Regularisation in Random Forests

Bruna Wundervald*[1], Katarina Domijan[1] and Andrew Parnell[1]

[1]Hamilton Institute, Maynooth University, Ireland
*Email: brunadaviesw@gmail.com

Abstract: Shrinkage is still a hard task to perform in tree-based methods. In this work, we evaluated and extended the current methods. Our extension has demonstrated to be better than what was proposed in Deng and Runger (2012), since we ended up with fewer variables but still a good performance. Nevertheless, we see that the method still has a lot of room for improvement, which defines our next steps.

Introduction

In real life problems, predictors can be hard or even economically expensive to obtain. Shrinkage methods, also known as regularisation, can make the coefficients regressions coefficients in a model to be close or exactly equal to zero (Hastie, Tibshirani, and Friedman (2009)), leading to variable selection. For tree-based methods, that do not have regression coefficients to be shrunken, there is not yet a standard regularisation procedure well established in the literature. The main goals of this work are to understand and explore regularisation approaches for trees and random forests, such as the one proposed in Deng and Runger (2012). We describe the test the methods and propose an extension to it.

Tree-based methods and shrinkage

Consider a variable of interest $Y_i \in \mathbb{R}$ and $\mathbf{x} = (x_{i1}, \ldots, x_{ip})'$, $1 \leq i \leq n$. A statistical framework for non-parametric regression characterizes their relationship as

$$y_i = f_0(\mathbf{x_i}) + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \tag{1}$$

where f_0 is an unknown regression function. A tree is a flexible non-parametric method, based on the estimation of a series of binary conditional splitting statements on the predictors space, that creates rules with the form: $x_j > x_{j,th}$, where x_j is the value of the feature at j and $x_{j,th}$ is the decision cut point. The final model predicts Y with a constant c_m in each splitted region R_m. Random forests (Breiman (2001)) are an extension of tree-based methods, combining B trees that are grown on bootstrapped samples, but for each tree only a sample $m \approx \sqrt{p}$ of the predictors is considered for the splitting nodes. The prediction is an average of all of the results in the training set.

An option for shrinkage in tree-based methods is presented in Deng and Runger (2012), where the variable importance is penalized for each tree when building a random forest. The authors also introduced the Guided Regularized Random Forests (GRRF), that leverages the importance scores calculated from a standard random forest model based on all the training data. In this case, the penalization coefficient depends on the previously obtained importance measures, that is

$$Gain_R(X_i, v) = \begin{cases} \lambda_i Gain(X_i, v), i \notin F \text{ and} \\ Gain(X_i, v), i \in F, \end{cases} \tag{2}$$

where $\lambda_i \in (0, 1]$ is the coefficient for $X_i (i \in \{1, ...P\})$.

In order to empirically verify the functionalities proposed in (Deng and Runger (2012)), we tested the methods in a real gene dataset with 48910 columns and proposed an extension for it. First, we used the real gene expression data to calculate the marginal correlations of all of the predictors and the response. With

that, we selected the ones with a minimum of correlation to the response in the train set, for which the cut point was determined as, for $\mathbf{x} = (x_{i1}, \ldots, x_{i48910})'$,

$$|corr(\mathbf{x_j}, \mathbf{y})| > 0.16, \text{ for } j = 1, \ldots, 48910. \tag{3}$$

This prior selection strategy produced a subset of 2626 variables. With this new smaller dataset, we run the GRRF model 100 times, using $\lambda = 0.8$ and $\gamma = 0.9$. We add a new weighting technique, proposed as

$$\lambda_i = \begin{cases} (1-\gamma)\lambda_0 + \gamma\, Imp_i' \tau, \; |corr(\mathbf{x_i}, \mathbf{y})| \leq 0.5 \\ (1-\gamma)\lambda_0 + \gamma\, Imp_i' \, |corr(\mathbf{x_i}, \mathbf{y})|, \; |corr(\mathbf{x_i}, \mathbf{y})| > 0.5 \end{cases} \tag{4}$$

where $|corr(\mathbf{x_i}, \mathbf{y})| \in [0, 1]$ and the new parameter $\tau \in [0, 1]$, where τ it's not expected to be bigger than 0.5 for regularisation. We compared the GRRF model, our proposed extension and a model that uses the absolute values of the marginal correlations as the weighting, as a benchmark. Results are in Figure 1.

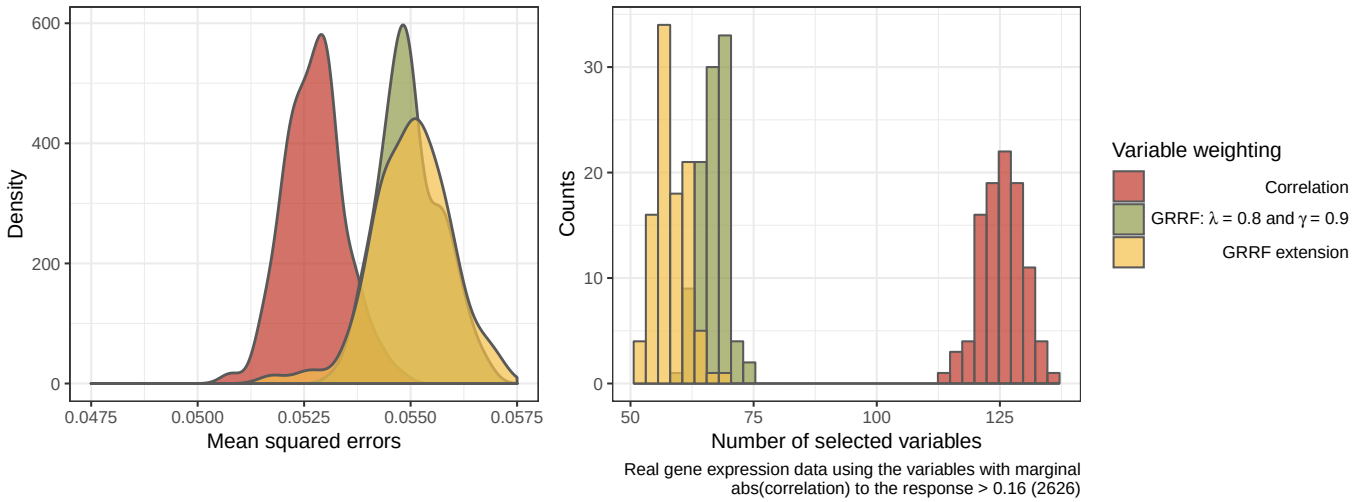

Figure 1: Comparison of the MSR (evaluated in the test set) densities and histograms of the number of selected variables for 100 re-runs of the three models: (i) in green, the standard GRRF; (ii) in red, using simply the correlation as the variable weighting; (iii) using the methodology proposed in Equation 4.

Conclusions

Shrinkage is still a hard task to perform in tree-based methods. As the current methods do not seem to shrink enough, an extension of Deng and Runger (2012) was proposed. In a 100 re-run of the model, our extension presented better results, by keeping the same performance in the test set but also having smaller numbers of selected variables, making it a promising approach. Next steps of the project include comparing the regularisation methods with variable selection, increasing the model reruns, adjusting our proposed extension and improving the validation techniques.

References

Breiman, Leo. 2001. "Random Forests." *Machine Learning*. https://doi.org/10.1017/CBO9781107415324.004.

Deng, Houtao, and George C. Runger. 2012. "Gene Selection with Guided Regularized Random Forest." *CoRR* abs/1209.6425. http://arxiv.org/abs/1209.6425.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. "The Elements of Statistical Learning." *Elements* 1: 337–87. https://doi.org/10.1007/b94608.