ISPRS Journal of Photogrammetry and Remote Sensing 105 (2015) 155–168

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin

Andrew Mellor[a,c,*], Samia Boukir[b], Andrew Haywood[c], Simon Jones[a]

[a] School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, VIC 3001, Australia
[b] G&E Laboratory (EA 4592), IPB/University of Bordeaux, 1 allée F. Daguin, 33670 Pessac, France
[c] Victorian Department of Environment, Land, Water and Planning, 8 Nicholson Street, East Melbourne, VIC 3002, Australia

ARTICLE INFO

Article history:
Received 25 March 2014
Received in revised form 15 January 2015
Accepted 24 March 2015
Available online 27 April 2015

Keywords:
Ensemble margin
Training data
Classification
Remote sensing
Imbalance
Mislabelling

ABSTRACT

Studies have demonstrated the robust performance of the ensemble machine learning classifier, random forests, for remote sensing land cover classification, particularly across complex landscapes. This study introduces new ensemble margin criteria to evaluate the performance of Random Forests (RF) in the context of large area land cover classification and examines the effect of different training data characteristics (imbalance and mislabelling) on classification accuracy and uncertainty. The study presents a new margin weighted confusion matrix, which used in combination with the traditional confusion matrix, provides confidence estimates associated with correctly and misclassified instances in the RF classification model. Landsat TM satellite imagery, topographic and climate ancillary data are used to build binary (forest/non-forest) and multiclass (forest canopy cover classes) classification models, trained using sample aerial photograph maps, across Victoria, Australia. Experiments were undertaken to reveal insights into the behaviour of RF over large and complex data, in which training data are not evenly distributed among classes (imbalance) and contain systematically mislabelled instances. Results of experiments reveal that while the error rate of the RF classifier is relatively insensitive to mislabelled training data (in the multiclass experiment, overall 78.3% Kappa with no mislabelled instances to 70.1% with 25% mislabelling in each class), the level of associated confidence falls at a faster rate than overall accuracy with increasing amounts of mislabelled training data. In general, balanced training data resulted in the lowest overall error rates for classification experiments (82.3% and 78.3% for the binary and multiclass experiments respectively). However, results of the study demonstrate that imbalance can be introduced to improve error rates of more difficult classes, without adversely affecting overall classification accuracy.

1. Introduction

Accurate spatially explicit classification maps are important sources of information for natural resource land managers and forest monitoring programs. Land management agencies typically monitor and report on large areas (i.e. regional or continental scale, covering millions of hectares) relying on the interpretation of large complex remotely sensed data, calibrated and validated using, typically, a limited amount of ground reference data (Lippitt et al., 2008). Studies have demonstrated the successful application of ensemble machine learning classifiers, such as Random Forests (RF), integrating remote sensing (satellite imagery) and ancillary spatial data, to improve supervised classification accuracy of forest and other natural environment land cover maps (Cutler et al., 2007; Mellor et al., 2013; Rodriguez-Galiano et al., 2012), for which conventional parametric statistical classification techniques might not be appropriate (Gislason et al., 2006). In ensemble classification, multiple (base) classifiers are constructed. From the ensemble, a final class is determined by, for example, averaging or a majority vote. In machine learning, the margin theory examines the proximity of data points to decision boundaries. Margin theory is a means by which to understand and evaluate ensemble classification and can be used to estimate confidence in the classification outcome (Schapire et al., 1998). Such ancillary information is important, particularly when relying on satellite image derived maps for scientific inference (McRoberts, 2011). The characteristics

* Corresponding author at: School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, VIC 3001, Australia.

E-mail address: andrew.mellor@rmit.edu.au (A. Mellor).

http://dx.doi.org/10.1016/j.isprsjprs.2015.03.014

of training data is a fundamental consideration when constructing any supervised classifier (including ensemble machine learning). Learning from imbalanced training data (i.e. unevenly distributed data between classes) is a common problem (Japkowicz and Stephen, 2002). Machine learning algorithms, such as RF, are constructed to minimize the overall classification error rate and imbalanced training data can result in poor accuracy for minority classes (Chen et al., 2004). Furthermore, it is assumed that, in its implementation, the classifier is run using data drawn from the same distribution as the training data (Provost, 2000). In RF, decision trees are induced using bootstrap samples of training data (Breiman, 2001) and in situations where training data includes only a minority of training data samples for a particular class (relative to other classes), it is likely that a bootstrap sample may include few or even no samples from this class and hence fewer leaves describing the minority class, resulting in poor classification accuracy for the minority class prediction (Chen et al., 2004) as well as weaker confidence estimates (He and Garcia, 2009).

The imbalance training data problem is common in large area natural resource applications using remote sensing (e.g. forest classification), whereby within reference data, rare land cover or forest classes may be under-represented relative to more abundant classes, due to the time and cost resource constraints of collecting enough representative training samples. Studies have shown balanced datasets improve overall classification compared to imbalanced data (Estabrooks et al., 2004; Weiss and Provost, 2003). Several techniques have been demonstrated to address the imbalance training data problem. These include down-sampling majority classes (Freeman et al., 2012) and weighting rare training observations more highly than common classes (Chen et al., 2004). Techniques involving over-sampling the minority class through replication of samples to match the quantity of majority class training samples (Ling and Li, 1998) and a combination of over-sampling (minority) and down-sampling (majority) training classes (Chawla et al., 2002) have also been explored.

Training data class mislabelling (or noise) is another important consideration in using bagging ensemble algorithms such as RF. This is an issue that often adversely affects machine learning algorithms (Guo, 2011). In large area remote sensing classification for forest monitoring programs, training data typically include ground-based (i.e. field data collection) (Lillesand and Kiefer, 1994) or data sampled from remote sensing imagery of a higher spatial resolution, such as very high resolution satellite imagery (e.g. Quickbird) or digital aerial photography (Wulder, 1998). Deriving training data using manual and semi-automated mapping from high spatial resolution imagery are methods which are prone to a variety of sources of labelling error and bias. These sources include interpreter bias and inconsistency, spatial resolution (scale), geometric and radiometric variability, and error associated with temporal discontinuity between training data (i.e. aerial photography acquisition date or season) and satellite imagery used for classification (Morgan et al., 2010). Other training data labelling errors are associated with inconsistency of vegetation classification methods, techniques and spatial resolution (Bradley and Friedl, 1996). In forest environments, common training data class mislabelling errors are caused by the similarity of forest types as their signatures appear in aerial photography (Delaney and Skidmore, 1998).

For their application in an operational setting (such as a large area forest monitoring program), it is important that machine learning classifiers are resilient to mislabelling in training data (Lippitt et al., 2008). Studies have demonstrated the relative resilience of bagging ensemble classifiers, such as RF, to training data noise (class mislabelling) (DeFries and Cheung-Wai Chan, 2000). In evaluating machine learning algorithms for land cover change mapping, Rogan et al. (2008) investigated the effect of artificially introduced training data noise to classification accuracy. Their study found the addition of 10% noise reduced accuracy of decision tree classifiers S-Plus and C4.5 by 7% and 20% respectively. In a land cover classification study, Rodriguez-Galiano et al. (2012) found the RF classifier performance (overall classification error) to be relatively insensitive for up to 20% deliberately mislabelled training instances, above which error rate increased exponentially. Na et al. (2009) reported a reduction in RF overall accuracy by almost 50% associated with a 30% increase in the amount of artificial noise.

In this paper, we examine how training data class imbalance and class mislabelling affect RF performance in the context of large area forest classification in an operational land management agency setting. This was achieved across diverse and complex forest ecosystems and topography, dominated by open canopy sclerophyll forests and woodland. We evaluate RF performance associated with training data characteristics through a new perspective involving ensemble margins. The magnitude of ensemble margin is usually interpreted as a measure of confidence in classification prediction and significant work has been published about bounding and reducing prediction error based on the classification margin (Guo, 2011; Schapire et al., 1998). The nature of a training set can have a major impact on classification accuracy (Foody, 1999) and the margin ensemble can be used to understand how training data characteristics can affect classification outcomes. Foody (2002) emphasizes the need for more accuracy assessment information (including confidence measures) to be provided with land cover and other remote sensing derived classification maps, to aid user interpretation and application. The value of very large area mapping is ultimately limited by poor quality accuracy assessment and reporting (Foody, 2002). In this study, we evaluate new ensemble margin statistics as a means of providing distinct information about margin distribution and classification prediction confidence and supplementing traditional measures of classification performance. Furthermore, we introduce a novel method for assessing classification uncertainty through the use of an ensemble margin weighted confusion matrix, that to the best of our knowledge is used for the first time in land cover classification using remote sensing and ancillary geospatial data.

2. Random Forests

Random Forests (RF) uses a bootstrap aggregation technique (bagging) (Breiman, 1996) to generate sub-sets of training data with which to build an ensemble of decision trees (base classifiers). The bagging process involves resampling the original training set with replacement, resulting in a greater diversity of decision trees, thereby improving classifier stability and accuracy. Moreover, in constructing trees, as some training data instances may be used more than once or not at all, correlation between trees is reduced, and as a result, RF is more robust to variations in input data and less sensitive to mislabeled training data or over-fitting (Pal, 2005; Rodriguez-Galiano et al., 2012).

In constructing each decision tree, at each node (split) a randomly selected subset of model predictor variables are evaluated for partitioning the data into increasingly homogeneous subsets – the variable used to split the data is that which results in the greatest increase in data purity. Increasing the number of predictor variables selected for tree construction results in stronger individual decision trees, but with increased correlation between trees, model accuracy is reduced (Rodriguez-Galiano et al., 2012). Therefore, to minimize the generalization error, it is necessary to optimize this parameter, together with the number of decision trees in the ensemble. Tree building continues until there are no further gains in purity. A response variable can be predicted as an average (continuous variable classification) or model vote