Improved Urban Scene Classification Using Full-Waveform Lidar

M. Azadbakht, C. S. Fraser, and K. Khoshelham

Abstract
Full-waveform lidar data provides supplementary radiometric as well as more accurate geometric target information, when compared to discrete return systems. In this research, a wide range of classes in an urban scene; including trees, medium vegetation, low vegetation (grass), water bodies, pitched roofs, flat roofs, asphalt, vehicles, power lines, walls (fences) and concrete are considered. In order to tackle the challenge of distinguishing geometrically similar classes and enhancing the separability of other targets, a new set of features based on deconvolved waveforms is introduced. The positive effect of the proposed feature dataset on classification accuracy in individual classes is shown using two ensemble classifiers (random forests and RUSBoost). Performance of the classifiers is improved by integration with sampling techniques, especially for the under-represented classes. The final output of the proposed method is a highly detailed land cover map of the urban scene, which affords good separability between critical classes.

Introduction
Full-waveform lidar systems provide an almost continuous representation of the received signal (Mallet and Breton, 2009) in the form of amplitudes of returned pulses along the laser beam. These can be advantageous in characterizing targets. The recorded signal contains information regarding the geometry and radiometric characteristics of targets, but it is affected by various additional parameters, including flying height and the laser energy. Postprocessing of waveforms allows retrieval of point clouds with higher accuracy than those produced by discrete lidar systems. Also, valuable radiometric attributes of targets, such as the cross-section and its derivatives, can be recovered. The retrieved range-corrected target cross-section signal is independent of the laser instrument characteristics, and it is thus more representative of the targets. Moreover, the signal shape itself contains information that is of interest in applications such as forestry (Muss et al., 2013) and urban environments.

Detailed spatial information of urban environments is advantageous in a variety of applications, including change detection, mapping specific targets, urban planning and management, and disaster management (e.g., flood modeling). Accurate mapping of these classes (e.g., power lines or roof types) is a time-consuming task without utilizing lidar data. Land cover classification using lidar, along with the application of different methodologies over diverse classes of targets, has already been reported by many researchers (Buján et al., 2012; Chehata et al., 2009; Mallet et al., 2011; Rodriguez-Galiano et al., 2012). A thorough review of proposed lidar-based classification methods in urban environments has been reported by Yan et al. (2015) and the relevance of specific features from full-waveform lidar data for classification in these environments has also been reported (Alexander et al., 2010; Mallet et al., 2011; Niemeyer et al., 2011). However, thus far, only a limited number of classes have been considered. For example, the three major classes of buildings, vegetation, and ground were investigated by Mallet et al. (2011) and Niemeyer et al. (2011). Alexander et al. (2010) reported classification using decision trees and waveform data over six targets in an urban scene, where a small number of samples was used to assess the classification result. A combined feature dataset of lidar and a satellite image was utilized by Guo et al. (2011), where per-class performance of the under-represented classes was not satisfactory.

Different target surfaces are in different abundance in complex urban environments, which causes data imbalance, in that small numbers of training samples are available for specific classes. This is of significant importance when training a classifier for a large number of classes, and it should be addressed in order to attain more acceptable results. In this study, two state-of-the-art classifiers, random forests and RUSBoost, have been selected in the investigation of a very detailed classification problem with 11 classes, in which the training dataset is heavily imbalanced towards a few of the classes. The feature dataset is comprised of geometric, radiometric and pulse shape attributes extracted from waveforms. In order to balance the training samples, different portions of majority classes are considered in the case of classification using random forests, while a random undersampling strategy is implemented in RUSBoost. The relevance of waveform features is shown by analyzing their impact on the classification accuracy.

Methodology
In this study, features extracted from waveform lidar are considered in order to classify a dense urban environment. These features range from geometric to radiometric, and extend to additional features extracted from the waveform itself. Retrieval of the target response (cross-section) is based upon a robust deconvolution method developed in previous work (Azadbakht et al., 2016). Two ensemble classifiers are then applied to examine the role of both waveform features and the size of the training dataset in the identification of various target classes.

Waveform Processing
In order to retrieve the target response from waveforms, a robust deconvolution method based upon sparsity-based regularization (Azadbakht et al., 2016; Azadbakht et al., 2014)
is applied to the received waveforms. Similar to other deconvolution methods, it does not demand the number of returns in advance and requires no assumption on the shape of pulses (Roncat et al., 2011; Wu et al., 2011). In addition, the regularization parameter in this method is estimated automatically with respect to the emitted pulse and the received waveform by adopting the L-curve method (Hansen and O’Leary, 1993). This is of importance in the retrieval of the appropriate signal when dealing with different levels of complexity in waveforms and in the occasional changes of the laser energy.

Suppose $P$ is a vector representing the waveform recorded by the receiver, $a$ a vector including the differential backscatter cross-section, $S$ the blur matrix with elements from the system waveform, and $\lambda$ the regularization parameter value. The cost function is then considered to be minimized in order to retrieve the proper target response with minimal oscillations (Azadbakht et al., 2016):

$$\hat{a} = \arg \min \|P - Sa\|_2 + \lambda \|L\|_2$$

s.t. $\sigma \geq 0$ \hspace{1cm} (1)

Here, $\|\cdot\|_2$ and $\|\cdot\|_1$ are the two-norm and one-norm, respectively, and $L$ is an operator that controls the smoothness of the result. The solution of the non-smooth convex optimization problem in Equation 1 is obtained using primal-dual interior point methods (Boyd and Vandenberghe, 2004).

Features of Interest

In order to extract features at a point level, two different neighborhoods, a cylindrical and a spherical, are defined over georeferenced points derived from the waveforms. For each neighborhood, the number of points selected is based upon two principal criteria: a fixed search radius for neighbors and a constant number of points within the neighborhood. Three different radius values of 1, 1.5, and 2 meters, and neighborhoods with 10 and 20 contiguous points, are considered so as to examine the role of the adjacent points in describing the dominant local geometry of the points in the neighborhood. Distribution of the points inside these neighborhoods can be described by the covariance matrix of the point coordinates (Mallet et al., 2011). Each of these neighborhoods is useful in recognizing specific targets, for example, the spherical environment could be advantageous in distinguishing power lines by a large vertical distance, e.g., points on the ground. On the other hand, 3D structures with vertical distribution of points, e.g., trees, can be appropriately described using cylindrical environments.

Eigenvalues ($\lambda_1, \lambda_2, \lambda_3$) are calculated for each local neighbourhood around the points, and they are normalized with respect to the summation of the three related eigenvalues ($\sigma = \lambda_1/\sum \lambda_i$). Apart from the normalized eigenvalues, the characteristics of other combinations can also be investigated, including anisotropy $A_A$, planarity $P_s$, omnivariance $O_s$, and scatter $S_s$, where each feature is appropriate in distinguishing a particular structure. A DTM and a DSM are calculated from a TIN, which is obtained from the retrieved 3D points, and a normalized DSM (nDSM) is obtained as a result of subtracting the DTM from the DSM. This feature is of significant importance in separating ground and non-ground targets. Additional geometric features are slope $S$, aspect $A_s$, curvature $C$, profile curvature $C_{nOP}$, and plan curvature $C_{OPEN}$, which are also extracted from the DSM of the aggregated points. These features can provide a local quantitative description of the target complexity, and they are beneficial when dealing with targets such as pitched or flat roofs and separating them from other targets, e.g., trees. Pulse (return) number $R_t$, total number of returns $N_r$ (in a single waveform) and the ratio of these two $N_r$ are also considered, with the potential for separating multi-return targets from single-return surfaces.

Radiometric features including the pulse amplitude $A$, the target cross-section $\sigma$, the total backscatter cross-section of the whole waveform $\sigma_t$, and the backscattering coefficient $\gamma$ are also extracted from the retrieved waveforms in Equation 1 after calibrating them for range, and also with regard to an asphalt target which is considered as a standard Lambertian surface with a specified reflectivity at the required wavelength (Alexander et al., 2010; Azadbakht et al., 2016; Mallet et al., 2011).

Additional features are extracted from the geometry of the restored target response. The features of interest are the centre of mass ($C_x, C_y$) for each pulse, the pulse length $L$, the pulse area fraction from each tail end to the maximum amplitude, restricted by the line connecting the centre of mass, and the maximum amplitude. The calculated areas are normalized by both the whole area under the curve ($A_L, A_R$) and the summation of these two left/right areas ($A_L, A_R$). Figure 1 shows a graphical description of these features for an asymmetric pulse. Two exponential functions $E_r, E_s$ are fitted to the left and right tails of the pulse to describe the growth and decay rates. These features can better describe the roughness of the surface, with the assumption of affecting the distribution of the returned energy, where higher $A_L$ (or $A_R$) represents the skewness of the pulse. The exponential decay (or growth) rate also can describe the surface geometry and also the imperviousness of the surface. Altogether, a set of 31 features was considered, for each neighborhood, in order to investigate their levels of significance in classification of targets in an urban scene.

Classification Strategies

Ensemble classification methods are known as a combination of individual classifiers, and the aim of these is to increase the classification accuracy by amalgamating decisions of individual classifiers to label classes more appropriately (Galar et al., 2012). The Random forest classifier was selected for application to the dataset because it is less affected by noise and outliers as a result of the randomness of the variable and sample selection, in comparison to a single tree classifier. A five-fold cross-validation was considered to partition the sample set into five splits, where four subsets at a time are used to train the classifier and the remaining subset is used as the test set. The total number of 100 trees was used in the random forest, where the number of features at each split was estimated according to the out-of-bag (OOB) error for each fold.
Targets in a complex urban area are neither evenly distributed nor cover similar area proportions of the study area. As a result, it is highly improbable to record a similar number of samples from each target using a laser scanner, and dominant targets (e.g., trees and buildings) will consequently comprise a far higher percentage of samples. This eventually causes data imbalance and impedes acceptable training of the classifier because the training will be biased towards those classes with a large number of samples and the random forest classifier is no exception to this rule. This is basically due to the intrinsic characteristics of classification methods (even ensembles), which try to increase the overall accuracy (Galar et al., 2012), while overall accuracy does not necessarily represent the classification performance for smaller classes with fewer samples.

Two major research directions to address the data imbalance are cost-sensitive approaches and sampling methods (He and Garcia, 2009). The former assigns high misclassification costs to the classes of fewer samples, while the latter alleviates the data imbalance by equalizing the training set using over-/under-sampling, thus providing a more stabilized number of samples for both the majority and minority classes. Undersampling the majority classes may lead to information deficiency when excluding samples of these classes from the training set. Oversampling the minority classes, on the other hand, can cause overfitting issues due to the use of a similar sample more frequently than others (Galar et al., 2012; Seiffert et al., 2010). In this study, the emphasis is on the sampling strategy, with both the random forest and RUSBoost classifiers.

RUSBoost is an ensemble classifier which brings boosting (Adaboost.M2 algorithm) together with random undersampling of majority classes in order to enhance the classification accuracy when data samples are skewed (Seiffert et al., 2010). This method appears to take less training time and is of a lower complexity than similar boosting methods that consider oversampling (e.g., SMOTEBoost (Chawla et al., 2003)), due to the smaller number of samples that are incorporated in the training. Moreover, the common problem of undersampling techniques (information insufficiency) can be relieved by the inherent characteristics of boosting (Seiffert et al., 2010).

The majority and minority classes are also separately considered to examine how both the overall and per-class performance of the classifier are affected. In this case, the random forest classifier is implemented on the two subsets, without imposing specific sampling techniques. For the evaluation of classification performance, in the case of imbalanced data, a single evaluation metric (e.g., overall accuracy) is inadequate and is not representative of smaller classes. Therefore, sensitivity (recall), specificity (1-FPR) and precision for each class, along with the geometric mean (G-mean) of the recall measures of all classes (Sun et al., 2006) as a standard evaluation measure for the entire classifier, are used instead.

Results of Experiments and Discussion

Study Area

The study area is an urban environment with a high variety of land cover types located in Woodridge, Queensland, Australia. The lidar dataset was acquired in October 2013 using a Riegl LMS-Q680i scanner. Of the whole 37 strips acquired, which cover a very large area (over Karawatha forest), only one has been selected for this study. A simultaneously captured aerial photograph has been used to cross check the sample points. An intensity map and the nDSM of the study area are shown in Figure 2. Figure 3 illustrates a perspective view of the sample points with their assigned true labels, which were determined manually by using a combination of the aerial photograph, a TIN and the nDSM.
Table 1. OA and Kappa Indices (in %) for Different Types of Neighborhoods; Cyl-im and Cyl-jN, and Sph-im, and Sph-jN are Cylindrical and Spherical Neighborhoods, Respectively, with i m Radius and j Neighbors

<table>
<thead>
<tr>
<th>Neighborhood</th>
<th>OA</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyl-1m</td>
<td>95.08</td>
<td>92.72</td>
</tr>
<tr>
<td>Cyl-1.5m</td>
<td>95.22</td>
<td>92.92</td>
</tr>
<tr>
<td>Cyl-2m</td>
<td>95.35</td>
<td>93.11</td>
</tr>
<tr>
<td>Cyl-10N</td>
<td>94.43</td>
<td>91.75</td>
</tr>
<tr>
<td>Cyl-20N</td>
<td>94.64</td>
<td>92.06</td>
</tr>
<tr>
<td>Sph-1m</td>
<td>95.20</td>
<td>92.91</td>
</tr>
<tr>
<td>Sph-1.5m</td>
<td>95.47</td>
<td>93.31</td>
</tr>
<tr>
<td>Sph-2m</td>
<td>95.45</td>
<td>93.27</td>
</tr>
<tr>
<td>Sph-10N</td>
<td>94.53</td>
<td>91.88</td>
</tr>
<tr>
<td>Sph-20N</td>
<td>94.73</td>
<td>92.19</td>
</tr>
</tbody>
</table>

Table 2. Sample Proportion (in %) of the Full Dataset

<table>
<thead>
<tr>
<th>Class</th>
<th>Sample proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asphalt</td>
<td>11.4</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.5</td>
</tr>
<tr>
<td>Concrete</td>
<td>1.2</td>
</tr>
<tr>
<td>Flat roof</td>
<td>2.9</td>
</tr>
<tr>
<td>Grass</td>
<td>16.7</td>
</tr>
<tr>
<td>Med-Veg.</td>
<td>1.4</td>
</tr>
<tr>
<td>Pitched roof</td>
<td>13.6</td>
</tr>
<tr>
<td>Power line</td>
<td>0.8</td>
</tr>
<tr>
<td>Tree</td>
<td>50.5</td>
</tr>
<tr>
<td>Wall</td>
<td>0.7</td>
</tr>
<tr>
<td>Water</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Table 3. Recall and Precision Values (in %) for Individual Classes of the Best Neighborhoods

<table>
<thead>
<tr>
<th>Class</th>
<th>Recall</th>
<th>Precision</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asphalt</td>
<td>98.36</td>
<td>97.80</td>
<td>98.49</td>
<td>97.27</td>
</tr>
<tr>
<td>Vehicle</td>
<td>51.74</td>
<td>86.88</td>
<td>68.40</td>
<td>76.36</td>
</tr>
<tr>
<td>Concrete</td>
<td>7.16</td>
<td>62.65</td>
<td>1.17</td>
<td>58.62</td>
</tr>
<tr>
<td>Flat roof</td>
<td>86.33</td>
<td>87.32</td>
<td>84.79</td>
<td>88.10</td>
</tr>
<tr>
<td>Grass</td>
<td>98.78</td>
<td>92.44</td>
<td>98.92</td>
<td>92.03</td>
</tr>
<tr>
<td>Med-Veg.</td>
<td>58.08</td>
<td>74.08</td>
<td>34.92</td>
<td>60.12</td>
</tr>
<tr>
<td>Pitched roof</td>
<td>95.84</td>
<td>96.71</td>
<td>96.82</td>
<td>97.36</td>
</tr>
<tr>
<td>Power line</td>
<td>24.35</td>
<td>89.27</td>
<td>69.28</td>
<td>86.10</td>
</tr>
<tr>
<td>Tree</td>
<td>99.39</td>
<td>96.78</td>
<td>99.46</td>
<td>97.27</td>
</tr>
<tr>
<td>Wall</td>
<td>54.08</td>
<td>71.47</td>
<td>50.12</td>
<td>70.02</td>
</tr>
<tr>
<td>Water</td>
<td>57.35</td>
<td>84.03</td>
<td>41.94</td>
<td>87.62</td>
</tr>
</tbody>
</table>

Results of Random Forest Classification

The random forest classification was performed on the dataset, with its performance being evaluated by a five-fold cross-validation. Overall accuracy (OA) and Kappa indices are presented in Table 1, for the different neighborhood types. At first glance, the results show a satisfactory classification with a very high level of accuracy. However, inspection of the prevalence of samples in each class reveals that major classes contain more than 92 percent of the total samples, with 50.5 percent of the samples from the trees class. Water bodies, walls, power lines, and vehicles, on the other hand, each comprise less than 1 percent of the observations (see Table 2). This shows that the dataset is heavily imbalanced and the overall accuracy and Kappa factor are not appropriately representative for all classes.

Table 3 presents the recall and precision values per class for two different types of neighborhoods with maximum overall accuracy. The table lists low accuracy (particularly for the recall values) for smaller classes, although the overall accuracy is higher than 95 percent in both cases. This is chiefly due to the low number of samples for these classes; the classifier is trying to train the model that maximizes overall accuracy, and therefore the training is biased towards the majority classes.

Eigenvalue features from the best two neighborhoods (Cyl-2m and Sph-1.5m) were aggregated together with all other features, as they can be complementary when considering a large number of classes with different characteristics. For example, two classes of power lines and vehicles were better distinguished by applying the spherical neighborhood, while some other classes (e.g., medium vegetation, water bodies, and walls) were better classified in the case of the cylindrical neighborhood. In this case, both the OA and Kappa values were increased, to 96.47 percent and 94.79 percent, respectively. The recall was also improved for all classes, with more significant improvements for water bodies, power lines, and medium vegetation. G-mean values of 55 percent and 48 percent were reported for Cyl-2m and Sph-1.5m neighborhoods, respectively, while this measure was increased to 65.01 percent in the case of the aggregated feature set.

The importance of lidar waveform features for individual classes is shown in Figure 4, where most of the features are with negative significance in the case of the concrete class. The most important feature for almost all classes, on the other hand, is the nDSM (shown in Figure 4 as Z). Among the waveform features, the total cross-section σt, the backscatter coefficient λ, the pulse length L, and Cq emerged with higher contribution weights. It is noticeable that the important values are normalized with regard to the norm of all values for each individual class.

Figure 5 illustrates the sensitivity and specificity per class for the aggregated feature set, as a result of performing the random forest classification. Even though high specificity for all classes results in marginal false positive rates, the sensitivity values for minority classes are low, indicating that samples of minority classes are misclassified as belonging to other classes. These results provide evidence that the dataset is heavily imbalanced. In order to balance the number...
of observations for the random forest classifier the majority classes were under-sampled by stratified sampling (Chen et al., 2004) of the majority strata, without changing the number of samples in the minority classes. This was also based on the fraction of samples in each class, where all the samples were selected from the classes of interest (e.g., classes with a limited number of samples) and different fractions of the majority samples were examined with the same classifier (random forests).

By considering only the majority classes (asphalt, flat roofs, grass, medium vegetation, pitched roofs and trees), and by applying the random forest classifier, the OA was calculated to be 98.53 percent when the Kappa index was 97.77 percent. Lower accuracy measures for the flat roofs and medium vegetation classes are mainly due to the limited number of samples in these two classes (see Figure 6). The G-mean measure was increased significantly to 92 percent, which indicates an increase of recall values for almost all classes. The features in the same set, as reported in Table 6, were among the most important that were ranked with regard to the mean decrease in accuracy index.

Separately, observations from minority classes were used in a classification in order to examine their separability (see Figure 7). Exclusion of samples from the majority classes (e.g., grass), which were the source of biases in training, resulted in a considerable improvement in classification for concrete since more than 90 percent of concrete samples were misclassified as grass in the main model (without undersampling). Smaller recall/precision values for both medium vegetation and wall classes were mainly due to their similarities and the overlap of their features. Also, vehicles showed a high level of misclassification, which is due to their characteristics being similar to both medium vegetation and wall classes. The OA and Kappa index values were 91.13 percent and 90.53 percent, respectively, and the G-mean measure was 90 percent.

An inspection of the feature rankings in this case reveals significant improvement for features from waveforms, particularly for $\sigma_\gamma$, $\gamma$, $\sigma$ and $C_m$, with almost the same importance level for $\sigma_\gamma$ and the nDSM.

**Results of Random Forest Classification with Sampling**

An advantage of using different fractions of samples in stratified sampling, with the random forest classification, is that it allows an assessment of feature importance in different conditions since feature importance can be affected by an excessive number of samples from majority classes (Chen et al., 2004). Another general advantage of using stratified fractions of samples is improvement of the model in terms of computational cost, provided that these fractions of samples result in at least similar accuracy as compared to that using whole samples. Table 4 shows the three sampling scenarios, with different fractions from the majority classes.

The OA (respectively Kappa) values for the three sampling scenarios in Table 4 are 96.49 percent (94.85 percent), 96.23 percent (94.46 percent), and 93.56 percent (90.71 percent), which highlights a gradual decrease in accuracy. On the other hand, the G-mean measures for these three cases were 71.82 percent, 75.13 percent, and 83.03 percent, compared to 65.01 percent in the case of using all samples. Significant increase of the G-mean measure for RFC-Str.3 is accompanied by nearly 3 percent and 4 percent decreases of the OA and Kappa measures, respectively, in comparison with RFC-Str.2. This also shows decreases of precision and recall values for minority and majority classes, respectively. Inclusion of adequate portions of samples from the majority classes (e.g., as in RFC-Str.1 and RFC-Str.2) thus resulted in improvement of the G-mean measure, while no deterioration of the classification performance in terms of the overall accuracy was evident.

An important consequence of this is that taking an identical number of samples from all classes during undersampling, when dealing with extremely skewed data sets, is avoided. This finding is in close agreement with the findings of other studies, in which a larger fraction of the majority class was recommended in the case of binary classification (Seiffert et al., 2010). The precision and recall measures are listed in Table 5 for RFC-Str.1 and RFC-Str.2.
As shown in Table 6, non-geometric features extracted from waveforms (radiometric and shape-based features) contribute to the classification model in the top ten features, which are highlighted in the Table in terms of the mean decrease in accuracy. Almost similar features were highlighted by considering different proportions of majority classes by stratified undersampling, with an increase of the contribution of waveform features in the case of Str.1 and Str.3. C1 particularly, emerges among the more important features in all three sampling scenarios. This measure, however, can be unreliable depending on the type of variables (or their scales).

Results of RUSBoost Classification

Boosting methods (including RUSBoost) are also capable of handling imbalanced samples due to their inherent characteristics of training sequentially and, therefore, improving the tree performance at each step. RUSBoost classification was implemented with different numbers of trees (100, 200, 500, and 1,000), since boosting methods can overfit when using a large number of trees.

Table 7 shows three different ratios to the number of samples in the water class, as the class with the smallest number of samples. Boosting may outperform random forests because of consideration of previously grown trees when the current tree is grown, and this can improve the model for those samples with poor results. Similar to the evaluation of random forest classification, a five-fold cross validation was used to partition the dataset as training and test samples for the RUSBoost classifier.

The recall values for the minority classes were significantly improved by adopting RUSBoost. The maximum improvements were reported in the case of concrete (from 9 percent to 90.2 percent), walls (from 55 percent to 84.4 percent), power lines (from 72 percent to 92.1 percent), water bodies (from 76 percent to 95 percent) and vehicles (from 69 percent to 86.8 percent), with nearly identical improvements for different numbers of trees. The precision and recall measures are listed in Table 8 for RUS-1 to 3. G-mean measures of 78.26 percent, 83.79 percent, and 85.80 percent were obtained for the case of RFC-Str.1 and RFC-Str.2, respectively, while the OA and Kappa measures for the random forest and RUSBoost (500 trees) classifiers.

Figure 8 also illustrates the overall comparison of the random forest and RUSBoost classifiers, under different sampling conditions. As seen in Figure 8, the RUSBoost outperforms the random forest classifier over the total dataset and it provides results of higher accuracy than that for RFC-Str. in terms of the G-mean measure. Similar to RFC-Str., undersampling majority classes to the number of samples in the smallest class (e.g., RUS3) causes a significant decrease in both OA and Kappa measures. This obviously indicates reductions in recall/precision values for majority/minority classes. The OA and Kappa measures for RUS2 are 94 percent and 91.30 percent, respectively, while 83.79 percent is reported as the G-mean value.

As shown in Table 6, non-geometric features extracted from waveforms (radiometric and shape-based features) contribute to the classification model in the top ten features, which are highlighted in the Table in terms of the mean decrease in accuracy. Almost similar features were highlighted by considering different proportions of majority classes by stratified undersampling, with an increase of the contribution of waveform features in the case of Str.1 and Str.3. C1 particularly, emerges among the more important features in all three sampling scenarios. This measure, however, can be unreliable depending on the type of variables (or their scales).

Results of RUSBoost Classification

Boosting methods (including RUSBoost) are also capable of handling imbalanced samples due to their inherent characteristics of training sequentially and, therefore, improving the tree performance at each step. RUSBoost classification was implemented with different numbers of trees (100, 200, 500, and 1,000), since boosting methods can overfit when using a large number of trees.

Table 7 shows three different ratios to the number of samples in the water class, as the class with the smallest number of samples. Boosting may outperform random forests because of consideration of previously grown trees when the current tree is grown, and this can improve the model for those samples with poor results. Similar to the evaluation of random forest classification, a five-fold cross validation was used to partition the dataset as training and test samples for the RUSBoost classifier.

The recall values for the minority classes were significantly improved by adopting RUSBoost. The maximum improvements were reported in the case of concrete (from 9 percent to 90.2 percent), walls (from 55 percent to 84.4 percent), power lines (from 72 percent to 92.1 percent), water bodies (from 76 percent to 95 percent) and vehicles (from 69 percent to 86.8 percent), with nearly identical improvements for different numbers of trees. The precision and recall measures are listed in Table 8 for RUS-1 to 3. G-mean measures of 78.26 percent, 83.79 percent, and 85.80 percent were obtained for the case of RFC-Str.1 and RFC-Str.2, respectively, while the OA and Kappa measures for the random forest and RUSBoost (500 trees) classifiers.

Figure 8 also illustrates the overall comparison of the random forest and RUSBoost classifiers, under different sampling conditions. As seen in Figure 8, the RUSBoost outperforms the random forest classifier over the total dataset and it provides results of higher accuracy than that for RFC-Str. in terms of the G-mean measure. Similar to RFC-Str., undersampling majority classes to the number of samples in the smallest class (e.g., RUS3) causes a significant decrease in both OA and Kappa measures. This obviously indicates reductions in recall/precision values for majority/minority classes. The OA and Kappa measures for RUS2 are 94 percent and 91.30 percent, respectively, while 83.79 percent is reported as the G-mean value.

As shown in Table 6, non-geometric features extracted from waveforms (radiometric and shape-based features) contribute to the classification model in the top ten features, which are highlighted in the Table in terms of the mean decrease in accuracy. Almost similar features were highlighted by considering different proportions of majority classes by stratified undersampling, with an increase of the contribution of waveform features in the case of Str.1 and Str.3. C1 particularly, emerges among the more important features in all three sampling scenarios. This measure, however, can be unreliable depending on the type of variables (or their scales).

Results of RUSBoost Classification

Boosting methods (including RUSBoost) are also capable of handling imbalanced samples due to their inherent characteristics of training sequentially and, therefore, improving the tree performance at each step. RUSBoost classification was implemented with different numbers of trees (100, 200, 500, and 1,000), since boosting methods can overfit when using a large number of trees.

Table 7 shows three different ratios to the number of samples in the water class, as the class with the smallest number of samples. Boosting may outperform random forests because of consideration of previously grown trees when the current tree is grown, and this can improve the model for those samples with poor results. Similar to the evaluation of random forest classification, a five-fold cross validation was used to partition the dataset as training and test samples for the RUSBoost classifier.

The recall values for the minority classes were significantly improved by adopting RUSBoost. The maximum improvements were reported in the case of concrete (from 9 percent to 90.2 percent), walls (from 55 percent to 84.4 percent), power lines (from 72 percent to 92.1 percent), water bodies (from 76 percent to 95 percent) and vehicles (from 69 percent to 86.8 percent), with nearly identical improvements for different numbers of trees. The precision and recall measures are listed in Table 8 for RUS-1 to 3. G-mean measures of 78.26 percent, 83.79 percent, and 85.80 percent were obtained for the case of RFC-Str.1 and RFC-Str.2, respectively, while the OA and Kappa measures for the random forest and RUSBoost (500 trees) classifiers.

Figure 8 also illustrates the overall comparison of the random forest and RUSBoost classifiers, under different sampling conditions. As seen in Figure 8, the RUSBoost outperforms the random forest classifier over the total dataset and it provides results of higher accuracy than that for RFC-Str. in terms of the G-mean measure. Similar to RFC-Str., undersampling majority classes to the number of samples in the smallest class (e.g., RUS3) causes a significant decrease in both OA and Kappa measures. This obviously indicates reductions in recall/precision values for majority/minority classes. The OA and Kappa measures for RUS2 are 94 percent and 91.30 percent, respectively, while 83.79 percent is reported as the G-mean value.

As shown in Table 6, non-geometric features extracted from waveforms (radiometric and shape-based features) contribute to the classification model in the top ten features, which are highlighted in the Table in terms of the mean decrease in accuracy. Almost similar features were highlighted by considering different proportions of majority classes by stratified undersampling, with an increase of the contribution of waveform features in the case of Str.1 and Str.3. C1 particularly, emerges among the more important features in all three sampling scenarios. This measure, however, can be unreliable depending on the type of variables (or their scales).

Results of RUSBoost Classification

Boosting methods (including RUSBoost) are also capable of handling imbalanced samples due to their inherent characteristics of training sequentially and, therefore, improving the tree performance at each step. RUSBoost classification was implemented with different numbers of trees (100, 200, 500, and 1,000), since boosting methods can overfit when using a large number of trees.

Table 7 shows three different ratios to the number of samples in the water class, as the class with the smallest number of samples. Boosting may outperform random forests because of consideration of previously grown trees when the current tree is grown, and this can improve the model for those samples with poor results. Similar to the evaluation of random forest classification, a five-fold cross validation was used to partition the dataset as training and test samples for the RUSBoost classifier.

The recall values for the minority classes were significantly improved by adopting RUSBoost. The maximum improvements were reported in the case of concrete (from 9 percent to 90.2 percent), walls (from 55 percent to 84.4 percent), power lines (from 72 percent to 92.1 percent), water bodies (from 76 percent to 95 percent) and vehicles (from 69 percent to 86.8 percent), with nearly identical improvements for different numbers of trees. The precision and recall measures are listed in Table 8 for RUS-1 to 3. G-mean measures of 78.26 percent, 83.79 percent, and 85.80 percent were obtained for the case of RFC-Str.1 and RFC-Str.2, respectively, while the OA and Kappa measures for the random forest and RUSBoost (500 trees) classifiers.

Figure 8 also illustrates the overall comparison of the random forest and RUSBoost classifiers, under different sampling conditions. As seen in Figure 8, the RUSBoost outperforms the random forest classifier over the total dataset and it provides results of higher accuracy than that for RFC-Str. in terms of the G-mean measure. Similar to RFC-Str., undersampling majority classes to the number of samples in the smallest class (e.g., RUS3) causes a significant decrease in both OA and Kappa measures. This obviously indicates reductions in recall/precision values for majority/minority classes. The OA and Kappa measures for RUS2 are 94 percent and 91.30 percent, respectively, while 83.79 percent is reported as the G-mean value.
misclassification rates for minority classes for the random forest classifier, while erroneously classified points in majority classes are detectable in the case of RUSBoost.

Conclusions

Land cover classification across 11 adopted classes from full-waveform lidar data in an urban environment has been investigated. The results in terms of different performance measures indicated high accuracy when the dataset was divided into majority and minority groups of classes. Each group contained six classes, with medium vegetation being considered in both groups. Imbalanced datasets were recognized through a classification result of high specificity and low sensitivity values, as was shown in Figure 5. This indicated the inappropriateness of the random forest classifier for minority classes. The recall values for those minority classes were improved after applying RFC-Str. and RUSBoost, which is a well-known boosting classifier for learning imbalanced datasets. The associated reduction in specificity values was shown to be marginal. High recall values were reported with the random forest classifier only for those classes with a higher number of samples, while RUSBoost performed much better in the case of minority classes, with the high recall values of majority classes being maintained. The geometric mean of recall values of individual classes (Figure 8) represented a consistent measure to assess the overall performance of the classifier, while OA and Kappa both ignored the inferior performance of the classifier for minority classes.

On the other hand, both OA and Kappa can be used, particularly to assess the accuracy in the case of majority classes, since a significant decrease of these two measures should similarly be avoided. Therefore, a trade-off between the G-mean, from one side, and these two measures from the other side can be considered as a criterion in selection of the optimal sampling scenario.

The improvements in the recall values for minority classes was due to the inherent differences of the random forest and boosting classifiers (e.g., RUSBoost). Boosting algorithms improved the classification result for those classes with weak performance from the previous step. The effects of undersampling prior to training was also shown to be an important factor. Inferior results for minority samples in the case of the random forest classifier were due to the classifier (similar to other classifiers) straining to increase the overall accuracy of the classification through the improvement of majority classes. This can be explained as a result of lower probability of the contribution of minority samples during the bootstrap sampling. On the other hand, both False Positive (FP) and False Negative (FN) parameters are considered in the case of boosting. Even though RUSBoost resulted in high recall values for minority classes, the Type I error increased on occasion for these classes, which caused samples from other (possibly majority) classes to be wrongly labeled as members of minority classes. This was especially so in the case of taking an identical number of samples from all classes (e.g., RUS3). Stratified sampling also showed improvements over the entire dataset, while both the OA and Kappa values were significantly decreased for the third subsampling strategy. Reducing the computation time is an obvious advantage of stratified
undersampling. The provision of a highly accurate classification result for an urban area with many classes requires further investigation into the extraction of additional features from either full-waveform Lidar data or from ancillary sources, along with an improved classification strategy. Integration of data mining techniques may well boost classifier performance and this is under investigation.

References


