A Forest Measure of Variable Importance Resistant to Correlations

Padraic G. Neville* Pei-Yi Tan†

Abstract

Variable importance estimates that are output from decision trees and random forests are often used to reduce the dimension of data, especially in the presence of many variables, because decision trees can process many variables quickly. However, trees typically inflate the importance of correlated variables and even promote irrelevant correlated variables above predictive independent variables. Strobl et al. (2008) analyze the cause and propose a remedy. Unfortunately, the remedy is too complex to be practical for a large number of observations. This paper presents a simple method, called *random branch assignments*, which conforms to the analysis of Strobl et al. and yet can handle many observations. Although the method still incorrectly ranks the variables when the signal-to-noise ratio is less than 1, it is dramatically less sensitive to correlation effects than the measures of variable importance in the randomForest() function in R.

Key Words: Decision Tree, Forest, Variable Importance

1. Introduction

Decision trees and random forests are often used for reducing the number of variables in a data set. Trees can handle many variables quickly and often find variables that have interaction effects. However, trees can inflate the importance of correlated variables and even promote irrelevant variables to a higher importance than variables that are predictive but independent. This problem exists both for the classical loss reduction (decrease in impurity) measure of importance introduced by Breiman et al. (1984) and for Breiman's (2001) permutation method. Strobl et al. (2008) diagnose the problem, formulate principles that are required of a solution, and propose a remedy. Their method works well on a thousand observations, but is too complex to handle hundreds of thousands of observations in an acceptable amount of time. This paper proposes a simple method that conforms to the principles set out in Strobl and that can process many observations quickly. When a variable Z is evaluated, a splitting rule that involves Z is replaced by a rule that randomly assigns an observation to a branch whose probability is proportional to the size of the branch. The method is called *random branch assignments* (RBA). In simulation studies RBA is more resistant to correlation effects than other measures.

The next section defines the various measures of importance, presents Strobl's explanation of the problem, and argues that RBA satisfies Strobl's requirements of a solution. Simulations are presented in Section 3. RBA is shown to be resistant to correlations and stable over a wide range of signal-to-noise ratios in the data.

2. Variable Importance

2.1 Loss reduction

Variable importance for decision trees originates from Breiman et al. (1984) and implemented in the CART software program. It has been called Gini increase, Gini importance,

*SAS, 28 Roselawn Ave, Fairport, NY 14450

†SAS, 1205 Williams Road, New Smyrna Beach, FL 32168

and impurity reduction. This papter calls it loss reduction to emphasize its relationship to the reduction of error from using the model.

A measure of impurity is computed in each node. For an interval response, a common measure of impurity is the sum of square residuals. If variable Z is used to split the node, then the reduction in impurity from the parent node to the child nodes is credited to Z. If the impurity is 600 in a parent node and 100 and 200 in the two child nodes respectively, then the reduction in impurity is 300. The variable importance for variable Z is proportional to the sum of the reduction in impurity over all nodes in the forest that split on Z.

2.2 Breiman's method

Breiman (2001) introduces a different measure of variable importance that uses random forests. A random forest consists of an ensemble of decision trees. Each tree is trained from a different sample of the data. The observations that are withheld from training a tree are called out-of-bag (OOB). In every node in the tree, a random subset of the variables compete to form the splitting rule. Randomization makes the trees different.

To compute the importance of Z, first permute the values of Z in the OOB sample of each tree, and then compute the OOB predictions of each observation. The OOB prediction of an observation is the average prediction from trees for which the observation is OOB. Use the predictions to compute a goodness-of-fit measure, such as mean square error (MSE). The importance of Z is proportional to the fit that is based on the permuted data minus the fit that is based on the data without permutations.

2.3 Conditional and marginal importance

To illustrate the difference between loss reduction and Breiman's method, this paper uses a simple data set that is constructed as follows. Generate X and Z from a bivariate normal distribution with zero means, unit variances, and correlation ρ. Let the response Y equal X. Given X, Y is known exactly, and Z can provide no additional information about Y. Z is said to have zero *conditional* importance. On the other hand, if X is not known, then Z provides some information about Y by virtue of the correlation, and Z is said to have some *marginal* importance. Different measures of importance can be placed on a spectrum from conditional to marginal importance (Grömping 2009).

For several values of ρ, generate 25 data sets of 500 observations each. For each data set, apply a random forest and plot the proportion of total importance that is credited to Z. Figure 1 shows the proportions plotted against of 11 values of ρ. The upper curve is based on loss reduction, and the curve below it is based on Breiman's method. Breiman's method is seen to be a more conditional measure of importance. This is desirable for dimension reduction, in which the goal is to retain few variables to cover the most information about the response. Note a another curve is below the Breiman curve. This curve is due to Strobl et al. and is more towards the conditional end of the spectrum than Breiman's.

2.4 Strobl's method

Strobl et al. obtain a more conditional measure of importance by analyzing the logic of Breiman's method and revealing an omission. The analysis centers around statistical independence. If Y and Z are independent, then for sets B and C,

$$P((Y \in B) \cap (Z \in C)) = P(Y \in B)P(Z \in C) \tag{1}$$

A bivariate histogram of the data would be approximately equal to the product of univariate histograms. If the values of Z are permuted, the univariate histogram of those values

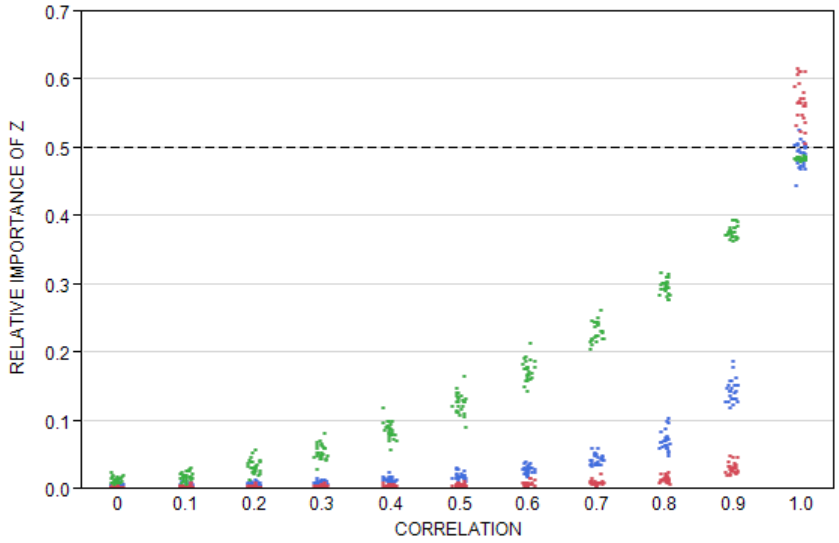

Figure 1: Proportion of total importance credited to Z for various correlation values ρ. The measures of importance are loss reduction (top), Breiman's method (middle), and Strobl's method (bottom).

would remain unchanged and so would the bivariate histogram. On the other hand, if permuting the values of Z does significantly change the bivariate histogram, then Y has some dependence on Z.

Strobl et al. say that this is the essence of Breiman's logic and that this logic ignores the influence of other variables. To incorporate the other variables, the reasoning should proceed from conditional independence:

$$P((Y \in B) \cap (Z \in C) \mid X \in A) = P(Y \in B \mid X \in A)P(Z \in C \mid X \in A) \quad (2)$$

Probabilities that involve Y and Z are computed separately for separate values of X. When the values of Z are permuted, only values from observations that have the same value of X are permuted. Strobl et al. develop an algorithm around conditional independence. The resulting variable importance measure is more towards the conditional end of the spectrum than Breiman's.

The method of Strobl et al. works well for data that contain a few hundred observations, but becomes prohibitively slow for hundreds of thousands of observations. Let V denote the number of variables, and let T denote the number trees. Then Breiman's method requires approximately $VT/3$ permutations. Strobl's method multiplies that number by some proportion of N.

2.5 Random branch assignments

This paper presents a simple algorithm, called random branch assignments (RBA), that satisfies the logic of Strobl et al. and that avoids all the permutations. When the trees are created, the number of observations in each node is saved. If preferred, the number of observations from a separate hold-out data set can be used. The argument assumes that the observation sizes that are saved in each node are proportional to the number of observations tat visit the node from the data set being evaluated. To compute the importance of a variable *Z*, randomize the branch assignment rules that involve *Z* and then apply the randomized model to the data and compute a goodness-of-fit measure. The randomized rule is one that

randomly assigns an observation to a branch with probability proportional to the number of observations in the branch. For example, suppose a node that contains 100 training observations is split by values of *Z* into two nodes: one contains 25 training observations and the other contains 75. When the importance of *Z* is evaluated, an observation that reaches the node is randomly assigned to the smaller branch with probability 0.25. As in the Breiman method, the importance of *Z* is proportional to the randomized fit minus the fit without randomization.

This paper claims that RBA satisfies the objectives of the methods of Breiman and of Strobl et al. First note that the purpose of permuting the values is to break any relationship between the response variable and Z without changing the univariate distribution of Z. The same thing can be accomplished by replacing an observation value with a value that is randomly chosen from the univariate distribution. Next note that the branch assignment rule lumps together the values that are assigned to the same branch. To assign a branch to an observation, all that is needed is the probability that Z is in a lump, not the probabilities of the individual values. The RBA method uses the lump probabilities conditioned on arriving at the node that is being split, which is exactly the conditional requirement of Strobl et al.

3. Simulations

3.1 Data and notation

Generate $X_i \sim$Normal(0,1), $i =$1, 2, ... 16, with correlation of 0.9 between pairs of X_1 to X_6, and all other pairs independent. The response Y is generated as

$$\begin{aligned} W &= 4X_1 + 4X_2 + 2X_3 + 2X_4 \\ &\quad - 4X_7 - 4X_8 - 2X_9 - 2X_{10} \\ Y &= W + \epsilon \\ \epsilon &\sim \text{Normal}(0, \sigma^2) \\ \omega^2 &= \text{Variance}(W) \\ \psi &= \omega^2/\sigma^2 \end{aligned}$$

ω^2 is the variance of W and equals 90.8. σ varies between data sets. ψ is the signal-to-noise ratio. The way variable importance measures depend on ψ is of interest. Let $\iota_m(X)$ denote the importance of X under measure m, and define the ratio,

$$\lambda_m(a, b; c, d) = \frac{\text{average}(\iota_m(X_a) + \iota_m(X_b))}{\text{average}(\iota_m(X_c) + \iota_m(X_d))} \tag{3}$$

where the average is taken over the 500 data sets in the simulation.

Variables X_1, X_2, X_7, and X_8 are the most important variables. A perfect conditional measure of importance m would have $\iota_m(X_1) = \iota_m(X_2) = \iota_m(X_7) = \iota_m(X_8)$ and $\lambda_m(1, 2; 7, 8) = 1$. However, because the importance of correlated variables are commonly inflated, it is anticipated that $\lambda_m(1, 2; 7, 8) > 1$. Similarly, X_3, X_4, X_9, and X_{10} are equally important but it is anticipated that $\lambda_m(3, 4; 9, 10) > 1$. Variables X_5 and X_6 have no conditional importance, but a measure of importance might elevate X_5 and X_6 above some variables that define W, yielding $\lambda_m(5, 6; 9, 10) > 1$, suggesting that some variables that do not determine Y are more important than some variables that do. This is most undesirable.

3.2 Software and importance measures

The high-performance HPFOREST procedure in SAS® Enterprise Miner™ computes loss reduction and will compute RBA in 2015. The definitive implementation of Breiman's method is in the randomForest() function in R (Liaw and Wiener 2002). This function invokes code that was originally written by Breiman. For an interval response, the randomForest() function outputs two measures of importance: Breiman's method with mean square error (MSE) and a decrease in MSE impurity which in this paper are called ι_Breiman and ι_impurity, respectively. Both are based on the training data.

The HPFOREST procedure outputs several measures of importance. This paper uses two:RBA and loss reduction, both of which are based on the training data MSE. They are denoted by ι_RBA and ι_loss, respectively. ι_impurity and ι_loss are different implementations of the same measure.

Both the HPFOREST procedure and the randomForest() function sample without replacement 0.6 of the training data to train a single tree. Other software specificiations: 100 trees in a forest, at least 10 observations in each leaf, and 8 randomly selected variables in each node to compete for splitting. In randomForest(), a splitting rule is found for each of the 8 variables, and the best split is used in the tree. In PROC HPFOREST, the variables compete on a measure of association described in Hothorn et al. (2006) and implemented in the R party package. A split is made with the winning variable.

3.3 Results comparing measures

Figure 2 shows the variable importance averaged over 500 data sets of 1,000 observations with signal-to-nose ratio $\psi = 182$. The numbers for measure m have been divided by $(\iota_m(X_9) + \iota_m(X_{10})/2$ to align the measure on X_9 and X_{10}.

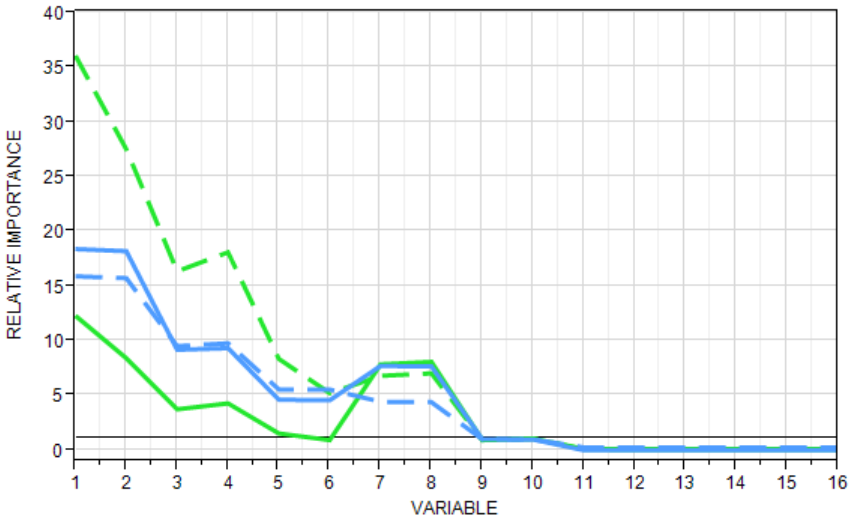

Figure 2: Average variable importance over 500 data sets of 1,000 observations with $\psi = 182$, scaled to 1 for variables X_9 and X_{10}. From top to bottom: ι_loss (green dash), ι_Breiman (blue solid), ι_impurity (blue dash), and ι_RBA (green solid)

The inflation of the importance of the first six variables compared to the second six is apparent. ι_loss (represented by the green dash line) is the most inflated. ι_Breiman (blue solid line) and ι_impurity (blue dash line) are in the middle. ι_RBA (green solid line) shows

Table 1: Ratios $\lambda_m(1,2;7,8)$

ψ	ι_RBA	ι_Breiman	ι_Loss	ι_Impurity
0.5	0.78	96.86	0.78	0.74
1	0.97	34.72	0.85	1.00
2	1.00	10.07	0.97	1.31
4	1.07	4.03	1.34	2.02
8	1.30	2.46	2.12	3.35
182	1.61	2.39	3.58	4.71

the least inflation. Perhaps most disturbing is elevation of variables X_5 and X_6, which have no influence on the response Y, up to and above the variables X_9 and X_{10}, which do.

Figure 3 shows the result from reducing the signal-to-nose ratio ψ from 182 to 8. ι_Breiman (blue solid) is the most inflated here, slightly worse than with $\psi = 182$. Inflation

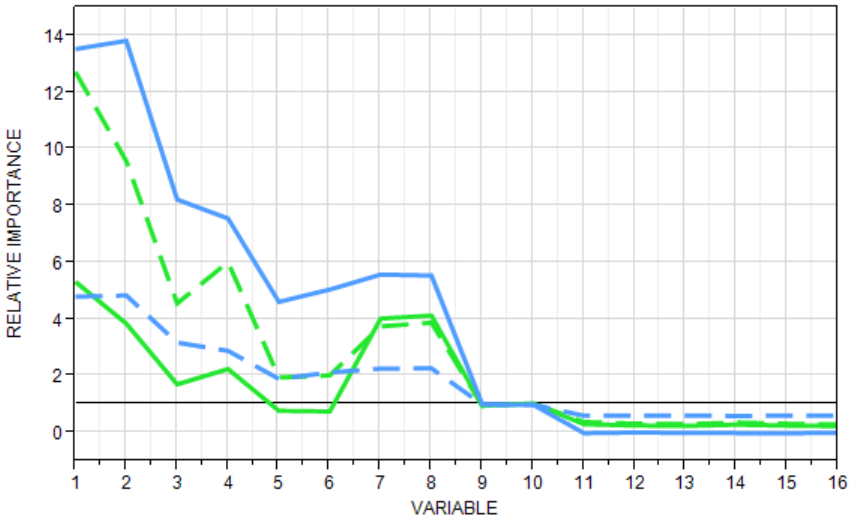

Figure 3: Average variable importance over 500 data sets of 1,000 observations with $\psi = 8$, scaled to 1 for variables X_9 and X_{10}. From top to bottom: ι_Breiman (blue solid), ι_loss (green dash), ι_impurity (blue dash), and ι_RBA (green solid)

from the other measures have decreased slightly. ι_RBA now gives the correct ranks, up to ties. $\lambda_{RBA}(5,6;9,10)$ is about 0.8, which is < 1 as desired.

Table 1 presents $\lambda_m(1,2;7,8)$, the inflation of $\iota_m(X_1)$ and $\iota_m(X_2)$ compared to $\iota_m(X_7)$ and $\iota_m(X_8)$, for different levels of the signal-to-noise ratio ψ. $\lambda_m(1,2;7,8) = 1$ for an ideal conditional measure of importance.

$\iota_m(X_1)$ and $\iota_m(X_2)$ become more inflated with increasing signal for all measures except ι_Breiman. However, ι_Breiman never falls below 2. For $\psi <= 4$, ι_Breiman is wildly large. ι_RBA varies the least, with a minimum of 0.78 and a maximum of 1.61. Except for the single case of ι_impurity at $\psi = 1$, ι_RBA is the closest to 1 at every ψ. This is the main reason for recommending ι_RBA.

3.4 RBA results for varying N and ψ

Now focus on ι_RBA. For 1,000 observations and $\psi < 1$, ι_RBA, ι_loss, and ι_impurity deflate the importance of correlated variables, which is opposite to the more well-known tendency of inflation. Figure 4 shows $\lambda_{RBA}(1, 2; 7, 8)$ for N = 1,000, 2,000, 4,000, 6,000, 8,000, and 10,000. The lines represent the ratios of the average over 500 data sets

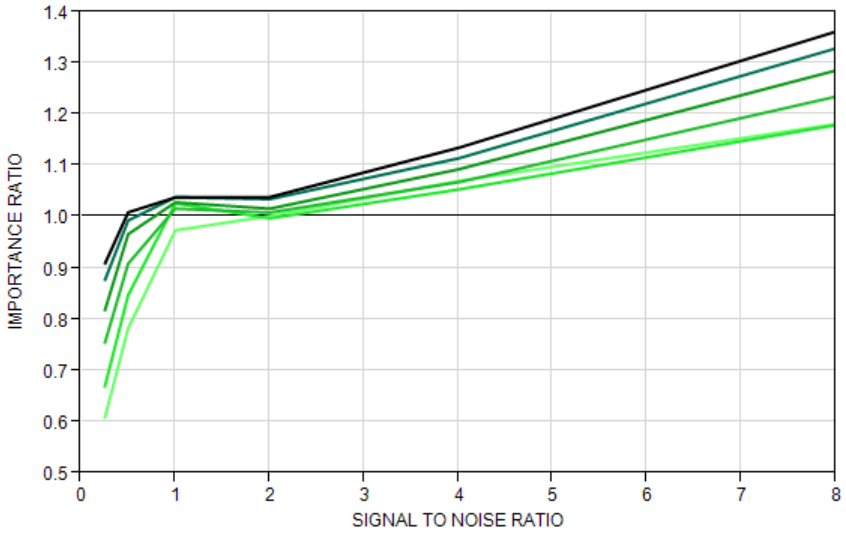

Figure 4: $\lambda_{RBA}(1, 2; 7, 8)$ for N =1 (lightest green line, bottom), 2, 4, 6, 8, and 10 thousand (black line, top).

of $\iota_{RBA}(X_1) + \iota_{RBA}(X_2)$ and $\iota_{RBA}(X_7) + \iota_{RBA}(X_8)$. The light green line at the bottom represents data sets that contain 1,000 observations. The top black line represents 10,000 observations. Generally, $\lambda_{RBA}(1, 2; 7, 8)$ falls below 1 rapidly as ψ decreases from 1, causing X_1 and X_2 to be underrated. The fall is less with more observations. For larger ψ, the opposite prevails: more observations result in more inflation.

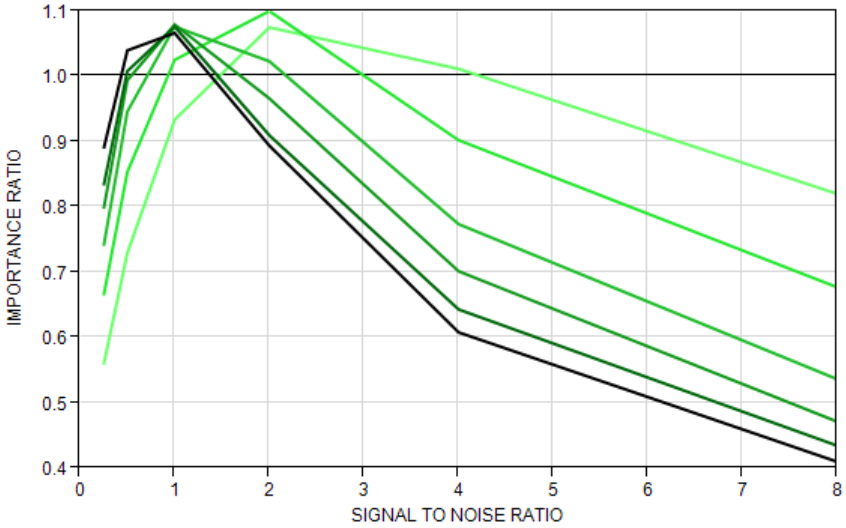

Figure 5: $\lambda_{RBA}(5, 6; 9, 10)$ for N =1 (lightest green line, top right), 2, 4, 6, 8, and 10 thousand (black line, bottom right).

Figure 5 shows $\lambda_{RBA}(5, 6; 9, 10)$ for the same data. Except for ψ near 1 or 2, $\lambda_{RBA}(5, 6; 9, 10) < 1$. A value < 1 indicates that the irrelevant but correlated variables X_5 and X_6 are ranked less than X_9 and X_{10}, which is good. Increasing N decreases $\lambda_{RBA}(5, 6; 9, 10)$ when $\psi > 1$, but not when $\psi < 1$. This is another reason for not trusting variable importance measures when there is more noise than signal in the data.

4. Discussion

After the 2001 article, Breiman wrote that the permutation variable importance method for misclassification was "too volatile" and he dropped it (Breiman 2003). He replaced the misclassification statistic with a function (margin) of the class probabilities. He does not discuss what to do with an interval response, and he might not support the MSE measure in this paper. However, permutation MSE importance is what the randomForest() function outputs today, and some reliable measure of importance is desired. This paper has shown that the permutation MSE measure of importance will often not rank variables consistent with their contribution to the generation of the response variable. That much was known. This paper presents a simple and practical method that does a much better job in the simulation experiments provided the signal-to-noise ratio is not too small.

REFERENCES

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32.

Breiman, L. (2003), "Manual–Setting Up, Using, And Understanding Random Forests V4.0," www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J (1984), *Classification and Regression Trees*, Belmont, California: Wadsworth.

Grömping, U. (2009), "Variable Importance Assessment in Regression: Linear Regression versus Random Forest," *The American Statistician*, 63, 309–319

Hothorn, T., Hornik, K., and Zeileis, A. (2006), "Unbiased Recursive Partitioning: A Conditional Inference Framework," *Journal of Computational and Graphical Statistics*, 15, 651–674.

Liaw, A., and Wiener, M. (2002), "Classification and regression by randomForest," *R News*, 2(3):18–22.

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., and Zeileis, A. (2008), "Conditional Variable Importance for Random Forests," *BMC Bioinformatics*, 9, 307–317.