Review Article

Random forest in remote sensing: A review of applications and future directions

Mariana Belgiu [a,*], Lucian Drăguţ [b]

[a] Department of Geoinformatics – Z_GIS, Salzburg University, Schillerstrasse 30, 5020 Salzburg, Austria
[b] West University of Timisoara, Department of Geography, Vasile Parvan Avenue, 300223 Timisoara, Romania

ABSTRACT

A random forest (RF) classifier is an ensemble classifier that produces multiple decision trees, using a randomly selected subset of training samples and variables. This classifier has become popular within the remote sensing community due to the accuracy of its classifications. The overall objective of this work was to review the utilization of RF classifier in remote sensing. This review has revealed that RF classifier can successfully handle high data dimensionality and multicolinearity, being both fast and insensitive to overfitting. It is, however, sensitive to the sampling design. The variable importance (VI) measurement provided by the RF classifier has been extensively exploited in different scenarios, for example to reduce the number of dimensions of hyperspectral data, to identify the most relevant multisource remote sensing and geographic data, and to select the most suitable season to classify particular target classes. Further investigations are required into less commonly exploited uses of this classifier, such as for sample proximity analysis to detect and remove outliers in the training samples.

1. Introduction

Remote sensing has proved its value in many fields but the success of any image classification depends on various factors, including the choice of a suitable classification procedure (Lu and Weng, 2007). Supervised classifiers are widely used since they are more robust than model-based approaches (Niemeyer et al., 2014). These classifiers are able to learn the characteristics of target classes from training samples and to identify these learned characteristics in the unclassified data. An efficient supervised classifier needs to address the challenges (Millard and Richardson, 2015) involved in (1) mitigating the Hughes phenomenon (i.e. the "curse of dimensionality"), which occurs when the number of variables is much larger than the number of training samples (Ghosh et al., 2014), (2) dealing with the nonlinearity of variables, (3) dealing with imbalanced training samples and noise in both training samples and unlabelled data, and (4) reducing computation time (Gislason et al., 2006).

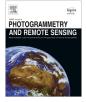Over the last two decades the use of the random forest (RF) classifier (Breiman, 2001) has received increasing attention due to the excellent classification results obtained and the speed of processing (Du et al., 2015; Pal, 2005; Rodriguez-Galiano et al., 2012). The RF classifier yields reliable classifications using predictions derived from an ensemble of decision trees (Breiman, 2001). Furthermore, this classifier can be successfully used to select and rank those variables with the greatest ability to discriminate between the target classes. This is an important asset given that the high dimensionality of remotely sensed data makes the selection of the most relevant variables a time-consuming (Körting et al., 2013), error prone, and subjective task (Belgiu et al., 2014a).

A number of studies have systematically investigated the utilization of the RF classifier for hyperspectral data classification (Ham et al., 2005) and land cover (LC) classification of Enhanced Thematic Mapper (ETM+) (Pal, 2005) or Multispectral Scanner (MSS) and Digital Elevation Model (DEM) data (Gislason et al., 2006). There has, however been no publication to date dedicated to summarizing the use of this versatile and efficient classifier in different application scenarios.

The objective of this work has therefore been to summarize the use of the RF classifier in remote sensing, with special attention to its parameterization and its sensitivities to changes in sampling procedures, to the size and representativeness of training sample

* Corresponding author at: Schillerstr. 30, 5020 Salzburg, Austria. Tel.: +43 662 8044 7518; fax: +43 662 8044 7560.

E-mail addresses: mariana.belgiu2@sbg.ac.at (M. Belgiu), lucian.dragut@fulbrightmail.com (L. Drăguţ).

sets, and to noise in the data. Following a brief overview of ensemble classifiers in remote sensing, we present the characteristics of the RF classifier (Section 2). The RF description is intended for readers who have limited experience with machine learning classifiers. Section 3 is dedicated to the use of the RF classifier with data from different sensors. In the following section we then address the sensitivity of the RF classifier to sampling procedures and to feature filtering. Section 5 includes an evaluation of the performance of the RF classifier compared to other mainstream classifiers. Possible objectives for future research are then discussed in Section 6 and conclusions are presented in Section 7. This article is concerned specifically with the use of RF classifier for classification tasks, and regression models are therefore not addressed. We use the terms 'variable' and 'feature' interchangeably to refer to class attributes or properties identified in the remotely sensed data.

2. Ensemble classifiers in remote sensing

Supervised parametric classifiers such as Maximum Likelihood Classification (MLC) deliver excellent results when dealing with unimodal data. However, they have limitations when dealing with multi-modal input datasets because these classifiers assume a normal data distribution (Liu et al., 2011). Non-parametric supervised classifiers such as the Classification and Regression Tree (CART), Support Vector Machine (SVM) (Mountrakis et al., 2011), and Artificial Neural Network (ANN) (Mas and Flores, 2007) classifiers do not make any assumptions regarding frequency distribution and have therefore become increasingly popular for classifying remotely sensed data, which rarely have normal distributions.

Because the nature and causes of spatial variation in images are not understood, the analysis has been limited to the empirical association between surface phenomenon and patterns in images (Woodcock et al., 1988), with the implicit assumption that reality has a consistent spectral response in imagery. This assumption is often violated, however, as a consequence of the complex interplay between factors like scene complexity, scale and aggregation (Marceau et al., 1994). Therefore, simple classifiers may reach their limits in many applications.

In the last years the attention of the remote sensing community has turned to ensemble classifiers (Miao et al., 2012; Gislason et al., 2006). These classifiers can be based on an individual supervised classifier or on a number of different supervised classifiers that are trained using bagging (Breiman, 1996) or boosting approaches (Schapire, 1990; Freund and Schapire, 1997), or variations of these approaches. In the bagging approach (also known as the bootstrap aggregation approach) each classifier in the ensemble is trained on a random subset of a training samples set, whereas in the boosting approach the ensemble classifiers are trained iteratively using all of the training samples, increasing the weightings for the incorrectly classified samples during the training procedure. Previous work has shown that using boosting and bagging ensemble methods achieved greater accuracy than using single classifiers such as decision tree classifiers (Briem et al., 2002; Miao et al., 2012), as well as being more stable and robust to noise in the training data (DeFries and Chan, 2000). In an experimental study using data from various application domains, Dietterich (2000) established that boosting is more accurate than bagging. Boosting approaches have been shown to reduce classification variance and bias (Gislason et al., 2006). However, they require large computational resources, overfit if there are insufficient training samples, and are sensitive to any outliers present in the training samples (Xu et al., 2014). Examples of boosting methods commonly used in remote sensing are AdaBoost (Chan and Paelinckx, 2008; Miao et al., 2012) and JointBoost (Guo et al., 2015). In contrast, bagging approaches reduce the classification variance but they have little

effect on the classification bias (Briem et al., 2002; DeFries and Chan, 2000).

2.1. The random forest classifier

The RF classifier is an ensemble classifier that uses a set of CARTs to make a prediction (Breiman, 2001). The trees are created by drawing a subset of training samples through replacement (a bagging approach). This means that the same sample can be selected several times, while others may not be selected at all (Fig. 1A).

About two thirds of the samples (referred to as *in-bag* samples) are used to train the trees (Fig. 1A) with the remaining one third (referred to as *out-of-the bag* samples) are used in an internal cross-validation technique for estimating how well the resulting RF model performs (Breiman, 2001).

This error estimate is known as the out-of-bag (OOB) error. Each decision tree is independently produced without any pruning and each node is split using a user-defined number of features ($Mtry$), selected at random. By growing the forest up to a user-defined number of trees ($Ntree$), the algorithm creates trees that have high variance and low bias (Breiman, 2001). The final classification decision is taken by averaging (using the arithmetic mean) the class assignment probabilities calculated by all produced trees. A new unlabelled data input is thus evaluated against all decision trees created in the ensemble and each tree votes for a class membership. The membership class with the maximum votes will be the one that is finally selected (Fig. 1B).

As mentioned above, two parameters need to be set in order to produce the forest trees: the number of decision trees to be generated ($Ntree$) and the number of variables to be selected and tested for the best split when growing the trees ($Mtry$). Theoretical and empirical research has highlighted that classification accuracy is less sensitive to $Ntree$ than to the $Mtry$ parameter (Ghosh et al., 2014; Kulkarni and Sinha, 2012). Since RF classifier is computationally efficient and does not overfit, $Ntree$ can be as large as possible (Guan et al., 2013). The majority of the studies reported in this review set the $Ntree$ value to 500 because the errors stabilize before this number of classification trees is achieved (Lawrence et al., 2006). Another reason for this value being commonly used could be the fact that 500 is the default value in the R package for random forests, which is called "*randomForest*" and is the most popular RF implementation. Other investigators have used different values to $Ntree$, such as 5000 (Adelabu et al., 2014a; Díaz-Uriarte and De Andres, 2006; Millard and Richardson, 2015; Nitze et al., 2015; Stumpf and Kerle, 2011), 1000 (Colditz, 2015; Reese et al., 2014; Sesnie et al., 2010), or 100 (Guan et al., 2013). In a study dedicated to polarimetric synthetic aperture radar (PolSAR) image classification, Du et al. (2015) investigated the sensitivity of the RF classifier to the number of trees (from 10 to 200 trees in steps of 10) and showed that this parameter has no influence on the classification results. Similar results were reported by Topouzelis and Psyllos (2012), who used random forests to classify oil spills from synthetic aperture radar (SAR) data and found that an ensemble of 70 trees was sufficient for this classification purpose as the classification no longer improved as the number of trees increased above this threshold. Based on the results published to date, we suggest that the default value of 500 for $Ntree$ is an acceptable value when using the RF classifier on remotely sensed data.

The $Mtry$ parameter is usually set to the square root of the number of input variables (Gislason et al., 2006). Ghosh et al. (2014) set $Mtry$ to the total number of available variables, but this approach increases the computation time as the algorithm has to compute the information gain contributed by all of the variables used to split the nodes.