Perfect Population Classification on Hapmap Data with a Small Number of SNPs

Nina Zhou[1] and Lipo Wang[2]

[1] College of Information Engineering, Xiangtan University,
Xiangtan, Hunan, China
[2] Nanyang Technological University
Block S1, 50 Nanyang Avenue, Singapore 639798

Abstract. The single nucleotide polymorphisms (SNPs) are believed to determine human differences and, to some degree, provide biomedical researchers a possibility of predicting risks of some diseases and explaining patients' different responses to drug regimens. With the availability of millions of SNPs in the Hapmap Project, although large amount of information about SNPs is available, the tremendous size also causes a major challenge for research on SNPs. Inspired from the recent research work on population classification by Park et al (2006), we attempt to find as few SNPs as possible from the original nearly 4 millions SNPs to classify the 3 populations in the Hapmap genotype data. In this paper, we propose to first use a modified t-test measure to rank SNPs, and then combine the ranking result with a classifier, e.g., the support vector machine, to find the optimal SNP subset. Compared with Park et al's result, our proposed method is more efficient in ranking features and classifying the three populations, i.e., we obtained perfect classification using only 11 SNPs in comparison with 82 SNPs used by Park et al.

1 Background

A single nucleotide polymorphism (SNP, pronounced as "snip") is a small genetic variation occurring within a person's DNA sequence. For example, when the DNA sequence $AAAT\underline{C}CGG$ is changed to $AAAT\underline{T}CGG$, the variation, i.e., the replacement of the single nucleotide C by the single nucleotide T, is called an SNP variation. SNPs are the most common type of genetic variations in the human genome, very stable from generation to generation [4] and are believed to determine the human difference between any two unrelated individuals, e.g., different physical traits, different predispositions to diseases, and different responses to medicine. Therefore, SNPs can be effective biological markers for scientists to diagnose disease and track population ancestry.

Much research work on SNPs has already been explored, such as searching for genetic regions associated with complex diseases [8, 17], summarizing and analyzing SNPs for cost-effective genotyping [1, 7, 32]. These work can be categorized into association studies on SNPs. Usually association studies are based on the fact that SNPs in close proximity on the same chromosome are often correlated, which is measured by 'linkage disequilibrium' (LD) [2]. Therefore, the

M. Ishikawa et al. (Eds.): ICONIP 2007, Part II, LNCS 4985, pp. 789–797, 2008.

correlation between SNPs is always used to selecting the optimal subset of SNPs (also referred as the tagging SNPs). For example, Bafna et al. [1] and Hall et al. [7] searched for the SNPs with predictive power and determined neighborhoods for those predictive SNPs based on the correlation between SNPs. Then Bafna et al. [1] and Hall et al. [7] proposed the notion of informativeness, which measures how well a single SNP or a set of SNPs predict another single SNP or another set of SNPs within the neighborhoods. Finally, based on the informativeness measure, Bafna et al. [1] and Hall et al. [7] optimally selected the most informative subset of SNPs (tag SNPs) with the minimum size. Eran et al. [8] proposed a prediction accuracy measure to measure how well the value of an SNP is predicted by the values of only two closest tag SNPs, and utilized dynamic programming to find the set of tag SNPs which had the maximum prediction accuracy. For quickly finding a small number of tag SNPs, Eran et al. [8] also utilized the random sampling algorithm to randomly generate some sets of tag SNPs and find the set of tag SNPs with the maximum prediction accuracy. Phuong et al. [17] proposed to select tag SNPs by discarding redundant features, which was based on the method of feature selection using feature similarity (FSFS) [15]. Phuong et al. [17] first grouped features into clusters in which each feature is similar by the linkage disequilibrium (LD) measure γ^2 [18] and then choose one feature from each cluster as the representation of the cluster.

Other research such as tracking population history also has been developed, which is categorized into population studies on SNPs. For example, Rosenberg et al. [19] proposed to select genetic markers with highest informativeness for inference of individual ancestry. In 2005, Rosenberg proposed to select informative marker panels for population assignment. He used genotypes from eight species, i.e., carp, cat, chicken etc., as the experiment data, and compared five proposed multivariate algorithms to select efficient marker panels. All the five approaches are based on a performance function, which is used to measure the probability of correctly assigning individuals to its populations. The probability is the optimal rate of correct assignment (ORCA) in [19]. Although this algorithm is approximately a performance function, he pointed out that the algorithm can not be realistically realized if some terms in the ORCA are large. With the development of the Hapmap Project (www.hapmap.org), Park et al [16] proposed a different way to select informative SNPs to classify the three populations, i.e., Utah residents with ancestry from Northern and Western Europe (CEU), Yoruban in Ibadan, Nigeria in West Africa (YRI), and Han Chinese in Beijing together with Japanese in Tokyo (CHB+JPT). They proposed to adopt the nearest shrunken centroid method (NSCM) to rank SNPs for each class. That is, each SNP has three ranking scores for three populations. If the three scores of one SNP have great variance, this SNP will have great power for classifying the three populations. Or otherwise. In this way, they obtain the result of using the top 82 SNPs from nearly 4 millions SNPs to completely classify the three populations.

Inspired from the research work by Park et al (2006), we attempt to find as few SNPs as possible from the original nearly 4 millions SNPs to classify the 3 populations in the Hapmap genotype data. In this paper, we propose to firstly

rank SNPs according to a feature importance ranking measure, i.e., a modified t-test, where the higher the ranking value, the stronger the corresponding classification power. Secondly, from the ranking list, we randomly choose different numbers of top ranked SNPs, e.g., 2, 5, 7, 10 and so on, test them through a classifier, e.g., the support vector machine (SVM) [25, 26] and determine the SNP subset which has the smaller size and highest classification accuracy.

2 Methods

In many existed feature selection algorithms, feature ranking is often used to show which input features are more important [6, 28] so as to improve the efficiency of feature selection process, especially when a great number of features are involved. Therefore, feature ranking is used in our experiment to determine each feature's classification power. In this paper, we will adopt a t-test ranking measure modified from [3, 23, 27].

2.1 Modified T-test

The t-test, also called as the student t-test [3], is originally used to evaluate whether the means of two classes are statistically different from each other by calculating a ratio between the difference of two class means and variability of the two classes. It was adopted by [11, 21] to rank features (genes) for microarray data and for mass spectrometry data [31, 13]. We notice that the original t-test is only limited to 2-class problems. In order to extend the original t-test to multi-class problems, Tibshirani et al. [23] developed the nearest shrunken centroid method, i.e., calculating a t-statistic value (1) for each gene of each class. This t-statistic value measured the difference between the mean of one class and the mean of all the classes. The difference is standardized by the within-class standard deviation.

$$t_{ic} = \frac{\overline{x}_{ic} - \overline{x}_i}{M_c \cdot (S_i + S_0)} \tag{1}$$

$$S_i^2 = \frac{1}{N - C} \sum_{c=1}^{C} \sum_{j \in c} (x_{ij} - \overline{x}_{ic})^2 \tag{2}$$

$$M_c = \sqrt{1/n_c + 1/N} \tag{3}$$

Here t_{ic} denotes the t-statistics value for the i-th feature of the c-th class. $\overline{x}_{ic}$ denotes the i-th feature's mean value in the c-th class and $\overline{x}_i$ indicates the i-th feature's mean value for all classes. x_{ij} represents the i-th feature of the j-th sample. N is the total number of all the samples for all the C classes and n_c is the number of samples for the c-th class. S_i is the within-class standard deviation and S_0 is set to be the median value of S_i for all the features. This t-statistic value of [23] measured the deviation between each class and the mean of all classes and was used to constitute a classifier. The authors did not refer to using

the t-statistic of each class to rank features for all the classes. In the following work [27], Wang et al. extended the t-statistic algorithm to rank features for all the classes. That is, the t-score (t-statistic value) of feature i is calculated as the greatest t-score for all classes:

$$t_i = max\left\{\frac{|\overline{x}_{ic} - \overline{x}_i|}{M_c S_i}, c = 1, 2, ...C\right\} \tag{4}$$

However, (4) still can not be used to deal with our data. It is because of the non-numerical type. For example, if two alleles for one SNP are A and T, its feature values are expressed as AA, AT and TT. If representing them simply by three numerical values, e.g., 1, 2 and 3, and making the calculation according to (4), it will be meaningless. We proposed to use vectors to represent different feature values and thereby obtained the modified t-test (5), which can deal with our problem. In the following, we generalized the t-score of each feature in 3 steps:

1. Suppose the feature set is $F = (f_1, ..., f_i, ..., f_g)$, and feature i has m_i different nominal values represented as $f_i = (x_i^{(1)}, x_i^{(2)}, ..., x_i^{(m_i)})$
2. Transform each nominal feature value into a vector with the dimension m_i, i.e., $x_i^{(1)} \Rightarrow \boldsymbol{X}_i^{(1)}$={0, ..., 0, 1}, $x_i^{(2)} \Rightarrow \boldsymbol{X}_i^{(2)}$={0, ..., 1, 0}, ..., $x_i^{(m_i)} \Rightarrow \boldsymbol{X}_i^{(m_i)}$={1, ..., 0, 0}.
3. Replace all the numerical features in (1) and (2) with the format of vectors (see (5) and (6)).

$$t_i = max\left\{\frac{\left|\overline{\boldsymbol{X}}_{ic} - \overline{\boldsymbol{X}}_i\right|}{M_c S_i}, c = 1, 2, ..., C\right\} \tag{5}$$

$\overline{\boldsymbol{X}}_{ic}$ and $\overline{\boldsymbol{X}}_i$ are two row vectors indicating the i-th SNP's mean status in the c-th class and mean status for all the classes. $\left|\overline{\boldsymbol{X}}_{ic} - \overline{\boldsymbol{X}}_i\right|$ denotes the Euclidean distance of the two vectors.

$$S_i^2 = \frac{1}{N - C}\sum_{c=1}^{C}\sum_{j \in c}(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_{ic})(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_{ic})^T \tag{6}$$

Here $\boldsymbol{X}_{ij}$ is a row vector denoting the i-th SNP in the j-th class. $(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_{ic})(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_{ic})^T$ is a scalar.

The ranking rule is: the greater the t-scores, the more relevant the features.

2.2 The Classifier

The classifier in our experiment will be used twice. In the first time, we will use a classifier to test different feature subsets, formed from the ranking list with top ranking values, and determine a candidate feature subset. In the second time, we again need to test different feature subsets generated from the candidate feature

subset, and find the optimal feature subset, i.e., the one with the best classification accuracy and minimum size. Considering the importance of the classifier, we would like to choose the support vector machine (SVM) [25] as the classifier because of its very good performance, such as effectively avoiding overfitting and accomodating large feature spaces, and successfully used in bioinformatics [14, 27]. Since Hsu et al. [9] indicated that RBF kernel is generally a first choice and Keerthi et al. [12] showed that the linear kernel is special case of the RBF kernel, we choose the RBF kernel for the SVM in our experiment. During the classification process, the kernel parameter γ and the penal parameter ν [9] are determined through a double cross-validation method [5]. For example, for a 10-fold cross validation method, we first separate the original samples into 10 equal subsets, each time having one subset as the testing set, and the other nine subsets as the training set. Then for the training set, we use the 10-fold cross validation one more time.

3 Experiments and Discussion

3.1 Experimental Data and Its Preprocessing

The genotype data is downloaded from the the directory of (/Index of/genotypes/latest_ncbi_build36/rs_strand/non-redundant) on the website (http://www.hapmap.org/genotypes/), which contains data files with genotypes submitted by HapMap genotyping centers to the HapMap Data Coordination Center (DCC) to date. From the column 12 to the last column, data files provide observed genotypes of samples (one genotype per column) with sample identifiers in column headers (Coriell catalog numbers, example: NA10847) and duplicate samples having .dup suffix. The genotypes are provided for each chromosome of each population, i.e., chromosomes 1-22, chromosome X and Y, and four populations: CEU, YRI, JPT and HCB. Here CEU represents Utah residents with ancestry from northern and western Europe. YRI represents Yoruba individuals from Ibadan and Nigeria. Each of the two populations has 90 reference individuals (samples) which are comprised of 30 father-mother-offspring trios. JPT represents Japanese individuals from Tokyo, and HCB means Han Chinese individuals from Beijing. Each of these two populations has 45 samples and the individuals in each of the populations are unrelated. For efficient experiments, we remove the children samples from the CEU and YRI populations to make sure all the samples involved in the experiment are unrelated. Thus the total number of samples used in our experiment is 210. Usually the JPT and CHB can be classified as one population (denoted as JPT+CHB) because of their similar DNA sequence. In this paper, we will carry out two respective classifications on the original 4 populations and the 3 populations, i.e., CEU, YRI, JPT+CHB.

Combining all the features together from the 24 chromosomes (Chromosome 1, 2, ..., 22, X and Y), we have nearly 4 million SNPs involved in the experiment. SNPs are usually expressed as strings of two or more alleles, e.g., *AT* or *ATCG*. SNPs with two alleles are called as bi-allelic SNPs and SNPs with 3 or 4 alleles are called as multi-allelic SNPs. If the alleles consisting of one SNP are the same,

e.g., AA or TT, this type of SNPs are called homozygous. Or they are called as hyterozygous, e.g., AT. Although some locus show us there may be 3 or 4 alleles at those positions, e.g., 4 alleles $A/T/C/G$ at one SNP position, their real feature values consist of only two alleles, e.g., A/T. These are the error descriptions existed in the data and have already been announced in the website. Therefore, all data samples are strings of bi-allelic SNPs. Besides, we notice some features have unknown value for some samples. In this case, instead of removing those features from our experiment, we will replace them according to the rule adopted by Park et al [16] in their experiment. That is, we replace the missing value with the major allele of each population class. After this preprocessing, we need to transform the nominal features into vectors as the modified t-test algorithm required. For example, according to description of the generalized steps, AA is represented by {0, 0, 1}, TT is represented by {0, 1, 0}, and AT is represented by {1, 0, 0}. The three bits of vectors represent three different features (SNPs). Therefore, the calculation between them will not lose the information of the three different feature values.

3.2 Implementation

After using the modified t-test ranking measure, we have two experiments to conduct, i.e., classification on 3 populations and 4 populations, respectively. From the 210 samples, we randomly choose 40 samples from YRI and CEU, respectively, and 30 samples from JPT and CHB, respectively, as the training set. The 70 samples left are used as the testing set.

We first rank the SNPs of 24 chromosomes, respectively. Then from the 23 ranking lists (except chromosome Y which only has 49 SNPs) we choose their top 100 SNPs to form a new feature subset with 2300 features together with the 49 features form Chromosome Y. According to their ranking scores, we re-rank them again. In this way, we greatly reduce the number of features involved in the experiment. Furthermore, this will not lead to loss of important information and instead will facilitate to improve the efficiency of the experiment.

3.3 Results and Discussions

In Table 1, we provide the ranking result from the modified t-test ranking measure. It includes four types of information. The first column (Ranking NO.) means the ranking order of top 11 features corresponding to their ranking scores. The second column lists the 11 features' names and the third column (Chromosome) provides the location of each SNP. The fourth and fifth column list the ranking score of each SNP, respectively for 3-population and 4-population problem. Although the ranking list for 3 populations is the same as the one for 4 populations, ranking values are different for the two conditions. Based on the ranking list in Table 1, we combine different number of features (see Table 2), i.e., 2, 5, 7, 10, 11 and 20, and input them into the classifier, respectively. From Table 2, we find out that 3 populations are completely classified (100% accuracy) when the top 11 features are input. While using the same 11 features to classify the 4 populations, we obtain the accuracy 78.57% (55/70), in which 55

Table 1. Top 11 features' ranking list for the modified t-test ranking measure. SNP names by boldface indicate that those SNPs' combination leads to best classification.

Ranking No.	Name of SNPs	Chromosome	Ranking value for 3 populations	Ranking value for 4 populations
1	**rs11499**	chr3	9.6017	9.5666
2	rs5825	chr4	8.1264	8.1022
3	rs4143483	chr4	7.2546	7.2281
4	rs1299386	chr7	7.2546	7.2281
5	rs1813166	chr7	6.7457	6.7210
6	rs2040513	chr7	6.7457	6.7210
7	rs4131595	chr7	6.7457	6.7210
8	**rs289632**	chr8	6.5661	6.5421
9	rs1785847	chr18	6.5661	6.5421
10	rs2474273	chrX	6.5661	6.5421
11	**rs4120141**	chrX	6.4379	6.4144

Table 2. Classification accuracy for different feature subsets formed from the ranking list in Table 1. The number of SNPs that leads to the best classification is indicated by boldface.

Number of features	Accuracy for 3 populations	Accuracy for 4 populations
2	70% (49/70)	55.71% (39/70)
5	70% (49/70)	55.71% (39/70)
7	70% (49/70)	57.14% (40/70)
10	98.57% (69/70)	57.14% (40/70)
11	100% (70/70)	78.57% (55/70)
20	100% (70/70)	78.57% (55/70)

of 70 testing samples are correctly classified. It means that CEU and YRI are completely recognized and JPT and CHB are recognized as the third class.

4 Conclusion

In this paper, we proposed a modified t-test ranking measure to rank a large amount of SNPs. This measure is able to deal with data with nominal features in the form of vectors, which is the major superiority over the original t-test ranking measure. Besides, we adopted the F-statistics ranking measure [24] on the genotype data and compared the results with those obtained from the modified t-test ranking measure. The comparisons showed that the modified t-test ranking measure is comparable with the F-statistics ranking measure. However, due to space limitation, we will not present the comparisons in this paper. After obtaining the ranked features, we utilize a classifier to determine an optimal feature subset, which has the minimum size but leads to the highest classification accuracy. The final results show that the modified t-test ranking method is

efficient on determining the importance of the SNPs. Compared to the classification method of Park et al[16], we obtained better result, i.e., perfect classification of the 3 populations using only 11 SNPs, compared to 82 SNPs used in [16].

References

[1] Bafna, V., Halldorsson, B., Schwartz, R., Clark, A., Istrail, S.: Haplotypes and Informative SNP selection: Don't block out information. In: Proc. of RECOMB, pp. 19–27 (2003)

[2] Celedon, J.C.: Candidate genes, SNPs, Haplotypes and linkage disequilibrium. Powerpoint presentation (2004), http://innateimmunity.net/files/CANDGENES/siframes.html

[3] Devore, J., Peck, R.: Statistics:the exploration and analysis of data, 3rd edn. Duxbury Press, Pacific Grove (1997)

[4] Duerinck, K.F.: (2001), http://www.duerinck.com/snp.html

[5] Francois, R., Langrognet, F.: Double Cross Validation for Model Based Classification, User (2006), http://www.r-project.org/user-2006/Abstracts/Francois+Langrognet.pdf

[6] Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3, 1157–1182 (2003)

[7] Halldrsson, B., Bafna, V., Lippert, R., Schwartz, R., de la Vega, F., Clark, A., Istrail, S.: Optimal haplotype blockfree selection of tagging snps for genome-wide association studies. Genome research 14, 1633–1640 (2004)

[8] Halperin, E., Kimmel, G., Shamir, R.: Tag SNP selection in genotype data for maximizig SNP prediction accuracy. Bioinformatics 199, 195–203 (2005)

[9] Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei (2003)

[10] Human genome project information (2006), http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.html

[11] Jaeger, J., Sengupta, R., Ruzzo, W.L.: Improved Gene Selection For Classification Of Microarrays. Pac. Symp. Biocomput., 53–64 (2003)

[12] Keerthi, S.S., Lin, C.-J.: Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Computation 15, 1667–1689 (2003)

[13] Levner, I.: Feature selection and nearest centroid classification for protein mass spectrometry. BMC Bioinformatics 6, 68 (2005)

[14] Liu, B., Wan, C.R., Wang, L.P.: An efficient semi-unsupervised gene selection method via spectral biclustering. IEEE Trans. on Nano-Bioscience 5, 110–114 (2006)

[15] Mitra, Pabitra, Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity. IEEE trans. on Pattern analysis and machine intelligence 3, 301–312 (2002)

[16] Park, J.S., Hwang, S.H., Lee, Y.S., Kim, S.C.: SNP@Ethnos: a database of ethnically variant single-nucleotide polymorphisms. Nucleic Acids Research 0, D1–D5 (2006)

[17] Phuong, T.M., Lin, Z., Altman, R.B.: Choosing SNPs using Feature Selection. In: Proc IEEE Comput Syst Bioinform Conf. 2005 (CSB 2005), pp. 301–309 (2005)

[18] Pritchard, J.K., Przeworski, M.: Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. 69, 1–14 (2001)

[19] Rosenberg, N.A., et al.: Informativeness of genetic markers for inference of ancestry. Am. J. Hum. Genet. 73, 1402–1422 (2003)
[20] Rosenberg, N.A.: Algorithms for selecting informative marker panels for population assignment. Journal of computational biology 9, 1183–1201 (2005)
[21] Su, Y., Murali, T.M., Pavlovic, V., Schaffer, M., Kasif, S.: RankGene: Identifcation of Diagnostic Genes Based on Expression Data. Bioinformatics 19, 1578–1579 (2003)
[22] The International HapMap Consortium: The international Hapmap Project. Nature 426, 789–796 (2003), www.hapmap.org/genotypes
[23] Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS 99, 6567–6572 (2002)
[24] Trochim, W.M.: The Research Methods Knowledge Base, 2nd edn. Atomic Dog Publishing (2004), http://www.socialresearchmethods.net/kb/
[25] Vapnik, V.: Statistical learning theory. Wiley, NewYork (1998)
[26] Wang, L.P.: Support Vector Machines: Theory and Applications. Springer, Heidelberg (2005)
[27] Wang, L.P., Chu, F., Xie, W.: Accurate cancer classification using expressions of very few genes. IEEE Transactions on Bioinformatics and Computational Biology 4, 40–53 (2007)
[28] Wang, L.P., Fu, X.J.: Data Mining with Computational Intelligence. Springer, Berlin (2005)
[29] Welch, B.L.: The generalizaition of student's problem when several different population are involved. Biomethika 34, 28–35 (1947)
[30] Wright, S.: The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 19, 395–420 (1965)
[31] Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., Zhao, H.: Comparison of statistical methods for classifcation of ovarian cancer using mass spectrometry data. BioInformatics 19, 1636–1643 (2003)
[32] Zhen, L., Altman, R.B.: Finding Haplotype Tagging SNPs by Use of Principle Components Analysis. Am. J. Hum. Genet. 75, 850–861 (2004)