Benchmarking random forest: a large scale experiment

Raphael Couronné[1], Philipp Probst[1], Anne-Laure Boulesteix[1]

[1] Department of Medical Informatics, Biometry and Epidemiology, University of Munich (LMU), Marchioninistr. 15, D-81377 Munich, Germany.

Abstract

1 Introduction

In the low dimensional world, logistic regression is considered a standard approach to binary classification. This is especially true in scientific fields such as medicine or psycho-social sciences where the focus is not only on prediction but also on explanation; see Schmuheli (Statistical Science 2010) for a discussion of this distinction. Since their invention 15 years ago, random forests Breiman (2001) have strongly gained in popularity and are increasingly becoming a common "standard tool" used by scientists without any strong background in statistics or machine learning. Our experience as authors, reviewers and readers is that random forest can now be used routinely without the audience strongly questioning this choice. While their use was in the early years limited to innovation-friendly scientists interested (or experts) in machine learning, it has now become commonplace. Random forests are well-known in various non-computational communities.

In this context, we think that the performance of the method should be systematically investigated in a large-scale benchmarking experiment and compared to the current standard: logistic regression. We make the—admittedly somewhat controversial—choice to consider the standard version of RF only, with default parameters, and logistic regression only as the standard approach which is very often considered in a first step for low dimensional binary classification problem. We also investigate the dependence of our conclusions on datasets' characteristics. In particular, as a important by-product of our study, we provide insights into the importance of inclusion criteria for datasets in benchmarking experiments and more generally critically discuss design issues and scientific practice in this context. This paper is structured as follows. After a short overview of LR and RF, as well as the associated VIM and partial dependance plots Friedman (2001), we present the methodology of the benchmark, including the criteria for the dataset's selection. The paper then emphasizes of the analysis of the performance in regards with the dataset's characteristics.

2 Methods

2.1 Logistic regression (LR)

2.1.1 Model

A population is divided into two classes, represented with the vector Y, $Y_i ? 0, 1$. Each observation is described with p features, $X = (X_1, , X_p)$. The logistic regression model estimates the probability of a binary response using the logistic function on a linear function from X. We use the following notations : π_i = $\mathbb{P}(Y = 1|X_i)$, and β the parameter be estimated with maximum likelihood. The regression logistic model is defined with the following formula :

$$logit(\pi_i) = ln\frac{\pi_i}{(1-\pi_i)} = \beta_0 + \sum_{j=1}^{p} \beta_j * X_{i,j} \tag{1}$$

2.2 Random forest (RF)

2.2.1 Brief overview

The random forest is an "ensemble learning" technique which extends the model of decision tree by aggregating a large number of decision trees, and averaging their result, resulting in a lower variance. We present here the random forest first described in Breiman (2001). Each tree of the forest is built using a bootstrap sample of the original dataset, using the CART method and the Decrease Gini Impurtiy (DGI) as the splitting criterion. At each split of the tree, only $mtry$ variables are considered (different subset at each split).

pitfalls see ... ???

2.2.2 Parameters

We present here the most important parameters for the random forest, and their commonly accepted default value.

Fore more details on hyperparameter tuning, see Philip Probst ???

$ntree$: the number of trees in the forest. Theorically, increasing the number of trees always yields more reliable results. Its value should be high enough so that each candidate predictor has enough opportunities to be selected. In practice, it is advised to gradually increase this parameter until the chosen measure of performance stabilizes. Default value is 500 in package $randomForest$.

$mtry$: the number of candidate predictors randomly considered at each split. A low value gives more opportunities to predictors with small effects which may contribute to accurate prediction. A high value reduces the risk of having only non-informative predictors at hand. Default value is $\sqrt{p}$ for classification and $\frac{p}{3}$ for regression.

$nodesize$: minimum size of terminal nodes. Setting this number larger causes smaller trees to be grown. Default value is 1 for classification and 5 for regression.

$replace$: each tree is built on a subsample of the observations. The resampling scheme is also a parameter. In general, the method is a bootstrap sample with or without replacement. This is chosen with the replace parameter. Note that the option without replacement is advised to decrease the bias in favor of predictors with many categories. Default value is $True$.

2.2.3 Variable importance measures

Random Forest can be used to rank internally the importance of the variables, as presented in Breiman (2001). The random forest computes two difference variable importance : the Gini VIM and the permutation VIM. The Gini importance corresponds to the sum of the DGI for all the splits of the forest corresponding to the chosen variable divided by the number of trees, it reflects the importance of the variable in the construction of the tree using DGI. The permutation VIM is based on the accuracy. For a variable, compute the difference of the OOB error before and after permuting randomly the values of the considered variable. An important predictor is expected to reduce more accuracy when neutralized with the permutation. Note that the Gini VIM has bias, for example with the number of candidates for the split, which issue is adressed in Strobl et al. (2007).

VIs are not sufficient to capture the patterns of dependency between predictors and response. They only indicate—in the form of a single number—whether there is such a dependency. Partial dependence plots can be used to address this shortcoming. They can essentially be applied to any prediction method but are particularly useful for black-box methods which (in contrast to, say, generalized linear models) do not yield any interpretable patterns.

2.3 Partial dependence plots

2.3.1 Principle

Partial dependence plots (PDPs) offer an insight of any black box machine learning model, visualizing how each feature influence the prediction by averaging the prediction on all the other features. The PDPs method was developed first introduced by Friedman (2001) for his gradient boosting machine. Let F denote the mathematical function associated with the model, X_t the chosen target and $\overline{X_t}$ the complement subset such that $X_t \subset X$ and $X_t \cup \overline{X_t} = X$. The partial dependance of F on X_t is

$$F_{X_t} = \boldsymbol{E}_{\overline{X_t}} F(X_t, \overline{X_t}) \tag{2}$$

which can be estimated from the data using the empirical distribution

$$F_{X_t} = \frac{1}{N} \sum_{i=1}^{N} F(X_t, \overline{X_{i,t}}). \tag{3}$$

As an illustration, we display in figure 1 the partial dependence plots obtained by logistic regression and random forest for three simulated datasets of size $n = 1000$. The first simulated dataset (first column) is simulated from the logistic model $logit(P(Y = 1)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ (with $\beta_0 = 1$, $\beta_1 = 5$, $\beta_2 = -2$). The second and third datasets are simulated from $logit(\mathbb{P}(Y = 1)) = \beta_0 + \beta_1 x_1 + \beta_2 x_1 x_2$ and $logit(P(Y = 1)) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$, respectively. For all three datasets the random vector $(X_1, X_2)^\top$ follows the distribution $\mathcal{N}(0, 1) \times \mathcal{N}(0, 1)$.

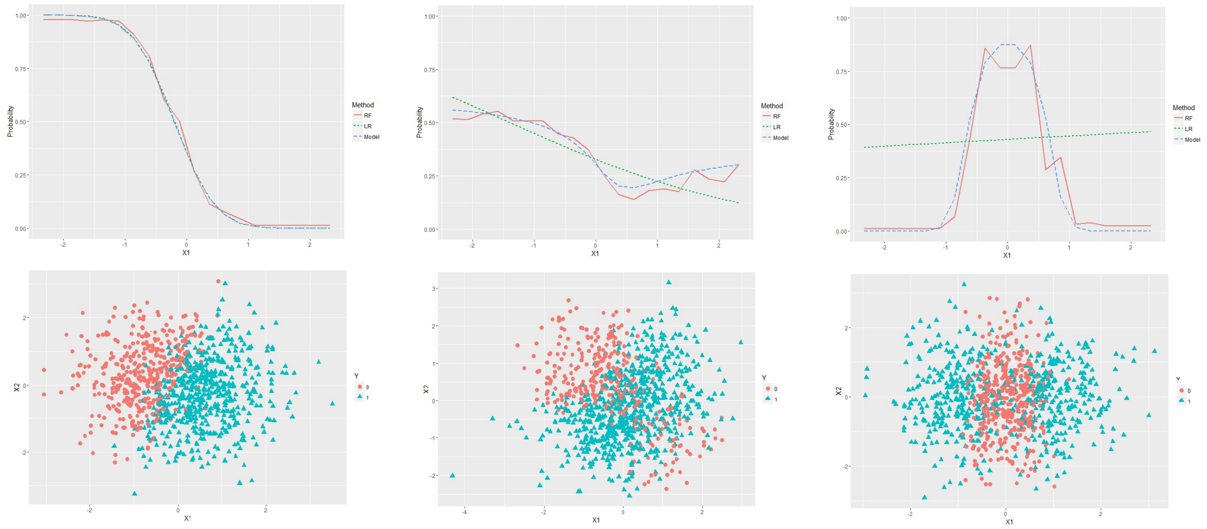

Figure 1: Plot of the PDP

2.4 Benchmarking with real data

In this section we present the design of our benchmarking experiment. Most importantly, the experiment is based on a collection of M real datasets (in contrast to other types of benchmarking experiments relying on simulated data). The prediction accuracy of LR and RF on real datasets is estimated through cross-validation as briefly presented in Section **??**. Issues related to the statistical analysis of the benchmarking results as discussed in Boulesteix et al. (The American Statistician 2015) are reviewed in Section **??**.

2.4.1 Cross-validation

- Brief overview of CV
- At the end we obtain results in the form of a $M \times 2$ data matrix containing the CV errors of LR (first column) and RF (second column) for the J considered datasets.

2.4.2 Accuracy measures

In our study, we consider the following measures quantifying prediction accuracy in the case of a binary classification problem:

- Accuracy :

$$Acc = \sum_{i=1}^{N} \sum_{j=1}^{c} f(i,j)C(i,j). \tag{4}$$

- AUC :
- ...

3 Test Bibliography

Je cite Durand and Durand (2007)

References

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Durand, P., Durand, R., jan 2007. Les tomates tueuses. Le beau journal, 24.

Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics 8 (1), 1.