A Study of Strength and Correlation in Random Forests

Simon Bernard, Laurent Heutte, Sébastien Adam

A Study of Strength and Correlation in Random Forests

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 classification methods, and more particularly on two important properties called *strength* and *correlation*. These two properties have been introduced by Breiman in the calculation of an upper bound of the generalization error. We thus propose to experimentally study the actual relation between these properties and the error rate in order to confirm and extend the Breiman theoretical results. We show that the error rate statistically decreases with the joint maximization of the *strength* and minimization of the *correlation*, and this for different sizes of RF.

Key words: Random Forests, Ensemble of Classifiers, Strength, Correlation

1 Introduction

Recently, Leo Breiman has proposed a new family of ensemble methods called Random Forest (RF) [1], that can be defined as a generic principle of classifier combination that uses L tree-structured base classifiers. The particularity of this kind of combination is that each decision tree is built from a random vector of parameters. It can be built for example by randomly sampling a feature subset (as in Random Subspace Method [2]), and/or by randomly sampling a training data subset (as in Bagging [3]).

RF is now known to be one of the most efficient classification methods ([1, 4, 5]). In a recent paper ([6]), we have shown the interest of designing a dynamic RF induction method, that would add trees to the ensemble by making them growing according to the trees already added to the committee. But for that purpose, it is crucial to find a criterion that could guide the tree induction so that it could suit to the rest of the ensemble. The main objective here is to minimize the error rate of the ensemble while adding more and more trees. For that purpose, we have followed Breiman's idea that the properties of *strength* and *correlation* could be very helpful for better understanding and controlling the behavior of RF ([1]). Our idea is to determine the relation between these two properties and the performance of RF. To do so, we have decided to generate a large set of forests that exhibit different error rates on the same test set, and to monitor their *strength* and *correlation*. Those forests are actually sub-forests generated thanks to a classifier selection method applied to a large RF: the principle is to grow a large pool of trees and to select different subsets from it, in order to simulate the growing of several RF, with different sizes and different performance. By this means, we show that error rates statistically decrease for a joint increasing *strength* and decreasing *correlation*.

The paper is thus organized as follows: we recall in section 2 the Forest-RI principles; in section 3, we explain our approach and describe our experimental protocol, the datasets used, and the results. 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 ([1])li 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 modifies the "traditional" 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.

In the literature, only few research works have focused on the number of trees that have to be grown in a RF. When introducing RF formalism in [1], Breiman demonstrated that for an increasing number of trees in the forest, the generalization 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. 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. Besides, in [6] we have shown that a subset of individual trees is able to outperform the whole ensemble. This led us to the conclusion that one could benefit from building trees in a more dependent way than it is actually done in traditional RF induction algorithm like Forest-RI for example. In this perspective, we have decided to focus on the two crucial properties of *strength* and *correlation*, that Breiman stated in [1] that they could be very helpful in a better understanding of RF behaviors. In the next section, we detail the definitions of these properties, and the result that led him to this conclusion.

3 Strength and Correlation

A RF is usually noted as an ensemble of individual 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. In [1], Leo Breiman introduces the *strength* and the *correlation* properties through an upper bound of the generalization error noted PE^*:

$$PE^* \leq \frac{\overline{\rho}(1 - s^2)}{s^2}$$

In this bound, s stands for the *strength* and $\overline{\rho}$ stands for the *correlation*. The main conclusion of this result is that the lower the ratio $\frac{\overline{\rho}}{s^2}$, the better the forest, since it gives better chances to obtain a low error rate.

Through the demonstration of this result, Breiman defines the margin function of a RF by the following equation:

$$mr(x, y) = P_{\Theta}(h(x, \Theta) = y) - \max_{j \neq y} P_{\Theta}(h(x, \Theta) = j)$$

where x is an input data, y its class, and where the subscripts Θ indicate that the probability is over the $\{\Theta_k\}$ family of random vectors. The *strength* is then defined as the expectation of this margin over the data space:

$$s = E_{x,y}[mr(x, y)]$$

Then, Breiman defines the raw margin function as

$$rm(\Theta, x, y) = I(h(x, \Theta) = y) - I(h(x, \Theta) = \hat{j}(x, y))$$

where $\hat{j}(x, y)$ represents the index of the "best" class among the wrong classes, defined by:

$$\hat{j}(x, y) = arg \max_{j \neq y} P_{\Theta}(h(x, \Theta) = j)$$

What is called here *correlation* is actually the statistical mean correlation between $rm(\Theta, x, y)$ and $rm(\Theta', x, y)$ over all pairs of (Θ,Θ').

The demonstration of these results strongly depends on the assumption that RF contains a "large" number of trees. However, we have already mentioned in the previous section that the number of trees does not have to be that large to obtain an accurate RF. As a consequence, we have decided in this work to focus on RF made of reasonably small number of trees, *i.e.* from 50 to 200 trees, in order to determine the actual relation between *strength*, *correlation* and error rates. The goal is to confirm Breiman's theoretical results for RF of different sizes, and thus to give elements of understanding of RF behaviors that could be helpful for guiding the RF induction. We detail our classifier selection approach for that purpose in the next section.

4 Tree Selection

For studying the relation between the ratio $\frac{\overline{\rho}}{s^2}$ and the error rate, we generate a pool of forests and we measure for all of them the *strength*, the *correlation* and the error rate. To be able to accurately determine this relation it is necessary to have at our disposal an ensemble of RF that exhibit error rates in a large range of values. For that purpose, we have chosen to apply a classifier selection approach to RF, in order to generate a large ensemble of sub-forests with different error rates on the same test set. Among the classifier selection approaches that we can be found in the literature, we are interested in what is called the *wrapper* approach, since it specifically consists in selecting the subset of classifiers that *a posteriori* optimizes the combination performance ([7]). Moreover, since our goal is to generate a large pool of sub-forests so that we could observe significant differences between these sub-forests in terms of error rates, we have decided to use Genetic Algorithms (GA) ([8]) for decision tree selection in RF. This approach allows to produce sub-forests that could potentially exhibit high and low error rates.

Hence, we have applied this classifier selection strategy on several datasets described in the next subsection.

Table 1. Datasets description

Dataset	# Samples	# Features	# Classes	Dataset	# Samples	# Features	# Classes
Diabetes	768	8	2	OptDigits	5620	64	10
Gamma	19020	10	2	Page-blocks	5473	10	5
Isolet	7797	616	26	Pendigits	10992	16	10
Letter	20000	16	26	Segment	2310	19	7
Madelon	2600	500	2	Spambase	4610	57	2
Mfeat-factors	2000	216	10	Vehicle	946	18	4
Mfeat-fourier	2000	76	10	Waveform	5000	40	3
Mfeat-karhunen	2000	64	10	Digits	38142	330	10
Mfeat-zernike	2000	47	10	DigReject	14733	330	2
Musk	6597	166	2	Mnist	60000	85	10

4.1 Datasets used

The 20 datasets used in our experiments are described in Table 1. The first 17 datasets in this table have been selected from the UCI repository [9]. The last 3 datasets are handwritten character recognition databases; the MNIST database ([10]) with a 85 multiresolution density feature set ($1 + 2 \times 2 + 4 \times 4 + 8 \times 8$) built from greyscale mean values; Digits and DigReject both described in [11], on which a 330-feature set has been extracted, made from three state-of-the-art kinds of descriptors, as detailed in [12].

4.2 Experimental protocol

First, each dataset has been randomly split into a training and a test subset, containing respectively two thirds and one third of the original dataset. We denote by T_r the training set and by T_s the test set. Then, a RF has been grown from T_r, with a number L of trees fixed to 500. The value of the hyperparameter K, which denotes the number of features randomly selected at each node of the trees, has been fixed to $\sqrt{M}$, M being the dimension of the feature space, which is a default value commonly used in the literature ([12]). A classifier selection process using a GA has been then applied to this forest. The size of sub-forests generated during this process is fixed, so that all of them could be fairly compared with each other, as the number of trees could affect the calculation of *strength* and *correlation*. The selection procedure through GA is conducted for the following sizes of sub-forests: 50, 100, 150 and 200. Concerning the GA parameters, they have been fixed to the following classical values: number of generations fixed to 300, population size to 50, mutation probability to 0.01, crossover probability to 0.60 and selection proportion to 0.80. At each new generation, several new sub-forests, are generated using the crossover and mutation operators. Therefore, for each size of sub-forest (50, 100, 150, 200) and for each dataset (among the 20 datasets used), a total of 15000 sub-forests are studied, *i.e.* 50 sub-forests for each generation step multiplied by 300 generations. For each of these 15000 RF, the error rate, the *strength* and the *correlation* have been measured on T_s. All these measures are analyzed and discussed in the next subsection.

4.3 Results and Discussion

The results obtained with the experimental protocol detailed in the previous subsection can be illustrated with a 2D-point cloud, each point being a sub-forest represented by its error rate and its value of $\frac{\overline{\rho}}{s^2}$. Figure 4.3 gives the results obtained with sub-forests made of 50 trees. Note that the tendancies observed on this figure and the conclusion drawn from them can be extended to all the sub-forest sizes tested in these experiments (please see [13] for other sizes). To have a better idea of the shape of the clouds, a regression line has been drawn on each diagram. These lines illustrate the relation between $\frac{\overline{\rho}}{s^2}$ and error rates. One can clearly observe a decrease of $\frac{\overline{\rho}}{s^2}$ for decreasing error rates. This observation is consistent with Breiman's theoretical result ([1]) since it indicates that the smaller the ratio, the better chances for the error rate to be low. Note however that this relation is thus still verified even for small forests, which was not demonstrated by Breiman since his result lied on the assumption of a large number of trees grown in the RF.

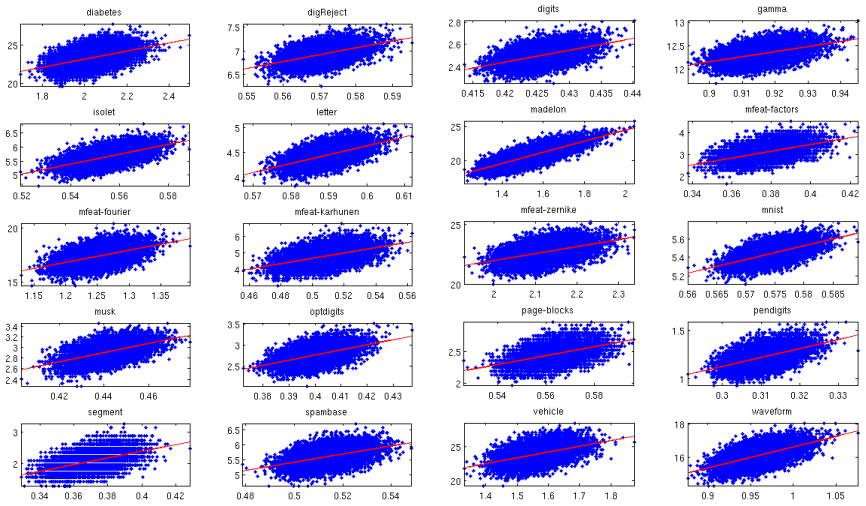

Fig. 1. Error rates (y-axis) according to $\frac{\overline{\rho}}{s^2}$ values (x-axis) for all the sub-forests of 50 trees, obtained during the selection process. The red line is the regression line of the cloud.

In the perspective of designing a dynamic RF induction algorithm, this result is interesting because it indicates that minimizing the ratio $\frac{\overline{\rho}}{s^2}$ should allow to minimize *a priori* the error rate. Thus, using *strength* and *correlation* as criteria for guiding the tree induction in a RF growing process should be a promising idea. Hence, a direct perspective of this work is to study a mean to guide the tree induction process to make it fit to the rest of the forest, by maximizing the current ensemble *strength* and jointly minimizing its *correlation*. This could be done for example by weighting the training

data so that each tree would focus its induction on previous errors, which is actually part of a work in progress that aims at designing a complete dynamic RF induction algorithm.

5 Conclusion

In this paper we have studied the *strength* and *correlation* properties of RF in a perspective of designing a dynamic RF induction process. The interest of inducting a RF in a more dependent way than it is traditionaly done in a RF induction algorithm has already been demonstrated in a recent paper ([6]), but it requires the use of a criterion for the guidance of such a sequential procedure. We have thus proposed in this paper a study on the feasability of using the *strength* and *correlation*, two crucial properties of RF, for that purpose. We have firstly shown that the relation between these two properties and the performance, theoretically established by Breiman for RF made of a large number of trees, is still verified with smaller RF, *i.e.* made of 50 to 200 trees. We have thus shown that such a dynamic algorithm can be designed according to the joint objectives of maximizing the *strength* and minimizing the *correlation*.

References

1. Breiman, L.: Random forests. Machine Learning **45**(1) (2001) 5–32
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. Breiman, L.: Bagging predictors. Machine Learning **24**(2) (1996) 123–140
4. Boinee, P., Angelis, A.D., Foresti, G.: Ensembling classifiers - an application to image data classification from cherenkov telescope experiment. World Academy of Science, Engineering and Technology **12** (2005) 66–70
5. Banfield, R., Hall, L., Bowyer, K., Kegelmeyer, W.: A comparison of decision tree ensemble creation techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(1) (2006) 173–180
6. Bernard, S., Heutte, L., Adam, S.: On the selection of decision trees in random forests. International Joint Conference on Neural Network (2009) 302–307
7. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence **97**(1-2) (1997) 273–324
8. 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
9. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
10. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324
11. Chatelain, C., Heutte, L., Paquet, T.: A two-stage outlier rejection strategy for numerical field extraction in handwritten documents. Internation Conference on Pattern Recognition, Honk Kong, China **3** (2006) 224–227
12. Bernard, S., Heutte, L., Adam, S.: Influence of hyperparameters on random forest accuracy. International Workshop on Multiple Classifier Systems (2009) 171–180
13. Bernard, S.: Forêts aléatoires : De l'analyse des mécanismes de fonctionnement à la construction dynamique. Thèse de Doctorat, Université de Rouen, France (2009)