Distribution Agreement

Signature:

______________________________ _______________

Yi Xiao Date

Applying Weighted Random Forest Algorithm on Metabolic Pathways Selection

By

Yi Xiao
Master of Science in Public Health

Biostatistics and Bioinformatics

____

Tianwei Yu, PhD
(Thesis Advisor)

____

Hao Wu, PhD
(Reader)

Applying Weighted Random Forest Algorithm on Metabolic Pathways Selection

By

Yi Xiao

B.S.
Wuhan University
2016

Thesis Committee Chair: Tianwei Yu, PhD
Reader: Hao Wu, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2019

Abstract

Applying Weighted Random Forest Algorithm on Metabolic Pathways Selection

By Yi Xiao

Background: Functional analysis using high-resolution liquid chromatography−mass spectrometry (LC−MS) data involves data analysis based on metabolic pathways or the genome-scale metabolic network. It is critical in feature selection and interpretation of metabolomics data. One of the main challenges is the lack of the feature identity in the LC−MS data. When matching mass-to-charge ratio (m/z) values of the features to theoretical values, some features can be matched to multiple known metabolites. When multiple matching occurs, usually only one of the matches can be true. Current network/pathway analysis methods ignore the uncertainty in metabolite identification, which could lead to some pathways that are not related to disease outcome being selected by including erroneously matched features.

Methods: We explored three potential methods based on Random Forest to address the multi-match issue. All the three approaches attempt to down-weight the contribution of multi-matched features to the pathway. (1) Weighted tree approach 1: lowering the tree weight if percent of multi-matched features used in the tree is high; (2) weighted tree approach 2: compute tree weight based on both feature importance score and the features' multi-match status; (3) weighted sampling approach: apply multi-match status of each feature in variable-importance Random Forest, which samples features at each node based on a prior probability.

Results: By conducting a series of simulation studies, we found that (1) using weighted tree approach 1, the differentiation between true/false pathways is not significantly different from unweighted random forest; (2) using weighted tree approach 2, the weighted random forest show significant lower MSE, but still doesn't out-perform unweighted Random Forest in pathway selection; (3) the weighted sampling approach works best on distinguishing between pathway with multi-match true features and pathway with no multi-match true features.

Conclusion: the random forest prediction accuracy is not sensitive to the change of tree weight based on feature information. The weighted sampling approach works better. We decided to use multi-match information and importance score to adjust sampling probability. We expect to see the false pathways with more multi-match features to have lower prediction accuracy than the true pathways in which only part of true features are multi-match.

Applying Weighted Random Forest Algorithm on Metabolic Pathways Selection

By

Yi Xiao

B.S.
Wuhan University
2016

Thesis Committee Chair: Tianwei Yu, PhD

Reader: Hao Wu, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2019

Table of Contents

1. Introduction

1.1 Development of statistical methods for Gene Set Analysis and its application in metabolomics

Given the rapid increase in gene expression, genome-wide association, and proteomics/metabolomics studies, a wide range of analytical tools have been developed to determine the property of a feature set's relevant to phenotypes of interest. Most of the methods were first developed for gene sets [1]. Owing to the complexity of gene-gene interaction, early methods focused more on identifying individual relevant genes functioning on phenotypes of interest [2]. However, evaluation of the genes set as a functional unit was importance because considering the pathways as a whole can offer easy functional interpretations and help guide the selection of genes.

In common approaches, genes are first ranked according to the evidence for differential expression. The top rank genes list is compared to predefined gene sets representing different pathways, thus determining which sets are overrepresented [3].

To determine gene set importance, a gene set statistic is defined to represent its significance in relation to the phenotype. Two types of null hypothesis are defined based on questions of interest. Q1: Does the average gene in the gene set show the same pattern of association to the phenotype as the rest of the gene sets? Q2: Does the gene set contain any gene related to phenotype [4]?

Because significant gene sets should be distinguished from equally sized randomly chosen gene lists, shuffling genes was used to find the background distribution to address Q1. For Q2, shuffling phenotypes was used because the focus is testing association with specific phenotypes compared to randomly selected phenotype [5].

In the common situation, Q2 is more favored than Q1 because it preserves relationship of genes in a gene set, thus addressing correlation to phenotype as a whole. The current commonly used gene-set level statistics include χ^2-test, mean test, median test and Wilcoxon rank sum test (WKS test). Moreover, significance measurement is based on the validity of analytical background distribution. Multiple testing correction also needs to be completed at the end of the process. Almost all the test statistics above ignore the correlation between genes within a gene set. However, gene regulation on downstream genes plays a crucial role in phenotype change owing to the complexity of pathway topology structure [2]. This correlation violates the independent assumption required by many statistical models.

In untargeted LC-MS metabolomics data, one of the main challenges is the lack of the feature identity. Features are matched by mass-to-charge ratio (m/z) values to theoretical values of known metabolites. Some features can be matched to multiple known metabolites. When multiple matching occurs, usually only one of the matches can be true. The multi-match between features and the complex correlation between features is a unique challenge when trying to apply gene set analysis methodology to metabolomics data [6].

1.2 Random forest application and advancement

Random forest is a tree-based ensemble learning method widely applied in ‘large p, small n’ problems, which account for high-dimension data with features interaction like genomic and metabolomic data [7]. To explore the genomic marker associated with phenotypes, variable selection or pathway ranking should be performed on high-dimensional correlated genomic data among which feature interaction effect is hard to pre-specify. Accordingly, random forest excels in performance because of the flexibility of the tree structure to capture interactions and construct complex response surface. Random forest also has a number of adjustable parameters which make the construction of the random forest easily controllable. In addition, it is easy to obtain cross-validation results with out-of-bag samples [8].

Given the training data, random forest is built on many different trees, which makes aggregation effective. the difference between the trees relies on two crucial factors: first is best split at each node is chosen from random subsets of predictors; the second is each tree is built on a bootstrap sample, leaving approximately one-third of the observations as out-of-bag data which then can be used for the estimation of accuracy. On the progress of random forest construction, much information is yield to interpret data.

Importance of a variable is defined as Gini index reduction for the variable summed over each tree, then divided by the number of trees. Likewise, permutation importance is also frequently used for a variable, which randomly permutes the given variable in OOB data and calculates the marginal decrease in accuracy. The larger the marginal decrease, the more predictive the variable. In addition, proximity is a measure of similarity between samples under unsupervised learning situation. This can be used for clustering and missing data imputation[9].

1.3 Weighted random forest

Equal weights for both variables and trees are not all appropriate in situations with some variable surpass others in predictive power. This especially true in genomics and metabolomics applications, where the majority of variables are irrelevant to the phenotype. A variety of weighting methods have been proposed to increase the accuracy of Random Forest. For example, when classes of response variable are not balanced, RF classifier tends to bias toward the majority class. Weights assigned to each class are applied to node splitting standard and solve the issue of imbalance. In addition, weighted random sampling to choose candidate variables at each node is also considered as an efficient approach to boost the overall predictive power. As importantly, weight-adjusted voting for ensembles which use iterative weights for sample and trees are a practical extension. Moreover, weight incorporating tree accuracy and variable importance has also been introduced. These all serve as stimulation for our study [10].

1.4 Applying Random Forest in pathway analysis

Pathway analysis can be conducted using predictive models. The predictive power of a pathway on the phenotype can be seen as a reflection of how strongly the pathway is associated with the phenotype. This approach, although not as statistical rigorous as some other pathway analysis methods, allows maximum flexibility in terms of allowing nonlinear and complex relations between features and the phenotype to be considered, which is especially suitable for metabolomics data, where dynamic regulations are abundant.

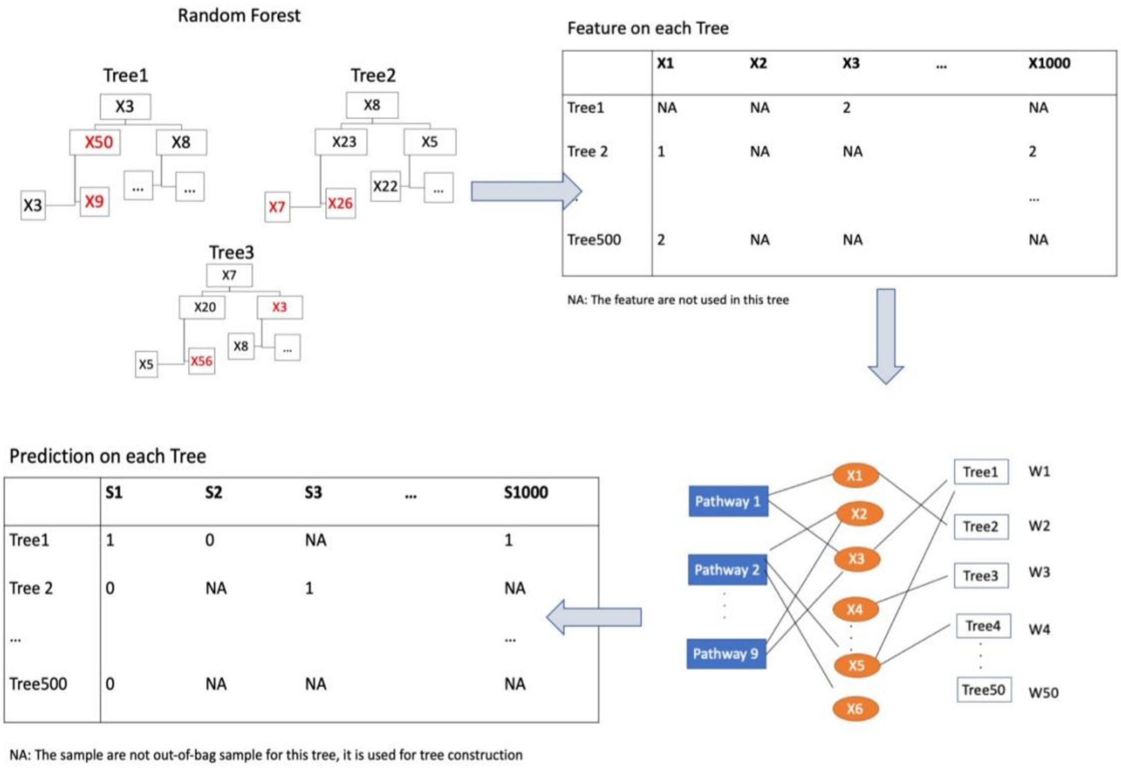
In this study, we employ Random Forest for pathway analysis, and attempt to incorporate feature multi-matching information into the building of the forest. The goal is to differentiate truly predictive pathways from those that appear predictive only because some erroneously annotated features. If a pathway is truly associated with the phenotype, we expect multiple features annotated to the pathway contribute to the prediction using Random Forest, and down-weighting those multi-matched features will have a limited impact on the predictive power of the pathway overall. On the other hand, if a pathway is not truly associated with the phenotype, and some multiple-matched features make it apparently predictive, then down-weighting multi-matched features will substantially reduce the predictive power. We explore three different approaches to construct weighted Random Forest, and test whether they can achieve our goal.

2. Method

2.1 Weighted random forest

Random forest is one kind of ensemble learning, which build a prediction model by combining a collection of weak learners. In this case, the task can be broken down into two: develop base learners from training data, combine them into a predictor model [11]. In traditional random forest algorithm, each tree draws a different bootstrapped sample of N subjects and select m variable at random, pick the best split point (variable), and then split the node into two daughter nodes. These steps are repeated at each node to form a tree. Then prediction is based on the average fit on each tree (regression), or by taking the majority votes from all trees [10].

However, in our situation, not all variables should contribute equally on the prediction for response variable. Accounting for that, the weights based on variables used should be applied on each tree. For the multiple match relationship between LC-MS features and disease-related pathway, what we aim to do is lower the weight of trees that contain multiple matched features. Hopefully, it will decrease the accuracy of falsely related pathways mostly due to multi-matching, and keep truly related pathway as they have highly predictive uniquely matched features. Figure 1 shows the overall workflow.



Feature on each Tree

	X1	X2	X3	...	X1000
Tree1	NA	NA	2		NA
Tree 2	1	NA	NA		2
.					...
Tree500	2	NA	NA		NA

NA: The feature are not used in this tree

Prediction on each Tree

	S1	S2	S3	...	S1000
Tree1	1	0	NA		1
Tree 2	0	NA	1		NA
...					...
Tree500	0	NA	NA		NA

NA: The sample are not out-of-bag sample for this tree, it is used for tree construction

Figure 1, overall workflow. From the random forest, we extract two pieces of detailed information. Each tree is comprised of multiple features, and one feature can be used on multiple nodes. We extract usage of features in each tree in the upper-right table. We generate tree weight using the table based on the multi-matching status of the features in each tree. We then extract prediction of each sample from each tree in the lower-left table, and combine tree weight and tree prediction to compute the weighted prediction for each sample.

2.2 Choice of weight

Assuming a data consists of *p* predictor variables, a binary response variable coded as 0 or 1, collected on *N* subjects.

For each tree $j = 1, \ldots, ntree, compute\ weight\ as\ w_j$

$n_{multi}: number\ of\ unique\ features\ matched\ to\ more\ than\ one\ pathway\ used\ in\ the\ tree$

$n_{total}: number\ of\ unique\ features\ in\ the\ tree$

$$w_j = (1 - \frac{n_{multi}}{n_{total}})^4$$

For each subject $i = 1, \ldots, N, compute\ prediction\ as\ v_i\ based\ on\ OBB\ prediction\ \ v_{ij}$

$$v_i = \frac{\sum_{j=1}^{ntree} w_j v_{ij} I(sample\ i\ is\ OOB\ in\ tree\ j)}{\sum_{j=1}^{ntree} w_j\, I(sample\ i\ is\ OOB\ in\ tree\ j)},$$

$$\hat{y}_i = I(v_i > 0.5),$$

Where *I()* is the identity function, which takes value 1 when the statement in the parenthesis is true, and 0 otherwise.

2.3 Simulation study

2.3.1 Feature level simulations

In order to compare the performance of weighted random forest (wRF) to traditional RF, we conducted a simulation study. We simulated dataset that consists of 2000 metabolic features and 1000 subjects (500 cases and 500 controls), with the disease status coded as (0, 1) representing the subject contracted a disease or not. Among the 2000 features, 200 are true predictors.

We wish to construct a regression model on metabolic feature ($X_i: i = 1, ..200$) in predicting disease (D_j) status, which is as follows:

$$D_j = \sum_{i=1}^{200} X_i \beta_i + \varepsilon,$$

$$y_j = I\left(D_j > median(\{D_k\}_{k=1,\ldots,1000})\right)$$

Where $\varepsilon \sim N(0,1)$, X_i is generated by random sampling from the standard normal distribution, β_i are from random sampling from the uniform distribution on (0, 1) interval.

2.3.2 *Simulating the Pathways*

We create 50 pathways, among which 10 are true pathway related to disease and the remaining are not. For true pathways, we randomly assign a fixed number of true features, with a pre-specified a number (k) of true features to be strictly non-overlapping with other pathways. The purpose of controlling this proportion is to make the proportion of multi-matched features to be roughly the same for all pathways. At the same time, a false pathway comprises of features randomly sampled from all features. We tune the k parameter and monitor the multi-match percentage to keep all pathways nearly the same.

2.3.3 *Tune Parameters*

In real data, we don't know which pathway is truly related to the disease and which are not. We only observe their proportion of multi-matched features and their predictive power. Thus, in simulation, we try to adjust the number of true predictors in the true pathways to check in what situation the weighted method works well. Also, when we sample randomly from the feature pool for each pathway, the multi-match feature percent should be monitored. If all features were randomly sampled, the true pathway will have more multi-matched future, as they all include a higher proportion of true features sampled from a small pool. So we control two main parameter:

$n_{feature}$: Number of true predictors in true pathways;

r : Ratio of non-multi match true features in true pathways, $k = n_{feature} * r$

We run the simulation on $n_{feature} = 5,10,15,20,25,30$; $r = 0.2,0.5,0.6$. All result are the average on 50 simulations.

3. Result

3.1 Tree level accuracy is impacted by the number of true predictors involved in the tree

To prove the feasibility of our algorithm, we need to assess whether trees involving more true predictors are indeed more predictive. We also observe how the tree accuracy based on OOB samples is related to different weights on the tree. We found that the accuracy has a positive trend as the percent of true predictor increase, though the relation is quite noisy and non-linear (Figure 2). The weights in the figure correspond to the following:

$$w\_tree\_pos1 = \ percent\ true \qquad w\_tree\_neg1 = \ 1 - \ percent\ true$$

$$w\_tree\_pos2 = \ percent\ true\hat{\ }2 \qquad w\_tree\_neg2 = \ (1 - percent\ true)\hat{\ }2$$

$$w\_tree\_pos3 = \ percent\ true\hat{\ }4 \qquad w\_tree\_neg3 = \ (1 - percent\ true)\hat{\ }4$$

As shown on the first row (Figure 2), the accuracy has positive association with percent of true predictors in tree. This proves potentially manipulating the tree level output can influence the overall accuracy of the random forest. However, the range of tree-level accuracy is widely varied.

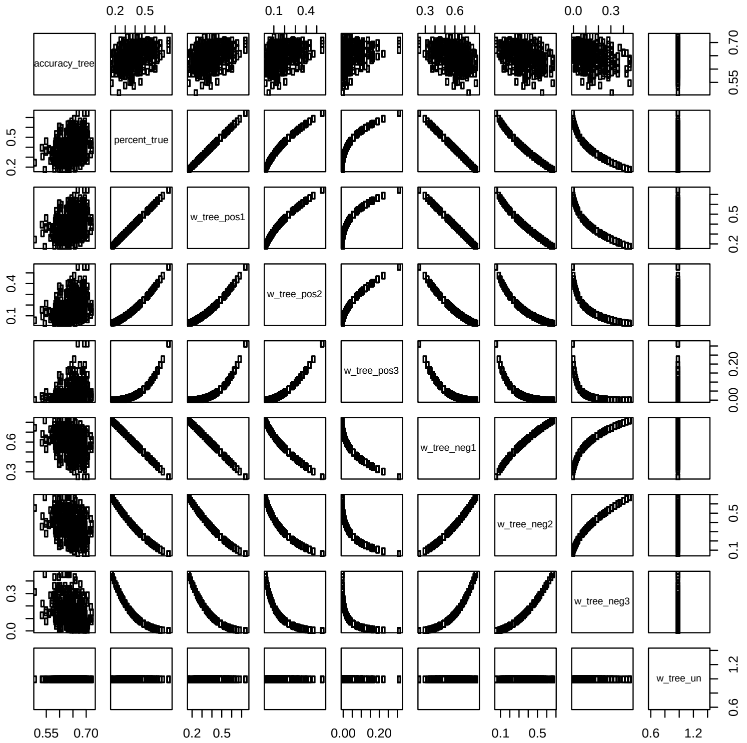


Figure 2, plot of tree-level accuracy, percent true features, and various weighting parameters.

3.2 Examine the number of multi-match feature and number of true predictors

We examine one multi count plot with r=0.6 (Figure 3). When true predictor=5, 10, 15, 20, all show similar level of multi count feature proportions among all pathways. When true predictor= 25,30, the true pathways tend to have larger multi count proportions. The situation may require larger r value. In real data, the number of true predictors in each pathway is expected to be <50%. Thus the latter situation may not be as relevant.

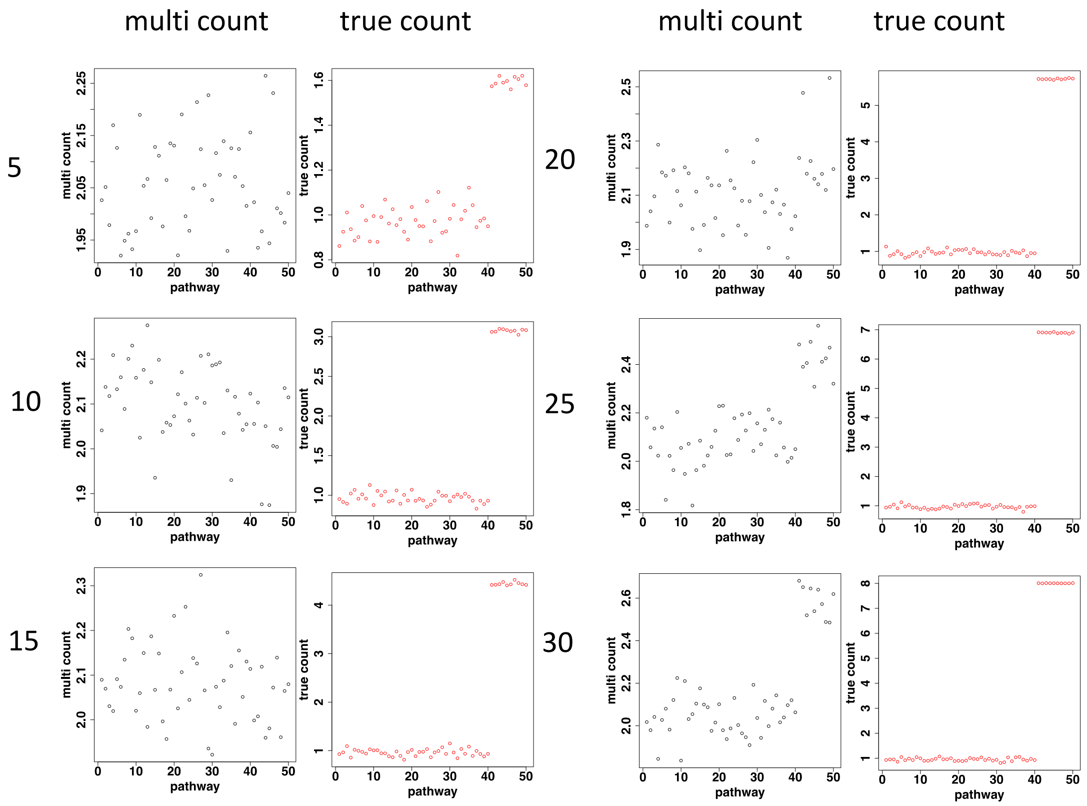


Figure 3 scatterplot of number of multi-match feature and true predictors for all pathways at r=0.6. The last 10 pathways are true pathways. The three columns from left to right are: number of true predictors per true pathway, the scatterplot of number of multi-match features for each pathway, the scatterplot of the number of true features for each pathway.

3.3 Comparison between wRF and RF in pathway selection

We next examine whether the weighted method show a better performance in terms of differentiating true pathways from false pathways. The OOB accuracy rate was found for each pathway, and compared with the pathway labels using Receiver Operating Characteristic (ROC) curves and Precision Recall (PR) curves.

At r=0.6, the boxplots (Figure 4) for both Precision recall and ROC curve AUC were created on 50 simulations for each $n_{feature} = 5,10,15,20,25,30$. In all cases, there is no clear difference between the weighted and unweighted results.

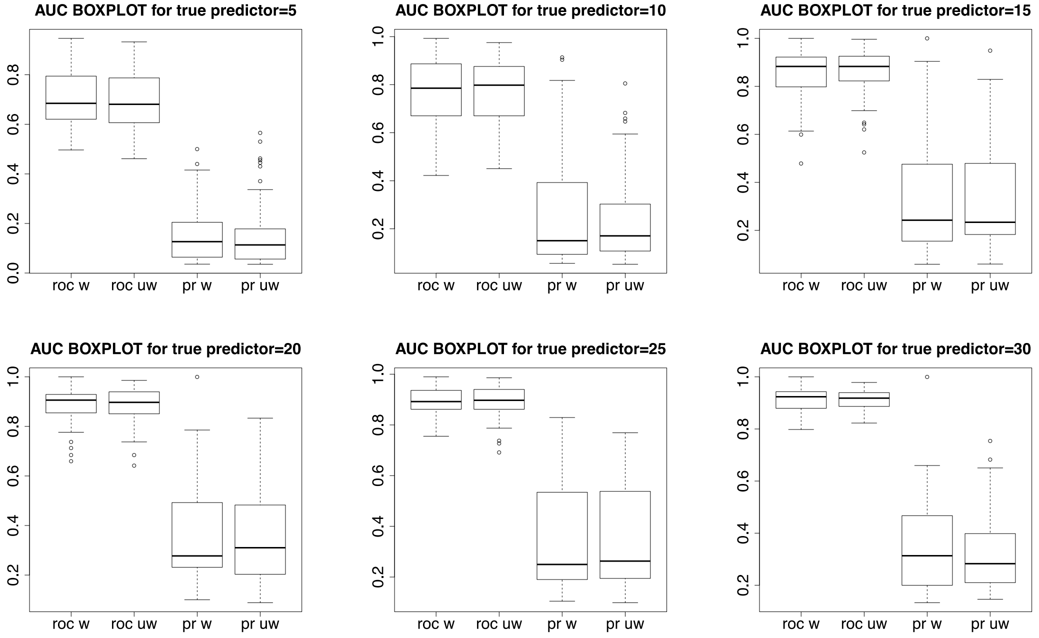


Figure 4, Boxplots of ROC-AUC and PR-AUC of weighted and unweighted RF. the six boxplot above from left to right, upper to lower are true predictors from 5 to 30 sequentially. The fix ratio of r=0.6 is used in all plots. In each box plot, the 4 box from left to right are sequentially weighted ROC AUC, unweighted ROC AUC, weighted Precision Recall AUC, and unweighted Precision Recall AUC.

This plots of PR-AUC and ROC-AUC for multiple combinations of #true predictors and fix ratios (r) shows the AUC has a significant overall increase when the number of true predictor increase (Figure 5). The AUC values can be compared on two levels: among different fix ratio and weight vs. traditional. Examining the AUC from ROC curve, the accuracy tends to converge as the true predictor increase. However, the AUC from the PR curve shows an inconsistent trend. More importantly, the weighted method seems not to have a significant improvement in accuracy.

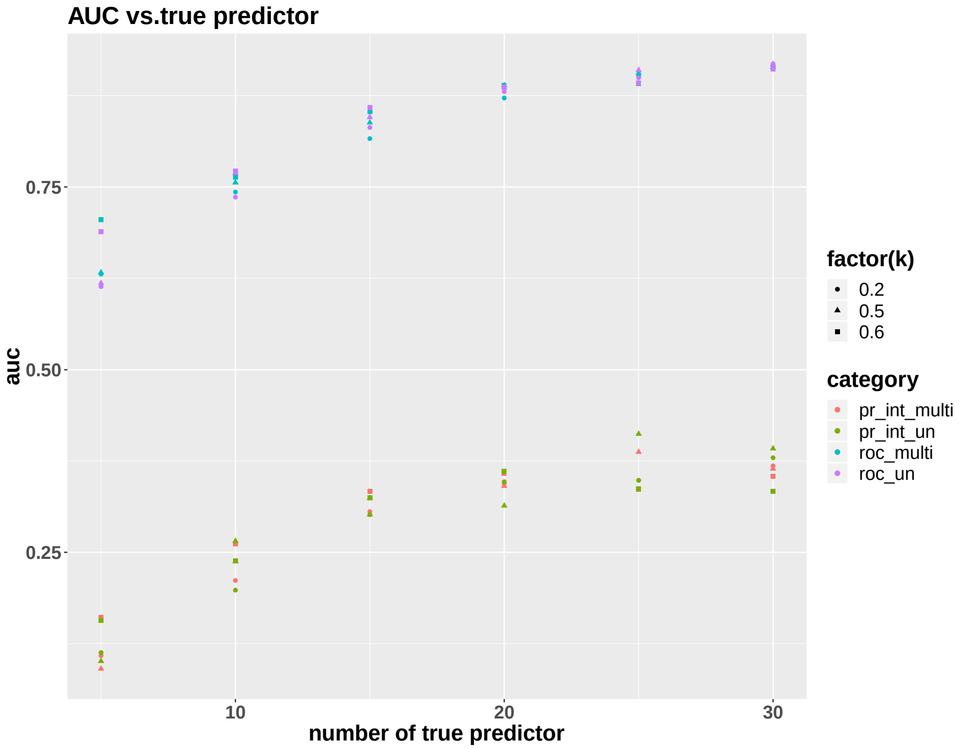


Figure 5, Comparison of ROC-AUC and PR-AUC of weighted and unweighted RF. The point and line shape represent different r values; the line color represent different category of AUC: pr_int_multi: the Precision-Recall AUC for multi-match weighted method; pr_int_un: the Precision-Recall AUC for unweighted method; roc_multi: the ROC curve AUC for multi-match weighted method; roc_un: the ROC curve AUC for unweighted method.

4. Discussion and Conclusion

The result does not show significant difference on AUC when we use multi-match information to calculate the tree weight. We think it is because most of trees contain one or two multi-match nodes. If we down-weight for all the trees, the overall accuracy will not change much. In addition, the multi-match information does not include the feature predictive power, it is an information we wish to apply on the model.

We may combine the feature importance score and number of pathway each feature matches to as feature weight, and design a formula to convert feature weight into tree weight. This would be more precise in utilizing the tree level info.

4.1 Adjust on the formula by adding feature importance

Scheme1: construct tree weight based on multi-match and importance for each feature

Assume $tree_j$ has used features $f_1, f_2, f_3 \dots, f_m$, and feature f_i matches to μ_i pathways, and its importance score calculated from RF is I_i, we can weigh the tree by

$$w_j = \frac{\log(I_1 * I_2 * \dots * I_n) - r * log(\mu_1 * \mu_2 * \dots * \mu_n\ )}{n}$$

Scheme 2: apply R package 'viRandomForests' and use multi-match in feature sampling probability

Brief introduction on '*viRandomForests*': comparing to traditional random forest which sample features with equal probability at each node, the '*viRandomForests*' samples according to variable importance [12]. The default setting samples based on variable importance. We can set sample probability in function by ourselves. This flexibility allows us to add multi-match info when constructing sample probability.

Extract feature importance score I_i, construct the sample probability $f_{prob,i}$ for the feature f_i, the number of pathways f_i matches is $n_{multi-i}$, d is the total number of features

$$w_i = \begin{cases} \frac{1}{d} + \frac{I_i}{\max(I_j)} & if \max(I_j) > 0 \\ \frac{1}{d} & otherwise \end{cases}$$

$$f_{prob,i} = \frac{w_i}{n_{multi-i}}$$

Our simple workflow summary:

Run traditional random forest

Extract feature importance score

Calculate $\{f_{prob,i}\}_{i=1,\dots,p}$

Run *viRandomForests* function using $\{f_{prob,i}\}_{i=1,\dots,p}$ as sampling probabilities

Scheme 3: A more extreme version of Scheme 2.

In order to yield more obvious comparison between weighted and unweighted methods, one extreme method to decrease the weight of multi-matched feature is to force the feature importance I_i to $\min(I_i, i = 1, \dots, p)$ if $n_{multi-i}$>1.

$$\widetilde{I_\imath} = \begin{cases} I_i & if\, n_{multi-i} = 1 \\ \min(I_i\ ) & if\, n_{multi-i} > 1 \end{cases}$$

$$f_{prob,i} = \begin{cases} \frac{1}{d} + \frac{\widetilde{I_\iota}}{\max(\widetilde{I_\iota})} & if \max(\widetilde{I_\iota}) > 0 \\ \frac{1}{d} & otherwise \end{cases}$$

The workflow follows that of Scheme 2.

4.2 Simple comparison between the schemes

We create a data contains 100 features among which 10 are true predictors, and a continuous outcome variable using a linear model. We calculate MSE using three weighting methods for each of the schemes mentioned in section 4.1.

- w: Weight using only importance score;
- w.m: Weight using both importance score and multi-match, artificially set all true predictor as multi-match, to check whether the MSE will decrease drastically;
- uw: Unweighted.

As shown in Figure 6, only scheme 3 showed a significant difference between weighting using importance score only and weighting using both importance score and multi-match status. Thus in future simulations, scheme 3 is the most promising to separate true pathways from the false ones, where false pathway have been assigned to true features but most of them are multi-matched to many pathways. We decided to generate multiple pathway data simulation using scheme3.

4.3 Conclusion

In conclusion, weighted Random Forest may be a promising method to select important pathways in the presence of multi-matching. We have found that weighting by trees is generally not sensitive enough for multi-matching, and weighting by feature sampling probability may be a viable approach.

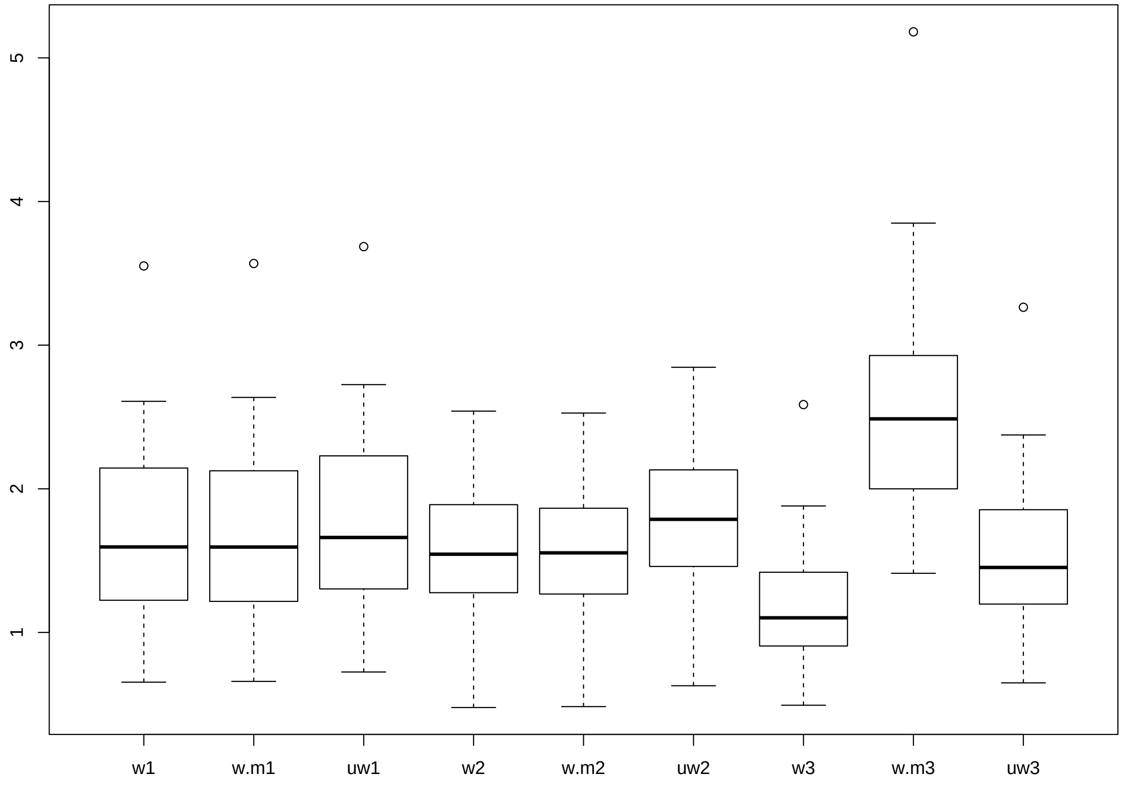


Figure 6, boxplot of MSE generated from the 3 schemes. The labels 1, 2 and 3 represent scheme 1,2,3 respectively. w: Weight using only importance score; w.m: weight using both importance score and multi-match, artificially setting all true predictor as multi-match; uw: unweighted.

Reference

1. de Leeuw, C.A., et al., *The statistical properties of gene-set analysis.* Nature Reviews Genetics, 2016. **17**(6): p. 353-364.

2. Hung, J.H., et al., *Gene set enrichment analysis: performance evaluation and usage guidelines.* Brief Bioinform, 2012. **13**(3): p. 281-91.

3. Wang, K., M. Li, and M. Bucan, *Pathway-based approaches for analysis of genomewide association studies.* Am J Hum Genet, 2007. **81**(6): p. 1278-83.

4. Wang, L., et al., *Gene set analysis of genome-wide association studies: methodological issues and perspectives.* Genomics, 2011. **98**(1): p. 1-8.

5. Tian, L., et al., *Discovering statistically significant pathways in expression profiling studies.* Proc Natl Acad Sci U S A, 2005. **102**(38): p. 13544-9.

6. Cai, Q., et al., *Network Marker Selection for Untargeted LC-MS Metabolomics Data.* J Proteome Res, 2017. **16**(3): p. 1261-1269.

7. Fawagreh, K., M.M. Gaber, and E. Elyan, *Random forests: from early developments to recent advancements.* Systems Science & Control Engineering, 2014. **2**(1): p. 602-609.

8. Chen, X. and H. Ishwaran, *Random forests for genomic data analysis.* Genomics, 2012. **99**(6): p. 323-9.

9. Pang, H., et al., *Pathway analysis using random forests classification and regression.* Bioinformatics, 2006. **22**(16): p. 2028-36.

10. Winham, S.J., R.R. Freimuth, and J.M. Biernacka, *A Weighted Random Forests Approach to Improve Predictive Performance.* Stat Anal Data Min, 2013. **6**(6): p. 496-505.

11. *<the elements of statistical learning.pdf>*.

12. Liu, Y. and H. Zhao, *Variable importance-weighted Random Forests.* Quant Biol, 2017. **5**(4): p. 338-351.