**Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006**

A Fast Boosting-Based Screening Method for Large-scale Association Study in Complex Traits with Genetic Heterogeneity

Lu-yong Wang*, Daniel Fasulo

Abstract—**Genome-wide association study for complex diseases will generate massive amount of single nucleotide polymorphisms (SNPs) data. Univariate statistical test (i.e. Fisher exact test) was used to single out non-associated SNPs. However, the disease-susceptible SNPs may have little marginal effects in population and are unlikely to retain after the univariate tests. Also, model-based methods are impractical for large-scale dataset. Moreover, genetic heterogeneity makes the traditional methods harder to identify the genetic causes of diseases. A more recent random forest method provides a more robust method for screening the SNPs in thousands scale. However, for more large-scale data, i.e., Affymetrix Human Mapping 100K GeneChip data, a faster screening method is required to screening SNPs in whole-genome large scale association analysis with genetic heterogeneity. We propose a boosting–based method for rapid screening in large-scale analysis of complex traits in the presence of genetic heterogeneity. It provides a relatively fast and fairly good tool for screening and limiting the candidate SNPs for further more complex computational modeling task.**

I. Introduction

Though there is great success in identifying the genetic causes of diseases, especially the monogenic Mendelian traits, the nature of common and complex diseases is still mysterious. Complex phenotypes are not typically the results of variation of a single genetic locus, but are the result of interplays among contributions of different genetic and environmental factors. Genome-wide association studies for complex diseases are generating tremendous amount of single nucleotide polymorphism (SNP) data. It is crucial and challenging to weed out the noise and identify the SNPs contributing to the complex traits.

A logical first approach is to conduct univariate association tests on each individual SNP. The goal is to prioritize the SNPs of variants that influence the disease susceptibility, rather than to prove that a particular set of SNPs that influence the diseases. However, using univariate association test will result in low power for SNPs with very small marginal effects in the population. Many model-based methods[1], *i.e.*, multivariate adaptive regression splines models, have also been investigated in genetic linkage analysis and association studies [2]. However, these model selection methods are limited in the number of predictors that can be included in one analysis, and the types of interactions must be specified in advance. These methods are not suitable for initial task of identifying a subset from a massive set of SNPs for further analysis.

In another problem of multi-locus interaction detection, combinatorial partitioning methods (CPM) [3] were developed to detect interactions among multiple loci to predict quantitative trait variation. Multifactor-dimensionality reduction method (MDR) seeks to reduce the dimensionality of multi-locus genotype space to facilitate the identification of gene-gene interactions [4]. It has been successfully applied to breast cancer case-control dataset. Recently, Bayesian network was also introduced into this mining the multi-locus interaction in complex traits [5], and it is applied on plasma apoE level genetic epidemiology data.

However, these methods are limited in the number of predictors, causing researchers still to resort to a two-stage approach. Univariate association screening only considers main effects at first, and the effects between loci with main effects are considered in the second stage. This approach could lead to the loss of important loci with only weak effects. A more accurate screening method are required for the further modeling works. More importantly, traditional methods have not manage to address genetic heterogeneity problem and not applicable for large-scale analysis [4].

Genetic heterogeneity is an important concern in genetic epidemiology, which means there are multiple disease-leading pathways involving different subsets of genes. One of possible reasons for genetic heterogeneity is the ethnic background among the population. A subgroup dividing or population clustering based on genetic profiles may be a powerful tool in analyzing the genetic causes of specific trait across the population. However, in most of the cases, genetic heterogeneity is seldom so simple. As prior knowledge regarding the cause of genetic heterogeneity is rarely known in the study, traditional methods based on estimation over the entire population, are unlikely to succeed in tackling the genetic causes of disease. Wang and Fasulo [6] explored the boosted generative modeling method to model the delicate relationship of genetic interaction network to address the genetic heterogeneity problems. However, these approaches are not intended to for large scale association analysis due to computational complexity. The problems remains: how can we limit from thousands of SNPs to a reasonable number that can be used for available modeling methods, especially in the presence of genetic heterogeneity.

Lunetta, *et al.*, first introduced random forest (RF) as a

Dr. Lu-yong, Wang, is with Integrated Data Systems Department, Siemens Corporate Research, Princeton, NJ, 08540 (phone: 609-734-3671; fax: 609-734-6565; email: luyong.wang@siemens.com).

1-4244-0033-3/06/$20.00 ©2006 IEEE.

screening tool for identifying SNPs associated with a disease with genetic heterogeneity[7]. A RF is a collection of classification trees grown on different bootstrap samples of the observations, using a random subset of predictors to define the best split at each node. Each tree is grown on the bootstrap samples of the observations. The bootstrap sample has the same number of individuals as the original sample, but some individuals are represented multiple times, while others are left out. The left-out ("out-of-bag") individuals are used to estimate prediction error. With a forest of resulting classification trees, each tree predicts the class of an individual. The predictions are counted across all trees for which the individual was out-of-bag. The class with the most votes is the individual's predicted class.

RF generates an importance score for each variable, which is quantified by the increase in classification error occurring when the values of the predictors are randomly permuted. This score can prioritize the variables. However, the computational complexity of RF method brings problem. It takes average 123 minutes for each replicate of dataset with 1K SNPs to complete analysis [7]. Considering the whole genome association analysis, *i.e.*, Affymetrix Human Mapping 100K GeneChip, it is impractical in certain situation which requires a quick answer.

To address this efficiency problem, we propose a fast screening method, which can perform well and fast enough for the rapid screening of the candidate SNPs in complex disease research.

II. Methods

A. Boosting

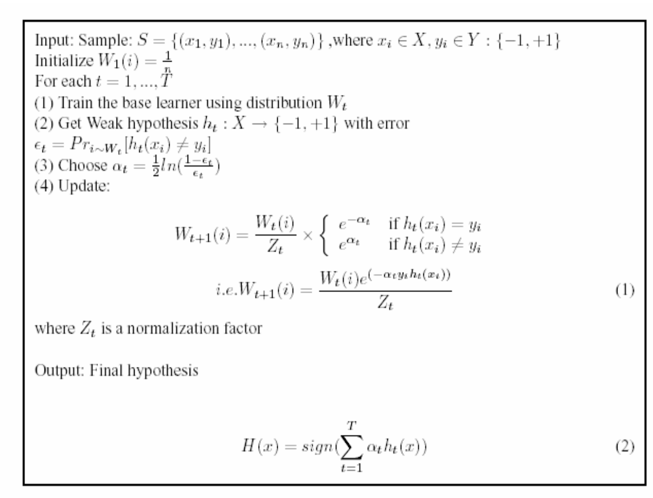
We introduce our method in a logical order of boosting and boosted decision stump variable selection (BDSVS) for clarity. Boosting is a method for improving the accuracy of any given learning algorithm [8]. AdaBoost repeatedly calls a given base learning algorithm in *t* rounds (Figure 1). $W_t(i)$ represents the weight of the distribution on training example i on round t. At each iteration t, the base learner is utilized to find a weak hypothesis h_t: $X \rightarrow \{-1, +1\}$ appropriate for the distribution. The weights will be updated. Usually, the weights of incorrectly classified examples are increased so that the base learner is forced to concentrate on the hard examples in the training set. The base learner is called again with new weights over the training examples, and the process repeats. Finally, all the weak hypotheses are combined into a single, strong hypothesis using a weighted majority vote.

B. Boosted Variable Selection

Decision stump is used as a "weak" learner, which is a decision tree with only one split. AdaBoost algorithm can combine those decision stumps into an accurate classifier. The combined classifier is a committee with decision stumps. The committee makes a decision by a weighted majority vote. The base classifiers are constructed on weighted examples. For the 1st round, set all samples' weights equal. For the next round, the weights are increased for the samples misclassified by the first decision stump and decreased for the samples correctly classified by the first decision stump. Therefore, the second decision stump focuses on the samples misclassified by the first stump. This procedure is repeated until a defined number of stumps have been created. Decision stump uses the information gain criteria for decision stump which variable to choose. The boosting algorithm is run for T rounds, which is total number of variables selected. At each round, it looks for the previously unselected variables with the highest information gain on the weighted distribution of training examples. Boosted decision stumps have many desirable properties as a method of variable selection. Boosting assign higher weights to examples misclassified in the previous rounds. A variable that correctly predicts the class label of examples that previously predicted wrong will have a higher information gain. This will eventually help guide the search for variables that are highly a predictive of a small regions of instance space get selected, because if that part of the instance space get selected, because if that part of the space is often misclassified by other variables, those example will keep increasing in weight. This heuristic is not theoretical optimal, but is proven to perform well on empirical studies[8].

C. Genetic models and Data Simulation

The complex diseases are simulated with sibling recurrence risk ratio for the disease (λ_s) fixed at 2.0 and population disease prevalence K_p at 0.10. These parameters are consistent with or lower than the estimate from known complex genetic traits, such as Alzheimer's disease, which are caused by multiple genetic and environmental factors. These complex traits are estimated of cumulative prevalence in siblings of affected range from 30%-40% compared to a population prevalence of 10% at age of 80. The genetic models include both genetic heterogeneity and multiplicative interaction as defined in [9]. The sets of 4, 8, 16, 32 risk SNPs ("rSNP") are simulated in linkage equilibrium. Each set is composed of disease-susceptible independent pairs or quartet to increase disease risk. For simplicity, the models are simulated such that each rSNP pair or quartet accounts for the same proportion of the genetic risk and each SNP within a pair of a pair or quartet contribute equally to the genetic risk. Thus, all the rSNPs simulated for a model have the same allele frequency and the same observed marginal effect in the population. The models are represented in the following format: HhMm, where h refers to the number of heterogeneous systems and m refers to the number of disease-susceptible SNPs within each system. The models in our analysis are H4M2, H8M2, H4M4, H8M4. Besides rSNPs contributing to the diseases, noise SNPs independent of the disease status are also simulated with allele



Input: Sample: $S = \{(x_1, y_1), ..., (x_n, y_n)\}$,where $x_i \in X, y_i \in Y : \{-1, +1\}$
Initialize $W_1(i) = \frac{1}{n}$
For each $t = 1, ..., T$
(1) Train the base learner using distribution W_t
(2) Get Weak hypothesis $h_t : X \rightarrow \{-1, +1\}$ with error
$\epsilon_t = Pr_{i \sim W_t}[h_t(x_i) \neq y_i]$
(3) Choose $\alpha_t = \frac{1}{2}ln(\frac{1-\epsilon_t}{\epsilon_t})$
(4) Update:

$$W_{t+1}(i) = \frac{W_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases}$$

$$i.e. W_{t+1}(i) = \frac{W_t(i)e^{(-\alpha_t y_i h_t(x_i))}}{Z_t} \qquad (1)$$

where Z_t is a normalization factor

Output: Final hypothesis

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \qquad (2)$$

Fig. 1. AdaBoost Algorithm

frequencies distributed equally across the range of 0.01-0.99. 500 cases and 500 controls are simulated for each of 100 replicate dataset. Moreover, in a real-world association study, we are unlikely to genotype all the polymorphisms related to the disease, only a subset of the total number of rSNPs are included in each analysis. The design is represented as KkSsNn, where k refers to the total number of rSNPs genotyped, s refers to the number of rSNPs within each system genotyped, and n is the total number of SNPs genotyped. For example, K4S2N100 in the design of H8M2 means that 4 rSNPs (2 rSNPs from each of 2 heterogeneity systems) are genotyped out of the total 16 rSNPs related to trait; 96 nSNPs are also genotyped additionally; there are total 100 SNPs genotyped for each individual.

III. Results and Discussions

A. Empirical evaluation of the screening performance on simulated complex disease data with genetic heterogeneity

Our BDSVS method, RF and univariate analysis (Fisher exact test) were carried out on the simulation data described in the Methods section. Our methods take the most significant variables (SNPs) sequentially, while RF takes its standardized Z_T variable importance scores, and Fisher exact test takes its p-value to rank SNPs. All the set of simulation data in Figure 2 and 3 are based on K4S2N100 and K4S2N1000 design, where 2 SNPs from each of the first 2 heterogeneity systems are included in the analysis. Figure 2 and Figure 3 show the percentage of replicates for which all rSNPs are among the top-ranking N SNPs for K4S2N100 and K4S2N100 analysis design.

For K4S2N100 design in Figure 2, both our BDSVS method and random forest method have a higher proportion of replicates among the top ranked SNPs than Fisher exact test. Thus, for a given probability of retaining all of the rSNPs, more SNPs can be eliminated using either our BDSVS method or the standardized Z_T variable importance scores of RF than the Fisher exact test p-value. For example, for the model H16M2, only 16 SNPs need to retain to have an 80% probability that the 4rSNPs are in the retained set for our BDSVS (15 SNPs for random forest), while 44 SNPs need to retain if the Fisher exact test p-value is used. In another words, for any number of retained SNPs, the probability that all the genotyped rSNPs is higher for Random forest and our BDSVS methods than that for the univariate Fisher exact test p-values. Our BDSVS method perform similarly with RF in H16M2N100 and H8M4N100 (some situations better than Random forest and some *vise vesa*), while in H8M2N100 and H4M4N100, although RF behaves slightly better than BDSVS method in the expense of computation complexity discussed in next section.

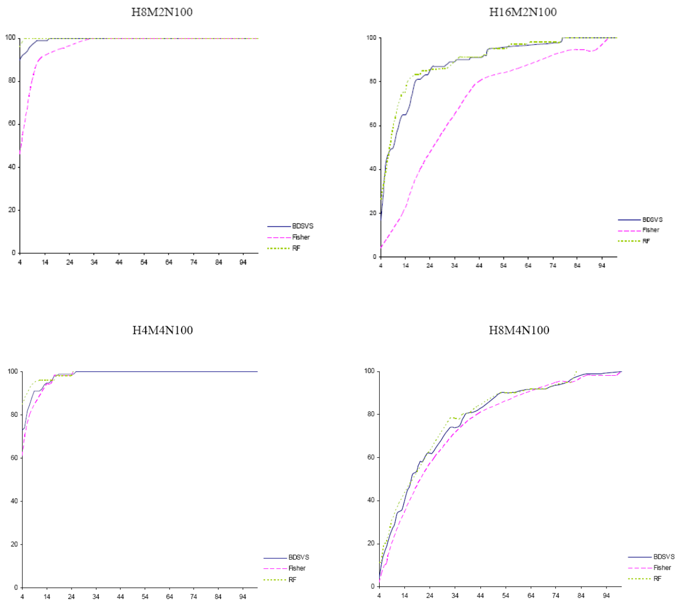


Figure 2 Proportion of the replicates for which all the rSNPs are among the top-ranking N SNPs for K4S2N100 analysis.

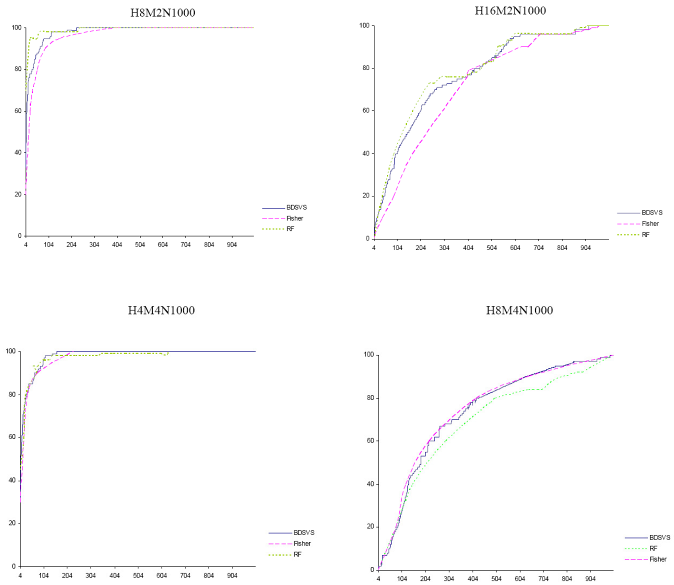


Figure 3 Proportion of the replicates for which all the rSNPs are among the top-ranking N SNPs for K4S2N1000 analysis.

For K4S2N1000 analysis design in Figure 3, our BDSVS gives better results than the Fisher exact test p-value criterion in H8M2N1000 and H16M2N1000, and H4M4N1000, where Random forest method gives also better result than the Fisher exact test p-value criterion. Since all the dataset were simulated according to genetic heterogeneity, the subpopulation strategy by both the boosting (weighted instance space) and random forest (bootstrap samples) will improve the detection of disease susceptible SNPs.

In H8M4N1000, BDSVS method gives similar performance as univariate Fisher exact test analysis, where Random Forest has been shown worse performance than these two methods. The advantage of BDSVS over Fisher exact test diminishes in this particular example, while RF performs worse than other 2 methods. It can be explained that the poor experimental sampling under over-complex situation (H8M4N1000) leading to the signal buried by noises. Ensemble techniques can not improve in this case.

Overall, our method of BDSVS method perform better than univariate Fisher exact test in most of the cases, while RF performs also better than univariate Fisher exact test except H8M4N1000. Its subpopulation strategy achieved by re-weighting inherent in BDSVS provides a good method to solve genetic heterogeneity problem among the population. Though in a few cases, RF performs slightly better than BDSVS method, it is balanced by RF's heavy computational complexity discussed later. The trade-off between these 2 methods depends on the requirements. It indicates that our BDSVS provide a good alternative method to the screening task of candidate loci in the presence genetic heterogeneity.

B. Computational Complexity Comparison

In many scenarios, a rapid method with good performance is required, while speed instead of accuracy is of the first priority. For example, we need a fast method that can display promptly the rudimentary results without computing for days or weeks.

We empirically compare the time complexity of the two methods: BDSVS and random forest. As expected, random forest is much more computationally expensive than our BDSVS method. On average, each 500 cases vs. 500 controls replicate dataset with total 100 SNPs took 40 minutes to complete on a 2.6 Ghz Intel Xeon Processor in the experiments. For a similar replicate of the dataset with 1K SNPs, it took more than 123 minutes on average on a 2.6Ghz Intel Xeon workstation. It will take days ~ weeks to finish the analysis on a whole-genome genetic association dataset, such as 100K SNPs, which makes it infeasible to carry out according to prompt requirement. On contrast, our BDSVS method takes only a second on 1K SNP dataset to complete the analysis on a 1.5GHz Pentium 4 PC, while the 100K SNP data can be estimated within minute.

Methods	100 SNPs	1000 SNPs	100K SNPs	Machine Used
BDSVS	< 1 sec	~sec	~mins	1.5GHz Intel Pentium IV PC
Random Forest	~40 mins	~123 mins	Days~ weeks	2.6GHz Intel Xeon Workstation

Table 1. Computational Complexity Comparison

IV. Conclusion

Genome-wide association study for complex diseases is ongoing and producing enormous SNPs data simultaneously in the experiments. Univariate statistical association test was used to first screen out non-associated SNPs, retaining only those meeting some criterion. Fisher exact test can provide p-value measurement for each SNP and it can be used as a standard for screening. However, the univariate tests may miss the disease susceptible SNPs, which may have small marginal effects in population. Moreover, genetic heterogeneity makes the problem harder. A recent screening method based RF provides a more robust, while computationally expensive method for screening the SNPs in thousands scale. However, in genome-wide large-scale association analysis, a faster screening method is required.

In this paper, we propose a fast screening method based boosted variable selection. This method is a quick screening alternative, which can also perform well and fast enough for the rapid identification and screening of the candidate SNPs and provide an assistant tool for further delicate modeling task. It uses boosting process to solve the genetic heterogeneity problem, which is inherent in many complex traits. It enables fast detection of the SNPs with little marginal effects. More importantly, its computational efficiency makes it a good candidate for screening large-scale association study as a speedy and excellent alternative to computationally expensive random forest screening.

References

[1] C. Oh, K. Q. Ye, Q. He, and N. R. Mendel, "Locating disease genes using Bayesian variable selection with the Haseman-Elston method," *BMC Genet*, vol. 4, pp. S69, 2003.

[2] T. P. York and L. J. Eaves, "Common disease analysis using Multivariate Adaptive Regression Splines (MARS): Genetic Analysis Workshop 12 simulated sequence data," *Genet Epidemiol*, vol. 21 Suppl 1, pp. S649 - 54, 2001.

[3] M. R. Nelson, S. L. Kardia, R. E. Ferrell, and C. F. Sing, "A combinatorial partitioning method to identify multi-locus genotypic partitions that predict the quantitive trait variation," *Genome Res*, vol. 11, pp. 2115-2119, 2001.

[4] M. Ritchie, L. Hahn, and J. Moore, "Power of multifactor dimensionality reduction in detecting gene-gene interactions in the presence of genotypic error, missing data, phenocopy, and genetic heterogeneity," *Genetic Epidemiol*, vol. 24, pp. 150-157, 2003.

[5] A. Rodin and E. Boerwinkle, "Mining genetic epidermiology data with Bayesian networks I: Bayesian networks and example application(plasma apoE levels)," *Bioinformatics*, vol. 21, pp. 3273-3278, 2005.

[6] L.-y. Wang and D. Fasulo, "Exploiting Gene-gene Interactions Contributing to Complex Disease Traits with Boosted Generative Modeling," *Manuscript in preparation*, 2006.

[7] K. Lunetta, L. Hayward, J. Segal, and P. Eerdewegh, "Screening large-scale association study data: exploiting interactions using random forests," *BMC genetics*, vol. 5, pp. 32, 2004.

[8] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," *Proceedings of the Thirteenth International Conference in Machine Learning*, pp. 148 - 156, 1996.

[9] N. Risch, "Linkage strategies for genetically complex traits. II. The power of affected relative pairs," *Am J Hum Genet*, vol. 46, pp. 229 - 241, 1990.