文章编号:1671-8860(2007)12-1099-04　　**文献标志码**:A

基于 CART 集成学习的城市不透水层百分比遥感估算

廖明生[1]　江利明[1,2]　林　珲[2]　杨立民[2]

(1　武汉大学测绘遥感信息工程国家重点实验室,武汉市珞喻路 129 号,430079)
(2　香港中文大学太空与地球信息科学研究所,香港新界沙田)

摘　要:利用 Landsat ETM$^+$ 遥感数据,提出了一种基于 CART 集成学习的 ISP 遥感亚像元估算方法,将 Boosting 重采样技术引入 CART 分析中,用于提高 ISP 估算的精度。实验结果表明,该方法的 ISP 估算性能优于传统的单一 CART 学习算法,从 ETM$^+$ 影像中估算的 ISP 值与真实值之间的相关系数达到 0.91,平均偏差为 11.16%。

关键词:城市不透水层;遥感影像;分类与回归树;Boosting 技术;集成学习

中图法分类号:P237.3

不透水层是城市地区的重要特征,被定义为诸如屋顶、沥青、水泥道路以及停车场等具有不透水性的地表面。作为城市环境的关键指数,不透水层百分比(impervious surfaces percent, ISP)广泛应用于城市水文过程模拟、水质面源污染以及城市专题制图等研究中[1,2]。

近年来,基于统计模型和机器学习的 ISP 亚像元遥感估算方法相继被提出,如多元回归法[3]、人工神经网络方法[4]、决策树方法[5,6]、光谱混合模型[7]等。其中,决策树方法通过一系列树型结构的决策规则来建立 ISP 预测模型,由于决策树在连续变量回归问题中具有非线性学习能力,且实现简单,运算效率高,该方法在美国地质调查局(USGS)的地学分析和监测计划(GAM)中得到了成功应用和推广,获取的不透水层数据已加入美国国家土地覆盖数据库[5,6]。然而,决策树是一种弱学习算法[8],受其学习能力的限制,这种 ISP 估算方法对数据噪声和训练样本误差比较敏感,在大量噪声存在的情况下将显著降低预测模型的估算精度;另外,对不均衡样本的欠学习也限制了其精度的进一步提高。研究表明,该方法在低 ISP 样本中(小于 20%)的估算性能并不理想[5,6]。针对上述问题,本文在基于 CART 分析的 ISP 估算方法中,引入目前在机器学习领域广泛采用的 Boosting 技术进行集成学习,以期达到改善 ISP 估算性能的目的。

1　CART 分析与 Boosting 技术

分类与回归树分析(classification and regression tree, CART)是一种通用的决策树构建算法,它可以实例化为各种不同的决策树,当因变量或目标变量为离散的分类类别值时称为分类树,而为连续值时则称为回归树[9]。CART 继承了一般决策树具备的所有优点,既可以用于分类研究,又能够进行连续变量的预测和回归,因此,在遥感应用领域表现出了巨大的优势,目前多用于遥感分类研究,并取得了不错的分类效果[10,11]。

Boosting 技术是一种在机器学习领域发展起来的重采样技术,其目的是通过某种弱学习算法循环产生数个简单的、精度比随机猜测略好的弱规则,再将这些规则进行集成,以提高给定的学习算法的分类精度和预测性能[12]。决策树和神经网络均为弱学习算法,样本训练集的较小波动都将导致它们产生的预测函数发生较大变化。各种

收稿日期:2007-09-28。
项目来源:国家自然科学基金资助项目(60472039,40523005)。

仿真数据的研究表明，在决策树方法中采用 Boosting 技术能够显著提高决策树的学习能力[8]。

AdaBoost(adaptive boosting)算法是当前最流行的一种 Boosting 技术[12]。该算法的主要思想是给定一弱学习算法和一训练样本集 $D=\{(x_1,y_1),\cdots,(x_n,y_n)\}$，这里 x_i 为一向量，y_i 对于分类问题为一类别标志，对于回归问题为一数值。初始化时，对每一个训练样本赋相等的权重 $1/n$，然后用该学习算法对训练集训练 $k_{\max}$ 轮，每次训练后，对训练失败的训练赋以较大的权重，也就是让学习算法能够聚焦于那些较困难的样本上，在后续的学习中集中对它们进行学习，从而得到一个预测函数序列 $(h_1,\cdots,h_k,\cdots,h_{k_{\max}})$，其中 h_k 也有一定的权重，预测效果好的预测函数权重较大，反之较小。最终的预测函数 $H(x)$ 对分类问题采用有权重的投票方式，回归问题则采用加权平均的方法对新数据进行判别，即

$$H(x)=\sum_{k=1}^{k_{\max}}\alpha_k h_k(x) \qquad (1)$$

其中，h_k 为弱学习算法每次迭代的预测函数；α_k 为预测函数权重。

在本文 ISP 估算中，将 CART 作为弱学习算法，尝试采用 AdaBoost 算法进行集成学习，以提高那些突出值样本的学习能力，并降低 CART 算法对数据噪声和训练样本误差的敏感性，最终达到改善 ISP 的整体估算性能的目的。为方便起见，文中 CART 集成学习算法均指这种基于 Boosting 技术的集成学习，而把未经重采样技术的 CART 算法称为单一 CART 算法。

2 实验结果与分析

在基于决策树的 ISP 估算方法中，首先利用高分辨率的航空影像获取 ISP 估算的训练数据和测试数据，通过决策树算法进行学习并建立回归预测模型，在此基础上，利用中分辨率的遥感数据进行大面积的不透水层百分比估算和制图。该方法的主要步骤包括：① 获取 ISP 训练数据和测试数据；② 建立 ISP 预测模型；③ 不透水层制图和精度评估。该方法技术流程如图 1 所示。

2.1 实验数据与预处理

本文选择上海浦东新区作为实验区，实验区位于北纬 30°08′20″至 31°23′22″，东经 121°27′18″至 121°48′43″间。该地区地形平坦，主要地物类型有商业用地、居民地、工业用地、道路、城市绿地、农田、水体和少量滩涂等。实验所用的中分辨率遥感数据为 2001 年 7 月 31 日采集的 Landsat TM/ETM$^+$ 影像，全景图像质量较好，无云和条带影响。另外，选取小块区域的 IKONOS 影像用于获取不透水层百分比估算的训练和测试数据，其获取时间为 2001 年 6 月，该 IKONOS 影像包括 1 m 分辨率的全色波段和 4 m 分辨率的多光谱数据(蓝、绿、红和近红外 4 个波段)。上述 TM 影像和 IKONOS 影像经辐射校正预处理后，以 1∶1万地形图为基准，选择一定数量的控制点，采用二次多项式和最近邻重采样方法对原始影像进行几何纠正和配准(误差控制在 0.5 个像元内)，纠正后的遥感影像具有 WGS84 UTM 投影坐标系，其中 TM 影像的空间分辨率为 30 m，IKONOS 影像为 1 m。

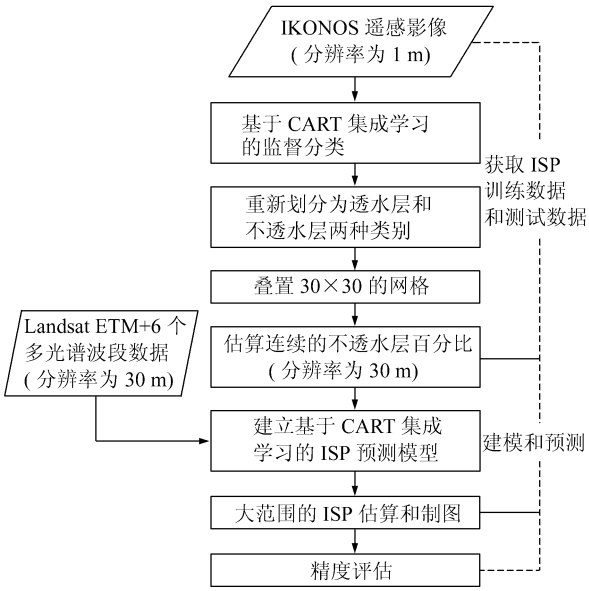

图 1 基于 CART 集成学习的 ISP 估算方法流程图

Fig. 1 Flowchart of Estimating Urban Impervious Surface Percent

2.2 ISP 训练数据的获取

ISP 训练数据的获取是本文估算方法的关键步骤，直接关系到 ISP 预测模型的质量。在同期 1∶1 万地形图和目视解译的基础上，利用 CART 集成学习算法对 IKONOS 影像进行监督分类，共提取出 7 类土地利用/覆盖地物类型，包括不透水层(主要由建筑物、道路等组成)、裸地、草地、树、水体和阴影。参考同期的地面资料，对分类结果进行精度评定，总分类精度为 89.15%，Kappa 系数为 0.904 1。

从该分类结果中计算落在 30 m 空间范围内(30×30 网格中)的不透水层像元总数(分辨率为 1 m)，进而可以计算得到分辨率为 30 m 的像元

不透水层百分比。需要说明的是,由于很难识别阴影的实际地物类型,被分为阴影的像元需要排除在百分比计算外。

2.3　ISP 估算结果

将利用 IKONS 影像获得的 ISP 结果(30 m 分辨率)作为 ISP 预测模型中的目标变量,Landsat TM/ETM$^+$影像的 6 个波段数据(第 6 波段除外)作为独立变量,采用随机分层采样方法从中抽取 1 887 组学习样本,采用 CART 集成学习算法对这些样本进行学习并建立 ISP 预测模型,其中算法中的学习迭代次数 k_{max} 为 25。

在 ISP 预测模型建立后,就可以通过 Landsat TM/ETM$^+$数据估算整个实验区的不透水层百分比,如图 2 所示。结合实地资料分析,图中 ISP 估算结果的空间分布模式从整体上来看比较合理,ISP 高于 60%的地区大部分集中在浦东新区的 7 大功能区内、主要城镇以及黄浦江沿岸地区,这主要是因为浦东新区城市化和工业化使得不透水层地表大为增加;在其他植被覆盖区(如城市绿地和菜地等)以及江(海)边滩涂地区,ISP 一般低于 30%,在城区和郊区结合地区则具有中等的 ISP。

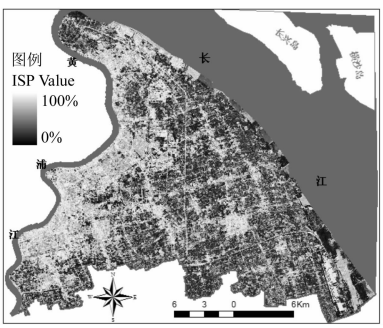

图 2　上海浦东新区的 ISP 估算结果

Fig. 2　Estimating ISP Result of Pudong New Area in Shanghai City by Using the Proposed Method

为了验证 CART 集成学习在 ISP 估算中的有效性,本文采用相同的训练样本进行了基于单一 CART 算法的 ISP 估算实验,图 3 是两种学习算法在金桥出口加工区附近的 ISP 估算结果比较。可以看出,单一 CART 方法受其学习能力的限制,估算结果并不理想,通过调查和判读,其整体估算效果尚可,但在植被地区的 ISP 估算结果虚高现象严重,如位于图中左下角的汤臣高尔夫球场和图右边的大片绿地(ISP 均达到 50%);另外,估算结果中存在较多噪声。相比较而言,CART 集成学习方法在整体上取得了令人满意的估算效果,对植被地区的 ISP 估算比较合理,并且呈现了更多的细节信息,见图 3(b)。

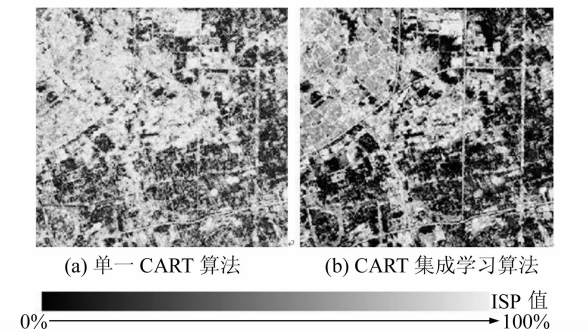

图 3　不同学习算法的 ISP 估算结果比较

Fig. 3　Comparison of ISP Result Estimated by Different Machine Learning Algorithms

2.4　精度评定

通过回归树模型预测的 ISP 是连续变量,因此,分类精度评估中常用的混淆矩阵和 Kappa 系数并不适用。本文采用统计回归分析中的 4 个常用评价指标来评估 ISP 预测模型的质量,即评估实际 ISP 值和预测 ISP 值之间拟合程度。这 4 个评价指标包括平均偏差、相对偏差和 Pearson 相关系数 r_2,这些指标已在 ISP 精度评估中得到了广泛的应用[3,5,6]。

随机抽取了 1 000 个测试样本用于定量评估和比较 ISP 预测模型的估算性能,这些测试样本完全独立于前面的 2 189 个训练样本。基于单一 CART 算法和 CART 集成学习算法的 ISP 预测模型评估结果见表 1。另外,本文也给出了测试样本中 ISP 预测值和参考值的线性回归优度拟合结果,如图 4 所示。

表 1　基于两种学习算法的 ISP 预测模型性能比较

Tab. 1　Performance Assessment of ISP Predicting Models

评价指标	单一 CART 学习	CART 集成学习
RMSE	13.72	10.24
平均偏差/(%)	14.47	11.16
相对偏差	0.56	0.31
r_2	0.82	0.91

测试样本的 ISP 预测值和参考值的回归分析结果表明,单一 CART 方法的回归直线较多地偏离了直线 $y=x$,其斜率和截距分别为 0.79 和 12.16,并且预测值和参考值在整个范围内波动较大,部分 ISP 预测值并不可靠,落在给定的 5%容忍误差边界之外,见图 4(a)。这主要是因为单一 CART 算法对噪声数据、样本突出值以及不均衡样本数据的学习能力有限,从而影响了其 ISP 估算性能。CART 集成学习算法由于在学习过程中采用了 Boosting 重采样技术,提高了预测的泛

化能力和稳健性,因此其 ISP 预测值和参考值的回归结果要优于单一 CART 算法,回归直线斜率高达 0.86,绝大部分预测值落在 5%容忍误差边界内,如图 4(b)所示。

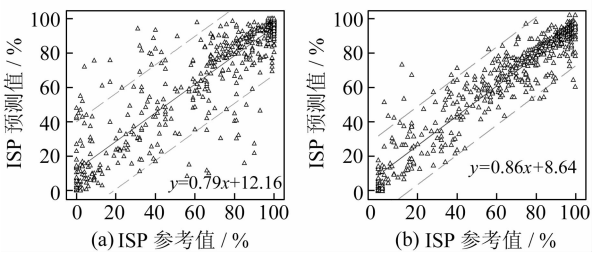

图 4 ISP 预测值和参考值的线性回归优度拟合结果

Fig. 4 Goodness-of-Fit of Linear Regression of Predicting and Reference ISP

3 结 语

本文提出了一种基于 CART 集成学习的 ISP 估算方法,该方法不仅继承了决策树方法的结构清晰、易于理解、运算效率高等优点,而且采用了基于 Boosting 重采样技术的集成学习方式,进一步提高了 ISP 估算的精度,满足了通过中等分辨率遥感影像进行城市 ISP 信息可靠提取的需要。然而,在光学遥感影像中,裸地(包括稀疏的草地)与一些人工建筑物(如停车场等)之间存在一定的光谱混淆,仍然导致了一些非人工地物的 ISP 估算偏高,ISP 遥感估算中的有效特征选取问题还有待进一步研究。

参 考 文 献

[1] Brabec E, Schulte S, Richards P L. Impervious Surfaces and Water Quality: a Review of Current Literature and Its Implications for Watershed Planning[J]. Journal of Planning Literature, 2002, 16(4):499-514

[2] Arnold J C A, Gibbons C J. Impervious Surface Coverage: the Emergence of a Key Urban Environmental Indicator[J]. Journal of the American Planning Association, 1996, 62(2):243-258

[3] Yang X. Estimating Landscape Imperviousness Index from Satellite Imagery[J]. Geoscience and Remote Sensing Letters, IEEE, 2006, 3(1):6-9

[4] Sangbum L, Lathrop R G. Subpixel Analysis of Landsat ETM/sup+/ Using Self-organizing Map (SOM) Neural Networks for Urban Land Cover Characterization[J]. IEEE Transactions on Geoscience and Remote Sensing, 2006, 44(6):1 642-1 654

[5] Yang L, Huang C, Homer C G, et al. An Approach for Mapping Large-Area Impervious Surfaces: Synergistic Use of Landsat 7 ETM+ and High Spatial Resolution Imagery [J]. Canadian Journal of Remote Sensing, 2003, 29(2):230-240

[6] Xian G, Crane M. Assessments of Urban Growth in the Tampa Bay Watershed Using Remote Sensing Data[J]. Remote Sensing of Environment, 2005, 97(22):203-215

[7] Wu C, Murray A T. Estimating Impervious Surface Distribution by Spectral Mixture Analysis[J]. Remote Sensing of Environment, 2003, 84(23):493-505

[8] Breiman L. Arcing Classifiers (with Discussion) [J]. Ann Statist, 1998, 26(3):801-849

[9] Breiman L, Friedman J, Olshen R. Classification and Regression Tree[M]. New York: Chapman and Hall, 1984

[10] Lawrence R, Bunn A, Powell S. Classification of Remotely Sensed Imagery Using Stochastic Gradient Boosting as a Refinement of Classification Tree Analysis[J]. Remote Sensing of Environment, 2004, 90(3):331-336

[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 Trans on Geoscience and Remote Sensing, 1999, 37(2):969-977

[12] Richard O D, Peter E H, David G S. 模式分类[M]. 李宏东, 姚天翔, 译. 北京: 机械工业出版社 & 中信出版社, 2003

第一作者简介:廖明生,教授,博士生导师。现主要从事遥感影像处理与分析、雷达干涉测量等方面的研究。

E-mail:liao@lmars.whu.edu.cn

(下转第 1106 页)