ISPRS Journal of Photogrammetry and Remote Sensing 129 (2017) 151–161

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Exploring diversity in ensemble classification: Applications in large area land cover mapping

Andrew Mellor [a,*], Samia Boukir [b]

[a] School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, VIC 3001, Australia
[b] Bordeaux INP, G&E, EA 4592, F-33600 Pessac, France

ARTICLE INFO

Article history:
Received 27 November 2016
Received in revised form 25 April 2017
Accepted 25 April 2017

Keywords:
Diversity
Ensemble
Margin
Random forests
Training data
Classification
Land cover
Remote sensing

ABSTRACT

Ensemble classifiers, such as random forests, are now commonly applied in the field of remote sensing, and have been shown to perform better than single classifier systems, resulting in reduced generalisation error. Diversity across the members of ensemble classifiers is known to have a strong influence on classification performance - whereby classifier errors are uncorrelated and more uniformly distributed across ensemble members. The relationship between ensemble diversity and classification performance has not yet been fully explored in the fields of information science and machine learning and has never been examined in the field of remote sensing. This study is a novel exploration of ensemble diversity and its link to classification performance, applied to a multi-class canopy cover classification problem using random forests and multisource remote sensing and ancillary GIS data, across seven million hectares of diverse dry-sclerophyll dominated public forests in Victoria Australia. A particular emphasis is placed on analysing the relationship between ensemble diversity and ensemble margin - two key concepts in ensemble learning. The main novelty of our work is on *boosting* diversity by emphasizing the contribution of lower margin instances used in the learning process. Exploring the influence of tree pruning on diversity is also a new empirical analysis that contributes to a better understanding of ensemble performance. Results reveal insights into the trade-off between ensemble classification accuracy and diversity, and through the ensemble margin, demonstrate how inducing diversity by targeting lower margin training samples is a means of achieving better classifier performance for more difficult or rarer classes and reducing information redundancy in classification problems. Our findings inform strategies for collecting training data and designing and parameterising ensemble classifiers, such as random forests. This is particularly important in large area remote sensing applications, for which training data is costly and resource intensive to collect.

1. Introduction

Across a broad range of applications, ensemble classification systems (also known as multiple or committee classifiers) have been shown to produce better results than single expert systems (Polikar, 2006) and achieve reduced generalisation error (Opitz and Maclin, 1999; Tumer and Ghosh, 1996). In remote sensing application areas, such as ecology and natural resource management, ensemble classifiers, like Random Forests (RF) (Breiman, 2001), have become increasingly popular. Incorporating remote sensing data and ancillary continuous and categorical biophysical spatial data, RF has been applied in a variety of large area land cover (Rodriguez-Galiano et al., 2012) and forest attribution studies, including biomass (Baccini et al., 2008), canopy height (Wilkes et al., 2015), canopy cover (Mellor et al., 2015) and species (Dalponte et al., 2013; Evans and Cushman, 2009). The RF classifier builds an ensemble of decision trees (known as base classifiers or ensemble members) and assigns classification through voting or averaging among these ensemble members.

Diversity between ensemble members is considered a key factor affecting overall classification performance (Ham et al., 2005; Kapp et al., 2007; Kuncheva and Whitaker, 2003; Melville and Mooney, 2005). Ensemble classifiers which achieve higher overall classification rates are those in which misclassified instances (errors) made by ensemble members are uncorrelated (Banfield et al., 2005; Elghazel et al., 2011). Ensemble classifiers are often

* Corresponding author.
E-mail addresses: andrew.mellor@rmit.edu.au (A. Mellor), samia.boukir@ipb.fr (S. Boukir).

http://dx.doi.org/10.1016/j.isprsjprs.2017.04.017

more accurate than their component (base) classifiers, and diversity is greater, if errors made by ensemble members are uncorrelated (Díez-Pastor et al., 2015; Hansen and Salamon, 1990) and more uniformly distributed (Banfield et al., 2005). While ensemble diversity has been studied in the fields of information science and machine learning, to the best of our knowledge, the relationship between ensemble diversity and classification performance has not been actively explored in remote sensing. Gaining a greater insight into the role of diversity in ensemble classification is important, not least because of the increasing popularity of ensemble classifiers, such as random forests in this field (Belgiu and Drăguţ, 2016). Moreover, while advances in remote sensing science and technology (such as new sensors and image analysis techniques) seek to address land cover mapping (classification) error, the availability of suitable reference (training and test) data is a fundamental requirement in supervised image classification (Foody et al., 2016). Training and test data are also expensive (Pflugmacher et al., 2012), and as such, there are significant benefits to designing classifiers which make more efficient use of training data, such as reducing class information redundancy and maximizing the application of training data for classes which are rarer or more difficult to classify.

In this paper, we explore the relationship between ensemble diversity and classification performance in the context of large area land cover classification across complex forest ecosystems and topography, using remote sensing and ancillary spatial data. We focus on the relationship between ensemble diversity and ensemble margin, two fundamental theories in ensemble learning. Applying the RF classifier, we evaluate different ways of inducing diversity in ensemble classification to improve classification performance and efficiency, and reduce training data redundancy. The main novelty of our work is on *boosting* diversity by targeting lower margin training samples (which represent class decision boundaries or more difficult or rarer classes) in the learning process. We also propose a new empirical analysis that explores the influence of tree pruning, and decision tree depth, on diversity, which leads to a better understanding of RF classifier performance. The findings of this work may be used to inform training data collection strategies and to design more efficient classification. Key concepts used in the paper are introduced in sections II through IV. Section V describes the study area and data, and experiments, results and discussion are included in sections VI through VIII.

2. Random forests

Random forests (Breiman, 2001) is a popular ensemble classifier (Belgiu and Drăguţ, 2016), which generates decision trees using sub-sets of bootstrap-aggregated training data (sampling with replacement), otherwise known as bagging. These decision trees represent diverse base classifiers, which are combined into an ensemble. In addition to bagging, diversity is induced through the random selection of a sub-set of input (explanatory or predictor) variables which are evaluated for partitioning data at each decision tree node (Elghazel et al., 2011). A response variable is predicted as a modal vote (for categorical data) or average (for continuous variables) among the ensemble decision trees. Studies have reported that the number of variables randomly sampled to split training data at decision tree nodes does not affect classification rates (and other RF performance measures) (Cutler et al., 2007).

3. Ensemble margin

The margin provides a measure of confidence in ensemble classification (Guo et al., 2011; Mellor et al., 2014, 2015) and is an important concept in ensemble methods (Schapire and Freund, 1998). The ensemble margin is calculated as the difference between the number of votes assigned to different classes by the base classifiers in an ensemble. The unsupervised version of Schapire's margin (Eq. (1)) of a sample x is the difference between the number of votes (respectively V_{c1} and V_{c2}) assigned to the first and second most popular classes (respectively c_1 and c_2), normalised by the number of base classifiers (T) in the ensemble, regardless of true class labels (Guo and Boukir, 2013). It has been used in large area remote sensing classification as an ancillary measure of random forest classifier performance (Mellor et al., 2014, 2015).

$$margin(x) = \frac{V_{c1} - V_{c2}}{T}, \; 0 \leqslant margin(x) \leqslant 1 \tag{1}$$

Correctly classified training instances with high margin values (i.e. close to 1) represent instances located away from class decision boundaries and can contain a high degree of redundant information in a classification problem. Conversely, training instances with low margin values (i.e. close to 0) are located near decision boundaries and are more informative in a classification task. Unlike Schapire's margin (Schapire and Freund, 1998), which is supervised and calculated as the difference between votes assigned to the true class and those assigned to the most voted class that is different from the true class, class labels in the unsupervised margin (Guo and Boukir, 2013) (applied in this study) are not of significance. As such, the unsupervised margin may be more robust to noise (Guo, 2011). The mean margin (Eq. (2)) is a descriptive statistic for the ensemble margin, calculated from the unsupervised margin values (Eq. (1)), which can be used as a confidence measure for model performance (Mellor et al., 2014, 2015). This measure ranges from −1 (weakest ensemble classifier) to +1 (strongest ensemble classifier).

$$\mu = \frac{(n_c\mu_c) - (n_m\mu_m)}{n_c + n_m}, \quad -1 \leqslant \mu \leqslant 1 \tag{2}$$

where n_c is the number of correctly classified instances, n_m is the number of misclassified instances, μ_c and μ_m are mean margins for correctly and misclassified instances respectively.

4. Ensemble diversity

Ensemble diversity is important for majority vote accuracy and aims at decreasing the probability of identical errors (correlation between ensemble members). While it is accepted that diversity improves overall ensemble classification performance, there is no general agreement on how it should be quantified or dealt with (Kapp et al., 2007), nor is there a widely perceived concept of diversity or theoretical framework which supports the development of methods to capture diversity among classifiers (Bi, 2012). A review by Kuncheva and Whitaker (Kuncheva and Whitaker, 2003) compared ten measures of pairwise and non pair-wise diversity, finding most to be highly correlated. In pairwise measures, the diversity values between all pairs of classifiers are initially calculated. The overall diversity measure value is then computed as the mean of all pair-wise values. Unlike pairwise measures, non-pairwise measures are calculated by counting a statistical value of all ensemble classifiers to measure the whole diversity. Therefore they generally run much faster than pairwise measures (Guo, 2011). Diversity can be measured at the output (prediction) level, the input (training data) level and at the structure or parameter level (Guo and Boukir, 2014). In this study, we measure diversity at the output level (i.e. diversity among the class labels assigned across each of the base classifiers in the ensemble), using *KW (Kohavi and Wolpert) variance* (Kohavi and Wolpert, 1996), a