

基于决策树的蓬莱市土地覆盖信息提取

邱向红¹, 王周龙¹, 张明明¹, 张淑萍²

(1. 鲁东大学地理与规划学院, 山东烟台 264025; 2. 山东省泗水县实验中学, 山东泗水 273200)

摘要:以蓬莱市为研究区,对TM影像经过预处理后,进行典型地物的光谱分析。在此基础上,运用决策树分类法,选取影像的光谱特征值、NDWI值、NDVI值、K-T变换信息和DEM值等数据作为测试变量,选择适当阈值设定判别规则,建立决策树模型进行土地利用覆盖信息提取并做出精度评价。将其提取精度与监督分类结果精度进行比较,结果表明其分类精度有很大提高,尤其在研究区是丘陵地形的情况下,DEM数据的使用使林地、果园的可分性大大加强。

关键词:土地覆盖;信息提取;决策树;光谱特征;NDWI;NDVI;K-T变换;蓬莱市

中图分类号:P208

文献标识码:A

0 引言

土地覆盖及其变化是全球环境变化研究的重点和热点。遥感技术因其有多波段、多时相、全天候和覆盖范围广等优点,在提取土地覆盖信息的应用中是必要的技术手段^[1]。利用遥感技术提取土地利用信息的关键是将土地覆被进行分类。传统的计算机自动分类方法都是基于地物光谱值进行的,例如监督分类中的最大似然法(Maximum likelihood classification,简称mLC)有着严密的数学统计基础,但它的应用前提是假设数据源基于正态分布,这一假设在实际应用中存在明显缺陷,而且这种单纯依靠光谱值进行的分类容易产生同物异谱、异物同谱现象,因此研究人员发展了很多改进的分类方法,如神经元网络分类法、决策树分类法、模糊分类法等进行土地利用遥感信息的提取^[2-6]。

决策树分类是一个分层次的处理结构,一个决策树模型由一个根节点(Root node)、一系列内部节点(Internal nodes 分支)和若干叶节点(Terminal nodes)组成。每个内部节点对应一个决策判别规则,向下的2个分支对应是或是否符合此判别规则的2个叶节点。每一个叶节点可以是最终结果,也可以继续向下分。依次向下进行,直到分出所需要的

所有叶节点,且每个叶节点对应一个土地类型。运用决策树模型进行土地覆盖分类,关键是构建模型时各判别规则的设定,目前众多研究者将各种技术运用到决策树分类中,以达到判定函数最优化的目的^[7-10]。该文应用决策树分类法,采用多个判别规则,对蓬莱市土地覆盖信息进行提取。

1 研究区概况与数据源选取

蓬莱市地处胶东北部,庙岛海峡南岸,北濒渤、黄二海。全市总面积1 129 km²,境内山岭岗丘绵延起伏,属低山丘陵地貌类型,地势南高北低,由南向北逐步倾斜。南部为深山区,中部和北部为浅山丘陵区,地形复杂。蓬莱属暖温带季风大陆性气候,气温适中,变化平稳。森林植被多以经济林和防护林为主,且经济林种类繁多。全市人口众多,工农业较为发达,旅游业为其主导产业。

数据源信息采用一幅Landsat 5提供的TM多波段影像,其空间分辨率为30 m。成像时间为2004年10月5日,所含的6个波段(1~5,7)包括可见光、近红外和短波红外波段,可反映地表各种地物的光谱信息,进行不同地表特征识别。同时收集该区1:50 000地形图、2002年TM影像目视解译图及其他该市的相关研究报告为辅助资料。

* 收稿日期:2009-03-27;修订日期:2009-08-25;编辑:陶卫卫

作者简介:邱向红(1982—),女,山东德州人,在读硕士,主要从事土地利用与遥感监测应用研究。

2 基于决策树的土地覆被信息提取

2.1 数据预处理

该次研究选用影像已经过辐射校正,需进行几何精校正。首先通过控制点-多项式拟合校正方法,以已有校正完毕的1:50 000 地形图作为地理参考校正图像,在校正图像中共选取 17 个控制点,采用二次多项式双线性内插法进行几何精校正,并将误差控制在 0.5 个像元以内。同时将地形图等高线进行矢量化生成 DEM 图。

2.2 典型地物光谱分析

地物间存在光谱差异性,是对遥感影像进行分析解译的理论基础,TM 影像计算机自动分类,就是建立在全面掌握分类对象各波段光谱特性的基础上^[11]。根据影像目视解译初步判读,结合蓬莱市实际土地利用状况,将土地覆盖系统分为水域、建设用地、果园、林地、水浇地、旱地、滩涂和裸地 8 种类型,其中建设用地又包括居民地、道路和工矿用地。

为获取不同土地类型的光谱特征,实现土地信息自动提取,对每种土地类型各选取 10 个样点,测定所有样点在 6 个波段的光谱值,进行典型地物样点光谱值统计(表 1)。

表 1 典型地物样点光谱值统计

地物	建设用地	水域	林地	果园	水浇地	旱地	滩涂	裸地
B1	最大值 99	79	89	99	90	85	88	151
B1	最小值 81	71	81	92	80	79	8	88
B1	平均值 89.9	76.7	86.3	95.7	83.4	82	82.7	102.5
B2	最大值 48	36	35	44	41	44	42	82
B2	最小值 36	31	31	38	36	38	34	46
B2	平均值 42.7	33	33.1	40.7	37.9	40.9	38.4	55
B3	最大值 77	42	36	59	62	73	58	140
B3	最小值 51	35	27	43	51	59	39	71
B3	平均值 64.5	38.3	31.6	49.6	55.6	66.6	50.7	90.2
B4	最大值 59	22	96	83	58	70	45	106
B4	最小值 40	17	70	73	46	60	22	66
B4	平均值 51.8	19.7	86.5	77.2	50.7	65.5	35.8	78.2
B5	最大值 116	21	83	115	102	139	68	172
B5	最小值 62	10	64	86	81	113	22	130
B5	平均值 84.3	13.5	73.7	100	90.1	127.8	51.2	144.1
B6	最大值 67	9	35	68	57	77	36	107
B6	最小值 31	4	19	40	38	54	11	69
B6	平均值 47.4	5.6	25.8	49.7	49.1	66.1	24.7	80

为便于直观分析,根据各土地类型平均值做出光谱特征响应曲线(图 1)。由表 1 和图 1 可得到各

地物在 6 个波段间变化的趋势规律,根据这些规律寻找不同土地类型在多个特征变量中的特征值,并进一步分析各特征参数可逐步将各地物区分开来。

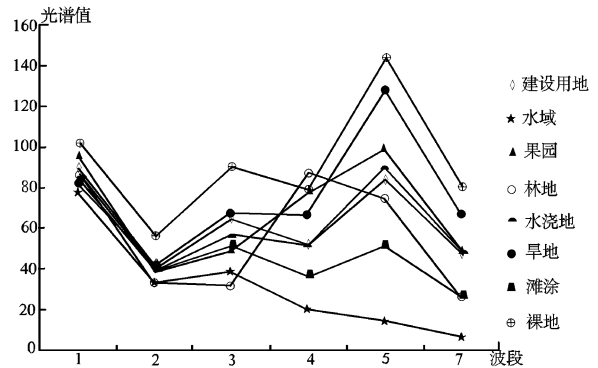


图 1 不同地物样点亮度值曲线

2.3 多种特征信息提取

在决策树分类模型中可以使用多种特征变量设立分类规则,如植被指数、光谱值、光谱值算术运算值、纹理信息和主成分分析等。选择合理的特征变量建立判别规则是运用决策树法分类的关键,该文在分析数据特征基础上,提出适合研究区提取土地覆被信息的特征变量。

2.3.1 归一化差异水体指数(NDWI)

由图 1,表 1 可知在绿光波段与近红外波段,只有水域和滩涂其 TM2 亮度值大于 TM4 亮度值,其他地物均表现相反。利用 1996 年由 McFeeters 提出的归一化差异水体指数(Normalized Difference Water Index, NDWI),基于 TM 影像其计算公式: $NDWI = (TM2 - TM4) / (TM2 + TM4)$,进行计算后统计各种土地类型样点结果,只有水域和滩涂的 NDWI 值大于 0,其他地物 NDWI 值都小于或等于 0。因此,利用 NDWI 指数差异性首先将水域、滩涂与其他地物区分开来。即首先根据 NDWI 值是否大于 0,将图像分为 2 大类,水域和滩涂一支,其他地物另一支。

在区分水域与滩涂时,利用 TM7 波段上水域最大亮度值小于滩涂最小亮度值这一规律,采用 $TM7 > 10$ 作为判别函数,设立阈值将两者分开,即图像上 $TM7 > 10$ 的像元归为滩涂,否则即为水域。

2.3.2 归一化植被指数(NDVI)

NDVI 指数常被用来进行植被信息提取,反映的是植被在红光波段和近红外波段反射与背景地物间的指标差异,在研究区内植被类型主要是林地和果园,利用 NDVI 指数将林地、果园与剩余地物进行

区分。经过多次试验对比,将阈值设为 0.1,即 $NDVI > 0.1$ 的像元属于植被。

研究区地处山地丘陵区,在区分林地、果园这样光谱反射较为相似的地物时,地形数据是有用的辅助数据^[12]。建立的模型中先采用生成的 DEM 栅格数据将山地区林地提取出来。果园生长海拔一般不足 100 m,因此取高程 100 m 为分类阈值,即高程大于 100 m,植被为林地。当海拔小于 100 m,植被兼有林地和果园 2 种,通过分析二者典型样点光谱值,发现果园的 TM5, TM7 波段光谱最小值之和为 126,林地在这 2 个波段最大值之和为 118,通过反复验证,取 $TM5 + TM7 > 120$ 为阈值,将平原区林地与果园分开。

2.3.3 区分裸地

裸地是直接出露的地表,地表为土质,表面没有植被或其他覆盖物,因此反射率高。通过对表 1 和图 1 研究发现,裸地反射值在各波段都较大,尤其在 TM5, TM7 波段表现明显。利用这一规律,经过反复实验,设定裸地的判别条件为: $TM5 > 140$ 且 $TM7 > 80$,满足这两个判别条件的像元归为裸地。

2.3.4 缨帽变换(K-T变换)

缨帽变换是一种特殊的主成分分析,是基于图像物理特征具有固定转换系数的处理方法。图像在经过缨帽变换处理后,共得到 6 个波段,其中前 3 个分量分别代表亮度(Brightness)、绿度(Greenness)和湿度(Wetness),分别可以提供亮度信息、植被覆盖生长情况和水分含量信息^[13]。利用 K-T 变换对建设用地、水浇地和旱地进行区分,在处理后的图像上选取以上 3 种地物类型样点进行统计(图 2)。

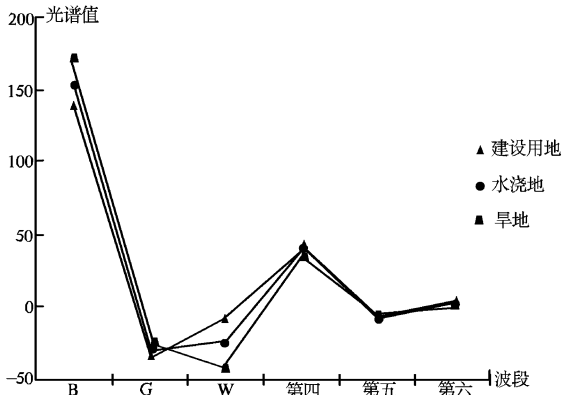


图 2 K-T 变换后建设用地、水浇地和旱地的亮度值曲线

通过分析,三者明显的变化主要发生在绿度和湿度

波段。从绿度过渡到湿度波段,建设用地光谱值呈上升趋势,旱地呈下降趋势,而水浇地曲线基本平行,变化幅度很小。根据上述分析,首先对绿、湿波段进行差值运算,符合 $K - T(3 - 2) > 10$ 条件的像元归为建设用地。区分剩余两类地物时,将绿、湿波段做比值运算,设定阈值为 0.8,即满足 $K - T2/3 < 0.8$ 判别条件的像元划为旱地,否则归于水浇地。

3 决策树试验分类及其精度评价

3.1 决策树分层设计方案

对各种地类的多种特征进行分析,选用适当特征设定判断条件,对蓬莱市土地覆被进行分类。如图 3 所示,在原始影像基础上,首先运用 NDWI 指数将水体和陆地分开, $NDWI > 0$ 的像元分到水体,否则为陆地,后利用 $TM7 > 10$ 这一波段阈值将水体分为滩涂与水域。运用 $NDVI > 0.1$ 这一判断条件,将林地和果园与剩余非植被类型分开,再利用 $DEM > 100$ m 和 $TM5 + TM7 > 120$ 这 2 个条件把林地与果园分开。因为裸地光谱反射率较高,设定 $TM5 > 140$ 和 $TM7 > 80$ 两个光谱阈值区分裸地。根据 K-T 变换的波段数据值, $K - T(3 - 2) > 10$ 的像元归为建设用地,差值小于等于 10 同时 $K - T2/3 < 0.8$ 的划为旱地,最终剩余水浇地。运行决策树分类模型后得到的土地信息提取结果如图 4。

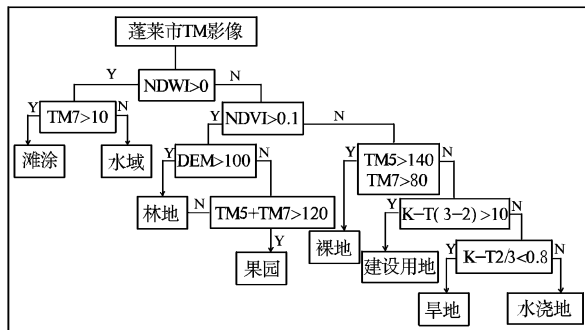


图 3 不同土地覆被信息自动提取流程图

3.2 精度评价及结果分析

3.2.1 精度检验

将由决策树方法提取的土地利用信息结果,在原始影像上进行目视判读检验。为保证评价精度,在采样时利用分层采样法。因为当影像上某些类别所占数量很少时,随机采样常会漏掉这些类别,采用分层采样即分别对每一类别进行随机采样。该次分类中,滩涂和裸地面积较小,采用分层采样法保证这

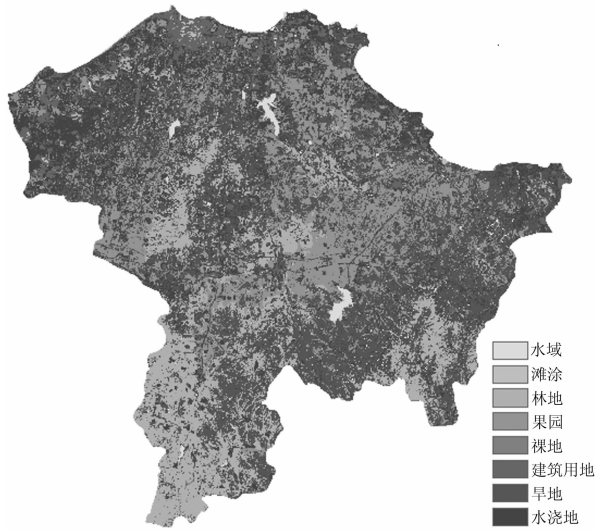


图4 决策树分类结果

2个类型至少分别有10个样点参与精度评价。最终8类地物共选取256个样点,建立混淆矩阵,以此计算各种地物的生产者精度、用户精度、总体精度和Kappa系数(表2)。同时将其与常规监督分类(图5)提取结果(表3)进行比较。

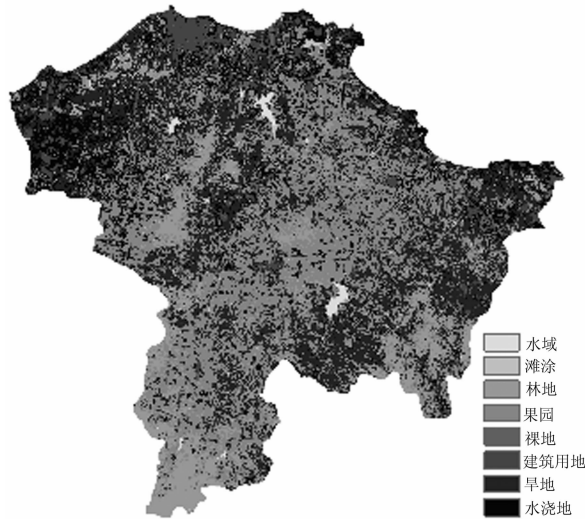


图5 监督分类结果

3.2.2 结果分析

通过精度检验的结果可以看出,决策树分类精度比监督分类有很大提高,总体精度提高了18.75%,Kappa系数提高了0.2332。在监督分类中因为林地和果园光谱特征相似,因此二者混分情况较多;而决策树分类模型中增加的DEM信息,使林地和果园在分类时不仅考虑光谱值还考虑高程因素,因此分类更加准确。区分裸地、建设用地、旱地、水浇地4种地类时,因为K-T变换的绿色和湿度

波段兼有植被覆盖和水分信息,利用4种地物在这两方面表现出特征值差异性,将其应用于分类比监督分类只考虑光谱信息分类精度有所提高。

表2 决策树分类法精度评价

类型	水域	滩涂	林地	果园	裸地	建设用地	旱地	水浇地	合计	生产者精度/%
水域	11	0	0	0	0	0	0	0	11	100
滩涂	0	10	0	0	0	1	0	0	11	90.91
林地	0	0	37	4	0	0	0	0	41	90.24
果园	0	0	1	43	0	0	0	0	44	97.73
裸地	0	0	0	0	10	0	0	0	10	100
建设用地	0	0	2	0	0	18	3	1	24	75
旱地	0	0	0	0	0	0	64	2	66	96.97
水浇地	0	0	0	1	0	3	11	34	49	69.39
合计	11	10	40	48	10	22	78	37	256	—
用户精度/%	100	100	92.5	89.58	100	81.82	82.05	91.89	-	-

注:总体精度=88.67%;Kappa系数=0.8624。

表3 监督分类法精度评价

类型	水域	滩涂	林地	果园	裸地	建设用地	旱地	水浇地	合计	生产者精度/%
水域	10	0	0	0	0	0	0	0	10	100
滩涂	5	5	0	0	0	0	0	1	11	45.45
林地	0	0	22	4	0	0	3	1	30	73.33
果园	0	0	13	45	0	0	12	4	74	60.81
裸地	0	0	0	0	4	0	3	5	12	33.33
建设用地	0	0	0	0	0	14	1	5	20	70
旱地	0	0	0	1	0	0	58	5	64	90.63
水浇地	0	0	0	2	0	5	7	21	35	60
合计	15	5	35	52	4	19	84	42	256	-
用户精度/%	66.67	100	62.86	86.54	100	73.68	69.05	50	-	-

注:总体精度=69.92%;Kappa系数=0.6292。

4 结束语

该次研究采用决策树分类法对蓬莱市土地覆盖信息进行提取,与监督分类法提取结果比较后显示出取得了较好的提取效果。该分类模型思路清晰,结构简单,易于理解和操作,在分类时考虑到研究区地处丘陵地带,地表起伏不均,地物破碎分布等特点,灵活运用各地物的多种特征定义分类条件。选用NDWI,NDVI,K-T变换波段值、DEM数据及波段光谱差值作为特征变量建立判别规则,具有一定的可行性。根据分类结果精度评价,表明建立的决策树模型具有很强的应用价值。

参考文献:

- [1] 罗湘华,倪晋仁. 土地利用/土地覆盖变化研究进展[J]. 应用基础与工程科学学报,2000,8(3):262-272.
- [2] 梁益同,胡江林. NOAA 卫星图像水体信息神经网络识别方法的探讨[J]. 应用气象学报,2001,12(1):85-90.
- [3] 张树清,陈春,万恩璞. 三江平原湿地遥感分类模式研究[J]. 遥感技术与应用,1999,14(1):54-58.
- [4] Carpenter G, Gail A. ART Neural Networks for Remote Sensing Vegetation Classification from Landsat TM and Terrain Data[J]. IEEE Transactions on Geoscience and Remote Sensing, 1997, 35(2):308-325.
- [5] Hansen M C, DeFries R S, Townshend J R G, et al. Global Percent Tree Cover at A Spatial Resolution of 500 Meters: First Results of The MODIS Vegetation Continuous Fields Algorithm[J]. Earth Interactions, 2003,7(10):1-15.
- [6] 王昆,戚浩平. 土地利用与土地覆盖遥感分类方法研究综述[J]. 山西建筑,2008,34(5):353-354.
- [7] 许卫东,尹球,匡定波. 小波变换在高光谱决策树分类中的应用研究[J]. 遥感学报,2006,10(2):204-210.
- [8] 张友静,高云霄,黄浩,等. 基于 SVM 决策支持树的城市植被类型遥感分类研究[J]. 遥感学报,2006,10(2):191-196.
- [9] 王圆圆,李京. 基于决策树的高光谱数据特征选择及其对分类结果的影响分析[J]. 遥感学报,2007,11(1):69-76.
- [10] 陈亮,张友静,陈波. 结合多尺度纹理的高分辨率遥感影像决策树分类[J]. 地理与地理信息科学,2007,23(4):18-21.
- [11] 李彤,吴骅. 采用决策树分类技术对北京市土地覆盖现状进行研究[J]. 遥感技术与应用,2004,19(6):485-487.
- [12] 王周龙,冯学智,刘晓玫,等. 秦淮河丘陵地区土地利用遥感信息提取及制图[J]. 遥感学报,2003,7(2):131-135.
- [13] 买买提沙吾提,塔西甫拉提·特依拜,丁建丽,等. 基于决策树分类法的塔克拉玛干南缘沙漠化信息提取方法研究[J]. 环境科学研究,2008,21(2):109-114.

Extraction of Land Covering Information Based on Decison Tree in Penglai City

DI Xiang-hong¹, WANG Zhou-long¹, ZHANG Ming-ming¹, ZHANG Shu-ping²

(1. Geographic and Planning College of Ludong University, Shandong Yantai 264025, China; 2. Sishui Shiyan Middle School in Shandong Province, Shandong Sishui 273200, China)

Abstract: In this paper, taking Penglai as a study area, after the TM images of this area are preprocessed, spectrum analysis of typical surface features are carried out. On these basis, by using decision tree classification, selecting spectral characteristics, NDWI, NDVI, K-T Transformation and DEM data as test variables, using proper thresholds for setting discriminating rules, a simple decision tree model is built. On the basis of these discriminating rules for extracting land covering information, an accuracy assessment is given to the result by stratified random samples. It is proved that the decision tree classification can get higher accuracy result compared with the supervised classification method. Especially, in the case of hilly topography in this area, by using DEM datas to classify model for distinguishing woodland and orchard, the result is more accurate.

Key words: Land covering; information extraction; decision tree; spectral characteristics; NDWI; NDVI; K-T transformation