CLUSTER VALIDATION USING GRAPH THEORETIC CONCEPTS

N. R. PAL* and J. BISWAS

Machine Intelligence Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta, India

(*Received* 31 *January* 1996; *in revised form* 29 *July* 1996; *received for publication* 9 *August* 1996)

Abstract– In this article we have generalized Dunn's index and the Davies–Bouldin index for cluster validation using graph structures, such as GG, RNG and MST. Unlike Dunn's index and the Davies–Bouldin index, the proposed indices are not sensitive to noisy points and are applicable to hyperspherical and structural clusters as well. The relationships between various indices have also been established. The effectiveness of the generalized indices and superiority over some existing cluster validity indices are established using eight data sets.

Cluster validity Gabriel graph Relative neighborhood graph Minimal spanning tree Structural clusters

1. INTRODUCTION

Clustering is a part of data analysis which is required in many fields related to pattern recognition, biological sciences, social sciences, psychology, etc. Usually, based on a similarity or dissimilarity measure, clustering divides a given data set into a number of classes so that data in a class have more similarity between them while data from different classes are dissimilar.

There are many methods for clustering but these methods are not universal. No method is good for all types of data, nor are all methods equally applicable to all problems. Even the ideas behind different clustering methods are not the same. The type of data normally determines the methodology to be effective.

Before we partition the data set, we must know whether there is any cluster structure in the data. Obtaining an answer to the question,"Is the data set clusterable?" is known as assessment of clustering tendency. If the outcome of the tendency assessment test is positive then only we should try to cluster the data. Most clustering algorithms need to know the number of classes to look for. This is an unsupervised method and in most cases, users will not have any prior knowledge about the number of clusters present in the data set. It may be so that the number of classes we are separating the data set into is larger or smaller than the actual number of classes. If it is larger then one or more good compact clusters may be broken. If it is smaller then more than one separate clusters may be merged. Thus, finding the right number of clusters is an important problem. Even when we know the right number of clusters, because of inappropriate choice of algorithmic parameters or wrong choice of the clustering algorithm itself, the generated partitions may not reflect the desired clustering of the data. The problem of choosing the right number of clusters (c), and given c, selecting the best partition are known as the cluster validity problem.

In this paper we propose some new indices for cluster validation. We have generalized Dunn's index and the Davies–Bouldin index, using the Minimal Spanning Tree (MST), Relative Neighborhood Graph (RNG) and Gabriel Graph (GG). We use the proposed indices for selecting the appropriate number of clusters, but they can also be used for choosing the best partition, given a fixed value of c.

The rest of the paper is organized as follows. In Section 2 we briefly describe the Hard c-means (HCM) algorithm as in all our simulation work we have used HCM as the clustering algorithm. Section 3 describes some existing cluster validity indices while Section 4 introduces the new indices. The usefulness of the proposed indices is established in Section 5 and the paper is concluded in Section 6.

2. HARD c-MEANS ALGORITHM (HCM)

In all our experiments we used the HCM algorithm as the clustering algorithm. For completeness, we now provide a brief description of HCM.

Let $X = \{x_1, x_2, \ldots, x_n\} \subset R^p$ be a set of unlabelled feature vectors in p-space. Clustering these vectors into c classes $(X_1, X_2, \ldots, X_c)$ means partitioning X in c subsets $X_1, X_2, \ldots, X_c$ such that

$$X = X_1 \cup X_2 \cup \cdots \cup X_i \cup \cdots \cup X_c,$$
$$X_i \cap X_j = \emptyset, \quad i \neq j, \; X_i \neq \emptyset.$$

We denote the set of all hard c-partitions of X as

$$M_{hcn} = \left\{ U \in R^{c \times n} : \text{ for } k = 1, \ldots, n; \; U_{(k)} = e_i \; \exists i. \right.$$
$$\left. 0 < \sum_{k=1}^{n} u_{ik} < n \; \forall i \right\},$$

* Author to whom correspondence should be addressed.

†The work was partially supported by University Grants Commission, India.

where

$$e_i = (0, 0, \ldots, 1, \ldots, 0)^T,$$

(here "1" occurs in the ith place) is the crisp label vector for class i. $U = [u_{ik}]_{c \times n}$, where u_{ik} is the degree to which the data point x_k belongs to class i, that is, the membership of x_k in class i. Here $u_{ik} \in \{0, 1\}$, $\sum_i u_{ik} = 1$ and U_{ik} = kth column of the partition matrix U.

The vector $U_{ik} = e_i$ indicates that the associated data point x_k completely belongs to class i (that is, full membership to class i and no membership to other classes). Hence such label vectors are often called "crisp label" vectors and M_{hcn} as the hard c-partition space for X. In the sequel, any clustering algorithm A, that can generate a hard partition $U_A \subset M_{hcn}$ of X, will be referred to as a "crisp" or "hard" clustering algorithm.

The HCM algorithm generates a crisp partition $U_{HCM} \subset M_{hcn}$ of X. The HCM model is the constrained optimization problem:

$$\min_{U,V} \left\{ J_1(U, V : X) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \|\mathbf{x}_k - \mathbf{v}_i\|_A^2 \right\},$$

where $U \in M_{hcn}$, $V = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c)$ is the set of cluster centers (prototypes). $\mathbf{v}_i \subset R^p$ for $1 \le i \le c$ and $\|\cdot\|_A$ is any inner product norm. Optimal partitions U^* are taken for pairs (U^*, V^*) that are local minimizers of J_1. An approximate solution to this optimization problem can be found by the HCM algorithm, which is based on the first-order necessary conditions for local extrema of J_1 stated as the HCM theorem.

2.1. HCM Theorem

$(U, V) \in M_{hcn} \times R^p$ may minimize J_1 only if

$$u_{ik} = \begin{cases} 1, & \|\mathbf{x}_k - \mathbf{v}_i\|_A \le \|\mathbf{x}_k - \mathbf{v}_j\|_A; \\ & j = 1, \ldots, c,\ j \ne i, \\ 0, & \text{otherwise}, \end{cases}$$
$$1 \le i \le c,\ 1 \le k \le n. \qquad (1)$$

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n} (u_{ik}) \mathbf{x}_k}{\sum_{k=1}^{n} (u_{ik})} = \frac{\sum_{\mathbf{x}_k \in X_i} \mathbf{x}_k}{n_i}, \quad 1 \le i \le c,$$

$\mathbf{v}_i$ is the mean vector of the points currently in the cluster i. n_i is the number of points in the ith cluster of X. Ties in equation (1) may be resolved arbitrarily. The HCM algorithm is an iterative process which iterates through the necessary conditions in equation (1) that attempts to minimize J_1. It produces crisp partitions of X by assigning all the membership of each $\mathbf{x}_k$ to class i when prototype $\mathbf{v}_i$ is nearest to it.

Algorithm HCM $(X, \varepsilon, c, T_{max}, A, V)$

Input: $X, \varepsilon, c, T_{max}, A$
Output: V
Here, X = data to be clustered, ε = terminating error, T_{max} = iteration limit, A = square matrix for the inner product norm

Begin Algorithm

Initialize $V_0 = (\mathbf{v}_{10}, \mathbf{v}_{20}, \ldots, \mathbf{v}_{c0}) \in R^{c \times p}$

For $t = 1$ to T_{max}

Calculate U_t with V_{t-1} and equation (1)
Calculate V_t with U_t and equation (1)
If $E_t = |V_t - V_{t-1}|_1 \le \varepsilon$

Stop

Else

Next t.

End Algorithm

3. SOME CLUSTER VALIDITY METHODS

In Section 1, we informally discussed the concept of cluster validity. Here, before describing some existing cluster validation methods, we formally introduce the notion of cluster validity.

As defined earlier, let $X = \{x_1, x_2, \ldots, x_n\} \subset R^p$ be a set of n feature vectors in p-space. Crisp clustering algorithms are formally represented as functions, $C : R^{p,n} \to M_{hcn}$.

Let $P = \{U_j | U_j \in M_{hcn},\ 1 \le j \le N\}$ be the set of N different partitions of the fixed data set X with an algorithm C_i at various values of its parameters, i.e.

$$U_j = C_i(X; (p_{i1}, p_{i2}, \ldots, p_{ik_i})).$$

Here $\{p_{ij}\}$ are the k_i parameters of algorithm C_i. For example, the parameters for HCM are $\{c, T, \varepsilon, \|\cdot\|_A, |\cdot|_{1t}, V_0\}$. Thus, using M algorithms we can feasibly generate many partitions. The number of classes, c, is a common parameter of every C_i and it exerts the most important effect on clustering. Thus, the most effective strategy for clustering is to first decide what seems to be the most reasonable estimate of the correct number of classes by choosing a C_i and fixing all its parameters except c.

Let $P = \{U_i(c) : U_i(c) = C_i(X; (c, p_{i2}, \ldots, p_{ik_i}))_i,\ c = 2, 3, \ldots, c_{max}\}$ be the set of partitions for different c for algorithm C_i.

Our problem here is to find the best value for c. There have been several attempts to solve the cluster validity problems[1-4]. Next, we briefly describe three of them which will be used later for comparison.

3.1. Modified Hubert's Statistic

The Modified Hubert's Statistic (MHS)[2] assesses the fit between a data set and the cluster structure imposed by C_i. Hubert's Γ statistic is the point serial correlation coefficient between any two matrices. When the two matrices are symmetric, Γ can be written in its raw form as

$$\Gamma(P, Q(U)) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_{ij} q_{ij}.$$

Here $P = [p_{ij}]$ is an $n \times n$ proximity matrix, p_{ij} being the observed proximity between the objects i and j (e.g. $p_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ in any norm). $Q=Q(U)$ is the unique adjacency matrix of the equivalence relation induced on X by U. $Q=[q_{ij}]$ is defined in terms of any hard c-partition U of X as

$$q_{ij} = \begin{cases} 0, & u_{ik} = u_{jk} = 1 \text{ for some } k, \\ 1, & \text{otherwise.} \end{cases}$$

Here u_{ik} is the membership of x_i in class k.

In its normalized form, Γ becomes the sample correlation coefficient between the entries of P and Q:

$$\hat{\Gamma}(P, Q(U)) = \frac{[(1/M)\sum_{i}^{n} \sum_{j=i+1}^{n} (p_{ij} - \bar{p})(q_{ij} - \bar{q})]}{(S_P S_Q)}, \tag{2}$$

where $M = n(n-1)/2$ is the total number of entries under the double summation,

$$\bar{p} = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_{ij} \qquad \bar{q} = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} q_{ij}$$

$$S_P^2 = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_{ij}^2 - \bar{p}^2, \qquad S_Q^2 = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} q_{ij}^2 - \bar{q}^2.$$

For normalized index $-1 \le \hat{\Gamma} \le 1$. If P and Q are not symmetric then all summations are extended over all n entries and $M = n^2$. $\hat{\Gamma}$ measures the degree of linear correspondence between the entries of P and Q. A positive value of Γ close to 1 indicates that P and Q are (more or less) linearly correlated.

Use of Γ or $\hat{\Gamma}$ for cluster validity requires knowledge of their distribution, which can be found by computing Γ or $\hat{\Gamma}$ for all $n!$ permutations. To avoid such a computationally expensive process, the MHS has been proposed and used for cluster validity. MHS is Hubert's raw or normalized statistic computed on P and $Q_v(U)$. The indices are

$$V_{\mathrm{MH}}(c) = \Gamma(P, Q_v(U_i(c))) = \text{Hubert's modified raw statistic.}$$

$$V_{\mathrm{MH}\hat{\Gamma}}(c) = \hat{\Gamma}(P, Q_v(U_i(c))) = \text{Hubert's modified normalized statistic.}$$

where $Q_v = Q_v(U) = [q_{v,ij}] = [\,|\mathbf{v}_{U(i)} - \mathbf{v}_{U(j)}|_t]$, $U(i)=k$ if ith object is in kth cluster. $|\mathbf{v}_i - \mathbf{v}_j|_t$ is the Euclidean distance between the cluster centroids. These indices are not defined for $c=1$ or $c=n$. It has been seen that these indices tend to increase with an increase of c. For this, the value of V_{MH} or $V_{\mathrm{MH}\hat{\Gamma}}$ is not used to select c; rather the change in the value as a function of c is examined. For well separated clusters a sharp knee is expected at the partition $U_i(c)$ which contains the number of clusters that provide the best fit to the data as measured by this statistic.

3.2. *Davies-Bouldin index*

The Davies–Bouldin Index (DBI)[11] is based on the idea that for a good partition inter cluster separation as well as intra cluster homogeneity and compactness should be high. According to Davies and Bouldin,[11] a real-valued function R of s_i, s_j and d_{ij} (where s_i, s_j are some measure of dispersion for class i and j, respectively and d_{ij} is the distance between the prototypes of clusters i and j) can be taken as a cluster similarity measure if it satisfies the following five conditions c_1–c_5:

c_1: $R_{ij} \ge 0$,
c_2: $R_{ij} = R_{ji}$,
c_3: $R_{ij} = 0$ if $s_i = 0$ and $s_j = 0$,
c_4: if $s_i = s_k$ and $d_{ij} < d_{ik}$ then $R_{ij} > R_{ik}$,
c_5: if $d_{ij} = d_{ik}$ and $s_j > s_i$ then $R_{ij} > R_{ik}$.

For notational simplicity we used R_{ij} for $R(s_i, s_j, d_{ij})$.

Conditions c_1–c_5 imply that the function R is nonnegative and symmetric. The similarity between clusters is zero only if their dispersion functions vanish. If the distance between clusters increases while their dispersion remains the same, then their similarity decreases. If the dispersion of clusters increases while distance remains constant, then their similarity increases. In reference (1), the dispersion s_i of ith cluster and the separation d_{ij} between ith and jth clusters are defined as

$$s_{i,q} = \left(\frac{1}{|X_i|}\sum_{\mathbf{x}\in X_i} |\mathbf{x} - \mathbf{v}_i|_t^q\right)^{1/q} \tag{3}$$

and

$$d_{ij,t} = \left(\sum_{l=1}^{n} |\mathbf{v}_{i,l} - \mathbf{v}_{j,l}|^t\right)^{1/t} = \text{the Minkowsky } q\text{-norm for } q = t. \tag{4}$$

For a given U, $\mathbf{v}_i$ is the centroid of ith class. $q, t \ge 1$, q is an integer and q, t can be selected independent of each other.

The DBI index is then defined as

$$V_{\mathrm{DB},qt}(c) = \frac{1}{c}\sum_{i=1}^{c} \max_{\substack{j \\ j\neq i}} \{R_{ij}\}, \tag{5}$$

where

$$R_{ij} = \frac{s_{i,q} + s_{j,q}}{d_{ij,t}}. \tag{6}$$

Since the goal is to achieve minimum within-cluster dispersion and maximum between-cluster separation, the number of clusters c that minimizes $V_{\mathrm{DB},qt}$ is taken as the optimal value of c. Note that R_{ij} in equation (6) satisfies c_1–c_5.

3.3. Dunn's indices

Dunn[5] defined the separation index for U, a hard c-partition of X, as

$$V_D(c) = \min_{1\le i\le c}\left\{\min_{\substack{1\le j\le c\\ j\ne i}}\left(\frac{\text{dist}(X_i, X_j)}{\max_{1\le k\le c}\{\text{diam}(X_k)\}}\right)\right\}, \quad (7)$$

where

$$\text{diam}(X_i) = \max_{x,y\in X_i}\{d(x,y)\} \text{ and}$$

$$\text{dist}(X_i, X_j) = \min_{x\in X_i, y\in Y_j}\{d(x,y)\}.$$

Dunn proved that X can be clustered into a compact and well separated (CWS) c-partition with respect to d if $\max_{1<c<M_{hc}}\{V_D(c)\} > 1$.

Intuitively, large values of V_D correspond to good clusters, so the number of clusters c that maximizes V_D can be taken as the optimal value of c. Note that $\text{diam}(X_k)$ and $\text{dist}(X_i, X_j)$ may be severely affected by one or two noisy points. Consequently, $V_D(c)$ may not perform satisfactorily as a cluster validity index. This issue has been addressed in reference (6). In reference (6), the authors have also provided some generalizations of Dunn's indices (DI) which are not as sensitive as DI to noisy points for hyperspherical type clusters. In this paper we propose some other generalizations of V_D which are expected to perform well for both hyperspherical and structural (like circles, shells, etc.) clusters.

Dunn[5] also defined a second index for CWS clusters as

$$V'_D(c) = \min_{1\le i\le c}\left\{\min_{\substack{1\le j\le c\\ j\ne i}}\left(\frac{\text{dist}(X_i, \text{conv}(X_j))}{\max_{1\le k\le c}\{\text{diam}(X_k)\}}\right)\right\}, \quad (8)$$

where $\text{conv}(S)$ is the convex hull of S. Dunn proved that X can be partitioned into CWS clusters if $\max_{1<c<M_{hc}}\{V'_D(c)\} > 1$. Since $\text{conv}(S)$ is computationally expensive, it is not used generally.

4. PROPOSED GRAPH THEORY BASED INDICES

Different types of geometric structures or graphs can be used to impose a structure on a multidimensional data set. The nodes of the graph are the data points and the edges represent the relation between the nodes.

Here three types of graph structures have been used.[3, 9] These are the Minimal Spanning Tree (MST), Relative Neighborhood Graph (RNG) and Gabriel Graph (GG). These graphs have been successfully used for clustering.[7–9] We use them for cluster validity. We impose these graph structures on the partitioned data set to obtain information about the quality of the partition or clusters. We have introduced two new indices, one of which is based on DI, while the other is defined using the concept of DBI. Before describing the proposed method, we first introduce the relevant graph types.

A graph $G = (V, E)$ is a pair where $V = \{v_1, v_2, \ldots, v_m\}$ is a set of vertices and $E = \{e_1, e_2, \ldots, e_n\}$ is a set of distinct edges. Each edge $e_q = \{v_i, v_j\}$ connects a pair of vertices. A graph $G_s = (S, E_s)$ is a subgraph of $G = (V, E)$ if $S \subset V$, $E_s \subset E$. In other words, $E_s = \{e_u = \{x_i, x_j\} \mid e_u \in E,\ x_i \in S,\ x_j \in S\}$. A graph where every two vertices are connected is called a complete graph. A path $e_1, e_2, \ldots, e_u$ joining two vertices x_i and x_j is a sequence of edges such that any two successive edges are adjacent, any vertex of the path is adjacent to two of its edges at most, x_i is adjacent to the first edge only and x_j is adjacent to the last edge only.

Minimal Spanning Tree (MST). A spanning tree in a graph G is a minimal subgraph connecting all the vertices of G. If G is a weighted graph (i.e. there is a real number associated with each edge of G), then the weight of the spanning tree T of G is defined as the sum of the weights of all the branches in T. In general, different spanning trees of G will have different weights. A spanning tree with the smallest weight in a weighted graph is called a Minimal Spanning Tree (MST). Figure 1(b) depicts an MST of the graph in Fig. 1(a).

Relative Neighborhood Graph (RNG). Let x_i, x_j be two data points. They are connected in the RNG if

$$d(x_i, x_j) \le \max\{d(x_i, x_k), d(x_j, x_k) \forall k,\ k \ne i,\ k \ne j\},$$

where $d(x_i, x_j)$ is the Euclidean distance between x_i and x_j. In other words x_i, x_j are connected in RNG if no other point falls in $\text{LUNE}(x_i, x_j)$, where $\text{LUNE}(x_i, x_j)$ is the intersection of the two disks of radius $d(x_i, x_j)$ and having centers at x_i and x_j. This is the region of influence of RNG. Figure 2(a) illustrates $\text{LUNE}(x_i, x_j)$ and Fig. 2(b) represents the RNG of Fig. 1(a).

Gabriel Graph (GG). The points x_i, x_j are connected in GG if

$$d^2(x_i, x_j) < d^2(x_i, x_k) + d^2(x_j, x_k) \quad \forall k,\ k \ne i,\ k \ne j,$$

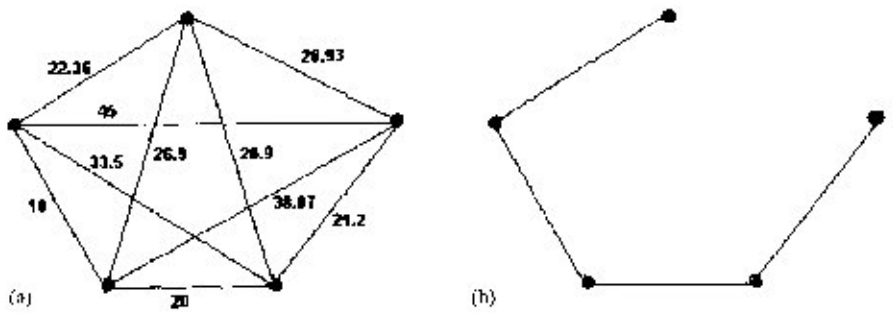

Fig. 1. (a) Graph G. (b) An MST of graph G.

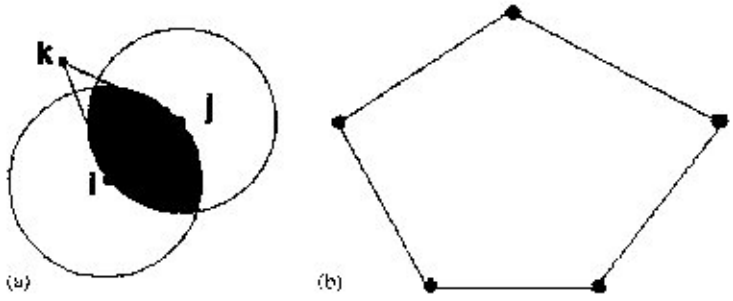

Fig. 2. (a) LUNE x_i, x_j. (b) RNG of graph G.

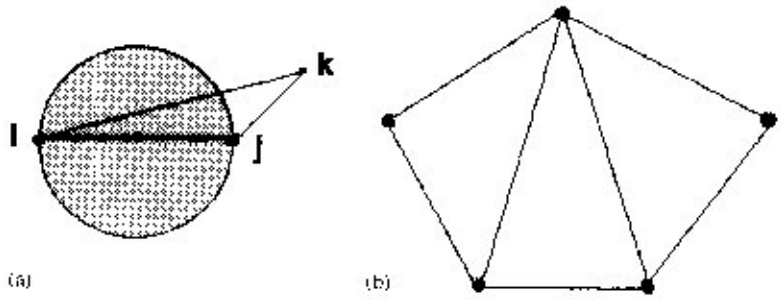

Fig. 3. (a) DISK x_i, x_j. (b) GG of graph G.

i.e. x_i and x_j are connected in GG if no other points lie in $\mathbf{DISK}(x_i, x_j)$, where $\mathrm{DISK}(x_i, x_j)$ is the disk with diameter $d(x_i, x_j)$ centered at the mid point of x_i and x_j. Figure 3(a) shows $\mathbf{DISK}(x_i, x_j)$ for a pair (x_i, x_j) and Fig. 3(b) displays the GG of Fig. 1(a).

Suppose we have a data set $X \in R^p$ which has been clustered into c classes, $X_1, X_2, \ldots, X_c$. In order to define cluster validity indices, we need to define diameter of a class and separation between classes. We propose to define these in terms of edges of the three special types of graphs just discussed. In the following we define different indices in terms of GG, but the indices can be written using MST and RNG, as well, in a similar manner.

Let $E_i^{GG} = \{e_{ij}^{GG},\ 1 \le j \le l_i,$ where l_i is $|E_i|;\ 1 \le i \le c\}$ be the set of edges of the GG computed on X_i.

We define the diameter d_i^{GG} of the cluster X_i as

$$d_i^{GG} = \max_j \{e_{ij}^{GG},\ j = 1, 2, \ldots, l_i\}. \tag{9}$$

The maximum of all diameters gives the maximum possible spread D of all clusters in the partition; i.e.

$$D^{GG} = \max\{d_i,\ 1 < i \le c\}.$$

Note that for a good partition, D will have a lower value as compared to a bad partition. If two separated clusters are merged in a partition, then D will be higher for that partition. The diameter of a set as defined by Dunn is heavily influenced by the presence of noisy points, while the effect of noisy points on equation (9), is not much.

We define the separation $d_{ij} = \mathrm{dist}(X_i, X_j)$ between two clusters X_i and X_j by the distance between the cluster centers of class i and j, i.e.

$$d_{ij} = \|\mathbf{v}_i - \mathbf{v}_j\|. \tag{10}$$

We can now define an index of cluster validity

$$V_D^{GG}(c) = \min_{1 \le i \le c} \left\{ \min_{\substack{1 \le j \le c \\ j \ne i}} \left\{ \frac{d_{ij}}{\max_{1 \le k \le c} \{d_k^{GG}\}} \right\} \right\}. \tag{11}$$

For good clustering of X, the numerator will be large and the denominator will be smaller, resulting in a higher value of $V_D^{GG}(c)$. Thus, to obtain the optimal c, we maximize V_D^{GG} with respect to c.

Similarly, we can define two other indices V_D^{RNG} and V_D^{MST} as

$$V_D^{RNG}(c) = \min_{1 \le i \le c} \left\{ \min_{\substack{1 \le j \le c \\ j \ne i}} \left\{ \frac{d_{ij}}{\max_{1 \le k \le c} \{d_k^{RNG}\}} \right\} \right\} \tag{12}$$

and

$$V_D^{MST}(c) = \min_{1 \le i \le c} \left\{ \min_{\substack{1 \le j \le c \\ j \ne i}} \left\{ \frac{d_{ij}}{\max_{1 \le k \le c} \{d_k^{MST}\}} \right\} \right\}. \tag{13}$$

Note that V_D^{GG}, V_D^{MST} and V_D^{RNG} are, in a sense, generalization of DI.

Next, we define another family of three indices using the paradigm of Davies and Bouldin.[1] Again we shall introduce the new measure using GG, but the definition could be written analogously for the other two types of graphs also.

d_i^{GG}, the maximum Gabriel edge length of class i, i.e. the diameter of class i, can also be viewed as a measure of dispersion s_i of the ith class, i.e. $d_i^{GG} = s_i$.

Thus, we define a cluster similarity measure for two clusters i and j, as

$$R_{ij}^{GG} = \frac{s_i + s_j}{d_{ij}} = \frac{d_i^{GG} + d_j^{GG}}{d_{ij}}, \tag{14}$$

R_{ij}^{GG} satisfies conditions c_1–c_5 as shown below.

Here $R_{ij}^{GG} \geq 0$ as $d_i^{GG} \geq 0$, $d_j^{GG} \geq 0$ and $d_{ij} > 0$, ignoring the pathological case of $d_{ij} = 0$;

$$R_{ij}^{GG} = 0 \Leftrightarrow d_i^{GG} = d_j^{GG} = 0 \Leftrightarrow d_i^{GG} = 0 \text{ and } d_j^{GG} = 0.$$

If $d_j^{GG} = d_i^{GG}$ and $d_{ij} \leq d_{ik}$, then

$$\frac{d_i^{GG} + d_j^{GG}}{d_{ij}} \geq \frac{d_i^{GG} + d_k^{GG}}{d_{ik}} \Rightarrow R_{ij}^{GG} \geq R_{ik}^{GG}.$$

Thus R_{ij}^{GG} can be used as a cluster similarity measure.

Hence the cluster validity index V_{DB}^{GG} can be defined as

$$V_{DB}^{GG} = \frac{1}{c}\sum_{i=1}^{c} \max_{\substack{j \\ j \neq i}} \left(\frac{d_i^{GG} + d_j^{GG}}{d_{ij}} \right), \tag{15}$$

Similarly, we can define V_{DB}^{RNG} and V_{DB}^{MST} as

$$V_{DB}^{RNG} = \frac{1}{c}\sum_{i=1}^{c} \max_{\substack{j \\ j \neq i}} \left(\frac{d_i^{RNG} + d_j^{RNG}}{d_{ij}} \right), \tag{16}$$

$$V_{DB}^{MST} = \frac{1}{c}\sum_{i=1}^{c} \max_{\substack{j \\ j \neq i}} \left(\frac{d_i^{MST} + d_j^{MST}}{d_{ij}} \right). \tag{17}$$

In order to get the right choice for the number of clusters c, we minimize $V_{DB}(* = GG, RNG, MST)$ with respect to c.

It is known that,[2]

$$e^{MST} \subset e^{RNG} \subset e^{GG}, \tag{18}$$

where e^x is the set of edges in x = MST or RNG or GG.

From equation (18) we can infer that,

$$\min_i \{e_i \in e^{GG}\} \leq \min_i \{e_i \in e^{RNG}\} \leq \min_i \{e_i \in e^{MST}\} \tag{19}$$

and

$$\max_i \{e_i \in e^{GG}\} \geq \max_i \{e_i \in e^{RNG}\} \geq \max_i \{e_i \in e^{MST}\}. \tag{20}$$

We now characterize the proposed indices through the following results.

Lemma 1. Given a partition with c clusters,

$$\begin{aligned} &d_i^{GG} \geq d_i^{RNG} \geq d_i^{MST} \quad \text{for } i = 1, 2, \ldots, c; \\ &\text{and } D^{GG} \geq D^{RNG} \geq D^{MST}. \end{aligned} \tag{21}$$

Proof. Diameter of the class i is defined as

$$d_i^* = \max_j \{e_{ij}^*,\ j = 0, 1, \ldots, l_i,\ l_i = |E_i|\},$$
$$i = 0, 1, \ldots, c,$$

where *—GG or RNG or MST.

From equation (20) we see that,

$$d_i^{GG} \geq d_i^{RNG} \geq d_i^{MST} \quad \text{for } i = 1, 2, \ldots, c.$$

Since $D^* = \max\{d_i^*,\ i = 1, 2, \ldots, c\}$, again from equation (20) we find that $D^{GG} \geq D^{RNG} \geq D^{MST}$.

Our next result shows the relationship between different cluster validity indices introduced earlier.

Theorem 1. For a given partition of the data set X, $V_D^{GG} \leq V_D^{RNG} \leq V_D^{MST}$.

Proof. Since the numerators of equations (11) (13) are the same and by Lemma 1, $D^{GG} \geq D^{RNG} \geq D^{MST}$, the result follows.

Theorem 2. For a given partition of the data set X, $V_{DB}^{GG} \geq V_{DB}^{RNG} \geq V_{DB}^{MST}$.

Proof. Since the denominators of equations (15) (17) are the same and by Lemma 1, $D^{GG} \geq D^{RNG} \geq D^{MST}$, the result follows.

5. RESULTS

We have used seven data sets $X_1, X_2, \ldots, X_7$.

X_1 = IRIS[10] is a four-dimensional (p=4) data set. It contains 150 data points. As these data are obtained from observations over three different physical classes of flowers, the data set is supposed to contain three classes. But in their numerical representation, two of the classes have a large overlap while the third is well separated from the other two. Thus, for IRIS data the number of classes can be taken as 3 as well as 2.

X_2 = CRUD-OIL[1-1] is a five-dimensional (p=5) data set which contains 56 points. This data set is also supposed to have three classes. Scatterplots of every pair of features indicated significant overlaps between two classes. Consequently, a clustering algorithm may fail to extract the clusters and the best guess from a cluster validity algorithm may be c=2.

The remaining data sets are synthetically generated.

X_3 = 3-BOXES is a two-dimensional (p=2) data set containing 60 points. It consists of three well separated squares in the first quadrant.

X_4 = NORMAL3[4] is a four-dimensional (p=4) data set with 400 points. It has been generated by drawing 100 points each from four multivariate normal distributions with population mean $\mu_i = 3e_i$ and covariance $\Sigma_i = I_4$, $i = 1, 2, 3, 4$.

X_5 = NORMAL2, X_6 = NORMAL1 and X_7 = NORMAL0.5[4] are generated by the same procedure as that for X_4 with the same covariance matrix but different mean vectors, namely, $\mu_i = 2e_i$ for X_5, $\mu_i = e_i$ for X_6, and $\mu_i = 0.5e_i$ for X_7.

Thus in X_4 conceptually we have four well-separated balls along the four axes. In X_5 the balls are closer and possibly the cluster structure can still be found. For X_6 and X_7 the cluster structure will be lost for all practical purpose.

For each of these data sets we ran the HCM algorithm with the same parameters but different values of c. We took $\epsilon = 0.0001$, $\|\cdot\|_A$ as the Euclidean norm, $\|\cdot\|_{err}$ as the 1-norm on $R^{n\times p}$, $V_0 = c$ randomly chosen distinct points in X.

For each data set, with the same c, the initial centroids are taken to be the same while computing different indices. In other words, for a given data set and a given c, all indices are computed on the *same* partition. The results with different initializations have been investigated and are found to be similar. Following reference (6) for all data sets we report results for $c=2$ to $c=10$. However, we ran the HCM algorithm up to $c=15$. Changing the maximum number of clusters from 10 to 15 does not alter the decision except for one or two cases which will be reported in the appropriate place. Note that, for MHS, the detection of the knee is a difficult task. We have reported our best guess about the position of the knee, and for some data sets we failed to locate any knee. In each table, the optimal value of each index is written in bold face. The value of MHS at the knee is also in bold face.

In Table 1, results obtained from IRIS data are shown. All the indices except DI and MHS show $c=2$ while DI indicates $c=3$ and MHS shows $c=4$. This result is satisfactory because in IRIS two of the three classes have a large overlap.

Results of CRUD-OIL data are depicted in Table 2. It is seen that except DI and MHS all other indices are showing $c=2$. The DI shows $c=8$ while MHS supports $c=4$. We mention here that for X_2, DI is found to be very sensitive to the initialization of the HCM algorithm. This could be for two reasons: sensitivity of DI on noisy points and strong dependence of the HCM partitions on initialization because of the overlapped structure in X_2.

Table 1. Values of different cluster validity indices for IRIS

c	V_D^{CC}	V_{DB}^{CC}	V_D^{RNG}	V_{DB}^{RNG}	V_D^{MST}	V_{DB}^{MST}	V_{DU}	V_D	V_{MH}
2	**2.395426**	**0.678279**	**2.395426**	**0.636418**	**2.395426**	**0.625805**	**0.474366**	0.005853	0.834592
3	1.753869	0.872109	2.089181	0.726676	2.195607	0.711208	0.725587	**0.009763**	0.917663
4	1.182951	1.194019	1.409113	0.912909	1.480895	0.905256	0.838556	0.006780	**0.948924**
5	0.775874	1.427597	0.924208	1.119402	0.971289	1.099600	0.977069	0.004651	0.950123
6	0.775874	1.531209	0.924208	1.269473	0.971289	1.262377	0.991749	0.007273	0.966005
7	0.707338	1.699257	0.842570	1.344238	0.885491	1.339239	1.002672	0.006873	0.968544
8	0.775874	1.727979	0.924208	1.324436	0.971289	1.315729	1.012749	0.007547	0.973041
9	0.492909	1.880016	0.587145	1.487421	0.617055	1.473506	1.015235	0.007547	0.974639
10	0.528389	1.988695	0.629409	1.606379	0.661472	1.534061	1.172512	0.009479	0.978235

Optimal value of each index is denoted in bold.

Table 2. Values of different cluster validity indices for CRUD-OIL

c	V_D^{CC}	V_{DB}^{CC}	V_D^{RNG}	V_{DB}^{RNG}	V_D^{MST}	V_{DB}^{MST}	V_{DU}	V_D	V_{MH}
2	**2.075988**	**0.836688**	**2.814733**	**0.685026**	**3.032582**	**0.648719**	**0.704056**	0.008920	0.678384
3	1.440645	1.112782	2.104480	0.889284	2.104480	0.865176	0.756643	0.012384	0.848407
4	0.818074	1.566213	1.195034	1.138765	1.195034	1.125589	0.932857	0.008080	**0.894894**
5	0.520710	2.248303	0.760648	1.636724	0.760648	1.618060	1.377626	0.008080	0.909040
6	0.770979	1.805026	0.829986	1.492055	0.829986	1.456706	1.146692	0.008080	0.905223
7	0.888399	1.569695	0.935800	1.425780	0.935800	1.399098	1.011990	0.025115	0.929155
8	0.786080	1.637396	0.828022	1.390225	0.828022	1.376193	0.951914	**0.045874**	0.937581
9	0.577050	1.640272	0.607839	1.524245	0.607839	1.492903	1.028606	0.045874	0.944498
10	0.481465	1.566633	0.507154	1.456654	0.507154	1.428446	0.969305	0.029725	0.944511

Optimal value of each index is denoted in bold.

Table 3. Values of different cluster validity indices for 3-BOXES

c	V_D^{CC}	V_{DB}^{CC}	V_D^{RNG}	V_{DB}^{RNG}	V_D^{MST}	V_{DB}^{MST}	V_{DU}	V_D	V_{MH}
2	2.601857	0.518548	2.601857	0.507463	2.601857	0.502441	0.433401	0.216326	0.748586
3	**4.689130**	**0.341970**	**5.111313**	**0.291948**	**5.328658**	**0.250309**	**0.297341**	**1.667148**	**0.977539**
4	0.917450	0.688878	1.000052	0.653155	1.042577	0.625998	0.705849	0.022822	0.979681
5	0.917450	1.005219	1.000052	0.912774	1.042577	0.888593	0.916286	0.022822	0.981854
6	0.917450	0.848283	1.000052	0.807018	1.042577	0.788065	0.751777	0.022822	0.984006
7	1.505797	0.891021	1.551977	0.850881	1.839470	0.798851	0.833029	0.022822	0.985337
8	1.102407	1.129484	1.445029	0.862773	1.712710	0.823885	0.784617	0.016904	0.986672
9	1.182384	0.996152	1.218646	0.911040	1.444391	0.876473	0.737104	0.022315	0.987254
10	1.156778	1.064423	1.192255	0.982782	1.413111	0.951672	0.722149	0.017381	0.987697

Optimal value of each index is denoted in bold.

Table 4. Values of different cluster validity indices for NORMAL3

c	V_D^{GG}	V_{DB}^{GG}	V_D^{RNG}	V_{DB}^{RNG}	V_D^{MST}	V_{DB}^{MST}	V_{DBI}	V_D	V_{MHI}
2	0.897570	2.030707	1.238693	1.558489	1.280038	1.419855	1.651551	0.001774	0.305835
3	1.221244	1.558235	1.387893	1.275885	1.358978	1.305016	1.309436	0.003610	0.479035
4	**1.365872**	**1.376641**	**1.550772**	**1.176334**	**1.555113**	**1.136158**	**0.960587**	0.003769	**0.617933**
5	0.705823	1.898983	0.744621	1.509851	0.748977	1.400830	1.310804	0.003769	0.628612
6	0.676002	2.228070	0.768247	1.692587	0.759655	1.650330	1.476371	**0.005265**	0.673828
7	0.612149	2.432249	0.759743	1.884236	0.677807	1.571043	1.615520	0.002615	0.655790
8	0.660693	2.417408	0.697010	1.869506	0.788915	1.693354	1.627074	0.004532	0.701720
9	0.676865	2.501661	0.697010	1.963815	0.718352	1.790657	1.636045	0.004506	0.697447
10	0.717900	2.530463	0.739267	1.951735	0.753321	1.751380	1.631872	0.004292	0.707763

Optimal value for each index is denoted in bold.

Table 5. Values of different cluster validity indices for NORMAL2

c	V_D^{GG}	V_{DB}^{GG}	V_D^{RNG}	V_{DB}^{RNG}	V_D^{MST}	V_{DB}^{MST}	V_{DBI}	V_D	V_{MHI}
2	0.705975	2.571567	1.105666	1.710855	1.107307	1.670509	1.882509	0.000501	0.327746
3	0.762700	2.255044	0.907033	1.795033	1.055449	1.594779	1.543743	0.003890	0.426844
4	**1.029115**	**1.781134**	**1.320175**	**1.423210**	**1.283738**	**1.389919**	**1.231225**	0.004174	0.484216
5	0.676996	2.216568	0.860837	1.774511	0.890555	1.655441	1.413089	0.002930	0.512098
6	0.731053	2.157618	0.887965	1.716933	0.833227	1.720372	1.468493	0.002657	0.522344
7	0.590565	2.404976	0.924916	1.762479	0.890133	1.775315	1.468717	0.002981	0.556307
8	0.562516	2.557078	0.880988	1.850478	0.864769	1.765378	1.538023	0.002412	0.571317
9	0.562939	2.529467	0.881649	1.854167	0.862264	1.646962	1.542558	0.006123	0.604699
10	0.439385	2.605717	0.596840	1.980370	0.691774	1.880642	1.620202	**0.008006**	0.616677

Optimal value for each index is denoted in bold.

Table 3 displays the results obtained for the 3-BOXES data. We see that all indices indicate $c=3$ to be the best choice. This is to be expected as this data set contains three well separated clusters.

From Tables 4 and 5 we find that for NORMAL3 and NORMAL2 all indices except V_D indicate the right value of c, i.e. $c=4$. Although in NORMAL3 and NORMAL2, the four clusters are measurably separated, DI fails to pick up the correct value of c. For X_4= NORMAL3, DI indicates $c=6$ while for X_3=NORMAL2 it suggests $c=8$. V_{MHI}, on the other hand, can pick up the correct value of c for X_4 but fails for X_5. As an illustration, Fig. 4(a) shows the plot of V_{MHI} for X_4.

For X_6=NORMAL1, Table 6 reveals that V_D^{GG} and V_D^{RNG} could find the correct value of $c=4$ while V_{DB}^{GG}, V_{DB}^{MST}, V_{DBI}, indicate $c=2$ and V_{DB}^{RNG} favors $c=3$. On the other hand, V_D^{MST} and V_D support $c=6$ and $c=10$, respectively. We have already mentioned that for X_7=NORMAL0.5, the cluster structure is lost. This is also indicated by the results reported in Table 7. Except for V_D^{GG}, all indices point out the incorrect number of clusters. For X_5, X_6 and X_7, MHS failed to show clear knee in the respective plots of V_{MHI}. Figure 4(b) shows the plot of V_{MHI} for X_6. It is almost impossible to detect any knee for Fig. 4(b). The graphs of V_{MHI} for X_5 and X_7 were similar. For NORMAL1 and NORMAL0.5, the

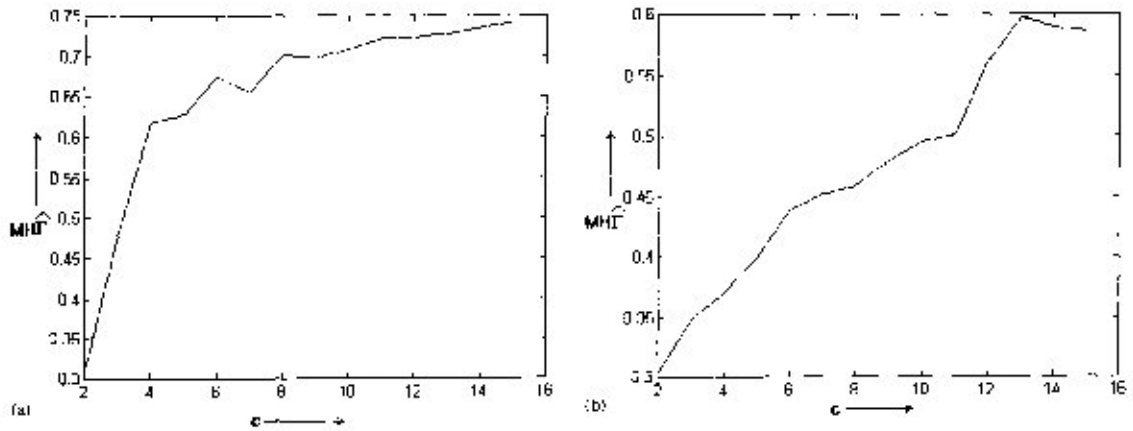

Fig. 4. (a) Plot of MHI for NORMAL3. (b) Plot of MHI for NORMAL1.

Table 6. Values of different cluster validity indices for NORMAL1

c	V_D^{GG}	V_{DB}^{GG}	V_D^{RNG}	V_{DB}^{RNG}	V_D^{MST}	V_{DB}^{MST}	V_{DBI}	V_D	V_{MH}
2	0.619355	3.102576	0.972360	1.849818	0.972360	1.847014	1.984221	0.002089	0.302763
3	0.626239	2.864982	0.983167	1.876096	0.972806	1.838714	1.776556	0.003006	0.348183
4	**0.718958**	2.390379	**1.040235**	1.826190	1.002248	1.755936	1.559612	0.001687	0.370320
5	0.674214	2.330367	0.894015	1.770653	1.010477	1.756109	1.459412	0.002768	0.410462
6	0.698493	**2.258769**	0.997535	**1.652606**	**1.018217**	**1.688157**	**1.434463**	0.002933	0.437350
7	0.668663	2.355577	0.920380	1.798702	0.769224	1.803021	1.492716	0.003550	0.451360
8	0.626221	2.352080	0.848597	1.741158	0.925492	1.750541	1.461776	0.002122	0.458742
9	0.556307	2.491958	0.785719	1.817097	0.831804	1.736978	1.508810	0.001734	0.478853
10	0.541747	2.526476	0.765155	1.800321	0.818739	1.706669	1.483891	**0.004483**	0.495291

Optimal value for each index is denoted in bold.

Table 7. Values of different cluster validity indices for NORMAL0.5

c	V_D^{GG}	V_{DB}^{GG}	V_D^{RNG}	V_{DB}^{RNG}	V_D^{MST}	V_{DB}^{MST}	V_{DBI}	V_D	V_{MH}
2	0.656798	2.873550	0.887640	1.990385	0.960382	1.827727	2.131290	0.001360	0.277681
3	0.732374	2.451159	0.989779	1.825374	1.174267	1.604411	1.740866	0.002933	0.342625
4	**0.793969**	2.352568	0.956010	1.820611	1.174335	1.498847	1.659138	0.001888	0.352638
5	0.753667	2.284601	**1.018556**	**1.636806**	**1.255538**	**1.491872**	1.448650	0.000882	0.378323
6	0.758244	**2.212553**	0.912993	1.761544	1.145176	1.533243	1.493035	0.001149	0.410200
7	0.648991	2.309624	0.902366	1.711041	1.055647	1.579245	1.470256	0.001632	0.451972
8	0.653544	2.300060	0.929834	1.716736	1.046422	1.624712	**1.365516**	0.001897	0.465437
9	0.616329	2.588835	0.835756	1.739146	0.931332	1.642556	1.435599	0.001953	0.510732
10	0.660475	2.549896	0.870203	1.766926	1.028333	1.643503	1.424026	**0.005776**	0.514679

Optimal value for each index is denoted in bold.

Table 8. Summary

Data set:	Value of selected c							No. of correct cases
Index	X_1	X_2	X_3	X_4	X_5	X_6	X_7	
V_D^{GG}	**2**	**2**	**3**	**4**	**4**	**4**	**4**	7
V_{DB}^{GG}	**2**	**2**	**3**	**4**	**4**	15	6	5
V_D^{RNG}	**2**	**2**	**3**	**4**	**4**	**4**	5	6
V_{DB}^{RNG}	**2**	**2**	**3**	**4**	**4**	6	5	5
V_D^{MST}	**2**	**2**	**3**	4	4	4	6	5
V_{DB}^{MST}	**2**	2	**3**	4	4	13	5	5
V_{DBI}	**2**	**2**	**3**	**4**	4	15	13	5
V_D	3	8	**3**	6	10	12	10	2
V_{MH}	**4**	**4**	**3**	**4**	*	*	*	2

Optimal value for each index is denoted in bold.
*, no knee detected on graph of V_{MH}.

Table 9. Values of different cluster validity indices for ARC

Partition	c	V_D^{GG}	V_{DB}^{GG}	V_D^{RNG}	V_D^{RNG}	V_D^{MST}	V_{DB}^{MST}	V_{DBI}	V_D
P_2	2	1.861816	1.021271	1.861816	1.021271	1.861865	1.021245	3.965332	**0.005126**
P_3	3	1.420194	0.868883	1.420194	0.868883	1.420239	0.868861	5.270424	0.004621
P_4	4	**3.200383**	**0.549078**	**3.200383**	**0.549078**	**3.200383**	**0.549078**	7.859345	0.004621
P_4'	4	1.890507	1.005771	1.890507	1.005771	1.890552	1.005746	2.010857	0.001284
P_5	5	0.852939	0.588470	0.883239	0.575744	0.883254	0.575737	4.146724	0.000488
P_5'	5	2.004905	0.834017	2.076102	0.832159	2.076137	0.832140	**1.445499**	0.001588

Optimal value for each index is denoted in bold.

optimal value changes for a few indices if we consider the maximum number of clusters to be 15 instead of 10. This behavior may be attributed to the highly overlapped structure of the two data sets. We emphasize here that when a data set has an overlapped structure, if c_{max} is far away from the actual number of clusters, most cluster validity indices will behave erratically when the partitions are generated by c-means type algorithms.

Table 8 compares the overall performance of the nine cluster validity indices. Table 8 reveals that all general-

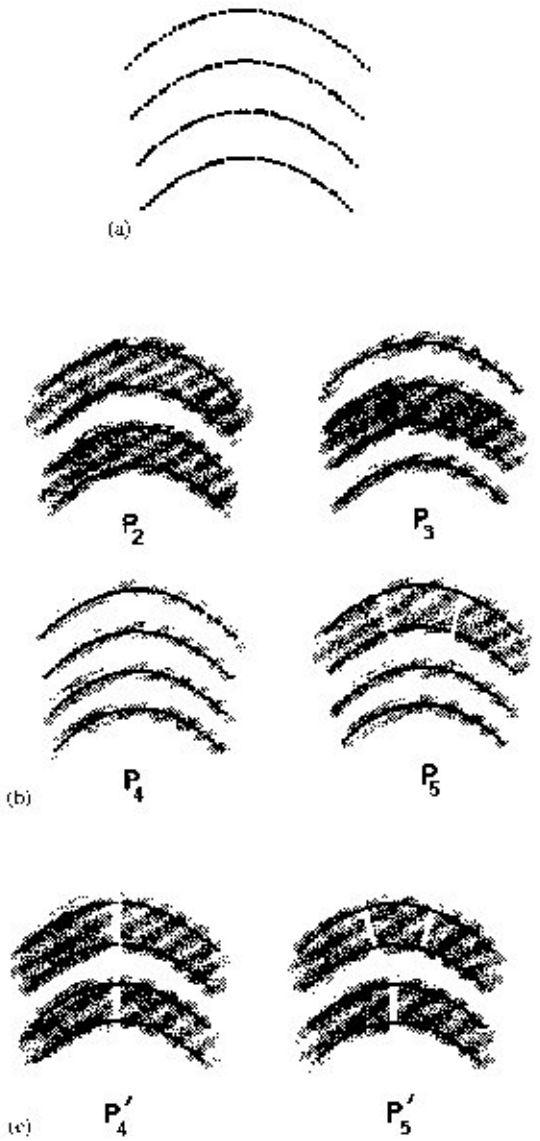

Fig. 5. (a) ARCS. (b) Partitions P_2, P_3, P_4, P_5. (c) Partitions P'_4 and P'_5.

izations of the DI perform much better than the original V_D. V_D^{GG} is found to be the best of the group which finds the most desired class for all data sets. On the other hand V_{MH} fails miserably. It selects the right number of clusters only for the two well-separated data sets. For three of the data sets it is difficult even to locate any knee on the graph of V_{MH}. These cases are indicated by asterisk in Table 8. Table 8 also reflects that the performance of the generalizations of V_{DBI} is comparable to the original index. This raises a question, how much do we achieve through the generalization of the Davies–Bouldin index. To answer this, we use another data set X_8=ARCS containing four circular arcs in two dimensions with 50 points on each arc [Fig. 5(a)]. Thus there are four chain-type clusters which can be easily extracted by a linkage algorithm. Note that any c-means type algorithm which seeks hyperspherical clusters will fail here. Suppose by some clustering algorithm we have partitioned the data set into 2, 3, 4 and 5 classes as shown in Fig. 5(b). We designate these partitions for $c = 2, 3, 4$ and 5 as P_2, P_3, P_4 and P_5, respectively. Of these four partitions, a cluster validity index should be able to pick up the right partition P_4 with c=4.

Table 9 depicts the values of various indices for X_8. It is interesting to see that V_{DBI} and V_D fail to pick up the best partition. To further illustrate this we have also computed the indices for the partitions P'_4 and P'_5 [Fig. 5(c)] with c=4 and c=5 incorrect clusters, respectively. Our generalized indices can easily select the correct partition P_4. The failure of V_{DBI} and V_D do not necessarily mean that they are bad indices but that they are suitable for spherical clusters and are sensitive to noisy points. On the other hand, the proposed indices are equally applicable for both structural (chain type) and spherical clusters, and are less sensitive to noise.

6. CONCLUSIONS

We addressed the problem of cluster validity. With a brief review of some existing cluster validity indices, we generalized two of them: the DI and DBI. These indices, although good for hyperspherical clusters, are very much more sensitive to the presence of a few noisy points. Moreover, they are not expected to work for structural clusters, such as lines, arcs, etc. Our generalizations of DBI and DI used concepts from graph theory. To be more precise, we have used information from GG, RNG and MST defined using inter data point Euclidean distances. Some properties of the indices were also proved. The new indices were tried on eight data sets and were found to outperform the existing indices.

REFERENCES

1. D. L. Davies and D. Bouldin, A cluster separation measure, *IEEE Trans. Pattern Analysis Mach. Intell.* **PAMI-1**, 224–227 (1979).
2. A. Jain and R. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, New Jersey (1988).
3. R. Dubes and A. K. Jain, Validity studies in clustering methodologies, *Pattern Recognition* **11**, 235–254 (1979).
4. N.R. Pal and J.C. Bezdek, On cluster validity for the fuzzy c-means model, *IEEE Trans. Fuzzy Systems* **3**, 370–379 (1995).
5. J. C. Dunn, Indices of partition fuzziness and the detection of clusters in large data sets, *Fuzzy Automata and Decision Processes*, M.M. Gupta, ed. Elsevier, New York (1976).
6. J. C. Bezdek and N. R. Pal, Some new indices of cluster validity, *IEEE Trans. Systems Man Cybernet.* (submitted).
7. C. T. Zahn, Graph theoretical methods for detecting and describing gestalt clusters, *IEEE Trans. Comput.* **C-20**, 68–86 (1971).
8. G. T. Toussaint, The relative neighborhood graph of a finite planner set, *Pattern Recognition* **12**, 261–268 (1980).
9. R. Urquhart, Graph theoretical clustering based on limited neighborhood sets, *Pattern Recognition* **15**, 173–187 (1982).
10. R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. Prentice-Hall, Englewood Cliffs, New Jersey (1982).
11. R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. Wiley Interscience, New York (1973).

About the Author — NIKHIL R. PAL obtained B.Sc. with honors in Physics and Master of Business Management from the University of Calcutta in 1979 and 1982, respectively. He obtained M.Tech. and Ph.D. in Computer Science from the Indian Statistical Institute in 1984 and 1991, respectively. Currently he is an Associate Professor in the Machine Intelligence Unit of Indian Statistical Institute, Calcutta. From September 1991 to February 1993 and July 1994 to December 1994 he was with the Computer Science Department of the University of West Florida. He was also a Guest Faculty of the University of Calcutta. His research interests include image processing, pattern recognition, fuzzy sets theory, measures of uncertainty, neural networks, genetic algorithms, and fuzzy logic controllers. He is an associate editor of the *International Journal of Approximate Reasoning* and *IEEE Transactions on Fuzzy Systems*.

About the Author — JHUMA BISWAS received an M.Sc. degree in Electronic Science from the University of Calcutta. She is currently working for her Ph.D. in the Machine Intelligence Unit of the Indian Statistical Institute, Calcutta. Her research interests include pattern recognition, neural networks and image processing.