A Balanced Accuracy Function for Epistasis Modeling in Imbalanced Datasets using Multifactor Dimensionality Reduction

Digna R. Velez,[1] Bill C. White,[2] Alison A. Motsinger,[1] William S. Bush,[1] Marylyn D. Ritchie,[1] Scott M. Williams,[1] and Jason H. Moore[2–5]*

[1] *Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, Tennessee*
[2] *Computational Genetics Laboratory, Departments of Genetics and Community and Family Medicine, Norris-Cotton Cancer Center, Dartmouth Medical School, Lebanon, New Hampshire*
[3] *Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire*
[4] *Department of Computer Science, University of New Hampshire, Durham, New Hampshire*
[5] *Department of Computer Science, University of Vermont, Burlington, Vermont*

Multifactor dimensionality reduction (MDR) was developed as a method for detecting statistical patterns of epistasis. The overall goal of MDR is to change the representation space of the data to make interactions easier to detect. It is well known that machine learning methods may not provide robust models when the class variable (e.g. case-control status) is imbalanced and accuracy is used as the fitness measure. This is because most methods learn patterns that are relevant for the larger of the two classes. The goal of this study was to evaluate three different strategies for improving the power of MDR to detect epistasis in imbalanced datasets. The methods evaluated were: (1) over-sampling that resamples with replacement the smaller class until the data are balanced, (2) under-sampling that randomly removes subjects from the larger class until the data are balanced, and (3) balanced accuracy [(sensitivity+specificity)/2] as the fitness function with and without an adjusted threshold. These three methods were compared using simulated data with two-locus epistatic interactions of varying heritability (0.01, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4) and minor allele frequency (0.2, 0.4) that were embedded in 100 replicate datasets of varying sample sizes (400, 800, 1600). Each dataset was generated with different ratios of cases to controls (1 : 1, 1 : 2, 1 : 4). We found that the balanced accuracy function with an adjusted threshold significantly outperformed both over-sampling and under-sampling and fully recovered the power. These results suggest that balanced accuracy should be used instead of accuracy for the MDR analysis of epistasis in imbalanced datasets. *Genet. Epidemiol.* 31:306–315, 2007. © 2007 Wiley-Liss, Inc.

The supplemental materials described in this article can be found at http://www.interscience.wiley.com/jpages/0741-0395/suppmat.
Contract grant sponsor: National Institutes of Health Grants; Contract grant numbers: AI59694, HD047447, LM009012, RR018787, GM62758, AG20135.

*Correspondence to: Jason H. Moore, Ph.D., 706 Rubin Bldg, HB7937, Dartmouth Medical School, One Medical Center Dr. Lebanon, NH 03756. E-mail: jason.h.moore@dartmouth.edu; Web:www.epistasis.org

Received 19 August 2006; Revised 21 October 2006; Accepted 28 December 2006
Published online 23 February 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20211

INTRODUCTION

A central goal of genetic epidemiology is to determine the genetic architecture of common human diseases. Success in this endeavor is highly dependent on assumptions made about the number of polymorphisms that play a role in determining disease susceptibility and the nature of the mapping relationship between genotype and phenotype. Several phenomena such as phenocopy, trait heterogeneity, locus heterogeneity, plastic reaction norms (gene-environment interaction), and epistasis (gene-gene interaction) may greatly complicate the genetic architecture of biomedical traits such as disease susceptibility [reviewed in Thornton-Wells et al., 2004]. The focus of this study is on nonlinearities in the genotype to phenotype relationship arising from epistasis. Epistasis is believed to play an important role in the genetic architecture of many common human diseases [e.g. Templeton, 2000; Cordell, 2002; Moore, 2003; Thornton-Wells et al., 2004; Moore and Williams, 2005; Nagel, 2005].

Epistasis has been recognized for many years as deviations from the simple inheritance patterns observed by Mendel [Bateson, 1909] or deviations from additivity in a linear statistical model [Fisher, 1918] and is likely due, in part, to canalization or mechanisms of stabilizing selection that evolve robust gene networks [Waddington, 1942, 1957; Gibson and Wagner, 2000; Proulx and Phillips, 2005]. Epistasis has been defined in multiple ways [e.g. Hollander et al., 1955; Philips, 1998; Brodie, 2000]. We have reviewed two types of epistasis, biological and statistical, and their relationship to each other [Moore and Williams, 2005]. Biological epistasis results from physical interactions between biomolecules (e.g. DNA, RNA, proteins, and enzymes) and occur at the cellular level in an individual. This is the type of epistasis that Bateson [1909] referred to when he coined the term. Statistical epistasis on the other hand occurs at the population level and is realized when there is interindividual variation in DNA sequences. The statistical phenomenon of epistasis is the focus of Fisher's [1918] definition. Understanding the relationship between biological and statistical epistasis is difficult but is important if we are to make biological inferences from statistical results [Moore and Williams, 2005].

Detecting statistical patterns of epistasis in human populations is difficult due to the sparseness of the available data in multiple dimensions. The combination of data sparseness and nonlinearity in the relationship between genotype and phenotype has motivated the development of multiple analytical approaches ranging from those based on parametric statistical models such as logistic regression [Millstein et al., 2005] to those based on machine learning and data mining methods [McKinney et al., 2006]. One of these methods, multifactor dimensionality reduction (MDR) was developed as a nonparametric (i.e. no parameters are estimated) and genetic model-free (i.e. no genetic model is assumed) data mining strategy for identifying combinations of single-nucleotide polymorphisms (SNPs) and other discrete factors such as smoking that are predictive of a discrete clinical endpoint [Ritchie et al., 2001, 2003; Hahn et al., 2003; Hahn and Moore, 2004; Moore, 2004; Moore et al., 2006; Moore, 2007]. The goal of MDR is to change the representation space of the data to make nonlinear interactions easier to detect and characterize. Thus, MDR can be seen in a broader sense as a data processing step that precedes classification [Moore et al., 2006; Moore, 2007]. At the heart of the MDR approach is a feature or attribute construction algorithm that creates a new discrete attribute by pooling levels from multiple discrete factors [Moore et al., 2006; Moore, 2007]. The process of defining a new attribute as a function of two or more other attributes is referred to as constructive induction or attribute construction and was first described by Michalski et al. [1983]. Constructive induction using MDR is accomplished in the following way. Given a threshold *T*, a multilocus genotype combination, for example, is considered "high risk" if the ratio of cases to controls exceeds *T*; otherwise it is considered low risk. Once genotypes are labeled "high risk" and "low risk" a new binary attribute is created with those two levels. Figure 1 illustrates this process for a dataset of 200 cases and 200 controls that was simulated using the penetrance function in Table I.

Once an MDR attribute is constructed, it can be statistically evaluated using any classification method such as naïve Bayes, decision trees, or logistic regression [Moore et al., 2006]. We have traditionally used a naive Bayes classifier since MDR attributes are single variables with two levels [Hahn and Moore, 2004]. It is also possible to add this new attribute back to the original dataset to be evaluated with all other attributes in a process called interleaving [Moore et al., 2006; Moore, 2007]. Computational methods such as bootstrapping, cross-validation, and permutation testing can be employed as wrappers to MDR-based constructive induction and classification to facilitate identification of a best set of predictors and their model. For the purposes of this study we change the representation of the data using MDR as shown in Figure 1 and use a naïve Bayes algorithm for classification. This approach will collectively be referred to simply as MDR for the remainder of the manuscript. The MDR method has been successfully applied to detecting gene-gene and gene-environment interactions for a variety of clinical endpoints including, for example, adverse drug reactions [Wilke et al., 2005a,b], Alzheimer disease [Martin et al., 2006], asthma [Chan et al., 2006; Millstein et al., 2006], atrial fibrillation [Tsai et al., 2004; Moore et al., 2006; Motsinger et al., 2006; Asselbergs et al., 2006], autism [Ma et al., 2005; Ashley-Koch et al., 2006], bladder cancer [Andrew et al., 2006], familial amyloid polyneuropathy [Soares et al., 2005], hypertension [Williams et al., 2004; Sanada et al., 2006], multiple sclerosis [Brassat et al., 2006], myocardial infarction [Coffey et al., 2004; Mannila et al., 2006], prostate cancer [Xu et al., 2005], schizophrenia [Qin et al., 2005], sporadic breast cancer [Ritchie et al., 2001], and type II diabetes [Cho et al., 2004].

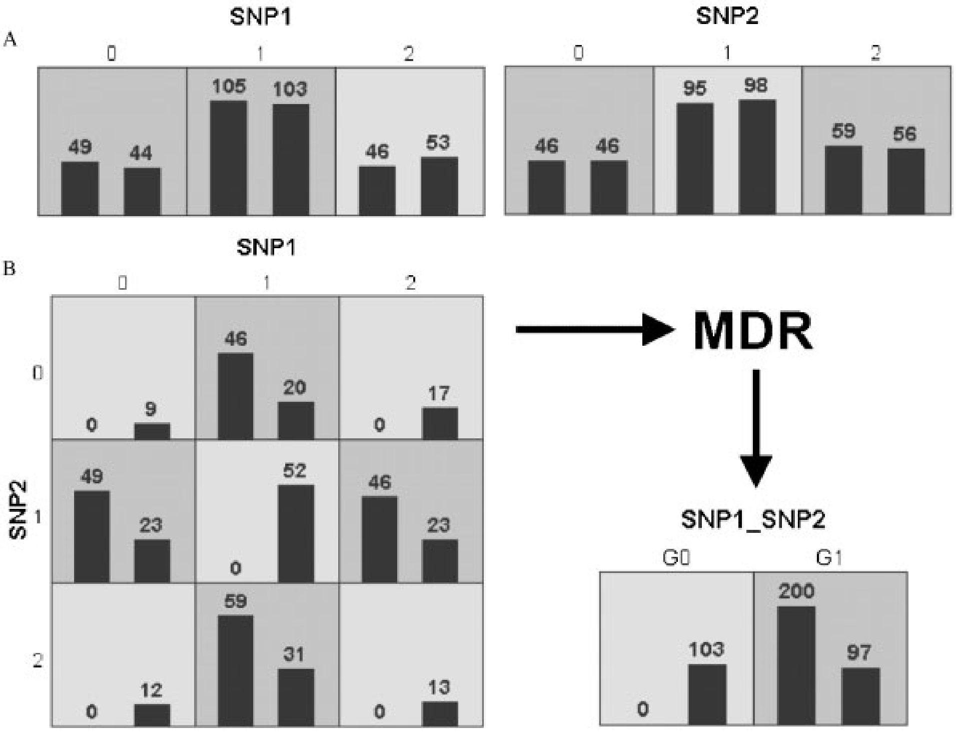

Fig. 1. MDR attribute construction. (A) Distribution of cases (left bars) and controls (right bars) for each of the three genotypes of SNP1 and SNP2. The dark-shaded cells have been labeled "high risk" using a threshold of $T = 1$. The light-shaded cells have been labeled "low risk". (B) Distribution of cases and controls when the two functional SNPs are considered jointly. A new single attribute is constructed by pooling the "high risk" genotype combinations into one group (G1) and the "low risk" into another group (G0).

TABLE I. Penetrance values for a two-locus M170 epistatic model

	AA (0.25)	Aa (0.50)	aa (0.25)
BB (0.25)	0.0	0.1	0.0
Bb (0.50)	0.1	0.0	0.1
bb (0.25)	0.0	0.1	0.0

Genotype frequencies labeled in parentheses.

Despite these successes, there remain a number of questions about the optimal way to use MDR for constructive induction and classification. For example, what is the optimal threshold, *T*, for defining high-risk and low-risk genotype combinations when the numbers of subjects in each class (e.g. cases and controls) are not equal? What is the optimal function for assessing the quality of a classifier when the classes are imbalanced? These issues are not unique to MDR. In fact, the presence of imbalanced classes is an issue for data mining and classification in multiple disciplines including business, industry, and scientific research [e.g. Chawla et al., 2004]. The two most common approaches to correct for class-imbalance involve reweighing or resampling data. Reweighing approaches assign a high cost or weight to the misclassification of the minority class to minimize overall cost sensitivity, an approach known as cost sensitivity learning [Pazzani et al., 1994; Domingos, 1999]. Resampling techniques seek to create balanced datasets by sampling the minority class with replacement (over-sampling) or randomly removing instances from the majority class (under-sampling). Studies by Chawla et al. [2003, 2004] and Japkowicz [2000] provide extensive examples of these resampling techniques. Resampling schemes include: over-sampling, under-sampling, directed over-sampling (the choice of samples to replace is informed, based on previous knowledge of the data), directed under-sampling (the choice of samples to remove is informed) and combinations of resampling procedures [Drummond and Holte, 2003; Chawla et al., 2004]. Other studies have examined the effect of resampling in the training set classification distribution on method performance instead of resampling the original data [Catlett, 1991; Chan and Stolfo, 1998]. These resampling techniques have been the most frequently studied candidate solutions to the class-imbalance problem.

Other approaches are based on alterations at the algorithmic level. These include: (1) adjusting the decision threshold and (2) recognition-based (learning from one class) as opposed to discrimination-based (two classes) learning [Japkowic, 2000; Chawla, 2003; Chawla et al., 2004]. Another approach utilizes balanced accuracy in weighing the performance of a model. This approach weighs the two classes equally and it is thought to be more powerful than using accuracy alone in circumstances of a rare affected class [Powers et al., 2005]. Balanced accuracy allows performance of a model to be weighed equally between the two classes, regardless of the class sizes [Mower, 2005; Powers et al., 2005].

The goal of this study was to evaluate several different approaches for detecting epistasis in imbalanced datasets using MDR. Using simulated data of varying epistatic pattern, heritability, sample size, and allele frequency, we evaluated the power of MDR to detect two-locus models of gene-gene interactions while accounting for imbalanced classes using (1) over-sampling, (2) under-sampling, and (3) balanced accuracy with and without a modified threshold (*T*). We show that combining the use of a threshold equal to the ratio of the classes with the balanced accuracy function has the highest power across a wide range of different datasets. The results of this study provide an important framework for MDR constructive induction and classification in genetic epidemiology datasets of all shapes and sizes.

METHODS

DATA SIMULATION

The goal of the simulation study was to generate artificial datasets that could be used to evaluate the power of MDR to detect gene-gene interactions when the datasets have imbalanced classes. We developed a total of 70 different penetrance functions that define a probabilistic relationship between genotype and phenotype where susceptibility to disease is dependent on genotypes from two loci in the absence of any marginal effects. The models were distributed evenly across seven broad-sense heritabilities (0.01, 0.025, 0.05, 0.1, 0.2, 0.3, and 0.4) and two different minor allele frequencies (0.2 and 0.4). A total of five models for each of the 14 heritability-allele frequency combinations were generated for a total of 70 models. The details of the 70 penetrance functions are available online in Supplementary Table 1. Genotype frequencies for all 70 epistasis models were consistent with Hardy–Weinberg proportions. One hundred data sets were generated for each model with three sample sizes (400, 800, and 1600 total individuals) and case-control proportions of 1 : 1, 1 : 2 and 1 : 4. Each pair of functional polymorphisms was embedded within a set of 10 independent SNPs. A total of 42,000 datasets were generated and analyzed. The data are available upon request.

In addition to the data described above, we also analyzed datasets generated using the penetrance function described in Table I. The epistatic pattern in this model is based on the nonlinear XOR function. This model (M170) was selected for illustrative purposes and has been previously described by Li and Reich [2000]. The M170 model used here has a minor allele frequency of 0.5 and a broad-sense heritability of 0.05.

EPISTASIS ANALYSIS USING MDR

As described above, the goal of MDR is to change the representation space of the data to make interactions easier to detect. This is accomplished by combining two or more attributes into a single attribute that can be modeled using a discrete data classifier. Here, we used a simple probabilistic classifier that is similar to naïve Bayes [Hahn and Moore, 2004]. Naïve Bayes works well with MDR since there is only a single variable with two levels. For each dataset we evaluated all possible MDR attributes that are functions of one to five SNPs. Five-fold cross validation [Motsinger and Ritchie, 2006] was used to select an MDR model with maximum testing accuracy (i.e. most likely to generalize) and maximum cross-validation consistency as described previously [Ritchie et al., 2001, 2003; Hahn et al., 2003; Moore, 2004]. In this paper the term accuracy is used in place of maximum testing accuracy. Power was estimated as the number of times MDR correctly identified the two functional SNPs out of each set of 100 balanced and imbalanced datasets. An open-source, freely available, and platform-independent MDR software package is available from www.epistasis.org. An MDR analysis tutorial is available at compgen.blogspot.com.

METHODS FOR ADDRESSING CLASS IMBALANCE

We evaluated the following three methods for addressing the imbalance of the classes: (1) over-sampling, (2) under-sampling, and (3) balanced accuracy with and without threshold (*T*) adjustments. Over-sampling and under-sampling were performed on the imbalanced data to achieve a 1 : 1 proportion of cases and controls. Over-sampled datasets were created by randomly resampling cases with replacement until the number of cases equaled the number controls. Under-sampled datasets were created by using all cases and randomly removing subjects from the control group until the number of controls equaled the number of cases.

Balanced accuracy was applied to an MDR attribute constructed with an adjusted threshold (adjusted to the ratio of cases and controls) and an unadjusted threshold of 1.0. Balanced accuracy is defined as the arithmetic mean of sensitivity and specificity:

$$\frac{(\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})+\mathrm{TN}/(\mathrm{TN}+\mathrm{TP}))}{2} = \frac{(\mathrm{sensitivity}+\mathrm{specificity})}{2},$$

where TP are true positives, TN are true negatives, and FN are false negatives. Weighing the two classes equally accounts for both positive and negative errors caused by the class imbalance. Balanced accuracy is algebraically identical to accuracy when datasets are completely balanced.

VISUALIZATION OF RESULTS

Results for each set of 100 simulated datasets are reported as power. For each sample size, heritability, and minor allele frequency, we averaged the power across the five models to produce one power estimate. We used segment diagrams to display the averaged power results. Each segment within a segment plot represents the particular method employed under a given imbalanced proportion of cases to controls (1 : 2 or 1 : 4). The radius of each segment represents the difference in power of MDR applied to datasets with a 1 : 1 proportion of cases to controls (i.e. optimal results) and MDR applied to imbalanced datasets using each of the methods for addressing the class imbalance. Shorter segments indicate there was less of a difference between the baseline power observed in balanced datasets, and the power observed in imbalanced datasets suggesting that the method had good performance. Longer segments indicate there was a larger difference in power, indicating a particular method performed poorly.

RESULTS

The power of MDR to detect the correct two functional SNPs for the M170 epistatic model is presented in Table II. For all sample sizes, accuracy (i.e. the proportion of subjects who are correctly classified by the model in the training dataset) and balanced accuracy with an unmodified threshold of 1.0 had 0% power to detect the correct model for a 1 : 4 ratio of cases to controls. The power was slightly higher when the ratio of imbalance was 1 : 2 with power ranging from 28 to 44% for accuracy and 17 to 28% for balanced accuracy. Adjusting the threshold to the ratio of

TABLE II. Power of MDR to detect the correct two-locus M170 epistatic model

		Power (%)					
						Thresh 1	
		Threshold 1		Adjusted threshold		Under-sample	Over-sample
Sample Size	Ratio	Accuracy	Balanced Accuracy	Accuracy	Balanced Accuracy	Accuracy	
	1 : 1	92	92	—	—	—	—
400	1 : 2	28	28	88	94	93	76
	1 : 4	0	0	0	96	95	61
	1 : 1	98	98	—	—	—	—
800	1 : 2	44	17	87	95	96	74
	1 : 4	0	0	0	92	97	66
	1 : 1	99	99	—	—	—	—
1600	1 : 2	42	26	94	98	98	89
	1 : 4	0	0	0	92	96	64

MDR, multifactor dimensionality reduction.

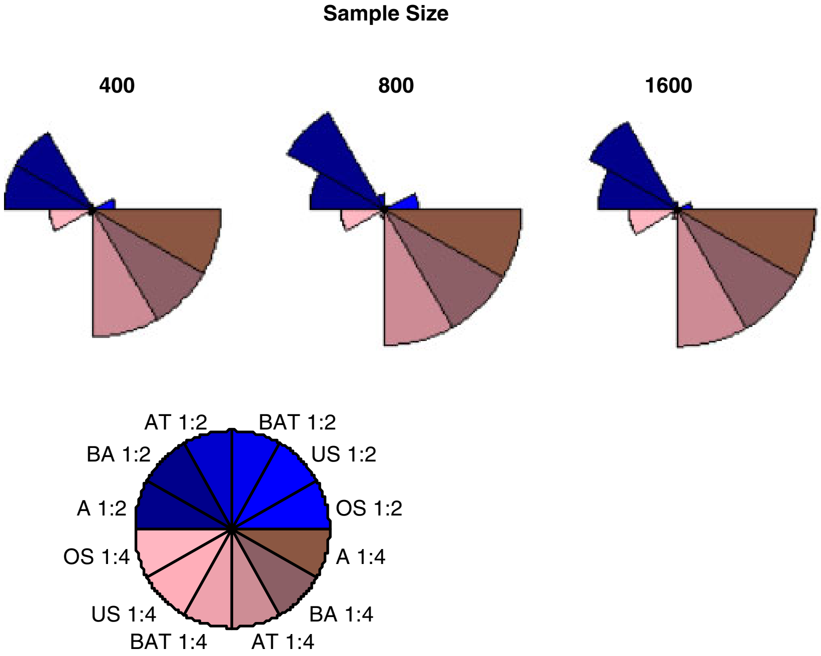

Fig. 2. Segment plots of the results for the M170 model. Blue indicates an imbalanced ratio of 1 : 2 and pink indicates a ratio of 1 : 4. The labels for the legend are: BAT, balanced accuracy with an adjusted threshold; BA, balanced accuracy with a threshold of 1; AT, accuracy with an adjusted threshold; A, accuracy with a threshold of 1; US, under-sampling; and OS, over-sampling. Note that the segments for BAT are the shortest suggesting that the balanced accuracy function recovers most of the power and outperforms the other methods.

cases to controls significantly increased the power of both functions to identify the correct functional SNPs when the ratio of imbalance was 1 : 2, with power ranging from 88 to 94% for accuracy and 94 to 98% for balanced accuracy. Accuracy with an adjusted threshold had 0% power with a 1 : 4 ratio of imbalance and performed poorly across all sample sizes. In contrast, balanced accuracy with an adjusted threshold substantially outperformed accuracy with an adjusted threshold for a 1 : 4 ratio, with power ranging from 92 to 96%. Note that small differences between accuracy and balanced accuracy are due to stochastic splits of the data during cross-validation.

Under-sampling also performed well with results comparable to balanced accuracy with an adjusted threshold. The power with under-sampling ranged from 93 to 98%. Over-sampling performed slightly worse than under-sampling and balanced accuracy with an adjusted threshold with power ranging from 61 to 89%.

The M170 power results are visually represented with segment plots in Figure 2 (1 : 2 imbalanced ratios are labeled in blue and 1 : 4 ratios are labeled in pink). Each segment represents the power for each class-imbalance approach to its corresponding power had the dataset been balanced. As can be seen in the figure, the largest segments (lowest power) are those with accuracy and balanced accuracy with an unmodified threshold for both imbalanced ratios. The approaches that recover the most power are the smallest segments on the plots. The best approaches across all sample sizes and ratios of imbalance are accuracy and balanced accuracy with an adjusted threshold and under-sampling.

Figure 3 visually illustrates the power results for the 70 penetrance models. Figure 3A represents the power results for a minor allele frequency of 0.2, whereas figure 3B represents the power results for a minor allele frequency of 0.4. Sample sizes are labeled at the top of each column while the heritabilities are labeled for each row. Heritability ranging from 0.01 to 0.1 had comparable segment patterns between both 0.2 and 0.4 minor allele frequencies. The largest segments for a 1 : 4 ratio, for these heritabilities, were accuracy and balanced accuracy with thresholds of 1, and accuracy with an adjusted threshold and for a 1 : 2 ratio the largest segments were accuracy and balanced accuracy with thresholds of 1. Larger heritabilities were more robust to class imbalance as is apparent from the decreasing segment sizes as heritabilites increases. Balanced accuracy with

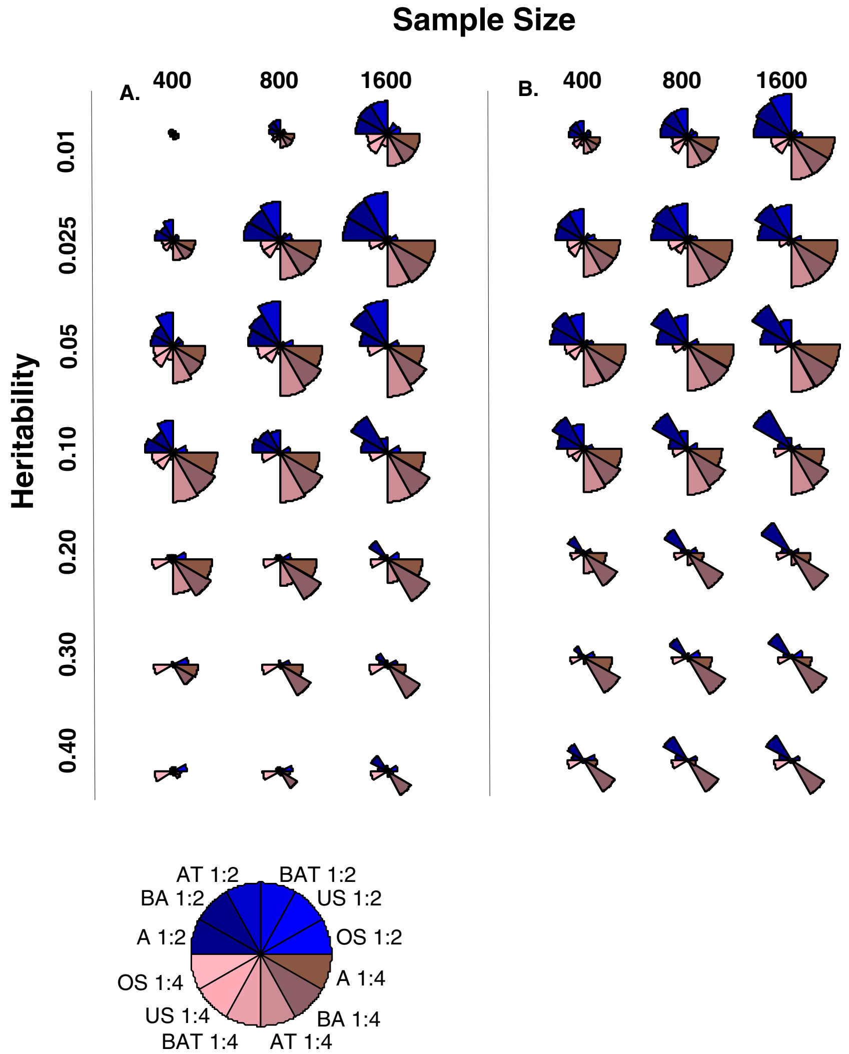

Fig. 3. Segment plots for all 70 penetrance models. Heritabilities are labeled vertically ranging from 0.01 through 0.4. (A) Minor allele frequency of 0.2. (B) Minor allele frequency of 0.4. Note that the segments for BAT are consistently short suggesting that the balanced accuracy function recovers most of the power and outperforms the other methods. Colors are defined in the legend for Figure 2.

an unmodified threshold was the worst approach, as it was always one of the largest segments across heritabilities. These results agree with the results from the M170 penetrance model. Balanced accuracy with an adjusted threshold seems to be the best approach, apparent from the smaller segment size, across heritabilities and sample sizes. The raw power results for all penetrance models are available on-line in Supplemental Tables 2–10.

DISCUSSION

MDR uses a novel attribute construction algorithm to facilitate the detection of nonlinear interactions among multiple discrete genetic and/or environmental factors that are predictive of a discrete endpoint such as case-control status. The goal of MDR is to change the representation space of the data to make high-order interactions easier to detect using computational and/or statistical classifiers. As such, MDR should be viewed as a method that complements, rather than replaces, other methods since the results of an MDR analysis can be used to change the dataset before analysis using neural networks, classification trees, or logistic regression.

The goal of this study was to determine the best approach for implementing MDR when the class labels are not equal (i.e. imbalanced). Our results suggest that the balanced accuracy function should be used instead of accuracy as the fitness measure when MDR is used with a classifier such as naïve Bayes to detect epistasis in imbalanced datasets. Although under-sampling was consistently better than over-sampling, neither was consistently better than using balanced accuracy with an adjusted threshold. The poor performance of over-sampling supports the conclusions of Chawla et al. [2002] demonstrating that over-sampling data spread out the decision regions making them larger and less specific thus decreasing the number of observations that can be correctly predicted. Under-sampling, on the other hand, proved to have more reliable results. As Drummond and Holte (2003) explain, under-sampling increases the generalizability of one class by removing instances of the other class. Our results suggest that under-sampling increases the sensitivity to the misclassification costs and class distribution. At face value resampling approaches are appealing. However, as alluded to above, there are some important pitfalls. For example, under-sampling throws out potentially useful data, whereas over-sampling increases the size of the training set and increases the time to build a classifier, as well as introducing a false sense of power and leading to possible type I errors. These represent real drawbacks to both procedures.

Chawla et al. [2002] and Weiss and Provost [2001] have both performed studies supporting adjustments to threshold and decision branches in an algorithm to avoid sampling techniques. However, our results do not completely support their conclusions. Modifying the threshold alone while still using accuracy did not work because accuracy allows over-represented groups (e.g. the controls) to have more weight in the calculation. Using balanced accuracy along with a threshold of 1.0 did not work either because high- and low-risk group would still be incorrectly classified. Adjusting the threshold and applying balanced accuracy simultaneously corrected for the decreased ratio of cases to controls. Balanced accuracy did this by averaging the proportion of cases correctly predicted by the model with the proportion of controls correctly predicted.

The class-imbalance issue has created debate in the data mining community as to whether a dataset should represent the naturally occurring class distribution. To minimize the effect that training set size has on classifier performance it is important to choose the training data carefully; which involves choosing an appropriate class distribution in the training dataset [Weiss and Provost, 2001]. Weiss and Provost [2001] argue that a naturally occurring class distribution produces "classifiers that perform poorly on minority-class examples" and a naturally occurring class distribution should not be used for learning in circumstances of imbalance. They performed studies that support the notion that when there is class-imbalance a class distribution other than the natural class distribution needs to be chosen. Thus, they are proponents of resampling techniques. Our results showing that under-sampling generally performed well supports this idea. However, our results show that resampling methods did not perform as well as using balanced accuracy with an appropriate threshold.

The MDR approach carries out constructive induction by first comparing the distribution of cases and controls for each multilocus genotype combination to a threshold (T). Genotype combinations are assigned to one of two groups based on this threshold and then pooled to form a new attribute or variable with two levels. This new variable is then evaluated using a discrete data classifier. The results of this study show that a threshold equal to the ratio of cases to controls should be used during the constructive induction phase and then combined with a classifier that uses balanced accuracy when the classes are not balanced. On the basis of these results, we have implemented this new approach in an open-source and freely available MDR software package that can be downloaded from www.epistasis.org. This approach substantially improves the power of MDR for detecting epistasis in genetic studies where the ratio of cases to controls is not equal.

ACKNOWLEDGMENTS

We thank Nate Barney and William Holden for their expert assistance with the design, development, programming, and support of the open-source MDR software package. We thank Scott Dudek and Todd Edwards for their technical assistance. We also thank the anonymous reviewers for their very helpful comments and suggestions.

REFERENCES

Andrew AS, Nelson HH, Kelsey KT, Moore JH, Meng AC, Casella DP, Tosteson TD, Schned AR, Karagas MR. 2006. Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility. Carcinogenesis 27:1030–1037.

Ashley-Koch AE, Mei H, Jaworski J, Ma DQ, Ritchie MD, Menold MM, Delong GR, Abramson RK, Wright HH, Hussman JP, Cuccaro ML, Gilbert JR, Martin ER, Pericak-Vance MA. 2006. An analysis paradigm for investigating multi-locus effects in complex disease: examination of three GABA receptor subunit genes on 15q11-q13 as risk factors for autistic disorder. Ann Hum Genet 70:281–292.

Asselbergs FW, Moore JH, van den Berg MP, Rimm EB, de Boer RA, Dullaart RP, Navis G, van Gilst WH. 2006. A role for CETP TaqIB polymorphism in determining susceptibility to atrial fibrillation: a nested case control study. BMC Med Genet 7:39.

Bateson W. 1909. Mendel's Principles of Heredity. Cambridge: Cambridge University Press.

Brassat D, Motsinger AA, Caillier SJ, Erlich HA, Walker K, Steiner LL, Cree BA, Barcellos LF, Pericak-Vance MA, Schmidt S, Gregory S, Hauser SL, Haines JL, Oksenberg JR, Ritchie MD. 2006. Multifactor dimensionality reduction reveals gene-gene interactions associated with multiple sclerosis susceptibility in African Americans. Genes Immun 7:310–315.

Brodie III ED. 2000. Why evolutionary genetics does not always add up. In: Wolf J, Brodie III B, Wade M, editors. Epistasis and the Evolutionary Process. New York: Oxford University Press. p 3–19.

Catlett J. 1991. Megainduction: machine learning on very large databases. Ph.D. thesis, Department of Computer Science, University of Sydney.

Chan IH, Leung TF, Tang NL, Li CY, Sung YM, Wong GW, Wong CK, Lam CW. 2006. Gene-gene interactions for asthma and plasma total IgE concentration in Chinese children. J Allergy Clin Immunol 117:127–133.

Chan P, Stolfo S. 1998. Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, p 164–168.

Chawla N, Japkowicz N, Kolcz A. editors, 2003. Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Data Sets http://www.site.uottawa.ca/~nat/Workshop2003/workshop 2003.html, August 2003.

Chawla N, Japkowicz N, Kolcz A. editors. 2004. SIGKDD Explorations, Special Issue on Class-imbalances. SIGKDD Explorations 6:1–6.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: Synthetic minority over-sampling technique. J Arti Intel Res 16: 321–357.

Cho YM, Ritchie MD, Moore JH, Park JY, Lee KU, Shin HD, Lee HK, Park KS. 2004. Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. Diabetologia 47:549–554.

Coffey CS, Hebert PR, Krumholz HM, Morgan TM, Williams SM, Moore JH. 2004. Reporting of model validation procedures in human studies of genetic interactions. Nutrition 20:69–73.

Cordell HJ. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11:2463–2468.

Domingos P. 1999. MetaCost: a general method for making classifiers cost-sensitive. Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, p 155–164.

Drummond C, Holte R. 2001. Explicitly representing expected cost: an alternative to ROC representation. Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge of Discovery and Data Mining, p 198–207.

Drummond C, Holte R. 2003. C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. Proceedings of the ICML-2003 Workshop: Learning with Imbalanced Data Sets II.

Fisher RA. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinburgh 52:399–433.

Gibson G, Wagner G. 2000. Canalization in evolutionary genetics: a stabilizing theory? Bioessays 22:372–380.

Hahn LW, Moore JH. 2004. Ideal discrimination of discrete clinical endpoints using multilocus genotypes. In Silico Biol 4:183–194.

Hahn LW, Ritchie MD, Moore JH. 2003. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19:376–382.

Hollander WF. 1955 Epistasis and hypostasis. J Hered 46:222–225.

Japkowicz N. 2000. The Class-imbalance Problem: Significance and Strategies. Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning.

Li W, Reich J. 2000. A complete enumeration and classification of two-locus disease models. Hum Hered 50:334–349.

Ma DQ, Whitehead PL, Menold MM, Martin ER, shley-Koch AE, Mei H, Ritchie MD, Delong GR, Abramson RK, Wright HH, Cuccaro ML, Hussman JP, Gilbert JR, Pericak-Vance MA. 2005. Identification of significant association and gene-gene interaction of GABA receptor subunit genes in autism. Am J Hum Genet 77:377–388.

Mannila MN, Eriksson P, Ericsson CG, Hamsten A, Silveira A. 2006. Epistatic and pleiotropic effects of polymorphisms in the fibrinogen and coagulation factor XIII genes on plasma fibrinogen concentration, fibrin gel structure and risk of myocardial infarction. Thromb Haemost 95:420–427.

Martin ER, Ritchie MD, Hahn L, Kang S, Moore JH. 2006. A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. Genet Epidemiol 30:111–123.

McKinney BA, Reif DM, Ritchie MD, Moore JH. 2006. Machine learning for detecting gene-gene interactions: a review. Appl Bioinformatics 5:77–88.

Michalski RS. 1983. A theory and methodology of inductive learning. Artif Intell 20:111–161.

Millstein J, Conti DV, Gilliland FD, Gauderman WJ. 2006. A testing framework for identifying susceptibility genes in the presence of epistasis. Am J Hum Genet 78:15–27.

Millstein J, Siegmund KD, Conti DV, Gauderman WJ. 2005. Identifying susceptibility genes by using joint tests of association and linkage and accounting for epistasis. BMC Genet 6 (Suppl 1):S147.

Moore JH. 2003. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 56: 73–82.

Moore JH. 2004. Computational analysis of gene-gene interactions using multifactor dimensionality reduction. Expert Rev Mol Diagn 4:795–803.

Moore JH, Williams SM. 2002. New strategies for identifying gene-gene interactions in hypertension. Ann Med 34:88–95.

Moore JH. 2007. Genome-wide analysis of epistasis using multifactor dimensionality reduction: feature selection and construction in the domain of human genetics. In: Zu X, Davidson I. editors. Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data. Hershey: IGI Press.

Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC. 2006. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J. Theor Biol 241:252–261.

Moore JH, Williams SM. 2005. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bioessays 27:637–646.

Motsinger AA, Donahue BS, Brown NJ, Roden DM, Ritchie MD. 2006. Risk factor interactions and genetic effects associated with post-operative atrial fibrillation. Pacific Symposium on Biocomputing, p 584–595.

Motsinger AA, Ritchie MD. 2006. The effect of reduction in cross-validation intervals on the performance of multifactor dimensionality reduction. Genet Epidemiol 30:546–555.

Mower JP. 2005. PREP-Mt: predictive RNA editor for plant mitochondrial genes. BMC Bioinforma 6:96.

Nagel RL. 2005. Epistasis and the genetics of human diseases. C R Biol 328:606–615.

Pazzani M, Merz C, Murphy P, Ali K, Hume T, Brunk C. 1994. Reducing misclassification costs. Proceedings of the 11th International Conference on Machine Learning, p 217–225.

Phillips PC. 1998. The language of gene interaction. Genetics 149:1167–1171.

Powers R, Goldszmidt M, Cohen I. 2005. Short term performance forcasting in enterprise systems. Hewlett-Packard Development Company Technical Reports, Computer Science Department, Stanford University, Stanford, CA 94305 http://www.hpl.hp.com/techreports/2005/HPL-2005-50.pdf

Proulx S, Phillips PC. 2005. The opportunity for canalization and the evolution of genetic networks. Am Nat 165:147–162.

Provost F, Fawcett T. 2001. Robust classification for imprecise environments. Machine Learning 42:203–231.

Qin S, Zhao X, Pan Y, Liu J, Feng G, Fu J, Bao J, He L. 2005. An association study of the N-methyl-D-aspartate receptor subunit gene (GRIN1) and NR2B subunit gene (GRIN2B) in schizophrenia with universal DNA microarray. Eur J Hum Genet 13:807–814.

Ritchie MD, Hahn LW, Moore JH. 2003. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24:150–157.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–147.

Sanada H, Yatabe J, Midorikawa S, Hashimoto S, Watanabe T, Moore JH, Ritchie MD, Williams SM, Pezzullo JC, Sasaki M, Eisner GM, Jose PA, Felder RA. 2006. Single-nucleotide polymorphisms for diagnosis of salt-sensitive hypertension. Clin Chem 52:352–360.

Soares ML, Coelho T, Sousa A, Batalov S, Conceicao I, Sales-Luis ML, Ritchie MD, Williams SM, Nievergelt CM, Schork NJ, Saraiva MJ, Buxbaum JN. 2005. Susceptibility and modifier genes in Portuguese transthyretin V30M amyloid polyneuropathy: complexity in a single-gene disease. Hum Mol Genet 14: 543–553.

Templeton AR. 2000. Epistasis and complex traits. In: Wolf J, Brodie III B, Wade M, editors. Epistasis and the Evolutionary Process. New York: Oxford University Press.

Thornton-Wells TA, Moore JH, Haines JL. 2004. Genetics, statistics and human disease: analytical retooling for complexity. Trends Genet 20:640–647.

Tsai CT, Lai LP, Lin JL, Chiang FT, Hwang JJ, Ritchie MD, Moore JH, Hsu KL, Tseng CD, Liau CS, Tseng YZ. 2004. Renin-angiotensin system gene polymorphisms and atrial fibrillation. Circulation 109:1640–1646.

Waddington CH. 1942. Canalization of development and the inheritance of acquired characters. Nature 150:563–565.

Waddington CH. 1957. The Strategy of the Genes. New York: MacMillan.

Weiss GM, Provost F. 2001. The effect of class distribution on classifier learning: an empirical study, Technical Report ML-TR 44. Department of Computer Science, Rutgers University.

Wilke RA, Reif DM, Moore JH. 2005a. Combinatorial pharmacogenetics. Nat Rev Drug Discov 4:911–918.

Wilke RA, Moore JH, Burmester JK. 2005b. Relative impact of CYP3A genotype and concomitant medication on the severity of atorvastatin-induced muscle damage. Pharmacogenet Genomics 15:415–421.

Williams SM, Ritchie MD, Phillips JA III, Dawson E, Prince M, Dzhura E, Willis A, Semenya A, Summar M, White BC, Addy JH, Kpodonu J, Wong LJ, Felder RA, Jose PA, Moore JH. 2004. Multilocus analysis of hypertension: a hierarchical approach. Hum Hered 57:28–38.

Xu J, Lowey J, Wiklund F, Sun J, Lindmark F, Hsu FC, Dimitrov L, Chang B, Turner AR, Liu W, Adami HO, Suh E, Moore JH, Zheng SL, Isaacs WB, Trent JM, Gronberg H. 2005. The interaction of four genes in the inflammation pathway significantly predicts prostate cancer risk. Cancer Epidemiol Biomarkers Prev 14:2563–2568.