AUSTRALIAN JOURNAL OF INTELLIGENT INFORMATION PROCESSING SYSTEMS

Volume 14, No. 1
2014

Editorial Policy

The Australian Journal of Intelligent Information Processing Systems is an interdisciplinary forum for providing the latest information on research developments and related activities in the design and implementation of intelligent information processing systems.

The areas of interest include, but are not limited to:

- artificial intelligence
- artificial neural networks
- human centred computing
- fuzzy systems
- virtual reality

The journal both theoretical and application-oriented papers in diverse areas related to intelligent systems. The journal also publishes survey articles, short notes, Ph.D. thesis abstracts and project reports.

Editorial Board

The Australian Journal of Intelligent Information Processing Systems is published quarterly. All correspondence including manuscripts and advertising inquiries should be addressed to the Editor in Chief: Professor Tom Gedeon, Faculty of Engineering and Information Technology, The Australian National University, Canberra ACT 0200, Australia.

Subscriptions:
Individual rate: A$80.
Institutional rate: A$220 for paper or online[1] subscription.

Individual copies can be purchased at A$15 per single copy.

Reprints of technical articles are available, quantities no less than 50 can be ordered. Orders should be addressed to: AJIIPS, Faculty of Engineering and Information Technology, The Australian National University, Canberra ACT 0200, Australia. Email address: ajiips@cs.anu.edu.au.

[1] Online subscription gives additional access to back issues without any fee.

Australian Journal of Intelligent Information Processing Systems

ISSN: 1321-2133

Volume 14, No. 1
2014

CONTENTS

This is a refereed journal.
Abstracted by INSPEC.

Publication Date November 2014

Detecting SNP Interactions in Balanced and Imbalanced Datasets using Associative Classification

Suneetha Uppu, Aneesh Krishna and Raj P.Gopalan

Department of Computing, Curtin University, Perth, Australia

Suneetha.uppu@postgrad.curtin.edu.au, A.Krishna@curtin.edu.au,
R.Gopalan@curtin.edu.au

Abstract. The genetic epidemiology behind the complex diseases are characterised by multiple factors acting together or independently. The complex network of these multiple factors induces pathological mechanisms which lead to disease manifestation. Advances in genotyping technology have dramatically increased the understanding of single nucleotide polymorphisms (SNPs) associated with complex diseases. The interactions between SNPs responsible for disease susceptibility are being intensively explored in this era of genome wide association studies (GWAS). Several machine learning and data mining approaches have been proposed to track the inheritance of the disease and its susceptibility towards the environmental factors. However, detecting these interactions continues to be a critical challenge due to bio-molecular complexities and computational limitations. The goal of this research is to study the effectiveness of associative classification for detecting the epistasis in balanced and imbalanced datasets. The proposed approach was evaluated for two locus epistasis interactions using simulated data. The datasets were generated for 5 different penetrance functions by varying heritability, minor allele frequency and sample size. In total, 23,400 datasets were generated and several experiments conducted to identify the disease causal SNP interactions. The accuracy of classification by the proposed approach was compared with the previous approaches. Though the associative classification showed small improvement in accuracy for balanced datasets, it outperformed existing approaches for higher order multi-locus interactions in imbalanced datasets.

Keywords: Epistasis, multi-locus, associative classification, SNP interactions

1 Introduction

The advances in technology helped biologists, geneticists and computer professionals to develop outstanding methods in exploring human genome. A major goal in this new era of genetics is to discover the susceptibility of diseases. A substantial number of GWA studies were carried out to identify the sources of complex diseases. SNPs have become most commonly used biomarkers in case-control based GWA studies. A polymorphism is a single variation in DNA sequence and is present in one percent of the population. Despite the success in identifying thousands of genetic variances associated with complex human diseases, these studies present a number of challenges to the researchers. These include population stratification, missing heritability, low effect size of associations, difficulties in addressing the role of rare variant, and genes involved in the locus and mechanism which lead to the disease[1].

Often in genetic epidemiology, the biological phenomenon is not merely interpreted just as linear models; it may also be due to interference of interaction between genes and gene-environmental factors. Interaction is a nonlinear effect that depends at least on two independent factors that may influence phenotype. Gene-gene interaction or epistasis is of two types, biological and statistical. Biological epistasis is a phenomenon of physical interactions between biomolecules such as DNA, RNA, proteins and enzymes. Statistical epistasis occurs at population level due to inter-individual variation in DNA sequences. It is intuitively difficult to produce biological interpretations from statistical results due to inherent nonlinearity. However, interaction studies are challenging mathematically and computationally. For example, in a study of 300,000 SNPs in GWA, there are 4.5*1010 pair wise two-way interactions. It further grows to 4.5*1015 for three way interactions[2]. Hence analyses of higher order interactions hit the limits of current computational technology. These challenges have been addressed by developing a number of tools and approaches.

Recent research has explored the use of varying and modifying logic regression[3], penalized logistic regression[4], classification and regression tree (CART)[5], multivariate adaptive regression splines (MARS)[6], focused interaction testing framework[7] and automated detection of informative combined effects (DICE)[8].The evolution of huge high dimensional data in genomics has led to the application of data mining and machine learning approaches including data reduction and pattern recognition. These approaches discover interesting interactions by considering all genomic variables in vast search spaces. Data reduction approaches reduce high dimensional data to low dimensional data. They include combinatorial partitioning method (CPM)[9], restricted partition method (RPM)[10], set association[11], and multifactor dimensionality reduction (MDR)[12]. Pattern recognition approaches extract patterns from the data using techniques such as cluster analysis[13], support vector machines (SVM)[14], self-organizing maps (SOM)[15]and neural networks (NNs) [14].

Variable selection, model building and model interpretation are the three primary challenges in these approaches. In addressing these challenges, new strategies and methods are developed to improve genome-wide interaction studies.

Tree based epistasis association mapping (TEAM), Boolean operation based screening and testing (BOOST), and GPU-based BOOST (GBOOST) are some of the exhaustive search approaches. Even though these approaches are feasible, they are computationally intensive and execution time increases exponentially by number of SNPs. In order to overcome these limitations, filtering approaches are used to analyse interesting SNPs. They include Random Forest (RF), epiFOREST, SNPInterForest, Random Jungle (RJ), forest based haplotype approach, bayesian approach, Bayesian epistasis association mapping (BEAM), fast epistatic interactions detection using markov blanket (FEPI-MB), Bayesian network based epistatic association studies (bNEAT), andMega SNPHunter and biological filters [2]. Though several after mentioned approaches are developed, SNPs with weak marginal effect may be filter out that may significantly contribute the disease. The contingency table will have many empty cells which may lead to unstable estimation with large variance.

Although research has progressed with application of new technologies and methods, none of the current approaches reveals the unexplained features of complex diseases due to interactions. Hence, efficient techniques have to be added to address the interaction effects between SNPs. Many researchers have shown that Associative Classification (AC) is more accurate than traditional classifiers[16].The rules generated in AC can be stored and can provide reasoning to the classification.AC is also suitable for both categorical and discrete data. Mapping SNP-SNP interactions to disease can be improved by integrating association rules and classification. In this paper, a new approach based on associative classifier is implemented to identify the interactions more effectively than the existing methods. The proposed approach will classify the subjects by determining the complexity of interactions and their associations with the disease. The goal of this study was to evaluate the proposed approach on the simulated data by varying heritability, minor allele frequencies and case control ratio. The study identified two way SNP-SNP interactions for both balanced and imbalanced datasets. Finally, the approach is validated in terms of accuracy and compared with previous methods under same simulated scenarios.

The associative classification is briefly reviewed and then applied to the present problem in Section 2.1. Data Simulation scenarios and Data Analysis are explained in Section 2.2 and Section 2.3. Results are demonstrated in Section 3. The discussion and conclusion are included in Section 4 and Section 5 respectively.

2 Methods

2.1 Associative classifier

Associative classification (AC) is a promising approach in data mining to build accurate and efficient classifiers for large datasets. The association rule mining is integrated into classification to obtain valuable rules that cannot be generated by other traditional classification approaches. These rules are easily interpretable and provide confidence probability to resolve the uncertainty of the classification problem. The AC algorithm is of three phases. They are: rule generation, building classifier and classification. In rule generation phase, association rule mining is used to generate frequent items from the large dataset. The association rule mining identifies interesting associations and correlations between attributes. In particular, class based association rules (CARs) are generated in this phase. The syntactic constraint of an association rule is that the consequents are restricted to be as class label and all other attributes are antecedents. In building classifier phase, redundant rules are removed and they are ordered to form a classifier. Ordering the rules is performed by criteria such as confidence, support, and rule length. Finally, test data is classified from the ordered rules in a classifier phase. Some of the common ACs are classification based on association (CBA)[17], classification based on multiple association rules (CMAR)[18], classification based on predictive association rules (CPAR)[19], lazy associative classification (LAC) and live and let live (L3)[20].

To formulate SNP interactions as an AC problem, let D be a relation of tuples, whose schema is represented by n distinct attributes SNP1, SNP2,…..,SNPn and a class attribute C. Let C be a finite set of class labels with case c1 and control c2 respectively, where, c ϵ C. The attributes are treated as categorical where the class labels are known in training data instances in D and the class labels are unknown in testing data instances. Each instance tuple in D is represented as ti = (vi1, vi2, …, vin, ci) where vi1 is an item value for SNP1, vi2 for SNP2, etc and ci is a class label.

Association rule R is generated in the form of $X \rightarrow Y$ which matches a tuple t ϵ D when $X \subseteq t$. X is the antecedent which represents interacting SNPs associated with class label and Y the consequent which represents case or control. Support and Confidence are the two parameters used to measure the quality of association rules. Support is the number of tuples in D containing $X \cup Y$ and confidence is the number of tuples matching $X \cup Y$ divided by the number of tuples containing X. The CARs are organised and ranked by computing support and confidence along with the rule cardinality (measure of number of elements of the rule). Redundant and noisy rules are discarded in the rule pruning phase that passes minimum support and confidence thresholds. Several pruning techniques (such as pessimistic error, database coverage, chi-square, redundant rule and lazy pruning) are adopted to reduce the size of AC. Hence, most significant and high quality rules are selected to form a more accurate and efficient classification model. These rules are used to predict the test data. Finally, the accuracy of the dataset is calculated as the ratio of number of objects correctly classified to the total number of objects in the test data.

Model 1, p=0.5, q=0.5

	BB	Bb	bb
AA	0	0.1	0
Aa	0.1	0	0.1
aa	0	0.1	0

Model 2, p=0.5, q=0.5

	BB	Bb	bb
AA	0	0	0.1
Aa	0	0.05	0
aa	0.1	0	0

Model 3, p=0.25, q=0.75

	BB	Bb	bb
AA	0.08	0.07	0.05
Aa	0.1	0	0.1
aa	0.3	0.1	0.04

Model 4, p=0.25, q=0.75

	BB	Bb	bb
AA	0	0.01	0.09
Aa	0.04	0.01	0.08
aa	0.07	0.09	0.03

Model 5, p=0.1, q=0.9

	BB	Bb	bb
AA	0.7	0.05	0.02
Aa	0.05	0.09	0.01
aa	0.02	0.01	0.03

Model 6, p=0.1, q=0.9

	BB	Bb	bb
AA	0.09	0.001	0.02
Aa	0.08	0.07	0.005
aa	0.003	0.007	0.02

Fig. 1. Penetrance functions and minor allele frequencies used to simulate case-control data exhibiting multi-locus SNP-SNP interactions in absence of main effects [21].

2.2 Data Simulation

The goal of this simulated study is to detect interactions between multi-locus SNPs using the AC approach. Two simulated scenarios are considered to evaluate the accuracy AC with the previous approaches in the absence of main effect. Absence of main effect is considered to provide a high degree of complexity while identifying the associations of interacting SNPs related to the disease.

Scenario I.

In first the scenario, six two locus epistasis (gene-gene interactions) models with different penetrance values are simulated for 20 SNPs with two functional SNPs (P1 and P2) and 18 independent non-functional SNPs. Case-control datasets are simulated with 200 cases and 200 controls in accordance to Hardy-Weinberg proportions. Figure 1 represents the overview of model dependent allele frequencies along with their penetrance tables. A simple model of two alleles (p and q) necessarily sums to unity. That is, $p+q = 1$ where p is minor allele frequency and q is the alternative allele frequency. Model 1 is based on nonlinear XOR function described by [21, 22] in which all high risk genotype combinations (AaBB, Aabb, AABb and aaBb) have a penetrance value of 0.1. Model 2 is described by [22, 23] in which high risk genotype combinations (AAbb, AaBb and aaBB) have penetrance values 0.1, 0.05 and 0.1 respectively. Other four models are described by [22] with Minor Allele Frequencies (MAFs) of 0.25, 0.25, 0.1 and 0.1 respectively. Ratios of 1:1, 1:2, 1:4, and 1:6 cases and controls are generated. 100 datasets are simulated for each model in order to evaluate the power of AC by estimating the number of times the approach successfully identified two functional SNPs. In total, 2,400 datasets are generated and analysed in the absence of main effect.

Scenario II.

In the second scenario, datasets are replicated as in the simulated study performed by Velez, D.R., [24] with 20 SNPs. Among these 2 SNPs are functional and 18 SNPs are non-functional. The two locus interaction models are generated from publicly available tool GAMETES [25]. The tool generates randomly pure and strict n-locus disease models with specified heritability, minor allele frequency and population. In this simulated scenario, two locus epistasis models are distributed across seven heritability (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). Five models for each 14 heritability-allele frequency combinations are generated to develop 70 models in accordance to Hardy-Weinberg proportions. The penetrance tables are generated for these 70 models in the absence of main effect. These are available in the appendix as Table 1. One hundred datasets are generated for each model with sample size of 400. The penetrance tables are generated for these 70 models in the absence of main effect. These are available in the appendix as Table 1. One hundred datasets are generated for each model with sample size of 400. The case-control ratios of the samples are 1:1, 1:2, and 1:4. In total, 21,000 datasets are generated and analysed to identify the two way interactions in the absence of main effect.

Table 2: Accuracy of 6 models with 1:1 ratio of cases and controls								
Model	MAF	MDR	RF	AC-CPAR	AC-CBA	SVM	NN	Naïve Bayes
1	0.5	76	57.75	59.5	70	48	59.5	47.75
2	0.5	81.5	57.25	62.5	74	55.5	64.75	57
3	0.25	63	58.25	58.75	60	56	56.5	60.5
4	0.25	74.5	66	69.25	65.75	62	61.5	62
5	0.1	47.75	50.5	53.75	52	49.75	51.25	48.25
6	0.1	57	52	55	54.5	57.75	53.25	56.25

Table 3: Accuracy of 6 models with 1:2 ratio of cases and controls								
Model	MAF	MDR	RF	AC-CPAR	AC-CBA	SVM	NN	Naïve Bayes
1	0.5	76.78	63.33	64.83	66.75	66.67	65.17	62.67
2	0.5	81.46	62.5	68.5	69.75	66.75	65.5	62
3	0.25	53.12	65.25	63	66.75	66.75	56.75	62.5
4	0.25	76.75	70.25	73	65.5	71.25	73.75	72.5
5	0.1	51.59	61.25	61.5	66.75	66.75	57	66
6	0.1	56.88	62.5	62.5	66.75	66.75	62.75	62

Table 4: Accuracy of 6 models with 1:4 ratio of cases and controls								
Model	MAF	MDR	RF	AC-CPAR	AC-CBA	SVM	NN	Naïve Bayes
1	0.5	76.25	78.6	76.6	80	79.25	73.9	79.8
2	0.5	82.66	79.75	79.5	80	79.5	72.75	79.25
3	0.25	53.28	78.25	78.25	69.75	80	66.75	77.75
4	0.25	76.09	79	78.25	80.25	72.25	78	79
5	0.1	57.66	76.5	77.5	80	67.5	68.75	78.25
6	0.1	54.84	75.5	78.5	80	68.5	75.5	79

Table 5: Accuracy of 6 models with 1:6 ratio of cases and controls								
Model	MAF	MDR	RF	AC-CPAR	AC-CBA	SVM	NN	Naïve Bayes
1	0.5	76.16	91.5	91.75	91.25	82.25	86.75	91.25
2	0.5	82.7	91.25	91.75	91.75	85.25	84	90.25
3	0.25	45.94	91.5	91.5	91.75	87	85.75	91.5
4	0.25	71.8	91	91.25	91.25	87	85	90
5	0.1	43.47	90.5	90	93.4	83	85.75	91.5
6	0.1	44.02	91	91.25	91.75	86	87.75	90.25

2.3 Data analysis

The datasets for both scenarios are analysed using the latest MDR software tool available from www.epistasis.org. All the possible two locus interactions between SNPs are exhaustively evaluated on the data using the naïve Bayes classifier. Naive Bayes classifier is assessed using balanced accuracy. Balanced accuracy (arithmetic mean of sensitivity and specificity) is estimated using 10 fold cross validation for both training and testing data. Finally, the best MDR model with maximum testing balanced accuracy and highest cross validation consistency is selected. The power of MDR to detect SNP interactions has been estimated by the number of times the functional SNPs are identified in 100 datasets of each model. The final results are statistically evaluated with a 1000 fold permutation test and whose p-values are compared with 0.05 in determining the significance of the findings. The datasets for both scenarios are analysed using associative classifier algorithms. The accuracy of AC algorithms is analysed using 10 fold cross validation and the disease causal interacting SNPs are investigated. There are numerous methods and software implementations that have been used to investigate the interactions between SNPs. The most prominent approaches for identifying genetic effects in the presence of interactions are MDR, RF, SVM and NN [14]. Further, Naïve Bayes algorithm is also considered in this paper as it is a well-established machine learning method and has been successfully applied in analysing GWAS data [26, 27]. Both scenarios are analysed using RF, SVM, NN and Naïve Bayes algorithms. Ten Fold cross validation is performed to reduce the possibility of biased estimation due to the division of data.

3 Results

Several experiments are performed over 23,400 datasets to evaluate the accuracy of AC over other approaches. The accuracy of AC for all six models in Scenario-I is presented in Table 2 to Table 4. Table 2 represents the accuracy of AC along with other previous approaches with 400 samples of 1:1 ratio of cases and controls. The accuracy of AC is higher for model 5 when compared to other approaches. Table 3 represents the results of AC and other current approaches for 1:2 ratios of cases and controls. The accuracy of AC is high for model 3, models 5 and model 6. Table 4 and Table 5 represent accuracy of AC for 1:4 and 1:6 ratios of cases and controls respectively. The results show that AC

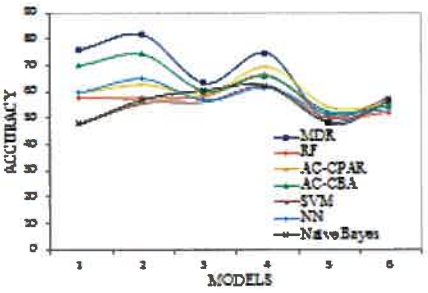

Fig. 2a. Accuracy of 6 models with 1:1 ratio

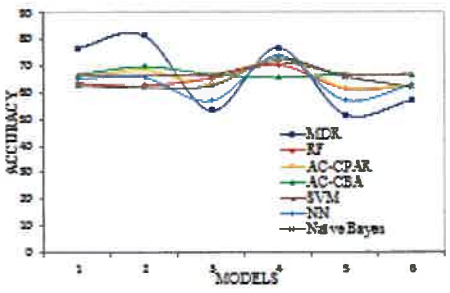

Fig. 2b. Accuracy of 6 models with 1:2 ratio

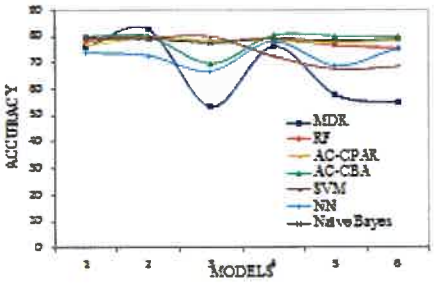

Fig. 2c. Accuracy of 6 models with 1:4 ratio

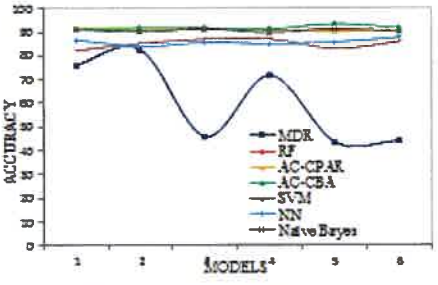

Fig. 2d. Accuracy of 6 models with 1:6 ratio

outperforms in all 6 models compared to the current approaches. The performance of AC is seen to be better with imbalanced data than with balanced data.

In scenario II, the accuracy of AC for 70 models is presented in Appendix from Table 6 to Table 8 and compared with previous methods. Table 6 represents the results of AC and other current approaches for 1:1 ratio of 400 samples. The accuracy of AC is higher only in two cases (MAF=0.2, H=0.01 and MAF=0.4, H=0.025) when compared with existing approaches. Table 7 and Table 8 shows the results of 1:2 and 1:4 ratios of cases and controls respectively. The accuracy of AC is higher in majority of the models in 1:2 ratios. AC performed well when the heritability is ≤ 0.1 for both cases of MAF = 0.2 and 0.4. However, as shown in Table 8, AC outperformed other approaches in all 70 models for the ratio of 1:4. The results of scenario II confirm that the accuracy of AC is higher over other current approaches in imbalanced datasets.

4 Discussion

The goal of this study is to determine whether AC is a better approach for identifying the higher order SNP interactions in absence of main effect. The approach considers the ratio of cases and controls for each SNP combination at different locus. It generates statistically significant genotype combinatorial associations in terms of rules based on cases and controls. Predicting class labels of test objects from these rules retains higher accuracy in genetic combinations that contribute to a disease. Despite the increase in accuracy, the approach will still reduce the false positive error by permutation testing under the null hypothesis. The results have been obtained on two simulated scenarios to identify complex associations between genotype and phenotype.

In the first scenario, as stated in the results, the approach is validated for both balanced and imbalanced datasets. Figure 2a shows the accuracy of AC over MDR, RF, SVM, NN and Naïve Bayes classifiers in 1:1 ratios of cases and controls for 400 samples. On an average of 100 datasets for each model, MDR significantly performed well for 1 to 4 models. However, AC performed better than other algorithms when allele frequencies are 0.1 and 0.9. SVM performed equally as MDR with difference of less than 1% in accuracy for model 6. Figure 2b shows the accuracy of AC over other algorithms in 1:2 ratios of cases and controls in a sample size of 400. On average, AC achieves an improvement in accuracy of 13% compared to MDR for model 3, 5 0.1. It is observed that both for balanced and imbalanced data, AC is more accurate when the allele frequencies are 0.1 and 0.9. Figure 2c exhibits accuracy of samples with 1:4 ratios. On average, the accuracy of AC is about 12% higher than MDR. However, it is observed that the accuracy of AC is slightly reduced by about 2% in model 2 where allele frequencies are equal. Figure 2d shows that the accuracy of AC is much higher than MDR in 1:6 ratios of all 6 models. Accuracy of AC is about 50% higher than MDR when MAF values are 0.1 and 0.25.

The results of second scenario of simulations, demonstrated that the AC performed well across a wide range of SNP-SNP interaction models. Figure 3a illustrates accuracy of AC over other approaches in balanced data of 400 samples. MDR predominantly outperformed AC and other approaches. However, AC is more accurate than other approaches up

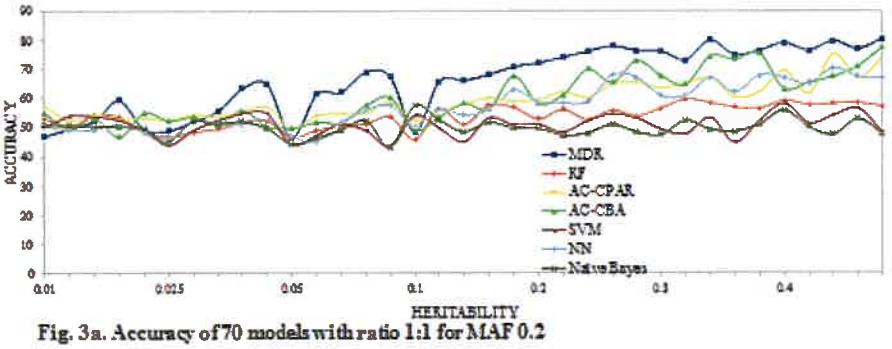

Fig. 3a. Accuracy of 70 models with ratio 1:1 for MAF 0.2

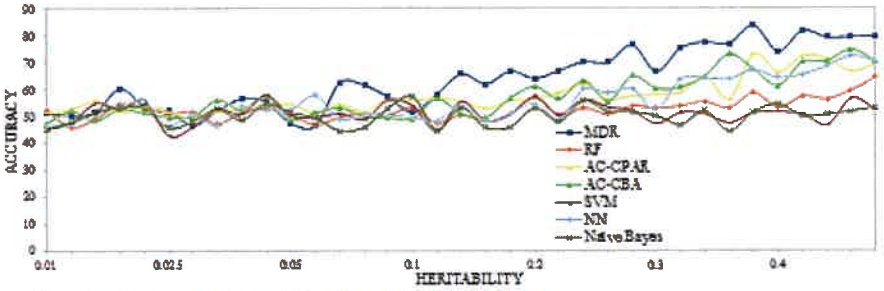

Fig. 3b. Accuracy of 70 models with ratio 1:1 for MAF 0.4

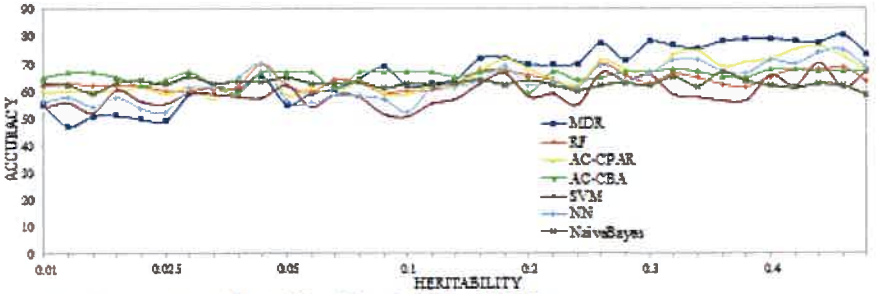

Fig. 4a. Accuracy of 70 models with ratio 1:2 for MAF 0.2

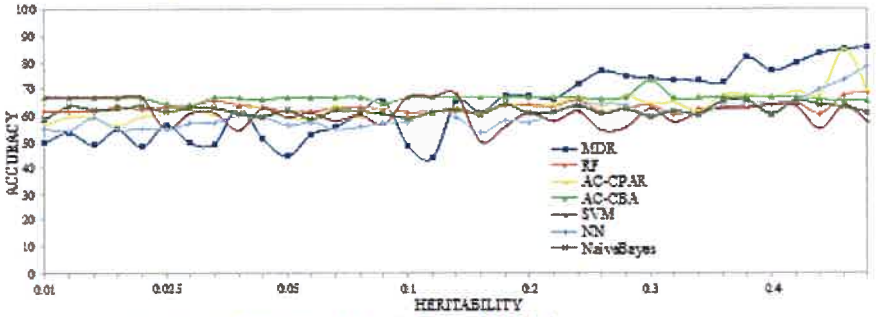

Fig. 4b. Accuracy of 70 models with ratio 1:2 for MAF 0.4

to 10% for allele frequencies 0.2 and 0.8 with heritability of 0.01. Further experiments were performed to observe the performance of AC when there is no genetic influence over the phenotype. It performed significantly better than all other approaches including MDR. Figure 3b illustrates the accuracy along y-axis and heritability along x-axis for 1:1 ratio with MAF equal to 0.4. MDR performed significantly better in balanced data compared to other methods. However, accuracy of AC improved up to 4% when heritability is 0.025. It also significantly performed better than other approaches when there is no genetic influence over the phenotype. Figure 4a and Figure 4b graphically represents accuracy of AC for 1:2 ratio of sample size 400 with MAF 0.2 and 0.4 respectively. For average of MAF 0.2 and 0.4, the accuracy of AC is higher by up to 14% and 16% respectively compared to MDR for heritability values of 0.01, 0.025, 0.05 and 0.1. It is also been observed that, Naive Bayes' algorithm significantly performed better than MDR. However, on average AC was more accurate than Naive Bayes' algorithm for MAF 0.2 and 0.4 by upto 3% and 5% respectively. AC had the same accuracy as MDR for heritability 0.2, 0.3 and 0.4 for both MAF values (0.2 and 0.4). Figure 5a and 5b illustrates accuracy of AC for 1:4 ratios with MAF 0.2 and 0.4 respectively. AC predominantly outperformed in all 70 models compared with other existing approaches. These results demonstrate that the power of AC increases in imbalanced data with higher proportions of controls than cases.

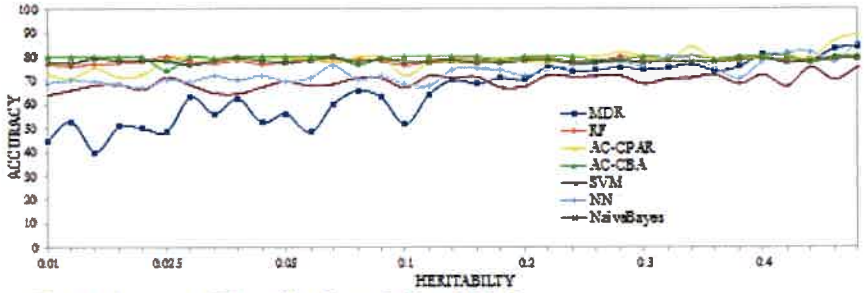

Fig. 5a. Accuracy of 70 models of ratio 1:4 for MAF 0.2

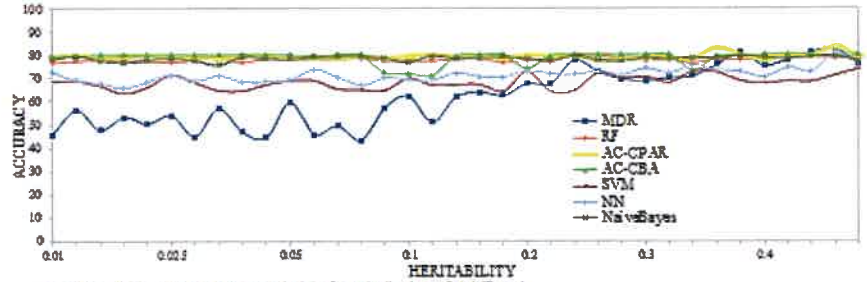

Fig. 5b. Accuracy of 70 models of ratio 1:4 for MAF 0.4

5 Conclusion and Future works

In this paper, association based classification approach was implemented for detecting interactions in balanced and unbalanced data. The approach was evaluated for two locus interactions using simulated data. The approach performed significantly better than the existing approaches in imbalanced data. However, the experimental results showed only small improvement in accuracy for balanced data. Further studies will investigate the performance of AC over three-way to ten-way genotype interactions and how these contribute to associated phenotype. The approach will be further applied to real data to confirm the success rate of identifying the interactions between SNPs in high dimensional genome. Further, the empirical power of the approach will be determined in the presence of genotyping error, missing data, phenocopy and genetic heterogeneity.

References

1. Maher, B., *The case of the missing heritability.* Nature, 2008. **456**(7218): p. 18-21.
2. Padyukov, L., *Between the Lines of Genetic Code: Genetic Interactions in Understanding Disease and Complex Phenotypes.* 2013: Academic Press.
3. Ruczinski, C.K.I., M.L. LeBlanc, and L. Hsu, *Sequence analysis using logic regression.* Genetic epidemiology, 2001. **21**(1): p. S626-S631.
4. Park, M.Y. and T. Hastie, *Penalized logistic regression for detecting gene interactions.* Biostatistics, 2008. **9**(1): p. 30-50.
5. Loh, W.Y., *Classification and regression trees.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011. **1**(1): p. 14-23.
6. Cook, N.R., R.Y. Zee, and P.M. Ridker, *Tree and spline based association analysis of gene–gene interaction models for ischemic stroke.* Statistics in medicine, 2004. **23**(9): p. 1439-1453.
7. Millstein, J., et al., *A testing framework for identifying susceptibility genes in the presence of epistasis.* The American Journal of Human Genetics, 2006. **78**(1): p. 15-27.
8. Tahri-Daizadeh, N., et al., *Automated detection of informative combined effects in genetic association studies of complex traits.* Genome Research, 2003. **13**(8): p. 1952-1960.
9. Nelson, M., et al., *A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation.* Genome Research, 2001. **11**(3): p. 458-470.
10. Culverhouse, R., T. Klein, and W. Shannon, *Detecting epistatic interactions contributing to quantitative traits.* Genetic epidemiology, 2004. **27**(2): p. 141-152.
11. Wille, A., J. Hoh, and J. Ott, *Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers.* Genetic epidemiology, 2003. **25**(4): p. 350-359.
12. Ritchie, M.D., et al., *Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.* The American Journal of Human Genetics, 2001. **69**(1): p. 138-147.

13. Kaufman, L. and P.J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. Vol. 344. 2009: John Wiley & Sons.
14. Upstill-Goddard, R., et al., *Machine learning approaches for the discovery of gene–gene interactions in disease data.* Briefings in bioinformatics, 2013. **14**(2): p. 251-260.
15. Kohonen, T., *Self-organizing maps*. Vol. 30. 2001: Springer.
16. Thabtah, F., *A review of associative classification mining.* The Knowledge Engineering Review, 2007. **22**(01): p. 37-65.
17. Ma, B.L.W.H.Y. *Integrating classification and association rule mining.* in *Proceedings of the 4th.* 1998.
18. Li, W., J. Han, and J. Pei. *CMAR: Accurate and efficient classification based on multiple class-association rules.* in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. 2001. IEEE.
19. Han, J. *CPAR: Classification based on predictive association rules.* in *Proceedings of the third SIAM international conference on data mining*. 2003.
20. Baralis, E., S. Chiusano, and P. Garza, *A lazy approach to associative classification.* Knowledge and Data Engineering, IEEE Transactions on, 2008. **20**(2): p. 156-171.
21. Li, W. and J. Reich, *A complete enumeration and classification of two-locus disease models.* Human heredity, 2000. **50**(6): p. 334-349.
22. Moore, J.H., et al. *Application of genetic algorithms to the discovery of complex models for simulation studies in human genetics.* in *Proceedings of the Genetic and Evolutionary Computation Conference/GECCO. Genetic and Evolutionary Computation Conference*. 2002. NIH Public Access.
23. Frankel, W.N. and N.J. Schork, *Who's afraid of epistasis?* Nature genetics, 1996. **14**(4): p. 371-373.
24. Velez, D.R., et al., *A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction.* Genetic epidemiology, 2007. **31**(4): p. 306-315.
25. Urbanowicz, R.J., et al., *GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures.* BioData mining, 2012. **5**(1): p. 1-14.
26. Moore, J.H., F.W. Asselbergs, and S.M. Williams, *Bioinformatics challenges for genome-wide association studies.* Bioinformatics, 2010. **26**(4): p. 445-455.
27. Bellazzi, R. and B. Zupan, *Predictive data mining in clinical medicine: current issues and guidelines.* International journal of medical informatics, 2008. **77**(2): p. 81-97.

Appendix

Table 1: Penetrance tables for 70 models in Scenario II

Model 1 — MAF = 0.2, H = 0.01, K=0.117

	AA	Aa	aa
BB	0.117	0.12	0.0954
Bb	0.117	0.123	0.0677
bb	0.119	0.02	0.8509

Model 2 — MAF = 0.2, H = 0.01, K=0.623

	AA	Aa	aa
BB	0.598	0.6701	0.661
Bb	0.664	0.5533	0.535
bb	0.71	0.4369	0.736

Model 3 — MAF = 0.2, H = 0.01, K=0.096

	AA	Aa	aa
BB	0.086	0.12	0.062
Bb	0.119	0.0485	0.107
bb	0.069	0.0936	0.544

Model 4 — MAF = 0.2, H = 0.01, K=0.079

	AA	Aa	aa
BB	0.092	0.056	0.062
Bb	0.058	0.124	0.0682
bb	0.048	0.097	0.4439

Model 5 — MAF = 0.2, H = 0.01, K=0.942

	AA	Aa	aa
BB	0.931	0.961	0.9613
Bb	0.961	0.902	0.9423
bb	0.959	0.947	0.6225

Model 6 — MAF = 0.2, H = 0.025, K=0.065

	AA	Aa	aa
BB	0.066	0.0649	0.039
Bb	0.067	0.0685	3.706
bb	0.022	0.034	0.995

Model 7 — MAF = 0.2, H = 0.025, K=0.611

	AA	Aa	aa
BB	0.644	0.5518	0.57
Bb	0.531	0.7593	0.705
bb	0.733	0.3785	0.529

Model 8 — MAF = 0.2, H = 0.025, K=0.046

	AA	Aa	aa
BB	0.056	0.03	0.0166
Bb	0.03	0.081	0.0186
bb	0.013	0.026	0.7333

Model 9 — MAF = 0.2, H = 0.025, K=0.089

	AA	Aa	aa
BB	0.11	0.048	0.0812
Bb	0.051	0.171	0.0363
bb	0.057	0.085	0.6268

Model 10 — MAF = 0.2, H = 0.025, K=0.936

	AA	Aa	aa
BB	0.919	0.9695	0.952
Bb	0.97	0.8668	0.967
bb	0.951	0.9694	0.449

Model 11 — MAF = 0.2, H = 0.05, K=0.677

	AA	Aa	aa
BB	0.663	0.6758	0.927
Bb	0.676	0.7472	0.153
bb	0.928	0.1517	0.883

Model 12 — MAF = 0.2, H = 0.05, K=0.593

	AA	Aa	aa
BB	0.544	0.703	0.5154
Bb	0.7	0.361	0.753
bb	0.538	0.708	0.5728

Model 13 — MAF = 0.2, H = 0.05, K=0.168

	AA	Aa	aa
BB	0.211	0.084	0.1493
Bb	0.083	0.34	0.153
bb	0.16	0.132	0.5883

Model 14 — MAF = 0.2, H = 0.5, K=0.098

	AA	Aa	aa
BB	0.132	0.0362	0.048
Bb	0.033	0.2245	0.125
bb	0.071	0.0779	0.694

Model 15 — MAF = 0.2, H = 0.5, K=0.052

	AA	Aa	aa
BB	0.075	0.0124	0.015
Bb	0.012	0.1355	0.038
bb	0.015	0.0388	0.778

Model 16 — MAF = 0.2, H = 0.5, K=0.052

	AA	Aa	aa
BB	0.075	0.012	0.0147
Bb	0.012	0.136	0.0385
bb	0.015	0.039	0.7783

Model 17 — MAF = 0.2, H = 0.1, K=0.364

	AA	Aa	aa
BB	0.44	0.204	0.4594
Bb	0.217	0.684	0.1818
bb	0.352	0.398	0.3171

Model 18 — MAF = 0.2, H = 0.1, K=0.178

	AA	Aa	aa
BB	0.242	0.0612	0.104
Bb	0.052	0.4223	0.255
bb	0.179	0.1039	0.762

Model 19 — MAF = 0.2, H = 0.1, K=0.118

	AA	Aa	aa
BB	0.172	0.016	0.091
Bb	0.021	0.3184	0.079
bb	0.047	0.1663	0.881

Model 20 — MAF = 0.2, H = 0.1, K=0.921

	AA	Aa	aa
BB	0.878	1	0.9841
Bb	0.999	0.768	0.9111
bb	0.992	0.896	0.0026

Model 21 — MAF = 0.2, H = 0.2, K=0.470

	AA	Aa	aa
BB	0.389	0.667	0.1984
Bb	0.666	0.016	0.9821
bb	0.209	0.961	0.7352

Model 22 — MAF = 0.2, H = 0.2, K=0.070

	AA	Aa	aa
BB	0.609	0.2485	0.424
Bb	0.265	0.9067	0.661
bb	0.295	0.9197	0.076

Model 23 — MAF = 0.2, H = 0.2, K=0.071

	AA	Aa	aa
BB	0.602	0.9223	0.743
Bb	0.918	0.2946	0.71
bb	0.775	0.6457	0.191

Model 24 — MAF = 0.2, H = 0.2, K=0.209

	AA	Aa	aa
BB	0.311	0.028	0.0328
Bb	0.025	0.545	0.4747
bb	0.059	0.422	0.9064

Model 25 — MAF = 0.2, H = 0.2, K=0.174

	AA	Aa	aa
BB	0.264	0.003	0.1174
Bb	0.001	0.519	0.1857
bb	0.13	0.16	0.9939

Model 26 — MAF = 0.2, H = 0.3, K=0.421

	AA	Aa	aa
BB	0.541	0.161	0.587
Bb	0.156	0.9935	0.077
bb	0.623	0.0047	0.516

Model 27 — MAF = 0.2, H = 0.3, K=0.658

	AA	Aa	aa
BB	0.513	0.9159	0.926
Bb	0.908	0.2235	0.147
bb	0.986	0.0269	0.477

Model 28 — MAF = 0.2, H = 0.3, K=0.662

	AA	Aa	aa
BB	0.528	0.928	0.6957
Bb	0.937	0.125	0.5846
bb	0.628	0.721	0.7591

Model 29
MAF = 0.2, H = 0.3, K=0.266

	AA	Aa	aa
BB	0.396	0.006	0.2733
Bb	0.04	0.734	0.1641
bb	0.007	0.696	0.9823

Model 30
MAF = 0.2, H = 0.3, K=0.752

	AA	Aa	aa
BB	0.623	0.9992	0.844
Bb	0.992	0.2837	0.662
bb	0.903	0.5439	0.003

Model 31
MAF = 0.2, H = 0.4, K=0.604

	AA	Aa	aa
BB	0.448	0.8803	0.902
Bb	0.942	0.0039	0.01
bb	0.409	0.9975	0.596

Model 32
MAF = 0.2, H = 0.4, K=0.627

	AA	Aa	aa
BB	0.473	0.96	0.447
Bb	0.913	0.012	0.9929
bb	0.825	0.236	0.6026

Model 33
MAF = 0.2, H = 0.4, K=0.656

	AA	Aa	aa
BB	0.505	0.946	0.7615
Bb	0.98	0.038	0.4295
bb	0.488	0.977	0.7962

Model 34
MAF = 0.2, H = 0.4, K=0.664

	AA	Aa	aa
BB	0.509	0.9718	0.696
Bb	0.978	0.0427	0.621
bb	0.643	0.7272	0.519

Model 35
MAF = 0.2, H = 0.4, K=0.299

	AA	Aa	aa
BB	0.449	5.4303	0.301
Bb	0.003	0.9049	0.209
bb	0.282	0.2477	0.999

Model 36
MAF = 0.4, H = 0.009, K=0.670

	AA	Aa	aa
BB	0.648	0.653	0.7766
Bb	0.67	0.673	0.6631
bb	0.723	0.703	0.4549

Model 37
MAF = 0.4, H = 0.009, K=0.429

	AA	Aa	aa
BB	0.4	0.456	0.4162
Bb	0.404	0.445	0.4388
bb	0.571	0.323	0.4302

Model 38
MAF = 0.4, H = 0.009, K=0.347

	AA	Aa	aa
BB	0.423	0.3147	0.276
Bb	0.312	0.3714	0.356
bb	0.285	0.3493	0.482

Model 39
MAF = 0.4, H = 0.009, K=0.398

	AA	Aa	aa
BB	0.352	0.4489	0.351
Bb	0.456	0.3552	0.4
bb	0.331	0.4153	0.501

Model 40
MAF = 0.4, H = 0.009, K=0.227

	AA	Aa	aa
BB	0.283	0.194	0.1999
Bb	0.189	0.263	0.2046
bb	0.215	0.193	0.3566

Model 41
MAF = 0.4, H = 0.025, K=0.343

	AA	Aa	aa
BB	0.387	0.369	0.1676
Bb	0.335	0.343	0.3629
bb	0.269	0.287	0.6793

Model 42
MAF = 0.4, H = 0.025, K=0.370

	AA	Aa	aa
BB	0.307	0.4246	0.351
Bb	0.344	0.3739	0.421
bb	0.593	0.2396	0.263

Model 43
MAF = 0.4, H = 0.025, K=0.514

	AA	Aa	aa
BB	0.633	0.4716	0.378
Bb	0.424	0.54	0.643
bb	0.52	0.5372	0.436

Model 44
MAF = 0.4, H = 0.025, K=0.471

	AA	Aa	aa
BB	0.356	0.574	0.4235
Bb	0.55	0.415	0.4643
bb	0.496	0.41	0.6008

Model 45
MAF = 0.4, H = 0.025, K=0.780

	AA	Aa	aa
BB	0.698	0.832	0.8079
Bb	0.837	0.721	0.8283
bb	0.794	0.839	0.5726

Model 46
MAF = 0.4, H = 0.05, K=0.307

	AA	Aa	aa
BB	0.358	0.3404	0.092
Bb	0.315	0.3013	0.307
bb	0.168	0.2498	0.793

Model 47
MAF = 0.4, H = 0.05, K=0.634

	AA	Aa	aa
BB	0.726	0.5405	0.711
Bb	0.607	0.6329	0.703
bb	0.513	0.8521	0.257

Model 48
MAF = 0.4, H = 0.05, K=0.395

	AA	Aa	aa
BB	0.41	0.29	0.678
Bb	0.357	0.486	0.2105
bb	0.478	0.361	0.3137

Model 49
MAF = 0.4, H = 0.05, K=0.746

	AA	Aa	aa
BB	0.616	0.831	0.7832
Bb	0.806	0.674	0.8288
bb	0.859	0.772	0.4142

Model 50
MAF = 0.4, H = 0.05, K=0.755

	AA	Aa	aa
BB	0.615	0.8437	0.807
Bb	0.839	0.6753	0.808
bb	0.821	0.7968	0.483

Model 51
MAF = 0.4, H = 0.1, K=0.675

	AA	Aa	aa
BB	0.583	0.6369	0.999
Bb	0.676	0.6812	0.657
bb	0.882	0.7452	0.001

Model 52
MAF = 0.4, H = 0.1, K=0.566

	AA	Aa	aa
BB	0.718	0.597	0.1351
Bb	0.476	0.525	0.8966
bb	0.499	0.624	0.5463

Model 53
MAF = 0.4, H = 0.1, K=0.558

	AA	Aa	aa
BB	0.527	0.704	0.1905
Bb	0.617	0.417	0.8491
bb	0.452	0.653	0.5118

Model 54
MAF = 0.4, H = 0.1, K=0.585

	AA	Aa	aa
BB	0.319	0.7223	0.774
Bb	0.774	0.4913	0.445
bb	0.62	0.5603	0.584

Model 55
MAF = 0.4, H = 0.1, K=0.757

	AA	Aa	aa
BB	0.573	0.8764	0.817
Bb	0.882	0.6363	0.842
bb	0.801	0.8547	0.369

Model 56
MAF = 0.4, H = 0.2, K=0.432

	AA	Aa	aa
BB	0.192	0.438	0.9564
Bb	0.46	0.498	0.1719
bb	0.888	0.223	0.0332

Model 57
MAF = 0.4, H = 0.2, K=0.403

	AA	Aa	aa
BB	0.108	0.543	0.6526
Bb	0.589	0.388	0.037
bb	0.515	0.14	0.9454

Model 58
MAF = 0.4, H = 0.2, K=0.484

	AA	Aa	aa
BB	0.645	0.3206	0.616
Bb	0.212	0.6739	0.529
bb	0.941	0.2852	0.055

Model 59
MAF = 0.4, H = 0.2, K=0.611

	AA	Aa	aa
BB	0.88	0.3895	0.674
Bb	0.332	0.7902	0.706
bb	0.847	0.5758	0.189

Model 60
MAF = 0.4, H = 0.2, K=0.828

	AA	Aa	aa
BB	0.622	0.944	0.9452
Bb	0.948	0.698	0.9522
bb	0.935	0.96	0.1951

Model 61
MAF = 0.4, H = 0.3, K=0.434

	AA	Aa	aa
BB	0.173	0.444	0.998
Bb	0.457	0.563	0.0016
bb	0.958	0.032	0.4684

Model 62
MAF = 0.4, H = 0.3, K=0.496

	AA	Aa	aa
BB	0.953	0.3002	0.059
Bb	0.304	0.5577	0.746
bb	0.046	0.7556	0.733

Model 63
MAF = 0.4, H = 0.3, K=0.460

	AA	Aa	aa
BB	0.052	0.8053	0.349
Bb	0.608	0.3085	0.587
bb	0.939	0.1436	0.337

Model 64
MAF = 0.4, H = 0.3, K=0.482

	AA	Aa	aa
BB	0.144	0.693	0.6135
Bb	0.846	0.227	0.4337
bb	0.156	0.777	0.336

Model 65
MAF = 0.4, H = 0.3, K=0.743

	AA	Aa	aa
BB	0.439	0.948	0.818
Bb	0.971	0.518	0.9084
bb	0.748	0.961	0.0829

Model 66
MAF = 0.4, H = 0.4, K=0.493

	AA	Aa	aa
BB	0.026	0.72	0.865
Bb	0.709	0.478	0.054
bb	0.899	0.0289	0.974

Model 67
MAF = 0.4, H = 0.4, K=0.423

	AA	Aa	aa
BB	0.077	0.8168	0.022
Bb	0.5	0.2441	0.787
bb	0.971	0.0754	0.235

Model 68
MAF = 0.4, H = 0.4, K=0.438

	AA	Aa	aa
BB	0.984	0.17	0.0122
Bb	0.092	0.625	0.6573
bb	0.247	0.481	0.7387

Model 69
MAF = 0.4, H = 0.4, K=0.575

	AA	Aa	aa
BB	0.047	0.887	0.8291
Bb	0.965	0.347	0.3838
bb	0.595	0.56	0.5796

Model 70
MAF = 0.4, H = 0.4, K=0.729

	AA	Aa	aa
BB	0.361	0.9957	0.758
Bb	0.992	0.459	0.947
bb	0.767	0.9397	0.012

Table 6: Accuracy of 70 models with 400 samples of 1:1 ratio of cases and controls

Model	MAF	Heritability	MDR	RF	AC-CPAR	AC-CBA	SVM	NN	Naïve Bayes
1	0.2	0.01	47	53	57	54.75	50.25	50.5	50.75
2	0.2	0.01	49.5	50.25	52.5	49.5	54	49	50.75
3	0.2	0.01	52.25	54.25	54.5	53	53.75	49.25	50.5
4	0.2	0.01	59.25	53.5	52.75	46.75	52.5	51.25	50.25
5	0.2	0.01	49.25	48.25	53.25	55	49	48.5	49.5
6	0.2	0.025	48.75	46.75	52.5	52.25	44	46.5	45.25
7	0.2	0.025	52	48.5	54.25	53.5	49	51.75	52.25
8	0.2	0.025	55.5	49.5	53.75	50.75	52.5	51.75	51.25
9	0.2	0.025	63.25	51.75	55.25	55.75	54.75	51.25	52
10	0.2	0.025	64.75	52.25	57	50.25	54.75	53.25	49.75
11	0.2	0.05	46.5	46.5	49.25	49.5	44.25	47.25	44.5
12	0.2	0.05	61.25	49	54	51.75	47.25	45.25	46.25
13	0.2	0.05	62	50.25	54.75	51.5	51.25	51.75	49.25
14	0.2	0.05	69	51.25	55	57.5	48.75	56	52.25
15	0.2	0.05	67.25	53.5	58.5	60.25	43.5	57.5	43.25
16	0.2	0.1	48.25	45.75	52	49	54	50.5	57.5
17	0.2	0.1	65.75	56.25	55.75	53.5	49.75	56.25	52.5
18	0.2	0.1	66	51	57.75	58	45	54	48.5
19	0.2	0.1	68	57.5	60	56.25	53.25	56	51.75
20	0.2	0.1	70.75	56.75	58.75	67.5	50.75	62.75	49.75
21	0.2	0.2	72	52.75	59.25	58	51	58.5	49.75
22	0.2	0.2	74	56	62	61	48.25	58.25	47.25
23	0.2	0.2	76	52.5	60.25	69.75	52.25	58.75	48
24	0.2	0.2	77.75	55.75	64.75	65.25	54.5	68	51
25	0.2	0.2	76.25	53.75	65.25	72.75	53	67	48.25
26	0.2	0.3	76	56.5	63.5	67.5	49.25	60.75	47.5
27	0.2	0.3	73	59.5	65	64.5	47.75	60.75	52.25
28	0.2	0.3	79.75	58.25	66.75	74.25	53.25	67	49.25
29	0.2	0.3	74.75	56.75	60.5	73.25	44.75	62	48.5
30	0.2	0.3	76.25	56.25	62	75.25	52.25	67.75	51
31	0.2	0.4	78.75	59	69.25	63	58	67	55.75
32	0.2	0.4	76.25	57.5	61.5	65.25	51.25	64.75	50
33	0.2	0.4	79.5	58	74.75	67.25	54	70.25	47.5
34	0.2	0.4	76.75	58.25	66.75	70.75	56.25	67.25	52.75
35	0.2	0.4	80.25	57	73.5	77.25	48.25	67	47.75
36	0.4	0.01	51.25	52.25	51.5	47.25	45.25	46	45
37	0.4	0.01	50.25	45.75	52.25	51.75	47.5	47.25	47.75
38	0.4	0.01	51.25	49	55.5	48.25	54.75	51.25	51.5
39	0.4	0.01	60.25	54.5	51.5	52.5	52.5	53.75	53.25
40	0.4	0.01	53.5	53	53.5	51	55.75	55.5	53
41	0.4	0.025	51.75	50.75	51.5	49.5	42.5	46.5	45.5
42	0.4	0.025	46.5	51.5	49	49.5	46	50.5	48
43	0.4	0.025	51.75	47	51.5	56	52.5	46	52.5
44	0.4	0.025	56.5	51	52.5	52.25	50.5	53.5	48.5
45	0.4	0.025	55.25	52.75	54.5	57.25	57.75	52	54.5
46	0.4	0.05	47.25	50.5	54.25	48.75	50	52.25	51.25
47	0.4	0.05	46.25	46.5	51	51	49.5	57.5	49.25
48	0.4	0.05	62.25	49.5	54	53	50.75	48.75	44.25
49	0.4	0.05	61.25	50.75	51.5	49.5	48.75	50.5	45.75
50	0.4	0.05	57	50.25	56	49.25	55.75	50.25	52.75
51	0.4	0.1	51.25	53	55.5	48.75	53.75	49.25	57.25
52	0.4	0.1	58	47.25	56.5	56.25	44.25	48	44.5
53	0.4	0.1	65.5	50.25	55.25	50.5	55.25	53.75	53
54	0.4	0.1	61.75	47.75	52.5	49.25	48	48	45.5
55	0.4	0.1	66.5	50.5	55.5	56.5	50.5	50.75	45.75
56	0.4	0.2	63.75	56.75	56.25	60.75	56.75	54.5	52.75
57	0.4	0.2	66.75	49.75	58.25	56.25	50.25	48.5	47.75
58	0.4	0.2	70	52.75	61.5	63	56	59.75	55.5
59	0.4	0.2	69.75	50.5	56	55.25	53	58.25	51.5
60	0.4	0.2	76.25	53.5	57.25	64.75	51.75	60	51.25
61	0.4	0.3	66.25	52.75	57.75	60	47	52	50
62	0.4	0.3	75	53.5	58.5	60.5	51	63.5	46.5
63	0.4	0.3	77.25	55	64.5	64.5	50.5	64	51.75
64	0.4	0.3	76.75	52.75	56	72.75	47.25	63.5	44.25
65	0.4	0.3	83.5	58.25	72.75	67.25	51	67	51.25
66	0.4	0.4	73.75	52.75	66	61	51.25	64	54.25
67	0.4	0.4	81.5	57	72	69.75	50.5	65.25	49.75
68	0.4	0.4	79	56	71	70.25	46.5	68.25	50.5
69	0.4	0.4	79.25	59.25	66.25	74.5	56.25	72.25	51.5
70	0.4	0.4	79.5	64	69	70	52	70	53

Table 7: Accuracy of 70models with 400 samples of 1:2 ratio of cases and controls

Model	MAF	Heritability	MDR	RF	AC-CPAR	AC-CBA	SVM	NN	NaiveBayes
1	0.2	0.01	54.63	61.75	59.5	65	54	56.25	62.75
2	0.2	0.01	47.09	62.75	59.75	66.75	55.75	57.75	62.25
3	0.2	0.01	50.84	61.75	59.75	66.75	52	54.25	59.25
4	0.2	0.01	51.05	62.25	59.75	65	60.25	57.75	63
5	0.2	0.01	49.71	61.75	61.25	62	56	53.75	63.75
6	0.2	0.025	49.14	60	58.75	64	55.25	52.25	62.5
7	0.2	0.025	58.9	60.25	61.5	66.75	59.25	61	65.25
8	0.2	0.025	61.37	60.25	57	62	58.75	61.5	62.75
9	0.2	0.025	60.62	61.5	65	59.25	58	65	63.55
10	0.2	0.025	65.45	70	69.5	66.75	57.5	70	63.25
11	0.2	0.05	54.98	62	58.5	66.75	62	56.5	65
12	0.2	0.05	59.1	59.5	61	66.75	54.25	55.25	62.75
13	0.2	0.05	60.63	64	58.5	61	58.75	58.75	62.75
14	0.2	0.05	64.33	63	63.75	66.75	57.75	58.25	63.25
15	0.2	0.05	68.7	59.25	58.75	66.75	51.5	57	61
16	0.2	0.1	61.74	60	58.75	66.75	50.5	52.25	62.5
17	0.2	0.1	62.31	60.25	61.5	66.75	55.25	61	62.25
18	0.2	0.1	64.33	62.5	61.25	65	57	61.5	62.75
19	0.2	0.1	71.89	67.25	67.25	66.75	62.5	63.75	64
20	0.2	0.1	72	67.25	71.5	66.75	66.25	68.75	62.25
21	0.2	0.2	69.59	65.5	67.75	59.25	57.75	62	63.75
22	0.2	0.2	69.43	63.75	64	66.75	58.75	63	62
23	0.2	0.2	69.94	61.25	62.25	63.75	54.75	61	60
24	0.2	0.2	77.46	69.75	71	64.75	66.5	69.25	62
25	0.2	0.2	71.25	65.5	67.75	66.75	63.5	63.5	62.75
26	0.2	0.3	77.88	62.75	65.75	66.75	65.75	66.25	62
27	0.2	0.3	76.53	66	73	67.75	58.5	71	65
28	0.2	0.3	75.44	64.5	74.5	66.75	57.5	71.25	61.5
29	0.2	0.3	78.08	62.25	68.75	65	56.25	67	66.75
30	0.2	0.3	78.83	61.25	70.5	64.75	56.5	66.25	63.5
31	0.2	0.4	79.01	65	71.5	67.5	65.5	71	62
32	0.2	0.4	78.07	67.5	75	67.75	61.5	69.75	61
33	0.2	0.4	77.7	67.25	76.25	66.75	70	73.75	62.5
34	0.2	0.4	80.32	68.5	72.5	67	61	75	61.75
35	0.2	0.4	73.32	63	67	66.75	67.25	68	58.25
36	0.4	0.01	49.54	61.5	56	61.5	55	54.75	58.5
37	0.4	0.01	53.45	61.5	59.5	66.75	49.5	54.25	63.5
38	0.4	0.01	48.59	61.5	59	66.75	65	59	62.25
39	0.4	0.01	54.42	63.25	56	66.75	58.5	54.5	62.5
40	0.4	0.01	47.85	62.25	59.5	66.5	55.5	55	63.5
41	0.4	0.025	56.47	63	60	64	54.75	54.5	61.5
42	0.4	0.025	49.34	63.5	62.5	63.25	60.75	56.75	62.75
43	0.4	0.025	49.15	65.25	60	66.75	61	57.25	62.75
44	0.4	0.025	62.65	63.75	62.25	66.25	54.25	59.75	60.75
45	0.4	0.025	51.02	62.75	62.5	65.75	62.5	59.5	59.5
46	0.4	0.05	44.27	61.25	59.25	66.75	59.25	56	61.75
47	0.4	0.05	52.7	61.25	58.75	66.5	60.75	57.25	58.5
48	0.4	0.05	55.34	62.75	63.25	66.75	57.75	54.5	61.75
49	0.4	0.05	60.97	62.75	59	66.75	60	55.5	61.5
50	0.4	0.05	65.3	62.25	62.75	64.25	56.5	56.75	60.5
51	0.4	0.1	48.42	61	59	66.5	66.75	57.25	58.75
52	0.4	0.1	43.53	61.25	60	66.75	66.75	61.25	61
53	0.4	0.1	64.92	61.5	63	66.75	67.75	59	62.25
54	0.4	0.1	60.98	60.75	60.25	66.75	49.75	53.25	60.25
55	0.4	0.1	67.16	63.5	63.75	66.75	55.5	57.75	64
56	0.4	0.2	66.96	63.5	63	66.5	61	57.25	60.5
57	0.4	0.2	66.06	63.25	63.75	66.75	57.5	60.5	61.25
58	0.4	0.2	71.48	65.5	66.5	66.5	61.25	63.25	63.5
59	0.4	0.2	76.54	61	63.25	65.75	54	64.25	60.5
60	0.4	0.2	74.69	62.5	67.25	66.75	55.25	63.5	62.25
61	0.4	0.3	73.7	63.5	64.25	73	62.25	59.75	59.25
62	0.4	0.3	73.21	60	64.75	66.25	57.25	61.25	61.5
63	0.4	0.3	72.81	62	61.25	66.25	60.75	59.75	60.25
64	0.4	0.3	72.04	62.75	67.5	66.75	62.5	63.75	65.25
65	0.4	0.3	81.6	62.75	67.5	66.75	62.5	63.75	65.25
66	0.4	0.4	76.59	63.75	66	66.75	63.75	64.75	60.25
67	0.4	0.4	79.56	64.25	69	66.5	63.25	65.25	65.25
68	0.4	0.4	83.32	60.25	67.5	65.75	54.5	69.5	63.75
69	0.4	0.4	84.8	67	84.75	65	63.25	73.25	63
70	0.4	0.4	85.72	68.25	68.5	65.25	57	78.5	60.5

Table 8: Accuracy of 70 models with 400 samples of with 1:4 ratio of cases and controls

Model	MAF	Heritability	MDR	RF	AC-CPAR	AC-CBA	SVM	NN	NaiveBayes
1	0.2	0.01	44.69	77	72.5	80	64	69	77.75
2	0.2	0.01	52.81	76.25	71	80	66	70.25	77.5
3	0.2	0.01	40	77	75.25	80	68.25	69.75	79.5
4	0.2	0.01	50.78	77.5	71.5	80	68.5	68.5	78.5
5	0.2	0.01	50	78.25	72.5	80	66.25	67.25	78.5
6	0.2	0.025	48.59	80.25	79	74.5	71.5	70.5	78.5
7	0.2	0.025	63.13	78.25	80	80	68.25	69.75	76.75
8	0.2	0.025	56.09	77.25	79.25	79.5	64.75	72	78.5
9	0.2	0.025	62.5	78.5	78.75	80	64.5	70	79.5
10	0.2	0.025	52.5	77	80	80	67.25	72.25	78.5
11	0.2	0.05	56.09	78	79.5	80	69.75	69.75	77.5
12	0.2	0.05	48.29	78.75	79.25	80	67.75	71.25	78.75
13	0.2	0.05	60.16	78	77.5	80	68.5	76.25	79.75
14	0.2	0.05	65.78	79	79.75	76.75	71.25	70.5	78
15	0.2	0.05	62.81	78.25	79.75	79.25	71	71.75	79
16	0.2	0.1	51.56	76.5	72	80	67	68.5	78
17	0.2	0.1	64.22	78	77.25	80	72	68	78
18	0.2	0.1	70.16	78.75	77.5	80	71	74.75	78.5
19	0.2	0.1	68.59	79	77.25	80	71.5	75	77.5
20	0.2	0.1	70.94	78	79	78.75	67	74.5	77.75
21	0.2	0.2	70.16	79.25	78	80	67.5	72	78.5
22	0.2	0.2	75.63	79.5	77.25	80	72.25	74	78.75
23	0.2	0.2	73.75	77.25	80	80	71.5	76	77.5
24	0.2	0.2	74.06	78	79.75	78	71.75	78	77.5
25	0.2	0.2	75.47	79.75	81.75	78.25	72	78.25	78.5
26	0.2	0.3	74.37	79	80	80	68.5	76	77.5
27	0.2	0.3	75	78.75	78	79.5	70.5	80	78.25
28	0.2	0.3	76.72	80.25	84	80	71	79	77.5
29	0.2	0.3	73.75	78.75	78.25	78.75	72.25	75	77.75
30	0.2	0.3	75.63	79	78	80	68.5	70.75	78.5
31	0.2	0.4	81.09	79.5	80.25	80	72.25	78	79.25
32	0.2	0.4	80	78.25	80.25	80	67.5	81.75	77.5
33	0.2	0.4	78.12	78.5	79.25	77.25	75.5	82	78.25
34	0.2	0.4	83.28	79.5	86.75	80	70.25	78	79
35	0.2	0.4	84.22	79.5	89.25	80	75.5	86	79.25
36	0.4	0.01	45.63	76.75	79.25	79.75	68.75	72.75	78.25
37	0.4	0.01	56.25	77.5	80.25	79.75	69	69.25	79.75
38	0.4	0.01	47.66	78	79	80	66.75	68	78
39	0.4	0.01	53.28	77.5	79.75	80	63.5	66	76.75
40	0.4	0.01	50.47	77.5	80	80	66.25	68.75	78
41	0.4	0.025	54.06	77	79.25	80	71.5	71.5	78.75
42	0.4	0.025	44.84	78	79.5	80	68.25	69	77.75
43	0.4	0.025	57.03	79.25	78.75	80	64.75	71.25	75.75
44	0.4	0.025	46.88	76.75	79.75	80	64.5	68.75	79.25
45	0.4	0.025	44.37	78.5	80	80	67.25	68.5	78.75
46	0.4	0.05	59.69	78.25	78.75	80	69	69.25	78.25
47	0.4	0.05	45.78	78.5	79	79.5	69	74	79.25
48	0.4	0.05	49.38	78.25	78.5	80	65.5	70.75	79.25
49	0.4	0.05	43.28	78.75	79.5	80	65	67	79.5
50	0.4	0.05	57.34	77.5	78.75	72.5	64.75	70.5	78.75
51	0.4	0.1	62.34	77.75	79.75	72.25	70	69.25	77
52	0.4	0.1	50.94	77.5	80.25	71.25	67	69.25	79.75
53	0.4	0.1	61.87	78.25	79.5	79.5	67.25	72.25	78.75
54	0.4	0.1	64.06	78.5	79.25	80	67.25	70.75	79
55	0.4	0.1	62.66	76.75	79	80	64.5	70.5	79.25
56	0.4	0.2	67.81	78	80	74.25	73	73.25	78.5
57	0.4	0.2	67.66	77.25	79.5	79.75	64.5	72.25	77.75
58	0.4	0.2	77.81	79.75	79.5	80	65	71.75	79.75
59	0.4	0.2	72.97	80	77.75	80.25	72.5	73.25	78
60	0.4	0.2	69.69	78.5	78.25	80	70.25	71.5	77.5
61	0.4	0.3	69.06	77.75	79	80	70.5	74.25	79
62	0.4	0.3	70	78	79	80	68.25	72.25	78.25
63	0.4	0.3	70.94	76.5	78.5	73.5	73	75.25	79
64	0.4	0.3	75.78	78.25	83	79.25	72.5	73.5	78.5
65	0.4	0.3	80.94	78	80	79.5	68.75	72.75	79.75
66	0.4	0.4	75.31	78.5	78	80	67.75	70.5	79
67	0.4	0.4	78.12	79.5	80	80.25	68.75	74.5	78.5
68	0.4	0.4	81.09	79.25	79.75	79.75	68.5	72.75	79.5
69	0.4	0.4	82.5	79	83.25	79.5	71.25	80.75	79.5
70	0.4	0.4	76.09	78.75	79.25	80.25	74	76.25	77.25