A gene-based permuted extreme gradient boost method for detecting gene-gene interactions of qualitative trait and application to XXX data

Yingjie GUO, Chenxi WU, Ao li, Junwei Zhang, (someone for real data analysis), Alon Keinan, Maozu GUO

Abstract: Boosted Tree is a popular and highly effective method in machine learning for modeling additive models with non-linear terms. In this paper, we propose a novel gene-based, permuted, extreme gradient boosting method called gpXGB to detect interactions between genes in qualitative traits, which has advantage in both statistical power and biological interpretability. The main idea is to permute the genotype within each class of the dataset in two ways, one keep the interaction between genes and another remove such interactions, then rank the AUC differences of the result of XGB after these two different types of permutation.

1. Introduction

Genome-wide association studies (GWAS) have identified over six thousand single-nucleotide polymorphisms (SNPs) associated with complex diseases or traits. Earlier GWAS analysis strategies were largely based on single locus models, which test the association between individual markers and a given phenotype independently. Although this type of approaches have successfully identified many regions of disease susceptibility, most of these SNPs identified have small effect sizes which failed to fully account for the heritability of complex traits. Genetic interaction has been hypothesized to play an important role in the genetic basis of complex diseases and traits, and to be one of the possible solutions to this problem of “missing heritability”. Even if genetic interaction explains only a fraction of “missing heritability”, they can still provide some biological insights on the pathway level by aiding the construction of novel gene pathway topologies.

The first investigations on genetic interactions have been at the SNP level, in which various statistical methods, including logic and logistic regressions, odds-ratio, linkage disequilibrium (LD) and entropy-based statistic, are employed to detect SNP-SNP interactions (i.e. epistasis). Other techniques that have been used to study SNP-SNP interactions include multifactor dimensionality reduction, Tuning RelieF, Random Jungle, BEAM, BOOST(Wan, Yang et al. 2010) and pRF(Li, Malley et al. 2016). These marker-based methods may encounter some common challenges, such as the complexity arising from the large number of pairwise or higher-order tests because all pairs or groups of SNPs have to be considered; the extensive burden of multiple-testing correction they entail. (What is this?) In this paper, we aim to improve the power of gene-gene interaction detection by moving beyond SNP level (is it true that gene level testing have greater statistical power than SNP level?), and instead consider all potential pairs of SNPs from each of a pair of genes in a single gene-based interaction detection.

Gene-based approaches have been successful for regular GWAS tests of main (marginal) associations, and there are several potential advantages in extending this methodology to gene-gene interaction detections. Firstly, a gene-based approach can substantially reduce the number of tests needed. For example for 20,000 genes, there are ~2×10^8 possible pairwise gene-based interactions to be tested, while for 3 million SNPs there are over ~5×10^{12} possible marker-based interactions to be tested. Secondly, a gene-based interaction test may have greater power, because when there are multiple interactions between features in the targeted genes (or other kind of regions), the effect of these interactions may be aggregated by the algorithm. Such aggregration has already been seen in gene-based GWAS tests for main association effect. Thirdly, a gene-based approach may be better at leveraging prior biological knowledge, which is often on the level of genes. For example, one may test pairs of genes that exhibit protein-protein interactions (PPI) or that participate in the same pathways.

In the work of Peng et al. (Peng, Zhao et al. 2010), canonical correlation analysis between two genes is done on both the case and the control group, and a U-statistic, called CCU, is used to measure the difference of the correlation between these two genes, which is used to indicate the presence of interaction. A limitation of this method is that in the correlation analysis only linear relations are considered. To overcome this limitation, (Yuan, Gao et al. 2012, Larson, Jenkins et al. 2014) extended CCU to KCCU, where the canonical correlation analysis is kernelized to account for possible non-linearity. Li et. al. (Li, Huang et al. 2015) introduced another method called GBIGM which is entropy-based and non-parametric. More recently, Emily (Emily 2016) developed a new method called AGGrGATOr which combines the p-values in marker-level interaction tests to measure the interaction between two genes. Earlier (Ma, Clark et al. 2013) this strategy was successfully used for the interaction detection for quantitative phenotypes.

In this paper, rather than designing a new dedicated statistic, we use a machine learning algorithm extreme gradient boosting (Xgboost) (Chen and Guestrin 2016) to propose a new approach, called gene-based permuted extreme gradient boost (gpXGB), to detect gene-gene interactions. The idea is to compare the performance of Xgboost on two different test datasets obtained from different permutation strategies, one keeping while another removing the interactions between selected gene pairs. An advantage of our new approach is that it is nonparametric, hence may be more flexible for data-driven exploratory genome-wideassociation studies.

2. Methods

In this section we first detail our gpXGB approach. We then describe the various simulation studies we conducted to assess the type-I error rate as well as the statistical power of our approach in gene-gene interaction detection. Finally, we apply our approach to the NESDA () dataset to evaluate the capability of our approach in a real-life situation.

2.1 Overview of gpXGB

Our method, gpXGB, is a machine learning based procedure for detecting the interaction between two genes in susceptibility with a binary phenotype, typically a case/control disease status. Let $y \in \{0, 1\}$ be the phenotype, where $y = 0$ stands for membership of the control group and $y = 1$ for membership of the case group. Let n be the number of instances in our sample, $\mathbf{Y} = \{y_1, \dots y_n\}$ be the vector consisting of their observed binary phenotypes. Let X_g, where $g = 1, \dots, G$, be the G genes in our gene list, each a collection of m_g SNP markers. The observed genotypes for gene X_g can be represented by an $n \times m_g$ matrix $\mathbf{X}_g = [x_{i,j}]_{1 \le i \le n, 1 \le j \le m_g}$ where $x_{i,j} \in \{0, 1, 2\}$ is the number of copies of the minor allele for SNP j carried by individual i. Let $\mathbf{X}_g^D$ and $\mathbf{X}_g^C$ be the matrices whose columns are those columns of $\mathbf{X}_g$ corresponding to samples in the case and control group, respectively. The genotype values in these matrices may also be adjusted to account for various covariates and population stratification.

We choose Xgboost as our classifier because gradient boosting decision tree (GBDT) is an effective and relatively model-agnostic way to approximate true target function which may have non-linear structure, and Xgboost is an implementation for GBDT known for its precision and computational efficiency.

Our approach consists of three steps: 1) training 2) permutating 3) testing and ranking. We start by training an Xgboost model with all the genes in the gene list and use cross-validation to choose a best model and save it. Then, for each selected pair of genes, we use two permutation strategies to generate two different test datasets, one keep while the other remove the interaction between the selected pair of genes. Lastly, we calculate the performance difference ΔAUC for the two test datasets on the well-trained model, which we use as a measurement for the strength of interaction between these two genes. The various steps of the gpXGB framework are illustrated in Figure1.

2.2 Overview of XGBoost

XGBoost (Chen and Guestrin 2016) is a scalable supervised machine learning system based on tree boosting, and recently has been dominating in applied machine learning as well as in Kaggle competitions. It is an advanced implementation of gradient boosted decision trees (GBDT) with speed and performance improvement.

2.2.1 Ensemble of CARTs

In this ensemble model, the base classifier is the CART (Classifying And Regression Tree), which is similar to decision trees, but on each leaf, instead of a classification, a real-valued score is assigned. This makes ensemble training easier and may also provide more information beyond classification.

Let $\mathcal{F}$ be the space of functions that can be represented by CARTs, the ensemble predictor is $\hat{y} = \sum_k f_k, f_k \in \mathcal{F}$. In our case we interpret it as in logistic regression, namely

$$p(y = 1|x) = \frac{1}{1 + e^{-\hat{y}(x)}} \tag{ 1}$$

Hence, the learning objective is

$$obj = \sum_i l(y_i, \hat{y}(x_i)) + \sum_k \Omega(f_k) \tag{2}$$

where $l(y, \hat{(y)}) = y \log(1 + e^{-\hat{y}}) + (1 - y) \log(1 + e^{\hat{y}})$ is the logistic regression loss function, and $\Omega(f_k)$ is the regularizor.

2.2.2 Gradient Boosting

It is not feasible to train all the trees in the ensemble together at once because it is hard to calculate the gradient as which is needed in traditional optimization methods. Instead, Xgboost use an additive training strategy: fix the trees have already learned, add new trees one at a time. Let $\hat{y}^{(t)}$ be the predictor at iteration t, then

$$\hat{y}^{(0)} = 0 \tag{ 3}$$

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + f_t \tag{ 4}$$

Where $f_t \in \mathcal{F}$ and optimizes the following target function, which is obtained by the Taylor expansion of the lost function for logistic regression to the second order.

$$obj^t = \sum_i \left(g_i(\hat{y}^{(t-1)}(x_i)) f_t(x_i) + \frac{h_i(\hat{y}^{(t-1)}(x_i))^2}{2} f_t^2(x_i) \right) + \Omega(f_i) \tag{ 5}$$

Here $g_i(\hat{y}) = \frac{d}{d\hat{y}} l(y_i, \hat{y}), h_i(\hat{y}) = \frac{d^2}{d\hat{y}^2} l(y_i, \hat{y})$.

2.2.3 Regularizer and Training strategy for CARTs

For any $f \in \mathcal{F}$, let T be the number of leaves in the tree representing f, $w_1, \cdots w_T$ be the scores on the leaves. Then the regularizer used in XGBoost is

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j} w_j^2 \tag{6}$$

The purpose of the second term is that it can smoothen the leaf scores.

To optimize f_t, firstly note that given a tree structure, obj^t is a quadratic function of the scores w_j, and the minimum of obj^tas well as the w_j that minimizes obj^t can be easily calculated given the tree structure. Now the tree can be constructed by a greedy algorithm in which one starts with a tree with one single node, and repeatedly split its leaves in a way that maximizes the decrease in obj^t in each step.

2.3 Permutation

The idea of detecting SNP interaction through permutation has previously been used by Greene (Greene, Himmelstein et al. 2010) and Jing (Li, Malley et al. 2016).
Greene et al. designed an explicit test of epistasis to reflect only nonlinear interaction or epistasis component of the model. They advanced the traditional permutation testing framework by shuffling each SNP column instead of randomizing the class label, which can generate permuted datasets for testing the null hypothesis that the only genetic associations in the data are linear or additive in nature and that any nonlinear interaction effects are only there by chance. This yields an explicit test of epistasis when combining with a method such as MDR, is capable of modeling nonlinear interactions. Jing et al. developed a permuted random forest (RF) method. They generated two test dataset by permutating the genotype of a pair of SNPs., In one dataset the SNPs are shuffled independently, while in the other they are shuffled as a pair hence keeping the pairwise association.. The difference of error rate between the two test dataset on a well-trained RF model is then used to measure the strength of the interaction of selected SNP pair. Other SNPs except for the selected pair are kept their original form in both datasets, so the interactions among other non-selected SNPs were preserved in both of the permutation framework.

Both methods above are marker-based. Motivated by them, we designed a gene-based permutation strategy for our interaction detection. For each pair of genes, we carried out two permutation strategies to generate two test datasets. Firstly, we divide the samples by class label into case and control groups. Then, in the first permutation strategy, we shuffle the genotypes of all the genes independently among the samples within each group, which removes all associations between genes within each group while keeping the association between SNPs within each gene. The independent margin effect of each gene is preserved due to the unchanged genotype frequencies of each gene within each group before and after permutation. In the second permutation strategy, within both the case and the control group, the genotypes of the two chosen genes are shuffled together as a group while the genotypes of all other genes are shuffled independently. Hence, the difference between the two test datasets is merely the presentation or deletion of the interaction between the pair of genes.

Compared with Jing's method, our approach is gene-based instead of marker-based, hence is more suitable for detecting interactions between genes. Also, in both our permutation strategies, the associations between the genes that are not being tested are destroyed, while in Jing's approach the associations between the SNPs not currently being tested are preserved.

XXXXXXXX(emphasize the difference between our permutation and Jing's)

Figure for permutation procedure ***(not finished)***

2.3 Testing and Ranking

We introduce a new approach to gene-based gene-gene interaction detection. It is based on comparing the performance of two test dataset keeping or removing interaction through permutation strategies (in section 2.2) on predictive model (in section 2.1). Our interaction estimation technique is based on the following observation. If and interact, then test data of the second permutation strategy have significant better predictive performance than test data of the first permutation strategy, because the latter one cannot reflect the true functional dependency between and .On the other hand, if the two gene do not interact, then the absence of the interaction in the test data should not hurts its performance. Hence in the absence of an interaction between and , the predictive performance on the first test data and the second test data should be comparable.

In this paper, we use AUC (Area Under Curve), the area under the a receiver operating characteristic (ROC) curve, to measure the classifier performance. In machine learning, ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. Consider the ROC plot of the true positive rate vs the false positive rate as the threshold value for classifying an item as 0 or is increased from 0 to 1: if the classifier is very good, the true positive rate will increase quickly and the area under the curve will be close to 1. If the classifier is no better than random guessing, the true positive rate will increase linearly with the false positive rate and the area under the curve will be around 0.5. Generally, the larger AUC is, the better performance the classifier has. Especially, the AUC is independent of the fraction of the test dataset which is class 0 or class 1 that it is useful for evaluating the performance of classifier even on unbalanced data sets.

In the third step of testing, for each pair of genes, both of the permuted datasets were tested using the well-trained Xgboost model to get their AUC scores. Permutation was repeated 100 times and the average AUC was calculated from all permutations. We named the average AUC from the first permutation strategy, AUC1, in which the test dataset only maintain the margin effect of genes in the gene list. And named the average AUC from the second permutation strategy, AUC2, in which the test dataset kept both interaction between selected pair of gene and margin effect of all the genes. Therefore, the subtraction of AUC2 and AUC1 would be the difference of classifier performance caused by the interaction.

In the last step, after each pair of genes was permuted using the two permutation schemes (in the section 2.2) and tested to get the AUC scores, the AUC difference was calculated and used as the measurement of the interaction strength for selected gene pair, since removing an important interaction could have a strong effect on classifier performance. The larger the is, the stronger interaction signal was indicated for that pair of genes. The s were ranked and the pair of genes with the largest having the strongest interaction among all the gene pairs. In practice, we can pick top 5 pairs as the candidate interaction pair or selected candidate pairs through the distribution of .

The ALGORITHM 2 is the framework of gpXGB. ***(not finished)***

ALGORITHM2: gpXGB

Input: SNPs on interested gene set considered in a case-control study, gene location information, buffer region size

Output: The ranking list sorted by for all the gene pairs in the dataset

Step1: Train the Xgboost model with grid search of the proper parameter, using 5-fold cross validation for each parameter combination. Select the best model with XX (criterion).

Step2:

Not finished below

2.4 Simulation study

To evaluate the performance of our gpXGB procedure for gene-based gene-gene interaction detection, we use GAMETES to generate case-control simulation datasets with different parameter setting. Simple description for GAMETES. We simulate two scenarios where the phenotype is as the sum of one pair or multiple pairs (we choose 5) of random, pure and strict 2-locus genetic models. In each sensoria, heritability varied [redacted] and MAF was either 0.2 or 0.4. For each of the 12 genetic constrains combinations, 1,000 models were ranked by CORs and the models with the highest, moderate and lowest EDMs were selected as the three models for data simulation. For each selected model, we simulated 100 replicate datasets under the sample size [redacted] with balanced cases and controls. All together, we generated a total of

2.5 Application with protein-protein interactions (PPI) to Netherlands Study of Depression and Anxiety (NESDA) GWAS dataset

3. Results

3.1 methods to compare

KCCU

The kernel canonical correlation-based U-statistic model (KCCU) is a gene-based gene-gene interaction detection method which can reflect nonlinear relationship between two genes in the case-control dataset. In KCCU, for given two genes [redacted] , such that . Consider the genotype matrices and , with corresponding reduced kernel representations and . Define and to be the respective maximal kernel canonical correlations for case and control between gene and . After transform the and to an analog of the Fisher's simple correlation coefficient transformation, we can obtain the KCCU statistic. The details of KCCU method can be found in the paper [].

3.2 Simulation studies

3.3 real dataset

4. Discussion

Abstract

Among the large of number of statistical methods that have been proposed to identify gene-gene interactions in case-control genome-wide association studies (GWAS), gene-based methods have recently grown in popularity as they confer advantage in both statistical power and biological interpretation.

Introduction

Sampling of cases and controls was then completed from a sufficiently large number of simulated genotype-phenotype pairs.

For both type I error and power simulations, we consider whether or not explicit marginal effects are included in the disease model. Each simulation scenario is conducted with case-control status sample sizes of 500, 1000, and 1500.

XXX data study

We applied the bagged-pRF approach to detect gene-gene interaction within the XXX pathway, using data from a case-control study of XXX.

Results

Type I error

Power

Application to XXX data

Discussion

XXX data findings

Conflict of interest

Acknowledgements

References

Chen, T. and C. Guestrin (2016). XGBoost: A Scalable Tree Boosting System. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Emily, M. (2016). "AGGrEGATOr: A Gene-based GEne-Gene interActTiOn test for case-control association studies." Stat Appl Genet Mol Biol **15**(2): 151-171.

Greene, C. S., D. S. Himmelstein, H. H. Nelson, K. T. Kelsey, S. M. Williams, A. S. Andrew, M. R. Karagas and J. H. Moore (2010). "Enabling personal genomics with an explicit test of epistasis." Pac Symp Biocomput: 327-336.

Larson, N. B., G. D. Jenkins, M. C. Larson, R. A. Vierkant, T. A. Sellers, C. M. Phelan, J. M. Schildkraut, R. Sutphen, P. P. Pharoah, S. A. Gayther, N. Wentzensen, C. Ovarian Cancer Association, E. L. Goode and B. L. Fridley (2014). "Kernel canonical correlation analysis for assessing gene-gene interactions and application to ovarian cancer." Eur J Hum Genet **22**(1): 126-131.

Li, J., D. Huang, M. Guo, X. Liu, C. Wang, Z. Teng, R. Zhang, Y. Jiang, H. Lv and L. Wang (2015). "A gene-based information gain method for detecting gene-gene interactions in case-control studies." Eur J Hum Genet **23**(11): 1566-1572.

Li, J., J. D. Malley, A. S. Andrew, M. R. Karagas and J. H. Moore (2016). "Detecting gene-gene interactions using a permutation-based random forest method." BioData Min **9**: 14.

Ma, L., A. G. Clark and A. Keinan (2013). "Gene-based testing of interactions in association studies of quantitative traits." PLoS Genet **9**(2): e1003321.

Peng, Q., J. Zhao and F. Xue (2010). "A gene-based method for detecting gene-gene co-association in a case-control association study." Eur J Hum Genet **18**(5): 582-587.

Wan, X., C. Yang, Q. Yang, H. Xue, X. Fan, N. L. Tang and W. Yu (2010). "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies." Am J Hum Genet **87**(3): 325-340.

Yuan, Z., Q. Gao, Y. He, X. Zhang, F. Li, J. Zhao and F. Xue (2012). "Detection for gene-gene co-association via kernel canonical correlation analysis." BMC Genet **13**: 83.