文章编号:1671-8860(2013)12-1460-05　　文献标志码:A

AdaTree 算法在遥感影像分类中的应用

张晓贺[1,2]　翟　亮[2]　张继贤[2]　杨享兵[3]

(1　68332 部队,渭南市,715200)
(2　中国测绘科学研究院,北京市莲花池西路 28 号,100830)
(3　承德市国土资源局,承德市东环路,067000)

摘　要:目前的遥感影像分类研究中,决策树的生成完全依赖于现有的数据挖掘软件,缺少对决策树算法的深入研究和改进。本文以遥感影像分类为背景,采用 BoostTree 算法作为模型,通过算法改进构建了一种新的复合决策树算法——AdaTree,并以该算法为基础,设计实现了决策树遥感影像分类系统。以 AdaTree 算法作为分类器,分别对 Landsat7 ETM+影像和 WordView2 影像进行了基于像元和面向对象的分类实验,并与 BoostTree 和 SVM 算法进行了比较。实验结果表明,AdaTree 算法在分类精度上要优于 BoostTree 和 SVM 算法,平均 Kappa 系数分别达到 0.905 2 和 0.939 8。

关键词:决策树;AdaBoost;分类;遥感
中图法分类号:P237.4;TP753

决策树算法是数据挖掘中获取分类规则的主要方法之一。遥感技术的发展,尤其是高光谱遥感数据的出现,使得决策树算法在遥感影像分类中得到了广泛应用。被广泛应用于遥感影像分类的决策树算法主要有 C4.5、CART 以及 C5.0(一种 BoostTree 算法)等[1-6]。但是,目前在大多数关于决策树遥感影像分类的研究中,决策树的生成完全依赖于现有的数据挖掘软件,缺少对决策树算法的深入研究,尤其是算法在遥感影像分类中的具体实现。这样不仅导致在分类的具体实施过程中,需要使用多款软件,进行多次数据格式转换,而且也不利于对决策树算法进行改进,使其更加适用于遥感影像分类。

为了解决上述问题,本文以 BoostTree 算法为模型,结合遥感影像分类的特点,使用 AdaBoost.M1 算法作为推进技术,采用重新构建的 C4.5* 决策树算法作为弱分类器,修改了 AdaBoost.M1 算法中的最终预测函数,构建了复合决策树算法 AdaTree,并以此算法作为分类器设计实现了决策树遥感影像分类系统。本文利用该系统分别对 Landsat7 ETM+(简称为 ETM+)和 WordView2(简称为 WV2)影像进行了基于像元和面向对象的分类实验,并与 BoostTree 和 SVM 分类算法进行了比较。

1　AdaTree 分类器构建

1.1　决策树生成算法

在决策树的构建过程中,往往采用自顶向下的递归方式,在树的内部结点进行属性值(预测变量)的比较,并根据不同的属性值判断从该结点向下的分支直至叶节点的形成。为了避免决策树过于复杂和庞大,同时防止过度拟合现象的发生,需要在生成决策树的过程中或者是结束后对决策树进行剪枝[7]。概括来讲,决策树生成算法的构建主要需要考虑 4 个方面:决策树的结构(二叉树或者多叉树);决策树属性选择度量,即分裂准则;决策树生长停止条件,即叶节点形成的条件;剪枝方法。

相对于其他应用领域,在遥感影像分类中,用于生成决策树的样本(训练元组)数量比较少,均为连续属性,但后期影像分类中需要预测的数据

收稿日期:2013-10-10。
项目来源:国家科技支撑计划资助项目(2012BAH28B01);地理空间信息工程国家测绘地理信息局重点实验室基金资助项目(777121801)。

量大，所以在决策树生成算法的构建中可以忽略时间复杂度的影响，而更加关注分类的精度与效率。和多叉树相比，二叉树生成效率低，但是预测精度高[7]，且转换后的规则描述简单，适合大规模实例元组的预测。同时，在不考虑时间复杂度的前提下，采用二叉树结构可以不进行连续属性的离散化而有利于提高分类精度，因此，本文以常被用作弱分类器的 C4.5 算法为基础，修改树的结构为二叉树，采用信息增益比率 GainRatio[8] 度量属性选择，利用错误剪枝算法 EBP[9] 进行剪枝，重新构建了 C4.5* 算法。算法具体描述如下。

输入：

数据划分（训练样本）D，训练元组和对应类标号的集合；

Attribute_list，候选属性的集合；

Attribute_selection_method，即分裂准则，这个准则在本文中由具有最高信息增益比率 GainRatio 的分裂属性和相应的分裂点组成。

输出：一棵决策树

方法：

1）创建一个节点 N；

2）如果 D 中的元组都是同一类 C，返回 N 作为叶节点，以类 C 标记；

3）如果 D 中的元组所有属性值都相同但所属类别不同（异物同谱）或者树的高度达到用户定义的高度，返回 N 作为叶节点，标记为 D 中的多数类（多数表决）；

4）计算每个属性的候选阈值，即每相邻两个值的中间值；

5）计算 D 中所有属性中每个候选阈值二路划分下的信息增益比率，找出增益比率最高的属性和对应的阈值；

6）二路划分 D，并对每个划分产生子树 D_L、D_R；

7）分别对子树 D_L 和 D_R 重复上述操作，直至全部到达叶节点；

8）执行错误剪枝 EBP 算法。

和原 C4.5 算法相比，C4.5* 算法用二叉树结构替换了多叉树结构，去掉了连续属性的离散化处理，制定了新的决策树生长停止条件，使得算法更加适用于遥感影像分类。

1.2　AdaBoost 算法

AdaBoost 算法[10] 的基本思想为：在弱分类器的迭代过程中，根据每次训练的预测结果，不断调整训练样本权重，对于预测失败的样本赋予较大的权重，从而使下一次的迭代更加关注那些出错的样本，最后得到一个预测函数系列，并根据每次训练的预测结果为每个预测函数赋予权重。在应用这些预测函数进行分类时，采用加权投票的方式决定被预测实例的类别归属。经过不断完善，该算法已成为目前最流行的 Boosting 算法，而 BoostTree 是将 Boosting 算法和决策树算法结合得最成功的算法。

在复合决策树的构建中，本文以 BoostTree 为模板，采用 AdaBoost.M1 算法作为推进技术，但在与 C4.5* 算法具体结合的时候，修改了其最终预测函数中的权重构成。AdaBoost.M1 算法描述如下。

输入：

m 个样本序列 $\{(x_1, y_1),(x_2, y_2),\cdots,(x_m, y_m)$；$x_i \in X\}$，其中，$X$ 表示某个实例空间，$y_i \in Y=\{1,\cdots,k\}$ 为类别标签；

弱分类学习器 WeakLearn；

迭代次数 T。

初始化：对于任意样本 i 赋予初始权重 $D_1(i)=1/m$

执行 $t=1, 2, \cdots, T$ 循环：

1）如果 $t>1$，根据样本权重 D_t 进行概率重采样，生成新的训练集；

2）调用弱分类学习器 WeakLearn；

3）得出预测函数 $h_i: X \to Y$；

4）计算预测函数 h_i 的错误：

$$\varepsilon_t = \sum_{i:h_t(x_i)\neq y_i} D_t(i) \tag{1}$$

如果 $\varepsilon_t>1/2$，$T=t-1$，结束循环；

5）计算 $\beta_t=\varepsilon_t/(1-\varepsilon_t)$；

6）更新样本权重 D_{t+1}：

$$D_{t+1}(i) = \frac{D_t(i)}{Z_i} \times \begin{cases} \beta_t, h_i(x_i) = y_i \\ 1, \text{其他} \end{cases} \tag{2}$$

其中，Z_i 是一个标准化因子，使得所有样本权重之和为 1；

输出最终预测函数：

$$h_{\text{fin}}(x) = \arg\max_{y\in Y} \sum_{t:h_t(x)=y} \lg\frac{1}{\beta_t} \tag{3}$$

在执行循环步骤 1 重采样中，本文采用了轮盘赌概率重采样方法[11]。

1.3　C4.5* 与 AdaBoost.M1 的结合

在传统的 BoostTree 算法中，Boosting 算法和决策树算法的结合是松散、相对独立的，两者结合的最终预测函数（式（3））中的权重由每棵决策树作为弱分类器整体的预测错误率给出，忽略了每个树叶预测精度对最终分类的影响。为了提高最终分类精度，本文修改了最终预测函数，为每个树叶制定了预测权重。同时，为了便于在遥感影像分类中使用生成的决策树，将最终生成的复合决策树转换为规则集 rules。在转换中，从根到每个树叶节点的每条路径创建一个规则。转换后，每条规则的预测权重由其准确率和所在单棵树的权重共同给出。

设 n_{cov} 为训练中规则 r 覆盖的元组数，n_{cor} 为 r

正确分类的元组数，则 r 的准确率定义为：

$$\mathrm{acc}(r) = \frac{n_{\mathrm{cor}}}{n_{\mathrm{cov}}} \tag{4}$$

设 r 的预测权重为 $\mathrm{pre}(r)$，结合式(3)，则有：

$$\mathrm{pre}(r) = \mathrm{acc}(r) \times \lg \frac{1}{\beta_t} \tag{5}$$

则 AdaTree 最终输出的预测函数为：

$$h_{\mathrm{fin}}(x) = \arg\max_{y \in Y} \sum_{r:h_r(x)=y} \mathrm{pre}(r) \tag{6}$$

当利用 rules 对某个元组(实例)进行分类时，共有 T 条规则命中该元组(每棵树转换成的规则集中会有一条规则命中该元组)，根据每条规则的预测权重统计每种类别的加权得票数，以得票数最高的类别作为最终的分类结果。

2 决策树遥感影像分类系统

2.1 需求分析

根据决策树分类原理，利用决策树算法进行分类一般需要经过样本采集、生成训练集、生成规则集和执行分类4个步骤。目前，常用的遥感影像分类方法分为基于像元和面向对象两种。其中，基于像元的分类方法适用于低分辨率影像，而面向对象的分类方法适用于高分辨率影像[10,12-13]。

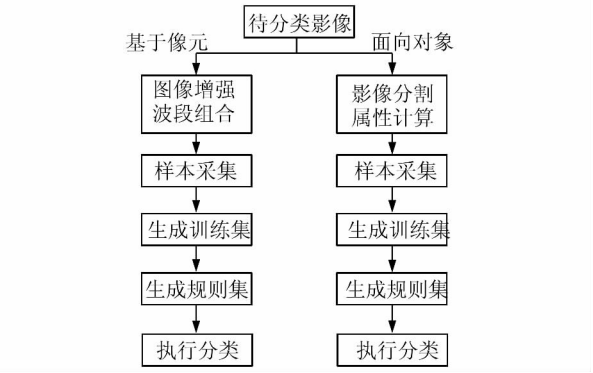

图1 两种分类方法的流程图

Fig. 1 Processes of Cells-Based and Object-Oriented Classification

如图1所示，对于分类器而言，两种方法的区别在于操作对象的不同，其工作原理是一样的。在利用决策树分类器实现基于像元的分类方法时，除生成规则集外，需要提供以像元为分类对象的操作功能，包括利用样本数据和波段组合生成标准训练集，以及根据每个像元对应的波段值利用生成的规则集对其进行分类。此外，还需要提供图像增强等辅助分类功能，如TC变换、NDVI计算、纹理特征计算等。对于面向对象的分类方法，则需要提供以分割对象(矢量多边形)为分类对象的操作功能，包括从分割对象中选取样本生成训练集，以及根据每个分割对象的属性值利用规则集对其进行分类。此外，还需要提供影像分割以及属性计算等功能。

2.2 设计与实现

本文以C#为开发语言，Visual Studio 2005和ArcEngine 9.3为开发平台，根据§2.1节中的需求分析实现了GLC_Info决策树遥感影像分类系统，该系统以AdaTree分类器为核心，提供了基于像元和面向对象两种遥感影像分类方法，同时提供了一些辅助分类功能：① 样本采集，提供了点、线、面三种样本采集方式。② 图像增强，提供了TC变换、NDVI计算和纹理特征计算等功能。③ AdaTree决策树分类，该模块为系统核心模块，包括生成训练集、生成规则集、和执行分类三部分。其中，生成训练集对于基于像元分类，用于输入波段组合(原始影像、图像增强结果)和样本数据，导出标准训练集；对于面向对象分类，用于输入作为样本的分割对象，输出标准训练集。生成规则集主要用于决策树的建立和规则集的生成，同时加载了十折交叉验证功能用于检验训练样本，并且可以通过设定阈值进行多次训练，自动选取高精度的分类规则。执行分类对于基于像元分类，用于输入规则集和波段组合，得到分类影像；对于面向对象分类，用于输入规则集和分割结果，对分割结果进行类别标识。④ 实现了精度评价、分类后修改、单类提取、矢栅互转等辅助功能。

此外，由于该系统影像分割功能尚在完善中，在实施面向对象分类时，需通过eCoginition或者ENVI EX获得分割结果。

3 实验与分析

为了检验本文中AdaTree算法和所研发分类系统的有效性，分别采用经过校正的10幅ETM+(分辨率30 m)和5幅WV2(分辨率为多光谱2 m，全色0.5 m)影像进行分类实验。实验根据分辨率的不同分别对ETM+和WV2影像采用了基于像元和面向对象两种分类方法。

如表1所示，共进行了4种实验，实验1、2采用基于像元的分类方法对10幅ETM+影像进行了分类，波段组合采用了TC变换、NDVI指数、一阶纹理变换加原始1～5、7波段，共14个预测变量，解译标志为人造覆盖、裸地、耕地、林地、草地、灌木、水体和湿地，分类器分别为Boost-

Tree(C4.5＋AdaBoost.M1)、AdaTree。实验 3、4 采用面向对象的分类方法对 5 幅 WV2 影像进行分类，通过 eCoginition(分割尺度 80，属性为 Customized、Layer Values、Geometry 三类 43 种)获得分割结果，解译标志为耕地、园地、草地、房屋、道路、裸地、水体，分类器分别为 SVM 和 AdaTree 算法。上述各实验中训练样本数量为 200～300，检验样本数量为 250 左右。

表 1　实验结果

Tab.1　Results of Experiment

序号	实验数据	分类方法	分类器	Kappa
1	ETM＋	基于像元	BoostTree	0.879 2
2	ETM＋	基于像元	AdaTree	0.905 2
3	WV2	面向对象	SVM	0.897 9
4	WV2	面向对象	AdaTree	0.939 8

实验 1、2 表明，在 ETM＋基于像元的分类实验中，AdaTree 算法的分类精度要高于原有的 BoostTree 算法。实验 3、4 表明，在 WV2 面向对象的分类中，AdaTree 算法的分类精度要远高于 SVM。此外，AdaTree 算法作为分类器不仅适用于基于像元的遥感影像分类，同时也适用于面向对象的遥感影像分类。实验中部分分类结果如图 2、3 所示。

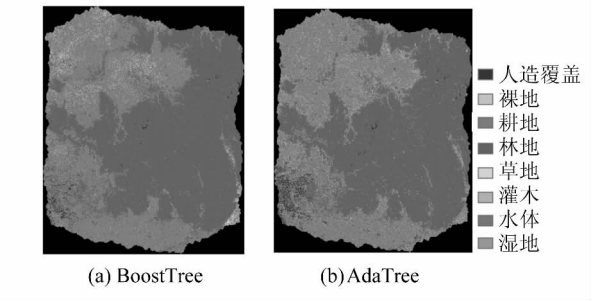

图 2　ETM＋分类结果

Fig.2　Results of Classification with ETM＋

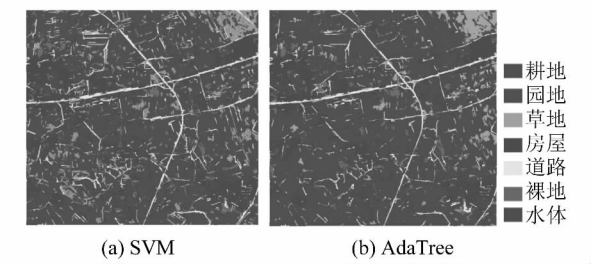

图 3　WV2 分类结果

Fig.3　Results of Classification with WV2

4　结　语

本文以 BoostTree 算法为基础，通过算法改进构建了 AdaTree 算法，然后，以该算法为基础研发了决策树遥感影像分类系统。该系统不仅实现了基于像元的遥感影像分类，并且可以在获得影像分割的基础上，实现对分割结果的自动分类。同时，该系统具有图像增强、精度评价以及分类后修改等一系列辅助功能。实验结果表明，本文所构建的 AdaTree 算法作为分类器在分类精度上要优于 BoostTree 和 SVM 分类算法，并且同时适用于基于像元和面向对象两种遥感影像分类方法。

参　考　文　献

[1] Joy S M, Reich R M, Reynokls R T. A Non-parametric Supervised Classification of Vegetation Types on the Kaibab National Forest Using Decision Trees [J]. Int J Remote Sensing, 2003, 24(9): 1 835-1 852

[2] Franklin S E, Stenhouse G B, Hansen M J, et al. An Integrated Decision Tree Approach (IDTA) to Mapping Landcover Using Satellite Remote Sensing in Support of Grizzly Bear Habitat Analysis in the Alberta Yellow Head Ecosystem [J]. Canadian Journal of Remote Sensing, 2001, 27: 579-592

[3] Marc S, Sasan S, De Gianfranco G. Use of Decision Tree and Multiscale Texture for Classification of JERS-1 SAR Data over Tropical Forest[J]. IEEE Transactions on Geoscience and Remote Sensing, 2000, 38: 2 310-2 321

[4] Zhai Liang, Xie Wenhan, Sang Huiyong, et al. Land Cover Mapping with Landsat Data: The Tasmania Case Study[C]. International Symposium on Image and Data Fusion, Tengchong, Yunnan, China, 2011

[5] Bittencourt H R, Clark R T. Use of Classification and Regression Trees (CART) to Classify Remotely-Sensed Digital in Ages[J]. IEEE International Geoscience and Remote Sensing, 2003,7(6): 3 751-3 753

[6] 温兴平，刘雪华，杨晓峰. 基于 C5.0 决策树算法的 ETM＋影像信息提取[J]. 地理与地理信息科学，2007，23(6)：26-29

[7] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2007

[9] Esposito F, Malerba D. A Comparative Analysis of Methods for Pruning Decision Trees [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1997, 19(5): 476-491

[8] 高潮，刘志雄. 基于轮盘赌编码和粒子群算法的并行机调度优化[J]. 机械制造，2006，48(550)：21-

29

[10] 龚建雅，姚璜，沈欣. 利用AdaBoost算法进行高分辨率影像的面向对象分类[J]. 武汉大学学报·信息科学版，2010，35(12)：1 440-1 443

[11] 高潮，刘志雄. 基于轮盘赌编码和粒子群算法的并行机调度优化[J]. 机械制造，2006，48(550)：21-29

[12] 陈云浩，冯通，史培军，等. 基于面向对象和规则的遥感影像分类研究[J]. 武汉大学学报·信息科学版，2006，31(4)：306-319

[13] 王卫红，夏列刚，骆剑承，等. 面向对象的遥感影像多层次迭代分类方法研究[J]. 武汉大学学报·信息科学版，2011，36(10)：1 154-1 158

第一作者简介：张晓贺，硕士生，主要从事遥感技术和图像处理研究。

E-mail：nwu_zxh@qq.com

Application of AdaTree Algorithm to Remote Sensing Image Classification

ZHANG Xiaohe[1,2] *ZHAI Liang*[2] *ZHANG Jixian*[2] *YANG Xiangbing*[3]

(1 Unit 68332, Weinan 715200, China)

(2 Chinese Academy of Surveying and Mapping, 28 West Lianhuachi Road, Beijing 100830, China)

(3 Chengde Land Resources Bureau, Donghuan Road, Chengde 067000, China)

Abstract: As one of main classification methods used in data mining, the decision tree algorithm is widely used in remote sensing image classification. However, in current studies of remote sensing image classification, the building of decision trees was found to be dependent on existing data mining software, with little research work focused on decision tree algorithms. Based on the BoostTree algorithm, we propose a new algorithm of decision tree ensembles for remote sensing image classification-AdaTree which is a combination of C4.5 and AdaBoost.M1 algorithms. In AdaTree, the structure of C4.5 and the final hypothesis of AdaBoost.M1 were modified. With the AdaTree classifier algorithm, a piece of software was developed for cell-based and object-oriented remote sensing image classification. An experiment with Landsat7 ETM+ and Wordview2 images showed accuracy and efficient improvements of the AdaTree classifier when compared with BoostTree and SVM, either in cells-based or object-oriented classification. Its average Kappa coefficients reached 0.905 2 and 0.939 8.

Key words: decision tree; AdaBoost; classification; remote sensing

About the first author: ZHANG Xiaohe, postgraduate, majors in remote sensing technology and image processing.

E-mail: nwu_zxh@qq.com