A Comparative Study of Different Classification Techniques on SNP Data

Mohammad Ali[1], Al Hasib Billah[2], Sutapa Dey Barna[2], Mahfuza Yesmin[2], Md. Menhazul Abedin[1] and N. A. M Faisal Ahmed[1]

[1]Lecturer, Statistics Discipline, Khulna University, Khulna-9208, Bangladesh

[2] Research Student, Statistics Discipline, Khulna University, Khulna-9208, Bangladesh

Corresponding author

Mohammad Ali

Email: ali.ru.stat@gmail.com

Abstract

Genetic factor is the most important part for consideration, often leads to the analysis of Single-nucleotide Polymorphism (SNP) which actually causes the trait. SNPs are the large fields of research that have been widely popular in today's modern world. Now-a-days, different classification methods become widely popular in the field of genome wide association study (GWAS) considering significance of SNPs. But there is no serious study in literature for comparing of different classification techniques. In this paper, it is compared four classical classification techniques with a machine learning technique for predicting binary trait given the genotypic information. For this purpose, considered simulated data with the help of R packages. The data sets are partitioned to train and test data consisting 70% and 30% respectively. Then different classification techniques are performed, namely logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), naive bayes (NB) and Support vector machine (SVM). The models are evaluated by the performance measure accuracy, sensitivity, specificity for training data as well as test data. Results suggest that SVM performs slightly better than other techniques.

Keywords: Single-nucleotide Polymorphism, Classification, Simulation Study.

1. Introduction

In recent years, there has been an unprecedented chain of discoveries in the genomic data sets in binary or complex traits. Since 2005, nearly 100 loci for as many as 40 common diseases and traits have been identified and replicated in Genome-wide association studies (GWAS), many in those genes not previously suspected of having any role in the disease under study and some other yet containing unknown genes. Though this limitation, many scientific and biological discoveries have been made through the experimental models of GWAS. The aim of this studies are at detecting variations at genomic level that are associated with the diseases under study. According to previous researches the huge variety of genetic variations associated with disease has grown exponentially. Such as coronary heart disease (CHD), the number of affecting genes as grown from a handful to more than 45 (Altshuler, D., et al 2000). Genome-wide characterization of the levels and patterns of human genetic variation has enabled the researcher possible to search the genes liable for. With the help of studies identifying the associated genes making the DNA markers related with the desired diseases ranging from acute one to chronic severe types make the perfections of finding those genes responsible for in reality. GWASs typically focus on associations between single-nucleotide polymorphisms (SNPs) (Lipka, A.E., G.P. McCabe, and Doerge R 2009) and traits like major human diseases, but can equally be applied to any other organism. There are small variations in the individual nucleotides of the genomes (SNPs) as well as many larger variations, such as deletions, insertions and copy number variations. Any of these may cause alterations in an individual's traits, or phenotype, which can be anything from disease risk to physical properties such as height. The most common methods are based on a case-control design (Clarke, G.M., et al (2011) and try to find marker loci associated to the disease by comparing

genotype frequencies between random samples of cases (diseased) and controls. These associations result in the way of categorical outcomes e.g. binary or any numerical category which eventually leads us to the use of logistic regression. It is mostly used and widely popular analyzing technique among the researchers. A handful number of researcher use a few other techniques. There is no serious study to compared classification technique on genome-wide association. Five classification techniques are performed on three simulated data sets. It is found that SVM provides slight better results than others.

2. Methods and materials

2.1. Classification

Classification is the process of identifying, naming and categorizing living subjects based on their physical and biological characteristics. It is the arrangement of organisms into orderly groups based on their similarities. Classification has two distinct meanings. It may be given a set of observations with the aim of establishing the existence of classes or clusters in the data. In this research logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), naive bayes (NB) and Support vector machine (SVM) are used.

2.2. Single-nucleotide Polymorphism (SNP)

A Single-nucleotide polymorphism or SNP (pronounced snip) is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual). For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide (Manuelidis, L 1982). In this case, we say that there are two alleles: C and T. Almost all common SNPs have only two alleles. Within a population, SNPs can be

assigned a minor allele frequency — the lowest allele frequency at a locus that is observed in a particular population. This is simply the lesser of the two allele frequencies for single-nucleotide polymorphisms (SNP) (Gunderson, K.L., et al (2005). There are variations between human populations, so a SNP allele that is common in one geographical or ethnic group may be much rarer in another. SNPs occur normally throughout a person's DNA. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome. Most commonly, these variations are found in the DNA between genes. They can act as biological markers, helping scientists locate genes that are associated with disease. When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the gene's function. Most SNPs have no effect on health or development. Some of these genetic differences, however, have proven to be very important in the study of human health. Researchers have found SNPs that may help predict an individual's response to certain drugs, susceptibility to environmental factors such as toxins, and risk of developing particular diseases. SNPs can also be used to track the inheritance of disease genes within families. Future studies will work to identify SNPs associated with complex diseases (Nachman, M.W. 2001) such as heart disease, diabetes, and cancer.

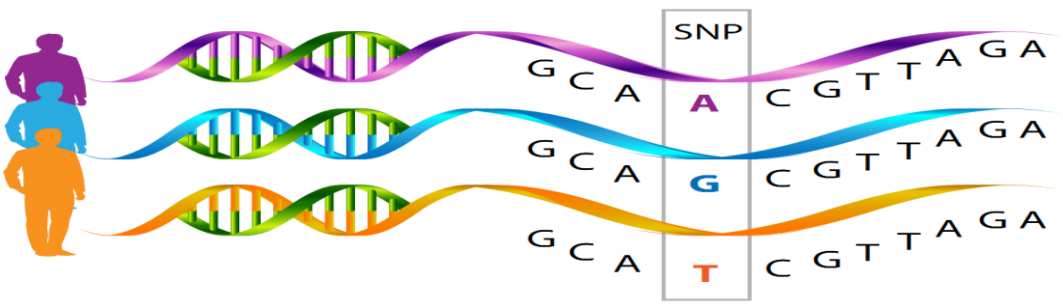

Figure 1 – A SNP

2.3. Simulation Study

To perform the classification with dataset collected from real genomic coding is quietly difficult to collect. The data may not be always available and sometimes can cut a huge budget. Furthermore, SNP data mostly rely on advanced lab facility which sometimes become very hard to manage. Get rid off this difficulties simulated data using "R" (scrime package) was used which one is most popular in this field.

Data of several observations (case-control) and different SNPs are simulated using the R function "simulateSNPglm" where each SNP exhibits minor allele frequency of 0.25 and 0.07. Following flowchart represent the whole simulation study.

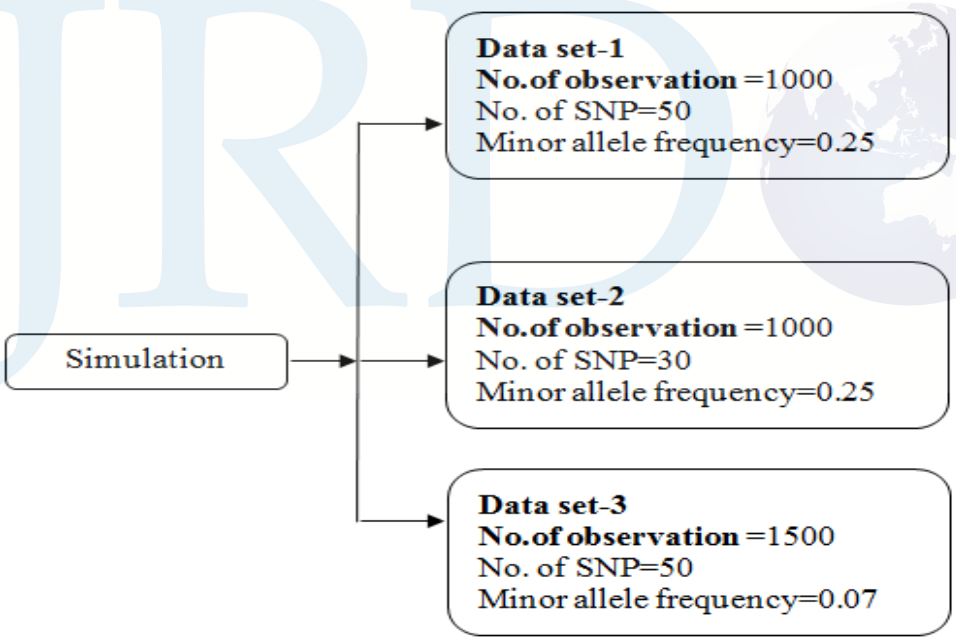

Figure 2- Simulation study

3. Results and Discussion

The emission results of some well-known classification techniques are presented. Data sets split into two-part named training (70%) and test (30%) and then apply Logistic, LDA, QDA, NB and SVM. Results are presented through tabular form.

3.1. Kernel choice for SVM on data set-1

For training data and test data the SVM with different kernels on data set-1 are shown in Table 1.

Table 1- SVM with different kernels results for data set-1

SVM Model	Performance Measure (training)			Performance Measure (test)		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Linear Kernel	0.5643	0.0000	1.0000	0.4833	0.0000	1.0000
Polynomial Kernel	0.5643	0.0000	1.0000	0.4833	0.0000	1.0000
Radial Kernel	1.0000	1.0000	1.0000	0.5933	0.4387	0.7586
Sigmoid Kernel	0.5443	0.0033	0.9620	0.4767	0.0129	0.9724

From Table 1 observed that that SVM with radial kernel performs better result than other kernels for both training and test data.

3.2. Comparison SVM with other classification techniques on data set 1

In order to compare the five-classification techniques: Logistic, NB, LDA, QDA and SVM the classification performance measures such as accuracy, sensitivity, specificity for 70% training and 30% test data set. SVM with radial kernel is chosen because it performs better seen in Table 1. These performance measures are shown in Table 2.

Table 2- Comparison of different classification techniques for data set 1

Classification Model	Performance Measure (training)			Performance Measure (test)		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SVM	1.0000	1.0000	1.0000	0.5933	0.4387	0.7586
Logistic	0.3143	0.3088	0.3219	0.42	0.3892	0.4696
NB	0.6757	0.5672	0.7595	0.5733	0.4452	0.7103
LDA	0.6714	0.5836	0.7392	0.4833	0.0129	0.9862
QDA	0.9286	0.9344	0.9241	0.5203	0.4065	0.6483

From Table 2 observed that SVM performs better result than others classifier.

3.3. Kernel choice for SVM on data set-2

For training data and test data the SVM with different kernels on data set 2 are shown in Table 3.

Table 3- SVM with different kernels results for data set-2

SVM Model	Performance Measure(training)			Performance Measure (test)		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Linear Kernel	0.4962	0.0000	1.0000	0.5200	0.0000	1.000

Polynomial Kernel	0.4962	0.0000	1.0000	0.5200	0.0000	1.000
Radial Kernel	0.9695	0.9395	1.0000	0.6378	0.6250	0.6496
Sigmoid Kernel	0.4943	0.0000	0.9962	0.5200	0.0000	1.000

From Table 3 observed that for training data set the SVM with radial kernel performs better result than other kernels.

3.4. Comparison SVM with other classification techniques on data set 2

In order to compare the five-classification techniques: Logistic, NB, LDA, QDA and SVM the classification performance measures such as accuracy, sensitivity, specificity for 70% training and 30% test data set. SVM with radial kernel is chosen because it performs better seen in Table 3. These performance measures are shown in Table 4.

Table 4- Comparison of different classification techniques for data set 2

Classification Model	Performance Measure(training)			Performance Measure (test)		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SVM	0.9695	0.9395	1.0000	0.6378	0.6250	0.6496
Logistic	0.3876	0.3904	0.3849	0.3622	0.3463	0.3790
NB	0.6086	0.6163	0.6008	0.6156	0.6250	0.6068
LDA	0.6124	0.6181	0.6065	0.6133	0.4306	0.7821
QDA	0.7752	0.7977	0.7524	0.6022	0.5741	0.6282

Table 4 conclude that SVM performs better result than Logistic, NB, LDA, and QDA.

3.5. Kernel choice for SVM on data set-3

For training data and test data the SVM with different kernels on data set 3 are shown in Table 5.

Table 5-SVM with different kernels results for data set-3

SVM Model	Performance Measure(training)			Performance Measure (test)		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Linear Kernel	0.6390	0.0000	1.0000	0.6267	0.0000	1.0000
Polynomial Kernel	0.6390	0.0000	1.0000	0.6267	0.0000	1.0000
Radial Kernel	0.8600	0.6121	1.0000	0.6289	0.09524	0.94681
Sigmoid Kernel	0.6314	0.0079	0.9836	0.6289	0.0104	0.01190

From Table 5 observed that for training data set the SVM with radial kernel performs better result than other kernels.

3.6. Comparison SVM with other classification techniques on data set 3

In order to compare the five-classification techniques: Logistic, NB, LDA, QDA and SVM the classification performance measures such as accuracy, sensitivity, specificity for 70% training and 30% test data set. SVM with radial kernel is chosen because it performs better seen in Table 5. These performance measures are shown in Table 6.

Table 6- Comparison of different classification techniques for data set 3

Classification Model	Performance Measure(training)			Performance Measure (test)		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Logistic	0.2971	0.2792	0.3586	0.3333	0.3047	0.4196
SVM	0.8600	0.6121	1.0000	0.6289	0.09524	0.94681
NB	0.7029	0.5435	0.7928	0.6556	0.5298	0.7305
LDA	0.7057	0.4116	0.8718	0.6733	0.4048	0.8333
QDA	0.7419	0.6596	0.7884	0.6289	0.5298	0.6879

From Table 6 it is seen that SVM gives highest accuracy and specificity than Logistic, NB, LDA and QDA but QDA gives the highest sensitivity. In the sense of research objective SVM is better classifier than others.

4. Conclusions

In order to compare different classification techniques three simulated SNP data is considered. Then perform five classification techniques. From findings, it is concluded that SVM gives better results than other techniques for predicting binary traits given genotype information. SVM tool should be emphasized more for better statistical analysis of classification problem on SNP data. It demands serious attention from bioinformatics community for its popularity.

Acknowledgments: Most acknowledge goes to Statistics Discipline, Khulna University, Khulna, where this study is conducted also authors are thanks to author of “scrime” package. Authors are very much thankful to different author whose articles are used in this paper.

References

1. Altshuler, D., et al (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing, Nature, 407(6803): p. 513-516.
2. Clarke, G.M., et al (2011). Basic statistical analysis in genetic case-control studies, Nature protocols. p. 121-133.
3. Gunderson, K.L., et al (2005). A genome-wide scalable SNP genotyping assay using microarray technology, Nature genetics. 37(5): p. 549-554.
4. Lipka, A.E., G.P. McCabe, and Doerge R (2009). Associating SNPs with binary traits.
5. Manuelidis, L (1982). Nucleotide sequence definition of a major human repeated DNA, the Hind III 1.9 kb family. Nucleic acids research,. 10(10): p. 3211-3219.
6. Nachman, M.W. (2001). Single nucleotide polymorphisms and recombination rate in humans. TRENDS in Genetics, 17(9): p. 481-485.