PARAMETRIC VALIDITY INDEX OF CLUSTERING FOR MICROARRAY GENE EXPRESSION DATA

Rui Fa and Asoke K Nandi

Signal Processing and Communications Research Group
Department of Electrical Engineering and Electronics, The University of Liverpool, L69 3GJ, UK

ABSTRACT

An important area of genomic signal processing is microarray gene expression data analysis, which employs clustering algorithms to group individual genes or samples in a population. Due to the non-unique nature of clustering, the cluster validation is necessary for evaluating the results of clustering algorithms. In this paper, we propose a parametric validity index (PVI) which employs two tunable parameters $\alpha$ and $\beta$ to control the proportions of objects being taken into account to calculate the dissimilarities. There are two advantages of the proposed PVI: on one hand, its computational complexity is low, and on the other hand, it has flexibility of tuning the parameters to meet different datasets, especially the microarray datasets. The PVI can be averaged over a range of values of $\alpha$ and $\beta$. We investigate the new PVI for assessing five clustering algorithms in four microarray datasets. The experimental results appear to suggest that the proposed PVI has relatively robust performance and provides fairly accurate judgements.

1. INTRODUCTION

Microarray is a comparatively new technology to investigate the expression levels of thousands of genes simultaneously. The major task of microarray data analysis is to identify genes with similar biological functions and their interactions [1, 2]. Gene expression data measured by microarrays are preprocessed using image analysis techniques and then are organized in a matrix where each row represents a gene and each column represents a sample. Clustering is a useful exploratory technique for the analysis of gene expression data. A number of clustering algorithms have been developed in the past, including the $k$-means, the $k$-medoids, the hierarchical clustering (HC), the self-organizing map (SOM) and the model-based clustering (MClust) [1, 3, 4].

The goal of the clustering analysis is to group individual genes or samples in a population within which the objects are more similar to each other than those in other clusters. Although there are many commonly used clustering algorithms, there is no existing guideline to guarantee that one clustering algorithm, which works well in one dataset, can perform also well in a different dataset. Even the same algorithm with different parameter settings or different initialization methods usually produce different clustering results. Thus, the task of assessing the clustering algorithms can be as important as the clustering algorithms themselves.

Since clustering is unsupervised classification, it is more difficult to assess than a supervised approach. The procedure for evaluating the results of a clustering algorithm is known as clustering validation. In general terms, clustering validation can be classified according to two different methods. On one hand, based on the approaches how to investigate cluster validity, there are three classes, namely external criteria, internal criteria and relative criteria [5, 6]; on the other hand, based on the methods how the validity indices are calculated, clustering validation can be classified into three classes. Class one includes cost-function based indices, class two includes density-based indices and class three includes geometric approaches [10]. Among these validity approaches, we are more interested in the relative criterion and geometric indices because of their simplicity and low computational load. There are many validity indices, which belong to both relative criterion and geometric indices, proposed to assess clustering results, including Dunn’s index (DI) [7], the $I$-index (II) [8], the Calinski Harabasz (CH) index [9], the geometrical index (GI) [10] and the validity index $V_i$ [11]. The basic principle behind these methods is to calculate the ratio of the intra-cluster scatter to the inter-cluster separation. However, none of these widely adopted methods can be claimed to work well for all types of data. In other words, the fact that the existing validity indices have relatively good performance in assessing the clustering algorithms in one dataset, say in one microarray dataset, does not mean that they will have good assessing performance in other datasets. It motivates us to propose and investigate a relatively robust validity index in assessing clustering algorithms for different microarray datasets.

In this paper, we propose a parametric validity index (PVI) which is one of geometric indices to calculate the ratio of the inter-cluster dissimilarity to the intra-cluster dissimilarity. We introduce two tunable parameters $\alpha$ and $\beta$ in the new index to control the proportions of objects being taken into account to calculate the dissimilarities. There are two advantages of the proposed PVI: on one hand, its computational complexity is low, and on the other hand, it has flexibility of tuning the parameters to meet different datasets, especially the microarray datasets. The PVI also can be averaged over a range of values of $\alpha$ and $\beta$ to avoid its dependency on the dataset structure. We investigate the new PVI for assessing five clustering algorithms, namely the $k$-means, the $k$-medoids, the HC, the SOM and the MClust, in three datasets, including a synthetic microarray gene expression data, Leukaemia and Yeast cell cycle datasets. The experimental results appear to suggest that the proposed PVI has relatively robust performance and provides fairly accurate judgements.

The rest of the paper is organised as follows, Sec. 2 presents the principle of the proposed PVI, Sec. 3 briefly reviews the clustering algorithms, namely the $k$-means, the $k$-medoids, the HC, the SOM and the MClust algorithms, and presents a variety of existing validity indices. In Sec. 4, the datasets, which help in demonstrating and examining the clustering algorithms and validity indices, is introduced. Sec. 5 presents the experimental results and finally...
the discussions and conclusions are made in Sec. 6.

2. PROPOSED PARAMETRIC VALIDITY INDEX

![Diagram showing different spaces: Inner space, Inter outer space, Intra outer space.](image)

Fig. 1. Illustration of the proposed PVI.

In this section, we detail the principle of the proposed Parametric Validity Index (PVI). The PVI belongs to the class of geometric indices to calculate the ratio of the intra-cluster dissimilarity to the inter-cluster dissimilarity.

2.1. \( \alpha \) and \( \beta \)

We introduce two tunable parameter \( \alpha \) and \( \beta \) in the new index to control the proportions of objects that are involved in the calculation of the intra-cluster dissimilarities and the inter-cluster dissimilarities. For the sake of simplification, let us look at a 2-D plane first, as depicted in Fig. 1. The objects marked by ‘×' and ‘o' belong to two different clusters. In this case, the intra-cluster dissimilarity is represented by the distance of two objects within one cluster while the inter-cluster dissimilarity is represented by the distance of two objects in different clusters. It is not necessary to take all objects into account to calculate the index, thus, in order to calculate the dissimilarities efficiently, we must determine those objects at both ends of the dissimilarities. We define the inner space as the objects in the cluster under test, which are used to calculate both intra-cluster and inter-cluster dissimilarities, as the objects in the area marked yellow in Fig. 1. The outer ends objects in the same cluster are called the intra outer space, which is the area marked navy-blue in Fig. 1, and the outer ends objects in the different clusters are called the inter outer space, which is the area marked brick-red in Fig. 1. Thus, let us define \( N_k^a, Na_k^a, Ne_k^a \) to denote the numbers of objects in the inner space, the intra outer space and the inter outer space, respectively, for the \( k \)-th cluster. The fractions, \( \alpha \) and \( \beta \), are used to control \( N_k^a, Na_k^a, Ne_k^a \), which can be expressed as

\[
N_k^a = \lceil\alpha N_k\rceil, \quad Na_k^a = \lceil\beta N_k\rceil, \quad Ne_k^a = \lceil\beta(N - N_k)\rceil,
\]

where \( N_k \) is the number of the objects in the \( k \)-th cluster, \( N \) is the number of all objects in the dataset and \( \lceil \cdot \rceil \) is the ceiling operator. Both \( \alpha \) and \( \beta \) can be chosen from the range of \((0, 1]\). Thus, \( N_k^a, Na_k^a, Ne_k^a \) can be any integer within the range of \([1, N_k], [1, N_k], \) and \([1, N - N_k], \) respectively.

There are a few steps to calculate the PVI. Firstly, we need to form the subset \( A_k \) for the inner ends. For each object in the \( k \)-th cluster, we can obtain a total dissimilarity by the summation of dissimilarities between it and all others in the \( k \)-th cluster, as

\[
D_n = \sum_{m=1, m\neq n}^{N_k} D(x_n, x_m), \quad n = 1, ..., N_k.
\]

where \( D(\cdot, \cdot) \) denotes the calculation of the dissimilarity of two objects. Note that there are many methods for dissimilarity measure (or similarity measure), such as Euclidean distance, Pearson’s correlation coefficient and so on [1]. In this paper, we calculate the dissimilarity based on Pearson’s correlation coefficient. Pearson’s correlation coefficient is defined as

\[
PCC(x_n, x_m) = \frac{\sum_{d=1}^{n}(x_{nd} - \mu_n)(x_{md} - \mu_m)}{\sqrt{\sum_{d=1}^{n}(x_{nd} - \mu_n)^2}\sqrt{\sum_{d=1}^{n}(x_{md} - \mu_m)^2}},
\]

where \( \mu_n \) and \( \mu_m \) are means for \( x_n \) and \( x_m \), respectively. Thus, the dissimilarity is obtained by

\[
D(x_n, x_m) = 1 - PCC(x_n, x_m).
\]

We pick \( N_k^a \) most central located objects, which have relatively smaller total dissimilarities, from the \( k \)-th cluster to form \( A_k \), which is \( \{a_k^a | a = 1, ..., N_k^a\} \). Secondly, for each object in the inner space \( a_k^a \), we need to form subsets \( B_k^a \) and \( C_k^a \) for the objects in the intra outer space and the inter outer space, respectively. The objects in \( B_k^a \), denoted as \( \{b_{k,a}^b | b = 1, ..., Na_k^a\} \), are those \( Na_k^a \) objects in the \( k \)-th cluster, which are farthest from \( a_k^a \), while the objects in \( C_k^a \), denoted as \( \{c_{k,a}^c | c = 1, ..., Ne_k^a\} \), are those \( Ne_k^a \) objects in the clusters but the \( k \)-th cluster, which are closest to \( a_k^a \). Afterwards, we need to do some calculations as follows:

\[
Da_k^a = \sum_{b=1}^{Na_k^a} D(a_k^a, b_{k,a}^b),
\]

\[
De_k^a = \sum_{c=1}^{Ne_k^a} D(a_k^a, c_{k,a}^c),
\]

where \( Da_k^a \) denotes a normalized intra-cluster dissimilarity and \( De_k^a \) denotes a normalized inter-cluster dissimilarity for the \( \alpha \)-th inner end in the \( k \)-th cluster. Finally, the PVI can be obtained by

\[
PVI(K, \alpha, \beta) = \sum_{k=1}^{K} \sum_{a=1}^{N_k} \left( \frac{Da_k^a}{De_k^a} \right).
\]

Obviously, there are two advantages for the new index. On one hand, the PVI only involves some of objects into the calculation rather than all objects, which reduces the computational complexity depending on the settings of \( \alpha \) and \( \beta \). On the other hand, the flexibility brought by two tunable parameters makes the index more robust and the PVI may meet many different datasets. However, fairly speaking, because the settings of the parameters are largely dependent on the dataset structure, the flexibility may be a disadvantage when the dataset structure is not available a priori. Thus, we explore more to develop an averaged PVI to overcome this shortcoming, which is presented in the next part.
2.2. Averaged PVI

As presented in the previous part, we obtain the PVI that is a function of \( \alpha \) and \( \beta \), namely \( \text{PVI}(\alpha, \beta) \), where \( \alpha \) and \( \beta \) can be any value in the range of (0,1]. If we evenly pick a number of values in (0,1] for \( \alpha \) and \( \beta \), for each pair of \( \alpha \) and \( \beta \), there will be a PVI indicating the validation value. We can obtain the averaged PVI (APVI) over a number of both \( \alpha \) and \( \beta \), which is mathematically written as

\[
\text{APVI} = \frac{\sum_{\alpha} \sum_{\beta} \text{PVI}(\alpha, \beta)}{N_{\alpha}N_{\beta}},
\]

where \( N_{\alpha} \) and \( N_{\beta} \) are the numbers of \( \alpha \) and \( \beta \), respectively. As we know, the index value indicates the confidence level on the given clustering result. The averaging of index values give us an impression of the overall confidence level of the clustering result.

3. CLUSTERING ALGORITHMS AND VALIDITY INDICES

In this section, we review five popular clustering algorithms, namely the \( k \)-means, the HC, the SOM and the model-based clustering (MClust) algorithms, as well as five validity indices, namely the DI, the II, the GI, the \( V_t \) and the \( CH \) indices. There are many other types of clustering algorithms and validity indices, but due to the limited space, we will not investigate all of them. The interested reader is advised to consult [1] and the references therein.

3.1. Clustering Algorithms

The \( k \)-means is one of the most popular clustering algorithms [1]. The basic idea is to produce \( K \) clusters from a set of \( n \) objects so as to minimise the dissimilarities between the memberships and the cluster center within a cluster. The dissimilarity is defined as the distance between two objects, which could be the Euclidean distance, the Pearson’s correlation coefficient, the Mahalanobis distance, etc. For the sake of simplicity, throughout the paper, we employ the Euclidean distance as the dissimilarity measure. Thus, it leads to the following objective function:

\[
E = \sum_{j=1}^{K} \sum_{x_i \in C_j} ||x_i - u_j||^2,
\]

where \( C_j \) denotes the \( j \)-th cluster, \( u_j \) is the center of the cluster \( j \) which is the mean of objects in \( C_j \), and \( x_i \) is the observations to be clustered. The \( k \)-means attempts to minimise \( E \) iteratively. It moves every objects to the closest cluster. Once the object is moved from one cluster to another cluster, the centres of both clusters need to be recalculated until the \( k \)-means converges.

The \( k \)-medoids is one of family members of partitioning algorithms, which is an extension of the \( k \)-means. The \( k \)-medoids is designed to handle the outliers efficiently. Instead of using the means, it chooses the medoids to represent the cluster centres. A medoid is the most centrally located object within a cluster, whose total distance to all other objects intra-cluster is the shortest. Similar to the \( k \)-means, the \( k \)-medoids keeps updating the medoids if cluster membership changes until the process converges.

In contrast to partition-based clustering, which directly divides the data set into disconnected parts, HC is path-based clustering algorithm, which generates a hierarchical series of nested and connected clusters. This nested cluster structure can be graphically represented in a tree by dendrogram. There are two approaches to implement the HC: one is called agglomerative which initially regards each data object as an individual cluster and merges the closest pair of clusters at each step, until all the groups are merged into one cluster; The other is called divisive, which starts with one cluster containing all the data objects and splits singleton clusters at each step. In this paper, we consider agglomerative HC with average linkage.

The SOM is one of most commonly used neural network algorithms [3]. It belongs to the category of competitive learning networks. The SOM is based on unsupervised learning, which chooses a geometry of nodes, say a \( m \times n \) grid (may be higher dimensional grid), to map the objects into \( k \)-dimensional space. Each node in the SOM contains two kinds of position information: one is the node weight \( w \) which is used to map the objects and the other is the node position \( p \) which indicates the position of the node in the grid. The node weights are initialized randomly. On the subsequent iterations, the data object \( x_i \) is simultaneously compared with the weights of all nodes at each successive instant of time \( t \), then the best matching node \( w_m(t) \) is moved towards the data \( x_i \) by

\[
w_m(t) = w_m(t-1) + \alpha(t)[x_i - w_m(t-1)], 0 < \alpha(t) < 1,
\]

where \( \alpha(t) \) is the step size, which shrinks as the number of iterations increases, and the nodes within the neighbourhood of the best matching node will also adjust their weights moving closer to the object in the similar fashion to (9). The radius of the neighbourhood also shrinks over time.

The model-based clustering (MClust) approaches assume that the gene expression data is a finite mixture of multivariate normal distributions and provide a statistical framework to estimate each component corresponding to a different cluster [4]. The MClust employs the Expectation-Maximization (EM) algorithm, which iterates between Expectation (E) steps and Maximization (M) steps. The probabilities that an observation belongs to each component \( \Omega = \{ \theta_k | k = 1, ..., K \} \), which is conditioned on the parameters \( \Theta = \{ \theta_k | k = 1, ..., K \} \), are estimated from the data in the E step. The parameters \( \Theta \) are estimated with the given probabilities \( \Omega \) as to maximize the likelihood of complete data.

3.2. Validity Indices

In this section, we list five validity indices which we are going to investigate and compare with our proposed PVI. The first is the \( V_t \) [11]. The validity index \( V_t \) is the ratio of the inter-cluster separation and the intra-cluster scatter measures, which is mathematically expressed as

\[
V_t(K) = \frac{\sum_{i=1}^{K} I_{e_i}}{\sum_{i=1}^{K} I_{a_i}},
\]

where \( K \) is the number of clusters. The \( V_t \) employs \( I_{a_i} \), the largest dissimilarity of the minimum spanning tree (MST) for cluster \( i \), as the intra-cluster scatter and \( I_{e_i} = \min_{1 \leq j \neq i} I_{e_{ij}} \), where \( I_{e_{ij}} \) is the largest dissimilarity of the MST for cluster \( i \) and cluster \( j \), as the inter-cluster separation.

The second is the \( DI \) [7], which is written as

\[
DI(K) = \min_{1 \leq i \leq K} \left\{ \min_{1 \leq j \leq K} \left\{ \frac{\delta(C_i, C_j)}{\max_{1 \leq k \leq K} |\Delta(C_k)|} \right\} \right\},
\]
Table 1. Description of the datasets employed in the experiments.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Clusters</th>
<th>Objects (Genes)</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic gene data [4, 14]</td>
<td>3</td>
<td>430</td>
<td>24</td>
</tr>
<tr>
<td>Leukemia [17]</td>
<td>3</td>
<td>999</td>
<td>38</td>
</tr>
<tr>
<td>Yeast cell cycle [12, 13]</td>
<td>5</td>
<td>384</td>
<td>17</td>
</tr>
</tbody>
</table>

where $\delta(C_i, C_j) = \min \|x_i - x_j\|_2$ is the maximum distance between cluster $i$ and cluster $j$, $\Delta(C_k)$ is the largest intra-cluster separation of cluster $k$. The third is the $H$ [8], which is written as

$$I(K) = \left( \frac{1}{K} \times \frac{E_1}{E_K} \times D_K \right)^P, \quad (12)$$

where $E_1 = \sum_j \|x_j - \mu\|_2$ where $\mu$ is the centroid of the whole dataset, $E_K = \sum_{k=1}^{K} \sum_{j \in C_k} \|x_j - u_k\|_2$, $D_K = \min_{i,j} \|u_i - u_j\|_2$ and power $P$ is constant, which is 2 in our experiments. The next one is the GI [10], which is expressed as

$$GI(K) = \max_{1 \leq K \leq n} \left\{ \frac{(2 \sum_{m=1}^{M} \sqrt{\lambda_{mk}})^2}{\min_{1 \leq j \leq K} \|u_i - u_j\|_2} \right\}, \quad (13)$$

where $M$ is the number of dimensions, $\lambda_{mk}$ are the eigenvalues of the covariance matrix of the $k$-th cluster. Note that the closest GI value to zero suggests the best number of clusters. The last one is the $CH$ [9] which is given by

$$CH(K) = \frac{\sum_{k=1}^{K} n_k \|u_k - \mu\|^2}{K - 1} / \left[ \frac{\sum_{k=1}^{K} \sum_{j=1}^{n_k} \|x_i - u_k\|^2}{n - K} \right], \quad (14)$$

where $n_k$ is the number of memberships in the cluster $k$ and $n$ is the total number of the objects.

4. DATASETS

The three microarray datasets are explored in this paper, including one synthetic gene data, leukemia data and Yeast cell cycle data, are listed in Table 1. The synthetic data set, which we investigate here, models gene expression data with cyclic behavior. Classes are modelled as genes that have peak times over the time courses as presented in [4, 14]. In this work, we generate a dataset with 450 genes and 24 samples, which has five clusters. The leukemia dataset [17] consists of 38 bone marrow samples obtained from acute leukemia patients at time of diagnosis. The samples include 11 acute myeloid leukemia (AML) samples, 8 T-lineage acute lymphoblastic leukemia (ALL) samples and 19 B-lineage ALL samples. There are 999 genes in the dataset. The yeast cell cycle data, which is available at http://faculty.washington.edu/kayee/model/, is also investigated. The yeast cell cycle dataset was published by Cho et al. [12]. It consisted of more than 6000 genes over 17 time points taken at 10 minutes intervals, where 383 genes were identified and these demonstrated consistent periodic changes in transcript level. It was commonly believed that the time course was divided into early G1, late G1, S, G2, and M phases, those 383 genes would peak at one of the five phases. In [4], a subset with 384 genes was investigated. Although we cannot tell what is the difference between these two subsets, those five phases were claimed in common. In this paper, we investigate the the same subset with 384 genes as [4].

Fig. 2. The principal component analysis for Leukemia dataset. Each axes represent one of three largest components.

5. EXPERIMENTAL RESULTS

We split this section into two parts. In the first part, we present the results of the proposed PVI with different parameters for three different microarray gene data sets. In the second part, we compare the averaged PVI to other existing validity indices using the same datasets as those in part one. Before investigating the validity indices, we study two real microarray gene datasets to find out the clustering structures. The principal component analysis (PCA) of the leukemia dataset in a 3-D plot is depicted in Fig. 2. The three axes represent the first, the second and the third principal components (PCs). The data is standardised and there are clearly three clusters, which is consistent with the description in [17]. The 3D PCA of the yeast dataset is shown in Fig. 3 (a) [16]. The 2-D projections on the x-y, x-z, and y-z planes are shown in Fig. 3 (b)-(d), respectively. The SOM is employed to separate the dataset into clusters. Although [4, 10] mentioned that there are five clusters in the dataset, obviously, the cluster marked with the pentagram,
which has only ten members, should be a separate cluster representing the sixth phase in terms of its geometrical position. Thus, the clustering with six clusters is more reasonable than the clustering with five clusters.

We employ the deterministic initialization for the \(k\)-means and the \(k\)-medoids algorithm as presented in [15]. For the SOM algorithms, which are initialised randomly, we obtain the validity indices by averaging 100 experimental results while for each of the \(k\)-means, \(k\)-medoids, the HC and the MClust, we only take one set of experimental result due to its deterministic nature.

### 5.1. PVI with Parameters

We investigate the PVI along with other existing indices assessing five clustering algorithms for three microarray gene data sets. We choose 3 values for both \(\alpha\) and \(\beta\), which are 0.1, 0.3 and 0.5. That is, there are nine pairs of \(\alpha\) and \(\beta\) in the investigation. The suggested best numbers of clusters of all five clustering algorithms are collected and shown as a list in Table 2. The results show that for the synthetic gene dataset, the most of PVIs with different parameters indicate that five is the best number of clusters for all five clustering algorithms. Note that the PVI with smaller \(\beta\) (\(\leq 0.3\)) normally provides accurate indication in this case. The \(DI\) and the \(CH\) indices also indicate the five clusters for all clustering algorithms. The \(V_I\) suggests that five is the best for the \(k\)-means, \(k\)-medoids and the MClust, but four is the best for the HC and the SOM, which is not accurate. The \(GI\) and the \(II\) indices give completely wrong indications. For the leukemia dataset, all validity indices indicate that the HC does not provide a good clustering result. In this dataset, note the PVIs with smaller \(\alpha\) and \(\beta\) (\(\leq 0.3\)) give wrong indications, which is opposite to the result in the synthetic dataset. That is, for the different datasets, the ranges of the parameters to indicate the correct results are different. We also note that the \(CH\) does not suggest the correct result in this case.

#### Table 2. The suggested best numbers of clusters of all five clustering algorithms in three microarray gene datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Clustering</th>
<th>((0.1,0.1))</th>
<th>((0.1,0.3))</th>
<th>((0.1,0.5))</th>
<th>((0.3,0.1))</th>
<th>((0.3,0.3))</th>
<th>((0.3,0.5))</th>
<th>((0.5,0.1))</th>
<th>((0.5,0.3))</th>
<th>((0.5,0.5))</th>
<th>(VI)</th>
<th>(II)</th>
<th>(GI)</th>
<th>(CH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic</td>
<td>(K)-Means</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>(K)-Medoids</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Hierarchical</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>SOM</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>MClust</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Leukemia</td>
<td>(K)-Means</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>(K)-Medoids</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Hierarchical</td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>SOM</td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>MClust</td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Yeast</td>
<td>(K)-Means</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(K)-Medoids</td>
<td>15</td>
<td>4</td>
<td>4</td>
<td>15</td>
<td>4</td>
<td>4</td>
<td>16</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Hierarchical</td>
<td>16</td>
<td>16</td>
<td>6</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>SOM</td>
<td>16</td>
<td>16</td>
<td>6</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>MClust</td>
<td>14</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>12</td>
<td>12</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

**Fig. 4.** Comparisons of validity indices assessing clustering algorithms for synthetic gene dataset.

**Fig. 5.** Comparisons of validity indices assessing clustering algorithms for leukemia dataset.
the yeast data, as we mentioned that six should be the more reasonable number of clusters, none of the existing validity indices suggest the correct result, while the PVIs show that the k-means provide six clusters and the PVIs with smaller $\alpha$ and $\beta$ also show that the SOM has correct clustering.

5.2. Averaged PVI

According to the results in Table 2, we know the assessment performance depends on the parameter settings, which depends on the dataset structure. For different dataset, the parameter settings may be different. To overcome such dependency and make the PVI somehow "adaptive", we average the PVIs over a range of $\alpha$ and $\beta$. In this work, we choose the range of $[0.1, 0.5]$ for both $\alpha$ and $\beta$. The step is 0.1, thus $N_\alpha$ and $N_\beta$ are both five. The comparison of the APVI with other validity indices are shown in Figs. 4, 5 and 6. The results are consistent with those presented in Table 2.

6. DISCUSSIONS AND CONCLUSIONS

In this paper, we proposed a new validity index, which employs two tunable parameters $\alpha$ and $\beta$ to control the numbers of objects being used to calculate the index. The most important advantage of the PVI is that it has flexibility of tuning the parameters to meet different datasets, especially the microarray datasets. The PVI also can be averaged over a range of $\alpha$ and $\beta$ to avoid its dependency on the dataset structure. We investigate the new PVI for assessing five clustering algorithms, namely the k-means, the k-medoids, the HC, the SOM and the MClust, in three datasets, including a synthetic microarray gene expression data, Leukaemia and Yeast cell cycle datasets. The experimental results appear to suggest that the proposed PVI has relatively robust performance and provides fairly correct judgements.

7. REFERENCES