Evaluation Metrics for Feature Selection in Population Genomic Data

Ioannis Kavakiotis[1], Alexandros Triantafyllidis[2], Grigorios Tsoumakas[1], Ioannis Vlahavas[1]

[1]Department of Computer Science, Aristotle University of Thessaloniki, 54124,Greece
{ikavak, greg, vlahavas}@csd.auth.gr
[2]Department of Genetics, Development and Molecular Biology, School of Biology, Aristotle University of Thessaloniki, 54124, Greece
atriant@bio.auth.gr

Abstract. Single Nucleotide Polymorphisms (SNPs) are considered nowadays one of the most important class of genetic markers with a wide range of applications with both scientific and economic interests. Although the advance of biotechnology has made feasible the production of genome wide SNP datasets, the cost of the production is still high. The transformation of the initial dataset into a smaller one with the same genetic information is a crucial task and it is performed through feature selection. Biologists evaluate features using methods originating from the field of population genetics. Although several studies have been performed in order to compare the existing biological methods, there is a lack of comparison between methods originating from the biology field with others originating from the machine learning. In this study we present some early results which support that biological methods perform slightly better than machine learning methods.

Keywords: Feature selection; Single nucleotide polymorphism; SNPs; Bioinformatics; Machine learning

1 Introduction

Single Nucleotide Polymorphisms (SNPs) are considered nowadays one of the most important classes of genetic markers in many scientific fields such as human genetics and molecular ecology. At the same time analyses of SNP data are nowadays gaining more popularity in many applications with great medical as well as economic interest, such as food traceability, discrimination between wild and framed population and forensic investigations [1].

In recent years, the advancement of biotechnology has made feasible the genotyping of thousands or even millions of SNPs and consequently the production of very large SNP datasets from a wide range of model and non model organisms. Two important tasks in terms of analyses, performed with SNP datasets are the individual assignment to groups of origin and the selection of the most informative markers

(SNPs). In terms of computer science these are the classification and the feature selection tasks respectively.

The importance of feature selection for SNP datasets is beyond dispute either for biology or machine learning. From the computational view, feature selection aims to improve the prediction performance of the predictors through defying the curse of dimensionality, to provide faster and more cost effective predictors and to facilitate data visualization and data understanding [2]. From the biology perspective, the importance of selecting the most informative markers from genome wide datasets has been stated in many scientific projects and publications [1, 3, 4] and are mainly economic. The main drawback in using genome wide datasets is the high cost to produce, in contrast to smaller panels (datasets) which are cheaper and more flexible and can occur through/after the feature selection process.

The methods to measure marker informativeness in population genetics and consequently to select features are not similar to those used in the field of machine learning. More specifically, the two most popular genetic methods depend on the allele frequencies of a particular marker i.e. attribute (explained in subsequent section). Moreover the assignment success of the reduced dataset is evaluated through genetic assignment approaches or software such as GENECLASS2 [5]

The purpose of this study is to compare, for the first time, the frequency based metrics which are used for feature selection purposes in population genetics with established methods from the field of machine learning and data mining and eventually determine which are more appropriate for feature selection when building models for the classification of individuals into groups of origin.

2 Background Knowledge

The Genetic Information: from DNA to Genome.
Deoxyribonucleic acid (DNA) is the primary hereditary material of all known living organisms. It is a very large molecule composed of smaller units linked together. These smaller units are four (for DNA) and are called nucleotides named adenine (A), cytosine (C), guanine (G), thymine (T).

Genes are made of nucleotides and each one contains a particular set of instructions. A chromosome is an organized structure of DNA which contains genes as well as many other nucleotide sequences, with gene regulatory or no (known) function. The genome is the entirety of hereditary information possessed by an organism and it is encoded mostly in DNA.

Single Nucleotide Polymorphisms – SNPs.

Single nucleotide polymorphism is the most common type of genetic variation. A SNP occurs when a Single Nucleotide (A, T, G or C) in the genome differs between members of a biological species or paired chromosomes. For instance, consider two DNA sequences from different individuals; *seq1:* AT**C**TG and *seq2:* AT**G**TG, which differ in the third nucleotide.

An allele is one of the possible alternative forms of the same gene or same genetic locus. In the previous example there were two alleles, which is the most common case in SNP variations. A marker with only two alleles is called biallelic.

Population Genomic Datasets – SNP Datasets.

The dimensionality of SNP datasets can vary a lot depending on the number of different SNPs analyzed multiplied by the number of samples analyzed. Animal datasets can reach a hundred thousand attributes (SNPs) whereas human dataset can easily contain over a million SNPs. Each attribute is a genotype which occurs from the combination of two nucleotides in the two chromosomal copies of a diploid organism and thus can have at most three values. For instance, one SNP genotype can have the following values AA, GG and AG (GA is considered same) which occur from the combination of the two alleles adenine and guanine.

The term allele frequency refers to the frequency of each allele in a population. For instance, consider a dataset with five individuals. Assume that the first attribute can have the following values TT, CC, TC. Three individuals have values TT, one CC and one TC. The allele frequencies are 0.7 for the allele T and 0.3 for the allele C.

3 Methods

3.1 Feature Selection via Filter Methods

Feature selection methods are divided in two major categories. The first category comprises of 'filter' methods, which evaluate the attributes based on general characteristics of the data. The second category contains "wrapper" methods which use a machine learning algorithm to evaluate the candidate subset of features [6]. The main advantage of wrappers is that they commonly offer better classification accuracy which is the most important aim. The main disadvantage of wrapper methods is that the algorithm will build a model many times in order to evaluate different subsets, which in some cases, such as SVMs, is too computationally expensive. This is an important drawback especially for SNP datasets which can be very large. On the contrary, the main advantage of filters is that they are much faster than wrapper methods.

3.2 Evaluation Metrics

Allele Frequency Based Methods

In our study we included two methods for evaluation of marker informativeness from the field of population genetics, although until now several methods have been proposed. The first is Delta [7] which is without dispute the most commonly used method. The second one is Pairwise Wright's F_{ST} [8] which is reported to be the most successful one [1]. For the two methods we use the same notations: p_A^i is the frequency of the allele A in the i^{th} population, p_A^j is the frequency of the same allele A in the j^{th} population and p_A is the frequency of allele A in all populations. The notations

with the second allele *B* occur similarly. Delta and pairwise Wright's F_{ST} are given by calculating the mean of their values out of all possible population combinations.

Delta

For a biallelic marker i.e a marker with two alleles the delta value is given by the following equation:

$$\delta = |p_A^i - p_A^j|$$

Pairwise Wright's F_{ST}.

For a biallelic marker the F_{ST} value is given by the following equation:

$$F_{ST} = (H_t - H_s) / H_t$$

where H_s is the average expected heterogygosity across subpopulations and H_t is the expected heterozygosity of the total population [9] and they are given by the following equations:

$$H_t = 2\, p_A\, p_B$$

$$H_s = p_A^i\, p_A^j + p_B^i\, p_B^j$$

Machine Learning Methods

The following two attribute evaluators from machine learning are implemented in Weka machine learning library [10].

InfoGain

This criterion evaluates features by measuring the information gain with respect to the class [11]. Information gain is given by

$$Infogain = H(Y) - H(Y|X)$$

where

$$H(Y) = -\sum_{y \epsilon Y} p(y) log_2 (p(y))$$

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) log_2 (p(y|x))$$

ReliefF

The general idea behind ReliefF [12] is that scores features according to their value's ability to distinguish better instances from different classes and group together instances of the same class. This is achieved by sampling instances randomly and checking neighboring instances of the same and different classes.

3.3 Classification Algorithm – Decision Trees

The comparison of the different feature ranking methods was made through the comparison of the classification accuracy of different subset of features, with decision trees (J48). Decision trees have many advantages such as simplicity and high interpretability. Moreover, a feature selection process can aid the algorithm to produce smaller and more predictive trees [11].

4 Evaluation

The experiments were conducted on a dataset [4], which consisted of 59436 attributes/SNPs and 446 instances/individuals that are almost equally distributed between 14 different classes/groups. For every feature selection method we followed the following procedure: We built twelve times the classifier using 10 – fold cross-validation with different number of attributes. Each time we selected the top-k most informative SNPs (20, 40, 60, 80, 100, 120, 160, 200, 300, 400, 500, 1000) from the training folds in order to avoid having upwardly biased results.

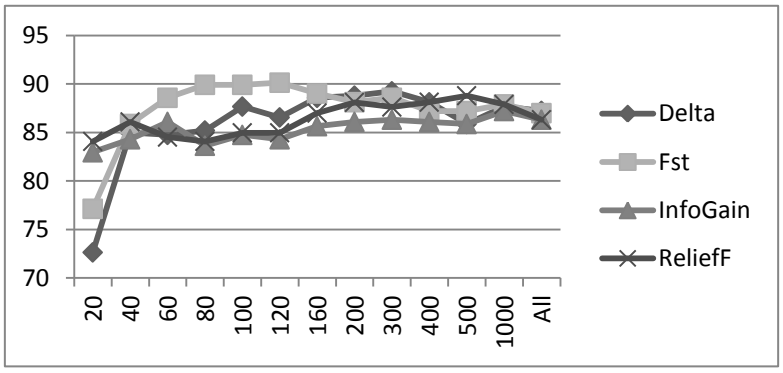

Fig. 1. Comparison of Feature selection methods through classification accuracy

The results clearly show that the best method for selecting the most informative SNPs is Pairwise Wright's F_{ST} which is the only method that reaches 90,1% accuracy so early (120 SNPs). The other methods need many more attributes to reach the same level of accuracy (Delta 89%/300SNPs, InfoGain 87/1000, ReliefF 89/500). Another interesting observation is that the first twenty SNPs selected by the machine learning methods are significantly more informative than those selected by the allele frequency methods.

Table 1. Score with first 20 SNP and best score for each method

	Delta	PairwiseF_{ST}	InfoGain	ReliefF
Best Score	89	90.1	87	89
20 first SNPs	72.6	77.13	82.96	84.08

5 Conclusion and Future Work

Although some studies have been conducted to compare biological methods for selecting the most informative markers from a SNP dataset, none of them compare biological with machine learning methods. In this study we present some early results of our attempt to compare the most successful biological with the most successful machine learning methods and determine which are the most appropriate for the task of informative SNP selection. In the future, we intend to broaden our study in various ways. Firstly, we intent to compare other more sophisticated techniques from the machine learning field. Secondly, we aim to evaluate methods with more classifiers in order to support our conclusions with more confidence and lastly, to run our experiments with more datasets. Those datasets will contain populations/groups with varying degrees of differentiation trying to respond to real life biological problems.

6 References

1. Wilkinson, S., Wiener, P., Archibald, AL., et al.: Evaluation of approaches for identifying population informative markers from high density SNP chips. BMC Genet, 12, 45 (2011)
2. Guyon, I., Elisseeff,A., An introduction to variable and feature selection. J. Mach Learn Res., 3, 1157–1182 (2003)
3. Nielsen, EE., Cariani, A., Mac Aoidh, E., et al.: Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. Nat. Com., 3, 851 doi:10.1038/ ncomms 1845 (2012).
4. Wilkinson, S., Archibald, AL., Haley, CS., et al. Development of a genetic tool for product regulation in the diverse British pig breed market. BMC Gen, 13, 580 (2012)
5. Piry, S., Alapetite, A., Cornuet, J.M., Petkau, D., Baudouin, L., Estoup. A., GENECLASS2: A software for genetic assignment and first generation migrant detection. *J Hered*, **95**, 536-539 (2004)
6. Witten, I.H., Frank, E., Hall, M.A., Data Mining: Practical Machine Learning Tools and Techniques (third edition). Morgan Kaufmann, Burlington, MA (2011)
7. Shriver, M.D., Smith, M.W., Jin, L., et al. Ethnic affiliation estimation by use of population-specific DNA markers. Am J Hum Genet, 60, 957-964 (1997).
8. Wright, S., The genetical structure of populations. Ann Eugenic, 15, 323 (1951)
9. Beebee, T., Rowe, G., An Introduction to Molecular Ecology. Oxford University Press, Oxford, UK (2004).
10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., The WEKA Data Mining Software: An Update; SIGKDD Explorations, 11, 10–18. (2009)
11. Wang,Y., et al. Gene selection from microarray data for cancer classification–a machine learning approach. Comput. Biol. Chem., 29, 37–46 (2005).
12. Robnik-Sikonja, M., Kononenko, I., Theoretical and empirical analysis of relief and relieff. Mach Lean, 53, 23–69 (2003)