文章编号: 1007-4619(2005)04-0405-08

基于 MODIS数据的决策树分类方法研究与应用

刘勇洪, 牛 铮, 王长耀
(中国科学院 遥感应用研究所, 遥感科学国家重点实验室, 北京 100101)

摘 要: 介绍了目前国际上流行的两种决策树算法——CART算法与 C4.5算法, 并引入了两种机器学习领域里的分类新技术——boosting和 bagging技术, 为探究这些决策树分类算法与新技术在遥感影像分类方面的潜力, 以中国华北地区 MODIS250m分辨率影像进行了土地覆盖决策树分类试验与分析。研究结果表明决策树在满足充分训练样本的条件下, 相对于传统方法如最大似然法(MLC)能明显提高分类精度, 而在样本量不足下决策树分类表现差于 MLC; 并发现在单一决策树生成中, 分类回归树 CART算法表现较 C4.5算法具有分类精度和树结构优势, 分类精度的提高取决于树结构的合理构建与剪枝处理; 另外在决策树 CART中引入 boosting技术, 能明显提高那些较难识别类别的分类准确率 18.5%到 25.6%。

关键词: 决策树; CART算法; C4.5算法; boosting和 bagging技术; 土地覆盖 MODIS250m
中图分类号: TN911.73 文献标识码: A

1 引 言

土地覆盖及其变化是全球环境变化过程中的重要因子, 而土地覆盖 植被类型的分布对于地球生态系统过程的物质和能量交换中起着非常重要的作用和地位, 也是全球变化和碳循环模拟、气候模拟等研究的重要内容。传统的土地覆盖 植被制图主要采用地面调查和测量的手段, 具有工作量大、更新周期长等缺点。遥感技术的发展, 特别是全球 1km NOAA/AVHRR数据集产品的提供, 为大范围的土地覆盖和森林制图提供了一条新途径, MODIS 250m分辨率全球数据的提供则兴起新一轮全球环境变化遥感研究的高潮。

近年来, 在全球及区域土地覆盖 植被覆盖遥感制图方法上, 决策树作为一种新兴的分类方法已得到成功应用。Hanson等人利用 NOAA/AVHRR全球 1°×1°数据进行了决策树与最大似然法的土地覆盖分类[1], 显示分类树法的精度优于最大似然法, 马里兰大学全球 8km的土地覆盖产品也采用了二元决策树分类算法进行监督分类[2], 目前分发的 MODIS 1km全球土地覆盖产品也把决策树作为一种主要分类方法[3], Muchoney等人利用 MODIS数据对美国中部进行土地覆盖分类[4], 比较了决策树、神经网络、最大似然法 3种分类方法效果, 结果显示决策树分类精度最高, 此外在小区域范围内, Joy等人利用 TM影像采用决策树对森林类型识别也取得较好的效果[5]。在国内, 决策树也开始得到应用, 王建等人利用地物的光谱统计特性结合纹理、形状等建立分层决策树有效提取荒漠化土地类型[6]; 张丰等人根据水稻的高光谱特性建立混合决策分类树[7], 达到总体分类精度 94.9%效果; 赵萍建立了基于光谱特征和形状特征的简单决策树来自动提取居民地信息[8]; 李爽则对 3种不同的决策树算法结构及理论进行了阐述[9]。

决策树作为一种监督分类方法, 由于它的非参数和树结构特性, 在处理遥感影像由于云覆盖和星下校正反射率 NBAR (Nadir BRDF-adjusted reflectance)数据不全造成的损失问题上具有良好的稳健性和鲁棒性, 并克服了最大似然法对数据分布要求的局限。同时, 决策树相对于另一种流行的分类方法——人工神经网络法具有以下几个优势: (1)分类树不含隐含层, 从而避免了神经网络方法的内在模糊性。(2)计算时间明显少于神经网络。

收稿日期: 2004-04-06; 修订日期: 2004-05-08
基金项目: 中国科学院知识创新工程重大项(KZCX1-SW-01)和国家高技术研究发展计划(863计划 2003AA131170)资助。
作者简介: 刘勇洪(1974—), 男, 在读硕士研究生, 1996年获南京气象学院农业气象专业学士学位, 主要从事遥感图像分类、土地覆盖等方面的研究。Email: liuyh7414@163.com.

(3)树的分割层次关系有利于数据结构的解释,有助于消除输入数据冗余和噪声,并能用于分类特征提取,例如 Borak 等人运用决策树从大量数据中进行分类特征选择[10],取得较好效果。

国内建立决策树的方法主要基于光谱统计特性生成的阈值以及相关先验知识,在实际工作中由于时间、地点变化较大而难以操作,结果往往与研究者的经验和专业知识密切相关。本文介绍了近年来进行遥感分类制图的常见两种决策树方法——CART算法与 C4.5算法,同时引进了两种应用到决策树中提高分类精度的新技术——boosting和 bagging技术,并尝试采用上述方法和新技术进行了区域尺度土地覆盖的遥感分类试验,目的在于探讨决策树分类器在遥感数据应用方面的技术问题,并通过与最大似然法的比较分析,挖掘决策树分类器在遥感应用方面的相对优势、局限性及其应用潜力。

2 决策树算法

决策树方法是多元统计分类中的一种方法。它利用树结构原则,按一定的分割原则把数据分为特征更为均质的子集,这些子集在数据结构中称为节点,其基本思想是利用一组自变量来预测每个样本最可能对应的类型即因变量。一个决策树包括一个根节点(Root node——输入变量),一系列内部节点(Internal nodes——分支)及终极节点(Terminal nodes——叶)。每个内部节点有一个父节点和两个或以上子节点,代表一个数据子集,每个终极节点代表树的预测结果即标识为不同的类别。这里主要介绍了一种在遥感应用中广泛使用的算法——分类回归树(CART),此外简单介绍了另一种决策树算法——C4.5算法。

2.1 分类回归树(CART)基本原理

分类回归树 CART(Classification and Regression Tree)为一种通用的树生长算法,由 Breiman等人提出[11],是一种监督分类方法,它利用训练样本来构造二叉树并进行决策分类。其特点是充分利用二叉树的结构(Binary Tree structured),即根节点包含所有样本,在一定的分割规则下根节点被分割为两个子节点,这个过程又在子节点上重复进行,成为一个回归过程,直至不可再分成为叶节点为止。构造CART树采用的思路:在整体样本数据的基础上,生成一个层次多、叶节点多的大树,以充分反映数据之间的联系(这时树生长未考虑噪声,往往反映的是训练过度情况下的数据联系),然后对其进行删减,产生一系列子树,从中选择适当大小的树,用于对数据进行分类,具体来讲,分为树生长和树剪枝两部分:

2.1.1 树生长

树节点处的一次判别称为一个分支,它对应于将训练样本划分成子集,根节点处的分支对应于全部训练样本,其后每一次判决都是一次训练子集划分过程,因此构造树的过程实际上是一个属性查询产生分割规则的过程。在本文中,CART采用了一种称为"节点不纯度[12]"的指标:用 $i(N)$表示节点 N的"不纯度",当节点上的模式数据均来自同一类别时,$i(N)=0$;而若数据所属类别均匀分布时,$i(N)$应当很大,分割规则即是基于不纯度函数的极小值而产生的。这里介绍两种当前流行的"不纯度"测量函数:

(1)"熵不纯度"(entropy impurity),亦称为信息量不纯度(information impurity):

$$i(N)=-\sum_j P(\omega_j)\log_2 P(\omega_j) \tag{1}$$

其中 $P(\omega_j)$是节点 N处属于 ω_j类模式样本数占总样本数的概率。根据众所周知的熵的特性,如果所有模式的样本都来自同一类别,则不纯度为零,否则是大于零的正值,当所有类别以等概率出现时,熵取最大值。

(2)方差不纯度——"Gini不纯度",根据多分类问题中节点样本来自不同类别与总体分布方差有关而提出:

$$i(N)=-\sum_{i\neq j}P(\omega_i)P(\omega_j)=1-\sum_j P^2(\omega_j) \tag{2}$$

"Gini不纯度"的意义即为当节点 N的类别标识任意选取时对应的误差率。

当给定一部分树,目前已生长到节点 N,要求对该节点作属性查询,一个明显的启发式的思路是选择那个能够使不纯度下降最快的那个查询,不纯度下降差可记为:

$$\Delta i(N)=i(N)-P_L\,i(N_L)-(1-P_L)\,i(N_R) \tag{3}$$

其中 N_L和 N_R分别是左、右子节点,$i(N_L)$,$i(N_R)$是相应的不纯度。P_L是当查询 T被采纳时,树由 N生长到 N_L的概率,这样最佳的查询值 S就是那个能最大化 $\Delta i(T)$的值。

2.1.2 树剪枝

如果我们持续生长树,直到所有的叶节点都达

到最小的不纯度为止, 数据一般将被 "过拟合", 那么分类树就退化为一个方便的查找表, 这样, 对有较大贝叶斯误差的噪声信号的推广性能就不可能很好, 相反, 如果分支停止的太早, 那么对训练样本的误差就不够小, 导致分类性能很差, 一种主要的停止分支方法就是剪枝 (*pruning*), 同时也是为了避免树生长的过分庞大。本文通过最小化如下这个定义的全局指标来达到目的:

$$cost = \alpha \cdot size + \sum_{叶节点} i(n) \qquad (4)$$

其中 *cost*可以理解为该树加权错分率与对复杂度处罚值之和的代价函数, *size*表示叶节点数量, 可以用于衡量这个树分类器的复杂度, α为一复杂度参数, $\sum_{叶节点} i(n)$表示为所有叶节点的不纯度的求和, 表征了使用该分类树对训练样本进行分类时的不确定性。

根据公式 (4), 树剪枝可由下面两步组成:

(1) 在所有互为兄弟的叶节点中, 比较重新合并叶节点后 *cost*值的变化。

(2) 删除使 *cost*值最大减少的叶节点, 若 *cost*值不减少, 则不做变化。

重复上述修剪过程, 直到修剪不能再进行。

在剪枝过程中, 训练误差随叶节点的数目增加而减少; 测试误差则最初减小, 达到最小值, 然后由于训练数据对树的过分影响, 测试误差又逐渐增加, 利用独立的测试数据集进行测试, 则选择具有最小测试误差的子树作为最优决策树。本文选择一种启发式验证技术——交叉验证技术 (cross validation)[3] 来进行最优树的选择: 10重交叉验证 (10-fold cross validation), 即训练集被随机划分为 10组相等数量不相交的数据子集, 分类器要训练 10次, 每次采用 9组数据子集进行训练, 余下的 1组子集用作验证集 (validation set), 用于评价测试误差, 估计出的测试误差是 10组误差的平均。

2.2　C4.5基本原理

C4.5为另一种广泛使用的单一决策树生成法, 采用 "信息获取率 (information gain ratio)"矩阵来实现分类。它利用训练集, 对每次选取信息获取率 (gain ratio)最大的但同时获取的信息增益又不低于所有属性平均值的属性, 作为树的结点, 将每一个可能的取值作为此节点的一个分支, 递归地形成决策树。树生成算法中也采用了 CART中的 "熵不纯度" 函数, 而信息增益相当于 CART中的 "不纯度下降差", 另外增加了 "获取率"这一指标, 主要是为了去除高分支属性的影响而对信息增益的一种改进。 "获取率"同时考虑了每一次划分所产生的子结点的个数和每个子结点的大小 (包含的数据实例的个数), 考虑的对象主要是一个个地划分, 而不再考虑分类所蕴涵的信息量。递归的结束条件是子集中的数据记录在主属性上取值都相同, 或没有属性可再供划分使用。

与 CART不同的是, C4.5利用了基于分支的统计显著性的误差概率技术来实现剪枝, 另一个显著差别是体现在对缺损模式的处理上, 在训练阶段, C4.5并没有象 CART以替代分支 (surrogate split)来解决分类数据的缺损, 而是以概率加权的方法来处理 "属性丢失"的问题[12]。

2.3　决策树中采用的新技术——boosting与 bagging方法

在决策树分类器设计中, 一种于 20世纪 90年代中后期在机器学习领域发展的被称之为 "boosting (增强法)[13,14]"的技术被广泛采用来提高分类精度。这种方法可以提高那些较难识别样本的分类准确率, 同时这种技术能降低分类算法对数据噪声和训练样本误差的敏感性。

本文采用了一种基于 AdaBoot (adaptive boosting——自适应增强)[15] 方法的 boosting技术, 它实际上是用训练样本来设计分类器的一种重采样技术。在 AdaBoot方法中, 每一个训练样本都被赋予一个权重, 表明它被某个分量分类器选入训练集的概率。如果每个样本点已经被准确地分类, 那么在构造下一个训练集中, 它被选中的概率就降低; 相反, 如果某个样本点没有被正确分类, 那么它的权重就得到提高。通过这样的方式, AdaBoot方法就能够 "聚焦于"那些较困难的样本上, 如此权重更新过的样本集被递归使用来训练下一个分类器, 直至分类误差小于某个阈值。

此外, 本文还采用了另一种分类器设计中的重采样技术——bagging算法[12] 来提高分类精度: 此名来自于 bootstrap aggregation (自助聚集), 它表示如下过程: 从大小为 n的原始数据集 D中, 分别独立地抽取 n'个数据 ($n' < n$)形成自助数据集, 并且将这个过程独立地进行许多次, 直到产生很多个独立的自助数据集。然后, 每一个自助数据集都被独立地用于训练一个 "分量分类器" (component classifier)。一般 bagging算法能提高 "不稳定"分类器的识别率, 因为它相当于对不连续处进行了平均

化处理。

这两种算法均产生多个“分量分类器”而不是最佳的单个分类器,没有剪枝过程发生,最终的分类结果将根据这些“分量分类器”各自的判决结果的投票来决定。

Boosting技术已成功地应用到MODIS 1km土地覆盖制图产品中[3], bagging技术的应用还未见报道,本文则对这两种技术与 CART决策树相结合进行了最新尝试。

3 应用试验

遥感影像土地覆盖分类是指通过对遥感影像上各种地物的光谱信息的分析,将像元划分为不同类型的土地覆盖单位,因此地物的光谱特性是土地覆盖分类的主要判别依据。为测试中分辨率影像MODIS 250m分辨率数据对区域尺度土地覆盖分类的性能,本文选取中国华北地区(包括北京、天津、河北、山东、河南、山西四省二市)作为宏观土地覆盖分类研究区域。研究区域的土地覆盖类型采用中国植被编码体系,主要基于 1∶400万《中国植被图》(1979,中国地图出版社,中国科学院与环境信息系统国家重点试验室数字化)所采用的中国植被编码体系并结合遥感数据的特点而设计的二级土地覆盖分类体系[16],依据本区实际情况,把华北主要土地覆盖类型分成了 8类:落叶阔叶林(记为阔叶林)、常绿针叶林(记为针叶林)、灌木矮林、草地、农田、沼泽草甸、水体、建筑居民地。

在遥感影像分类中,一个好的监督分类过程应考虑以下 3个方面:(1)代表性较好的训练样本;(2)较佳的分类特征;(3)设计良好的分类器。

3.1 样本选取

训练数据的质量在很大程度上影响着制图精度[9]。由于缺乏相应区域足够的地面实际样本,为最大限度的保证选取样本的代表性,以华北地区 1∶400万比例尺植被类型图为基础,参考 1990—1995年的地面森林抽样清查数据,并结合 1∶100万土地利用数据库(1995,基于 TM影像解译得到的全国 1∶100万比例尺的中国资源环境遥感数据库),所形成的矢量数据以统一的地理坐标方式投影到遥感影像图上,结合遥感影像目视解译选出各类有代表性的样点 4605个,把总样本按比例 7∶3分成 2部分,一部分为训练样本 3226个,另一部分验证样本 1379个。

3.2 分类特征选择及数据获取、处理

为获取较优的分类特征,本文从土地覆盖类型的光谱特征空间、物候特征空间、地形特征空间 3方面进行选择:

首先选取了美国 LP DAAC(Land Process Distributed Active Archive Center)的 MODIS数据产品:2001年 4月 7日—14日 MODIS 8天合成的波段反射率产品 MOD09Q1(1—2波段 250m分辨率)和 MOD09A1(3—7波段 500m分辨率)图像,代表各类别的反射率波谱特征;

其次,根据本区各类别的物候特性,选取了 2001年中代表四季的 4个时相 16天最大值合成的 MODIS250m分辨率植被指数产品 MOD13A1,分别是 1月 1日—16日,4月 7日—22日,7月 12—27日,9月 30日—10月 15日,本产品包括两种植被指数:归一化植被指数 *NDVI*和增强性植被指数 *EVI*[17]。由于 *EVI*对植被类型季节性变化较 *NDVI*更为敏感[18],而且克服了一些 *NDVI*的不足[19],因此本文采用多时相 *EVI*植被指数作为波谱物候特征。

考虑到本区各类别在地形上的差异,地形特征选用了中国 1km格网的 DEM(地面数字高程模型)影像。

对以上特征影像分别进行地理几何校正与重采样,采样方法为邻近法,投影体系为双标准纬线等积圆锥投影(ALBERS),椭球为 Krosovsky体系,分辨率统一到 250m,最终影像大小为 4725列×5543行。

由于原始数据数据类型为 16位,占用计算机资源较大。为有效压缩数据,充分利用 7个反射率波段的波谱信息,对 7个波段进行了主成分变换,取前 3个主成分波段(代表原始 7个波段方差的 94.2%)参加分类运算,然后与其它 4波段 *EVI*数据、1波段 *DEM*数据进行了 0—255之间数据类型到 8位的归一化,转换公式为:

$$DN_{\text{变换后}}=\frac{DN_{\text{变换前}}-DN_{\min}}{DN_{\max}-DN_{\min}}\times 255$$

其中,$DN_{\text{变换前}}$,$DN_{\min}$,$DN_{\max}$分别为每波段的实际值、最小值、最大值。

3.3 试验方法与结果分析

3.3.1 样本数敏感性分析

为测试决策树分类器对样本数据大小的敏感分析,我们从训练样本总数 3226中随机抽取 3226,2800,2400,2000,1600,1200,800,400共 8组训练样本,每组类别样本数见表 1,对以上 8组训练样本分

别采用 3种不同的决策树算法(分别是 CART-GINI不纯度、CART-Entropy不纯度、C4.5算法)来训练单一决策数分类器的生成,然后统一用验证样本数1379进行总体分类精度的评价,并将分类结果与传统分类方法——最大似然法(MLC)进行相互比较,结果见图 1。

表 1　各类别下不同大小的训练样本与验证样本数

Table 1　The number of samples of each class for various training samples groups

类别	训练样本数								验证样本数
阔叶林	335	287	253	204	168	128	83	39	143
针叶林	349	298	259	222	172	126	86	48	149
灌丛矮林	573	498	425	355	285	205	133	75	246
草地	505	431	370	317	246	190	135	75	216
农田	652	589	485	405	333	246	158	68	280
水体	228	193	175	125	120	91	75	31	97
建筑用地	254	220	196	158	117	91	51	24	109
沼泽草甸	330	284	237	214	159	123	79	40	139
合计	3226	2800	2400	2000	1600	1200	800	400	1379

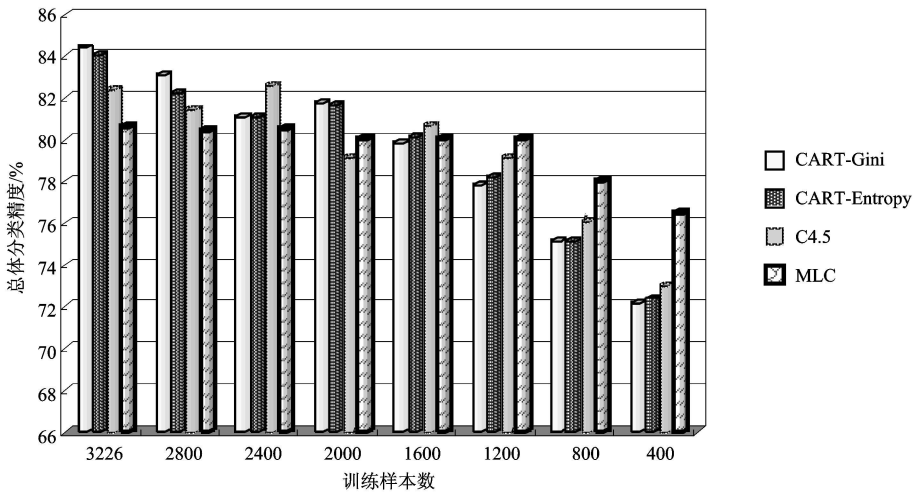

图 1　不同分类方法下的总体分类精度

Fig 1　The overall classification accuracy of different methods

从图 1可知,在所有的 8组训练样本与 4种分类方法的组合中,总体分类精度最高的是所有训练样本数 3226均参加分类下的 CART-Gini方法,较同组训练样本下的最大似然法(MLC)提高了约 4个百分点。对不同训练样本, CART-Entropy分类效果与 CART-Gini差别很小,表明对 CART方法,选择不同形式的不纯度函数对最终分类效果影响很小。但 CART方法随着训练样本数的下降,分类精度也逐步降低,尤其是训练样本数在 800以下,即一半以上的类别样本数不足 100的情形下(表 1),分类精度明显偏低(75%以下)。对 C4.5方法,分类效果则不稳定,两次出现随样本数减少而分类精度增加现象(分别在样本数 2400、1600两组里),表明 C4.5对样本变化较敏感,分类效果不稳定,与 CART方法比较,在大样本量(2800及以上)分类效果差于 CART,在较小样本量(1600及以下),分类效果则好于 CART。而对最大似然法(MLC),只要样本量在 1200及以上(见表 1:大多类样本数在 100以上),分类精度则比较稳定(精度在 80%上下 1个百分点内),与样本大小没有关系,这与文献[3]的结论一致,表明只要能满足最大似然法近似高斯分布条件的样本量,分类精度就相对稳定。

从图 1中还可看出,对小样本量(1200及以下),最大似然法精度都明显高于其它三种决策树

方法, 可见, 决策树对分类精度的提高是基于较大数量的训练样本之上的, 从表 1得知, 决策树要获得较好的分类精度(不低于最大似然法精度), 总训练样本数需在 1600以上, 即平均每类样本数需大于 200.

3.3.2 树结构分析

为测试决策树结构与分类精度的关系, 我们对决策树分类器产生叶子节点数(表征分类器的复杂程度)和分类精度进行了分析, C4.5算法只给出了剪枝后的叶子节点数, 结果见表 2.

表 2 两种决策数算法树结构比较

Table 2 The comparison of structure for the two decision-tree methods

训练样本数		3226	2800	2400	2000	1600	1200	800	400
CART-Gini	剪枝前叶节点数	109	95	103	74	82	85	59	38
	剪枝后叶节点数	81	76	69	69	58	81	55	30
	剪枝前最大相对误差	0.197	0.207	0.230	0.206	0.235	0.261	0.283	0.318
	剪枝后最佳相对误差	0.193	0.200	0.220	0.205	0.228	0.258	0.273	0.315
	分类精度 /%	84.4	83.1	81.1	81.8	79.8	77.8	75.1	72.2
C4.5	剪枝后叶节点数	137	106	115	101	98	84	65	49
	分类精度 /%	82.4	81.5	82.6	79.2	80.7	79.2	76.1	73

从表 2看出, 不同训练样本下产生的决策树大小不同, 训练样本数在不足到充足之间变化时, 树的叶节点数也发生着不同变化: 在 CART-Gini方法中, 训练样本分别为 1200和 3226且剪枝后的叶节点均为 81的情况下, 分类精度却为 77.8%和 84.4%; 在 C4.5方法中, 训练样本分别为 2400与 3226且剪枝后的相应的叶节点数为 137与 115的情况下, 分类精度则非常接近(分别为 82.6%和 82.4%), 可见对决策树而言, 分类精度与树结构的复杂程度(叶节点数)并无明显关系, 只要决策数构建得越合理, 分类精度就越高。

从表 2中还可看出, CART-Gini采用最小代价函数(*cost*)与交叉验证(cross-validation)的方法剪枝效果明显: 在 3种训练样本分别为 3226, 2800, 2400而产生较高分类精度条件下, 剪枝前与剪枝后的叶点数相差十分明显, 叶节点数分别减少了 28, 19, 34, 相对误差也相应减少, 表明在决策树的合理构建中, 剪枝技术也十分重要。另外, 在同样条件下, CART-Gini叶节点数明显少于 C4.5, 而且在大样本量(2000及以上)条件下分类效果也好于对方, 表明 CART-Gini算法在单一决策树分类中相对于 C4.5具有一定优势。

3.3.3 加入 boosting和 bagging技术

在上述结果基础上, 对两种最高分类精度的单一决策数分类器方法 CART-Gini和 CART-Entropy分别加入 boosting和 bagging技术进行复合决策树分类器的产生, 并与最大似然法 MLC进行了分类效果比较, 见表 3.

表 3 CART中加入 boosting和 bagging技术的分类效果比较

Table 3 The comparison of classification accuracy for the addition of boosting and bagging in CART

分类算法	总体分类精度 /%		
	原始	boosting	bagging
CART-Gini	84.4	87.5	83.1
CART-Entropy	84.1	87.2	85.5
MLC	80.6		

从表中可知, 在 CART两种算法中加入 boosting技术能明显提高决策树分类精度, 相对于原始单一决策树分类精度提高了 3个百分点, 而 bagging技术则表现不稳定, 例如 CART-Gini加入 bagging技术分类精度为 83.1%, 不及剪枝处理的单一决策树分类精度 84.4%, 但总的来说, 都高于最大似然法 MLC分类精度 80.6%。

3.3.4 各类别精度分析

我们对最大似然法 MLC、单一决策树 CART-Gini、复合决策树 CART-Gini-boosting的各类别的分类精度(预测结果与实际类别之比)和总体分类精度(表 4)进行了比较。

从表 4得知, 最大似然法 MLC对植被类型阔叶林、针叶林识别率很低, 分类精度均不及 50%, 另外由于水体与沼泽易混分, 其识别率也偏低(仅为 72.2%)。单一决策树 CART-Gini则由于改善了这 3类的识别而使总体分类精度得到增加, 加入 boosting技术的复合决策树由于“聚焦于”那些较难识别的样本上, 相对最大似然法明显提高了阔叶林、

表 4　MLC, CART-Gini, CART-Gini boosting分类结果比较

Table 4　Classification accuracy of MLC, CART-Gini and CART-Gini boosting

分类方法	分类精度 /%								
	阔叶林	针叶林	灌丛矮林	草地	农田	水体	建筑用地	沼泽草甸	总体
MLC	46.2	47.7	79.3	99.5	91.8	72.2	99.1	93.5	80.6
CART-Gini	56.6	70.5	72.4	94.0	98.6	85.6	97.2	95.0	84.4
CART-Gini boosting	64.7	66.2	80.6	97.7	99.3	97.8	97.3	95.7	87.5

针叶林以及水体类别的识别率, 分类精度提高了 18.5%到 25.6%, 总体精度则较 MLC提高了近 7个百分点。可见 boosting技术对提高分类精度尤其是提高那些在最大似然法中难以识别的类别精度具有明显的作用。

4　结　论

在介绍近年来遥感应用领域流行的两种决策树基本原理及两种分类新技术的基础上, 我们对中国华北地区进行了基于 MODIS数据的决策树土地覆盖分类试验, 其中两种新技术 boosting和 bagging与 CART方法的结合尚未见报道。从试验中我们可得出: (1)决策树在满足充分的训练样本条件下, 相对传统方法如最大似然法 (MLC)能明显提高分类精度。(2)在大的训练样本即平均每类在 200以上, 决策树分类法表现优于最大似然法 MLC, 而在小样本量(每类 50—150)下 MLC分类表现优于决策树。(3)在单一决策树生成中, 分类回归树 CART算法较 C4.5算法具有一定优势。(4)分类精度的高低取决于决策树结构的合理性, 而与复杂程度无关, 且剪枝处理对树结构十分重要。(5)在决策树中引入一种机器学习领域内的增强技术 —— boosting技术, 能明显提高那些较难识别类别的分类准确率 18.5%到 25.6%。

上述结论均是针对中分辨率 MODIS250m 数据而言, 选定的区域也仅在华北地区, 而且样本选取均基于矢量数据和影像目视解译, 包含着许多不确定性, 对其它遥感影像数据如高分辨率 TM, SPOT等数据以及应用区域尺度的推广, 还需其它试验进一步检验。

决策树算法用于遥感分类的优势在于对数字影像数据特征空间的分割上, 其分类结构简单明了, 尤其是二叉树结构的单一决策树结构十分容易解释。由于它属于严格“非参”, 对于输入数据空间特征和分类标识, 具有更好的弹性和鲁棒性, 但正如本文所述的, 它的算法基础比较复杂, 而且需要大量的训练样本来探究各类别属性间的复杂关系, 在针对空间数据特征比较简单而且样本量不足的情况下, 其表现并不一定比传统方法如 MLC好, 甚至可能更差。但当遥感影像数据特征的空间分布很复杂, 或者源数据各维具有不同的统计分布和尺度时, 对数据源要求服从某一分布的最大似然法 MLC就显得力不从心, 而用决策树分类法能获得理想的分类结果。

参 考 文 献 (References)

[1] Hanson M C, Dubayah R, DeFries R S. Classification Trees: an Alternative to Traditional Land Cover Classifers[J]. *INT. J. Remote Sensing*, 1996, **17**: 1075—1081.

[2] DeFries R S, Hansen M C, Townsend J G R, *et al*. Global Land Cover Classifications at 8 km Spatial Resolution: The Use of Training Data Derived from Landsat Imagery in Decision Tree Classifiers[J]. *INT. J. Remote Sensing*, 1998, **19**: 3141—3168.

[3] Friedl M A, McIver D K, Hodges J C F, *et al*. Global Land Cover Mapping from MODIS: Algorithms and Early Results[J]. *Remote Sensing of Environment*, 2002, **83**: 287—302.

[4] Muchoney D, Borak J, Borak H C, *et al*. Application of the MODIS Global Supervised Classification to Vegetation and Land Cover Mapping of Central America [J]. *INT. J. Remote Sensing*, 2000, **21**: 1115—1138.

[5] Joy S M, Reich R M, Reynolds 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): 1835—1852.

[6] Wang J, Dong G R, Li W J, *et al*. Primary Study on the Multi-Layer remote Sensing Information Extraction of Desertion of Desertification Land Types by Using Decision Tree Technology [J]. *Journal of Desert Research*, 2000, **20**(3): 243—247. [王建, 董光荣, 李文君等. 利用遥感信息决策树方法分层提取荒漠化土地类型的研究探讨[J]. 中国沙漠, 2000, **20**(3): 243—247.]

[7] Zhang F, Xiong Z, Kou N. Research on Rice Fine Classification using Hyper-spectral Remote Sensing Imagines[J]. *Journal of Wuhan University of Technology*, 2002, **24**(10): 36—39. [张

丰, 熊桢, 寇宁. 高光谱遥感数据用于水稻精细分类研究[J]. 武汉理工大学学报, 2002, 24(10): 36—39.]

[8] Zhao P, Feng X Z, Lin G F. The Decision Tree Algorithm of Automatically Extracting Residential Information from SPOT Images[J]. *Journal of Remote Sensing*, 2003, 7(4): 309—315. [赵萍, 冯学智, 林广发. SPOT卫星影像居民地信息自动提取的决策树方法研究[J]. 遥感学报, 2003, 7(4): 309—315.]

[9] Li Sh, Zhang E X. The Decision Tree Classification and Its Application in Land Cover[J]. *Areal Research and Development*, 2003, 22(1): 17—21. [李爽, 张二勋. 基于决策树遥感影像分类方法[J]. 地域研究与开发, 2003, 22(1): 17—21.]

[10] Borak J S, Strahler A H. Feature Selection and Land Cover Classification of a MODIS-like Data Set for a Semiarid Environment[J]. *INT. J. Remote Sensing*, 1999, 20: 919—938.

[11] Breiman L, Friedman J H, Olshen R A, *et al*. Classification and Regression Tree[M]. Wadsworth Inc. 1984.

[12] Richard O D, Peter E H, David G S. Pattern Classification[M]. Beijing: China Machine Press, CITIC Publishing House, 2003. [Richard O Duda, Peter E Hart, David G Stor著, 李宏东, 姚天翔 等译, 模式分类[M]. 北京, 机械工业出版社, 中信出版社, 2003.]

[13] Friedl M A, McIver D K, Brodley C E. Integration of Domain Knowledge in the Form of Ancillary Map into Supervised Classification of Remotely Sensed Data[J]. *Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS)*, 2002, 2: 1038—1040.

[14] McIver D K, Friedl M A. Using Prior Probabilities in Decision-tree Classification of Remotely Sensed Data[J]. *Remote Sensing of Environment*, 2002, 81: 253—261.

[15] Friedl M A, Brodley C E, Strahler A H, *et al*. Maximizing Land Cover Classification Accuracies Produced by Decision Trees at Continental to Global Scales[J]. *IEEE Trans Geosci Remote Sensing*, 1999, 37: 969—977.

[16] Yan H. Remote Sensing Study of Land Cover Change and Its Environmental Effects in China[D]. PhD Dissertation, Institute of Remote Sensing Application, CAS, 2002. [延昊, 中国土地覆盖变化与环境影响遥感研究[D]. 博士论文, 中国科学院遥感应用研究所, 2002.]

[17] Enhanced Vegetation Index (EVI)[R]. http://tbrs.arizona.edu/project/MODIS/evi.php, 2003.

[18] Fereeira L G, Yoshioka H, Huete A, *et al*. Seasonal Landscape and Spectral Vegetation Index Dynamics in the Brazilian Cerrado, An Annlysis within the Large-Scale Biosphere Atmosphere Experiment in Amzaonia (LBA)[J]. *Remote Sensing of Environment*, 2003, 87: 534—550.

[19] Wang Z X, Liu C, Huete A. From AVHRR-NDVI to MODIS-EVI: Advances in vegetation index research[J]. *Journal of Ecology*, 2003, 23(5): 979—987. [王正兴, 刘闯, HUETE Alfredo. 植被指数研究进展: 从AVHRR-NDVI到MODIS-EVI[J]. 生态学报, 2003, 23(5): 979—987.]

Research and Application of the Decision Tree Classification Using MODIS Data

LIU Yong-hong, NIU Zheng, WANG Chang-yao

(*The State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing Applications, CAS, Beijing 100101, China*)

Abstract: Decision tree classification algorithms have significant potential in remote sensing data classification. In this research, two popular decision tree algorithms—— CART and C4.5 are presented, and two techniques known as boosting and bagging in machine learning area are introduced. We examined these methods to maximize classification accuracies using these decision trees and techniques to map land cover of Huabei area in China from MODIS 250m data. The result indicates that decision tree with abundance training samples has higher classification accuracy than maximum likelihood classifier(MLC) in the land cover classification test, whereas insufficient samples resulted in a lower accuracy for decision tree than MLC. The result also shows CART algorithm has more advantageous than C4.5 algorithm in classification accuracy and tree structure. And the decision tree classification accuracy depends on the optimal structure and pruning process. We also tested the behaviour of boosting and bagging techniques combined with CART and the result shows that adding boosting technique to decision tree can increase classification accuracies by 18.5%—25.6% for the poorly separable classes in MLC.

Key words: decision tree; CART; C4.5; boosting and bagging; land cover; MODIS 250m