World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:4, No:5, 2010

A Novel Microarray Biclustering Algorithm

Chieh-Yuan Tsai and Chuang-Cheng Chiu

***Abstract*—**Biclustering aims at identifying several biclusters that reveal potential local patterns from a microarray matrix. A bicluster is a sub-matrix of the microarray consisting of only a subset of genes co-regulates in a subset of conditions. In this study, we extend the motif of subspace clustering to present a *K*-biclusters clustering (KBC) algorithm for the microarray biclustering issue. Besides minimizing the dissimilarities between genes and bicluster centers within all biclusters, the objective function of the KBC algorithm additionally takes into account how to minimize the residues within all biclusters based on the mean square residue model. In addition, the objective function also maximizes the entropy of conditions to stimulate more conditions to contribute the identification of biclusters. The KBC algorithm adopts the *K*-means type clustering process to efficiently make the partition of *K* biclusters be optimized. A set of experiments on a practical microarray dataset are demonstrated to show the performance of the proposed KBC algorithm.

***Keywords*—**Microarray, Biclustering, Subspace clustering, Mean square residue model.

I. Introduction

Generally, the measurements of a microarray experiment are organized using a matrix format, called microarray matrix, whose rows represent genes and columns represent various specific experimental conditions. Each element in the matrix records a numeric value that represents the expression level of a particular gene under a given experimental condition [1]. Given a microarray matrix, biclustering aims at identifying subgroups of genes and subgroups of conditions by performing simultaneous clustering of both genes and conditions instead of clustering these two dimensions separately [2]. In each identified bicluster, each gene is identified based on only a subset of all conditions while each condition is selected using only a subset of all genes [3]. Through biclustering, biologists can efficiently annotate the genes with unknown functions and discover the functional relationships between genes [4].

A type of clustering methods, named subspace clustering [5], can find clusters from subspaces of data instead of the entire data space so that each found cluster is a set of objects identified by a subset of dimensions and different clusters are represented in different subsets of dimensions. Therefore, subspace clustering is suitable to address the microarray biclustering issue. The simultaneous clustering and attribute discrimination (SCAD) algorithm [6], one of the classical subspace clustering methods, adopts the *K*-means type clustering process [7] to efficiently make the partition of *K* clusters converge toward an optimal solution. During its clustering process, the SCAD algorithm adds an additional step to compute weights of dimensions in different clusters. The weight of each dimension in each cluster is adaptively adjusted to reflect the contribution of the dimension in forming the particular cluster. Through the SCAD algorithm, the subspaces of all *K* clusters can be respectively identified by their corresponding dimension weights.

In this study, we aim at extending the motif of the SCAD algorithm to develop a novel *K*-biclusters clustering (KBC) algorithm. The KBC algorithm is more appropriate for the biclustering issue than the SCAD algorithm since it adopts a more meaningful objective function. Besides minimizing the dissimilarities between genes and bicluster centers within all biclusters, the objective function of the KBC algorithm additionally takes into account how to minimize the residues within all biclusters. The residue is an indicator that measures the difference between the observed expression level of a particular gene under an experimental condition and its corresponding expected value [2]. Based on the notion of residue, the Mean Square Residue (MSR) model is accordingly defined to quantify the value coherence of each bicluster. Therefore, we attempt to integrate the MSR model into the objective function of the KBC algorithm. To accomplish the integration successfully, the MSR model is refined as a generalized form based on the memberships of a gene and a condition that belong to each bicluster respectively. In addition, the objective function also maximizes the entropy of conditions to stimulate more conditions to contribute the identification of biclusters. Similar to the SCAD algorithm, the proposed KBC algorithm adopts the *K*-means type clustering process to efficiently make the partition of *K* biclusters be optimized simultaneously through finite iterations. As a result, the biologically meaningful information hidden in a microarray matrix can be successfully uncovered through the proposed KBC algorithm.

II. Related works

A. Mean squared residue model for biclustering

Assume a set of genes $\boldsymbol{G}=\{\boldsymbol{g}_i|i=1,\ldots,I\}$ is experimented under a set of experimental conditions $\boldsymbol{C}=\{\boldsymbol{c}_j|j=1,\ldots,J\}$, and then the microarray matrix $\boldsymbol{E}$ with size $I \times J$, is accordingly obtained. An element e_{ij} in $\boldsymbol{E}$ records the observed expression level of the gene $\boldsymbol{g}_i$ under the condition $\boldsymbol{c}_j$. $\boldsymbol{g}_i=(e_{i1},\ldots,e_{ij},\ldots,e_{iJ})$ represents its all expression levels under all J conditions is represented as. Similarly, $\boldsymbol{c}_j=(e_{1j},\ldots,e_{ij},\ldots,e_{Ij})$ represents its all expression levels for all I genes. A bicluster $\boldsymbol{B}_k$ essentially

Chieh-Yuan Tsai is with the Department of Industrial Engineering and Management, Yuan-Ze University, Taiwan. (corresponding author to provide e-mail: cytsai@saturn.yzu.edu.tw).

corresponds to a sub-matrix of $\boldsymbol{E}$ that exhibits some coherent tendency. Note that k=1,…,K where K is the number of biclusters determined by users in advance. Each bicluster $\boldsymbol{B}_k$ is consisting of expression levels derived from a subset of $\boldsymbol{G}$, denoted as $\boldsymbol{G}_k \in \boldsymbol{G}$, and a subset of $\boldsymbol{C}$, denoted as $\boldsymbol{C}_k \in \boldsymbol{C}$. Assume the number of genes in $\boldsymbol{G}_k$ is I_k, $I_k \leq I$, and the number of conditions in $\boldsymbol{C}_k$ is J_k, $J_k \leq J$, so that the bicluster size of $\boldsymbol{B}_k$ equals to I_k multiplied by J_k. Cheng and Church [2] proposed an additive-based value-coherent bicluster model to translate an expression level e_{ij} in $\boldsymbol{B}_k$ into the sum of different effects, shown as Equation (1):

$$e_{ij} = \mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} \quad \text{for } \forall e_{ij} \in \boldsymbol{B}_k \tag{1}$$

where $\mu^{(k)} = \sum_{g_i \in G_k} \sum_{c_j \in C_k} e_{ij} / I_k \times J_k$ is the background effect for $\boldsymbol{B}_k$, $\alpha_i^{(k)} = \sum_{c_j \in C_k} e_{ij} / J_k - \mu^{(k)} = e_{iJ}^{(k)} - \mu^{(k)}$ is the ith gene effect for $\boldsymbol{B}_k$, and $\beta_j^{(k)} = \sum_{g_i \in G_k} e_{ij} / I_k - \mu^{(k)} = e_{Ij}^{(k)} - \mu^{(k)}$ is the jth condition effect for $\boldsymbol{B}_k$. In theory, a bicluster $\boldsymbol{B}_k$ is fully value-coherent if all e_{ij} in $\boldsymbol{B}_k$ satisfies:

$$e_{ij} = e_{iJ}^{(k)} + e_{Ij}^{(k)} - \mu^{(k)} \quad \text{for } \forall e_{ij} \in \boldsymbol{B}_k \tag{2}$$

However, given an arbitrary sub-matrix of $\boldsymbol{E}$, it might not be a fully value-coherent bicluster. To quantify the value coherence of a bicluster, the notion of residue is introduced to calculate the difference between the observed and expected values of e_{ij}. The residue $r_{ij}^{(k)}$ with respect to e_{ij} in $\boldsymbol{B}_k$ can be calculated as follows:

$$\begin{aligned} r_{ij}^{(k)} &= e_{ij} - (e_{iJ}^{(k)} + e_{Ij}^{(k)} - \mu^{(k)}) \\ &= e_{ij} - e_{iJ}^{(k)} - e_{Ij}^{(k)} + \mu^{(k)} \quad \text{for } \forall e_{ij} \in \boldsymbol{B}_k \end{aligned} \tag{3}$$

The lower the residue $r_{ij}^{(k)}$, the stronger the coherence of e_{ij} in the bicluster $\boldsymbol{B}_k$ is. Accordingly, the total incoherence of the genes $\boldsymbol{G}_k$ and the conditions $\boldsymbol{C}_k$ in $\boldsymbol{B}_k$ can be measured using the mean squared residue (MSR) defined as Equation (4):

$$\begin{aligned} MSR^{(k)} &= \sum_{g_i \in G_k} \sum_{c_j \in C_k} \left(r_{ij}^{(k)}\right)^2 \Big/ I_k J_k \\ &= \sum_{g_i \in G_k} \sum_{c_j \in C_k} \left(e_{ij} - e_{iJ}^{(k)} - e_{Ij}^{(k)} + \mu^{(k)}\right)^2 \Big/ I_k J_k \end{aligned} \tag{4}$$

The lower the MSR value of $\boldsymbol{B}_k$, the stronger the coherence exhibited by $\boldsymbol{B}_k$ is, i.e., the better the quality of $\boldsymbol{B}_k$ is. As a result, the biclustering problem can be regarded as an optimization model to find out the most appropriate K biclusters so that the sum of the K mean squared residues of all K biclusters can be minimized, shown as follows:

$$Minimize \sum_{k=1}^{K} MSR^{(k)} = \sum_{k=1}^{K} \left[\sum_{g_i \in G_k} \sum_{c_j \in C_k} \frac{\left(e_{ij} - e_{iJ}^{(k)} - e_{Ij}^{(k)} + \mu^{(k)}\right)^2}{I_k \times J_k} \right] \tag{5}$$

In the past decade, many scholars proposed their individual methods to optimize this MSR model [8, 9, 10, 11, 12, 13, 14].

B. Simultaneous clustering and attribute discrimination algorithm

Frigui and Nasraoui [6] proposed the simultaneous clustering and attribute discrimination (SCAD) algorithm for subspace clustering of large high-dimensional sparse data. When applying the SCAD algorithm to the biclustering issue, not only all I genes in $\boldsymbol{G}$ can be partitioned into K biclusters but also different weighting values will be assigned to different conditions for each bicluster based on the importance of the conditions in identifying the corresponding biclusters. Assume the all I genes have to be clustered as K biclusters $\boldsymbol{B}=\{\boldsymbol{B}_k | k=1,\ldots,K\}$ where $\boldsymbol{B}_k$ represents the kth bicluster with a center $\boldsymbol{o}_k=(o_{k1},\ldots,o_{kj},\ldots,o_{kJ})$ also contains J values. Note that the all K bicluster centers are organized by a set $\boldsymbol{O}=\{\boldsymbol{o}_k | k=1,\ldots,K\}$. Therefore, the objective of SCAD is to minimize the sum of dissimilarities among all I genes and all K centers. The dissimilarity between $\boldsymbol{g}_i$ and $\boldsymbol{o}_k$ is defined as:

$$diss(\boldsymbol{g}_i, \boldsymbol{o}_k) = \sum_{j=1}^{J} (w_{kj})^{\beta} (e_{ij} - o_{kj})^2 \tag{6}$$

where w_{kj} is the weight of the jth condition for the kth bicluster, and β (>1) is the parameter that controls the power of condition weight and is given by users. Based on Equation (6), the objective function of the SCAD algorithm is presented as:

$$\begin{aligned} &Minimize\ J_{SCAD}(\boldsymbol{U}, \boldsymbol{O}, \boldsymbol{W}) = \sum_{k=1}^{K} \sum_{i=1}^{I} (u_{ik})^{\alpha} diss(\boldsymbol{g}_i, \boldsymbol{o}_k) \\ &= \sum_{k=1}^{K} \sum_{i=1}^{I} \sum_{j=1}^{J} (u_{ik})^{\alpha} (w_{kj})^{\beta} (e_{ij} - o_{kj})^2 \end{aligned} \tag{7}$$

subject to

$$\begin{cases} u_{ik} \in [0,1], \ 1 \leq i \leq I \text{ and } 1 \leq k \leq K \\ \sum_{k=1}^{K} u_{ik} = 1, \ 1 \leq i \leq I \\ w_{kj} \in [0,1], \ 1 \leq k \leq K \text{ and } 1 \leq j \leq J \\ \sum_{j=1}^{J} w_{kj} = 1, \ 1 \leq k \leq K \end{cases} \tag{8}$$

where $\boldsymbol{U}$ is a $I \times K$ matrix that records the gene-bicluster memberships and u_{ik} is an element in $\boldsymbol{U}$ that represents the membership of gene $\boldsymbol{g}_i$ belongs to bicluster $\boldsymbol{B}_k$. In addition, $\boldsymbol{W}=\{w_{kj} | k=1,\ldots,K \text{ and } j=1,\ldots,J\}$ is a $K \times J$ matrix that records the weights of all J conditions for the K biclusters. α (>1) is the fuzzifier parameter given by users that controls the fuzziness of the memberships. When $\boldsymbol{O}$ and $\boldsymbol{W}$ are fixed, the update equation of u_{ik} for Equation (7) is shown as:

$$u_{ik} = \left(\sum_{j=1}^{J} (w_{kj})^{\beta} \times (e_{ij} - o_{kj})^2 \right)^{\frac{1}{1-\alpha}} \Bigg/ \sum_{t=1}^{K} \left[\left(\sum_{j=1}^{J} (w_{ij})^{\beta} \times (e_{ij} - o_{tj})^2 \right)^{\frac{1}{1-\alpha}} \right] \quad 1 \leq i \leq I, 1 \leq k \leq K \tag{9}$$

When $\boldsymbol{U}$ and $\boldsymbol{W}$ are fixed, the update equation of o_{kj} for Equation (7) is given by:

$$o_{kj} = \sum_{i=1}^{I}(u_{ik})^{\alpha} \times e_{ij} \Big/ \sum_{i=1}^{I}(u_{ik})^{\alpha} \quad 1 \le k \le K, 1 \le j \le J \tag{10}$$

Finally, the update equation of w_{kj} is shown as Equation (10) when $\boldsymbol{U}$ and $\boldsymbol{O}$ are fixed:

$$w_{kj} = \frac{\left(\sum_{i=1}^{I}(u_{ik})^{\alpha} \times (e_{ij} - o_{kj})^2\right)^{\frac{1}{1-\beta}}}{\sum_{t=1}^{J}\left[\left(\sum_{i=1}^{I}(u_{ik})^{\alpha} \times (e_{it} - o_{kt})^2\right)^{\frac{1}{1-\beta}}\right]} \quad 1 \le k \le K, 1 \le j \le J \tag{11}$$

Initially, $\boldsymbol{O}$ and $\boldsymbol{W}$ are randomly generated randomly. Afterward, the SCAD algorithm repeats to execute the three update equations (9), (10) and (11) until all K bicluster centers in $\boldsymbol{O}$ remain the same without being changed. The pseudo-code of the SCAD algorithm is illustrated in Fig. 1. Through the SCAD algorithm, which genes belong to the bicluster $\boldsymbol{B}_k$ can be recognized based on the elements u_{ik} for i=1,…,I in $\boldsymbol{U}$. Similarly, which conditions are used for the bicluster $\boldsymbol{B}_k$ can be identified based on the elements w_{kj} for j=1,…,J in $\boldsymbol{W}$. Therefore, the goal of biclustering can be achieved by the SCAD algorithm.

Input: a gene set G of I genes where each gene has the expression values under the J conditions; the number of biclusters, K;
Randomly initialize O and W.
Repeat {
Calculate the bicluster memberships of the all I genes for all K biclusters using Equation (9).
Find the K bicluster centers for the K biclusters respectively using Equation (10).
Re-evaluate the weights of the J conditions for the K biclusters using Equation (11).
} Until all the K bicluster centers in O remain the same without being changed.

Fig. 1 The pseudo-code of the SCAD algorithm

III. The Proposed K-Biclusters Clustering Algorithm

In this study, we present a novel K-biclusters clustering (KBC) algorithm which extends the motif of the SCAD algorithm for effective microarray biclustering. First, we introduce how to re-express the MSR model used to quantify the value coherence of each bicluster as a refined form. The refined MSR model is integrated with the original objective of SCAD to construct the objective function of the KBC algorithm. Therefore, the new objective function not only minimizes the dissimilarities among genes and centers within biclusters but also minimizes the total residues within biclusters. Its corresponding computational procedures to optimize the objective function are then expounded.

A. Refinement of the MSR model

Through using the matrixes $\boldsymbol{U}$ and $\boldsymbol{W}$ to represent the gene-bicluster and condition-bicluster memberships, the base $e_{iJ}^{(k)} = \sum_{\boldsymbol{c}_j \in \boldsymbol{C}_k} e_{ij} / J_k$ of $\boldsymbol{g}_i$, the base $e_{Ij}^{(k)} = \sum_{\boldsymbol{g}_i \in \boldsymbol{G}_k} e_{ij} / I_k$ of $\boldsymbol{c}_j$ and the background effort $\mu^{(k)} = \sum_{\boldsymbol{g}_i \in \boldsymbol{G}_k} \sum_{\boldsymbol{c}_j \in \boldsymbol{C}_k} e_{ij} / I_k \times J_k$ within the bicluster $\boldsymbol{B}_k$ can be respectively redefined as:

$$e_{iJ}^{(k)} = \sum_{j=1}^{J} w_{kj} \times e_{ij} \Big/ \sum_{j=1}^{J} w_{kj} \quad 1 \le k \le K, 1 \le i \le I \tag{12}$$

$$e_{Ij}^{(k)} = \sum_{i=1}^{I} u_{ik} \times e_{ij} \Big/ \sum_{i=1}^{I} u_{ik} \quad 1 \le k \le K, 1 \le j \le J \tag{13}$$

$$\mu^{(k)} = \sum_{i=1}^{I}\sum_{j=1}^{J} u_{ik} \times w_{kj} \times e_{ij} \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J} u_{ik} \times w_{kj} \quad 1 \le k \le K \tag{14}$$

Accordingly, the residue $r_{ij}^{(k)}$ with respect to e_{ij} in $\boldsymbol{B}_k$ for i=1,…,I and j=1,…,J is also re-defined as:

$$r_{ij}^{(k)} = e_{ij} - e_{iJ}^{(k)} - e_{Ij}^{(k)} + \mu^{(k)} = e_{ij} - \frac{\sum_{b=1}^{J} w_{kb} e_{ib}}{\sum_{b=1}^{J} w_{kb}} - \frac{\sum_{a=1}^{I} u_{ak} e_{aj}}{\sum_{a=1}^{I} u_{ak}} + \frac{\sum_{a=1}^{I}\sum_{b=1}^{J} u_{ak} w_{kb} e_{ab}}{\sum_{a=1}^{I}\sum_{b=1}^{J} u_{ak} w_{kb}}$$
$$1 \le k \le K, 1 \le i \le I, 1 \le j \le J \tag{15}$$

where $u_{ik} \in \{0,1\}$ is an element in $\boldsymbol{U}$ that represents the membership of gene $\boldsymbol{g}_i$ belongs to bicluster $\boldsymbol{B}_k$, and $w_{kj} \in [0,1]$ is an element in $\boldsymbol{W}$ that represents the membership of condition $\boldsymbol{c}_j$ belongs to bicluster $\boldsymbol{B}_k$. When the matrixes $\boldsymbol{U}$ and $\boldsymbol{W}$ are known, the value of each residue $r_{ij}^{(k)}$ can be effortlessly calculated using Equation (15). Based on Equation (15), we define a novel form of MSR model used to measure the incoherence of $\boldsymbol{B}_k$ as:

$$MSR^{(k)} = \sum_{i=1}^{I}\sum_{j=1}^{J} u_{ik} w_{kj} (r_{ij}^{(k)})^2 \quad 1 \le k \le K \tag{16}$$

The lower the MSR value of $\boldsymbol{B}_k$, the stronger the coherence exhibited by $\boldsymbol{B}_k$ is. The refined MSR model that minimizes the sum of K MSR values of all K biclusters is shown as:

$$\begin{aligned} Minimize\ J_{MSR}(\boldsymbol{U},\boldsymbol{W}) &= \sum_{k=1}^{K} MSR^{(k)} \\ &= \sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J} u_{ik} w_{kj} (r_{ij}^{(k)})^2 \end{aligned} \tag{17}$$

When the value of $\sum_{k=1}^{K} MSR^{(k)}$ is small, each of the K biclusters is compact or separated from other biclusters. When the value of $\sum_{k=1}^{K} MSR^{(k)}$ is large, by contrast, some of these K biclusters are not compact or separated from other biclusters. Through the above analysis, the purpose of refined MSR model is close to the purpose of the SCAD algorithm. Therefore, the objective function in Equation (17) can be integrated with the objective function of SCAD in Equation (7) for effective microarray biclustering.

B. The computational procedure of the KBC algorithm

By integrating the two objectives Equation (7) and Equation (17) together, the new objective function used in the proposed

KBC algorithm is written as follows:

$$\begin{aligned} &Minimize\ J_{KB}(\boldsymbol{U},\boldsymbol{W},\boldsymbol{O}) \\ &=\left(\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J}u_{ik}w_{kj}(e_{ij}-o_{kj})^2\right)+\left(\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J}u_{ik}w_{kj}(r_{ij}^{(k)})^2\right) \end{aligned} \tag{18}$$

subject to

$$\begin{cases} u_{ik}\in\{0,1\},\ 1\le i\le I \text{ and } 1\le k\le K \\ \sum_{k=1}^{K}u_{ik}=1,\ 1\le i\le I \\ w_{kj}\in[0,1],\ 1\le j\le J \text{ and } 1\le k\le K \\ \sum_{j=1}^{J}w_{kj}=1,\ 1\le k\le K \end{cases} \tag{19}$$

The first term in Equation (18) is used to minimize the within-bicluster dispersion, i.e. to minimize the dissimilarities among genes within biclusters. The second term in Equation (18) is used to minimize the within-bicluster incoherence, i.e. to minimize the total residues within biclusters. For stimulating more conditions to contribute the identification of biclusters, further, the new objective function in Equation (18) should involve an additional term used to maximize the entropy of conditions. Entropy is the measurement of information and uncertainty on a random variable [15]. In this study, we assume the condition weight is a random variable. Therefore, if the values of condition weights are distributed uniformly, the condition weights have the maximum entropy, i.e. maximal uncertainty or minimal information to know the value of a condition weight. The entropy of all condition weights is formulated as $-\sum_{k=1}^{K}\sum_{j=1}^{J}w_{kj}\ln w_{kj}$. To stimulate more conditions, i.e. to make the condition weights be distributed as uniformly as possible, the entropy of all condition weights should be maximized. Because the objective function in Equation (18) is to achieve a minimization goal, we have to add the negative term of $-\sum_{k=1}^{K}\sum_{j=1}^{J}w_{kj}\ln w_{kj}$ into Equation (18) when taking into account the entropy of all condition weights. As a result, the new objective function is updated as Equation (20):

$$\begin{aligned} &Minimize\ J_{KB}(\boldsymbol{U},\boldsymbol{W},\boldsymbol{O}) \\ &=\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J}u_{ik}w_{kj}(e_{ij}-o_{kj})^2+\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J}u_{ik}w_{kj}(r_{ij}^{(k)})^2+\delta\sum_{k=1}^{K}\sum_{j=1}^{J}w_{kj}\ln w_{kj} \\ &=\left(\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J}u_{ik}w_{kj}[(e_{ij}-o_{kj})^2+(r_{ij}^{(k)})^2]\right)+\delta\left(\sum_{k=1}^{K}\sum_{j=1}^{J}w_{kj}\ln w_{kj}\right) \end{aligned} \tag{20}$$

where its corresponding constraints is the same with Equation (19). The parameter δ, δ>0, controls the incentive strength for biclustering on more conditions, and is given by users.

Minimizing the above objective function J_{KB} with the constraints is considered as a constrained nonlinear optimization problem. By introducing a vector of K Lagrangian multipliers $\boldsymbol{\lambda}=[\lambda_1,\ldots,\lambda_k,\ldots,\lambda_K]$ to the constraints $\sum_{j=1}^{J}w_{kj}=1$ for k=1,…,K, the minimization of Equation (20) can be transformed to minimize a Lagrangian function shown as Equation (21):

$$\begin{aligned} &Minimize\ \Psi(\boldsymbol{U},\boldsymbol{W},\boldsymbol{O},\boldsymbol{\lambda}) \\ &=\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J}(u_{ik})(w_{kj})[(e_{ij}-o_{kj})^2+(r_{ij}^{(k)})^2]+\delta\sum_{k=1}^{K}\sum_{j=1}^{J}w_{kj}\ln w_{kj} \\ &-\sum_{k=1}^{K}\sum_{j=1}^{J}\lambda_k(w_{kj}-1) \end{aligned} \tag{21}$$

If $(\hat{\boldsymbol{U}},\hat{\boldsymbol{W}},\hat{\boldsymbol{O}},\hat{\boldsymbol{\lambda}})$ is able to minimize $\Psi(\boldsymbol{U},\boldsymbol{W},\boldsymbol{O},\boldsymbol{\lambda})$, the gradient of $\Psi(\boldsymbol{U},\boldsymbol{W},\boldsymbol{O},\boldsymbol{\lambda})$ in all variables must vanish. That is, the first partial differentiation of $\Psi(\boldsymbol{U},\boldsymbol{W},\boldsymbol{O},\boldsymbol{\lambda})$ with respect to u_{ik}, w_{kj}, o_{kj} and λ_k must equal to zero. Therefore, we first fix $\boldsymbol{U}=\hat{\boldsymbol{U}}$, $\boldsymbol{W}=\hat{\boldsymbol{W}}$ and $\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}$ to make the gradient of $\Psi(\boldsymbol{U},\boldsymbol{W},\boldsymbol{O},\boldsymbol{\lambda})$ with respect to o_{kj} be zero, and then we obtain:

$$\frac{\partial\Psi(\boldsymbol{U},\boldsymbol{W},\boldsymbol{O},\boldsymbol{\lambda})}{\partial o_{kj}}=2\times\sum_{i=1}^{I}u_{ik}w_{kj}(o_{kj}-e_{ij})=0 \tag{22}$$

From Equation (22), we obtain the Equation for updating bicluster centers:

$$o_{kj}=\sum_{i=1}^{I}u_{ik}\times e_{ij}\Big/\sum_{i=1}^{I}u_{ik}\quad 1\le k\le K,1\le j\le J \tag{23}$$

Similarly, we afterward fix $\boldsymbol{W}=\hat{\boldsymbol{W}}$, $\boldsymbol{O}=\hat{\boldsymbol{O}}$ and $\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}$ to acquire the Equation for assigning each gene to a proper bicluster:

$$u_{ik}=\begin{cases}1,\text{ if }\sum_{j=1}^{J}w_{kj}[(e_{ij}-o_{kj})^2+(r_{ij}^{(k)})^2]<\sum_{j=1}^{J}w_{kj}[(e_{pj}-o_{kj})^2+(r_{pj}^{(k)})^2] \\ 0,\text{ otherwise}\end{cases}$$
$$1\le i\le I,1\le p\le I,p\ne i,1\le k\le K \tag{24}$$

Finally, the minimization problem in Equation (21) can be decomposed into K independent minimization sub-problems:

$$\begin{aligned} Minimize\ \Psi_k=&\sum_{i=1}^{I}\sum_{j=1}^{J}u_{ik}w_{kj}[(e_{ij}-o_{kj})^2+(r_{ij}^{(k)})^2] \\ &+\delta\sum_{j=1}^{J}w_{kj}\ln w_{kj}-\lambda_k(\sum_{j=1}^{J}w_{kj}-1) \end{aligned} \tag{25}$$

When each of the K independent sub-problems can be minimized, minimizing Equation (21) is accordingly accomplished. Therefore, by setting the gradient of Ψ_k with respect to λ_k and w_{kj} to zero, we obtain

$$\partial\Psi_k/\partial\lambda_k=\sum_{j=1}^{J}w_{kj}-1=0\quad 1\le k\le K \tag{26}$$

$$\frac{\partial \Psi_k}{\partial w_{kj}} = \left(\sum_{i=1}^{I} u_{ik} [(e_{ij} - o_{kj})^2 + (r_{ij}^{(k)})^2] \right) + \delta(1 + \ln w_{kj}) - \lambda_k = 0 \quad (27)$$
$$1 \le k \le K, 1 \le j \le J$$

From Equation (27), we obtain

$$w_{kj} = \exp\left(\frac{\lambda_k - \left(\sum_{i=1}^{I} u_{ik} [(e_{ij} - o_{kj})^2 + (r_{ij}^{(k)})^2] \right) - \delta}{\delta} \right) = \exp\left(\frac{\lambda_k - \delta}{\delta} \right) \exp\left(- \sum_{i=1}^{I} u_{ik} [(e_{ij} - o_{kj})^2 + (r_{ij}^{(k)})^2] \Big/ \delta \right) \quad (28)$$

Substituting Equation (28) into Equation (26), we have

$$\sum_{j=1}^{J} w_{kj} = \exp\left(\frac{\lambda_k - \delta}{\delta} \right) \sum_{j=1}^{J} \exp\left(- \sum_{i=1}^{I} u_{ik} [(e_{ij} - o_{kj})^2 + (r_{ij}^{(k)})^2] \Big/ \delta \right) = 1 \quad (29)$$

It follows that

$$\exp\left(\frac{\lambda_k - \delta}{\delta} \right) = 1 \Big/ \sum_{j=1}^{J} \exp\left(- \sum_{i=1}^{I} u_{ik} [(e_{ij} - o_{kj})^2 + (r_{ij}^{(k)})^2] \Big/ \delta \right) \quad (30)$$

Substituting Equation (30) back to Equation (28), we obtain

$$w_{kj} = \exp\left(\frac{\lambda_k - \delta}{\delta} \right) \exp\left(- \sum_{i=1}^{I} u_{ik} [(e_{ij} - o_{kj})^2 + (r_{ij}^{(k)})^2] \Big/ \delta \right) = \frac{\exp\left(- \sum_{i=1}^{I} u_{ik} [(e_{ij} - o_{kj})^2 + (r_{ij}^{(k)})^2] \Big/ \delta \right)}{\sum_{b=1}^{J} \exp\left(- \sum_{i=1}^{I} u_{ik} [(e_{ib} - o_{kb})^2 + (r_{ib}^{(k)})^2] \Big/ \delta \right)}$$
$$1 \le k \le K, 1 \le j \le J \quad (31)$$

As shown in Equation (31), w_{kj} is inversely proportional to $\sum_{i=1}^{I} u_{ik} [(e_{ij} - o_{kj})^2 + (r_{ij}^{(k)})^2]$ if δ>0. It means the condition $\boldsymbol{c}_j$ is more important for the bicluster $\boldsymbol{B}_k$, i.e. w_{kj} is larger, when all genes are closer to the center $\boldsymbol{o}_k$ and all residues are smaller in terms of the condition $\boldsymbol{c}_j$ for the bicluster $\boldsymbol{B}_k$. If δ=0, only one condition weight will equals to one when its corresponding value of $\sum_{i=1}^{I} u_{ik} [(e_{ij} - o_{kj})^2 + (r_{ij}^{(k)})^2]$ is smallest among all condition weights. That is, each bicluster contains only one important condition. It may not be desirable for the biclustering issue. If δ>0, w_{kj} is proportional to $\sum_{i=1}^{I} u_{ik} [(e_{ij} - o_{kj})^2 + (r_{ij}^{(k)})^2]$. It is also obviously contradictory to the original essence of condition weighting. Therefore, δ is suggested to be δ>0.

Given a microarray matrix $\boldsymbol{E}$ consisting of the gene set $\boldsymbol{G}$ and the condition set $\boldsymbol{C}$, the KBC algorithm initially sets the J condition weights for each bicluster by the same value, i.e. w_{kj}=1/J for k=1,…,K and j=1,…,J. Simultaneously, each of the all I genes is randomly assigned to one of the all K biclusters, i.e. the values of u_{ik} for k=1,…,K and i=1,…,I are determined randomly at the beginning. Subsequently, the KBC algorithm adopts the K-means type clustering procedures to make the K biclusters be monotonously optimized by iteratively executing the following four steps through finite iterations:

1. Calculate the values of all residues $r_{ij}^{(k)}$ for k=1,…,K, i=1,…,I and j=1,…,J based on the all obtained u_{ik} and w_{kj} using Equation (15).
2. Update the K bicluster centers $\boldsymbol{O}$={$\boldsymbol{o}_k$|k=1,…,K}, i.e. to update o_{kj} for k=1,…,K and j=1,…,J using Equation (23).
3. Assign each gene $\boldsymbol{g}_i$ to a proper bicluster $\boldsymbol{B}_k$, i.e. to determine u_{ik} for k=1,…,K and i=1,…,I using Equation (24).
4. Evaluate the new weights of all conditions for each bicluster, i.e. to calculate w_{kj} for for k=1,…,K and j=1,…,J using Equation (31).

After finite iterations, assume the all K bicluster centers have not be changed at the tth iteration. It means that each center $\boldsymbol{o}_k(t)$ at the tth iteration is equal to $\boldsymbol{o}_k(t+1)$ at the (t+1)th iteration for k=1,…,K where t is the number index of iteration, i.e. $\max_{k=1,\cdots,K} \|\boldsymbol{o}_k(t+1) - \boldsymbol{o}_k(t)\| < \varepsilon$ where ε is the user-specified threshold parameter. Note that $\varepsilon = 0.1$ is adopted in this study. In this situation, it is declared that the objective of KBC algorithm has been converged on a minimum state at the tth iteration. The total computational complexity of the KBC algorithm is $O(KIJt)$. It shows that the computational complexity increases linearly as the number of genes, conditions, and biclusters increases. Therefore, the KBC algorithm is scalable for a large microarray matrix. The pseudo-code of the KBC algorithm is shown as Fig. 2.

Input: a gene set G of I genes where each gene has the expression values under the J conditions;
the number of biclusters K;
the parameter δ that controls the incentive strength for biclustering on more conditions;
the parameter ε as the threshold to stop the K-Biclusters clustering algorithm.
Initialize W by setting w_{kj}=1/J for k=1,…,K and j=1,…,J.
Initialize U randomly.
Set t = 0 where t is the index number of iteration.
Repeat {
Calculate all residues $r_{ij}^{(k)}$ for k=1,…,K, i=1,…,I and j=1,…,J using Equation (15).
Update the K bicluster centers o_{kj} for k=1,…,K and j=1,…,J using Equation (23).
Assign each gene to a proper bicluster by determining u_{ik} for k=1,…,K and i=1,…,I using Equation (24).
Evaluate the weights of all conditions for each bicluster w_{kj} for k=1,…,K and j=1,…,J using Equation (31).
$t = t + 1$.
} Until all the K bicluster centers in O remain the same without being changed.
That is, the maximum of ($\|\boldsymbol{o}_k(t+1) - \boldsymbol{o}_k(t)\|$) for k=1,…,K is less than ε.

Fig. 2 The pseudo-code of the proposed KBC clustering algorithm

IV. Experiments

The performance of the proposed K-biclusters clustering (KBC) algorithm is compared with that of the simultaneous clustering and attribute discrimination (SCAD) algorithm through several experiments. The yeast cell cycle dataset [16] commonly used in the literatures serves as the benchmark dataset in our experiments. The dataset records the expression

levels of 384 genes under 17 conditions. Each gene in the dataset has been biologically characterized and classified to one of the five phases of cell cycles, including early G1 phase, late G1 phase, S phase, G2 phase and M phase [17]. The genes in each phase of the cell cycle present higher expression levels at specific conditions than those at other conditions, so each phase of cell cycle can be considered as a bicluster. It means there are five biclusters existing in this dataset in which the bicluster level of each gene has been known. Its data and legend can be downloaded from http://faculty.washington.edu/kayee/model/.

After executing the KBC or SCAD algorithms, their generated matrixes $\boldsymbol{U}$ and $\boldsymbol{W}$ can reveal the biclustering results to users. For the KBC algorithm, if u_{ik}=1 in $\boldsymbol{U}$, the gene $\boldsymbol{g}_i$ belongs to the bicluster $\boldsymbol{B}_k$. If u_{ik}=0, otherwise, $\boldsymbol{g}_i$ does not belong to $\boldsymbol{B}_k$. Meanwhile, if $w_{kj} \geq 1/J$ in $\boldsymbol{W}$, i.e. the condition weight w_{kj} is larger than the average of all J condition weights, then the condition $\boldsymbol{c}_j$ belongs to the bicluster $\boldsymbol{B}_k$. If $w_{kj}<1/J$, otherwise, $\boldsymbol{c}_j$ does not belong to $\boldsymbol{B}_k$. For the SCAD algorithm, if $u_{ik} \geq 1/K$ in $\boldsymbol{U}$, then the gene $\boldsymbol{g}_i$ belongs to the bicluster $\boldsymbol{B}_k$. If $u_{ik}<1/K$, otherwise, $\boldsymbol{g}_i$ does not belong to $\boldsymbol{B}_k$. Meanwhile, if $w_{kj} \geq 1/J$ in $\boldsymbol{W}$, then the condition $\boldsymbol{c}_j$ belongs to the bicluster $\boldsymbol{B}_k$. If $w_{kj}<1/J$, otherwise, $\boldsymbol{c}_j$ does not belong to $\boldsymbol{B}_k$. From the biclustering result generated by an algorithm, therefore, we know each bicluster contains which genes and which conditions according to these principles.

Because the yeast cell cycle dataset is a labeled microarray dataset, the performance of an algorithm for the biclustering issue can be evaluated based on the accuracy rates of bicluster recognition. The accuracy rate is the number of genes that are correctly classified into the corresponding biclusters divided by the number of all genes. The larger the accuracy rate, the better the effectiveness of an algorithm is. The number of biclusters K=5 is directly assigned to the two algorithms. Further, two parameters α and β have to be assigned for the SCAD algorithm. We take six values of α=0.5, 1, 1.5, 2, 2.5, 3 and six values of β=0.5, 1, 1.5, 2, 2.5, 3 to form 36 parameter pairs of (α,β) for the SCAD algorithm. By contrast, only one parameter δ is specified for the proposed KBC algorithm. Therefore, eight values of δ=0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4 are concerned in our experiments. The SCAD and KBC algorithms both randomly generate initial bicluster centers. For sake of fair comparison, they both need to be performed five times for each parameter setting and the average of experiment results in the five runs is finally recorded.

Fig. 3 shows the accuracy rates using the SCAD algorithm with different parameter settings for this dataset. As shown in Fig. 3, the relationship between the accuracy rate and the parameters (α,β) is seen as a mountain function in which the peak (i.e. corresponding to the highest accuracy rate) occurs at (α,β)=(2,1.5). On the other hand, the accuracy rates using the KBC algorithm with different values of δ for this dataset is shown as Fig. 4. When δ=0, i.e. disregarding the effort of the term $\sum_{k=1}^{K}\sum_{j=1}^{J} w_{kj} \times \ln w_{kj}$ in Equation (20), the generated accuracy rate is relatively poor. When $\delta \geq 1$ is given, the generated accuracy rate will become relatively good.

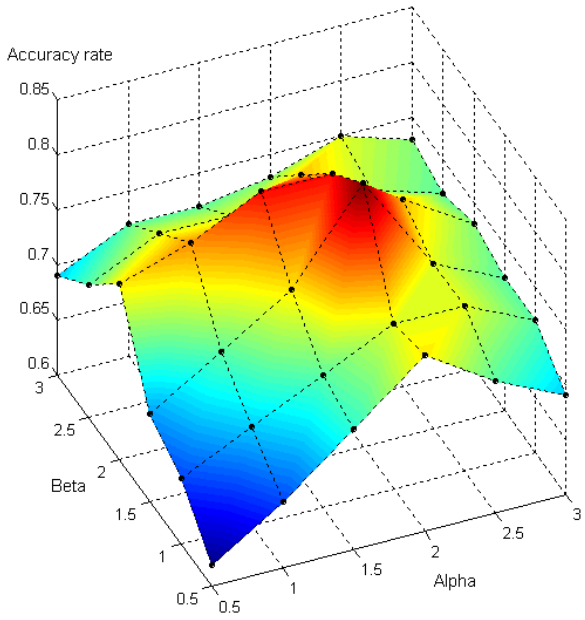

Fig. 3 The accuracy rates using the SCAD algorithm with different settings of (α,β) for the yeast cell cycle dataset

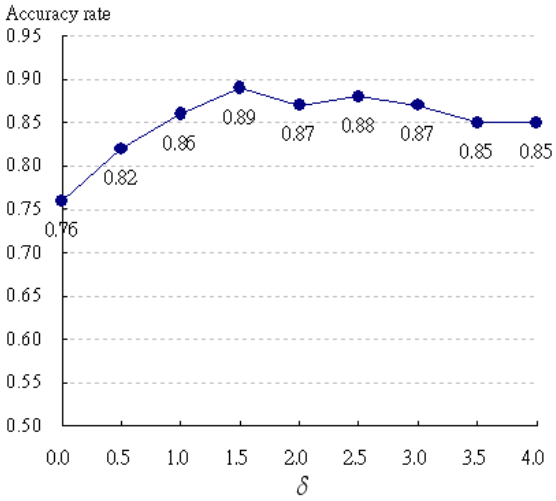

Fig 4. The accuracy rates using the KBC algorithm with different values of δ for the yeast cell cycle dataset

The comparisons between the two algorithms based on accuracy rate are summarized in Table 1. From Table 1, we know that the effectiveness of the proposed KBC algorithm is superior to the effectiveness of the SCAD algorithm because its average of accuracy rates is relatively high. In addition, using the KBC algorithm can get the smaller standard deviation of accuracy rates than using the SCAD algorithm. It means that the effectiveness of the KBC algorithm is more insensitive to the parameter settings than the SCAD algorithm. Figure 4 also reveals this contention since the high accuracy rates can be obtained as long as $\delta \geq 1$. The insensitivity to the parameter settings will facilitate the manipulation of the KBC algorithm for biologists.

Table I The summarization of accuracy rates generated by the SCAD and KBC algorithms for the yeast cell cycle dataset

Algorithms		SCAD	KBC
Accuracy rate	Maximum	0.84	0.89
	Minimum	0.62	0.76
	Average	0.733	0.850
	Standard deviation	0.043	0.039

V. Conclusion

In this study, we present the *K*-biclusters clustering (KBC) algorithm for microarray biclustering. To achieve effective biclustering, the KBC algorithm attempts to minimize the dissimilarities between genes and bicluster centers, to minimize the residues within all biclusters, and to involve conditions as many as possible. Through our experiments on the yeast cell cycle dataset, the KBC algorithm is indeed effectively conduct the microarray biclustering issue. By following the K-means type clustering process, furthermore, the algorithm is capable of efficiently finding several biclusters simultaneously. It prevents from the more duplication degrees among biclusters occurred in traditional bilcustering approaches. In addition to the number of biclusters, only one parameter δ needs to be assigned to the KBC algorithm. Our experiment results show that the performance of the algorithm is insensitive to the parameter setting, which facilitates the manipulation of the algorithm for users.

How to assign a proper number of biclusters *K* is still a challenge for the KBC algorithm, especially when applying it to real applications. It is well-known that the biclustering result will be highly influenced by the parameter. In the future, therefore, we will exploit the notion of data self-agglomeration to refine the KBC algorithm so that the KBC algorithm will automatically determine the proper number of biclusters without manual setting.

References

[1] J. L. DeRisi, V. R. Iyer, and P. O. Brown, 1997. "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, pp. 680–686, 1997.

[2] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pp. 93–103, 2000.

[3] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, pp.24–45, 2004.

[4] D. P. Berrer, W. Dubitzky and M. Granzow, *A Practical Approach to Microarray Data Analysis*. Kluwer, Norwell, pp. 15–19, 2003.

[5] L. Parsons, E. Haque and H. Liu, "Subspace clustering for high dimensional data: A review," *SIGKDD Explorations*, vol. 6, pp. 90–105, 2004.

[6] H. Frigui and O. Nasraoui, "Unsupervised learning of prototypes and attribute weights," *Pattern Recognition*, vol. 37, pp. 567–581, 2004.

[7] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[8] K. Bryan, P. Cunningham and N. Bolshakova, "Application of simulated annealing to the biclustering of gene expression data," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, pp. 519–525, 2006.

[9] J. Yang, W. Wang, H. Wang and P. Yu, "δ-clusters: Capturing subspace correlation in a large data set," in *Proceedings of the 18th IEEE International Conference on Data Engineering*, pp. 517–528, 2002.

[10] J. Yang, H. Wang, W. Wang and P. Yu, "Enhanced biclustering on expression data," in *Proceedings of the Third IEEE Symposium on BioInformatics and Bioengineering*, pp. 1–7, 2003.

[11] A. Tanay, R. Sharan and R., Shamir, "Discovering statistically significant biclusters in gene expression data," *Bioinformatics*, vol. 18, pp. 136–144, 2002.

[12] S. Mitra and H. Banka, "Multi-objective evolutionary biclustering of gene expression data," *Pattern Recognition*, vol. 39, pp. 2464–2477, 2006.

[13] M. Filippone, F. Masulli, S. Rovetta, S. Mitra and H. Banka, "Possibilistic approach to biclustering: An application to oligonucleotide microarray data analysis," *Lecture Notes in Computer Science*, vol. 4210, pp. 312–322, 2006.

[14] S. Mitra, H. Banka and J. H. Paik, "Evolutionary fuzzy biclustering of gene expression data," *Lecture Notes in Artificial Intelligence*, vol. 4481, pp. 284–291, 2007.

[15] C. E. Shannon, "A mathematical theory of communication," The Bell System Technical Journal, vol. 27, pp. 379–423, 623–656, 1948.

[16] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart and R. W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, pp. 65–73, 1998.

[17] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein and B. Futcher, "Comprehensive identification of cell cycle regulated genes of the yeast Saccharomyces Cerevisiae by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273–3297, 1998.