HAL
archives-ouvertes.fr

Towards a better understanding of random forests through the study of strength and correlation

Simon Bernard, Laurent Heutte, Sébastien Adam

▶ **To cite this version:**

Simon Bernard, Laurent Heutte, Sébastien Adam. Towards a better understanding of random forests through the study of strength and correlation. 5th International Conference on Intelligent Computing (ICIC), Sep 2009, Ulsan, South Korea. pp.536-545, 10.1007/978-3-642-04020-7_57 . hal-00436361

HAL Id: hal-00436361

https://hal.archives-ouvertes.fr/hal-00436361

Submitted on 26 Nov 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Towards a Better Understanding of Random Forests Through the Study of Strength and Correlation

Simon Bernard, Laurent Heutte, and Sébastien Adam

Université de Rouen, LITIS EA 4108
BP 12 - 76801 Saint-Etienne du Rouvray, France.
{simon.bernard,laurent.heutte,sebastien.adam}@univ-rouen.fr

Abstract. In this paper we present a study on the Random Forest (RF) family of ensemble methods. From our point of view, a "classical" RF induction process presents two main drawbacks : (i) the number of trees has to be *a priori* fixed (ii) trees are independently, thus arbitrarily, added to the ensemble due to the randomization principle. Hence, this kind of process offers no guarantee that all the trees will well cooperate into the same committee. In this work we thus propose to study the RF mechanisms that explain this cooperation by analysing, for particular subsets of trees called sub-forests, the link between accuracy and properties such as Strength and Correlation. We show that these properties, through the Correlation/Strengh^2 ratio, should be taken into account to explain the sub-forest performance.

Key words: Classification, Ensemble Method, Ensemble of Classifiers, Classifier Selection, Random Forests, Decision Trees

1 Introduction

Random Forest is a family of classifier ensemble methods that use randomization to produce a diverse pool of individual classifiers, as for Bagging [1] or Random Subspaces methods [2]. It can be defined as a generic principle of classifier combination that uses L tree-structured base classifiers $\{h(x, \Theta_k),\ k = 1, ...L\}$ where $\{\Theta_k\}$ is a family of independent identically distributed random vectors, and x is an input data. The particularity of this kind of ensemble is that each decision tree is built from a random vector of parameters. A Random Forest can be built for example by randomly sampling a feature subset for each decision tree (as in Random Subspaces [2]), and/or by randomly sampling a training data subset for each decision tree (as in Bagging [1]). Since they have been introduced in 2001, RF have been studied in many ways, both theoretically and experimentally [3–11]. In most of those works, it has been shown that RF are particularly competitive with one of the most efficient learning principles, i.e. boosting [5, 7, 10]. However, the mechanisms that explain the good performance of RF are not clearly identified. For example, it has been theoretically proved in [5] and experimentally confirmed in [9], that above a certain number of trees, adding more trees in the forest does not improve the accuracy. Yet, no research work has studied the way each tree contributes to the performance of a RF, and how they cooperate.

In this paper we propose to go one step further in the understanding of RF mechanisms, by studying properties of different subsets of decision trees, with different performance. Thus our aim is not to find better subsets of trees that can outperform a larger ensemble of trees, but rather to study properties that can explain those differences in the sub-RF performance. Therefore, as we will discuss in section 3, we have decided to use a sub-optimal classifier selection technique, *i.e.* SFS (Sequential Forward Selection)[12], to generate different sub-RF that exhibit better accuracies on a test set than the initial forest from which sub-RF have been obtained. By monitoring the accuracies and by focusing on some particular properties shared by these sub-forests, we bring some primary responses for explaining the differences in performance. Strength and correlation properties, as defined in [5], have been chosen to study the evolution of sub-RF accuracy during the classifier selection process. We show that these properties, through the Correlation/Strengh^2 ratio, are important criteria that should be taken into account for explaining the performance evolution according to the number of trees in the sub-RF.

The paper is thus organized as follows: we recall in section 2 the Forest-RI principles; in section 3, we first explain our approach for studying the RF mechanisms, and then describe our experimental protocol, the datasets used, and the results obtained. We finally draw some conclusions and future works in the last section.

2 The Forest-RI algorithm

One can see Random Forests as a family of methods, made of different decision trees ensemble induction algorithms, such as the Breiman Forest-RI method often cited as the reference algorithm in the literature. In this algorithm the Bagging principle is used with another randomization technique called Random Feature Selection. The training step consists in building an ensemble of decision trees, each one trained from a bootstrap sample of the original training set — *i.e.* applying the Bagging principle — and with a decision tree induction method called Random Tree. This induction algorithm, usually based on the CART algorithm [13], modifies the splitting procedure for each node, in such a way that the selection of the feature used for the splitting criterion is partially randomized. That is to say, for each node, a feature subset is randomly drawn, from which the best splitting criterion is then selected.

To sum up, in the Forest-RI method, a decision tree is grown by using the following process :

- Let N be the size of the original training set. N instances are randomly drawn with replacement, to form the bootstrap sample, which is then used to build the tree.
- Let M be the dimensionality of the original feature space, and K a preliminary fixed parameter so that $K \in [1..M]$. For each node of the tree, a subset of K features is randomly drawn without replacement, among which the best split is then selected.
- The tree is thus built to reach its maximum size. No pruning is performed.

In this process the tree induction is directed by a single hyperparameter, *i.e.* the number K of randomly selected features. This number allows to introduce more or less randomization in the induction. Consequently, except when $K = M$, in which case the

tree induction is not randomized at all, each tree of a RF presents structure and properties that can not be foreseen *a priori*. With the introduction of randomization in the RF induction, one hopes to take benefits of complementarities of individual trees, but there is no guarantee that adding a tree in a RF will allow to improve the performance of the ensemble. One can even imagine that some trees of a RF make the accuracy of the ensemble decrease. This idea has led us to study how trees cooperate in the same committee to ensure a better accuracy.

In the literature, only few research works have focused on the way trees have to be grown in a RF. When introducing RF formalism in [5], Breiman demonstrated that above a certain number of trees, adding more trees does not allow to improve the performance. Precisely he stated that for an increasing number of trees in the forest, the generalisation error converges to a maximum. This result indicates that the number of trees in a forest does not have to be as large as possible to produce an accurate RF. The works of Latinne et al. in [9], and of Bernard et al. in [3] experimentally confirm this statement. However, noting that above a certain number of trees no improvement can be obtained by adding more "arbitrary" trees in the forest does not mean obviously that the optimal performance has been reached. It does neither give explanation of the good cooperation of trees in the ensemble. Thus the idea of our experimental work is to lead an analysis of the evolution of accuracies, in order to understand this cooperation. We present in the next section our experimental approach for those purposes.

Notice that in the rest of this paper, the term Random Forest (RF) will always stand for a forest built with the Forest-RI algorithm.

3 Analysing strength and correlation in sub-RF

The principle of our experiments is to generate different subsets of trees and to evaluate their accuracies on the same test set so that it will be possible to examine some properties shared by the "best" sub-RF regarding the performance of the initial RF. The idea is firstly to use a classifier selection technique to generate those sub-RF and then to measure and monitor the properties on which we have decided to focus on. First, two main choices have thus to be made: a selection criterion and a selection method.

Selection criteria for classifier selection can be divided into two main approaches: the filter approach and the wrapper approach [14]. On the one hand the filter approach consists in selecting a subset of classifiers according to an *a priori* evaluation that does not take into account the combination performance. On the other hand, the wrapper approach attempts to select the subset of classifiers that *a posteriori* optimizes the combination performance. As we intend to find a correlation between properties and accuracy the wrapper principle has been adopted for our experiments. Thus classifiers have been selected by optimizing the accuracy on a test set — i.e. minimizing the error rate. It is obvious that using the test set for the selection criterion, rather than an independent validation set, does not allow to evaluate the generalisation error rate of the resulting subsets and thus will not allow to conclude on the overall performance of the sub-RF comparing to the initial forest. However our goal is not actually to perform classifier selection to find better sub-RF in terms of generalisation capacities, but rather to focus on

a possible link between accuracy and some properties regarding a particular prediction set.

Concerning the selection method, for the reasons mentioned above, the optimality of the selection method is not a priority here. That is the reason why the well-known classifier selection algorithm SFS (Sequential Forward Selection) has been chosen. This method is known to be sub-optimal because the sequential process makes each iteration depend on the previous one, and finally not all the possible solutions are explored. However it presents the advantage to be fast and simple. This selection technique iteratively builds a sub-optimal subset from an ensemble of classifiers according to a given criterion [12]. At each iteration of the SFS process, each remaining classifier is added to the current subset and the one that optimizes the performance of the ensemble is retained. The stopping criterion in such an iterative process is commonly based on the convergence of the accuracy, but it can also be defined for example by a maximum number of iterations that determines the number of classifiers in the final subset [15]. For our experiments we have decided to let the selection algorithm explore all the possible iterations, *i.e.* for a number L', from 1 to L, of trees in the final subset, where L is the size of the original RF.

Then, our goal is thus to bring elements of explanation for this evolution of accuracy. In [5], Breiman introduced two crucial notions for inducing accurate RF : the *strength*, noted s, and the *correlation*, noted $\overline{\rho}$. The definition of the *strength* is based on the margin function of a RF, that measures the extent to which the average number of votes for the right class exceeds the average vote for any other class. The strength is consequently defined as the expectation of the margin function over all the training samples. The *correlation* is the "classical" pairwise correlation, averaged over all the pairs of decision trees in the forest. This pairwise correlation is however computed through the raw margin function of each tree which gives three possible answers for a given training sample: 1 if the tree predicts the right label; -1 if the tree predicts the most popular of the wrong label; and 0 if the tree does not predict any of these two labels.

Breiman proved that an upper bound of the generalisation error of RF is given thanks to the ratio $\frac{\overline{\rho}}{s^2}$. He conjectured that *in understanding the functioning of random forests, this ratio will be a helpful guide — the smaller it is, the better*. We thus propose to experimentally study this statement by measuring this ratio for each of the sub-RF obtained during the selection process. This would allow to match this property with the accuracy, and to determine whether or not it can totally or partially explain its evolution.

We first describe in the following subsection the datasets used. We then detail our experimental protocol and results in the next two subsections.

3.1 Datasets

The 18 datasets that have been used in these experiments are described in Table 1. The first 13 datasets have been selected from the UCI repository [16], because they concern different machine learning issues in terms of number of classes, number of features and number of samples. Twonorm and Ringnorm are two synthetic datasets designed by Breiman and described in [17]. Three additional datasets on different handwritten digit recognition problems have been used: (i) the well-known MNIST database [18] with a

85 multiresolution density feature set ($1 + 2 \times 2 + 4 \times 4 + 8 \times 8$) built from greyscale mean values as explained in [3]; (ii) Digits and DigReject both described in [19], on which a 330-feature set has been extracted, made from three state-of-the-art kinds of descriptors, *i.e.* a 117-statistical/structural feature set [20], a 128-feature set extracted from the chaincode (contour-based) [21], and the same 85-feature set as for MNIST.

Table 1. Datasets description

Dataset	Size	Features	Classes	Dataset	Size	Features	Classes
Digits	38142	330	10	Mnist	60000	84	10
DigReject	14733	330	2	Musk	6597	166	2
Gamma	19020	10	2	Pendigits	10992	16	10
Letter	20000	16	26	Ringnorm	7400	20	2
Madelon	2600	500	2	Segment	2310	19	7
Mfeat-factors	2000	216	10	Spambase	4610	57	2
Mfeat-fourier	2000	76	10	Twonorm	7400	20	2
Mfeat-karhunen	2000	64	10	Vehicle	946	18	4
Mfeat-zernike	2000	47	10	Waveform	5000	40	3

3.2 Experimental protocol

In this section we describe the full experimental protocol.

First, each dataset has been divided into a training and a testing subset, with respectively two thirds of the samples used for training, and the other third for testing. We denote this split by $T = (T_r, T_s)$ where T_r and T_s stand respectively for the training set and the testing set. Then, a RF is grown from T_r, with a number L of trees fixed to 300. The value of the hyperparameter K has been fixed to $\sqrt{M}$, which is a default value commonly used in the literature. An experimental work on the parametrization of RF, presented in [22], has shown that this value of K is a good compromise to induct accurate RF. Thus, SFS method is applied on the RF, so that at each iteration the tree to add is the one that allows to obtain the most accurate sub-forest on the test set. For each of these sub-RF three measures have been monitored : (i) the error rate measured on T_s (ii) the strength value (iii) and the correlation value.

Finally a statistical test of significance has been performed at each iteration of the selection procedure, in order to state whether or not the resulting subset outperforms the initial forest regarding to a particular prediction set. For that purpose we lean on the comparison of five approximate statistical tests, compared in [23]. In this paper, it is recommended to use McNemar's test [24], for which it is shown that it better suits to experimental protocols like ours. This test is used here to determine whether or not two learning algorithms differ significantly according to their sets of predictions. Three answers can thus be obtained through the McNemar test:

- H_0 is rejected and $n_{01} > n_{10}$: Algorithm B produces significantly more accurate classifiers than algorithm A.

- H_0 is rejected and $n_{01} < n_{10}$: Algorithm A produces significantly more accurate classifiers than algorithm B.
- H_0 is accepted: The two algorithms do not produce classifiers significantly different in term of accuracy.

This procedure has been run for all the sub-RF in order to determine whether or not they outperform the initial RF on the test set. Algorithm 1 summarizes the whole experimental protocol applied to each dataset. Results are presented and discussed in the next subsection.

Algorithm 1 Experimental Protocol

Input: N: # samples in the original dataset.
Input: M: # features in the original dataset.
Input: L: # trees in the original forest.
Output: $\epsilon[L]$: 1D table for storing error rates.
Output: $s[L]$: 1D table for storing strength values.
Output: $\rho[L]$: 1D table for storing correlation values.
Output: $\mathcal{M}[L]$: 1D table for storing McNemar test answers.

Randomly draw without replacement $\frac{2}{3} \times N$ samples from the original dataset to form the training subset T_r. The remaining samples form the testing subset T_s.

$h \leftarrow$ Forest-RI($L = 300, K = \sqrt{M}, T_r$).

$h^{(0)}_{SFS} \leftarrow \emptyset$.

for $i = 1$ *to* L **do**

$h^{(i)}_{SFS} \leftarrow h^{(i-1)}_{SFS} \cup h(k)$ where $k = argmin_{h(j) \notin h^{(i-1)}_{SFS}} \{\text{error}(h^{(i-1)}_{SFS} \cup h(j), T_s)\}$.

$\epsilon(i) \leftarrow$ Test $h^{(i)}_{SFS}$ on T_s.

$s(i) \leftarrow$ compute strength for $h^{(i)}_{SFS}$ on T_s.

$\rho(i) \leftarrow$ compute correlation for $h^{(i)}_{SFS}$ on T_s.

$\mathcal{M}(i) \leftarrow$ run a McNemar test of significance with classifiers $(h, h^{(i)}_{SFS})$ on the testing set T_s

end for

3.3 Results

Table 2 presents the best error rates obtained during the selection process on each dataset, and the number of trees of the corresponding subset. Figure 1 presents 18 diagrams of our results for the 18 datasets used. For each of them, a curve has been plotted, representing the error rate obtained during the selection procedure, according to the number of trees in the subset. For comparison, a line has also been drawn on each diagram that represents the error rate obtained with the initial forest of 300 trees. Finally the McNemar test results have also been plotted on the same diagrams. For each sub-RF of each size a McNemar answer has been obtained that indicates whether or not the sub-RF has outperformed the initial forest. If so, a mark has been drawn on the

x-axis of the diagrams. In that way sub-RF with significant improvement over the initial RF are highlighted.

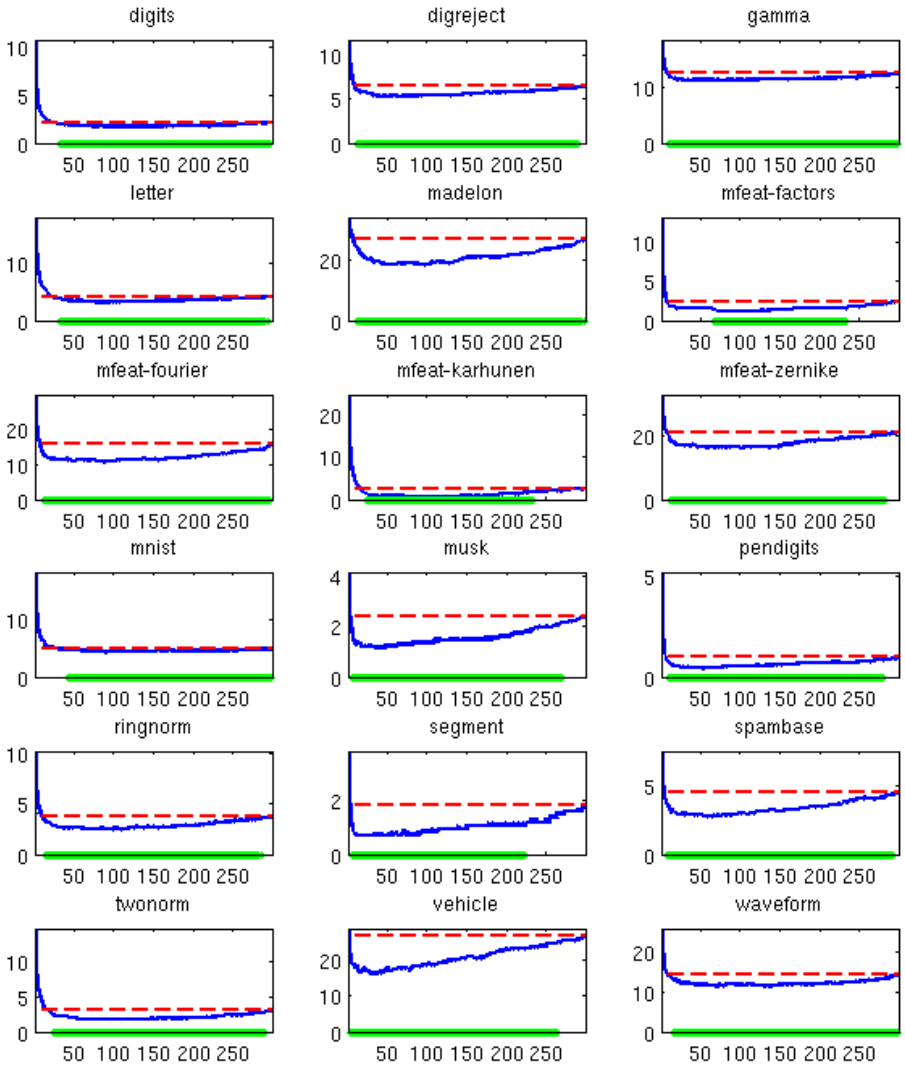

Fig. 1. Error Rates obtained during the selection process on the 18 datasets, according to the number of trees in the subsets. The dashed lines represent the error rates obtained with the initial RF made of 300 trees. The marks on the x-axis represent the sub-RFs for which McNemar test indicates that the accuracy improvement is statistically significant.

Table 2. Best error rates and number of trees of the corresponding selected subsets

Dataset	SFS		Forest-RI	Dataset	SFS		Forest-RI
	error rates	# trees	300 trees		error rates	# trees	300 trees
Digits	1.68	101	2.18	MNIST	4.44	109	4.92
DigReject	5.27	43	6.52	Musk	1.13	35	2.36
Gamma	11.14	76	12.45	Pendigits	0.41	41	0.96
Letter	3.17	83	4.15	Ringnorm	2.47	99	3.73
Madelon	18.13	59	26.67	Segment	0.65	11	1.83
Mfeat-fac	1.21	70	2.42	Spambase	2.73	58	4.56
Mfeat-fou	10.60	51	15.76	Twonorm	1.74	82	3.01
Mfeat-kar	0.60	73	2.57	Vehicle	16.07	31	26.79
Mfeat-zer	16.06	102	21.21	Waveform	11.35	138	14.48

One can first observe from Table 2 that in spite of the sub-optimality of SFS, it always finds a subset of trees that exhibits a lower error rate on the test set than the initial RF, induced with Forest-RI. Results of McNemar statistical test presented in Figure 1 strongly confirm this statement. Though it does not concern generalisation performance, a surprising result is that the number of trees in the "best" subset found during the selection process is often very small regarding to the size of the initial forest, sometimes even approaching 30% of the amount of available trees (Musk, Segment and Vehicle). Classifier selection has already shown to be a powerful tool for obtaining significant improvement with ensemble of classifiers [25–27], but this result leads us to think that it would be interesting to further focus on the number of trees that have to be grown in a forest to obtain significant improvement comparing to RF induced with Forest-RI, and rather according to generalisation accuracy.

Additional diagrams are then presented in Figures 2 and 3: (i) the 9 diagrams in figure 2 represent the error rates according to the ratio $\frac{\bar{\rho}}{s^2}$ (ii) The 9 diagrams in figure 3 represent this ratio according to the size of the sub-RF obtained during the selection process. In a concern to be clearer, only 9 of the 18 datasets have been used to illustrate our results, but tendencies discussed in this section can be expanded to the rest of the datasets since all the curves obtained follow the same global behavior.

Figure 2 highlights the link between $\frac{\bar{\rho}}{s^2}$ and RF accuracy. One can firstly remark that the points on these diagrams follow the same global distribution for all the datasets, with a strong rise of error rates for increasing values of the ratio. From our point of view, this illustrates the fact that error rates are strongly related to the ratio and that this former measure tends to explain, at least partially, the variation of performance from a sub-RF to another. However, according to Breiman's theoretical results, the values of this ratio should decrease jointly with the error rate, and one can observe that it is not strictly the case in these diagrams. The minimal error rate is never reached for the minimal value of $\frac{\bar{\rho}}{s^2}$, and a small rise is obtained for decreasing values on the x-axis. It seems to us that those few points for which the ratio is weak but for which the error rate is not minimal as expected, are particular cases for which the ratio computation has been biased. An explanation of these few points that do not fit with Breiman's

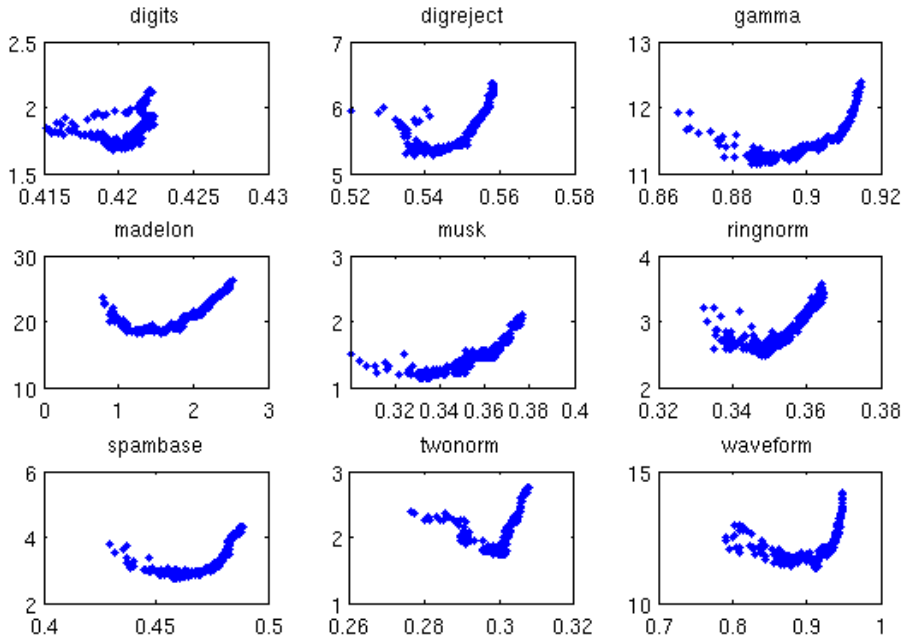

Fig. 2. Error rates according to the ratio $\frac{\bar{\rho}}{s^2}$

theoretical results can probably be found in figure 3. One can notice indeed that all the curves in this figure surprisingly tend to monotonically decrease as the number of trees in sub-RF decreases. This decreasement is firstly consistent with our previous results that show that best error rates are obtained for a small subset of trees. But values of $\frac{\bar{\rho}}{s^2}$ seem particularly unstable for sub-RF made of less than about 50 trees, and this unstability has systematically been obtained for all the datasets used in our experiments. We think in the light of these results that the size of the sub-forests obtained during the selection process plays a role in the explanation of performance variations. Although the SFS method is interesting for evaluating the extent to which RF performance can be enhanced by selecting particular subsets of trees, the sequentiality of such a procedure seems to introduce a bias in the computation of Strength and Correlation measures. To continue the analysis of RF mechanisms via the study of sub-RF and of their accuracy evolution, we think that it would be more judicious to use a selection method such as Genetic Algorithm and to fix the number of trees selected, so that resulting sub-RF would be of the same size.

4 Conclusions

In this paper different sub-RF have been generated from a pool of random trees, in order to analyse RF mechanisms that could explain performance variation from a forest

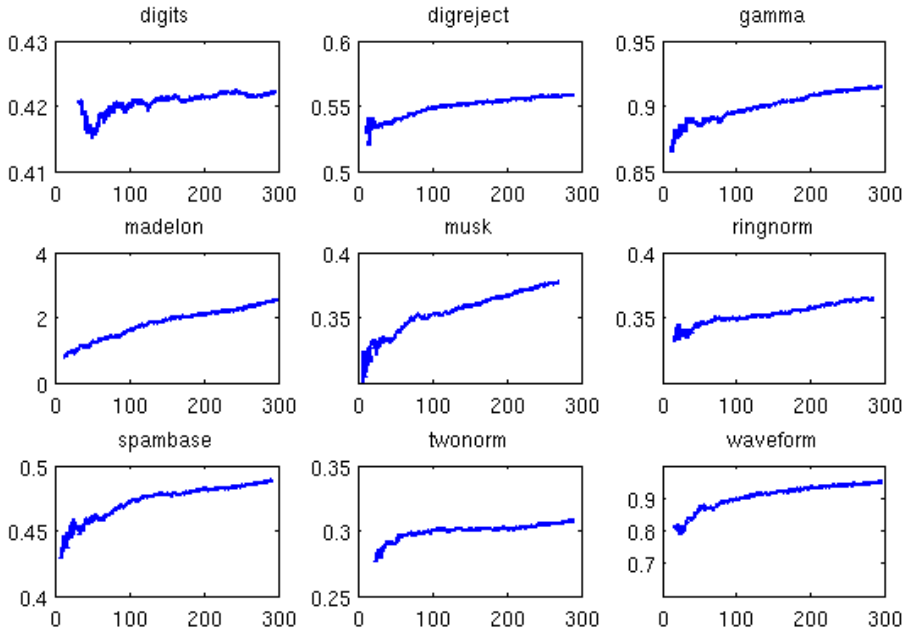

Fig. 3. the ratio $\frac{\bar{\rho}}{s^2}$ according to the size of the sub-RF

to another. A classifier selection method, *i.e.* the Sequential Forward Selection method (SFS), has been used for generating those sub-forests. The goal was firstly to obtain subsets of trees that exhibit lower error rates than the initial forest on a particular test set, in an attempt to correlate *strength* and *correlation* properties with those performance variations. A surprising result is that the "best" subsets of decision trees on the test set have shown to contain a small number of trees regarding to the amount of trees available in the initial forest, *i.e.* sometimes about 30%. Though this result is not significant in terms of generalisation performance, this statement gives the intuition that in a "traditional" RF induction algorithm several trees may deteriorate the performance of the ensemble, and that improvement could be obtained by inducing decision trees in a less arbitrary way in order to build accurate RF.

However for being able to design a RF induction procedure that could avoid this deterioration, it is essential to firstly identify and understand mechanisms that could explain the variations in sub-RF performance. We have thus propose in this paper a primary analysis that aimed at studying Strength and Correlation properties according to sub-RF accuracy obtained during the selection process on a particular test set. This experimental work has shown that the ratio $\frac{\bar{\rho}}{s^2}$, introduced by Breiman in [5], is linked to performance variation of RF. The nature of this link still remains an open issue but we believe that this classifier selection approach will be very helpful in future works to identify it and to better understand Random Forest mechanisms.

References

1. Breiman, L.: Bagging predictors. Machine Learning **24**(2) (1996) 123–140
2. Ho, T.: The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(8) (1998) 832–844
3. Bernard, S., Heutte, L., Adam, S.: Using random forests for handwritten digit recognition. International Conference on Document Analysis and Recognition (2007) 1043–1047
4. Boinee, P., Angelis, A.D., Foresti, G.: Meta random forests. International Journal of Computationnal Intelligence **2**(3) (2005) 138–147
5. Breiman, L.: Random forests. Machine Learning **45**(1) (2001) 5–32
6. Breiman, L.: Consistency of random forests and other averaging classifiers. Technical Report (2004)
7. Cutler, A., Zhao, G.: Pert - perfect random tree ensembles. Computing Science and Statistics **33** (2001)
8. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning **36**(1) (2006) 3–42
9. Latinne, P., Debeir, O., Decaestecker, C.: Limiting the number of trees in random forests. 2nd International Workshop on Multiple Classifier Systems (2001) 178–187
10. Rodriguez, J., Kuncheva, L., Alonso, C.: Rotation forest : A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(10) (2006) 1619–1630
11. Robnik-Sikonja, M.: Improving random forests. European Conference on Machine Learning, LNAI 3210, Springer, Berlin (2004) 359–370
12. Hao, H., Liu, C., Sako, H.: Comparison of genetic algorithm and sequential search methods for classifier subset selection. Seventh International Conference on Document Analysis and Recognition **2** (2003) 765–769
13. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Chapman and Hall (Wadsworth, Inc.): New York (1984)
14. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence **97**(1-2) (1997) 273–324
15. Roli, F., Giacinto, G., Vernazza, G.: Methods for designing multiple classifier systems. Multiple Classifiers Systems (2001) 78–87
16. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
17. Breiman, L.: Arcing classifiers. The Annals of Statistics **26**(3) (1998) 801–849
18. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324
19. Chatelain, C., Heutte, L., Paquet, T.: A two-stage outlier rejection strategy for numerical field extraction in handwritten documents. International Conference on Pattern Recognition, Honk Kong, China **3** (2006) 224–227
20. Heutte, L., Paquet, T., Moreau, J., Lecourtier, Y., Olivier, C.: A structural/statistical feature based vector for handwritten character recognition. Pattern Recognition Letters **19**(7) (1998) 629–641
21. Kimura, F., Tsuruoka, S., Miyake, Y., Shridhar, M.: A lexicon directed algorithm for recognition of unconstrained handwritten words. IEICE Transaction on Information and System **E77-D**(7) (1994) 785–793
22. Bernard, S., Heutte, L., Adam, S.: Forest-rk : A new random forest induction method. International Conference on Intelligent Computing (2008) 430–437
23. Dietterich, T.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation **10** (1998) 1895–1923
24. Everitt, B.: The Analysis of Contingency Tables. Chapman and Hall, London (1977)

25. Aksela, M.: Comparison of classifier selection methods for improving committee performance. 4th International Workshop on Multiple Classifier Systems **2709** (2003) 84–93
26. Banfield, R., Hall, L., Bowyer, K., Kegelmeyer, W.: A new ensemble diversity measure applied to thinning ensembles. 4th International Workshop on Multiple Classifier Systems **2709** (2003) 306–316
27. Santos, E.D., Sabourin, R., Maupin, P.: A dynamic overproduce-and-choose strategy for the selection of classifier ensembles. Pattern Recognition 41 (2008) 2993 – 3009 **41** (2008) 2993–3009