Journal of Biomedical Informatics 43 (2010) 233–239

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

SKM-SNP: SNP markers detection method

Yang Liu [a], Mark Li [a], Yiu M. Cheung [b], Pak C. Sham [c], Michael K. Ng [a,*]

[a] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
[b] Department of Computer Science, Hong Kong Baptist University, Hong Kong
[c] Department of Psychiatry, The University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 3 April 2009
Available online 17 November 2009

Keywords:
Single nucleotide polymorphism
Subspace clustering
SKM-SNP
K-mode

ABSTRACT

SKM-SNP, SNP markers detection program, is proposed to identify a set of relevant SNPs for the association between a disease and multiple marker genotypes. We employ a subspace categorical clustering algorithm to compute a weight for each SNP in the group of patient samples and the group of normal samples, and use the weights to identify the subsets of relevant SNPs that categorize these two groups. The experiments on both Schizophrenia and Parkinson Disease data sets containing genome-wide SNPs are reported to demonstrate the program. Results indicate that our method can find some relevant SNPs that categorize the disease samples. The online SKM-SNP program is available at http://www.math.hkbu.edu.hk/~mng/SKM-SNP/SKM-SNP.html.

1. Introduction

Variations (e.g., insertions, deletions, and mutations) in the DNA sequences of humans have a major impact on genetic diseases and phenotypic differences. Single nucleotide polymorphism (SNP) is one of the most common DNA sequence variation occurring when a single nucleotide—A, T, C, or G—in the genome (or other shared sequence) differs between members of species (or between paired chromosomes in an individual). In SNPs data sets, the association between a disease and a set of relevant SNPs are investigated. Patients and normals are often categorized in groups according to their SNPs. Thousands of SNPs in different regions of chromosomes are used to describe characteristics of patient/normal samples.

High-dimensional data is a phenomenon in the field of bioinformatics. Above SNP data set is a typical example. Clearly, clustering of high-dimensional categorical data requires special treatment. There are two key properties of data sets of such data mining tasks: high-dimensional and categorical.

To tackle high-dimensional data, some subspace clustering methods are proposed and studied, see [1] for details. The basic idea of the methods is to find clusters from subspaces of data instead of the entire data space. In subspace data clustering, each cluster is a set of objects identified by a subset of dimensions and different clusters are represented in different subsets of dimensions. The major challenge of subspace clustering, which makes it distinctive from traditional clustering, is the simultaneous determination of both cluster memberships of objects and the subspace of each cluster.

Cluster memberships are determined by the similarities of objects measured with respect to subspaces. According to the ways that the subspaces of clusters are determined, subspace clustering methods can be divided into two types. The first type is to find out the exact subspaces of different clusters [2–8]. The second type is to cluster data objects in the entire data space but assign different weights to different dimensions of clusters in the clustering process, based on the importance of the dimensions in identifying the corresponding clusters [9–16,**?**]. However, all these methods are developed to handle numerical data sets.

One widely used SNP selection approach for candidate gene studies is based on potential impact on protein functions or gene regulations [18–21]. A problem with these methods is that such biological information is rarely available or still unknown to human beings. In this paper, we develop SKM-SNP, a SNP markers detection program, which employ a subspace clustering algorithm to determine a set of relevant SNPs for the association between a disease and multiple marker genotypes. We consider that different SNPs to be categorical dimensions and they make different contributions to identification of (patient or normal) samples in clusters. The difference of contribution of a SNP (categorical dimension) is represented as a weight that can be treated as the degree of the dimension in contribution to the cluster. In subspace clustering, the decrease of the weight entropy in a cluster implies the increase of certainty of a subset of dimensions with larger weights in determination of the cluster. Therefore, in the clustering process, we simultaneously minimize the within cluster dispersion and the weight entropy to stimulate more dimensions to contribute to the identification of a cluster. A formula for computing a dimension weight is implemented to the clustering process as an additional step in each iteration, so the cluster memberships of samples and

* Corresponding author. Fax: +852 3411 5811.
E-mail address: mng@hkbu.edu.hk (M.K. Ng).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.11.004

the weights of SNPs in each cluster can be obtained simultaneously.

2. Methods

2.1. The algorithm

SKM-SNP is a new K-mode-type algorithm for soft subspace clustering of high-dimensional categorical data. In this algorithm, we consider that the weight of a dimension in a cluster represents the probability of contribution of that dimension in forming the cluster. The entropy of the dimension weights represents the certainty of dimensions in identification of a cluster. Therefore, we consider the K-mode objective function by adding the weight entropy term to it so that we can simultaneously minimize the within cluster dispersion and the weight entropy to stimulate more dimensions to contribute to the identification of clusters. The SKM-SNP program is based on the minimization of an objective function (1):

$$\sum_{l=1}^{k}\sum_{j=1}^{n}\sum_{i=1}^{m}\omega_{l,j}\lambda_{l,i}\delta(z_{l,i},x_{j,i})+\gamma\sum_{l=1}^{k}\sum_{i=1}^{m}\lambda_{l,i}\log\lambda_{l,i} \quad (1)$$

subject to

$$\begin{cases}\sum_{l=1}^{k} w_{lj}=1, \quad 1\leqslant j\leqslant n, \quad 1\leqslant l\leqslant k, \quad w_{lj}\in\{0,1\}\\ \sum_{i=1}^{m}\lambda_{li}=1, \quad 1\leqslant l\leqslant k, \quad 1\leqslant i\leqslant m, \quad 0\leqslant\lambda_{li}\leqslant 1.\end{cases}$$

Here n is the number of samples, k is the number of clusters, m is the number of SNPs, $x_{j,i}$ is the ith SNP of the jth sample, $z_{l,i}$ is the ith SNP of the lth center mode, $\delta(z_{l,i},x_{j,i})$ is a distance function which is equal to one if both genotypes $z_{l,i}$ and $x_{j,i}$ are the same, or equal to zero if they are different, $\lambda_{l,i}$ is the weight of the ith SNP of the lth center mode, $w_{l,j}$ is the degree of membership of the jth sample to the lth cluster.

The idea of the minimization of (1) is to partition the samples to the correct group (to determine $w_{l,j}$); to find the representatives of normal and disease groups (to determine $z_{l,i}$); and to find the relevance of SNPs in each group (to determine $\lambda_{l,i}$). We note that the first term in (1) is the sum of the within cluster dispersions and the second term the weight entropy. The positive parameter γ controls the strength of the incentive for clustering on more SNPs. For detail about subspace clustering for numerical data only, we refer to the paper [22].

To minimize the objective function in (1), we first initialize the center modes of genotypes of SNPs randomly and all the weights of SNPs to $1/m$. Afterward we can start an iterative process of partitioning the samples, updating cluster centers modes of genotypes, and calculating the weights of the SNPs. The iterative loop is repeated until the objective function value does not improve. In each step, we have explicit formulae to handle the computation.

- The partitioning of the sample is given as follows:

$$w_{lj}=\begin{cases}1, & \text{if } \sum_{i=1}^{m}\lambda_{li}\phi(z_{l,i},x_{j,i})\leqslant\sum_{i=1}^{m}\lambda_{ri}\phi(z_{r,i},x_{j,i})\ \forall r,\\ 0, & \text{otherwise},\end{cases} \quad (2)$$

  i.e., when there are more genotypes of a sample consistent with the representative of a normal/disease group, the sample is assigned to that group.
- The center modes $z_{l,i}$ is set to be the genotype $a_i^{(r)}$ if

$$|\{w_{l,j}|x_{j,i}=a_i^{(r)},w_{l,j}=1\}|\geqslant|\{w_{l,j}|x_{j,i}=a_i^{(t)},w_{l,j}=1\}|\ \forall t, \quad (3)$$

  where t is an index for the number of possible combination of genotypes of the ith SNP. Here the SNP genotype of the representative in the normal/disease group is the most frequent SNP genotype of the samples in that group.
- The dimension weights can be calculated as follows:

$$\lambda_{l,i}=\frac{\exp\left(\frac{-D_{l,i}}{\gamma}\right)}{\sum_{t=1}^{m}\exp\left(\frac{-D_{l,t}}{\gamma}\right)} \text{ where } D_{l,t}=\sum_{j=1}^{n}w_{l,j}\delta(z_{l,t},x_{j,t}). \quad (4)$$

  Here the ith SNP weight calculation is based on its relevance with respect to other SNPs, where the relevance is measured by the aggregated difference between the SNP genotype of the samples in the group and that of the corresponding center mode. The input parameter γ is used to control the size of the weights as follows:

  (i) $\gamma>0$: In this case, according to (4), λ_{li} is inversely proportional to D_{li}. The smaller D_{li}, the larger λ_{li}, the more important the corresponding SNP.
  (ii) $\gamma=0$: $\lambda_{li'}$ is equal to one, indicating that the index i' has the smallest value of $D_{li'}$. The other weights λ_{li} for $i\neq i'$ are equal to zero. Each cluster contains only one important dimension. It may not be desirable for high-dimensional data sets.
  (iii) $\gamma<0$: In this case, according to (4), λ_{li} is proportional to D_{li}. The larger D_{li}, the larger λ_{li}. This is contradictory to the original idea of dimension weighting. Therefore, γ cannot be smaller than zero.

2.2. Convergency and complexity analysis

The proposed algorithm converges in a finite number of iterations. To divide a data set into k clusters, the number of possible partitions is finite. We can show that each possible partition W only occurs once in the clustering process. Assume that $W^{h_1}=W^{h_2}$, where h is the iteration index and $h_1\neq h_2$. We note that given W^h, we can compute the minimizer Z^h which is independent of Λ^h according to (3). For W^{h_1} and W^{h_2}, we have the minimizers Z^{h_1} and Z^{h_2}, respectively. It is clear that $Z^{h_1}=Z^{h_2}$ since $W^{h_1}=W^{h_2}$. Using W^{h_1} and Z^{h_1}, and W^{h_2} and Z^{h_2}, we can compute the minimizers Λ^{h_1} and Λ^{h_2} respectively according to (4). It is clear that $\Lambda^{h_1}=\Lambda^{h_2}$. Therefore, we obtain

$$F(W^{h_1},Z^{h_1},\Lambda^{h_1})=F(W^{h_2},Z^{h_2},\Lambda^{h_2}).$$

However, the sequence $F(\cdot,\cdot,\cdot)$ generated by the algorithm is strictly decreasing. Therefore, the proposed algorithm converges in a finite number of iterations.

The complexity of the algorithm in each step depends on the number of SNPs, the number of possible combination of genotypes, the number of samples and the number of clusters. This is because it only adds a new step to the K-mode clustering process to calculate the dimension weights of each cluster. The run-time complexity can be analyzed as follows. We only consider the three major computational steps:

- **Partitioning the objects**: After initialization of the dimension weights of each cluster and the cluster centers, a cluster membership is assigned to each object. This process simply compares the summation of

$$\sum_{i=1}^{m}\lambda_{li}\phi(z_{l,i},x_{j,i})$$

  in (2) for each object in all k clusters. Thus, the complexity for this step is $O(mnk)$ operations.
- **Updating cluster centers**: Given the partition matrix W, updating cluster centers is to find the means of the objects in the same cluster. Thus, for k clusters, the computational complexity for this step is $O(mnk)$.

- **Calculating dimensions weights**: The last phase of this algorithm is to calculate the dimensions weights for all clusters based on the partition matrix W and Z. In this step, we only go through the whole data set once to update the dimensions weights. The computational complexity of this step is also $O(mnk)$.

If the clustering process needs h iterations to converge, the total computational complexity of this algorithm is $O(hmnk)$. This shows that the computational complexity increases linearly as the number of dimensions, or objects or clusters increases.

2.3. An example

In this subsection, we make use of the following example to demonstrate the proposed algorithm. In this example, there are six samples and four SNPs. Below symbols "A" and "a" represent the two alleles. The symbol "M" refers to the missing percentages of genotypes.

Sample	SNP			
	1	2	3	4
I	AA	AA	M	Aa
II	AA	AA	Aa	AA
III	AA	AA	aa	M
IV	Aa	aa	aa	aa
V	aa	AA	aa	aa
VI	M	Aa	aa	aa

We first initialize the two center modes of genotypes of SNPs, for example,

$$[z_{11}, z_{12}, z_{13}, z_{14}] = [AA, AA, AA, AA] \quad \text{and} \quad [z_{21}, z_{22}, z_{23}, z_{24}] = [aa, aa, aa, aa]$$

We also set the γ to be 1 and all the initial weights of SNPs to 1/4. Next we partition the samples by computing the distance between the sample and the center modes as in (2):

Sample	The distance	
	Center mode 1	Center mode 2
I	1/2 (∗)	1
II	1/4 (∗)	1
III	1/4 (∗)	3/4
IV	1	1/4 (∗)
V	3/4	1/4 (∗)
VI	1	1/2 (∗)

Here $*$ in the bracket refers to the sample belonging the particular center mode. Now we can use the partitioning of the samples to calculate $D_{l,t}$ and then the weights of SNPs as in (4):

SNP	$D_{l,i}$		$\lambda_{l,i}$	
	Center mode 1	Center mode 2	Center mode 1	Center mode 2
1	0	2	0.88	0.12
2	0	2	0.88	0.12
3	3	0	0.95	0.05
4	2	0	0.95	0.05

The above results tell us that SNP 1 and SNP 2 are relevant to the first cluster, while SNP 3 and SNP 4 are relevant to the second cluster. In the next step, we need to update center mode 1 and center mode 2 by using (3):

$$[z_{11}, z_{12}, z_{13}, z_{14}] = [AA, AA, \star, \star] \quad \text{and} \quad [z_{21}, z_{22}, z_{23}, z_{24}] = [\star, \star, aa, aa]$$

Since there is no dominant category in SNP 3 and SNP 4 for the first cluster and no dominant category in SNP 1 and SNP 2 for the second cluster, the category can be assigned arbitrary. Indeed, they are not relevant SNPs in the clusters. As the partitioning of samples is the same as before, the algorithm can be stopped and the clustering results are obtained.

3. Experimental results

3.1. Schizophrenia SNPs data set

In this subsection, we analyze the case/control populations of patients served in a data set from Genome Research Center, The University of Hong Kong. The data is related to Schizophrenia and is consisted of 488 cases (patients) recruited from hospitals in Hong Kong and 520 controls (normal) recruited from the community. There are 144 SNPs in the data set. Schizophrenia is a serious mental disorder affecting close to 1% of the population world-wide. Such disease incurs huge economic burden and human suffering. A large genetic component has been demonstrated by family, twin and adoption studies. The mode of inheritance is complex with multiple genes all contributing to the overall liability of developing the disorder. Recent molecular genetic studies have revealed a number of possible susceptibility loci. By using the genotype data already generated for the HapMap project [23] based on Asian samples. There are 144 SNPs on chromosome 3p that are picked by CLUSTAG developed by Ao et al. [24] making an average marker density of 1 tagging SNP per 25 kb.

The accuracy measure is used to evaluate the performance of the clustering algorithm. Objects in a lth cluster are assumed to be classified either correctly or incorrectly with respect to a given class of objects. Let the number of correctly classified objects be n_l, we can calculate the clustering accuracy as:

$$r = \frac{\sum_{l=1}^{k} n_l}{n} \tag{5}$$

where n is the total number of objects. Table 1 shows the clustering accuracy results of different algorithms: SKM-SNP, K-mode [25], COSA [15] and PROCLUS [3] algorithms. We can see from the table that the SKM-SNP algorithm is better than the other algorithms. The focus of our study is to determine the relevant SNPs associated to the case/control populations. In Fig. 1, we show the weights of SNPs for the case and control groups. It is clear from the figure that there are some weights of SNPs for the case and control populations that are about the same, however, there are some SNPs where they have significant different patterns of weights.

In Table 2, we show the top ten weights of SNPs in the control group and their corresponding genotypes distributions where A and a represent the major and minor alleles. The column under "M" refers to the missing percentages of genotypes in the group. As a comparison, we also list the genotypes distribution of the selected SNPs of the case group in Table 2. Similarly, in Table 3, we show the top ten weights of SNPs in the case group and the corre-

Table 1
Clustering accuracy results for different algorithms for Schizophrenia dataset.

Algorithm	Clustering accuracy
SKM-SNP	0.6824
K-mode	0.5203
COSA	0.5000
PROCLUS	0.5100

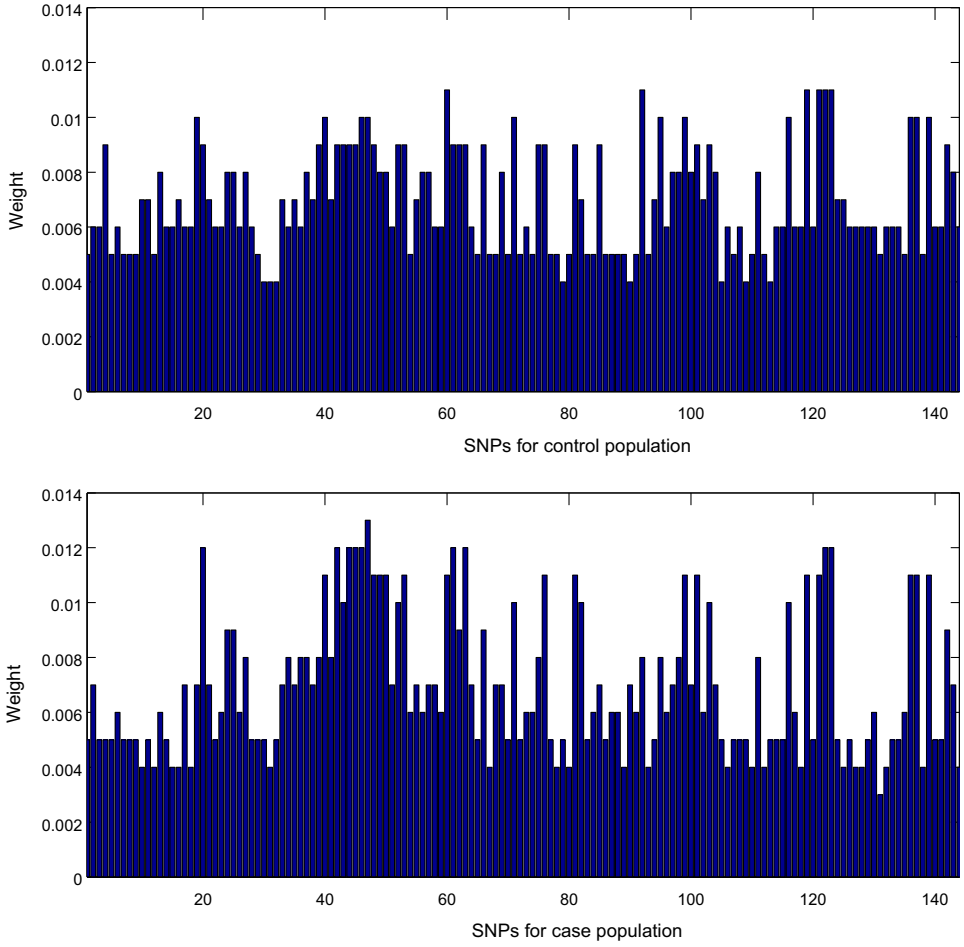

Fig. 1. The weights of SNPs for Schizophrenia dataset: control population (upper) and case population (lower).

Table 2
The top ten weights of SNPs in the control group and their genotypes distributions for Schizophrenia dataset.

SNPs	Control					Case				
	Weight	AA	Aa	aa	M	Weight	AA	Aa	aa	M
60	0.011	100.0	0.0	0.0	0.0	0.011	89.5	3.5	0.0	7.0
92	0.011	96.8	3.2	0.0	0.0	0.008	74.4	24.4	0.0	1.2
119	0.011	100.0	0.0	0.0	0.0	0.011	89.5	2.4	0.0	8.1
121	0.011	98.4	0.0	0.0	1.6	0.011	91.9	3.4	0.0	4.7
122	0.011	100.0	0.0	0.0	0.0	0.012	93.0	1.2	0.0	5.8
123	0.011	100.0	0.0	0.0	0.0	0.012	93.0	2.3	0.0	4.7
19	0.010	100.0	0.0	0.0	0.0	0.007	66.3	19.7	0.0	14.0
40	0.010	95.3	3.1	0.0	1.6	0.011	89.5	5.8	0.0	4.7
46	0.010	91.9	8.1	0.0	0.0	0.012	95.3	0.0	0.0	4.7
47	0.010	91.9	6.5	0.0	1.6	0.013	97.7	0.0	0.0	2.3

Table 3
The top ten weights of SNPs in the case group and their genotypes distributions for Schizophrenia dataset.

SNPs	Case					Control				
	Weight	AA	Aa	aa	M	Weight	AA	Aa	aa	M
47	0.013	97.7	0.0	0.0	2.3	0.010	91.9	6.5	0.0	1.6
20	0.012	93.0	1.2	0.0	5.8	0.009	90.3	9.7	0.0	0.0
42	0.012	95.3	0.0	0.0	4.7	0.009	90.3	8.1	0.0	1.6
44	0.012	93.0	0.0	0.0	7.0	0.009	91.9	8.1	0.0	0.0
45	0.012	95.3	0.0	0.0	4.7	0.009	90.3	8.1	1.6	0.0
46	0.012	95.3	0.0	0.0	4.7	0.010	91.9	8.1	0.0	0.0
61	0.012	93.0	0.0	0.0	7.0	0.009	87.1	4.8	0.0	8.1
63	0.012	95.3	1.2	0.0	3.5	0.009	90.3	4.9	0.0	4.8
122	0.012	93.0	1.2	0.0	5.8	0.011	100	0.0	0.0	0.0
123	0.012	93.0	2.3	0.0	4.7	0.011	100	0.0	0.0	0.0

sponding SNPs in the control group for a comparison. We find from the two tables that there are more missing value categories in the case population than those in the control population. Two populations have the four same SNPs (46, 47, 122, 123) in their lists. For these four SNPs, their genotypes distributions are quite different. Even we remove the missing value category, their genotypes distributions are also quite different. These results demonstrate that these four SNPs may associate to the case/control populations. For the other SNPs (19, 60, 92, 119, 121) in the tables, we can find the genotypes distributions of two populations are quite different. Among these five SNPs, the most significant one is 19. Further biological investigations on the above SNPs can determine their genetic functions and study how they are relevant to the disease.

3.2. Parkinson Disease SNPs data set

The other SNPs data set is 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). The original data set has two parts: cc (Caucasian controls) and pd (Parkinson Disease), each of them consists of two sub-parts: map file and pre file. The whole file of cc contains 271 of the actual genotypes for Caucasian controls, while the pd contains 270 of the genotypes for Parkinson Disease patients. Each SNP map file contains the chromosome, NCBI Build 35 position, marker name, major allele, minor allele, major allele frequency, minor allele frequency and number of missing genotypes for the marker; where the frequencies and number of missing genotypes are based upon the 271/270 individuals in the datasets. All the pre file (or genotypes file) contains the individual ID (in the format of ND-XXXX), affectation status (1 for unaffected, 2 for affected, can also be a class label) followed by unencoded allele calls (0 is a missing genotypes) for each genotype in the order specified within the map file. The alleles are called in forward orientation according to dbSNP.

We do the data reprocessing as follows: each chromosome will be a separate file, that is to say, we combine each chromosome of $cc\_*.pre$ file and $pd\_*.pre$ file to one file. For Individual Id, we just ignore it and for class Label, we read them out and consider them as the final class label in order to have a reference for clustering accuracy comparisons. All the remaining genotypes, we combine every two of them together to make up a SNP. For each list of SNPs, we compute the allele that appear with the most frequency and label it as the major in this particular list and all the other alleles will be the minor. For SNPs that appear with major/major, we label it as 0. For SNPs that with major/minor or minor/major, we label it as 1. For SNPs that with all the combinations of minors, we label it as 2. Missing values will be represented as 3. Details of data processing step can be seen in our webpage at http://www.math.hkbu.edu.hk/~mng/SKM-SNP/SKM-SNP.html.

Table 4 shows the clustering accuracy results (correctly classified samples) for 22 chromosomes by using the SKM-SNP program, compared with traditional *K*-mode algorithm. We use the most frequent genotypes in case and control groups to be the initial modes for the program. The parameter γ is tuned in each chromosome to obtain the highest accuracy in the test. We find that the average clustering accuracy of our proposed algorithm is higher than that of *K*-mode algorithm which is non-subspace-type [25] by 3.0%. In the table, we show the computation time of the proposed algorithm, and find that it only takes about a minute to generate the weights of the SNPs and the clustering results.

Table 4
Clustering accuracy results. The number in the bracket refers to the number of SNPs in the chromosome.

Chromosome	*K*-mode	SKM-SNP		
	Accuracy	Accuracy	γ	Time
1 (31,532)	0.8614	0.9020	2200	45.7
2 (32,706)	0.9150	0.9298	700	47.8
3 (27,691)	0.8152	0.8281	3000	46.4
4 (24,193)	0.7930	0.8262	5000	65.7
5 (24,570)	0.8207	0.8521	500	30.9
6 (26,372)	0.7079	0.8706	1000	39.1
7 (21,382)	0.7560	0.8115	1400	44.8
8 (22,434)	0.7431	0.7523	2200	63.2
9 (19,542)	0.7800	0.8022	7000	40.8
10 (20,007)	0.7283	0.7449	2300	42.9
11 (19,539)	0.7689	0.7911	3000	41.0
12 (19,572)	0.7616	0.7745	2000	48.8
13 (14,123)	0.7264	0.7726	4000	26.5
14 (12,645)	0.6802	0.7061	3400	23.7
15 (11,618)	0.6433	0.6525	2500	31.6
16 (11,767)	0.7006	0.7523	2800	24.7
17 (11,619)	0.6266	0.6266	2000	22.9
18 (12,613)	0.7375	0.7505	2000	26.6
19 (8608)	0.6451	0.6710	2300	23.5
20 (10,375)	0.6451	0.6617	1500	30.2
21 (6612)	0.5749	0.6081	1900	14.2
22 (7071)	0.6118	0.6266	3000	14.9

Because of the curse of dimensionality, we must select important and relevant dimensions for clustering and classification in high-dimensional data sets. Experimental results in [2–4,14,15,17] have shown that the selection of dimensions is very important for obtaining good results for high-dimensional numerical data sets in clustering and classification. For categorical data clustering, the existing *k*-mode algorithm is not capable in selection of dimensions. Therefore we expect *k*-mode algorithm may not work very well for high-dimensional data sets. However, the dimensions weighting procedure is incorporated in the proposed algorithm, we expect that the SKM-SNP method would be a useful tool for SNP marker selection in the data sets discussed.

In addition, we choose chromosome 2, which has the highest accuracy as an example to demonstrate the weights of SNPs for SNP markers detection. Fig. 2 shows the accuracies obtained when we increase γ value from zero. We can see from this figure that SKM-SNP can get a reasonably good accuracy of 92.98% when γ is between 700 and 1000, 1.5% higher than the traditional *K*-mode algorithm ($\gamma = 0$).

As there are thirty thousands of SNPs, we further filter out a portion of them to do a detailed analysis. We first filter out those SNPs whose initial modes or final modes are the same after we run the SKM-SNP program once. The remaining number of SNPs is 1887, 5.8% of the original chromosome 2 data set. With this much smaller data set, we can still obtain a satisfactory accuracy

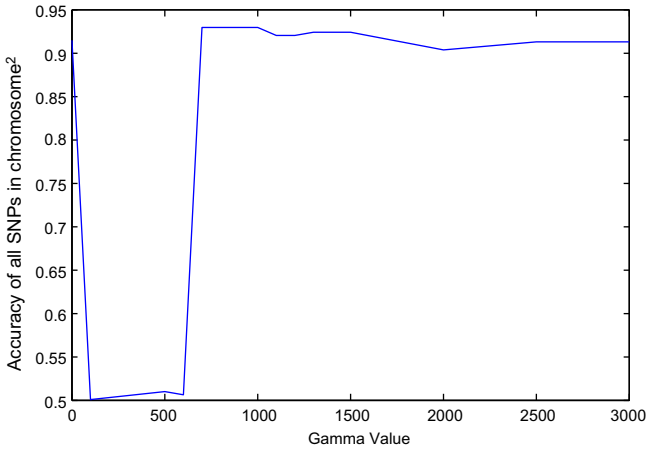

Fig. 2. Relationship between γ and accuracy of all SNPs in chromosome 2 using SKM-SNP.

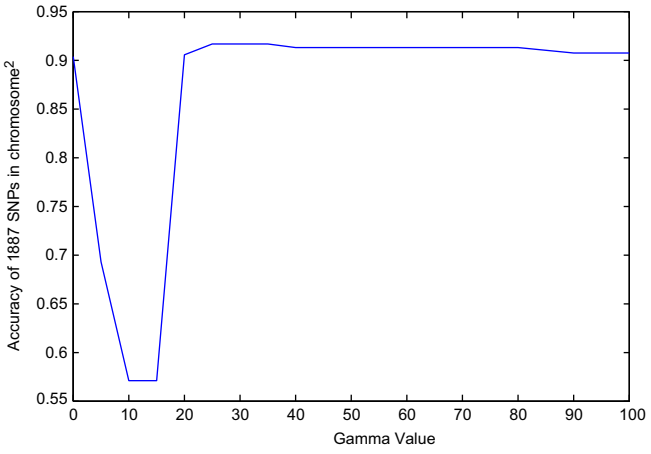

Fig. 3. Relationship between γ and accuracy of selected 1887 SNPs in chromosome 2 using SKM-SNP.

Table 5
The top ten weights of SNPs in the case group and their genotypes distributions for Parkinson dataset.

SNPs	Case					Control				
	Weight	AA	Aa	aa	M	Weight	AA	Aa	aa	M
rs7565244	15.55	59.3	35.2	5.5	0.0	4.70	45.8	46.9	7.3	0.0
rs2017444	12.74	56.4	33.3	9.6	0.7	4.40	43.6	43.9	12.5	0.0
rs6736992	12.32	56.3	34.1	9.6	0.0	4.25	43.2	43.9	12.9	0.0
rs353111	11.91	55.6	33.0	8.1	3.3	7.01	44.3	49.1	5.9	0.7
rs6547378	11.15	53.3	37.1	9.6	0.0	5.37	43.9	46.1	10.0	0.0
rs7605630	10.78	54.4	37.8	7.8	0.0	4.86	44.7	46.1	9.2	0.0
rs3731714	10.78	55.9	34.8	9.3	0.0	5.19	42.8	47.2	10.0	0.0
rs935415	10.78	54.4	39.3	6.3	0.0	6.56	41.7	49.1	9.2	0.0
rs1427682	10.43	56.3	37.4	6.3	0.0	4.40	43.9	46.1	10.0	0.0
rs9653591	10.43	54.1	38.5	7.4	0.0	7.50	42.8	50.2	7.0	0.0

Table 6
The top ten weights of SNPs in the control group and their genotypes distributions for Parkinson dataset.

SNPs	Control					Case				
	Weight	AA	Aa	aa	M	Weight	AA	Aa	aa	M
rs2306676	14.12	57.2	33.6	9.2	0.0	6.12	45.6	48.5	5.9	0.0
rs10497018	13.21	56.1	34.3	8.5	1.1	5.92	41.9	47.8	9.2	1.1
rs1113958	12.78	55.7	35.8	8.5	0.0	6.76	41.9	48.4	9.3	0.4
rs1364658	12.78	55.7	35.4	8.9	0.0	7.23	39.6	49.6	10.4	0.4
rs6761958	12.36	55.4	37.6	7.0	0.0	5.53	45.9	47.0	6.7	0.4
rs4849987	12.36	55.4	38.0	6.6	0.0	6.12	45.9	47.8	6.3	0.0
rs12987286	12.36	55.4	35.8	8.8	0.0	7.23	39.6	49.6	10.8	0.0
rs2580823	11.95	54.6	38.0	7.4	0.0	6.32	45.9	47.8	6.3	0.0
rs6708081	11.56	55.7	38.7	5.6	0.0	5.92	46.7	48.1	5.2	0.0
rs867014	11.56	56.5	36.5	7.0	0.0	6.12	50.0	44.1	5.9	0.0

of 91.68% when γ is between 25 and 35, 1.3% higher compared with *K*-mode of 90.39%, see Fig. 3. Then we apply SKM-SNP program again for this data set. In Table 5 and Table 6, we show the SNPs whose weights (the magnitude 10^{-4}) ranking the top ten in the case and control groups. Their corresponding percentages of genotype distributions are also shown in the table where "A" and "a" represent the major and minor alleles. The column under "M" refers to the missing percentages of genotypes in the groups. We see from the two tables that the SNPs of the top ten weights are different in the two groups. These results indicate their subspace structure of two clusters are different. Based on the weights, we can identify some relevant SNPs associated with case and control groups in a data set. Therefore we can further study these SNPs for disease-related genetic analysis.

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 [26] 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 SKM-SNP. These results not only validate the efficiency of our program, but also demonstrate that SNPs selected by our program associate to control/case populations. Biologists can further investigate on the above important SNPs to determine their genetic functions and study how they are relevant to the Parkinson Disease.

4. Conclusions

In this paper, we have developed SKM-SNP, a new SNP markers detection method to identify the subset of SNPs. SKM-SNP utilizes subspace categorical clustering techniques by adding a weight value to each SNP and includes weight entropy in the objective function so that each subspace cluster can be formed by several relevant SNPs that are similar within a cluster and dissimilar among clusters. This program has been efficiently and successfully used in real Schizophrenia and Parkinson Disease SNP data sets.

Acknowledgments

This work was financially supported by Research Grant Council 201508 and HKBU FRGs. We thank the participants and the submitters for depositing samples at the NINDS Neurogenetics repository. The samples for this study are derived from the NINDS Neurogenetics repository at Coriell Cell Repositories. Access to the samples and to these data are available from the website: http://ccr.coriell.org/Sections/BrowseCatalog/DiseaseDetail.aspx?PgId=403&omim=PAR40000&coll=.

References

[1] Parsons L, Haque E, Liu H. Subspace clustering for high dimensional data: a review. ACM SIGKDD Explor Newslett 2004;6:90–105.
[2] Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data for data mining application. Proc ACM SIGMOD 1998:94–105.
[3] Aggarwal C, Procopiuc C, Wolf J, Yu P, Park J. Fast algorithms for projected clustering. Proc ACM SIGMOD 1999:61–72.
[4] Aggarwal C, Yu P. Finding generalized projected clusters in high dimensional spaces. Proc ACM SIGMOD 2000:70–81.
[5] Chakrabarti K, Mehrotra S. Local dimensionality reduction: a new approach to indexing high dimensional spaces. Proc. of 26th Intel. Conf. on very large data bases 2000:89–100.
[6] Procopiuc C, Jones M, Agarwal P, Murali T. A monte carlo algorithm for fast projective clustering. Proc ACM SIGMOD 2002:418–27.
[7] Yip KY, Cheung DW, Ng MK. A practical projected clustering algorithm. IEEE Trans Knowl Data Eng 2004;16:1387–97.
[8] Yip KY, Cheung DW, Ng MK. On discovery of extremely low-dimensional clusters using semi-supervised projected clustering. Proc. of the 21st international conference on data engineering 2005:329–40.

[9] Desarbo W, Carroll J, Clark L, Green P. Synthesized clustering: a method for amalgamating clustering bases with differential weighting variables. Psychometrika 1984;49:57–78.
[10] Milligan G. A validation study of a variable weighting algorithm for cluster analysis. J Classif 1989;6:53–71.
[11] Modha D, Spangler W. Feature weighting in *k*-means clustering. Mach Learn 2003;52:217–37.
[12] Chan Y, Ching W, Ng MK, Huang ZX. An optimization algorithm for clustering using weighted dissimilarity measures. Pattern Recogn 2004;37:943–52.
[13] Frigui H, Nasraoui O. Unsupervised learning of prototypes and attribute weights. Pattern Recogn 2004;37:567–81.
[14] Domeniconi C, Papadopoulos D, Gunopulos D, Ma S. Subspace clustering of high dimensional data. Proc. of SIAM international conference on data mining 2004:517–21.
[15] Friedman J, Meulman J. Clustering objects on subsets of attributes. J R Statist Soc B 2004;66:815–49.
[16] Huang J, Ng MK, Rong H, Li Z. Automated variable weighting in *k*-means type clustering. IEEE Trans Pattern Anal Mach Intell 2005;27:1–12.
[17] Jing LP, Ng MK, Xu J, Huang ZX. Subspace clustering of text documents with feature weighting *k*-means algorithm. Proc. of the 9th Pacific-Asia conference on knowledge discovery and data mining 2005:802–12.
[18] Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, et al. Linkage disequilibrium in the human genome. Nature 2001;411:199–204.
[19] Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. Nat Genet 2001;29:229–32.
[20] Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Genova GD, et al. Haplotype tagging for the identification of common disease genes. Nat Genet 2001;29:233–7.
[21] Zollner S, Haeseler AV. A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. Am J Hum Genet 2000;66:615–28.
[22] Jing LP, Ng MK, Huang ZX. An entropy weighting *k*-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Trans Knowl Data Eng 2007;19:1026–41.
[23] HapMap Project, http://www.hapmap.org/2008.
[24] Ao SI, Yip K, Ng MK, Cheung D, Fong PY, Melhado I, et al. CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs. Bioinformatics 2004;21:1735–6.
[25] Huang ZX. Extensions to the *k*-means algorithm for clustering large data sets with categorical values. Data Min Knowl Discov 1998;2:283–304.
[26] Hicks AA, Petursson H, Jonsson T, Stefansson H, Johannsdottir HS, Sainz J, et al. A susceptibility gene for late-onset idiopathic Parkinson's disease. Ann Neurol 2002;52:549–55.