The Informative Extremes: Using Both Nearest and Farthest Individuals Can Improve Relief Algorithms in the Domain of Human Genetics

Casey S. Greene, Daniel S. Himmelstein, Jeff Kiralis, and Jason H. Moore

Dartmouth Medical School, Lebanon, NH 03756, USA
Jason.H.Moore@dartmouth.edu
http://www.epistasis.org

Abstract. A primary goal of human genetics is the discovery of genetic factors that influence individual susceptibility to common human diseases. This problem is difficult because common diseases are likely the result of joint failure of two or more interacting components instead of single component failures. Efficient algorithms that can detect interacting attributes are needed. The Relief family of machine learning algorithms, which use nearest neighbors to weight attributes, are a promising approach. Recently an improved Relief algorithm called Spatially Uniform ReliefF (SURF) has been developed that significantly increases the ability of these algorithms to detect interacting attributes. Here we introduce an algorithm called SURF* which uses distant instances along with the usual nearby ones to weight attributes. The weighting depends on whether the instances are are nearby or distant. We show this new algorithm significantly outperforms both ReliefF and SURF for genetic analysis in the presence of attribute interactions. We make SURF* freely available in the open source MDR software package. MDR is a cross-platform Java application which features a user friendly graphical interface.

1 Introduction

New genotyping technologies are allowing human geneticists to routinely measure individual genetic variation on a vast "genome-wide" scale [1,2,3]. It is now feasible to measure more than one million variations from across the human genome. Here we focus on a particular type of variation, the single nucleotide polymorphism or SNP. Each SNP is a single point in a DNA sequence that differs between individuals. A major goal of human genetics is to link these genetic variations to disease risk [4]. Currently this problem is approached as a set of independent steps. The first step in the process is to discover SNPs that reliably predict disease susceptibility across many samples [5], but discovery of these robust single predictors has proven difficult [6,7,8]. Furthermore even the reliable and robust disease-associated SNPs that have been discovered often cannot be combined into effective classifiers of disease risk [9]. These association studies, by their nature, ignore complex interactions that may lead to disease susceptibility.

C. Pizzuti, M.D. Ritchie, and M. Giacobini (Eds.): EvoBIO 2010, LNCS 6023, pp. 182–193, 2010.

The term for complex gene-gene interactions that influence a trait such as disease susceptibility is epistasis. It is becoming apparent that studies ignoring epistasis are also likely to be neglecting informative markers [10,11]. Because of the complexity present in cellular and biological systems, epistasis is thought to be fundamental to an individual's risk for common human diseases [12]. This knowledge, combined with the inability of single-marker approaches to offer predictive models of individual disease risk, suggests that researchers should also carefully examine gene-gene interactions for associations with disease. Unfortunately examining the joint effect of these polymorphisms is difficult because commonly used methods are combinatorial.

Relief algorithms [13], which use nearest neighbors, have successfully detected gene-gene interactions in genetic association studies [14]. Here we introduce a novel Relief algorithm called SURF*. SURF* is better able than other Relief algorithms to detect SNPs which participate in epistatic interactions that relate to disease risk. The novel feature of SURF* is that it uses distant individuals, as well as the usual near ones, to adjust the scores of SNPs. Using these distant individuals has the effect of increasing sample size considerably.

This paper is organized as follows. Section 1.1 discusses approaches used in genetic association studies. Section 2 discusses intuitively how Relief algorithms can, in linear time with respect to the number of SNPs, detect epistatic interactions. Section 2.1 examines how SURF specifically is able to detect these interactions. This is important because we improve SURF with a novel approach, SURF*. A theoretical assessment of the improvement provided by SURF* is described in Section 3. We evaluate the new SURF* method empirically using a study design described in Section 4. This framework allows us to directly assess the success rate of the method. The results of the simulation are discussed in Section 5 and we discuss their implications in Section 6.

1.1 Related Work

The state of the art in this field still relies on only the analysis of single SNPs as in a recent large study of 17,000 individuals and seven common diseases from the Wellcome Trust Case Control Consortium [15]. While some approaches do consider complexity, these often condition on the effect of single SNPs or require combinatorial methods to exhaustively examine all potential interactions [16,17]. In the first case, these have the potential to miss interactions without main effects. In the second, the time to analyze large datasets becomes prohibitive because this type of analysis requires the consideration of the joint effect of attributes, here SNPs, a combinatorial challenge which has been previously described [18,19,20]. When datasets contain many SNPs, such combinatorial methods are infeasible.

2 Relief Algorithms

Cordell [21] provides a recent and thorough review of current analysis methods, including Relief algorithms, for these studies as well as the potential benefits and drawbacks of each. To this point the use of Relief algorithms in this field has been relatively limited [14,22,23], probably because previous small scale studies have not required these types of algorithms, and large scale studies have, thus far, often ignored epistasis. Given the difficulty of detecting predictive interacting SNPs, our novel and more effective Relief algorithm could greatly enhance the state of the field.

Relief algorithms, the first of which was developed by Kira and Rendell [13], are a natural fit for large scale genetic association studies designed to detect epistasis. They are fast and scale linearly with the number of SNPs and quadratically with the number of individuals. Furthermore these algorithms are able to detect interacting pairs of attributes that contribute to disease susceptibility. We have previously discussed how Relief algorithms do this from a mathematical point of view [24]. In summary, the Relief algorithm returns a weight for each SNP. Higher scores indicate that a SNP is more likely to be predictive of disease status. The adjustment of these scores is performed using the genetically most similar individuals. Here the inter-individual distance is the number of SNPs with differing genotypes between two individuals. Therefore, nearest individuals share the greatest number of genotypes. Relief works on the assumption that the SNPs of nearby individuals with different genotypes are most useful for assessing the predictiveness of the SNP. The algorithm adjusts the scores of these SNPs– upward if the two individuals have different disease status, and downward by the same amount if they have the same status. More precisely, for each individual I_i, SNP scores are adjusted using its nearest hit (the individual which is closest to I_i and in the same class as I_i) and its nearest miss, (the individual which is closest to I_i and in the other class from I_i). ReliefF [25] differs from Relief largely because it uses multiple neighbors for weighting instead of only the single nearest neighbor.

Table 1. Penetrance values for an example epistasis model with a heritability of 0.2.

		SNP_1		
		AA (0.36)	Aa (0.48)	aa (0.16)
SNP_2	BB (0.36)	0.393	0.764	0.664
	Bb (0.48)	0.850	0.398	0.733
	bb (0.16)	0.406	0.927	0.147

2.1 Spatially Uniform ReliefF (SURF)

Spatially Uniform ReliefF (SURF), developed by Greene et al. [24], detects attribute interactions in the same manner as Relief and ReliefF. SURF, like ReliefF, uses multiple nearest neighbors, but, instead of using a fixed number of nearest neighbors, SURF uses all neighbors within a specific similarity threshold, T. Instances may not be uniformly distributed in space and some instances may have more informative neighbors than other instances. SURF uses all neighbors more similar than the threshold, T, for weighting, while Relief and ReliefF may use either more or fewer neighbors. This can cause ReliefF to potentially include uninformative neighbors or to neglect informative ones. This swaps the number-of-neighbors used by Relief for the similarity-threshold used by SURF. For this we use the mean of the distances between all pairs of individuals, which can be easily computed from the data [24].

Here we will briefly outline how SURF is capable of detecting interacting pairs of attributes. This is thoroughly discussed in the appendix to Greene et al. [24] but here we highlight the parts necessary to understand how SURF* improves on SURF and adjust the notation to accommodate both the nearest and furthest individuals. To understand these algorithms, it is first necessary to understand the problem. We illustrate the situation of interacting pairs of SNPs using the penetrance table given in Table 1. According to this example, if an individual has genotype BB, the probability she is sick is $.36 \cdot .393 + .48 \cdot .764 + .16 \cdot .664 \approx .614$. If she has genotype Bb, this probability is the same, and likewise if she has genotype bb. Thus just SNP 2's genotype is not predictive of disease status. Similarly if SNP 1's genotype is known, but not SNP 2's, the probability she is sick is as before, $.36 \cdot .764 + .48 \cdot .398 + .16 \cdot .927 \approx .614$. Thus the genotypes of SNPs 1 and 2 are together predictive of disease status, but neither is individually. This is what makes SNPs 1 and 2 an epistatic pair of SNPs. In our study we employ 9000 datasets from 30 of these genetic models. In all models there are pairs of SNPs which are jointly predictive but no singly informative SNPs. Detecting these epistatic pairs is much more difficult then detecting SNPs which alone have an effect.

A basic fact we will use about epistatic pairs is that

$$|H_{2\Delta}| - |M_{2\Delta}| = \frac{1}{2}(|M_{1\Delta}| - |H_{1\Delta}|) = |H_{0\Delta}| - |M_{0\Delta}|. \tag{1}$$

This is discussed in sections 1 and 2 of the appendix in the paper first describing SURF [24].

Now let I_i be a random, but fixed, individual and let T be the threshold distance. Then each miss with distance less than T from I_i is in one of the three sets $M_{0\Delta}$, $M_{1\Delta}$ or $M_{2\Delta}$. For $k = 0$, 1 and 2, let $CM_{k\Delta}$ be the subset of $M_{k\Delta}$ consisting of those individuals with distance $< T$ from I_i. The notation $CM_{k\Delta}$ might be read as "close misses involving k changes of the relevant SNPs". Using analogous notation for hits with H in place of M, the contribution of individual I_i to the (SURF) score of a relevant SNP is

$$\begin{aligned} S_i^C &= \frac{1}{2}(|CM_{1\Delta}| - |CH_{1\Delta}|) + (|CM_{2\Delta}| - |CH_{2\Delta}|) \\ &= \frac{1}{2}(|CM_{1\Delta}| - |CH_{1\Delta}|) - (|CH_{2\Delta}| - |CM_{2\Delta}|). \end{aligned} \tag{2}$$

The $\frac{1}{2}$ is here since each individual in $CM_{1\Delta}$ and $CH_{1\Delta}$ changes the score of a relevant SNP by $\frac{1}{2}$, on the average. The total SURF score of a relevant SNP is the sum of the S_i^C over all individuals.

It follows from equation (1) that if arbitrary neighbors are used, rather than nearest ones, the expected score of a relevant SNP would be 0 since

$$\frac{1}{2}(|M_{1\varDelta}| - |H_{1\varDelta}|) = |H_{2\varDelta}| - |M_{2\varDelta}|.$$

The score S_i^C tends to be positive though because close neighbors are more apt to lie in the sets $M_{1\varDelta}$ and $H_{1\varDelta}$ rather than in $M_{2\varDelta}$ and $H_{2\varDelta}$, making

$$\frac{1}{2}(|CM_{1\varDelta}| - |CH_{1\varDelta}|) > |CH_{2\varDelta}| - |CM_{2\varDelta}|.$$

The reason close neighbors are more apt to lie in the $1\varDelta$-sets than in the $2\varDelta$ ones is that relevant SNPs of individuals in the $1\varDelta$-sets contribute one to the distance from I_i, while those in the $2\varDelta$-sets contribute two.

3 The Value of Both Nearest and Farthest

It is clear that the assumption made by Relief algorithms such as SURF, that the SNPs of nearby individuals with different genotypes are useful for assessing the predictiveness of the SNP, is correct as these algorithms are successful. It is not clear that distant individuals are not also useful. Our analysis suggests that using the states of genotypes for these most distant individuals can substantially improve the success rates of these algorithms. Using this information effectively increases the sample size available to SURF greatly improving its ability to detect epistatic SNPs when sample sizes are limited. We call the algorithm SURF* because using distant individuals is the opposite of SURF and because, in mathematics, * indicates opposite. Strictly speaking, the SURF* that we discuss includes both SURF and this additional opposite component.

We outline how this approach using both closest and farthest individuals outperforms the nearest neighbor approaches. The SURF* algorithm we introduce uses nearby neighbors in the same way the SURF algorithm does. The new part of the SURF* algorithm using distant individuals identifies those SNPs of distant individuals in different states and adjusts their scores–downward by one if the two individuals have different disease status, and upward by the same amount if they have the same status. (This is the same as with Relief algorithms, but upward and downward have been interchanged.) Specifically, we define subsets $DM_{k\varDelta}$ made up of distant misses of $M_{k\varDelta}$ consisting of those misses with distance $> T$ from I_i. Subsets $DH_{k\varDelta}$ made up of distant hits of $H_{k\varDelta}$ consist of those hits with distance $> T$ from I_i Then the contribution of individual I_i to the (distant individuals) score of a relevant SNP is

$$\begin{aligned} S_i^D &= -\frac{1}{2}(|DM_{1\varDelta}| - |DH_{1\varDelta}|) - (|DM_{2\varDelta}| - |DH_{2\varDelta}|) \\ &= -\frac{1}{2}(|DM_{1\varDelta}| - |DH_{1\varDelta}|) + (|DH_{2\varDelta}| - |DM_{2\varDelta}|). \end{aligned} \tag{3}$$

The mean of this is positive for essentially the same reason that the mean of S_i^C is. Namely, individuals in the 2Δ-group tend to be one farther from I_i than those in the 1Δ-group.

The means of S_i^C and S_i^D are the same, or nearly so. So the mean of the overall score $\Sigma_i S_i^C + \Sigma_i S_i^D$ of a relevant SNP is doubled by using distant individuals along with the usual close ones. We suspect that $\Sigma_i S_i^C$ and $\Sigma_i S_i^D$ are not independent. If so, with V denoting variance, we have

$$V(\Sigma_i S_i^C + \Sigma_i S_i^D) > V(\Sigma_i S_i^C) + V(\Sigma_i S_i^D).$$

Thus using distant individuals does not quite have the effect of doubling the sample size, but it does substantially increase the success rate. This improvement in success rate indicates that these methods are more likely to detect interacting relevant SNPs in these genetic association studies.

4 Experimental Design

Here we evaluate these methods in the context of a simulation study. The goal of our simulation study is to generate artificial datasets with high concept difficulty to evaluate these methods in the domain of human genetics. Our dataset characteristics were chosen to closely match common genetic association study designs from human genetics. We first develop 30 different penetrance functions (i.e. genetic models) which determine the relationship between genotype and phenotype in our simulated data. These functions determine the probability that an individual has the studied disease given his or her genotype. This probability depends only on the genotypes of the two interacting SNPs, not on the genotype of any one SNP. This case where there are no single SNP effects is thought to be the most difficult. Single SNP effects are easily found with other methods. The 30 penetrance functions consist of six groups of five with heritabilities of 0.025, 0.05, 0.1, 0.2, 0.3, or 0.4. Each of the six heritabilities is realized by all five models in one group. These heritabilities range from very small to large genetic effect sizes and thus test the algorithms across a broad swathe of scenarios.

SNPs are chosen for genotyping such that each SNP has two alleles due to technological constraints and such that these alleles are both common in the population. Here each model contains SNPs with two alleles which have frequencies of 0.4 and 0.6. Each of the models is used to generate 100 datasets with sample sizes of 800, 1600, and 3200. Studies with 800 individuals would be considered small relative to other genetic association studies while studies with 3200 individuals would be considered large. Each consists of an equal number of case and control subjects because genetic association studies are frequently designed to be balanced. Each pair of relevant SNPs is added to a set of 998 irrelevant SNPs for a total of 1000 attributes. This is similar to the size seen in association studies using custom SNP arrays to perform genotyping. A total of 9,000 datasets are generated and analyzed. This large number of datasets and study design allows us to rigorously evaluate and compare these methods across situations likely to be encountered. Due to the difficulty of detecting and characterizing epistasis, well studied real datasets of these sizes where epistatic interactions have been validated are not widely available.

By performing a simulation study it is possible to determine the success rate of a method. This is possible because the relevant SNPs are known before the algorithm is applied to the data. The success rate is the percentage of time that a method scores both relevant SNPs above a given threshold. To estimate it, we use all 100 datasets for each of the 30 models. Specifically, the percentage of datasets in which a method ranks the two relevant SNPs above the N^{th} percentile of all SNPs is the estimate of the method's success rate. We examine the 95^{th} percentile because this is likely to be useful in practice and because ReliefF has been used in the genetic analysis of complex diseases in this fashion [14]. This represents the situation where the method filters a dataset with 1000 SNPs to 50 SNPs before a combinatorial analysis is performed on this manageable subset.

It is also important to understand whether differences observed between the estimates of success rates for the various methods are due to chance or are due to a true performance difference. To determine whether differences between success rates at these thresholds are likely due to chance, we apply Fisher's exact test to assess the significance of these differences. Fisher's exact test is a significance test appropriate for categorical count data such as success rate [26]. The resulting p-value for this test can be interpreted as the likelihood of seeing a difference of the size observed among success rates when the methods do not differ. We consider results statistically significant when $p \leq 0.05$. Additionally, we graphically show results for filtering to each percentile from the 99^{th} to the 50^{th}. Highly significant results indicate that the observed differences are unlikely to be due to chance.

We test each method using parameter settings from Greene et al. [24]. ReliefF requires that a number of neighbors be specified. In 2003 Robnik-Sikonja and Kononenko [27] performed a comprehensive analysis and determined that ten neighbors was an appropriate number for ReliefF, so we use ten neighbors here. Similarly, SURF requires a distance threshold. Greene et al. [24] suggest that the mean distance can be used as an acceptable threshold and thus we use the mean distance in this situation. To facilitate comparison between these methods we do not use a distance decay, although in future studies altering this parameter could allow for further improvement in success rate because the distance decay increases the influence of the most extreme individuals.

5 Empirical Results

The novel method, SURF*, which uses both near and far individuals for weighting, significantly outperforms both the SURF and ReliefF methods that use only nearby individuals. Figure 1 shows an example plot for a specific sample size (1600) and heritability (0.2) combination. This figure summarizes the success rate estimated from analysis of 500 simulated independent datasets with this heritability and sample size. The arrows on the right side of the graph indicate whether the methods varied significantly in their abilities to successfully filter a dataset to the 95th percentile (i.e. filter a dataset of 1000 SNPs to 50 SNPs without removing either relevant SNP). In this case the differences between all three

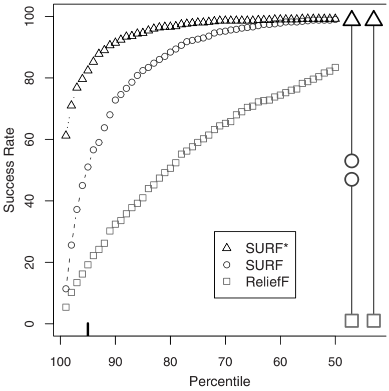

Fig. 1. This figure shows success rate analysis results for an example heritability (0.2) and sample size (1600). The arrows on the right side of the graph indicate whether the methods varied significantly in their abilities to successfully filter a dataset to the 95th percentile (shown by the tick mark above the x-axis between the 100^{th} and 90^{th} percentiles). The caps of the arrows illustrate which two methods are being compared, and the line connecting these caps indicates the level of significance of the differences between this pair of methods (no line represents $p \geq 0.05$, a dotted line represents $0.01 \leq p < 0.05$, a dashed line represents $0.001 \leq p < 0.01$, and a solid line represents $p < 0.001$. In this case the differences between all three methods were highly significant.

methods were highly significant. These results indicate clear differences between these methods for this heritability and sample size. Furthermore the differences observed were highly significant ($p \leq 0.001$) indicating that differences of this magnitude are likely to be observed by chance less than one time out of 1000. While this figure shows clear differences in success rate at this heritability and sample size, it is most informative to consider an algorithm's performance a wide range of potential use cases.

Figure 2 shows results as small multiples of the example shown in Figure 1 across all tested sample sizes and heritabilities. Each plot represents results for 500 datasets with the specified sample size and heritability. None of the methods perform particularly well at the lowest sample sizes and heritabilities. That is, when the genetic effect is smallest, a larger study would be needed to discover the relevant SNPs. This is well known in genetics and, fortunately, studies aiming to detect smaller effects are designed to contain more individuals. Also as expected, at the highest sample sizes and heritabilties all of the methods perform well.

The range where results are most similar to what would be seen in practice, the simulations with 1600 individuals and modest heritabilities, 800 individuals and high heritabilities, and 3200 individuals with lower heritabilities are also the areas where SURF* outperforms other methods by the widest margin. In these bands the differences between SURF* and the other methods are highly statistically significant. These results indicate that SURF* greatly improves upon currently used approaches. SURF*'s consistently high performance indicates that

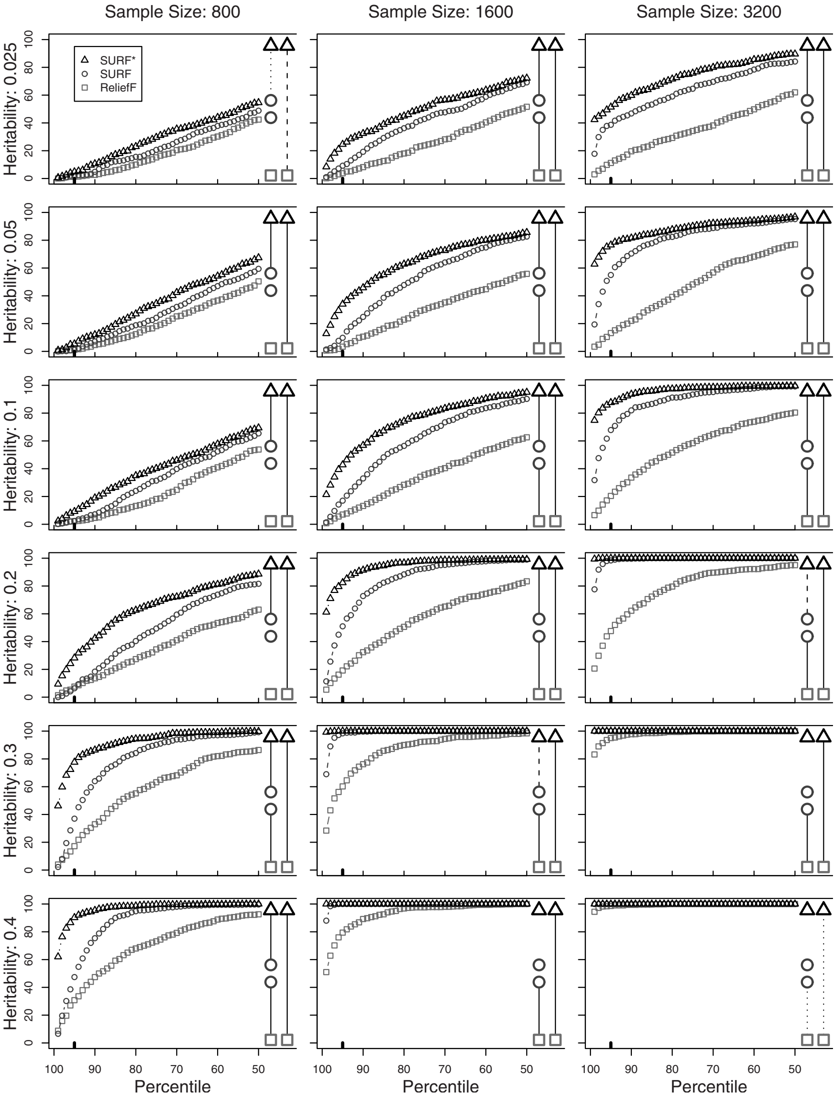

Fig. 2. This is a summary of success rates as shown in Figure 1 across a wide range of sample sizes and heritabilities. The x-axis for each plot corresponds to the percentiles as in Figure 1. The y-axis corresponds to the success rate. Significance is shown with arrows as described in Figure 1. Across these situations, SURF* outperforms both existing methods.

it should be used in place of SURF when the goal is to detect SNPs predictive of disease through epistatic interactions. While here we are most interested in the ability to filter a dataset of 1000 SNPs to a smaller dataset of 50 SNPs which can be combinatorially analyzed, it is important to note that across the entire range of percentiles examined, SURF* outperforms currently used methods. This indicates that when used for more or less stringent filtering, SURF* is still more effective than currently existing methods. Using both the nearest and farthest individuals greatly and significantly improves SURF's ability to detect SNPs which interact to predict disease.

6 Discussion and Conclusions

Epistatic interactions have often been shown to affect complex traits in model organisms, and thus it would be prudent to consider the potential role of epistasis on the complex traits of common human disease susceptibilities [28,29]. Unfortunately epistasis is not often considered because an exhaustive analysis is computationally intractable [20]. Machine learning methods such as SURF offer promise but these approaches must be modified to cope with the small sample sizes and large number of attributes present in high throughput genetic datasets. Our theoretical work in Section 3 suggests that SURF*, which uses a greater number of individuals for attribute weighting than SURF, will be a more powerful way to approach this problem. We observe this effect in our empirical results (Section 5). Using the farthest individuals in addition to the nearest ones greatly increases the success rates of these methods at moderate sample sizes and heritabilities. Additionally, these improvements may generalize to other Relief algorithms and could increase their ability to detect interactions.

Here we examine the role of these Relief algorithms in isolation, but it is important to note that these can be used in conjunction with other information sources as well during a genetic analysis [23]. Improved Relief algorithms should offer an immediate increase in success rate to detect interactions when they are used in place of current algorithms as information sources for these methods. SURF* does perform more weighting due to the increased number of individuals that are used, but with SURF* it is no longer necessary to find the nearest individuals so the computational costs remain relatively similar. A method which provides a significant increase in success rate is likely to improve our understanding of common human diseases.

Future work should focus on effective and efficient methods to assess the significance of discovered SNPs. Relief methods return scores which are a measure of SNP quality but which are not easily converted to statistical significance. Additionally, work should be done to develop powerful Relief methods capable of detecting interactions between discrete and continuous variables and endpoints. Genetic association studies often include SNPs, which are discrete, in addition to measures of the environment, which are continuous. Methods capable of detecting gene-gene, gene-environment, and environment-environment interactions will therefore be useful. Relief methods capable of examining continuous data exist [30,27], but they should be rigorously evaluated for their ability to detect interactions between discrete and continuous attributes. The impact of including farthest individuals on the success of those algorithms should also be examined.

7 Method Availability

SURF* is freely available in the open source MDR software package from `http://sourceforge.net/projects/mdr/`. MDR is a cross-platform Java application which features a user friendly graphical interface.

Acknowledgement

This work was supported by NIH grants LM009012, AI59694, HD047447, and ES007373.

References

1. Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G., Chee, M.S.: A genome-wide scalable SNP genotyping assay using microarray technology. Nat. Genet. 37(5), 549–554 (2005)
2. Steemers, F.J., Gunderson, K.L.: Whole genome genotyping technologies on the BeadArray platform. Biotechnology Journal 2(1), 41–49 (2007)
3. Thomas, D.C., Haile, R.W., Duggan, D.: Recent developments in genomewide association scans: A workshop summary and review. Am. J. Hum. Genet. 77(3), 337–345 (2005)
4. Chanock, S., Taylor, J.G.: Using genetic variation to study immunomodulation. Current Opinion in Pharmacology 2(4), 463–469 (2002)
5. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., Hirschhorn, J.N.: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9(5), 356–369 (2008)
6. Hirschhorn, J.N., Lohmueller, K., Byrne, E., Hirschhorn, K.: A comprehensive review of genetic association studies. Genet. Med. 4, 45–61 (2002)
7. Shriner, D., Vaughan, L.K., Padilla, M.A., Tiwari, H.K.: Problems with Genome-Wide association studies. Science 316(5833), 1840–1841 (2007)
8. Williams, S.M., Canter, J.A., Crawford, D.C., Moore, J.H., Ritchie, M.D., Haines, J.L.: Problems with Genome-Wide association studies. Science 316(5833), 1841–1842 (2007)
9. Jakobsdottir, J., Gorin, M.B., Conley, Y.P., Ferrell, R.E., Weeks, D.E.: Interpretation of genetic association studies: Markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genetics 5(2), e1000337 (2009)
10. Moore, J.H.: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human Heredity 56, 73–82 (2003)
11. Phillips, P.C.: Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems. Nat. Rev. Genet. 9(11), 855–867 (2008)
12. Tyler, A.L., Asselbergs, F.W., Williams, S.M., Moore, J.H.: Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. BioEssays 31(2), 220–227 (2009)
13. Kira, K., Rendell, L.A.: A practical approach to feature selection, pp. 249–256 (1992)
14. Beretta, L., Cappiello, F., Moore, J.H., Barili, M., Greene, C.S., Scorza, R.: Ability of epistatic interactions of cytokine single-nucleotide polymorphisms to predict susceptibility to disease subsets in systemic sclerosis patients. Arthritis and Rheumatism 59(7), 974–983 (2008)

15. The Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145), 661–678 (2007)
16. Gayan, J., Gonzalez-Perez, A., Bermudo, F., Saez, M., Royo, J., Quintas, A., Galan, J., Moron, F., Ramirez-Lorca, R., Real, L., Ruiz, A.: A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. BMC Genomics 9(1), 360 (2008)
17. 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)
18. 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)
19. Freitas, A.A.: Understanding the crucial role of attribute interaction in data mining. Artif. Intell. Rev. 16(3), 177–199 (2001)
20. Moore, J.H., Ritchie, M.D.: The challenges of Whole-Genome approaches to common diseases. JAMA 291(13), 1642–1643 (2004)
21. Cordell, H.: Detecting gene-gene interactions that underlie human diseases. Nature Reviews Genetics 10(6), 392–404 (2009)
22. McKinney, B., Reif, D., White, B., Crowe, J., Moore, J.: Evaporative cooling feature selection for genotypic data involving interactions. Bioinformatics 23(16), 2113–2120 (2007)
23. McKinney, B.A., Crowe, J.E., Guo, J., Tian, D.: Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. PLoS Genet. 5(3), e1000432 (2009)
24. Greene, C.S., Penrod, N.M., Kiralis, J., Moore, J.H.: Spatially Uniform ReliefF (SURF) for Computationally-efficient Filtering of Gene-gene Interactions. BioData Mining 2, 5 (2009)
25. 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)
26. Sokal, R.R., Rohlf, F.J.: Biometry: the principles and practice of statistics in biological research, 3rd edn. W. H. Freeman and Co., New York (1995)
27. Robnik-Sikonja, M., Kononenko, I.: Theoretical and empirical analysis of relieff and rrelieff. Mach. Learn. 53, 23–69 (2003)
28. Kroymann, J., Mitchell-Olds, T.: Epistasis and balanced polymorphism influencing complex trait variation. Nature 435(7038), 95–98 (2005)
29. Shao, H., Burrage, L.C., Sinasac, D.S., Hill, A.E., Ernest, S.R., O'Brien, W., Courtland, H., Jepsen, K.J., Kirby, A., Kulbokas, E.J., Daly, M.J., Broman, K.W., Lander, E.S., Nadeau, J.H.: Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. Proc. Nat. Acad. Sci. 105(50), 19910–19914 (2008)
30. Robnik-Sikonja, M., Kononenko, I.: An adaptation of relief for attribute estimation in regression. In: ICML 1997: Proceedings of the Fourteenth International Conference on Machine Learning, pp. 296–304 (1997)