地矿测绘
Geological and Mineral Surveying and Mapping
第 3 卷◆第 3 期◆版本 1.0◆2020 年
文章类型：论文|刊号（ISSN）：2630-4732 /（中图刊号）：561GL001

基于加权随机森林的机载LiDAR点云建筑物提取

洪绍轩　袁枫　齐吉婧
航天宏图信息技术股份有限公司
DOI:10.32629/gmsm.v3i3.761

[摘　要] 针对传统建筑物点云提取算法中先滤波后提取依赖滤波精度,离散点云组织方式与参数设置困难,泛化能力弱等问题,利用改进的加权随机森林模型对LiDAR点云建筑物进行提取。首先针对传统随机森林算法等权投票方式导致分类器整体性能下降的问题,对随机森林分类的投票预测模型进行改进,并实现基于改进的加权随机森林模型对LiDAR点云建筑物的粗提取,然后运用改进的迭代三角网方法对初始建筑物点云进行精确提取,最后利用栅格化连通性原则实现建筑物点云的单体分割。实验选取国际摄影测量与遥感协会提供的典型区域的LiDAR点云数据进行建筑物提取,并与传统随机森林算法进行建筑物提取进行比较,结果表明本文算法可以实现建筑物的高精度提取,验证了算法的可靠性与适用性。

[关键词] 机载LiDAR；建筑物提取；加权随机森林；单体化分割

机载LiDAR是一种快速获得高精度地表三维空间地理信息的测量技术,为测绘事业发展提供了便捷、准确、有利的技术支持[1]。

近年来,针对机载激光雷达点云数据的建筑物检测问题,国内外学者做了很多相关的研究工作,依据建筑物点云检测的策略,可以将现有的建筑物点云检测方法分为传统式与智能式两类分类方法。传统式建筑物提取方法,大多基于滤波和检测建筑物两个步骤。赵宗泽提出一种基于数学形态学的LiDAR建筑物点云区域检测算法[2]。Mohammad等提出一种将多光谱图像与机载激光雷达点云数据相结合的建筑物点云自动检测算法[3]。汪禹芹通过TIN点云分割算法检测建筑物点云[4]。Ben等提出一种基于三角形的区域生长算法对建筑物点云进行检测[5]。上述传统的算法中,基于图像处理技术的算法需要经过重采样处理,会影响原始LiDAR点云数据的基本特征与初始结构；基于多源数据融合的策略,需要对点云以05应的影像进行智能有准确的叠加处理；采用聚类分析的策略,需要计算庞大的距离矩阵,导致运算效率低。基于人工智能方法有随机森林(Random Forest, RF)、支持向量机(Support Vector Machine, SVM)、贝叶斯神经网络AdaBoost等机器学习方法[6 7],目前这些人工智能方法大多用在点云的分类技术中,其中最常用的是随机森林方法,例如,熊艳等人提出了通过构造随机森林分类器进行机载LiDAR点云分类的方法[8]；孙杰等人利用随机森林算法对城区机载LiDAR点云数据进行分类；Guan等人提出了基于分割算法的随机森林地物分类方法[9]。上述都是属于点云分类方法,如果直接利用分类结果提取出建筑物,由于没有后处理技术,无法保证建筑物的精度。

传统随机森林算法是将所有的预测结果按照统一的权重进行投票,这样就会忽略不同分类决策树对不同地物的分类能力的差别,强分类器会产生效果更好的预测,而弱分类器则会影响分类结果的准确性,因此这样的投票方式会影响随机森林分类器的分类性能。鉴于此,很多相关的研究人员在随机森林的投票方式上做了很多改进,Croux等通过计算决策树分类的好坏,判断决策树的分类能力,将其中分类效果不好的的决策树移除,不参与投票[10]。杨彪等通过再次选取训练样本作为预测数据,根据预测数这下睡的舒服了吧据的分类结果,对每棵树的投票权重进行修改[11]。

综合以上考虑,针对传统随机森林算法投票方式进行改进,通过计算不同决策树对每一个类别的分类错误率,定义该决策树对某一类别的投票权重,充分考虑决策树的分类能力,使得投票模型更加注重决策树分类性能,对那些对某一类别地物的分类错误率低的决策树赋予得更高权重,对某一类别地物的分类错误率高的决策树赋予较低权重,为了随机森林模型的分类精度更高。其次对分类后的建筑物增加相应的后处理技术,更精确的提取建筑物,并实现建筑物的单体化分割。

1 原理

本文提出的建筑物提取方法,是基于点的分类方式初步检测建筑物,再对建筑物点云进行精细化提取。首先基于加权随机森林算法,对粗差去除后的的点云进行建筑物检测,然后迭代构建Delaunay三角网,根据连通性准则,去除误分的非建筑物点云,得到精细化的建筑物点云,最后利用栅格连通性原则对建筑物点云进行单体化分割。具体流程图如图1所示。

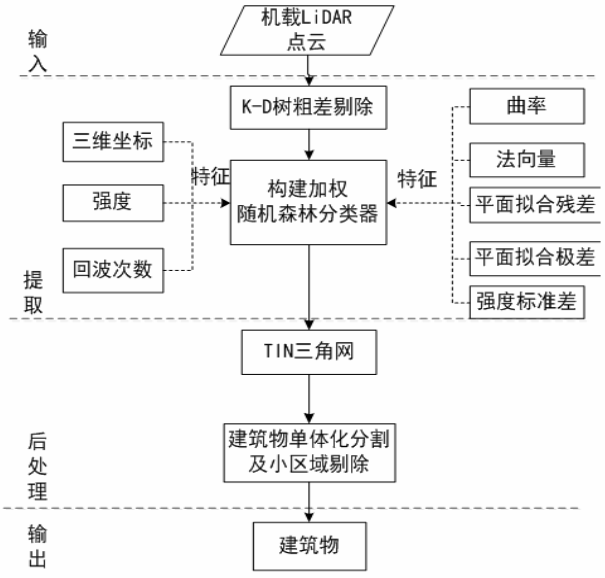


图1　研究技术路线

1.1特征提取

1.1.1高程分布征

高程分布特征可以描述不同地物的高程在空间上的分布特点[12]。定义所有空间三维点云集合为：

$$E = \{N_i(x_i, y_i, z_i) \mid i = 1, 2, ..., n\} \quad (1)$$

式中,N表示第i个点的空间三维坐标。

设集合中任意一点为：

$$P = N_1 \in E \quad (2)$$

则以该点为中心半径为R的邻域点集可表示为：

$$P=\{P_j \mid \left\| P_j-P\le R \right\| P_j\ne P, i=1,2,\ ...,n\} \tag{3}$$

通过对任意一点与其邻域范围内点做相关统计，选取高程的极差和标准差作为属性特征库的子集，其计算公式如下。

(1)Z坐标极差H_r：

$$H_r=\max(Z_i)-\min(Z_i) \tag{4}$$

(2)Z坐标标准差H_{STD}：

$$H_{STD}=\sqrt{\frac{\sum_{i=1}^{n}\left(Z_i-\tilde{Z}\right)^2}{n}} \tag{5}$$

式中，Z_i表示第i个邻域点的高程值；$\tilde{Z}$表示任意点与其邻域范围内的点的高程。

1.1.2空间统计特征

空间统计特征主要描述点集中任意一点与其在邻域范围内的点在一维、二维、三维空间内分布的程度，主要计算的是邻域点云的平面曲率与法向量。

定义任一点P_i的邻域点集合M：

$$M=\{Q_i((x_i,y_i,z_i)\mid i=1,2,...,k\} \tag{6}$$

式中，Q为第i个点的三维坐标。

可以表现出P_i的邻域特征，如P_i点的邻域样本协方差矩阵可表现为：

$$C_{P_i}=\frac{1}{k}\sum_{i=1}^{k}(Q_i-\mu)(Q_i-\mu)^T \tag{7}$$

式中，

$$\mu=\frac{1}{k}\sum_{i=1}^{k}Q_i \tag{8}$$

因为激光点云读取的是空间三维坐标，因此C_{P_i}对应三个特征值。其主成分为：

$$\begin{cases}R_1=\psi_1^{\ T}M\\R_2=\psi_2^{\ T}M\\R_3=\psi_3^{\ T}M\end{cases} \tag{9}$$

得到对应于当前点的3个主成分系数ψ_1、ψ_2、ψ_3（$\psi_1\le\psi_2\le\psi_3$），进一步对这3个主成分系数进行归一化：

$$\lambda_1=\psi_i\Big/\sum_{i=1}^{3}\psi_i \tag{10}$$

点集M相似于一个椭球体，点集M的三个主成分类似于椭球的三个短轴[13]，特征根λ_1、λ_2、λ_3可表示为轴的长度。其中最小的λ对应的ψ就是P_i所在平面的法向量，因此可以判断点的P_i邻域离散点是否共面。设定点的P_i表面曲率为：

$$f=\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} \tag{11}$$

1.1.3表面相关特征

地物表面特征可以定义为拟合面的粗糙度、拟合面极差（S_r）以及拟合平面标准差（S_{STD}）。表面特征示意图如图2所示。利用最小二乘原理对邻域内所有点集做平面拟合，拟合面粗糙度定义为任一中心点到该拟合平面的距离；拟合面极差定义为邻域点至拟合平面最大的距离差；拟合面标准差定义为邻域点至拟合平面距离的标准差。

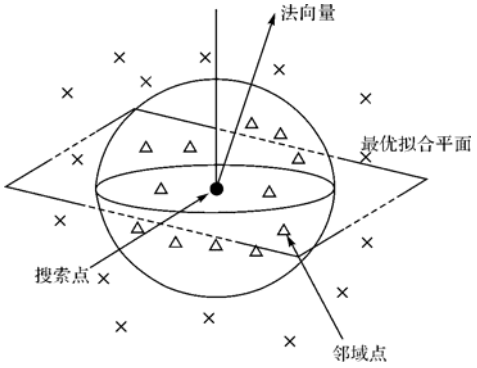


图2　表面相关特征示意图

1.1.4激光回波特征

平均强度值定义为：邻域点集内强度的平均值，同时计算邻域点的强度标准差。

1.2基于加权随机森林算法建筑物提取

基于加权随机森林算法初步检测建筑物点云的具体步骤如下：

(1)设有训练样本集。

$$Q=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\} \tag{12}$$

共包含有M个特征，样本类别$C\in\{L_1,\ L_2,\ ...,\ L_t\}$，其中$x_i\in X$，$y_i\in Y$，$i=1,2,...,n$。

(2)特征选择。对每棵决策树的叶节点处，按照基尼系数准则进行最优特征的选择，通过叶节点分裂前后基尼系数下降量达到最大的准则，选取最优特征的特征分裂量。

(3)剪枝。对每个节点的样本数设置限制阈值，即当某个节点的样本数大于限制阈值$n_{threshold}$时，才可以继续分裂，最终构建决策树$h_t(x)$。

(4)随机森林模型就由以上步骤生成的T棵分类所树组成，根据决策树对每一个类别的分类能力计算决策树对该地物类别的分类权重。对于类别$C\in\{L_1,L_2,\ ...,\ L_n\}$，决策树的对每一类地物的初始权重设置为：

$$w_i^1=\frac{1}{n},\ \ i=1,2,...,n \tag{13}$$

文章类型：论文|刊号（ISSN）：2630-4732 /（中图刊号）：561GL001

①计算决策树 $h_t(x)$ 对类别C的误分率 ε_t^C：

$$\varepsilon_t^C = \sum_{i=1}^{n} w_i^t \bullet sign\left(y_i = l, h_t(x_i) \neq y_i\right) \tag{14}$$

②若 $\varepsilon_t^C \leq 1 - \frac{1}{l}$，跳过次决策树，结束此次循环。

③若 $\varepsilon_t^C > 1 - \frac{1}{l}$，计算 $h_t(x)$ 对C类的权重 a_t^C：

$$a_t^C = \frac{1}{2}\log\frac{1-\varepsilon_t^C}{\varepsilon_t^C} \tag{15}$$

④更新样本权重 w_i^{t+1}：

$$w_i^{t+1} = \frac{w_i^t}{Z_t}\exp\left(a_t^C \bullet \text{sign}\left(y_i = l, h_t(x_i) \neq y_i\right)\right) \tag{16}$$

式中，Z_t 是标准化因子：

$$Z_t = \sum_{i=1}^{n} w_i^t \bullet \exp(a_t^C \bullet \text{sign}(y_i = l, h_t(x_i) \neq y_i)) \tag{17}$$

⑤对t=1:T循环步骤(a)到步骤(d)。

最终加权随机森林的分类器预测投票模型为：

$$H_y(x) = \arg\max_Y \sum_{t=1}^{k} a_t^y \bullet \text{sign}(h_t(x) = Y) \tag{18}$$

式中，a_t^y 为不同决策树对不同地物类别的投票权重，$h_t(x)$ 为不同决策树预测结果 $\text{sign}(\bullet)$，取值为-1或1。

1.3建筑物点云后处理

1.3.1 TIN三角网优化建筑物

(1)对初步检测的建筑物脚点构建Delaunay三角网[14]。

(2)对所有三角形进行判断，若三角形中至少有一条边大于阈 T_m 值，删除整个三角形，图3(a)、(b)分别为原始Delaunay三角网和优化后的Delaunay三角网。

(3)确定每个激光脚点的邻接三角形的个数，若大于2，则认为该点为建筑物脚点；反之，若小于或者等于2，则认为该点为非建筑物脚点，将其剔除。

(4)重复步骤(1)~(3)，直到没有新的非建筑物脚点，迭代终止，得到最终的建筑物点群。

(a)原始Delaunay三角网　　(b) 优化后Delaunay三角网

图3　Delaunay三角网

1.3.2建筑物单体分割

(1)首先将提取到的建筑物点云进行栅格化处理，将其投影到二维虚拟格网里，格网间距如下公式：

$$d = \sqrt{n \bullet \frac{S}{N}} \tag{19}$$

式中，d表示格网间距大小；n表示网格内激光脚点个数；S表示数据区域总面积；N表示数据中总的点数。

(2)基于格网八邻域连通性对建筑物点云进行区域分割。

设定建筑物最小面积阈 T_S 值。

(3)对所有的相互独立的小区域分颜色显示，得到最终单体化分割的建筑物激光脚点。

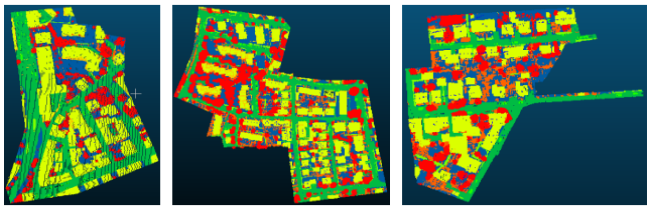
2 实验与分析

2.1实验数据

实验选取国际摄影测量与遥感学会(ISPRS)提供用于城市分类的测试数据集[15]。测试区域正射影像如图4所示。试验场地(a)位于瓦伊根市中心，具有密集、复杂的高大建筑物和树木。该标准数据集中包含179 997个点。试验场地(b)中地面比较平坦，建筑物顶面形态各异，并且其中的一些建筑物周围有很多树木，会有遮挡的情况。该测试数据集中包含753 876个激光脚点。试验场地(c)地形起伏比较大，建筑物与树木紧邻，树木与灌木林比较多，该数据集共有激光脚点231 725个。实验中将原始点云分为两类，一类是建筑物，将其他类别统一归算为非建筑物点云，上述三组测区标准建筑物点云数据如图6所示。

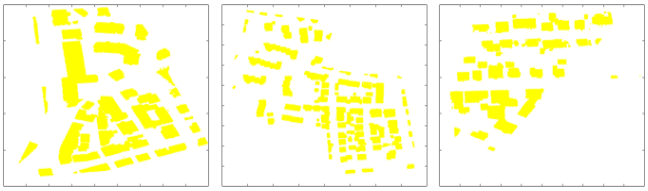
(a)Test1　　(b) Test2　　(c) Test3

图4　测试区域正射影像图



(a)Test1　　(b) Test2　　(c) Test3

图5　测试区域激光点云俯瞰图



(a)Test1　　(b) Test2　　(c) Test3

图6　测试区域标准建筑物点云

2.2建筑物提取

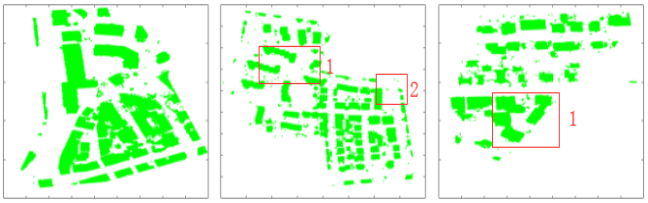
2.2.1初始建筑物检测

对粗差去除后的点云数据，利用加权随机森林算法对实验区域进行建

文章类型：论文|刊号（ISSN）：2630-4732 /（中图刊号）：561GL001

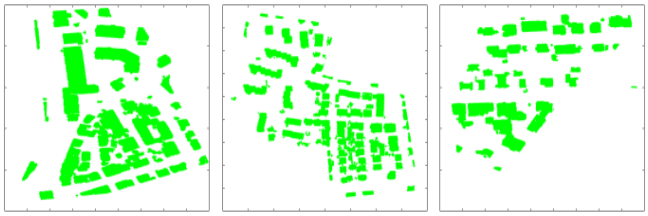
筑物检测。离散点云数据量很大，因此在提取的点云特征的时候，采取了K-D树最近邻查询算法，以提高最近邻点的搜索效率。

利用加权随机森林算法对三组测区进行二分类，得到初步的建筑物点云如图7所示。经过后处理技术优化后得到的建筑物如图8所示。最后建筑物单体化分割结果如图9所示，图中不同的颜色分别代表每栋独立的建筑物区域。



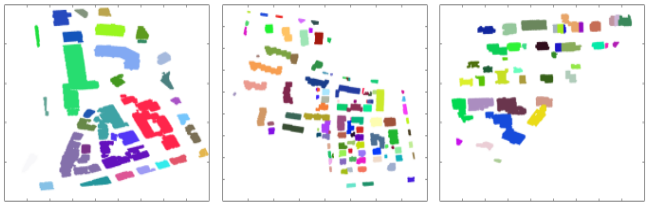
(a)Test1　　(b) Test2　　(c) Test3

图7　加权随机森林检测建筑物点云



(a)Test1　　(b) Test2　　(c) Test3

图8　TIN三角网优化建筑物点云



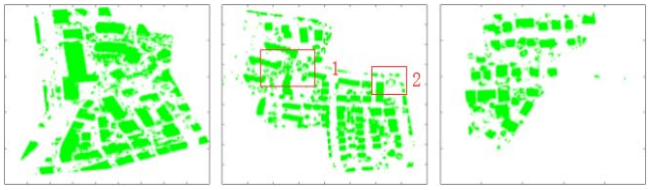
(a)Test1　　(b) Test2　　(c) Test3

图9　建筑物点云单体化分割

2.2.2对比分析

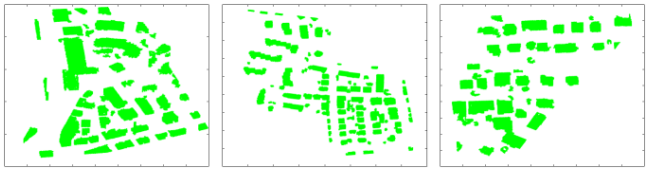
将本文算法与传统的随机森林算法进行对比。

利用传统的随机森林算法初步检测建筑物点云如图10所示，应用TIN三角网对建筑物点云进行优化处理，结果如图11所示，建筑物单体化分割结果如图12所示。



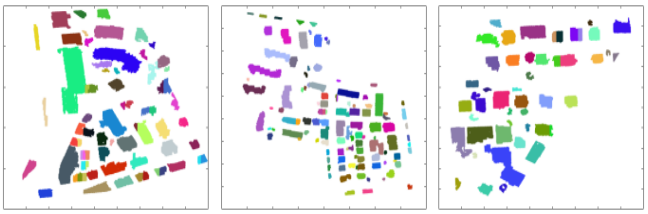
(a)Test1　　(b) Test2　　(c) Test3

图10　传统随机森林算法检测建筑物点云



(a)Test1　　(b) Test2　　(c) Test3

图11　传统随机森林算法检测最终建筑物点云



(a)Test1　　(b) Test2　　(c) Test3

图12　传统随机森林算法建筑物点云单体化分割

2.2.3精度评定

本文建立误差混淆矩阵方法对建筑物提取结果进行的定量评价，如表1所示，提取第Ⅰ类误差、第II类误差、总误差以及Kappa系数四个指标对结果进行定量分析。

表1中S_{11}表示正确提取的建筑物点数目，S_{12}表示误提取的建筑物点数目，S_{21}表示误提取的非建筑物点数目，S_{22}表示正确提取的非建筑物点数目，S_{1+}和S_{2+}分别表示参考数据中建筑物点与非建筑物点数目，S_{+1}和S_{+2}分别表示算法提取出的建筑物点与非建筑物点数目，N为原始点云总数目。

基于混淆矩阵，计算第Ⅰ类误差、第II类误差、总误差以及Kappa系数计算公式如下，

①第I类误差 δ_1：

$$\delta_1 = \frac{s_{12}}{s_{1+}} \times 100\% \tag{20}$$

②第II类误 δ_2：

$$\delta_2 = \frac{s_{21}}{s_{2+}} \times 100\% \tag{21}$$

③总误差 ω：

$$\omega = \frac{(s_{12}+s_{21})}{N} \times 100\% \tag{22}$$

④Kappa系数：

$$\text{Kappa} = \frac{(s_{11}+s_{22})/N-(s_{1+}\times s_{+1}+s_{2+}\times s_{+2})/N^2}{1-(s_{1+}\times s_{+1}+s_{2+}\times s_{+2})/N^2} \tag{23}$$

表1　混淆矩阵表示形式

参考结果	算法提取结果		
	建筑物	非建筑物	总和
建筑物	S11	S12	S1+
非建筑物	S21	S22	S2+
总和	S+1	S+2	N

表2　加权随机森林算法建筑物提取结果统计

测区	参考数据		加权随机森林算法					
	建筑物点数(个)	非建筑物点数(个)	建筑物点数(个)			非建筑物点数(个)		
			总点数	正确	误分	总点数	正确	误分
Test1	53 445	126 552	51 453	49 941	1 512	128 544	125 054	3 490
Test2	152 045	601 831	149 384	139 128	10 256	604 492	589 653	14 839
Test3	55 603	176 122	52 289	51 041	1 248	179 436	174 857	4 579

表3　传统随机森林算法建筑物提取结果统计

测区	参考数据		传统随机森林算法					
	建筑物点数(个)	非建筑物点数(个)	建筑物点数(个)			非建筑物点数(个)		
			总点数	正确	误分	总点数	正确	误分
Test1	53 445	126 552	51 453	49 941	1 512	128 544	125 054	3 490
Test2	152 045	601 831	134 088	122 232	11 856	619 788	592 546	27 242
Test3	55 603	176 122	47 658	43 241	4 417	184 067	171 705	12 362

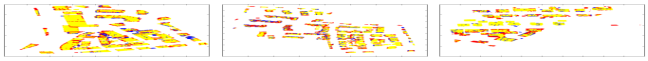
文章类型：论文|刊号（ISSN）：2630-4732 /（中图刊号）：561GL001

表4　不同算法精度对比

实验测区	误差类型	T_RF 算法	W_RF 算法
Test1	I 类误差(%)	14.01	6.53
	II 类误差(%)	4.51	1.19
	总误差(%)	7.33	2.78
	Kappa(%)	82.27	93.27
Test2	I 类误差(%)	17.92	9.76
	II 类误差(%)	2.47	1.70
	总误差(%)	5.19	3.33
	Kappa(%)	83.15	89.59
Test3	I 类误差(%)	22.23	8.23
	II 类误差(%)	2.51	0.71
	总误差(%)	7.24	0.71
	Kappa(%)	79.13	92.96

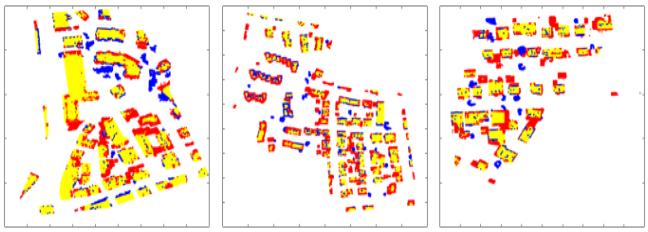
对上述传统随机森林与加权随机森林两种算法建筑物提取结果进行统计，分别如表2、表3、表4所示，通过对比发现，统随机森林算法检测建筑物精度较低。表4中统随机森林算法检测的三组测试区域的I类误差明显比较大，说明该算法对于建筑物漏提取现象比较严重，II类误差比较低，表明统随机森林算法对于植被的过滤效果比较好。本文加权随机森林算法的I类误差维持在7%左右，说明了随机森林算法对于建筑物点云检测准确性较高，而且还会比较好的保持建筑物的完整性。II类误差维持在0.71%~1.7%范围内，充分说明了随机森林算法对于非建筑物的过滤效果很好，而且在建筑物点云的优化过程中，可以将某些成簇的非建筑物点云过滤掉，降低误分率。三组测区的平均Kappa系数大致在90%，可以说明基于加权随机森林算法提取建筑物的相对准确性，通过对实验结果的统计分析，可以验证加权随机森林算法提取建筑物的可行性。

图13与图14分别列出了两种对比算法检测建筑物点云的误差统计图，图中黄色的点云表示的正确检测的建筑物点云，红色的点云表示的是I类误差，蓝色点云表示的是II类误差。通过两种算法的区域误差图的比较，会发现统随机森林算法检测结果中有很多的蓝色区域，说明了统随机森林算法将非建筑物误分为建筑物。本文加权随机森林算法提取建筑物总体上看可以将建筑物周围的低矮建筑物较好的提取出来，而且蓝色区域较少，充分的说明了本文的加权随机森林算法提取效果更好。



(a) Test1　　(b) Test2　　(c) Test3

图13　本文算法提取建筑物点云误差图



(a) Test1　　(b) Test2　　(c) Test3

图14　传统随机森林算法提取建筑物点云误差图

3 结论

本文在分析机载LiDAR数据结构与特点以及现有的机载LiDAR建筑物点云提取算法的基础上，重点介绍了对传统随机森林算法的改进，以及将其应用在LiDAR点云建筑物提取中。该算法具备以下特点：①对传统的随机森林模型投票方式进行改进，提出了加权随机森林模型；②利用加权随机森林模型对点云数据实行二分类，利用相应的后处理技术对建筑物进行优化。③加权随机森林模型提取建筑物可以解决传统先滤波后提取算法依赖滤波精度等问题，也避免了参数的不合理选择。

基金项目：

高分辨率对地观测系统重大专项(民用部分)：项目编号：(06-Y20A17-9001-17/18)。

[参考文献]

[1]李德仁.展望大数据时代的地球空间信息学[J].测绘学报,2016,45(4):379-384.

[2]赵宗泽.基于数学形态学的机载LiDAR点云建筑物区域提取[D].武汉:武汉大学,2016.

[3]Mohammad Awrangjeb, Mehdi Ravanbakhsh, Clive S, et al. Automatic detection of residential buildings using LIDAR data and multispectral imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing. 2010,(5).

[4]汪禹芹.机载LiDAR点云数据的建筑物提取和模型规范化研究[D].南京:南京大学,2013.

[5]Gorte B. Segmentation of TIN-Structured Surface Models[J]. Proceedings Joint International Symposium on Geospatial Theory Processing & Applications on Cdrom,2002.

[6]乔纪纲,陈明辉,艾彬,等.SVM用于LiDAR数据的地物分类[J].测绘通报,2013,(7):35-38.

[7]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),2014,39(11):1310-1313.

[8]熊艳,高仁强,徐战亚.机载LiDAR点云数据降维与分类的随机森林方法[J].测绘学报,2018,47(04):508-518.

[9]Guan H, Yu J, Li J, et al. Random Forests-based feature selection for land-use classification using LIDAR data and orthoimagery[J]. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2012, XXXIX-B7: 203-208.

[10]Croux C, Joossens K, Lemmens A. Trimmed bagging[J]. Computational Statistics & Data Analysis, 2007, 52(1):362-368.

[11]杨飚,尚秀伟.加权随机森林算法研究[J].微型机与应用,2016,35(03):28-30.

[12]张良.基于多时相机载LiDAR数据的三维变化检测关键技术研究[D].武汉大学,2014.

[13]王竞雪,洪绍轩.结合区域生长及主成分分析的机载LiDAR建筑物点云提取[J].信号处理,2018,34(09):1094-1104.

[14]洪绍轩,王竞雪.结合OTSU与迭代三角网的机载LiDAR建筑物点云提取[J].遥感信息,2018,33(06):79-85.

[15]李云帆,谭德宝,高广,等.双阈值Alpha-Shape算法提取点云建筑物轮廓研究[J].长江科学院院报,2016,33(11):1-4.

作者简介：

洪绍轩(1993--),男,辽宁省鞍山市人,硕士,算法工程师,研究方向：机载LiDAR点云分类与建模,从事工作：遥感研发。