




Feature Selection for Detecting Gene-Gene Interactions in Genome-Wide Association Studies

Faramarz Dorani and Ting Hu^(✉) 

Department of Computer Science, Memorial University,
St. John's, NL A1B 3X5, Canada
{faramarz.dorani,ting.hu}@mun.ca

Abstract. Disease association studies aim at finding the genetic variations underlying complex human diseases in order to better understand the etiology of the disease and to provide better diagnoses, treatment, and even prevention. The non-linear interactions among multiple genetic factors play an important role in finding those genetic variations, but have not always been taken fully into account. This is due to the fact that searching combinations of interacting genetic factors becomes inhibitive as its complexity grows exponentially with the size of data. It is especially challenging for genome-wide association studies (GWAS) where typically more than a million single-nucleotide polymorphisms (SNPs) are under consideration. Dimensionality reduction is thus needed to allow us to investigate only a subset of genetic attributes that most likely have interaction effects. In this article, we conduct a comprehensive study by examining six widely used feature selection methods in machine learning for filtering interacting SNPs rather than the ones with strong individual main effects. Those six feature selection methods include chi-square, logistic regression, odds ratio, and three Relief-based algorithms. By applying all six feature selection methods to both a simulated and a real GWAS datasets, we report that Relief-based methods perform the best in filtering SNPs associated with a disease in terms of strong interaction effects.

Keywords: Feature selection · Relief algorithms · Information gain
Gene-gene interactions · Genome-wide association studies

1 Introduction

The fundamental task of genetic association studies is to detect genetic variations that contribute to a disease status. In genome-wide association studies (GWAS), partial or all of the human genome is genotyped for discovering the associations between genetic factors and a disease or a phenotypic trait [1]. GWAS first began as a consequent of the HapMap Project [2] in 2005 aiming at discovering new treatments for common human diseases such as cancers. GWAS investigate the

genetic variations in two phenotypically distinguished populations, healthy and diseased, to find the variants that can explain the disease. There are two types of genetic variation: single nucleotide polymorphism (SNP) and copy number variation (CNV). In GWAS the genetic variants under consideration are SNPs, the most common type of variation among people. SNPs occur within a person's DNA in almost every 300 nucleotides, meaning that there are around ten million SNPs in the whole human genome. A SNP generally refers to a base-pair (or locus) in the DNA sequence which has a variation higher than 1% in a population [3]. Variations represent different alleles at a bi-allelic locus. In GWAS, genome data of a group of healthy individuals (i.e., controls) and diseased individuals (i.e., cases) are collected and genotyped, which usually contain more than one million SNPs and thus are regarded as *high dimensional* data.

It is a challenging task to analyze high dimensional SNP data for GWAS. The number of variables, i.e., SNPs, brings an extensive computational burden for informatics methods [4, 5]. Moreover, in the studies of common human diseases, it has been accepted that the non-additive effects of multiple interacting genetic variables play an important role explaining the risk of a disease [6, 7]. The traditional one-gene-at-a-time strategies likely overlook important interacting genes that have moderate individual effects. Therefore, powerful data mining and machine learning methods are needed in order to examine multiple variables at a time and to search for gene-gene interactions that contribute to a disease. A GWAS dataset with a million variables can be prohibitive for the application of any machine learning algorithms for detecting gene-gene interactions, since enumerating all possible combinations of variables is impossible. In addition, many of those variables can be redundant or irrelevant for the disease under consideration. Thus the selection of a subset of relevant and potential variables to be included in the subsequent analysis, i.e., *feature selection*, is usually needed [4].

Feature selection is frequently used as a pre-processing step in machine learning when the original data contain noisy or irrelevant features that could compromise the prediction power of learning algorithms [8]. Feature selection methods choose only a subset of the most important features, and thus reduce the dimensionality of the data, speed up the learning process, simplify the learned model, and improve the prediction performance [9, 10].

Feature selection involves two main objectives, i.e., to maximize the prediction accuracy and to minimize the number of features. There are two general approaches for selecting features for predictive models: filter and wrapper. The key difference between these two is that in filter approaches the learning algorithm has no influence in selecting features. That is, features are selected based on a filtering criterion independent of the learning model. Both filter and wrapper approaches have wide applications. Filter approaches have the advantage of high speed while wrapper approaches generally can achieve better prediction accuracies [11]. Of those two, filter approaches are often used in bioinformatics studies given the fact that they can easily scale to very high-dimensional data, that they are computationally simple and fast, and that they are independent of the classification algorithm [12].

There have been studies investigating the performance of feature selection methods on high dimensional datasets in bioinformatics. Hua et al. [13] evaluated the performance of several filter and wrapper feature selection methods on both synthetic and real gene-expression microarrays data with around 20,000 features (genes) and 180 samples. Shah and Kusiak [14] used a genetic algorithm (GA) to search for the best subset of SNPs in a dataset with 172 SNPs. The feature subset was then evaluated by a baseline classifier to compare with using the whole feature set. Wu et al. [15] proposed an SNP selection and classification approach based on random forest (RF) for GWAS. Their stratified random forest (SRF) method was tested on Parkinson and Alzheimer’s case-control data and was shown to outperform other methods including the original RF and support vector machines (SVM) in terms of test-error and run time. Brown et al. [16] proposed a framework of using mutual information for feature selection. Their objective was to select the smallest feature subset that has the highest mutual information with the phenotypic outcome.

However, most existing studies used the classification accuracy as the indicator for feature selection performance. The contribution of a feature to a phenotypic outcome could be its individual main effect or its interacting effect combined with other features. Using the overall classification accuracy was not able to distinguish the interaction effects of multiple variables and the individual main effects.

In our study, we focus on searching for features (SNPs) that have strong associations with the disease outcome in terms of gene-gene interactions. This differentiates our work from many existing studies that mostly focus on SNPs with high main-effects. We apply information gain to quantify the pair-wise synergy of SNPs and use that to evaluate various feature selection methods in order to identify the ones that can find subsets of SNPs with high synergistic effects on the disease status. We investigate six most popular filter algorithms, and test them on both simulated and real GWAS datasets. Our findings can be helpful for the recommendation of feature selection methods for detecting gene-gene interactions in GWAS.

2 Methods

In this section, we first discuss the data that will be used in this study, which include a simulated and a real population-based GWAS datasets. Then we introduce the information gain measure that will be employed as the quantification of the synergistic interaction effect of pairs of SNPs. Last, we present the six feature selection algorithms that will be investigated and compared.

2.1 Datasets

GWAS collect DNA sequencing data from two phenotypically distinguished populations, namely the diseased cases and healthy controls. A few thousand to a million of SNPs are usually genotyped for each sample. Each SNP can be

regarded as a bi-allelic variable, i.e., it has two different variations, with the common allele among a population called the *reference* and the other called *variant*. Given the fact that human chromosomes are paired, three categorical values are usually used to code for each SNP, i.e., 0 for homozygous reference, 1 for heterozygous variant, and 2 for homozygous variant.

For this study, we use a simulated genetic association dataset generated by the genetic architecture model emulator for testing and evaluating software (GAMETES) [17, 18]. GAMETES is a fast algorithm for generating simulation data of complex genetic models. Particularly, in addition to additive models, GAMETES is specialized for generating pure interaction models, i.e., interacting features without the existence of any main effects. Each n -locus model is generated deterministically, based on a set of random parameters and specified values of heritability, minor allele frequencies, and population disease prevalence. Since we focus on pairwise SNP interactions, we use GAMETES to generate a population of 500 samples with half being cases and half being controls. The dataset has 1000 SNPs, where 15 pairs are two-locus interacting models with a minor allele frequency of 0.2 and another 970 are random SNPs. We set the heritability to 0.2 and population prevalence to 0.5.

In addition, we use a real GWAS dataset collected for a case-control study on colorectal cancer (CRC) from the Colorectal Transdisciplinary (CORECT) consortium [19]. The dataset has over two million genetic variants of 1152 individuals of which 656 are CRC cases and 496 are healthy controls. Quality control [20] is first conducted to remove low-quality samples and sub-standard SNPs from the dataset. Then we remove redundant SNPs that are in linkage disequilibrium (LD). After quality control and LD pruning steps, 186,251 SNPs and 944 samples pass various filters. In this remaining population, 472 samples are cases and 472 are controls. The minimum and maximum minor allele frequency (MAF) of the SNPs are 0.04737 and 0.5 respectively.

2.2 Quantification of Pairwise Interactions Using Information Gain

Information theoretic measures such as entropy and mutual information [21] quantify the uncertainty of single random variables and the dependence of two variables, and have seen increasing applications in genetic association studies [22–25]. In such a context, the *entropy* $H(C)$ of the disease class C measures the unpredictability of the disease, and the conditional entropy $H(C|A)$ measures the uncertainty of C given the knowledge of SNP A . Subtracting $H(C|A)$ from $H(C)$ gives the *mutual information* of A and C , and is the reduction in the uncertainty of the class C due to the knowledge about SNP A 's genotype, defined as

$$I(A; C) = H(C) - H(C|A). \quad (1)$$

Mutual information $I(A; C)$ essentially captures the main effect of SNP A on the disease status C .

When two SNPs, A and B , are considered, mutual information $I(A, B; C)$ measures how much the disease status C can be explained by combining both A and B . The *information gain* $IG(A; B; C)$, calculated as

$$IG(A; B; C) = I(A, B; C) - I(A; C) - I(B; C), \quad (2)$$

is the information gained about the class C from the genotypes of SNPs A and B considered together minus that from each of these SNPs considered separately. In brief, $IG(A; B; C)$ measures the amount of synergetic influence SNPs A and B have on class C . Thus, information gain IG can be used to evaluate the pairwise interaction effect between two SNPs in association with the disease.

2.3 Feature Selection Algorithms

We choose six most widely used feature selection algorithms in our comparative study, and investigate their performance on searching variables that contribute to the disease in terms of gene-gene interactions. These six feature selection algorithms include three uni-variate approaches, chi-square, logistic regression, and odds ratio, and three Relief-based algorithms, ReliefF, TuRF, and SURF. They will be applied to both simulated and real GWAS datasets and provide rankings of all the SNPs in the data.

Chi-square: The chi-square (χ^2) test of independence [26] is commonly used in human genetics and genetic epidemiology [4] for categorical data. A χ^2 test estimates how likely different alleles of a SNP can differentiate the disease status. It is a very efficient filtering method for assessing the independent effect of individual SNPs on disease susceptibility.

Logistic regression: Logistic regression measures the relationship between the categorical outcome and multiple independent variables by estimating probabilities using a logistic function. A linear relationship between variables and the categorical outcome is usually assumed, and a coefficient is estimated for each variable when such a linear relationship is trained to best predict the outcome. The variable coefficient can then be used as a quantification of the importance of each variable.

Odds-ratio: Odds ratio (OR) is the most commonly used statistic in case-control studies. It measures the association between an exposure (e.g., health characteristic) and an outcome (e.g., disease status). The OR represents the odds that a disease status will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure [27].

ReliefF: Relief is able to detect complex attribute dependencies even in the absence of main effects [28]. It estimates the quality of attributes using a nearest-neighbor algorithm. While Relief uses, for each individual, a single nearest neighbor in each class, ReliefF, a variant of Relief, uses multiple, usually 10, nearest neighbors, and thus is more robust when a dataset contains noise [29, 30]. The basic idea of Relief-based algorithms is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give more

Table 1. Ranks of the 30 known interacting SNPs by feature selection algorithms.

	Logit	χ^2	OR	ReliefF	TuRF	SURF
Mean	549.16	548.30	444.10	202.63	166.96	233.16
SD	277.99	267.18	287.04	201.74	259.74	212.13
Median	617.50	536.50	346.50	130.00	21.50	183.50

weights to features that discriminate the instance from its neighbors of different classes. Comparing to uni-variate feature selection algorithms, ReliefF is able to capture attribute interactions because it selects nearest neighbors using the entire vector of values across all attributes [4, 30].

Tuned ReliefF (TuRF): It is an extension of ReliefF specifically for large-scale genetics data [31]. This method systematically and iteratively removes attributes that have low-quality estimates so that the remaining attributes can be re-estimated more accurately. It improves the estimation of weights in noisy data but does not fundamentally change the underlying ReliefF algorithm. It is useful when data contain a large number of non-relevant SNPs. It is also more computationally intense because of the iterative process of removing attributes.

Spatially Uniform ReliefF (SURF): SURF is also an extension of the ReliefF algorithm [32]. It incorporates the spatial information when assesses neighbors. Instead of using a fixed number of neighbors as the threshold in ReliefF, SURF uses a fixed distance threshold for choosing neighbors. It is reported to be able to improve the sensitivity detecting small interaction effects.

3 Results

3.1 Feature Selection Algorithms on the Simulated Data

First, we apply all six feature selection algorithms to the simulated dataset that contains 30 known SNPs with pairwise interactions and 970 random SNPs. The chi-square, odd-ratio, ReliefF, TuRF, and SURF algorithms are implemented using the multifactor dimensionality reduction (MDR) software with default parameter settings [33]. Logistic regression is implemented using the Python *scikit-learn* package [34].

Each algorithm yields a ranking of all 1000 SNPs. Table 1 shows the statistics of the ranks of those 30 known SNPs by each feature selection algorithm. We see that TuRF has both the highest mean and median ranks among all the methods, and the differences are significant. ReliefF performs the second best, followed by SURF.

Figure 1 shows the recall-at- k for all six feature selection algorithms. The y-axis shows the fraction of those 30 known SNPs detected by the top k SNPs ranked by each feature selection algorithm. We can see that for all values of k , TuRF has the highest recalls. In addition, all three Relief-based algorithms outperform the other methods.

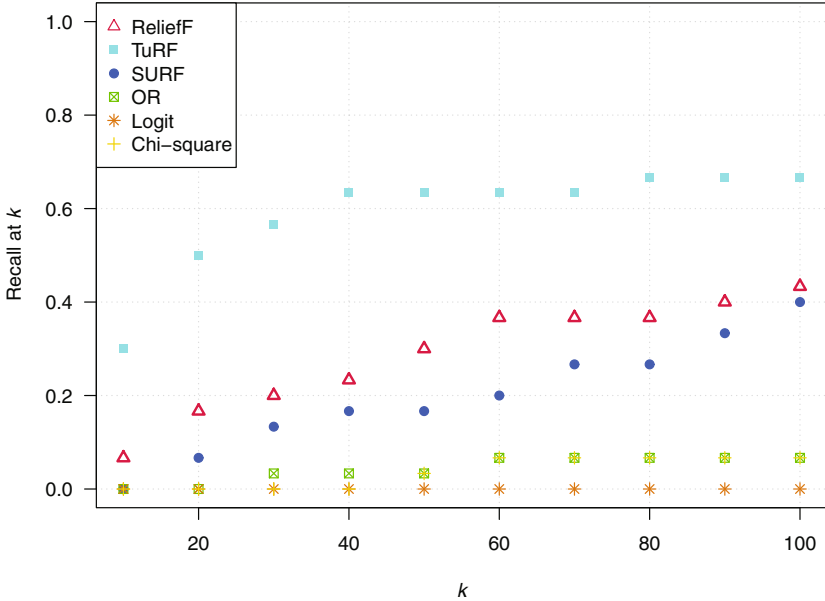


Fig. 1. Diagram of recall-at- k for six feature selection algorithms applied to the simulated dataset. Recall-at- k is the fraction of the 30 known interacting SNPs detected by the top k ranked SNPs using each feature selection algorithm.

Figure 2 shows the distributions of the ranks of those 30 known interacting SNPs using different feature selection algorithms. The x-axis is the rank of SNPs and the y-axis is the density. Again, TuRF has the highest density around high ranks, meaning that it produces the highest ranks for those 30 known SNPs. SURF and ReliefF also have better ranking performance comparing to the other three methods. Odds-ratio, logistic regression, and chi-square have flat distributions across the entire rank range, which indicates their inability to identify those 30 interacting SNPs.

3.2 Feature Selection Algorithms on the CRC Data

We then compare the performance of those six feature selection algorithms using the CRC GWAS dataset. The CRC GWAS dataset is processed using PLINK software [35]. PLINK can conduct some fundamental association tests by comparing allele frequencies of SNPs between cases and controls. We use the command `--assoc` to compute chi-square and odds-ratio scores for each SNP, and the command `--logistic` for logistic regression analysis. Again, we used the MDR software [33] to implement ReliefF, TuRF, and SURF algorithms.

Each feature selection algorithm generates a ranking of all the 186,251 SNPs in the dataset. For detecting gene-gene interactions, exhaustive enumeration of all possible combinations of SNPs is usually considered. Even for pairwise

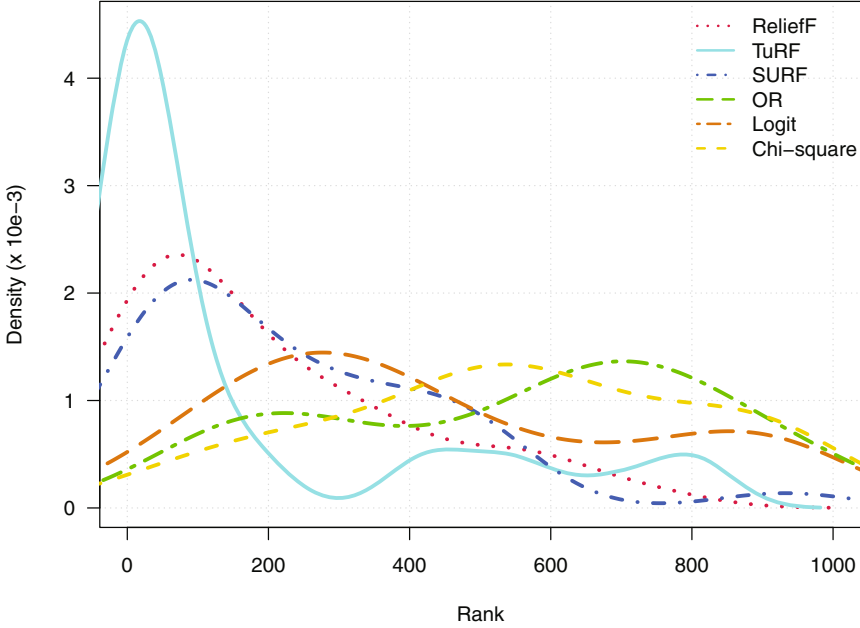


Fig. 2. Density of the ranks of the 30 known interaction SNPs using different feature selection algorithms on the simulated dataset.

interactions, the total number of possible pairs $\binom{n}{2}$ grows fast with the number of SNPs n . Therefore, we can only consider a moderate subset of SNPs for interaction analysis, and we use the rankings estimated using feature selection algorithms to filter those potentially more important SNPs. We choose the subset of the top 10,000 SNPs by each feature selection algorithm. Then, for the six subsets of filtered 10,000 SNPs, we evaluate their pairwise interactions separately using the information gain (IG) measure.

Table 2. Statistics of the information gain values of all $\binom{10,000}{2}$ SNP pairs filtered by each feature selection algorithm ($\times 10^{-3}$).

	Logit	χ^2	OR	ReliefF	TuRF	SURF
Max	27.4	27.6	27.4	30.2	28.9	28.2
Min	-4	-5.1	-4	-3.2	-2.9	-5.7
Mean	2.760	3.047	2.776	3.190	3.191	3.056
SD	2.117	2.221	2.120	2.243	2.251	2.224
Median	2.3	2.6	2.3	2.7	2.7	2.6

Table 2 shows the maximum, minimum, mean, standard deviation, and median values of the information gain calculated using all $\binom{10,000}{2}$ pairs of the

10,000 SNPs filtered by the six feature selection algorithms. As we can see, ReliefF finds the SNP pair with the highest interaction strength, and TuRF has the best overall distribution.

Figure 3 shows the distribution of the interaction strength of all $\binom{10,000}{2}$ pairs of SNPs selected by each feature selection algorithm. We see that the distributions of ReliefF and TuRF have overall more SNP pairs with higher IG values. The distributions of SURF and chi-square are comparable, and logistic regression and odds ratio have the lowest overall IG values.

The significance of the IG value of each pair of SNPs can be assessed using permutation testing. For each permutation, we randomly shuffle the case/control labels of all the samples in the data in order to remove the association between the genotypes of SNPs and the disease status. Repeating such a permutation multiple times generates a null distribution of what can be observed by chance. For each permuted dataset, we compute the IG value of each pair of SNPs. In this study, we perform a 100-fold permutation test. The significance level (p -value) of the IG of each SNP pair can be assessed by comparing the IG value of the pair calculated using the real dataset to the IG values calculated using the 100 permuted datasets (see Algorithm 1).

Algorithm 1. Permutation testing algorithm

```

1: procedure COMPUTEPVALUE
2:    $D \leftarrow$  original dataset
3:    $n \leftarrow$  number of permutations
4:    $m \leftarrow$  number of SNP pairs
5:    $C \leftarrow$  counter for each SNP pair
6:    $i \leftarrow 1$ 
7:   while  $i < n$  do
8:     Generate a random permutation  $D'$  of the original dataset  $D$ 
9:      $j \leftarrow 1$ 
10:    while  $j < m$  do
11:      calculate  $IG_j^{D'}$  for the  $j$ -th SNP pair
12:      increase  $C_j$  by 1 if  $IG_j^{D'}$  is greater than the real observed  $IG_j^D$ 
13:       $j \leftarrow j + 1$ 
14:     $i \leftarrow i + 1$ 
15:    compute the significance level  $p_k$  for each SNP pair  $k$  as  $\frac{C_k}{n}$ 

```

We apply permutation testing to all six subsets of $\binom{10,000}{2}$ pairs of SNPs selected by each feature selection algorithm, such that their significance level p -values can be assessed. Figure 4 shows the number of SNP pairs that pass two different p -value thresholds, 0.01 and 0.05. TuRF has more SNP pairs with significant interaction strength using both thresholds. All three Relief-based algorithms have higher numbers of significant SNP pairs than the other three methods. Logistic regression and odds ratio find the least numbers of significant interacting SNP pairs.

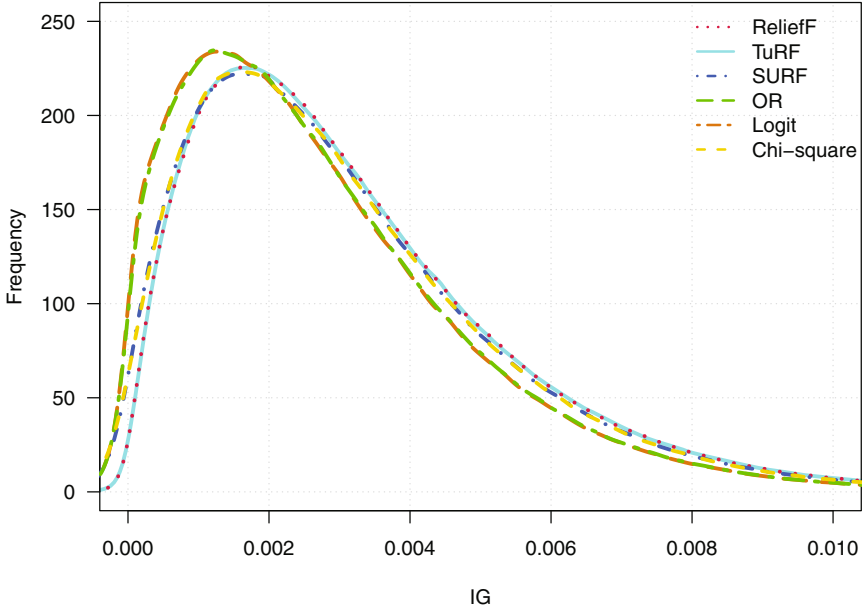


Fig. 3. Distribution of the information gain (IG) values of all pairs of filtered 10,000 SNPs by each feature selection algorithm.

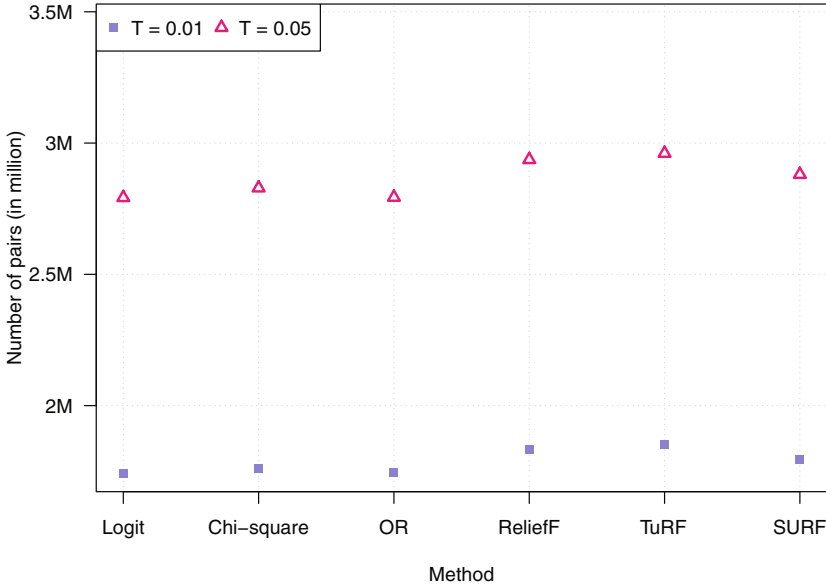


Fig. 4. The number of SNP pairs with significant interaction strengths using p -value cutoff T . Red triangles show the results using cutoff $p \leq 0.01$, and blue squares show the results with cutoff $p \leq 0.05$. (Color figure online)

4 Discussion

The goal of genome-wide association studies (GWAS) is to identify genetic markers that can explain complex human diseases. Most existing analyses for GWAS look at one gene at a time due to the limitation of analytical methodologies and computational resources. Such a strategy very likely overlook potentially important genetic attributes that have low main effects but contribute to a disease outcome through multifactorial interactions. Detecting such non-additive gene-gene interactions help us better understand the underlying genetic background of common diseases and better develop new strategies to treat, diagnose, and prevent them.

Detecting gene-gene interactions for GWAS imposes computational challenges since enumerating combinations of genetic attributes becomes inhibitive when up to a million variables are under consideration. Thus, feature selection becomes a necessity for the task.

In this study, we investigated the performance of six widely used feature selection algorithms for detecting potentially interacting single nucleotide polymorphisms (SNPs) for GWAS. We used both a simulated and real genetic datasets. We adopted information gain as a measure for quantifying pairwise interaction strength of SNPs in order to evaluate the filtering performance of those six feature selection algorithms. Among the investigated feature selection methods, three are single variable feature scoring methods. That is, they only consider individual main effects of SNP on the disease status. Three other methods are extensions of the Relief algorithm which is a multivariate feature selection algorithm.

For the simulated dataset, we generated a population-based dataset with 1000 SNPs including 15 pairs of interacting SNPs and 970 random ones. We applied all six feature selection algorithms to rank those 1000 SNPs and look into the recall-at- k of detecting those 30 known interacting SNPs. The TuRF algorithm has the highest recall-at- k for all k values, followed by ReliefF and SURF. All three Relief-based algorithms perform better than odds ratio, logistic regression, and chi-square.

We also tested the feature selection algorithms using a real GWAS dataset on colorectal cancer (CRC). We used information gain to quantify pairwise interaction strength of SNPs in order to evaluate the filtering performance of the feature selection algorithms. We chose 10,000 top-ranked SNPs by each feature selection algorithm and applied information gain measure and permutation testing to compute the interaction strengths and their significance levels of all pairs of SNPs. We found that TuRF again was able to filter more significant interacting SNPs than the rest of the feature selection algorithms. All three Relief-based algorithms outperformed the other three methods.

TuRF and ReliefF had comparable performance on the application to the real CRC dataset. By looking at their top 10,000 SNPs, we saw that only 1474 were overlapped. That is, only 14% of their top 10K SNPs are the same. This is interesting that they seemed to be able to find different sets of interacting SNPs.

There is no general rule for selecting the best feature selection method in machine learning studies. The decision mostly depends on the data and research question of the investigation. For the purpose of detecting gene-gene interactions, Relief-based methods were shown to have better performance than the common univariate methods. Gene-gene interactions can be very challenging to detect by univariate methods since interacting genetic factors may not show significant individual main effects. By evaluating sample similarity using all genetic attributes, Relief-base algorithms are able to capture the non-addition interaction effects among multiple attributes, and are recommended for detecting gene-gene interactions for GWAS.

In future studies, we expect to explore more sophisticated feature selection algorithms, especially wrapper and embedded methods, and test their utilities in genetic association and bioinformatics studies.

Acknowledgments. This research was supported by Newfoundland and Labrador Research and Development Corporation (RDC) Ignite Grant 5404.1942.101 and the Natural Science and Engineering Research Council (NSERC) of Canada Discovery Grant RGPIN-2016-04699 to TH.

References

1. Wellcome Trust Case Control Consortium, et al.: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**(7145), 661 (2007)
2. Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'ang, L.Y., Huang, W., Liu, B., Shen, Y., et al.: The international HapMap project. *Nature* **426**(6968), 789–796 (2003)
3. The 1000 Genomes Project Consortium, et al.: A map of human genome variation from population scale sequencing. *Nature* **467**(7319), 1061 (2010)
4. Moore, J.H., Asselbergs, F.W., Williams, S.M.: Bioinformatics challenges for genome-wide association studies. *Bioinformatics* **26**(4), 445–455 (2010)
5. Hu, T., Andrew, A.S., Karagas, M.R., Moore, J.H.: Statistical epistasis networks reduce the computational complexity of searching three-locus genetic models. *Proc. Pac. Symp. Biocomput.* **18**, 397–408 (2013)
6. Cordell, H.J.: Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **11**(20), 2463–2468 (2002)
7. Hu, T., Chen, Y., Kiralis, J.W., Moore, J.H.: ViSEN: methodology and software for visualization of statistical epistasis networks. *Genet. Epidemiol.* **37**, 283–285 (2013)
8. Yu, L., Liu, H.: Feature selection for high-dimensional data: a fast correlation-based filter solution. *ICML* **3**, 856–863 (2003)
9. Dash, M., Liu, H.: Feature selection for classification. *Intell. Data Anal.* **1**(1–4), 131–156 (1997)
10. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**(Mar), 1157–1182 (2003)
11. Freitas, A.A.: *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer Science & Business Media, Heidelberg (2013)

12. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**(19), 2507–2517 (2007)
13. Hua, J., Tembe, W.D., Dougherty, E.R.: Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recogn.* **42**(3), 409–424 (2009)
14. Shah, S.C., Kusiak, A.: Data mining and genetic algorithm based gene/SNP selection. *Artif. Intell. Med.* **31**(3), 183–196 (2004)
15. Wu, Q., Ye, Y., Liu, Y., Ng, M.K.: SNP selection and classification of genome-wide SNP data using stratified sampling random forests. *IEEE Trans. Nanobiosci.* **11**(3), 216–227 (2012)
16. Brown, G., Pocock, A., Zhao, M.J., Luján, M.: Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* **13**(Jan), 27–66 (2012)
17. Urbanowicz, R.J., Kiralis, J.W., Fisher, J.M., Moore, J.H.: Predicting the difficulty of pure, strict, epistatic models: metrics for simulated model selection. *BioData Min.* **5**, 15 (2012)
18. Urbanowicz, R.J., Kiralis, J., Sinnott-Armstrong, N.A., Heberling, T., Fisher, J.M., Moore, J.H.: Gametes: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData Min.* **5**(1), 16 (2012)
19. Schumacher, F.R., Schmit, S.L., Jiao, S., Edlund, C.K., Wang, H., Zhang, B., Hsu, L., Huang, S.C., Fischer, C.P., et al.: Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nature Commun.* **6**, 7138 (2015)
20. Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., Zondervan, K.T.: Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**(9), 1564–1573 (2010)
21. Cover, T.M., Thomas, J.A.: *Elements of Information Theory*, 2nd edn. Wiley, Hoboken (2006)
22. Hu, T., Sinnott-Armstrong, N.A., Kiralis, J.W., Andrew, A.S., Karagas, M.R., Moore, J.H.: Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinform.* **12**, 364 (2011)
23. Fan, R., Zhong, M., Wang, S., Zhang, Y., Andrew, A., Karagas, M., Chen, H., Amos, C.I., Xiong, M., Moore, J.H.: Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment interactions/correlations of complex diseases. *Genet. Epidemiol.* **35**(7), 706–721 (2011)
24. Li, H., Lee, Y., Chen, J.L., Rebman, E., Li, J., Lussier, Y.A.: Complex-disease networks of trait-associated single-nucleotide polymorphisms (SNPs) unveiled by information theory. *J. Am. Med. Inform. Assoc.* **19**, 295–305 (2012)
25. Hu, T., Chen, Y., Kiralis, J.W., Collins, R.L., Wejse, C., Sirugo, G., Williams, S.M., Moore, J.H.: An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *J. Am. Med. Inform. Assoc.* **20**(4), 630–636 (2013)
26. Yates, F.: Contingency tables involving small numbers and the χ^2 test. *Suppl. J. Roy. Stat. Soc.* **1**(2), 217–235 (1934)
27. Szumilas, M.: Explaining odds ratios. *J. Can. Acad. Child Adolesc. Psychiatry* **19**(3), 227 (2010)
28. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: *Proceedings of the Ninth International Workshop on Machine Learning*, pp. 249–256 (1992)
29. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In: Bergadano, F., De Raedt, L. (eds.) *ECML 1994*. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994). https://doi.org/10.1007/3-540-57868-4_57
30. Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of relief and rrelieff. *Mach. Learn.* **53**(1–2), 23–69 (2003)

31. Moore, J.H., White, B.C.: Tuning ReliefF for genome-wide genetic analysis. In: Marchiori, E., Moore, J.H., Rajapakse, J.C. (eds.) *EvoBIO 2007*. LNCS, vol. 4447, pp. 166–175. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-71783-6_16
32. Greene, C.S., Penrod, N.M., Kiralis, J., Moore, J.H.: Spatially uniform relief (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Min.* **2**(1), 5 (2009)
33. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., Moore, J.H.: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **69**(1), 138–147 (2001)
34. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
35. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J., et al.: Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**(3), 559–575 (2007)