doi 10.18699/bgrs2024-4.3-08

The performance of machine learning approach in genome-wide association study of disease

Khvorykh G.[1]*, Belousov M.[2], Limborska S.[1], Khrunin A.[1]

[1] *National Research Centre "Kurchatov Institute", Moscow, Russia*

[2] *National Research University Higher School of Economics, Moscow, Russia*

** gennady.khvorykh@gmail.com*

Key words: genome-wide association study; single nucleotide polymorphism; machine learning; random forest; xgboos; shapley values

Motivation and Aim: The most common diseases like heart disease, strokes and diabetes, cancer, mental health, and autoimmune disorders are the most expensive and worried about in society. Individual differences in susceptibility to them are known to be substantially genetically determined. In spite of the tremendous development of data collection technologies in the field of molecular genetics occurring last decades, the dissection of genetic architecture of common diseases still remains a challenging task. Problems with sample sizes, environmental factors, pleiotropy are the main challenges in understanding common diseases [1]. Certain progress in this field is ascertained with machine learning (ML) algorithms. They may substitute traditional genome-wide association studies (GWAS) known for almost 20 years [2]. One ML-based approach consists of ranking single-nucleotide polymorphisms (SNPs) by their ability to predict case or control group the individual belongs to. The higher the rank, the more evident the association of SNP with a disease. However, the efficacy of this strategy demands both theoretical and empirical examinations. In this research, we investigated computationally the influence of data sizes on ranking of SNPs. Two metrics to rank the loci were considered: the feature impurity importance and SHapley Additive exPlanations (SHAP) [3]. The models were trained with Random Forest (RF) [4] and eXtreme Gradient Boosted trees (XGBoost) [5] using three data sets, each generated under the assumption of uncorrelated or correlated loci. The performance of the approach was characterized by accuracy, specificity and sensitivity. The results obtained were compared with those of classical GWAS. The observations thus made contributed to proper application of ML in associative studies.

Methods and Algorithms: Assuming alleles were uncorrelated, we generated genotypes for 25100 biallelic loci with Plink 1.9 tool (argument --simulate) [6]. The allele frequencies were uniformly distributed within the interval [0.05, 0.95]. One hundred loci were made to be associated with a disease. Another set of genotypes accounted for LD in groups of individuals of European ancestry. It was generated from haplotypes of CEU population (N = 95) from 1000 Genomes project [7] with Hapgen 2 [8] and included 23583 SNPs from chromosome 22. In this case, there were five SNPs associated with disease (rs4823464, rs137425, rs3761422, rs9606478, and rs7292279). The odds ratios (ORs) equaled to 1.5 and 2.25 for heterozygous and homozygous allele pairs, respectively. The disease loci were validated with Plink 1.9 tool (arguments --assoc). The case/control groups included 652/652, 4929/652, and 4929/4929 individuals, respectively. They are refereed further as small, unbalanced, and large datasets. The sizes of the unbalanced data matched those of the actual genotype-phenotype data that we had

previously used for GWAS of ischemic stroke [9]. The genotypes represented by symbols were converted into binary variables by one-hot encoding. The synthetic datasets thus obtained are available at https://github.com/inzilico/synthetic-data.git. The models were trained with RandomForestClassifier() from module scikit-learn 1.4 [10] and XGBClassifier() from package XGBoost 2.0.3 using the default arguments. The feature impurity importances were obtained from feature_importances_ attribute of trained models, the SHAP values were estimated with TreeExplainer() and Explainer() functions from shap package [11]. The performance of ML-based approach was assessed by accuracy, sensitivity, and specificity, where loci associated to disease were selected by percentiles varying from 95.0 till 99.9. The data were visualized with built-in functions from packages mentioned above and ggplot2 R package [12]. To automate the data processing, the custom scripts were written in Python 3.10 [13], R 4.2.1 [14], and GNU Bash 5.1.16 [15].

Results: The features sorted in descending way by feature impurity importance and SHAP values revealed the expected monotonic decrease in metrics with the increase in the rank for all datasets. However, the shape of the curves depended on the sample size and the type of metrics. The model fitted on unbalanced data did not predict the control samples (minor class). In the case of balanced datasets, the mean accuracies were around 0.5. Nevertheless, it allowed identifying the disease loci. However, the sensitivity estimated with 99 percentiles of SHAP values were 0.12, 0.27, and 0.60 for small, unbalanced and large datasets. The lists of selected loci for three datasets had less than 2.4 % of common items in pairwise comparisons. The distribution of neutral and disease loci in importance vs. SHAP space for three datasets (Fig. . **1**) demonstrated that it is not almost possible to subset the disease ones in the case of small dataset because of strong mixing them with the neutral ones. In the case of unbalanced dataset, the separation of a small portion of disease loci is possible. In the case of large dataset, an essential part of disease loci can be potentially separated, once the threshold values of metrics are estimated. The similar results were observed in the case of LD aware datasets.

Conclusion: The size and ratio of classes are crucial for the application of ML-based approach in associative studies. Its effective usage with ensemble algorithms

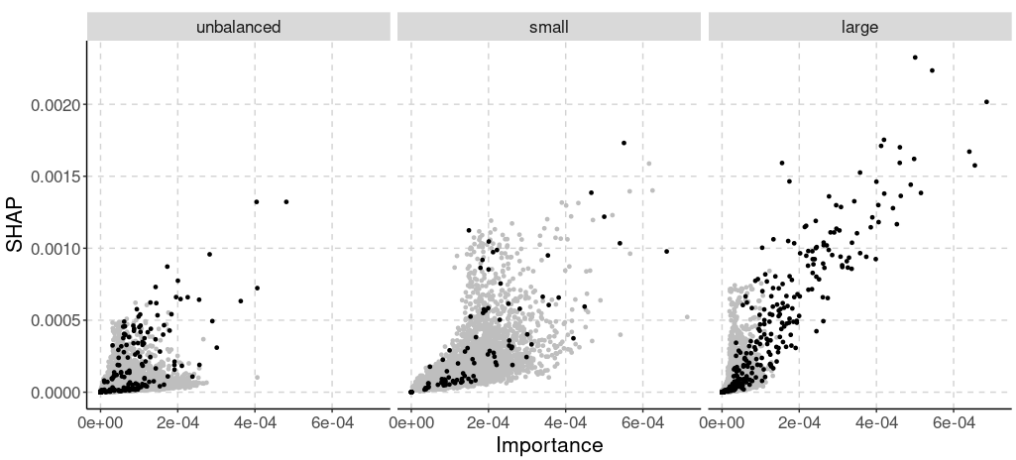

Fig. 1. The distribution of neutral (gray) and disease (black) loci in importance vs. SHAP space for three datasets without LD

(RF, XGBoost) would likely require balanced datasets with a total size of about 10000 individuals or more.

Funding: This research was supported by the Thematic plan of the National Research Centre "Kurchatov Institute" (5f.5.9., simulation of synthetic datasets) and by the Russian Science Foundation, grant number 23-14-00131 (fitting ML models and assessment of approach performance).

References

1. Visscher P.M. Challenges in understanding common disease. *Genome Med*. 2017;9:112. doi 10.1186/s13073-017-0506-1
2. Nicholls H.L., John C.R., Watson D.S., Munroe P.B., Barnes M.R., Cabrera C.P. Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. *Front Genet.* 2020;11:350. doi 10.3389/fgene.2020.00350
3. Lundberg S.M., Lee S.-I. A Unified approach to interpreting model predictions. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. 2017. [http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf]
4. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32. doi 10.1023/a:1010933404324
5. Chen T., Guestrin C. XGBoost: A Scalable Tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016. doi 10.1145/2939672.2939785
6. Chang C.C., Chow C.C. et al. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7. doi 10.1186/s13742-015-0047-8
7. Auton A., Abecasis G.R., Altshuler D.M., Durbin R.M., Bentley D.R. et al. A global reference for human genetic variation. *Nature*. 2015;526:68-74. doi 10.1038/nature15393
8. Su Z., Marchini J., Donnelly P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*. 2011;27:2304-2305. doi 10.1093/bioinformatics/btr341
9. Khvorykh G.V., Sapozhnikov N.A., Limborska S.A., Khrunin A.V. Evaluation of density-based spatial clustering for identifying genomic loci associated with ischemic stroke in genome-wide data. *Int J Mol Sci*. 2023;24:15355. doi 10.3390/ijms242015355
10. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O. et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830
11. Lundberg S.M., Erion G., Chen H., DeGrave A., Prutkin J.M., Nair B. et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2:56-67. doi 10.1038/s42256-019-0138-9
12. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag, 2016
13. Van Rossum G., Drake F.L. Python 3 reference manual. Scotts Valley, CA: CreateSpace, 2009
14. R core team. R: a language and environment for statistical computing. 2021
15. GNU Bash (5.1) [Unix shell program]. 2020. https://www.gnu.org/software/bash/