Hoang Minh Nguyen, Begüm Demir, Michele Dalponte

Weighted Support Vector Machines for Tree Species Classification Using Lidar Data

Open Access via institutional repository of Technische Universität Berlin

Document type

Conference paper | Accepted version

(i. e. final author-created version that incorporates referee comments and is the version accepted for publication; also known as: Author's Accepted Manuscript (AAM), Final Draft, Postprint)

This version is available at

<https://doi.org/10.14279/depositonce-9350>

Citation details

Nguyen, H. M., Demir, B., & Dalponte, M. (2019). Weighted Support Vector Machines for Tree Species Classification Using Lidar Data. In: 2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2019). pp. 6740–6743 <https://doi.org/10.1109/igarss.2019.8900398>

© © 20xx IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Terms of use

This work is protected by copyright and/or related rights. You are free to use this work in any way permitted by the copyright and related rights legislation that applies to your usage. For other uses, you must obtain permission from the rights-holder(s).

WEIGHTED SUPPORT VECTOR MACHINES FOR TREE SPECIES CLASSIFICATION USING LIDAR DATA

Hoang Minh Nguyen^{1,3}, Begüm Demir², and Michele Dalponte¹

¹ Dept. of Sustainable Agro-ecosystems and Bioresources, Research and Innovation Centre, Fondazione E. Mach, San Michele all'Adige, Italy.

² Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Berlin, Germany.

³ Dept. of Information Engineering and Computer Science, University of Trento, Trento, Italy.

ABSTRACT

Tree species classification at individual tree crowns (ITCs) level using remote sensing data requires the availability of a sufficient number of reliable reference samples. Two main issues that affect the classification performance are: a) an imbalanced distribution of the tree species classes; and b) the presence of unreliable samples due to field collection errors, coordinates misalignments, etc. In this study, we present a weighted Support Vector Machine (WSVM) classifier that addresses these problems by considering: 1) different weights for different classes of tree species to mitigate the effects of the class imbalance distribution; and 2) different weights for different training samples according to their importance for the considered classification problem. Experimental results obtained on a study area located in the Italian Alps showed that the proposed method increased the overall and kappa accuracies of about 2%, and the mean class accuracy of about 10% with respect to a standard SVM.

Index Terms— tree species classification, support vector machine, lidar data, sample weighting, remote sensing

1. INTRODUCTION

An accurate tree species mapping has an important role in sustainable forest management as this information can contribute to the construction of management policies for each tree species. Specifically, tree species are related to many ecosystem parameters such as biodiversity and biomass. Thus, knowing the species distribution could increase the value of forest inventories. Recently, hyperspectral images have been widely used as remote sensing data source, since they provide useful information for tree species classification through the measurement of tree spectral signatures. However, hyperspectral images are rarely available over large areas, and in order to be effectively used, they need proper weather conditions to be acquired. In the forestry and ecology domains, a type of data

that is widely used to predict forest structural characteristics is light detection and ranging (lidar) data. The information contained in lidar data about the elevation and intensity of trees canopy could be used to characterize different tree species. Thus, it can be useful to extract a wide range of features for tree species classification [1].

In order to get a detailed tree species classification map, well-known approaches are based on working at either pixel level or individual tree crown (ITC) level. In ITC level classification all the pixels belonging to a particular ITC are associated to the same information. ITCs should be automatically delineated on remote sensing data and then a unique species should be assigned to each ITC. Regarding the automatic delineation of the ITCs, a wide range of literature is existing, and the reader is referred to [2], [3] for further details.

Techniques for classification of tree species require the availability of reliable reference samples to be used in the learning phase of the classification algorithm. Reliability of the labeled training samples depends on both the quantity and quality of the available samples. In the case of tree species classification, the training set can be poor due to the following reasons:

- *Imbalanced classes*: In a forest the tree species are not present in the same amount. In details, there are: i) majority species that represent usually the dominant species and cover the majority of the canopy; and ii) minority species for which in the extreme cases only few trees per hectare are present.
- *Inaccurate positions of field data*: It is a particularly difficult task to locate the exact position of a tree in the forest and it is usually done using a global positioning system (GPS) device. In some cases, the accuracy could be particularly low, especially in a dense forest or in mountain areas where the GPS accuracy is usually low.
- *Errors in delineation of ITCs*: Methods for automatic ITCs delineations are not perfect, and thus they are associated to a delineation error that could be quite high in the case of broadleaves trees.

- *Matching errors between field data and remote sensing data:* Trees recognized in the field should be associated to an ITC delineated on the remote sensing data. Errors could occur due to positioning and also because multiple adjacent trees measured in the field could be identified as just one crown detected by analyzing the remote sensing data.
- *Biased training sets:* The quality of training samples depends on sampling schemes adopted for field surveys to collect the reference values of samples (i.e., species of the trees). In general, species of the trees from the same area are taken (i.e., nearby tree samples are collected) through the ground surveys, not representing the whole considered forest. This results in biased training sets that lead to i) poor estimations of the true underlying distributions of the samples; and ii) reduced information given to the classification algorithm by the considered training samples.

To address the above-mentioned problems, in this paper we introduce a weighted Support Vector Machine (WSVM) algorithm in the framework of tree species classification using lidar data. The WSVM gives more importance to the labeled training samples with high weights and less importance with those of lower weights while modeling the SVM separating hyperplane. To this end a strategy that defines a weight value for each training sample according to the distribution of the unlabeled samples is introduced. Experiments carried out on one dataset located in the Italian Alps show the effectiveness of our approach.

2. DATA SET DESCRIPTION

The study area (4 km²) is located in the municipality of Lavarone in the Italian Alps. The altitude varies from 1200 to 1600 m above the sea level. Dominant tree species are Norway spruce (*Picea abies* Karst.), about 47% of the total stem volume, Silver Fir (*Abies Alba* Mill.), about 36%, and European beech (*Fagus Sylvatica* L.), about 13%. Other relevant species are European larch (*Larix decidua* Mill.), and Scots pine (*Pinus sylvestris* L.) that account for about 4%.

The species, height, and locations of 5655 trees were collected in the summers of 2016 and 2018. Due to the low number of field samples for some species, the tree species were grouped into five classes: i) Sliver fir (2164 trees); ii) European larch (113 trees); iii) broadleaves (1795 trees; all the broadleaves species), iv) Norway spruce (1437 trees); and v) Scots pine (146 trees). The lidar data were acquired by an Optech ALTM 3100EA sensor with a maximum scan angle of 21 degree. The mean point density was 21.5 points per square meter for the first return while the pulse density was 14.4 pulses/m². Up to four returns per pulse were measured. Then the following steps are considered: i) pre-processing of the Lidar data, 2) delineation of the ITCs, 3) extraction of the features. Each step is explained in the following.

2.1. Lidar data preprocessing

A digital terrain model (DTM) was generated from the lidar points by the vendor using TerraScan software with a grid size of 0.5 m. The lidar point cloud was normalized to create a canopy height model (CHM) by subtracting the DTM from the z values of the lidar pulses. The intensity values were range calibrated using the same method as in [4].

2.2. ITCs delineation

The automatic ITCs delineation was performed using the method implemented in the R package *itcSegment* and used in [5]. For further details the reader is referred to [6]. The delineated ITCs were matched with the field measured trees, and the matched ITCs were divided into two sets: training set and test set (Table 1).

Table 1: Number of samples of each class in the training and test sets.

Class name	Training set	Test set
Sliver fir	232	231
European larch	35	34
Broadleaves	108	108
Norway spruce	253	252
Scots pine	37	36

2.3. Features extraction

From each delineated ITC, features were extracted in order to model the trees. In particular, two sets of features are considered: i) 46 elevation and intensity features derived directly from the lidar point cloud; and ii) three topography features derived from the DTM. The description of features are given in Table 2.

Table 2: Description of extracted features.

Feature	Description
Zmax	Maximum Z
Zmean	Mean Z
Zsd	Standard deviation of Z distribution
Zskew	Skewness of Z distribution
Zkurt	Kurtosis of Z distribution
Zentropy	Entropy of Z distribution
ZqP	Ph percentile of height distribution, with P from 5 to 95 at steps of 5
ZpcumP	Cumulative percentage of points in the Pth layer, with P from 5 to 95 at steps of 5
Itot	Sum of intensities for each return
Imax	Maximum intensity
Imean	Mean intensity
Isd	Standard deviation of intensity
Iskew	Skewness of intensity distribution
Ikurt	Kurtosis of intensity distribution
IpcumzqP	Percentage of intensity returned below the Pth percentile of Z, with P from 5 to 95
pRth	Percentage of Rth return, with R from 1 to 4
Slope	Properties of the tangent plane to the surface
DTM	Digital terrain model value (m)
aspect	Compass direction in which the slope faces to

3. WEIGHTED SVMs FOR TREE SPECIES CLASSIFICATION

In the recent years, SVM has been widely applied to classification problems in forestry applications. In the standard SVM all training samples have equal importance, and thus the incomplete (poor) training sets with the wrongly labeled training samples, which are common in tree species classification, may result in reduced classification performances. In this paper we present a weighted SVM (WSVM) algorithm, which extracts and exploits different weights for different training samples. The WSVM extends the SVM to the case for which different training samples make different contributions to the learning of the SVM separating hyperplane. WSVM was previously defined in [7][8].

Let $S = \{(\mathbf{x}_i, y_i, s_i)\}_{i=1}^N$ be a set of labeled training samples, where $\mathbf{x}_i$ is the i -th training sample, y_i is the corresponding class label from the pool of the classes $K = k_1, \dots, k_\Psi$, and s_i is its weight. Then, in the WSVM the optimization problem is defined as:

$$\begin{aligned} & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^N s_i \xi_i \\ \text{subject to: } & \begin{cases} y_i(\omega \cdot \Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \\ \xi_i \geq 0, \quad i = 1, \dots, N \end{cases} \end{aligned} \quad (1)$$

where $\Phi()$ is the mapping function that projects the samples from the original variable space to a higher dimensional space and b is the bias. C is a regularization (i.e., penalty) parameter. From (1), it can be seen that in the WSVM, the penalty value C of mis-classification for each training sample has a different weighting effect that is driven by s_i . A solution for imbalanced classes in tree species classification problems is to allow a different penalty parameter for each class of tree species. Specifically, samples of each class can be associated to different penalty values defined by class weights so that the decision boundary can pay more attention to the minority classes. To this end, we define a weight value for each sample as follows:

$$s_i = CW_k \times SW_i \quad (2)$$

where CW_k is the class weight of the class k to which the sample $\mathbf{x}_i$ is belonging, and SW_i is the weight of $\mathbf{x}_i$. Hence, the penalty parameter includes two components: i) the class weight for the different classes; and ii) the weight of each sample to reflect their importance to the class that it belongs to.

For a training set that has Ψ classes, the class weight is defined as:

$$CW_k = \frac{\max(N_k)}{N_k}, \quad k = 1, \dots, \Psi \quad (3)$$

where N_k is the number of samples that belong to the k -th class.

In the cases in which this ratio is very large, we can lose information from the majority classes. This problem can be addressed by setting the class weight of majority classes as:

$$CW_{major} = \text{mean}(CW_k) \quad (4)$$

The choice of the sample weights is important and in this paper we propose a strategy to estimate the sample weights based on the unlabeled samples. We consider that training samples associated to the highest density regions of the feature space are much more important than those located in the low density regions. This is due to the fact that i) samples in high density regions are statistically very representative of the underlying sample distribution; and ii) the classification results on samples located in the high density patterns of the feature space affect more the overall accuracy of the classification process than those obtained on samples within low density regions. Thus, we describe a strategy that defines higher weights for the samples located in the high density regions of the feature space while giving reduced weights to those that fall into the low density regions of the feature space. To this end, the importance of each training sample is adapted according to the distribution of the unlabeled samples. The proposed strategy requires measuring the density d_i around the training sample $\mathbf{x}_i$ in the feature space. The density d_i is estimated by considering the mean Euclidean distance of the training sample $\mathbf{x}_i$ to the P -nearest neighbour samples. By this way, we consider as high density region, a region where the mean Euclidean distance among the samples is small. In greater detail the steps of estimation of the sample weights are as follows:

1. The Euclidean distance between each training sample and all the available unlabeled samples is computed.
2. The mean value of the distance between each training sample $\mathbf{x}_i$ and the P nearest unlabeled samples in the feature space is computed. The value of P varies depending on the class and is estimated as the square root of the total number of samples of each class in the training set.
3. Weights are assigned according to the mean distance calculated at step 2: a small mean distance is related to the fact that the density around that sample is high, and thus, a high value of the weight is assigned to that sample.
4. Weights are normalized in the range $[0,1]$.

By considering the proposed class and sample weighing in the framework of the WSVM, the WSVM adopts the highest penalty values for wrong classifications of the most informative labeled samples and lower penalty values for the less informative samples. As a result, the WSVM

generates a more robust separating hyperplane skewed towards the important samples by allowing some misclassification for the less important samples.

4. EXPERIMENTAL RESULTS

Results achieved with WSVM are compared with those obtained by the standard SVM. Results of each method are provided in terms of: i) overall accuracy; ii) mean class accuracy; iii) kappa coefficient; and iv) producer's accuracy of each class. Radial Basis Function (RBF) kernel was used for both methods. The values for the regularization parameter and the spread of the RBF kernel were chosen performing a model selection. Table 3 shows the classification accuracies obtained by the SVM and the WSVM. From the table, one can observe that the accuracy obtained by the SVM is significantly improved by using the WSVM. The proposed WSVM provides 2.4% higher overall accuracy, 3.9% higher kappa coefficient and 10.2% higher mean class accuracy with respect to the traditional SVM. These results clearly show that the sample and class weighting strategies implemented in the WSVM rectifies the performance loss of the SVM.

Table 3: Overall accuracy (OA), mean class accuracy (MCA) and kappa coefficient (KC) [in %] obtained by the standard SVM and the WSVM.

Method	OA	MCA	KC
SVM	80.2	68.8	71.1
WSVM	82.6	79.0	75.0

Table 4: Producer's accuracy [in %] for each class.

Tree Species	SVM	WSVM
Sliver fir	86.1	88.7
Broadleaves	84.3	76.9
European larch	38.2	67.6
Norway spruce	82.5	82.9
Scots pine	52.8	72.2

Table 4 shows the producer's accuracy (PA) associated to each class. European larch and Scots pine (which are the minority classes in the considered dataset) were not properly classified by the standard SVM. WSVM improves significantly the PAs of these two classes. In details, the improvements are 29.4% for European larch and 19.4% for Scots pine with respect to the standard SVM classifier. On the other hand, the accuracy of the Broadleaves class decreased by 7.4%.

5. CONCLUSIONS

In this paper, a weighted SVM (WSVM) classifier has been presented for tree species classification at ITC level using

lidar data. The considered WSVM extracts and exploits both class weights and sample weights defined based on a strategy that relies on the distribution of the unlabeled samples in the feature space. According to the adopted strategy, higher weights are given for the samples located in the high density regions of the feature space while giving reduced weights to those that fall into the low density regions of the feature space. Experimental results show that that the WSVM algorithm with class and sample weights strategy allows one: i) to significantly improve the classification accuracy of minority classes and ii) efficiently solves the limitations of the imbalanced classes and of the presence of unreliable training samples. As a future work we plan to improve the weighting scheme in order to keep the improvement on the minority classes without reducing the accuracy on the majority ones.

6. REFERENCES

- [1] I. Korpela, H. O. Ørka, M. Maltamo, T. Tokola, and J. Hyyppä, "Tree Species classification using airborne LiDAR – Effects of Stand and Tree Parameters, Downsizing of Training Set, Intensity Normalization, and Sensor Type," *Silva Fenn.*, vol. 44, no. 1432, pp. 319–339, 2010.
- [2] Z. Zhen, L. Quackenbush, and L. Zhang, "Trends in Automatic Individual Tree Crown Detection and Delineation—Evolution of LiDAR Data," *Remote Sens.*, vol. 8, no. 4, p. 333, Apr. 2016.
- [3] E. Lindberg and J. Holmgren, "Individual Tree Crown Methods for 3D Data from Remote Sensing," *Curr. For. Reports*, vol. 3, no. 1, pp. 19–31, Mar. 2017.
- [4] M. Dalponte, L. Ene, T. Gobakken, E. Næsset, and D. Gianelle, "Predicting Selected Forest Stand Characteristics with Multispectral ALS Data," *Remote Sens.*, vol. 10, no. 4, p. 586, Apr. 2018.
- [5] M. Dalponte and D. A. Coomes, "Tree-centric mapping of forest carbon density from airborne laser scanning and hyperspectral data," *Methods Ecol. Evol.*, vol. 7, no. 10, pp. 1236–1245, Oct. 2016.
- [6] M. Dalponte, L. Frizzera, H. O. Ørka, T. Gobakken, E. Næsset, and D. Gianelle, "Predicting stem diameters and aboveground biomass of individual trees using remote sensing data," *Ecol. Indic.*, vol. 85, no. May 2017, pp. 367–376, 2018.
- [7] R. Batuwita and V. Palade, "FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 3, pp. 558–571, Jun. 2010.
- [8] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013.