Improving the interpretability of random forest models of genetic association in the presence of non-additive interactions

DEPARTMENT of BIOSTATISTICS EPIDEMIOLOGY & INFORMATICS

Alena Orlenko, Jason H Moore

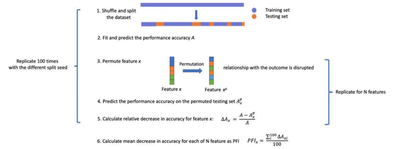
Penn Institute for Biomedical Informatics

Background

- Non-additive interactions among genes are frequently associated with a number of phenotypes, including known complex diseases such as Alzheimer's, diabetes, and cardiovascular disease.
- Detecting interactions requires careful selection of analytical methods, and some machine learning algorithms are unable or underpowered to detect or model feature interactions that exhibit non-additivity.
- The Random Forest (RF) method is often employed in these efforts due to its ability to detect and model non-additive interactions. RF has the built-in ability to estimate feature importance scores, a characteristic that allows the model to be interpreted with the order and effect size of the feature association with the outcome. This characteristic is very important for epidemiological and clinical studies where results of predictive modeling could be used to define the future direction of the research efforts.
- An alternative way to interpret the model is with a permutation feature importance metric which employs a permutation approach and with the Shapely additive explanations which employ cooperative game theory approach.
- Currently, it is unclear which RF feature importance metric provides a superior estimation of the true informative contribution of features in genetic association analysis.

Methods

- We compared three feature importance metrics: RF's built-in feature importance coefficients (BIC), mean SHAP values, and PFI coefficients in real and simulated datasets with non-additive interactions



Scheme of the permutation feature importance custom implementation

- We used Heuristic Identification of Biological Architectures for simulating Complex Hierarchical Interactions (HIBACHI) software to simulate genetic datasets with non-additive epistatic interactions of different complexity.
- The HIBACHI framework has the ability to consider any desirable biological concept in the form of mathematical expressions that define the genotype-phenotype relationship and evolve models that can be used to simulate data consistent with that relationship. We set up a simulation goal to maximize two- or three-way interactions among features and compared RF's feature importance metrics with the sensitivity analysis results of the simulated data that provided us with the ground truth information about the feature ranks
- To examine the convergence of the RF's feature importance metrics we used two real-world datasets with evidence for non-additive interactions (genome-wide association study of Alzheimer's Disease and a genome-wide association study of Primary Open Angle Glaucoma).
- We used the visualization of the statistical interaction network (ViSEN) method to analyze and visualize SNP main effects, and two-way and three-way gene-gene interactions among SNPs for real-world datasets via the mutual information and information gain terms.

Results

Evaluation of the feature importances in real-world datasets with non-additive interactions

- The most powerful predictor of Alzheimer's disease at this time is ApoE E4 gene variation: one or two copies of ApoE is associated with an increased risk of disease onset. Some carriers of ApoE E4 variation haven't developed an Alzheimer's disease so it is very likely that other genetic factors are involved in disease's pathophysiology. ViSEN entropy-based analysis revealed several strong pairwise genetic interactions, along with the known largest independent signal from the ApoE variant (rs429358) (Fig.1A)
- ViSEN method allocated non-additive interactions within the Glaucoma disease dataset: several strong pairwise interactions in addition to the independent main effect contribution from the SNP affiliated with retinal ganglion cells pathology (rs2157719) have been confirmed (Fig.1B)

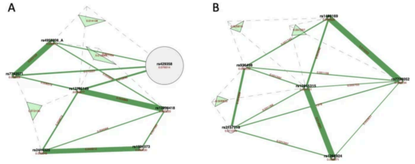


Figure 1. ViSEN plots for selected SNPs in Alzheimer's (**a**) and Glaucoma (**b**) datasets. SNPs main effects, two-way and three-way IG values are noted respectively

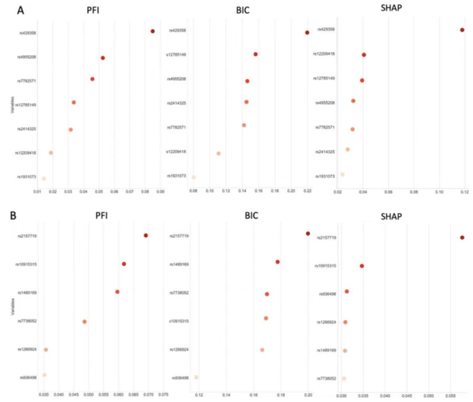


Figure 2. PFI and BIC estimates for Alzheimer's (**a**) and Glaucoma (**b**) datasets. *PFI -permutation feature importance, BIC – build-in coefficients, SHAP - shapley additive explanations*

- Three feature importance metrics were considered, PFI, BIC and SHAP, and each was compared after RF analysis of data derived from genome-wide association studies of Glaucoma and Alzheimer's.
- The resulting feature ranking confirms the lack of consensus between the studied metrics (Fig.2 A,B)

Evaluation of feature importance metrics performance with simulated datasets

Sample size 1000												
Two-Way IG							Three-Way IG					
% of cases:	25%			50%			25%			50%		
Metrics:	PFI	BIC	SHAP	PFI	BIC	SHAP	PFI	BIC	SHAP	PFI	BIC	SHAP
F1	70%	41%	**71%**	**91%**	42%	82%	**80%**	38%	57%	**79%**	18%	68%
F2	**63%**	41%	62%	**90%**	42%	82%	**69%**	33%	45%	**60%**	36%	52%
F3	**93%**	84%	89%	78%	**82%**	81%	**79%**	33%	53%	71%	18%	**73%**
F4	17%	17%	15%	15%	17%	16%	**74%**	55%	58%	11%	**14%**	11%
F5	5%	4%	5%	4%	5%	3%	33%	31%	29%	5%	6%	4%
Sample size 10,000												
Two-Way IG							Three-Way IG					
% of cases:	25%			50%			25%			50%		
Metrics:	PFI	BIC	SHAP	PFI	BIC	SHAP	PFI	BIC	SHAP	PFI	BIC	SHAP
F1	**79%**	36%	73%	**83%**	39%	75%	**89%**	42%	68%	**86%**	19%	76%
F2	**79%**	32%	73%	**83%**	39%	75%	**80%**	34%	49%	**75%**	34%	56%
F3	**100%**	87%	99%	**81%**	79%	81%	**89%**	33%	66%	**85%**	22%	70%
F4	13%	12%	13%	44%	38%	40%	**84%**	62%	69%	**53%**	51%	52%
F5	5%	4%	5%	15%	13%	14%	40%	38%	37%	11%	9%	10%

F1, F2, etc. – feature ranks, *PFI* permutation feature importance, *BIC* build-in coefficients, *SHAP* shapley additive explanations, *IG* information gain

Table 1. PFI, BIC and SHAP success in identification of feature ranks in datasets with two-way and three-way epistatic interactions. It is expressed as the percentage of a match of a metric rank's estimate with the true feature rank that was retrieved with the HIBACHI sensitivity analysis

- Prediction uncertainty has been associated with RF predictions in the past and we attempted to reveal the true interpretation with the computational experiments driven by HIBACHI simulations.
- We set up a simulation goal to maximize two- or three-way interactions among features and compared RF's feature importance metrics with the sensitivity analysis results of the simulated data that provided us with the ground truth information about the feature ranks.
- In all HIBACHI experimental setups, which included such factors as the proportion of cases and controls, sample size and interaction complexity, PFI metrics produced the most precise feature ranking (Table 1, Fig. 3).
- Although BIC and SHAP metrics misplaced feature ranks for the large percentage of replicates with BIC failed to identify the majority of them, it correctly identified features that belong to the interactive pair or trio by putting them as a top-ranked two and three features correspondingly.
- While BIC and SHAP metrics can still be useful, when there is a need for an absolute precision, PFI estimation method should be used.

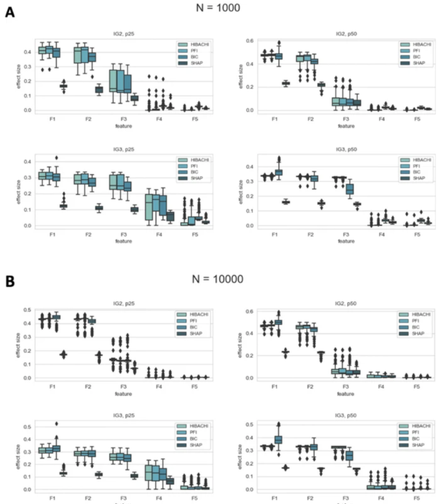


Figure 3. Effect size per feature rank estimated by PFI, BIC, SHAP and HIBACHI sensitivity analysis for sample size 1000 (**a**) and 10,000 (**b**). *F1, F2,* etc. *– feature ranks, PFI -permutation feature importance, BIC – build-in coefficients, SHAP - shapley additive explanations, IG-Information Gain, p25, p50 – percentage of cases*

Conclusion

- We performed a comparative analysis of feature importance metrics with the aim to improve Random Forest's interpretability in datasets with complex interactions.
- By analyzing both real and simulated data, we established that the permutation feature importance metric provides more precise feature importance rank estimation in the presence of non-additive interactions.

Reference: Orlenko A, Moore JH. A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions. BioData Min. 2021 Jan 29;14(1):9. PMID: 33514397

Funding: This work was supported by National Institutes of Health (USA) grants LM010098 and AI116794.