Variable Ranking by Random Forests Model for Genome-Wide Association Study

Atsushi Kawaguchi

***Abstract*—An important step in the genome-wide association study (GWAS) is the ranking of single nucleotide polymorphisms (SNPs). We propose a method based on the variable importance measure from the random forests model. SNPs in the entire genome region are randomly divided into subsets. We then fit the random forests model to each subset to compute subranks for the SNPs. The ranks of the SNPs are defined based on these subranks and then iteratively improved. We study the impact of the parameters and show that our method performs well in comparison to popular existing methods. We apply our method to select SNPs in a real-data study of the link between SNPs and human fingerprint ridge counts.**

***Index Terms*—Ensemble Trees, Genome-Wide Association Study, Screening, SNP Ranking, Statistical Interaction .**

I. Introduction

THE genome-wide association study (GWAS) is an attempt to unravel the genetic basis of complex genetic diseases. Specifically, GWAS uses single nucleotide polymorphisms (SNPs) as genetic markers because they are easy to type and abundant in the human genome. The goal of GWAS is to search for genetic factors that influence common complex traits and to characterize the effects of those factors. [1] gave an overview of statistical approaches to GWAS. One of the steps in the analysis is to select informative SNPs, usually based on certain association measures between the SNPs and a phenotype. Incorporating screened SNPs based on a preliminary ranking would be useful for the selection and could be faster than an exhaustive search. Most studies have used a univariate analysis to rank SNPs; however, the interactions between SNPs should be considered [2]. Traditional regression methods can deal with interactions but require an exhaustive search, so it is difficult to analyze higher-order interactions. Recursive partitioning approaches have been used as an alternative to traditional regression methods to detect the genetic loci and their interactions that influence the aphenotypic outcome. For high-dimensional structures such as GWAS, random forests (RF, [3]) offer substantial improvements in accuracy over the use of a single tree. [4] and [5] provide reviews of the application of RF to GWAS. RF also provides a natural approach for handling collinearity among SNPs and the nonlinear effect of SNPs. The measure of variable importance resulting from the application of the approach provides a general measure of the contribution of each potential predictor variable to the observed variability in the trait under investigation. Therefore, we propose a ranking method for SNPs based on RF that can subsequently be used for the screening and selection. In the preliminary ranking of SNPs, the target is on the entire genome region, which has a few million SNPs. A direct fit of RF would therefore pose a large computational burden. We propose randomly splitting the SNPs into subsets and fitting the RF model to each subset. The subranks for the SNPs are computed based on the variable importance from each fitted RF. The rank for the entire genome region is then defined based on these subranks. Moreover, to ensure that the final ranking is independent of the splitting process, we introduce an iterative process that uses different random splits of the SNPs. The numerical results indicate that our procedure improves the ranking of SNPs; see Section 3 for a detailed discussion of the results. The remainder of the paper is organized as follows. The proposed algorithm is described in Section II. In Section III, an extensive simulation study is presented. Section IV discusses an application of our method to real data pertaining to human fingerprint ridge counts. Concluding remarks are given in Section V.

II. Method

Suppose we have a sample consisting of a continuous phenotype and SNPs. Let Y_i denote the phenotype of the i-th individual $(i = 1, \ldots, n)$ and $Y = (Y_1, \ldots, Y_n)$. Let $\boldsymbol{X}_i$ denote the observed p-dimensional genotypic data (SNPs) for the i-th individual, and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. We consider a regression problem, that is, for individual i, let X_i represent the vector of predictor variable values and Y_i its response. In this section, we first give a brief review of random forests (RF) and then describe our method.

A. Random Forests

Generating the RF involves generating a set of classification or regression trees. A summary of the RF-generation algorithm is given below. In the first step, we randomly select approximately two-thirds of our sample. This is called the learning sample (LS) since it is used to grow a tree. The remaining individuals constitute the out-of-bag (OOB) data, and they are used to evaluate how well the tree applies to data that were not used to generate it. In the second step of the algorithm, we grow a tree based on the LS data. Instead of determining the best split among all potential predictors, we choose a random sample of these variables (one-third of the variables) to consider as potential splitting variables. Note that, in the context of RFs, no prune is implemented. Finally, we repeat these steps B times. The resulting output is the B trees. We do not generate a final tree that can be interpreted as a model for the genotype-trait association. Instead, we use the measures of variable importance described in the next section.

B. Variable Importance

We use measures of variable importance to determine which SNPs or sets of SNPs are important in the prediction.

A. Kawaguchi is with Biostatistics Center, Kurume University, 67 Asahi-machi, Kurume, Fukuoka 830-0011 Japan.

We define the variable importance as follows. Let $V_j(\boldsymbol{X}_i)$ denote the average response in tree j and t_{ij} be an indicator with value 1 when individual i is OOB for tree j and 0 otherwise. Let $\boldsymbol{X}^{(A,j)} = (\boldsymbol{X}_1^{(A,j)}, ..., \boldsymbol{X}_n^{(A,j)})$ be the vector of predictor variables with the value of variable A randomly permuted among the OOB observations for tree j, and $\boldsymbol{X}^{(A)}$ the set of $\boldsymbol{X}^{(A,j)}$ for all trees where n is the total number of individuals in the sample. The variable importance is averaged over the trees of the forest; it is defined as:

$$VI(A) = \frac{1}{B}\sum_{j=1}^{B}\frac{1}{N_j}\sum_{i=1}^{n} t_{ij}\left\{V_j(\boldsymbol{X}_i) - V_j(\boldsymbol{X}_i^{(A,j)})\right\}^2$$

where N_j is the number of OOB individuals for tree j, and B is the number of trees.

C. Main Method

We now introduce our ranking method. We randomly split the SNPs into m subsets of roughly equal size. Call this set of SNPs G_j $(j = 1, 2, \cdots, m)$. We fit a random forests to each subset G_j, and compute the variable importance and the OOB error ($OOBerr_j$) for each fitted random forests model. We rank the SNPs according to the VI to give the subrank. Let $\tilde{R}_k$ denote the rank in the subset for the k-th SNP in the j-th subset. The entire rank for the k-th SNP is defined as the rank of each subset weighted by the OOB error from the fitted RF in the j-th subset:

$$R_k = w_j \tilde{R}_k$$

where $w_j = \exp(OOBerr_j / \min_\ell OOBerr_\ell)$. This process is repeated K times (we set K=10). The entire ranks for the individual SNPs found at each iteration are then averaged to give the final rank.

III. Simulation Study

We use a simulation study to analyze the impact of the parameters of our method and to provide a comparison with other ranking methods including the non-splitting (direct fit) method. We used SNPs from the International HapMap project [6].

A. Data

The HapMap genotype CEU samples (release 22) are available for download on the website (http://hapmap.ncbi.nlm.nih.gov/downloads/). Prior to inclusion, the individual samples and SNPs were subjected to a rigorous quality control. Individual SNPs were required to have a call rate $\geq$ 95%, to have a minor allele frequency $<$ 1%, and to be in Hardy-Weinberg equilibrium (p $<$ 0.001). Missing SNPs were imputed. The data were quality-controlled using PLINK v1.02 [7]. After this process, 60 samples and 2,275,723 SNPs remained.

To generate response (phenotype) values with a complex structure, we applied the following steps. First, generate random numbers following a normal distribution, say, temporal responses. Next, randomly select 50 samples from the 60 and fix these, and test the association between the temporal response and each individual SNP. Select the 5 SNPs with the strongest association and 10 neighborhoods for each, that is, the number of associated SNPs is 50. Now fit the RF to the data with the temporal response and the selected SNPs. Using the fitted model, compute the predicted values from the SNPs and treat these values as responses. The fitted RF model is regarded as true complex structure between the SNPs and the response. By repeating this process, we generated 10 simulation data sets.

TABLE I
The mean of the resulting ranking for 50 true SNPs (multiplied by 10^{-4}) for impact of parameters in simulation study

	fold size (nfold)				
ntree	500	1000	2000	3000	5000
50	74.2	70.0	69.2	65.5	68.5
100	73.1	68.6	67.4	**63.6**	66.9
150	68.2	67.3	66.8	64.5	67.0

B. Impact of Parameters

The parameters of our method are the fold size (nfold) and the number of trees (ntree) in the RF model. We evaluated these using the mean of the resulting ranking for 50 true SNPs; smaller means are better. If we had the true ranking, the mean would be $(1+2+\cdots+50)/50 = 25.5$. We test nfold = 500, 1000, 2000, 3000, and 5000 and ntree = 50, 100, and 150. Table I shows that the case where nfold = 3000 and ntree = 100 has the smallest mean. The means are considerably larger than the true value of 25.5 probably because our true contained some SNPs with a poor relationship to the response.

C. Comparison

We compare our method with the original RF (direct fit without splitting) with the ranking based on the variable importance (namely, SNPs with high variable importance are ranked highly). We also compare with a simple linear regression. As in Section III-B, we evaluated the results using the mean of the computed ranking for 50 true SNPs; smaller is better. In our method, we used the parameters identified in the previous section (nfold = 3000 and ntree = 100). For the original RF, the number of trees was set to 50 since we could not implement more with 8 GB of memory. Fig. 1 shows that our method gave the smallest mean rank. Our method with ntree = 50, as for the original RF, was also better than the other two methods.

IV. Real-Data Application

We now present an application of our method to real data. We used data from a study at the Department of Forensic Medicine and Human Genetics at Kurume University. It investigated the relationship between the human fingerprint ridge count and the genotype. We performed a quality control of the SNPs as described in Section III-A. There were 182 patients in this study, and 333,510 SNPs were used in the analysis. We ranked the SNPs using the three methods (proposed, original, and simple linear regression) compared in the previous section. We iteratively fitted RF models, at each iteration building a new forest after discarding the SNPs with the lowest ranks. We computed the OOB errors for each model. Fig. 2 shows that our method has the lowest number of OOB errors for all the sets of SNPs.

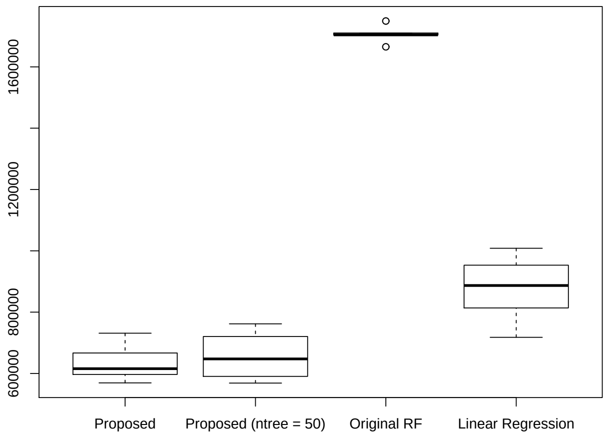

Fig. 1. Comparison of different methods via simulation study. Horizontal line represents mean rank for true SNPs.

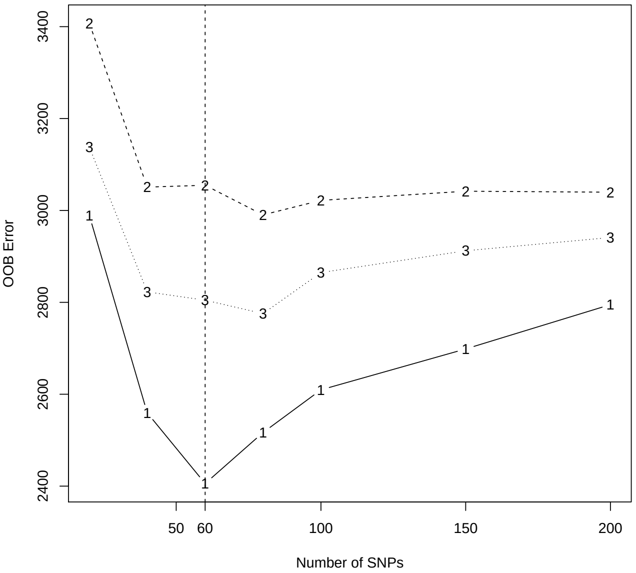

Fig. 2. OOB errors for each set of SNPs based on ranking from three methods (1 = proposed, 2 = original RF, 3 = simple linear regression). Vertical dashed line shows minimum from proposed method.

From the results for our method, we selected the 60 SNPs with the highest ranks and the smallest OOB errors and computed the variable importance from the fitted RF model. Fig. 3 shows that rs1612154 was the most important SNP in this analysis, followed by rs4130590 and rs2297602.

V. Discussion

We have proposed a new method for SNP ranking for the genome-wide association study. We have compared our method with conventional methods using simulated and real data. Our method performed well in both cases.

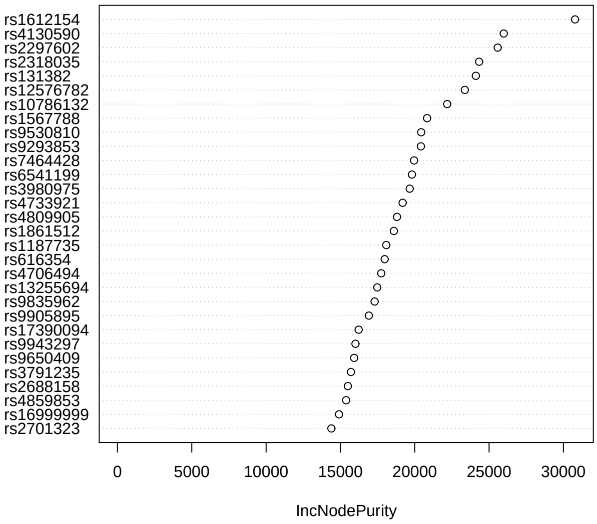

Fig. 3. Variable importance computed from RF model with 60 SNPs selected by proposed method.

The key features of our method are the division of the SNPs into subgroups and the iterative rank update. We required the response to be continuous. Of course, the method could be applied to a discrete variable such as a binary variable representing the presence or absence of disease. The size of the subgroup and the number of iterations should be determined before the algorithm is applied. In our simulation, a size of 3000 was optimal. On the other hand, the number of iterations was set to 10. More iterations would improve the performance, but we believe that the improvement would be small. In the future, the appropriate number of iterations should be studied.

Ranking SNPs with high accuracy helps us to select predictive SNPs, as shown in our real-data analysis. In conclusion, our ranking method will help to construct a new association scheme that better assesses the relationship between genotype and phenotype.

Acknowledgment

The author appreciates the constructive comments of anonymous referees. The author also appreciates Dr. Yoshiro Koda for the data and his support. This study was partly supported by a grant from the Kurume University Millennium Box Foundation for the Promotion of Science.

References

[1] D. J. Balding, “A tutorial on statistical methods for population association studies,” *Nature Reviews Genetics*, vol. 7, pp. 781–791, 2006.

[2] H. J. Cordell, “Detecting gene-gene interactions that underlie human diseases,” *Nature Reviews Genetics*, vol. 10, pp. 392–404, 2009.

[3] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.

[4] B. A. Goldstein, E. Copley, and F. B. S. Briggs, “Random forests for genetic association studies,” *Statistical Applications in Genetics and Molecular Biology*, vol. 10, no. 1, p. Article 32, 2011.

[5] Y. V. Sun, “Multigenic modeling of complex disease by random forests,” *Advances in Genetics*, vol. 72, no. 10, pp. 73–99, 2010.

[6] International HapMap Consortium, “A second-generation human haplotype map of over 3.1 million snps,” *Nature*, vol. 449, pp. 851–862, 2007.

[7] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. De Bakker, M. J. Daly, and et al., “Plink: A tool set for whole-genome association and population-based linkage analyses,” *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.