

基于 ReliefF-GA 和随机森林算法的遥感影像面向对象分类

肖艳^{1*} 刘项¹ 王斌²

1 长春工程学院, 吉林长春, 130012;

2 长春市测绘院地理信息分院, 吉林长春, 130021;

摘要: 为实现土地覆盖信息准确、快速的提取, 提出了一种基于 ReliefF-GA 和随机森林算法的面向对象分类方法。本文以吉林省长春市典型城区为研究对象, 基于 Landsat8 多光谱遥感影像开展地表覆盖分类方法探索。研究流程主要包括三个阶段: 首先通过分割构建对象级分析单元, 继而采用光谱-纹理-空间多维特征融合策略, 形成包含 83 维特征参数的初始特征空间; 随后引入 ReliefF-GA 混合特征优选模型, 通过 ReliefF 算法的相关权重排序与遗传算法 (Genetic Algorithm, GA) 的全局寻优机制协同作用, 筛选出具有显著判别力的 9 维特征组合; 最终基于优化后的特征集构建随机森林分类器, 实验结果显示总体分类精度达 94.83%, Kappa 一致性系数达 0.9364。为验证方法有效性, 将其与特征空间优化特征选择算法和 K 最近邻分类算法进行比较, 结果表明, 提出的方法能更有效的用于面向对象土地覆盖分类。

关键词: 随机森林; 面向对象; 土地覆盖分类; 特征优选; 遗传算法

DOI: 10.69979/3029-2727.25.01.049

引言

土地规划管理、环境保护等基础工作的顺利开展, 离不开实时且可靠的土地覆盖信息, 其中遥感技术是关键的技术支撑。遥感技术凭借着宏观性、综合性、动态性以及快速性等优势, 已经成为获取土地覆盖信息的不二之选, 是最为行之有效的方法。近年来, 遥感影像分类领域的研究趋势呈现出向对象导向方法转移的显著趋势^[1], 区别与传统像素级处理模式, 该技术将具备丰富语义特征的地理实体作为基本单元, 通过整合空间结构与光谱特性, 有效提升了地物识别精度并实现了更符合认知规律的地表信息提取^[2]。

遥感影像分类研究中, 基于智能算法的数据解析方法备受关注。在众多分类模型中, 集成学习技术展现出显著优势, 特别是以决策树为基础的组合分类器。这类算法中, 以 RF 为代表的集成方法因其独特的性能特点而获得学界青睐。相较于传统分类器, 该算法不仅能够有效处理含有噪声的样本数据, 对不完整数据也表现出良好的适应性。目前, 这种分类技术已在多个学科领域得到成功应用, 其应用范围从最初的遥感分析逐步扩展到其它相关研究领域^[3]。

在面向对象分类中, 能够从影像对象中提取的特征数量很多^[1]。特征维数过高不仅会增加运算复杂度, 而且还可能导致分类精度的降低^[4]。为此, 本文引入了 Re

liefF-GA 特征选择算法。该方法采用两阶段筛选机制: 第一阶段运用 ReliefF 算法作为过滤式方法, 通过评估特征权重来识别并去除不相关特征和噪声数据; 第二阶段引入遗传算法作为封装式方法, 通过模拟生物进化过程, 利用选择、交叉和变异等操作对特征空间进行全局搜索, 最终确定最优特征组合^[5]。这种混合策略充分发挥了过滤式方法的高效性和封装式方法的精确性, 有效提升了特征选择的性能。

综上, 本文提出了一种基于 ReliefF-GA 和随机森林算法的面向对象分类方法, 并将其用于土地覆盖分类当中, 以验证该方法的有效性。

1 研究区概况及数据准备

1.1 研究区概况

研究区位于吉林省长春市内 (43° 39' 44" N~43° 58' 55" N, 125° 18' 28" E~125° 45' 24" E), 研究区的土地覆盖类型有居民地、耕地、林地、草地、道路、水体^[6]。

1.2 数据源及预处理

本研究所选遥感数据为 landsat8 影像, 获取时间为 2014 年 07 月 15 日, 空间分辨率 30m, 影像质量良好。在预处理阶段, 基于 ENVI5.1 软件平台, 采用辐射校准模块与 FLAASH 大气校正组件对 Landsat8 卫星影像实施

辐射量标定与大气干扰消除。随后,根据研究区域的空间边界对校正后的遥感数据进行地理裁切,确保数据范围与目标区域完全吻合。最后,选择第 2 至第 7 波段进行多光谱数据融合,生成包含六个特征通道的复合数据集,为后续空间分析构建多维度的光谱信息基础^[6]。

本次野外采样时间为 2014 年 7 月 12 日-2014 年 8 月 3 日,共收集样本点 847 个。把样本随机分为 2 组,分别是训练组和验证组,其中训练组样本用于训练分类模型,验证组样本用于计算分类结果精度。训练组和验证组的样本数量如表 1 所示。

表 1 各地类样本的数量

地类	数量	
	训练组	验证组
耕地	100	43
林地	100	43
草地	92	40
居民地	100	43
道路	100	43
水体	100	43
总计	592	255

2 理论背景

考虑到随机森林算法目前发展已较为成熟,所以在此不做赘述,关于其详细介绍见文献^[7]。下面是对 Relief-GA 特征选择算法的介绍。

2.1 Relief 算法

Relief 算法是属于特征权重算法,能够去除冗余特征,并且可以有效解决有噪声的回归问题及多类问题。其利用“假设间隔”对特征的分类能力进行评价,假设间隔 θ 的计算公式如下

$$\theta = \frac{1}{2} (\|x - M(x)\| - \|x - H(x)\|), \quad (1)$$

式中 $M(x)$ 、 $H(x)$ 分别是与样本 x 的异类和同类最邻近样本点。

2.2 封装算法

封装特征选择算法由分类器和搜索算法两部分构成。本文建立的封装算法是以遗传算法作为搜索算法,随机森林算法的分类精度作为评估函数,具体步骤如下:

Step1 使用 0、1 二进制字符串 $(x_1, \dots, x_N)$ 表示一个个体(染色体),该个体对应 N 维特征向量。

Step2 采用随机法生成 N_0 个初始串 $(\chi_0^1, \dots, \chi_0^{N_0})$ 作为初始群体。

Step3 以用随机森林算法分类的总体精度作为目标函数计算第 k 代种群中各个体适应值 $f(\chi_k^i)$ $i=1, \dots, N_0$ 。

Step4 针对第 k 代种群,利用最优解保存策略进行选择操作。

Step5 在第 k 代种群中,寻找适应值最高的前 (N_0-1) 个个体,对其进行交叉变异操作。将交叉变异后的 (N_0-1) 个个体与交叉变异前适应值最高的个体进行合并,形成新一代群体 $(\chi_{k+1}^1, \dots, \chi_{k+1}^{N_0})$ 。

Step6 若遗传代数满足终止条件,则计算第 $(k+1)$ 代种群中每个个体的适应值,将适应值最高的个体输出;否则令 $k=k+1$ 转 Step3。

3 试验结果与分析

3.1 影像分割

本文基于 eCognition 9.0 平台的多尺度分割模块,采用分形网络演化算法对 Landsat8 影像实施对象化处理。技术参数配置中,光谱通道权重均等分配(波段 2-7 各设 1),形态特征参数设置为形状因子 0.1、光谱因子 0.9,分割尺度定为 30,该参数组合是通过前期实验优选确定。

3.2 特征提取

本文构建了包含 83 维特征的空间分析数据集,涵盖光谱、纹理、几何三大类特征空间。光谱维度(24 维)通过计算各波段(B2-B7)的均值、标准差、最大像素值和最小像素值来表征地物光谱反射特性;纹理维度(48 维)采用灰度共生矩阵(GLCM)提取 6 个波段的熵值、同质度等 8 类纹理指标,刻画像元间的空间关联特征;几何维度(11 维)则通过形态学运算获取密度、长宽比等形状特征,量化地物轮廓的几何属性。

3.3 特征选择

本文采用 Relief-GA 混合特征选择策略优化分类特征集。首先通过 Weka 机器学习平台的 Relief 过滤式方法,基于特征权重评估对 83 维原始特征进行初筛(剔除冗余特征及噪声数据),获得 40 维候选特征子集;然后基于 Matlab 平台构建封装算法,通过选择、

交叉、变异等进化操作对候选集进行全局寻优，最终确定 9 维最优特征组合，如表 2 所示。

表 2 选择的特征

类别	特征			总计
光谱特征	2、7 波段的对象均值	4、5 波段的对象的最小像素值	2 波段的对象的标准差	5
纹理特征	3 波段的对象的非相似性	7 波段的对象的熵	5 和 7 波段的对象同质度	4

3.4 分类

基于选择的 9 维最优特征组合，利用随机森林算法对研究区进行土地覆盖分类，结果如图 1（a）所示。当利用随机森林算法分类时，需要预先设定生长树数目 N 的取值。为确定一个合适的 N 值，本研究进行了多次试验，最终设定 N 为 400。

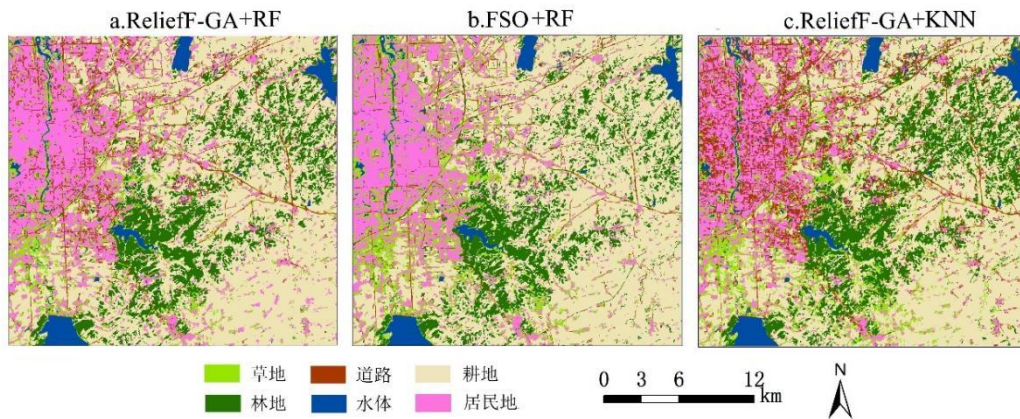


图 1 不同分类方法的土地覆盖分类图

3.5 精度分析

本文采用混淆矩阵对分类精度进行评估，基于验证

样本集构建了包含四类统计评估指标的精度评价体系：总体分类精度、Kappa 一致性系数、生产者精度和用户精度，具体计算结果详见表 3。

表 3 本文提出方法的分类精度混淆矩阵

分类数据	参考数据					
	居民地	道路	耕地	林地	水体	草地
居民地	1931	175	0	0	0	14
道路	0	2861	39	0	0	0
耕地	0	0	4215	138	10	142
林地	0	0	0	2643	0	47
水体	0	0	0	0	4083	0
草地	0	0	224	138	0	1257
生产者精度/%	100.00	94.24	94.13	90.54	99.76	86.10
用户精度/%	91.08	98.66	93.56	98.25	100.00	77.64
总体精度/%	94.83					
Kappa	0.9364					

结果如图 1（a）和表 3 所示，ReliefF-GA 与随机森林的算法组合在土地覆盖分类中展现出显著优势，总体分类精度达 94.83%（Kappa=0.9364）。单一地物类型（居民地、道路、耕地、林地、水体）的生产者精度与用户精度均突破 90% 阈值。但数据表明，草地、耕地、林地存在明显混淆现象，究其原因，7 月获取的 Landsat8 影像恰逢植被生长旺季，植被及物候特征趋同导致光谱可分性降低，增加了分类器判别难度。

3.6 与其它分类方法对比

为进一步验证提出方法的有效性，本文构建两种其它分类方法作为对比方法：①采用特征空间优化（feature space optimization, FSO）算法进行特征选择，随机森林算法进行分类，记为“FSO+RF”分类方法；②采用 ReliefF-GA 算法进行特征选择，K 最近邻（k-Nearest Neighbor, KNN）算法进行分类，记为“ReliefF-GA+KNN”分类方法，其中特征空间优化和 K 最近邻分别

是面向对象分类中比较有代表性的特征选择和分类方法。利用“FSO+RF”和“ReliefF-GA +KNN”对研究区影像进行分类的结果如图 1 (b) 和 (c) 所示。“FSO+RF”方法的总体精度为 91.44%、Kappa 系数为 0.8945; “ReliefF-GA +KNN”方法的总体精度为 88.00%、Kappa 系数为 0.8520。可见, 本文提出方法的总体精度和 Kappa 系数最高, 该对比结果进一步证明了本文提出的基于 ReliefF-GA 和随机森林的分类方法在遥感影像面向对象分类中的优越性。

4 结论

本文提出了一种基于 ReliefF-GA 和随机森林算法的面向对象分类方法, 并以 landsat8 遥感影像为数据源, 对吉林省长春市部分区域进行了土地覆盖分类, 取得了较好的分类效果, 总体精度为 94.83%、Kappa 系数为 0.9364。为了进一步证明提出方法的有效性, 将其与“FSO+RF”和“ReliefF-GA+KNN”分类方法进行比较, 对比结果进一步证明了本文提出的方法在遥感影像面向对象分类中的优越性。

参考文献

- [1] 肖艳, 王斌. 基于面向对象的极化雷达影像分类[J]. 红外与毫米波学报, 2020, 39(4): 505-512.
- [2] Bui Q T, Chou T Y, Hoang T V, et al. Gradient Boosting Machine and Object-Based CNN for L

and Cover Classification[J]. Remote Sensing, 2021, 13(14): 2709.

[3] Li H, Yu C, Xia J, et al. A Model Output Machine Learning Method for Grid Temperature Forecasts in the Beijing Area[J]. Advances in Atmospheric Sciences, 2019, 36(10): 1156-1170.

[4] 薛敬. 基于群体智能优化的高光谱遥感影像光谱特征选择研究[D]. 山西大学, 2024.

[5] Naghibi S A, Ahmadi K, Daneshi A. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping[J]. Water Resources Management, 2017, 31: 2761-2775.

[6] 肖艳, 姜琦刚, 王斌, 等. 基于 ReliefF 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.

[7] 董兰芳, 张军挺. 基于深度学习与随机森林的人脸年龄与性别分类研究[J]. 计算机工程, 2018, 48(5): 252-257.

基金项目: 吉林省科技厅项目 (YDZJ202201ZYTS499)

作者简介: 肖艳 (1988), 女, 黑龙江双城人, 博士, 副教授, 主要从事遥感影像分类研究。

通讯作者: 肖艳。