ENSEMBLE DIVERSITY ANALYSIS ON REMOTE SENSING DATA CLASSIFICATION USING RANDOM FORESTS

Samia Boukir[1], *Andrew Mellor*[2]

[1]Bordeaux INP, G&E, EA 4592, F-33600 Pessac, France
[2]School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, VIC 3001, Australia
Email: samia.boukir@ipb.fr , andrew.mellor@rmit.edu.au

ABSTRACT

Ensemble classifiers perform better than single classifiers and result in reduced generalisation error. Diversity across ensemble members is a key factor affecting classification performance. Here, an original exploration of the relationship between ensemble diversity and classification performance applied to large area remote sensing classification, using random forests, is undertaken. Results demonstrate how targeting lower margin training samples is both a strategy for inducing diversity in ensemble classifiers and achieving better classifier performance for difficult or rare classes, and a way to reduce data redundancy.

Index Terms **—** Ensemble diversity, ensemble margin, classification, training data selection, land cover

1. INTRODUCTION

Ensemble classification systems have been shown to produce better results than single expert systems [1] and achieve reduced generalisation error [2], [3]. In the field of remote sensing, ensemble classifiers, such as Random Forests (RF) [4], have become increasingly popular. RF has been applied in a variety of land cover [5] and forest attribution studies [6], [7]. The RF classifier [4] builds an ensemble of decision trees (known as base classifiers) using sub-sets of bootstrap-aggregated training data (sampling with replacement), or bagging. These decision trees represent diverse base classifiers, which are combined in a single ensemble. In addition to bagging, diversity is induced through the random selection of a sub-set of input variables evaluated for partitioning data at each decision tree node [8]. RF assigns classification through majority voting among its ensemble members.

Diversity between ensemble members is considered a key factor affecting overall classification performance [9]–[11]. Ensemble classifiers achieve higher classification rates and diversity is greater if misclassified instances made by ensemble members are uncorrelated [8], [12], [13].

In this paper, we present a novel exploration of ensemble diversity and its link to classification performance in the context of large area (i.e. millions of hectares) land cover classification using the RF classifier. 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 (which represent class decision boundaries or more difficult or rare classes) used in the learning process.

2. ENSEMBLE MARGIN

The margin is an important concept in ensemble methods [14] such as RF, providing a measure of classification confidence [7], [15], [16]. An alternative margin function, which does not require the true class labels, is an unsupervised version of the classic margin [14] which may be more robust to class noise [17], [18]. This ensemble margin is calculated as the difference between the maximum number of base classifier votes assigned to a class and the number of votes assigned to the second most voted for class by the ensemble. Equation 1 shows how it is calculated where V_{c1} represents the number of votes for the most voted class c_1 for instance x, V_{c2} the number of votes for the second most popular class c_2 and T the number of base classifiers in the ensemble [17], [18]. This margin ranges from 0 to 1. In previous work [7], [16], it has been used in large area remote sensing classification as an ancillary measure of random forest classifier performance.

$$\mathrm{margin(x)} = \frac{\mathrm{V_{c1}} - \mathrm{V_{c2}}}{\mathrm{T}} \qquad (1)$$

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 (close to 0) are often located near decision boundaries and are more informative in a classification task.

The mean margin (equation 2) is a descriptive statistic, introduced in previous work [7], [16], for the ensemble margin, calculated from the unsupervised margin (equation 1) values, which can be used as a confidence measure of

model performance. This measure ranges from -1 (weakest ensemble classifier) to +1 (strongest ensemble classifier).

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

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

3. 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 ensemble classification performance, there is no general agreement on how it should be quantified or dealt with [9], nor a widely perceived concept of diversity or theoretical framework which supports the development of methods to capture diversity among classifiers [19]. A review by Kuncheva and Whitaker [10] 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. Then, the overall diversity measure value is computed as the mean of all pair-wise values. Unlike pairwise measures, non-pairwise 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. Diversity can be measured at the output level, the input level and at the structure or parameter level [20].

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 [21], a popular non-pairwise diversity measure, which can be expressed as equation (3) [9].

$$KW = \frac{1}{NT^2}\sum_{j=1}^{N} t(x_j)\ (T - t(x_j)) \quad (3)$$

where diversity increases with KW variance, T is the size of the ensemble of classifiers, $t(x_j)$ the number of classifiers that correctly recognise sample x_j and N represents the number of samples.

The minimum value for KW diversity is 0 (lowest diversity), it occurs when all the T ensemble members correctly classify all of the samples (overall accuracy of 100% and mean margin μ of 1), or conversely, when all of the T ensemble members misclassify all of the samples (overall accuracy of 0% and negative mean margins μ ranging from -1, in binary classification, to 0). KW diversity is maximum (KW = 0.25) when half of the T ensemble members correctly classify each of the samples (mean margin μ ranging from 0 (binary classification) to 0.5). In this case, underlying events are equiprobable, i.e. the probability of an instance being correctly classified and misclassified are the same, such as in random prediction.

A good diversity measure would have the ability to find the extent of diversity among classifiers and estimate the improvement or deterioration in accuracy of individual classifiers when they have been combined [19]. An optimal ensemble classifier achieves the right balance between the accuracy of base classifiers and the diversity of the ensemble. Over-fitting can occur if diversity is too low and there is too much correlation between base classifiers. Too much diversity however, can reduce the accuracy of the ensemble. For example, an ensemble classifier with random prediction has the highest diversity but the lowest accuracy. This accuracy-diversity trade-off will be investigated in this study. A particular emphasis is placed on analysing the relationship between diversity and ensemble margin which play a key role in majority vote performance.

4. APPLICATION TO LAND COVER MAPPING

In this section, applying the RF classifier, we evaluate different ways of inducing diversity in ensemble classification to improve classification performance on a large area land cover classification. The main originality of this empirical analysis lies in how the ensemble margin is explicitly involved in the learning process, to induce greater diversity in the ensemble and therefore significantly influence its performance. We run experiments using multiclass (forest canopy cover classes: woodland, open, closed, shrub and non-forest) land cover classification problems. The study area for the experiments comprises 7.2 million hectares of public forests in the state of Victoria, south east Australia.

4.1 Remote sensing and reference data

Sources of remote sensing data included in the model were multi-spectral (Landsat TM and MODIS NDVI), together with derived texture indices and topographic and biophysical climate data. Forest cover maps, derived from Aerial Photographic Interpretation of 30-50 cm resolution colour aerial photographs were used as reference data [22]. A detailed description of the study site and data sources can be found in [7] and [24].

Following our previous work [7], reference data were divided into training and test subsets, comprising 100,000 (20,000 per class) and 25,000 (5,000 per class) samples respectively. Training data were used to calculate margin values (calculated by equation 1) then mean margin (calculated by equation 2). Test data were used to calculate RF model overall and per-class accuracies and KW diversity. For a clearer illustration of results, all diversity values were normalised, to range from 0 to 1, rather than 0 to 0.25. Following our previous work [7], 150 base classifiers were used in each experiment.

4.2 Ensemble diversity analysis

4.2.1 Influence of the number of variables on diversity

The number *m* of variables randomly sampled as candidates to partition training data at each decision tree node was adjusted to evaluate this parameter's effect on diversity and classification performance. Starting with 2, *m* was increased (in single increments) for each RF ensemble model, up to 17 (the maximum number of variables available). A quite similar experiment has been carried out in previous work [20] but without involving margin. This experiment is particularly relevant to exhibit the link between diversity and margin.

Figure 1 shows the effect of the number of variables on both diversity and margin as well as on ensemble and (mean) base classifier accuracies. In this figure, diversity decreases as the number of variables *m* increases. Indeed, the less variables used, the more uncertainty is introduced and diversity is achieved [20]. The mean margin has, expectedly, the opposite behaviour, increasing with the number of variables and hence, classification confidence. Note that a standard RF model would use 4 variables ($m = \sqrt{17}$), which appears here to result in optimal classification performance.

While the mean tree accuracy is reduced with less variables, the difference between ensemble and tree accuracies is higher for 2 variables than for the 17 variables (15.5% and 12% respectively). This illustrates how a loss in tree accuracy is compensated for by higher diversity.

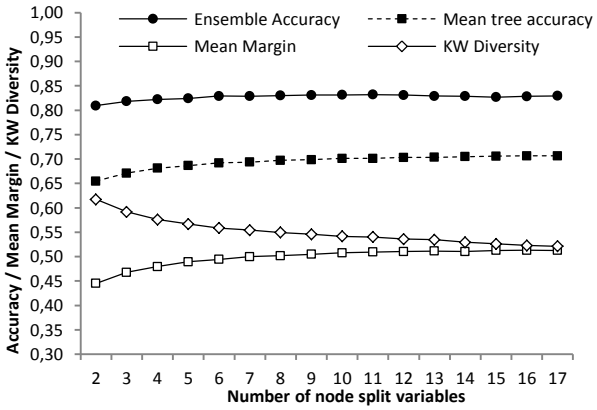

Fig 1. Effect of the number of variables on diversity, margin, ensemble and mean base classifier accuracies

4.2.2 Inducing diversity through low margin sampling

This second experiment constitutes the major contribution of our exploration of ensemble diversity - by investigating a new means of inducing diversity in ensemble learning. This consists of emphasizing the role of lower margin samples in the learning process. For this experiment, the margin (equation 1) was first calculated for each training instance. Percentile distributions were then calculated from the margin values of the training set. RF classifications were run on sub-sets of the original training set, using only training instances in the bottom (lowest margins) and top (highest margins) 50th, 60th, 70th, 80th and 90th percentiles, as well as all training instances. These results were compared to ensemble classifiers generated using a random subset (50%, 60%, 70%, 80% and 90%) of all training instances.

Figures 2 to 4 show results from this experiment, mean base classifier accuracy, ensemble accuracy and KW diversity as a function of training set size, selected by training instances in the bottom (lowest margins) and top (highest margins) 50th to 90th percentiles, and randomly selected training instances (equivalent proportions of the total training set).

Lower margin sampling models result in lower mean decision tree accuracies than from higher margin sampling models (Figure 2). Highest margin generated models exhibit the highest mean tree accuracy (Figure 2), but apart from the 50th margin percentile case, return the poorest ensemble accuracies compared to equivalent training set size models from bottom margin percentiles and random sampling (Figure 3). It is worth highlighting that for the 70th lowest margin percentile, the overall accuracy achieved is the same as that of the entire training set. Hence, the 30% highest margin samples that have been discarded from the training set are redundant. Redundancy not only slows down the training task, but it also weakens bagging performance, affecting the rarer and most difficult classes. The lowest margin training sample selection approach minimises data redundancy. Figure 4 shows lower margin sampling models also exhibit the highest diversity compared to random and highest margin sampling models.

These results, comparing two opposite margin sampling strategies, show that targeting lower margin training data (which represent samples closer to class boundaries and/or more difficult than higher margin samples) is a means of inducing further diversity among base classifiers in an ensemble. This low margin sampling strategy while decreasing mean base classifier accuracy, demonstrates improved ensemble performance induced by the underlying increase in diversity.

The effect of low margin sampling on ensemble accuracy is more pronounced on the open canopy class, the most challenging class. Unsurprisingly, this class returns its highest accuracy (74%) in the bottom 50th percentile margins model and its lowest accuracy (53%) in the top 50th percentile margins model. Furthermore, there is a greater than 5% increase in accuracy between lowest margin and random sampling for 50% training set size. Indeed, open canopy has the highest proportion of low margin samples. Consequently, as any hard or rare class, it is favoured by a lower margin training data selection approach. This strategy reduces data redundancy and increases information significance. Therefore, it designs stronger ensemble classifiers with an increased capability for handling hard or rare classes.

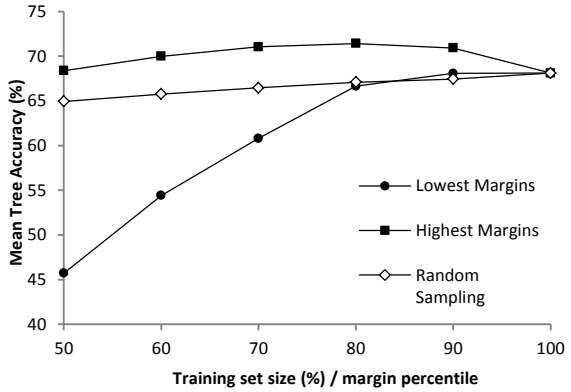

Fig 2. Effect of lowest margin, highest margin and random sampling strategies on mean base classifier accuracy

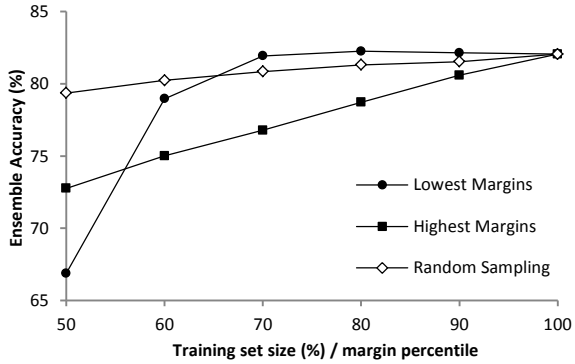

Fig 3. Effect of lowest margin, highest margin and random sampling strategies on ensemble accuracy

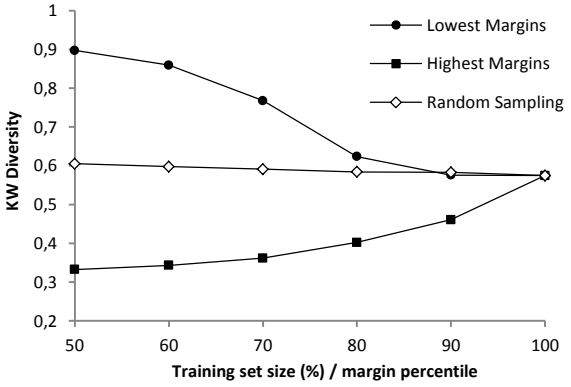

Fig 4. Effect of lowest margin, highest margin and random sampling strategies on diversity

4.2.3 Influence of tree pruning on diversity

This last experiment is a new empirical analysis which aims at investigating the influence of tree pruning (and therefore decision tree depth) on diversity, for a better understanding of ensemble performance in general and RF performance in particular. The minimum node size is a model parameter used to control the minimum size of terminal nodes in each decision tree, and therefore, the depth of decision trees. By default, in RF ensembles (and the other experiments applied in this study), the minimum node size is set to 1. In this experiment, the minimum node size was increased for each RF ensemble model from 1 up to 250.

Figure 5 shows the effect of tree pruning (through increasing the minimum node size) on diversity as well as on ensemble and (mean) base classifier accuracies. Results reveal ensemble accuracy to be highest where trees are grown to their greatest depth (minimum node size of 1), such as in RF ensembles which use unpruned trees. Decreasing diversity is associated with lower ensemble accuracy and increasing minimum node size (shallower trees). Mean tree accuracy is relatively stable for minimum node size under 50. Hence, the loss in ensemble accuracy in this range is mainly due to the loss in diversity. A minimum node size over 50 also affects mean tree accuracy and therefore induces a steeper drop in ensemble accuracy.

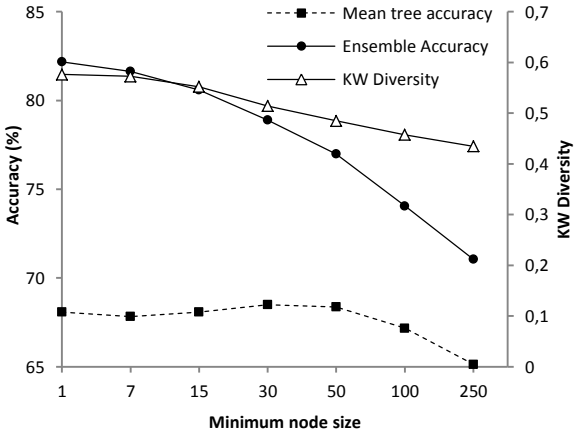

Fig 5. Effect of tree pruning on diversity, ensemble and mean base classifier accuracies

5. CONCLUSION

This work provides insights into the relationship between ensemble diversity and classification performance, in a large area classification problem context using the random forest ensemble classifier. Investigating the effect of the number of decision tree splitting variables on classification performance, showed how lower single tree classification performance associated with fewer splitting variables is compensated for through higher diversity. Targeting lower margin training samples is a way to increase uncertainty and consequently induce diversity in ensemble learning - a strategy which reduces data redundancy and increases the significance of training information. Exploring the influence of tree pruning on classification performance demonstrated that unpruned trees achieve both the highest single tree classification accuracy and the highest diversity among ensemble members, two ingredients for optimal ensemble classification performance. This result partly explains the superiority of random forests, which use unpruned trees, over other tree-based ensembles such as bagging which involve tree pruning.

6. REFERENCES

[1] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, pp. 21–45, 2006.

[2] K. Tumer and J. Ghosh, "Error Correlation and Error Reduction in Ensemble Classifiers," *Conn. Sci.*, vol. 8, no. 3–4, pp. 385–404, Dec. 1996.

[3] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *J. Artif. Intell. Res.*, vol. 11, pp. 169–198, 1999.

[4] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[5] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 67, pp. 93–104, Jan. 2012.

[6] P. Wilkes, S. Jones, L. Suarez, and A. Mellor, "Mapping Forest Canopy Height Across Large Areas by Upscaling ALS Estimates with Freely Available Satellite Data," *Remote Sens.*, vol. 7, pp. 1–25, 2015.

[7] A. Mellor, S. Boukir, A. Haywood, and S. Jones, "Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin," *ISPRS J. Photogramm. Remote Sens.*, vol. 105, pp. 155–168, Jul. 2015.

[8] H. Elghazel, A. Aussem, and F. Perraud, "Trading-Off Diversity and Accuracy for Optimal Ensemble Tree Selection in Random Forests," in *Ensembles in Machine Learning Applications*, vol. 373, O. Okun, G. Valentini, and M. Re, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 169–179.

[9] M. N. Kapp, R. Sabourin, and P. Maupin, "An empirical study on diversity measures and margin theory for ensembles of classifiers," in *2007 10th International Conference on Information Fusion*, 2007, pp. 1–8.

[10] L. Kuncheva and C. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach. Learn.*, vol. 51, pp. 181–207, 2003.

[11] P. Melville and R. J. Mooney, "Creating diversity in ensembles using artificial data," *Inf. Fusion*, vol. 6, no. 1, pp. 99–111, Mar. 2005.

[12] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "Ensemble diversity measures and their application to thinning," *Inf. Fusion*, vol. 6, no. 1, pp. 49–62, Mar. 2005.

[13] L. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. October, pp. 993–1001, 1990.

[14] R. Schapire, Y. Freund, P. Barlett, and W.S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Ann. Stat.*, vol. 26, no. 5, pp. 1651–1686, Oct. 1998.

[15] L. Guo, N. Chehata, C. Mallet, and S. Boukir, "Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests," *ISPRS J. Photogramm. Remote Sens.*, vol. 66, no. 1, pp. 56–66, Jan. 2011.

[16] A. Mellor, S. Boukir, A. Haywood, and S. Jones, "Using ensemble margin to explore issues of training data imbalance and mislabeling on large area land cover classification," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 5067–5071.

[17] L. Guo and S. Boukir, "Margin-based ordered aggregation for ensemble pruning," *Pattern Recognit. Lett.*, vol. 34, no. 6, pp. 603–609, Apr. 2013.

[18] L. Guo, S. Boukir, and N. Chehata, "Support Vectors Selection for Supervised Learning Using an Ensemble Approach," in *20th International Conference on Pattern Recognition*, 2010, pp. 37–40.

[19] Y. Bi, "The impact of diversity on the accuracy of evidential classifier ensembles," *Int. J. Approx. Reason.*, vol. 53, pp. 584–607, 2012.

[20] L. Guo and S. Boukir, "Ensemble margin framework for image classification," in *ICIP 2014, IEEE International Conference on Image Processing*, 2014, pp. 4231–4235.

[21] R. Kohavi and D. Wolpert, "Bias plus variance decomposition for zero-one loss functions," in *13th International Conference of Machine Learning, ICML '96*, 1996, pp. 275–283.

[22] E. Farmer, S. Jones, C. Clarke, L. Buxton, M. Soto-Berelov, S. Page, A. Mellor, and A. Haywood, "Creating a large area landcover dataset for public land monitoring and reporting," in *Progress in Geospatial Science Research*, C. Arrowsmith, C. Bellman, W. Cartwright, S. Jones, and M. Shortis, Eds. Melbourne: Publishing Solutions, 2013, pp. 85–98.

[23] A. Mellor, A. Haywood, C. Stone, and S. Jones, "The Performance of Random Forests in an Operational Setting for Large Area Sclerophyll Forest Classification," *Remote Sens.*, vol. 5, no. 6, pp. 2838–2856, Jun. 2013.

[24] A. Haywood, A. Mellor, and C. Stone, "A strategic forest inventory for public land in Victoria, Australia," *For. Ecol. Manage.*, vol. 367, pp. 86–96, 2016.