DOI:10.19651/j.cnki.emt.2312627

# 基于特征重要性加权的随机森林点云分类研究*

吴 冬 阎卫东 王井利
(沈阳建筑大学土木工程学院 沈阳 110168)

**摘 要**:针对传统的随机森林模型构建时样本选取的随机性导致随机森林中包含了大量分类精度较低、分类性能相似的决策树分类器,进而影响整体随机森林模型分类精度与效率的问题,该文提出了一种基于特征重要性加权投票的随机森林算法。从决策树分类精度、不一致度量两方面剔除分类精度较低、分类性能相似的决策树,依据整体随机森林与单棵决策树特征重要性之间的相似性,计算每棵决策树的投票权重,提高了三维点云分类精度与分类效率。实验表明,改进后的随机森林分类算法照比传统的随机森林、支持向量机、决策树、神经网络、基于点特征分类方法分别提高了 0.20%、15.159%、5.893%、6.316%、28.935%。在分类效率上,改进的随机森林照比传统的随机森林减少了约 75%的时间。

**关键词**:点云分类;随机森林;不一致度量;特征重要性;加权投票
**中图分类号**:TP391 **文献标识码**:A **国家标准学科分类代码**:510.4

# Random forest point cloud classification algorithm based on feature importance weighting

Wu Dong Yan Weidong Wang Jingli
(School of Civil Engineering, Shenyang Jianzhu University, Shenyang 110168, China)

**Abstract**: The randomness of sample selection during the construction of traditional random forest model leads to a large number of decision tree classifiers with low classification accuracy and similar classification performance in random forest, which affects the accuracy and efficiency of the overall random forest model classification. In order to improve the accuracy and efficiency of random forest model in point cloud classification, a random forest algorithm based on feature importance weighted voting was proposed. Firstly, decision trees with low classification accuracy and similar classification performance are eliminated from the aspects of classification accuracy and inconsistency measurement of decision trees. Secondly, the voting weight of each decision tree is calculated based on the similarity between random forest and decision tree feature importance. In this paper, three sets of densely matched point clouds are taken as examples to compare the improved stochastic forest classification model with the traditional stochastic forest, support vector machine classifier (SVM), neural network and decision tree. The experiments show that the improved random forest classification algorithm is 0.20%, 15.159%, 5.893%, 6.316% and 28.935% higher than the traditional random forest, support vector machine, decision tree, neural network and point-based feature classification method, respectively. In terms of classification efficiency, the improved random forest classification algorithm takes about 75% less time than the traditional random forest.

**Keywords**: point cloud classification; random forest; inconsistency measure; feature importance; weighted voting

## 0 引 言

密集匹配点云因具有丰富的色彩信息、三维信息而被广泛关注[1],在进行密集匹配点云数据处理方面,点云分类一直是研究的热点。然而点云的稀疏性、多噪点、分布不均匀等缺陷给点云分类带来了挑战。

国内外专家学者对点云分类的研究主要集中于三个方面:基于点特征的分类[2]、基于深度学习的分类和基于传统机器学习的分类[3]。在基于点特征的分类方面,Aijazi 等[4]提出将点云多种特征融合成超级体素,基于体素信息完成

收稿日期:2023-01-13
* 基金项目:辽宁省科技厅项目(2021JH2/10100005)资助

点云分类;Strimbu 等[5]基于点云高程信息提出了一种低矮植被树冠分类策略,较好地实现了对树冠点云的分割。此类基于点特征的分类方式要求研究人员能够准确找出各类地物点云的特征,对点云特征提取要求较高。在深度学习点云分类方面,PointNet 网络[6]的提出使点云分类技术得到了重大革新,该网络也成为了研究人员的研究热点;随后在 PointNet 网络的基础上 Qi 等[7]团队提出了多层次特征提取结构形成了 PointNet++网络,使得点云分类的精度得到大幅度提升。

基于传统机器学习的分类方法主要有:支持向量机[8]、随机森林[9]等。Park 等[10]利用随机森林模型实现了建筑物屋顶、墙壁、地面点云的有效分割,但受到点云高程异常的影响存在着部分误分类点云;Huan 等[11]将点云进行了聚类处理,然后利用随机森林模型对聚类后的点云进行重分类,该方法虽然提高了分类精度但运算量大,时间成本高。针对随机森林模型预测时间长的问题,Xue 等[12]提出了一种基于改进的布料模拟滤波(improved cloth simulation filterin, ICSF)和弱相关随机森林的点云分类算法,基于最大互信息系数进行决策树筛选,改进后的随机森林缩短了建模时间。胡海瑛等[13]针对分类精度低提出利用多基元特征训练随机森林模型,该方法主要依靠增加点云特征来提高其分类精度,但特征数量的增多使得建模速度降低。Sun 等[14]从随机森林训练的特征子集出发对随机森林模型进行改进,实现了密集匹配点云的高精度分类,但其未对模型的效率进行改进。

现有的基于随机森林分类方法存在误分类、时间成本高等缺点。本文提出了一种基于特征重要性加权投票的随机森林分类算法,首先针对分类精度低的问题,根据随机森林中单棵决策树分类精度对较低分类精度的决策树进行剔除;其次依据不一致度量标准对分类性能相似的决策树进行剔除;最后改变传统随机森林中的分类结果投票规则,依据整体随机森林模型与单棵决策树特征重要性之间的相似性,计算每棵决策树的投票权重,对分类结果进行加权投票。该方法在提高点云分类精度的同时提高点云分类效率。

## 1　点云特征提取

密集匹配点云与机载 Lidar 点云相比,具有良好的色彩信息、地物纹理信息,可据此进行地物的判读,有利于后期的点云分类处理。本文主要从点云的颜色特征、邻域特征以及与高程有关的信息特征三方面进行点云特征提取。

### 1.1　点云颜色特征提取

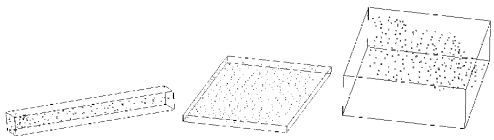
密集匹配点云所具有的颜色信息使得不同类型的地物表现出较为明显的色彩差异,所以本文将点云的颜色的红(Red,R)绿(Green,G)蓝(Blue,B)3 个通道的信息作为输入特征。然而不同的光照条件使得地物的颜色特征发生变化,因此考虑将点云的颜色特征从 RGB 空间转换到 HSV(Hue, Saturation, Value)空间,该颜色空间下的 $H$ 分量可以有效地避免光照变化的影响[13]。对于树木、植被等地物,其区别于其它地物的一个重要特征是 RGB 色彩空间中绿色($G$)通道的分量较大,故本文引入点云绿度 $Gr$ 来区分植被[15]其表达式如式(1)所示。

$$Gr = G/(R+G+B) \tag{1}$$

式中:$R$、$G$、$B$ 分别代表红、绿、蓝 3 个通道分量值。

### 1.2　邻域特征

邻域特征是指对点云局部信息的提取。不同地物点的邻近点云所呈现的形态存在着明显差异。如图 1 所示,线状地物(如电线杆、电力线等)邻近点云的最小外接长方体呈现长条状,如图 1(a)所示;面状地物(如建筑物立面、屋顶面)在一定的邻域范围内邻近点的最小外接长方体呈现扁平状,如图 1(b)所示;对于杂乱(如树木)点云在一定的邻域范围内邻近点的最小外接长方体近似立方体,如图 1(c)所示。



(a) 线状点云　(b) 面状点云　(c) 杂乱点云

图 1　不同形态点云空间分布特征

本文首先利用 kd-tree 对预处理后的点云建立起拓扑关系。其次,提取点云中离散点 $P$ 的球体半径 $R$(本研究中 $R=5$ cm)内的所有近邻点 $\Omega=\{p_1、p_2、\cdots、p_i、\cdots、p_k\}$,通过研究近邻点构建的协方差张量 $C_x$ 确定离散点 $P$ 的线性、平面性以及立体性结构特征。

离散点 $P$ 的近邻点 $\Omega=\{p_1、p_2、\cdots、p_i、\cdots、p_k\}$ 构建的邻近点协方差张量 $C_x$ 为:

$$C_x = \frac{1}{k}\sum_{i=1}^{k}(p_i-\overline{p})(p_i-\overline{p})^{\mathrm{T}} \tag{2}$$

式中:$\overline{p}$ 为 $k$ 个邻域点的中心点,其表达式为:

$$\overline{p} = \mathrm{argmin}_p \sum_{i=1}^{k} \| pi - \sum_{i=1}^{k} pi \| \tag{3}$$

由协方差张量可计算得到其 3 个特征值 $\lambda_1>\lambda_2>\lambda_3>0$,为了消除点云密度不均匀对分类精度的影响,对计算得到的特征值进行归一化处理使得 $\lambda_1+\lambda_2+\lambda_3=1$。

离散点 $P$ 的线性 $Linearity$、平面性 $Planarity$ 以及立体性 $Scatter$ 结构特征可分别表示为:

$$\begin{cases} Linearity = (\lambda_1-\lambda_2)/\lambda_1 \\ Planarity = (\lambda_2-\lambda_3)/\lambda_1 \\ Scatter = \lambda_3/\lambda_1 \end{cases} \tag{4}$$

### 1.3　与高程有关的信息特征

在同一场景中,不同的地物间存在着高程差异,点云高程可以作为地物分类的重要指标。不同地物点云的邻域高程呈现不同的特征,规则地物与不规则地物高程标准差和偏度也存在着差异。文中以离散点 $P$ 邻域内的平均高程

$\bar{h}$、高程标准差 $\sigma_h$ 以及高程偏度 $sk_h$ 作为点云的高程信息特征[16]。

$$\begin{cases} \bar{h} = \dfrac{1}{k}\sum_{i=1}^{k} h_i \\ \sigma_h = \sqrt{\dfrac{1}{k}\sum_{i-1}^{k}(h_i - \bar{h})^2} \\ sk_h = \dfrac{\sum_{i=1}^{k}(h_i - \bar{h})^4}{[\sum_{i=1}^{k}(h_i - \bar{h})^2]^2 - 3} \end{cases} \tag{5}$$

式中：$k$ 表示点 $P$ 在球半径 $R$（本研究中 $R=5$ cm）内的邻域点数量，$h_i$ 表示点 $P$ 邻域内第 $i$ 个点的高程值。

## 2　基于特征重要性加权的随机森林点云分类

随机森林（random forest，RF）算法[17]是由多棵决策树组成的一种分类器。该算法首先用 bootstrap 抽样方法生成多个训练数据集，每个训练集通过训练得到一棵决策树。最后输入测试样本，对所有决策树预测得到的类别归属进行投票，得到最终预测结果。

### 2.1　决策树选取

由于随机森林中构建与训练决策树所用到的子训练集和数据集特征是随机选取产生的，所以最后生成的随机森林中可能会出现分类性能较差的决策树分类器，此类决策树分类器的存在将影响最终随机森林模型预测的精度[18]。为了提高随机森林模型整体的分类精度，王诚[19]根据决策树的 $AUC$ 值对决策树进行筛选，剔除分类性能较差的决策树。评价决策树分类性能的参数除了 $AUC$ 值外，还包括分类精度（classification accuracy）、精确率（precision）、召回率（recall）等，其中分类精度是评价分类器分类性能的最常用的指标。本文以分类精度作为分类器性能衡量指标，计算传统随机森林中每棵决策树 $\{T_1, T_2, \cdots, T_n\}$ 的分类精度 $P_i (i = 1, 2, \cdots n)$，并据此对决策树进行排序，同时剔除分类精度低于平均分类精度的决策树，得到高精度的决策树 $\{T_1, T_2, \cdots, T_p\}$，$(p<n)$。

单棵决策树分类模型比多棵决策树集成后的随机森林来说，其分类性能及泛化能力较差。但如果集成分类模型中含有多个分类性能相似或相近的分类器，则整个继承分类模型的泛化能力也会有所降低[18]。为此，本文对随机森林中决策树进行多样性度量。其主要的衡量指标包括 $Q$ 统计值、相关系数、不一致度量 $dis$、双次失败度量 $DF$[20]。通过实验对比分析，发现不一致度量 $dis$ 在不同的决策树之间的差异性较为明显，因此笔者选用了不一致度量 $dis$ 作为决策树多样性衡量指标。

假设有 $L$ 个分类器，$Ci$ 和 $Cj (i, j = 1, 2, \cdots, L, i \neq j)$ 分别为两个不同的分类器，$N^{11}$（$N^{00}$）为分类器 $Ci$ 和 $Cj$ 都对其正确（错误）分类的样例数目，$N^{10}$（$N^{01}$）为满足以下要求的样例数目：分类器 $Ci$（$Cj$）对其正确分类而分类器 $Cj$（$Ci$）对其错误分类，由此总的样例数目 $N$ 可以表示式(6)：

$$N = N^{11} + N^{10} + N^{01} + N^{00} \tag{6}$$

不一致度量关注两个分类器 $Ci$ 和 $Cj$ 分类结果不同的样例，他们之间的不一致度量 $dis$ 定义为：

$$dis_{ij} = (N^{10} + N^{01})/N \tag{7}$$

两个分类器的不一致度量 $dis$ 的取值范围是 $[0, 1]$。不一致度量 $dis$ 数值越大，两个分类器之间的多样性程度就会越高；反之的话多样性程度就会越低[21]。

利用式(7)计算出的随机森林中任意两棵决策树间的不一致度量 $dis_{ij}$，组成不一致度量矩阵 $\boldsymbol{dis}$，如式(8)所示。

$$\boldsymbol{dis} = \begin{bmatrix} dis_{T1,T1} & dis_{T1,T2} & \cdots & dis_{T1,Tp} \\ dis_{T2,T1} & dis_{T2,T2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ dis_{Tp,T1} & \cdots & \cdots & dis_{Tp,Tp} \end{bmatrix} \tag{8}$$

在计算出随机森林不一致度量矩阵后，通过设定不一致度量阈值 $dis_{Thresh}$（本文设为 0.65），将 $dis_{Ti,Tj} < dis_{Thresh}$ 对应的决策树 $Ti$，$Tj$ 归为一类，最终将选取的 $p$ 棵决策树 $\{T_1, T_2, \cdots, T_p\}$，$(p<n)$ 分为 $q$ 个类。

最后，从每个类别中挑选出分类精度最高的决策树，组成一个含有 $q$ 个决策树的随机森林分类器 $\{T_1, T_2, \cdots, T_q\}$，$(q<p<n)$。改进后随机森林构建的流程伪代码如算法 1 所示。

**算法 1　改进后随机森林生成伪代码**

算法描述——决策树选取

Input：$D$ 数据集
　　$F$ 特征集
　　$K$ 传统随机模型包含的决策树棵数
　　$dis_{Thresh}$ 不一致度量阈值

Output：决策树筛选后的随机森林

算法：1）利用 Bagging 算法将数据集 $D$ 分为训练集与测试集；

2）从训练集中使用 Bootstrapping 方法选取 $k$ 个子样本集，从特征集 $F$ 中随机选取子样本集的特征，并分别训练出 $k$ 个决策树模型；

3）利用测试数据集对生成的每棵决策树进行分类精度评定，选取分类精度优于平均分类精度的 $p$ 棵决策树；

4）for $i=1$ to $p$
　　for $j=1$ to $p$
　　按公式计算决策树 $Ti$ 与 $Tj$ 之间的不一致度量 $dis_{ij}$

5）根据不一致度量阈值 $dis_{Thresh}$ 将 $p$ 棵决策树分为 $q$ 类，并选取每个类中分类精度最高的决策树组成新的随机森林。

### 2.2　基于特征重要性的加权投票

随机森林中,特征重要性是衡量每个特征对模型分类过程中的贡献值大小。特征重要性评估主要分为基于基尼指数、基于袋外数据两种。本文基于各个决策树的特征重要性与随机森林总体特征重要性的相关程度提出了一种加权投票方式。若决策树 $Ti$ 与随机森林的特征重要性相关程度越高,则该决策树 $Ti$ 的权重值越大。

首先,基于袋外数据计算上一小节构建出的随机森林的特征重要性;基于袋外数据计算随机森林中每棵决策树的特征重要性;

其次,分别绘制出随机森林和每棵决策树的特征重要性直方图 $H$、$Hi(i=1,2,\cdots,q)$;

图 2、3 为随机森林以及随机森林中第 $i$ 棵决策树的特征重要性直方图,其横轴 0～12 表示第 1 节描述的样本的 13 个不同特征。

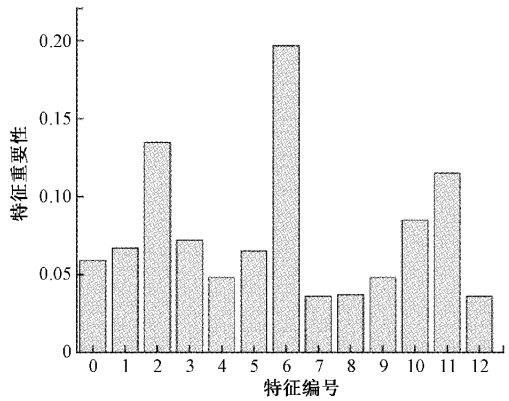



图 2　随机森林特征重要性直方图

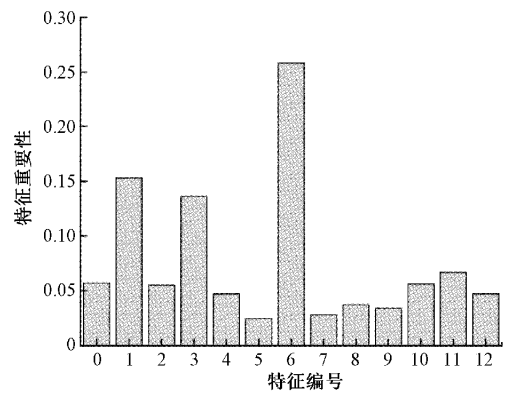



图 3　第 $i$ 棵决策树的特征重要性直方图

接着,依据直方图相关性计算方法计算出每一棵决策树与随机森林间的相关系数 $d_{(H,Hi)}$,$(i=1,2,\cdots,q)$,其表达式如式(9)所示。

$$d_{(H,H_i)}=\frac{\sum_I(H_1(I)-\overline{H_1})(H_2(I)-\overline{H_2})}{\sqrt{\sum_I(H_1(I)-\overline{H_1})^2\sum_I(H_2(I)-\overline{H_2})^2}} \tag{9}$$

式中:$\overline{H_k}=\frac{1}{N}\sum_I H_k(I)$,$N$ 为直方图中 bin 的数目。

特征重要性直方图的相关系数 $d_{(H,Hi)}$ 取值范围为 $[0,+1]$,相关系数越趋近于 1,表明该决策树与随机森林的特征重要性越相似,故在投票决策时该决策树所占的权重也相应增大。

最后,将计算出来的相关系数作为决策树的权重值,依据权重值对分类结果进行预测。引入特征重要性加权投票的随机森林模型 $\{T_1,T_2,\cdots,T_q\}$ 的投票结果最终表示为式(10)[22]:

$$\max\left\{c\mid ci=\sum_{j-1}^{k}d_{(H,Hi)}*I(Tj=l),l\in C\right\} \tag{10}$$

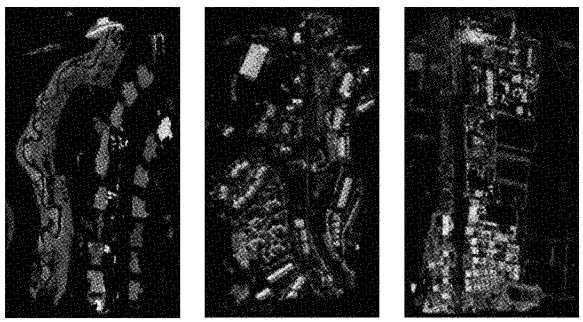
式中:$C$ 为所有分类标签的集合,$I(*)$为示性函数。

利用式(10)统计计算出所有类别标签加权投票后的数量,将数量最多的类别标签作为最终改进后模型的预测结果进行输出。

## 3　实验与结果分析

### 3.1　数据集

为验证本文算法的精度及预测效率,选用了三组数据集进行实验,分别是来自 Pix 4d 官网(https://www.pix4d.com/)的摄影测量点云 ankeny 和 cadastre 数据集以及来自 DPCloud 官网(http://www.dpgrid.com/)的昊龙训练数据。其中昊龙数据集为无人机影像数据,研究中通过空中三角测量、多视影像密集匹配、纹理映射等将该影像数据集构建成三维点云模型。研究中分类对象为非地面点云,所以在分类前采用布料模拟滤波算法[23]对三组数据集进行滤波处理,处理后得到的非地面点如图 4 所示。



(a) Ankeny数据集　(b) Cadastre数据集　(c) 昊龙数据集

图 4　滤波后的非地面点云

### 3.2　精度分析

随机选取每一组非地面点云的 1/3 数据作为模型的训练与验证数据,其中训练集、验证集、测试集的比例为 8∶1∶1。由于研究中缺少分类真值,故对该模型训练使用到的数据进行手工分类,共计分为 5 类,建筑物、电力线、树木、车辆、其它。

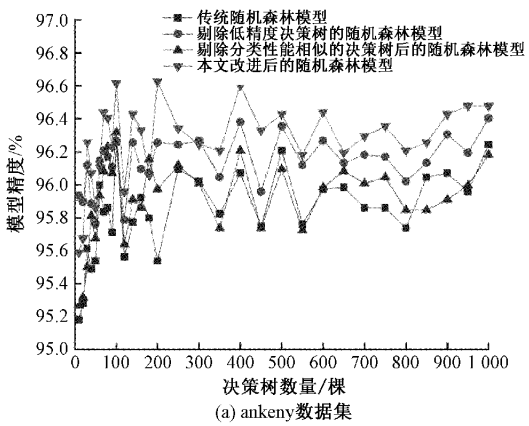
本文剔除了传统随机森林中低精度决策树以及分类性能相似的决策树,为验证剔除低精度决策树和分类性能相

似的决策树对模型精度和分类效率都有提高作用，研究中将改进后的随机森林分类模型(ImproveRF)与传统的随机森林模型(RF)、剔除低精度决策树的随机森林(记为 RF1)以及基于不一致度量筛选分类性能相似的决策树后的随机森林(记为 RF2)进行精度和分类效率的比较分析。各分类模型在不同数量决策树($n\_tree$)下的模型精度如表 1。

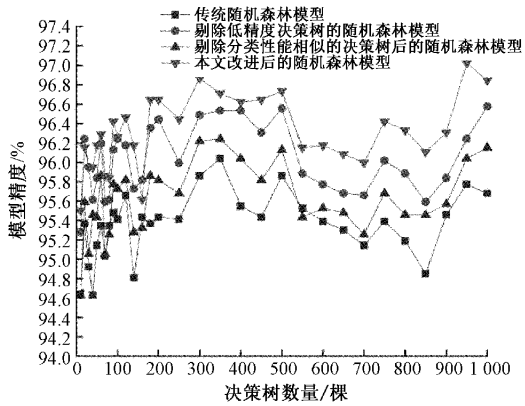
**表 1　精度对比结果**　　%

| $n\_tree$ | ankeny 数据集 | | | | cadastre 数据集 | | | | 昊龙数据集 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RF | RF1 | RF2 | ImproveRF | RF | RF1 | RF2 | ImproveRF | RF | RF1 | RF2 | ImproveRF |
| 10 | 95.180 | 95.934 | 95.266 | 95.588 | 94.652 | 95.279 | 94.630 | 95.502 | 97.249 | 97.945 | 97.551 | 97.912 |
| 50 | 95.538 | 95.761 | 95.674 | 95.860 | 95.144 | 95.838 | 95.435 | 96.174 | 98.180 | 98.499 | 98.197 | 98.465 |
| 100 | 96.255 | 96.267 | 96.317 | 96.614 | 95.413 | 96.241 | 95.726 | 96.263 | 98.390 | 98.633 | 98.390 | 98.633 |
| 200 | 95.538 | 96.255 | 95.971 | 96.626 | 95.435 | 96.442 | 95.816 | 96.644 | 98.524 | 98.725 | 98.490 | 98.817 |
| 300 | 96.020 | 96.267 | 96.008 | 96.243 | 95.860 | 96.487 | 96.218 | 96.867 | 98.205 | 98.415 | 98.264 | 98.557 |
| 400 | 96.070 | 96.379 | 96.206 | 96.601 | 95.547 | 96.532 | 96.039 | 96.621 | 98.381 | 98.507 | 98.432 | 98.650 |
| 500 | 96.206 | 96.354 | 96.094 | 96.428 | 95.860 | 96.554 | 96.129 | 96.733 | 98.515 | 98.725 | 98.641 | 98.851 |
| 600 | 95.971 | 96.267 | 95.983 | 96.440 | 95.390 | 95.771 | 95.525 | 96.174 | 98.641 | 98.826 | 98.759 | 98.952 |
| 700 | 95.860 | 96.181 | 96.008 | 96.292 | 95.144 | 95.659 | 95.256 | 95.995 | 98.625 | 98.692 | 98.557 | 98.767 |
| 800 | 95.736 | 96.020 | 95.847 | 96.206 | 95.189 | 95.883 | 95.458 | 96.330 | 98.549 | 98.767 | 98.717 | 98.868 |
| 900 | 96.070 | 96.305 | 95.909 | 96.428 | 95.458 | 95.838 | 95.569 | 96.308 | 98.641 | 98.683 | 98.524 | 98.801 |
| 1 000 | 96.243 | 96.403 | 96.181 | 96.478 | 95.681 | 96.576 | 96.151 | 96.845 | 98.306 | 98.499 | 98.390 | 98.591 |

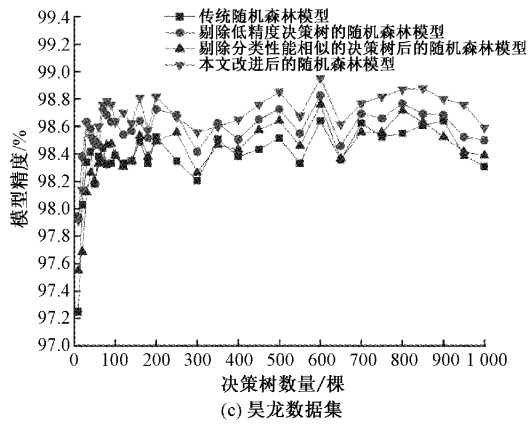
图 5(a)～(c)分别表示不同分类模型在 3 组数据集上的分类精度随决策树棵树的变化趋势。从图 5(a)～(c)中可以看出随着决策树棵树的增多，本文改进的随机森林模型(ImproveRF)分类精度明显高于其他 3 种模型。剔除低精度决策树的随机森林(RF1)模型的分类精度高于传统的随机森林模型(RF)，说明剔除低精度决策树对整个模型精度的提高起到了促进作用。而基于不一致度量筛选出分类性能相似的决策树的主要目的在于减少模型中决策树的棵树，提高模型效率，从图 5(a)～(c)中可以看出 3 组数据集中剔除分类性能相似的决策树后的随机森林(RF2)模型分类精度与传统随机森林模型分类精度相差较小。利用 3 组数据集对最终改进后的随机森林模型精度验证，最终模型精度比传统的随机森林高出 0.2%左右。



(a) ankeny数据集



(b) cadastre数据集



(c) 昊龙数据集

图 5　不同分类模型精度随决策树棵数变化

此外,笔者还比较了支持向量机(support vector machine,SVM)、决策树、神经网络(PointNet 网络为例)等不同分类方法与本文改进后的随机森林模型($n\_tree=500$)的分类精度。本文算法和支持向量机、决策树的训练样本为单个离散点,其训练样本为 64 510 个离散点,验证数据集数量为 7 036 个,测试样本个数为 7 512 个。PointNet 网络的训练样本为独立地物的点云模型,本文所用的研究数据中没有足够多的单体地物点云支撑 PointNet 模型的训练,故进行了地物点云模型增强等操作增加样本数量。最终 PointNet 网络训练样本数据集为 1 034 个点云模型(包括建筑物、树木、植被、车辆、电力线、其他),验证数据集数量为 200 个,测试样本个数为 200 个。

不同分类方法分类精度对比结果如表 2。从表中可以看出本文算法精度明显优于其他几种分类方法,3 组数据集的平均精度相比于支持向量机、决策树、PointNet 网络、基于点特征分类以及文献[12]的方法分别提高了 15.159%、5.893%、6.316%、28.935%、1.168%。在此实验中,PointNet 网络分类精度较低的原因可能是受到样本量的影响,而本文方法利用单个点的相关特征进行训练,样本量较大,训练得到的模型分类精度较高。由此可以看出在利用相同数据集对本文算法和 PointNet 网络进行训练,本文算法在精度上优势较为明显。

**表 2　分类精度对比**　　%

| 数据集 | 本文算法/% | SVM | | 决策树 | | 神经网络(PointNet) | | 基于点特征分类 | | 文献[12]方法 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 精度 | 提高率 | 精度 | 提高率 | 精度 | 提高率 | 精度 | 提高率 | 精度 | 提高率 |
| Ankeny | 96.034 | 86.312 | 9.722 | 90.583 | 5.451 | 90.942 | 5.092 | 70.431 | 25.603 | 94.942 | 1.092 |
| Cadastre | 96.172 | 72.953 | 23.219 | 91.537 | 4.635 | 83.921 | 12.251 | 64.749 | 31.423 | 96.325 | −0.153 |
| 吴龙 | 97.103 | 84.566 | 12.537 | 89.511 | 7.592 | 95.499 | 1.604 | 67.325 | 29.778 | 94.537 | 2.566 |
| 平均值 | 96.436 | 81.277 | 15.159 | 90.544 | 5.893 | 90.121 | 6.316 | 67.502 | 28.935 | 95.268 | 1.168 |

几种分类算法中基于点特征的分类方法较差,其原因在于该方法主要利用点云中各个离散点的特征阈值进行分类,而确定阈值较为困难,故引起分类精度的降低。文献[12]的方法在 Cadastre 数据集上分类精度略优于本文算法,在 Ankeny 数据集和吴龙数据集上本文方法分类精度更高。其原因在于文献[12]中所提方法仅保留了相关系数最小、精度最高的决策树,整体模型的精度受到少数几棵决策树的影响。当样本量过小或者训练不充分时模型的精度将会受到较大影响。而本文算法仅剔除了分类精度低于平均精度以及少量分类性能相似的决策树,改进后的随机森林模型中的决策树仍具备多样性,在提高分类效率的同时提升其分类精度。

图 6～8 为几种主流分类方法对 3 组数据集进行分类得到的分类结果。

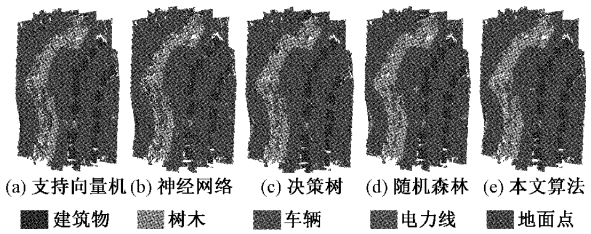



图 6　Ankeny 数据集分类结果

### 3.3　模型效率分析

为验证本文算法的高效性,笔者利用本文改进后的随机森林模型(ImproveRF)与传统的随机森林模型(RF)及剔除低精度决策树的随机森林(RF1)对待分类点云进行预测,表 3 为本文改进后的随机森林模型(ImproveRF)与传统的随机森林模型(RF)在不同决策树数量下对 3 组数据进行预测所花费的时间对比结果。

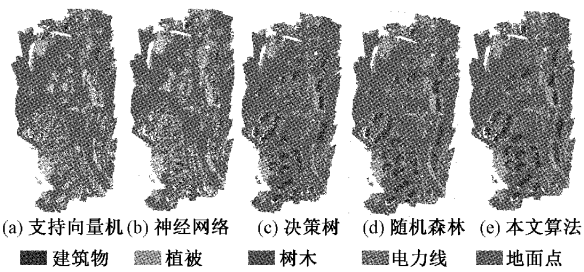



图 7　Cadastre 数据集分类结果

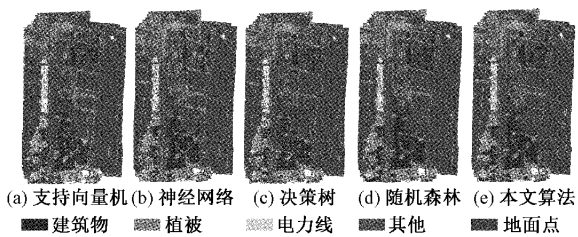



图 8　吴龙数据集分类结果

图 9 为 3 种分类模型对 3 组数据集进行预测所用时间的对比图,从图 9 中可以明显看出最终改进的随机森林模型照比其他两种模型在分类预测时间上明显缩短。图 10 为最终改进的随机森林模型与传统随机森林模型相比,预测时间缩短的比率 $rate$:

$$rate=(time_{RF}-time_{ImproveRF})/time_{RF} \tag{11}$$

式中:$time_{RF}$ 表示传统随机森林模型预测花费的时间,$time_{ImproveRF}$ 表示最终改进的随机森林模型预测花费的时间。

**表 3　改进后随机森林模型时间对比**

| *n_tree* | ankeny 数据集 | | | cadastre 数据集 | | | 昊龙数据集 | | |
|---|---|---|---|---|---|---|---|---|---|
| | RF/s | ImproveRF/s | 缩减率/% | RF/s | ImproveRF/s | 缩减率/% | RF/s | ImproveRF/s | 缩减率/% |
| 10 | 0.250 | 0.191 | 23.5 | 0.238 | 0.134 | 43.7 | 0.479 | 0.287 | 40.1 |
| 50 | 1.122 | 0.573 | 48.9 | 0.937 | 0.453 | 51.7 | 1.942 | 0.626 | 67.7 |
| 100 | 2.142 | 1.040 | 51.4 | 2.281 | 0.805 | 64.7 | 3.403 | 1.041 | 69.4 |
| 200 | 4.027 | 1.379 | 65.7 | 3.963 | 1.140 | 71.2 | 8.581 | 2.092 | 75.6 |
| 250 | 4.983 | 1.629 | 67.3 | 5.142 | 1.412 | 72.5 | 9.219 | 2.226 | 75.9 |
| 300 | 7.022 | 1.789 | 74.5 | 5.903 | 1.531 | 74.1 | 10.932 | 2.708 | 75.2 |
| 350 | 6.965 | 2.015 | 71.1 | 6.578 | 1.833 | 72.1 | 15.059 | 3.167 | 79.0 |
| 400 | 9.172 | 2.251 | 75.5 | 7.748 | 2.093 | 73.0 | 14.522 | 3.530 | 75.7 |
| 450 | 9.899 | 2.629 | 73.4 | 8.396 | 2.189 | 73.9 | 16.468 | 3.864 | 76.5 |
| 500 | 10.596 | 2.738 | 74.2 | 9.729 | 2.559 | 73.7 | 18.634 | 4.429 | 76.2 |
| 750 | 21.095 | 5.737 | 72.8 | 13.750 | 3.338 | 75.7 | 31.462 | 6.701 | 78.7 |
| 1 000 | 23.305 | 6.325 | 72.9 | 17.290 | 4.141 | 76.1 | 36.072 | 8.052 | 77.7 |

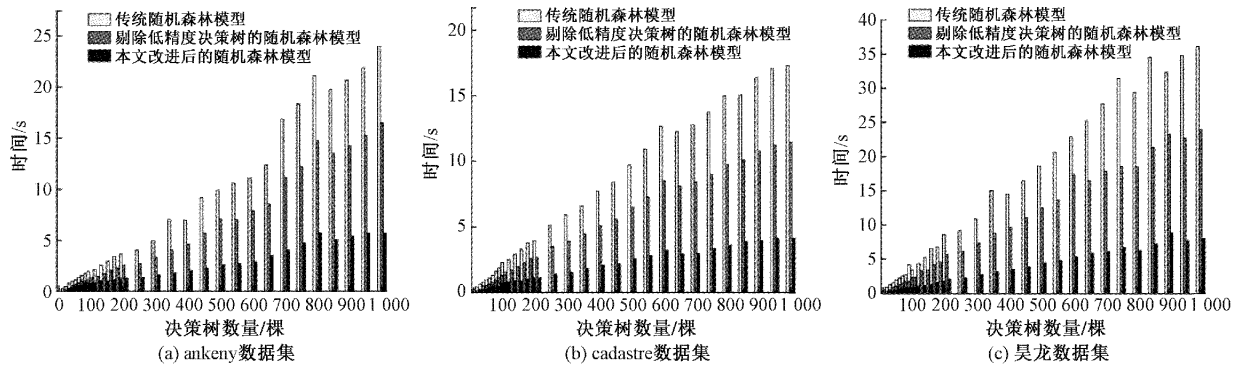



图 9　不同分类模型对 3 组数据集分类预测所用时间

改进后的随机森林模型需要依据构建好的随机森林模型中的决策树精度等指标进行构建，其建模速度相对于传统的随机森林会有所降低，但改进后的随机森林模型训练和分类效率得到了提升。从图 10 中可以看出，随着决策树数量的增多，时间缩短率明显上升，当决策树数量达到 400 棵时，时间缩短率基本保持平稳。3 组数据集最大时间缩短率分别能达到 76.4%、76.5%、79.0%。其主要原因在于：本文方法首先在传统随机森林模型的基础上剔除了分类精度低于平均分类精度的决策树，其次在高精度决策树的基础上进一步剔除了分类性能相似的决策树。模型中决策树的减少使得模型整体效率的提升和分类时间的缩短。同理文献[13-14]中的方法未改变决策树数量，故在模型分类效率上，本文方法优于文献[13-14]改进后的模型。

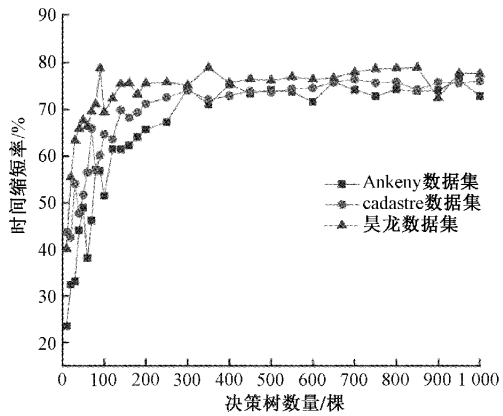



图 10　本文算法与传统随机森林相比时间缩短率

综上所述，改进后的随机森林不仅提高了模型的整体精度，而且在时间效率上也得到了明显提升。

## 4　结　　论

本文首先从随机森林中单棵决策树的分类精度及决策树之间分类性能的相似性出发，进行决策树筛选，在保证分类器多样性的前提下剔除低精度决策树；然后提出基于特征重要性的加权投票方法实现密集匹配点云的高精度分类。通过实验分析得到以下几点结论：

剔除低精度决策树的随机森林及基于不一致度量筛选后的随机森林模型的分类精度比传统的随机森林差，而本文改进后的模型分类精度比传统的随机森林高出 0.2%

左右。故改进后的模型并未使得模型分类精度降低;

本文改进后的随机森林模型相比于支持向量机(SVM)、决策树、神经网络、基于点特征分类等 4 种分类模型而言,分类精度分别提高了 15.159%、5.893%、6.316%、28.935%;

随着改进后模型中决策树数量的增多,模型分类时间缩短率明显上升。3 组数据集最大时间缩短率分别能达到 76.4%、76.5%、79.0%,分类预测效率得到大幅度提升。

## 参考文献


[1] 魏峰. 利用密集匹配点云的建筑物结构矢量化方法[J]. 遥感信息,2022,37(1):119-124.

[2] 王瑞,杨风暴. 基于表面粗糙度聚类的机载雷达点云数据地物分类方法研究[J]. 电子测量技术,2021,44(20):137-141.

[3] 戴莫凡,邢帅,徐青,等. 多特征融合与几何卷积的机载 LiDAR 点云地物分类[J]. 中国图象图形学报,2022,27(2):574-585.

[4] AIJAZI A, CHECCHIN P, TRASSOUDAINE L. Segmentation based classification of 3D urban point clouds:A super-voxel based approach with evaluation[J]. Remote Sensing, 2013, 5(4): 1624-1650.

[5] 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.

[6] QI C R, SU H, MO K, et al. PointNet: Deep learning on point sets for 3D classification and segmentation[J]. NeuRIPS, 2017, DOI: 10.1109/CVPR.2017.16.

[7] QI C R, LI YI, SU H, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space[J]. NeuRIPS, 2017, DOI: 10.48550/arXiv.1706.02413.

[8] 释小松,程英蕾,赵中阳,等. 基于三角网滤波和支持向量机的点云分类算法[J]. 激光与光电子学进展,2019,56(16): 32-40.

[9] NIEMEYER J, ROTTENSTENDTEINER F, SOERGEL U. Classification of urban LiDAR data using conditional random field and random forests[C]. Urban Remote Sensing Event, IEEE, 2013, DOI: 10.1109/jurse.2013.6550685.

[10] PARK Y J, JEAN-M G. Creating 3D city models with building footprints and LIDAR point cloud classification: A machine learning approach [J]. Computers, Environment and Urban Systems, 2019, 75:76-89.

[11] HUAN N, LIN X, ZHANG J. Classification of ALS point cloud with improved point cloud segmentation and random forests[J]. Remote Sens,2017,9(3):288.

[12] XUE D, CHENG Y, SHI X, et al. An improved random forest model applied to point cloud classification[J]. IOP Conference Series Materials Science and Engineering, 2020, 768(7):072037.

[13] 胡海瑛,惠振阳,李娜. 基于多基元特征向量融合的机载 LiDAR 点云分类[J]. 中国激光,2020,47(8):237-247.

[14] SUN T B, JINHAO L, JIANGMING K, et al. Research on target classification method for dense matching point cloud based on improved random forest algorithm[J]. International Journal of Information and Communication Technology, 2022, 21(3): 290-303.

[15] 王旭东,段福洲,屈新原,等. 面向对象和 SVM 结合的无人机数据建筑物提取[J]. 国土资源遥感,2017,29(1):97-103.

[16] HUAN N, XIANGGUO L, JIXIAN Z. Classification of ALS point cloud with improved point cloud segmentation and random forests[J]. Remote Sensing Technology & Application, 2017, 9(3):288.

[17] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.

[18] 王日升. 基于 Spark 的一种改进的随机森林算法研究[D]. 太原:太原理工大学,2017.

[19] 王诚. 一种基于聚类约简决策树的改进随机森林算法[J]. 南京邮电大学学报:自然科学版,2019,39(3):91-97.

[20] 章宁,陈钦. 基于 AUC 及 Q 统计值的集成学习训练方法[J]. 计算机应用,2019,39(4):935-939.

[21] 张璐璐. 多尺度分类挖掘方法[D]. 石家庄:河北师范大学,2020. 37-38.

[22] 马晓东. 基于加权决策树的随机森林模型优化[D]. 武汉:华中师范大学,2017.

[23] ZHANG W, JIANBO Q, PENG W, et al. An easy-to-use airborne LiDAR data filtering method based on cloth simulation [J]. Remote Sensing, 2016, 8(6):501.


## 作者简介


**吴冬**,博士研究生,主要研究方向为三维点云处理。

E-mail:1159918446@qq.com

**王井利**(通信作者),教授,硕士生导师,主要研究方向为三维激光扫描及精密测量技术。

E-mail:cehui0129@126.com