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

Comments on: A Random Forest Guided Tour

Peter Bühlmann · Florencia Leonardi

Received: date / Accepted: date

We congratulate Gérard Biau and Erwan Scornet for an interesting paper on an important topic, namely towards better understanding of Random Forests and related ensemble schemes.

1 Some further thoughts on the paper

Biau and Scornet (referred in the sequel as "BS") present a nice overview on recent developments for Random Forests. As mentioned in BS, already Amit and Geman (1997) proposed to randomly select covariables (or "features") during the process of learning decision trees, and average at the final stage: their motivation was mainly of computational nature, in order to deal with very many features. In fact, Breiman (2001) is referring to the paper by Amit and Geman, but he certainly is the person who has made a pioneering contribution with Random Forests pointing out its stunning accuracy in a wide range of problems, its versatility and introducing also concepts of variable importance.

Peter Bühlmann
Seminar for Statistics, ETH Zürich, CH-8092 Zürich
Tel.: +41 44 632 7338
Fax: +41 11 632 1228
E-mail: buhlmann@stat.math.ethz.ch

Florencia Leonardi
Departamento de Estatística, Universidade de São Paulo
São Paulo - SP - Brasil
Tel.: +55 11 3091 6119
Fax: +55 11 3091 6130
E-mail: florencia@usp.br

1.1 Improving the performance of Random Forests?

Random Forests as proposed by Breiman (2001) is surprisingly accurate for regression and classification problems, and there is empirical support that it is among the "best off-the-shelf classifiers/estimators". In particular, the fact that the performance of the algorithm is rather insensitive to the choice of the tuning parameters justifies its use without the need to carefully choose a regularization parameter with e.g. cross-validation. We note that this finding is very different from other nonparametric or high-dimensional estimation schemes such as the Lasso (Tibshirani, 1996) or versions thereof.

Improving Random Forests, with respect to a wide range of applications and datasets, has been found to be very difficult. Recently, Cannings and Samworth (2015) proposed a random projection ensemble method for classification, and they report a couple of scenarios where Random Forests can be outperformed by some of their random projection ensemble classifiers. However, when Random Forests competes against *one* of their proposed random projection methods, the empirical results do not point very clearly in favor of one or the other method.

1.2 Subsampling and bootstrapping

The theoretical arguments for Random Forests seem to be much better developed for its version with subsampling instead of bootstrapping. We point here to an older result of Freedman (1977), saying that subsampling with subsample size a_n is closest to bootstrap resampling with respect to the total variation norm for $a_n = \lfloor n/2 \rfloor$. This fact has been also empirically exploited in Bühlmann and Yu (2002) with their subagging procedure when compared to bagging (Breiman, 1996). The theory described in BS is interesting: for example, that the median forest is consistent if $a_n = o(n)$ (Scornet, 2015), and that asymptotic normality holds if $a_n = o(\sqrt{n})$ (Mentch and Hooker, 2015). All these results exclude the range where $a_n \sim Cn$ for some $0 < C < 1$. Is it a fundamental limitation that $a_n = o(n)$ (e.g. for consistency) or is it rather a lack of techniques to deal with U-statistics of order a_n where a_n is very large and asymptotically proportional to n?

1.3 Variable importance

Variable importance is very crucial for many practical applications. Constructing importance measures based on Random Forests is very interesting as it enables applications with mixed data-types (continuous and categorical data): for example, Fellinghauer et al (2013) make use of Random Forests variable importance for constructing conditional independence graphs.

The importance measure which have been proposed so far seem to work "reasonably well", but they seem to lack a more rigorous multivariate justification. For example, the $\widehat{MDA}$ measure and its population version MDA^*

are reflecting a marginal aspect, as $\widehat{MDA}$ is based on marginal permutation of the variable of interest. A multivariate interpretation, e.g. as for a regression coefficient in a (generalized) linear model, cannot be easily achieved unless one adopts the computationally cumbersome conditional importance measure (Strobl et al, 2008).

This remains a topic of future research.

1.4 Extensions based on Random Forests

As BS point out, Random Forests can be used in other settings than classification or regression.

Missing data problems are briefly mentioned in BS. Another powerful method is MissForest (Stekhoven and Bühlmann, 2012): it uses Random Forests regression iteratively to impute the missing values. The method enjoys the same advantage as Random Forests for regression and classification, namely that it can be easily used for mixed type data, taking nonlinearities and interactions into account.

For survival problems, Random Forests and other ensemble methods have been proposed also in Hothorn et al (2006), based on the idea of weighted resampling with inverse probability of censoring (IPC) weights (van der Laan and Robins, 2003, cf.).

2 Some outlook: inhomogeneous large-scale data

The Random Forests methodology and also the presented theory in BS rely on the assumption that the data are i.i.d. realizations of $(X_1, Y_1), \ldots, (X_n, Y_n)$. In particular for large-scale data (or "big data") where n is large, the i.i.d. assumption is questionable. It can be weakened by assuming that the data comes from G *unknown* groups, with i.i.d. realizations within each group. If the groups are completely unstructured, this corresponds to a mixture model with G components. For example, a mixture of high-dimensional regression models has been considered by Städler et al (2010). In view of no further structural assumption about the groups, the problem of estimating the mixture components is rather difficult. An easier case occurs when the groups contain consecutive observations: such a scenario then corresponds to a change point problem. Especially for large-scale ("big") data, we believe that a change point model of the following form is a reasonable approximation:

$$\begin{aligned} &Y_i = f_i(X_i) + \sigma\varepsilon_i \ (i = 1, \ldots, n), \\ &\varepsilon_1, \ldots, \varepsilon_n \text{ i.i.d. with } \mathbb{E}[\varepsilon_i] = 0, \ \mathrm{Var}(\varepsilon_i) = 1, \end{aligned} \tag{1}$$

where $f_i(\cdot)$ is piecewise constant as a function of i, and ε_i is independent of X_i. That is, when having G segments with corresponding change points scaled

to the interval $[0,1]$, namely, $\alpha_0 = 0 < \alpha_1 < \ldots < \alpha_G = 1$,

$$f_i(x) = \sum_{g=1}^{G} f(g;x)\mathbf{I}(i/n \in (\alpha_{g-1}, \alpha_g]),$$

involving G nonparametric regression functions $f(1;\cdot), \ldots, f(G;\cdot)$ for the G different segments. A related but high-dimensional parametric linear change point model has been proposed by Leonardi and Bühlmann (2016). There, the regression function $f(g;x)$ is fitted with an ℓ_1-norm penalized linear function $\beta_g^T x$: replacing it by a Random Forests regression function, we propose the following joint estimator for the change points and the nonparametric regression functions: for $\gamma > 0$

$$\hat{\alpha} \;=\; \operatorname*{arg\,min}_{G,\,\alpha} \Big\{ \sum_{g=1}^{G} L_n(\alpha_{g-1}, \alpha_g) + \gamma G \Big\}, \tag{2}$$

where the loss function L_n is given by

$$L_n(\alpha_{g-1}, \alpha_g) = n^{-1} \sum_{i=1}^{n} (y_i - \hat{f}(g;x_i))^2 \, \mathbf{I}(i/n \in (\alpha_{g-1}, \alpha_g]), \tag{3}$$

the functions $\hat{f}(g;\cdot)$ are the Random Forest regression functions estimated on the subsample $\{(X_i, Y_i)\colon \mathbf{I}(i/n \in (\alpha_{g-1}, \alpha_g]\}$ from a segment $(\alpha_{g-1}, \alpha_g]$, and the minimization in (2) is over the set of all vectors $\alpha = (\alpha_0, \ldots, \alpha_G)$ satisfying $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_G = 1$ and $\alpha_g - \alpha_{g-1} \geq \delta$ for all g, with $\delta > 0$. The parameter δ ensures that the estimated segments will not become too small, containing at least $\delta \cdot n$ data points.

2.1 Binary Segmentation algorithm

It is proved in Leonardi and Bühlmann (2016) for a high-dimensional linear change point model, that a binary segmentation algorithm leads to an estimator which has the same statistical properties as the global estimator in (2), in terms of an oracle inequality which bears some similarity to the one from high-dimensional regression (Bühlmann and van de Geer, 2011, cf.). The algorithm here is as follows.

For $0 \leq u < v \leq 1$ define

$$h(u,v) = \operatorname*{arg\,min}_{s \in \{u\} \cup [u+\delta, v-\delta]} \{ L_n(u,s) + L_n(s,v) + \gamma(1 + \mathbf{I}(s > u)) \}. \tag{4}$$

The binary segmentation algorithm works by computing the best single change point for the interval $(0,1]$ (obtained when $h(0,1) \neq 0$) and then to iterate this criterion on both segments separated by this point, until no more change points are found (due to the penalty in the objective function). We can describe this algorithm by using a binary tree structure T with nodes labeled by subintervals $(u,v] \subset (0,1]$. The steps of the algorithm are given by:

1. Initialize T to the tree with a single root node labeled by $(0, 1]$.
2. For each terminal node $(u, v]$ in T compute $s = h(u, v)$. If $s > u$ add to T the additional nodes $(u, s]$ and $(s, v]$ as descendants of node $(u, v]$.
3. Repeat 2. until no more nodes can be added to T.

The set of terminal nodes in T, denoted by T^0, can be identified with the estimated change point vector $\hat{\alpha}$, by picking up the extremes in these intervals; that is

$$\hat{\alpha} = \bigcup_{(u,v]\in T^0} \{u, v\}\,.$$

2.2 Some numerical illustrations

Consider the following regression model with change points as in (1):

$$\begin{aligned}
&\varepsilon_i \sim \mathcal{N}(0,1),\ \sigma = 1;\quad X_i \sim \mathcal{N}_p(0, \Sigma),\ \Sigma_{ij} = 0.8^{|i-j|}\ \forall i, j;\\
&\alpha_0 = 0, \alpha_1 = 0.3, \alpha_2 = 0.7, \alpha_3 = 1;\\
&f(g = 1; x) = \sin(x^{(1)}) + x^{(2)} + x^{(3)} + x^{(1)}x^{(2)},\\
&f(g = 2; x) = \sin(x^{(1)}) + x^{(p-1)} + x^{(p)} + x^{(p-2)}x^{(p-1)},\\
&f(g = 3; x) = x^{(1)} + x^{(1)}x^{(20)} + \sin(x^{(50)})
\end{aligned} \tag{5}$$

with $p = 2n$ and $n \in \{50, 100, \ldots, 250\}$. For 20 independent replications of sample size n we compute the estimated change points and the number of groups given by the binary segmentation algorithm of Section 2.1, using the loss function based on Random Forests defined in (3). The boxplots corresponding to the first estimated change point and the barplots for the estimated number of groups for each sample size n are summarized in Figure 1, suggesting asymptotic consistency as $n \to \infty$ (while $p = 2n \to \infty$ as well). For the simulations we used $\delta = 0.1$ and a γ parameter depending on n and p, given by $\gamma(n) = \sqrt{\log(p)/n}$. In practice the number of groups can be selected by a cross-validation procedure as proposed in Leonardi and Bühlmann (2016).

2.3 Maximin aggregation: Magging

Besides the nonparametric segmentation which is interesting in its own right, we have access to the output statistics of the Random Forests estimates, namely the functions $\hat{f}(g; \cdot)$ as well as the importance measures $\widehat{MDA}(g; X^{(j)})$ (see formula (6) in the the paper by BS).

We can aggregate the estimates of the different segments to a single estimated regression function or a single variable importance measure. Instead of mean aggregation (Breiman, 1996), it is perhaps more interesting to ask for some sort of "stability" across all the G groups. This can be achieved by maximin aggregation called "magging" (Meinshausen and Bühlmann, 2015; Bühlmann and Meinshausen, 2016). The idea is to find the convex combination of $\hat{f}(1; \cdot), \ldots, \hat{f}(G; \cdot)$ which optimizes the explained variance in the worst

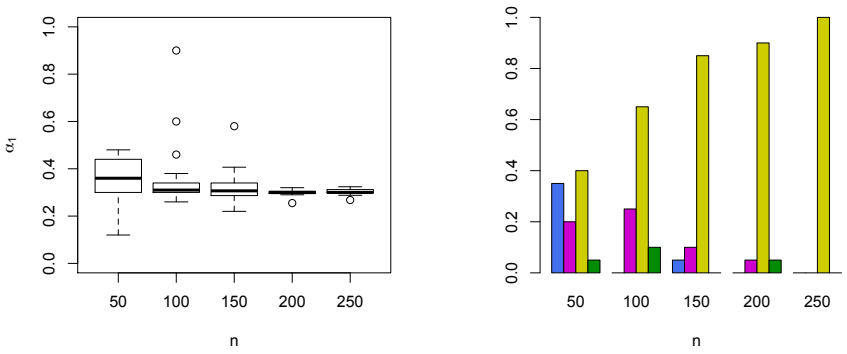

Fig. 1 First estimated change point fraction $\hat{\alpha}_1$ (left panel) and number of groups G (right panel), as a function of sample size n and $p = 2n$. We used $\delta = 0.1$ and $\gamma = \sqrt{\log(p)/n}$

case scenario across the G groups. It can be shown that this corresponds to a convex aggregation whose ℓ_2-norm is minimized:

$$\hat{f}_{\text{magging}} = \sum_{g=1}^{G} \hat{w}_g \hat{f}(g; \cdot), \tag{6}$$

where the convex combination weights can be computed from a quadratic program

$$\hat{w} = \text{argmin}_{w \in C_G} \| \sum_{g=1}^{G} w_g (\hat{f}(g; X_1), \ldots, \hat{f}(g; X_n))^T \|_2^2, \tag{7}$$

with $C_G = \{w \in \mathbb{R}^G;\ w_g \geq 0, \sum_{g=1}^{G} w_g = 1\}$. The aggregated regression estimator with magging, as in (6), is useful for predicting response variables at *new* X-variables: not on average but with some robustness against the worst case. Furthermore, the aggregated variable importance measure with magging

$$\widehat{MDA}_{\text{magging}}(X^{(j)}) = \sum_{g=1}^{G} \hat{w}_g \widehat{MDA}(g; X^{(j)}),$$

with $\hat{w}$ as in (7), for each variable $X^{(j)}$ $(j = 1 \ldots, p)$, is summarizing the variable importance across all G segments. If a variable is important in all the segments, it should be picked up by a large value of $\widehat{MDA}_{\text{magging}}(\cdot)$.

If we do not have access to the groups, we can estimate them as described in (2) and Section 2.1, and then plug-in the estimated version into the magging procedure.

We compute the importance measure for the model in Section 2.2, for a single sample of size $n = 250$ points and with $p = 500$ covariables, with the

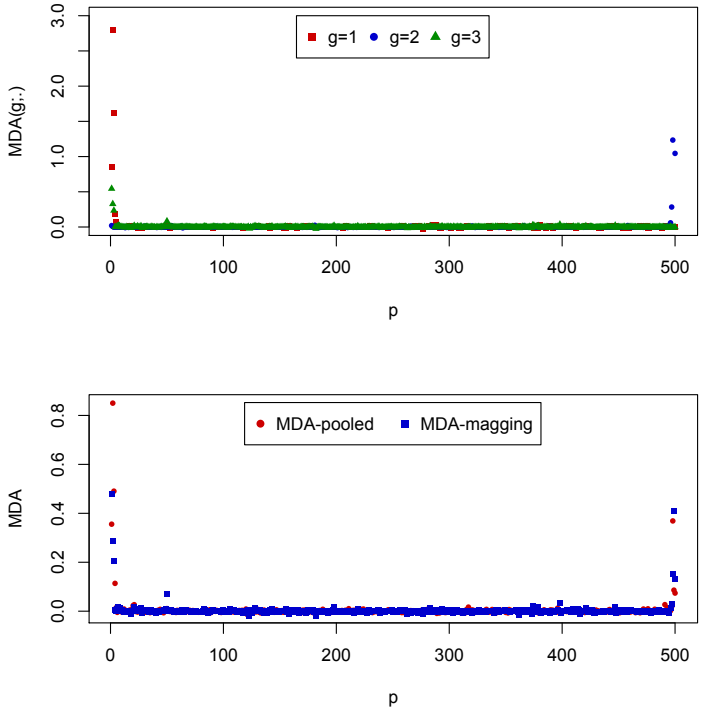

Fig. 2 Top: importance measure $\widehat{MDA}(g;\cdot)$ for each estimated group $g = 1, 2, 3$ of model (5). Bottom: "Pooled" importance measure $\widehat{MDA}$ given by Random Forests on the whole data and aggregated importance measure $\widehat{MDA}_{\text{magging}}$, with weights given by (7). The vector of weights $\hat{w}$ in (7) is $\hat{w} = (0.000, 0.128, 0.872)$. For $\widehat{MDA}$, the top 7 most important variables are 2, 3, 498, 1, 4, 499, 500; and for $\widehat{MDA}_{\text{magging}}$ the top 7 important variables are 1, 499, 2, 3, 498, 500, 50.

estimated change points using the method described above. The values of the $\widehat{MDA}(g;\cdot)$ importance measure for each group and each covariable and the aggregated $\widehat{MDA}_{\text{magging}}$ measure based on magging are shown in Figure 2. The weights $\hat{w}$ computed by (7) are $\hat{w} = (0.000, 0.128, 0.872)$, saying in particular that the magging estimate for optimizing the worst-case performance (across groups) is putting weight zero to the first group with $g = 1$. In contrast, if we would pool all the data without taking the groups, this would correspond approximately to the average weights, each having the value $1/3$, which would provide a rather different solution which does not protect against worst case performance as argued in Meinshausen and Bühlmann (2015) and Bühlmann and Meinshausen (2016). As described in the caption of Figure 2, the true

active variable 50 is found to be important in magging while it does not appear to be "relevant" in the pooled data.

3 Conclusions

Biau and Scornet have provided a very insightful review of the theory and methodology of Random Forests. The theory and methodology is assuming that the data is homogeneous, being i.i.d. realizations from the same distribution or realizations from a stationary stochastic process. We propose here that for heterogeneous data, which is rather rule than exception in large-scale problems, one should segment or group the data first, then use Random Forests (or other flexible and powerful regression or classification methods), and finally aggregate the estimates from each estimated segment or group. The latter step can be done with magging (maximin aggregation) which optimizes the predictive performance in a worst case scenario.

References

Amit Y, Geman D (1997) Shape quantization and recognition with randomized trees. Neural Computation 9:1545–1588

Breiman L (1996) Bagging predictors. Machine Learning 24:123–140

Breiman L (2001) Random Forests. Machine Learning 45:5–32

Bühlmann P, van de Geer S (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Verlag

Bühlmann P, Meinshausen N (2016) Magging: maximin aggregation for inhomogeneous large-scale data. Proceedings of the IEEE 104:126–135

Bühlmann P, Yu B (2002) Analyzing bagging. The Annals of Statistics 30:927–961

Cannings T, Samworth R (2015) Random projection ensemble classification. Preprint arXiv:1504.04595

Fellinghauer B, Bühlmann P, Ryffel M, von Rhein M, Reinhardt J (2013) Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables. Computational Statistics & Data Analysis 64:132–152

Freedman D (1977) A remark on the difference between sampling with and without replacement. Journal of the American Statistical Association 72:681

Hothorn T, Bühlmann P, Dudoit S, Molinaro A, van der Laan M (2006) Survival ensembles. Biostatistics 7:355–373

van der Laan M, Robins J (2003) Unified Methods for Censored Longitudinal Data and Causality. Springer Verlag

Leonardi F, Bühlmann P (2016) Computationally efficient change point detection for high-dimensional regression. Preprint arXiv:1601.03704

Meinshausen N, Bühlmann P (2015) Maximin effects in inhomogeneous large-scale data. Annals of Statistics 43:1801–1830

Mentch L, Hooker G (2015) Ensemble trees and CLTs: Statistical inference for supervised learning. To appear in the Journal of Machine Learning Research

Scornet E (2015) On the asymptotics of Random Forests. To appear in the Journal of Multivariate Analysis

Städler N, Bühlmann P, van de Geer S (2010) ℓ_1-penalization for mixture regression models (with discussion). Test 19:209–285

Stekhoven D, Bühlmann P (2012) Missforest - nonparametric missing value imputation for mixed-type data. Bioinformatics 28:112–118

Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for Random Forests. BMC Bioinformatics 9:307

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58:267–288