

一种机载 LiDAR 点云分类的自适应特征选择方法

张爱武 肖涛 段乙好

首都师范大学三维信息获取与应用教育部重点实验室, 北京 100048

摘要 不同地形条件下,不同的特征组合、特征维数对点云的分类效率及分类结果有不同的影响。提出了一种机载 LiDAR 点云分类的自适应特征选择方法,该方法依据地形起伏情况对整体点云数据进行区域划分,自适应选择适宜该区域 LiDAR 点云分类的特征集合。为了验证这种特征选择方法的有效性,利用优选后的特征集合,分别采用随机森林和支持向量机算法进行分类实验验证,实验结果表明,在不同地形条件的区域里,适合 LiDAR 点云分类的特征集合不同。该方法可以有效地降低特征维数,缩短运算时间,且分类精度较高。

关键词 遥感; 机载 LiDAR; 自适应特征选择; 点云分类; 随机森林; 支持向量机

中图分类号 TN959.3 文献标识码 A

doi: 10.3788/LOP53.082802

A Method of Adaptive Feature Selection for Airborne LiDAR Point Cloud Classification

Zhang Aiwu Xiao Tao Duan Yihao

Ministry of Education Key Laboratory of 3D Information Acquisition & Application,
Capital Normal University, Beijing 100048, China

Abstract In different terrain conditions, different feature combinations and dimensions have different influences on the effectiveness and accuracy of classification. A method is proposed to select the airborne LiDAR point cloud classification with adaptively feature selection. The whole point cloud is divided into different regions in accordance with the terrain conditions, and the suitable feature set is selected adaptively for classification. In order to evaluate the effective of this method, the random forest method and support vector machine classification method are used to classify the experimental data with the feature set after optimization. Experimental result shows that the suitable feature set for classification in different areas are different. The proposed method can reduce the feature dimensions effectively, shorten time consumption, and achieve high classification accuracy.

Key words remote sensing; airborne LiDAR; adaptive features selection; point cloud classification; random forest; support vector machine

OCIS codes 280.3420; 280.3640; 280.5600

1 引言

机载 LiDAR 可以快速获取大面积地物的高精度、高密度三维坐标信息,在林业调查、灾害预警、数字城市建设等领域有广泛的应用。由于地形起伏、地物之间拓扑关系的复杂性,如何快速有效地从点云数据中提取出地物信息是实现机载 LiDAR 后续应用的关键步骤之一。

国内外学者一般采用以下三种方法进行点云分类:1)利用 LiDAR 数据的几何及地形特征提取出特定的地物。左志权等^[1]以栅格高程图像为基础,依据自动聚类以及拓扑模型实现了地面图斑与其他地物图斑的分离,采用回波比率和建筑物高度等条件分别提取出树木和建筑物。张志伟等^[2]通过分析机载 LiDAR

收稿日期: 2016-03-23; 收到修改稿日期: 2016-04-05; 网络出版日期: 2016-07-29

基金项目: 国家自然科学基金(41571369)、教育部博士点基金(20131108110005)、北京市长城学者(CIT&TCD20150323)

作者简介: 张爱武(1972—),女,教授,博士生导师,主要从事地理信息工程、激光雷达、计算机视觉等方面的研究。

E-mail: zhangaw98@163.com

点云数据中道路信息的连通性、起伏程度以及反射强度等既有知识,设置不同的阈值,利用数学形态学中的开运算方法滤除树木及建筑物数据,从而提取出道路数据。卢维欣等^[3]利用点云的集合特征分离地面点,对非地面点按距离聚类,通过对聚类得到的点云簇进行整体分析,识别出了建筑物点云数据。苗启广等^[4]对点云数据建立规则格网,利用各网格中点云的高程标准差预测地形坡度,从而通过渐进式形态滤波提取出了地面激光点云。Strimbu等^[5]提出获取森林 LiDAR 点云的分层拓扑结构,通过量化树冠组成部分的加权图的拓扑关系进行图像分割,进而分割出单棵树冠。以上这种单纯利用 LiDAR 数据的几何及地形特征提取出特定的地物方法往往容易受地形起伏等因素影响,而且分类地物类型单一,同种分类算法不能同时应用于提取其他地物。2)融合多源特征,采用机器学习的方法进行监督分类。龚亮等^[6]提出一种 LiDAR 数据与航空影像融合的方法,将每个点云数据都赋予 RGB 值,根据赋予的光谱信息提取出植被,然后通过高程纹理以及回波次数分别提取出建筑物及剩余的植被点云。Gong等^[7]通过聚类算法分离出道路和植被,然后将 LiDAR 与航空遥感影像融合,设置点云强度和光谱范围阈值,最终分离出道路点云。刘丽娟等^[8]将机载 LiDAR 数据与高光谱数据进行融合,采用支持向量机(SVM)分类方法对融合的数据进行分类,实现了温带天然林不同树种的识别。以上几种分类方法仅强调了点云特征对于分类的重要性,忽略了特征向量中可能存在的冗余特征以及分类能力较差的特征对于分类精度和效率的影响。3)对点云特征进行优选,获取最优特征集合,从而达到降低分类复杂性,提高分类精度的目的。Mallet等^[9]利用 Kononenko等^[10-11]算法对点云特征重要性进行分析,利用 SVM 分类方法对特征进行优选直至得到最优特征集。Guo等^[12]在 Mallet等^[9]基础上加入了多光谱特征,分析了光谱特征对于分类的重要性,采用随机森林算法对点云分类,取得了良好的效果。孙杰等^[13]采用获取特征的平均置换精度的方式得到特征的重要性排名,利用后向特征消除法得到最优特征集合然后进行分类。郭波等^[14]提取了点云分类的必要特征,再结合地物在空间上的相互关联关系,利用 JointBoost 分类器^[15]实现了点云分类和特征降维。范士俊等^[16]融合了激光点云的振幅、回波次数、波宽等波形特征并用随机森林算法对实验区域点云进行分类。上述几种点云特征优选方法在地形起伏不大的情况下,可以取得良好效果,但是在地形复杂的区域,由于 LiDAR 点云特征对不同地形分类的贡献程度不同,如果按照同一标准进行特征选择则很难取得理想分类结果。

基于此,本文提出基于不同地形区域的自适应特征选择方法。依据地形起伏情况对整体点云数据进行分块使得各区域的地形起伏相似,以减小地形对分类结果的影响;通过随机森林特征变量重要性算法评估每个区域中点云各特征的分类能力,并对特征进行排名,去除能力较弱特征;再计算特征集合中各特征之间的相关系数,去除特征之间高度相关且重要性排名较低的特征,从而得到比较适宜该地形区域 LiDAR 点云分类的特征集合;利用随机森林(RF)和 SVM 分类方法验证本文方法的有效性。

2 特征提取

机载 LiDAR 点云数据具有的几何特征和波形特征可以作为 LiDAR 点云监督分类的重要依据。点云的几何特征可以通过对该点的三维邻域内所有点进行统计分析得到。邻域一般有三种表示方法^[12]:1) v_k 最近邻 K 邻域,即以距离待分类点最近的 K 个点形成的点的集合;2) v_s 球体邻域,即以待分类点为中心,半径为 r 的球体内的所有点的集合;3) v_c 柱状邻域,即以待分类点为中心,半径为 r 形成的垂直圆柱体中所有点的集合。本文采用 v_s 球体邻域提取点云的几何特征(如邻域高差、竖直角等),可以通过非线性最优化算法提取点云的波形特征(振幅、波宽、回波次数等)。

1) 高程相关的特征:高度特征能够有效区分地面点与非地面点。高度 H 可以反映出点云的垂直信息,如图 1 所示,一般地面点高度最低、树木的冠层以及建筑物的点云高度值均高于地面点;邻域高差 ΔZ ,即待分类点与邻域内最低点的差值,可以反映出邻域点云内部垂直结构,树木点云的邻域高差数值较大。

2) 投影相关的特征:将待分类点的球形邻域 v_s 中的所有点投影到 xy 平面上进行一定的格网划分,统计出非空网格的个数作为待分类点云的投影特征 P_a ,不同地物类型的点云数据 P_a 值不同。如图 2(a)、(b)所示,树木顶层的点云 P_a 值一般较大,而建筑物边缘点云的 P_a 值则相对较小。

3) 面相关特征:

①竖直角 θ ,即对象点所在的邻域内所有点拟合平面的法向量与竖直方向的夹角。图 3 反映了不同地

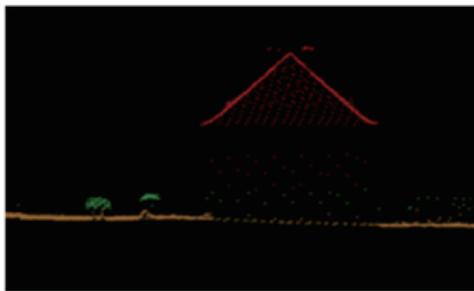


图1 局部点云按高程显示图

Fig. 1 Display of local point cloud according to elevation

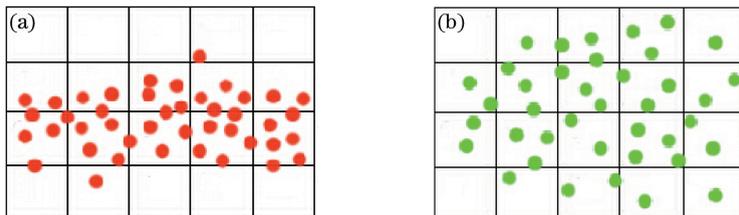


图2 不同地物点云的投影特征。(a)建筑物边缘;(b)树木

Fig. 2 Projection features of different objects point cloud. (a) Edge of building; (b) tree

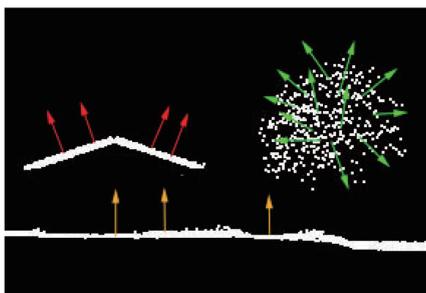


图3 点云法向示意图

Fig. 3 Diagram of the point cloud vector

物类型的点云的法向基本分布,地面点云和建筑物点云的竖直角值相对比较稳定,而树木点云的竖直角则比较散乱。

②平均竖直角 θ_a ,即邻域内点的竖直角平均值。法向量是几何体表面的重要属性,一般采用主成分分析(PCA)方法得到^[17]。首先对象点与 v_s 邻域内点创建协方差矩阵 C 表示为

$$C = \frac{1}{k} \sum_{i=1}^k (\mathbf{P}_i - \bar{\mathbf{P}}) \cdot (\mathbf{P}_i - \bar{\mathbf{P}})^T, \quad (1)$$

式中 k 为 v_s 邻域内点的个数, \mathbf{P}_i 表示每个点云对象, $\bar{\mathbf{P}}$ 表示最近邻元素的三维质心。协方差矩阵 C 的最小特征值对应的特征向量即为拟合平面的法向量 $\mathbf{n} = (n_x, n_y, n_z)$, 由法向量 \mathbf{n} 可以计算得到 θ_a 为

$$\theta_a = \arctan(\sqrt{n_x^2 + n_y^2} / n_z). \quad (2)$$

③法向散射系数 N_s ,即 v_s 邻域内点竖直角方差。

④点到拟合平面距离 d_p ,利用邻域 v_s 内的点的坐标,通过最小二乘法可以得到拟合平面的方程,计算得到 d_p 为

$$d_p = \frac{|ax_i + by_i + cz_i + d|}{\sqrt{a^2 + b^2 + c^2}}, \quad (3)$$

式中 a, b, c, d 为拟合平面 $ax + by + cz + d = 0$ 的参数。

⑤平面拟合残差 δ ,即邻域所有对象点到拟合平面距离之和。通过此特征可以将有多个面的对象与仅

有单个面的对象分离。

⑥表面系数 $S^{[12]}$, 即对象点到邻域拟合平面距离的方差的平均值。研究区域内的建筑物既有平面屋顶也有“人”字形屋顶, 平面拟合残差对于“人”字形屋脊位置不够敏感, 因此通过表面系数可以减小屋脊位置点云数据被错分的概率。表面系数的计算表达式为

$$S = \frac{1}{n} \sum_{i=1}^n \sqrt{\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d}_p)^2}, \quad (4)$$

式中 n 表示 v_s 邻域内点的个数。 d_i 表示当前点到拟合平面的距离, \bar{d}_p 表示邻域内点到拟合平面的距离的平均值。

4) 协方差矩阵相关特征: 协方差矩阵可以作为三维点云的结构张量, 能够反映出点集的空间分布特征。协方差矩阵的特征组合可以有效区分城区地物^[18]。通过计算邻域内点云协方差矩阵获取对象特征值 λ_1 、 λ_2 、 λ_3 , 按照降序排列 $\lambda_1 > \lambda_2 > \lambda_3$ 。进而获取用于区分平面边缘和线等特征。协方差矩阵的相关特征^[19]:

①特征值 λ_1 、 λ_2 、 λ_3 。其中 λ_3 在平面区域值较低, 在非平面区域值较高。当 $\lambda_1 \approx \lambda_2 \approx \lambda_3$ 时点集是一个空间体结构; 当 $\lambda_1 \gg \lambda_2 \approx \lambda_3$ 时点集近似一条直线; 当 $\lambda_1 \approx \lambda_2 \gg \lambda_3$ 时点集近似一个面。

②各向异性系数 $A_\lambda = (\lambda_1 - \lambda_3) / \lambda_1$ 和平面指数 $P_\lambda = (\lambda_2 - \lambda_3) / \lambda_1$ 能够体现出点云邻域的平面特性。线性指数 $L_\lambda = (\lambda_1 - \lambda_2) / \lambda_1$ 能够表现出邻域点云的线性特征。球面指数 $S_\lambda = \lambda_3 / \lambda_1$, 信息熵 $E_\lambda = -\sum_{i=1}^3 \lambda_i \ln(\lambda_i)$, 结构张量总变化指数 $O_\lambda = \sqrt[3]{\lambda_1 \lambda_2 \lambda_3}$ 反映出邻域点云的三维结构特征。

5) 波形特征: 全波形机载 LiDAR 数据中包含地物后向散射的全部波形信息, 记录了一束脉冲沿途遇到的地物回波信号的总和。一束激光脉冲打到每个地物后返回的回波波形理论上呈高斯分布, 一般认为仪器接收的回波信号波形是由多个高斯函数叠加而成, 因此可以通过多个高斯函数的集合进行模拟全波形的回波信号。根据回波高斯函数的振幅初值、波峰位置以及高斯函数的标准差, 通过非线性最优化算法^[20]进行波形分解得到振幅 A , 波宽 W 和回波次数 N 。

综上分别提取了点云的高度 H 、邻域高差 ΔZ 、投影面积 P_a 、竖直角 θ 、平均竖直角 θ_a 、点到拟合平面距离 d_p 、平面拟合残差 δ 、表面系数 S 、协方差矩阵归一化特征值 λ_1 、 λ_2 、 λ_3 、各向异性系数 A_λ 、平面指数 P_λ 、线性指数 L_λ 、球面指数 S_λ 、信息熵 E_λ 、结构张量总变化指数 O_λ 、振幅 A 、波宽 W 和回波次数 N , 共计 20 个特征。

3 自适应特征选择方法

特征选择的目的是去除分类能力较弱的特征和特征集合中的冗余特征, 降低分类的复杂性、提高点云的分类效率和精度。本文特征选择过程主要采用随机森林变量重要性算法来评估特征的分类能力, 并结合相关系数衡量特征之间的相关程度, 设定相关系数阈值, 去除特征之间高度相关但分类能力较弱的特征。

3.1 特征选择流程

本文特征选择主要包括两个步骤: 1) 根据随机森林特征变量重要性算法获取每个特征对于类别的重要程度系数, 并依据重要程度系数的高低对特征进行排序, 重要程度系数作为衡量特征分类能力依据, 其中排名较低的特征被认为是分类能力较弱的特征; 2) 计算特征之间的相关系数, 假如两特征的相关系数较高, 那么其中分类重要程度系数排名较低的特征即被认为是冗余特征, 可以考虑删除。经过设置相关系数阈值筛选后得到的特征组合即为最终点云分类依赖的训练特征向量集合。

图 4 为特征选择流程图。具体流程如下:

1) 将所有点云特征组合在一起构成原始特征集合 F , 利用随机森林特征重要性算法计算 F 中每个特征的分类重要程度系数, 并依据分类重要程度系数从大到小排序;

2) 根据特征分类重要程度系数的高低, 依次将特征加入到特征集合中, 并统计分类精度, 当分类精度达到最高时, 设定特征分类重要程度系数阈值 I , 去除特征集 F 中排名靠后的特征, 得到特征集合 F' ;

3) 构建特征集合 F' 中的相关系数系数矩阵;

4) 设置相关系数阈值 T , 从对称矩阵的第一列开始比较对角线下方的值与 T 的大小, 若小于 T , 则保留该特征, 否则删除该特征, 已删除的特征与其他特征的相关系数不再跟阈值 T 进行比较;

5) 重复步骤 4), 直到所有的特征分析完成, 即可得到优选的特征集合 F'' 。

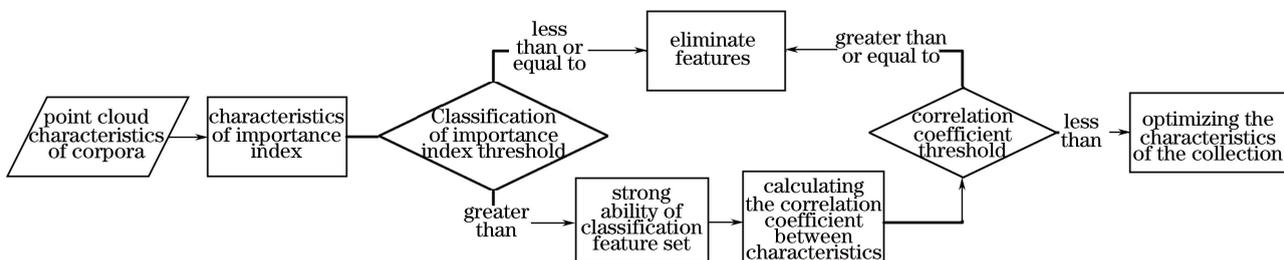


图 4 特征选择流程图

Fig. 4 Flowchart of feature selection

3.2 随机森林特征变量重要性评估算法

随机森林^[21]算法是由 Leo Breiman 和 Adele Cutler 等提出的一种基于分类和决策树(CART)的组合算法。它是由一系列决策树 $\{h(x, \theta_k), k=1, 2, 3, \dots, K\}$ 组合构成的强分类器, 其中 K 表示随机森林中决策树的个数, $\{\theta_k\}$ 为独立同分布的随机变量。随机森林算法在训练中有两次随机选择的过程。首先通过有放回的抽样方式随机选取样本数据, 然后再通过随机选择每个样本集的特征进行数据集的分割。

随机森林对于噪声和缺失数据的稳健性较强, 通过度量特征变量重要性提取出较重要的特征可以对高维数据进行降维^[22]。假设随机森林共有 K 棵决策树, 训练样本个数为 N , 每个样本数据有 M 个特征变量, 特征变量 $m (m=1, 2, 3, \dots, M)$ 的重要程度指数用 I_m 表示。随机森林衡量特征变量重要性算法步骤^[23]如下:

1) 从训练样本中有放回地随机抽取 N 个样本, N 为训练样本数据的总数。重复 K 次, 得到 K 个 bootstrap 数据集。在总训练样本中建立决策树 $T_k, k=1, 2, 3, \dots, K$, 并利用该树中的 bootstrap 数据集 C_k 作为训练数据对其袋外数据 B_k 进行分类, 统计正确分类的个数 R_k 。

2) 随机改变袋外数据 B_k 中的特征变量 $V_m (m=1, 2, 3, \dots, M)$ 的值, 即对袋外数据 B_k 加入了噪声扰动。假设扰动后的袋外数据为 B_k^m , 使用决策树 T_k 对 B_k^m 进行分类, 统计正确分类的个数 R_k^m 。重复本步骤直到所有的特征全部分析完毕。

3) 重复 1)、2) 步骤, 直到所有的决策树分析完毕。

4) 特征变量 m 的重要程度指数 I_m 可以表示为

$$I_m = \frac{1}{K} \sum_{k=1}^K (R_k - R_k^m) \quad (5)$$

变量重要程度指数 I_m 能够反映出该变量的分类能力的强弱。将各特征按照其特征变量重要程度指数的高低进行排序, 排名靠后的特征分类能力较弱, 可以认为该特征对分类贡献较小而予以剔除。这样做一方面可以减少特征数量, 降低分类复杂性提高分类速度, 另一方面也可以提高分类的正确率。

3.3 相关系数

一般采用皮尔逊相关系数计算两个连续型变量的相关性, 它是反映两变量相关关系的方向和密切程度的指标。由于特征计算值在一定的区间范围内可以认为是连续的, 所以利用皮尔逊相关系数来衡量 LiDAR 点云的特征之间的相关程度。特征 f_A, f_B 的相关系数用 r_{AB} 表示

$$r_{AB} = \frac{\sum_{i=1}^n (f_{Ai} - \bar{f}_A) (f_{Bi} - \bar{f}_B)}{\sqrt{\sum_{i=1}^n (f_{Ai} - \bar{f}_A)^2 \sum_{i=1}^n (f_{Bi} - \bar{f}_B)^2}} \quad (6)$$

式中 n 表示样本个数。相关系数 r_{AB} 的值域为 $[-1, 1]$, $r_{AB} > 0$ 表示两变量正相关, $r_{AB} = 0$ 表示两变量不相关, $r_{AB} < 0$ 表示两变量负相关。 $|r_{AB}|$ 的值越接近 1, 表示两变量的相关程度越高。

通过计算特征集合中各特征之间的相关系数,构建特征相关系数矩阵,设置相关系数阈值,降低特征集合的冗余度,从而达到特征降维和进一步减少分类时间的目的。

4 实验及分析

所采用的实验数据是由 RIGEL 激光扫描仪获取的,共 763320 个激光点。图 5 为点云按高程显示的效果图。从图中可以看出,该区域有一定的地形起伏,整体西部区域的高度低于东部区域。为尽量减小地形的起伏对整体点云分类的影响,将点云分成了地形变化不大的 A、B 两块区域如图 6(a)、(b)所示,其中 A 区域共 352578 个激光点,B 区域共 410742 个激光点。

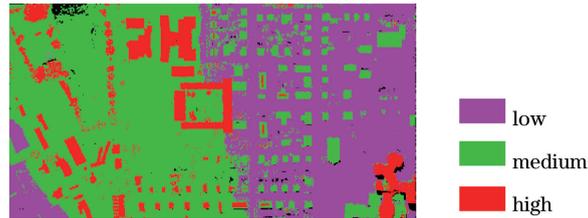


图 5 整体点云数据按高程显示

Fig. 5 Display of the whole area according to elevation

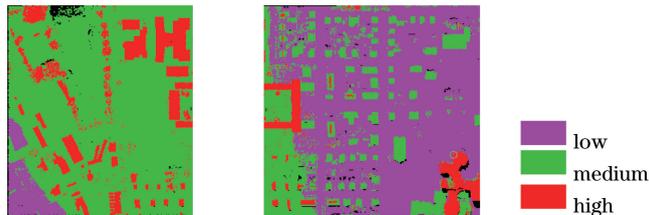


图 6 实验区点云划分。(a) A 区域;(b) B 区域

Fig. 6 Point cloud classification in test area. (a) Area A; (b) area B

将上文提取的点云特征置于随机森林算法模型中,在模型中共设置了 200 棵树,每个数据集中随机选取 5 个特征用于分割节点。用随机森林特征变量重要性算法分析 A、B 两区域每个特征对于类别的重要程度,并对其进行降序排列,结果如图 7(a)、(b)所示。从图中可以看出,各个特征对于分类的重要程度指数均高于 0.047。由于两研究区域都有地形起伏,因此在图 7(a)、(b)中可以看到能反映出邻域点云的高低情况的邻域高差 ΔZ 排名相对靠前,而点云数据的高度值 H 排名却相对靠后。波形特征排名比较靠前,说明波形特征有助于点云分类。协方差矩阵相关的特征中 λ_3 分类能力最强,而 λ_1 、 λ_2 以及各向异性系数 A_λ 分类能力则相对较弱。

按照特征变量的排名,依次将特征加入到训练样本特征集中,采用随机森林算法进行点云分类,分别统计出 A、B 两区域的不同特征个数情况下分类精度。结果图 8(a)、(b)所示,其中 A 区域分类精度最高达到 89.506%,此时特征数为 18,重要程度系数为 0.053,B 区域特征数为 19 时整体分类精度最高达到 89.203%,此时特征变量重要程度系数为 0.052。因此 A 区域重要程度系数阈值 I_A 设为 0.053,B 区域特征变量重要程度系数阈值 I_B 设为 0.052。此时 A 区域点云的特征 λ_2 、平面拟合残差 δ 以及 B 区域点云的各向异性系数 A_λ 被认为是分类能力较弱的特征并予以剔除。

图 8 中的特征集合是按照每个特征的分类重要程度系数由高到低依次加入的,因此特征变量重要程度系数的高低能够反映出该特征分类能力的强弱。若特征集中没有冗余特征,那么点云的分类精度曲线的斜率应该是随着特征的增加而逐渐降低,而且分类精度是逐渐增加的。但是从图 8(a)、(b)中可以看出,随着特征的逐渐加入,曲线的斜率趋势却是杂乱的,甚至出现了分类精度降低的情况。这说明 A、B 两区域的训练特征集中均存在不同程度的冗余特征。

为了去除点云特征集合中的冗余特征,通过构建特征相关系数矩阵,设置相关系数阈值,以达到进一步精简适宜点云分类的特征集合的目的。不同的相关系数阈值条件下各区域所选取的特征集合不同,得到分

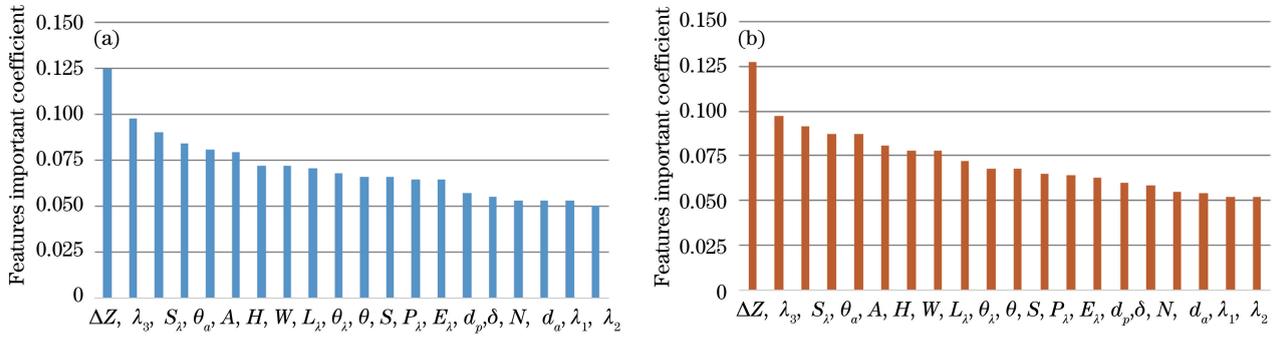


图 7 不同区域特征重要程度系数图。(a) A 区域；(b) B 区域

Fig. 7 Features important coefficient of different areas. (a) Area A; (b) area B

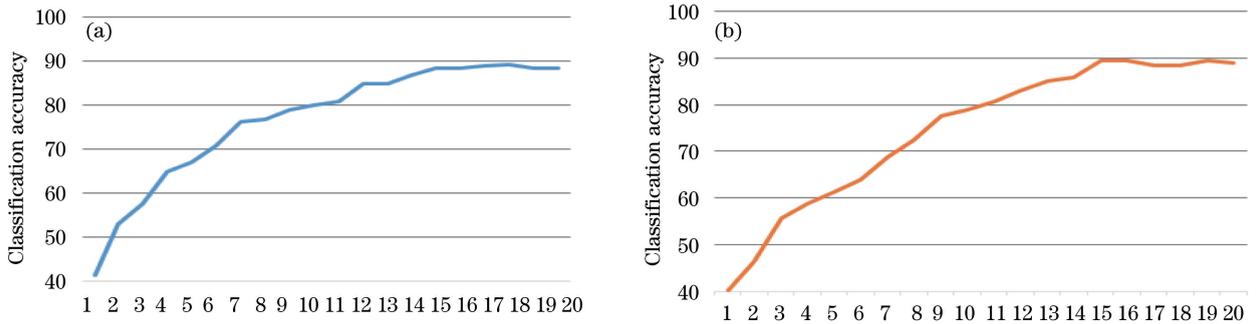


图 8 不同区域分类精度曲线。(a) A 区域；(b) B 区域

Fig. 8 Classification accuracy curves of different areas. (a) Area A; (b) area B

	ΔZ	λ_3	θ_a	A	S_x	O_x	W	H	L_x	θ	N	S	P_x	E_x	P_a	A_x	d_p	λ_1	λ_2	d
ΔZ	1	0.839123	0.834744	-0.503549	0.751814	0.898489	0.584196	0.302224	0.731409	0.879584	0.433879	0.754815	-0.294673	-0.646519	0.296591	0.137461	0.558346	0.558139	0.544484	0.671248
λ_3	0.839123	1	0.798466	-0.520574	0.81655	0.833146	0.559836	0.279176	0.752716	0.723649	0.515139	0.725186	-0.567597	-0.525246	0.40106	0.078912	0.507602	0.384655	0.41384	0.67936
θ_a	0.834744	0.798466	1	-0.510171	0.758213	0.844693	0.616729	0.313512	0.691395	0.848139	0.422515	0.751261	-0.35194	-0.526065	0.215246	0.131069	0.531228	0.440741	0.5364	0.69875
A	-0.503549	-0.520574	-0.510171	1	-0.510137	-0.362578	-0.134687	-0.483847	-0.435504	-0.285297	-0.473191	0.286818	0.262441	-0.167386	-0.021706	-0.32036	-0.197083	-0.22405	-0.447057	
S_x	0.751814	0.81655	0.758213	-0.507508	1	0.794379	0.626651	0.279759	0.757567	0.467478	0.448305	0.462564	-0.900075	-0.24675	0.134449	-0.073214	0.433312	0.109579	0.224971	0.570346
O_x	0.898489	0.833146	0.844693	-0.510137	0.794379	1	0.55897	0.307593	0.751684	0.759176	0.469647	0.759214	-0.376823	-0.676435	0.478298	0.141016	0.582347	0.568797	0.558876	0.743581
W	0.584196	0.559836	0.616729	-0.362578	0.626651	0.55897	1	0.191964	0.581518	0.520795	0.154385	0.575164	-0.43139	-0.175127	0.015202	0.049719	0.410533	0.134653	0.104703	0.477954
H	0.302224	0.279176	0.313512	-0.134687	0.279759	0.307593	0.191964	1	0.250795	0.280969	0.085932	0.273274	-0.10621	-0.10621	0.039971	0.051323	0.184793	0.075961	0.063319	0.223746
L_x	0.731409	0.752716	0.691395	-0.483847	0.757567	0.751684	0.581518	0.250795	1	0.630765	0.424449	0.658605	-0.481493	-0.397961	0.208422	0.134764	0.479104	0.317018	0.272339	0.88598
θ	0.879584	0.723649	0.848139	-0.435504	0.467478	0.759176	0.520795	0.280969	0.630765	1	0.375139	0.652604	-0.279416	-0.49892	0.095441	0.132835	0.495364	0.425739	0.42131	0.576252
N	0.433879	0.515139	0.422515	-0.285297	0.448305	0.489647	0.154385	0.085932	0.424449	0.375139	1	0.265151	-0.179891	-0.330512	0.273514	-0.012058	0.134796	0.244234	0.31848	0.26329
S	0.754815	0.725186	0.751261	-0.473191	0.462564	0.759214	0.575164	0.273274	0.658605	0.626515	0.265151	1	-0.4093	-0.429319	0.258317	0.176913	0.675522	0.375228	0.302224	0.43637
P_x	-0.294673	-0.567597	-0.35194	0.286818	-0.900075	-0.376823	-0.43139	-0.224733	-0.481493	-0.279416	-0.179891	-0.4093	1	-0.187669	0.141406	-0.351315	-0.268267	0.157155	0.51841	-0.343461
E_x	-0.646519	-0.525246	-0.526065	0.262441	-0.24675	-0.676435	-0.175127	-0.10621	-0.10621	-0.49892	-0.330512	-0.429319	-0.187669	1	-0.625103	-0.238997	-0.334979	-0.85251	-0.758329	-0.445612
P_a	0.296591	0.40106	0.215246	-0.167386	0.134449	0.478298	0.015202	0.039971	0.208422	0.095441	0.273514	0.258317	0.141406	-0.625103	1	0.093393	0.194081	0.575264	0.571302	0.240893
A_x	0.137461	0.078912	0.131069	-0.021706	-0.073214	0.141016	0.049719	0.051323	0.134764	0.132835	-0.012058	0.176913	-0.351315	-0.238997	0.093393	1	0.178215	0.472241	-0.300075	0.196584
d_p	0.558346	0.507602	0.531228	-0.32036	0.433312	0.582347	0.410533	0.184793	0.479104	0.495364	0.134796	0.675522	-0.268267	-0.334979	0.194081	0.178215	1	0.314021	0.232398	0.693294
λ_1	0.558139	0.384655	0.440741	-0.197083	0.109579	0.134653	0.075961	0.317018	0.425739	0.244234	0.375228	0.157155	-0.85251	0.575264	0.472241	0.314021	1	0.574702	0.346479	
λ_2	0.546484	0.414384	0.51364	-0.225405	0.220871	0.558876	0.106703	0.063319	0.273274	0.424131	0.518948	0.399824	0.518944	0.766329	0.674362	0.586875	0.232366	0.970702	1	0.573503
d	0.671248	0.647936	0.649875	-0.407057	0.570346	0.703581	0.487354	0.223746	0.58598	0.576252	0.203229	0.83637	-0.383451	-0.106614	0.260892	0.190684	0.690804	0.946479	0.274593	1

图 9 A 区域相关系数矩阵

Fig. 9 Correlation coefficient matrix of area A

类结果也不相同。图 9、图 10 分别为 A、B 两区域的相关系数矩阵，特征的放置顺序依据每个特征的分类重要程度系数从高到低进行排列。其中红色直线划掉的部分表示分类能力较弱的特征，不再参与相关系数阈值筛选。

经过一系列的实验，得到了如图 11 所示的不同相关系数阈值条件下 A、B 两区域的分类精度曲线图，从图 11(a)中可以看到，当 A 区域的特征相关系数阈值 T_A 设置为 0.90 时，A 区域的分类精度达到最高，最高精度为 91.206%、B 区域的相关系数阈值 $T_B=0.85$ 时，分类精度达到最高，最高精度为 89.914%。A、B 两区域的整体精度达到 90.657%。

以去除 A 区域冗余特征为例，参考图 9 中 A 区域相关系数矩阵图，当相关系数阈值为 0.9 时，从相关系数矩阵的第一列 ΔZ 开始比较对角线下方的相关系数绝对值与阈值 0.9 的大小，若小于阈值，则保留该特征，否则删除该特征。已删除的特征参与相关系数阈值的比较。可以看到第一列相关系数阈值均低于 0.90，然后开始比较第二列 λ_3 所在列的值与阈值的大小，依次类推，当分析至特征 S_x 时，发现 S_x 与 P_x 相关系数超过阈值，而

ΔZ	ΔZ	λ_3	S_λ	θ_a	A	H	W	L_λ	O_λ	θ	S	P_λ	E_λ	d_p	δ	N	P_a	λ_1	λ_2	A
ΔZ	1	0.66938	0.532917	0.799309	-0.446041	0.150868	0.504249	0.511748	0.812422	0.865048	0.653751	0.023927	-0.633988	0.447976	0.576506	0.315716	0.227997	0.594852	0.533425	0.269169
λ_3	0.66938	1	0.80122	0.634731	-0.578646	0.025526	0.509398	0.867845	0.849891	0.569244	0.725163	-0.505632	-0.291173	0.491076	0.651156	0.472094	0.36456	0.205057	0.179242	0.164879
S_λ	0.532917	0.80122	1	0.545396	-0.555408	-0.000903	0.575269	0.703298	0.710481	0.477836	0.632344	-0.614718	-0.047324	0.400282	0.547481	0.428197	0.123179	-0.050357	-0.01963	0.003568
θ_a	0.799309	0.634731	0.545396	1	-0.443391	0.183819	0.515119	0.448954	0.723369	0.917238	0.639232	-0.043692	-0.474507	0.408042	0.537677	0.314633	0.116829	0.439216	0.430109	0.234068
A	-0.446041	-0.578646	-0.555408	-0.443391	1	-0.067505	-0.393839	-0.502197	-0.556672	-0.384394	-0.517274	0.27728	0.187408	-0.339662	-0.443286	-0.211877	-0.148713	-0.121301	-0.111721	-0.114721
H	0.150868	0.025526	-0.000903	0.183819	-0.067505	1	0.00347	-0.02128	0.047221	0.145069	0.043215	0.072523	-0.079934	-0.011869	0.033454	-0.033267	0.06466	0.078382	0.102011	0.010087
W	0.504249	0.509398	0.575269	0.515119	-0.393839	0.00347	1	0.495006	0.48979	0.433863	0.511046	-0.27349	-0.14203	0.323029	0.423692	0.15983	-0.000793	0.102928	0.073263	0.11413
L_λ	0.511748	0.867845	0.703298	0.448954	-0.502197	-0.02128	0.495006	1	0.658974	0.434021	0.584408	-0.441376	-0.175816	0.399959	0.501255	0.376203	0.192387	0.865896	0.031845	0.184981
O_λ	0.812422	0.849891	0.710481	0.723369	-0.556672	0.047221	0.48979	0.658974	1	0.655396	0.768455	-0.218371	-0.560504	0.587859	0.696054	0.411996	0.480532	0.764703	0.426106	0.253598
θ	0.865048	0.569244	0.477836	0.917238	-0.384394	0.145069	0.433863	0.434021	0.655396	1	0.555043	-0.024301	-0.451512	0.389562	0.480437	0.283607	0.025258	0.421575	0.402706	0.21487
S	0.653751	0.725163	0.632344	0.639232	-0.517274	0.043215	0.511046	0.584408	0.768455	0.555043	1	-0.327483	-0.289074	0.635575	0.843068	0.219681	0.0267231	0.25755	0.174831	0.252252
P_λ	0.023927	-0.505632	-0.614718	-0.043692	0.27728	0.072523	-0.27349	-0.441376	-0.216371	-0.024301	-0.327483	1	-0.476169	-0.233496	-0.299371	-0.211071	0.142378	0.413482	0.857996	-0.239762
E_λ	-0.633988	-0.291173	-0.047324	-0.474507	0.187408	-0.079934	-0.14203	-0.175816	-0.560504	-0.451512	-0.289074	-0.476169	1	-0.2002	-0.26852	-0.148618	-0.467828	-0.767982	-0.863219	-0.23437
d_p	0.447976	0.491076	0.400282	0.408042	-0.339662	-0.011869	0.033454	0.399959	0.587859	0.389562	0.635575	-0.233496	-0.2002	1	0.657769	0.107979	0.19947	0.193739	0.105021	0.232333
δ	0.576506	0.651156	0.547481	0.537677	-0.443286	0.033454	0.423692	0.501255	0.696054	0.480437	0.843068	-0.233496	-0.2002	0.657769	1	0.159568	0.27172	0.245343	0.150934	0.246577
N	0.315716	0.472094	0.428197	0.314633	-0.304548	-0.033267	0.15983	0.376203	0.411996	0.283607	0.219681	-0.21071	-0.148618	0.107979	0.159568	1	0.193082	0.099916	0.116243	0.08524
P_a	0.227997	0.36456	0.123179	0.116829	-0.211877	0.06466	-0.000793	0.192387	0.480532	0.025258	0.287231	0.142378	-0.467828	0.19947	0.27172	0.193082	1	0.423801	0.453154	0.104957
λ_1	0.594852	0.205057	-0.050357	0.439216	-0.148713	0.078382	0.102928	0.865896	0.764703	0.421575	0.25755	0.413482	-0.767982	0.193739	0.245343	0.099916	0.423801	1	0.557613	0.464287
λ_2	0.533425	0.179242	-0.01963	0.430109	-0.121301	0.102011	0.073263	0.031845	0.426106	0.402706	0.174831	0.857996	-0.863219	0.150521	0.150934	0.116243	0.453154	0.557613	1	-0.216919
A	0.269169	0.164879	0.003568	0.234068	-0.114721	0.010087	0.11413	0.184981	0.253598	0.21487	0.252252	-0.239762	-0.23437	0.232333	0.249577	0.03524	0.104957	0.464287	-0.214919	1

图 10 B 区域相关系数矩阵
Fig. 10 Correlation coefficient matrix of area B

且 S_λ 特征重要程度系数高于 P_λ , 因此, 特征集合中保留 S_λ , 去除特征 P_λ 。然后继续分析下一列数据直到所有特征相关系数分析完成。使用同样的方法去除 B 区域的特征集合中的冗余特征。最终, A 区域中的平面指数 P_λ 以及 B 区域的特征 λ_1, λ_2 、竖直角 θ 和线性指数 L_λ 被认为是冗余特征而被去除。因此 A 区域中最终参与分类的特征集合为 $\{\Delta Z, \lambda_3, \theta_a, A, S_\lambda, O_\lambda, W, H, \theta, N, S, L_\lambda, E_\lambda, P_a, A_\lambda, d_p, \lambda_1\}$ 共计 17 个特征, B 区域中最终参与分类的特征集合为 $\{\Delta Z, \lambda_3, S_\lambda, \theta_a, A, H, W, O_\lambda, S, P_\lambda, E_\lambda, d_p, \delta, N, P_a\}$ 共计 15 个特征。

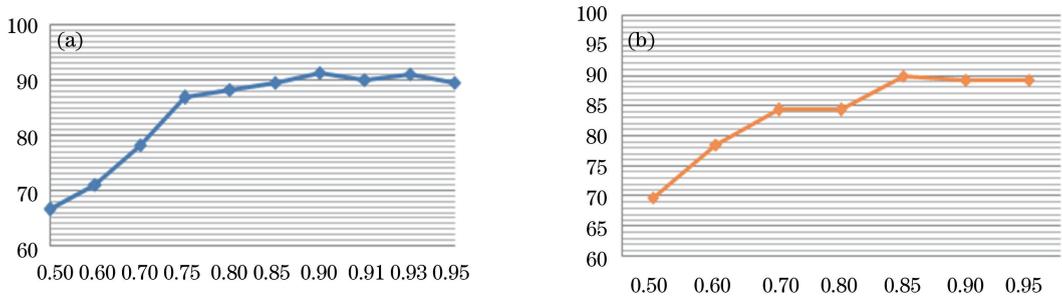


图 11 不同相关系数阈值条件下各区域分类精度图。(a) A 区域; (b) B 区域

Fig. 11 Classification accuracy of each area with different correlation coefficient thresholds. (a) Area A; (b) area B

实验共选取了 37022 个验证样本用来测试分类精度。表 1、表 2 分别为未进行特征优选和经过特征优选后使用随机森林分类算法得到的混淆矩阵, 分类整体精度分别达到 90.657% 和 88.947%, 各类地物点云的分类精度均有一定提升, 其中建筑物和地面的分类精度提高较明显, 分别提高了 2.15% 和 2.28%。地形起伏不同的 A、B 两区域最终参与点云分类的特征集合不同, 说明不同的地形, LiDAR 点云的特征分类能力不同, 参与点云分类的有效特征集合也不相同。经过特征选择后, 点云的分类精度都有所提高。图 12(a)、(b) 分别为特征优选前后使用随机森林分类效果图。实验证明了本文算法对提高点云分类精度的有效性, 说明即使存在地形的高低起伏, 本文算法依然可以取得较好的分类效果。

表 1 未进行特征选择的 RF 分类混淆矩阵

Table 1 Confusion matrix of RF classification without feature selection

	Building	Car	Ground	Tree	Precision / %
Building	9802	251	467	409	89.688
Car	105	3221	169	213	86.866
Ground	793	276	10940	263	89.390
Tree	430	341	375	8967	88.668

(Overall accuracy: 88.947%)

表 2 利用优选的特征集进行 RF 分类混淆矩阵

Table 2 Confusion matrix of RF classification using the optimal feature set

	Building	Car	Ground	Tree	Precision /%
Building	10037	183	326	383	91.838
Car	115	3234	156	203	87.217
Ground	491	153	11250	378	91.672
Tree	427	335	309	9042	89.410

(Overallaccuracy: 90.657%)

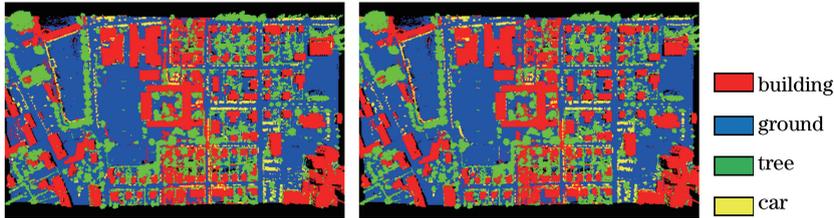


图 12 特征选择前后随机森林分类结果。(a)未进行特征选择;(b)利用优选特征集的分类结果

Fig. 12 Result of RF before and after feature selection. (a) Without feature selection;

(b) classification result using the optimal feature set

为了测试本文算法的稳健性,采用 SVM 分类方法进行对比。SVM 是由 Vapnik^[24]提出的一种比较优秀的分类算法。以统计学习理论为基础,将样本空间映射到更高维的特征空间中,并寻找样本在高维空间中最优线性区分超平面,该算法具有很好的泛化性^[25]。图 13(a)、(b)分别为特征选择前后使用支持向量机分类算法得到的效果图。表 3、表 4 为特征优选前后利用 SVM 分类的混淆矩阵,可以看到地物的分类精度都有不同程度的提高,整体精度提高了 2.118%。说明本文算法稳健性较好,得到的特征集合可以适用于不同的分类器中。

表 3 未进行特征选择的 SVM 分类混淆矩阵

Table 3 Confusion matrix of SVM classification without feature selection

	Building	Car	Ground	Tree	Precision /%
Building	9759	264	481	395	89.295
Car	131	3225	167	185	86.974
Ground	840	305	10815	312	88.127
Tree	776	226	378	8733	86.354

(Overallaccuracy: 87.926%)

表 4 利用优选的特征集进行 SVM 分类混淆矩阵

Table 4 Confusion matrix of SVM classification using the optimal feature set

	Building	Car	Ground	Tree	Precision /%
Building	9876	262	373	418	90.365
Car	108	3227	162	211	87.028
Ground	594	283	11195	200	91.224
Tree	458	349	271	9035	89.340

(Overallaccuracy: 90.044%)

表 5 对比了特征选择前后的 SVM 分类算法和随机森林(RF)分类算法的分类时间和精度。由于随机森林(RF)有个多次交叉验证的过程,因此分类时间比 SVM 长,但是 RF 分类精度略高于 SVM,另外,通过特征选择后,SVM 和 RF 算法的分类时间均有所降低,分类精度都有不同程度的提高,说明本文算法普适性较好,能够有效提高分类精度和效率。

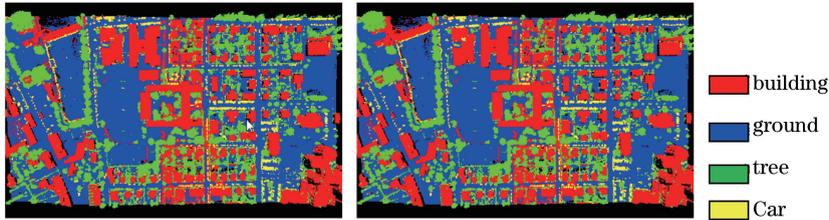


图 13 特征选择前后 SVM 分类结果。(a)未进行特征选择;(b)利用优选的特征集的分类结果

Fig. 13 Classification results of SVM before and after feature selection.

(a) Without feature selection; (b) classification result using the optimal feature set

表 5 分类时间与精度

Table 5 Classification of time and accuracy

	Time /s	Accuracy /%
SVM	307	87.926
SVM+feature selection	302	90.044
RF	381	88.947
RF+feature selection	368	90.657

5 结 论

通过区域划分的方式将实验区域划分为地形起伏不太明显的 A、B 两区域以降低地形对于分类的影响,根据特征的分类能力和特征之间的相关性自适应地对特征进行优选。特征选择过程表明 A 区域点云的 λ_2 、平面拟合残差 δ 以及 B 区域点云的各向异性系数 A_λ 的分类能力较低,A 区域中的平面指数 P_λ 以及 B 区域的特征 λ_1 、 λ_2 、竖直角 θ 和线性指数 L_λ 则被认为是训练样本中的冗余特征。采用 RF 和 SVM 算法对点云进行分类,结果表明,通过剔除训练样本特征向量中的分类能力较弱的特征和冗余特征的特征选择方法在提高分类效率的情况下,分类精度也有所提高。实验结果表明,在不同地形条件的区域里,适合 LiDAR 点云分类的特征集合不同。寻求一个能准确反映特征变量分类能力的指数可能会进一步提高点云数据分类精度,这也是以后研究工作的重点。

参 考 文 献

- Zuo Zhiqian, Zhang Zuxun, Zhang Jianqing. Classification of LiDAR point clouds for urban area based on multi-echo region ratio and recognition topology model[J]. Chinese J Lasers, 2012, 39(4): 0414001.
左志权, 张祖勋, 张剑清. 区域回波比率与拓扑识别模型结合的城区激光雷达点云分类方法[J]. 中国激光, 2012, 39(4): 0414001.
- Zhang Zhiwei, Liu Zhigang. Method extraction road information in LIDAR data based on the existing knowledge of asymptotic mathematical morphology[J]. Science of Surveying and Mapping, 2010, 35(4): 154-156.
张志伟, 刘志刚. 利用既有知识渐近数学形态学提取 LiDAR 数据中道路信息方法研究[J]. 测绘科学, 2010, 35(4): 154-156.
- Lu Weixin, Wan Youchun, He Peipei, et al. Extracting and plane segmenting building from large scene point cloud[J]. Chinese J Lasers, 2015, 42(9): 0914004.
卢维欣, 万幼川, 何培培, 等. 大场景内建筑物点云提取及平面分割算法[J]. 中国激光, 2015, 42(9): 0914004.
- Miao Qiguang, Guo Xue, Song Jianfeng, et al. LiDAR point cloud data with morphological filter algorithm based on region prediction[J]. Laser & Optoelectronics Progress, 2015, 52(1): 011003.
苗启广, 郭雪, 宋建锋, 等. 基于区域预测的 LiDAR 点云数据形态学滤波算法[J]. 激光与光电子学进展, 2015, 52(1): 011003.
- Strimbu V F, Strimbu B M. A graph-based segmentation algorithm for tree crown extraction using airborne LiDAR data [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2015, 104: 30-43.
- Gong Liang, Li Zhengguo, Bao Quanfu. Classification of LiDAR object points by fusing aerial image[J]. Engineering of

- Surveying and Mapping, 2012, 21(1): 34-38.
- 龚亮, 李正国, 包全福. 融合航空影像的 LiDAR 地物点云分类[J]. 测绘工程, 2012, 21(1): 34-38.
- 7 Gong L, Zhang Y, Li Z, *et al.* Automated road extraction from LiDAR data based on intensity and aerial photo[C]. 2010 3rd International Congress on Image and Signal Processing (CISP), 2010: 2130-2133.
- 8 Liu Lijuan, Pang Yong, Fan Wenyi, *et al.* Fused airborne LiDAR and hyperspectral data for tree species identification in a natural temperate forest[J]. Journal of Remote Sensing, 2013, 17(3): 679-695.
- 刘丽娟, 庞勇, 范文义, 等. 机载 LiDAR 和高光谱融合实现温带天然林树种识别[J]. 遥感学报, 2013, 17(3): 679-695.
- 9 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.
- 10 Kononenko I. Estimating attributes: Analysis and extensions of RELIEF[J]. Machine Learning, 2005, 784: 171-182.
- 11 Chen Y W, Lin C J. Combining SVMs with various feature selection strategies[M]. Berlin: Springer, 2006: 315-324.
- 12 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.
- 13 Sun Jie, Lai Zulong. Airborne LiDAR feature selection for urban classification using random forests[J]. Geomatics and Information Science of Wuhan University, 2014, 39(11): 1310-1313.
- 孙杰, 赖祖龙. 利用随机森林的城区机载 LiDAR 数据特征选择与分类[J]. 武汉大学学报·信息科学版, 2014, 11(39): 1310-1313.
- 14 Guo Bo, Huang Xianfeng, Zhang Fan, *et al.* Points cloud classification using jointboost combined with contextual information for feature reduction[J]. Acta Geodaetica et Cartographica Sinica, 2013, 42(5): 715-721.
- 郭波, 黄先锋, 张帆, 等. 顾及空间上下文关系的 Joint Boost 点云分类及特征降维[J]. 测绘学报, 2014, 42(5): 715-821.
- 15 Torralba A, Murphy K P, Freeman W T. Sharing visual features for multiclass and multiview object detection[J]. Pattern Analysis and Machine Intelligence, 2007, 29(5): 854-869.
- 16 Fan Shijun, Zhang Aiwu, Hu Shaoxin, *et al.* A method of classification for airborne full waveform LiDAR data based on random forest[J]. Chinese J Lasers, 2013, 40(9): 0914001.
- 范士俊, 张爱武, 胡少兴, 等. 基于随机森林的机载激光全波形点云数据分类方法[J]. 中国激光, 2013, 40(9): 0914001.
- 17 Meng Fang. The study of interactive rendering for large scale three dimensional point cloud data[D]. Beijing: Beijing University, 2005.
- 孟放. 大型三维点云数据的交互绘制研究[D]. 北京: 北京大学, 2005.
- 18 Carlberg M, Gao P, Chen G, *et al.* Classifying urban landscape in aerial LiDAR using 3D shape analysis[C]. Image Processing (ICIP), 2009: 1681-1684.
- 19 Gross H, Thoennessen U. Extraction of lines from laser point clouds[C]. Symposium of ISPRS Commission III: Photogrammetric Computer Vision PCV06 International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2006: 86-91.
- 20 Li Hanlun, Zhang Aiwu, Liu Zhao, *et al.* A LiDAR point classification method based on SVM and waveform decomposition[J]. Bulletin of Surveying and Mapping, 2014, 1: 28-32.
- 李含伦, 张爱武, 刘诏, 等. 基于 LiDAR 波形分解的点云 SVM 分类方法研究[J]. 测绘通报, 2014, 1: 28-32.
- 21 Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.
- 22 Yao Dengju, Yang Jing, Zhan Xiaojuan. Feature selection algorithm based on random forest[J]. Journal of Jilin Univerdity (Engineering and Technology Edition), 2014, 44(1): 137-141.
- 姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工学版), 2014, 44(1): 137-141.
- 23 Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: A survey and results of new tests[J]. Pattern Recognition, 2011, 44(2): 330-349.
- 24 Vapnik V N. Statistical learning theory[M]. New York: Wiley, 1998.
- 25 Luts J, Ojeda F, van de Plas D, *et al.* A tutorial on support vector machine-based methods for classification problems in chemometrics[J]. Analytica Chimica Acta, 2010, 665(2): 129-145.