Ensemble Feature Selection using Rank Aggregation Methods for Population Genomic Data

Ioannis Kavakiotis
Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
ikavak@csd.auth.gr

Alexandros Triantafyllidis
Department of Genetics, Development & Molecular Biology, School of Biology, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
atriant@bio.auth.gr

Grigorios Tsoumakas, Ioannis Vlahavas
Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
{greg,vlahavas}@csd.auth.gr

ABSTRACT
Single Nucleotide Polymorphisms (SNPs) constitute important genetic markers with numerous medical and biological applications of high scientific and economic interest. SNP datasets are typically high dimensional, containing up to million features. Reasons originating from both biology and machine learning, dictate to perform feature selection which is mainly performed after feature evaluation. In this paper we present methods for SNP evaluation and eventually selection, based on combining results obtained from established genetic marker evaluation methods originating from the field of population genetics. To achieve this we have formulated the feature selection task as a ranking aggregation problem, which is a classical problem in social choice and voting theory.

CCS Concepts
• **Applied computing → Life and medical sciences→ Bioinformatics**

• **Applied computing →Life and medical sciences → Genetics → Population genetics**

• **Computing methodologies → Machine learning → Machine learning algorithms → Feature selection**

Keywords
Bioinformatics; Data Mining; Feature Selection; Rank aggregation; Single Nucleotide Polymorphism; Informative Marker Selection; Population genomics

1. INTRODUCTION
Nowadays, many scientific fields, including biology, have entered the era of big data [1]. The term big data is a broad term to describe very high dimensional datasets and all related difficulties to process and analyze them [2]. Terabyte sized datasets are now common in biology [3], due to significant advances in biotechnology and more specifically high-throughput technologies, which have enabled even small laboratories to become big data generators [1].

SNPs are nowadays considered one of the most important biological markers due to the fact that SNP data analysis is used in applications such as identification of disease related mutations, Qualitative Trait Loci detection (QLT), food traceability, brand authentication, discrimination between wild and/or farmed populations and anthropological forensic investigations [4].

SNP datasets are considered high dimensional. Animal and plant datasets can contain hundreds of thousands loci whereas human datasets millions [5]. High dimensional datasets come along with many difficulties both in processing and analysis. High computational cost in training and the decrease of the generalization ability due to the curse of dimensionality [6] are two of the most important ones from the data mining perspective. From a biology perspective the production of genome wide dataset is too expensive. Fortunately, for many high dimensional datasets most features are irrelevant to the outcome and consequently the elimination of those redundant features can improve the classification accuracy [6]. This is also the case for SNP datasets. An important task in SNP analysis is the assignment of individuals, or groups of individuals, to their population of origin, based on their (multi-locus) genotypes [7], which is the classification task performed with specialized classifiers for genetic data. It is evident from the above that a feature selection step before the classification is an essential step. Feature selection in SNP datasets is based on ranking SNPs according to their informativeness i.e. the marker information content, which is the amount of information that a locus holds regarding the ancestry of an individual [8].

Although many methods exist in the field of population genetics to perform the feature evaluation task, no method can be broadly accepted as the most successful one because none outperforms the others in all circumstances [5]. In this paper we present methods for SNP evaluation and eventually selection, based on combining established SNP selection methods from the field of population genetics. These methods are inspired by the ranking aggregation problem which is a classical problem in social choice and voting theory.

2. BACKGROUND KNOWLEDGE

2.1 Single Nucleotide Polymorphisms – SNPs
Single nucleotide polymorphisms are the most common type of genetic variation among living organisms. Each SNP represents a specific difference in a single nucleotide. For instance, consider two genotyped DNA fragments from different individuals*seq1:* AT**C**TG and *seq2:* AT**G**TG. Those two differ in one nucleotide in the third position. SNPs can be extremely harmful especially when

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
SETN '16, May 18-20, 2016, Thessaloniki, Greece
© 2016 ACM. ISBN 978-1-4503-3734-2/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2903220.2903233

they are located in genes or gene regulatory regions and harmless, mainly when they are located in non-coding regions i.e. not in genes.

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 (C and G), which is the most common case in SNP variations. A marker with only two alleles is called biallelic.

2.2 Population Genomic Datasets – SNP Datasets

SNP datasets are high-dimensional. Usually animal datasets can reach a hundred thousand attributes (SNPs), whereas human datasets can contain over a million SNPs. Each attribute of a SNP dataset is a biallelic marker genotype i.e. a variation between two nucleotides. Consequently, each attribute can have at most three values, occurring from the combination of two nucleotides. For instance, one SNP can have the following genotype values AA, GG and AG (GA is considered the same) which occur from Adenine and Guanine alleles.

3. SNP EVALUATION METHODS

Comparisons of the different metrics have been published many times and results are contradictory. In general, no method outperforms all others in all cases and differences between metrics are marginal [5].Based on the findings of Ding et al.[9] and Wilkinson et al. [4] we decided to use Pairwise Wright's F_{ST}[10] and Informativeness for Assignment [8]in the experimental process described later. These metrics are probably the most informative and Delta [11], is probably the most commonly used measure of marker informativeness.

3.1.1 Delta

For a biallelic marker the delta value is given by the following equation:

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

where p_A^i is the frequency of the allele A in the i^{th} population and p_A^j is the frequency of the same allele in the j^{th} population. It is important to mention that delta is calculated only between two populations, so if there are more than two populations, the delta value is computed for each one of every possible combination between existing populations and their average is subsequently calculated in order to produce a value for each SNP marker.

3.1.2 Pairwise Wright's F_{ST}

Pairwise Wright's F_{ST} for more than two populations is computed with the same approach as outlined for delta. For a biallelic marker the F_{ST} value is given by the following equation:

$$F_{ST} = \frac{H_t - H_s}{H_t}$$

where H_s is the average expected heterozygosity across subpopulations and H_t is the expected heterozygosity of the total population [36] and they are given by the following equations:

$$H_t = 2 * p_A * p_B$$

$$H_s = p_A^i * p_A^j + p_B^i * p_B^j$$

where 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. Notations for allele B are defined similarly.

5.3.2 Informativeness for Assignment (I_n)

I_n is a mutual information-based statistics that takes into account self-reported ancestry information from sampled individuals [8].

$$In = \sum_{j=0}^{N} (-p_j \log_2 p_j + \sum_{i=0}^{K} (p_{ij} \log_2 p_{ij})/K)$$

where i= 1,2,…,K are the populations with K≥2 and N are the loci, p_{ij} denotes the frequency of allele j in population I and p_j denotes the average frequency of allele j over K populations.

In all cases allele frequencies are calculated as in [9].

4. VOTING SYSTEMS AND RANKING AGGREGATION PROBLEM

Voting system is called a method by which voters choose between available options, with its most common application being the election process. Social choice theory or voting theory is a scientific subfield combing economics, political sciences and mathematics. There are numerous voting systems and they are mainly characterized by the method that voters use to express their preferences. The voting systems we used in this study belong to the class of preferential voting systems or ranked voting systems in which voters rank options in a hierarchy on the ordinal scale. The core of each preferential voting system is how those votes can be aggregated to yield a final result which is the so called ranking aggregation problem. More formally, the ranking aggregation problem is the problem of computing a consensus ranking of the alternatives, given the individual ranking preferences of several voters [12].

In recent years problems regarding aggregation of few long lists have gained much popularity due to two important fields of their applications. The first field is the World Wide Web and the aggregation of results from different search engines, to provide a more successful ranking. The second field is bioinformatics and the analysis of high-throughput data mainly combining results from different gene expression experiments [13]. Considering that marker informativeness measures produce long lists of ranked SNPs, this kind of high level meta-analysis can be performed to combine results obtained from different measures and therefore produce a more successful list in terms of classification accuracy.

4.1 Ranking Aggregation Methods

All implemented methods used in this study belong to heuristic methods [13] and can be divided in two categories, the borda methods and those which use Markov chains.

4.1.1 Borda Methods

Borda methods are intuitive, easy to understand and are based on the initial method proposed by French mathematician and political scientist Jean-Charles de Borda in 1781[14]. Each one of the L voters ranks options in a hierarchy producing a rank list of choices. In our setting voters are represented by the three genetic methods (Delta, Pairwise F_{ST} and I_n) and the available options are considered

the SNPs. The score of each choice (u) at the i^{th} ranked list (Borda Score – $B(u_i)$) is its position in the list. The final score of each option is calculated from the following function:

$$B(u) = f(B_1(u), B_2(u), \dots, B_L(u))$$

It is obvious that the function f can be any function that can combine the scores. In our study we used the following functions:

- $f(x_1, x_2, \dots, x_L) = median\{|x_1|, |x_2|, \dots, |x_L|\}$ $(median)$
- $f(x_1, x_2, \dots, x_L) = (\prod_{l=1}^{L} |x_l|)^{1/L}$ $(geometric\ mean)$
- $f(x_1, x_2, \dots, x_L) = \sum_{l=1}^{L} |x_l|^p / L$ $(p - norm)$
- $f(x_1, x_2, \dots, x_L) = \sum_{l=1}^{L} |x_l|^1 / L$ $(arithmetic\ mean = p - norm, p = 1)$

4.1.2 Markov Chain Methods

The other two methods are based on Markov chains and they are less intuitive. The main idea is to construct the transition matrix of an ergodic Markov chain. Its stationary distribution will assign a larger probability to a state that is ranked higher [13].The two Markov methods differ in the construction of the transition matrix. The chain will move to a state with better ranking in at least one of the input lists in MC1, whereas half of the input lists in MC2. The exact description of the method can be found in [13].

5. EXPERIMENTAL SETUP

In this subsection we present the experimental process and the evaluation of the proposed methods.

5.1 Dataset

For the experimental analysis we used the dataset which is described in detail in the work of Wilkinson et al. [15]. The dataset is a complete coverage of the types of pigs in the UK and consists of 446 pigs from 7 traditional British breeds, 5 commercial purebreds, an imported European breed and an imported Asian breed that were genotyped with the PorcineSNP60 BeadChip (59436 SNPs)[16].

5.2 Data Preprocessing

The first step in data analysis is the preprocessing of the data. The preprocessing step includes a series of analyses in order to ensure high quality of data submitted to further analysis and consequently to ensure the high quality of the final results. During the preprocessing step, the following analyses were performed: Firstly, missing value detection. Afterwards, validation of the dataset to contain only biallelic markers, i.e. containing no more than two alleles, which is a common problem occurring mainly in non-curated data. Moreover, conversion of the data in all the formats used in the analysis (PED and ARFF files) using the PED Converter offered in software TRES [5]. Finally, splitting the dataset into train and test set (70/30), again using TRES.

5.3 SNP Evaluation and Rank Aggregation

The main analysis started with the three established methods (Delta, F_{ST} and I_n). We first obtained the three ranked SNP lists which correspond to the three different evaluation algorithms. The SNP evaluation task was performed using TRES. Afterwards, the three lists were used as input to the six different ranking aggregation algorithms. Consequently, six new SNP rankings were produced.

5.4 Method Comparison

For the evaluation step we used GENECLASS2.0 [17]. GENECLASS2.0 is a software that offers an extensive list of genetic assignment methods. Such methods assign an individual to the population of origin based on its genotype. In our case we used the Bayesian assignment method of Rannala and Mountain [18] which has been extensively used.

For the evaluation of each method (the three genetic and six proposed) we followed the same procedure. Firstly, we generated reduced datasets (in GENEPOP format) using TRES. Each generated dataset was based on the original training and evaluation datasets (train and test) and contained a subset of the original 59,436 SNPs. The containing attributes/SNPs of each dataset were based on each method's ranked SNP list. Each subset contained some of the most informative SNPs starting from the 20 most informative up to 200 with 20 SNPs step. Therefore for each method we created 10pairs of data sets (train and test) which were evaluated separately in Geneclass2 using the genetic assignment test proposed by Rannala and Mountain.

6. RESULTS AND DISCUSSION

The following figures (figure 1) depict the experimental procedure's results. More specifically, figure 1 shows the assignment accuracy for the first 100 SNP. Assignment accuracies for subsets of 120 – 200 SNPs are omitted intentionally because all methods have managed to surpass the threshold of 95%, which is the desired in similar studies, with less than 100 SNP. Methods illustrated in figure 1 are the Arithmetic Mean (ARM), Geometric Mean (GEM), Median (MED), p - norm for p = 2 (L2N), Markov Chain 1 (MC1), Markov Chain 2 (MC2), Delta, Pairwise Wright's F_{ST} (PWFst) and Informativeness for Assignment (In).

SNPs	ARM	GEM	MED	L2N	MC1	MC2	Delta	PWFst	In
20	0.76	0.76	0.76	0.78	0.76	0.78	0.62	0.84	0.78
40	0.93	0.93	0.92	0.94	0.92	0.94	0.93	0.91	0.93
60	0.94	0.95	0.93	0.95	0.94	0.93	0.93	0.92	0.94
80	0.96	0.97	0.96	0.96	0.97	0.97	0.96	0.96	0.94
100	0.99	0.99	0.99	0.99	0.99	0.98	0.97	0.96	0.96

Figure 1: Assignment accuracy of the methods for 5 SNP subsets. Green and red cells depict the best assignment accuracy achieved at each SNP subset from a proposed method or from an established method respectively. Blue cells depict the first time a method surpassed the 95% accuracy.

The green color cells in Figure 1 indicate those cases where proposed methods had better assignment accuracy than the three established genetic ones for a certain number of SNPs. Contrariwise, red cells depict cases where established genetic methods performed better. Finally, blue cells depict cases where the desired threshold of 95% has been surpassed for the first time.

Figure 1 shows that PWF_{ST} surpasses all other methods in assignment accuracy at 20 SNPs reaching 84%. At 40 SNP there is a completely different situation where all metrics exceed the assignment accuracy of PWF_{ST}. Moreover, two of the proposed methods exhibit best performance (L2N and MC2). At 60 SNPs two of the proposed methods surpass the desired limit of 95%. At 80 SNPs, all methods except I_n have exceeded the threshold of 95%. Again, two of the proposed methods achieved the best performance. Finally, at 100 SNP all methods surpass the 95% threshold, whereas all the proposed methods achieve almost 100%.

Results clearly show that aggregation methods, although slightly, performed better compared to the established genetic methods almost in all SNP subsets except for initial 20 SNP based subset, where PWF_{ST} performed better, although failing to exceed the desired threshold of 95%. A general conclusion is that ranking aggregation methods tend to increase the performance of the methods. Considering that these aggregation methods have been implemented intact, without any adaptation to the examined problem of SNP selection, it is evident that this study can be the

basis on which amendments would give greater results compared to the established genetic methods.

7. CONCLUSION AND FUTURE WORK

This study is the first attempt in literature, to our knowledge, to apply ranking aggregation methods for the informative marker selection task. The initial purpose of this study is to decide whether the aggregation of results obtained by the established genetic methods can produce comparable or better results. Initially, we implemented six different aggregation algorithms producing six different metrics for SNP evaluation. Afterwards, we performed a comparative study with the three most commonly used algorithms in the area in order to evaluate their accuracy. Initial results showed that the proposed methods performed better than the established methods, but they are not able to replace them. Moreover, experiment should be performed in more datasets, because the use of a single dataset is a limit for the validity of the conclusions. The fact that the results are marginally better, is a great sign that the area is fruitful for research, and that appropriate amendments to the proposed methods which can capture the individual nature of the genetic data can give significantly better results.

Firstly, following the observation that the metric PWF_{ST} picks by far the first 20 most informative SNPs, a possible amendment could be to modify the aggregation algorithm in order to give greater weight to the methods had the largest increase in accuracy between subsets of SNP. For example, a higher weight could be given to the 20 first SNPs selected based on the PWF_{ST} because the classification accuracy is at 84%. In the next 20 (20-40) the increase in accuracy is not so great in PWF_{ST}, compared with the Delta and In. Therefore in the 20-40 SNP interval greater weight could be given in the selected SNP from both methods and not just by PWF_{ST}.

As mentioned before, the rating of each SNP is the ranking order within the list, ignoring the information of the relative position of the SNP in the list. In the future study emphasis should be given on the relative position of the SNP within each list through the scores obtained from each metric (eg F_{ST} score, Delta Score). Such an approach would give more information about the assignment accuracy of each SNP aiming to achieve better results after method combination.

Acknowledgement

Ioannis Kavakiotis gratefully acknowledges financial support from the Hellenic Artificial Intelligence Society (EETN) for attending this conference.

8. REFERENCES

[1] Marx, V. Biology: the big challenges of big data. Nature 498, 255–260 (2013).

[2] Peralta, D., del Río, S., Ramírez-Gallego, S., Triguero, I., Benitez, J.M, Herrera F. Evolutionary Feature Selection for Big Data Classification: A MapReduce Approach. *Mathematical Problems in Engineering* Volume 2015 (2015), Article ID 246139, 11 pages

[3] Mattmann CA. 2013. Computing: A vision for data science. *Nature.* 2013 Jan 24;493(7433):473-5. doi: 10.1038/493473a.

[4] Wilkinson, S., Wiener, P., Archibald, A.L., et al. (2011) Evaluation of approaches for identifying population informative markers from high density SNP chips. *BMC Genetics*, 12, 45.

[5] Kavakiotis, I., Triantafyllidis, A., Ntelidou, D., Alexandri P, Megens. H.J., Crooijmans R.P., Groenen M.A., Tsoumakas G., Vlahavas I. (2015) "TRES: Identification of Discriminatory and Informative SNPs from Population Genomic Data.",*J Hered.* 2015 Sep-Oct; 106(5):672-6.

[6] Tan, M., Tsang, I. W., and L. Wang, “Towards ultrahigh dimensional feature selection for big data,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1371–1429, 2014

[7] Manel, S., Gaggiotti,OE.,Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology and Evolution*, 20, 136–142

[8] Rosenberg, NA., Li, L.M., Ward, R., Pritchard, J.K. (2003) Informativeness of genetic markers for inference of ancestry. American Journal of Human Genetics, 73, 1402-1422.

[9] Ding, L., Wiener, H., Abebe, T., et al. (2011) Comparison of measures of marker informativeness for ancestry and admixture mapping *BMC Genomics*, 12, 622.

[10] Wright S (1951) Thegenetical structure of populations. *Annals Eugenics*, 15, 323

[11] Shriver, M.D, Smith, M.W., Jin L et al. (1997) Ethnic affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics*, 60, 957-964.

[12] Cynthia Dwork, C., Kumar, R.,Naor, M.,Sivakumar D., Rank aggregation methods for the web. In Proceedings of the Tenth International World Wide Web Conference, 2001

[13] Lin S. Rank aggregation methods. Wiley interdisciplinary Reviews: Computational Statistics, 2:555- 570, 2010

[14] Borda P. Memoire sur les elections au scrutin, Histoire de l’Academie Royale des Sciences, 1781.

[15] Wilkinson, S., Archibald, A.L., Haley, C.S., et al. (2012) Development of a genetic tool for product regulation in the diverse British pig breed market. *BMC Genomics*, 13, 580.

[16] Ramos. A.M., Crooijmans, R., Affara N.A., et al. 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by Next Generation Sequencing technology. PLoS One. 4: 8.

[17] Piry, S., Alapetite, A., Cornuet, J.M., Petkau, D., Baudouin, L., Estoup A. 2004. GENECLASS2: A software for genetic assignment and first generation migrant detection. *Journal of Heredity*, 95: 536-539.

[18] Rannala B., Mountain J. L.1997 Detecting immigration by using multilocusgenotypes.*Proc. Natl. Acad. Sci.* USA 94: 9197–9201