ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

The stability of feature selection and class prediction from ensemble tree classifiers

Jérôme Paul, Michel Verleysen and Pierre Dupont

Université catholique de Louvain - ICTEAM/Machine Learning Group
Place Sainte Barbe 2, 1348 Louvain-la-Neuve - Belgium

Abstract. The bootstrap aggregating procedure at the core of ensemble tree classifiers reduces, in most cases, the variance of such models while offering good generalization capabilities. The average predictive performance of those ensembles is known to improve up to a certain point while increasing the ensemble size. The present work studies this convergence in contrast to the stability of the class prediction and the variable selection performed while and after growing the ensemble. Experiments on several biomedical datasets, using random forests or bagging of decision trees, show that class prediction and, most notably, variable selection typically require orders of magnitude more trees to get stable.

1 Motivation

The generalization capabilities of Random Forests (RF) [1], and similar tree ensemble classifiers, is known to increase up to a certain point with the number of trees in the forest. This number is commonly chosen, typically through an internal cross-validation, to reach a plateau of the predictive performance. However, stability issues of such ensemble have not been extensively studied so far. Our first objective is to assess to which extent the predictive performance convergence appears earlier than a stable class prediction. In other words, while the *average* predictive performance no longer changes significantly once this plateau has been reached, the specific labels assigned to each test example can still vary. The bootstrap mechanism at the core of the estimation of such classifiers is known to reduce variance in most cases and indeed stabilizing the individual class prediction is expected but possibly with a larger number of trees.

Tree ensemble techniques also perform an embedded variable selection. Such a selection already occurs at each node while growing the various trees. It can also be performed globally once the forest is built. Breiman suggests in particular a permutation test to select the most relevant features from a Random Forest [1]. In the present work, we study the stability of this variable selection in contrast to the convergence of the average predictive performance and of the class predictions. Our central question of interest is to assess to which extent the variable selection is more brittle than the individual class predictions. Our experiments conducted on various biomedical datasets, with RF and bagging of decision trees or stumps, show that orders of magnitude more trees are typically required to get a stable variable selection as compared to reaching a stable class prediction.

2 Ensemble of tree classifiers

Bagging of decision trees is arguably among the simplest approaches to overcome the strong tendency of a single decision tree to over-fit the learning data.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Bagging (bootstrap aggregating) builds a set of classifiers from successive bootstrap samples of the original training set. The final classifier combines individual decision trees by a majority vote, a form of unweighted averaging [2]. The diversity of the ensemble combined with the final averaging is known to increase the robustness of the aggregated classifier. Random Forests (RF) go one step further to promote the ensemble diversity by randomly sub-sampling the set of features to be evaluated at each node while growing the trees [1].

The approaches mentioned so far mostly focus on improving the predictive accuracy of the tree ensemble under various settings. The number of trees is typically tuned to the smallest possible ensemble size while still reaching the asymptotic predictive performance. Pruning techniques can also be used to further reduce the ensemble size with a marginal loss, or sometimes even a gain, in predictive accuracy [3].

A distinct but related issue is the stability of the class prediction, that is to which extent the class label predicted for each test example stays the same over different data resamplings. The experiments reported in section 4 show that stable class predictions can be observed but at the cost of increasing the ensemble size beyond the convergence of the prediction accuracy.

Ensembles of tree classifiers are also commonly used to select features. The very process of learning decision trees includes a greedy selection of a most discriminant feature at each node, according to a relevance index such as information gain, gain ratio or the Gini index. However the diversity of the various trees, even though beneficial for the predictive accuracy, may result in a large set of used features, some of them only marginally present in the ensemble. Further selecting the most prominent variables increases the interpretability of the combined classifier, a key aspect for applications such as medical diagnosis from gene expression measurements. Such a selection can be performed according to the number of times a given feature appears in the forest, possibly weighting each feature occurrence by its relevance at the corresponding node. An even better alternative to estimate the importance of each feature to classify unseen examples relies on a permutation test computed on the out-of-bag examples from each bootstrap round [1]. For each variable, one compares the out-of-bag classification accuracy after permuting the feature values on those examples with the accuracy obtained from the original values. The more the classification performance drops after permutation the more important is estimated the corresponding feature. We study here the stability of the most prominent features resulting from this additional selection and we compare it to the class prediction stability.

3 Experimental design and stability assessment

We aim at assessing the predictive and stability performances of ensemble tree classifiers while growing the number T of trees in the ensemble. For the sake of this study, we compare a representative set of such classification methods: bagging of unpruned CART trees [4], Random Forests [1], as well as bagging of decision stumps and RF of decision stumps. Practical experiments are conducted on several biomedical datasets described below. Most of them fall into the small n (number of examples), large p (number of features) setting. In those cases,

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

a standard 10-CV protocol is likely to show highly variable results when the number of examples is limited to a few tens. Hence, we rely on $K = 200$ random splittings of the data into train (90%) and test (10%). Each data partitioning results from uniform sampling without replacement (a significance test for such a protocol is proposed in [5]) and we report average predictive performances over all resamplings.

Predictive performance is measured by the *Balanced Classification Rate* (BCR), which is the per class accuracy, averaged over the various classes. BCR is preferred to accuracy for classification problems with unequal class priors. BCR is also simpler than ROC analysis for multi-class problems. For binary classification problems, BCR simply reduces to the arithmetic average between specificity and sensitivity.

The *stability of the class prediction* measures to which extent each individual test example is assigned the same class label across various resamplings. For each example x_i, let c^* denote the most commonly predicted class label (across all resamplings for which that example appears in a test fold); let $n_{x_i}^{c^*}$ be the number of times such a prediction occurs out of the n_{x_i} occurrences of x_i in a test fold. The class prediction stability is given by:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{n_{x_i}^{c^*}}{n_{x_i}}\ , \tag{1}$$

where n denotes the total number of examples, each of them appearing approximately $0.10 \times K = 20$ times in a test fold. Such a stability index falls in the interval $[\frac{1}{|C|}, 1]$ with $|C|$ classes. The stability is equal to 1 when each test example is always assigned the same, although not necessarily correct, class label.

The *stability of the feature selection* can be measured according to the Kuncheva Index [6]. This index measures to which extent K sets S_i of s selected features share common features:

$$K(\{S_1, ..., S_K\}) = \frac{2}{K(K-1)}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{|S_i \cap S_j| - \frac{s^2}{p}}{s - \frac{s^2}{p}}\ , \tag{2}$$

where p is the total number of features, and feature selection is performed on each of the K training folds. The s^2/p term corrects a bias due to the chance of selecting common features among two sets chosen at random and motivates our choice of this specific stability index. The Kuncheva index ranges within (-1,1] and the greater its value the larger the number of common features across the K feature sets. In the experiments reported in Section 4, s was set equal to 25 to stick to a small subset of the most important features. Additional experiments (not reported) show that our conclusions remain essentially identical for larger values of s. In some marginal cases however, such as bagging of a few decision stumps, the number of selected features is bound to be lower than the prescribed s, which tends to increase the stability.

Table 1 presents the main characteristics of the datasets used in the present study: the number of instances, number of continuous/categorical features and class priors. We focus on biomedical data for which the number of features often largely exceeds the number of samples. This is particularly true for gene

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

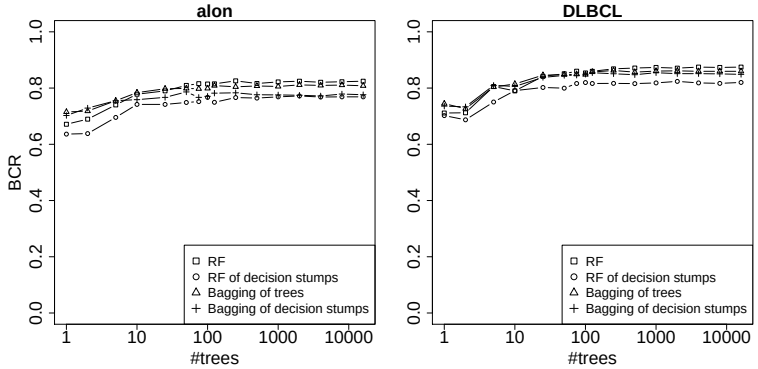
expression data (`Alon`, `DLBCL` and `van't Veer`) although we also consider electrocardiogram data (`Arrhythmia`) to broaden the scope of our study. Those are binary classification tasks but our main conclusions are likely applying as well to multi-class problems. `Alon` [7] task aims at discriminating between normal and colon tumor tissues, `DLBCL` [8] concerns the prediction of tissue type from diffuse large B-cell lymphoma, while `van'Veer` [9] aims at predicting distant metastasis from breast cancer samples. The `Arrhythmia` [10] problem discriminates between normal and arrhythmic ECG.

Name	#examples	# cont. feat.	# cat. feat.	class priors
`Alon` [7]	62	2000	0	40/22
`DLBCL` [8]	77	7129	0	58/19
`van't Veer` [9]	77	4353	2	44/33
`Arrhythmia` [10]	430	198	64	185/245

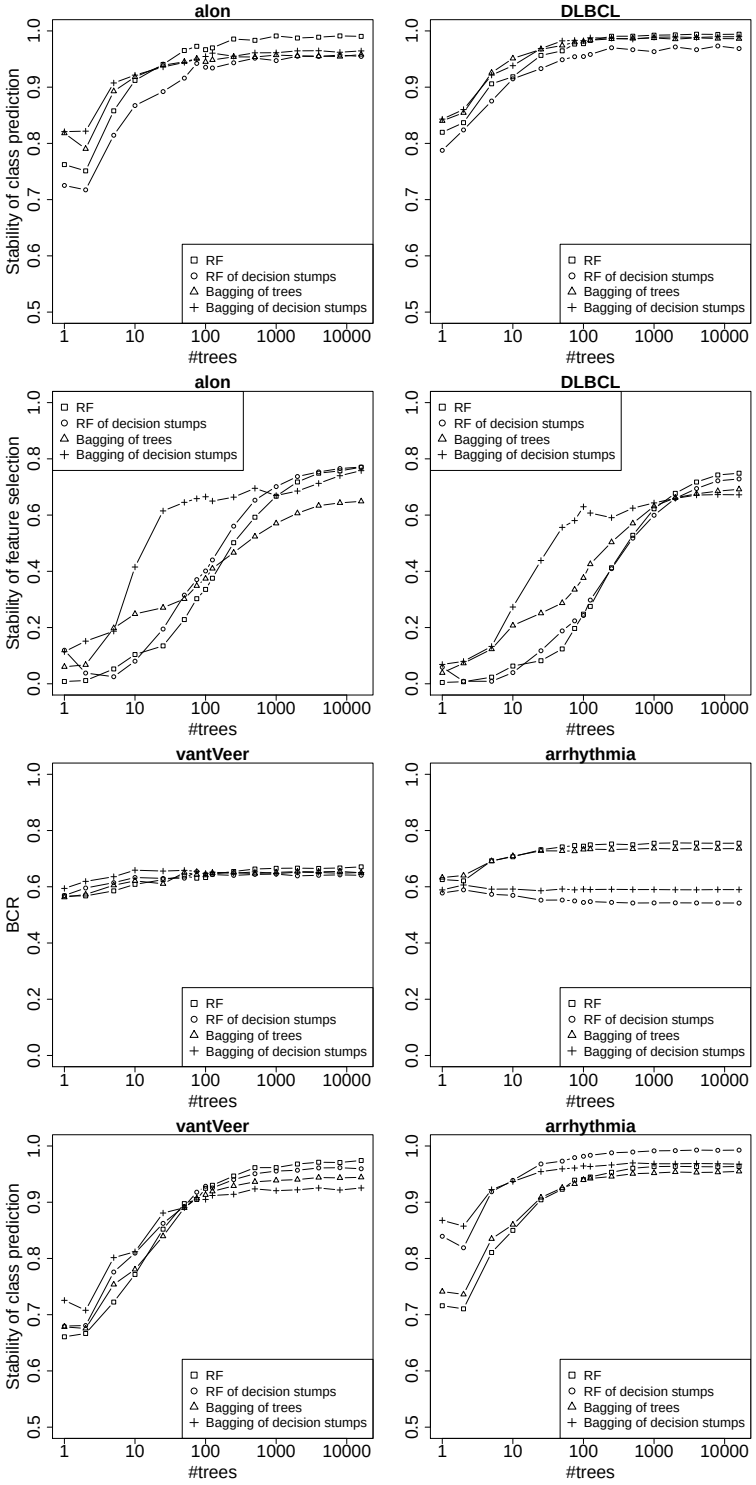
Table 1: Datasets overview

4 Results

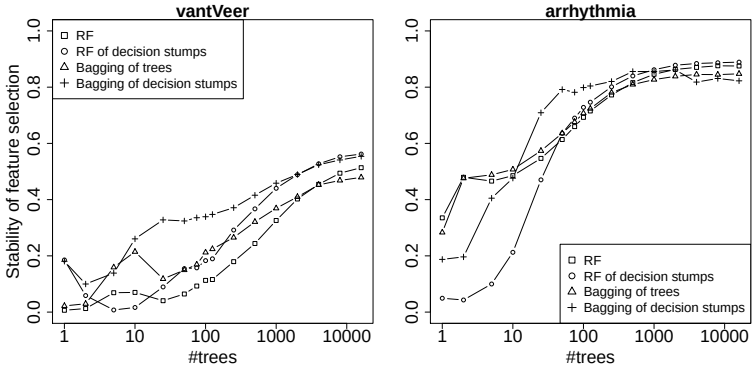
The experimental results are reported in the plots below. They show no significant BCR differences across the various ensemble classifiers, but for `Arrhythmia` on which stumps have significantly lower results ($p-$value $< 2.10^{-10}$ according to the corrected resampled t-test [5]). For all methods, the convergence of the predictive performance is typically reached after 10 or 20 trees for the four datasets. The same conclusions can be drawn when the predictive performance is estimated from the classification accuracy instead of the BCR (results not shown). The class predictions typically require an order of magnitude more trees (100...200 trees) to get stable. Feature selection only get stable, and yet not perfectly, from at least an order of magnitude more trees ($\geq 1,000$) on the 3 genomic datasets. An earlier convergence is obtained for `Arrhythmia` as a natural consequence of a smaller total number of features and more training samples. Bagging of decision stumps also tends to offer an earlier convergence of the feature selection stability. This makes sense as there is only one feature selected for each stump without any random sampling of the feature space.



ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.



ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.



5 Conclusion and perspectives

Our experimental study demonstrates, for a variety of ensemble tree classifiers, that stable class predictions and, most notably, stable feature selection require orders of magnitude more trees than those needed to reach the asymptotic predictive performance. Our future work includes a more formal analysis (similarly to [11]) of such behaviors, and possibly ways to promote an earlier stability without losing the benefits of the ensemble.

References

[1] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[2] Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996. 10.1007/BF00058655.

[3] G. Martinez-Muñoz, D. Hernández-Lobato, and A. Suárez. An analysis of ensemble pruning techniques based on ordered aggregation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):245 –259, feb. 2009.

[4] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth Inc., 1984.

[5] C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine Learning*, 52:239–281, 2003.

[6] Ludmila I. Kuncheva. A stability index for feature selection. In *AIAP'07: Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference*, pages 390–395, Anaheim, CA, USA, 2007. ACTA Press.

[7] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12):6745–6750, June 1999.

[8] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med*, 8(1):68–74, January 2002.

[9] Laura J. van 't Veer, Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, René Bernards, and Stephen H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, January 2002.

[10] A. Frank and A. Asuncion. UCI ML repository - http://archive.ics.uci.edu/ml, 2010.

[11] D. Hernández-Lobato, G. Martínez-Muñoz, and A. Suárez. Inference on the prediction of ensembles of infinite size. *Pattern Recognition*, 44:1426–1434, 2011.