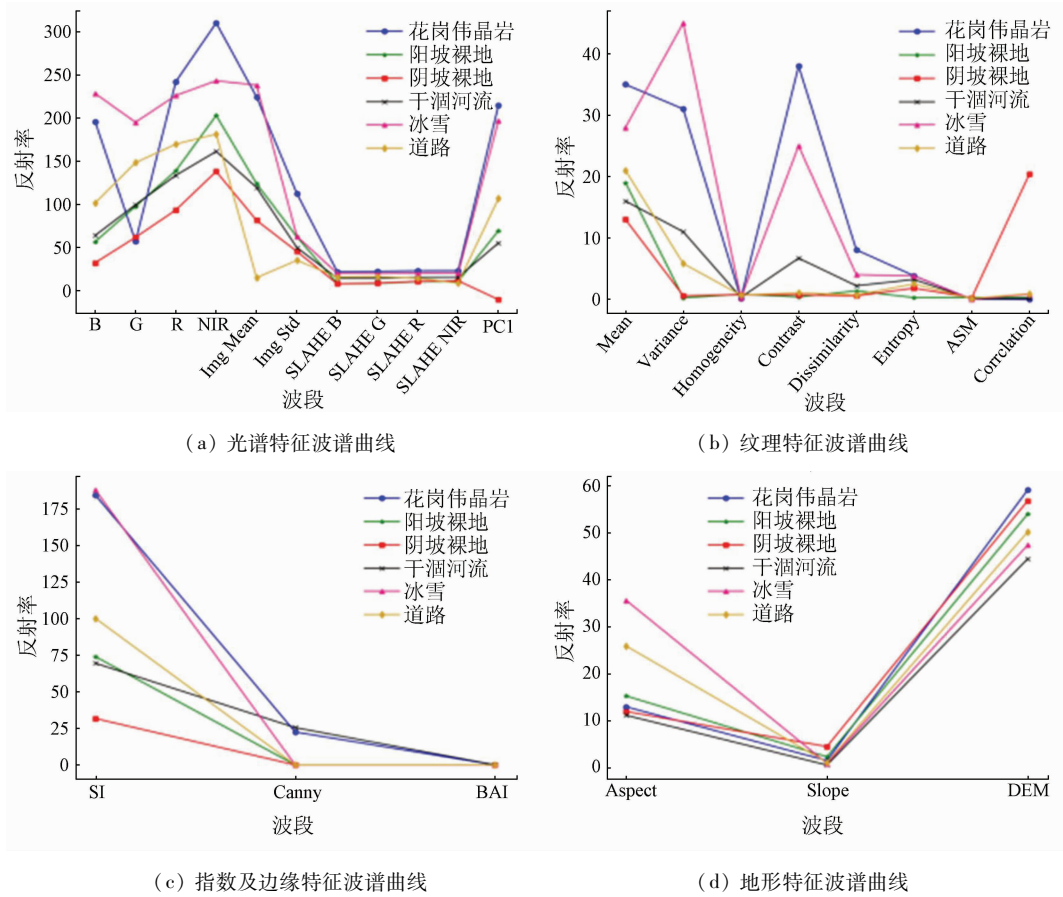


图 5 地物波谱曲线
Fig. 5 Curve of landmark spectrum

3.3 分类结果与精度分析

3.3.1 随机森林分类结果与精度分析

基于随机森林算法对研究区的 GF-2 高空间分辨率影像数据进行花岗伟晶岩提取,其各类型地物分类结果如图 6 所示。此次实验结果共提取出 84 处花岗伟晶岩,其形状大小不一,以脉状产出为主,成群出现,且轮廓较为明显,与其他类型地物区分较好。在研究区内随机均匀生成 1 000 个地面验证点,其中花岗伟晶岩验证点 104 个,道路验证点 19 个,冰雪验证点 51 个,阳坡裸地验证点 414 个,干涸河流验证点 87 个,阴坡裸地验证点 325 个,通

过计算地面验证点与真实地物类型之间的混淆矩阵获得分类结果的精度统计(表 3)评价,其中总体精度达到 93.1%; Kappa 系数达到 0.902; 花岗伟晶岩、冰雪、阳坡裸地及阴坡裸地共 4 种地物的用户精度和生产者精度都高于 93%、道路高于 81%、干涸河流高于 73%。花岗伟晶岩由于其自身的色调、纹理,以及分布地形海拔都较高等优势,在分类结果上效果较好,用户精度高达 94.24%、生产者精度高达 98%,以上结果说明基于多特征的随机森林算法在提取花岗伟晶岩信息中有较好的适用性。

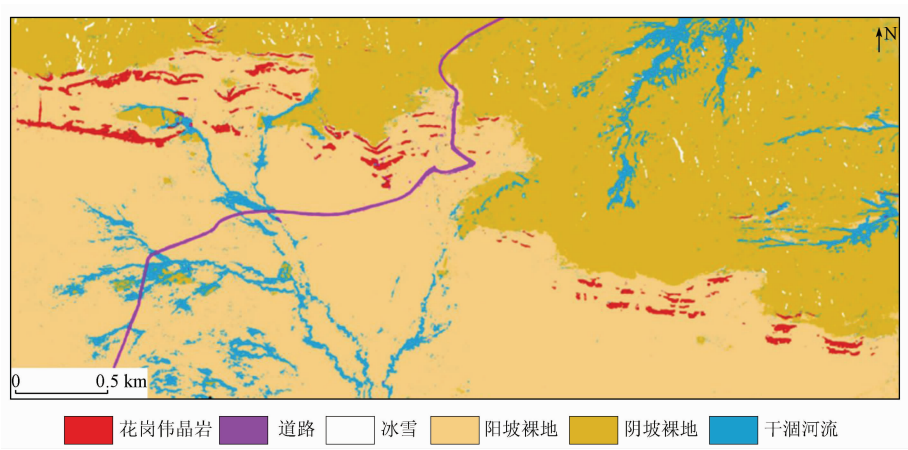


图 6 研究区地物分类结果

Fig. 6 Results of landmark classification in the study area

表 3 分类结果的精度统计

Tab. 3 Precision evaluation of the classification result

地物类别	用户 精度/%	生产者 精度/%	错分 误差/%	漏分 误差/%	总体 精度/%	Kappa 系数
花 岗 伟 晶 岩	94.24	98.00	5.76	2.00	93.10	0.902
道路	94.73	81.81	5.27	18.19		
冰雪	98.04	98.04	1.96	1.96		
阳坡裸地	94.44	95.13	5.56	4.87		
干涸河流	81.61	73.19	18.39	26.81		
阴坡裸地	93.23	94.98	6.77	5.02		

3.3.2 CLAHE 特征层讨论

本文新引入的 4 个 CLAHE 特征层中,CLAHE B 与 CLAHE R 作为特征变量参与了随机森林分类,为验证 CLAHE B 和 CLAHE R 对分类结果产生的影响,故对优选的 16 个特征变量中去除 CLAHE B 和 CLAHE R 这 2 个特征变量,只保留余下的 14 个特征变量参与随机森林分类,并通过计算验证相同的 1 000 个地面验证点与真实地物类型之间的混淆矩阵,并获得分类结果的精度(表 4)评价,其中总体精

表 4 14 个特征的分类结果精度统计

Tab. 4 Precision evaluation of the 14 feature classification result

地物类别	用户 精度/%	生产者 精度/%	错分 误差/%	漏分 误差/%	总体 精度/%	Kappa 系数
花 岗 伟 晶 岩	93.27	97.00	6.73	3.00	90.40	0.864
道路	94.74	85.71	5.26	14.29		
冰雪	98.04	98.04	1.96	1.96		
阳坡裸地	90.58	93.28	9.42	6.72		
干涸河流	78.16	62.39	21.84	37.61		
阴坡裸地	91.08	93.38	8.92	6.62		

度为 90.4%,较最佳分类结果的总体精度下降 2.7 百分点; Kappa 系数为 0.864,较最佳分类结果的 Kappa 系数下降 0.038;各地物类型的用户精度与

生产者精度都有不同程度的下降。

CLAHE 方法的优越性在于可以通过改变不同地物间的对比度,以突出色调上的差异,相比普通的直方图均衡化方法只增强全局影像对比度,CLAHE 方法在此基础上引入了自适应性与限制性,自适应性改进了增强全局对比度中只考虑整体影像区域,而忽略局部影像区域的情况,进而注重影像的局部像素差异,在突出局部影像对比度的同时获得更多细节特征;限制性是在自适应性基础上的进一步改进,目的是为了 避免均衡化过程中产生的色调不连续与对比度过度增强等问题。以精度下降最为明显的干涸河流为例,且为了效果更加直观,故采用真彩色的方式展示(图 7),CLAHE 真彩色样本影像中(图 7(b)),干涸河流的亮度要明显亮于阳坡裸地,且对比度更加突出,边缘轮廓更加清晰,2 类地物在色调上有明显差异,有利于分类精度的提升,证实了 CLAHE 方法在区分色调较为相近地物时的有效性,同时 CLAHE 特征增强了花岗伟晶岩与其他类型地物间亮度与对比度的差异,有助于提升花岗伟晶岩的提取效果,使其用户精度和生产者精度分别上升 0.97 百分点与 1 百分点。



(a) 真彩色影像 (b) CLAHE 真彩色影像

图 7 干涸河流影像对比

Fig. 7 Image comparison of dry rivers

4 结论

1) 本文基于 GF-2 高空间分辨率影像数据结合特征优选的随机森林算法提出花岗伟晶岩的提取方法。地物分类结果总体精度为 93.10%, Kappa 系数为 0.902; 花岗伟晶岩的用户精度及生产者精度分别为 94.24% 和 98.00%, 证实了该方法的有效性。

2) 特征变量个数持续增多会产生数据冗余及计算量过大等问题, 导致过拟合现象发生。此次研究在特征优化中确定特征变量个数为 16 个时分类精度最高。

3) 本文新引入的限制对比度自适应直方图均衡化(CLAHE)特征层中的 CLAHE B 和 CLAHE R 的特征重要性得分较高, 并参与随机森林分类。CLAHE 特征变量具有提升影像对比度、突出不同类型地物间色调差异等优势, 有助于区分色调相近的地物, 有利于地物分类精度的提升。

参考文献(References):

[1] 王登红. 关键矿产的研究意义、矿种厘定、资源属性、找矿进展、存在问题及主攻方向[J]. 地质学报, 2019, 93(6): 1189-1209.
Wang D H. Study on critical mineral resources: Significance of research, determination of types, attributes of resources, progress of prospecting, problems of utilization, and direction of exploitation [J]. Acta Geologica Sinica, 2019, 93(6): 1189-1209.

[2] 王学求, 刘汉粮, 王 玮, 等. 中国锂矿地球化学背景与空间分布: 远景区预测[J]. 地球学报, 2020, 41(6): 797-806.
Wang X Q, Liu H L, Wang W, et al. Geochemical abundance and spatial distribution of lithium in China: Implications for potential prospects[J]. Acta Geoscientica Sinica, 2020, 41(6): 797-806.

[3] 唐 军. 浅析典型伟晶岩型锂矿床成矿地质特征[J]. 世界有色金属, 2020(14): 143-144.
Tang J. A brief analysis of metallogenic characteristics of typical pegmatite type lithium deposits[J]. World Nonferrous Metals, 2020(14): 143-144.

[4] 姚佛军, 徐兴旺, 杨建民, 等. 戈壁浅覆盖区花岗岩中锂铍伟晶岩的 ASTER 遥感识别技术——以新疆镜儿泉地区为例[J]. 矿床地质, 2020, 39(4): 686-696.
Yao F J, Xu X W, Yang J M, et al. A technology for identifying Li-Be pegmatite using ASTER remote sensing data in granite of Gobi shallow-covered area: A case study of recognition and prediction of Li-Be pegmatite in Jingerquan, Xinjiang[J]. Mineral Deposits, 2020, 39(4): 686-696.

[5] 代晶晶, 王登红, 代鸿章, 等. 遥感技术在川西甲基卡大型锂矿基地找矿填图中的应用[J]. 中国地质, 2017, 44(2): 389-398.
Dai J J, Wang D H, Dai H Z, et al. Geological mapping and ore-prospecting study using remote sensing technology in Jiajika area of

Western Sichuan Province[J]. Geology in China, 2017, 44(2): 389-398.

[6] 代晶晶, 王登红, 代鸿章, 等. 川西甲基卡锂矿基地典型岩石及矿物反射波谱特征研究[J]. 岩矿测试, 2018, 37(5): 507-517.
Dai J J, Wang D H, Dai H Z, et al. Reflectance spectral characteristics of rocks and minerals in Jiajika lithium deposits in West Sichuan[J]. Rock and Mineral Analysis, 2018, 37(5): 507-517.

[7] Cardoso-Fernandes J, Teodoro A C, Lima A. Remote sensing data in lithium (Li) exploration: A new approach for the detection of Li-bearing pegmatites[J]. International Journal of Applied Earth Observation and Geoinformation, 2019, 76: 10-25.

[8] 王海宇. 基于深度语义分割的遥感影像伟晶岩脉信息提取研究[D]. 北京: 中国地质大学(北京), 2021.
Wang H Y. Research on extracting pegmatite dike information of remote sensing image based on depth semantic segmentation [D]. Beijing: China University of Geosciences(Beijing), 2021.

[9] Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: A survey and results of new tests[J]. Pattern Recognition, 2011, 44(2): 330-349.

[10] 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 S Y, Zhang Y W, Yu Z H. Classification of Honghe wetland remote sensing image based on random forests[J]. Geomatics & Spatial Information Technology, 2014, 37(4): 83-85, 93.

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

[12] 陈海洋, 孟令奎, 周 元. 基于随机森林的遥感影像雪冰云信息检测方法[J]. 测绘地理信息, 2022, 47(2): 105-110.
Chen H Y, Meng L K, Zhou Y. Ice snow and cloud detection in remote sensing images based on random forest[J]. Journal of Geomatics, 2022, 47(2): 105-110.

[13] Feng Q, Liu J, Gong J. UAV remote sensing for urban vegetation mapping using random forest and texture analysis [J]. Remote Sensing, 2015, 7(1): 1074-1094.

[14] Dell'Acqua F, Gamba P. Texture-based characterization of urban environments on satellite SAR images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2003, 41(1): 153-159.

[15] Stasolla M, Gamba P. Spatial indexes for the extraction of formal and informal human settlements from high-resolution SAR images [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2008, 1(2): 98-106.

[16] Feng Q, Gong J, Liu J et al. Flood mapping based on multiple endmember spectral mixture analysis and random forest classifier: The case of Yuyao, China [J]. Remote Sensing, 2018, 7(9): 12539-12562.

[17] Ali A, Pour A. Lithological mapping and hydrothermal alteration using Landsat 8 data: A case study in ariab mining district, red sea hills, Sudan[J]. International Journal of Sciences Basic and Applied Sciences, 2014, 3(3): 199-208.

[18] 郭 霞. 基于 Landsat-8 OLI 遥感数据的随机森林分类优化研究[D]. 秦皇岛: 燕山大学, 2021.
Guo X. Optimization research of random forest classification based on Landsat-8 OLI remote sensing images[D]. Qinghuangdao: Yanshan University, 2021.

[19] 刘 念. 基于坎尼边缘检测算法的遥感影像自动绘图研究[J]. 测绘地理信息, 2016, 41(6): 75 – 78.
Liu N. Automatic mapdrawing of remote sensing images based on Canny edge detection algorithm[J]. Journal of Geomatics, 2016, 41(6): 75 – 78.

[20] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5 – 32.

[21] 张佳华, 姚宜斌, 曹 娜. 基于决策树对有无降水进行预测[J]. 测绘地理信息, 2017, 42(5): 107 – 109.
Zhang J H, Yao Y B, Cao N. Prediction of whether precipitation based on decision tree[J]. Journal of Geomatics, 2017, 42(5): 107 – 109.

[22] Lerman P M. Fitting segmented regression models by grid search [J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1980, 29(1): 77 – 84.

[23] 陈红星. 基于多源数据和随机森林算法的建筑物尺度人口估算[D]. 上海: 华东师范大学, 2021.
Chen H X. Population estimation at the building level based on multi – source data and random forest algorithm[D]. Shanghai: East China Normal University, 2021.

[24] 黄东山. 特征选择及半监督分类方法研究[D]. 武汉: 华中科技大学, 2011.
Huang D S. Research on feature selection and semi – supervised classification[D]. Wuhan: Huazhong University of Science and Technology, 2011.

[25] 袁 帅, 杨宏晖, 申 昇. 基于互信息的顺序向前特征选择算法 [J]. 声学技术, 2014, 33(4): 359 – 362.
Yuan S, Yang H H, Shen S. Forward order feature selection algorithm based on mutual information[J]. Technical Acoustics, 2014, 33(4): 359 – 362.

Extracting granite pegmatite information based on GF – 2 images and the random forest algorithm

DU Xiaochuan^{1,2}, LOU Debo², XU Lingang¹, FAN Yinglin^{1,2}, ZHANG Lin¹, LI Wanyue^{1,2}

(1. School of Earth Sciences and Resources, China University of Geosciences (Beijing), Beijing 100083, China; 2. MNR Key Laboratory of Metallogeny and Mineral Assessment, Institute of Mineral Resources, Chinese Academy of Geological Sciences, Beijing 100037, China)

Abstract: Granite pegmatites serve as a significant carrier and prospecting marker of granite pegmatite – type lithium deposits. The southeastern Zhaka area in Tianjun County, Qinghai Province demonstrates considerable prospecting potential for lithium deposits. Nevertheless, its high altitudes and deep cross – cutting characteristics pose challenges in surface surveys. Hence, this study extracted the granite pegmatite information within the study area from remote sensing images using the random forest algorithm. With high – spatial – resolution GF – 2 remote sensing images as the primary data source, it extracted the spectral, texture, exponential, topographic, and edge features from various ground objects within the study area. These features, together with the newly introduced contrast limited adaptive histogram equalization (CLAHE) features, constituted 25 feature variables, forming a feature subset. Then, feature variables in the subset were evaluated for their feature importance, and their importance scores were used for feature selection, determining the optimal feature combination for extracting granite pegmatite information. Ultimately, 16 feature variables were chosen for random forest classification, with the accuracy of the classification results assessed. The study indicates that: ①The CLAHE feature variables emphasize the tonal variations among ground objects, thereby enhancing the classification accuracy, with the overall accuracy increased by 2.7 percentage points and the Kappa coefficient increased by 0.035; ②The classification results for granite pegmatites based on GF – 2 images and the random forest algorithm exhibited overall accuracy of 93.1%, with a Kappa coefficient of 0.902, user accuracy of 94.24%, and producer accuracy of 98.00%, confirming the effectiveness of the method used in this study. Moreover, this study provides reliable data for future research in the study area.

Keywords: granite pegmatite; random forest algorithm; CLAHE; feature importance; Qinghai Province
(责任编辑: 张 仙)