Jurgen Palsma University of Kent
jurgen.palsma@gmail.com 25/03/2018

Measures of variable importance in Random Forests

If random forests can be particularly accurate by combining multiple decision trees together through an ensembling method, they do not, unfortunately, reflect the comprehensibility that single decision trees offer. Because of this, indicators have been developed to allow these complex models to be a bit more interpretable by quantifying the importance of each attribute (variable) of a dataset used in the classifier. These indicators measure *variable importance*(VI). In the following paragraphs, we will explore and review some of the literature concerning these indicators.

Mean Decrease Accuracy

Historically, Breiman[1] was the first to propose measures of variable importance. His first proposition, which is called *Mean Decrease Accuracy* (MDA) by Louppe et al,[2] or "*Permutation importance"* by others[3], consists of evaluating the importance of a variable in a single tree by applying noise to a variable which is used as a condition in the tree to evaluate (mis)classification this might induce. The motivation for this method is that if an individual is misclassified when one of its attributes has noise applied to it, then the attribute must play a significant role in the classification of the individual, and thus has *high importance*.

To calculate this, imagine that for each X_i attribute of a data set, you select the trees that incorporate a decision on this attribute and average the percent change in accuracy C_i when applying noise to attribute X_i in each tree compared to the accuracy of the tree when the attribute has no noise added to it, by running "*out of bag*"[7] samples through that tree. You can then rank the changes in accuracy C to have an overview of the variables that impact classification the most.

This method is a good start for evaluating variable importance as it is easy to implement but tends to be biased for attributes which have low variation in numerical data or a small number of categories in categorical data (low variation/categories means higher sensitivity to noise) and does not directly take into account variable correlations: if you permute a variable for a tree in which another correlated variable is conditioned upon, you will not see a decrease in accuracy. As a result, correlated variables tend to have a low MDA score even if they play a big part in classification in your model on their own (without the other correlated variables). This also applies to conditional variables: if your classification depends highly on, say, two correlated variables, and you permute one of the variables, you will see a high decrease in accuracy for that variable without necessarily knowing that it is important because of the conditionality of the other variable.

Mean Decrease Impurity

Breiman also proposed with his CART[1] algorithm to "[add] up the *gini decreases*[8] *for each individual variable over all trees in the forest"* as a measure of variable importance. Louppe et al.[2], named this measure "*Mean Decrease Gini"* and proposed to generalise it to all impurity measures such as the gini importance through the term "*Mean Decrease Impurity"* (MDI). As a result, the MDI of a variable X_i is summation of the decrease of impurity for each node where this variable is conditioned upon, for each tree, divided by the number of nodes in all trees where this variable is conditioned upon. Although Strobl et al.[4] deems this method as "biased", in the cases where either the attributes of the data set are all uninformative (called "*null case*" in the paper[4]) or are all equally informative(called "power case"[4]), this method can still be seen as reliable as these cases very rarely happen in real-life scenarios,

and additionally, this technique it is widely used today, relatively easy to compute, and applicable to all impurity measures.
Since then, a lot of other VI measures have been proposed. They can be as simple as counting the number of nodes on which a variable is conditioned upon, which is obviously not very accurate but can be used as a very quick to compute and implement measure, or other more problem-specific, or field specific measures of VI:

Proximity matrix difference

For example, in the field of genetics, Zhou et al.[5] propose the "*proximity matrix difference*"[9] as a measure of VI, which is shown to yield better results for "*large p, small n* problems"[10]. More precisely, this technique allows less significant variables (of which there can be a lot in genetic data) to be more visible in importance measures. The technique is similar to MDA: instead of computing the change of predictive accuracy when permuting a variable, you quantify the misclassification you created by noising up the variable by calculating the change of proximity of your attributes relative to one another. This change of proximity is calculated through what Zhou et al. name the "*proximity ratio"* [5], which is the inner class proximity of samples (how many samples of the same class end in the same leaves of your trees) divided by the outer class proximity of samples (how many samples of different classes end in the same leaves of your trees) for a model.

VI with missing values

Another interesting measure of variable importance proposed by Hapfelmeier et al.[6] , focuses on the importance of missing variables. In their paper, they argue that the classical MDA and MDI are biased when used with data having missing values. For example, in the case of MDA, it is not possible to compute a MDA score on samples that have missing values of the noised up variable. As a result, the authors propose to follow Breiman's idea but instead of adding noise to an attribute, we introduce stochasticity when a split is done on a target variable. The procedure is roughly the same: for each tree and for each attribute X_i, you pass the OOB samples through the tree, but instead of adding noise to the variable, you "randomly assign each observation with [a random probability] to the child nodes of a node *k* that uses [X_i] for its primary[11] split"[6]. Just as with MDA, if the OOB sample used is misclassified when being split randomly according to that variable, or if the mean decrease in accuracy over all trees is high, it must mean that this specific variable plays an important role in classification. Unlike MDA however, this can be applied for samples in which the variable is missing, and can thus cope with datasets having a lot of missing values (which is very handy in real-life situations).

Conclusion

As you can see, even with the few VI measures that we have covered in these paragraph, one can already have a list of measures handy for multiple use cases: depending on your needs, VI can be computed with naive and easy measures, such as summing up nodes in all trees where a variable is used, with more intricate and widely used methods such as MDI and MDA, or even for field-specific use of random forests such as the proximity matrix difference. As a result, one can have a scalable strategy, depending on the need of comprehension of individual variables, and time/effort availability, to restore the comprehensibility of individual decision trees when constructing random forests models.

Annex

References

1.Breiman, Leo. “Machine Learning.” *Machine Learning*, vol. 45, no. 3, 2001, pp. 261–277., doi:10.1023/a:1017934522171.

1.5 “Random Forests by Leo Breiman and Adele Cutler.” *Random Forests - Classification Description*, www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html

2.Louppe, G., Wehenkel, L., Sutera, A. and Geurts, P., 2013. Understanding variable importances in forests of randomized trees.

3.Strobl, Carolin, et al. “Conditional Variable Importance for Random Forests.” *BMC Bioinformatics*, vol. 9, no. 1, 2008, p. 307. doi:10.1186/1471-2105-9-307.

4.Strobl, Carolin, et al. “Bias in random forest variable importance measures : Illustrations, sources and a solution.” *BMC Bioinformatics*, vol. 8, no. 1, 2007, doi:10.1186/1471-2105-8-25

5.Zhou, Qifeng, et al. “Gene Selection Using Random Forest and Proximity Differences Criterion on DNA Microarray Data.” *Journal of Convergence Information Technology*, vol. 5, no. 6, 2010, pp. 161–170., doi:10.4156/jcit.vol5.issue6.17.

6.Hapfelmeier, Alexander, et al. “A New Variable Importance Measure for Random Forests with Missing Data.” *Statistics and Computing*, vol. 24, no. 1, 2012, pp. 21–34., doi:10.1007/s11222-012-9349-1.

Terminology

7. *Out of bag* - an out of bag sample, or *“OOB”*, is a sample which was not used in constructed the tree when we use methods such as bagging to create trees in our random forest.

8. *Gini* - when we talk about *gini decrease* or *gini impurity*, we refer to the impurity measure used by Breiman to decide which variable to condition upon when constructing a node in a tree. (see ref. 1)

9. *Proximity matrix* - quoting Breiman (ref 1.5) ; “*A*fter each tree is built, all of the data are run down the tree, and proximities are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one”. - the matrix of proximities is thus a matrix representing these proximities.

10. *Large p, small n* - a problem is said to be “large p, small n” when its dataset has a lot of attributes but a small number of samples.

11. *Primary split* - this terminology is used when we use an *attribute surrogation* technique; I.E., when we define a list of conditions in a node in case a sample reaches this node but cannot be conditioned upon because it has a missing value for the attribute of used in the *primary split*. If that is the case, we will use the nth condition, using a nth *surrogate* variable in our node, for which our sample has a value.