Variable Selection using Intersection and Average of Random Forests

Faraz Niyaghi[1], Sharmodeep Bhattacharyya[1], Sarah C. Emerson[1]
[1]Department of Statistics, Oregon State University, 239 Weniger Hall, Corvallis, OR 97331

Abstract
Random forest (RF) has demonstrated the ability to select important variables and model complex data. However, due to the random sampling of data points and variables within RF algorithm, rankings of the selected variables can alter among fitted models to the same data set. This can result in selecting a noise variable over a main variable. This research investigates intersection and average methods to stabilize RF's variable selection. First, multiple RF models are fitted to the data, and ranking of variables and their relative importance are evaluated for each model. Average method ranks the variables based on their mean relative importance. Intersection method iteratively selects variables that are in common among top-ranked variables of these models. These methods also showed potential in detecting main effects in interaction terms.

Key Words: Variable Selection, Random Forest, Intersection of Forests, Stability

1. Introduction

Variable selection is a statistical procedure with applications in many problems, and its importance has grown in tandem with the increasing size of datasets to which regression or classification analysis is applied. One of its applications is in genomics where the number of features (genes) is much larger than the number of observations in a sample. For example, Guyon et al. investigated the performance of support vector machines in identifying the important variables which can be used in cancer diagnosis (2002). There are several other techniques for variable selection including lasso, principal component analysis, and RF (Genuer et al., 2010; Tibshirani, 1996; Wold et al., 1987). We focus on variable selection using RF in this paper.

RF is one of the most widely used machine learning methods which was introduced by Brieman (2001). It can be used for both classification and regression problems. To fully understand the underlying algorithm of RF, we need to introduce classification and regression trees (CART) which was developed in the late nineties (Breiman et al., 1984). CART consist of nodes which recursively partition the feature space such that similar dependent variables fall into same regions. CART are simple estimators with low computation cost but they usually have a high variance. RF has the same bias as CART, however, it significantly reduces the variance by fitting several trees and averaging their predicted values. In order to achieve a small variance, RF generates uncorrelated trees in two steps: using a bootstrap sample from the data in each tree and considering only a random sample of features at each node of the trees.

Variable importance (VI) is one of the outputs of RF which can be used for variable selection and ranking of variables. There are multiple ways of calculating VI including mean squared error and Gini index. The choice of importance criterion depends on the nature of the response variable (e.g. categorical or numerical). Regardless of importance criterion, VI of features changes with every run of RF due to the bootstrapping of data points and random selection of variables within RF algorithm. This variability is more evident when a small number of trees are used. This can result in inconsistent variable selection and picking a noise variable over main variables in some cases. Some studies have compared the performance and stability of different importance criteria (Calle and Urrea, 2010; Nicodemus, 2011) but there is room for further research on this topic. In this paper, we propose two methods to stabilize RF's variable selection.

2. Average and Intersection Methods

Average and Intersection methods are described in this section. Both methods start with splitting data into same-sized groups based on response value rankings. Then, within each response group, they form clusters using features only. K-means and Silhouette score are used to make clusters and to determine the best number of clusters to use, respectively (Rousseeuw, 1987). Next step is to fit separate RFs to each feature-based cluster and obtain relative VI from each model. Note that all of these steps are in common in average and intersection methods. The idea behind these steps is to make it easier to see the effect of each feature on the response variable. Especially, in cases where an independent variable has a significantly larger impact on the response variable, and as a result, masks the effect of other variables. Average method ranks the variables based on their mean VI obtained from RFs fitted in the last step. The average method steps are summarized in Algorithm 1.

```
Algorithm 1: Average Method
Input: Number of response bins M; Maximum number of feature-based clusters in each response bin K; Number of features to output N.
Output: Top N important features.

    1. Split data into M same-sized bins based on response value rankings.
    2. Within each response bin, use only features to make clusters and find optimal number of clusters in range 2 to K using Silhouette score.
    3. Fit separate RFs to clusters in step 2 and obtain relative VI from each model.
    4. Calculate average VI for each feature.
    5. Rank features based on their mean VI.
    6. Output top N features.
```

Intersection method uses the ranking of variables based on their VI from the fitted RFs. It iteratively selects variables that are in common among top-ranked variables of these models. The steps of this method are shown in Algorithm 2. Result section compares the performance of these methods with regular RF for two sets of simulated data.

Algorithm 2: Intersection Method
Input: Number of response bins M; Maximum number of feature-based clusters in each response bin K; Number of features to output N.
Output: Top N important features.

1. Split data into M same-sized bins based on response value rankings.
2. Within each response bin, use only features to make clusters and find optimal number of clusters in range 2 to K using Silhouette score.
3. Fit separate RFs to clusters in step 2 and obtain ranking of features based on their relative VI for each model.
4. Iteratively, select variables that are in common among the top-ranked variables in step 3.
5. In case of a tie, use mean VI to break the tie.
6. Output top N features.

3. Results

3.1 Highly Non-linear Data

A sample of size 100000 is simulated using the formula below:

$$Response = X + Y + Z^2 + W + Q + sin(XY^2) + e^{ZWQ^2} + \varepsilon$$

Where features and error term follow a standard normal distribution. In addition to X, Y, Z, W, and Q, 50 independent noise variables with standard normal distribution are added to the dataset. Regular RF, average method, and intersection method are compared in terms of their ability to select main variables.

Data are split into 5 same-sized bins based on response value rankings. Next, Silhouette score is used to choose the number of feature-based clusters in each response bin. In total, 19 clusters are formed and separate RF with 100 trees are fitted to them. VI obtained from these RFs is used by intersection and average methods for variable selection as shown in Table 1. Regarding regular RF, for the sake of fairness, a RF with 1900 trees is fitted to the whole data. The results from these methods are shown in table 1.

Table 1: Top 9 variables selected by each method in simulation 1

Regular RF	Average Method	Intersection Method
Q	Q	X
W	X	Y
Noise	Y	W
Noise	W	Z
Z	Z	Q
Noise	Noise	Noise
Noise	Noise	Noise
Noise	Noise	Noise
Noise	Noise	Noise

3.2 Low Signal-to-Noise Ratio Data

10000 data points are generated using the formula below:

$$Response = XY + ZWQ + 10\varepsilon$$

Where features and error term follow a standard normal distribution. Similar to simulation 1, 50 independent noise variables with standard normal distribution are added to the dataset. The exact same steps as in simulation 1 are taken in this simulation. The only difference is the optimal number of feature-based clusters. In this simulation, a total 24 clusters are formed. Thus, a RF with 2400 trees is fitted to the whole data to represent regular RF. The results from all 3 methods are shown in table 2.

Table 2: Top 9 variables selected by each method in simulation 2

Regular RF	Average Method	Intersection Method
X	X	X
Noise	Y	Y
Noise	Z	W
Noise	W	Z
Noise	Q	Q
Noise	Noise	Noise
Z	Noise	Noise
Noise	Noise	Noise
Noise	Noise	Noise

4. Discussion

RF’s variable selection is considered unstable due to the randomness within its algorithm. As shown in our simulations, this results in selecting redundant variables over main variables in some cases. In this paper, average and intersection methods are proposed to stabilize RF’s variable selection. The variable selection performance of these methods is compared to regular RF in two simulations.

In simulation 1, results show the average and intersection methods perform better in variable selection. Regular RF fails to pick X and Y in its top 9 selected variables. It is possibly due to masking effect of the exponential term in the simulation function. Presence of Q, W, and Z in the exponential term overpowers the effect of other variables. It is hypothesized that the success of average and intersection methods, in this case, is a result of the clustering step in their algorithm. Some clusters are probably formed in regions where values of Q, W, and Z are small, and this makes it possible to isolate and detect the effect of X and Y in these regions.

Simulation 2 is meant to illustrate how the 3 methods perform when applied to low signal-to-noise ratio data. Average and intersections method are found to be superior to the regular RF in this simulation. These results also show the potential of our proposed methods in detecting features in interaction terms. It is notable that our results are only generalizable to simulation settings included in this paper. Further research is required before inferring our findings to a broader population of data structures.

References

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

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and regression trees. CRC press.

Calle, M.L., Urrea, V., 2010. Letter to the editor: Stability of random forest importance measures. Briefings in bioinformatics 12, 86–89.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern recognition letters 31, 2225–2236.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine learning 46, 389–422.

Nicodemus, K.K., 2011. Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. Briefings in bioinformatics 12, 369–373.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20, 53–65.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58, 267–288.

Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometrics and intelligent laboratory systems 2, 37–52.