TEST manuscript No.
(will be inserted by the editor)

Comments on: A random forest guided tour

Pierre Geurts · Louis Wehenkel

Received: date / Accepted: date

We would like to congratulate the authors for having put together a very interesting, thoughtful, and timely review paper about Random Forests (RF). This paper will definitely serve as a reference on RF and foster further theoretical and applied research about this family of algorithms.

In our comments below, we will refer to Biau and Scornet's review paper as BS. Based on our own experience with tree-based methods, we bring up a few complementary points that were not explicitly addressed in BS, and which we believe to be of importance to well position some relevant related work and also to foster some interesting directions for further work.

Alternative randomization schemes

Research on tree-based ensemble methods dates back to the 19-nineties and was motivated at that time by the success of generic ensemble methods for supervised learning, such as Breiman (1996)'s Bagging and Freund and Schapire (1997)'s Adaboost algorithms, that can be combined with any kind of base-model. It was soon found that these ensemble methods are verry effective when applied on top of decision or regression trees, and this success fostered further research, in particular towards the design of alternative randomization techniques exploiting the specificities of tree-based models. Among the methods that were published at that time, Breiman's Random Forest method (BRF) was designed by combining two randomization techniques previously proposed in two generic ensemble methods: Breiman (1996)'s own bagging idea, by using bootstrap resampling of the learning

P. Geurts
Institut Montefiore, University of Liège, Sart Tilman B28, B4000 Liège, Belgium
Tel.: +32-4-3664815
E-mail: p.geurts@ulg.ac.be

L. Wehenkel
Institut Montefiore, University of Liège, Sart Tilman B28, B4000 Liège, Belgium
Tel.: +32-4-3662684
E-mail: l.wehenkel@ulg.ac.be

sample, and Ho (1998)'s *global* random subspace idea, by introducing randomization through the *local* pre-selection of a random subset of features used for deciding how to split at any tree node. The excellent performances of the BRF method has since been confirmed empirically in many contexts. This algorithm has also been implemented in several widely used and publicly available software packages, and has become the *de facto* standard RF method, well known and used in many application fields.

Since the publication - in 2001 - of Breiman's RF algorithm, a number of alternative randomization methods have also been proposed in the literature (see Section 4.2 of BS for a discussion of some of them), each one having been shown empirically to perform better than BRF on a significant number of benchmark problems. We contributed ourselves to this research effort by proposing the extremely randomized tree algorithm (ERT) (Geurts et al 2006a). With respect to Breiman's RF, ERT does not use bootstrap sampling, keeps the random subselection of features at each tree node, and it replaces the optimal cut-point selection by a purely random cut-point selection (as in the *uniform random forests* model discussed in Section 3.1 of BS). Due to the random cut-point selection, the ERT algorithm is typically much faster than BRF, while its accuracy has been shown empirically to be on many problems on par or better than that of BRF.

While often very effective, such alternative randomization methods, including those of BRF and ERT, have been designed on the basis of intuitions rather than from strong theoretical analyses and, therefore appear as heuristic. Surprisingly, while significant advances have recently been made towards the *theoretical characterization* of RF models, these characterizations do not yet provide any support for understanding and comparing these existing alternative randomization methods. In addition, none of these characterizations has proven that BRF was the best way to go, nor has led to the development of novel alternative randomization schemes with better properties than BRF. We believe that this lack of deeper understanding is due to several reasons. First, the most recent theoretical works focus on the *consistency* analysis of the RF estimators and thus on *asymptotic* settings, while the choice of a specific randomization scheme has mostly an impact on *convergence rates*. For example, purely random forests models (see Section 3.1 of BS) are consistent, while they have been shown to be typically less accurate in finite sample conditions than less strongly randomized models, in particular in the presence of irrelevant variables (Geurts et al 2006a). Second, most theoretical works focus on understanding Breiman's Random Forests *in its standard form*, while they do not analyse the impact of these *alternative randomization schemes* (with the exception of some simplifications made for mathematical convenience). Third, the impact of a randomization scheme in finite sample conditions is *problem dependent* (Geurts et al 2006a), and this makes theoretical analyses much more difficult.

Given this state of affairs, we believe that it would be very interesting to widen theoretical analyses around random forests so that they include a broader range of possible tree-specific or generic randomization schemes and, to target these analyses towards a better understanding of the optimal (problem-specific) way of carrying out this randomization. These questions are also directly related to the issue of choosing the optimal tuning parameters of Breiman's RF, which, as stated in BS's conclusion, is still an open theoretical question. Note that there already exists some literature trying to address these questions, mainly by studying the impact of different randomization schemes on the "diversity" of the result-

ing ensemble models, with the main argument that increasing "diversity" should be beneficial in the context of ensembles (Kuncheva and Whitaker 2003; Brown et al 2005). We think that this line of research is interesting, since it has lead to the design of new randomization schemes purposely targeting an increase in "diversity". However, as shown by Breiman's strength and correlation analysis (or, equivalently, by the bias/variance analysis made in Geurts et al 2006a), decreasing correlation will improve accuracy only if it does not come with a too important decrease of strength. Relating theoretically changes in correlation and strength to specific randomization schemes, under some useful assumptions about the nature of the target problem, is an open and potentially difficult question that should however be addressed in future research.

As a final thought, let us note that the choice of a randomization scheme should not be guided only by predictive performances. Indeed, in the "Big Data" era, this degree of freedom in RF methods can be also advantageously exploited to improve their computing times and scalability. For example, while *locally* randomizing the set of variables at each tree node is an essential ingredient of Breiman's RF, on the other hand, building each tree on a subset of variables randomly picked *globally* before building the tree (as in Ho (1998)'s original random subspace method) has the advantage that only this subset of variables needs to be accessible when training an individual tree. Therefore, although sampling the features globally instead of locally affects the asymptotic properties of RF (e.g., it of course prevents individual trees to be consistent), it neverthless results in a more convenient algorithm when it comes to distribute data and computations over different computing nodes and to address problems where all data can not fit into memory. On our side (Louppe and Geurts 2012), we have shown empirically that a "random-patching" approach, consisting in growing individual trees on a sub-dataset obtained by subsampling both features and observations, was often quite competitive in terms of accuracy. Further research is thus also needed to analyse theoretically the impact of the randomization scheme on the large sample properties of the resulting estimator, and so to design new randomization schemes providing potentially better tradeoffs between accuracy and scalability in such conditions.

Further extensions

With respect to the list of extensions of the RF models discussed in Section 6 of BS, we would like to comment below on three other important lines of research of interest, namely i.) multivariate and output kernelized regression trees, ii) soft tree models, iii) bayesian approaches for decision tree and forest learning.

Multivariate and output kernelized regression trees

While BS focuses on standard univariate regression and classification settings, tree models, and therefore forests, can be very naturally extended to handle multivariate regression problems, where one tree or forest is built to predict several outcomes at once (see e.g. Blockeel et al 1998; Geurts et al 2006b; Segal and Xiao 2011; Kocev et al 2013). These extensions are obtained by changing the CART-split criterion and by attaching vectorial predictions at tree leaves. A simple adaptation

of the split criterion consists in taking the sum over all outcomes of the standard CART criterion but other more sophisticated criteria have also been proposed (Segal and Xiao 2011). Adapting theoretical results related to standard RFs to the multivariate setting would be very interesting. One additional open question in this domain is to identify the conditions under which a single multivariate forest is expected to perform better (or worse) than several individual forests each grown separately to predict one specific outcome. So far, this question has only been addressed empirically in the literature (e.g., in Kocev et al 2013).

Let us also mention the generalization of multivariate forests that we proposed in (Geurts et al 2006b) to handle kernelized outputs, i.e., outputs that are not expressed efficiently in the form of a vector of real numbers, but that can be efficiently compared using a kernel (such as strings, images, graphs, for example). The idea behind this generalization is that the empirical variance over the vector output-space induced by the output-kernel, and used in the multi-output CART split criterion discussed above, can be rewritten by using only the output-kernel (i.e. the "kernel-trick"). This extension allows to grow and make predictions with trees in very complex output spaces. Interestingly, this kernelization of the output-space of forests builds on the input-space kernel interpretation of forests explained in Section 3.5 of BS. Indeed, it can be shown that a random forest can be used to compute an approximation of the output-space kernel among two test instances, in the form of a convolution of the input-space kernel of the forest with the output-space kernel values among the the training instances (Geurts et al 2006b).

Soft tree models

One of the trademarks of tree learning algorithms is the recursive partitioning that they induce over the input space. This recursive partitioning makes the tree construction algorithm very efficient and it is also at the heart of the interpretability of trees and forests. Indeed, as a result of recursive partitioning, each tree leaf is defined by a conjunction of simple tests based each one on a single predictor variable, and thus making a single tree very easy to understand as a set of (mutually exclusive and exhaustive) rules. The greedy approach to build this recursive partitioning also leads to the local selection of the most relevant variables, which further enhances interpretability. Despite these advantages, it has been argued that recursive *partitioning* is one of the reasons for the rather low accuracy of single trees, since it renders choices at deeper nodes of the trees dependent on only few samples and thus highly unstable (Geurts 2002). Geometrically, it also leads to piecewise constant approximations that require huge trees, and thus huge learning samples, to approximate smooth functions with sufficient accuracy.

Building a forest of random trees is one way to circumvent these issues and it indeed very efficiently does so. An alternative approach is to exploit soft tree models, namely trees where discrete binary splits are replaced by soft splits that send a test example down to all successor nodes with weights that depend on the value of the input variables for this example (Jordan 1994; Olaru and Wehenkel 2003; Geurts and Wehenkel 2005; Yildiz and Alpaydin 2013). For a given test example, tree propagation with soft splits thus computes a weight for each tree leaf and a final prediction is obtained by aggregating all leaf predictions by taking into account these weights. Such weights have been interpreted as the probability

that an example will reach a leaf given its input values (e.g. Jordan 1994; Geurts and Wehenkel 2005), but they have also been interpreted as fuzzy set membership degrees in the literature about fuzzy decision trees (Olaru and Wehenkel 2003). Essentially two approaches have been proposed to grow soft tree models: the first one is to start from a standard CART tree and then to soften its splits in a post-processing stage, e.g. (Geurts and Wehenkel 2005). The second one is to directly grow a soft tree model by adapting the tree construction algorithm, e.g. (Jordan 1994; Olaru and Wehenkel 2003).

Soft trees turn out to be often much more accurate than "crisp" CART trees. It also is easy to realize that a soft split using a sigmoid function can be simulated by an infinite ensemble of standard crisp splits where the attribute would be fixed and the cut-point would be randomly drawn from a logistic distribution (since the cumulative distribution function of a logistic distribution is a sigmoid). Softening decision trees thus has a similar effect as building an ensemble by only randomizing the cut-points in a post-processing stage applied on top of a single tree (Geurts et al 2001; Geurts and Wehenkel 2005). Of course ensembles of soft trees could also be built by randomizing their choice of splitting variables.

In addition to the improvement of accuracy, another advantage of soft tree models is their smoothness. A soft tree indeed represents a smooth function that can be differentiated with respect to its arguments and parameters. This opens the door to gradient descent techniques based on the back-propagation algorithm for refitting decision tree parameters (Olaru and Wehenkel 2003), with all their associated benefits such as the possibility to regularize and to incorporate prior constraints on model predictions. The price to pay with respect to crisp trees is however a significant increase of computing times both for training these models and for making predictions, and also some loss of interpretability as every test example can possibly follow all paths in the tree, which make these models less transparent than crisp trees.

Actually, soft trees share many of the features of multi-layer artificial neural networks. This parallel has been noted and exploited in the literature, where softened decision trees have been used for example to initialize the structure of neural networks (Sethi 1990). As a matter of fact, although soft tree models have been proposed already a long time ago, there seems to be a renewed interest for these models in the recent machine learning literature (see, e.g., Kontschieder et al 2015; Ioannou et al 2016), and in particular about their link with deep neural networks, certainly motivated by the recent successes of these latter methods. We believe that analysing more thoroughly the link between random forests and soft trees, on the one hand, and between soft trees and deep networks, on the other hand, is a very exciting direction for future research and should improve our understanding of all these seemingly different families of methods.

Bayesian approaches for decision tree and forest learning

To further broaden the perspective for understanding RF methods, let us briefly highlight the literature on bayesian approaches for learning decision trees or forests that have been proposed by several authors since more than twenty years ago (see, e.g., Buntine 1992; Chipman et al 1998, 2010; Lakshminarayanan et al 2013; Quadrianto and Ghahramani 2014; Matthew et al 2015). The common idea of these

methods is to define a prior distribution on the set of all possible decision trees, which typically penalizes more complex trees, and then to derive from this prior and the data a posterior distribution over all possible trees. A prediction is then obtained either by using a single tree of maximal posterior probability or, in the full bayesian setting, by computing the conditional expectation of predictions over the posterior distribution of all trees. The exact computation of this latter expectation being generally intractable, it is typically approximated by using sophisticated Monte Carlo methods, focusing on the modes of the posterior distribution, i.e. on the trees having a good compromize between their complexity and their fit to the dataset. There are obvious connections between forests of randomized trees and these bayesian approaches for decision tree averaging and exploring further these connections could potentially improve our understanding of random forests and, related to the first section of this comment, could also lead to the design of more principled randomization schemes to construct such forests while exploiting prior information about the considered problem.

References

Blockeel H, Raedt LD, Ramon J (1998) Top-down induction of clustering trees. In: Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '98, pp 55–63

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Brown G, Wyatt J, Harris R, Yao X (2005) Diversity creation methods: a survey and categorisation. Information Fusion 6(1):5 – 20, diversity in Multiple Classifier Systems

Buntine W (1992) Learning classification trees. Statistics and Computing 2:63–73

Chipman HA, George EI, , McCulloch RE (1998) Bayesian cart model search. Journal of the American Statistical Association 93(443):935–960

Chipman HA, George EI, McCulloch RE (2010) Bart: Bayesian additive regression trees. Ann Appl Stat 4(1):266–298

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139

Geurts P (2002) Contributions to decision tree induction: bias/variance tradeoff and time series classification. PhD thesis, University of Liège, Belgium

Geurts P, Wehenkel L (2005) Closed-form dual perturb and combine for tree-based models. In: Proceedings of the 22Nd International Conference on Machine Learning, ACM, New York, NY, USA, ICML '05, pp 233–240

Geurts P, Olaru C, Wehenkel L (2001) Improving the bias/variance tradeoff of decision trees - towards soft tree induction. Engineering intelligent systems 9:195–204

Geurts P, Ernst D, Wehenkel L (2006a) Extremely randomized trees. Mach Learn 63(1):3–42

Geurts P, Wehenkel L, d'Alché Buc F (2006b) Kernelizing the output of tree-based methods. In: Cohen WW, Moore A (eds) ICML, ACM, ACM International Conference Proceeding Series, vol 148, pp 345–352

Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20(8):832–844

Ioannou Y, Robertson D, Zikic D (2016) Decision Forests, Convolutional Networks and the Models in-Between. arXiv:160301250

Jordan MI (1994) Hierarchical mixtures of experts and the em algorithm. Neural Computation 6:181–214

Kocev D, Vens C, Struyf J, Deroski S (2013) Tree ensembles for predicting structured outputs. Pattern Recognition 46(3):817 – 833

Kontschieder P, Fiterau M, Criminisi A, Bulò SR (2015) Deep Neural Decision Forests. ICCV pp 1467–1475

Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 51(2):181–207

Lakshminarayanan B, Roy DM, Teh YW (2013) Top-down particle filtering for Bayesian decision trees. ICML pp 280–288

Louppe G, Geurts P (2012) Ensembles on random patches. In: Flach PA, Bie TD, Cristianini N (eds) ECML/PKDD (1), Springer, Lecture Notes in Computer Science, vol 7523, pp 346–361

Matthew T, Chen CS, Yu J, Wyle M (2015) Bayesian and Empirical Bayesian Forests. ICML pp 967–976

Olaru C, Wehenkel L (2003) A complete fuzzy decision tree technique. Fuzzy Sets Syst 138(2):221–254

Quadrianto N, Ghahramani Z (2014) A Very Simple Safe-Bayesian Random Forest. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(6):1297–1303

Segal M, Xiao Y (2011) Multivariate random forests. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(1):80–87

Sethi IK (1990) Entropy nets: from decision trees to neural networks. Proceedings of the IEEE 78(10):1605–1613, DOI 10.1109/5.58346

Yildiz OT, Alpaydin E (2013) Regularizing soft decision trees. In: Information Sciences and Systems 2013 - Proceedings of the 28th International Symposium on Computer and Information Sciences, ISCIS 2013, Paris, France, October 28-29, 2013, pp 15–21