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

Y. Brostaux[1]

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

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

Random forests have been introduced by Leo Breiman (2001) as a new learning algorithm, extending the capabilities of decision trees by aggregating and randomizing 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 dimensions (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.