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 (add reference) 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...

2 Methods

2.1 Logistic regression (LR)

2.1.1 Model

Standard logistic regression...

2.1.2 L_1-penalized logistic regression

Do we want to include it?

2.1.3 L_2-penalized logistic regression

Do we want to include it?

2.2 Random forest (RF)

2.2.1 Brief overview

2.2.2 Variable importance measures

- Short introduction into permutation-based VI
- Transition: 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

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

2.3.2 Measuring differences between partial dependence plots

- In the context of the comparison between LR and RF as potential standard classification tools, we are interested in the differences between the patterns of dependency

underlying the prediction rules yielded by the two methods.

- In this paper, we suggest to investigate this question by comparing the partial dependence plots for the two methods and by quantifying the difference between them through the use of discrepancy measures based on the L_1-norm or L_2-norm of the difference between the two plots.

- More precisely, we define the L_1- and L_2-based criteria as

$$formula(VIweights, etc)$$

- As can be seen from Figure XXX, the partial dependence plots of LR and RF are similar for the first dataset ($L_1 = XXX$ and $L_2 = XXX$). This was expected since the logistic regression model is the true model underlying the data. The similarity observed between the partial dependence plots of RF and LR in this setting indicates that RF can successfully recover this model structure in this case.

- In contrast, the plots are clearly different for the second and the third simulated datasets (middle and bottom rows).

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 J 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 2.4.1. Issues related to the statistical analysis of the benchmarking results as discussed in Boulesteix et al. (The American Statistician 2015) are reviewed in Section 2.4.3.

2.4.1 Cross-validation

- Brief overview of CV

- At the end we obtain results in the form of a $J \times 2$ 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:

- error rate...
- area under the curve...
- ...

2.4.3 Statistical analysis

summary of Boulesteix et al. (The American Statistician 2015)

2.4.4 The OpenML database

So far we have said that the benchmarking experiment used a collection of J real datasets without specifying which ones. In practice, one often uses already formatted datasets from public databases for this purpose. Many of them offer a user-friendly interface and a good documentation which facilitate to some extent the preliminary steps of the benchmarking experiment (search for datasets, data download, preprocessing). One of the most well-known such databases is UCI repository (add references). Specific scientific areas may have their own databases, such as ArrayExpress and GEO for molecular data from high-throughput experiments (add references). Most recently, the OpenML database (add reference) has been initiated as an exchange platform allowing machine learning scientists to share their data and results. This database includes as many as XXX datasets as of September 2016, a non-negligible proportion of which are relevant as example datasets for benchmarking classification methods.

2.4.5 Inclusion criteria

When using a huge database of datasets, it becomes obvious that one has to define criteria for inclusion in the benchmarking experiment. Inclusion criteria in this context do not have any long tradition in computational science. The criteria used by researchers to select datasets are most often completely non-transparent. It is often the fact that they select a number of datasets which were found to somehow fit the scope of the investigated methods, but without clear definition of this scope.

We conjecture that datasets are occasionally removed from the experiment *a posteriori* because the results do not meet the expectations/hopes of the researcher. While the vast majority of researchers certainly do not cheat consciously, such practices may substantially bias the conclusion of a benchmarking experiment; see for instance Yousefi et al. (Bioinformatics 2010) for theoretical and empirical investigation of this problem. In a word, "fishing for datasets" should be prohibited (see Rule 4 of Boulesteix PLOS Computational Biology 2015).

Even if fishing for datasets is prohibited, it is important that criteria for inclusion in the benchmarking experiment are clearly stated; see Boulesteix, Wilson and Hapfelmeier (technical report 2016, coming soon) for an extensive discussion of this issue.

In our study, we consider the following datasets' characteristics to define inclusion criteria:

- n: ...
- ...
- ...

Based on these datasets' characteristics, we define several sets of inclusion criteria and investigate the impact of this choice on the results of the benchmarking experiment. In the same vein, one can also analyse the results of benchmarking experiments for different subsets of datasets successively, following the principle of subgroup analyses performed

in clinical trials. For example, one could analyse the results for "large" datasets ($n > 1000$) and "small datasets" ($n \leq 1000$) separately.

2.4.6 Meta-learning

Going one step further, one can try to model the difference between the methods' performances based on the datasets' characteristics. Such a modelling approach can be seen as a simple form of *meta-learning*—a well-known task in machine learning.

3 Results

3.1 Datasets

- basic inclusion criteria
- flow-chart displaying the number of datasets excluded

3.2 Overall results

Boxplots of the performance of RF and LR for the basic criteria.

3.3 Explaining differences: datasets' characteristics

While it is obvious to any computational scientist that the performance of methods may depend on some datasets' characteristics, this issue is not easy to investigate in real data settings because i) it requires a large number of datasets—a condition that is often not fulfilled in practice; ii) this problem is enhanced by the correlations between characteristics. In our benchmarking experiment, however, we consider such a huge number of datasets that an investigation of the relationship between methods' performances and datasets' characteristic becomes possible to some extent.

- Results with varying inclusion criteria

- Subgroup analyses as additional file?
- Meta-learning

3.4 Explaining differences: partial dependence plots

- In the previous section we have investigated the impact of datasets' characteristics on the results of benchmarking and simply modeled the difference between methods' performance based on these characteristics.
- In this section, we take a different approach to the explanation of differences. We use partial dependence plots as a technique to assess the dependency pattern between response and predictors underlying the prediction rule. When the methods' performances are different, we intuitively expect these dependency patterns to be different, sampling variations put aside. We typically expect this to be the case when the true joint distribution of response and predictors is far from the logistic regression model or when the logistic regression model holds but RF fails to recover it (due to, e.g., too small sample size).
-

4 Discussion

- Summarizing results...
- Bias of standard random forest (literature by Strobl et al., Boulesteix et al.)
- In this paper we investigated only the basic version of RF as implemented in the package randomForest, with default parameter values. The reason for this choice lies in the fact that we wanted to investigate the ability of standard RF as available to any naive user to compete with the standard logistic regression approach—which

can be used without needing strong statistical/technical expertise. The random forest approach, however, has the potential to yield better accuracy than suggested by the standard version with default values. Smart tuning procedures may help to identify optimal values for the various parameters defining the tree and forest structure. In order to compete as a potential "standard tool", however, the whole RF method—including tuning—should run in a completely automatized manner. Non-automatic steps may be extremely useful in practice to achieve optimal performance in specific applications, but they disqualify the method as a standard method to be used for naive users; see Duin (1996) for a discussion of the difference between automatic and non-automatic procedures.

- An important problem related to the many possible variants of RF is that expert users may be tempted to try several variants successively and select the result that better fits their expectations/hopes—a form of fishing for significance. In this context, we thus insist that RF should be either used with the default parameter settings or a correct tuning procedure should be applied. By correct procedure, we mean that tuning is performed internally, i.e. the parameter settings are not chosen based on the final accuracy results. blabla (add references).

- Last but not least, a requirement that RF currently miss to fulfill in practice is *transportability* in the sense that the constructed prediction rule should be easily applicable (or in other words, *transportable*) to another dataset by another scientist. Provided the fitted coefficients $\hat{\beta}_0, \hat{\beta}_1, ;\hat{\beta}_p$ are stated somewhere (e.g., in a table in the paper describing the study), prediction rules constructed by fitting a logistic regression model are easily applicable by computing $\hat{P}(Y = 1)$ as $\hat{P}(Y = 1) = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p)/(1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p))$. For RF things are not so easy. One typically has to make software objects or data& code available to the potential users of the prediction rule. Furthermore, problems may occur due to software incompabilities. This topic is extensively discussed in

Boulesteix et al. (2016, transportability paper) including a survey of recent articles presenting prediction rules constructed by RF. While the situation may improve in the future through the development of interface tools and universal software-independent languages, for the moment one should currently keep in mind that making an RF prediction rule sustainably applicable to a wide non-expert audience may require time, efforts, pedagogical skills and organisation, while this is not the case for logistic regression.

- Conclusion: Standard random forest with default values performs well. But a number of practical and methodological issues have to be addressed.

5 Test Bibliography

Je cite Durand and Durand (2007)

References

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