The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

Shrunken Dissimilarity Measure for Genome-wide SNP Data Classification*

Haiyong Liao[1,†] Yang Liu[1,‡] Michael K. Ng[1,§]

[1]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

Abstract Recent development of high-resolution single-nucleotide polymorphism (SNP) arrays allows detailed assessment of genome-wide human genome variations. However, SNP data typically has a large number of SNPs (e.g., 400 thousand SNPs in genome-wide Parkinson disease SNP data) and a few hundred of samples. Conventional classification methods may not be effective when applied to such genome-wide SNP data. In this paper, we propose to develop and use shrunken dissimilarity measure to analyze and select relevant SNPs for classification problems. Examples for HapMap data and Parkinson data are given to demonstrate the effectiveness of the proposed method and illustrate it has the potential to become a useful analysis tool for SNP data sets. In particular, we find some SNPs in chromosome 2 that they contain in some genes which is relevant to Parkinson disease.

Keywords Shrinkage; Dissimilarity measure; Categorical centroids; Single nucleotide polymorphism; Genome-wide; Classification

1 Introduction

Single Nucleotide Polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide - A, C, G, or T - differs at the same position between individuals [1]. SNPs are believed to result in differences between individuals, such as susceptibility to diseases [2]. They are abundant in human genome [3, 4], which are considered as invaluable markers and potential powerful tools for both of genetic researches and applications in practice [5, 6]. For example, disease gene discovery [7], drug development [8], clinical treatment [9], etc. It is believed that more and more genetic researches and practical applications combined with machine learning or statistical or data mining methods will be investigated based on SNP data as SNPs will provide more useful information which is not shown by other methods. In SNP data, the association between a disease and a set of relevant SNPs are investigated. Patients and normals are often categorized in groups according to their SNP genotypes (categorical values). Thousands of SNPs in different regions of chromosomes are used to describe characteristics of patient/normal samples. There are two key properties of data sets for such classification task: high-dimensional and categorical.

*Research supported in part by RGC 201508 and HKBU FRGs.
†Email: 06459358@hkbu.edu.hk
‡Email: 08466246@hkbu.edu.hk
§Email: mng@math.hkbu.edu.hk

When many SNPs are used to detect the association between a disease and multiple marker genotypes, we expect in a typical data set that contains the genotype data of several thousands of SNPs in different individuals. It is common to find only several numbers of SNPs having genotype patterns that are highly specific to each group of individuals. The SNPs are called the relevant SNPs, as opposed to the irrelevant SNPs that do not help much in identifying the group (i.e., individuals of the same type). Due to the large number of SNPs being irrelevant to each group, two individuals in the same group could have low similarity when measured by a simple similarity function that consider the genotypes of all SNPs. The groups may thus be undetectable by classification algorithms. The classification problem is defined for such a scenario, see for instance [10]. Each group is a set of individuals with an associated set of relevant SNPs such that in the group formed by the relevant SNPs, the individuals are similar to each other but dissimilar to individuals outside the group. In this paper, we are interested in the development of high-dimensional categorical classification algorithm that can identify group of individuals and their relevant SNPs, i.e., detect association between a disease and multiple marker genotypes.

The outline of this paper is given as follows. In Section 2, we propose to develop and use shrunken dissimilarity measure to analyze SNP data classification. In Section 3, we present experimental results on two real SNP data sets: HapMap data and Parkinson data. We give concluding remarks in Section 4.

2 Shrunken Dissimilarity Measure

The nearest shrinkage centroid [11] has been developed to handle numerical microarray data sets. The main difference between gene expression and SNP data is that the expression values are continuous and SNPs are categorical [12]. In the literature, Park et al. [13] selected SNPs using the nearest shrunken centroid method. Their method is to represent genotypes by numerical numbers directly. In [14], Schwender has developed SAM for analysis of SNP data. Their method is to study contingency table for testing if the distribution of the genotypes of SNPs differs between different groups. The Pearson χ^2 statistic is used to handle rejection hypothesis. Shrunken χ^2 statistics are further constructed to analyze relevant SNPs. In this paper, we also make use of the shrinkage idea and extend the algorithm for categorical SNP data by using a genotype distribution measuring for categorical objects and modes instead of means for groups. These extensions will remove the numeric-only limitation of the nearest shrunken method and enable the classification process to be used to efficiently deal with genome-wide categorical SNP data sets.

Let x_{ij} be the categorical value for SNP $i = 1,2,\ldots,p$ and samples $j = 1,2,\ldots,n$. There are K classes and let C_k be indices of the n_k samples in class k. The centroid of the ith SNP in class k is defined as: $\bar{x}_{ik} = \text{mode}(i\text{th SNP in class } k)$, and the overall centroid for SNP i is: $\bar{x}_i = \text{mode}(i\text{th SNP in all classes})$. Let

$$d_{ik} = \frac{\text{dist}(\bar{x}_{ik}, \bar{x}_i)}{m_k(s_i + s_c)} \tag{1}$$

where s_i is the pooled within-class standard deviation for SNP i:

$$s_i^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{j \in C_k} \text{dist}(x_{ij}, \bar{x}_{ik})^2 \tag{2}$$

In (1), we consider the distance from class centroid to overall centroid for the ith SNP is given by

$$\text{dist}(\bar{x}_{ik}, \bar{x}_i) = \text{norm}(\bar{\mathbf{v}}_{ik} - \bar{\mathbf{v}}_i) \tag{3}$$

where $\bar{\mathbf{v}}_{ik}$ is the genotype distribution vector associated with ith SNP centroid in class k, and $\bar{\mathbf{v}}_i$ is the genotype distribution vector associated with ith SNP overall centroid. The soft thresholding d'_{ik} can be defined similarly by:

$$d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+. \tag{4}$$

In (1),we can see that if the difference between class centroid and overall centroid is small or large, it demonstrates that the difference is insignificant or is just some noise. Therefore, we shrinkage the value toward zero in (4) and the corresponding SNP does not contribute to the classification task. At the shrinkage step, the categorical centroids are not shrunken (compared with the procedure for the numerical microarray data). Here our task is to drop the categorical centroids whose d_{ik} is less than the threshold. Let $\mathbf{t}$ be a test sample, the class label of $\mathbf{t}$ is determined by:

$$C(\mathbf{t}) = \arg\min_k(\delta_k(\mathbf{t})), \quad \delta_k(\mathbf{t}) = \sum_{i \in \{i||d_{ik}|>\Delta\}} \frac{\text{dist}(t_i, \bar{x}_{ik})^2}{(s_i + s_0)^2} - 2\log(\pi_k) \tag{5}$$

where π_k is the prior probability of class k. It is the proportion of class k in the population. If it is unknown, it can be set to $\frac{1}{K}$.

3 Experimental Results

3.1 HapMap Data

We test the nearest shrunken categorical centroids method on HapMap SNP data [15]. Data are downloaded from the HapMap[1]. According to the LD map of chromosome 22 (see [16]), 200 SNPs from chromosome 22 of 4 populations CEU, CHB, JPT and YRI are picked out randomly from a region from 3.44e7-3.5e7 kb in Figure (1), which shows a great difference of SNP positions on the LD map over 4 populations. Here the LD map shows the intensity of linkage disequilibrium of SNPs. In the map, the "flat" curve means that the SNPs are in strong linkage disequilibrium, i.e., the recombination rarely occur between them, while the "steep" curve means the recombination occurs frequently in this part of chromosome. Missing data are considered as a category in the calculation.

In the first experiment, we take any two out of four populations to set up two-class classification problems. Cross-validation is used to employed The results are shown in Figures 2, 3 and 4. As shown in the figures, we can see that all have a high accuracy of more than 90 percent, except the CHB-JPT classification problem, only about 50 percent, when the threshold Δ is less than 2. Then accuracy decreases as the amount of shrinkage increases since less SNPs are used in the prediction. The reason for the poor accuracy of CHB-JPT classification is that these two populations are quite similar on their SNPs, see Figure 6.

[1] HapMap Website `http://www.hapmap.org/cgi-perl/gbrowse/hapmap_B35/?name=Chr22`

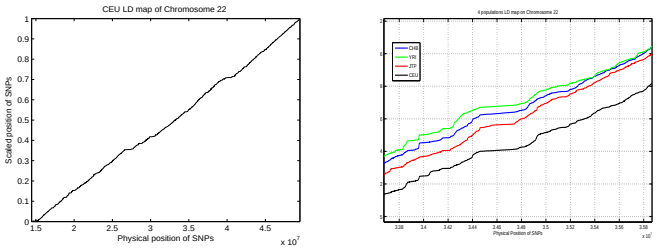

Figure 1: (left) the LD map of the whole chromosome and (right) the LD map of 4 populations on the part chromosome 22.

In the second experiment, we consider a four-class classification problem, i.e., to classify the four populations: CEU, CHB, JPT and YRI. The setting is the same as that in the first experiment. Figure 5 (left) shows the cross-validation classification accuracy using different values of Δ for 200 SNPs. The best accuracy is 77.78 percent when $\Delta = 1.5$. When $\Delta < 1.5$, there are a lot of SNPs to be used in the classification, but some of them are likely redundant. When $\Delta > 1.5$, a lot of SNPs are not used, we may throw away some useful SNPs in the classification process. The confusion matrix in Table 1 shows that the prediction for CEU and YRI is quite good, but bad for CHB and JPT. In these two cases, the accuracy is not high. When we use all 51793 SNPs in chromosome 22 to perform the classification, the best accuracy is 94.44 percent ($\Delta = 0.5$), see Figure 5 (right).

Table 1. Confusion matrix when $\Delta = 1.5$.

	CEU	YRI	CHB	JPT
CEU	43	0	1	1
YRI	0	45	0	0
CHB	0	0	30	15
JPT	0	0	23	22

By shrinkage (Δ is set to 1.5), the number of SNPs used for classification is decreased from 200 to 143, 143, 142 and 142 for CEU, YRI, CHB, and JPT respectively. In figure (6), we show the SNPs used in prediction and their value of d'_{ik}. The values of d'_{ik} in blue in the figure mean that its corresponding SNP appears in all four populations, while the values of d'_{ik} in red represents its corresponding SNP shows in only one population. Next we show the centroid genotype distribution vector corresponding to the d'_{ik} in red in the following two tables.

Table 2. genotype distribution vector of 12th SNP (left) and 127th SNP (right)

	aa	aA	AA
CEU	0	0.0667	0.9333
YRI	0.0667	0.5111	0.4222
CHB	0	0.0444	0.9556
JPT	0	0.0222	0.9778

	aa	aA	AA
CEU	0.1556	0.4	0.3778
YRI	0.0222	0.1333	0.8444
CHB	0	0	1
JPT	0	0	1

As shown in the above two tables, at 12th SNP, the genotype distribution vector of YRI is quite different from the others, similarly, at 127th SNP, the genotype distribution

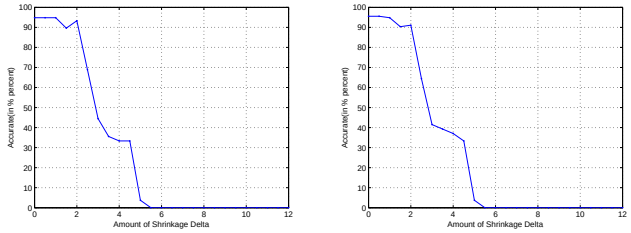

Figure 2: CEU-CHB classification (left) and CEU-JPT classification (right)

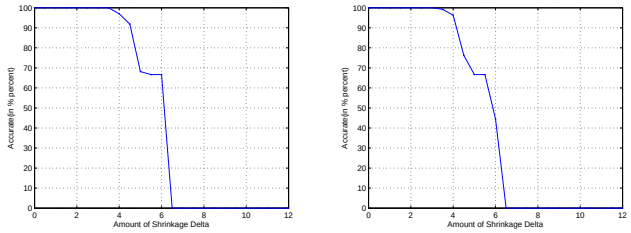

Figure 3: YRI-CHB classification (left) and YRI-JPT classification (right)

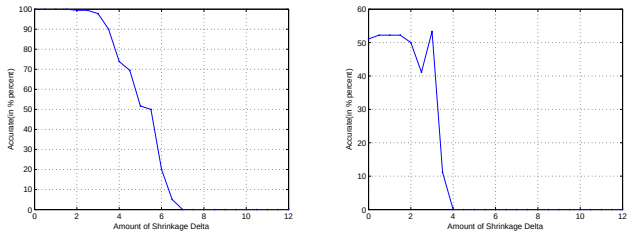

Figure 4: CEU-YRI classification (left) and CHB-JPT classification (right)

vector of CEU differs from those of the other three populations. The reason is that the mode of YRI "aA", while that of whole population is "AA", and therefore YRI population has more variation and has a large value of d'_{ik}.

3.2 Parkinson Disease SNP Data

We test the Parkinson disease genome-wide SNPs data set downloaded from the Coriell Institute for Medical Research. The genotyping was performed using the Illumina Infinium I and Infinium II assays. The Illumina Infinium I assay asseses 109,365 unique gene-centric SNPs while the Infinium II assay assesses 317,511 haplotype taggings SNPs based upon Phase I of the International HapMap Project. The Illumina Infinium I and II assays share 18,073 SNPs in common, so in combination the two assays represent 408,803 unique SNPs. The genotype data posted consists of these 408,803 SNPs for 270 individuals with idiopathic Parkinson Disease (case) and 271 neurologically normal control individuals (control).

Table 3 shows the average classification accuracy results (correctly classified sam-

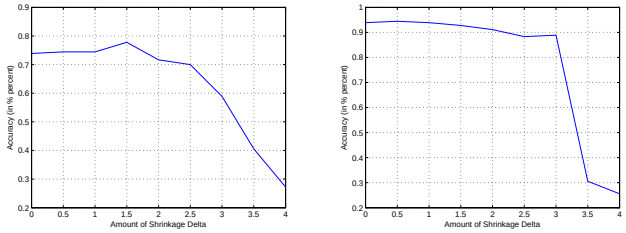

Figure 5: Classification accuracy for four classes problem using the part of chromosome 22 (left) and all 51793 SNPs in Chromosome 22 (right).

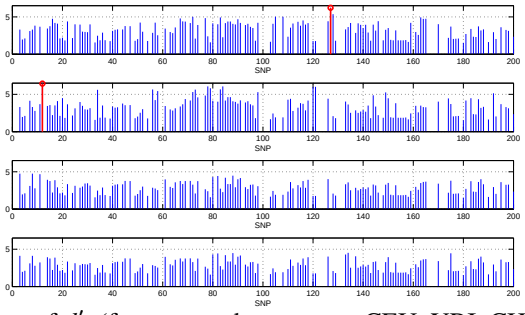

Figure 6: The values of d'_{ik} (from top to bottom are: CEU, YRI, CHB, JPT, $\Delta = 1.5$)

ples) for 22 chromosomes by using the nearest shrunken centroid program after 10-fold cross validation. We use the most frequent genotypes in case and control groups to be the modes for the program. The parameter Δ is tuned in each chromosome to obtain the highest accuracy in the test. Missing category is also considered in this experiment. Although the cause of Parkinson disease is still unknown to us, some of the genetic factors have been discovered. We know that there are many monogenes cloned or mapped on different chromosomes. The first gene to be isolated was *PARK*1 located in chromosome 4, two additional loci *PARK*3 and *PARK*4, on chromosome 2 and chromosome 4 respectively have been discovered in 1998 and 1999. Furthermore, four loci on chromosome 1, *PARK*6, *PARK*7, *PARK*9 and *PARK*10 have been reported to contain susceptibility genes, see [17] for details. We also find that the above chromosomes which have been reported to be associated with Parkinson disease also have relatively high accuracy in our method.

We choose chromosome 2 (with the highest classification accuracy) as an example to demonstrate the SNPs selected by the proposed method. Fig. 7 shows the accuracies obtained when we increase Δ value from zero to four. We can see from the figure that our method can get a reasonably good accuracy of 90.91% when Δ is equal to 1.0. By shrinkage, the number of SNPs selected for the classification is decreased from 32706 to 28. We also choose Chromosome 2 as an example to demonstrate the effectiveness of our method compared with Park's [13], see Table 4. We use the numerical values (0,1,2,3) to represent different genotypes. The results show that the proposed method is more effective.

Table 3: Classification accuracy results.

Chromosome	No. of SNPs	Accuracy	Δ	Chromosome	No. of SNPs	Accuracy	Δ
1	31532	0.8364	1.6	12	19572	0.8727	1.1
2	32706	0.9091	1.0	13	14123	0.8000	1.0
3	27691	0.8909	0.9	14	12645	0.8364	1.1
4	24193	0.8545	1.0	15	11618	0.8545	1.2
5	24570	0.7273	1.2	16	11767	0.7636	1.0
6	26372	0.8364	1.1	17	11619	0.7273	1.1
7	21382	0.8545	0.8	18	12613	0.8364	0.9
8	22434	0.8571	1.0	19	8608	0.7455	1.0
9	19542	0.8545	1.0	20	10375	0.6364	1.0
10	20007	0.8545	1.3	21	6612	0.8182	0.9
11	19539	0.8000	1.6	22	7071	0.6364	0.7

Table 4: Comparison between our own method and Park's method.

Own			Park		
Accuracy	Δ	No of SNPs	Accuracy	Δ	No of SNPs
0.9091	1.0	28	0.8364	0.9	221

4 Concluding Remarks

The main contribution of this paper is to develop a shrunken dissimilarity measure to handle SNP data classification problems. The method can be implemented on a PC very efficiently. The relevant SNPs are selected for HapMap data sets and Parkinson disease data sets. Experimental results are also reported to show the effectiveness of the method. In particular, we find some SNPs in chromosome 2 that they contain in some genes which is relevant to Parkinson disease. In the future, we study the following problems. (i) Our aim is to develop statistical analysis of the proposed shrunken dissimilarity measure so that a detailed statistical study of selected SNPs can be carried out. (ii) Detailed biological analysis of SNPs of other genome-wide SNP data sets will be studied. The genomic variation of data sets can take account of functional as well as linkage disequilibrium information. More importance is attached to some SNPs than others, based on their positions within the coding or regulatory regions or splice sites.

References

[1] Brookes A.J.: The essence of SNPs, *Gene*, **234**, 177-186, 1999.

[2] Risch N., Merikangas K.: The future of genetic studies of complex human diseases, *Science*, **273**, 1516-1517, 1996.

[3] Cargill M., Altshuler D., Ireland J. Sklar P., et al: Characterization of single nucleotide polymorphisms in coding regions of human genes, *Nature Genet.*, **22**, 231-238, 1999.

[4] The International SNP Map Working Group: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, *Nature*, **409**, 928-933, 2001.

[5] Hirschhorn J.N., Daly M.J.: Genome-wide association studies for common diseases and complex traits, *Nat. Rev. Genet.* **6**, 95-108, 2005.

[6] Syvanen A. C.: Toward genome-wide SNP genotyping, *Nature Genetics*, **37**, S5-S10, 2005.

[7] Ozaki K., Ohnishi Y., Iida A., et al: Functional SNPs in the lymphotoxin- gene that are associated with susceptibility to myocardial infarction, *Nature Genetics*, **32**, 650-654, 2002.

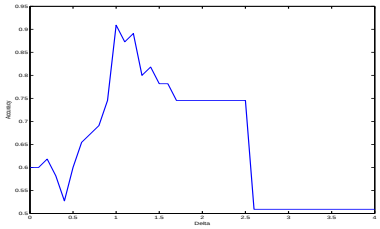

Figure 7: Relationship between Δ and accuracy in Chromosome 2

[8] Rothberg B.E.G.: Mapping a role for SNPs in drug development, *Nature Biotechnology*, **19**, 209-211, 2001.

[9] Erichsen H.C., Chanock S.J.: SNPs in cancer research and treatment, *British Journal of Cancer*, **90**, 747-751, 2004.

[10] Khlestkina E.K., Salina E.A.: SNP markers: Methods of analysis, ways of development, and comparison on an example of common wheat, *Russian Journal of Genetics*, **42**, 585-594, 2006.

[11] Bair E., Tibshirani R.: Machine learning method applied to DNA microarray data can improve the diagnosis of cancer, *SIGKDD Explorations*, **5**, 48-55, 2003.

[12] Schwender H., Ickstadt K., Rahnenfuhrer J.: Classification with high-dimensional genetic data: assigning patients and genetic features to known classes, *Biometrical Journal*, **50**, 911-926, 2008.

[13] Park J., Hwang S., Lee Y.S., Kim S.C., Lee D.: SNP Ethnos: a database of ethnically variant single-nucleotide polymorphisms, *Nucleic Acids Research*, **35, Database Issue**, D711-D715, 2007.

[14] Schwender H.: Modifying microarray analysis methods for categorical data-SAM and PAM for SNPs, *In: Classification-The Ubiquitous Challenge*, 370-377, 2005.

[15] The International HapMap consortium: The international Hapmap project, *Nature*, **426**, 789-796, 2003.

[16] Liao H., Ng M., Fung E., Sham P.: Unidimensional nonnegative scaling for genome-wide linkage disequilibrium maps, *International Journal of Bioinformatics Research and Applications*, **4**, 417-434, 2008.

[17] Hicks A.A., Petursson H., Jonsson T., Stefansson H., Johannsdottir H.S., Sainz J., et al: A susceptibility gene for late-onset idiopathic Parkinson's disease, *Annals of Neurology*, **52**, 549-555, 2002.

[18] Piluso G., Mirabella M., Ricci E., Belsito A., Abbondanza C., Servidei S., Puca A.A., Tonali P., Puca G.A., Nigro V.: Gamma1- and gamma2-syntrophins, two novel dystrophin-binding proteins localized in neuronal cells, *The Journal of biological chemistry*, **275**, 15851-15860, 2000.

[19] Watt F.M., Hogan B.L.: Out of eden: stem cells and their niches, *Science*, **287**, 1427-1430, 2000.