Unsupervised Pixel Classification in Satellite Imagery: A Two-stage Fuzzy Clustering Approach

Article *in* Fundamenta Informaticae · January 2008

Source: DBLP

CITATIONS
3

READS
108

2 authors:

Anirban Mukhopadhyay
University of Kalyani
188 PUBLICATIONS **2,919** CITATIONS
SEE PROFILE

Ujjwal Maulik
Jadavpur University
410 PUBLICATIONS **11,074** CITATIONS
SEE PROFILE

Fundamenta Informaticae 86 (2008) 411–428

IOS Press

Unsupervised Pixel Classification in Satellite Imagery: A Two-stage Fuzzy Clustering Approach

Anirban Mukhopadhyay*
Department of Computer Science and Engineering
University of Kalyani, Kalyani - 741235, India
anirbanbuba@yahoo.com

Ujjwal Maulik
Department of Computer Science and Engineering
Jadavpur University, Kolkata - 700032, India
drumaulik@cse.jdvu.ac.in

Abstract. A popular approach for landcover classification in remotely sensed satellite images is clustering the pixels in the spectral domain into several fuzzy partitions. It has been observed that performance of the clustering algorithms deteriorate with more and more overlaps in the data sets. Motivated by this observation, in this article a two-stage fuzzy clustering algorithm is described that utilizes the concept of points having significant membership to multiple classes. The points situated in the overlapped regions of different clusters are first identified and excluded from consideration while clustering. Thereafter, these points are given class labels based on Support vector Machine classifier which is trained by the remaining points. The well known fuzzy C-Means algorithm and some recently proposed genetic clustering schemes are utilized in the process. The effectiveness of the two-stage clustering technique has been demonstrated on IRS remote sensing satellite images of the cities of Bombay and Calcutta and compared with other well known clustering techniques. Also statistical significance test has been carried out to establish the statistical significance of the clustering results.

Keywords: Unsupervised pixel classification, significant multiclass membership, cluster validity index, variable string length genetic algorithm, multiobjective genetic algorithm, Support Vector Machine.

1. Introduction

Clustering [21, 27, 15, 24] is a popular unsupervised pattern classification approach in which a given data set is partitioned into K distinct groups based on some similarity/dissimilarity measures. Clustering can be either crisp or fuzzy. In case of crisp clustering, each point in the given data set belongs to a particular class, whereas fuzzy partitioning schemes associate with each data point a degree of membership to each class. Fuzzy C-Means [8, 9] (FCM) is a widely used technique that uses the principles of fuzzy sets to evolve a partition matrix $U(X)$ while minimizing the global cluster variance. The FCM algorithm often gets stuck at suboptimal solutions based on the initial configuration of the system. In order to overcome this, some genetic algorithm (GA) [17, 12, 13] based fuzzy clustering techniques have been proposed in [23, 25, 3].

In remote sensing applications, one of the important tasks is to classify the pixels in the images into homogeneous regions, each of which corresponds to some landcover type. The problem of pixel classification is often modeled as clustering in the intensity space [7, 25, 1, 4, 2]. In a satellite image, each pixel represents a landcover area, which may not necessarily belong to a single landcover type. Thus in remote sensing images, large number of pixels may have significant belongingness to multiple classes. Therefore a large amount of uncertainty is associated with the pixels in a remotely sensed image. In the unsupervised pixel classification framework, various clustering algorithms like fuzzy c-means (FCM) [10], split-and-merge [22], scale space techniques [29] and neural networks based methods [6] have been used for the purpose of satellite image segmentation.

It has been observed that in general, the performance of clustering algorithms degrade if the clusters in a data set are overlapped, i.e., there are many points in the data set that have significant membership to multiple classes (SiMM). It leads to a lot of confusion regarding their cluster assignments. Hence, it may be beneficial if these points are first determined and excluded from consideration while clustering the data set. They, in the subsequent stage, can be assigned to a cluster using some classifier which is trained by the remaining points. In this article, a Support vector Machines (SVM) based classifier [28] has been utilized for this purpose.

In this article, a two-stage fuzzy clustering algorithm has been described that utilizes the concept of points having significant membership to more than one cluster. Performance of the two-stage clustering method is compared to that of the conventional FCM, K-means, hierarchical average linkage [27] and the genetic algorithm based method on two artificial data sets. Also it is applied to cluster two Indian remote sensing (IRS) satellite images [19] of parts of the cities of Bombay and Calcutta. The superiority of the clustering technique described here, as compared to the well known FCM, K-means and single stage GA based clustering is demonstrated both visually (by showing the clustered images) and using a recently proposed cluster validity index $\mathcal{I}$ [24]. Moreover, statistical significance tests have been conducted to establish the significant superiority of the proposed technique.

2. Fuzzy Clustering Techniques

In this section some fuzzy clustering algorithms used in this article, *viz.,* fuzzy C-means (FCM), variable string length single objective genetic clustering and multiobjective GA based fuzzy clustering algorithm are discussed in brief.

2.1. Fuzzy C-means

Fuzzy C-Means (FCM) [8, 26] is a widely used partitional clustering algorithm. The objective of FCM technique is to use the principles of fuzzy sets to evolve a partition matrix $U(X, K)$ while minimizing the measure

$$J_m = \sum_{j=1}^{n} \sum_{k=1}^{K} u_{kj}^m D^2(z_k, x_j), \tag{1}$$

where n is the number of data objects, K represents number of clusters, $U = [u_{kj}]$ is the fuzzy membership matrix (partition matrix) and m denotes the fuzzy exponent. Here, $D(z_k, x_j)$ denotes the distance of point x_j from the center of the k^{th} cluster. In this article, Euclidean distance measure has been used. It is known that FCM algorithm sometimes gets stuck at some suboptimal solution [20, 18].

2.2. Variable String Length GA-based Fuzzy Clustering

A variable string length GA (VGA)-based fuzzy clustering technique has been proposed in [25] where real valued encoding of cluster centers is used. The algorithm automatically evolves the number of clusters as well as the partitioning and minimizes the Xie-Beni (XB) [30] cluster validity index. Let $\{z_1, z_2, \ldots, z_K\}$ be the set of K cluster centers encoded in a chromosome. Now, the XB index is defined as a function of the ratio of the total variation σ ($= \sum_{i=1}^{K} \sum_{k=1}^{n} u_{ik}^2 D^2(z_i, x_k)$) to the minimum separation sep ($= \min_{i \neq j}\{D^2(z_i, z_j)\}$) of the clusters, i.e.,

$$XB = \frac{\sigma}{n \times sep} = \frac{\sum_{i=1}^{K}(\sum_{k=1}^{n} u_{ik}^2 D^2(z_i, x_k))}{n \times (\min_{i \neq j}\{D^2(z_i, z_j)\})}. \tag{2}$$

Note that for compact and well separated clusters, σ should be low while sep should be high, thereby yielding lower values of the XB index. The objective is therefore to minimize the XB index for achieving proper clustering. Since the number of clusters is considered to be variable, the string lengths of different chromosomes in the same population are allowed to vary [25]. Elitism has been incorporated in order to keep track of the best chromosome obtained so far. The algorithm has been run for a fixed number of generations.

2.3. Multiobjective GA-based Fuzzy Clustering

An efficient multiobjective GA [13, 11, 5] based fuzzy clustering technique has been proposed recently in [2] that uses fixed length chromosomes encoding the real valued encoding of cluster centers. This algorithm optimizes two validity indices, J_m and XB simultaneously resulting a set of nondominated near-Pareto optimal clustering solutions from which the user can choose one depending on problem requirements. A popular Multiobjective GA, called Nondominated Sorting Genetic Algorithm-II (NSGA-II) [14] has been used as the underlying framework for designing the proposed multiobjective GA (MOGA) based fuzzy clustering scheme.

3. Support vector Machines

Support vector machine (SVM) classifiers are inspired by statistical learning theory and they perform structural risk minimization on a nested set structure of separating hyperplanes [28]. A training data set

is used to train the SVM classifier to obtain the optimal separating hyperplane in terms of generalization error. The SVM design algorithm is described below for a two-class problem. It can be extended to handle multi-class problems by designing a number of one-against-all on one-against-one two-class SVMs.

Suppose a data set consists of n feature vectors $< x_i, y_i >$, where $y_i \in \{+1, -1\}$, denotes the class label for the data point x_i. The problem of finding the weight vector w can be formulated as minimizing the following function:

$$L(w) = \frac{1}{2}||w||^2, \tag{3}$$

subject to

$$y_i[w.\phi(x_i) + b] \geq 1, i = 1, \ldots, n. \tag{4}$$

Here, b is the bias and the function $\phi(x)$ maps the input vector to the feature vector. The dual formulation is given by maximizing the following:

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j), \tag{5}$$

subject to

$$\sum_{i=1}^{n} y_i \alpha_i = 0 \text{ and } 0 \leq \alpha_i \leq C, i = 1, \ldots, n. \tag{6}$$

Only a small fraction of the α_i coefficients are nonzero. The corresponding pairs of x_i entries are known as support vectors and they fully define the decision function. Geometrically, the support vectors are the points lying near the separating hyperplane. $K(x_i, x_j) = \phi(x_i).\phi(x_j)$ is called the *kernel function*.

Kernel functions may be linear or non-linear, like polynomial, sigmoidal, radial basis functions (RBF), etc. As in remote sensing images, classes are usually spherical shaped, hence the use of spherical RBF kernel is most appropriate. RBF kernels are of the following form:

$$K(x_i, x_j) = e^{-w|x_i - x_j|^2}. \tag{7}$$

In this article, above mentioned RBF kernel is used. Also, the extended version of the aforesaid two-class SVM that deals with multi-class classification problem by designing a number of one-against-all two-class SVMs, is used here. For example, a K-class problem is handled with K two-class SVMs.

4. Two-stage Clustering Technique

In this section, SiMM points-based two-stage (SiMM-TS) clustering algorithm is discussed. First, the technique for identifying the SiMM points has been discussed. Subsequently, the clustering algorithm based on SiMM point identification has been described.

4.1. Identification of SiMM Points

The matrix $U(X, K)$ produced by some fuzzy clustering technique is used to find out the points which have significant multi-class membership (SiMM), i.e., the points which are situated at the overlapping

regions of two or more clusters, and hence cannot be assigned to any cluster with a reasonable amount of certainty. Let us assume that a particular point $x_j \in X$ has the highest membership value in cluster C_q, and next highest membership value in cluster C_r, i.e., $u_{qj} \geq u_{rj} \geq u_{kj}$ where $k = 1, \ldots, K$, and $k \neq q$, and $k \neq r$. Suppose the difference in the membership values u_{qj} and u_{rj} is δ, i.e., $\delta = u_{qj} - u_{rj}$. Let $\mathcal{B}$ be the set of points lying on the overlapping regions of two or more clusters (SiMM points) and $Prob(x_j \in \mathcal{B})$ denotes the probability of x_j belonging to $\mathcal{B}$. Evidently, as δ increases, x_j can be assigned more confidently to cluster C_q and hence less confidently to $\mathcal{B}$. Therefore,

$$Prob(x_j \in \mathcal{B}) \quad \propto \quad \frac{1}{\delta}. \tag{8}$$

For each point in the data set X, the value δ is calculated. Now the data points are sorted in ascending order of their δ values. Hence, in the sorted list, the probability $Prob(x_i \in \mathcal{B})$ decreases as we move towards the tail of the list. A tuning parameter $\mathcal{P}$ is defined on the sorted list such that the first $\mathcal{P}\%$ points from the sorted list are chosen to be the SiMM points. The value of $\mathcal{P}$ should be chosen carefully so that the appropriate set of SiMM points is identified.

4.2. SiMM-TS Clustering

The algorithm described here has two different stages.

4.2.1. Stage-I

In the first stage, the underlying data set is clustered using either an iterated version of FCM (IFCM) or VGA-based fuzzy clustering algorithm to evolve the number of clusters K as well as the fuzzy partition matrix $U(X, K)$. In IFCM, the data set X is clustered using FCM with different values of K starting from 2 to $\sqrt{n}$, where n is the number of points in the data set. The solution producing the best XB index value is considered and the corresponding partition matrix $U(X, K)$ is used for further processing. Using the resulting partition matrix obtained by IFCM or VGA, the SiMM points are identified using the technique discussed in Section 4.1.

4.2.2. Stage-II

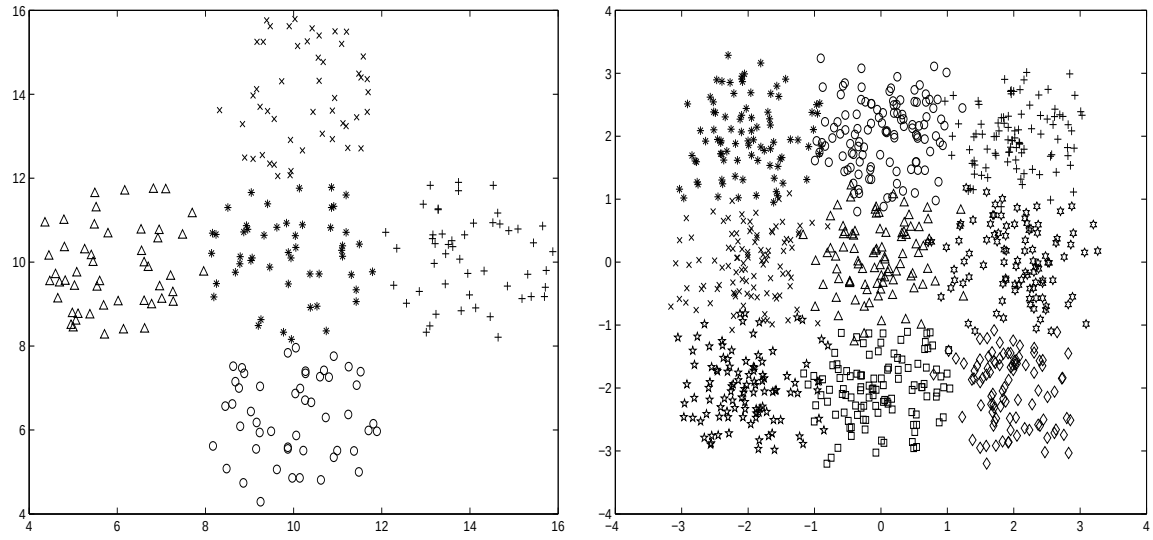
In the second stage of SiMM-TS, the SiMM points are excluded from the data set and the remaining points are re-clustered into K clusters using any of the FCM, fixed length GA and MOGA-based clustering algorithm. The clustering is defuzzified by assigning each of these remaining points to the cluster to which it has the highest membership degree. Next, the SVM classifier is trained by these points. Thereafter, each of the SiMM points that were identified in the first stage, and excluded from consideration in the second, is assigned to a cluster as predicted by the trained SVM classifier.

5. Illustration of SiMM-TS Clustering Process

The experimental results of clustering using the proposed SiMM-TS clustering scheme are provided for two artificial data sets. These are first described below, followed by the performance measure used for comparison. Finally, the results are provided.

5.1. Data Sets

Sp250_2_5: This is a overlapping two dimensional data set where the number of clusters is five. It has 250 points. The value of K is chosen to be 5. The data set is shown in Fig. 1(a).
St900_2_9: This is an overlapping two dimensional triangular distribution of data points having nine classes where all the classes are assumed to have equal *a priori* probabilities $(= \frac{1}{9})$. It has 900 data points. The $X - Y$ ranges for the nine classes are as follows:
Class 1: [-3.3, -0.7] $\times$ [0.7, 3.3],
Class 2: [-1.3, 1.3] $\times$ [0.7, 3.3],
Class 3: [0.7, 3.3] $\times$ [0.7, 3.3],
Class 4: [-3.3, -0.7] $\times$ [-1.3, 1.3],
Class 5: [-1.3, 1.3] $\times$ [-1.3, 1.3],
Class 6: [0.7, 3.3] $\times$ [-1.3, 1.3],
Class 7: [-3.3, -0.7] $\times$ [- 3.3, -0.7],
Class 8: [-1.3, 1.3] $\times$ [- 3.3, -0.7],
Class 9: [0.7, 3.3] $\times$ [- 3.3, -0.7].
The domain of the triangular distribution for each class and for each axis is 2.6. Consequently, the height will be $\frac{1}{1.3}$ (since $12 \times 2.6 \times height = 1$). This data set is shown in Fig. 1(b).



(a) (b)

Figure 1. Artificial data sets: (a) *Sp250_2_5* data, (b) *St900_2_9* data

5.2. Performance Metric

Here, the performances of the clustering algorithms are evaluated in terms of the *Minkowski score*. A clustering solution for a set of n elements can be represented by an $n \times n$ matrix C, where $C_{i,j} = 1$ if point i and j are in the same cluster according to the solution, and $C_{i,j} = 0$ otherwise. The Minkowski score (MS) of a clustering result C with reference to T, the matrix corresponding to the true clustering,

is defined as

$$MS(T, C) = \frac{\| T - C \|}{\| T \|} \tag{9}$$

where $\| T \| = \sqrt{\sum_i \sum_j T_{i,j}}$.

The Minkowski score is the normalized distance between the two matrices. Lower Minkowski score implies better clustering solution, and a perfect solution will have a score zero.

5.3. Input Parameters

The GA based algorithms have been run for 100 generations with population size 50. The crossover and mutation probabilities were taken as 0.8 and 0.1, respectively. The FCM and K-means algorithms have been executed for 100 iterations unless they converge before that. The fuzziness parameter m is taken as 2.0. For SiMM-TS clustering, the tuning parameter $\mathcal{P}$ is chosen experimentally, i.e., SiMM-TS is run for different values of $\mathcal{P}$ ranging from 5% to 15% and the best solution has been chosen. Each algorithm has been run for 20 times and the best performance index value has been reported.

5.4. Results

Table 1 and 2 show the comparative results in terms of Minkowski scores obtained by the different algorithms for the two data sets, respectively. As can be seen from tables, irrespective of the clustering method used in the proposed algorithm, the performance gets improved after the application of the second stage of clustering. For example, in case of *Sp250_2_5*, the Minkowski score after the application of VGA in the first stage of SiMM-TS is 0.4398 while this gets improved to 0.3418 (with MOGA at the second stage) at the end. Similarly, when IFCM is applied in the first stage, the score is 0.4404 which gets improved to 0.3527 (with MOGA at the second stage). It also appears that both VGA and IFCM correctly identify the number of clusters in the data sets. The final Minkowski scores are also better than those obtained using the average linkage and k-means clustering methods. Similar results are found for *St900_2_9* data set also. The results demonstrate the utility of adopting the approach presented in this paper, irrespective of the clustering method used.

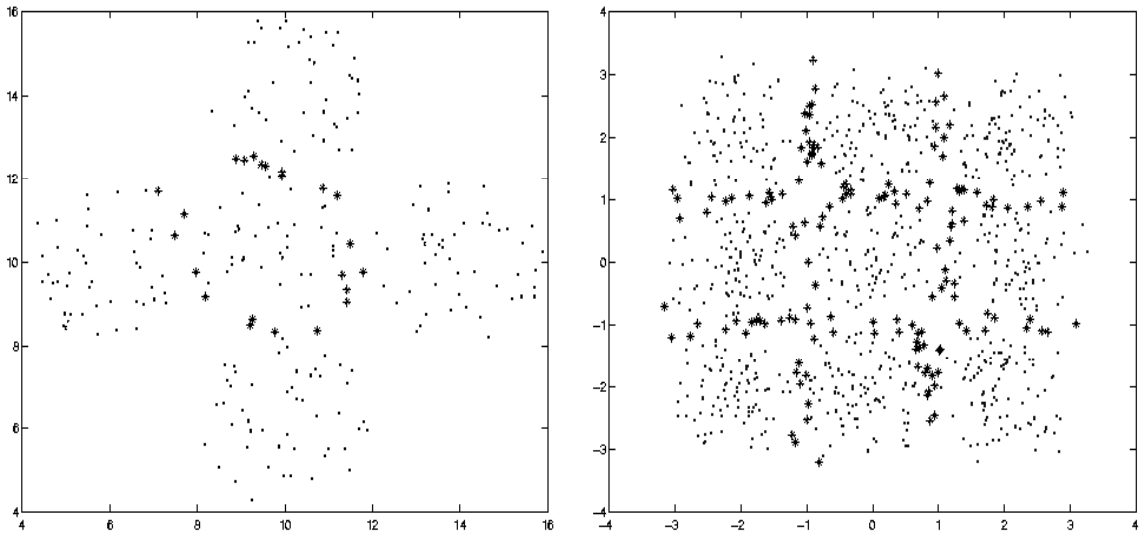
Table 1. Results for Sp250_2_5 data set

Stage-I Algorithm	K	Stage-I MS	Stage-II Algorithm	Final MS	Average linkage (K=5)	K-means (K=5)
VGA	5	0.4398	FCM	0.4037	0.4360	0.4243
			GA	0.3911		
			MOGA	0.3418		
IFCM	5	0.4404	FCM	0.4146		
			GA	0.3661		
			MOGA	0.3527		

Table 2. Results for St900_2_9 data set

Stage-I Algorithm	K	Stage-I MS	Stage-II Algorithm	Final MS	Average linkage (K=9)	K-means (K=9)
VGA	9	0.5348	FCM	0.5205	0.6516	0.6161
			GA	0.5128		
			MOGA	0.4418		
IFCM	9	0.5314	FCM	0.5022		
			GA	0.5022		
			MOGA	0.4418		

Fig. 2 shows, for the purpose of illustration, the SiMM points identified by one of the runs of the proposed method for *Sp250_2_5* and *St900_2_9* data sets. As is evident from the figure, these points are situated at the overlapping regions of two or more clusters.



(a) (b)

Figure 2. Artificial data set with the points identified as belonging to multiple classes marked as '*': (a) *Sp250_2_5* data, (b) *St900_2_9* data

6. Application to Pixel Classification

This section presents the results of application of the proposed SiMM-TS clustering on two remote sensing satellite images of parts of the cities of Bombay and Calcutta. Results are demonstrated both visually and using a cluster validity index $\mathcal{I}$ [24] (described later).

6.1. IRS Bombay Image

The data used here was acquired from Indian Remote Sensing Satellite (IRS-1A) [19] using the LISS-II sensor that has a resolution of $36.25m \times 36.25m$. The image contains four spectral bands namely, blue band of wavelength 0.45-0.52 μm, green band of wavelength 0.52-0.59 μm, red band of wavelength 0.62-0.68 μm, and near infrared band of wavelength 0.77-0.86 μm. Fig. 3 shows the IRS image of a part of Bombay in the near infrared band. As can be seen, the city area is enclosed on three sides by the Arabian sea. Towards the bottom right of the image, there are many islands, including the famous Elephanta island. The dockyard is situated on the south eastern part of Bombay, which can be seen as a set of three finger like structure. In the first stage of SiMM-TS, both VGA and IFCM identified 7 clusters. As per our ground knowledge, these clusters correspond to 7 landcover regions namely concrete (Concr.), open spaces (OS1 and OS2), vegetation (Veg), habitation (Hab), and turbid water (TW1 and TW2) [25].

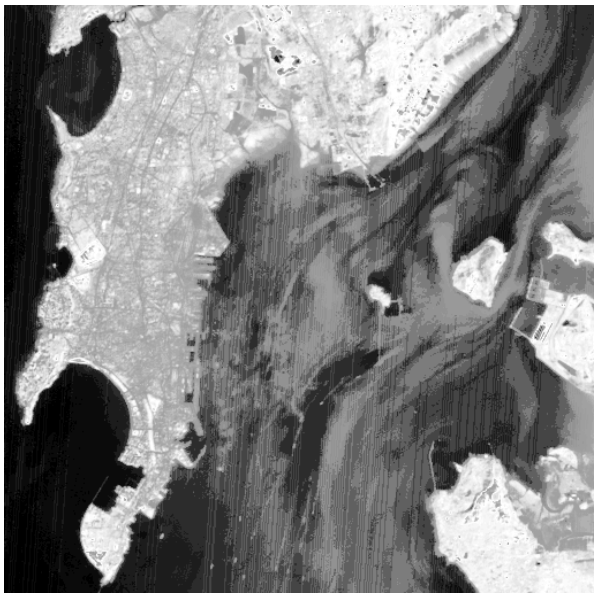


Figure 3. IRS image of Bombay in the NIR band with histogram equalization

The result of the application of the proposed SiMM-TS (VGA-MOGA combination) clustering technique on the Bombay image is shown in Fig. 4. The southern part of the city, which is heavily industrialized, has been classified as primarily belonging to habitation and concrete. Here, the class habitation represents the regions having concrete structures and buildings, but with relatively lower density than the class Concrete. Hence these two classes share common properties. From the result, it appears that the large water body of Arabian sea is grouped into two classes (TW1 and TW2). It is evident from the figure that the sea water has two distinct regions with different spectral properties. Hence the clustering result providing two partitions for this region is quite expected. Most of the islands, dockyard, several road structures have been correctly identified in the image. As expected, there is a high proportion of open space and vegetation within the islands.

Fig. 5 demonstrates the Bombay image clustered using the FCM clustering algorithm. It can be noted

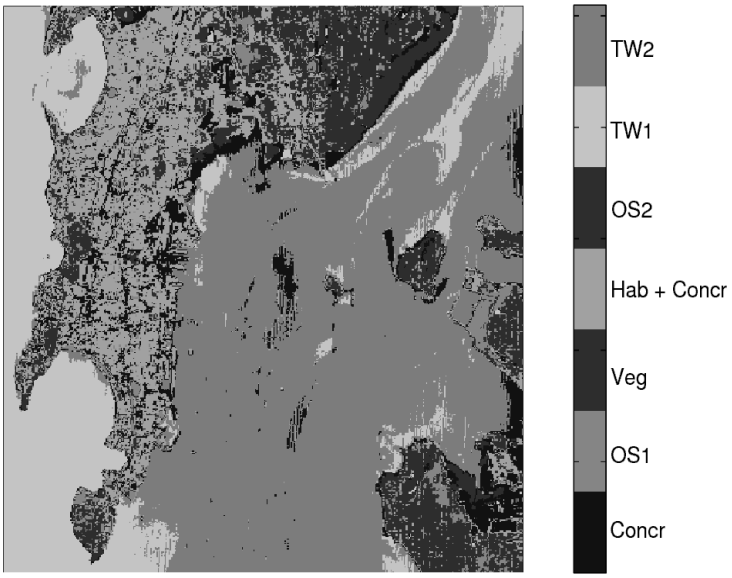


Figure 4. Clustered IRS image of Bombay using SiMM-TS (VGA-MOGA) clustering scheme

from the figure that the water of the Arabian sea has been wrongly clustered into three regions, rather than two as obtained earlier. It appears that the other regions in the image have been classified more or less correctly for this data.

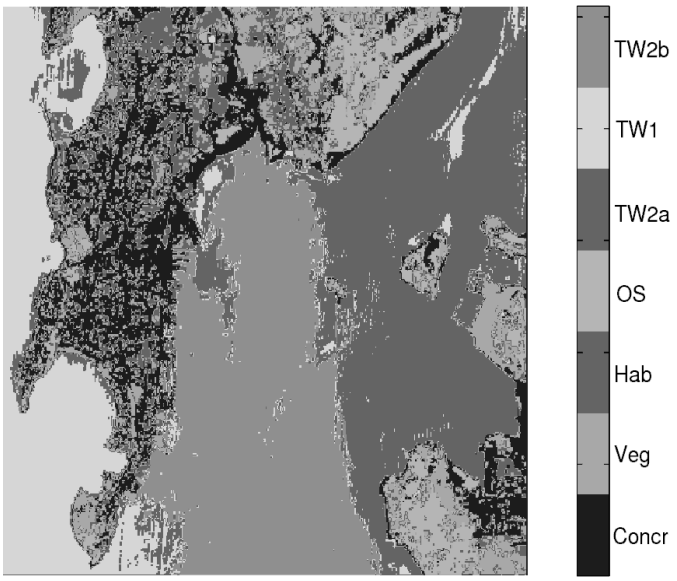


Figure 5. Clustered IRS image of Bombay using FCM clustering

In Fig. 6, the Bombay image clustered using K-means clustering has been shown. It appears from the figure that the sea area is wrongly classified into four different regions. Also there are overlapping

between the classes turbid water and concrete, as well as between open space and vegetation.

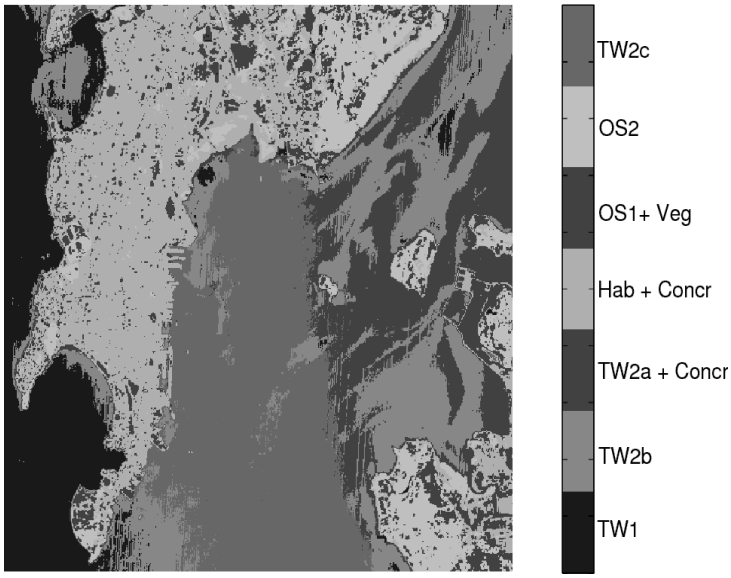


Figure 6. Clustered IRS image of Bombay using K-means clustering

6.2. IRS Calcutta Image

The IRS image of Calcutta was also obtained using the LISS-II sensor. It is available in four bands, viz., blue, green, red and near infrared. Fig. 7 shows the Calcutta image in the near infra red band. Some characteristic regions in the image are the river Hooghly cutting across the middle of the image, several fisheries observed towards the lower-right portion, a township, Saltlake, to the upper-left hand side of the fisheries. This township is surrounded on the top by a canal. Two parallel lines observed towards the upper right hand side of the image correspond to the airstrips in the Dumdum airport. Other than these there are several water bodies, roads etc. in the image. In the first stage of SiMM-TS, both VGA and IFCM determined 4 clusters, which, according to our ground knowledge, correspond to the classes turbid water (TW), pond water (PW), concrete (Concr.) and open space (OS) [25].

Fig. 8 shows the Calcutta image partitioned using the two-stage fuzzy algorithm (VGA-MOGA combination in the two stages). It appears that the water class has been differentiated into turbid water (the river Hooghly) and pond water (canal, fisheries etc.). This is expected because they differ in their spectral properties. Here, the class turbid water contains sea water, river water etc., where the soil content is more than that of pond water. The Saltlake township has come out as a combination of concrete and open space, which appears to be correct, as the township is known to have several open spaces. The canal bounding Saltlake from the upper portion has also been correctly classified as PW. The airstrips of Dumdum airport are classified correctly as belonging to the class concrete. Presence of some small areas of PW beside the airstrips is also correct as these correspond to the several ponds. The high proportion of concrete on both sides of the river, particularly towards the lower region of the image is also correct. This region corresponds to the central part of the city of Calcutta.

Figure 7. IRS image of Calcutta in the NIR band with histogram equalization

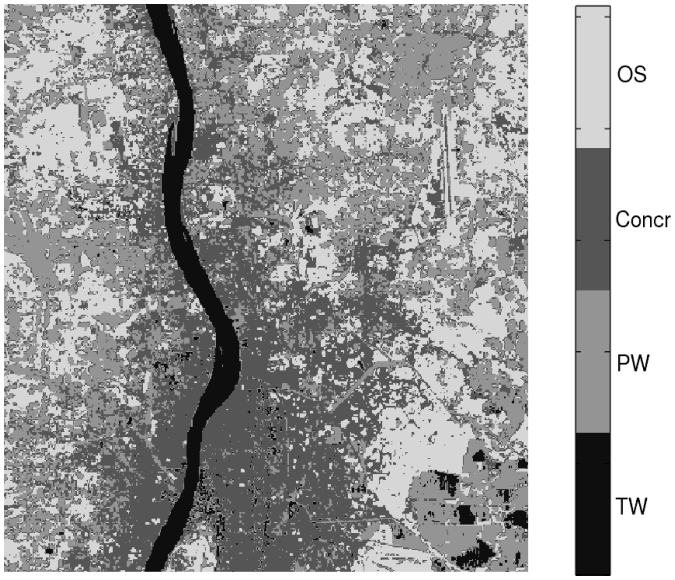


Figure 8. Clustered IRS image of Calcutta using SiMM-TS (VGA-MOGA) clustering scheme

Fig. 9 demonstrates the clustering result of the FCM algorithm on the IRS Calcutta image. It appears From the figure that the river Hooghly and the city region has been incorrectly classified into the same class. It is also evident that the whole Saltlake city has been wrongly put into one class. It is also apparent that although some portions like the fisheries, canals, parts of airstrip are identified correctly, there is a significant amount of confusion in the FCM clustering result.

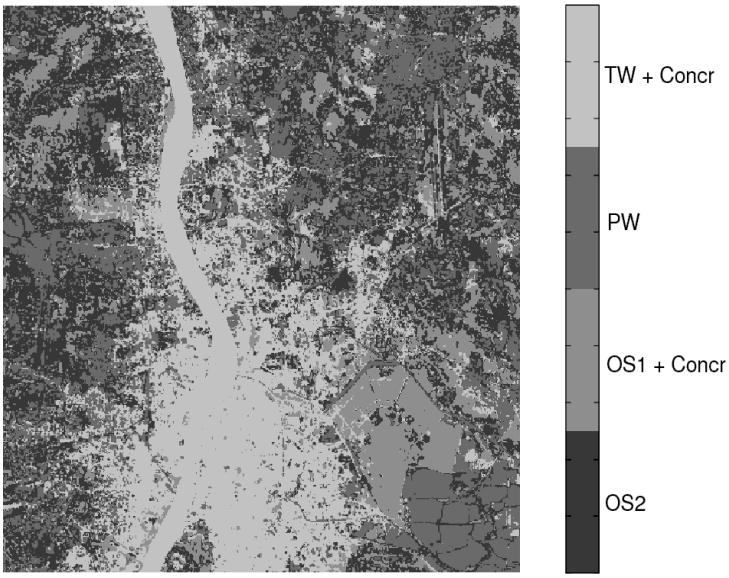


Figure 9. Clustered IRS image of Calcutta using FCM clustering

The K-means clustering result for Calcutta image is shown in Fig. 10. The figure implies that for K-means clustering also, a lot of confusion exists between the classes turbid water and concrete. Also open spaces and concrete classes have some amount of overlapping. It appears also that the airstrips of Dumdum airport are not very clear.

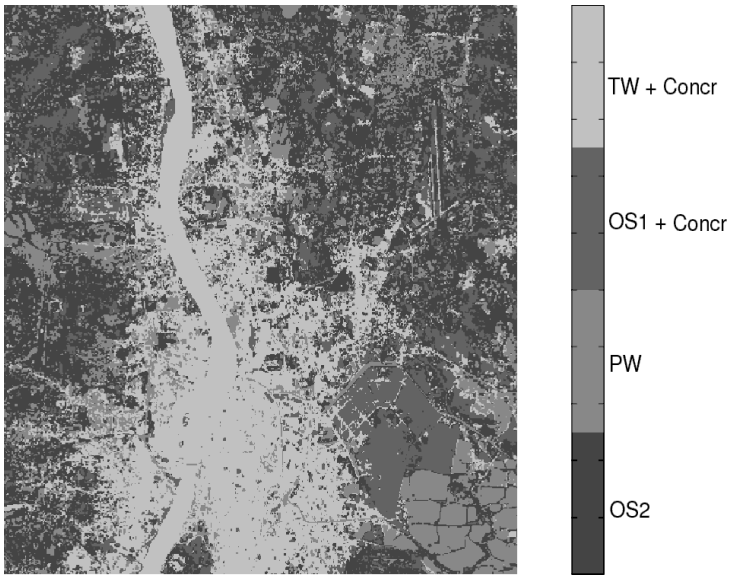


Figure 10. Clustered IRS image of Calcutta using K-means clustering

6.3. Numerical Results

To show the effectiveness of the SiMM-TS clustering technique, a cluster validity index [24] $\mathcal{I}$ has been calculated for the clustering solutions provided by different clustering algorithms. The index $\mathcal{I}$, proposed as a measure of indicating the goodness/validity of a clustering solution, is defined as follows:

$$\mathcal{I}(K) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times \mathcal{D}_K\right)^p, \tag{10}$$

where

$$E_K = \sum_{k=1}^{K} \sum_{j=1}^{n} u_{kj} D(x_j, z_k), \tag{11}$$

and

$$\mathcal{D}_K = \max_{i,j=1}^{K} \{D(z_i, z_j)\}. \tag{12}$$

The different terms are defined earlier. The index $\mathcal{I}$ is a composition of three factors, namely, $\frac{1}{K}$, $\frac{E_1}{E_K}$ and $\mathcal{D}_K$. These three factors are found to compete with and balance each other critically. $\mathcal{I}$ index has been shown to provide superior performance when compared to several other validity indices [24]. In this article, we have taken $p = 2$. Larger value of $\mathcal{I}$ index implies better solution. Note that for computing the index $\mathcal{I}$, knowledge about the true partitioning of the data is not necessary.

In Table 3, the best $\mathcal{I}$ index values over 20 consecutive runs of the SiMM-TS clustering technique (with different combinations of VGA and IFCM in the first stage and FCM, GA and MOGA in the second stage), along with single stage VGA, FCM and K-means algorithms are tabulated. It is evident from the table that the two-stage clustering algorithm with VGA-MOGA combination performs the best, far outperforming the single stage VGA, FCM and K-means algorithms. Hence it follows from the results that the SiMM-TS clustering algorithm outperforms both the single-stage VGA, FCM and K-means clustering algorithms.

Table 3. Best $\mathcal{I}$ index values for IRS Bombay and Calcutta images for different algorithms over 20 runs

Method	Bombay	Calcutta
Two-stage (VGA-FCM) clustering	193.9961	102.3039
Two-stage (VGA-GA) clustering	198.1145	103.8810
Two-stage (VGA-MOGA) clustering	**209.8299**	**111.3328**
Two-stage (IFCM-FCM) clustering	188.3759	85.5597
Two-stage (IFCM-GA) clustering	191.8271	89.2094
Two-stage (IFCM-MOGA) clustering	192.7360	94.3211
VGA clustering	180.4512	81.5934
FCM clustering	178.0322	31.1697
K-means clustering	161.3783	54.3139

7. Statistical Significance Test

Table 4 reports the average values of the $\mathcal{I}$ index scores produced by different algorithms over 20 consecutive runs for Bombay and Calcutta images. It is evident from the table that for both the images, SiMM-TS clustering technique with the VGA-MOGA combination produces the best average $\mathcal{I}$ index scores. To establish that the better average $\mathcal{I}$ index scores provided by VGA-MOGA is statistically significant and does not come by chance, it is necessary to conduct a statistical significance test.

Table 4. Average $\mathcal{I}$ index values for IRS Bombay and Calcutta images for different algorithms over 20 runs

Method	Bombay	Calcutta
Two-stage (VGA-FCM) clustering	190.9300	98.3677
Two-stage (VGA-GA) clustering	194.6330	99.1122
Two-stage (VGA-MOGA) clustering	**204.2614**	**105.2473**
Two-stage (IFCM-FCM) clustering	185.6765	82.1449
Two-stage (IFCM-GA) clustering	187.7561	86.0098
Two-stage (IFCM-MOGA) clustering	189.9294	90.9404
VGA clustering	177.5296	80.1971
FCM clustering	173.8489	28.5166
K-means clustering	157.5376	50.8785

In this article, a statistical significance test called t-test [16] has been carried out at the 5% significance level, to compare the average $\mathcal{I}$ index scores produced by different algorithms. Nine groups, corresponding to the nine algorithms (VGA-FCM, VGA-GA, VGA-MOGA, IFCM-FCM, IFCM-GA, IFCM-MOGA, VGA, FCM and K-means), have been created for each of the two image data considered here. Each group consists of $\mathcal{I}$ index scores produced by 20 consecutive runs of the corresponding algorithm. Two groups are compared at a time, one corresponding to the SiMM-TS (VGA-MOGA) scheme and the other corresponding to some other algorithm considered in this article.

Tables 5 and 6 report the results of the t-test for the IRS Bombay image and IRS Calcutta image, respectively. The null hypothesis (The means of two groups are equal) are shown in the tables. The alternative hypothesis is that the mean of the first group is larger than the mean of the second group. For each test, the degree of freedom is $M + N - 2$, where M and N are the sizes of two groups considered. Here $M = N = 20$. Hence the degree of freedom is 38. Also the values of t-statistic and the probability (*P-value*) of accepting the null hypothesis are shown in the tables. It is clear from the tables that the *P-values* are much less than 0.05 (5% significance level) which are strong evidences for rejecting the null hypothesis. This proves that the better average $\mathcal{I}$ index values produced by the SiMM-TS clustering with VGA-MOGA scheme is statistically significant and has not come by chance.

8. Discussion and Conclusions

In this article a two-stage clustering method (SiMM-TS) that identifies the points with the maximum confusion regarding their cluster assignments has been described. Widely used FCM algorithm and a

Table 5. The t-test results for the IRS Bombay image

Test No.	Null hypothesis ($H_0 : \mu_1 = \mu_2$)	t-test statistic	P-value	Accept/Reject
1	$\mu_{VGA-MOGA} = \mu_{VGA-FCM}$	12.7033	2.0065e-010	Reject
2	$\mu_{VGA-MOGA} = \mu_{VGA-GA}$	9.0797	3.8615e-008	Reject
3	$\mu_{VGA-MOGA} = \mu_{IFCM-FCM}$	19.4688	1.5321e-013	Reject
4	$\mu_{VGA-MOGA} = \mu_{IFCM-GA}$	14.2710	2.9582e-011	Reject
5	$\mu_{VGA-MOGA} = \mu_{IFCM-MOGA}$	14.6817	1.8445e-011	Reject
6	$\mu_{VGA-MOGA} = \mu_{VGA}$	16.0858	3.9810e-012	Reject
7	$\mu_{VGA-MOGA} = \mu_{IFCM}$	26.1787	8.8818e-016	Reject
8	$\mu_{VGA-MOGA} = \mu_{K-means}$	30.2918	5.1039e-029	Reject

Table 6. The t-test results for the IRS Calcutta image

Test No.	Null hypothesis ($H_0 : \mu_1 = \mu_2$)	t-test statistic	P-value	Accept/Reject
1	$\mu_{VGA-MOGA} = \mu_{VGA-FCM}$	4.7217	1.7016e-004	Reject
2	$\mu_{VGA-MOGA} = \mu_{VGA-GA}$	4.9114	1.1251e-004	Reject
3	$\mu_{VGA-MOGA} = \mu_{IFCM-FCM}$	16.7409	2.0255e-012	Reject
4	$\mu_{VGA-MOGA} = \mu_{IFCM-GA}$	14.2634	2.9843e-011	Reject
5	$\mu_{VGA-MOGA} = \mu_{IFCM-MOGA}$	10.1479	7.1216e-009	Reject
6	$\mu_{VGA-MOGA} = \mu_{VGA}$	20.4079	6.7946e-014	Reject
7	$\mu_{VGA-MOGA} = \mu_{IFCM}$	59.3758	6.8433e-042	Reject
8	$\mu_{VGA-MOGA} = \mu_{K-means}$	36.1834	4.2241e-030	Reject

recently developed VGA and MOGA based fuzzy clustering methods and an SVM classifier have been used for this purpose. Experimental results indicate that this approach, with a suitable choice of a single parameter $\mathcal{P}$, is likely to yield better results irrespective of the actual clustering technique adopted. Also the developed two-stage clustering algorithm has been applied to cluster remote sensing satellite imagery into unknown number of regions. Results obtained from this are validated both visually and using a cluster validity index $\mathcal{I}$, and are compared with other algorithms to demonstrate the effectiveness of the proposed two-stage technique. Also statistical significance tests based on t-statistic have been conducted to establish the fact that the better performance of the SiMM-TS clustering with VGA-MOGA combination is statistically significant.

Since in many real life situations, information about the number of clusters is not available, applying VGA clustering in the first stage of the algorithm SiMM-TS followed by the multiobjective genetic clustering scheme in the second stage is recommended for best clustering results. As a future work, more studies need to be made to determine a good choice of parameter $\mathcal{P}$. Authors are working in this direction.

References

[1] Bandyopadhyay, S., Maulik, U.: Genetic Clustering for Automatic Evolution of Clusters and Application to Image Classification, *Pattern Recognition*, **35**(2), 2002, 1197–1208.

[2] Bandyopadhyay, S., Maulik, U., Mukhopadhyay, A.: Multiobjective Genetic Clustering for Pixel Classification in Remote Sensing Imagery, *IEEE Transactions on Geoscience and Remote Sensing*, **45**(5), 2007, 1506–1511.

[3] Bandyopadhyay, S., Mukhopadhyay, A., Maulik, U.: An Improved Algorithm for Clustering Gene Expression Data, *Bioinformatics*, **23**(21), 2007, 2859–2865.

[4] Bandyopadhyay, S., Pal, S. K.: Pixel Classification Using Variable String Genetic Algorithms with Chromosome Differentiation, *IEEE Transactions on Geoscience and Remote Sensing*, **39**(2), 2001, 303– 308.

[5] Bandyopadhyay, S., Saha, S., Maulik, U., Deb, K.: A Simulated Annealing based Multi-objective Optimization Algorithm: AMOSA, *IEEE Transactions on Evolutionary Computation*, (in press).

[6] Baraldi, A., Parmiggiani, F.: A Neural Network for Unsupervised Categorization of Multivalued Input Pattern: An Application to Satellite Image Clustering, *IEEE Transactions on Geoscience and Remote Sensing*, **33**, 1995, 305–316.

[7] Bensaid, A. M., Hall, L. O., Bezdek, J. C., Clarke, L. P., Silbiger, M. L., Arrington, J. A., Murtagh, R. F.: Validity-Guided (Re)Clustering with Applications to Image Segmentation, *IEEE Transactions on Fuzzy Systems*, **4**(2), 1996, 112–123.

[8] Bezdek, J. C.: *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, New York, 1981.

[9] Bezdek, J. C., Pal, N. R.: Some New Indexes of Cluster Validity, *IEEE Transactions on Systems, Man and Cybernetics*, **28**, 1998, 301–315.

[10] Cannon, R. L., Dave, R., Bezdek, J. C., Trivedi, M.: Segmentation of a Thematic Mapper Image using Fuzzy c-means Clustering Algorithm, *IEEE Transactions on Geoscience and Remote Sensing*, **24**, 1986, 400– 408.

[11] Coello Coello, C. A.: A comprehensive survey of evolutionary-based multiobjective optimization techniques, *Knowledge and Information Systems*, **1**(3), 1999, 129–156.

[12] Davis, L., Ed.: *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York, 1991.

[13] Deb, K.: *Multi-objective Optimization Using Evolutionary Algorithms*, John Wiley and Sons, Ltd, England, 2001.

[14] Deb, K., Pratap, A., Agrawal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II, *IEEE Transactions on Evolutionary Computation*, **6**, 2002, 182–197.

[15] Everitt, B. S.: *Cluster Analysis*, Third edition, Halsted Press, 1993.

[16] Ferguson, G. A., Takane, Y.: *Statistical Analysis in Psychology and Education*, Sixth edition, McGraw-Hill Ryerson Limited, 2005.

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

[18] Groll, L., Jakel, J.: A New Convergence Proof of Fuzzy c-Means, *IEEE Transactions on Fuzzy Systems*, **13**(5), 2005, 717–720.

[19] data users handbook, I.: *NRSA, Hyderabad, India, Rep. IRS/NRSA/NDC/HB-01/86*, 1986.

[20] Hoppner, F., Klawonn, F.: A contribution to convergence theory of fuzzy c-means and derivatives, *IEEE Transactions on Fuzzy Systems*, **11**(5), 2003, 682–694.

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

[22] Laprade, R. H.: Split-and-merge Segmentation of Aerial Photographs, *Computer Vision Graphics and Image Processing*, **48**, 1988, 77–86.

[23] Maulik, U., Bandyopadhyay, S.: Genetic Algorithm Based Clustering Technique, *Pattern Recognition*, **33**, 2000, 1455–1465.

[24] Maulik, U., Bandyopadhyay, S.: Performance Evaluation of Some Clustering Algorithms and Validity Indices, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(12), 2002, 1650–1654.

[25] Maulik, U., Bandyopadhyay, S.: Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification, *IEEE Transactions on Geoscience and Remote Sensing*, **41**(5), 2003, 1075– 1081.

[26] Pal, N. R., Bezdek, J. C.: On Cluster Validity for the Fuzzy c-Means Model, *IEEE Transactions on Fuzzy Systems*, **3**, 1995, 370–379.

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

[28] Vapnik, V.: *Statistical Learning Theory*, Wiley, New York, USA, 1998.

[29] Wong, Y. F., Posner, E. C.: A New Clustering Algorithm Applicable to Polarimetric and SAR Images, *IEEE Transactions on Geoscience and Remote Sensing*, **31**(3), 1993, 634–644.

[30] Xie, X. L., Beni, G.: A Validity Measure for Fuzzy Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13**, 1991, 841–847.