BIOINFORMATICS ORIGINAL PAPER
Vol. 23 no. 21 2007, pages 2888–2896
doi:10.1093/bioinformatics/btm463

Gene expression

Graph-based consensus clustering for class discovery from gene expression data

Zhiwen Yu*, Hau-San Wong and Hongqiang Wang

Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

Received on April 11, 2007; revised on August 15, 2007; accepted on September 8, 2007

Advance Access publication September 14, 2007

Associate Editor: David Rocke

ABSTRACT

Motivation: Consensus clustering, also known as cluster ensemble, is one of the important techniques for microarray data analysis, and is particularly useful for class discovery from microarray data. Compared with traditional clustering algorithms, consensus clustering approaches have the ability to integrate multiple partitions from different cluster solutions to improve the robustness, stability, scalability and parallelization of the clustering algorithms. By consensus clustering, one can discover the underlying classes of the samples in gene expression data.

Results: In addition to exploring a graph-based consensus clustering (GCC) algorithm to estimate the underlying classes of the samples in microarray data, we also design a new validation index to determine the number of classes in microarray data. To our knowledge, this is the first time in which GCC is applied to class discovery for microarray data. Given a pre specified maximum number of classes (denoted as $K_{\max}$ in this article), our algorithm can discover the true number of classes for the samples in microarray data according to a new cluster validation index called the Modified Rand Index. Experiments on gene expression data indicate that our new algorithm can (i) outperform most of the existing algorithms, (ii) identify the number of classes correctly in real cancer datasets, and (iii) discover the classes of samples with biological meaning.

Availability: Matlab source code for the GCC algorithm is available upon request from Zhiwen Yu.

Contact: yuzhiwen@cs.cityu.edu.hk and cshswong@cityu.edu.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Recently, the problem of discovering the underlying classes from microarray data receives more and more attention due to its important applications in cancer diagnosis (Golub *et al*., 1999), gene expression analysis (Alizadeh *et al*., 2000) and related areas (Su *et al*., 2002). The adoption of microarray techniques allow the acquisition, through a series of experiments, the expression profile of a genome in a number of different experiment conditions (Baldi and Hatfield, 2002). Information acquired through microarray is usually characterized through two dimensions: the gene dimension and the sample dimension. We can either observe and characterize the expression level distribution of the complete set of genes included in the array under a particular experimental condition, or track the variation of the expression level of a single gene or a selected subset of genes across the different samples. In this article, we mainly focus on the categorization of the samples through our newly proposed graph-based consensus clustering (GCC) approach.

Most of the existing approaches in cancer classification can be categorized into two types: class discovery and class prediction. In general, class discovery consists of two steps: (1) a clustering algorithm is first adopted to partition the samples into K parts, when given a new set of microarray data with unknown number of classes, (2) a cluster validity index is then applied to determine the optimal K value, which corresponds to the final number of classes. Our new approach belongs to this category. For the class discovery problem, the researchers mainly focus on discovering the underlying classes from the samples. In Golub *et al*. (1999), two types of human acute leukemia were discovered with the help of self organizing feature maps and neighborhood analysis. In Alizadeh *et al*. (2000), two subtypes of diffused large B-cell lymphoma that are distinct at the molecular level are identified with centroid average hierarchical clustering. In Yeung *et al*. 2001, the Gaussian mixture model is applied to discover the underlying distribution of the data in microarray. Wigle *et al*. 2002, identified lung cancer cases are distinguished from the normal cases through statistical analysis and clustering approaches. In Dudoit and Fridlyand (2002), a new prediction-based resampling approach is designed to estimate the number of clusters in microarray data. In Dudoit and Fridlyand (2003), two new resampling approaches based on bagging are proposed to improve the accuracy of the clustering procedure. The silhouette index is adopted as a cluster validity index to determine the optimal number of clusters. In Smolkin and Ghosh (2003), a cluster stability score for gene expression data is proposed to assess the stability of individual clusters based on the random subspace techniques, which can be used to identify the number of clusters by combining with any clustering algorithm. In Handl *et al*. (2005), the authors performed a survey of cluster validity indices when applied to gene expression data analysis. In particular, they pointed out the benefits and the problems of computational cluster validation

*To whom correspondence should be addressed.

techniques. In Datta and Datta (2006), the authors designed two performance measures called biological homogeneity index (BHI) and biological stability index (BSI). BHI measures the biological homogeneity of the clusters, while BSI measures the stability of the clusters. In their work, the main focus is on the functionality of the genes, not the categories of the samples. In addition, since some of the genes in their dataset are labeled using biological tools, their approach belongs to the semi-supervized learning category. In Bertoni and Valentini (2006), randomized maps based on the Johnson–Lindenstrauss theory are designed to project the high-dimensional gene expression data on to a lower dimensional subspace. Then, a stability measure based on the projected data is proposed to discover the underlying structure from microarray data. In Bertoni and Valentini (2007a), a general framework for the assessment of clustering solutions is designed based on the random subspace technique. The true number of clusters is estimated by a χ^2-based statistical test, which makes use of the distribution of the similarity measure between pairs of clusterings.

Recently, researchers are paying more attention to class discovery based on the consensus clustering approaches. Consensus clustering approaches consist of two major steps: generating a cluster ensemble based on a clustering algorithm, and finding a consensus partition based on this ensemble. The existing consensus clustering approaches that are applied to gene expression data can be categorized into five types: (i) using different clustering algorithms as the basic clustering algorithms to obtain different solutions (Strehl and Ghosh, 2002). (ii) using random initializations of a single clustering algorithm (Grotkjaer *et al*., 2006), e.g. adopting different initial centers for *K*-means or EM. (iii) sub-sampling, re-sampling or adding noise to the original data (McShane, 2002, Monti *et al*., 2003, Valentini, 2007). (iv) using selected subsets of features (Bertoni and Valentini, 2005, Bertoni and Valentini, 2007a, Bertoni and Valentini, 2007b, Smolkin and Ghosh, 2003, Topchy *et al*., 2005, Valentini, 2006). (v) using different K values to generate different clustering solutions, where K is the number of clusters.

The current consensus clustering approaches have limitations. Specifically, two aspects of the current consensus clustering approaches can be improved, namely, the diversity of the ensemble, and the accuracy of the partitioning results obtained from the consensus matrix. In this article, we propose a new consensus clustering approach, known as GCC, to discover biologically meaningful classes automatically from gene expression data. Our new approach belongs to category (iv), in which the cluster ensemble is generated using different gene subsets. Compared with the previous approaches, in particular the approaches proposed by Bertoni and Valentini (2007a) and Smolkin and Ghosh (2003), the main features of the GCC approach include (1) the adoption of the normalized cut algorithm (Shi and Malik, 2000) to partition the consensus matrix, (2) the adoption of the random subspace technique, combined with the correlation clustering algorithm or *K*-means, to enhance the diversity of the cluster ensemble and (3) the design of a new cluster validity index called the Modified Rand Index, which measures the degree of agreement between two consensus matrices based on a penalty term.

The work most related to ours is described in Monti *et al.* (2003). They proposed a consensus clustering approach to identify the underlying types of cancers in a number of datasets. Unlike previous consensus clustering approaches, their approach can estimate the number of classes. However, although the results obtained using six cancer datasets are acceptable, there is a need to further improve the estimation performance to allow more accurate diagnosis. Compared with this approach, our proposed algorithm has two differences: (1) GCC only considers subsets of genes in the process of generating the clustering solution. (2) GCC adopts the new Modified Rand Index, which allows more accurate characterization of the class structure. Specifically, GCC first generates a set of clustering solutions in the gene subspace. Then, it creates a consensus matrix that integrates the partitions coming from the different clustering solutions. Finally, a new validation index, called the Modified Rand Index, is designed to estimate the number of classes automatically from gene expression data. Our experiments show that GCC successfully identifies the number of classes in real cancer datasets.

2 METHODS

We formulate the GCC problem for cancer classification as follows: Given a set of samples $\mathbf{S} = \{S_1, S_2, \ldots, S_{n_s}\}$ where n_s is the number of samples, the GCC constructor first randomly selects a subset of genes and obtains a clustering solution I^u by partitioning the samples into K disjoint classes ($I^u = \{C_1^u, C_2^u, \ldots, C_K^u\}, \cup_k C_k^u = \mathbf{S}, k \in \{1, \ldots, K\}$) based on the sampled subset of genes $\mathbf{G}$ [$\mathbf{G} = \{g_{i_1}, g_{i_2}, \ldots, g_{i_{ngs}}\}$ (where g_{i_h} $(1 \leq h \leq n_{gs})$] represents the hth sampled gene from the original set of genes, and n_{gs} is the total number of sampled genes]. Then, the above process is repeated B times to create a cluster ensemble $\mathbf{I}$ which consists of B clustering solutions I^u ($I^u \in \mathbf{I} = \{I^1, I^2, \ldots, I^B\}$). Finally, a consensus function ϕ is applied to obtain the final clustering solution I^{final} based on the cluster ensemble $\mathbf{I}$ and the pre specified parameter K.

Figure 1 provides an overview of the framework for GCC algorithm. Specifically, GCC first selects a subset of genes from the gene space. Then, the clustering algorithm is applied to partition the samples into K disjoint classes. GCC repeats the first two steps B times to obtain B clustering solutions. In the third step, it constructs a consensus matrix, partitions the consensus matrix by the normalized cut algorithm (Shi and Malik 2000) and obtains the final results. Finally, GCC estimates a suitable K value by a new cluster validation index.

2.1 Subspace generation

In the first step, GCC selects a subset of genes G by random sampling. Specifically, a constant n_{gs} ($n_{\min} \leq n_{gs} \leq n_{\max}$), which represents the number of genes in the subspace, is randomly generated by the following equation:

$$n_{gs} = n_{\min} + \lfloor \nu(n_{\max} - n_{\min}) \rfloor \quad (1)$$

where $\nu(\nu \in [0, 1])$ is a uniform random variable. $n_{\min}$ and $n_{\max}$, which are pre specified parameters by the user, controls the number of dimensions of the subspace. The default settings for $n_{\min}$ and $n_{\max}$ are $0.75n_g$ and $0.85n_g$ respectively, where $n_{\max} \leq n_g$. We have followed the settings of $n_{\min} = 0.75n_g$ and $n_{\max} = 0.85n_g$ in Smolkin and Ghosh (2003).

Then, it selects the gene one by one until n_{gs} genes are obtained. The index of each randomly selected gene is determined as follows:

$$h = \lfloor 1 + \nu' n_g \rfloor \quad (2)$$

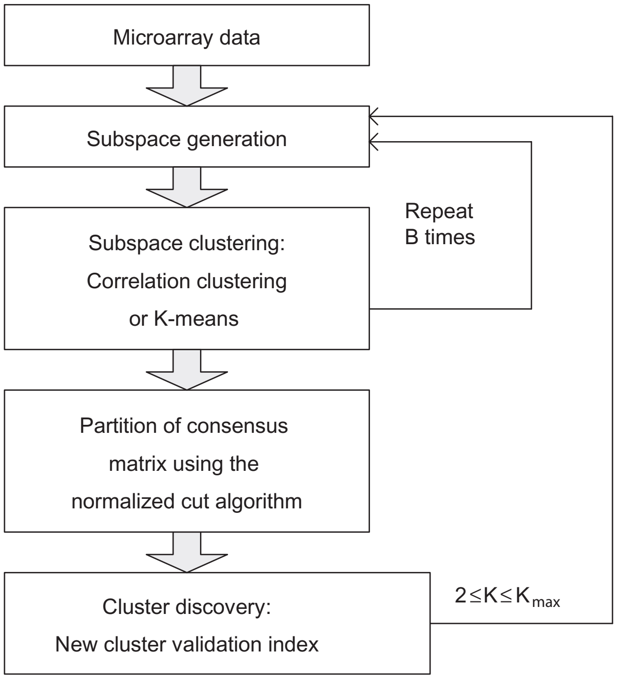

Fig. 1. The framework for GCC.

where h denotes the hth gene in microarray data, n_g is the total number of the genes and ν' is a uniform random variable. Finally, the randomly selected n_{gs} genes are used to construct a subspace.

2.2 Subspace clustering

GCC performs clustering in the selected subspace by two approaches: correlation clustering and K-means. The characteristics of the two approaches are summarized as follows:

Correlation clustering combines correlation analysis and graph partition. We first calculate the correlation matrix (**CM**) whose entries r_{ij} $(i, j \in \{1, \ldots, n_s\}$, n_s is the number of samples) are determined as follows:

$$r_{ij} = \frac{n_{gs}\sum_h s_{i,h}s_{j,h} - \sum_h s_{i,h}\sum_h s_{j,h}}{\sqrt{n_{gs}\sum_h s_{i,h}^2 - (\sum_h s_{i,h})^2} \times \sqrt{n_{gs}\sum_h s_{j,h}^2 - (\sum_h s_{j,h})^2}} \tag{3}$$

where s_i and s_j denotes the ith and jth samples, respectively.

Then, the normalized cut algorithm (Shi and Malik, 2000) is applied to partition the samples into K classes based on the **CM**. We can construct a graph $(G = (\mathbf{S}, \mathbf{CM}))$ whose vertices correspond to the samples, and whose edges denote the correlation between the samples. The normalized cut approach is applied to partition the graph recursively until K classes are obtained. We assume that the normalized cut first partitions the vertex set **S** of the graph **G** into two subsets **P** and **Q**. The cost function $Ncut(\mathbf{P}, \mathbf{Q})$, which represents a disassociation measure between **P** and **Q**, is defined as:

$$Ncut(\mathbf{P}, \mathbf{Q}) = \frac{cut(\mathbf{P}, \mathbf{Q})}{assoc(\mathbf{P}, \mathbf{S})} + \frac{cut(\mathbf{P}, \mathbf{Q})}{assoc(\mathbf{Q}, \mathbf{S})} \tag{4}$$

$$cut(\mathbf{P}, \mathbf{Q}) = \sum_{s_i \in \mathbf{P}, s_j \in \mathbf{Q}} r_{ij} \tag{5}$$

$$assoc(\mathbf{P}, \mathbf{S}) = \sum_{s_i \in \mathbf{P}, s_l \in \mathbf{S}} r_{il} \tag{6}$$

where r_{ij} denotes the weight of the edge between the vertices s_i and s_j, which is the value of the entry in the **CM**. An alternative representation for the above cost measure is as follows:

$$Ncut(\mathbf{P}, \mathbf{Q}) = \frac{\sum_{(v_i>0, v_j<0)} -r_{ij}v_iv_j}{\sum_{v_i>0}\varphi_i} + \frac{\sum_{(v_i<0, v_j>0)} -r_{ij}v_iv_j}{\sum_{v_i<0}\varphi_i} \tag{7}$$

where $\mathbf{v} = [v_1, \ldots, v_{n_s}]^T$ is an n_s-dimensional indicator vector ($n_s = |S|$, $|S|$ is the cardinality of the sample set), and $\varphi_i = \sum_j r_{ij}$ is the sum of the weights from the vertex s_i to all other vertices. In this way, the normalized cut problem can be formulated as an optimization problem, in which $Ncut(\mathbf{v})$ is minimized as follows:

$$\min_{\mathbf{v}} Ncut(\mathbf{v}) = \min_{\gamma} \frac{\gamma^T(\mathbf{D} - \mathbf{CM})\gamma}{\gamma^T\mathbf{D}\gamma} \tag{8}$$

$$\gamma = (1 + \mathbf{v}) - \tau(1 - \mathbf{v}) \tag{9}$$

$$\tau = \frac{\sum_{v_i>0}\varphi_i}{\sum_{v_i<0}\varphi_i} \tag{10}$$

with the constraints:

$$\gamma_i \in \{-\tau, 1\}, \gamma^T\mathbf{DI} = 0 \tag{11}$$

where **D** is an $n_s \times n_s$ diagonal matrix with φ_i $(i \in \{1, \ldots, n_s\})$ on its diagonal, **CM** is an $n_s \times n_s$ symmetric matrix with the elements r_{ij}, **I** denotes the identity matrix and γ_i is the ith component of γ.

Although finding the normalized cut is an NP-complete problem, an approximate discrete solution can be found efficiently by extending the domain of the variables from discrete to continuous. Based on this constraint relaxation, the above optimization problem can be solved using the following generalized eigenvalue system:

$$(\mathbf{D} - \mathbf{CM})\gamma = \lambda\mathbf{D}\gamma \tag{12}$$

where λ denotes the eigenvalue. If γ can take real values, the second smallest eigenvector of the generalized eigenvalue system is the solution to the normalized cut problem (Shi and Malik, 2000).

K-means is a popular clustering algorithm which partitions the samples into K classes by maximizing an objective function $\psi(\mathbf{Z}, \mathbf{C})$.

$$\psi(\mathbf{Z}, \mathbf{C}) = \sum_{k=1}^{K}\sum_{i=1}^{n_s} z_{i,k} \cdot \chi(s_i, c_k) \tag{13}$$

subject to

$$\sum_{k=1}^{K} z_{i,k} = 1 \tag{14}$$

where **Z** is an $n_s \times K$ partition matrix, and $z_{i,k}$ is an indicator variable: If $z_{i,k} = 1$, the sample s_i belongs to the kth cluster. **C** is a set of cluster centers $(\mathbf{C} = \{c_1, \ldots, c_K\})$, and $\chi(s_i, c_k)$ denotes the cosine distance between the sample s_i and the center c_k of the kth cluster, which is defined as:

$$\chi(s_i, c_k) = \frac{<s_i, c_k>}{|s_i| \cdot |c_k|} = \frac{\sum_{j=1}^{n_{gs}} s_{i,j} \times c_{k,j}}{\sqrt{\sum_{j=1}^{n_{gs}} s_{i,j}^2} \times \sqrt{\sum_{j=1}^{n_{gs}} c_{k,j}^2}} \tag{15}$$

where n_{gs} is the number of selected genes in the subspace. We adopt the cosine distance as the distance metric since the cosine distance can eliminate the effect of different magnitudes among the genes.

Through subspace clustering, GCC obtains the predicted labels of the samples. The adjacency matrix **M** is constructed by the predicted labels (Dudoit and Fridlyand 2003), whose elements m_{ij} are defined as:

$$m_{ij} = \begin{cases} 1 & \text{if } y_i = y_j, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \tag{16}$$

where y_i and y_j denote the predicted labels of the samples s_i and s_j, respectively.

2.3 Cluster ensemble

Given the number of classes K, GCC repeats the above two steps (subspace generation and subspace clustering) B times and obtains B clustering solutions $(I_K^1, I_K^2, ..., I_K^B)$ with B adjacency matrices $(\mathbf{M}_K^1, \mathbf{M}_K^2, ..., \mathbf{M}_K^B)$. Then, GCC constructs an $n_s \times n_s$ consensus matrix $(\mathbf{M}_K)$ by merging the adjacency matrix $\mathbf{M}_K^u$ $(u \in \{1, \ldots, B\})$ as follows:

$$\mathbf{M}_K = \frac{1}{B}\sum_{u=1}^{B}\mathbf{M}_K^u \tag{17}$$

where B is the number of adjacency matrices, and the element m_{ij} in the consensus matrix $\mathbf{M}_K$ denotes the frequencies that the ith sample s_i and the jth sample s_j appear in the same class.

Afterwards, GCC constructs a graph $(G_K = (\mathbf{S}, \mathbf{M}_K))$ whose vertices correspond to the samples in $\mathbf{S}$, and whose edges represent the probability that two samples appear in the same class, which is denoted as $m_{K,ij}$ in the consensus matrix $\mathbf{M}_K$. Then, the normalized cut algorithm is used to partition the samples into K classes based on $\mathbf{M}_K$ by minimizing the following cost function:

$$\min_{\mathbf{v}} Ncut(\mathbf{v}) = \min_{\gamma}\frac{\gamma^T(\mathbf{D}-\mathbf{M}_K)\gamma}{\gamma^T\mathbf{D}\gamma} \tag{18}$$

where the definitions of the parameters are the same as the above subsection, except that the consensus matrix $\mathbf{M}_K$ replaces the $\mathbf{CM}$.

2.4 Cluster discovery

We further define an aggregated consensus matrix $\mathbf{R}$ as follows:

$$\mathbf{R} = \frac{1}{(K_{\max}-1)\cdot B}\sum_{K=2}^{K_{\max}}\sum_{u=1}^{B}\mathbf{M}_K^u \tag{19}$$

where $\mathbf{M}_K^u$ denotes the adjacency matrix obtained by the uth clustering solution corresponding to a particular K value. Each entry in the aggregated consensus matrix denotes the probability of two samples appearing in the same class.

GCC further converts the aggregated consensus matrix to a binary matrix $\mathbf{R}^b$:

$$r_{ij}^b = \begin{cases} 1 & \text{if } r_{ij} \geq 0.5, \\ 0 & \text{if } r_{ij} < 0.5, \end{cases} \tag{20}$$

where r_{ij} and r_{ij}^b denotes the entry in the ith-row and the jth column of $\mathbf{R}$ and $\mathbf{R}^b$, respectively. By the same way, GCC converts the consensus matrix $\mathbf{M}_K$ to a binary matrix $\mathbf{M}_K^b$ with entries $m_{K,ij}^b \in \{0, 1\}$.

We propose a new cluster validity index $\zeta(\mathbf{M}_K^b, \mathbf{R}^b)$, known as the Modified Rand Index, as follows:

$$\zeta(\mathbf{M}_K^b, \mathbf{R}^b) = \frac{\sum_{i<j} 1\{m_{K,ij}^b = r_{ij}^b\}}{n_s(n_s-1)} + \frac{1}{K^2} \tag{21}$$

where $m_{K,ij}^b$ and r_{ij}^b are the entries of $\mathbf{M}_K^b$ and $\mathbf{R}^b$, respectively, and n_s is the number of samples. This index balances the degree of agreement between the two matrices $\mathbf{M}_K^b$ and $\mathbf{R}^b$ against the term $\frac{1}{K^2}$, which penalizes a large set of clusters.

The optimal number of classes K^* is selected as follows:

$$K^* = \text{argmax}_{K\in\{2,\ldots,K_{\max}\}}\zeta(\mathbf{M}_K^b, \mathbf{R}^b) \tag{22}$$

In general, cluster validity indices can be categorized into three types: internal indices, external indices and information-theory-based criteria (Sergios and Konstantinos, 2006). The Modified Rand Index can be considered a modified form of the external index. Unlike conventional external indices, the Modified Rand Index does not measure the discrepancy between the ground-truth partition and the partition obtained by the clustering algorithm, but the average of the discrepancies between the partition obtained by the consensus clustering approach and the partitions obtained by the clustering algorithm in a cluster ensemble. It also belongs to the category of indices for optimizing the predictive power/stability of the resulting solution (Handl *et al.*, 05). With the help of the random subspace technique and the consensus clustering approach, the Modified Rand Index can be used to estimate the number of clusters.

3 RESULTS

3.1 Experiment setting

In the experiment, we compare GCC with the consensus clustering (CC) approach described in Monti *et al.* (2003). Based on the adoption of different feature spaces, sampling approaches, clustering algorithms and consensus functions, we consider the following four combinations: GCC_{corr} (subspace + non-resampling + correlation clustering + normalized cut), $GCC_{K-means}$ (subspace + non-resampling + K-means + normalized cut), CC_{HC} [complete space + resampling + hierarchical clustering with average linkage (HC) + HC] and CC_{SOM} [complete space + resampling + self organizing map (SOM) + SOM]. The experiment settings for the CC approaches are the same as those in Monti *et al.* (2003). The experiment settings for the GCC approaches are shown in Table 1.

To evaluate the performance of these different approaches, we adopt the Adjusted Rand Index (ARI) (Milligan and Cooper, 1986) to measure the degree of agreement between different partitions with different numbers of clusters. If the partitions are identical, the value of the ARI is 1. If the partitions are drawn independently from one another, the index takes on an expected value of 0. Let (1) L^T with K^T classes be the true partition of the samples S, and L^P with K^P classes be the predicted partition of the same set of samples S. (2) n_k^T be the number of samples in the k-th class in the partition L^T, n_l^P be the number of samples in the l-th class in the partition L^P and n_{kl}^{TP} be the number of samples in both class k in L^T and class l in L^P. (3) n_s be the number of samples. ARI is computed as follows:

$$\text{ARI} = \frac{\sum_{k=1}^{K^T}\sum_{l=1}^{K^P}\binom{n_{kl}^{TP}}{2} - \eta}{\frac{1}{2}(\rho+\vartheta)-\eta} \tag{23}$$

Table 1. The experiment setting for the GCC approaches (n_{gs} denotes the number of genes of the subspace and B denotes the number of clustering solutions)

Parameter	Value
n_{gs}	$n_{gs} \in \{\lfloor 0.75n_g \rfloor, \ldots, \lfloor 0.85n_g \rfloor\}$
B	500

$$\rho = \sum_{k=1}^{K^T} \binom{n_k^T}{2}, \vartheta = \sum_{l=1}^{K^P} \binom{n_l^P}{2}, \eta = \frac{2\rho\vartheta}{n_s(n_s-1)} \qquad (24)$$

Table 2 provides a summary of the datasets used in our experiments. We generate three synthetic data sets in the space $[0, 1]^d$ (d is the number of features) using a set of Gaussian distributions with randomly selected centers, and with the covariance matrices of all distributions set to $0.25\mathbf{I}$ ($\mathbf{I}$ denotes the identity matrix). The points in Synthetic1, Synthetic2 and Synthetic3 originate from 3, 4, 7 Gaussian clusters, respectively. Synthetic1 is a 1000-gene by 75-sample dataset, in which each class contains 25 samples. Synthetic2 is a 1000-gene by 100-sample dataset, which is used to simulate microarray data with noisy genes. Specifically, 200 noisy genes are included among the 1000 genes. Similar to Synthetic1, each class consists of 25 samples. Synthetic3 is 1000-gene by 100-sample dataset, which is used to simulate microarray data with unequal-size clusters and noisy genes. In addition to including the noisy genes, we vary the number of samples in each class to include 8, 12, 16, 20, 24, 28, 32 samples in the seven classes, respectively. For the real datasets, the data preprocessing procedure is the same as that described in Monti *et al.* (2003). In Table 2, the abbreviation SRBCT stands for small round blue cell tumors, while CNS tumors refers to embryonal tumors of the central nervous system. In the Leukemia dataset, bone marrow samples are obtained from acute leukemia patients at the time of diagnosis, while diagnostic bone marrow samples in the St.Jude dataset are from pediatric patients with acute leukemia. The ranges of K for Synthetic3 and the St.Jude leukemia dataset are both set to $\{2, \ldots, 15\}$, while the ranges of K for the other datasets are set to $\{2, \ldots, 9\}$.

In the experiments, we first explore the relationship between ζ (the Modified Rand Index) and ARI. Then, the optimal K value is selected based on ζ. Finally, we compare the performances of the different approaches.

3.2 Relationship between ARI and ζ

We observe an interesting relationship between ARI and ζ, by which we can estimate the optimal K value. Figure 2 illustrates the change of ARI with respect to K in the different datasets, while Figure 3 shows the change of the Modified Rand Index ζ corresponding to the different K values in the various datasets.

Table 2. Summary of the datasets

Dataset	Source	No. of classes	No. of samples	No. of genes
Synthetic1	by the authors	3	75	1000
Synthetic2	by the authors	4	100	1000
Synthetic3	by the authors	7	140	1000
Breast	Hedenfalk *et al.* (2001)	3	22	351
CNS tumors	Pomeroy *et al.* (2002)	5	48	1000
Leukemia	Golub *et al.* (1999)	3	38	999
Lung cancer	Bhattacharjee *et al.* (2001)	4	197	1000
SRBCT	Khan *et al.* (2001)	4	63	227
St.Jude	Yeoh *et al.* (2002)	6	248	985

An interesting observation is that the trend of the curve for ARI and that of the curve for ζ are very similar. We perform correlation analysis on the curve of ARI and the corresponding curve of ζ in the same dataset. Table 3 lists the results of the correlation analysis between ARI and ζ in all datasets corresponding to the different approaches. All the correlation values in Table 3 are greater than 0.65. This implies that the degree of dependence between ARI and ζ is high.

To estimate the true K value of a dataset, we further study the relationship between ζ, ARI and the K values. If the predicted number of classes is the same as the true number of classes, the value of ARI attains its maximum as shown in Figure 2, since the value of ARI is directly related to the true number of clusters in the dataset. It can also be seen that the peak of the ARI curve in Figure 2 corresponds to that of ζ in the same dataset as shown in Figure 3. In other words, the maximum value of ζ also corresponds to the optimal K value. Given a new set of microarray data with unknown clusters, ARI cannot be calculated, while ζ can be computed when $K_{\max}$ is given. As a result, ζ can be considered as an alternative measure to discover the underlying clusters.

3.3 Experiment results

In general, the GCC approaches GCC_{corr} and $GCC_{K-means}$ outperform the consensus clustering approaches CC_{HC} and CC_{SOM} when applied to the gene expression data as shown in Table 4.

GCC_{corr} outperforms CC_{HC} and CC_{SOM} and correctly discovers the true number of clusters in the three synthetic datasets and the following five real datasets: Breast (BRCA1-mutation-positive samples, BRCA2-mutation-positive samples, Sporadic samples), CNS tumors (medulloblastomas, primitive neuroectodermal tumors, atypical teratoid/rhabdoid tumors, malignant gliomas and normal cerebellum), leukemia (acute myeloid leukemia samples, T-lineage acute lymphoblastic leukemia samples and B-lineage ALL samples), Lung cancer (adenocarcinomas, squamous cell carcinomas, carcinoids and normal lung) and SRBCT [neuroblastoma (NB), nonHodgkin lymphoma (in this case Burkitt's lymphoma (BL)), rhabdomyosarcoma (RMS) and Ewing's family of tumors (EWS)]. In the St.Jude dataset, the estimated number of GCC_{corr} is 5, while the true K value is 6. In fact, the values of ζ corresponding to $K = 5$ and $K = 6$ are nearly the same. When $K = 5$, GCC_{corr} identifies four important leukemia sub-types in the St.Jude dataset: T-lineage ALL, E2A-PBX1, TELAML1 and MLL rearrangements, while the two sub-types BCR-ABL and '*hyperdiploid* > 50' chromosomes are merged into a single class by GCC_{corr}.

The performance of $GCC_{K-means}$ is also better than those of CC_{HC} and CC_{SOM}, and is comparable with that of GCC_{corr}, since $GCC_{K-means}$ successfully discovers the underlying classes in three synthetic datasets and the following five real datasets: Breast cancer, CNS tumors, leukemia, Lung cancer and the St. Jude leukemia dataset (T-lineage ALL, E2A-PBX1, BCR-ABL, TELAML1, MLL rearrangements and '*hyperdiploid* > 50' chromosomes, where ALL denotes acute lymphoblastic leukemia). $GCC_{K-means}$ correctly discovers three classes in the SRBCT dataset: NB, nonHodgkin lymphoma (in this case

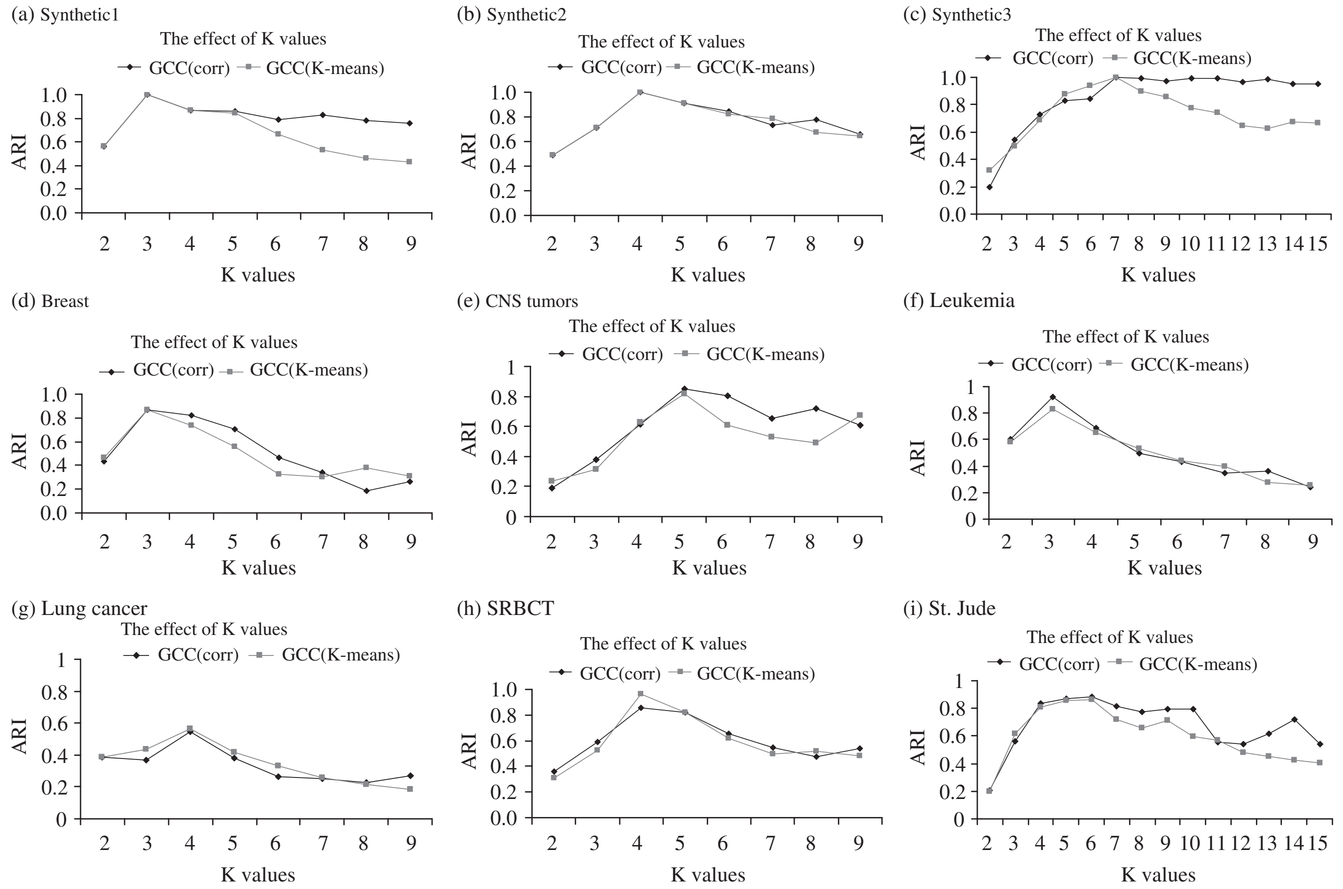

Fig. 2. The change of ARI with respect to different *K* values.

BL) and RMS, while subdividing the class called EWS into two subtypes.

CC_{HC} estimates the correct K value for the following datasets: Synthetic1, Synthetic2, Breast, CNS tumor and SRBCT, while CC_{SOM} discovers the true underlying classes in Synthetic1, Synthetic2, Synthetic3, Breast and CNS tumor.

In general, to address the problems associated with the high-dimensional and small sample nature of gene expression data, together with their high noise levels and large biological variabilities, the GCC approaches adopt the random subspace technique, together with the correlation clustering algorithm or *K*-means, to generate a more diverse set of clustering solutions when compared with existing CC approaches, such that a more accurate solution can be obtained. In addition, the new consensus function in GCC performs better than those in existing CC methods due to the adoption of the normalized cut algorithm, which results in a more accurate partition of the consensus matrix.

Table 5 lists the corresponding values of ARI with respect to the estimated K value in Table 4. The GCC approaches clearly outperform the CC approaches, especially in the Leukemia dataset and the Lung cancer dataset. To provide a further comparison of the results in Table 5, we design a statistical table as shown in Table 6. If the ARI value obtained by the first approach is significantly better/worse [the level of significance is set at a difference of 0.05 in the ARI value, as adopted in Kuncheva and Vetrov (2006)] than that obtained by the second approach in one of the datasets, the win/lose count of the first approach is incremented once. If the difference between the ARI values based on the two approaches is smaller than the level of significance, the tie count is incremented instead. As shown in Table 6, GCC approaches GCC_{corr} and $GCC_{K-means}$ clearly outperform the other approaches and achieve good results in most of the datasets.

In the following experiments, we adopt GCC_{corr} to illustrate the properties of our proposed consensus clustering approaches, since (1) GCC_{corr} achieves good performance in most of the datasets, and (2) there is a higher correlation between ARI and ζ in the case of GCC_{corr} for most of the datasets, as shown in Table 3. To further explore the robustness and the stability of our approach against different numbers of noisy components, we generate three more synthetic datasets (Synthetic4, Synthetic5 and Synthetic6) by varying the number of noisy genes in Synthetic2. The numbers of noisy genes in Synthetic4, Synthetic5 and Synthetic6 are 250, 300 and 350, respectively. The performance of GCC when applied to the three new datasets is illustrated in Figure 4. It can be seen that GCC is robust and stable against the changes in the

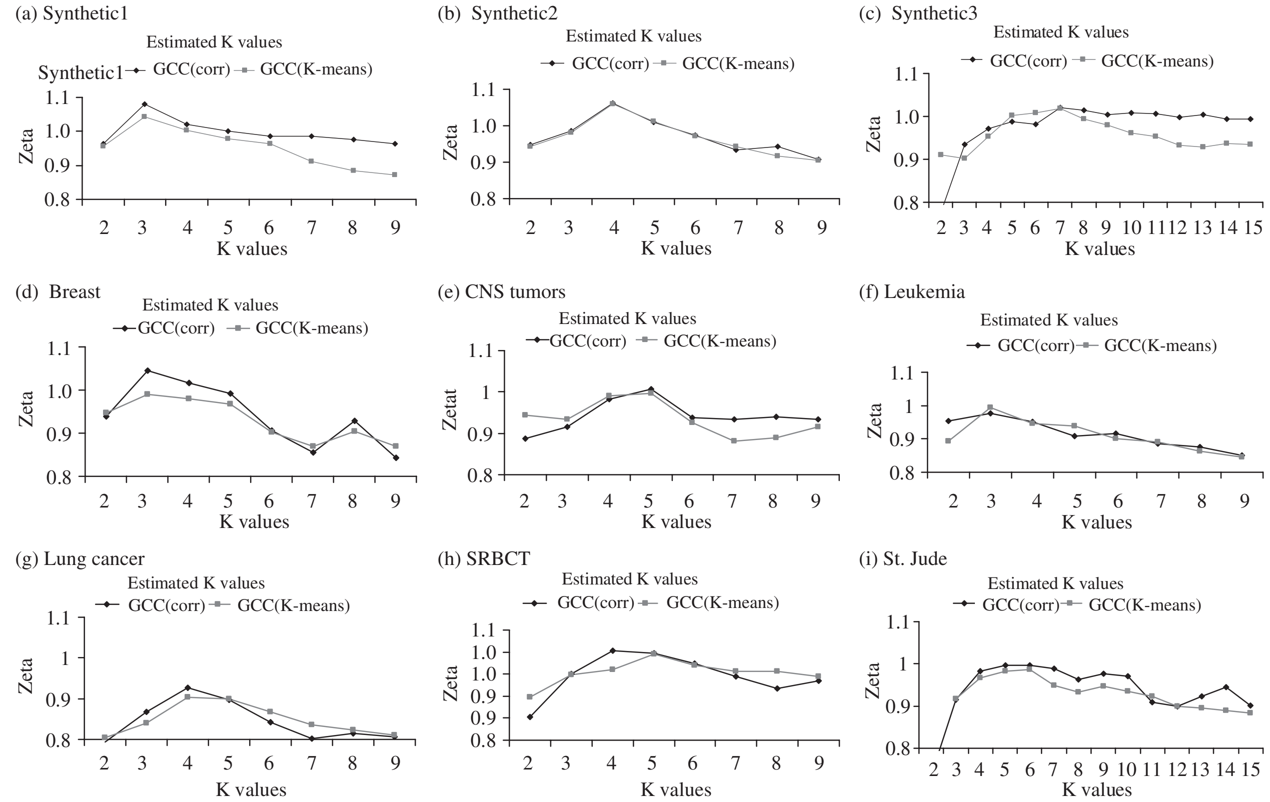

Fig. 3. The change of ζ (named zeta, which is the Modified Rand Index) with respect to different K values.

Table 3. The correlation analysis between ARI and ζ in all datasets

Dataset	GCC_{corr}	$GCC_{K\text{-}means}$
Synthetic1	0.8362	0.9566
Synthetic2	0.7496	0.7958
Synthetic3	0.9623	0.9424
Breast	0.8723	0.9262
CNS tumors	0.7744	0.6542
Leukemia	0.9598	0.9411
Lung cancer	0.7536	0.6824
SRBCT	0.9532	0.7274
St.Jude	0.9783	0.9150

Table 4. Estimated optimal K value by different approaches

Dataset	GCC_{corr}	$GCC_{K\text{-}means}$	CC_{HC}	CC_{SOM}	K_{true}
Synthetic1	3	3	3	3	3
Synthetic2	4	4	4	4	4
Synthetic3	7	7	6	7	7
Breast	3	3	3	3	3
CNS tumors	5	5	5	5	5
Leukemia	3	3	5	4	3
Lung cancer	4	4	5	5	4
SRBCT	4	5	4	5	4
St.Jude	5	6	5	5	6

Table 5. The corresponding values of ARI w.r.t the estimated K-values

Dataset	GCC_{corr}	$GCC_{K\text{-}means}$	CC_{HC}	CC_{SOM}
Synthetic1	1	1	1	1
Synthetic2	1	1	1	1
Synthetic3	1	1	0.968	0.976
Breast	0.866	0.867	0.756	0.854
CNS tumors	0.658	0.718	0.549	0.429
Leukemia	0.831	0.831	0.648	0.721
Lung cancer	0.544	0.562	0.310	0.233
SRBCT	0.858	0.819	0.864	0.772
St.Jude	0.873	0.860	0.948	0.825

Table 6. Statistical results by comparing different approaches

	GCC_{corr}	$GCC_{K\text{-}means}$	CC_{HC}	CC_{SOM}
GCC_{corr}	–	0/8/1	4/4/1	4/5/0
$GCC_{K\text{-}means}$	1/8/0	–	4/4/1	4/5/0
CC_{HC}	1/4/4	1/4/4	–	4/3/2
CC_{SOM}	0/5/4	0/5/4	2/3/4	–

The entry e_{ij} in the table denotes the count for Win/Tie/Lose. If the approach corresponding to row i is significantly better (at a level of significance 0.05) than the approach corresponding to column j, the first approach wins, and vice versa. Otherwise, the two approaches tie with each other.

number of noisy components, and can successfully estimate the four clusters in the datasets.

We further investigate the effect of the maximum K value ($K_{\max}$) on the Modified Rand Index, and the relationship between $K_{\max}$ and the sparseness of the discretized aggregated consensus matrix $\mathbf{R}^b$. The GCC algorithm GCC_{corr} is applied to the Synthetic2 dataset and the Leukemia dataset using different $K_{\max}$ values. It is observed that when $K_{\max}$ increases, GCC_{corr} still correctly estimates the number of clusters in the Synthetic2 dataset and the Leukemia dataset, as indicated in Tables 7 and 8. In addition, we observe that the ARI associated with GCC_{corr} is not affected by $K_{\max}$, while the value of the peak ζ decreases slightly when $K_{\max}$ increases. In general, when $K_{\max}$ is large, the number of entries in $\mathbf{R}^b$ whose values exceed the threshold in Equation (20) are small, which leads to sparseness of the binary matrix. We further observe that this sparseness causes a slight decrease of the peak ζ value for both datasets. However, this does not affect the capability of ζ to identify the correct number of clusters, since the set of ζ values corresponding to the different K values change in such a way that the position of the maximum point on the ζ versus K curve is maintained, as indicated in Figures 5 and 6. Although the peaks of the curves in Figures 5 and 6 become less prominent as $K_{\max}$ increases, our proposed approach still correctly estimates the optimal K value from each dataset.

4 CONCLUSION

In this article, we investigate the problem of class discovery in gene expression data. The major contribution of this article is in the design of a new framework, known as GCC, to discover the underlying classes of the samples in gene expression data.

Table 7. The effect of $K_{\max}$ on the Synthetic2 dataset

$K_{\max}$	6	7	8	9	10	11
Estimated K-value	4	4	4	4	4	4
ζ	1.0635	1.063	1.0623	1.0615	1.0589	1.0577
ARI	1	1	1	1	1	1

Table 8. The effect of $K_{\max}$ on the Leukemia dataset

$K_{\max}$	6	7	8	9	10	11
Estimated K-value	3	3	3	3	3	3
ζ	1.061	1.04	1.036	1.01	0.973	0.962
ARI	0.831	0.831	0.831	0.831	0.831	0.831

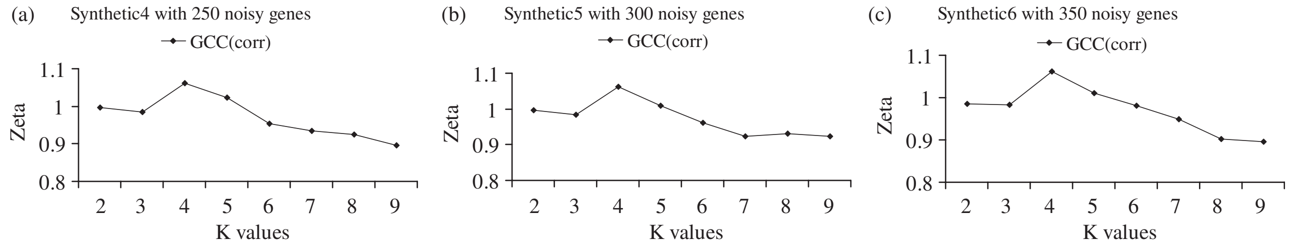

Fig. 4. Effect of different numbers of noisy genes.

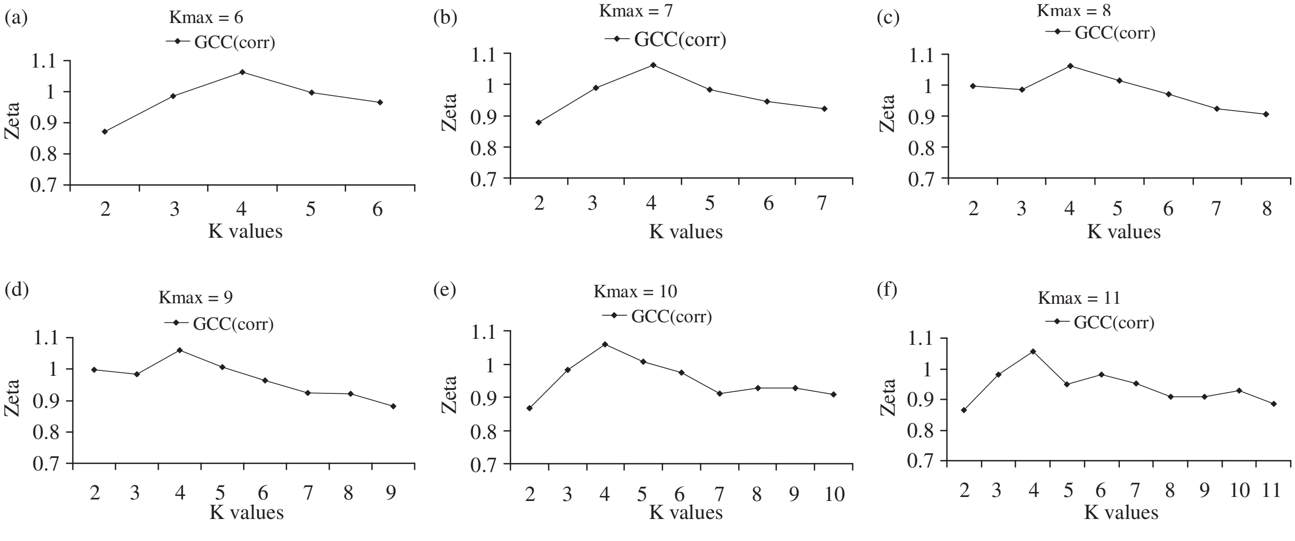

Fig. 5. The relationship between ζ and $K_{\max}$ (Synthetic2 dataset).

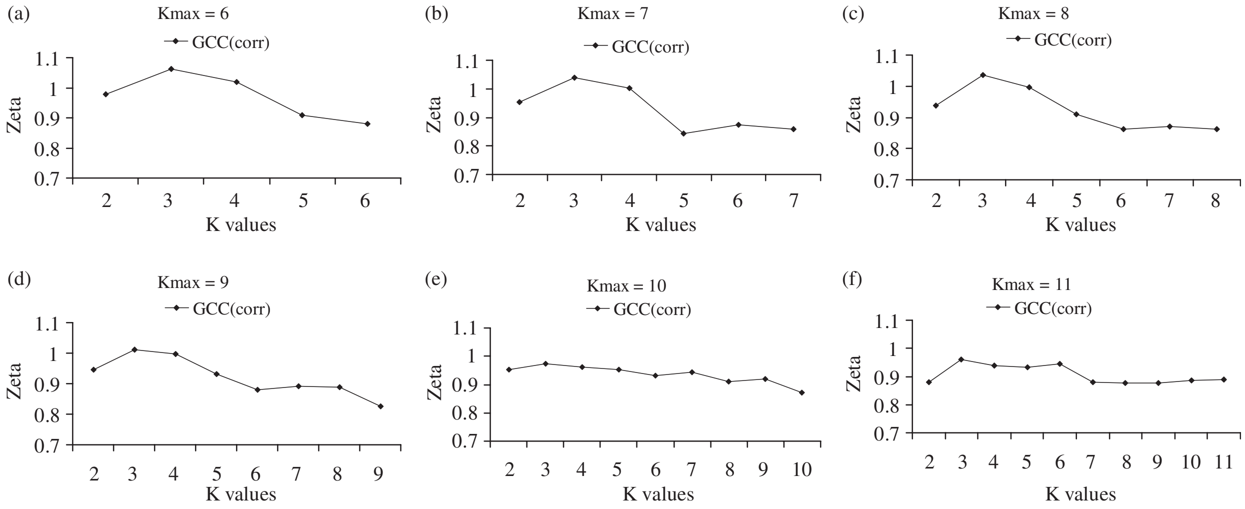

Fig. 6. The relationship between ζ and $K_{\max}$ (Leukemia dataset).

Our new approach can successfully estimate the true number of classes for the datasets in our experiments. In addition, based on our experiment results, we also observe that our new approach outperforms the consensus clustering approaches proposed in Monti *et al.* (2003) when applied to the characterization of gene expression data.

Conflict of Interest: none declared.

REFERENCES

Alizadeh,A.A. *et al.* (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Baldi,P. and Hatfield,G.W. (2002) *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*, Cambridge University Press, Cambridge.

Bertoni,A. and Valentini,G. (2005) Ensembles based on random projections to improve the accuracy of clustering algorithms. *Neural Nets, (WIRN 2005), LNCS*, **3931**, 31–37.

Bertoni,A. and Valentini,G. (2006) Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses. *Artif. Intell. in Med.*, **37**, 85–109.

Bertoni,A. and Valentini,G. (2007a) Model order selection for biomolecular data clustering. *BMC Bioinformatics*, **8** (Suppl. 2), S7.

Bertoni,A. and Valentini,G. (2007b) Randomized Embedding Clustering Ensembles for gene expression data analysis. *In SETIT 2007 – Proceedings of IEEE International Conference on Sciences of Electronic*, Technologies of Information and Telecommunications, Hammamet, Tunisia.

Bhattacharjee,A. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinomas sub-classes. *Proc. Natl Acad. Sci.*, **98**, 13790–13795.

Datta,S. and Datta,S. (2006) Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics*, **7**, 397.

Dudoit,S. and Fridlyand,J. (2002) A prediction-based resampling method to estimate the number of clusters in a dataset. *Genome Bio.*, **3**, 0036.1–0036.21.

Dudoit,S. and Fridlyand,J. (2003) Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, **19**, 1090–1099.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression. *Science*, **286**, 531–537.

Grotkjaer,T. *et al.* (2006) Robust multi-scale clustering of large DNA microarray datasets with the consensus algorithm. *Bioinformatics*, **22**, 58–67.

Handl,J. *et al.* (2005) Computational cluster validation in post-genomic data analysis Bioinformatics. *Bioinformatics*, **21**, 3201–3212.

Hedenfalk,I. *et al.* (2001) Gene-expression profiles in hereditary breast cancer. *New Engl. J. of Med.*, **344**, 539–548.

Khan,J. *et al.* (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Kuncheva,L.I. and Vetrov,D.P. (2006) Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE Trans. Pattern Anal. Mach. Intell.*, **28**, 1798–1808.

Mc Shane,L.M. (2002) Method for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, **18**, 1462–1469.

Milligan,G. and Cooper,M. (1986) A study of the comparability of external criteria for hierarchical cluster analysis. *Multivar. Behav. Res.*, **21**, 441–458.

Monti,S. *et al.* (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*, **52**, 91–118.

Pomeroy,S. *et al.* (2002) Gene expression-based classification and outcome prediction of central nervous system embryonal tumors. *Nature*, **415**, 436–442.

Sergios,T. and Konstantinos,K. (2006) *Pattern Recognition*. 3rd edn. Academic press, Elsevier, UK pp. 733–765.

Shi,J. and Malik,J. (2000) Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **22**, 888–905.

Strehl,A. and Ghosh,J. (2002) Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, **3**, 583–617.

Smolkin,M. and Ghosh,D. (2003) Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, **4**, 36.

Su,A.I. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci.*, **99**, 4465–4470.

Topchy,A. *et al.* (2005) Clustering ensembles: models of consensus and weak partitions. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 1866–1881.

Valentini,G. (2006) Clusterv: a tool for assessing the reliability of clusters discovered in DNA microarray data. *Bioinformatics*, **22**, 369–370.

Valentini,G. (2007) Mosclust: a software library for discovering significant structures in bio-molecular data. *Bioinformatics*, **23**, 387–389.

Wigle,D. *et al.* (2002) Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res.*, **62**, 3005–3008.

Yeoh,E.-J. *et al.* (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.

Yeung,K.Y. *et al.* (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.