Application of a New Symmetry-Based Cluster Validity Index for Satellite Image Segmentation

Sriparna Saha and Sanghamitra Bandyopadhyay

***Abstract*—An important approach for image segmentation is clustering pixels based on their spectral properties. In particular, satellite images contain land cover types, some of which cover significantly large areas, while some (e.g., bridges and roads) occupy relatively much smaller regions. Automatically detecting regions or clusters of such widely varying sizes presents a challenging task. In this letter, a symmetry-based cluster validity index, named Sym-index (Symmetry distance-based index), is proposed. It is able to correctly indicate the presence of clusters of different sizes as long as they are internally symmetrical. A genetic-algorithm-based clustering technique that optimizes the Sym-index is used for image segmentation where the number of clusters is determined automatically. The superiority of the proposed index, as compared to other indices, is established for automatically segmenting the land cover types from SPOT and Indian Remote Sensing satellite images of two different cities in India.**

***Index Terms*—Cluster validity index, Kd tree, point-symmetry (PS)-based distance, remote sensing imagery.**

I. Introduction

An IMPORTANT task in remote sensing applications is the classification of pixels in the images into homogeneous regions, each of which corresponds to some particular land cover type. This problem has often been modeled as a segmentation problem [1], and clustering methods have been used to solve it. However, since it is difficult to have *a priori* information about the number of clusters in satellite images, the clustering algorithms should be able to automatically determine this value. Moreover, in satellite images it is often the case that some regions occupy only a few pixels, while the neighboring regions are significantly large. Thus, automatically detecting regions or clusters of such widely varying sizes presents a challenge in designing segmentation algorithms.

One of the basic features of shapes and objects is symmetry [2]. Based on this observation, a new cluster validity index based on point symmetry (PS), Sym-index, is proposed in this letter, which is able to detect clusters of any shape and size as long as they possess symmetry. A PS-based measure of similarity called d_{ps} (which was developed in [3] to overcome the limitations of some existing PS-distances proposed in [2] and [4]) and a genetic PS-based clustering technique (GAPS) [3] are used for this purpose. The number of clusters K is manually varied from $K_{\max}$ to $K_{\max}$, and for each K, Sym-index is computed for the partitioning resulting from the application of GAPS-clustering. As a result, a total of $(K_{\max} - K_{\min} + 1)$ partitions will be generated denoted by $U^*_{K_{\min}}, U^*_{K_{\min}+1}, \ldots, U^*_{K_{\max}}$. Let the corresponding Sym-index values be $\mathrm{Sym}_{K_{\min}}, \mathrm{Sym}_{K_{\min}+1}, \ldots, \mathrm{Sym}_{K_{\max}}$. Let $K^* = \mathrm{argmax}_{i=K_{\min},\ldots,K_{\max}}[\mathrm{Sym}_i]$. Therefore, according to the Sym-index, K^* is the optimal number of clusters present in the data. The tuple $\langle U^*_{K^*}, K^* \rangle$ is presented as a solution to the segmentation problem.

The effectiveness of the newly proposed cluster validity index, Sym-index, in conjunction with the GAPS-clustering [3] for automatically detecting different types of regions is demonstrated on one simulated and two satellite images. The segmentation results are compared with those obtained by two other recently proposed cluster validity indices, namely, PS-index [4] and $\mathcal{I}$-index [5], and another well-known XB index [6].

II. Sym-Index: The Proposed Symmetry-Based Cluster Validity Index

A. *Definition of the PS Distance*

In this section, the PS distance [3], $d_{\mathrm{ps}}(\overline{x}, \overline{c})$, associated with point $\overline{x}$ with respect to a center $\overline{c}$ is described. As shown in [3], $d_{\mathrm{ps}}(\overline{x}, \overline{c})$ is able to overcome some serious limitations of an earlier PS distance [2]. Let a point be $\overline{x}$. The symmetrical (reflected) point of $\overline{x}$ with respect to a particular center $\overline{c}$ 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, $i = 1, 2, \ldots, \mathrm{knear}$. Then

$$d_{\mathrm{ps}}(\overline{x}, \overline{c}) = d_{\mathrm{sym}}(\overline{x}, \overline{c}) \times d_{\mathrm{e}}(\overline{x}, \overline{c}) \tag{1}$$

$$= \frac{\sum_{i=1}^{\mathrm{knear}} d_i}{\mathrm{knear}} \times d_{\mathrm{e}}(\overline{x}, \overline{c}) \tag{2}$$

where $d_{\mathrm{e}}(\overline{x}, \overline{c})$ is the Euclidean distance between the point $\overline{x}$ and $\overline{c}$. It can be seen from (2) that knear cannot be chosen equal to 1, since if $\overline{x'}$ exists in the data set then $d_{\mathrm{ps}}(\overline{x}, \overline{c}) = 0$ 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 = 2.

Note that $d_{\mathrm{ps}}(\overline{x}, \overline{c})$, which is a nonmetric, 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 complexity of computing $d_{\mathrm{ps}}(\overline{x}, \overline{c})$ is of order n, where n is the total number of data points. For all the n points and K clusters, the complexity becomes of order $n^2 K$. Thus, to reduce this, we have used Kd-tree based nearest neighbor search approximate nearest neighbor library (ANNlib), which is a library written in C++ (obtained from http://www.cs.umd.edu/~mount/ANN). Here, ANNlib is used to find d_i, $i = 1$ to knear, in (2) efficiently.

Manuscript received December 1, 2006; revised August 7, 2007.

The authors are with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India (e-mail: sriparna_r@isical.ac.in; sanghami@isical.ac.in).

Digital Object Identifier 10.1109/LGRS.2008.915595

B. Newly Proposed Cluster Validity Measure

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

$$\mathrm{Sym}(K) = \frac{1}{K} \times \frac{1}{\mathcal{E}_K} \times D_K. \tag{3}$$

Here, $\mathcal{E}_K = \sum_{k=1}^{K} E_k$, such that $E_k = \sum_{j=1}^{n_k} d^*_{\mathrm{ps}}(\overline{x}_{kj}, \overline{c}_k)$, and $D_K = \max_{i,j=1}^{K} \|\overline{c}_i - \overline{c}_j\|$. D_K is the maximum Euclidean distance between two cluster centers among all centers. $d^*_{\mathrm{ps}}(\overline{x_j}, \overline{c_i})$ is computed by (2) with some constraint. Here, first knear nearest neighbors of $\overline{x'}_j = 2 \times \overline{c}_i - \overline{x}_j$ 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 $\overline{c}_i$ 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. It may be mentioned that Sym-index is inspired by the $\mathcal{I}$-index developed in [5].

2) Explanation: As formulated in (3), 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 $d_{\mathrm{e}}(\overline{x}, \overline{c})$ in (2) also decreases, E_k, in general, decreases, resulting in an increase in the value of Sym-index. This will continue until the clusters do not get the symmetric shape. But after that as K increases, $\mathcal{E}_K$ will increase, because its total number of components will increase. Finally the third factor, D_K, measuring the maximum separation between a pair of clusters, increases with the value of K. Note that the 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 partition.

C. Mathematical Justification

In this section, we mathematically justify the new validity index by establishing its relationship to the well-known validity measure proposed by Dunn [7] for hard partitions. This is inspired by a proof of optimality of the Xie–Beni index [6].

1) Uniqueness and Global Optimality of the K-Partition: The separation index D_1 is a hard K-partition cluster validity criterion. It is known that if $D_1 > 1$, unique, compact, and separated hard clusters have been found [7]. Here, we will prove that if the optimal solution D_1 becomes sufficiently large, the optimal validity function Sym will also be large, which means that a unique K-partition has been found. The proof of this is as follows.

2) Theorem 1: For any $K = 2, \ldots, n-1$, let Sym be the overall Sym-index value of any hard partition, and D_1 be the separation index of the corresponding partition. Then we have

$$\mathrm{Sym} \geq \frac{D_1}{n \times K \times 0.5 \times \mathrm{knear} \times d_{\mathrm{NN}}^{\max}}$$

where n is the total number of data points, K is the total number of clusters and knear is the number of nearest neighbors considered while computing d_{ps} as defined in (2). $d_{\mathrm{NN}}^{\max}$ is the maximum nearest neighbor distance in the data set. That is $d_{\mathrm{NN}}^{\max} = \max_{i=1,\ldots,n} d_{\mathrm{NN}}(\overline{x}_i)$, where $d_{\mathrm{NN}}(\overline{x}_i)$ is the nearest neighbor distance of $\overline{x}_i$.

Proof: Let the hard K-partition be an optimal partition of the data set $X = \{\overline{x}_j; j = 1, 2, \ldots, n\}$ with $c_i (i = 1, 2, \ldots, K)$ being the centroids of each class u_i. The total symmetrical variation $\mathcal{E}_K$ of the optimal hard K-partition is defined in (3). Thus

$$\mathcal{E}_K = \sum_{i=1}^{K} \sum_{\overline{x}_j \in u_i} d_{\mathrm{ps}}(\overline{x}_j, \overline{c}_i) \tag{4}$$

$$= \sum_{i=1}^{K} \sum_{\overline{x}_j \in u_i} \frac{\sum_{ii=1}^{\mathrm{knear}} d_{ii}}{\mathrm{knear}} d_{\mathrm{e}}(\overline{x}_j, \overline{c}_i). \tag{5}$$

Assuming that $\overline{x}_j^*$ (the symmetrical point of $\overline{x}_j$ with respect to cluster center $\overline{c}_i$) lies within the data space, it may be noted that $d_1 \leq d_{\mathrm{NN}}^{\max}/2$, $d_2 \leq 3d_{\mathrm{NN}}^{\max}/2, \ldots, d_i \leq (2i-1)d_{\mathrm{NN}}^{\max}/2$, where d_i is the distance of ith nearest neighbor of $\overline{x}_j^*$. Considering the term $\sum_{ii=1}^{\mathrm{knear}} d_{ii}/\mathrm{knear}$, we can write

$$\frac{\sum_{ii=1}^{\mathrm{knear}} d_{ii}}{\mathrm{knear}} \leq \frac{d_{\mathrm{NN}}^{\max}}{2\mathrm{knear}} \left(\sum_{ii=1}^{\mathrm{knear}} (2 \times ii - 1) \right). \tag{6}$$

The right-hand side of the inequality may be written as

$$\frac{d_{\mathrm{NN}}^{\max}}{2\mathrm{knear}} \times \frac{(\mathrm{knear} \times (2 + (\mathrm{knear} - 1)2))}{2} = \frac{\mathrm{knear} \times d_{\mathrm{NN}}^{\max}}{2}. \tag{7}$$

So, combining (5), (6), and (7), we can write

$$\mathcal{E}_K \leq \sum_{i=1}^{K} \sum_{\overline{x}_j \in u_i} 0.5 \times \mathrm{knear} \times d_{\mathrm{NN}}^{\max} \times d_{\mathrm{e}}(\overline{x}_j, \overline{c}_i)$$

$$\leq 0.5 \times \mathrm{knear} \times d_{\mathrm{NN}}^{\max} \sum_{i=1}^{K} \sum_{x_j \in u_i} d_{\mathrm{e}}(\overline{x}_j, \overline{c}_i).$$

Suppose that the centroid $\overline{c}_i$ is inside the boundary of cluster i, for $i = 1$ to K. Then $d_{\mathrm{e}}(\overline{x}_j, \overline{c}_i) \leq \mathrm{dia}(u_i)$, for $\overline{x}_j \in u_i$ where $\mathrm{dia}(u_i) = \max_{\overline{x}_k, \overline{x}_j \in u_i} d_{\mathrm{e}}(\overline{x}_k, \overline{x}_j)$. We thus have

$$\mathcal{E}_K \leq 0.5 \times \mathrm{knear} \times d_{\mathrm{NN}}^{\max} \sum_{i=1}^{K} \sum_{\overline{x}_j \in u_i} \mathrm{dia}(u_i)$$

$$\leq 0.5 \times \mathrm{knear} \times d_{\mathrm{NN}}^{\max} \sum_{i=1}^{K} n_i \mathrm{dia}(u_i)$$

$$\leq 0.5 \times \mathrm{knear} \times d_{\mathrm{NN}}^{\max} \times n \times \max_i \mathrm{dia}(u_i).$$

Here, n_i denotes the total number of data points in cluster i. So, $1/\mathcal{E}_K \geq 1/0.5 \times \mathrm{knear} \times d_{\mathrm{NN}}^{\max} \times n \times \max_i \mathrm{dia}(u_i)$. We also have that $\min_{i,j,i \neq j} \operatorname{dis}(u_i, u_j) \leq D_K$ where $\operatorname{dis}(u_i, u_j) = \min_{\overline{x}_i \in u_i, \overline{x}_j \in u_j} d_{\mathrm{e}}(\overline{x}_i, \overline{x}_j)$ and $D_K = \max_{i,j=1}^{K} d_{\mathrm{e}}(\overline{c}_i, \overline{c}_j)$. Thus

$$\begin{aligned}\operatorname{Sym}(K) &= \frac{D_K}{K \times \mathcal{E}_K} \\ &\geq \frac{\min_{i,j} \operatorname{dis}(u_i, u_j)}{K \times 0.5 \times \text{knear} \times d_{\mathrm{NN}}^{\max} \times n \times \max_i \operatorname{dia}(u_i)}\end{aligned}$$

i.e.,

$$\operatorname{Sym}(K) = \frac{\min_{1 \leq i \leq K-1} \left\{ \min_{i+1 \leq j \leq K} \frac{\operatorname{dis}(u_i, u_j)}{\max_{1 \leq k \leq K} \operatorname{dia}(u_k)} \right\}}{K \times 0.5 \times \text{knear} \times d_{\mathrm{NN}}^{\max} \times n}. \tag{8}$$

The separation index D_1 (Dunn [7]) is defined as

$$D_1 = \min_{1 \leq i \leq K-1} \left\{ \min_{i+1 \leq j \leq K} \left\{ \frac{\operatorname{dis}(u_i, u_j)}{\max_{1 \leq k \leq K} \operatorname{dia}(u_k)} \right\} \right\}. \tag{9}$$

So, combining (8) and (9), we get

$$\operatorname{Sym}(K) \geq \frac{D_1}{K \times 0.5 \times \text{knear} \times d_{\mathrm{NN}}^{\max} \times n}.$$

Since the denominator is constant for a given K, Sym increases as D_1 grows without bound. As mentioned earlier, it has been proved by Dunn [7] that if $D_1 > 1$, the hard K-partition is unique. Thus, if the data set has a distinct substructure and the partitioning algorithm has found it, then the corresponding Sym-index value will be the maximum.

III. GAPS: Segmentation Algorithm Used

A genetic algorithm with point symmetry distance based clustering technique, GAPS, proposed in [3], is used as the underlying segmentation method. A brief overview of the basic steps of GAPS, which closely follow those of the conventional genetic algorithm (GA), are enumerated below. Note that details of GAPS are available in [3]. Given a particular value of number of clusters K, GAPS partitions the data in K symmetrical shaped clusters.

A. Initialization

The centers of the clusters are encoded in the fixed-length chromosome as a series of real numbers that correspond to each dimension per cluster. Thus, for a d-dimensional data set and for K number of clusters, the length of each chromosome is $K \times d$. The K cluster centers encoded in each chromosome are initialized to K randomly chosen points from the data set. Thereafter, five iterations of the K-means algorithm is executed to make the centers separated initially. Although five iterations do not guarantee that K-means will converge, the aim here was to select the initial set of cluster centers in a better way than just randomly initializing it. The GA component of GAPS in any case finally optimizes the centers in a better way than K-means.

B. Assignment of Points

Here, a point $\overline{x}_i$, $1 \leq i \leq n$, is assigned to cluster k if and only if $d_{\mathrm{ps}}(\overline{x}_i, \overline{c}_k) \leq d_{\mathrm{ps}}(\overline{x}_i, \overline{c}_j)$, $j = 1, \ldots, K$, $j \neq k$ and $d_{\mathrm{sym}}(\overline{x}_i, \overline{c}_k) \leq \theta$. For $d_{\mathrm{sym}}(\overline{x}_i, \overline{c}_k) > \theta$, point $\overline{x}_i$ is assigned to some cluster m if and only if $d_{\mathrm{e}}(\overline{x}_i, \overline{c}_m) \leq d_{\mathrm{e}}(\overline{x}_i, \overline{c}_j)$, $j = 1, 2 \ldots K$, $j \neq m$. The value of θ is kept equal to the maximum nearest neighbor distance among all the points in the data set, as here knear $= 2$.

C. Update of Centers

After the assignments are done, the cluster centers encoded in the chromosome are replaced by the mean points of the respective clusters.

D. Fitness Computation

For each chromosome, clustering metric, M, is calculated as defined below

$$M = \sum_{k=1}^{K} \sum_{\overline{x}_i \in k\text{th cluster}} d_{\mathrm{ps}}(\overline{x}_i, \overline{c}_k).$$

It is evident that smaller values of M, the objective function, correspond to partitions having symmetrical shaped clusters. The fitness function, fit, of a chromosome is defined as: fit $= 1/M$. So minimization of M means maximization of function, fit.

E. Genetic Operators Used

Roulette wheel selection [8] is used to implement the proportional selection strategy. The normal single-point crossover operation with crossover probability selected adaptively as in [9] is used here. The expressions for crossover probabilities are given below. Let $f_{\max}$ be the maximum fitness value of the current population, $\overline{f}$ be the average fitness value of the population and f' be the larger of the fitness values of the solutions to be crossed. Then the probability of crossover, μ_{c}, is calculated as

$$\begin{aligned}\mu_{\mathrm{c}} &= k_1 \times \frac{(f_{\max} - f')}{(f_{\max} - \overline{f})}, && \text{if } f' > \overline{f} \\ \mu_{\mathrm{c}} &= k_3, && \text{if } f' \leq \overline{f}.\end{aligned}$$

Here, as in [9], the values of k_1 and k_3 are kept equal to 1.0.

Each chromosome undergoes mutation with a probability μ_{m}, selected adaptively as like [9]. The expression for mutation probability, μ_{m}, is given below

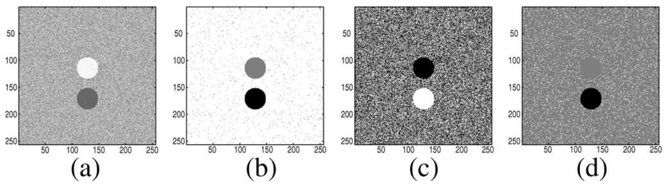
$$\begin{aligned}\mu_{\mathrm{m}} &= k_2 \times \frac{(f_{\max} - f)}{(f_{\max} - \overline{f})}, && \text{if } f > \overline{f} \\ \mu_{\mathrm{m}} &= k_4, && \text{if } f \leq \overline{f}.\end{aligned}$$

Here, values of k_2 and k_4 are kept equal to 0.5. The reason behind such type of adaptation is available in [3] as well as in [9].

F. Termination

Steps 2–5 are executed for a maximum number of generations. The best string seen up to the last generation provides the solution to the segmentation problem.

The parameters of the GAPS-clustering are as follows: population size $= 20$, number of generations $= 20$.



(a) (b) (c) (d)

Fig. 1. (a) SCI. (b) Segmented SCI obtained by GAPS-clustering with Sym-index (provides $K^* = 3$). (c) Segmented SCI obtained by K-means-clustering for $K = 3$. (d) Segmented SCI obtained by EM-clustering for $K = 3$.

IV. Application to Image Segmentation

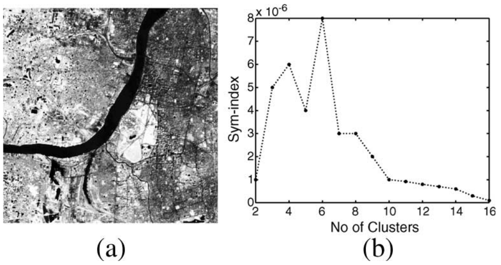
As mentioned earlier, Sym-index is used in conjunction with GAPS-clustering to segment an image in the intensity space. Three other indices, namely PS-index [4], $\mathcal{I}$-index [5], and XB-index [6] are considered for comparison. One artificially generated image and two remote sensing satellite images of parts of the cities of Kolkata and Mumbai are used. For each image, K, is varied from 2 to 16.

A. SCI

To show the effectiveness of the proposed Sym-index in identifying small clusters from much larger ones where there is a significant overlap of the small clusters with the bigger one, we first generate an artificial image of size 256 × 256 shown in Fig. 1(a). There are two small circles of radius 20 each, centered at (113, 128) and (170, 128), respectively. The pixels of these two small circles take gray values randomly in the range [160–170] and [65–75], respectively. The background pixels take values randomly in the range [70–166]. Here also K is varied from 2 to 16. Fig. 1(b) shows the segmented image using GAPS-clustering with Sym-index, when three clusters were automatically found. We have calculated Minkowski score (MS) [10] of the segmented image provided by the Sym-index. Smaller value of MS indicates better segmentation. The corresponding MS value is 0.177026. In contrast, PS-index, $\mathcal{I}$-index and XB-index attained their optimum values for $K^* = 9$, $K^* = 5$, and $K^* = 9$, respectively, i.e., they are not at all able to detect the proper number of clusters. K-means (with $K = 3$) is not able to find out the proper clustering from this data set [shown in Fig. 1(c)]. MS value in this case is 0.806444. The expectation–maximization (EM) algorithm is also not able to find out the proper clustering from this overlapping data set [Fig. 1(d)]. MS value in this case is 0.82.

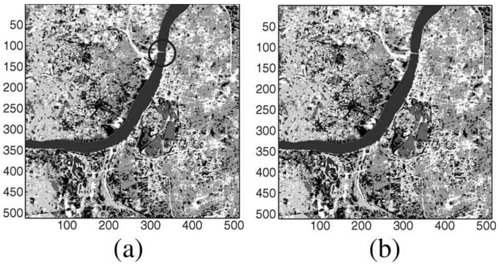
B. SPOT Image of Kolkata

Fig. 2(a) shows the 512 × 512 Satellite Pour l'Observation de la Terre (SPOT) [11] image of a part of the city of Kolkata (available in three spectral bands) in the near-infrared (NIR) band. GAPS-clustering is applied on this image data set while varying the number of clusters K from 2 to 16. For each obtained partitioning, the values of four cluster validity indices (Sym-index, PS-index, $\mathcal{I}$-index, and XB-index) are calculated. Sym-index obtained its optimal value for $K^* = 6$. The corresponding segmented image is shown in Fig. 3(a). Similarly, the $\mathcal{I}$-index, PS-index, and XB-index obtained their optimum values for $K^* = 8$, $K^* = 3$, and $K^* = 2$, respectively, and the corresponding segmented images are shown in Figs. 3(b), 4(a), and 4(b), respectively. The segmentations corresponding to the optimum values of Sym-index and $\mathcal{I}$-index are able to separate almost all the regions equally well [Fig. 3(a) and (b)]. Even the thin outline of the bridge on the river has been automatically identified [encircled in Fig. 3(a)]. This again illustrates the superiority of symmetry-based distance for detecting a small cluster. To validate the results, 932 pixel positions were manually selected from seven different land cover types which were labeled accordingly. For these points the MS [10] is calculated after application of GAPS-clustering for the optimal cluster number indicated by each of the indices. Smaller value of MS indicates better segmentation. The MS scores corresponding to Sym-index, $\mathcal{I}$-index, PS-index, and XB-index are 0.865, 0.8799, 1.36921, and 1.4319, respectively, again demonstrating the superior result obtained by GAPS-clustering in conjunction with the Sym-index. PS-index and XB-index perform poorly for this image. GAPS-clustering is also applied on these selected 932 points with $K = 7$. The class wise accuracies are 1, 0.83, 0.87, 0.82, 0.86, 0.86, and 0.88, respectively. The overall accuracy is 0.87. For segmented SPOT Kolkata image, Davies–Bouldin (DB) index [12] has been calculated corresponding to the optimal values of Sym-index, $\mathcal{I}$-index, PS-index, and XB-index. The values are listed



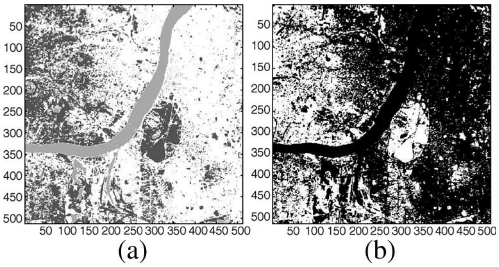
(a) (b)

Fig. 2. (a) SPOT image of Kolkata in the NIR band with histogram equalization. (b) Variation of Sym-index with number of clusters for Kolkata image using GAPS.



(a) (b)

Fig. 3. Segmented Kolkata image obtained by GAPS-clustering with (a) Sym-index (provides $K^* = 6$), (b) $\mathcal{I}$-index (provides $K^* = 8$).

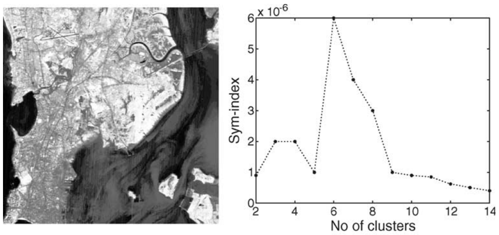


(a) (b)

Fig. 4. Segmented Kolkata image obtained by GAPS-clustering with (a) PS-index (provides $K^* = 3$) and (b) XB-index (provides $K^* = 2$).

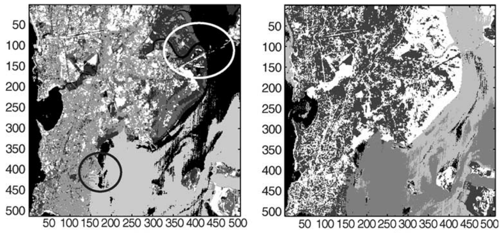
TABLE I
DB-INDEX VALUES OF THE SEGMENTED KOLKATA AND MUMBAI SATELLITE IMAGES CORRESPONDING TO THE OPTIMAL VALUES OF FOUR CLUSTER VALIDITY INDICES

Validity index	SPOT image of Kolkata	IRS image of Mumbai
Sym-index	0.669	1.586
$\mathcal{I}$ index	0.775	1.979
PS-index	0.800	1.586
XB-index	0.724	5.196



(a) (b)

Fig. 5. (a) IRS image of Mumbai in the NIR band with histogram equalization. (b) Variation of Sym-index with number of clusters for IRS image of Mumbai using GAPS.



(a) (b)

Fig. 6. Segmented Mumbai image obtained by GAPS-clustering with (a) Sym-index/PS-index (provides $K^* = 6$) and (b) $\mathcal{I}$-index (provides $K^* = 5$).

in Table I. As smaller values of DB are preferable, it signifies that segmentation corresponding to Sym-index is the best. Fig. 2(b) shows the variation of Sym-index with the number of clusters for this data set.

C. IRS Image of Mumbai

The Indian Remote Sensing (IRS) image of Mumbai was obtained using the Linear Imaging Self-Scanning System II sensor, available in four bands, viz., blue, green, red, and NIR. Fig. 5(a) shows the IRS image of a part of Mumbai in the NIR band. Here also, the number of clusters (K) is varied from 2 to 16. GAPS-clustering with Sym-index and PS-index get their optimal values for $K^* = 6$ whereas GAPS-clustering with $\mathcal{I}$-index and XB-index get their optimum values for $K^* = 5$ and $K^* = 3$, respectively. The segmented images corresponding to the optimum values of Sym-index and $\mathcal{I}$-index are shown in Fig. 6(a) and (b), respectively. It can be seen from the figures that the results using Sym-index and PS-index are the same (since GAPS-clustering provides the same partitioning for $K^* = 6$ in both the cases). The water (Arabian Sea) surrounding Mumbai gets differentiated into two distinct regions, based on the difference in their spectral properties. The other landmarks, e.g., the river above the bridge (north) and dockyard (south) [encircled in Fig. 6(a)], are well detected. In the segmentation obtained by GAPS-clustering using $\mathcal{I}$-index, some landmarks, e.g., the river just above the bridge on its left end, are not so well delineated. The DB index [12] values corresponding to the segmented images of Mumbai given by the optimal values of Sym-index/PS-index, $\mathcal{I}$-index, and XB-index are listed in Table I. As smaller values of DB indicates better clustering, the result signifies that the segmentation corresponding to Sym-index is the best. Fig. 5(b) shows the variation of Sym-index with the number of clusters for this data set.

V. DISCUSSION AND CONCLUSION

In this letter, the application of the proposed symmetry-based cluster validity index and GAPS-clustering technique is described for image segmentation. Its effectiveness vis-a-vis other well-known validity indices is first established for segmenting one artificially generated image. Thereafter, it is used for classifying different land covers in two multispectral satellite images. The choice of the underlying clustering technique is important. Although the Sym-index has the capability of indicating the proper symmetric clusters, the underlying clustering technique should be able to first detect them. For example, both well-known K-means and EM clustering algorithms are unable to find out the proper clustering from the data sets like the synthetic image. In contrast, GAPS-clustering [3] is able to tackle such situations as is evident from its consistently good performance. Effectiveness of other clustering techniques can also be investigated in the future. A detailed sensitivity study of different parameters of GAPS constitutes an important direction of future research.

REFERENCES

[1] U. Maulik and S. Bandyopadhyay, “Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 5, pp. 1075–1081, May 2003.

[2] M.-C. Su and C.-H. Chou, “A modified version of the K-means algorithm with a distance based on cluster symmetry,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 674–680, Jun. 2001.

[3] S. Bandyopadhyay and S. Saha, “GAPS: A clustering method using a new point symmetry based distance measure,” *Pattern Recognit.*, vol. 40, no. 12, pp. 3430–3451, Dec. 2007.

[4] C. H. Chou, M. C. Su, and E. Lai, “Symmetry as a new measure for cluster validity,” in *Proc. 2nd WSEAS Int. Conf. Sci. Comput. Soft Comput.*, Crete, Greece, 2002, pp. 209–213.

[5] U. Maulik and S. Bandyopadhyay, “Performance evaluation of some clustering algorithms and validity indices,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.

[6] X. L. Xie and G. Beni, “A validity measure for fuzzy clustering,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 841–847, Aug. 1991.

[7] J. C. Dunn, “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters,” *J. Cybern.*, vol. 3, no. 3, pp. 32–57, 1973.

[8] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. New York: Addison-Wesley, 1989.

[9] M. Srinivas and L. Patnaik, “Adaptive probabilities of crossover and mutation in genetic algorithms,” *IEEE Trans. Syst., Man, Cybern.*, vol. 24, no. 4, pp. 656–667, Apr. 1994.

[10] A. Ben-Hur and I. Guyon, *Detecting Stable Clusters Using Principal Component Analysis in Methods in Molecular Biology*, M. Brownstein and A. Kohodursky, Eds. Totowa, NJ: Humana, 2003, pp. 159–182.

[11] J. A. Richards, *Remote Sensing Digital Image Analysis: An Introduction*. New York: Springer-Verlag, 1993.

[12] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.