林业科学研究 2014,27(1):001～005
Forest Research

文章编号:1001-1498(2014)01-0001-05

TM 影像决策树分类中的影响因素研究

张连华[1,2],庞 勇[1*],岳彩荣[2],李增元[1],范应龙[1],谭炳香[1],车学俭[1]

(1. 中国林业科学研究院资源信息研究所,北京 100091; 2. 西南林业大学林学院,云南 昆明 650224)

摘要: 以云南省西双版纳州一景 TM 影像为例,分析了影响分类回归树方法的主要因素。结果表明在其他因素均一致的情况下,训练数据如果使用涵盖各类别的外业调查数据比使用系统布设的训练数据分类精度更高,并且多种参数波段的选择也会有效地提高分类的精度。

关键词: 决策树分类; TM 影像; 训练数据; 植被指数; 波段组合

中图分类号: TP751.2 **文献标识码:** A

DOI:10.13275/j.cnki.lykxyj.2014.01.001

Factors Affecting Decision Tree Classification Method over TM Image

ZHANG Lian-hua[1,2], *PANG Yong*[1], *YUE Cai-rong*[2], *LI Zeng-yuan*[1],
FAN Ying-long[1], *TAN Bing-xiang*[1], *CHE Xue-jian*[1]

(1. Research Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China;
2. College of Forest, Southwest Forestry University, Kunming 650224, Yunnan, China)

Abstract: Taking one scene TM image of Xishuangbanna of Yunnan as an example, the main factors affecting the classification and regression tree method were analyzed. The results show that in the parameters under identical circumstances, the training data has higher classification accuracy if the field investigation data covering all the classification data were used rather than the system layout data. It also shows that selecting various bands of parameters can also improve the precision of classification effectively.

Key words: decision tree; TM image; training data; vegetation index; bands combination

由于遥感技术具有覆盖范围大、周期短、能反映动态变化、受地面条件限制少等优点,现代土地类型、森林类型制图多通过野外实地调查和室内遥感信息判读来分析进行[1]。其中分类一直是遥感技术领域研究的一项重要内容[2]。近年来随着计算机技术的飞速发展,计算机自动识别分类已经逐渐代替了早期的分类技术,成为了遥感应用的一个重要组成部分,也是当前遥感发展的前沿[3]。

目前遥感影像分类的方法有多种,如传统的基于数理统计的最大似然分类法,近期出现的神经网络分类法、支撑向量机分类法、专家系统分类法、面向对象分类法等。但这些方法或者算法过于复杂、难以理解,或者对分类者有较高的遥感和地学知识要求,而且方法受遥感影像本身的空间分辨率以及同物异谱、异物同谱现象的限制,会出现较多的漏分、错分,并导致分类精度降低,未能在更大领域得到推广与应用[4-5]。而分类回归树算法能够基于数据集中任何可用的属性特征来搭建一系列的二叉决策树,进而确定每一个像素所属的类型。该算法无须相关领域知识,且具有更高的分类精度和更快的

收稿日期: 2013-03-20

基金项目: 亚太森林恢复与可持续管理网络项目"Forest Cover and Aboveground Biomass Mapping in the Greater Mekong Subregion and Malaysia"(编号:2011PA004) 和国家 863 课题"全球森林生物量和碳储量遥感估测关键技术(编号:2012AA12A306)"资助。

作者简介: 张连华(1989—),男,山东聊城人,硕士研究生。主要研究方向:遥感数据处理与分析、3S 技术在林业中的应用等。电话:010-62888640,E-mail: sdzhanglh@126.com

* 通讯作者:主要研究方向:激光雷达森林参数反演、林业遥感机理模型等。电话:010-62888847,E-mail: caf.pang@gmail.com

处理速度,已成为目前遥感影像分类研究中一种重要的方法途径[6]。

本文以一景 TM 影像为例,结合森林资源一类清查数据以及外业实测调查数据,分析了不同的训练数据对决策树方法的影响。同时讨论了波段参数的选择对决策树方法的贡献。

1 研究区及数据

1.1 研究区概况

研究区位于我国云南省勐海县、景洪市大部以及勐腊县(如图 1 所示),地理位置为 99°44′～101°58′ E,20°44′～22°37′ N。近些年随着版纳州社会经济的发展,其经济发展方式发生明显变化,尤其是茶树以及橡胶树的大片种植导致该地区土地覆盖类型发生了巨大变化,因此以该研究区为试验区分析影响决策树遥感影像分类的因素。

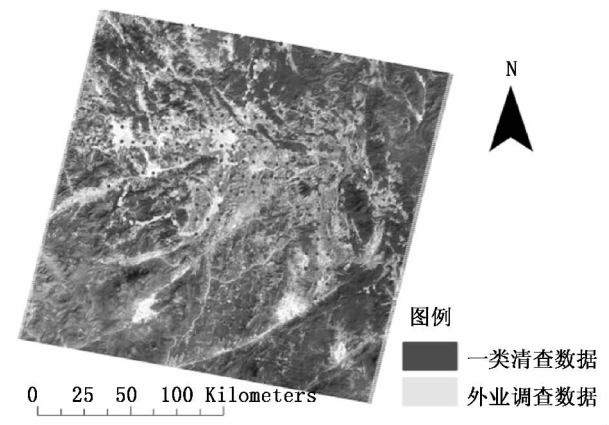

图 1 研究区域及所使用数据

1.2 试验数据

从美国地质调查局(United States Geological Survey)获得了版纳州 2010 年 2 月 14 日的一景无云覆盖 TM 影像,轨道号为 p130r045,成像时间对应该地区的干季。使用该景影像覆盖的一类调查样本的土地类型与森林类型等信息作为分类训练数据。

此外,于 2012 年 10 月 25 日—11 月 17 日在云南省版纳州进行了外业调查,外业根据 TM 遥感影像的光谱信息,利用 GPS 逐一考察每一斑块的土地类型信息,并且详细记录森林覆盖斑块的树种、龄级等信息,为确保样本的准确可靠性,同时使用数码相机分别取东、西、南、北 4 个方向进行拍照记录,为之后内业的分类研究提供真实的外业样地数据。为保证样本不受遥感影像分辨率的影响,将采集的每一样本再次进行了内业编辑以消除混合像元的影响。

2 影像特征波段选择

首先对 TM 数据进行预处理: 包括辐射定标以及基于“MODIS/6S”模型的大气校正,从而得到当地真实地表反射率影像。然后对于 TM 反射率影像进行特征波段提取与选择。

2.1 特征波段信息提取

由于实验区位于云南省南部,研究范围内不仅分布着橡胶林、茶园、农田等,还广泛分布着原始森林,对于这些复杂的植被类型仅使用个别波段或多个单波段数据分析对比是相当局限的[7],因此本研究提取多种不同的波段信息来有效地度量地表的覆盖状况[8]。

2.1.1 缨帽变换 缨帽变换是指在多维光谱空间中,通过线性变换、多维空间的旋转,将植物、土壤信息投影到多维空间的一个平面上,在这个平面上使植被生长状况的时间轨迹(光谱图形)和土壤亮度轴相互垂直。其中植被生长过程的光谱图形呈所谓的“缨帽”图形,而土壤光谱则构成一条土壤亮度线。本研究中选用的是缨帽变换的前三分量亮度、绿度与湿度[7]。

2.1.2 植被指数及其他参数波段提取 除缨帽变换之外,研究提取了差值植被指数、比值植被指数、归一化植被指数、归一化差异绿度指数与双差植被指数[7-9]。

除上述多种植被指数之外,还提取了一些用于提取其他地物类型的指数,如归一化建筑指数、改进型归一化水体指数、居民地旱地指数、居民地道路指数、缨帽变换湿度与绿度的比值、第一主成分变换、可见光三波段之和、红外三波段之和等[8-14]。

2.2 特征波段选择

由于上述提取的各特征波段之间难免会存在一定的相关性,因此还应该对这些特征波段进一步进行筛选以减少数据冗余。

2.2.1 特征波段选择原则 一般来说,选择最佳特征波段的原则有 3 点:(1)所选的波段信息量要大;(2)波段之间的相关性要小;(3)波段组合对所研究地物类型的光谱差异要大,可分性要大。目前比较广泛的选择方法有各波段信息量的比较、各波段间信息的相关性比较、最佳指数法(OIF)、各波段数据的熵和联合熵等方法[15]。

2.2.2 最佳指数法 最佳指数法(OIF)不仅考虑信息量而且考虑信息冗余,是目前波段组合选择的

理论最优方法。其定义如下:

$$OIF = \frac{\sum_{i=1}^{n} S_i}{\sum_{i=1,i<j=1}^{n} |R_{ij}|}$$

其中: S_i 为第 i 个波段的标准差, R_{ij} 为 i、j 两波段的相关系数。n 为所需的波段组合的个数。由该公式可知, OIF 越大,则相应组合波段的信息量就越大[15]。

使用该方法从上述提取的特征波段中共选择了 7 个最优的波段组合用于后续的分类研究之中,如表 1 所示。

表 1　最佳指数法选取的特征波段组合

特征波段	表达式
湿度	$WI = 0.144\,6TM1 + 0.176\,1TM2 + 0.332\,2TM3 + 0.339\,6TM4 - 0.621\,0TM5 - 0.418\,6TM7$
差值植被指数	$DVI = TM4 - TM3$
比值植被指数	$RVI = TM4/TM3$
归一化差异绿度植被指数	$NDGI = (TM2 - TM5)/(TM2 + TM5)$
双差植被指数	$DDVI = TM4 - TM3 - (TM3 - TM2)$
居民地道路指数	$TM5 - TM3$
湿度绿度比值	WI/GI

其中 GI 表示缨帽变换的绿地信息,表达式为 $GI = -0.272\,8TM1 - 0.217\,4TM2 - 0.550\,8TM3 + 0.772\,1TM4 + 0.073\,3TM5 - 0.164\,8TM7$

3　决策树法遥感影像分类

3.1　决策树方法

CART 是 Breiman 于 1984 年提出的决策树构建算法,其基本原理是通过对由测试变量和目标变量构成的训练数据集的循环分析而形成二叉树形式的决策树结构。CART 算法采用经济学中的基尼系数(Gini Index)作为选择最佳测试变量的准则。其中基尼系数的定义如下:

$$Gini = 1 - \sum_{j}^{J} P^2\left(\frac{j}{h}\right) \qquad P\left(\frac{j}{h}\right) = \frac{n_j(h)}{n(h)}$$

式中: $P\left(\frac{j}{h}\right)$ 是从训练样本集中随机抽取一个样本,当某一测试变量值为 h 时属于第 J 类的概率; $n_j(h)$ 为训练样本中测试变量值为 h 时属于第 J 类的样本个数; $n(h)$ 为训练样本中该测试变量值为 h 的样本个数; J 为类别个数[7]。

3.2　决策树生成

CART 算法从众多的预测属性中选择一个属性或多个属性的组合,作为树节点的分裂变量,把测试变量分到各个分支中,重复该过程建立一棵充分大的分类树,然后用剪枝算法对该树进行剪枝,得到一系列嵌套的分类树,最后用测试数据对该一系列分类树进行测试,从中选择最优的分类树[7]。

为了避免生成的决策树过大,定义最大树深度为 7,同时结合训练样本的大小定义剪枝算法中子节点最小个案数为 15,父节点最小个案数为 30,以保证后期不同训练样本、不同波段组合分类结果之间的可比性。

3.3　精度验证

首先 CART 决策树生成过程中可以使用交叉验证的方法来验证决策树的精度。即将输入的训练样本按一定百分比进一步分为训练样本与验证样本。但是由于遥感影像相邻像素之间光谱信息会相互影响,因此这种方法不仅依赖于训练样本的质量,而且会因为验证样本与训练样本相邻而导致过高估计整幅影像的分类精度。即交叉验证的方法只能说明使用该训练数据生成的决策树精度,而不能表达利用该决策树规则生成的整幅影像的分类精度。

为了避免上述问题,利用遥感软件中验证点分层随机生成的方法,使这些验证点平均分布在影像上并且保证各类别验证点的比例与分类类别的像元数比例一致,同时保证最小类别随机点数[16]。利用这一套验证数据分别验证不同训练样本、不同波段组合产生的分类结果。

4　分类结果比较与分析

在保证上述决策树生成以及验证数据一致的条件下,分别改变训练数据以及波段组合方式来进行决策树法分类,并对结果进行比较分析。

4.1　不同训练数据的分类结果比较分析

研究使用原始的 6 个 TM 反射率波段为输入波段,而训练数据分别使用一类清查数据以及外业调查数据。如图 1 所示,一类清查数据是系统布设,每隔 6 × 8 km 一个样本点; 而外业调查数据则根据遥感影像光谱特性以及交通线路进行布设。由表 2 中两种不同训练数据得到的精度验证结果可知: 利用外业调查数据作为训练样本的分类精度明显高于利用一类清查数据的分类精度,尤其是橡胶林、农田、城镇的分类精度。分析其原因主要是因为一类清查是以森林资源调查为主,而且布设方式为 6 × 8 km 的系统布设,对于以森林为主要类型的西双版纳州而言其样本基本全部是森林,而其他的地物类型样本则相对较少,无法表达整景影像的光谱与地物信

息，尤其是光谱异质性明显的灌木林、橡胶林、城镇等，其样本量反而较少。因此虽然一类清查数据具有权威性真实性，但是并不能当作训练数据用于分类研究。

表 2 不同训练数据的分类精度比较

类别	一类清查数据			外业调查数据		
	样本量	生产者精度/%	用户精度/%	样本量	生产者精度/%	用户精度/%
针叶林	3	0	0	3	0	0
常绿阔叶林	122	80.52	91.13	71	87.41	92.93
橡胶林	12	60.78	52.25	131	78.06	65.41
灌木林	6	36.11	32.50	35	44.93	54.39
农田	46	62.50	53.19	136	82.93	80.95
城镇	3	58.33	53.85	68	75.00	75.00
水体	2	81.81	90.00	21	81.81	90.00
总体	194	70.08		465	80.34	

4.2 不同参数波段的分类结果比较分析

利用外业调查数据作为训练数据，分别选取不同的波段组合方式进行分类，使用同一套验证样本对分类结果进行验证，如表 3 所示，精度比较结果说明通过最佳指数法选择参数波段的分类精度比使用原始影像的分类结果精度要高。

表 3 不同波段参数的分类精度比较 %

类别	原始反射率波段		最佳指数法选择波段		缨帽变换三波段		缨帽变换三波段结合原始反射率六波段	
	生产者精度	用户精度	生产者精度	用户精度	生产者精度	用户精度	生产者精度	用户精度
针叶林	0	0	33.33	50.00	0	0	33.33	100
常绿阔叶林	87.41	92.93	85.99	95.51	76.25	96.98	86.22	94.53
橡胶林	78.06	65.41	85.81	61.29	91.61	48.97	85.81	67.51
灌木林	44.93	54.39	42.03	65.91	18.84	92.86	57.97	68.97
农田	82.93	80.95	82.93	77.27	78.05	74.42	78.05	72.73
城镇	75.00	75.00	75.00	56.25	83.33	47.62	75.00	47.37
水体	81.81	90.00	81.81	90.00	90.91	90.91	81.81	100
总体	80.34		81.04		74.16		82.44	

由于该方法是依据数学理论得出，而缨帽变换是根据光谱空间变换将原始 6 个反射率波段变换为亮度、绿度、湿度，明显地增强了不同地物之间的差异性，依据的是遥感光谱理论，因此使用其与 6 个反射率波段结合进行分类实验，发现得到的精度比仅使用最佳指数法的精度要高[8]。但是仅使用缨帽变换的 3 个波段进行分类精度则明显降低。分析其原因可能是因为缨帽变换 3 波段的光谱信息较少，不足以表达整景影像的光谱差异性。

此外通过表中不同地物类型的精度发现影响分类总体精度的主要是针叶林、灌木林以及城镇。其中针叶林精度很低主要是因为研究区位于热带地区，其针叶林分布区域较小，一般为山顶的狭长区域，而其他地方的分布面积更小。而 TM 影像分辨率为 30 m，这些小的范围以及狭长区域极易受到遥感影像混合像元的影响。城镇精度较低主要也是因为遥感影像的分辨率不足以支持区分到小的村庄、寨子等区域，使其受周围农田、灌木林等其他地物的光谱影响较大。灌木林分类精度较低则主要是因为其与橡胶林、常绿阔叶林的光谱较为相似，尽管使用较多的波段组合其精度提高也十分有限，仅达到 57.97%。下一步研究将分析若使用多时相(如干季与湿季)的遥感影像进行分类看精度是否会提高[10]。

此外，研究还发现分类后处理会对分类精度有明显的改善作用，一般精度会提高 7% 左右。图 2 为使用缨帽变换三波段与原始反射率六波段结合得到的分类结果。

5 结论

决策树方法目前已广泛应用于遥感影像的分类中，该方法在拥有可靠的训练数据时能快速准确的进行机器学习并对遥感影像进行自动分类，并具有可靠的精度。本文对影响该方法分类精度的几个因

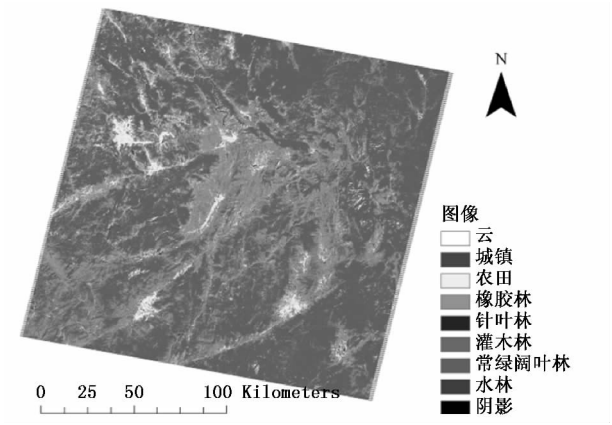

图 2　CART 分类结果

素进行了分析,得到如下结论:

(1) CART 生成过程中定义的最大树深度以及剪枝算法中子节点、父节点的最小个案数会影响最终的决策树精度,其值需要根据训练样本数据以及实际分类要求而定。

(2) 由于一类清查数据使用的是 6 × 8 km 的系统布设方法,对于以森林为主要类型的西双版纳州而言其样本基本全部是森林,而其他的地物类型样本则相对较少,尤其是光谱异质性明显的灌木林、橡胶林、城镇等,无法表达整景影像的光谱与地物信息。因此使用依据遥感影像的光谱差异性进行外业调查得到的训练数据比仅使用一类清查数据进行分类得到的结果精度要高。

(3) 由于最佳指数法在一定程度上减少了数据冗余,并且多种波段的组合也增加了信息量,因此依据最佳指数法选择的参数波段进行分类比仅使用原始反射率影像进行分类得到的精度要高。

(4) 缨帽变换方法是根据遥感光谱空间理论进行的数据变换,在一定程度上增强了不同地物之间的光谱差异性,研究发现在西双版纳州影像中缨帽变换三主成分亮度、绿度、湿度与原始六反射率波段结合共同参与分类得到的精度比最佳指数法进行分类得到的精度要高。

(5) 分类后处理可以减少分类过程中产生的斑点噪声、孤岛等现象,对分类的精度影响比较大,研究发现该景影像经过后处理的精度会提高 7% 左右。

参考文献:

[1] 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

[2] 陈　云,戴锦芳,李俊杰. 基于影像多种特征的 CART 决策树分类方法及其应用 [J]. 地理与地理信息科学,2008,24(2):33 - 36

[3] 罗来平,宫辉力,赵文吉,等. 遥感图像决策树分类器研究与实现 [J]. 遥感信息,2006(3):13 - 15

[4] Huang X, Jensen J R. A Machine-learning Approach to Automated Knowledge-based Building for Remote Sensing Image Analysis with GIS Data [J]. Photogrammetric Engineering and Remote Sensing, 1997,63(10):1185 - 1194

[5] Wickham J D, Stehman S V, Smith J H, *et al*. Thematic Accuracy of the 1992 National Land-Cover Data for the Western United States [J]. Remote Sensing of Environment, 2004,91:452 - 468

[6] Lawrence R L, Wright A. Rule-Based Classification Systems Using Classification and Regression Tree (CART) Analysis [J]. Photogrammetric Engineering and Remote Sensing, 2001,67(10):1137 - 1142

[7] 齐　乐,岳彩荣. 基于 CART 决策树方法的遥感影像分类 [J]. 林业调查规划,2011,36(2):62 - 66

[8] 赵英时. 遥感应用分析原理与方法 [M]. 北京: 科学出版社,2003

[9] 韩　涛. 用 TM 资料对祁连山部分地区进行针叶林、灌木林分类研究 [J]. 遥感技术与应用,2002,17(6):317 - 321

[10] 权维俊,郭文利,叶采华,等. 基于 TM 卫星影像获取北京市水体密度指数与植被覆盖指数的方法 [J]. 南京气象学院学报,2007,30(5):610 - 616

[11] 韩丛丛,逄杰武,吴泉源,等. TM 影像中居民地提取的决策树方法研究——以烟台市为例 [J]. 遥感信息,2007(6):73 - 75

[12] 邸向红,王周龙,张明明,等. 基于决策树的蓬莱市土地覆盖信息提取 [J]. 山东国土资源,2009,25(11):52 - 55

[13] 吴　见,彭道黎. 基于 TM 影像的多伦县土地利用信息提取 [J]. 东北林业大学学报,2004,38(10):88 - 94

[14] Baraldi A, Puzzolo V, Blonda P, *et al*. Automatic Spectral Rule-based Preliminary Mapping of Calibrated Landsat TM and ETM + Images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2006,49(9):2563 - 2586

[15] 李石华,王金亮,陈　姚,等. 多光谱遥感数据最佳波段选择方法试验研究 [J]. 云南地理环境研究,2005,17(6):29 - 33

[16] 杨阿强. CBERS 数据用于老挝土地覆被分类的方法论研究 [D]. 北京: 中国科学院地理科学与资源研究所,2009