DOI:10.13203/j.whugis20130206　　文章编号:1671-8860(2014)11-1310-04

利用随机森林的城区机载 LiDAR 数据特征选择与分类

孙　杰[1]　赖祖龙[1,2]

1　中国地质大学(武汉)信息工程学院,湖北 武汉,430074
2　武汉大学遥感信息工程学院,湖北 武汉,430079

摘　要:针对机载 LiDAR 系统数据分类中多源特征与城区分类目标相关性不明确的问题,在面向对象的数据特征挖掘基础上,提出了一种基于随机森林的机载 LiDAR 系统特征选择与分类方法,利用不同地区数据实验证明:本文方法能对机载 LiDAR 系统数据多源特征的重要性进行正确评估,通过特征选择,在减少特征的情况下仍能够达到较高的分类精度。

关键词:面向对象;LiDAR;随机森林;相关性;分类

中图法分类号:P237.3　　**文献标志码**:A

机载 LiDAR(light detection and ranging)是近年来出现的一种新型的主动式遥感对地观测系统,能够直接获取地表密集的三维坐标点集(点云)及影像信息,已广泛应用于城区地物分类。

众多学者将 LiDAR 系统多种数据特征用于分类,譬如点云几何、纹理特征,点云强度特征,基于几何特征值的特征和影像光谱特征等。针对挖掘出的多种特征,出现了如 ISODATA[1]、贝叶斯网络[2]、Dempster shafer 融合法[3]、支持向量机、分类树[4]等分类方法。然而,已有的分类方法大多关注于分类的过程,对于 LiDAR 系统多种特征与分类目标间的相关性研究较少。与目标相关性较小的特征会增加分类复杂度,甚至影响分类精度。Mallet 对 LiDAR 系统数据特征的相关性进行了初步研究[5],通过 ReliefF 或者 F-score 对特征重要性进行排序,然后利用 SVM 分类器依次对特征进行筛选直至达到最优,这类方法在计算特征相关性时隔离了其他特征的影响,并且与分类算法未建立关联。Chehata 等[6]首次将随机森林(random forest, RF)引入 LiDAR 数据分类,对部分点云特征进行了讨论,Guo 等在此基础上增加了光谱特征重要性的讨论[7-8]。然而,已有的研究对特征的讨论都是基于单个点或者单个像素的,未充分考虑目标对象的纹理、空间特征在分类过程中的重要性。

针对以上问题,本文基于随机森林,提出了一种面向对象的城区机载 LiDAR 数据特征选择与分类方法,对城区目标对象的几何、光谱、纹理等特征进行相关性评估,筛选出合适的特征用于城区地物分类,并利用支持向量机(SVM)分类方法验证本文方法的有效性。其主要步骤如下:① 首先结合 LiDAR 系统数据中的点云高程信息对正射影像进行分割,提取分割对象影像和配准区域点云的多种特征;② 基于随机森林对各种特征相关性进行评估;③ 特征选择;④ 不同特征集下,RF 与 SVM 分类比较。

1　对象特征提取

机载 LiDAR 系统能够同时获取测区点云与 CCD 影像数据,利用检校后的 POS 数据对 CCD 影像进行正射纠正,可实现点云与影像的初步配准,结合点云数据对纠正后的影像进行分割,基于分割结果提取影像和配准区域点云中的多种特征,主要包括点云几何特征、影像光谱特征、影像纹理特征、空间特征等。

1) 几何特征。(1)基于高度的特征:平均归一化高度 μ_H 可以去除地形影响,由 DSM 与 DEM

收稿日期:2013-06-05
项目来源:地理信息系统软件及其应用教育部工程研究中心开放研究基金资助项目(20111104);中国博士后科学面上基金资助项目(2012M511300)。

第一作者:孙杰,博士,主要从事 LiDAR 数据分析处理、模式识别和计算机视觉研究。E-mail:sunjie_cug@163.com

的差值表示;高度方差 V_H,能够反应对象内部点云的垂直分布。(2) 基于平面的特征:① 点云法向量偏角 N_z;② 点云法向量偏离角方差 V_{NZ};③ 对象点云平面残差 R_Z。(3)基于特征值的特征:通过计算协方差矩阵获取对象特征值:$\lambda_1>\lambda_2>\lambda_3$,进而获取额外特征用于区分平面、边缘和线等特征,其中 λ_3 在平面区域值较低,非平面区域值较高:各向异性 $A_\lambda=(\lambda_1-\lambda_3)/\lambda_1$;平面性 $P_\lambda=(\lambda_2-\lambda_3)/\lambda_1$;球面性 $S_\lambda=(\lambda_3/\lambda_1)$;线性 $L_\lambda=(\lambda_1-\lambda_2)/\lambda_1$。

2) 回波特征。平均回波次数 EN(echo number)

3) 光谱特征:① 影像灰度均值 μ_R、μ_G、μ_B 分别为 3 个波段灰度均值;② 点云强度均值 μ_I。

4) 纹理特征。分割对象具有一定的纹理特征,灰度共生矩阵(GLCM)是一种纹理度量工具。本文选取两种代表性的纹理特征:方差 V_{RGB} 和信息熵 E_{RGB}。

5) 空间特征。对象的空间特征选用对象多边形紧密度 C(Compactness)衡量,$C=\mathrm{Sqrt}(4\times A/\mathrm{PI})$,$A$ 为对象面积。

2　基于随机森林的特征选择

2.1　随机森林

随机森林是由 Breiman 等提出的一种机器学习方法[9],是一种基于决策树的分类器,分类效果与 Boosting 和 SVM 相当。在不增加原样本集样本的情况下通过拔靴法选择样本子集构建一组分量分类器,然后利用投票机制综合分量分类器的结果得到最终分类结果,在构建分量分类器时,未被选中的样本组成袋外(out of bag,OOB)数据集,用袋外数据进行测试得到袋外误差。除了具有不需要对数据预处理、对多类问题处理方便快捷、分类结果稳定等优点,随机森林一个重要的特点是可以在样本训练的过程中实现特征重要性的评估,特征 f 的重要性通过随机置换 OOB 中的采样特征后,统计置换前后分类精度差异确定,也称作平均置换精度差异。

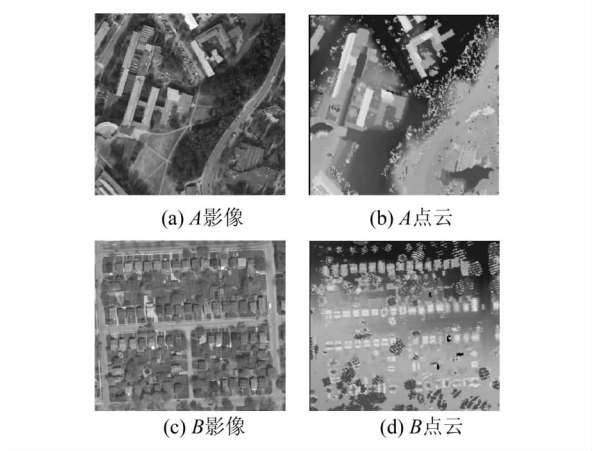
2.2　特征选择

特征选择的目的是从众多特征中标识出一个子集并使其能达到较好的分类效果。本文引入了一种逆向特征消除的方法进行特征选择,该方法已成功应用于基因芯片中的基因选择[10]。为了选择相关性最强的特征,对随机森林进行迭代拟合,在每次迭代过程中,通过设定比例因子 η(综合考虑计算复杂度和迭代删除比例通常设置为 0.2),对相关性较低的特征进行消除,在对森林进行完全拟合后,计算整个森林的最小标准误差,选择袋外误差小于 U 倍森林最小标准误差时的解。U 的理想值为 0,此时达到最小袋外误差,当 $U=1$ 时,可以得到最小特征集,但是能达到 $U=0$ 时相当的错误率,即分类树中的"1 Standard Error rule"规则[11]。本文选用 $U=1$ 时的特征集作为最终选取特征。

3　实　验

3.1　实验数据

为了验证本文方法,分别选用 A、B 两块不同区域的数据进行实验。数据 A 于 2002 年在赫尔辛基科技大学校区中心区域获取,采集工具为直升机搭载的 TopEye 激光扫描系统,平均航高 200 m,平均点云密度 2.5/m^2,相机为 2 000×3 000 像素,影像分辨率 0.5 m;数据 B 于 2004 年在尼亚加拉获取,采集工具为 Optech ALTM 3100 激光扫描系统,平均航高 1 190 m,平均点云密度 2.7/m^2,相机为 DSS 301 SN0039,影像分辨率为 0.2 m。如图 1,两个区域的数据均已进行检校,点云数据和纠正后的影像已可以较好地配准。



(a) A影像　(b) A点云　(c) B影像　(d) B点云

图 1　A、B 区域实验数据

Fig. 1　Expermental Data of Areas A、B

3.2　特征相关性

如图 1,A、B 两个区域数据都主要包含道路、树木、人工地、建筑等 4 大类,利用 eCognition 7.0 选取合适的尺度参数、光谱和形状因子,结合点云分别对 A、B 区域和系统正射影像进行分割,对分割后的影像对象及配准区域的点云进行特征挖掘,主要特征包括 §2 中所列出特征。选取一定数量的训练样本,利用 Liaw 等的 R 工具包组建

随机森林[12]，随机森林需要控制的参数有树的数目 Ntree 和每个分离点包含的变量数 Mtry，这里每个分类树的 Mtry 设为特征数的均方根约等于4，分类树个数 Ntree 设为 100。通过 Bootstrap 方法选择的训练集包含约 2/3 的原数据集样本，其余样本组成袋外数据，最终结果由所有决策树投票得出，这与交叉检验类似，可以限制过拟合。通过统计平均置换精度差异，20 种特征重要性如图 2 所示。

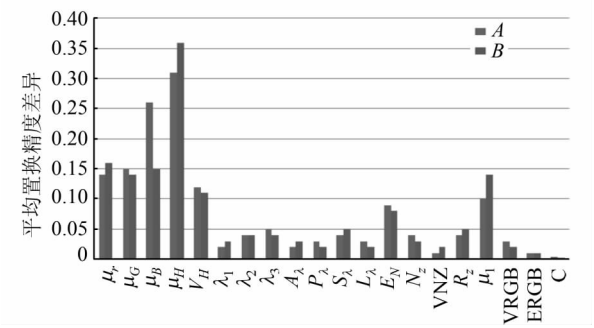


图 2 特征重要性分析图

Fig. 2 Features' Relevance Analysis

图 2 中条状数据分别为 A、B 区域中各种特征的平均置换精度差异。对于 A 区域，最重要的几类特征为对象平均归一化高度 μ_H 和影像波段灰度均值 μ_B、μ_R、μ_G，高度方差 V_H。对于 B 区域，最重要的几类特征为对象平均归一化高度 μ_H、影像波段灰度均值 μ_B、μ_R、μ_G，点云强度均值 μ_I、高度方差 V_H。对于 A、B 区域，对象平均归一化高度 μ_H 和影像波段灰度均值 μ_B、μ_R、μ_G 均为最重要的特征，这一结果的原因主要是由于 μ_H 代表了地物相对于地表的高度，能够直接用于区分大面积地表与非地表，光谱特征虽然包含丰富的地物信息，由于比较容易受阴影影响，重要性略弱，但是 μ_B 对于植被有较强的分辨力而 μ_G 对于植被分辨力较弱。回波数 EN 能较好的区分出树木等高植被，但是由于引入了 V_H，EN 的重要性显得稍弱。对于基于特征值的特征，实验区内由于非平面区域占较大比重，所以 λ_3 和 S_λ 在基于特征值的特征中表现出稍强重要性。此外，点云平面残差 R_Z 和法向量偏角 N_z 有利于建筑物顶面的分割，但是对于坡度较为敏感，重要性较弱，法向量方差特征 V_{N_Z}、空间紧密度特征 C 在两个区域中的的重要性都较弱。特别值得注意的是纹理特征 V_{RGB}、E_{RGB} 在 LiDAR 系统数据分类中的重要性较低，可能在较为复杂的地物分类中更能体现其重要性。

3.3 特征选择与分类

根据 A、B 区域分类特征重要性，采用逆向特征消除的方法分别去除冗余特征，图 3 和图 4 为这一过程的正向表达，图中横轴从左至右为按照重要性由强至弱依次增加的相关特征。

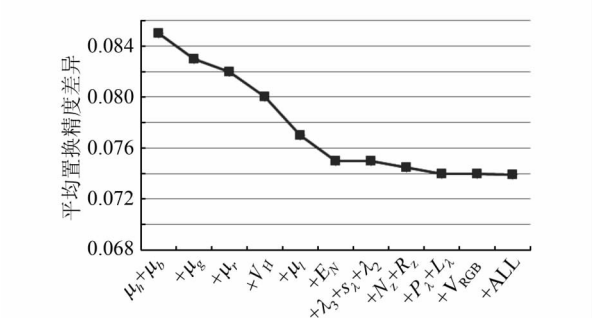


图 3 A 区特征迭代选择

Fig. 3 Feature Selection of Area A

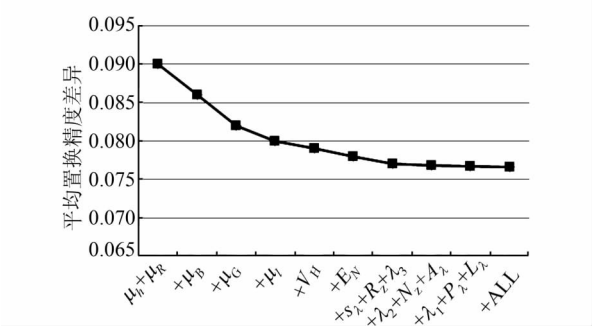


图 4 B 区特征迭代选择

Fig. 4 Feature Selection of Area B

A 区域通过特征选择后返回 15 个特征集 A_S，$[\lambda_1, A_\lambda, V_{N_Z}, E_{\mathrm{RGB}}, C]$ 被消除，B 区域特征选择后返回 16 个特征集 B_S，$[V_{N_Z}, V_{\mathrm{RGB}}, E_{\mathrm{RGB}}, C]$ 被消除。

为了验证特征选择与分类的有效性，利用 SVM 分类器分别对两个区域选择特征（A_S、B_S）和所有特征（ALL）进行分类。A、B 区域的验证数据通过专业人员结合影像目视解译与点云特征解译获取，对分类结果分别进行精度评定并与 RF 分类结果进行对比，结果见表 1。特征选择后，RF 分类精度与 SVM 相当，而 SVM 分类器在使用选择后的特征与使用所有特征相比，A 区数据分类全局精度由 92.45% 提升至 93.12%，B 区数据分类全局精度由 90.12% 提升至 91.25%，特征选择去除了部分特征后的分类精度略优于利用所有特征的分类精度。

表 1 不同分类特征分类精度

Tab. 1 Classificatoin Accuracy of Different Features

区域	方法_特征	全局精度/%
A	SVM_ALL	92.45
A	SVM_A_S	93.12
A	RF_A_S	93.25
B	SVM_ALL	90.12
B	SVM_B_S	91.25
B	RF_B_S	91.33

4 结　语

本文利用面向对象的方法挖掘机载 LiDAR 系统中多种特征。通过随机森林对各种特征与分类目标的重要性进行了全面评估,通过逆向迭代消除可以定量选择出与所分类目标最相关的特征。以 SVM 和 RF 分类器针对不同特征集进行实验,证明通过本文方法筛选出的特征分类精度略优于所有特征的分类精度,从而也证明了评估方法的有效性。随着面向对象的机载 LiDAR 系统数据分类在不同领域的不断广泛和深入的应用,本文方法将为不同地物分类中的特征评估与选择提供有力支撑。

参　考　文　献

[1] Haala N, Brenner C. Extraction of Buildings and Trees in Urban Environments[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 1999, 54(2): 130-137

[2] Stassopoulou A, Caelli T. Building Detection Using Bayesian Networks[J]. *International Journal of Pattern Recognition and Artificial Intelligence*, 2000, 14(6): 715-733

[3] Rottensteiner F, Trinder J, Clode S, et al. Using the Dempster-Shafer Method for the Fusion of LiDAR Data and Multi-spectral Images for Building Detection[J]. *Information Fusion*, 2005, 6(4): 283-300

[4] Ducic V, Hollaus M, Ullrich A, et al. 3D Vegetation Mapping and Classification Using Full-waveform Laser Scanning[C]. EARSel and ISPRS Workshop on 3D Remote Sensing in Forestry, Vienna, Austria, 2006

[5] Mallet C, Bretar F, Soergel U. Analysis of Full-waveform LiDAR Data for Classification of Urban Areas[J]. *Photogrammetrie Fernerkundung GeoInformation (PFG)*, 2008, 5: 337-349

[6] Chehata N, Guo L, Mallet C. Airborne LiDAR Feature Selection for Urban Classification Using Random Forests[J]. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2009, 39(3/W8): 207-212

[7] Mallet C, Bretar F, Roux M, et al. Relevance Assessment of Full-waveform LiDAR Data for Urban Area Classification[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2011, 66(6): S71-S84

[8] Guo L, Chehata N, Mallet C, et al. Relevance of Airborne LiDAR and Multispectral Image Data for Urban Scene Classification Using Random Forests[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2011, 66(1): 56-66

[9] Breiman L. Random Forests[J]. *Machine Learning*, 2001, 45(1): 5-32

[10] Díaz-Uriarte R, de Andres S A. Gene Selection and Classification of Microarray Data Using Random Forest[J]. *BMC Bioinformatics*, 2006, 7(1): 1-13

[11] Quinlan J R. Induction of Decision Trees[J]. *Machine Learning*, 1986, 1(1): 81-106

[12] Liaw A, Wiener M. Classification and Regression by Random Forest[J]. *R News*, 2002, 2(3): 18-22

Airborne LiDAR Feature Selection for Urban Classification Using Random Forests

SUN Jie[1]　*LAI Zulong*[1,2]

1　Institute of Information Engineering, China Geoscience University, Wuhan 430074, China

2　Institute of Remote Sensing Information Engineering, Wuhan University, Wuhan 430079, China

Abstract: To the question that multisource features' contribution to classification is not explicit in airborne LiDAR system data, based on object oriented data mining, this paper proposed a method to select features for classification using Random Forest. It's proved that the features' contribution can be evaluated correctly and the selected features can still make a high classification accuracy.

Key words: object oriented; LiDAR; random forest; relevance; classification

First author: SUN Jie, PhD, specializes in LiDAR data analysis and processing. E-mail:sunjie_cug@163.com

Foundation support: The Open Fund of Geographic Information System Software and Application Engineering Research Center of the Education Ministry, No. 20111104; the China Postdoctoral Science Foundation Projects, No. 2012M511300.