Daniel Kiser
STAT 5443
Dr. Jyotishka Datta
8 May 2018

Review of *Evaluation of variable selection methods for random forests and omics data sets*

In “omics” fields, such as genomics, proteomics, and transcriptomics, researchers are often more interested in finding all of the variables that are relevant to the response (such as a disease state) than they are in predicting the response. This is because finding the relevant variables is important for elucidating the specific genes or biological pathways that lead to a disease. Hence, such researchers require tools that are somewhat different from many of the commonly used machine learning methods, which tend to be geared towards prediction rather than discerning relationships between responses and variables.

One of the machine learning methods that omics researchers are attempting to adapt to their purposes is Random Forest (RF). RFs provide a measure of variable importance that is calculated based on the effects of removing each variable from the model and measuring the change in prediction error. However, it is difficult to use this measure of variable importance to distinguish between relevant and irrelevant variables since it is not clear where the cutoff between relevant and irrelevant variables should be. In *Evaluation of variable selection methods for random forests and omics data sets* (2017), Frauke Degenhardt, Stephan Seifert, and Silke Szymczak compare six different adaptations of RF that can be used to determine which variables are relevant. A brief summary of these adaptations is as follows:

1. Recursive Feature Elimination (RFE): RFE begins by generating a RF that includes all of the available variables. The user defines a certain number of the least important variables to be removed, and a new RF is generated with the remaining variables. The process is repeated until only one variable is left in the model. The prediction errors of each RF are compared, and the variables used to generate the RF with the minimal amount of error above a certain threshold are chosen. (The goal of this method is to find the minimal set of variables that are the best predictors. Since it is not seeking to determine all relevant variables, it seems that the authors are using this method as a baseline by which to measure the other methods.)

2. Recurrent relative variable importance – r2VIM: r2VIM assumes that most of the variables in the dataset are unimportant. A number of RFs are generated using different random seeds, and the average minimum importance value from all the RFs is selected as the baseline value. The other importance values are divided by this baseline, giving their relative importance values. Usually, a relative importance value of 3 is the threshold used to identify relevant variables.

3. Boruta: For each variable, Boruta creates a "shadow variable," or copy, of the original variable. The shadow variable has the same values as the original variable, but these values are permuted so that the shadow variable has no relationship to the response variable. When the RF is generated, the importance measures of the original variables are compared with the maximum importance measure of the shadow variables, and a statistical test is used to determine the probability that a particular variable is relevant to the response. Unlike other methods compared in this study, Boruta can be used on low-dimensional data sets.

4. Permutation: In the permutation method, the response values are repeatedly permuted while the values of the feature variables remain the same. For each permutation, the importance measure for each of the feature variables is determine. These measures are then used to determine the empirical distribution of the importance measures under the null hypothesis. Once the distribution under the null hypothesis has been determined, probabilities can be assigned to the importance measures produced by a RF generated from the unpermuted data set.

5. Altmann: Altmann is similar to the permutation method, but fits a parametric model to the distribution of importance measures in order to reduce the computational cost.

6. Vita: Vita creates an empirical distribution for the null hypothesis by assuming that the distribution of the negative importance measures are the mirror image of the distribution of the positive importance measures. It combines these two distributions into a single empirical distribution for the importance measures of the irrelevant variables. It then uses this distribution to determine the probability that a given variable in the data set is relevant.

To compare these six different methods, the authors tested them on both simulated data and experimental data. The measures they used to compare model performance were: sensitivity, false discovery rate (FDR), stability, empirical power, and root mean square error (RMSE).

First simulated data set:

The first simulated data set consisted of 100 observations and 5000 features and was generated by a nonlinear regression model. The three variables that were relevant to the response were used to create groups of either 10 (scenario 1) or 50 (scenario 2) correlated variables. It was found that in scenario 1, all the methods had about the same level of sensitivity. However, the sensitivity varied tremendously in scenario 2, with Boruta having the largest sensitivity (about 0.8) and RFE having the lowest (about 0.1). In both scenarios, Boruta and Vita were the most stable. RFE was by far the least stable, but this was expected since the purpose of RFE is to find the minimal set of variables that make good predictions. This makes the set of features chosen by RFE highly variable when the features are correlated with each other. The RMSE for all methods were very similar to each other.

In both scenarios, r2VIM had a larger false discovery rate (FDR) than the other methods, though this effect was most visible in scenario 1. In scenario 2, the FDR was significantly reduced for all the methods.

In terms of empirical power (measured by the frequency with which variables were marked as relevant), Boruta appeared to most consistently choose the correct relevant variables.

In order to determine how likely the different methods are to produce false positives, the authors used the same group sizes as before, except this time there was no relationship between the response and the features (scenarios 3 and 4). It was found that the permutation approach had the highest number of false positives, while Vita had zero. All the other methods fell somewhere in between.

Second simulated data set:

To generate the second simulated data set, the authors used the estimated dispersion matrix from a breast cancer data set to create a realistic correlation structure. The simulated data set had 200 observations and 12,592 features, and they randomly assigned an effect size from the set {-3, -2, -1,

-0.5, 0.5, 1, 2, 3} to 200 variables. They were then able to determine the minimum effect size that each of the methods were able to detect. They found that Boruta was the best at detecting the smallest effect sizes, followed by Vita and r2VIM.

Experimental data results:

The authors chose four data sets. Two were methylation data sets that could be used to predict sex, and two were gene expression data sets that could be used to predict breast cancer. They used one of each set of data sets for training and one for testing. The comparison measures they used were stability, classification error, and run-time.

With the exception of RFE on the methylation data, all of the methods had a similar classification error for both types of data sets. While stability between the methods was nearly identical for the methylation data, the stability varied greatly for the breast cancer data. The authors thought this had to do with a difference in effect size between the two data sets. However, Vita appeared to be the most stable, followed by Boruta. It seems that because of the way that stability was calculated in this case, the standard error was unable to be calculated, making it difficult to make informed comparisons between the methods. However, the authors reported that if the method used to calculate stability for the simulated data sets is used, the stability of all the methods is less than 20%. They believe that this is due to small effect sizes in the breast cancer data set.

When run-times are compared between the different methods, the permutation approach was by far the most computationally intensive, taking over 70 hours to run on both the methylation and breast cancer data sets, while Vita was the fastest (taking less than 8 minutes for both data sets!). The run-times for Boruta, r2VIM, and RFE were all within 5 hours for both data sets.

Discussion:

The authors concluded that Boruta and Vita are the best RF methods for variable selection. It seems to me that r2VIM was not far behind Boruta and Vita in many respects, such as in its ability to find smaller effect sizes, its empirical power to identify relevant variables, and its computational cost. It also had the highest stability in the second simulated data set that had a realistic correlation structure. While r2VIM had a higher FDR when the feature variables had a relationship with the response, it had

the second lowest false positive rate (after Vita) for the simulation of the null data set where the variables had no relationship to the response. This strongly suggests that there may be situations in which r2VIM would be the best variable selection method of the ones covered. Of course, given the necessarily limited nature of the simulated data sets, and given that only two types of experimental data sets were used, it would be unwise to suggest that one method is always better than the others. It is likely that all of the methods have their place in some situation. However, it does seem clear that the permutation approach is impractical for the high-dimensional data that is often found in omics studies, due to its long run-time.

Clearly, a big limitation of this study is that it is limited to RF methods. The authors do not compare any of the RF methods with regression methods such as LASSO, Elastic Net and logistic regression, which are also commonly used in omics studies. It would also be interesting to test the ability of Neural Networks (NNs) to find relevant variables in omics data, since in many prediction challenges NNs have outperformed RFs. However, just like RFs, NNs would have to be modified appropriately to be of any use in variable selection, since much of the work done with NNs is focused on prediction. Another downside is that NNs would probably be more computationally intensive to train and would require the tuning of many more parameters than RFs and regression models. Nevertheless, given the demonstrated capacity of NNs to analyze high dimensional data, it would be interesting to compare their capability for variable selection in omics with other models.

As the authors mentioned, it would also be interesting to study ways of incorporating prior biological knowledge into the models that could aid in identifying relevant genes. The most obvious way to do this would be to use some form of prior probabilities and Bayesian reasoning. However, it is difficult for me to see how Bayesian reasoning could be used with RFs, since they are nonparametric. Another approach would probably have to be used.