Noise Resistant Generalized Parametric Validity Index of Clustering for Gene Expression Data

Rui Fa and Asoke K. Nandi

Abstract—Validity indices have been investigated for decades. However, since there is no study of noise-resistance performance of these indices in the literature, there is no guideline for determining the best clustering in noisy data sets, especially microarray data sets. In this paper, we propose a generalized parametric validity (GPV) index which employs two tunable parameters α and β to control the proportions of objects being considered to calculate the dissimilarities. The greatest advantage of the proposed GPV index is its noise-resistance ability, which results from the flexibility of tuning the parameters. Several rules are set to guide the selection of parameter values. To illustrate the noise-resistance performance of the proposed index, we evaluate the GPV index for assessing five clustering algorithms in two gene expression data simulation models with different noise levels and compare the ability of determining the number of clusters with eight existing indices. We also test the GPV in three groups of real gene expression data sets. The experimental results suggest that the proposed GPV index has superior noise-resistance ability and provides fairly accurate judgements.

Index Terms—Clustering validity index, noise resistance, gene expression analysis, microarray

1 Introduction

Clustering analysis has been extensively employed in many scientific fields, including biology, physics, image and vision processing, and medical research [1], [2], [3], [4]. In particular, gene expression data and protein expression data analysis has used clustering as one of main exploratory tools for nearly one and a half decades [5], [6], [7], [8]. Gene expression data measured by high-throughput methods, including microarray or ribosomal nucleic acid sequencing (RNA-seq), are organized in a matrix where each row represents a gene and each column represents sample values at the same time. The goal of the clustering analysis is to group individual genes or samples from a population into a cluster within which the objects are more similar to each other than those in other clusters [3], [5], [6], [7], [8], [4], [9], [11], [12], [13], [14], [15]. However, due to its unsupervised nature, there is no existing guideline to guarantee an optimal clustering and is still an open question how to tell that a clustering algorithm or a clustering result is better. Thus, the task of assessing the results of clustering algorithms can be as important as the clustering algorithms themselves. The procedure for evaluating clustering algorithms and their results is known as *clustering validation* [16], [17].

There has been a lot of clustering validation algorithms in the literature since 1960s [18]. The most of clustering validation algorithms can be classified into three classes [16], [17], namely *external criterion*, *internal criterion* and *relative criterion*. The classification hierarchy of clustering validation algorithms is illustrated in Suppl. Fig. 1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2014.2312006. External criterion implies that the results of a clustering algorithm are evaluated based on a pre-specified structure, which is imposed on a data set to reflect the clustering structure of the data set. The best examples are Rand index (RI) [19] and adjusted Rand index (ARI) [20]. However, the demarcation of *internal criterion* and *relative criterion* is vague and sometimes they are mixed up [21], [22]. Internal criterion evaluates the clustering algorithms in terms of the inner structures of the data sets themselves, for example resampling based methods, like figure of merit (FOM) [23] and Clest [24]. This class of algorithms may have good estimates of number of clusters in a data set and also have good indication of the effectiveness of clustering algorithms. But due to their nature of re-sampling approach, the two main issues are that 1) they are computationally expensive, 2) they cannot validate individual partition result. Relative criterion evaluates the clustering partitions by the relative relationship between compactness and separation. We are more interested in the relative criterion because of the following three reasons: 1) their simplicity, 2) low computational load, and 3) their ability of judge the quality of respective clustering partitions. Relative criterion can be further classified into two main subclasses: a) model-based or information theoretic validation, e.g., minimum description length (MDL) [25], minimum message length (MML) [26], [27], Bayesian information criterion (BIC) [28], Akaike's information criterion (AIC), the informational complexity criterion (ICOMP), classification likelihood criterion (CLC), and the normalized entropy criterion (NEC); b) geometry-based validation, which considers the ratio of within-group distance to between-group distance (or its reciprocal), also

- *R. Fa is with the Department of Electronic and Computer Engineering, Brunel University, Uxbridge, Middlesex UB8 3PH, United Kingdom. E-mail: rui.fa@brunel.ac.uk.*
- *A.K. Nandi is with the Department of Electronic and Computer Engineering, Brunel University, Uxbridge, Middlesex UB8 3PH, United Kingdom, and the Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland. E-mail: asoke.nandi@brunel.ac.uk.*

Manuscript received 28 Mar. 2013; revised 21 Jan. 2014; accepted 24 Feb. 2014. Date of publication 16 Mar. 2014; date of current version 4 Aug. 2014. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TCBB.2014.2312006

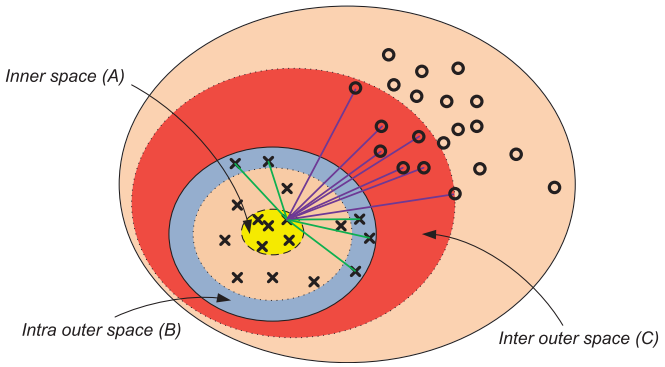

Fig. 1. Illustration of the proposed GPV. Symbols "x" and "o" represent two clusters. Considering "x" cluster, the yellow area is the inner space, labelled by "A"; the blue area is the intra outer space, labelled by "B"; and the red area is inter outer space, labelled by "C".

known as validity index. There are two further sub-subclasses in geometry-based validation i.e., fuzzy and crisp. Fuzzy validity indices include partition coefficient (PC), partition entropy (PE) [29], [30], Fukuyama-Sugeno (FS) index [31], Xie-Beni (XB) index and Kwon's extended XB (KEXB) index [30], [32]. Crisp validity indices include Calinski-Harabasz (CH) index [33], *Dunn's* index (DI) [34], Davies-Bouldin (DB) index [35] (DB is the counterpart of XB with crisp clustering), *I* index [36], Silhouettes [37], Krzanowski and Lai index (KL) [38], the geometrical index (GI) [39], and the validity index V_I [40]. Since the real gene expression data, especially microarray data, has high background noise, the noise resistance ability is important to the clustering algorithms and the validity indices. To the best of our knowledge, although some of these indices have been applied in the clustering analysis of gene expression data, there is no study to investigate their noise resistance ability and there is no validity index claimed to have this property.

Besides aforementioned criteria, there has been some other effort devoted into the research of integrating *a priori* biological knowledge into clustering and clustering validation [42], [43]. However, one should be very cautious to exploit those *a priori* biological knowledge in clustering analysis, especially clustering validation, because two serious issues may be easily neglected. First, is the *a priori* knowledge enough to judge the clustering? Second, even though we have strong enough *a priori* knowledge, can we exploit them to guide or judge the clustering of the newly collected data, especially noisy data? These two issues limit the practical use of this type of methods. In this paper, we do not integrate any *a priori* knowledge into the algorithm.

Recently, Lingras et al. proposed a decision-theoretic measure of rough cluster quality [44]. Although this method is designed for rough clustering specifically and not applicable to other types of clustering, it reveals an idea of parametric clustering validation. In this method, there is an important parameter, threshold, to control the ranges of lower bound, upper bound and boundary areas. The most important advantage of the parametric validation is the robustness, i.e., the validation can deal with various clustering algorithms for various data sets by adjusting the parameters. However, there are also two challenges in developing such parametric validation: one is the choice of the metric for cluster validation that is influenced by the parameters and the other, most essentially, are the optimal values for these parameters for different clustering algorithms and different data sets. In [44], the first challenge was tackled well by proposing a risk measurement strategy based on decision-theoretic framework; however, the second problem was left unaddressed.

In this paper, we propose a generalized parametric validity (GPV) index which is one of geometry-based indices to calculate the ratio of the inter-cluster dissimilarity to the the intra-cluster dissimilarity. We introduce two tunable parameter α and β in the new index to control the proportions of objects being taken into account to calculate the dissimilarities. The greatest advantage of the proposed GPV index is its noise-resistance ability. The noise-resistance ability results from the flexibility of tuning the parameters, which in turn leads to its robustness, to meet different data sets (especially the microarray data sets). Several rules are set to guide the selection of parameter values. By examining the dissimilarity densities of different data sets, the maximal appropriate values of the parameters for individual data set can be obtained. To validate our validity index, we must have: (1) some test data sets whose structures are already known, (2) some reference clustering results which has been evaluated by external indices, say ARI, (in this paper, we use five clustering algorithms, namely K-Means, K-Medoids, hierarchical clustering (HC), self-organizing map (SOM) and model-based clustering (MCLUST)), and (3) one or more metrics to indicate the effectiveness of the tested indices. We first obtain reference clustering results by clustering the test data sets and evaluating the results with ARI; then we test the proposed index with these reference clustering results and compare its ability to indicate the correct structure of the given data set with other existing validity indices. In this paper, we evaluate the new GPV in both simulated and real gene expression data sets. Although we focus on the gene expression data, our proposed index can also be applied in other expression data analysis like protein expression data. We employ simulated data because, first, we can control the noise levels by tuning the noise-controlling parameters in the model; second, both the true membership and the true number of clusters can be employed to evaluate the performance of validity indices in the simulated data sets. We investigate our proposed index in many real data sets. The first group contains three data sets from different experiments or species, which have different noise levels. We also consider other two data sets, namely Gasch set [45] and Ogawa set [46], which has been widely used for testing clustering algorithms [47], [48], and extract two subsets (one tight set and one loose set) from each of them. We demonstrate that some existing validation algorithms work well in one "clean" data set, but are problematic in the other noisy data set. In this case, the contribution of this paper turns to be twofold: on one hand, we propose GPV and the methods determining the parameters; on the other hand, to evaluate the proposed GPV, we investigate its noise-resistance ability and compare it with eight other existing validity indices, namely, *Dunn's* index [34], the *I*-index (II) [36], the *Calinski Harabasz* index [33], the geometrical index ([39], the validity index V_I [40], Davies-Bouldin index [35], Silhouettes [37], and the Krzanowski and Lai index [38]. The experimental results suggest that the proposed GPV

has relatively good noise-resistance performance and provides fairly accurate judgements.

The rest of the paper is organised as follows, In Section 2, both the synthetic data sets and real data sets are introduced first since an example data set will be used in demonstrating the development of the algorithm. Section 3 presents the principle of the proposed GPV and the selection rules for the parameters. Section 4 presents the experimental results and conducts a detail discussion regarding the clustering validations. Finally, conclusions are made in Section 5.

2 Data Sets

2.1 Simulated Data Sets

We employ two microarray gene expression data models to simulate or synthesize gene expression data. One simulates the state-based gene expression data [13] and another one simulates the periodic behaviour of yeast cell cycle [12], [49]. The advantages of using simulated data are that the ground truth is known and we have the freedom to manipulate the noise level of the data by tuning a few parameters.

The first simulating model (S1) is a stochastic model which simulates the state-based gene expression data [13]. Unlike [13], we only simulate gene expression data sets without scatter genes. There are 11 clusters $\{\mathcal{C}_k|k = 1,\ldots,11\}$ of genes with $M = 50$ samples in the simulated data. Random noise from normal distribution with standard deviations $\sigma_n = 0, 0.05, 0.1, 0.2, 0.4, 0.8,$ and 1.2 is added. The interested readers are referred to [13] for the details. The parameters used in this model are set as: $\mu = 6, \sigma = 1$, $\sigma_s = 1.0, \sigma_0 = 0.1$, and $\lambda = 10$. We generate 100 data sets for each σ_n.

The second simulating model (S2) was originally proposed in [49]. We employ it to generate a number of synthetic gene expression data sets with 500 synthetic genes in each data set and 24 samples for each gene. These 500 genes belong to $K = 5$ clusters and each cluster has 100 members. The model of cyclic gene expression is given by

$$x_{ij} = r + [d + yr](r + [d + yr]\sin(2\pi j/8 - \omega_i + zr), \qquad (1)$$

where x_{ij} is the expression value of the ith gene at the jth time point, each instant of r is an independent random number from the standard normal distribution $\mathcal{N}(0, 1)$, d controls the magnitude of the sinusoid and it is fixed to three here, y controls the random component added to the magnitude, z controls the random component added to the phase, and ω_i is the phase shift of the ith gene. ω_i will determine which cluster the gene i will be in. Since the noise in this model is not additive, we have to couple y and z to be a pair, and raise their values to change the noise power. By increasing values of y and z will increase the noise power. The paired parameters are listed as $(y, z) \in \{(0.1, 0.01), (0.3, 0.03), (0.5, 0.05), (0.7, 0.07), (0.9, 0.09), (1.1, 0.11), (1.3, 0.13), (1.5, 0.15), (1.7, 0.17), (1.9, 0.19), (2.1, 0.21), (2.3, 0.23), (2.5, 0.25)\}$. Thus, there are 13 parameter pairs (PPs) from PP1 to PP13, representing 13 noise levels from low to high. For each pair of parameters, we generate 1,000 data sets, and subsequently, we get 1,000 clustering results from each clustering algorithm.

2.2 Real Data Sets

2.2.1 Group 1

This group contains three real gene expression data sets, one Leukaemia data set and two Yeast cell cycle data sets. The leukemia data set [7] consists of 38 bone marrow samples obtained from acute leukemia patients at the time of diagnosis. The samples include 11 acute myeloid leukemia (AML) samples, eight T-lineage acute lymphoblastic leukemia (ALL) samples, and 19 B-lineage ALL samples. There are 999 genes in the data set. One of two Yeast cell cycle data sets is α-38, which was presented in [52]. The data set employed in the paper consists of 500 genes with highest periodicity scores and each gene has 25 time samples. Additionally, the peak times in (0, 100 percent] of these 500 genes in the cell cycle are provided, as the whole cell cycle is 100 percent. Another yeast cell cycle data set cdc-28 was published by Cho et al. [5]. It consisted of more than 6,000 genes over 17 time points taken at 10 minutes intervals. The data set we investigate in this paper is a subset of 384 genes out of 6,000 genes, which were demonstrated consistent periodic changes in transcript level [12]. It is available at http://faculty.washington.edu/kayee/model/. It was commonly believed that the time course was divided into early G1, late G1, S, G2, and M phases biologically, and those 384 genes would peak at one of the five phases. In the recent research [53], an extra statistical cluster called Q (questioned) phase, was identified. However, the members in the Q phase are more likely to belong to different biological groups in other recent data sets. We choose these three data sets because they have different noise levels: the leukemia data set has the lowest noise among three data sets, α-38 has moderate noise, and cdc-28 has highest noise.

2.2.2 Group 2

This group contains two subsets, one tight set and one loose set, which are extracted from Ogawa set [46]. The original Ogawa set about the phosphate accumulation and the polyphosphate metabolism of yeast *S. cerevisiae* contains 5,783 genes and eight samples after removing the genes with missing values. We borrow the idea in the Bi-CoPaM method [14], [15] to extract the test subsets. What Bi-CoPaM did is to fuse many clustering results from different clustering algorithms, generate a fuzzy consensus partition matrix and then binarize the fuzzy consensus partition matrix according to the threshold. Different values of the threshold will lead to clustering results with different levels of tightness. In this work, we used five clustering algorithms, namely K-Means, K-Medoids, HC, SOM and MCLUST, to partition the whole gene expression data with the number of clusters equal to 20 and generated the fuzzy consensus partition matrix based on the normalised votes of each gene to each cluster by individual clustering algorithm. In the fuzzy consensus partition matrix, "1" means all algorithms vote the gene to the cluster, "0" means no algorithm votes the gene to the cluster, and the value between 0 and 1 means that at least one algorithm votes, in this case, which could be 0.2, 0.4, 0.6 and 0.8. We employed a binarization technique

TABLE 1
Structure Summary of Data Sets in Group 2 and Group 3

	Group 2 (Ogawa set)				Group 3 (Gasch H_2O_2 set)			
	C1	C2	C3	C4	C1	C2	C3	C4
Tight subset	219	257	62	60	119	236	77	60
Loose subset	368	266	220	186	228	426	360	195

Both Groups 2 and 3 respectively contain one tight subset and one loose subset, which are extracted from Ogawa set and Gasch set respectively. The subsets of both groups have four clusters.

called difference threshold binarization (DTB), which assigns each gene to the maximum membership value cluster only if the difference between the maximum membership value and its closet membership value is larger than or equal to the threshold, otherwise does not assign the gene to any cluster. Thus, clusters lose genes significantly and many clusters are empty in the tightening. We extract four clusters which survive in the tightening with the threshold equal to 0.6 and form the tight set and extract the loose set choosing the same four clusters as the tight set with the threshold equal to 0.2. We summarise the subsets in Table 1. In total, the tight set contains 498 genes and the loose set contains 1,040 genes. Their profiles in four clusters are displayed in the supplementary, available in the online supplemental material.

2.2.3 Group 3

Using the same process in Group 2, we extract the tight subset and loose subset from Gasch set [45], which also have four clusters. The original Gasch set contains 6,153 genes and 178 experimental samples. In this group, we only consider the condition with H_2O_2 osmotic shock, which contains 10 samples. The structure summary is given in Table 1. The tight set totally contains 492 genes and the loose set totally contains 1,209 genes. Their profiles in four clusters are also displayed in the supplementary, available in the online supplemental material.

3 Generalized Parametric Validity Index

Suppose that gene expression data objects are formalized as numerical vectors $\boldsymbol{x}_i = \{x_{ij} | 1 \leq j \leq p\}$, where p is the number of features and x_{ij} is the value of the jth feature and the ith object. To be specific to gene expression data, "object" means gene and "feature" means sample. There are n objects in the data sets. In this section, we detail the principle of the proposed generalized parametric validity index. The GPV belongs to the class of geometry-based indices to calculate the ratio of the intra-cluster dissimilarity to the inter-cluster dissimilarity.

3.1 Proposed GPV Index

We introduce two tunable parameter α and β 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 '$\times$' and '$\circ$' 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 two different clusters. In some noisy scenario, the boundaries among clusters are blurred, and moreover, the boundary areas are more dominant in determining the quality of a clustering result than center areas. Thus, in order to calculate the dissimilarities efficiently, we must determine those objects at both ends of the dissimilarities. We define the inner space representing 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 "A" in Fig. 1. The outer ends objects in the same cluster lie in the intra outer space, which is the area marked "B" in Fig. 1, and the outer ends objects in the different clusters lie in the inter outer space, which is the area marked "C" in Fig. 1. Thus, let us define N_k^i, Na_k^o, Ne_k^o to denote the numbers of objects in the inner space, the intra outer space and the inter outer space, respectively, for the kth cluster. The fractions, α and β, are used to control N_k^i, Na_k^o, Ne_k^o, which can be expressed as

$$N_k^i = \lceil \alpha N_k \rceil, Na_k^o = \lceil \beta N_k \rceil, Ne_k^o = \lceil \beta (N - N_k) \rceil, \quad (2)$$

where N_k is the number of the objects in the kth cluster, N is the number of all objects in the data set and $\lceil \cdot \rceil$ is the ceiling operator. Both α and β can be chosen from the range of (0, 1]. Thus, N_k^i, Na_k^o, Ne_k^o 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 GPV. First, we need to form the subset $\mathcal{A}_k$ for the objects in the inner space. For each object in the kth cluster, we can obtain a total dissimilarity by the summation of dissimilarities between it and all others in the kth cluster, as

$$\mathcal{D}_n = \sum_{m=1}^{N_k} \mathcal{D}(\boldsymbol{x}_n, \boldsymbol{x}_m), n = 1, \ldots, N_k, \quad (3)$$

where $\mathcal{D}(\cdot, \cdot)$ 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 [3], [4]. In this paper, we calculate the dissimilarity for the proposed GPV based on Pearson's correlation coefficient. Pearson's correlation coefficient is defined as

$$\mathrm{PCC}(\boldsymbol{x}_n, \boldsymbol{x}_m) = \frac{\sum_{d=1}^{p} (x_{nd} - \mu_n)(x_{md} - \mu_m)}{\sqrt{\sum_{d=1}^{p} (x_{nd} - \mu_n)^2} \sqrt{\sum_{d=1}^{p} (x_{md} - \mu_m)^2}}, \quad (4)$$

where μ_n and μ_m are means for $\boldsymbol{x}_n$ and $\boldsymbol{x}_m$, respectively. Thus, the dissimilarity is obtained by

$$\mathcal{D}(\boldsymbol{x}_n, \boldsymbol{x}_m) = 1 - \mathrm{PCC}(\boldsymbol{x}_n, \boldsymbol{x}_m). \quad (5)$$

We pick N_k^i most centrally located objects, which have relatively smaller total dissimilarities, from the kth cluster to form $\mathcal{A}_k$, which is $\{\boldsymbol{a}_k^a | a = 1, \ldots, N_k^i\}$. Second, for each object in the inner space $\boldsymbol{a}_k^a$, we need to form subsets $\mathcal{B}_k^a$ and $\mathcal{C}_k^a$ for the object in the intra outer space and the inter outer space, respectively. The objects in $\mathcal{B}_k^a$, denoted as $\{\boldsymbol{b}_k^{a,b} | b = 1, \ldots, Na_k^o\}$, are those Na_k^o objects in the kth cluster,

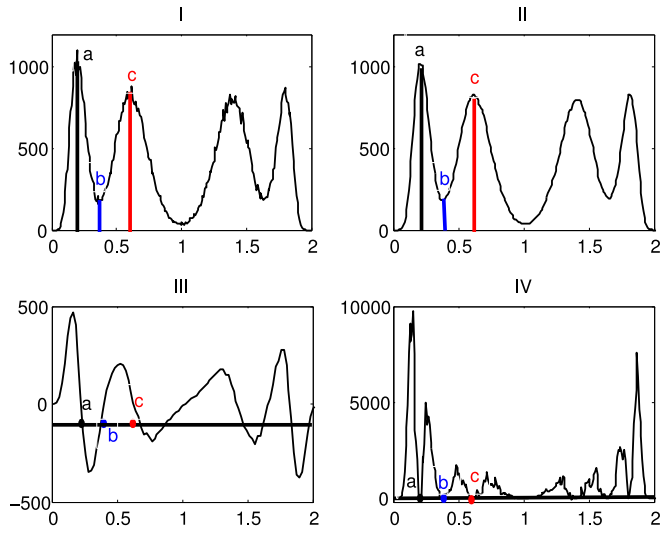

Fig. 2. The process to calculate the bounds of α and β. (I) Dissimilarity distribution of the synthetic gene data set and points a, b and c are what we need to find out to calculate the bounds. (II) The smoothed distribution curve filtered by a rectangular window with the size of 10. (III) The derivative of the distribution function. (IV) A methods to find the zero points in the derivative of the distribution function.

which are farthest from $\boldsymbol{a}_k^a$, while the objects in $\mathcal{C}_k^a$, denoted as $\{\boldsymbol{c}_k^{a,c}|c=1,\ldots,Ne_k^o\}$, are those Ne_k^o objects in the clusters but the kth cluster, which are closest to $\boldsymbol{a}_k^a$. Afterwards, we need to do some calculations as follows:

$$\begin{aligned}\mathcal{D}a_k^a &= \frac{\sum_{b=1}^{Na_k^o}\mathcal{D}\left(\boldsymbol{a}_k^a,\boldsymbol{b}_k^{a,b}\right)}{Na_k^o}\\ \mathcal{D}e_k^a &= \frac{\sum_{c=1}^{Ne_k^o}\mathcal{D}\left(\boldsymbol{a}_k^a,\boldsymbol{c}_k^{a,c}\right)}{Ne_k^o},\end{aligned} \tag{6}$$

where $\mathcal{D}a_k^a$ denotes a normalized intra-cluster dissimilarity and $\mathcal{D}e_k^a$ denotes a normalized inter-cluster dissimilarity for the ath inner end in the kth cluster. Finally, the GPV can be obtained by

$$\mathrm{GPV}(K,\alpha,\beta)=\sum_{k=1}^{K}\sum_{a=1}^{N_k^i}\left(\frac{\mathcal{D}e_k^a}{\mathcal{D}a_k^a}\right). \tag{7}$$

There are two advantages for choosing the new index. On one hand, the GPV only involves some objects into the calculation rather than all objects, which reduces the computational complexity depending on the settings of α and β. On the other hand, the flexibility obtained through two tunable parameters leads to its robustness and the GPV may be useful in many different data sets. As the settings of the parameters are largely dependent on the data set structure, it is crucial to explore the data set structure to obtain the optimal parameters. Thus, we develop a strategy to obtain the optimal α and β from any given data set, which is presented in the next part.

3.2 Selection of Parameters

In this part, we discuss the selection of parameters α and β. To this end, we have to first specify the physical meaning of these two parameters. Let us draw an analogy between a cluster and any objects with core and shell, say a fruit like an apple or a planet like the Earth. The "A" area shown in Fig. 1, which is determined by α, is the core of the cluster. The "B" area is analogous to the "shell", the "skin" or the "crust," and the "C" area is analogous to the "environment" and the "atmosphere", which is the interconnection between clusters. That is, the values of α and β are dependent on the definition of the "core", the "shell" and the "atmosphere". Undoubtedly, it is fairly easy to distinguish two clusters when the noise is relatively low. In this case, the inner space can be reasonably large (α is large) while the intra outer space and the inter outer space can be very thin and only small number of objects in there (β is small). However, when the noise increases and clusters spread out, the inner space becomes sparser while the intra outer space and the inter outer space expand and eventually mingle (α is going downwards, while β is going upwards). Thus in such scenario, the boundary areas among clusters are more dominant in determining the quality of the clustering results. Taking all within-cluster distances and between-cluster distances many diminish the subtle difference between within-cluster and between-cluster distances, and consequently degrade the performance of the validity index.

Since it is difficult to find the optimum value for α and β, what we can do is to find their upper or lower bounds. The first general rule is that there should be a good number of objects in all three spaces, otherwise GPV may be vulnerable to the outliers and its performance will degrade significantly. Our experience is that the minimum number of objects in any of these three spaces is five. The second general rule is that we avoid the overlapping between core and the shell, that is, $\alpha+\beta$ is smaller than or equal to one.

We further compute the upper bound of the parameter values, since the above two general rules are too loose. We investigate the dissimilarity density (obtained by histogram), which describes the distribution of all pairwise dissimilarities between the objects in the data set. Let us take a synthetic data set as an example, whose dissimilarity density, say $f(x)$, is shown in Fig. 2I. Note that there are clearly four wave shapes which means that there are at least four classes in the data set. Let us look at the figure from the left side to the right side, with the dissimilarity values ranging from 0 to 2. The first wave represents all the dissimilarities within a cluster, whichever cluster the objects belong to and its peak indicates the dissimilarity level with relatively high density. The second wave represents the cluster(s) closest to the cluster which the current object belongs to. The others waves we do not care about. In Fig. 2I, we mark a, b, c to indicate the dissimilarity values of the first peak, the first trough, and the second peak of the density, respectively. The area under the density is normalised by the total number of pairwise dissimilarities to be 1. Let us define $\mathcal{A}(x,y)$ to denote the area under the density from x to y, thus $\mathcal{A}(0,2)=1$. The boundary values of parameters α and β can be obtained by

$$B_\alpha=\frac{\mathcal{A}(0,a)}{\mathcal{A}(0,b)}\ \text{and}\ B_\beta=\frac{\mathcal{A}(a,c)}{\mathcal{A}(0,2)}. \tag{8}$$

TABLE 2
The List of Functions and Platforms to Implement the Clustering Algorithms

Algorithms	Platform	Functions	Options
KMeans	MATLAB	kmeans	KA initialization
HC	MATLAB	linkage& cluster	'complete'
SOM	MATLAB	newsom & train & sim	-
MCLUST	R	mclust	-
KMedoids	MATLAB	-	KA initialization

Thus, we get **Rule 3**, $\alpha \leq B_\alpha$ and $\beta \leq B_\beta$. Fig. 2II, 2III, 2IV show the process that we get the turning points a, b, and c. The details can be found in the supplementary, available in the online supplemental material.

4 Results and Discussions

4.1 Clustering

We use "kmeans" function in MATLAB to implement KMeans clustering. However, instead of a random initialization, we use a deterministic initialization, Kaufman approach (KA) [54], which was reported to be superior to other initialization approaches [50]. To implement HC clustering, we use "linkage" function with "complete" option and "cluster" function in MATLAB. To implement SOM clustering, we use "newsom", "train" and "sim" functions in MATLAB. Library "mclust" in R [55] is used to implement MCLUST. We implement KMedoids using MATLAB codes and employ the deterministic KA initialization. The list of functions and platform to implement the clustering algorithms is shown in Table 2.

4.2 Parameter Selection for GPV

Parameter selection is the first step of GPV. In this section, we determine the parameters based on the selection rules we developed in Section 3.2. Note that different data sets have their own proper parameters. For the S1 data sets, we calculate B_α and B_β based on (8), whose error bar plots are shown in Fig. 3a. We find that the mean value of B_α decreases while the mean value of B_β increases with the increase of noise, which means that the proportion of members in the core reduces and the boundary enlarges. We also find the variances of both B_α and B_β increase when the noise goes higher. These observations match what we expected as while the noise is getting heavier, clusters are spreading out and boundaries are getting blurred, so that α is going downwards, and on the contrary, β is going upwards. To guarantee the overall consistent validation performance, we select α and β for each σ_n based on B_α and B_β as $\alpha = [0.75, 0.7, 0.65, 0.55, 0.5, 0.45, 0.4]$ and $\beta = [0.05, 0.05, 0.05, 0.05, 0.2, 0.4, 0.4]$.

We synthesize 1,000 S2 data sets for each of 13 parameter pairs, corresponding to 13 different noise levels, that is, there are 13,000 data sets being examined. Similar to S1 data sets, using (8), we calculate all the values of B_α and B_β, and the error bar plots of B_α and B_β are shown in Fig. 3b. Note that although the noise is not linearly increase in this case, the estimation variances of B_α and B_β increase with the increase of the noise level. Thus, we can set $\alpha = 0.4$ and $\beta = 0.2$ for all S2 data sets.

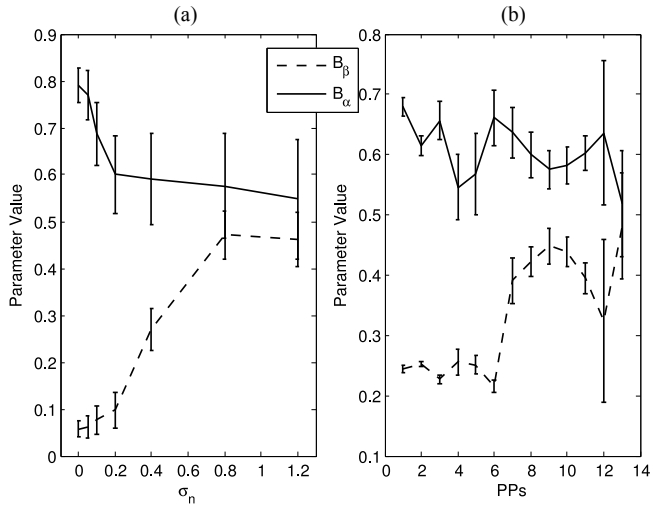

Fig. 3. (a) The error bar plots of B_α and B_β values against σ_n in the S1 data sets. (b) The error bar plots of B_α and B_β values for all PPs in the S2 data sets. The vertical bar represents the standard deviation.

We did the same with different real data sets to find B_α and B_β. In the leukemia data set case, we obtain $B_\alpha = 0.78$ and $B_\beta = 0.21$. Thus, we choose $\alpha = 0.75$ and $\beta = 0.2$. In the yeast cdc-28 data set case, we obtain $B_\alpha = 0.43$ and $B_\beta = 0.26$, and then we choose $\alpha = 0.4$ and $\beta = 0.2$. In the yeast α-38 data set case, we obtain $B_\alpha = 0.49$ and $B_\beta = 0.33$, and then we choose $\alpha = 0.45$ and $\beta = 0.3$.

4.3 Synthetic Data Sets

Since the clustering validity indices have to work in an unsupervised situation, the ground truth of the data sets, like the number of clusters and the membership of each cluster, is not available to judge the quality of the clustering in the most of real data sets. To "validate" these validity indices, we have to test these indices in the data sets whose ground truth is available, and make use of the ground truth of these data sets. It is logical to deduce that the best index will also work well in the similar type of data set when the ground truth is not available. To this end, we first assess the clustering algorithms by using an external criterion, namely ARI [20], which is a modification of original RI [19]. We also define accurate estimate ratio of number of clusters (AERNC) to measure the performance of determining the number of clusters for validity indices. Mathematically, AERNC is given by

$$\text{AERNC} = \frac{N_{\hat{K}=K}}{N_{\text{total}}}, \tag{9}$$

where $N_{\hat{K}=K}$ is the number of accurately estimating the number of clusters in N_{total} total experiments. Note that the AERNC performance of the indices are sensible only if the clustering results are sensible, which is a critical point to use AERNC as a metric to judge the quality of the validation results.

In the first place, we show the results of ARI and its AERNC against σ_n of S1 data sets in Fig. 4 as a reference. Overall, MCLUST has the best clustering performance while HC has the poorest among the five clustering algorithms.

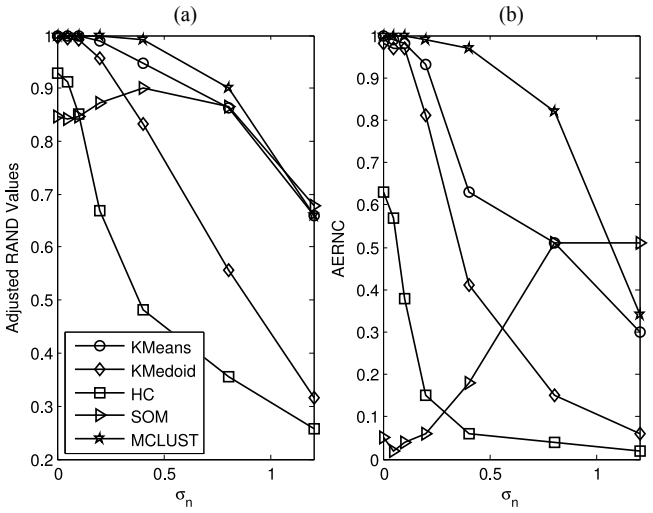

Fig. 4. Results for S1 data sets: (a) ARI versus σ_n for five clustering algorithms when $K = 11$. (b) AERNC versus σ_n of ARI for five clustering algorithms.

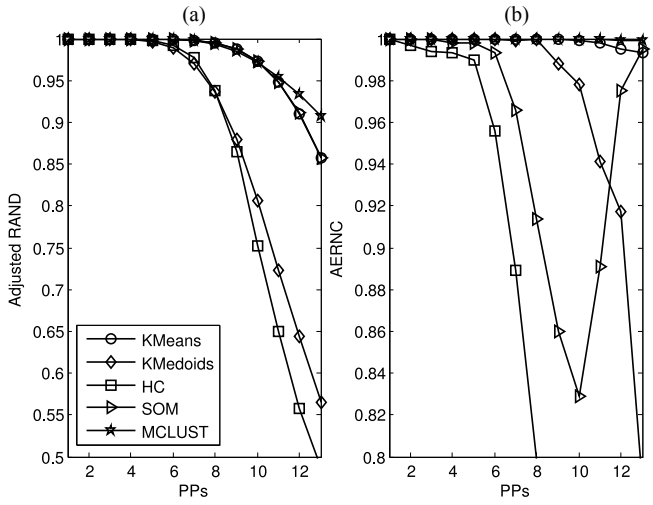

Fig. 6. Results for S2 data sets: (a) ARI versus PPs for five clustering algorithms when $K = 5$. (b) AERNC versus PPs of ARI for five clustering algorithms.

SOM has a very strange behavior: its ARI value is low when the noise is low, subsequently increase while the noise increases, and then falls down while the noise keeps increasing. Its AERNC performance, which goes up with the increase of the noise, is also odd. The reason for this observation is that the performance of SOM is highly related to the grid configuration. Let us look at the performance of AERNC against σ_n for all compared validity indices. In the main body of the paper, we only show KMeans and MCLUST results, which are depicted in Figs. 5a and 5b respectively, since they are top best clustering algorithms. Other results can be found in the supplementary, available online. In terms of the AERNC result, the proposed GPV index has the best performance among the compared indices in both clustering results. It is slightly better than CH when the noise is low. With the increase of noise, CH degrades sharply while GPV degrades gradually. GI, II, DI and KL do not perform well even at very low noise in S1 case. For comparison purposes, we also plot the AERNC performance of ARI for KMeans and MCLUST in bold lines.

It has reasonably good AERNC performance since external knowledge is employed.

We depict the results of ARI and its AERNC against noise levels of S2 data sets in Fig. 6. In Fig. 6a, ARI suggests that MCLUST performs best, KMeans and SOM are relatively good and HC is the worst one. The results of AERNC performance in Fig. 6b are consistent with those in Fig. 6a, except that SOM has an odd "V" shape. The results of AERNC against noise levels of all compared indices for KMeans and MCLUST are shown in Figs. 7a and 7b respectively. Same as in S1, GPV has the superior noise-resistance performance among all indices. KL is in the second place in the moderate noise but it has small estimation errors in the low noise. CH and Silhouette have good performance in the low noise, but degrade sharply while the noise rises.

In summary, the results in both simulated data sets strongly support that our proposed GPV has the superior noise-resistant performance among all compared indices. Although ARI has much better performance than validity indices, it cannot work without external knowledge. In this case, GPV is the best choice.

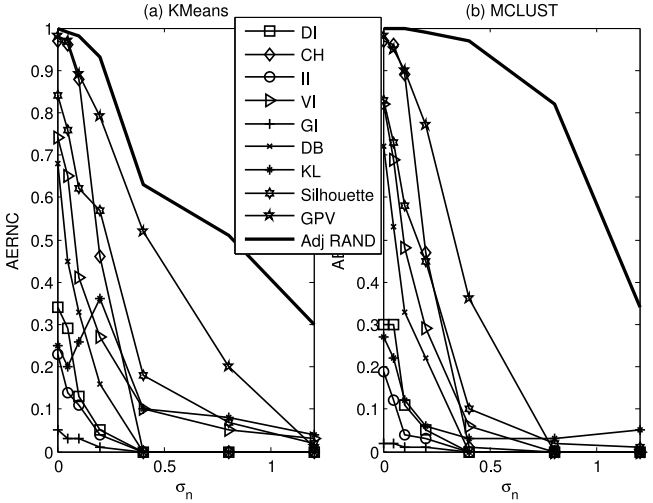

Fig. 5. AERNC versus σ_n of all compared indices for KMeans and MCLUST in S1 data sets. (a) KMeans and (b) MCLUST.

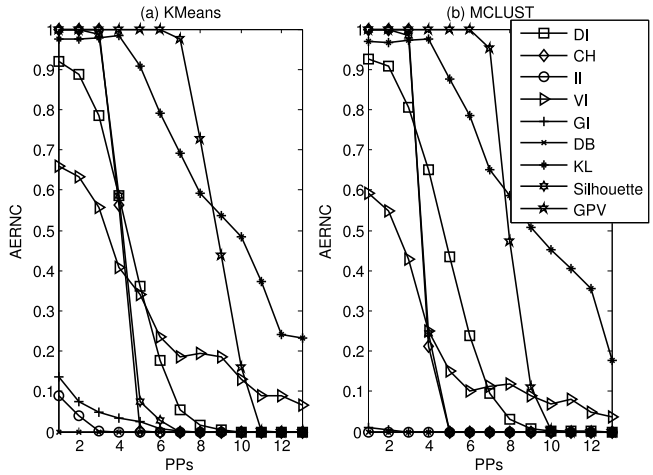

Fig. 7. AERNC versus PPs of all compared indices for KMeans and MCLUST in S2 data sets. (a) KMeans and (b) MCLUST.

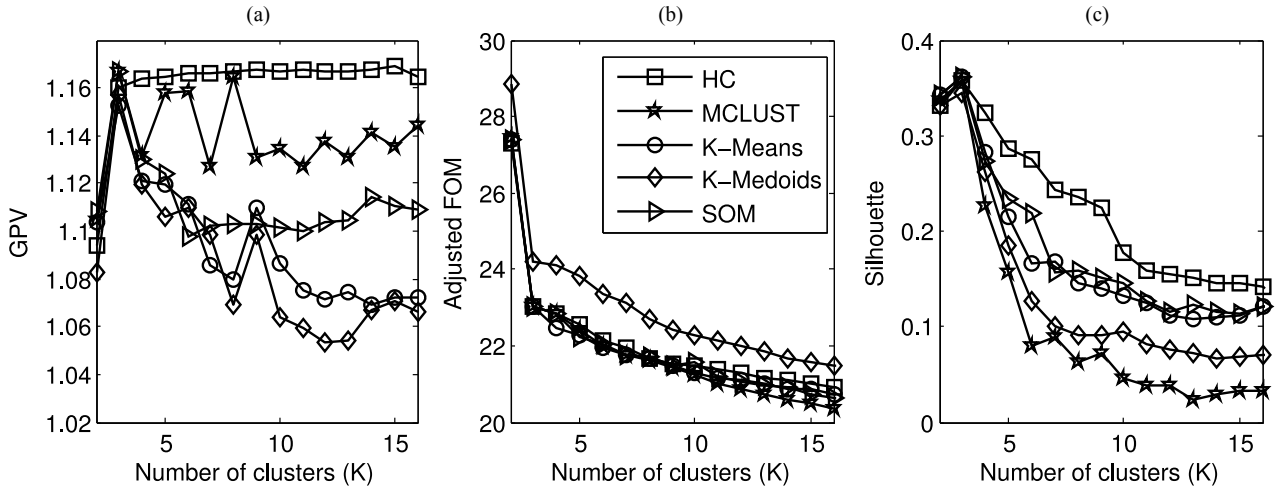

Fig. 8. Comparison of validity indices against the number of cluster K in Leukemia data for five clutering algorithms: (a) GPV, (b) Adjusted FOM, and (c) Silhouette.

4.4 Real Data Sets

4.4.1 Group 1

In this part, we test the proposed GPV index in three real data sets and compare its performance with existing indices. We employ adjusted FOM, which is one of internal validation algorithms [23], as reference to indicate the actual number of clusters. However, adjusted FOM might not be effective in some high noise data sets. In that case, a deeper investigation into the data set has to be done to determine the number of clusters.

1) Leukemia data set. In Fig. 8, we only depict the results of GPV, adjusted FOM and Silhouette due to the space limit. Other results can be found in Suppl. Figs. 4 and 5, available in the online supplemental material. We notice from the results that for the leukemia data set, many indices, namely DI, DB, GI, KL, VI, II, Silhouette and GPV, except CH, indicate that there are clearly three clusters more or less, which is consistent with the description in [7]. Adjusted FOM is not an objective and automatic index so that we have to determine the number of clusters subjectively based on the "knee" shape. In this case, Adjusted FOM also indicates the best number of clusters equal to three. The low noise level in Leukemia data is the main reason that many indices work well.

2) α-38 yeast cell cycle data set. We then test the indices in α-38 Yeast cell cycle data set. The well established biological knowledge tell us that there are four phases in the cell cycle, namely, G1, S, G2 and M phases. We depict the results of GPV, adjusted FOM and GI in Figs. 9a, 9b and 9c, respectively. Only these three indices indicate that the best number of clusters is four while most of others, which can be found in the Suppl. Figs. 6 and 7, available in the online supplemental material, indicate that three is the best number of clusters. We believe that in this data set, four is best number of cluster because of two reasons: 1) it is consistent with biological knowledge, 2) GPV is consistent with adjusted FOM, which has more reliability than other indices.

3) cdc-28 yeast cycle data set. Next, we discuss the results of cdc-28 yeast cycle data. At first, we notice that cdc-28 is such a very noisy data set that our results shows a great disagreement among the compared indices. To determine which indices indicate the correct number of clusters, we have

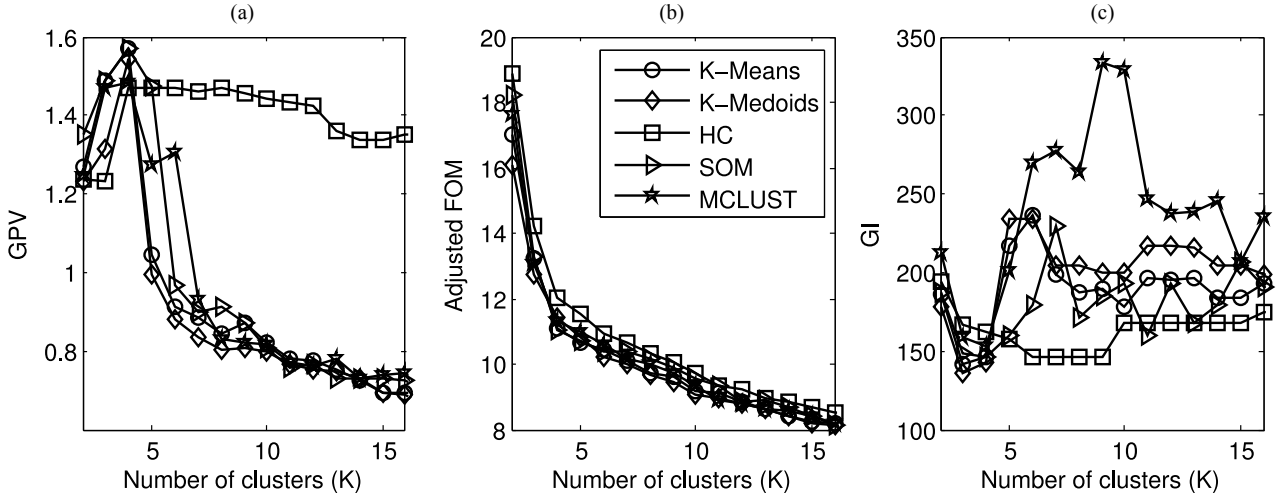

Fig. 9. Comparison of validity indices against the number of cluster K in α-38 data for five clutering algorithms: (a) GPV, (b) Adjusted FOM, and (c) DI.

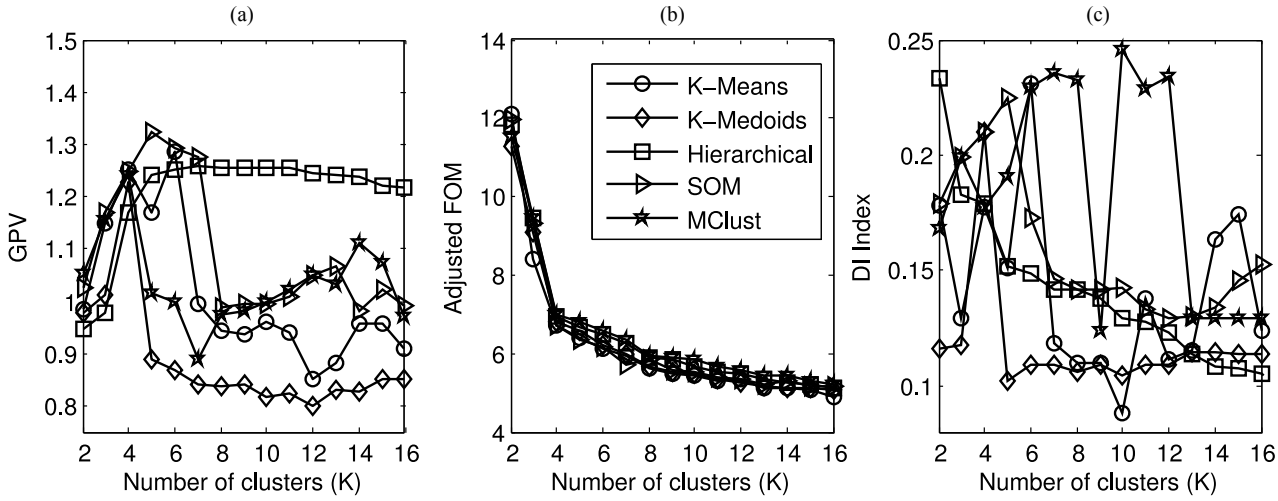

Fig. 10. Comparison of validity indices against the number of cluster K in cdc-28 data for five clutering algorithms: (a) The GPV index, (b) Adjusted FOM, and (c) DI.

taken a close look into the clustering memberships for the data set. A clustering is given with the data set by [12] showing that one complete cell cycle consists of five phases, namely Early G1, Late G1, S, G2 and M phases. However note that the memberships of the clustering in [12] and other clustering results are much different, i.e., KMedoids and MCLUS split Late G1 into many sub-groups and many clustering algorithms, namely HC, SOM and KMeans, discover a separated group apart from the five phases, called Q phase in [53]. The PCA of the yeast data is depicted in Suppl. Fig. 8, available in the online supplemental material, in which the Q phase remarked by "★" obviously stands out. We also examine the mean profile of each cluster shown in Suppl. Fig. 9, available in the online supplemental material. Thus, we can claim that Q phase is a numerically separate cluster, but we still do not know if these genes in Q phase belong to same functional group. The 12 genes in the Q phase (union of the results of HC, SOM and KMeans) are listed in Suppl. Table 1, available in the online supplemental material. Their profiles are individually depicted in Suppl. Fig. 10. The profiles of the same genes in another yeast time-course data set, yeast metabolic cycle (YMC) [56], are plotted in Suppl, available in the online supplemental material. Fig. 11. Comparing their patterns, we may find that three out of 12 genes, namely YBR067c (TIP1), YDL124w and YPL186c (UIP4), are most likely to be in same group, while another three out of 12, namely YLR015w (BRE2), YLR014c (PPR1) and YOR274w (MOD5) are most likely to be in another group. It turns out that these genes locate in the Q phase, which is not a biological group, maybe because of the temperature-induced effect, which was also mentioned in [5]. Thus, the Q phase is an oddity.

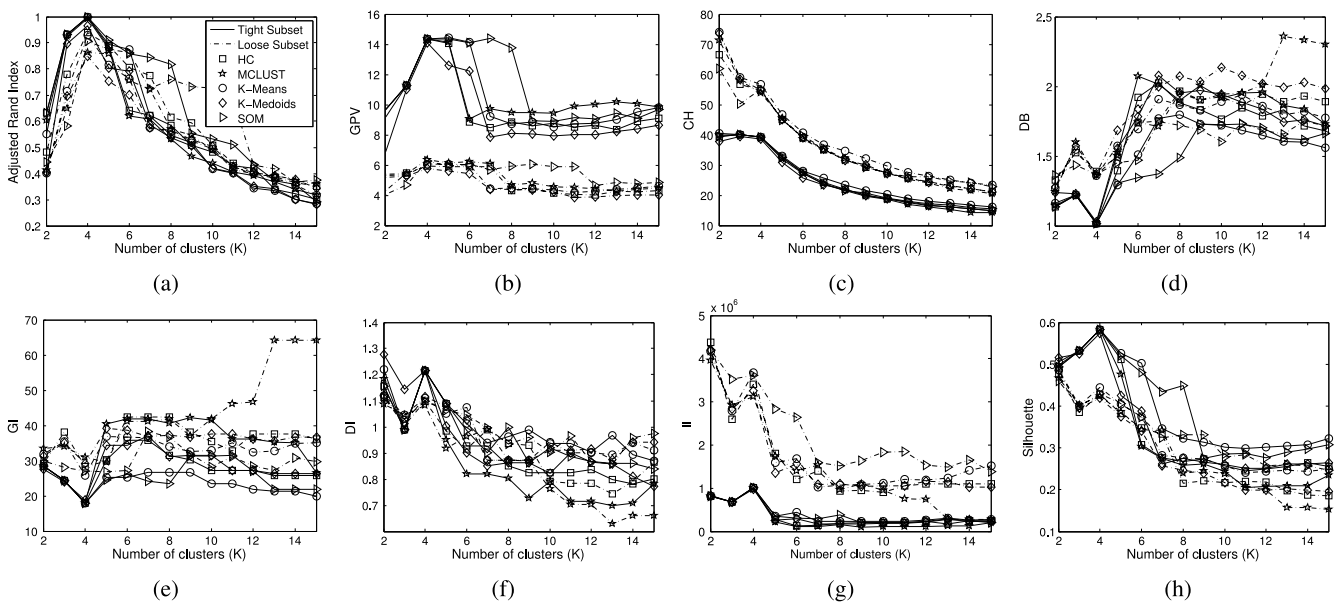

Fig. 11. Performance of index values against the number of clusters K for Group 2 data sets. (a) ARI, (b) GPV, (c) CH, (d) DB, (e) GI, (f) DI, (g) II, and (h) Silhouette. Legend: solid lines represent the performance in the tight subset; broken lines represent the performance in the loose subset; the curves with square markers, for both solid and broken lines, stand for HC; pentagon markers stand for MCLUST; circle markers stand for K-Means; diamond markers stand for K-Medoids; and triangular markers stand for SOM.

TABLE 3
The Number of Memberships for Clustering Algorithms, Where L G1 and E. G1 Denote Late G1 and Early G1, Respectively, Q Denotes the Questioned Phase

Algorithms	E. G1	L. G1	S	G2	M	Q
[12]	67	135	75	52	55	-
HC (K=4)	73	135	138			11
HC (K=5)	73	135	88		77	11
HC (K=6)	73	135	88		77	3 / 8
SOM (K=5)	72	148	77		77	10
SOM (K=6)	72	144	36	53	69	10
K-Means (K=5)	73	142	49	50	70	-
K-Means (K=6)	72	141	46	47	67	11
K-Medoids (K=5)	71	76/74	77/86			-
K-Medoids (K=6)	69	35/48/84	70/78			-
MCLUST (K=5)	68	36/128	69/83			-
MCLUST (K=6)	70	33/125	29	52	75	-

Normally, the logic behind validating the validity indices using real biological data is that the validity index, which assigns high values to those clustering algorithms with higher accuracy membership by referring to the knowledge of biological ground truth, is the better index. That is, there is an implicit assumption that the numerical data is consistent with the biological truth. However, it is not always the case; for example, the co-expressed genes are not necessarily co-regulated, which means that two functionally unconnected genes might be co-expressed in a certain situation (or maybe faulty experimental conditions). In our case, the Q phase exists as an independent numerical cluster no matter whether its members belong to same biological group or not. A good validity index have to indicate its existence. GPV, adjusted FOM and DI are presented in Figs. 10a, 10b and 10c, respectively. The proposed GPV, shown in Fig. 10a, indicates that the HC with five and six clusters, the SOM the five and six clusters, and the KMeans with six clusters have relatively high index values. The Q phase appears in all these clustering results as shown in Table 3. In Fig. 10c, DI shows that the highest value among all clustering results is MCLUST with 10 clusters, which does not make any sense. It shows that KMeans with six clusters, where the Q phase appears, has a high index value; while it also shows a high value of the SOM with five clusters but a low value of the SOM with six clusters, when Q phase appears in both results. All these observations at least lead to one fact that DI is not sensitive to the Q phase. Note that since adjusted FOM is a re-sampling algorithm, its result does not make any sense to individual clustering partition. Thus, our proposed GPV, which assigns high values to the clustering algorithms that discover the Q phase, is faithful to the underlying statistical patterns rather than the biological knowledge.

4.4.2 Group 2

In this part, we investigate the performance of all indices in Group 2 data sets, which are extracted from Ogawa set [46]. In Fig. 11, performance of all index values against the number of clusters K are shown respectively. Fig. 11a shows the ARI performance as a reference, which reflects that there are four clusters in the data sets and all clustering algorithms work reasonably well in both tight and loose subsets. Comparing all validity indices, we find that GPV and GI provide correct indications in both tight and loose subsets; DB, II, and silhouette have correct indications only in the tight subset, but fail in the loose subset; CH and DI fail to provide any sensible results in both subsets.

4.4.3 Group 3

All performance of index values against the number of clusters K for Group 3 data sets are shown in Fig. 12 respectively. Group 3 data sets are extracted from Gasch H_2O_2 set [45]. Fig. 12a shows the ARI performance, as we did for Group 2 data sets. In this group, only GPV provides correct indications in both tight and loose subsets, while all other indices fail to provide sensible results, even in the tight subset.

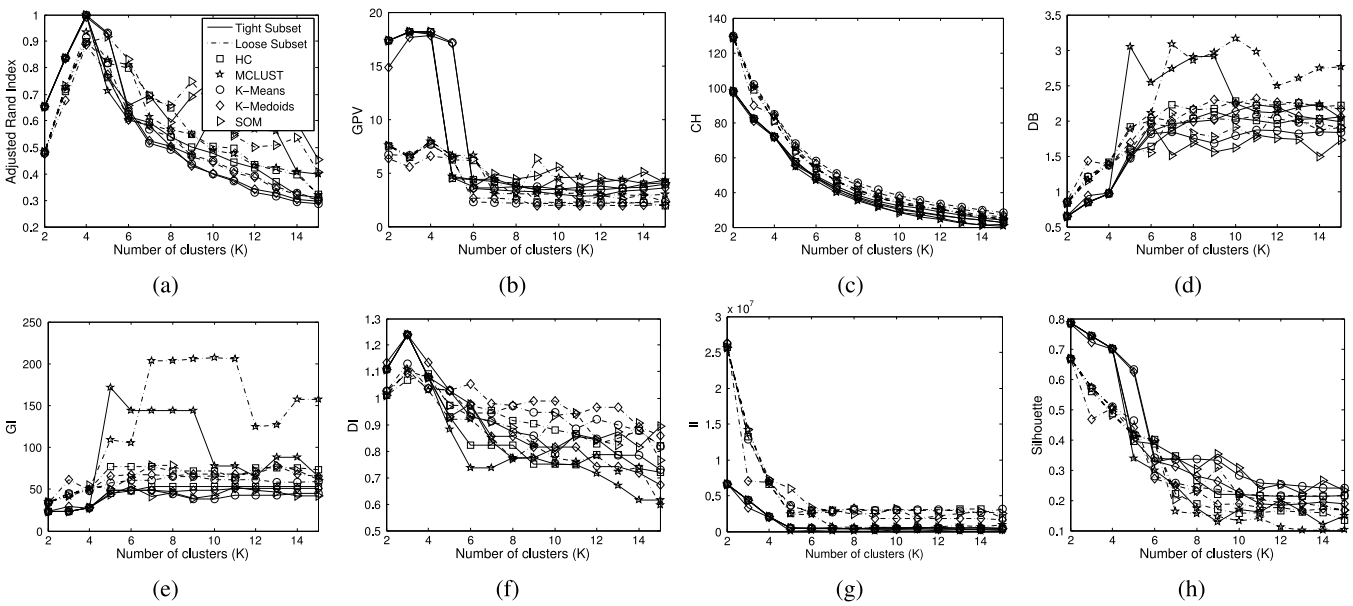

Fig. 12. Performance of index values against the number of clusters K for Group 3 data sets. (a) ARI, (b) GPV, (c) CH, (d) DB, (e) GI, (f) DI, (g) II, and (h) Silhouette. Legend: solid lines represent the performance in the tight subset; broken lines represent the performance in the loose subset; the curves with square markers, for both solid and broken lines, stand for HC; pentagon markers stand for MCLUST; circle markers stand for K-Means; diamond markers stand for K-Medoids; and triangular markers stand for SOM.

5 Conclusions and Discussions

In this paper, we proposed a new validity index, GPV, which employs two tunable parameter α and β to control the numbers of objects being used to calculate the index. The most important advantage of the GPV is that it has flexibility of tuning the parameters, which leads to its noise-resistance, in dealing with different data sets, especially microarray data sets. We discussed the physical properties of the parameters and developed several rules for the selections of the bounds of parameter values rather than parameters themselves. The rationale behind the proposed index is that in the noisy scenario, the boundary areas among clusters are more dominant in determining the quality of the clustering results. Taking all within-cluster distances and between-cluster distances may diminish the subtle difference between within-cluster and between-cluster distances, and consequently degrade the performance of the validity index.

To validate our proposed index, we first obtain reference clustering results by clustering the test data sets and evaluating them with external validation, say ARI; then we test the proposed index with these reference clustering results and compare its ability to indicate correct structure of the given data set with many other indices. By varying the noise level of two types of simulated gene expression data, we conducted a set of experiments to investigate the validation performance of all compared indices. The results suggested that the proposed GPV index has superior noise-resistance performance among all indices. We also tested the proposed GPV in three groups of real microarray gene expression data sets. Group 1 contains three data sets where one of them is relatively "clean" data, one is with moderate noise and the other one is somewhat noisy. Group 2 and Group 3 are groups of subsets extracted from Ogawa set and Gasch H_2O_2 set respectively and contain two data sets, one tight set and loose set, in each group. The experimental results support that the proposed GPV has high noise-resistant ability and high fidelity to the numerical data. In different circumstances, the GPV always has relatively robust performance and provides fairly correct judgements.

Acknowledgments

The project (Ref. NIHR-RP-PG-0310-1004-AN) is supported by National Institute for Health Research (NIHR), UK. This article summarises independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Grant Reference Number RP-PG-0310-1004). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. Prof. A. K. Nandi would like to thank TEKES for their award of the Finland Distinguished Professorship.

References

[1] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ, USA: Prentice Hall, 1988.

[2] A. Jain, R. Duin, and J. Ma, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 2, pp. 4–37, Jan. 2000.

[3] D. X. Jiang, C. Tang, and A. D. Zhan, "Cluster analysis for gene expression data: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1370–1386, Nov. 2004.

[4] R. Xu and D. II Wunsc, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[5] 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. Davi, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mol. Cell*, vol. 2, no. 1, pp. 65–73, 1998.

[6] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futche, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces Cerevisiae by microarray hybridization," *Mol. Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[7] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[8] M. P. Washburn, A. Koller, G. Oshiro, R. R. Ulaszek, D. Plouffe, C. Deciu, E. Winzeler, and J. R. Yates, "Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces Cerevisiae," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 6, pp. 3107–3112, 2003.

[9] S. A. Salem, L. B. Jack, and A. K. Nand, "Investigation of self-organizing oscillator networks for use in clustering microarray data," *IEEE Trans. NanoBiosci.*, vol. 7, no. 1, p. 65-79, Mar. 2008.

[10] J. Y, "General C-means clustering model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, p. 1197-1211, Aug. 2005.

[11] T. Kohone, "The self-organizing maps," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, Sep. 1990.

[12] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzz, "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, vol. 17, no. 10, pp. 977–987, 2001.

[13] A. Thalamuthu, I. Mukhopadhyay, X. J. Zheng, and G. C. Tsen, "Evaluation and comparison of gene clustering methods in microarray analysis," *Bioinformatics*, vol. 22, no. 19, pp. 2405–2412, 2006.

[14] B. Abu-Jamous, R. Fa, D. J. Roberts, and A. K. Nandi, "Paradigm of tunable clustering using binarization of consensus partition matrices (Bi-CoPaM) for gene discovery," *PloS One*, vol. 8, no. 2, p. e56432, 2013.

[15] B. Abu-Jamous, R. Fa, D. J. Roberts, and A. K. Nandi, "Yeast gene CMR1/YDL156W is consistently co-expressed with genes participating in DNA-metabolic processes in a variety of stringent clustering experiments," *J. R. Soc. Interface*, vol. 10, no. 81, 20120990, 2013.

[16] M. Halkidi, Y. Batistakis, and M. Vazirgianni, "Cluster validity methods: Part I," *ACM SIGMOD Rec.*, vol. 31, no. 2, pp. 40–45, 2002.

[17] M. Halkidi, Y. Batistakis, and M. Vazirgianni, "Cluster validity methods: Part II," *ACM SIGMOD Rec.*, vol. 31, no. 3, pp. 19–27, 2002.

[18] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[19] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.

[20] L. Hubert and P. Arabie, "Comparing Partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[21] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, 2005.

[22] R. Giancarlo, D. Scaturro, and F. Utro, "Computational cluster validation for microarray data analysis: Experimental assessment of clest, consensus clustering, figure of merit, gap statistics and model explorer," *BMC Bioinform*, vol. 9, no. 1, p. 462, 2008.

[23] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo, "Validating clustering for gene expression data," *Bioinform.*, vol. 17, no. 4, pp. 309–318, 2001.

[24] S. Dudoit and J. Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset," *Genome Biol.*, vol. 3, no. 7, research0036, 2002.

[25] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[26] J. Oliver, R. A. Baxter, and C. S. Wallace, "Unsupervised learning using MML," in *Proc. 13th Int. Conf. Mach. Learn.*, 1996, pp. 364–372.

[27] C. S. Wallace and D. L. Dowe, "Minimum message length and Kolmogorov complexity," *Comput. J.*, vol. 42, no. 4, pp. 270–283, 1999.
[28] C. Fraley and A. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," Dept. Statistics, Univ. Washington, Seattle, WA, USA, Tech. Rep. 329, 1998.
[29] J. C. Bezdek, "Cluster validity with fuzzy sets," *J. Cybern.*, vol. 3, pp. 58–73, 1974.
[30] W. Wang and Y. Zhang, "On fuzzy cluster validity indices," *Fuzzy Sets and Syst.*, vol. 158, pp. 2095–2117, 2007.
[31] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method," in *Proc. 5th Fuzzy Syst. Symp.*, 1989, vol. 247, pp. 247–250.
[32] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 841–847, Aug. 1991.
[33] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat. - Theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.
[34] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, 1973.
[35] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
[36] U. Maulik and S. Bandyopadhya, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.
[37] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.
[38] W. J. Krzanowski and Y. T. Lai, "A criterion for determining the number of groups in a data set using sum-of-squares clustering," *Biometrics*, vol. 44, no. 1, pp. 23–34, Mar. 1988.
[39] B. S. Y. Lam and H. Yan, "Assessment of microarray data clustering results based on a new geometrical index for cluster validity," *Soft Comput.*, vol. 11, no. 4, pp. 341–348, 2007.
[40] S. A. Salem and A. K. Nandi, "Development of assessment criteria for clustering algorithms," *Pattern Anal. Appl.*, vol. 12, no. 1, pp. 79–98, 2009.
[41] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. Royal Stat. Soc., Series B (Stat. Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.
[42] F. Gibbons and F. Roth, "Judging the quality of gene expression-based clustering methods using gene annotation," *Genome Res.*, vol. 12, pp. 1574–1581, 2002.
[43] G. Stegmayer, D. H. Milone, L. Kamenetzky, M. G. Lopez, and F. Carrari, "A biologically inspired validity measure for comparison of clustering methods over metabolic data sets," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 9, no. 3, pp. 706–716, May 2012.
[44] P. Lingras, M. Chen, and D. Q. Miao, "Rough cluster quality index based on decision theory," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 7, pp. 1014–1026, Jul. 2009.
[45] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown, "Genomic expression programs in the response of yeast cells to environmental changes," *Mol. Biol. Cell*, vol. 11, no. 12, pp. 4241–4257, 2000.
[46] N. Ogawa, J. DeRisi, and P. O. Brown, "New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces Cerevisiae revealed by genomic expression analysis," *Mol. Biol. Cell*, vol. 11, no. 12, pp. 4309–4321, 2000.
[47] A. G. de Brevern, S. Hazout, and A. Malpertuy, "Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering," *BMC bioinform.*, vol. 5, no. 1, p. 114, 2004.
[48] M. Celton, A. Malpertuy, G. Lelandais, and A. de Brevern, "Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments," *BMC Genomics*, vol. 11, no. 1, p. 15, 2010.
[49] L. P. Zhao, R. Presntice, and L. Breeden, "Statistical modelling of large microarray data sets to identify stimulus-response profiles," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 10, pp. 5631–5636, 2001.
[50] J. M. Peña, J. A. Lozano, and P. Larrañaga, "An empirical comparison of four initialization methods for the K-means algorithm," *Pattern Recognit. Lett.*, vol. 20, no. 10, pp. 1027–1040, 1999.
[51] K. Y. Yeung and W. L. Ruzzo, "Principlal component analysis for clustering gene expression data," *Bioinform.*, vol. 17, no. 9, pp. 763–774, 2001.
[52] T. Pramila, W. Wu, S. Miles, W. S. Noble, and L. L. Breeden, "The Forkhead transcription factor HCM1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle," *Genes Develop.*, vol. 20, pp. 2266–2278, 2006.
[53] R. Fa and A. K. Nandi, "Comparisons of validation criteria for clustering algorithms in microarray gene expression data analysis," in *Proc. 2nd Int. Workshop Genomic Sig. Proc.*, 2011, pp. 101–106.
[54] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY, USA: Wiley, 1990.
[55] C. Fraley and A. E. Raftery, "MCLUST: Software for model-based clustering, density estimation and discriminant analysis," Dept. Statistics, Univ. Washington, Seattle, WA, USA, Tech. Rep. 415, 2002.
[56] B. P. Tu, A. Kudlicki, M. Rowicka, and S. L. McKnigh, "Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes," *Science*, vol. 310, no. 5751, pp. 1152–1158, 2005.

Rui Fa received the bachelor's and master's degrees in electronic and electrical engineering from the Nanjing University of Science and Technology, China, in 2000 and 2003, respectively, and the PhD degree in electrical engineering from the University of Newcastle, United Kingdom, in 2007. He has held research positions in the University of York and the University of Leeds working in radar signal processing and wireless communication, since 2008. From October 2010, he started his research in bioinformatics field at the University of Liverpool and involved in a collaborative research project with the Universities of Oxford, Cambridge, and Bristol, which is funded by National Institute for Health Research (NIHR). Since 2013, he has been a senior research fellow at Brunel University. His current research interests include bioinformatics, machine learning, Bayesian statistics, statistical signal processing, and network science. He is a member of the IEEE.

Asoke K. Nandi received the PhD degree from the University of Cambridge (Trinity College), Cambridge, United Kingdom. He held academic positions in several universities, including Oxford, United Kingdom, Imperial College London, United Kingdom, Strathclyde, United Kingdom, and Liverpool, United Kingdom. In 2013, he moved to Brunel University, United Kingdom, to the Chair and Head of Electronic and Computer Engineering. He is a Finland Distinguished professor at the University of Jyvaskyla, Finland, and an adjunct professor at the University of Calgary, Canada. In 1983, he contributed to the discovery of the three fundamental particles known as W^+, W^-, and Z^0 (by the UA1 team at CERN) providing the evidence for the unification of the electromagnetic and weak forces, which was recognized by the Nobel Committee for Physics in 1984. His current research interests lie in the areas of signal processing and machine learning, with applications to biomedical data, gene expression data, functional magnetic resonance data, and communications research. He has made many fundamental and algorithmic contributions to many aspects of machine learning. He has much expertise in big data, dealing with heterogeneous data, and extracting information from multiple data sets obtained in different laboratories and different times. He has authored more than 490 technical publications, including 190 journal papers. For example, recently he published in *Blood* (2011), *PLOS ONE* (2013), *Royal Society Interface* (2013), and *NeuroImage* (2013). The Google Scholar h-index of his publications is 55. He was awarded the Institute of Electrical and Electronics Engineers (USA) Heinrich Hertz Award in 2012, the Water Arbitration Prize of the Institution of Mechanical Engineers (UK) in 1999, and the Mounbatten Premium, Division Award of the Electronics and Communications Division, of the Institution of Electrical Engineers (UK) in 1998. He is a fellow of seven institutions, including the Institute of Electrical and Electronics Engineers. He is a fellow of the IEEE.