On Some Symmetry Based Validity Indices

Sriparna Saha, *Student Member, IEEE* and Sanghamitra Bandyopadhyay, *Senior Member, IEEE*
Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India-700108
Email: {sriparna_r, sanghami}@isical.ac.in

***Abstract*—Identification of the correct number of clusters and the corresponding partitioning are two important considerations in clustering. In this paper, a newly developed point symmetry based distance is used to propose symmetry based versions of six cluster validity indices namely, DB-index, Dunn-index, Generalized Dunn-index, PS-index, I-index and XB-index. These indices provide measures of "symmetricity" of the different partitionings of a data set. A Kd-tree-based data structure is used to reduce the complexity of computing the symmetry distance. A newly developed genetic point symmetry based clustering technique, GAPS-clustering is used as the underlying partitioning algorithm. The number of clusters are varied from 2 to $\sqrt{n}$ where n is the total number of data points present in the data set and the values of all the validity indices are noted down. The optimum value of a validity index over these $\sqrt{n} - 1$ partitions corresponds to the appropriate partitioning and the number of partitions as indicated by the validity index. Results on five artificially generated and four real-life data sets show that symmetry distance based I-index performs the best compared to all the other five indices.**

***Index Terms*—Unsupervised classification, cluster validity index, symmetry property, point symmetry based distance, Kd-tree**

I. Introduction

Clustering [1] is a core problem in data-mining with innumerable applications spanning many fields. In order to mathematically identify clusters in a data set, it is usually necessary to first define a measure of similarity or proximity which will establish a rule for assigning patterns to the domain of a particular cluster centroid. The measure of similarity is usually data dependent. It may be noted that one of the basic feature of shapes and objects is symmetry. As symmetry is so common in the natural world, it can be assumed that some kind of symmetry exists in the clusters also. Based on this, Su and Chou have proposed a new type of non-metric distance, based on point symmetry. This work is extended in [2] in order to overcome some of the limitations existing in [3]. It has been shown in [4] that the PS distance proposed in [2] has some serious drawbacks. In order to overcome these limitations, a new point symmetry based distance d_{ps} (PS-distance) is developed in [4]. For reducing the complexity of computing the PS-distance, use of Kd-tree [5] is also proposed there. This proposed distance is then used to develop a genetic algorithm based clustering technique, GAPS [4].

The two fundamental questions that need to be addressed in any typical clustering scenario are: (i) how many clusters are actually present in the data, and (ii) how real or good the clustering itself. That is, whatever may be the clustering technique, one has to determine the number of clusters and also the validity of the clusters formed [6]. 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 $V_{k1} \geq V_{k2} \geq \ldots V_{km}, \forall ki \in 1, 2, \ldots, m,\ i = 1, 2, \ldots, m$ will indicate that $U_{k1} \uparrow \ldots \uparrow U_{km}$. 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. These are Davies-Bouldin (DB) index [7], Dunn's index [8], Xie-Beni (XB) index [9], I-index [10], CS-index [11], etc., to name just a few. 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 [12] 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 [10] evaluated the performance of four validity indices, namely, the Davies-Bouldin index [7], Dunn's index [8], Calinski-Harabasz index [10], and a recently developed index $\mathcal{I}$, in conjunction with three different algorithms viz. the well-known K-means [1], single-linkage algorithm [1] and a SA-based clustering method [10].

All the above mentioned indices use the Euclidean distances in their computation. They are therefore able to characterize only convex clusters. It has been shown in [4] that the symmetry based distance is effective not only for convex clusters, but also in cases where the clusters are non-convex, but satisfy the property of point-symmetry. In this article we conjecture that incorporation of the symmetry measure in the above mentioned validity indices will impart the property of characterizing non-convex, symmetric clusters to them. Thus, here we have developed six cluster validity indices using the newly proposed point symmetry based distance rather than the Euclidean distance. These indices follow the definitions of the six well-known existing cluster validity indices, namely, Davies-Bouldin index (DB-index) [7], Dunn index [8], Generalized Dunn's index [13], PS-index [2], $\mathcal{I}$-index [10], Xie-Beni index (XB index) [9]. The newly proposed point symmetry based distance is substituted in place of Euclidean distance in the definitions of these well-known validity indices and their performances are evaluated.

A newly developed genetic point symmetry based clustering technique, GAPS-clustering [4] is used as the underlying clustering algorithm. The number of clusters is varied from K_{min} to K_{max}. As a result, total $(K_{max} - K_{min} + 1)$ partitions will be generated, $U^*_{K_{min}}$, $U^*_{K_{min}+1} \ldots U^*_{K_{max}}$, with the correspondng validity index values computed as $V_{K_{min}}$, $V_{K_{min}+1} \ldots V_{K_{max}}$. Let $K^* = argopt_{i=K_{min} \ldots K_{max}}[V_i]$. Therefore, according to index V, K^* is the correct number of clusters present in the data. The corresponding U^*_K may be obtained by using a suitable clustering technique with the number of clusters set to K^*. The tuple $< U^*_{K^*}, K^* >$ is presented as the solution to the clustering problem. The effectiveness of the newly proposed point symmetry based cluster validity indices namely, *Sym-DB* index, *Sym-Dunn* index, *Sym-GDunn* index, *Sym-PS* index, *Sym-I* index and *Sym-XB* index, are shown in identifying number of clusters from five artificially generated and four real-life data sets of varying complexities. Experimental results show that the *Sym-I* index performs the best compared to all the other five indices.

II. The Existing Point Symmetry (PS)- based Distance Measures[2]

Motivated by the property of point symmetry that clusters often exhibit, a PS-distance was proposed in [3] which was further modified in [2]. The modified distance is defined as follows:
Given N patterns, $\overline{x}_j$, $j = 1, \ldots N$, and a reference vector $\overline{c}$ (e.g., a cluster centroid), the "point symmetry distance" between a pattern $\overline{x}_j$ and the reference vector $\overline{c}$ is defined as

$$d_c(\overline{x}_j, \overline{c}) = d_s(\overline{x}_j, \overline{c}) \times d_e(\overline{x}_j, \overline{c}) \qquad (1)$$

where

$$d_s(\overline{x}_j, \overline{c}) = \min_{i=1,\ldots N \text{ and } i \neq j} \left(\frac{\|(\overline{x}_j - \overline{c}) + (\overline{x}_i - \overline{c})\|}{\|(\overline{x}_j - \overline{c})\| + \|(\overline{x}_i - \overline{c})\|} \right) \qquad (2)$$

and $d_e(\overline{x}_j, \overline{c})$ denotes the Euclidean distance between $\overline{x}_j$ and $\overline{c}$. The value of $\overline{x}_i$, say $\overline{x}^*_j$, for which the quantity within brackets on the right hand side of Equation 2 attains its minimum value, is referred to as the symmetrical point of $\overline{x}_j$ with respect to $\overline{c}$. Note that if $\overline{x}^*_j$ is the same as the reflected point of $\overline{x}_j$ with respect to $\overline{c}$, then the numerator on the right hand side of Equation 2 will be equal to zero, and hence $d_s(\overline{x}_j, \overline{c}) = d_c(\overline{x}_j, \overline{c}) = 0$.

A. Limitations of the PS-distance

It is evident from Equation 1 that the PS-distance measure can be useful to detect clusters which have symmetrical shapes. But it will fail for datasets where clusters themselves are symmetrical with respect to some intermediate point. From equation 1, it can be noted that as $d_e(\overline{x}_j, \overline{c}) \approx d_e(\overline{x}^*_j, \overline{c})$, $d_c(\overline{x}_j, \overline{c}) \approx \frac{d_{symm}(\overline{x}_j, \overline{c})}{2}$, where $d_{symm}(\overline{x}_j, \overline{c}) = \|(\overline{x}_j - \overline{c}) + (\overline{x}^*_j - \overline{c})\|$. In effect, if a point $\overline{x}_j$ is almost equally symmetrical with respect to two centroids $\overline{c}_1$ and $\overline{c}_2$, it will be assigned to that cluster with respect to which it is more symmetric irrespective of the Euclidean distance between the cluster center and the particular point. This is intuitively unappealing. This is demonstrated in Figure 1. The centres of the three clusters are denoted by $\overline{c}_1$, $\overline{c}_2$ and $\overline{c}_3$, respectively. Let us take the point $\overline{x}$. The symmetrical point of $\overline{x}$ with respect to $\overline{c}_1$ is $\overline{x}_1$ as it is the first nearest neighbor of the point $\overline{x}^*_1 = (2 \times \overline{c}_1 - \overline{x})$. Let the Euclidean distance between $\overline{x}^*_1$ and $\overline{x}_1$ be d_1. So the symmetrical distance of $\overline{x}$ with respect to $\overline{c}_1$ is $d_c(\overline{x}, \overline{c}_1) = \frac{d_1}{d_e(\overline{x}, \overline{c}_1) + d_e(\overline{x}_1, \overline{c}_1)} \times d_e(\overline{x}, \overline{c}_1)$. Similarly symmetrical point of $\overline{x}$ with respect to $\overline{c}_2$ is $\overline{x}_2$, and the symmetrical distance of $\overline{x}$ with respect to $\overline{c}_2$ becomes $d_c(\overline{x}, \overline{c}_2) = \frac{d_2}{d_e(\overline{x}, \overline{c}_2) + d_e(\overline{x}_2, \overline{c}_2)} \times d_e(\overline{x}, \overline{c}_2)$. Let $d_2 < d_1$; Now as $d_e(\overline{x}, \overline{c}_2) \approx d_e(\overline{x}_2, \overline{c}_2)$ and $d_e(\overline{x}, \overline{c}_1) \approx d_e(\overline{x}_1, \overline{c}_2)$, therefore $d_s(\overline{x}, \overline{c}_1) \approx d_1/2$ and $d_s(\overline{x}, \overline{c}_2) \approx d_2/2$. Therefore $d_s(\overline{x}, \overline{c}_1) > d_s(\overline{x}, \overline{c}_2)$ and $\overline{x}$ is assigned to $\overline{c}_2$ even though $d_e(\overline{x}, \overline{c}_2) \gg d_e(\overline{x}, \overline{c}_1)$. This will happen for the other points also, finally resulting in merging of the three clusters. This is intuitively unappealing. From the above observations, it can be concluded that the PS-distance measure [2] has two limitations:

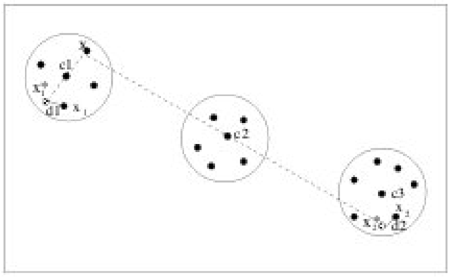

Fig. 1. Example where point symmetry distance proposed by Su and Chou fail

Observation 1 : *The PS-distance measure lacks the Euclidean distance difference property*. Here Euclidean distance difference (EDD) property is defined as follows:
Let $\overline{x}$ be a data point, $\overline{c}_1$ and $\overline{c}_2$ be two cluster centers, and θ be a distance measure. Let $\theta_1 = \theta(\overline{x}, \overline{c}_1)$, $\theta_2 = \theta(\overline{x}, \overline{c}_2)$, $d_{e_1} = d_e(\overline{x}, \overline{c}_1)$ and $d_{e_2} = d_e(\overline{x}, \overline{c}_2)$. Then θ is said to satisfy EDD property if for $\theta_1 \approx \theta_2$, point $\overline{x}$ is assigned to $\overline{c}_1$ if $d_{e1} < d_{e2}$, otherwise it is assigned to $\overline{c}_2$.

It is evident from Figure 1 and from the above discussion that in the PS-distance measure defined in Equation 1, there is no impact of the Euclidean distance. (Although a term $d_e(\overline{x}_j, \overline{c})$ is present, its effect gets almost neutralized by the denominator of the other term, $d_s(\overline{x}_j, \overline{c})$). It only measures the amount of symmetry of a particular point with respect to a particular cluster center. As a result a point might be assigned to a very far off cluster centre, if it happens to be marginally more symmetric with respect to it.

Observation 2: *The PSD measure leads to an unsatisfactory clustering result for the case of symmetrical interclusters*. If two clusters are symmetrical to each other with respect to a third cluster center, then these clusters are called "symmetrical interclusters".

In Figure 1 the first and the third clusters are "symmetrical interclusters" with respect to the middle one. As explained in the example, the three clusters get merged into one cluster since the PS-distance lacks the EDD property. This shows the limitation of the PS-distance in detecting symmetrical interclusters which is also experimentally demonstrated in this paper.

B. A New Definition of the Point Symmetry Distance

As discussed in Section 2, both the PS-based distances, d_s and d_c, will fail when the clusters themselves are symmetrical with respect to some intermediate point. It has been shown, in such cases the points are assigned to the farthest cluster. In order to overcome this limitation, we describe a new PS distance [4], $d_{ps}(\overline{x}, \overline{c})$ associated with point $\overline{x}$ with respect to a center $\overline{c}$. The proposed point symmetry distance is defined as follows: Let a point be $\overline{x}$. The symmetrical (reflected) point of $\overline{x}$ with respect to a particular centre $\overline{c}$ is $2 \times \overline{c} - \overline{x}$. Let us denote this by $\overline{x}^*$. Let $knear$ unique nearest neighbors of $\overline{x}^*$ be at Euclidean distances of d_i, $i = 1, 2, \dots knear$. Then

$$d_{ps}(\overline{x}, \overline{c}) = d_{sym}(\overline{x}, \overline{c}) \times d_e(\overline{x}, \overline{c}), \tag{3}$$

$$= \frac{\sum_{i=1}^{knear} d_i}{knear} \times d_e(\overline{x}, \overline{c}), \tag{4}$$

where $d_e(\overline{x}, \overline{c})$ is the Euclidean distance between the point $\overline{x}$ and $\overline{c}$. It can be seen from Equation 4 that $knear$ cannot be chosen equal to 1, since if $\overline{x}^*$ exists in the data set then $d_{ps}(\overline{x}, \overline{c}) = 0$ and hence there will be no impact of the Euclidean distance. On the contrary, large values of $knear$ may not be suitable because it may overestimate the amount of symmetry of a point with respect to a particular cluster center. Here $knear$ is chosen equal to 2.

Note that $d_{ps}(\overline{x}, \overline{c})$, which is a non-metric, is a way of measuring the amount of symmetry between a point and a cluster center, rather than the distance like any Minkowski distance.

The basic differences between the PS-distances in [3] and [2], and the proposed $d_{ps}(\overline{x}, \overline{c})$ are follows:

1) Instead of computing Euclidean distance between the original reflected point $\overline{x}^* = 2 \times \overline{c} - \overline{x}$ and its first nearest neighbor as in [3] and [2], here the average distance between $\overline{x}^*$ and its $knear$ unique nearest neighbors have been taken. Consequently this term will never be equal to 0, and the effect of $d_e(\overline{x}, \overline{c})$, the Euclidean distance, will always be considered. Note that if only the nearest neighbor of $\overline{x}^*$ is considered and this happens to coincide with $\overline{x}^*$, then this term will be 0, making the distance insensitive to $d_e(\overline{x}, \overline{c})$. But considering $knear$ nearest neighbors will reduce the problems discussed in Figure 1.
2) Considering the $knear$ nearest neighbors in the computation of d_{ps} makes the PS-distance more robust and noise resistant. From an intuitive point of view, if this term is less, then the likelihood that $\overline{x}$ is symmetrical with respect to $\overline{c}$ increases. This is not the case when only the first nearest neighbor is considered which could mislead the method in noisy situations.
3) In the PS-distances (in Equation 2) the denominator term is used to normalize the point symmetry distance so as to make it insensible to the Euclidean distance. But as shown earlier this will lead to lack of EDD property. As a result, d_c can not identify symmetrical interclusters. Unlike this, in d_{ps} (Equation 3), no denominator term is incorporated to normalize d_{sym}.

Observation: The proposed d_{ps} measure will, in general, work well for symmetrical interclusters. Using $knear = 2$, let the two nearest neighbors of the reflected point of $\overline{x}$ (in Figure 1) with respect to center $\overline{c}_1$ are at distances of d_1 and d_1^1 respectively. Then $d_{ps}(\overline{x}, \overline{c}_1) = d_{sym}(\overline{x}, \overline{c}_1) \times d_{e1} = \frac{d_1 + d_1^1}{2} \times d_{e1}$, where d_{e1} is the Euclidean distance between $\overline{x}$ and $\overline{c}_1$. Let the two nearest neighbors of the reflected point of $\overline{x}$ with respect to center $\overline{c}_2$ be at distances of d_2 and d_2^1 respectively. Hence, $d_{ps}(\overline{x}, \overline{c}_2) = d_{sym}(\overline{x}, \overline{c}_2) \times d_{e2} = \frac{d_2 + d_2^1}{2} \times d_{e2}$, where d_{e2} is the Euclidean distance between $\overline{x}$ and $\overline{c}_2$. Now in order to preserve the Euclidean distance difference property (EDD), i.e., to avoid merging of symmetrical interclusters, $d_{ps}(\overline{x}, \overline{c}_1)$ should be less than $d_{ps}(\overline{x}, \overline{c}_2)$ even when $d_{sym}(\overline{x}, \overline{c}_1) \approx d_{sym}(\overline{x}, \overline{c}_2)$. Now,

$$\begin{aligned} d_{ps}(\overline{x}, \overline{c}_1) &< d_{ps}(\overline{x}, \overline{c}_2) \\ \Longrightarrow \frac{d_1 + d_1^1}{2} \times d_{e1} &< \frac{d_2 + d_2^1}{2} \times d_{e2} \\ \Longrightarrow \frac{d_{e1}}{d_{e2}} &< \frac{d_2 + d_2^1}{d_1 + d_1^1}. \end{aligned} \tag{5}$$

From Figure 1, it is evident that, $d_{e2} >> d_{e1}$, so $\frac{d_{e1}}{d_{e2}} << 1$. Thus even when $(d_2 + d_2^1) \approx (d_1 + d_1^1)$, the inequality in Equation 5 is satisfied. Therefore the proposed distance satisfies EDD property and avoids merging of symmetrical interclusters. The experimental results provided in [4] also support the fact that the proposed measure is robust even in the presence of symmetrical interclusters since it obeys EDD property.

It is evident that the symmetrical distance computation is very time consuming because it involves the computation of the nearest neighbors. Computation of $d_{ps}(\overline{x_i}, \overline{c})$ is of complexity $O(N)$. Hence for N points and K clusters, the complexity of assigning the points to the different clusters is $O(N^2K)$. In order to reduce the computational complexity, an approximate nearest neighbor search using the Kd-tree approach is adopted in this article.

C. Kd-tree Based Nearest Neighbor Computation

A K-dimensional tree, or Kd-tree is a space-partitioning data structure for organizing points in a K-dimensional space. ANN (Approximate Nearest Neighbor) is a library written in C++ [14], which supports data structures and algorithms for both exact and approximate nearest neighbor searching in arbitrarily high dimensions. In this article ANN is used to find exact d_is, where $i = 1, \dots, knear$, in Equation 4 efficiently. The ANN library implements a number of different data structures, based on Kd-trees and box-decomposition trees, and employs a couple of different search strategies. ANN allows the user to specify a maximum approximation error bound, thus allowing the user to control the tradeoff between accuracy and running time. For the purpose of this article, we have kept the error bound=0, calculating the exact d_is. The Kd-tree structure can be constructed in $O(nlogn)$ time and takes $O(n)$ space [5].

III. Cluster Validity Indices

In this section, the six point symmetry distance based cluster validity indices are defined. Note that the definitions of these indices are inspired by those of six well-known existing cluster validity indices.

A. Symmetry Based Davies-Bouldin index (Sym-DB index)

This index is along the lines of the popular Davies-Bouldin (DB) index [7]. This is a function of the ratio of the sum of *within-cluster symmetry* to *between cluster separation*. The scatter within the ith cluster, S_i, is computed as

$$S_i = \frac{\sum_{\overline{x} \in C_i} d^*_{ps}(\overline{x}, \overline{z}_i)}{|C_i|},$$

where $\overline{z}_i$ represents the center of cluster i and $d^*_{ps}(\overline{x}, \overline{z}_i)$ is computed using Equation 4 with some constraint. Note that here the $knear$ nearest neighbors of the reflected point $\overline{x}^*$ of the point $\overline{x}$ with respect to $\overline{z_i}$ and $\overline{x}$ should belong to the ith cluster, i.e., the first $knear$ nearest neighbors of $\overline{x}^* = 2 \times \overline{z}_i - \overline{x}$ are searched among the points which are already in cluster i. The distance between cluster C_i and C_j, denoted by d_{ij}, is defined as $d_{ij} = d_e(\overline{z}_i, \overline{z}_j)$, where d_e stands for Euclidean distance computation. Then Symmetry Based DB index, *Sym-DB* index, is defined as

$$\textit{Sym-DB} = \frac{\sum_{i=1}^{K} R_i}{K}.$$

where $R_i = \max_{j, j \neq i} \{\frac{S_i + S_j}{d_{ij}}\}$. The objective is to minimize the *Sym-DB* index for achieving proper clustering.

B. Symmetry Based Dunn's Index (Sym-Dunn index)

This index is along the lines of popular Dunn's index [8]. Let S and T be two nonempty subsets of R^N. Then the radius $\triangle$ of S is defined as

$$\triangle(S) = \max_{x \in S} \{d^*_{ps}(\overline{x}, \overline{z})\},$$

where $\overline{z}$ represents the center of set S and $d^*_{ps}(\overline{x}, \overline{z})$ is computed using Equation 4. Note that here the $knear$ nearest neighbors of the reflected point $\overline{x}^*$ of the point $\overline{x}$ with respect to $\overline{z}$ and $\overline{x}$ should belong to the set S. The set distance δ between S and T is defined as

$$\delta(S, T) = \min_{\overline{x} \in S, \overline{y} \in T} \{d_e(\overline{x}, \overline{y})\}.$$

Here, $d_e(\overline{x}, \overline{y})$ indicates the Euclidean distance between points $\overline{x}$ and $\overline{y}$. For any partition, *Sym-Dunn* index is defined as follows

$$\textit{Sym-Dunn} = \min_{1 \leq i \leq K} \min_{1 \leq j \leq K, j \neq i} \{\frac{\delta(C_i, C_j)}{\max_{1 \leq k \leq K} \triangle(C_k)}\}.$$

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

C. Symmetry Based Generalized Dunn's Index (Sym-GDunn index)

This index is developed along the lines of Generalized Dunn's index [13]. The generalized Dunn's index was developed after demonstrating the sensitivity of the original Dunn's index [8], to changes in cluster structure, since not all of the data points were involved in the computation of the index. Let δ_i be any positive, semi-definite, symmetric set distance function and $\triangle_j$ be any positive, semi-definite diameter function. Then the generalized Dunn's index, *Sym-GDunn* index is defined as

$$\textit{Sym-GDunn} = \min_{1 \leq s \leq K} \{ \min_{1 \leq t \leq K, t \neq s} \{\frac{\delta_i(C_s, C_t)}{\max_{1 \leq k \leq K} \triangle_j(C_k)}\}\}.$$

As like [13], five set distance functions and three diameter functions can be defined. But here we have used δ_3 and $\triangle_3$. These two measures δ_3 and $\triangle_3$ are defined as follows:

$$\triangle_3(S) = 2(\frac{\sum_{x \in S} d^*_{ps}(\overline{x}, \overline{z}_S)}{|S|})$$

and

$$\delta_3(S, T) = \frac{1}{|S||T|} \sum_{\overline{x} \in S, \overline{y} \in T} d_e(\overline{x}, \overline{y}).$$

Here $\overline{z}_S$ and $\overline{z}_T$ are the centers of the sets S and T, respectively. Here, $d^*_{ps}(\overline{x}, \overline{z}_S)$ is computed by Equation 4 with some constraint. Note that here the $knear$ nearest neighbors of the reflected point $\overline{x}^*$ of the point $\overline{x}$ with respect to $\overline{z}_S$, and $\overline{x}$ should belong to the set S. Larger values of *Sym-GDunn* correspond to good clusters, and the number of clusters that maximizes *Sym-GDunn* is taken as the optimal number of clusters.

D. Newly proposed symmetry distance based PS-index (Sym-PS index)

This index is developed along the lines of PS-index [2]. The cluster validity index, *Sym-PS* index, is defined as

$$\textit{Sym-PS}(K) = \frac{1}{K} \sum_{i=1}^{K} \frac{1}{n_i} \sum_{\overline{x} \in S_i} \frac{d^*_{ps}(\overline{x}, \overline{z}_i)}{min_{m,n=1,\ldots,K,\ m \neq n} d_e(\overline{z}_m, \overline{z}_n)}$$

$$\Rightarrow \frac{1}{K} \sum_{i=1}^{K} \frac{1}{n_i} \sum_{\overline{x} \in S_i} \frac{d^*_{ps}(\overline{x}, \overline{z}_i)}{d_{min}} \tag{6}$$

where 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 Euclidean distance between any two cluster centers and $d^*_{ps}(\overline{x}, \overline{z}_i)$ is computed by Equation 4 with some constraint. Note that the $knear$ nearest neighbors of the reflected point $\overline{x}^*$ of the point $\overline{x}$ with respect to $\overline{z}_i$ and $\overline{x}$ should belong to the ith cluster. The smallest *Sym-PS*(K^*) indicates a valid optimal partition with the optimal cluster number K^*.

E. Symmetry distance based I-index (Sym-I index)

This is inspired by the I-index developed in [10]. Consider a partition of the data set $X = \{\overline{x}_j : j = 1, 2, \ldots n\}$ and the center of each cluster $\overline{z}_i$ can be computed by using $\overline{z}_i = \frac{\sum_{j=1}^{n_i} \overline{x}_j^i}{n_i}$ where n_i $(i = 1, 2, \ldots, K)$ is the number of points in cluster i. The new cluster validity function *Sym-I* index is defined as:

$$\textit{Sym-I}(K) = \left(\frac{1}{K} \times \frac{1}{\mathcal{E}_K} \times D_K \right), \qquad (7)$$

where K is the number of clusters. Here,

$$\mathcal{E}_K = \sum_{i=1}^{K} E_i$$

such that $E_i = \sum_{j=1}^{n_k} d_{ps}^*(\overline{x}_j^i, \overline{z}_i)$ and $D_K = max_{i,j=1}^{K} ||\overline{z}_i - \overline{z}_j||$. D_K is the maximum Euclidean distance between two cluster centres among all centres. $d_{ps}^*(\overline{x_j^i}, \overline{z_i})$ is computed by Equation 4 with some constraint. Here, first $knear$ nearest neighbors of $\overline{x}_j^* = 2 \times \overline{z}_i - \overline{x}_j^i$ are searched among the points which are already in cluster i, i.e., now the $knear$ nearest neighbors of the reflected point $\overline{x}_j^*$ of the point $\overline{x_j}$ with respect to $\overline{z}_i$ and $\overline{x_j}$ should belong to the ith cluster. The objective is to maximize this index in order to obtain the actual number of clusters.

F. Symmetry Distance Based Xie-Beni index (Sym-XB index)

Xie and Beni proposed a validity index (V_{XB}) that focussed on two properties: compactness and separation [9]. Here we have developed a new validity index, named *Sym-XB* index, along the lines of XB-index using newly developed point symmetry based distance. It is defined as follows:

$$\textit{Sym-XB} = \frac{\sum_{i=1}^{K} (\sum_{\overline{x} \in C_i} d_{ps}^{*2}(\overline{x}, \overline{z}_i))}{n(\min_{i,k=1,\ldots K, i \neq k} d_e^2(\overline{z}_i, \overline{z}_k))}.$$

$d_{ps}^*(\overline{x}, \overline{z_i})$ is computed by Equation 4. Note that here also the $knear$ nearest neighbors of the reflected point $\overline{x}^*$ of the point $\overline{x}$ with respect to $\overline{z_i}$ and $\overline{x}$ should belong to the ith cluster. The most desirable partition (or an optimal value of K) is obtained by minimizing *Sym-XB* index over $K = 2, 3, \ldots K_{max}$.

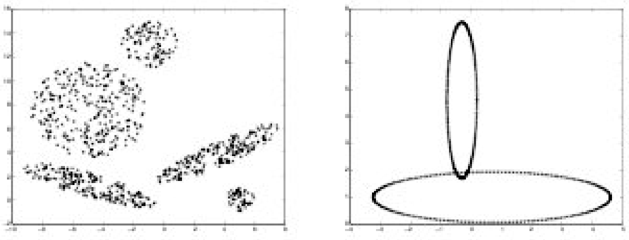
(a) (b)

Fig. 2. (a) Data1 (b) Data2

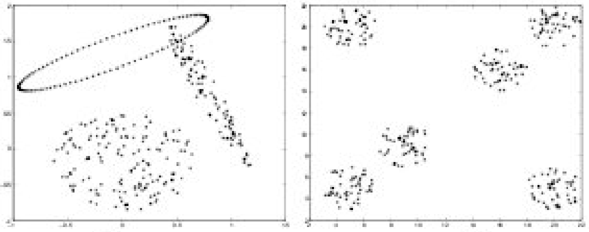
(a) (b)

Fig. 3. (a) Data3 (b) Data4

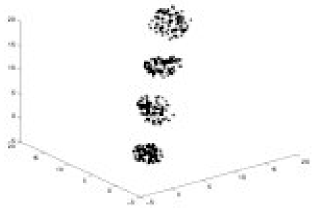

Fig. 4. Data5

G. GAPS: Clustering Algorithm Used for Segmentation

A newly proposed point symmetry based genetic clustering technique (GAPS) [4] is used as the underlying segmentation method. This algorithm uses the newly proposed point symmetry based distance for assigning the points to different clusters. The basic steps of GAPS closely follow those of the conventional GA. Here centre based encoding of the chromosome is used. Each string is a sequence of real numbers representing the *K* cluster centres. The *K* cluster centres encoded in each chromosome are initialized to *K* randomly chosen points from the data set. This process is repeated for each of the $Popsize$ chromosomes in the population, where $Popsize$ is the size of the population. Thereafter five iterations of the *K*-means algorithm are executed with the set of centers encoded in each chromosome. The resultant centers are used to replace the centers in the corresponding chromosomes. This makes the centers separated initially. In order to compute the fitness of the chromosomes, firstly assignment of points to different clusters are done. Here a point $\overline{x}_i$, $1 \leq i \leq n$, is assigned to cluster k iff $d_{ps}(\overline{x}_i, \overline{c}_k) \leq d_{ps}(\overline{x}_i, \overline{c}_j), \quad j = 1, \ldots, K, \ j \neq k$ and $d_{sym}(\overline{x}_i, \overline{c}_k) \leq \theta$. For $d_{sym}(\overline{x}_i, \overline{c}_k) > \theta$, point $\overline{x}_i$ is assigned to some cluster m iff $d_e(\overline{x}_i, \overline{c}_m) \leq d_e(\overline{x}_i, \overline{c}_j), \ j = 1, 2 \ldots K, \ j \neq m$. In other words, point $\overline{x}_i$ is assigned to that cluster with respect to whose centers its PS-distance is the minimum, provided the amount of symmetricity with respect to that cluster center is less than some threshold θ. Otherwise assignment is done based on the minimum Euclidean distance criterion as normally used in [15] or the *K*-means algorithm. We also provide a rough guideline of the choice of θ, the threshold value on the PS-distance. It is to be noted that if a point is indeed symmetric with respect to some cluster centre then

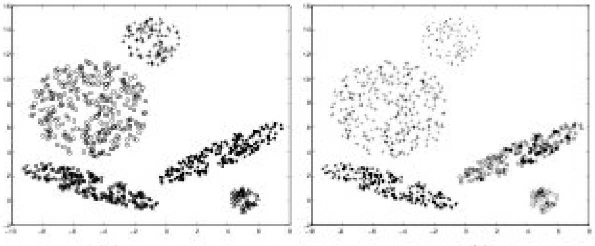
(a) (b)

Fig. 5. Clustered Data1 after application of GAPS (a) for $K = 5$ (b) for $K = 6$

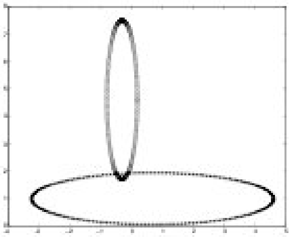

Fig. 6. Clustered Data2 after application of GAPS for $K = 2$

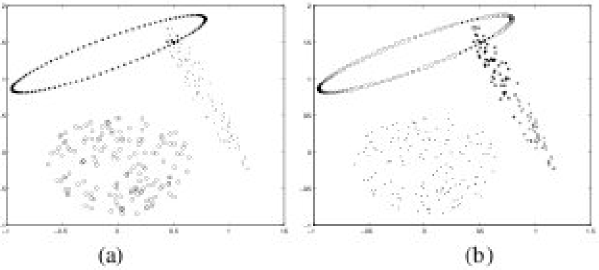
(a) (b)

Fig. 7. Clustered Data3 after application of GAPS (a) for $K = 3$ (b) for $K = 6$

TABLE I
DATA SETS USED

Name	No of points	dimension	Actual No. Clusters
Data1	850	2	5
Data2	400	2	2
Data3	350	2	3
Data4	300	2	6
Data5	400	3	4
Iris	150	4	3
Cancer	683	9	2
Newthyroid	215	5	3
glass	214	9	6

the symmetrical distance computed in the above way will be small, and can be bounded as follows. Let d_{NN}^{max} be the maximum nearest neighbor distance in the data set. That is $d_{NN}^{max} = \max_{i=1,\ldots N} d_{NN}(\overline{x}_i)$, where $d_{NN}(\overline{x}_i)$ is the nearest neighbor distance of $\overline{x}_i$. Assuming that $\overline{x}^*$ lies within the data space, it may be noted that $d_1 \leq \frac{d_{NN}^{max}}{2}$ and $d_2 \leq \frac{3d_{NN}^{max}}{2}$, resulted in, $\frac{d_1+d_2}{2} \leq d_{NN}^{max}$. Ideally, a point $\overline{x}$ is exactly symmetrical with respect to some $\overline{c}$ if $d_1 = 0$. However considering the uncertainty of the location of a point as the sphere of radius d_{NN}^{max} around $\overline{x}$, we have kept the threshold θ equals to d_{NN}^{max}. Thus the computation of θ is automatic and does not require user intervention.

After the assignments are done, the cluster centres encoded in the chromosome are replaced by the mean points of the respective clusters. Subsequently for each chromosome, *clustering_metric*, M is calculated as defined below:

```
$M = 0$
For $k = 1$ to $K$ do
    For all data points $\overline{x}_i$, $i = 1$ to $n$ and $\overline{x}_i \in k$th cluster, do
        $M = M + d_{ps}(\overline{x}_i, \overline{c}_k)$
```

Then the fitness function of that chromosome, fit, is defined as the inverse of M, i.e., $fit = \frac{1}{M}$. This fitness function, fit, will be maximized by using genetic algorithm. Roulette wheel selection is used to implement the proportional selection strategy. The normal single point crossover operation is used here. Crossover probability is selected adaptively as in [4]. Each chromosome undergoes mutation with a probability μ_m. The mutation probability is also selected adaptively for each chromosome as in [4]. In GAPS, the processes of fitness computation, selection, crossover, and mutation are executed for a maximum number of generations. The best string seen upto the last generation provides the solution to the clustering problem.

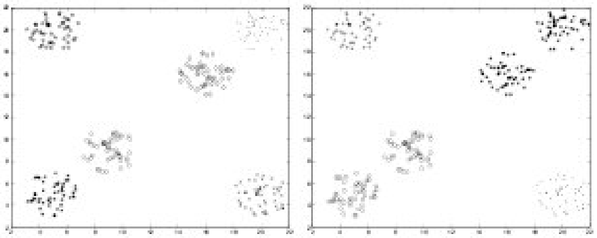
(a) (b)

Fig. 8. Clustered Data4 after application of GAPS (a) for $K = 6$ (b) for $K = 4$

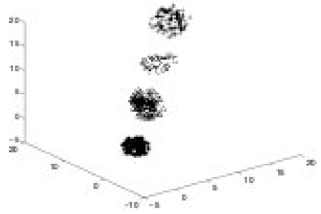

Fig. 9. Clustered Data5 after application of GAPS for $K = 4$

IV. EXPERIMENTAL RESULTS

The six data sets that are used for the experiments are divided into 2 different groups. The first group consists of five artificially generated 2/3-dimensional data

TABLE II

OPTIMAL VALUES OF THE FIVE SYMMETRY BASED VALIDITY INDICES FOR NINE DATA SETS, SEGMENTED USING GAPS CLUSTERING ALGORITHM WHERE K IS VARIED FROM 2 TO $\sqrt{n}$. THE NUMBER WITHIN BRACKETS IS THE CORRESPONDING CLUSTER NUMBER.

Data Set	Sym-DB Value (K^*)	Sym-Dunn Value (K^*)	Sym-GDunn Value (K^*)	Sym-PS Value (K^*)	Sym-I Value (K^*)	Sym-XB Value (K^*)
Data1	0.14(5)	0.577(5)	3.4218(5)	0.082(6)	0.0076(5)	0.0198(5)
Data2	0.05(2)	0.1456(2)	17.9135(2)	0.026(2)	0.049(2)	0.000836(2)
Data3	0.04(6)	0.565239(3)	22.558568(3)	0.024(3)	0.057(3)	0.000855(3)
Data4	0.13(6)	2.038660(4)	6.079382(6)	0.056928(4)	0.027737(6)	0.006497(4)
Data5	0.28(4)	1.020440(4)	2.606002(4)	0.137344(4)	0.013723(4)	0.024270(4)
Iris	0.18(2)	0.502211(8)	5.148168(2)	0.091630(2)	0.049885(3)	0.029(2)
Cancer	3.68(2)	0.033323(8)	0.147575(2)	1.840008(2)	0.000622(2)	4.88(2)
Newthyroid	5.206076(10)	0.008218(2)	0.177528(6)	2.923031(8)	0.001339(3)	18.678121(4)
Glass	1.299(5)	0.044829(8)	0.473918(2)	0.904809(2)	0.006686(6)	1.12(8)

sets. Figures 2(a), 2(b), 3(a), 3(b) and 4 show Data1, Data2, Data3, Data4 and Data5, respectively. The second group consists of four real-life data sets. These are *Iris*, *Cancer*, *Newthyroid* and *Glass* obtained from (http://www.ics.uci.edu/mlearn/MLRepository.html). The description of the data sets used for the experiment are shown in Table I. *Iris* data set represents different categories of irises characterized by four feature values. It has three classes Setosa, Versicolor and Virginica. The Wisconsin Breast Cancer data set has two categories in it: malignant and benign. The two classes are known to be linearly separable. The *Newthyroid* is the 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. The *Glass* identification data consists of 9 attributes. There are six classes present in the data. The study of classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence, if it is correctly identified.

The parameters of the GAPS are as follows: population size=100 and maximum number of generations=30. For GAPS, the crossover and mutation probabilities are chosen adaptively as in [4]. The results reported in the table are the average values obtained over ten runs of the algorithm. Here K_{min} is kept equal to 2 and K_{max} is set equal to $\sqrt{n}$ where n is the total number of data points present in the data set.

Figures 5(a), 6, 7(a), 8(a) and 9(a) show, respectively, the partitions obtained after application of GAPS-clustering on Data1, Data2, Data3, Data4 and Data5, respectively, for actual number of clusters present in the data sets. Table II shows the optimum values of the newly proposed point symmetry distance based six cluster validity indices, namely, *Sym-DB*, *Sym-Dunn*, *Sym-GDunn*, *Sym-PS*, *Sym-I* and *Sym-XB* indices, over $K = 1$ to $\sqrt{n}$ number of partitions obtained after application of GAPS-clustering and the indicating partition number for all the nine data sets used here. It can be seen that for Data1, all the indices except *Sym-PS* is able to find the proper clustering and the proper cluster number. *Sym-PS* provides $K^* = 6$ as the proper cluster number. The corresponding segmentation is shown in Figure 5(b). For Data2, GAPS-clustering is able to detect the proper clustering for $K = 2$ (shown in Figure 6) and all the indices are able to identify this. As like the previous case, for Data3 also GAPS-clustering is able to find the proper partitioning for $K = 3$ (the corresponding segmentation result is shown in Figure 7(a)) and all the indices except *Sym-DB* is able to detect this. Optimum value of *Sym-DB* wrongly indicates six as the proper cluster number. The corresponding partitioning is shown in Figure 7(b). But for Data4, *Sym-DB*, *Sym-GDunn* and *Sym-I* are able to detect the proper clustering (Figure 8(a)) and the proper partition number. *Sym-Dunn*, *Sym-PS* and *Sym-XB* merges two pairs of clusters into two clusters. The corresponding partitioning is shown in Figure 8(b). For Data5, all the indices are able to detect the proper clustering after application of GAPS (the corresponding partitioning is shown in Figure 9).

For the four real-life data sets, *Iris*, *Cancer*, *New-thyroid* and *Glass*, no visualization is possible as these are high-dimensional data sets. For these four data sets, the *Minkowski Score* [16] is calculated after application of GAPS-clustering algorithm. 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:

$$MS(T,S) = \sqrt{\frac{n_{01} + n_{10}}{n_{11} + n_{10}}}. \qquad (8)$$

For MS, the optimum score is 0, with lower scores being "better". For *Iris* data set, MS value corresponding to the partitioning obtained by GAPS-clustering for $K = 3$ is 0.62. As can be seen from Table II, only *Sym-I* index is able to detect the proper partition number for this data set. Optimum values of *Sym-DB*, *Sym-GDunn*, *Sym-PS* and *Sym-XB* indices indicate two clusters, which is also often obtained for many other methods for *Iris*. The performance *Sym-Dunn* is the worst. For *Cancer* dataset, MS value corresponding to the

TABLE III
EXECUTION TIME IN SECONDS BY GAPS WITH AND WITHOUT KD TREE BASED SEARCH

Data set	GAPS with Kd tree	GAPS with out Kd tree
Data3	77	5280
Data4	90	3870
Data5	128	6112

partitioning obtained by GAPS-clustering for $K = 2$ is 0.367056. *Sym-DB*, *Sym-GDunn*, *Sym-PS*, *Sym-I* and *Sym-XB* indices are able to indicate this partitioning. But again for this data set, the performances *Sym-Dunn* is the worst. For *Newthyroid* data set, only *Sym-I*-index is able to detect the proper cluster number (3 in this case) along with GAPS-clustering (see Table II). The corresponding MS value is 0.58. No other validity indices are able to detect the proper cluster number along with GAPS-clustering. For *Glass* data set, again only *Sym-I* index is able to detect the proper cluster number. The MS score of the corresponding partitioning is 0.7223. No other indices are able to detect the proper partitioning or the proper partition number. From the results on nine data sets, obtained by six newly developed point symmetry distance based cluster validity indices along with the GAPS-clustering technique, it is revealed that *Sym-I* index is able to detect the proper cluster number in almost all the cases. It may be noted that for real-life data sets having higher number of dimensions, most of the symmetry based cluster validity indices do not perform well. This may be due to the inability of the most of the cluster validity indices to handle higher dimensional data sets. More experiments have to be done in order to find out the proper reason of the inability of the proposed indices for detecting number of clusters from data sets having higher number of dimensions.

A. Effectiveness of Using Kd-tree for Nearest Neighbor Search

Note that the proposed implementations of GAPS and point symmetry based distance utilize a Kd-tree structure to reduce the time required for identifying the nearest neighbors. In order to demonstrate the time gain obtained, GAPS is also executed without using the Kd-tree data structure. GAPS is implemented in C and executed on a PIV processor, 1.6GHz speed, running Linux. Table III provides the time required for the two cases for three data sets Data3, Data4 and Data5. As is evident, incorporation of Kd-tree significantly reduces the computational burden of the process.

V. DISCUSSION AND CONCLUSION

Identifying the proper number of clusters and the proper partitioning from a data set are two crucial issues in unsupervised classification. Six newly proposed point symmetry distance based cluster validity indices which mimic the existing six cluster validity indices are proposed in this article. These newly developed indices exploit the property of point based symmetry to indicate both the appropriate number of clusters as well as the appropriate partitioning. Here point symmetry based distance is used in place of Euclidean distance in the definitions of the well-known six cluster validity indices, namely, DB-index, Dunn-index, Generalized Dunn's index, PS-index, I-index and XB-index. The performance of these six newly developed point symmetry based indices, named as *Sym-DB* index, *Sym-Dunn* index, *Sym-GDunn* index, *Sym-PS* index, *Sym-I* index and *Sym-XB* index, respectively, are evaluated on five artificially generated and four real-life data sets. Results show that *Sym-I* index is more effective than the other five in finding the proper cluster number and the proper partitioning from datasets with different shapes and convexity as long as the clusters present in them are symmetric in nature. As the point symmetry based distance computation is a time consuming process, Kd-tree based nearest neighbor search is used to reduce its time complexity. Future work includes incorporation of the newly developed point symmetry based distance in many other existing cluster validity indices. Performance of these symmetry based validity indices need to be checked along with many other existing clustering algorithms on many other real-world data sets.

REFERENCES

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

[2] C. H. Chou, M. C. Su, and E. Lai, "Symmetry as a new measure for cluster validity," in *2nd WSEAS Int. Conf. on Scientific Computation and Soft Computing*, Crete, Greece, 2002, pp. 209–213.

[3] M.-C. Su and C.-H. Chou, "A modified version of the k-means algorithm with a distance based on cluster symmetry," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 674–680, 2001.

[4] S. Bandyopadhyay and S. Saha, "GAPS: A clustering method using a new point symmetry based distance measure," *Pattern Recog.*, Accepted (March, 2007), URL: http://dx.doi.org/10.1016/j.patcog.2007.03.026.

[5] M. R. Anderberg, *Computational Geometry: Algorithms and Applications*. Springer, 2000.

[6] R. C. Dubes and A. K. Jain, "Clustering techniques : The user's dilemma," *Pattern Recognition*, vol. 8, pp. 247–260, 1976.

[7] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions Patt. Anal. Mach. Intell.*, vol. 1, pp. 224–227, 1979.

[8] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cyberns.*, vol. 3, pp. 32–57, 1973.

[9] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 841–847, 1991.

[10] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

[11] C. H. Chou, M. C. Su, and E. Lai, "A new cluster validity measure and its application to image compression," *Pattern Analysis and Applications*, vol. 7, pp. 205–220, 2004.

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

[13] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Transactions Systs., Man Cyberns.*, vol. 28, pp. 301–315, 1998.

[14] D. M. Mount and S. Arya, "ANN: A library for approximate nearest neighbor searching," 2005, http://www.cs.umd.edu/∼mount/ANN.

[15] U. Maulik and S. Bandyopadhyay, "Genetic algorithm based clustering technique," *Pattern Recog.*, vol. 33, pp. 1455–1465, 2000.

[16] A. Ben-Hur and I. Guyon, *Detecting Stable Clusters using Principal Component Analysis in Methods in Molecular Biology*, M. Brownstein and A. Kohodursky, Eds. Humana press, 2003.