Subdimension-based similarity measure for DNA microarray data clustering

Benson S. Y. Lam[1] and Hong Yan[1,2]
[1]*Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong*
[2]*School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia*
(Received 29 March 2006; published 9 October 2006)

Microarray data analysis is useful for understanding biological processes. A number of clustering algorithms have been used to achieve this task. However, the performance of these methods can be significantly degraded due to the presence of nonsignificant conditions. In this paper, we propose a robust clustering algorithm based on a similarity measure. The key concept of the proposed similarity measure is to measure the similarity between two data points by their subdimensions. For example, assume that $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ are ten-dimensional data vectors. The data point $\mathbf{x}_3$ is said to be closer to $\mathbf{x}_1$ than $\mathbf{x}_2$ if more than half of the dimensions of $\mathbf{x}_1$ and $\mathbf{x}_3$ are closer to $\mathbf{x}_1$ than $\mathbf{x}_2$. Thus, if two patterns are very similar except for a small amount of features, this measure will preserve the similarity. We have performed eight experiments to test the robustness of the proposed method, including three synthetic data sets, three real world data sets, and two microarray data sets. We also have compared the proposed method with four different clustering algorithms. Experimental results show that the proposed method yields better results than existing clustering algorithms.

DOI: 10.1103/PhysRevE.74.041906 PACS number(s): 87.10.+e

I. INTRODUCTION

Microarray data analysis is useful for understanding biological processes from gene expressions. Genes having similar expression patterns imply that they are coregulated and may have a common function. For example, Chu *et al.* studied the sporulation, which consists of meiosis and spore morphogenesis [1]. Cho *et al.* used similar technique to study the mitotic cell cycle [2]. Agrawal analyzed the cancers data set by measuring the similarity among genes [3]. A microarray data set can be viewed as an $n \times d$ matrix, where n is the number of genes and d is the number of conditions (features). Usually, n is a very large number, in the order of thousands [2,4]. If the microarray data is analyzed manually, the procedure will be labor intensive. This leads to the problem of automated clustering by the computer.

A clustering algorithm can be used to group objects, which have similar patterns, into the same class. There are many different clustering algorithms. They include partition-based clustering algorithm such as fuzzy c-means (FCM) algorithm [5], hierarchical-based method such as hierarchical clustering algorithm with complete link (HC) (Ref. [6]), distribution-based method such as the Gaussian mixture model (GMM) (Ref. [7]) and a nonparametric method [8]. They have been adopted to analyze gene expression data [9–12]. All of these methods have the same idea in that they make use of all the conditions in similarity or distance measures. A comprehensive review can be found in the papers by Jiang *et al.* [4], Jain *et al.* [13], and Kaufman and Rousseeuw [14]. For simplicity, we call these methods traditional clustering algorithms. There is also another kind of clustering method to analyze gene expression, subspace clustering, which only takes some of the conditions into account. For example, if there is a data set with 20 conditions, the subspace clustering may only take the first ten conditions for data clustering, which is different from traditional clustering algorithm taking all 20 conditions. The motivation of subspace clustering method is that the data set may contain nonsignificant conditions that can influence the clustering results. The subspace clustering method can be roughly separated into two classes. They are top-down [e.g., PROCLUS (Ref. [15]) and HARP (Ref. [16])] and bottom-up [e.g., CLIQUE (Ref. [17]) and ENCLUS (Ref. [18])] based methods. A comprehensive review can be found in the paper written by Parsons *et al.* [19]. Recently, biclustering algorithms also have drawn great attention and they can be viewed as a form of subspace clustering algorithm as well [20]. In this review, Bergmann *et al.* introduce a biclustering method to extract a subset of genes for data analysis [21]. Most of these methods employ the same philosophy to choose the conditions used for clustering. The condition is selected if the density of the group in the data is large enough. However, the largeness of the density needs a prior knowledge of the data and this cannot always be determined automatically. It employs user-defined parameters and the result may highly depend on such parameters. Recently, Yip *et al.* introduced a subspace clustering method called HARP, which outperforms many existing subspace clustering algorithms. However, Yip *et al.* also reported that their subspace clustering algorithm may not be able to yield more accurate results than the traditional clustering algorithms [16]. Thus, the problem of selecting the conditions is still a critical one.

In this paper, we introduce a proposed clustering algorithm to handle this problem. The key concept of the proposed algorithm is to measure the similarity between two objects in several subdimensions. Here, we introduce a new concept called subdimension. A data set is separated into p parts, which are not disjoint. Each part has the same number of input samples, namely genes in microarray data analysis, as the original data, but a smaller number of dimensions, namely conditions. In our formulation, each part has the same number of dimensions and we call each of these dimensions a subdimension. If more than half the conditions between two objects belong to the same group, these two objects are said to belong to the same group. The idea of this similarity measure has been demonstrated in our conference presentation [22]. Experimental results show that the cluster-

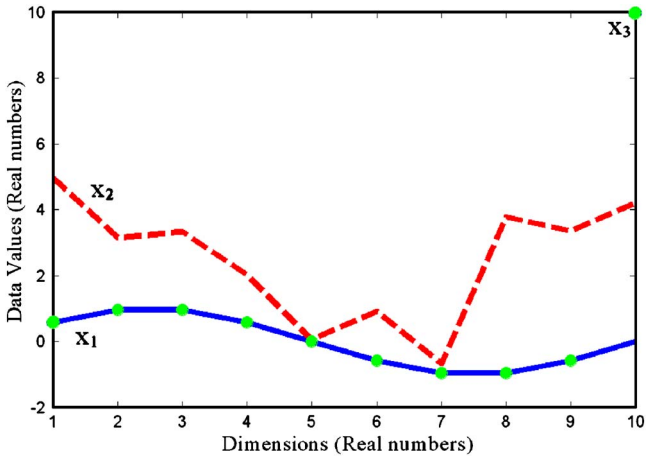

FIG. 1. (Color online) Illustration of the effect of nonsignificant condition to the similarity measure.

ing algorithm using this method gives better results than other methods.

The organization of this paper is as follows. In Sec. II, we briefly review the fuzzy c-means clustering algorithm. Then, we introduce the proposed similarity measure in Sec. III. After that, we introduce the proposed clustering method in Sec. IV. Experimental results are given in Sec. V, and discussions and conclusions are given in Sec. VI.

II. EXISTING CLUSTERING METHOD

In this section, we review the fuzzy c-means (FCM) clustering algorithm [5]. Given a data set $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbf{R}^d$, $\mathbf{x}_i=[x_i^1,\ldots x_i^d]^T$ and d is the dimension of each data vector. In the fuzzy c-means clustering algorithm, the following objective function is minimized:

$$J(U,V;\mathbf{X}) = \sum_{i=1}^{n}\sum_{k=1}^{c} \mu_{ik}^m |\mathbf{x}_i - \mathbf{v}_k|^2, \tag{1}$$

where $|\cdot|$ refers to the 1_2 norm, m is a constant that is usually set to 2, c is the total number of groups, and the membership values satisfy the following two conditions:

$$0 < \mu_{ik} < 1, \quad 0 \leqslant i \leqslant n \text{ and } 0 \leqslant k \leqslant c, \tag{2}$$

$$\sum_{k=1}^{c} \mu_{ik} = 1, \quad 0 \leqslant i \leqslant n. \tag{3}$$

Setting the derivative of $J(U,V;\mathbf{X})$ with respect to the unknowns $\mathbf{v}_k$ and μ_{ik} equal to zero, we obtain the following update equations:

$$\mu_{ik} = \frac{|\mathbf{x}_i - \mathbf{v}_k|^{-2/(m-1)}}{\sum_{k=1}^{c} |\mathbf{x}_i - \mathbf{v}_k|^{-2/(m-1)}}, \tag{4}$$

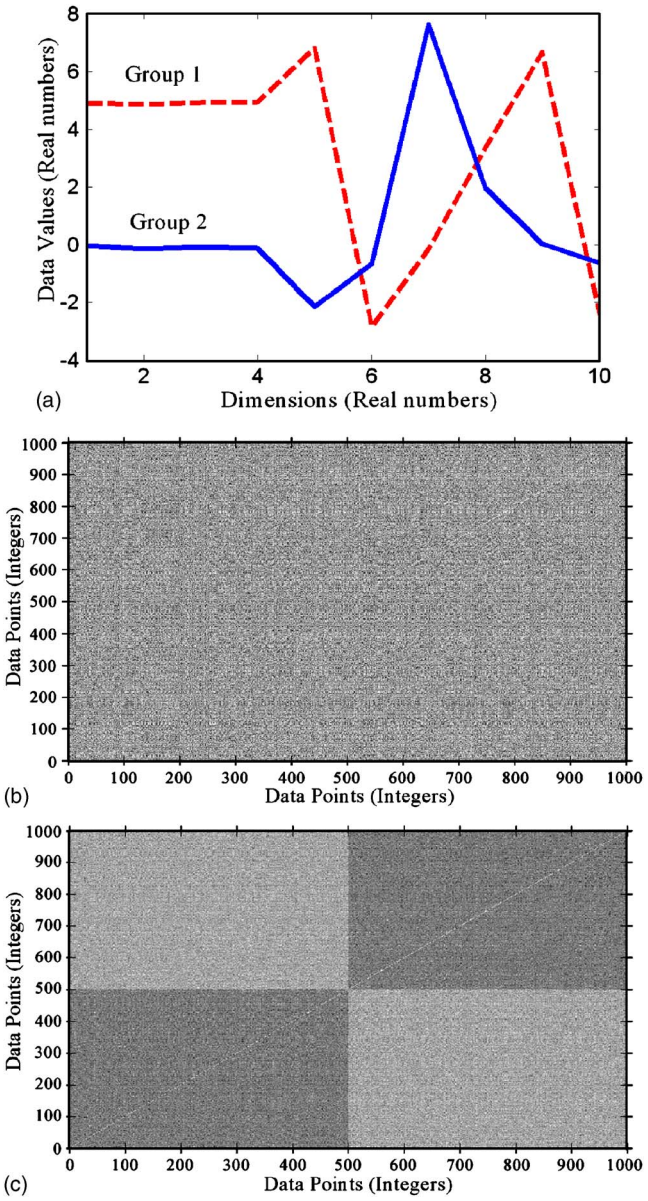

FIG. 2. (Color online) The mean values of the synthetic data set 1 and its partition matrices.

$$\mathbf{v}_k = \frac{\sum_{i=1}^{n} \mu_{ik}^m \mathbf{x}_i}{\sum_{i=1}^{n} \mu_{ik}^m}, \tag{5}$$

where $\mathbf{v}_k$ represents the kth cluster center. The clustering algorithm consists of two steps, updating the membership values from Eq. (4) and finding the center of each cluster from Eq. (5). The k-means clustering algorithm is a special case of the FCM algorithm. Although we use the FCM algorithm in this paper, our proposed method is applicable to other clustering techniques.

TABLE I. Summary of the proposed method.

Algorithm:	
1.	The data set is sorted according to its dimensions. For $s=2$ to 3, perform steps 2 to 4:
2.	The data set $\mathbf{X}$ is divided into several subdimensional sets, each with dimension s.
3.	The FCM algorithm is applied to each subdimensional set with the input parameter c varied from $c=2$ to $c=c_{\text{total}}$. Then, in each subdimensional set, the clustering results with largest $I_m(c)$ value will be considered.
4.	By making use of these clustering results, the matrix $\mathbf{P}_s$ can be obtained.
5.	The average partition matrix $\mathbf{P}=(\mathbf{P}_2+\mathbf{P}_3)/2$ is computed. By taking $\mathbf{P}$ as a distance matrix, the hierarchical clustering algorithm is applied to produce the final result.

III. PROBLEM STATEMENT

In this section, we indicate the problem of existing similarity. Then, we introduce the idea of the proposed method. We now consider three ten-dimension data points $\mathbf{x}_1=[0.5878, 0.9511, 0.9511, 0.5878, 0.0000, -0.5878, -0.9511, -0.9511, -0.5878, -0.0000]^{\mathrm{T}}$, $\mathbf{x}_2=[4.9550, 3.1490, 3.3364, 2.0429, 0.0620, 0.9109, -0.6682, 3.7762, 3.3466, 4.2073]^{\mathrm{T}}$ and $\mathbf{x}_3=[0.5878, 0.9511, 0.9511, 0.5878, 0.0000, -0.5878, -0.9511, -0.9511, -0.5878, 10.0000]^{\mathrm{T}}$. These three patterns are shown in Fig. 1. The one at the bottom is $\mathbf{x}_1$ while the one above $\mathbf{x}_1$ is $\mathbf{x}_2$. The third vector $\mathbf{x}_3$ (marked by ·) is almost the same as $\mathbf{x}_1$ except the tenth dimension, which is far away from $\mathbf{x}_1$. The variation in the tenth dimension of $\mathbf{x}_3$ is due to the presence of nonsignificant conditions. $\mathbf{x}_1$ and $\mathbf{x}_3$ should belong to the same group. However, if we measure their similarity using the 1_2 norm, we will find that $\mathbf{x}_1$ and $\mathbf{x}_2$ are more likely to be in the same group. The 1_2 norm distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ is $\|\mathbf{x}_1-\mathbf{x}_2\|=9.4641$ while the 1_2 norm distance between $\mathbf{x}_1$ and $\mathbf{x}_3$ is $\|\mathbf{x}_1-\mathbf{x}_3\|=10$. Because of this, a clustering algorithm may produce unreliable results.

Now, we introduce the proposed method by reformulating the distance measure as follows. Let $\mathbf{X}=\mathbf{A}_1\times\mathbf{A}_2\times\cdots\times\mathbf{A}_{d-1}\times\mathbf{A}_d$, where $\mathbf{A}_j$ represents the jth dimension of the data with $1\leqslant j\leqslant d$. We redefine the dimension of $\mathbf{X}$ as follows: $\mathbf{X}=\mathbf{B}_1\times\mathbf{B}_2\times\cdots\times\mathbf{B}_{p-1}\times\mathbf{B}_p$, where $p\leqslant d$, and $\mathbf{B}_j=\mathbf{A}_{j1}\times\mathbf{A}_{j2}\times\cdots\times\mathbf{A}_{j(s-1)}\times\mathbf{A}_{js}$ where $1\leqslant j\leqslant p$, $1\leqslant j_1, j_2, \ldots, j_s\leqslant d$ and s is the number of conditions in each subdimension and $s\leqslant d$. Here the original data set $\mathbf{X}$ is represented by d nonoverlapping subsets $\mathbf{A}_1$, $\mathbf{A}_2$, … and $\mathbf{A}_d$, where $\mathbf{A}_j$ simply represents the set of data values of all genes along dimension (condition) j. To work with subdimensions, we decompose $\mathbf{X}$ into many overlapping subsets $\mathbf{B}_1$, $\mathbf{B}_2$, …and $\mathbf{B}_p$, where $\mathbf{B}_j$ is the union of s subsets $\mathbf{A}_{j1}$, $\mathbf{A}_{j2}$, … and $\mathbf{A}_{js}$. An input data vector $\mathbf{x}_i$ is now decomposed into p vectors, $\mathbf{x}_{i(Bj)}$. In the subdimension-based similarity measure, $\mathbf{x}_a$ is closer to $\mathbf{x}_b$ than to $\mathbf{x}_c$ if

$$\begin{aligned}&\mathrm{Card}(\{\mathrm{j}:\|\mathbf{x}_{a(B_j)}-\mathbf{x}_{b(B_j)}\|\\&\quad<\|\mathbf{x}_{a(B_j)}-\mathbf{x}_{c(B_j)}\|\})\\&\quad>\mathrm{Card}(\{\mathrm{j}:\|\mathbf{x}_{a(B_j)}-\mathbf{x}_{b(B_j)}\|\geqslant\|\mathbf{x}_{a(B_j)}-\mathbf{x}_{c(B_j)}\|\}),\end{aligned}$$

where card($\mathbf{S}$) refers to the cardinality (or the number of elements) of the set $\mathbf{S}$. The above equation defines a new similarity measure between two objects, under which $\mathbf{x}_a$ is closer to $\mathbf{x}_b$ than to $\mathbf{x}_c$ if there $\mathbf{x}_{a(Bj)}$ is closer to $\mathbf{x}_{b(Bj)}$ than to $\mathbf{x}_{c(Bj)}$ in more subdimensions. The above classification criterion can also be written as

$$\frac{\mathrm{Card}(\{\mathrm{j}:\|\mathbf{x}_{a(B_j)}-\mathbf{x}_{b(B_j)}\|<\|\mathbf{x}_{a(B_j)}-\mathbf{x}_{c(B_j)}\|\})}{p}>\frac{1}{2},\qquad(6)$$

since

$$\begin{aligned}&\mathrm{Card}(\{\mathrm{j}:\|\mathbf{x}_{a(B_j)}-\mathbf{x}_{b(B_j)}\|\\&\quad<\|\mathbf{x}_{a(B_j)}-\mathbf{x}_{c(B_j)}\|\})+\mathrm{Card}(\{\mathrm{j}:\|\mathbf{x}_{a(B_j)}-\mathbf{x}_{b(B_j)}\|\\&\quad\geqslant\|\mathbf{x}_{a(B_j)}-\mathbf{x}_{c(B_j)}\|\})\end{aligned}$$

equals the number of subdimension vector sets$=p$.

This means that $\mathbf{x}_a$ is classified into the class of $\mathbf{x}_b$ if for more than 50% of subdimensions $\mathbf{x}_a$ is closer to $\mathbf{x}_b$ than to $\mathbf{x}_c$; otherwise, it is classified into the class of $\mathbf{x}_c$. To have a more reliable classification, we can require this ratio to be greater than 50%, for example, we can set it at 60%. However, in doing so, we may have to reject $\mathbf{x}_a$, that is, we cannot make a decision with enough confidence, if the ratio is between 50% and 60%. In practical applications, we can adjust this ratio to trade off between false positive and rejection rates in a pattern classification system.

Now, we apply this concept to the three patterns $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$. We first decompose these data vectors into subdimensional ones. For example, we can decompose $\mathbf{x}_1$ into eight subdimensional vectors, each of which has three dimensions, $[x_1^1, x_1^2, x_1^3]^{\mathrm{T}}, [x_1^2, x_1^3, x_1^4]^{\mathrm{T}}, \ldots, [x_1^8, x_1^9, x_1^{10}]^T$. Then we measure the similarity between all corresponding subdimensional vectors using the 1_2 norm. After the calculation, we say objects $\mathbf{x}_1$ and $\mathbf{x}_3$ are closer than $\mathbf{x}_1$ and $\mathbf{x}_2$ if more subdimensional vectors between $\mathbf{x}_1$ and $\mathbf{x}_3$ suggest they are closer. Obviously, for the three patterns $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$, all the subdimensional vectors between $\mathbf{x}_1$ and $\mathbf{x}_3$ give a smaller value than $\mathbf{x}_1$ and $\mathbf{x}_2$ except the last one, which contains the eighth, ninth, and tenth dimensions. Thus, we say that $\mathbf{x}_1$ and $\mathbf{x}_3$ are closer than $\mathbf{x}_1$ and $\mathbf{x}_2$.

IV. PROPOSED CLUSTERING ALGORITHM

In this section, we introduce the proposed clustering method. There are five steps in the proposed method, with the number of group c_{total} given by the user. These steps are given as follows.

Step 1: Let $\mathbf{X}=[x_i^1, x_i^2, \ldots, x_i^{d-1}, x_i^d]^T$ (for $1\leqslant i\leqslant n$) be the original data set sorted in ascending order with respect to the standard derivation of each dimension. This data set is divided into lower and upper halves: $\mathbf{G}^1=[x_i^1, x_i^2, \ldots, x_i^{d/2}]^T$ and $\mathbf{G}^2=[x_i^d, x_i^{d-1}, \ldots, x_i^{d/2+1}]^T$. Then, $\mathbf{G}^1$ and $\mathbf{G}^2$ are mixed in an

TABLE II. Clustering results for synthetic data set 1.

Groups	GMM	FCM	HC	HARP	Proposed method
1	297	255	421	478	500
2	206	249	99	492	500
Total (Max)	503	504	520	970	1000

alternative manner: $\mathbf{G}_s=[x_i^d, x_i^{d/2}x_i^{d-1}, x_i^{d/2-1}, \ldots, x_i^{d/2+1}, x_i^1]^T$. If the difference of standard derivation between two consecutive dimensions of $\mathbf{G}_s$ is larger than a threshold θ (which is taken as 2 in all the experiments), we will take $\mathbf{G}_s=\mathbf{X}$. We take $\mathbf{G}_s$ as a sorted data set in the second step of the method.

Step 2: We divide the data sets $\mathbf{X}\in\mathbf{R}^d$ into several subdimensional sets which have smaller dimensions: $\mathbf{X}_{(j)}=\{\mathbf{x}_{i(j)}\}$ where $\mathbf{x}_{i(j)}\in\mathbf{R}^s$, j is the jth subdimension of the data $1\leqslant j\leqslant p$ and $s\leqslant d$. In this paper, $s=2$ and $s=3$ are adopted. For example, $\mathbf{x}_i$ has ten dimensions, its $\mathbf{R}^2$ and $\mathbf{R}^3$ subdimensional sets will be $[x_i^1,x_i^2]^T,[x_i^2,x_i^3]^T,\ldots,[x_i^9,x_i^{10}]^T$, and $[x_i^1,x_i^2,x_i^3]^T,[x_i^2,x_i^3,x_i^4]^T,\ldots,[x_i^8,x_i^9,x_i^{10}]^T$, respectively.

Step 3: We apply the FCM algorithm to each of the subdimensional sets with the input parameter from $c=2$ to $c=c_{total}$ and evaluate the clustering results. This is equivalent to conducting cluster validity with each subdimensional set. For each subdimensional set, only the clustering result with the largest $I_m(c)$ is considered. The original I index was proposed by Maulik and Bandyopadhyay [23]. $I_m(c)$ is a modified version of the I index. The equation for $I_m(c)$ is given as follows:

$$I_m(c)=\left(\frac{1}{c}\times\frac{E_1}{E_c}\times D_c\right)^Q, \tag{7}$$

where the power Q is used to control the contrast between different cluster configurations. In this paper, we take $Q=1$. E_c and D_c are defined as

$$E_c=\sum_{i=1}^{n}\sum_{k=1}^{c}\delta_{ik}\|\mathbf{x}_i-\mathbf{v}_k\|^2, \tag{8}$$

$$D_c=\max_{i,j=1}^{c}\|\mathbf{v}_i-\mathbf{v}_j\|, \tag{9}$$

where $\mathbf{v}_k$ is the prototype of class k generated by the clustering algorithm. δ_{ik} is a binary variable. If $\mathbf{x}_i$ is a data point closest to $\mathbf{v}_k$, $\delta_{ik}=1$. Otherwise, $\delta_{ik}=0$. The difference between the proposed modified I index and the original I index is that the function E_c has a square power in the modified I index while no square power in the original I index.

Step 4: Based on the results in step 3, we are able to get a partition matrix $\mathbf{P}^j{}_s$. This partition matrix is a binary matrix. If the subdimensional points $\mathbf{x}_{p(j)}$ and $\mathbf{x}_{q(j)}$ belong to the same group, $\mathbf{P}^j{}_s(p,q)=1$. Otherwise, $\mathbf{P}^j{}_s(p,q)=0$. Now, we define the variable $\mathbf{P}_s$ as the mean of these partition matrices,

$$\mathbf{P}_s=1-\frac{1}{d-s}\sum_{j=1}^{d-s+1}\mathbf{P}^j_s. \tag{10}$$

As we adopt $s=2$ and $s=3$ for subdimensional sets, there are totally two variables $\mathbf{P}_2$ and $\mathbf{P}_3$. The average partition matrix $\mathbf{P}$ is defined as $\mathbf{P}=(\mathbf{P}_2+\mathbf{P}_3)/2$. Thus, if there are two data points $\mathbf{x}_p$ and $\mathbf{x}_q$ and their conditions are very similar to each other, the value $\mathbf{P}(p,q)$ will be small.

Step 5: We consider the average partition matrix $\mathbf{P}$ as a similarity matrix in hierarchical clustering algorithm. We adopt the complete link method in the hierarchical clustering algorithm to get the clustering result. Table I summarized these five steps.

V. EXPERIMENTAL RESULTS

In this section, we conduct eight experiments to test the robustness of the proposed method. Four different clustering algorithms are chosen to compare the performance of the proposed method. They are the GMM, the FCM algorithm, the hierarchical clustering method with complete link (HC), and the HARP algorithm, which is a subspace clustering method.

Each algorithm except HC will be performed ten times to each real world data set. In all the real world data sets we adopted in this paper, data samples have class labels. We make use of these labels for evaluating the algorithms. For example, after applying the FCM algorithm, we get c partitions $C_1,\ldots C_c$. In each original group, we find the number of

TABLE III. Clustering results for synthetic data set 2.

Groups	GMM	FCM	HC	HARP	Proposed method
1	200	236	109	402	494
2	209	192	253	498	497
3	134	113	155	205	494
Total (Max)	534	541	517	1105	1485

TABLE IV. Clustering results for synthetic data set 3.

Groups	GMM	FCM	HC	HARP	Proposed method
1	200	198	117	471	500
2	209	173	323	470	500
3	134	155	66	295	500
Total (Max)	534	526	506	1236	1500

objects correctly recognized in $C_1, \ldots C_c$ so that the sum of these numbers reaches maximum. Based on the number of correctly classified objects, we will compare the algorithms in three ways. They are the maximum, mean, and standard derivation of the number of correctly recognized objects in the ten runs. Also, we will show the number of correctly recognized objects in each group for the best clustering result among ten runs.

A. Synthetic data set

We perform three experiments based on three synthetic data sets. In each of these data sets, some of the dimensions have very large variance compared with other dimensions. This situation is similar to the one we introduced in Sec. III. This can make the conventional distance measure error prone.

Synthetic data set 1: Now, we consider a ten-dimensional data set with two groups. The first four dimensions are the same and they are generated by two normally distributed functions $N(0,1)$ and $N(5,1)$. Each of them consists of 500 points. The last six dimensions are generated by a normally distributed function which is $N(0,100)$ with 1000 points. Thus, the data matrix has a size of 1000×10. In this example, there are six components that are nonsignificant conditions for the two groups, while there are four components that contain information of the two groups. Figure 2(a) shows mean values of the two groups. The two groups can be clearly separated in terms of the first four dimensions but not in terms of the last six dimensions. If we apply traditional clustering algorithms such as the FCM algorithm to the first four dimensions of the data set, 100% accuracy will be obtained. However, the insertion of nonsignificant information from the extra six dimensions degrades the clustering result significantly. The clustering results for this data set are given in Table II. We can see that the GMM, FCM, and HC clustering algorithms have only half accuracy. For the HARP algorithm, it has a much higher accuracy than the nonsubspace clustering algorithm. As the higher dimensions are pruned in HARP, the subspace clustering algorithm has a better performance than the nonsubspace clustering algorithm. For the proposed method, it has 100% accuracy. One may think that the clustering result of the proposed method may be unreliable since the total number of dimensions that do not contain the information of the two groups are more than the total number of dimensions that contain the information of the two groups. However, in the proposed method, we divide the data set into several subdimensional sets and conduct data clustering for each of them. In the third and fourth steps of the proposed method, we found that only the first four dimensions shared the same clustering results. However, the last six dimensions produced very different clustering results in each subdimensional set. Thus, the variable $\mathbf{P}_2$ for the last six dimensions is just a random matrix and does not contribute much to the average partition matrix $\mathbf{P}$. The matrix $\mathbf{P}_2$ is given in Fig. 2(b). The darker pixels represent larger values in the partition matrix and vice versa. Figure 2(c) shows the average partition matrix $\mathbf{P}$ for synthetic data 1 after permuting the matrix $\mathbf{P}$ so that the first group consists of the first 500 elements. We can clearly see that the proposed method can detect 500 points in one group and another 500 points in another group.

Synthetic data set 2: Now, we consider a 24-dimensional data set with three groups. The first four dimensions are generated by three normally distributed functions $N(0,1)$, $N(5,1)$, and $N(10,1)$. Each of them consists of 500 points. The last 21 dimensions are generated by the normally distributed function with different variance from 5 in the fifth dimension to 100 in the 24th dimension, and the variances between consecutive two dimensions have a difference of 5. If we put the variances of the fifth to 24th dimensions in a

TABLE V. Clustering results for the iris data.

Groups	GMM	FCM	HC	HARP	Proposed method
1	50	50	50	49	50
2	40	47	49	38	48
3	49	37	27	12	43
Total (Max)	139	134	126	99	141
Mean	124.9	134	/	99	141
Std	19.1512	0	/	0	0

TABLE VI. Clustering results for the wine data.

Groups	GMM	FCM	HC	HARP	Proposed method
1	57	50	56	55	58
2	48	45	43	40	62
3	24	27	21	44	45
Total (Max)	129	122	120	139	165
Mean	124.7	122	/	139	165
Std	3.0569	0	/	0	0

vector, it will become $[5,10,15,20,\ldots,100]^T$. Thus, the data matrix has a size of 1500×24. This synthetic data set is different from the previous one. The elements of the synthetic data set 1 that are nonsignificant conditions have exactly the same variances. However, in this data set, the variances are not the same but monotonic increasing. The clustering results for this data set are given in Table III. Again, we can see that the GMM, FCM, and HC algorithms have only half accuracy. For the HARP algorithm, the result is not as good as the one given in synthetic data 1 (Table II). Its accuracy is reduced to around 75%. This shows that the HARP algorithm could not prune the noise dimensions well if they are very different. For the proposed method, it has 99% accuracy. In this experiment, we can see that the proposed technique is able to yield more accurate results than conventional methods.

Synthetic data set 3: Now, we consider a ten-dimensional data set with three groups. The first four dimensions are generated by three normally distributed functions $N(0,1)$, $N(5,1)$, and $N(10,1)$. Each of them consists of 500 points. The last six dimensions are generated by the normally distributed function with two different variances. The fifth to eighth dimensions are generated by normal distribution function with variance 10 while the ninth to tenth dimensions are generated by normal distribution function with variance 10 000. Thus, the data matrix has a size of 1500×10. The clustering results for this data set are given in Table IV. Similar to synthetic data set 2, the GMM, FCM, and HC clustering algorithms have only half accuracy. The HARP algorithm has 83% accuracy. The proposed method has 100% accuracy. In these experiments, we can see that the proposed method is able to yield more accurate results although both subspace and traditional clustering algorithms cannot.

B. Real world data

In this section, we will perform three other tests on three real world data sets. They are iris data, wine data, and Wisconsin diagnostic breast cancer (wdbc) data. These data sets can be found on the website [24]. The iris data set contains three groups and four features. Each group contains 50 objects. The wine data set has 178 objects with 13 features. This data set contains three groups. The wdbc data set has 576 objects with 30 features. It contains two groups. The clustering results for these three data sets are given in Tables V–VII, respectively. Except the wine data, the HARP algorithm yielded better results than other three traditional clustering algorithms. For the traditional algorithms, the FCM algorithm has a better performance. The proposed method yields the largest numbers of correctly classified objects in all cases.

C. Microarray data

In this subsection, we will perform two tests based on two microarray data sets. They are yeast cell cycle data set and sporulation data set.

Yeast cell cycle data: This data set was published by Cho *et al.* [2]. It consisted of 6220 genes with 17 time points taken at ten-minute intervals. In the study of Yeung *et al.* [11], a subset of 384 genes is adopted. This subset of data set can be found on the website [25]. We normalized each gene expression profile with zero mean and unit variance. This data set has five cycle phases. They are early $G1$ phase, late $G1$ phase, S phase, $S2$ phase, and M phase.

The clustering results are given in Table VIII. The number of genes classified correctly using GMM is 252 and its mean is nearly 200 with standard derivation nearly 30. The standard derivation is very high and this implies that the perfor-

TABLE VII. Clustering results for the wdbc data.

Groups	GMM	FCM	HC	HARP	Proposed method
1	351	356	357	354	344
2	155	130	20	67	192
Total (Max)	506	486	377	421	536
Mean	484.1	486	/	421	516.9
Std	23.8768	0	/	0	13.4367

TABLE VIII. Clustering results for the yeast cell cycle data.

Phases	GMM	FCM	HC	HARP	Proposed method
Early $G1$	60	50	48	41	48
Late $G1$	115	67	112	116	120
S phase	31	10	1	23	28
$G2$	24	38	46	31	35
M phase	22	51	49	32	52
Total (Max)	252	216	256	243	283
Mean	199.9	216	/	243	282.3
Std	29.6403	0	/	0	0.9487

mance of the GMM is highly dependent on the initial guess. The FCM algorithm produces exactly the same results in the ten runs and classifies 216 genes correctly. The HC method classifies 256 genes correctly, and this is better than the GMM and the FCM algorithms. The HARP classifies 243 genes correctly and it is not better than either partition- or hierarchical-based method. The proposed method produces 283, the largest number of correctly classified objects. Its standard derivation is 0.9487, which is very small.

Sporulation data: This data set consists of 6118 genes [1] and can found on the website [26]. We only take the genes with the value of root mean square of the $\log_2$ transformed data greater than 1.13. After the preprocessing, we get a subset of the data, which contains 1136 genes of the following seven phases: rapid transient induction ("metabolic"), early I induction, early II induction, early-middle induction, middle induction, mid-late induction, and late induction.

The clustering result is given in Table IX. The GMM method yields a good result, with 324 genes classified correctly. This is better than the HARP, FCM, and HC algorithms. However, it has a very large standard derivation. The FCM and HC algorithms are very stable and produce the same results, with 280 genes classified correctly. The HARP algorithm classifies 256 genes correctly, which has the poorest performance. The proposed method classifies 353 genes correctly, which offers the best performance. Although the results have a large standard derivation, the number of correctly classified genes is still larger than that produced by the GMM algorithm. Furthermore, in the table, we can see that most of the methods failed to detect genes in the mid-late and late stages, but the proposed method successfully detected three and 33 genes in these two stages, respectively.

VI. CONCLUSIONS AND DISCUSSIONS

Clustering is an important procedure used in microarray data analysis. A major goal of microarray data analysis is to identify genes with similar functions. There are many methods proposed to handle this problem. The Gaussian mixture model and the fuzzy c-means algorithm are two well-known clustering algorithms. These methods assume that all features are significant for gene classification. In these and other commonly used clustering methods, the key idea is to identify two genes belonging to a group by measuring the distance between them using all features or conditions. However, in microarray data analysis, some of the conditions may be nonsignificant and noisy. They may degrade the classification performance. There is another kind of clustering algorithm called subspace clustering, which assumes that there are some nonsignificant conditions for gene classification. This method prunes some conditions and conducts clustering for the remaining conditions. However, the choice of the suitable conditions is still a critical problem. For the real world data, such as the iris and wdbc data sets, the subspace clustering

TABLE IX. Clustering results for the sporulation data.

	GMM	FCM	HC	HARP	Proposed method
Metabolic	4	1	1	0	3
Early I	172	172	244	241	173
Early II	24	7	4	0	38
Early-mid	95	66	8	5	43
Middle	29	32	21	9	60
Mid-late	0	2	2	1	3
Late	0	0	0	0	33
Total (Max)	324	280	280	256	353
Mean	306.9	280	/	256	340.4
Std	13.4449	0	/	0	21.8744

algorithm does not yield a better result than the traditional ones.

In this paper, we have introduced a concept for data clustering. The proposed algorithm does not prune any conditions in the data set and does not need any prior knowledge for selecting significant conditions. The key concept of the proposed algorithm is to measure the similarity between two objects in a number of subdimensions. Such a similarity measure reduces the effects of noise and outliers in the data. Experiments showed that the proposed idea gives more accurate results. We conclude that the concept of measuring the similarity between two objects using subdimensions is more robust than pruning nonsignificant conditions. In the future, we will apply this concept to supervised classification problems.

ACKNOWLEDGMENTS

We would like to thank K. Y. Yip [16] for sharing the HARP program for comparison. The work described in this paper is fully supported by Hong Kong Research Grant Council (Projects CityU No. 1035/02E and CityU No. 122005).

[1] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. Brown, and I. Herskowitz, Science **282**, 699 (1998).
[2] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart, and R. Davis, Mol. Cell **2**, 65 (1998).
[3] H. Agrawal, Phys. Rev. Lett. **89**, 268702 (2002).
[4] D. Jiang, T. Chun, and A. Zhang, IEEE Trans. Knowl. Data Eng. **16**, 1370 (2004).
[5] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms* (Plenum, New York, 1981).
[6] R. Johnson and D. Wichern, *Applied Multivariate Statistical Analysis*, 3rd ed. (Prentice Hall, New York, 1992).
[7] J. Banfield and A. Raftery, Biometrics **49**, 803 (1993).
[8] M. Blatt, S. Wiseman, and E. Domany, Phys. Rev. Lett. **76**, 3251 (1996).
[9] D. Dembele and P. Kastner, Bioinformatics **19**, 973 (2003).
[10] D. Kim, K. Lee, and D. Lee, Bioinformatics **21**, 1927 (2005).
[11] C. Giurcaneanu and I. Tabus, EURASIP J. Appl. Signal Process. **1**, 64 (2004).
[12] K. Yeung, C. Fraley, A. Murua, A. Raftery, and W. Ruzzo, Bioinformatics **17**, 977 (2001).
[13] A. Jain, M. Murty, and P. Flynn, ACM Comput. Surv. **31**, 264 (1999).
[14] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley, New York, 1990).
[15] C. Aggarwal, J. Wolf, P. Yu, C. Procopius, and J. Park, in *Proceedings of the ACM SIGMOD International Conference on Management of Data* (ACM Press, New York, 1999), p. 61.
[16] K. Y. Yip, D. Cheng, and M. Ng, IEEE Trans. Knowl. Data Eng. **16**, 1387 (2004).
[17] C. Cheng, A. Fu, and Y. Zhang, in *Proceedings of the ACM SIGMOD International Conference on Management of Data* (ACM Press, New York, 1999), p. 84.
[18] R. Agrawal, J. Gehrke, D. Gunopulous, and P Ragahavan, in *Proceedings of the ACM SIGMOD International Conference on Management of Data* (ACM Press, New York, 1999), pp. 94–105.
[19] L. Parsons, E. Haque, and H. Liu, SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining 6, 90 (2004).
[20] S. Madeira and A. Oliveira, IEEE/ACM Trans. Comput. Biol. Bioinf. **1**, 4 (2004).
[21] S. Bergmann, J. Ihmels, and N. Barkai, Phys. Rev. E **67**, 031902 (2003).
[22] B. Lam and H. Yan, *International Symposium on Intelligent Signal Processing and Communication Systems* (2005), p. 461–464.
[23] U. Maulik and S. Bandyopadhyay, IEEE Trans. Pattern Anal. Mach. Intell. **24**, 1650 (2002).
[24] http://www.ics.uci.edu/~mlearn/MLSummary.html
[25] http://faculty.washington.edu/kayee/model/
[26] http://cmgm.stanford.edu/pbrown/sporulation