RANDOM FORESTS AND DECISION TREES CLASSIFIERS : EFFECTS OF DATA QUALITY ON THE LEARNING CURVE

gembloux
faculté universitaire
des sciences agronomiques

Y. Brostaux[1]

[1]*Gembloux Agricultural University, Gembloux, Belgium*

Email: *brostaux.y@fsagx.ac.be*

SIMa

Abstract

Random forests have been introduced by Leo Breiman (2001) as a new learning algorithm, extending the capabilities of decision trees by aggregating and randomising them. We explored the effects of the introduction of noise and irrelevant variables in the training set on the learning curve of a random forest classifier and compared them to the results of a classical decision tree algorithm inspired by Breiman's CART (1984). This study was realized by simulating 23 artificial binary concepts presenting a wide range of complexity and dimension (4 to 10 relevant variables), adding different noise and irrelevant variables rates to learning samples of various sizes (50 to 5000 examples). It appeared that random forests and individual decision trees have different sensitivities to those perturbation factors. The initial slope of the learning curve is more affected by irrelevant variables than by noise on both algorithms, but counterintuitively random forests show a greater sensitivity to noise than decision trees for this parameter. Globally, average learning speed is quite similar between the two algorithms but random forests better exploit both small and big samples : their learning curve starts lower and is not affected by the asymptotical limitation showed by single decision trees.

Introduction

In 2001, Leo Breiman published a new learning algorithm consisting of aggregation and randomisation of decision trees constructed on the same learning set, the Random Forests. Researches have been conducted to compare this method with existing ones, such has bagging, SVM, neural networks, but we lack a systematic view on the effects of the quality of the learning set on its intrinsic performances. This study put its emphasis on such effects, taking a single CART-like decision tree classifier as a control.

Methods

We simulated 23 artificial binary concepts presenting a wide range of complexity and dimension (4 to 10 relevant attributes). Samples were extracted from these theoretical spaces to form the learning sets used to train the random forests (500 trees, random selection of $\log_2(M + 1)$ attributes at each node).

We explored five sample sizes (50, 100, 500, 1000, 5000) and two types of data perturbation : addition of irrelevant attributes (random Bernouilli variables, 0%, 25%, 50%, 100%, 200% of the number of relevant attributes) and noise (substitution of the target variable by a random Bernouilli variable, 0%, 5%, 10%, 25%, 50% of the learning sample). Each combination of samples' parameters and concepts has been replicated 20 times.

We trained a random forest and a single pruned CART-like tree on each random sample and recorded the confusion matrix of the corresponding entire theoretical domain space, which was used to estimate the true misclassification rate of the classifiers.

Simulation and analysis were conducted under the R statistical programming environment (Ihaka & Gentleman, 1996).

Results

Examination of the learning curves showed that the initial decrease of the misclassification rate of both algorithms is approximately linear with the $\log_{10}$ of the training sample size, but is sometimes affected by an asymptotical limitation for the biggest samples, essentially for the CART-like trees. To take this behaviour into account, learning curves were adjusted by segmented linear regression models (Muggeo, 2003).

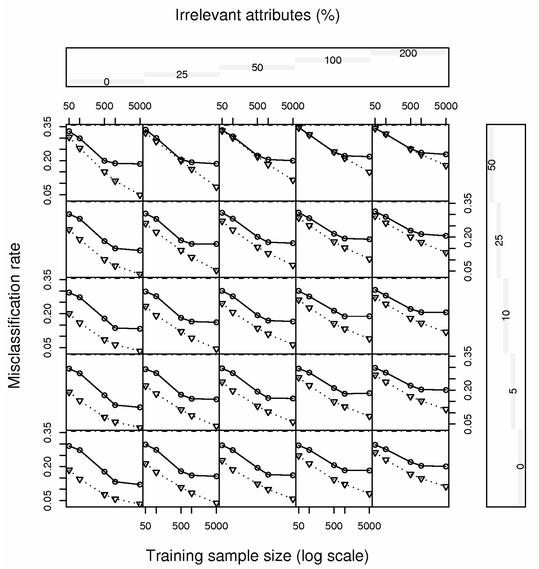

Figure 1. Evolution of misclassification rate with the training sample size (–o– single decision tree, ··∇·· RF, rows = noise levels (%), columns = irrelevant attributes (%).

Average initial learning speed is quite similar in both algorithms. Addition of irrelevant attributes reduced initial learning speed of both RF and single trees (dilution effect), but the noise level only affected RF algorithms.

Similarly, initial error rate increased with noise level in both algorithms, but surprisingly single trees were unaffected by addition of irrelevant variables.

However average initial error rate was as expected lower for the RF than for the single decision tree, so that the learning curves of RF always stayed under single tree's ones.

Conclusions

Despite their reputation of robustness attributed to bagging-based methods, performances of random forests were significantly deteriorated by degradation of the quality of the learning sample, sometimes more then single tree classifiers. But in average, the aggregated classifiers outperformed clearly the latter, extracting more information from a given sample. Moreover, random forests are not affected by the asymptotical limitation of the misclassification rate showed by single decision tree classifiers, and hence take better advantage of the additional information contained in biggest learning samples.

References

BREIMAN L. [2001]. Random forests. *Machine Learning*, **45**(1), 5-32.

BREIMAN L., FRIEDMAN J.H., OLSHEN R. et STONE C. [1984]. *Classification and Regression trees*. Belmont, CA, Wadsworth International Group.

IHAKA R. et GENTLEMAN R. [1996]. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **5**(3), 299-314.

MUGGEO V.M.R. [2003]. Estimating regression models with unknown break-points. *Statist. Med.* , **22**(19), 3055-3071.