# A Validity Index Based on Cluster Symmetry

Sriparna Saha, *Student Member, IEEE* and Sanghamitra Bandyopadhyay, *Senior Member, IEEE*

***Abstract*— An important consideration in clustering is the determination of the correct number of clusters and the appropriate partitioning of a given data set. In this paper, a newly developed point symmetry distance is used to propose a new cluster validity index named *Sym*-index which provides a measure of "symmetricity" of the different partitionings of a data set. The index is able to address all the above mentioned issues, viz., determining the number of clusters and evolving the proper partitioning as long as the clusters possess the property of symmetry. A Kd-tree-based data structure is used to reduce the complexity of computing the symmetry distance. Results demonstrating the superiority of the *Sym*-index in appropriately determining the proper partitioning and the number of clusters, as compared to two other recently proposed measures, namely the PS-index and $\mathcal{I}$-index, are provided for three clustering methods viz., two recently developed genetic algorithm based clustering techniques and the average linkage clustering algorithm. Four artificial data sets and two real life data sets are considered for this purpose. The effectiveness of the proposed validity index is then demonstrated for automatically classifying different landcover regions in remote sensing imagery.**

***Index Terms*— Unsupervised classification, cluster validity index, symmetry, point symmetry based distance, Kd tree, remote sensing imagery**

## 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. One of the basic feature of shapes and objects is symmetry. Su and Chou have proposed a point symmetry (PS) distance based similarity measure [2]. This work is extended in [3] to overcome some of the limitations existing in [2].

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 [4]. 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 $V_{k1}, \ldots, V_{km}$. The measure of validity of clusters should be such that it will be able to impose an ordering of the partitions in terms of their goodness. Several cluster validity indices have been proposed in the literature. 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.

Most of the validity measures usually assume a certain geometrical structure in the cluster shapes. But if several different structures exist in the same data set, these have often been found to fail. In [3], Chou et al. proposed a validity measure called PS-index, which is based on modified PS distance and it is capable of taking into account the variability of the cluster shapes. It has been shown in [5] that the PS distance proposed in [3] has some serious drawbacks. Consequently we conjecture here that the PS-index will be unable to identify the proper clustering in certain situations since it is based on the PS-distance which itself has some limitations (this is also demonstrated here experimentally). Therefore it would be challenging to design a cluster validity index that is able to detect not only the correct number of clusters but also indicate the appropriate partitioning. This article presents an attempt in this direction. Here we propose a cluster validity index named *Sym*-index (symmetry based cluster validity index) that uses a new definition of PS distance ($d_{ps}$). This distance, $d_{ps}$, is able to remove the drawbacks of the PS distances proposed in [2] and [3]. If the number of clusters, $K$, is varied within some range, then the value of $K$ corresponding to the maximum value of *Sym*-index will indicate the correct number of clusters for the data.

The superiority of this index as compared to PS-index [2] and a recently proposed $\mathcal{I}$-index [6] is demonstrated for four artificially generated data sets with different characteristics and two real-life data sets. Automatic classification of landcover regions in remote sensing image is used as another real-life application for demonstrating the effectiveness of *Sym*-index.

## II. THE EXISTING POINT SYMMETRY BASED CLUSTER VALIDITY INDEX [3]

In this section the existing PS-distances as proposed in [2] [3] are first described, and their limitations are discussed. The existing symmetry based cluster validity index is then described in detail.

authors are with Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India,Email:{sriparna_r,sanghami}@isical.ac.in

### A. The Point Symmetry (PS)- based Distance Measures

Motivated by the property of point symmetry that clusters often exhibit, a PS-distance was proposed in [2] which was further modified in [3]. 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

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

where

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

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 $d_s(\overline{x}_j, \overline{c}) = d_c(\overline{x}_j, \overline{c}) = 0$.

### B. 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 [3] has two limitations.

**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 $\theta$ 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 $\theta$ 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 $\overline{c}_2$.

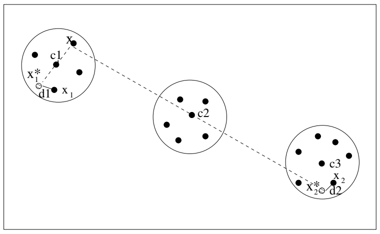

Fig. 1. Example where point symmetry distance proposed by Su and Chou fail

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.

### C. PS-index[3]

The cluster validity index, PS-index, based on the PS-distance defined above, is defined as

$$\begin{aligned} PS(K) &= \frac{1}{K} \sum_{i=1}^{K} \frac{1}{n_i} \sum_{j \in S_i} \frac{d_s(\overline{x}_j, \overline{c}_i) \times d_e(\overline{x}_j, \overline{c}_i)}{min_{m,n=1,\ldots,K \text{ and } m \neq n} d_e(\overline{c}_m, \overline{c}_n)} \\ &= \frac{1}{K} \sum_{i=1}^{K} \frac{1}{n_i} \sum_{j \in S_i} \frac{d_c(\overline{x}_j, \overline{c}_i)}{d_{min}} \end{aligned} \tag{3}$$

where $S_i$ is the set whose elements are the data points assigned to the $i$th 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 $d_c(\overline{x}_j, \overline{c}_i)$ is computed by Equation 1. The smallest $PS(K^*)$ indicates a valid optimal partition with the optimal cluster number $K^*$.

Since the point symmetry based distance $d_c$ [3] of Equation 1 has some inherent problems, hence the PS-index that is based on $d_c$, also suffers from similar drawbacks. In order to overcome this, a new point symmetry based distance is used to propose a new symmetry based cluster validity index.

## III. *Sym*-INDEX: THE PROPOSED SYMMETRY BASED CLUSTER VALIDITY INDEX

A newly developed symmetry based distance, $d_{ps}$, is first described in this section. A technique for reducing the computational complexity of $d_{ps}$ is mentioned. Finally a new cluster validity index is proposed that is based on $d_{ps}$. This is followed by an explanation of the interaction among the different components of the index so that it can indicate the proper partitioning of the data.

### A. 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 cluster center. It has been shown, in such cases the points are assigned to the farthest cluster. In order to overcome this limitation, we propose a new PS distance in this article which is called $d_{ps}(\overline{x}, \overline{c})$ 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 the first and second unique nearest neighbors of $\overline{x}^*$ be at Euclidean distances of $d_1$ and $d_2$ respectively. Then

$$d_{ps}(\overline{x}, \overline{c}) = \frac{(d_1 + d_2)}{2} \times d_e(\overline{x}, \overline{c}) \quad (4)$$

where $d_e(\overline{x}, \overline{c})$ is the Euclidean distance between the point $\overline{x}$ and $\overline{c}$. 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 based distances in [2] and [3], and the proposed point symmetry distance, $d_{ps}(\overline{x}, \overline{c})$, are as 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 [2] and [3], here the average distance between $\overline{x}^*$ and its first and the second unique nearest neighbors have been taken. Consequently the term, $(d_1 + d_2)/2$ will never be equal to 0, and the effect of $d_e(\overline{x}, \overline{c})$, the Euclidean distance, will always be considered. This will reduce the problems discussed in Figure 1.
2) Considering both $d_1$ and $d_2$ in the computation of $d_{ps}$ makes the PS-distance more robust and noise resistant. From an intuitive point of view, if both $d_1$ and $d_2$ of $\overline{x}$ with respect to $\overline{c}$ 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 4), no denominator term is incorporated to normalize it.

**Observation**: The proposed $d_{ps}$ measure will, in general, work well for symmetrical interclusters. Let the two nearest neighbors of the reflected point of $\overline{x}$ (in Figure 1) with respect to center $\overline{c}_1$ be 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, $d_{ps}(\overline{x}, \overline{c}_1) < d_{ps}(\overline{x}, \overline{c}_2) \implies \frac{d_1 + d_1^1}{2} \times d_{e1} < \frac{d_2 + d_2^1}{2} \times d_{e2} \implies \frac{d_{e1}}{d_{e2}} < \frac{d_2 + d_2^1}{d_1 + d_1^1}$. 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 $\frac{d_{e1}}{d_{e2}} < \frac{d_2 + d_2^1}{d_1 + d_1^1}$ is satisfied. Therefore the proposed distance satisfies EDD property and avoids merging of symmetrical interclusters. The experimental results provided in [5] also support the fact that the proposed measure is robust even in the presence of symmetrical interclusters since it obeys EDD property.

The computation of point symmetry based distance is highly complex. In order to compute the nearest neighbor distance of the reflected point of a particular data point with respect to a cluster center efficiently, we have used Kd-tree based nearest neighbor search. ANN (Approximate Nearest Neighbor), which is a library written in C++ [7], is used for this purpose. Here ANN is used to find $d_1$ and $d_2$ in Equation 4 efficiently. The Kd-tree structure can be constructed in $O(nlogn)$ time and takes $O(n)$ space.

### B. The Proposed Cluster Validity Measure

*1) Definition:* The newly developed PS distance is used to define a cluster validity function which measures the overall average symmetry with respect to the cluster centers. This is inspired by the $\mathcal{I}$-index developed in [6], i.e., it follows the definition of $\mathcal{I}$-index but the Euclidean distance replaced by the newly proposed point symmetry based distance.

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

$$Sym(K) = \left( \frac{1}{K} \times \frac{1}{\mathcal{E}_K} \times D_K \right), \quad (5)$$

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

*2) Explanation:* As formulated in Equation 5, *Sym*-index is a composition of three factors, $1/K$, $1/\mathcal{E}_K$ and $D_K$. The first factor increases as $K$ decreases; as *Sym*-index needs to be maximized for optimal clustering, this factor prefers to decrease the value of $K$. The second factor is a measure of the total within cluster symmetry. For clusters which have good symmetrical structures, $\mathcal{E}_K$ value is less. Note that as $K$ increases, in general, the clusters tend to become more symmetric. Moreover, as $d_e(\overline{x}, \overline{c})$ in Equation 4 also decreases, $\mathcal{E}_K$ decreases, resulting in an increase in the value of the *Sym*-index. Since *Sym*-index needs to be maximized, it will prefer to increase the value of $K$. Finally the third factor, $D_K$, measuring the maximum separation between a pair of clusters, increases with the value of $K$. Note that value of $D_K$ is bounded by the maximum separation between a pair of points in the data set. As these three factors are complementary in nature, so they are expected to compete and balance each other critically for determining the proper partitioning.

## IV. Experimental Results

Several artificially generated and real-life data sets were used to experimentally demonstrate that the *Sym*-index is not only able to find the proper cluster number for different types of data sets, but is also able to indicate the suitable clustering method. Due to lack of space results have been shown here only for two real-life and four artificially generated data sets. Three clustering algorithms viz., a newly developed point symmetry based genetic clustering technique (GAPS) [5], GAK-means algorithm [8] and the Average-linkage clustering algorithm [1] are used as the underlying partitioning techniques. The number of clusters, $K$ is varied from 2 to $\sqrt{n}$ for each algorithm, and the variation of the *Sym*-index is noted. Its maximum value indicates the appropriate algorithm and the appropriate number of clusters. Finally comparisons are made with two other recently developed cluster validity indices, i.e., a point symmetry based PS-index [3] and $\mathcal{I}$-index [6] in terms of the number of clusters and the clusterings obtained. The parameters of the genetic algorithms (GAPS and GAK-means) are as follows: population size is equal to 100, crossover and mutation probabilities are kept to be 0.8 and 0.01 respectively. The algorithms are executed for a maximum of 30 generations. Table I shows the optimum values of three validity indices, *Sym*-index, PS-index and $\mathcal{I}$-index and the number of clusters obtained after application of the three algorithms GAPS, GAK-means and Average Linkage on different data sets.

1) *Data1*: This data set contains 400 points distributed on two crossed ellipsoidal shells. The clustering result obtained after application of GAPS on this data set is shown in Figure 2(a). As expected, GAPS is able to detect the proper clustering since the data is symmetrical. The values of *Sym*-index and PS-index are the optimum for $K = 2$ (see Table I). $\mathcal{I}$-index could not identify the optimal clustering with any of the algorithms. Irrespective of the index used, GAK-means fails here since the clusters are non-convex. Again, Average linkage also fails here as the clusters have a little overlap.
2) *Data2*: This data set, consisting of 350 points, is a combination of ring-shaped, spherically compact and linear clusters. The clustering result obtained after application of GAPS on this data set is shown in Figure 2(b). As the clusters present here are symmetric, GAPS performs well for this data set. Again GAK-means and Average linkage are found to fail here. *Sym*-index is able to detect the proper clustering after application of GAPS with $K = 3$ (see Table I). $\mathcal{I}$ and PS both could not find proper clustering with any of the algorithms.
3) *Data3*: This data set consists of 250 points distributed over 5 spherically shaped highly overlapping clusters, each consisting of 50 points [9]. The clustering results obtained after application of GAK-means and GAPS on this data set are shown in Figure 3(a) and 3(b) respectively. Although, *Sym*-index is able to detect 5 clusters for all the three algorithms (see Table I), it attains the maximum value after application of GAK-means (Figure 3(a)). Its value for $K = 5$ with GAPS is poorer. Indeed, the clustering obtained here (Figure 3(b)) is not completely perfect. This again reveals the fact that the *Sym*-index is able to indicate the suitable clustering algorithm for a given data set. The Best value of PS-index corresponds to $K = 7$ with GAK-means. As this data set contains some symmetrical interclusters, PS-index should prefer the partition where some symmetrical interclusters are merged. But due to the denominator of its definition, it tries to maximize the minimal separation between two cluster centers. Thus optimal value of PS-index corresponds to $K = 7$ where some clusters are splitted rather than merged in order to maximize the minimal separation between any two cluster centers.
4) *Data4*: This data set contains 850 data points distributed over five clusters. The clustering result obtained after application of Average linkage on this data set is shown in Figure 4. As the clusters present here are symmetric and nonoverlapping, GAPS and Average linkage perform well. GAK-means fails here as all the clusters are not hyper-spherical in shape. *Sym*, $\mathcal{I}$ and PS indices are able to find the proper clustering with GAPS and Average linkage with $K = 5$ (see Table I). But PS-index attains its optimum value with GAK-means for $K = 7$.
5) 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

TABLE I

OPTIMUM VALUES OF THE *Sym*, PS AND $\mathcal{I}$ INDICES FOR ALL THE DATA SETS USING THE THREE ALGORITHMS (ENTRIES IN THE BRACKETS INDICATE THE NUMBER OF CLUSTERS CORRESPONDING TO WHICH THE INDEX GETS ITS OPTIMUM VALUE). THE BEST VALUE OF THE INDICES, ACROSS THE ALGORITHMS, ARE MENTIONED IN BOLD FACE.

| Data set | GAPS | | | GAK-means | | | Average linkage | | |
|---|---|---|---|---|---|---|---|---|---|
| | $Sym$ | PS | $\mathcal{I}$ | $Sym$ | PS | $\mathcal{I}$ | $Sym$ | PS | $\mathcal{I}$ |
| *Data1* | **0.049(2)** | **0.0024(2)** | 663.85(8) | 0.014(8) | 0.029(8) | **1101.25(5)** | 0.011(7) | 0.04(7) | 993.05(6) |
| *Data2* | **0.057(3)** | **0.018(6)** | 7.2(8) | 0.051(9) | 0.05(8) | **7.24(6)** | 0.019(4) | 0.04(4) | 3.27(5) |
| *Data3* | 0.012(5) | 0.037(6) | **1315.88(6)** | **0.014(5)** | **0.031(7)** | 1276.29(5) | 0.013(5) | 0.039(5) | 1240.83(4) |
| *Data4* | **0.0076(5)** | 0.022(5) | **12095.52(5)** | 0.004(7) | **0.015(7)** | 10259.18(8) | **0.0076(5)** | 0.022(5) | **12095.52(5)** |
| Iris | **0.049(3)** | **0.084(7)** | **691.29(3)** | 0.046(4) | 0.107(4) | 633.82(3) | 0.046(3) | 0.088(2) | 653.95(3) |
| Cancer | **0.00052(2)** | **0.125(2)** | 27662.41(2) | 0.0005(2) | 0.131(5) | **28055.57(3)** | 0.00048(2) | 0.093(3) | 27058.82(3) |

characterized by four feature values [10]. For this data set we have calculated the *Minkowski Score* (MS) [11] of the clustering result obtained after application of all three algorithms with $K = 3$ since it was not possible to demonstrate the clustering results for this 4-d data set pictorially. Smaller value of MS means better clustering. The MS scores are $0.58$, $0.61$ and $0.62$ for GAPS, GAK-means and Average linkage respectively. From the obtained MS values it is clear that GAPS is able to find the best clustering among the three algorithms and for this particular partitioning, *Sym* and $\mathcal{I}$ indices obtained their best values. PS-index is unable to find proper clustering with any algorithms.

6) Cancer: Here we use the Wisconsin Breast cancer data set consisting of 683 sample points. Each pattern has nine features. There are two categories in the data: malignant and benign. The two classes are known to be linearly separable. For this data set also we have calculated the *Minkowski Score* (MS) [11] of the clustering result obtained after application of all three algorithms with $K = 2$ since it is of 9-dimensional. The MS scores are $0.368$, $0.368$ and $0.445080$ for GAPS, GAK-means and Average linkage respectively. From the obtained MS values it is clear that GAPS and GAK-means perform almost similarly. For the partitioning obtained by GAPS for $K = 2$ *Sym*-index obtained its best value. $\mathcal{I}$-index obtained its best value with GAK-means for $K = 3$. Optimum value of PS-index indicates Average Linkage as the proper clustering algorithm where as 3 as the proper cluster number.

Interestingly, it was observed that for all the data sets, *Sym*-index was able to detect the proper number of clusters as well as the suitable clustering algorithm. For example, for *Data2* where GAPS should perform the best for $K = 3$, the value of *Sym*-index is the maximum for GAPS (0.57) as compared to those for GAK-means (0.51) and Average linkage (0.019) thereby indicating the suitable clustering technique. Again, for *Data3*, where GAK-means should perform the best for $K = 5$, the value of the *Sym*-index is the maximum (see Table I) for this case. The other indices are sometimes misled in this regard. For example, for *Data3*, $\mathcal{I}$ value is the more for GAPS with 6 clusters, (=1315.88) as compared to GAK-means with 5 clusters (=1276.29). Again, for *Data4*, GAPS with $K = 5$ should be the appropriate choice (one that is correctly indicated by *Sym*-index), PS-index attains its minimum value for GAK-means with $K = 7$. These results, therefore, point at the significant superiority of the proposed index.

## V. Application to Image Segmentation

The newly proposed cluster validity index along with the two other indices are used in conjunction with GAK-means [8] clustering algorithm for segmenting remote sensing satellite images of parts of the Mumbai.

The IRS image of Mumbai was obtained using the LISS-II sensor. It is available in four bands, viz., blue, green, red and near infra-red. Fig. 5(a) shows the *IRS* image of a part of Mumbai in the near infra red band. As can be seen, the elongated city area is surrounded on three sides by the Arabian sea. Towards the bottom right of the image, there are several islands, including the well known *Elephanta islands*. The dockyard is situated on the south eastern part of Mumbai, which can be seen as a set of three finger like structure.

After application of GAK-means algorithm on this image, *Sym*-index gets its optimal value for $K = 6$ where as PS-index and $\mathcal{I}$-index get their optimum values for $K = 3$ and $K = 4$ respectively. The partitionings corresponding to the optimum values of PS-index, *Sym*-index and $\mathcal{I}$-index are shown in Figures 5(b), 6(a) and 6(b) respectively. It can be seen from the figures that partition corresponding to optimum value of *Sym*-index is able to differenciate more regions much better than that of the partition corresponding to two other indices. Interestingly, in this case the bridge connecting Mumbai to the mainland has also been identified reasonably well (Figure 6(a)), while this is missed in the other two (Figures 5(b) and 6(b)).

## VI. Conclusion

A new symmetry based cluster validity index is proposed in this article that is able to indicate both the appropriate number of clusters as well as the appropriate partitioning. Its effectiveness is demonstrated for four artificially generated data sets, two real life data sets and for also one remote sensing image where determining the different types of landcovers is of great importance. GAPS, a newly proposed symmetry distance based genetic clustering technique, GAK-means and Average Linkage algorithms are used as the underlying partitioning methods. The experimental results establish the superiority of the newly proposed *Sym*-index in appropriately determining the number of clusters as well as to indicate the appropriate clustering technique, as compared to two other recently developed validity indices, PS-index and $\mathcal{I}$-index. As a part of future work, the effectiveness of the proposed index needs to be studied extensively with more data sets and algorithms. In the present study, some GA based clustering algorithms are used for the comparison purpose. Performance of some simple greedy search or other stochastic optimization such as simulated annealing will also be provided as a comparison in future. The authors are currently working in this direction.

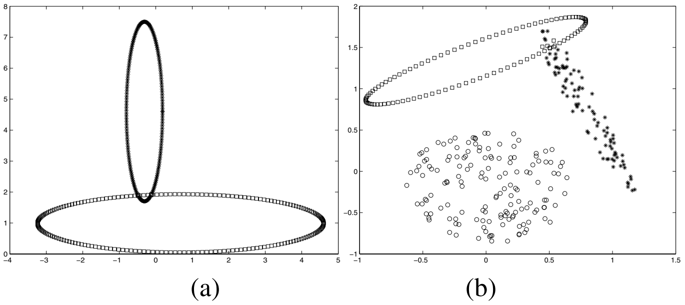

Fig. 2. Clustering obtained by GAPS on (a) *Data1* for $K = 2$ (b) *Data2* for $K = 3$

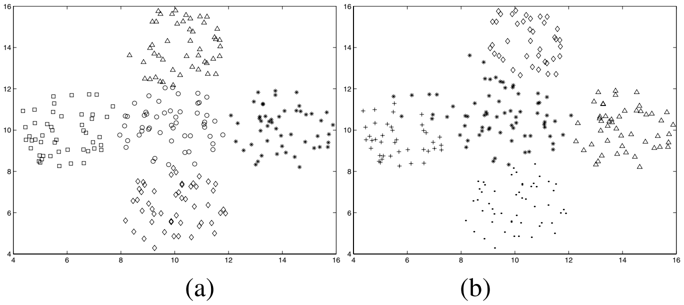

Fig. 3. Clustering on *Data3* (a) obtained by GAKmeans for $K = 5$ (b) obtained by GAPS-clustering for $K = 5$

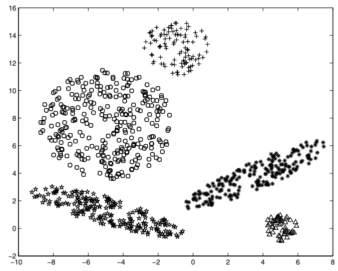

Fig. 4. Clustering obtained by Average linkage on *Data4* for $K = 5$

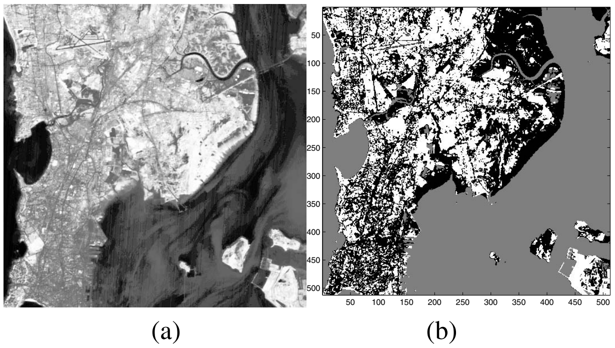

Fig. 5. (a) IRS image of Mumbai in the NIR band with histogram equalization (b) Clustered image of Mumbai corresponding to optimal value of PS-index attained for K=3

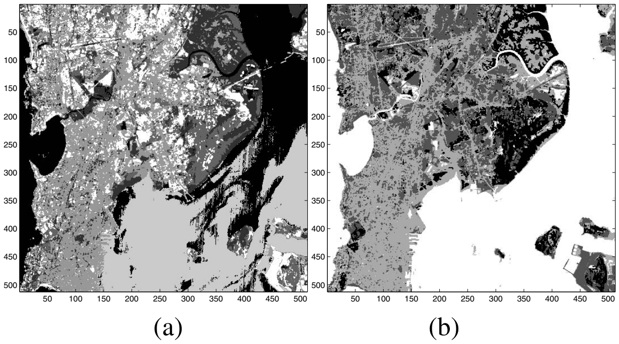

Fig. 6. Clustered image of Mumbai corresponding to optimal value of (a) *Sym*-index attained for K=6 (b) $\mathcal{I}$-index attained for K=4

## References

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

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

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

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

[9] S. Bandyopadhyay and U. Maulik, “Genetic clustering for automatic evolution of clusters and application to image classification,” *Pattern Recognition*, no. 2, pp. 1197–1208, 2002.

[10] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 3, pp. 179–188, 1936.

[11] 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.