35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Comparison of Aggregators for Multi-Objective SNP Selection

Zeliha Gormez, Ergun Gumus, Ahmet Sertbas and Olcay Kursun

Abstract— **SNPs (Single Nucleotide Polymorphisms) are genomic variants that associate with many genetic characteristics. These variants can also be utilized to track the on-going mutation in population genetics. The goal of this study was to select the most relevant SNP subsets for discriminating ethnic groups. Each SNP was evaluated by its: i) Mutual information, ii) Relief-F score, iii) Loadings of the first principal component, iv) Loadings of the second principal component. Combining these four feature ranking criteria in different ways, three different aggregation methods (Pareto Optimal, Condorcet, MC4) were compared with respect to their SNP selection accuracies. Results showed that SNP subsets chosen with Pareto Optimal yielded better classification accuracy.**

I. Introduction

Variations in genome, like Single Nucleotide Polymorphisms (SNPs), cause genetic diversity forming the genotype. Sometimes these variations cause undesirable results like genetic diseases passing along generations. On the other hand, they also determine the phenotype of a person and his/her ancestry. Among hundreds of thousands SNPs, choosing the most relevant ones serving to the task of phenotype forecasting is a serious challenge known as "feature selection" in machine learning terminology.

While selecting qualified features in high dimensional datasets, classical brute-force techniques or wrapper methods (i.e. forward/backward feature selection) suffers from high computation cost and time complexity. Contrarily, supervised filtering methods like Mutual Information [1] or Relief-F [2] perform well in such cases. Although, they assess each feature individually and might overlook feature groups coping well, they are commonly referred because of their low computation cost in high dimensional datasets. By these methods, finding the most valuable feature subset for high classification accuracy, which will be referred as "Objective 1", becomes possible.

Another way to weigh each feature is the well-known Principal Component Analysis (PCA) method. PCA not only helps dimension reduction by eliminating projection directions with low variance but also gives an opportunity to visualize the data processed. Weighing each feature using PCA loadings does not always yield to best discrimination especially in multi-class datasets. However, projections using these loadings can present other hidden characteristics of the data. Novembre et al. [3], showed that top-two principal components can be used to form a projection showing the genetic layout of nations all around Europe. This layout, which was formed by using the genomic data of individuals (SNPs), was highly correlated with the geographical layout of nations. This correlation, which will be denoted as "geo-genomic correlation", is a known fact arising from the migration paths of nations [4]. Geographically close groups are expected to have more common genetic variations than far groups. By using PCA, finding the most valuable feature (SNP) subsets for providing high geo-genomic correlation, which will be referred as "Objective 2", becomes possible.

Note that, maximizing either objective does not guarantee maximization of the other. SNP subsets that give high classification accuracy, may fail reflecting the geographical distances between groups. On the other hand, as the simple geo-genomic correlation utilizes average genomic distances between individuals across groups and ignoring the genomic variations within groups, classification accuracy may not be as good. The main aim of this study, similar to that of [11], is to determine feature (SNP) subsets of a well-known dataset formed after "Human Genome Diversity Project" (HGDP). Selected features should provide not only high classification accuracy (Objective 1) but also high geo-genomic correlation (Objective 2).

Mutual information and Relief-F scores of each SNP were calculated to favor Objective 1. Also feature loadings from top-two principal components (PC1&PC2) were used to favor Objective 2. This yielded a four criteria selection problem. Choosing highly ranked top-D SNPs for both objectives brings along the necessity of using aggregation methods working on four criteria. Three aggregators; i) Pareto Optimal, ii) Condorcet ranking, iii) MC4 ranking, were used for this task and their results were evaluated.

II. Material

A. Dataset

Human Genome Diversity Project (HGDP) aimed to identify the evolution based variance among people from different nations and track the on-going mutation in population genetics. Project was administrated by Luigi Luca Cavalli-Sforza from Stanford University and got contribution from many researchers/donators all over the world [5].

Raw data [6] contained 660918 SNPs of 1043 individuals from 52 ethnic groups. First 163 SNPs were from mitochondrial DNA and got excluded from the study.

B. Pre-processing

Each SNP possessed three combinations of two nucleotides. These combinations were digitized as 1 for homozygous pair of major allele, 0 for heterozygous pair, -1 for homozygous pair of minor allele.

After this step, SNPs were gathered in different groups according to their chromosomal origin. Then, they were

The authors are with the Computer Engineering Department of Istanbul University, Istanbul, TURKEY (e-mails: {zgormez, egumus, okursun, asertbas}@istanbul.edu.tr).

Z. Gormez is also with the Scientific and Technological Research Council of Turkey (TUBITAK), Kocaeli, TURKEY (e-mail: zeliha.gormez@tubitak.gov.tr).

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

reordered according to their nucleotide position in the corresponding chromosome.

Last step was to eliminate SNPs with minor allele frequency (MAF) values smaller than 5%. Such SNPs were considered to have no discriminative power.

III. Methods

Brief information about some of applied machine learning methods and aggregators is given in this section.

A. Mutual Information

Mutual information is a metric for assessing the dependency of two random variables *X* and *Y* [1]. The mutual information $I(X;Y)$ is also known as the relative entropy between joint probability function and product of marginal probabilities of these distributions.

B. Relief-F

Relief-F [2] gives more weight to features that discriminate neighboring instances of different classes. It estimates the following probability to assign as the weight, *w*, for each feature *f*:

$$w(f) = \text{P(different value of } f \mid \text{different class)} - \text{P(different value of } f \mid \text{same class)}$$

C. Pareto Optimal (PO)

In many types of multi criteria decision problems, there is not a unique optimum solution. Contrarily, for at least one criterion, there exist alternative solutions with much more optimum values than other solutions. These alternative solutions form "Pareto Optimal Solution Set" [7, 11].

Multi objective decision problem is defined as a maximization/minimization task of function *f*. Function *f* involves m decision variables and n objectives. Depending on this, optimization problem can be defined as follows:

$$\text{Maximize/minimize} \quad y = (f_1(x), f_2(x), ..., f_m(x))$$

$$\text{Subject to} \quad \begin{aligned} x &= (x_1, x_2, ..., x_m,) \in X \\ y &= (y_1, y_2, ..., y_n,) \in Y \end{aligned}$$

Here, *x* and *y* represent decision vector and objective vector, respectively. *X* is called as "decision space" and *Y* is known as "objective space". The solution set contains the decision vectors which have not been dominated by any other decision vector. Each decision vector in Pareto Optimal Set has an optimal value for at least one objective. These non-dominated solutions are known as "Pareto Optimal Solutions".

D. Condorcet Ranking

In Condorcet ranking [8], voters are asked to make a ranking list of candidates (Put the most preferred one to head of the list and last preferred one to the end of the list). Then, each candidate's score is calculated according to its comparison to each of other candidates one at a time. Calculated scores are sorted in descending order to find final rank of each candidate. In our scenario, candidates are SNPs and voters are criteria which are Mutual Information value, Relief-F score, PC1 and PC2 loadings.

E. MC4 Ranking

The MC4 is a Markov Chain (MC) based ranking algorithm. It can be defined briefly as follows [9]:

i) Construct the set *U* that consists of all items that appear within the top-k in at least one list.

ii) For each pair of items *i* and *j* in *U*, let the preference for *j* over *i*, m_{ij}^*, equal to 1 if the majority (≥ 50%) of lists that rank both *i* and *j* rank *j* above *i* and 0 otherwise. Let $m_{ij}^* = m_{ji}^* = 0.5$ if items *i* and *j* are never directly compared in any list.

iii) Define the transition matrix $M = \{ m_{ij} \}$ as follows: for $i \neq j$ set m_{ij} to $m_{ij}^* / |U|$ and let $m_{ii} = 1 - \Sigma_{j \neq i} m_{ij}$.

iv) Make the transition matrix *M* ergodic by multiplying each element by $1 - \varepsilon$ and then adding $\varepsilon / |U|$ to each element, where ε is a small, positive number. In practice, ε is chosen as 0.15.

IV. Experimental results

264 individuals from 12 ethnic groups were chosen as experiment subjects. List of these groups can be seen in Table 1.

TABLE 1. List of selected groups

Continent	Ethnic Group		Coordinates	Count
Africa	G1	Mozabite	32N, 3E	30
	G2	Biaka_Pygmies	4N, 17E	11
	G3	Yoruba	6-10N, 2-8E	24
	G4	Mandenka	12N, 12W	24
Asia	G5	Cambodia	12N, 105E	11
	G6	Japanese	38N, 138E	29
	G7	Balochi	30-31N, 66-67E	25
	G8	Yakut	62-64N,129-130E	25
Europe	G9	Adygei	44N, 39E	17
	G10	Orcadian	59N, 3W	16
	G11	French_Basque	43N, 0	24
	G12	Sardinian	40N, 9E	28

Main consideration in this choice was to find groups as close as possible so that both objectives could be maximized with small noise. Distribution of these groups over the world can be seen in Fig. 1 [10].

First half of each group was added to training set and the rest formed the test set. Mutual Information score, Relief-F score, PC1 and PC2 loadings (loadings from top-two principal components) of each SNP was calculated according to the training set. Absolute values of PC loadings were used because an SNP with a loading that has high absolute value can effect variance in either positive or negative way. But an SNP with a loading close to zero has no effect on total variance and so on classification.

After that, these SNPs were ranked according to three aggregation methods individually. Classification accuracy (Objective 1) and geo-genomic correlation (Objective 2) calculation were performed on test set samples according to top-D SNPs chosen by aggregators. As classifier, a K-nearest neighbor classifier with K = 5 was used. In order to calculate geo-genomic correlation, mean genomic distances (Euclidean distance) and geographical distances of each group pair were calculated. Accuracy and correlation values calculated over SNPs chosen by three aggregators can be seen in Table 2.

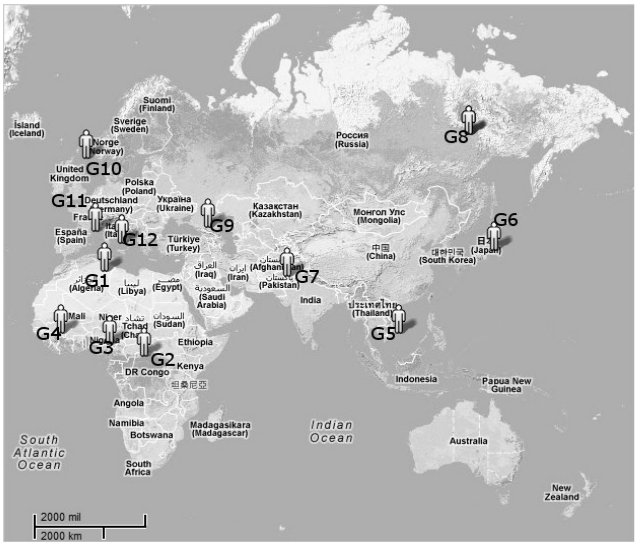

Fig. 1. Geographical locations of selected groups

TABLE 2. ACCURACY / CORRELATION VALUES (%) FOR SNPS CHOSEN BY THREE AGGREGATORS

Chr. ID	PO	Condorcet	MC4	# of SNPs
1	**48 / 49**	18 / 43	40 / 19	76
2	**40 / 56**	21 / 23	25 / 44	87
3	**52 / 50**	21 / 5	27 / 45	103
4	**44 / 54**	19 / 22	39 / 48	66
5	**50 / 54**	16 / 35	23 / 40	82
6	**43 / 66**	21 / 5	31 / 61	53
7	**59 / 49**	26 / 21	47 / **49**	106
8	**47 / 55**	22 / 36	31 / 35	56
9	**49 / 46**	20 / 26	36 / 23	81
10	**52** / 63	21 / 0	44 / **73**	76
11	**41** / 52	19 / 19	34 / **58**	81
12	**53** / 49	19 / 28	36 / **55**	67
13	**36 / 41**	22 / 37	**36** / 19	46
14	**37** / 46	19 / 27	28 / **47**	54
15	**38 / 55**	19 / 29	17 / 19	42
16	**44 / 54**	21 / 34	27 / 36	45
17	**44 / 51**	13 / 21	12 / 9	36
18	**39 / 52**	13 / 32	21 / 15	51
19	**41 / 56**	12 / 38	25 / 12	35
20	**49 / 48**	17 / 27	20 / 27	57
21	**36 / 49**	24 / 10	26 / 16	42
22	**42** / 61	19 / 10	39 / **67**	44
X	**41 / 51**	10 / 16	28 / 29	79
Mean	45 / 52	19 / 24	30 / 37	64

a. Highest values are marked bold and lowest values are marked underlined

As seen in Table 2, considering mean values, SNPs chosen by PO serves to both objectives better than SNPs chosen by other aggregation methods. Accuracy values of SNPs chosen by PO are always highest and correlation values are never lowest. Here one should note that, number of SNPs chosen from each chromosome was determined by Pareto Optimal (SNPs in first Pareto Layer) and same amount of top-D SNPs from other aggregators were chosen for comparison.

In order to increase classification accuracy and geo-genomic correlation more SNPs can be added to feature set. This can be achieved by using sub-layers of Pareto Optimal approach. Table 3 presents mean accuracy and correlation values (for all chromosomes) obtained by three aggregators. Here, first 15 layers of Pareto Optimal were used. Also, same amounts of top-D SNPs from Condorcet and MC4 ranking chosen.

TABLE 3. MEAN ACCURACY / CORRELATION VALUES (%) FOR INCREASING NUMBERS OF SNPS

Layer ID	PO	Condorcet	MC4	Mean # of SNPs
1	45 / 52	19 / 24	30 / 37	64
2	52 / 54	27 / 32	39 / 45	176
3	58 / 54	35 / 39	47 / 49	346
4	61 / 54	41 / 44	54 / 54	565
5	63 / 54	46 / 48	59 / 56	833
6	66 / 54	51 / 51	62 / 57	1157
7	69 / 54	56 / 52	66 / 57	1549
8	72 / 54	61 / 54	67 / 58	1995
9	73 / 54	63 / 55	70 / 58	2476
10	74 / 55	66 / 56	72 / 58	3014
11	76 / 55	69 / 57	73 / 58	3598
12	78 / 55	72 / 58	74 / 58	4238
13	79 / 55	73 / 58	75 / 58	4923
14	80 / 55	75 / 58	76 / 59	5634
15	80 / 55	77 / 58	78 / 59	6382

As seen in Table 3, as number of layers increase, mean accuracy for classification increases. However, this increase is limited for mean correlation values. Among all chromosomes, chromosome 11 was found to possess the most valuable SNPs considering mean accuracy and geo-genomic correlation values for selected 12 ethnic groups. Fig. 2 and 3 show progress of classification accuracy and geo-genomic correlation for increasing number of layers, respectively.

V. CONCLUSION

Genome wide association studies aim to find out the relations between bio-markers (i.e. SNPs) and phenotypes. Thus, SNP selection may involve choosing a small subset of SNPs that can relate with multiple objectives like disease associations, ethnicity grouping, geo-genomic correlation, migratory routes, etc. Given several distinct SNP rankings for such different objectives, aggregation methods such as Pareto Optimal, Condorcet, MC4 can be used to produce a final relevant set of SNPs. In this study, while assessing the utility of selected SNP subsets, two criteria: classification accuracy of ethnic groups and geo-genomic correlations. These criteria were measured by Mutual information and Relief-F scores (serving the first criterion) and principal component loadings (serving the second criterion). The results have shown that Pareto Optimal yielded better classification accuracy and geo-genomic correlation than the other two aggregators. Evaluating each chromosome individually, it has been found that chromosome 11 had the best performance. Our further studies have shown that better metrics can be developed for geo-genomic correspondence.

ACKNOWLEDGMENTS

This work was supported by Scientific Research Projects Coordination Unit of Istanbul University under the grant YADOP-27343.

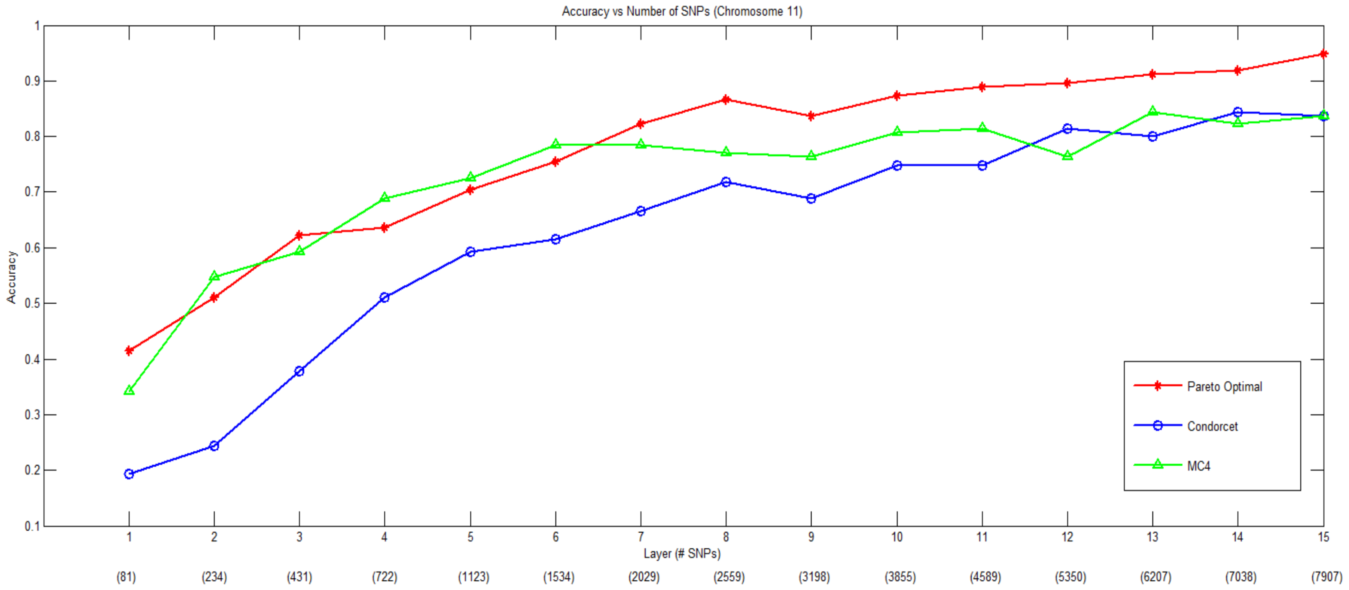

Fig. 2. Progress of classification accuracies for chromosome 11 according to increasing number of layers

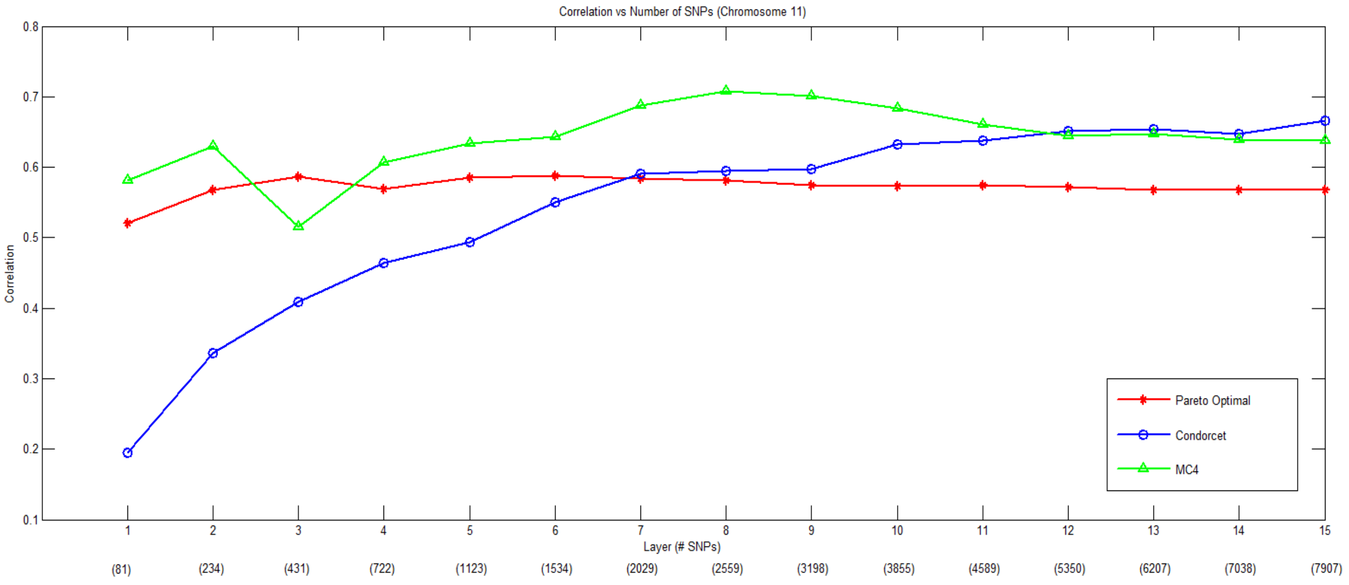

Fig. 3. Progress of geo-genomic correlation for chromosome 11 according to increasing number of layers

REFERENCES

[1] T.M. Cover, J.A. Thomas, “Elements of information theory”, John Wiley & Sons, New York, 1991.

[2] M. Robnik-Sikonja, I. Kononenko, “Theoretical and empirical analysis of ReliefF and RReliefF”. Machine Learning , vol. 53, pp. 23-69, 2003.

[3] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergmann, M.R. Nelson, M. Stephens, C.D. Bustamante, “Genes Mirror Geography within Europe”, Nature (Letter), vol. 456, pp. 98-101, 2008.

[4] R.A. Kittles, K.M. Weiss, “Race, ancestry, and genes: implications for defining disease risk”, Annual Review of Genomics and Human Genetics, vol. 4, pp. 33-67, 2003.

[5] L.L. Sforza, “The Human Genome Diversity Project: Past, Present and Future” Nature Reviews-Genetics, vol. 6, pp. 333-340, 2005.

[6] Human Genome Diversity Project, URL: http://hagsc.org/hgdp/files.html, online.

[7] E. Zitzler, L. Thiele, “Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach”, IEEE Transactions on Evolutionary Computation,vol. 3(4), pp. 257-271, 1999.

[8] R.C. Prati, “Combining feature ranking algorithms through rank aggregation”, International Joint Conference on Neural Networks (IJCNN), pp. 1-8, 2012.

[9] R. P. DeConde, S. Hawley, S. Falcon, N. Clegg, B. Knudsen, R. Etzioni, “Combining results of microarray experiments: A rank aggregation approach,” Stat. Appl. Genet. Mol. Biol., vol. 5(1), 2006.

[10] My Interactive Map, URL: http://www.pinmaps.net/, online.

[11] E. Gumus, Z. Gormez, O. Kursun, “Multi objective SNP selection using pareto optimality”, Computational Biology and Chemistry, vol:43, pp: 23–28, 2013.