A TWO-PASS RANDOM FORESTS CLASSIFICATION OF AIRBORNE LIDAR AND IMAGE DATA ON URBAN SCENES

Li Guo1, Nesrine Chehata1,2, Samia Boukir1

1- Institut EGID, University of Bordeaux, GHYMAC laboratory
1 allée F. Daguin 33670 Pessac, France.
E-mail: name.surname@egid.u-bordeaux3.fr
2- IGN, MATIS laboratory
4 av Pasteur 94165 Saint-Mandé, France.
E-mail: name.surname@ign.fr

ABSTRACT

Random forests ensemble classifier showed to be suitable for classifying multisource data such as lidar and RGB image for urban scene mapping. However, two major problems remain: (1) the class boundaries are not well classified, a common issue in classification (2) the data are highly imbalanced raising another issue more specific to urban scenes. In this paper, we propose a new ensemble method based on the margin paradigm to improve the classification accuracy of minor classes. Random forests classifier is used in a two-pass methodology with an improved capability for classifying imbalanced data.

Index Terms—Lidar, Urban, Classification, Margin, Random Forests

1. INTRODUCTION

Airborne Laser Scanning (ALS) is an active remote sensing technique providing direct range measurements between the laser scanner and the earth topography. Many authors showed the potential of multi-echo lidar data for urban area analysis and building extraction [1, 2, 3]. The joint use of lidar and multispectral data has been studied in [4, 5]. Since 2004, full-waveform (FW) lidar have emerged with the ability to record the complete waveform of the backscattered 1D-signal. The potential of such data has been studied [6, 7, 8] for urban mapping. Our previous contribution in this context [5] has been devoted to the evaluation of the importance of various image and lidar features to classify urban scenes using random forests. However, two major problems remain: (1) the class boundaries are not well classified, a common issue in classification (2) the data are highly imbalanced raising another issue though more specific to dense urban scenes. In this paper, we propose a new ensemble method based on the margin paradigm to improve the classification accuracy of minor classes. More exactly, we use a random forests classifier with an improved capability for classifying imbalanced and/or complex data.

This paper is organized as follows. The lidar and image features are recalled in Section 2. Section 3 presents ensemble classifiers and more specifically random forests and the margin definition. The low margin classifier is detailed in section 4. Results and discussion are then presented in Section 5.

2. LIDAR AND AIRBORNE IMAGE DATA FEATURES

Our multisource data are composed of RGB airborne orthoimage, multiecho (ME) and full-waveform (FW) lidar data.

In order to combine lidar and image data, lidar points are projected into 2D image geometry using a bilinear interpolation. A previous work measured these feature importance using random forests [5]. The most relevant features that led to a good total classification accuracy were five multisource features \([R, B, \Delta z, A, \sigma]\) three of them being lidar characteristics thus emphasizing the contribution of lidar data to urban scene classification. Lidar features are summarized in Table 1.

<table>
<thead>
<tr>
<th>Type</th>
<th>Symbol</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME lidar</td>
<td>(\Delta z)</td>
<td>Height diff. to the ground</td>
</tr>
<tr>
<td>FW Lidar features</td>
<td>(\sigma)</td>
<td>Echo cross-section</td>
</tr>
</tbody>
</table>

Table 1. Lidar features for classification

3. ENSEMBLE CLASSIFIER

Ensemble learning is a popular learning paradigm, which builds a classification model by integrating multiple component learners. Bagging [9] is one of the most widely used and successful ensemble methods. Bagging is the acronym of bootstrap aggregating. It is made of the ensemble of bootstrap-inspired classifiers produced by sampling with replacement from training instances and uses these classifiers to get an aggregated classifier.

3.1. Random Forests

Random Forests is a variant of bagging proposed by Breiman [9]. It is a decision tree based ensemble classifier. It runs fast and efficiently on large data sets. It does not require assumptions on the distribution of the data which is interesting when different types or scales of input attributes are used. For airborne multisource classification, we showed in previous work [5] that it achieves a classification accuracy comparable to Support Vector Machines (SVMs) precision with a shorter training time.

Random Forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [9]. In training, the algorithm creates multiple bootstrapped samples
of the original training data to build a number of no pruning Classification and Regression Trees. For classification, each tree gives a unit vote for the most popular class at each input instance. The final label is determined by a majority vote of the trees. The number of variables $M$, randomly chosen at each split, is considered as the single user-defined adjustable parameter. It is often set to the square root of the number of inputs.

3.2. Margin definition

The margin theory of ensemble methods was firstly proposed by Schapire et al. [10] to explain the success of boosting. Following Schapire’s definition, the margin of a sample $x$ is computed by equation 1, where $v_y$ is the number of votes for the true class $y$ and $v_c$ is the number of votes for any other class $c$. The range of the margin is from -1 to +1. A positive margin value of a sample indicates this sample has been correctly classified, a negative value means the sample has been misclassified. The larger the margin, the more confidence in the classification.

$$
\text{margin}(x, y) = \frac{v_y - \max_{c_1, ..., c_L \neq y} (v_{c_1})}{\sum_{c=1}^L (v_c)}
$$

In addition, the margin of a sample reveals some of its characteristics. A large positive value means most of the trees classified this sample correctly, which implies this sample is just in the centre of the distribution of the related class. On the contrary, a large negative value shows that only a few trees classified the corresponding sample correctly, which probably represents noise or outliers of the related class. A value close to 0 indicates that the sample is likely on the boundary between these two candidate classes.

In classification problems, most errors occur on the samples lying on the boundaries of classes. However, these samples contain more specific information about the classes. Therefore, we propose a new definition of the margin defined in equation 2, where $c_1$ is the most voted class for sample $x$, $c_2$ is the second most popular class. Our margin’s range is from 0 to +1. The smaller the margin, the closer is the priori the related sample to the boundary of the classes, and therefore the more information they provide. Furthermore, our margin concept does not require the true class label of sample $x$.

$$
\text{margin}(x) = \frac{\max(v_{c_1}) - \max_{c_2, ..., c_L \neq c_1} (v_{c_2})}{\sum_{c=1}^L (v_c)}
$$

In the following, we exploit the margin of samples to evaluate the quality of classification and then to build a better classifier involving this key concept in ensemble methods.

4. LOW MARGIN CLASSIFIER

Multisource lidar and image classification for dense urban scenes present two main problems: (1) Boundary points are not well classified (2) The data are highly imbalanced as will be shown in our experiments later. To deal with these issues, the new margin definition is used to first evaluate the quality of the classification and then build a low margin classifier to improve classification accuracy on minor classes.

We propose a two-pass algorithm as depicted in figure 1. The data is split into training and test samples. The first pass is run with the whole test data. A Random Forests classifier is built and the margin is processed for both training and data samples. The margin helps evaluating the achieved classification quality. High margin samples are likely to be well classified. Conversely, low margin samples are likely to correspond to misclassifications. Besides, in a highly imbalanced data set, the RF classifier tend to better classify the major classes. Therefore, most errors occur on minor classes samples leading to low classification margins.

Based on this observation, we propose to process a new classification involving significantly more low margin training samples to better classify minor classes. Consequently, in the first pass, we keep the labels of high margin test samples ($\text{margin} > T_M$) and we reclassify the minor margin test samples in the second pass with an appropriate RF classifier. A classifier that is designed to efficiently handle minor classes has to train more minor classes samples. Based on this, the lowest margin training samples ($\text{margin} \leq T_M$) are kept to build the second classifier. The first classifier labels high margin test samples while the second classifier labels low margin test samples.

![Fig. 1. Two-pass classification scheme](image)

This algorithm improves the classification on minor classes while keeping a similar global accuracy as will be demonstrated by our experiments. First, boundary points are reclassified with an appropriate classifier built from lower margin training samples. Secondly, as mostly low margin test samples are used in the second pass, test data is more balanced since errors (low margins) mainly occur on minor classes which also improves classification accuracy for these classes.

In our method, one threshold is introduced: $T_M$ on training and test data margins. The best parameter maximizes minor classes accuracy as shown in the next section.

5. RESULTS AND DISCUSSION

The Random Forests implementation software by L. Breiman and A. Cutler ([http://www.r-project.org](http://www.r-project.org)) was used in experiments. Underlying parameters have been fixed to $M = 2$ which means that two variables are considered at each split and the number of trees was set to 100.

5.1. Data set

The data acquisition was carried out with the RIEGL LMS-Q560 system over the city of Biberach (Germany). The lidar point cloud has a density of 2.5 pts/m² with a footprint size of 0.25 m. The orthophotography has a resolution of 0.25 m. The city of Biberach includes four classes: artificial grounds, natural grounds, vegetation and buildings. The number of available reference samples is 195660. 20% randomly selected samples are used as a training set (cf. Table 2). The ground truth is processed manually, based on a region oversegmentation of the orthophoto.
Table 2. Data Set.

<table>
<thead>
<tr>
<th>Class</th>
<th>Training samples</th>
<th>Test samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building</td>
<td>17617</td>
<td>71396</td>
</tr>
<tr>
<td>Vegetation</td>
<td>1616</td>
<td>6752</td>
</tr>
<tr>
<td>Art. Grnd</td>
<td>18671</td>
<td>75283</td>
</tr>
<tr>
<td>Nat. Grnd</td>
<td>860</td>
<td>3463</td>
</tr>
<tr>
<td>Total samples</td>
<td>35764</td>
<td>140896</td>
</tr>
</tbody>
</table>

5.2. Margin results

The margin helps evaluating the quality of classification. Figure 2 depicts the study area, the ground truth and the margin image. This image was produced when building the random forests, using all data as training samples.

Fig. 2. Margin image based on 5 best features Random Forests classification (T=100, M=2). No labels are provided on black regions on the ground truth.

One can observe that the margin values are significantly higher in the center of classes (Buildings, roads). The smaller values correspond mainly to the class boundaries. In our case, they are located on building facades which are a transition between building and artificial ground classes and also on vegetation boundaries which are a mix between artificial ground and vegetation. In fact, lidar data are characterized by multiple echos on building facades and vegetation. Moreover, lidar pulse can penetrate vegetation to reach the ground underneath. Consequently, lidar data on these areas are a combination of pixels of different classes which make them harder to classify leading to a low margin value. Other errors with low margin occur on artificial ground, essentially in shadowed streets: they are due to the use of RGB channels that have uniform irrelevant values in urban corridors.

5.3. Classification results

The Random Forest classifier is first run on the whole test set. The confusion matrix is given in Table 3. The training dataset is highly imbalanced with two major classes (building and artificial ground) that are more than 10 times larger than vegetation and natural ground classes. We can notice that artificial ground and buildings are well classified with lower error rate. However, the algorithm has more difficulties in classifying natural ground and vegetation which suffer from smaller training sets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Building</td>
<td>69128</td>
<td>242</td>
<td>1986</td>
<td>42</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td>Vegetation</td>
<td>468</td>
<td>4882</td>
<td>1344</td>
<td>58</td>
<td>27.7</td>
<td></td>
</tr>
<tr>
<td>Art.Grnd</td>
<td>318</td>
<td>1917</td>
<td>589</td>
<td>72559</td>
<td>318</td>
<td></td>
</tr>
<tr>
<td>Nat.Grnd</td>
<td>2438</td>
<td>59</td>
<td>10</td>
<td>956</td>
<td>29.6</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Confusion matrix for test data using a RF classifier with 5 best features, 100 trees, and 2 split variables. Total error rate=5.03%.

5.4. Margin and classification confidence

Figure 3 illustrates the test data margin histograms for respectively well and wrongly classified samples. The margin distribution for training and test data are very similar. We can notice that 88% of well classified samples have a high margin (margin ≥ 0.9). However, for misclassified samples, the margin distribution is more dispersed and decreasing from low margin values to high values. One can observe that some samples, despite their high margin value, correspond to misclassifications. In fact, a high margin value means most of the trees classified this sample to the same class. However, this class does not necessarily correspond to the ground truth label. This may be due to data noise, outliers, errors in the ground truth or the non-suitability of input features. For instance, the number of misclassified samples should be decreasing with an increasing margin value. This is not verified for margin = 1. In this extreme case, it is probably due to errors in the ground truth itself. These margin distributions confirm the suitability of our two-pass classification strategy. In the first pass, high margin samples are likely to be well classified, so their labels are kept. Besides, to build a new classifier which suits more minor classes, lower margin training samples are involved. Thus, the labels of lower margin test samples are reprocessed.

Fig. 3. Margin histograms for well classified, misclassified test data.

5.5. Two-pass classification results

The margin threshold \(M_T\) on training and test data is optimized to maximize the classification accuracy on minor classes (i.e. natural ground and vegetation classes). The best parameter is found to be: \(M_T = 0.7\). With regards to margin histograms (Figure 3), the margin threshold on test data should be high indeed to reconsider the dispersion of low margin samples. Besides, the margin threshold on training data has also to be high enough to ensure the design of a low margin classifier that not only involves low margin instances but higher margin samples as well, as demonstrated in [11].
Figure 4 illustrates the size of training data used for both passes with the best margin threshold. One can observe that the second pass consists of more balanced data, thus favoring minor classes samples.

![Figure 4](image)

**Fig. 4.** Training data size for initial and two-pass RF classification with $T_M = 0.7$

Figure 5 depicts the margin threshold impact on classification accuracy for minor classes (natural ground and vegetation). With a small margin ($M_T \leq 0.4$), the classifier involves solely low margin instances while excluding higher margin instances, leading to a poor classification accuracy on minor classes. When increasing margin threshold ($0.4 \leq M_T \leq 0.7$) and therefore using more high margin training samples, the classification accuracy on minor classes is improved. In fact, a good classifier for minor classes should also be trained with high margin samples [11]. However, when the margin threshold is near 0.9, the second-pass classifier is too similar to the first classifier, training data are imbalanced and consequently accuracy decreases on minor classes.

![Figure 5](image)

**Fig. 5.** Margin threshold impact on classification accuracy.

After combining high margin test data from the first classification step and the low margin test data labels from the second step, the classification accuracy is improved on minor classes while keeping a good global accuracy (cf. Table 4). One can observe that vegetation classification accuracy is improved by 1.2% while natural ground is improved by 2%. The global classification error slightly increased by 0.04%.

<table>
<thead>
<tr>
<th>Predict. Class</th>
<th>Building</th>
<th>Veget</th>
<th>Art.Grnd</th>
<th>Nat.Grnd</th>
<th>Error %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building</td>
<td>60103</td>
<td>296</td>
<td>1956</td>
<td>43</td>
<td>3.2</td>
</tr>
<tr>
<td>Vegetation</td>
<td>444</td>
<td>4961</td>
<td>1295</td>
<td>52</td>
<td>26.5</td>
</tr>
<tr>
<td>Art.Grnd</td>
<td>1900</td>
<td>637</td>
<td>72373</td>
<td>373</td>
<td>3.9</td>
</tr>
<tr>
<td>Nat.Grnd</td>
<td>54</td>
<td>16</td>
<td>887</td>
<td>2506</td>
<td>27.6</td>
</tr>
</tbody>
</table>

Table 4. Two-pass RF classification : Confusion matrix for test data using 5 best features, 100 trees, and 2 split variables. Total error rate=5.07%.

6. CONCLUSION

In this paper, we have presented a new ensemble method based on the margin paradigm to improve the classification accuracy of minor classes. Random forests classifier is used in a two-pass methodology with an improved capability for classifying imbalanced data. It was applied to multisource lidar and image data on urban scenes. The first pass labels the high margin test data, while the second pass builds a low margin classifier to label low margin test samples. Combined labeled test data has an improved precision on minor classes. The method introduces only one parameter which is the margin threshold. It can be applied to both test and training samples since their distributions are very similar. Through our experiments, several observations have been made: (1) Our method performs better in terms of classification accuracy of difficult classes. (2) Our method is suitable to imbalanced dataset. (3) Selecting the training data on low margin values leads to a more balanced training set. (4) A low margin classifier has to be built not only with low margin instances but higher margin samples as well.

7. ACKNOWLEDGMENTS

The authors would like to thank Clément mallet from institute IGN-lab MATIS for providing lidar data and the ground truth.

8. REFERENCES


