改进随机决策树群算法在监督分类中的应用

胥 海 威[1], 何 宽[2]
(1. 中南大学信息物理工程学院, 湖南 长沙 410083; 2. 黄河水利职业技术学院测绘工程系, 河南 开封 475004)

摘要: 通过添加树平衡系数、设定节点不纯度和区分样本类型, 对现有的随机决策树群算法进行了改进, 提出了改进的随机决策树群算法。以广东省龙门县土地覆盖的 ALOS 遥感影像为研究对象, 利用改进的随机决策树群算法对研究对象进行遥感监督分类, 并将研究结果同传统的最大似然分类方法的结果进行对比, 发现分类总体精度从 81. 46% 提高至 92 45%, Kappa 系数达 0. 9091。改进的随机决策树群算法考虑了极不均衡决策树、节点不纯度和训练样本区分对随机决策过程运行效率的影响, 可有效提高遥感分类效率和分类精度。

关键词: 遥感影像; 自动分类; 土地覆盖分类; 随机决策树群算法; 模式分类
中图分类号: TP75 文献标识码: A 文章编号: 1672- 0504(2010)06- 0038- 03

图像分类目的是将图像中每个像元根据其在不同波段的光谱亮度、空间结构特征或者其他信息, 按照某种规则或算法分为不同的类别[1]。决策树算法作为一种监督分类算法, 其非参数性和树结构特性使之具有灵活、直观、清晰、运算效率高、良好的稳健性和鲁棒性等特点[2]。目前决策树分类的研究成果很多: 如刘勇洪等评价多种分类器在华北地区土地覆盖遥感分类中的性能[2]; 陈君颖等[3- 5] 结合纹理信息使分类精度得到进一步提高, 但该算法需要人工干预, 针对每幅影像都要做大量的试验, 如果数据源发生变更, 需重新对新的影像进行试验, 其可移植性较差。本文针对上述缺点, 在前人研究的基础上, 使用随机决策树群算法并以龙门地区影像为研究区进行试验。

1 研究方法的提出

1.1 随机决策树群算法

针对决策树的上述缺陷, Breiman 于 1996 年提出 Tree Bagging 算法[6], Ho 于 1998 年提出 Random Subspace 算法[7], Breiman 于 2001 年提出 Random Forests 算法[8], 对上述问题做了改进; 在此基础上, Pierre 等于 2006 年提出了随机决策树群算法。与普通随机算法相同, 随机决策树群也是根据影像亮度特征, 以树形结构表示分类集合, 产生规则与发现规律[9]。每个随机决策树都由根节点、内部节点和叶节点组成。该算法最大的特点就是对专家知识的依赖较少, 但其运算量增加, 在对数据集进行分割时, 所选择的属性和阈值都是随机所得。在极端情况下, 该算法所生成的决策树可以不受训练样本的影响, 而且其随机程度可以通过对参数的调整进行控制[10]。对该算法的详细描述如下[10]:

创建随机决策树群(S)
输入: 训练样本集 S
输出: 树群 $T = \{t_1, t_2 \cdots, t_M\}$
- For i= 1 to M
  t_i= 生成随机树(S);
- 返回 T

生成随机树(S)
输入: 训练样本 S
输出: 决策树 t
- 如果下面情形之一, 返回叶子节点:
(1) 训练样本 S 的个数< 最小分类数
(2) 节点不纯度< 阈值(改进 2)
- 否则:
1、从属性集中选取 k 个属性$\{a_1, a_2, \cdots, a_k\}$;
2、对全部属性 a_i(i= 1, 2, …, k)(改进 4), 随机划分数据(S, a_i); 生成 K 个集合$\{s_1, s_2, \cdots, s_k\}$
3、对 K 个划分集分别计算得分 Score(改进 1), 取分值最高的划分方法 s^*;
4、根据划分 s^* 将样本集 S 划分为左数据集 Sl 和右数据集 Sr;
5、针对 Sl 和 Sr, tl= 生成随机树(Sl), tr= 生成随机树(Sr);
6、依照 s^* 创建分割节点, tl 和 tr 分别为左右子树;
7、最终返回随机决策树 t.

随机划分数据(S, a)
输入: 训练样本 S 和属性 a
输出: 划分方法
- 计算样本集 S 中属性 a 的最大值 $a^s_{\max}$ 和最小值 $a^s_{\min}$
- 在$[a^s_{\min}, a^s_{\max}]$中随机选择切点 a_c
- 返回划分集$[a< a_c]$

$$Score = \frac{2I^s_c(S)}{H_s(S) + H_c(S)}$$

收稿日期: 2010- 08- 24; 修订日期: 2010- 10- 13
作者简介: 胥海威(1981-), 男, 博士研究生, 研究方向为遥感分类与数字图像处理等。E- mail: haiweixu@126.com

$I_c^s(S)$表示划分结果 s 和类别 c 的互信息
$H_s(S)$表示训练样本中针对划分 s 的信息熵
$H_c(S)$表示训练样本中针对类别 c 的信息熵

1.2　随机决策树群算法的改进

本试验在该算法的基础上, 使用遥感影像数据进行决策树的生成和验证。在将随机决策树群算法运用到遥感影像分类的过程中, 需要对算法进行改进, 以提高运行效率。主要包括: 1) 在应用过程中发现根据上述算法生成的树群中会出现极不均衡的决策树, 即训练样本划分的数据过早地成为叶节点, 但另一分支却达不到节点不纯度的要求。针对这一情况, 在计算相关划分得分 Score 时, 添加树平衡系数 TreeBalance, 大大减少了上述情况的发生。2) 在训练样本数目较大时, 如果设定节点不纯度的阈值为 0, 将会出现许多样本个数仅为 1 的数据集, 很大程度上降低了运行效率, 所以本试验中设定节点不纯度为 0. 1。3) 随机决策树群算法将全部样本作为训练样本, 使用这些样本进行精度验证。本研究样本的数目足够多, 为了提高验证样本的独立性和分类精度的可信性, 随机将样本按 3∶ 1 的比例分为训练样本和验证样本(表 1)。4) 根据算法的表述, 为了保证每次生成决策树的独立性, 参与分类的属性应为全部属性的一部分, 而不宜使用全部属性。但是本次试验所用遥感影像波段数较少, 将不能保证决策树的完全生长, 因此使全部的波段参与分类。

表 1　各类别训练样本与验证样本数
Table 1　The training samples and validation samples in all categories

类别	水体	水田	植被	旱地	居民地	道路	合计
训练样本数	373	235	380	501	380	363	2 232
验证样本数	139	100	147	118	118	133	755

2　试验与分析

2.1　研究区概况与数据预处理

研究区位于广东省惠州市龙门县(东经 114°9′ ~ 114°20′, 北纬 23°44 ~ 23°52′), 地形以山地和丘陵为主。本试验使用影像为 ALOS 遥感影像, 空间分辨率为 10 m, 提取样本所用的影像大小为 1 909 × 1 612(图 1)。

训练数据的质量在很大程度上影响着制图精度[11]。由于缺少与影像对应区域地面的实际样本, 为了保证所选择样本的代表性, 参考 1∶ 100 万土地利用数据库, 结合遥感目视解译, 选出各类有代表性的样本 2 987 个, 并随机将样本按 3∶ 1 的比例分为训练样本与验证样本, 表 1 为试验精度较好的各类别训练样本与验证样本数。

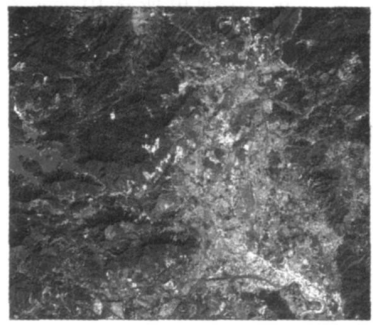


图 1　试验区影像
Fig. 1　ALOS image for the study area

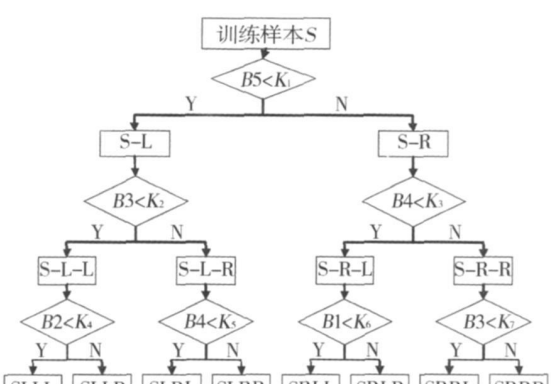
本试验所采用的 ALOS 影像包括 4 个波段, 由于样本类别包括植被和水体, 因此采用常见的归一化植被指数(NDVI)提取植被, 同时结合水体遥感, 定义水体指数 $WI = DN_B/DN_R$(蓝光波段数值与红光波段数值的比值), 作为提取水体的一种尝试。将上述的 $NDVI$ 和 WI 作为两个波段数据, 与影像的 4 个波段一起组成 6 个波段参与影像的分类。

2.2　分类结果与精度评价

由于算法随机性很强, 所以程序的分类精度也相差颇大, Kappa 系数低至 0. 41, 高至 0. 85 左右, 表 2 为分类结果较好、精度较高的一次试验精度评价, 相应的分类规则见图 2。由于该决策树较复杂, 本文

表 2　分类精度评价
Table 2　Accuracy assessment of classification

样本个数		实际类别						
		水体	水田	植被	旱地	居民地	道路	小计
分类类别	水体	139	0	0	0	0	0	139
	水田	0	80	0	0	4	11	95
	植被	0	3	147	0	0	3	153
	旱地	0	9	0	116	5	6	136
	居民地	0	8	0	0	104	1	113
	道路	0	0	0	2	5	112	119
	小计	139	100	147	118	118	133	755
总体精度= 92. 45%　Kappa 系数= 0. 9091								



注: $B1$、$B2$、$B3$ 分别代表红光、绿光、蓝光波段亮度值, $B4$ 代表近红外波段亮度值, $B5$ 代表 $NDVI$ 值, $B1$ 代表 WI 值。K_1 = 0 3095, K_2 = 94 2381, K_3 = 43 5495, K_4 = 72 8879, K_5 = 101. 1330, K_6 = 47. 2406, K_7 = 82 8564, K_8 = 52 6128, K_9 = 64 0968, K_{10} = 56 0059, K_{11} = 84 1442, K_{12} = 65. 1799, K_{13} = 90 0207, K_{14} = 82 6105, K_{15} = 0 4846

图 2　分类方法流程
Fig. 2　Classification flow chart

仅将部分流程画出, 并列出所选用的属性与阈值。与经典的最大似然法、最小距离法、平行六面体法、Mahalanobis 距离法、非监督分类方法(ISODATA TA 法)相比(表 3), 本文分类方法的精度有所提高。

表 3　不同分类方法结果精度比较
Table 3　Comparison between different classification methods

分类方法	总分类精度	Kappa 系数	运行效率
ISODATA TA 法	—	—	较高
最大似然法	81.46%	0.7838	一般
平行六面体法	63.97%	0.5725	一般
Mahalanobis 距离法	74.17%	0.6906	一般
最小距离法	76.16%	0.7139	一般
随机决策树群方法	92.45%	0.9091	高

使用该算法所得的分类规则处理具有代表性的区域影像, 将分类结果图与分类精度最高的 Envi 最大似然法分类结果比较如图 3(见封 3)。

3　结语

本文在前人研究的基础上, 提出改进的随机决策树群算法, 并对广东省龙门市 ALOS 遥感影像进行高分辨率影像监督分类, 分类精度从 81.46% 提高至 92.45%, Kappa 系数达 0.9091。可见改进后的随机决策树群算法能更好地用于遥感分类, 有助于提高分类精度和效率。但是, 高分辨率影像的波段数较少, 分类效果仍然不理想, 需要进一步研究如何结合高分辨影像和高光谱影像以提高影像的分类精度和效率。

参考文献:

[1] 赵英时. 遥感应用分析原理与方法[M]. 北京: 科学出版社, 2003. 194- 208.

[2] 刘勇洪, 牛铮, 徐文明, 等. 多种分类器在华北地区土地覆盖遥感分类中的性能评价[J]. 中国科学院研究生院学报, 2005, 22(6): 724- 731.

[3] 陈君颍, 田庆久. 高分辨率遥感植被分类研究[J]. 遥感学报, 2007(11): 221- 227.

[4] 申文明, 王文杰, 罗海江, 等. 基于决策树分类技术的遥感影像分类方法研究[J]. 遥感技术与应用, 2007, 22(3): 333- 338.

[5] 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.

[6] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123- 140.

[7] HO T. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832- 844.

[8] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45: 5- 32.

[9] QUINLAN J. Introduction of decision trees[J]. Machine Learning, 1986, 5: 239- 266.

[10] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees[J]. Machine Learning, 2006, 63: 3- 42.

[11] FRIEDL M A, BRODLEY C E, STRAHLER A H. Maximizing land cover classification accuracies produced by decision trees at continental to global scales[J]. IEEE Transactions on Geoscience and Remote Sensing, 1999, 37(2): 969- 977.

An Improved Random Decision Trees Algorithm with Application to Supervised Classification

XU Hai- wei[1], HE Kuan[2]

(1. School of Infophysical and Geometrics Engineering, Central South University, Changsha 410083;

2. Department of Surveying and Mapping, Yellow River Conservancy Technical Institute, Kaifeng 475004, China)

Abstract: An improved Random Decision Trees algorithm with application to land cover remote sensing classification was proposed in this paper. Firstly, an improved Random Decision Trees algorithm was presented by adding tree balance factor, setting node impurity and distinguishing sample types. Secondly, by taking the ALOS images of Longmen City of Guangdong Province in China as study area, the remote sensing classification was conducted using the improved Random Decision Trees algorithm. Finally, a comparison study was proceeded to compare the improved Random Decision Trees algorithm with Maximum Likelihood Classification method. The results indicate that the classification precision is improved from 81.46% to 92.45% and Kappa coefficient is up to 0.9091. The improved Random Decision Trees algorithm can improve the efficiency and accuracy of land cover remote sensing classification.

Key words: remote sensing image; automatic classification; land cover classification; Random Decision Trees; pattern classification

第 25 页插图

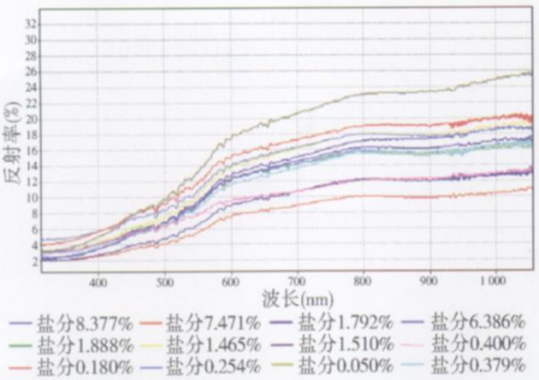


图 1　不同盐分土壤的光谱曲线

Fig. 1　Spectrum curves of soils with different salinity

第 40 页插图

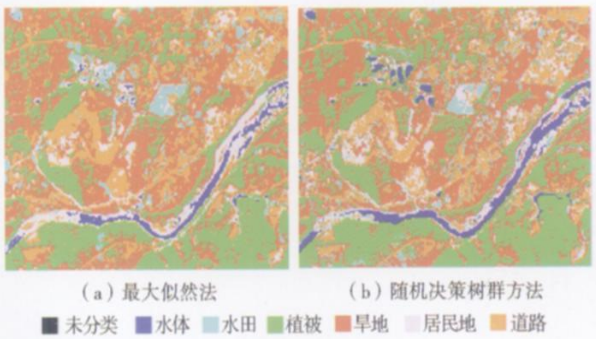


图 3　最大似然法和随机决策树群方法分类结果比较

Fig. 3　The classification results by Maximum Likelihood and Random Decision Trees

第 111 页插图

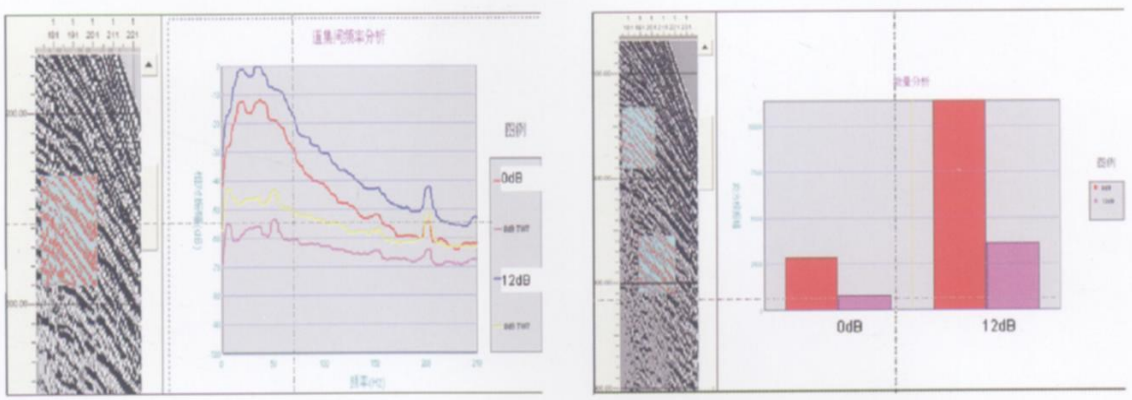


图 3　前放增益对比分析

Fig. 3　Comparative analysis of pre-amp gain

第 111 页插图

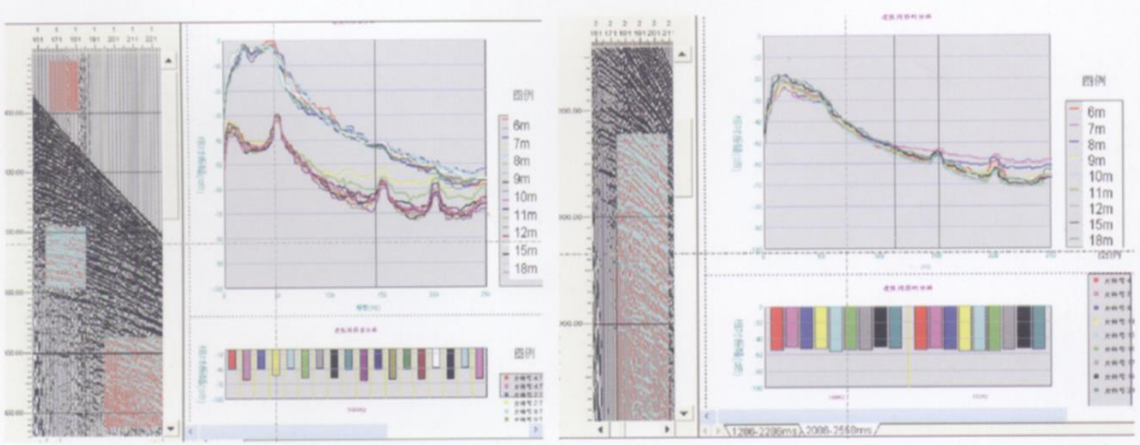


图 4　试验点Ⅰ井深对比分析

Fig. 4　Comparative analysis of well-depth at experimental location Ⅰ