Applied Soft Computing 12 (2012) 1555–1565

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Some connectivity based cluster validity indices

Sriparna Saha[a,*], Sanghamitra Bandyopadhyay[b]

[a] *Department of Computer Science and Engineering, Indian Institute of Technology, Patna, India*
[b] *Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India*

ARTICLE INFO

Article history:
Received 4 March 2011
Received in revised form 8 September 2011
Accepted 12 December 2011
Available online 2 February 2012

Keywords:
Clustering
Cluster validity index
Connectivity
Relative neighborhood graph
Single linkage clustering technique
K-means clustering technique

ABSTRACT

Identification of the correct number of clusters and the appropriate partitioning technique are some important considerations in clustering where several cluster validity indices, primarily utilizing the Euclidean distance, have been used in the literature. In this paper a new measure of connectivity is incorporated in the definitions of seven cluster validity indices namely, DB-index, Dunn-index, Generalized Dunn-index, PS-index, *I*-index, XB-index and SV-index, thereby yielding seven new cluster validity indices which are able to automatically detect clusters of any shape, size or convexity as long as they are well-separated. Here connectivity is measured using a novel approach following the concept of relative neighborhood graph. It is empirically established that incorporation of the property of connectivity significantly improves the capabilities of these indices in identifying the appropriate number of clusters. The well-known clustering techniques, single linkage clustering technique and *K*-means clustering technique are used as the underlying partitioning algorithms. Results on eight artificially generated and three real-life data sets show that connectivity based Dunn-index performs the best as compared to all the other six indices. Comparisons are made with the original versions of these seven cluster validity indices.

1. Introduction

Clustering [1–6] is an important technique in data-mining with several applications spanning many fields. It [7,8] is a popular unsupervised pattern classification technique which partitions the input space into *K* regions based on some similarity/dissimilarity metric where the value of *K* may or may not be known a priori. The three fundamental questions that need to be addressed in any typical clustering scenario are: (i) what is the model of a data set or what is a good clustering technique suitable for a given data set, (ii) what is the model order of the data, i.e., how many clusters are actually present in the data, and (iii) how real or good is the clustering itself.

Model selection in clustering consists of two steps. In the first step the proper clustering method for a particular data set has to be identified. Once this choice is made, the model order remains to be determined for the given data set in the second step. The task of determining the number of clusters and also the validity of the clusters formed [9] are generally addressed by providing several definitions of validity indices. The measure of validity of clusters should be such that it will be able to impose an ordering of the clusters in terms of its goodness. In other words, if $U_1, U_2, \ldots, U_m$ be the m partitions of X, and the corresponding values of a validity measure be $V_1, V_2, \ldots V_m$, then $V_{k1} \geq V_{k2} \geq \cdots V_{km}$, $\forall\, ki \in 1, 2, \ldots, m$, $i = 1, 2, \ldots, m$ will indicate that $U_{k1} \uparrow \cdots \uparrow U_{km}$. Here '$U_i \uparrow U_j$' indicates that partition U_i is a better clustering than U_j. Note that a validity measure may also define a decreasing sequence instead of an increasing sequence of $V_{k1}, \ldots, V_{km}$. Several cluster validity indices have been proposed in the literature e.g., Davies–Bouldin (DB) index [10], Dunn's index [11], Xie–Beni (XB) index [12], *I*-index [9], CS-index [13] to name just a few. A good review of the cluster validity indices and their categorization can be found in [14]. Some of these indices have been found to be able to detect the correct partitioning for a given number of clusters, while some can determine the appropriate number of clusters as well. Milligan and Cooper [15] have provided a comparison of several validity indices for data sets containing distinct non-overlapping clusters while using only hierarchical clustering algorithms. Maulik and Bandyopadhyay [9] evaluated the performance of four validity indices, namely, the Davies–Bouldin index [10], Dunn's index [11], Calinski–Harabasz index [9], and a recently developed index *I*, in conjunction with three different algorithms viz. the well-known *K*-means [1], single-linkage algorithm [1] and a SA-based clustering method [9]. However, the effectiveness of these indices in determining the proper clustering algorithm has seldom been studied. Such an attempt has been made in the present paper.

There exists a large number of cluster validation methods which employ the notion of the stability of clustering solutions. The main idea behind such approach to cluster validation requires that solutions are similar for two different data sets that have been

[*] Corresponding author. Tel.: +91 8809559190.
E-mail addresses: sriparna.saha@gmail.com, sriparna@iitp.ac.in (S. Saha), sanghami@isical.ac.in (S. Bandyopadhyay).

1568-4946/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2011.12.013

generated by the same (probabilistic) source. Breckenridge [16] has proposed a measure of stability by estimating the agreement of clustering solutions generated by a clustering algorithm and by a classifier trained using a second (clustered) data set. But that work did not lead to a specific implementation procedure, in particular not for model order selection. But later on many methods (such as [17–20]) for cluster validation came out based on the idea of Breckenridge. In [21], experiments have been performed to investigate the effectiveness of the stability-based index proposed in Ref. [19]. It has been shown experimentally that stability-index is not able to determine the appropriate model order from the data sets for which the clustering algorithm provides stable solutions for several values of K.

All the above mentioned indices use the Euclidean distances in their computation. They are therefore able to characterize only compact clusters. The concept of relative neighborhood graph (RNG) [22] has been successfully applied for solving several pattern recognition problems. One unsupervised clustering technique based on the concepts of RNG is developed in Ref. [23]. Another clustering technique is also developed in [24] which is based on the same concept of connectivity. These two works [23,24] are contemporary and are based on same concept. Some theoretical results are also available in the context of elongated clusters obtained by Normalized Cut method [25]. In this article the concepts of relative neighborhood graph [22] are used to calculate the amount of connectivity among a set of points. Here we conjecture that incorporation of the connectivity measure in the above mentioned validity indices will impart the property of characterizing non-compact, connected clusters to them. Thus, here we incorporate the newly proposed connectivity measure, rather than the Euclidean distance, to develop connectivity based versions of Davies–Bouldin index (DB-index) [10], Dunn's index [11], Generalized Dunn's index [26], PS-index [27], I-index [9], Xie–Beni index (XB index) [12] and SV-index [28]. The single linkage [29] and K-means clustering techniques are used as the underlying partitioning techniques. The effectiveness of the newly proposed connectivity based cluster validity indices in identifying the number of clusters along with the appropriate clustering technique is demonstrated for eight artificially generated and three real-life data sets of varying complexities. For the purpose of comparison, the number of clusters and the appropriate partitioning technique indicated by the original seven cluster validity indices using the Euclidean distance in their computation are also provided for all the artificial and real-life data sets. Experimental results show that incorporation of the concept of connectivity improves the capabilities of these indices to detect any type of clusters irrespective of their shapes and sizes, as long as they are well-separated. Results also reveal that *connect-Dunn* index performs the best compared to all the other six indices.

2. Newly proposed connectivity based cluster validity indices

In this article seven new cluster validity indices based on the concept of connectedness of the clusters are developed. These indices mimic the definitions of seven existing cluster validity indices. Incorporation of the measure of connectivity in the definitions of these cluster validity indices makes them capable of detecting the appropriate partitioning from data sets having clusters of any shape, size or convexity as long as they are well-separated. The concept of relative neighborhood graph (RNG) [22] has been successfully applied for solving several pattern recognition problems. An unsupervised clustering technique based on the concepts of RNG is developed in Ref. [23]. In this article, RNG is used to develop some cluster validity indices those quantify the degree of connectivity of well-separated clusters.

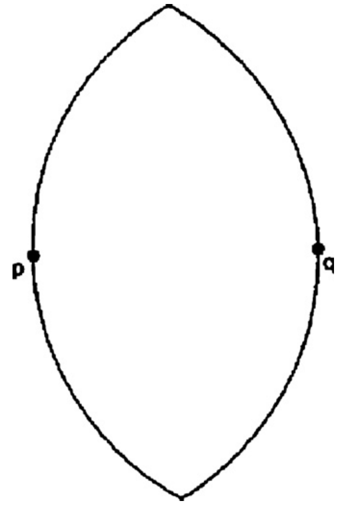

Fig. 1. The lune of two points **p** and **q** is the region between the two arcs, not including the boundary.

2.1. Relative neighborhood graph [22]

Suppose r is an integer and $\mathbf{p}$, $\mathbf{q}$ are two points in r-dimensional Euclidean space. Then the lune of $\mathbf{p}$ and $\mathbf{q}$ (denoted by $lun(\mathbf{p}, \mathbf{q})$ or $lun(\mathbf{pq})$) is the set of points

$$\{z \in R^r : d(\mathbf{p}, \mathbf{z}) < d(\mathbf{p} \cdot \mathbf{q}) \text{ and } d(\mathbf{q}, \mathbf{z}) < d(\mathbf{p}, \mathbf{q})\},$$

where d denotes the Euclidean distance. Alternatively, $lun(\mathbf{p}, \mathbf{q})$ denotes the interior of the region formed by the intersection of two r-dimensional hyperspheres of radius $d(\mathbf{p}, \mathbf{q})$, one of the hyperspheres being centered at $\mathbf{p}$ and the other at $\mathbf{q}$. This is illustrated in Fig. 1 which shows the lune of two points $\mathbf{p}$, $\mathbf{q}$ in the plane. If V is a set of n points in r-space, then define the relative neighborhood graph of V (denoted RNG(V) or simply RNG when V is understood) to be the undirected graph with vertices V such that for each pair $\mathbf{p}, \mathbf{q} \in V$, $\mathbf{pq}$ is an edge of RNG(V) if and only if $lun(\mathbf{p}, \mathbf{q}) \cap V = \emptyset$. Here the edge weight of a particular edge ($\mathbf{pq}$) is kept equal to $d(\mathbf{p}, \mathbf{q})$, the Euclidean distance between the points $\mathbf{p}$ and $\mathbf{q}$.

Fig. 2(a) shows a set V of points in the plane; Fig. 2(b) shows the RNG of this set of points V. The RNG problem is: Given a set V, find RNG(V).

2.2. Measuring the connectivity among a set of points

Here we propose a novel way of measuring the connectivity among a set of points using the above discussed RNG. The distance between a pair of points is measured in the following way.

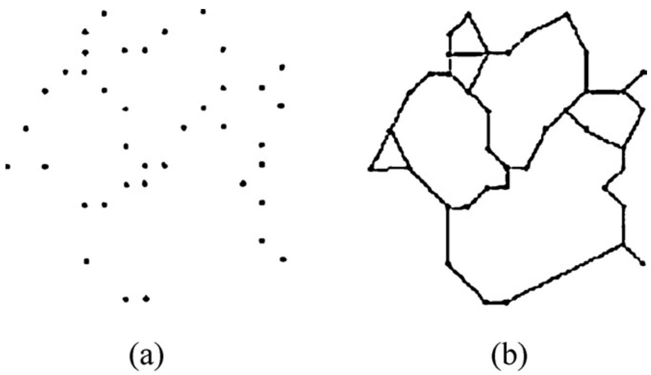

Fig. 2. (a) A set of points in the plane and (b) RNG of the points in (a).

- Construct the relative neighborhood graph of the whole data set.
- The distance between any two points, **x** and **y**, denoted as $d_{short}(\mathbf{x}, \mathbf{y})$, is measured along the relative neighborhood graph. Find all possible paths among these two points along the RNG. Suppose there are total p paths between **x** and **y**, and the number of edges along the ith path is n_i, for $i = 1, \ldots, p$. If the edges along the ith path are denoted as $ed_1^i, \ldots, ed_{n_i}^i$ and the corresponding edge weights are $w(ed_1^i), \ldots, w(ed_{n_i}^i)$, then the shortest distance between **x** and **y** is defined as follows:

$$d_{short}(\mathbf{p}, \mathbf{q}) = \min_{i=1}^{p}\max_{j=1}^{n_i} w(ed_j^i). \quad (1)$$

In order to improve the efficiency of computing d_{short}, we adopt the following pruning strategy. The maximum value of $w(ed_j^i)$ corresponding to the first path is stored in a temporary variable max. If in any of the next path being traced, a weight value greater than max is obtained, that path is pruned. However, if a smaller value of the maximum weight is found in any of the subsequent paths, then max is updated to this smaller value and the process repeats.

2.3. Definitions of connectivity based cluster validity indices

In this section we will describe in detail the newly proposed connectivity based cluster validity indices. Consider a partition of the data set $X = \{\mathbf{x}_j : j = 1, 2, \ldots n\}$ into K clusters. Then the medoid of the kth cluster, denoted by $\mathbf{z}_k$, is the point of that cluster which has the minimum average distance to all the other points in that cluster. Suppose the point which has the minimum average distance to all the points in the kth cluster is denoted by $\overline{x}_{\text{minindex}}^k$. Then,

$$\text{minindex} = argmin_{i=1}^{n_k} \frac{\sum_{j=1}^{n_k} d_e(\mathbf{x}_i^k, \mathbf{x}_j^k)}{n_k},$$

where n_k is the total number of points in the kth cluster, d_e stands for Euclidean distance that means $d_e(\mathbf{x}_i^k, \mathbf{x}_j^k)$ denotes the Euclidean distance between two points $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$, $\mathbf{x}_i^k$ denotes the ith point of the kth cluster. Then

$$\mathbf{z}_k = \mathbf{x}_{\text{minindex}}^k.$$

The relative neighborhood graph (RNG) of the whole data set is constructed beforehand. A distance matrix named $distance_{short}$ is kept of the size of $n \times n$ where n is the size of the data set. $distance_{short}$ is defined as follows:

$$distance_{short} = [[d_{short}(\mathbf{x}_i, \mathbf{x}_j)]_{i=1}^n]_{j=1}^n.$$

Here $d_{short}(\mathbf{x}_i, \mathbf{x}_j)$ is measured along the RNG using the procedure mentioned in Section 2.2.

2.3.1. Connectivity based Davies–Bouldin index (connect-DB index)

This index is developed along the lines of the popular Davies–Bouldin (DB) index [10]. This is a function of the ratio of the sum of *within-cluster connectivity* to *between cluster separation*. The scatter within the ith cluster, S_i, is computed as $S_i = (\sum_{\mathbf{x}\in C_i} d_{short}(\mathbf{x}, \mathbf{z}_i))/n_i$, where $\mathbf{z}_i$ represents the medoid of cluster i, n_i denotes the number of points present in cluster i, and $d_{short}(\mathbf{x}, \mathbf{z}_i)$ is obtained from $distance_{short}$ matrix. The distance between cluster C_i and C_j, denoted by d_{ij}, is defined as $d_{ij} = d_{short}(\mathbf{z}_i, \mathbf{z}_j)$. Then connectivity based DB index, *connect-DB* index, is defined as

$$connect\text{-}DB(K) = \frac{\sum_{i=1}^{K} R_i}{K}. \quad (2)$$

Here $R_i = \max_{j,j\neq i}\{(S_i + S_j)/d_{ij}\}$. The objective is to minimize the *connect-DB* index for achieving the proper clustering.

2.3.2. Connectivity based Dunn's index (connect-Dunn index)

This index is developed along the lines of the popular Dunn's index [11]. Let S and T be two nonempty subsets of R^N. Then the diameter Δ of S is defined as $\Delta(S) = \max_{x,y\in S}\{d_{short}(\mathbf{x}, \mathbf{y})\}$, where **x** and **y** are two points belonging to set S and $d_{short}(\mathbf{x}, \mathbf{y})$ is computed using Eq. (1). The set distance δ between S and T is defined as $\delta(S, T) = \min_{\mathbf{x}\in S,\mathbf{y}\in T}\{d_{short}(\mathbf{x}, \mathbf{y})\}$. Here, $d_{short}(\mathbf{x}, \mathbf{y})$ is obtained from $distance_{short}$ matrix. For any partition, *connect-Dunn* index is defined as follows

$$connect\text{-}Dunn(K) = \min_{1\leq i\leq K} \min_{1\leq j\leq K, j\neq i} \left\{ \frac{\delta(C_i, C_j)}{\max_{1\leq k\leq K} \Delta(C_k)} \right\}. \quad (3)$$

Larger value of *connect-Dunn* index corresponds to good clustering, and the number of clusters that maximizes this index is taken as the optimal number of clusters.

2.3.3. Connectivity based Generalized Dunn's index (connect-GDunn index)

. This index is developed along the lines of the Generalized Dunn's index [26]. The generalized Dunn's index was developed after demonstrating the sensitivity of the original Dunn's index [11], to changes in cluster structure, since not all of the data points were involved in the computation of the index. The symmetry based GDunn cluster validity index, *connect-GDunn* index, is defined as

$$connect\text{-}GDunn(K) = \min_{1\leq s\leq K} \left\{ \min_{1\leq t\leq K, t\neq s} \left\{ \frac{\delta(C_s, C_t)}{\max_{1\leq k\leq K} \Delta(C_k)} \right\} \right\}. \quad (4)$$

The two measures δ and Δ are defined as follows: $\Delta(S) = 2 \times (\sum_{x\in S} d_{short}(\mathbf{x}, \mathbf{z}_S))/|S|$ and $\delta(S, T) = (1/(|S||T|)) \sum_{\mathbf{x}\in S,\mathbf{y}\in T} d_{short}(\mathbf{x}, \mathbf{y})$. Here $\mathbf{z}_S$ and $\mathbf{z}_T$ are the medoids of the sets S and T, respectively. $d_{short}(\mathbf{x}, \mathbf{z}_S)$ is computed by Eq. (1). Larger values of *connect-GDunn* correspond to good clusters, and the number of clusters that maximizes *connect-GDunn* is taken as the optimal number of clusters.

2.3.4. Connectivity based PS-index (connect-PS index)

This index is developed along the lines of PS-index [27]. The cluster validity index, *connect-PS* index, is defined as

$$connect\text{-}PS(K) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i}\sum_{\mathbf{x}\in S_i}\frac{d_{short}(\mathbf{x}, \mathbf{z}_i)}{d_{min}} \quad (5)$$

where $d_{min} = \min_{m,n=1,\ldots,K,\, m\neq n} d_{short}(\mathbf{z}_m, \mathbf{z}_n)$, where $\mathbf{z}_m$ and $\mathbf{z}_n$ are the medoids of the two clusters m and n, respectively. S_i is the set whose elements are the data points assigned to the ith cluster, n_i is the number of elements in S_i, or, $n_i = |S_i|$, d_{min} is the minimum shortest distance between any two cluster medoids and $d_{short}(\mathbf{x}, \mathbf{z}_i)$ is computed by Eq. (1). The smallest *connect-PS*(K^*) indicates a valid optimal partition with the optimal number of clusters K^*.

2.3.5. Connectivity based I-index (connect-I index)

This index is developed along the lines of the *I*-index [9]. The new cluster validity function *connect-I* is defined as:

$$connect\text{-}I(K) = \left(\frac{1}{K} \times \frac{1}{\mathcal{E}_K} \times D_K\right). \quad (6)$$

Here, $\mathcal{E}_K = \sum_{i=1}^{K} E_i$, such that $E_i = \sum_{j=1}^{n_i} d_{short}(\mathbf{x}_j^i, \mathbf{z}_i)$ and $D_K = max_{i,j=1}^{K} d_{short}(\mathbf{z}_i, \mathbf{z}_j)$. D_K is the maximum shortest distance between two cluster medoids among all the pairs of medoids. $d_{short}(\mathbf{x}_j^i, \mathbf{z}_i)$ is computed by Eq. (1) where $\mathbf{z}_i$ is the medoid of cluster i and $\mathbf{x}_j^i$

denotes the *j*th point of the *i*th cluster. The objective is to maximize this index in order to obtain the actual number of clusters.

2.3.6. Connectivity based Xie–Beni index (connect-XB index)

This index is developed along the lines of the popular XB-index [12]. It is defined as follows:

$$connect\text{-}XB(K) = \frac{\sum_{i=1}^{K}\left(\sum_{\mathbf{x}\in C_i} d_{short}(\mathbf{x},\mathbf{z}_i)\right)}{n(\min_{i,k=1,\ldots,K,i\neq k} d_{short}(\mathbf{z}_i,\mathbf{z}_k))}. \tag{7}$$

$d_{short}(\mathbf{x},\mathbf{z}_i)$ is computed by Eq. (1). Here C_i denotes the cluster *i*, $\mathbf{z}_i$ is the medoid of cluster *i*, *n* is the size of the whole data set. The most desirable partition (or an optimal value of *K*) is obtained by minimizing *connect-XB* index over $K = 2, 3, \ldots, K_{\max}$.

2.3.7. Connectivity based SV index (connect-SV index)

Kim attempted to determine the optimal number of clusters by measuring the status of the given partition with both an under-partition index and an over-partition index [28]. Here, the newly developed *connect-SV* index is defined along the lines of SV index proposed by Kim [28].

$$connect\text{-}SV(K) = v_{under}(Z:X) + v_{over}(Z) = \frac{1}{K}\sum_{i=1}^{K}\sum_{\mathbf{x}\in C_i}\frac{d_{short}(\mathbf{x},\mathbf{z}_i)}{n_i} + \frac{K}{d_{min}}. \tag{8}$$

Here $d_{min} = \min_{i\neq j} d_{short}(\mathbf{z}_i,\mathbf{z}_j)$. Here $\mathbf{z}_i$ and $\mathbf{z}_j$ are the medoids of two clusters *i* and *j*, respectively and n_i denotes the number of points in cluster *i*. A minimum value of *connect-SV* index indicates the optimal number of clusters.

3. Data sets used

Eight artificial data sets and three real-life data sets are used for the experiments. A description of the data sets in terms of the number of points present, dimension of the data set and the number of clusters is presented in Table 1. These data sets are divided into two groups.

Table 1
Description of the data sets. Here AC denotes the actual number of clusters present in it.

Name	No. of points	Dimension	AC
Pat1	557	2	3
Pat2	417	2	2
Spiral	1000	2	2
Rect_3_2	400	2	3
Mixed_5_2	850	2	5
Sph_4_3	400	3	4
Sph_5_2	250	2	5
Sph_10_2	500	2	10
Iris	150	4	3
Cancer	683	9	2
Newthyroid	215	5	3

(1) Group 1: This group of data sets contains well-separated clusters of different shapes, sizes and convexities.
   (a) *Pat1*: This data, used in Ref. [30], consists of 880 patterns. There are three non convex clusters present in this data set. This is shown in Fig. 3(a).
   (b) *Pat2*: This data set, used in [31], consists of 2 non-linear, non-overlapping and non-symmetric clusters. The data set is shown in Fig. 3(b).
   (c) *Spiral*: This data set, used in Ref. [32], consists of 1000 data points distributed over 2 spiral clusters. This is shown in Fig. 4(a).
   (d) *Rect_3_2*: This data set, used in Ref. [33], is a combination of ring-shaped, compact and linear clusters shown in Fig. 4(b).
   (e) *Mixed_5_2*: This data set, used in Ref. [34], contains 850 data points distributed on five clusters, as shown in Fig. 5(a).
   (f) *Sph_4_3*: This data set, used in [35], consists of 400 points distributed over 4 hyperspherical shaped clusters. This data set is shown in Fig. 5(b). The clusters present in this data set are well-separated, each consisting of 100 data points.
   (g) *Sph_5_2*: This data set, used in [35], consists of 250 two dimensional data points distributed over 5 spherically shaped clusters. The clusters present in this data set are highly overlapping, each consisting of 50 data points. This data set is shown in Fig. 6(a).
(4) *Sph_10_2*: This data set, used in Ref. [36], consists of 500 two dimensional data points distributed over 10 different clusters. Some clusters are overlapping in nature. Each

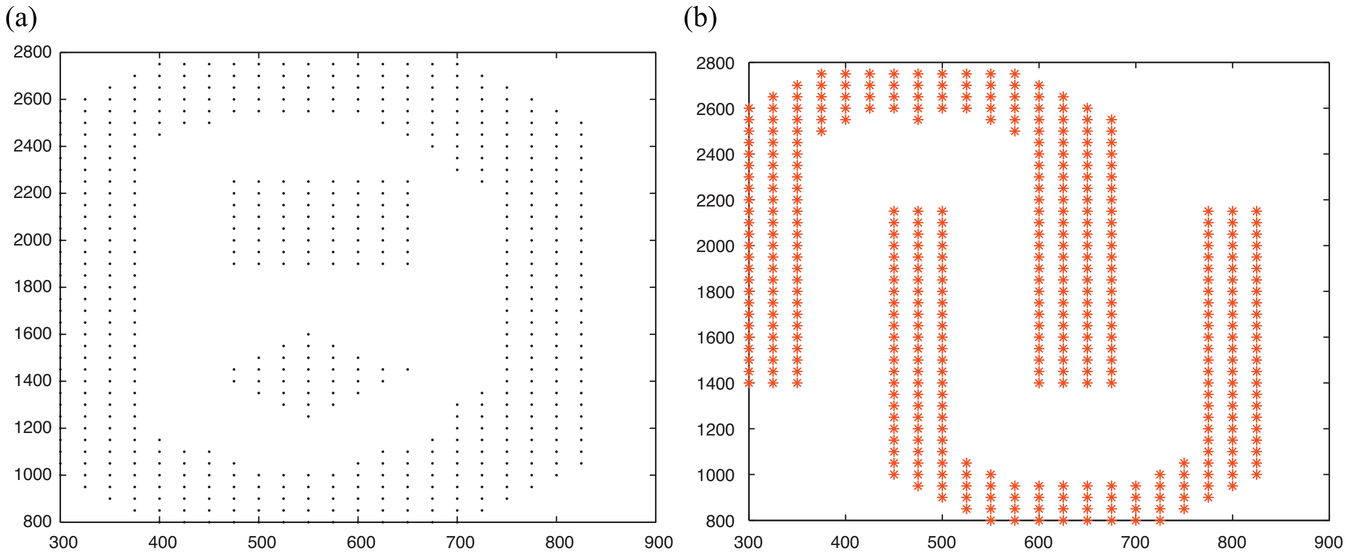

Fig. 3. (a) *Pat1* and (b) *Pat2*.

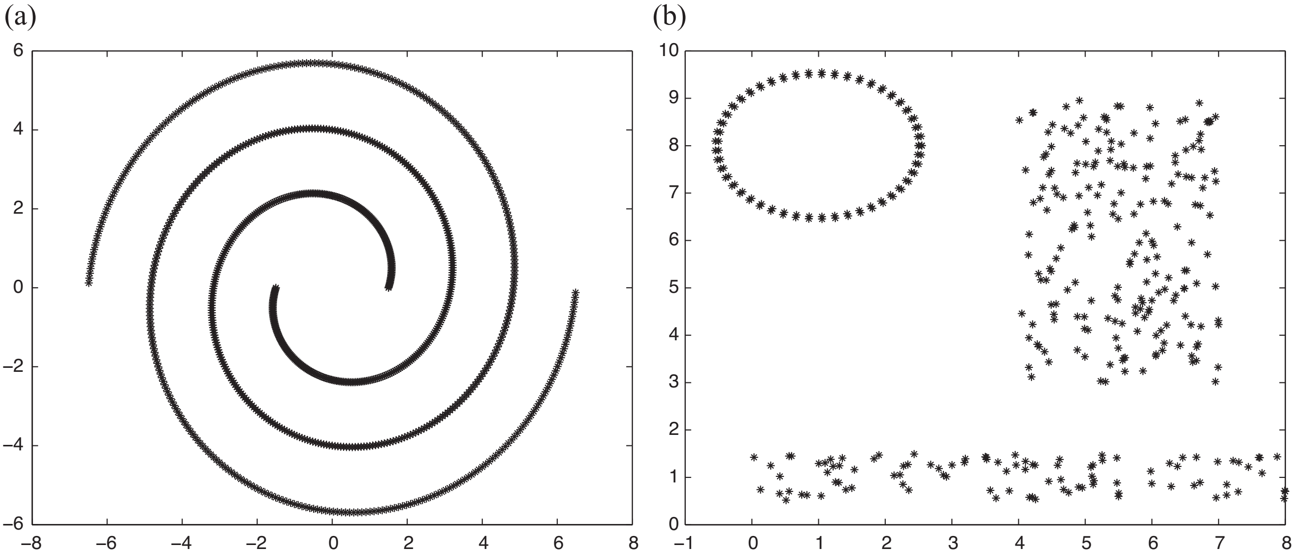

Fig. 4. (a) *Spiral* and (b) *Rect_3_2*.

cluster consists of 50 data points. This data set is shown in Fig. 6(b).

(2) Group 2: This group consists of three real-life data sets obtained from Ref. [37].

(a) *Iris*: This data set consists of 150 data points distributed over 3 clusters. Each cluster consists of 50 points. This data set represents different categories of irises characterized by four feature values [38]. It has three classes Setosa, Versicolor and Virginica. It is known that two classes (Versicolor and Virginica) have a large amount of overlap while the class Setosa is linearly separable from the other two.

(b) *Cancer*: Here we use the Wisconsin Breast *Cancer* data set, consists of 683 sample points. Each pattern has nine features corresponding to clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. There are two categories in the data: malignant and benign. The two classes are known to be linearly separable.

(c) *Newthyroid*: The original database from where it has been collected is titled as thyroid gland data ('normal', 'hypo' and 'hyper' functioning). Five laboratory tests are used to predict whether a patient's thyroid belongs to the class euthyroidism, hypothyroidism or hyperthyroidism. There are a total of 215 instances and the number of attributes is five.

4. Experimental results

The effectiveness of the connectivity based cluster validity indices in determining the appropriate number of clusters and the appropriate partitioning is established for the above-mentioned eleven data sets. Single linkage clustering technique [1] and K-means clustering technique [1] are used as the underlying partitioning techniques. The number of clusters, K, is varied from 2 to $\sqrt{n}$, where n is the number of data points. The method of choosing the optimum K and A_i (optimum partitioning technique) by a particular cluster validity index (V) are provided below.

Let S denote the set of clustering algorithms (models) and $CV(A, l)$ denotes the value of some cluster validity index CV for $K=l$ provided by a clustering algorithm $A \in S$. Then the

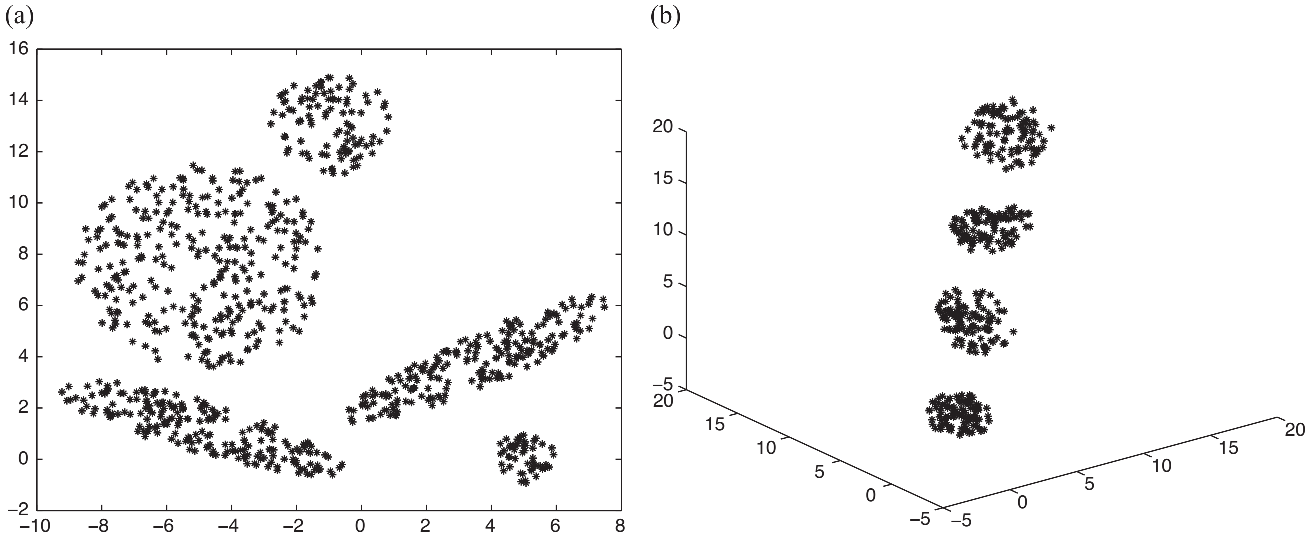

Fig. 5. (a) *Mixed_5_2* and (b) *Sph_4_3*.

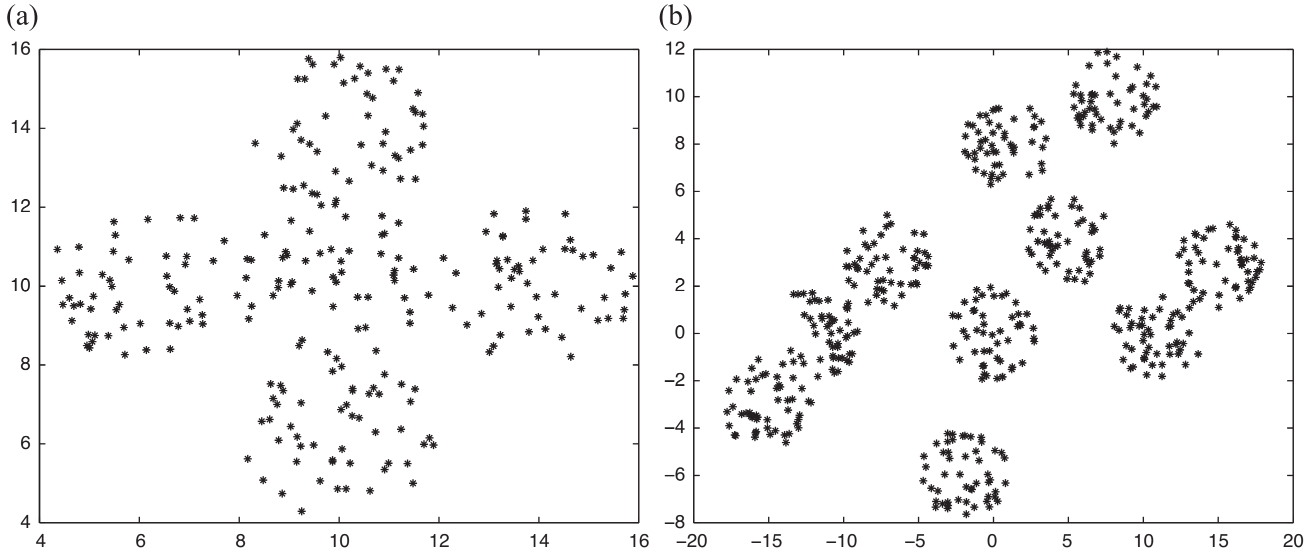

Fig. 6. (a) *Sph_5_2* and (b) *Sph_10_2*.

most appropriate algorithm (model) and the corresponding $K = K^*$ (model order), denoted by the tuple (A*, K*), is given by (A*, K*) = $\text{argopt}_{\forall A \in S \text{ and } l=2,3,\ldots,\sqrt{n}}\{CV(A, l)\}$. Table 2 shows the overall (A*, K*) values obtained using the different indices for all the data sets. The results reported in the table are the average values obtained over ten runs of the algorithm. It has also been noted that there is a significant difference between the best value and the second best value obtained by any of the proposed validity indices for all the data sets used here for experiment.

Figs. 7(a), 8(a), 9(a), 10(a), 11(a), 12, 13(a) and 14(a) show, respectively, the partitionings obtained after application of single linkage clustering technique on the eight artificial data sets, respectively, for actual number of clusters present in the data sets. Similarly, Figs. 7(b), 8(b), 9(b), 10(b), 11(b), 12, 13(b) and 14(b) show, respectively, the partitionings obtained after application of *K*-means clustering technique on the eight artificial data sets for actual number of clusters present in the data sets. These figures show that for first five data sets, *Pat1*, *Pat2*, *Spiral*, *Rect_3_2* and *Mixed_5_2*, single linkage clustering is the best suitable partitioning technique. For *Sph_4_3* both single linkage and *K*-means perform good. But for *Sph_5_2* and *Sph_10_2* *K*-means is the best partitioning technique. Table 2 shows the optimum number of clusters and the optimum partitioning technique identified by the seven newly proposed connectivity based cluster validity indices, namely, *connect-DB*, *connect-Dunn*, *connect-GDunn*, *connect-PS*, *connect-I*, *connect-XB* and *connect-SV* indices for all the data sets used here for experiment.

For *Pat1* data set, *connect-Dunn* and *connect-XB* indices are able to find the proper partitioning, the proper number of partitions and the proper partitioning technique (the corresponding partitioning is shown in Fig. 7(a)). Optimum values of *connect-GDunn*, *connect-I* and *connect-SV* indices indicate $K = 2$ as the proper number of clusters with Single linkage clustering technique whereas that of *connect-PS* indicates $K = 10$ as the proper number of clusters along with Single linkage clustering technique (refer to Table 2). *connect-DB* index wrongly indicates $K = 2$ as the proper number of clusters along with *K*-means clustering technique (refer to Table 2).

For *Pat2* data set, all the connectivity based indices except *connect-SV* are able to detect the proper number of clusters, the appropriate partitioning and the proper partitioning technique

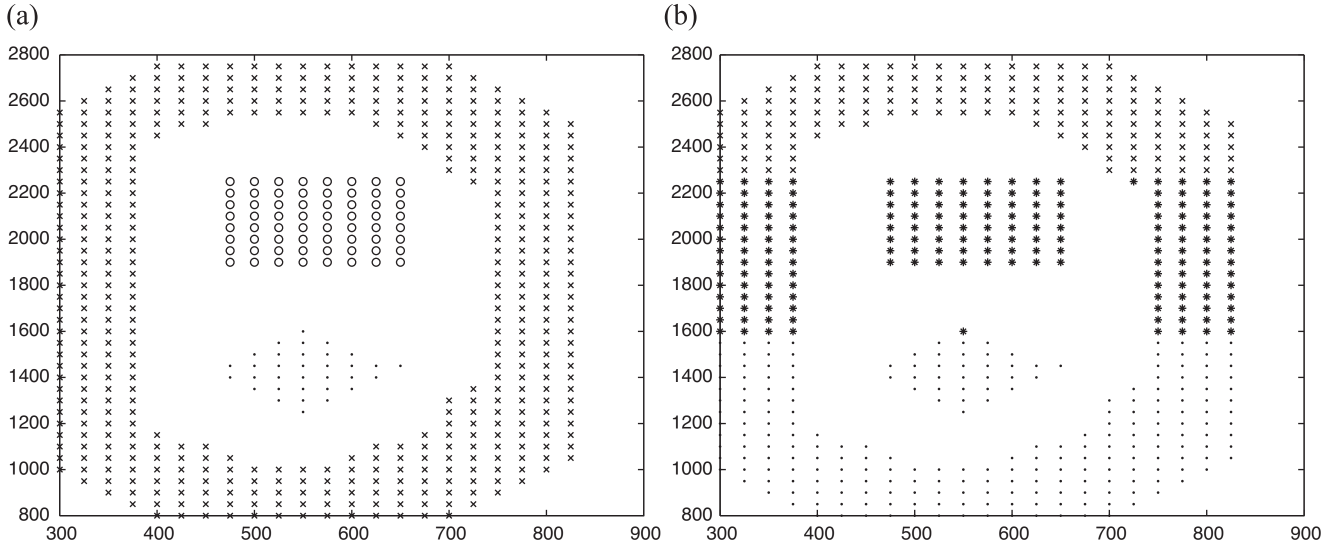

Fig. 7. Optimal partitioning on *Pat1* for $K = 3$ obtained by (a) single linkage clustering technique and (b) *K*-means clustering technique.

Table 2
Optimal number of clusters and the optimal partitioning technique identified by the newly proposed connectivity based version and the original version of seven cluster validity indices for eleven data sets, segmented using single linkage and *K*-means clustering algorithms where *K* is varied from 2 to $\sqrt{n}$. Here AC denotes the actual number of clusters present in the particular data set. The name within bracket indicates the optimal partitioning technique identified by a particular validity index. Here 'SL' stands for single linkage clustering technique and KM stands for *K*-means clustering technique. Success rates (defined in Section 4.2) of two different versions of seven cluster validity indices in detecting the proper partitioning and the proper number of partitions are also provided.

Data set	AC	DB		Dunn		GDunn		PS		*I*		XB		SV	
		Conn.	Org.	Conn.	Org.	Conn.	Org.	Conn.	Org.	Conn.	Org.	Conn.	Org.	Conn.	Org.
Pat1	3 (SL)	2 (KM)	2 (KM)	3 (SL)	6 (KM)	2 (SL)	6 (KM)	10 (SL)	2 (KM)	2 (SL)	3 (KM)	3 (SL)	2 (KM)	2 (KM)	10 (SL)
Pat2	2 (SL)	2 (SL)	3 (KM)	2 (SL)	8 (KM)	2 (SL)	8 (KM)	2 (SL)	2 (KM)	2 (SL)	3 (KM)	2 (SL)	3 (KM)	2 (KM)	10 (SL)
Spiral	2 (SL)	2 (SL)	5 (KM)	2 (SL)	2 (SL)	2 (SL)	10 (KM)	2 (SL)	4 (KM)	2 (SL)	4 (KM)	2 (SL)	6 (KM)	2 (SL)	3 (KM)
Rect_3_2	3 (SL)	3 (SL)	10 (SL)	3 (SL)	3 (SL)	3 (SL)	7 (SL)	3 (SL)	4 (SL)	3 (SL)	4 (SL)	3 (SL)	3 (KM)	2 (SL)	3 (KM)
Mixed_5_2	5 (SL)	2 (KM)	2 (SL)	5 (SL)	4 (SL)	3 (KM)	4 (SL)	3 (KM)	8 (SL)	2 (KM)	5 (SL)	3 (KM)	3 (KM)	2 (KM)	2 (SL)
Sph_4_3	4 (SL/KM)	3 (SL)	4 (SL/KM)	4 (SL/KM)	3 (KM)	4 (SL/KM)	3 (KM)	4 (SL/KM)	2 (SL)	4 (SL/KM)	4 (SL/KM)	4 (SL/KM)	4 (SL/KM)	2 (SL/KM)	4 (SL/KM)
Sph_5_2	5 (KM)	2 (KM)	5 (KM)	3 (SL)	7 (KM)	3 (SL)	4 (SL)	8 (SL)	2 (SL)	2 (SL)	5 (KM)	3 (SL)	4 (KM)	2 (SL)	2 (SL)
Sph_10_2	10 (KM)	6 (SL)	7 (SL)	6 (SL)	4 (SL)	6 (SL)	5 (KM)	6 (SL)	4 (SL)	2 (SL)	11 (KM)	6 (SL)	4 (KM)	2 (KM)	4 (SL)
Iris	3 (KM)	2 (SL)	2 (SL)	2 (SL)	10 (SL)	2 (SL)	2 (SL)	2 (SL)	3 (SL)	2 (SL)	3 (KM)	2 (SL)	2 (SL)	2 (SL)	2 (SL)
Cancer	2 (KM)	2 (KM)	2 (SL)	3 (SL)	10 (SL)	2 (SL)	8 (KM)	10 (SL)	10 (SL)	2 (SL)	2 (KM)	2 (SL)	2 (KM)	2 (KM)	5 (SL)
Newthyroid	3 (KM)	3 (KM)	3 (SL)	2 (SL)	4 (SL)	3 (SL)	2 (SL)	2 (SL)	9 (SL)	2 (SL)	2 (SL)	2 (SL)	2 (SL)	2 (KM)	4 (SL)
Success rate		0.45 (5/11)	0.18 (2/11)	0.54 (6/11)	0.18 (2/11)	0.36 (4/11)	0.0 (0/11)	0.36 (4/11)	0.0 (0/11)	0.36 (4/11)	0.45 (5/11)	0.45 (5/11)	0.18 (2/11)	0.18 (2/11)	0.09 (1/11)

(refer to Table 2). The corresponding partitioning is shown in Fig. 8(a). *connect-SV* wrongly takes its optimal value along with *K*-means clustering technique (refer to Table 2). Again for *Spiral* data set, all the connectivity based indices are able to detect the appropriate partitioning, appropriate partitioning technique (single linkage in this case) and the appropriate number of clusters (the corresponding partitioning is shown in Fig. 9(a)). For *Rect_3_2* data set, except for *connect-SV* all the connectivity based indices are able to detect the appropriate partitioning, appropriate partitioning technique and the appropriate number of clusters (refer to Table 2). The corresponding partitioning is shown in Fig. 10(a). Optimum value of *connect-SV* indicates $K=2$ number of clusters along with Single linkage clustering technique. But for *Mixed_5_2* data set, only *connect-Dunn* index is able to detect the appropriate partitioning, the appropriate number of clusters and the appropriate partitioning technique (refer to Table 2). The corresponding partitioning is shown in Fig. 11(a). For *Sph_4_3* data set, all the indices except *connect-DB* and *connect-SV* are able to detect the proper number of clusters, proper partitioning technique and the appropriate partitioning from this data set (refer to Table 2). The corresponding partitioning is shown in Fig. 12. *connect-DB* indicates $K=3$ as the optimum number of clusters along with single linkage clustering technique where as *connect-SV* indicates $K=2$ as the correct number of clusters along with both single linkage clustering technique *K*-means clustering technique. For *Sph_5_2* data set, none of the connectivity based indices are able to detect the appropriate partitioning, appropriate number of clusters and the proper partitioning technique (refer to Table 2). Again for *Sph_10_2* data sets none of the connectivity based cluster validity indices are able to detect the appropriate number of partitions and appropriate partitioning technique (refer to Table 2). Results on *Sph_5_2* and *Sph_10_2* data sets show that these connectivity based indices fail for overlapping clusters. These are not capable of detecting the appropriate number of clusters from data sets having overlapping clusters.

For the three real-life data sets, *Iris*, *Cancer* and *Newthyroid*, no visualization is possible as these are high-dimensional data sets. For these three data sets, the *Minkowski Score* [33] is calculated after applications of single linkage and *K*-means clustering algorithms. This is a measure of the quality of a solution given the true clustering. Let *T* be the "true" solution and *S* the solution we wish to measure. Denote by n_{11} the number of pairs of elements that are in the same cluster in both *S* and *T*. Denote by n_{01} the number of pairs that are in the same cluster only in S, and by n_{10} the number of pairs that are in the same cluster in T. *Minkowski Score* (MS) is then defined as: $\mathrm{MS}(T,S)=\sqrt{(n_{01}+n_{10})/(n_{11}+n_{10})}$. For MS, the optimum score is 0, with lower scores being "better". For *Iris* data set, MS values corresponding to the partitionings obtained by single linkage clustering technique and *K*-means clustering techniques for $K=3$ are 0.82 and 0.62, respectively. Thus *K*-means is the best clustering technique for this data set. As can be seen from Table 2, all the connectivity based indices are able to detect $K=2$ as the proper number of partitions for this data set along with Single linkage clustering technique. In *Iris* data set there are three clusters among which two are highly overlapping to each other. Thus many other methods of *Iris* also provides $K=2$ as the optimal number of clusters. The corresponding MS value is 0.82. For *Cancer* data set, the MS values corresponding to the partitionings obtained by Single linkage clustering technique and *K*-means clustering technique are 0.91 and 0.37, respectively. Thus *K*-means is again most suitable partitioning technique for this data set. *connect-GDunn*, *connect-I* and *connect-XB* indices are able to detect the appropriate number of clusters along with Single linkage clustering technique (refer to Table 2). The corresponding MS value is 0.91. *connect-Dunn* and *connect-PS* indices indicate $K=3$ and $K=10$ as the optimal number of clusters from this data set, respectively. The corresponding MS values are 0.91 and 0.91, respectively.

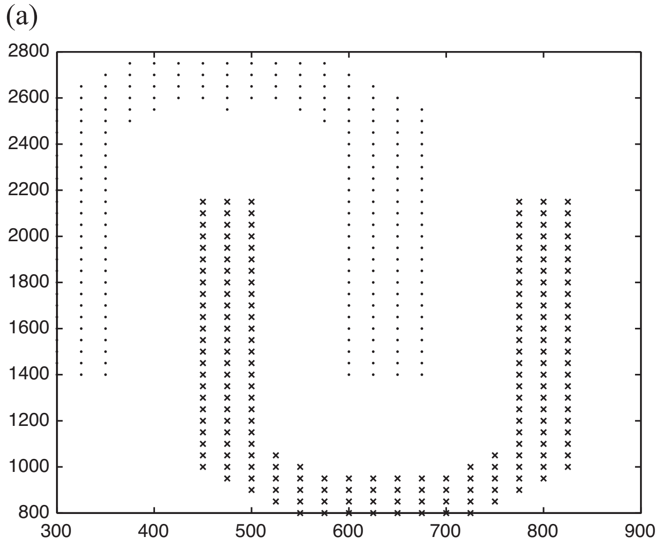

Fig. 8. Optimal partitioning on *Pat2* for $K = 2$ obtained by (a) single linkage clustering technique and (b) K-means clustering technique.

connect-DB and *connect-SV* indices indicate $K = 2$ number of clusters along with K-means clustering technique. The corresponding MS value is 0.37. For *Newthyroid* data, the MS values obtained by Single linkage and K-means clustering techniques are 0.93 and 0.74, respectively. Thus K-means is again the most suitable partitioning technique to segment this data set. Only *connect-DB* index is able to detect the proper number of clusters and the proper partitioning technique. The corresponding MS value is 0.74. *connect-Dunn* index is able to detect the proper number of clusters along with Single linkage clustering technique. The corresponding MS value is 0.91. Other connectivity based indices, *connect-DB*, *connect-Dunn*, *connect-PS*, *connect-I* and *connect-XB* indices determine $K = 2$ as the optimal number of clusters along with Single linkage clustering technique. The corresponding MS value is 0.92.

The above mentioned results show that *connect-DB* is able to detect the appropriate partitioning from five out of eleven data sets used here for the experiment. Similarly, *connect-Dunn*, *connect-GDunn*, *connect-PS*, *connect-I*, *connect-XB*, and *connect-SV* indices are able to detect the proper partitioning from six, four, four, four, five and two out of eleven data sets, respectively. Thus, it can be easily concluded that the proposed *connect-Dunn* index performs the best compared to other six indices for detecting the proper number of clusters and the proper partitioning from data sets having well-separated clusters of any shape, size or convexity. But these indices fail for data sets with overlapping clusters. This is evident from the results on *Sph_5_2*, *Sph_10_2* and three real-life data sets.

4.1. Computational time of the proposed cluster validity indices

In this section we have performed the complexity analysis of the proposed connectivity based indices. The main time consuming component of all the above mentioned cluster validity indices is the computation of d_{short}. In order to compute this at first the relative neighbor graph has to be constructed. If there are n number of points in the data set then construction of RNG takes a total of $O(n)$ time [39]. Thereafter the paths from the medoid of a cluster to all the points are found. This takes total $O(n^3)$ time where n is the total number of points in the data set. Thus the total complexity of finding d_{short} from all the medoids to all the

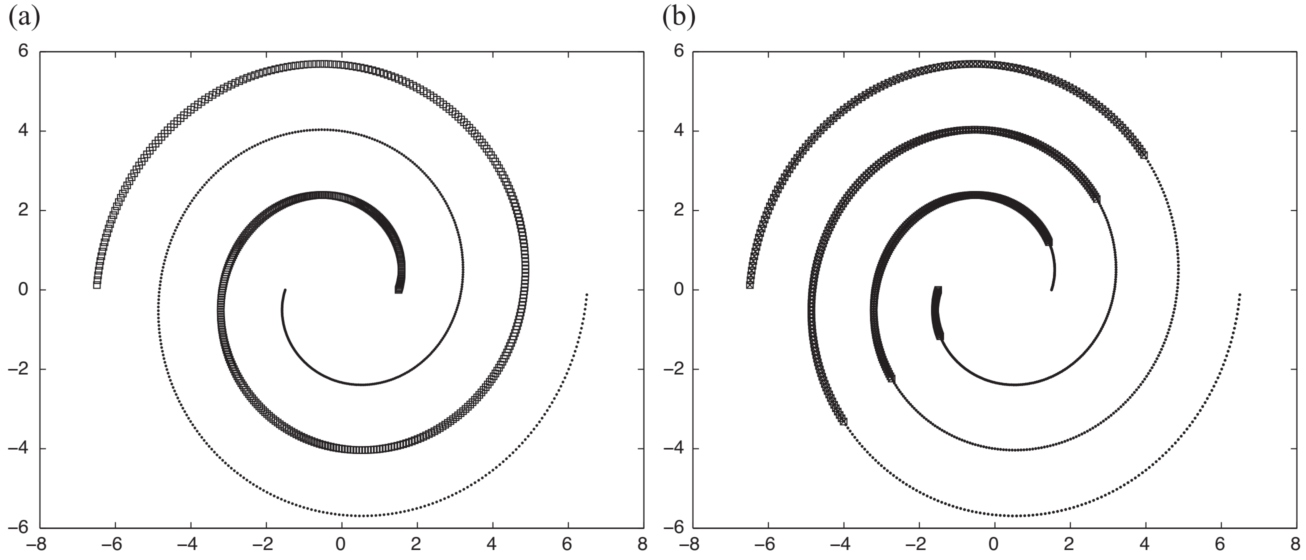

Fig. 9. Optimal partitioning on *Spiral* for $K = 2$ obtained by (a) single linkage clustering technique and (b) K-means clustering technique.

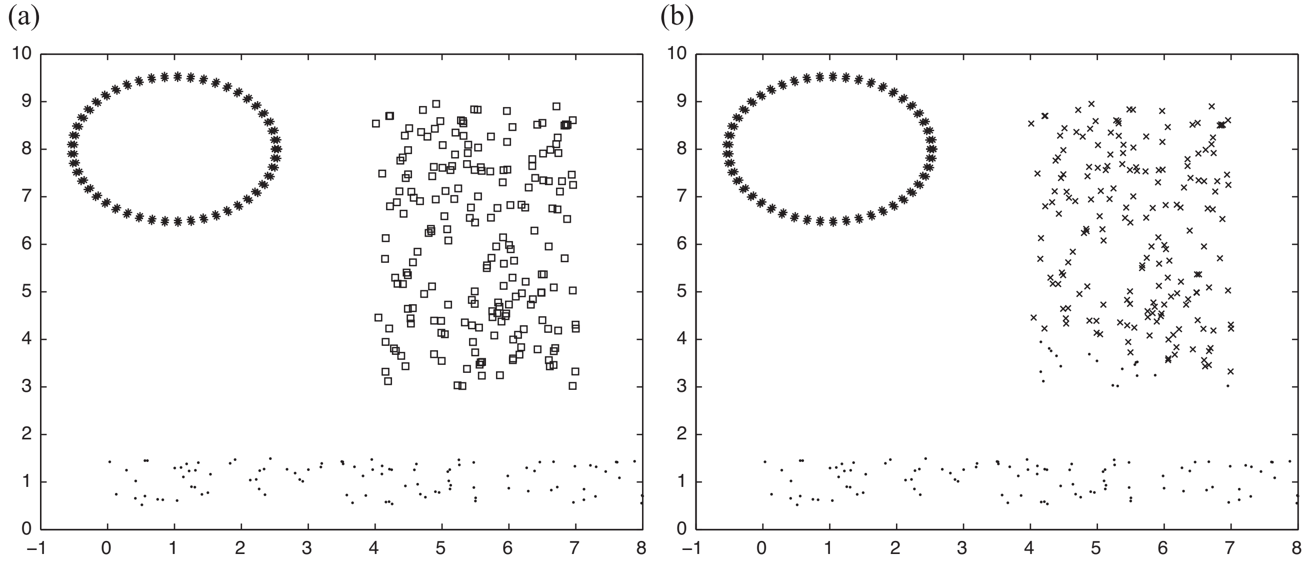

Fig. 10. Optimal partitioning on *Rect_3_2* for $K = 3$ obtained by (a) single linkage clustering technique and (b) *K*-means clustering technique.

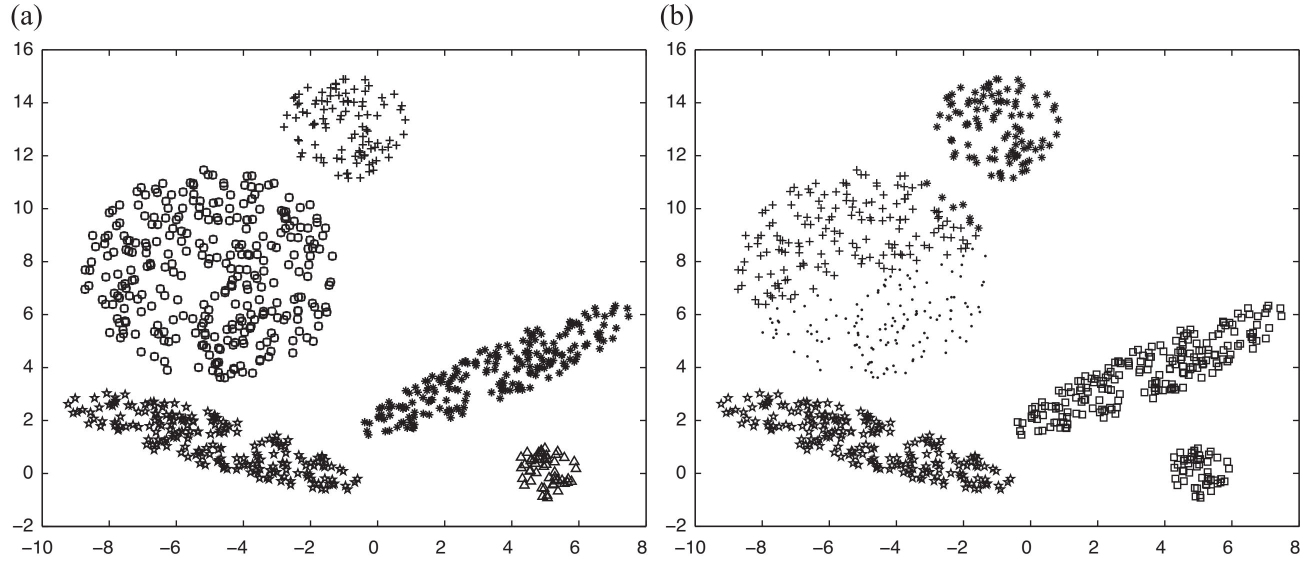

Fig. 11. Optimal partitioning on *Mixed_5_2* for $K = 5$ obtained by (a) single linkage clustering technique and (b) *K*-means clustering technique.

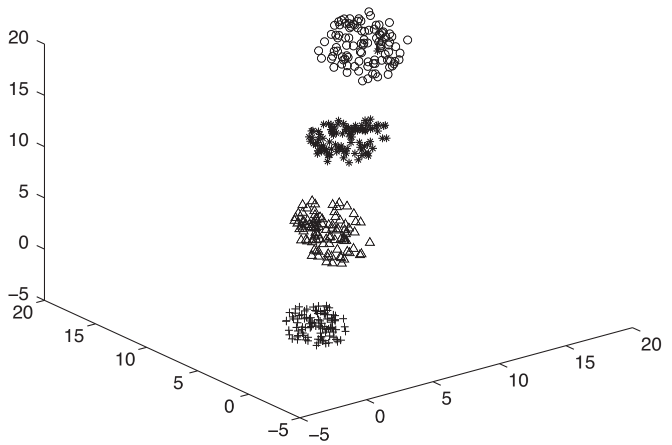

Fig. 12. Optimal partitioning on *Sph_4_3* for $K = 4$ obtained by (a) single linkage clustering technique and (b) *K*-means clustering technique.

other points is $O(n^3)$. As this shortest distance computation overrides other computations related to cluster validity indices, the time complexity of all the above mentioned cluster validity indices is $O(n^3)$.

Once we have constructed RNG and found all paths among the set of points, the total time taken by the cluster validity indices are more or less same. We have executed the proposed technique on HP Server running red hat linux enterprize 4 operating system with 6 GB RAM and 250 GB Hard disk. For *Pat1* data set, the seven cluster validity indices took, respectively, 2, 3, 3, 2, 2, 2 and 2 min of time. For *Pat2* data set, total time requirements of seven cluster validity indices are 1.5, 2.5, 2.5, 1.5, 1.5, 1.5 and 1.5, respectively. Similarly for *Spiral* data set, seven cluster validity indices took, respectively, 3.5, 5, 5, 3.5, 3.5, 3.5 and 3.5 minutes of time. For *Cancer* data set, the cluster validity indices took, respectively, 5, 6, 6, 5, 5, 5, and 5 min of time.

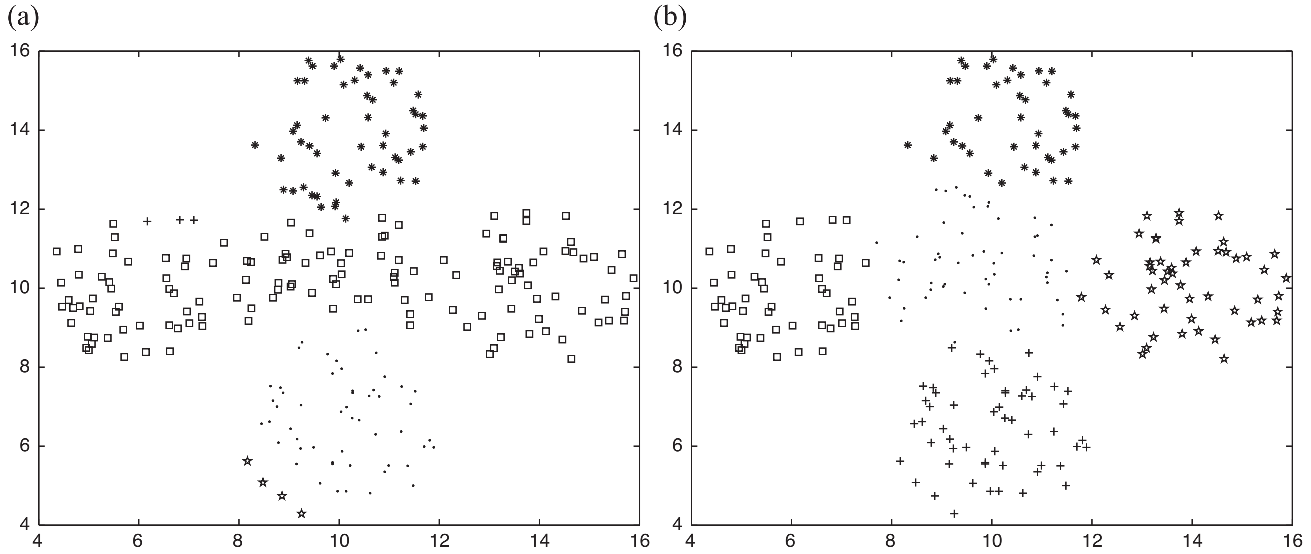

Fig. 13. Optimal partitioning on *Sph_5_2* for $K = 5$ obtained by (a) single linkage clustering technique and (b) K-means clustering technique.

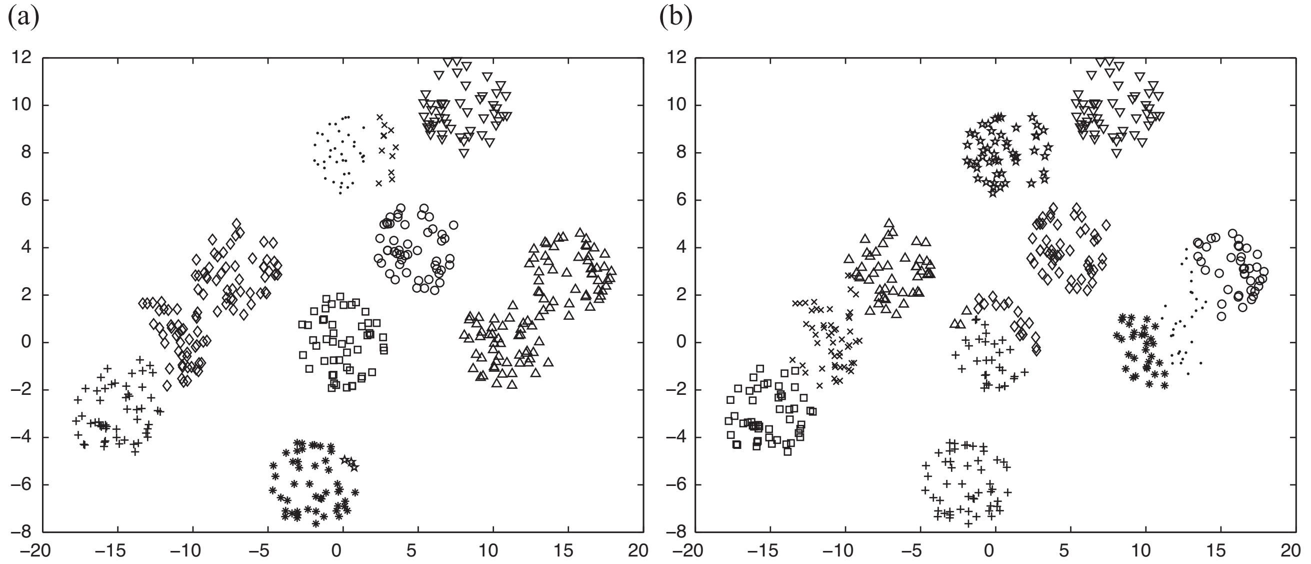

Fig. 14. Optimal partitioning on *Sph_10_2* for $K = 10$ obtained by (a) single linkage clustering technique and (b) K-means clustering technique.

4.2. *Comparison with original versions of the cluster validity indices*

For the purpose of comparison, the values of original seven cluster validity indices, DB-index, Dunn-index, GDunn-index, PS-index, *I*-index, XB-index, SV-index, are also calculated after application of single linkage and *K*-means clustering algorithms on the above mentioned data sets with *K* varies from 2 to $\sqrt{n}$. The number of clusters and the clustering technique identified by these indices on the above mentioned data sets are shown in Table 2.

The success rates of the two versions of seven cluster validity indices (original version and the connectivity based version) in detecting the proper number of partitions and the proper partitioning are shown in Table 2. Here $Success\text{-}Rate(i) = A/\text{total number of data sets}$, where A = Number of data sets for which index i succeeds in determining the appropriate number of clusters. From the results provided in Table 2, it is easy to conclude that incorporation of d_{short} distance in the definitions of existing cluster validity indices make them more effective in detecting any type of clusters from a data set irrespective of their shape, size or convexity as long as they are well-separated. This is more evident from the results on first five artificial data sets having clusters of different shapes but well-separated structures. While the original versions of the seven cluster validity indices mostly fail in detecting the proper number of partitions from these five data sets, incorporation of d_{short} distance impart the property of characterizing these non-compact or non-convex well-separated clusters of any shapes to them.

5. Discussion and conclusion

Identifying the proper number of clusters, appropriate partitioning and the appropriate partitioning technique from a data set are three crucial issues in unsupervised classification. Seven newly proposed connectivity based cluster validity indices which mimic the existing seven cluster validity indices are proposed in this article. These indices exploit the property of connectivity to indicate both the appropriate number of clusters and the appropriate partitioning technique. The effectiveness of these seven newly developed connectivity based indices in comparison with the original seven cluster validity indices is provided for eight artificially generated and three real-life data sets. Results show that incorporation of the concept of connectivity in the definitions of existing seven cluster validity indices makes them more effective in determining the proper number of clusters, proper partitioning and the appropriate partitioning technique from data sets having clusters of different shapes, sizes and convexity as long as they are well-separated. Note that our proposed validity indices will not be able to determine the appropriate partitioning from data sets having overlapping clusters. If the clusters are well-separated then only the proposed indices are useful.

In the newly proposed connectivity based cluster validity indices we have retained whatever normalization was used in the original version. However appropriate normalization is important and a study needs to be conducted in future. Moreover, the utility of the proposed indices needs to be studied on some more real life problems such as segmentation of MR brain images and segmentation of remote sensing satellite imagery.

References

[1] B.S. Everitt, S. Landau, M. Leese, Cluster Analysis, Arnold, London, 2001.

[2] T. Niknam, B. Amiri, An efficient hybrid approach based on pso, aco and *k*-means for cluster analysis, Applied Soft Computing 10 (1) (2010) 183–197.

[3] A. Graaff, A. Engelbrecht, Using sequential deviation to dynamically determine the number of clusters found by a local network neighbourhood artificial immune system, Applied Soft Computing 11 (2) (2011) 2698–2713 (The Impact of Soft Computing for the Progress of Artificial Intelligence).

[4] I. Saha, U. Maulik, D. Plewczynski, A new multi-objective technique for differential fuzzy clustering, Applied Soft Computing 11 (2) (2011) 2765–2776 (The Impact of Soft Computing for the Progress of Artificial Intelligence).

[5] U. Boryczka, Finding groups in data: cluster analysis with ants, Applied Soft Computing 9 (1) (2009) 61–70.

[6] L.K. Ming, L.C. Kiong, L.W. Soong, Autonomous and deterministic supervised fuzzy clustering with data imputation capabilities, Applied Soft Computing 11 (1) (2011) 1117–1125.

[7] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[8] J.T. Tou, R.C. Gonzalez, Pattern Recognition Principles, Addison-Wesley, Reading, 1974.

[9] U. Maulik, S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (12) (2002) 1650–1654.

[10] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (1979) 224–227.

[11] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics 3 (1973) 32–57.

[12] X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (1991) 841–847.

[13] C.H. Chou, M.C. Su, E. Lai, A new cluster validity measure and its application to image compression, Pattern Analysis and Applications 7 (2004) 205–220.

[14] M. Kim, R.S. Ramakrishna, New indices for cluster validity assessment, Pattern Recognition Letters 26 (15) (2005) 2353–2363.

[15] G.W. Milligan, C. Cooper, An examination of procedures for determining the number of clusters in a data set, Psychometrika 50 (2) (1985) 159–179.

[16] J. Breckenridge, Replicating cluster analysis: Method, consistency and validity, Multivariate Behavioral Research 24 (1989) 147–161.

[17] S. Dudoit, J. Fridlyand, A prediction-based resampling method for estimating the number of clusters in a data set, Genome Biology 3 (7) (2002) 1299–1323.

[18] R. Tibshirani, G. Walther, D. Botstein, P. Brown, Cluster Validation by Prediction Strength, Tech. Rep., Statistics Department, Stanford University, Stanford, CA, 2001.

[19] T. Lange, V. Roth, M.L. Braun, J.M. Buhmann, Stability-based validation of clustering solutions, Neural Computation 16 (6) (2004) 1299–1323.

[20] S. Ben-David, U. Luxburg, Dávid Pál, A sober look on clustering stability, in: G. Lugosi, H. Simon (Eds.), Proceedings of the 19th Annual Conference on Learning Theory (COLT), 2006.

[21] S. Saha, U. Maulik, Use of symmetry and stability for data clustering, Evolutionary Intelligence 3 (3–4) (2010) 103–122.

[22] G.T. Toussaint, The realtive neighborhood graph of a finite planar set, Pattern Recognition 12 (1980) 261–268.

[23] S. Bandyopadhyay, An automatic shape independent clustering technique, Pattern Recognition 37 (2004) 33–45.

[24] B. Fischer, V. Roth, J.M. Buhmann, Clustering with the connectivity kernel, NIPS, 16 (2004).

[25] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (August 2000) 888–905.

[26] J.C. Bezdek, N.R. Pal, Some new indexes of cluster validity, IEEE Transactions on Systems, Man and Cybernetics 28 (1998) 301–315.

[27] C.H. Chou, M.C. Su, E. Lai, Symmetry as a new measure for cluster validity, in: 2nd WSEAS International Conference on Scientific Computation and Soft Computing, 2002, pp. 209–213.

[28] D.J. Kim, Y.W. Park, D.J. Park, A novel validity index for determination of the optimal number of clusters, IEICE Transactions on Information and Systems D-E84 (2) (2001) 281–285.

[29] B.S. Everitt, Cluster Analysis, third ed., Halsted Press, 1993.

[30] S.K. Pal, S. Mitra, Fuzzy versions of kohonen's net and mlp-based classification: performance evaluation for certain nonconvex decision regions, Information Sciences 76 (1994) 297–337.

[31] S. Mitra, S.K. Pal, Fuzzy multi-layer perceptron, inferencing and rule generation, IEEE Transactions on Neural Networks 6 (1995) 51–63.

[32] J. Handl, J. Knowles, An evolutionary approach to multiobjective clustering, IEEE Transactions on Evolutionary Computation 11 (1) (2007) 56–76.

[33] S. Bandyopadhyay, S. Saha, GAPS: a clustering method using a new point symmetry based distance measure, Pattern Recognition 40 (2007) 3430–3451.

[34] S. Bandyopadhyay, S. Saha, A point symmetry based clustering technique for automatic evolution of clusters, IEEE Transactions on Knowledge and Data Engineering 20 (11) (November 2008) 1–17.

[35] S. Bandyopadhyay, U. Maulik, Genetic clustering for automatic evolution of clusters and application to image classification, Pattern Recognition (2) (2002) 1197–1208.

[36] S. Bandyopadhyay, S.K. Pal, Classification and Learning Using Genetic Algorithms Applications in Bioinformatics and Web Intelligence, Springer-Verlag, Heidelberg, Germany, 2007.

[37] http://www.ics.uci.edu/~mlearn/MLRepository.html.

[38] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 3 (1936) 179–188.

[39] A. Lingas, A linear-time construction of the relative neighborhood graph from the delaunay triangulation, Computational Geometry 4 (4) (1994) 199–208.