Journal of Basic and Applied Engineering Research
Print ISSN: 2350-0077; Online ISSN: 2350-0255; Volume 2, Number 16; April-June, 2015 pp. 1411-1416

http://www.krishisanskriti.org/jbaer.html

# Similarity and Dissimilarity Measures with Fuzzy Classifiers: SMIC Tool

**Uma Tyagi[1] and Anil Kumar[2]**

[1]*M.I.E.T, Meerut*
[2]*IIRS, ISRO, Dehradun*
*E-mail: [1]cse0906810103@gmail.com, [2]anil@iirs.gov.in*

**Abstract—***Remotely sensed images are dominated by mixed pixels. Hard classifiers may not handle mixed pixels. In literature, fuzzy theory based algorithms like Fuzzy c-Means, Possibilistic c-Means, Noise clustering with or without Entropy and Fuzzy c-Means with Entropy have been studied with Euclidean and Mahalonobis norms only, to handle mixed pixels. This paper presents the capabilities of SMIC tool with similarity and dissimilarity measures in fuzzy classifiers. Landsat 8 has been used to test implemented norms. Canberra and Manhattan dissimilarity norms generate poor results as assigning low membership value with high variance to favourable class. While Cosine and Correlation similarity norms provide best results as assigning high membership value with low variance to favourable class.*

**Keywords***: Similarity, Dissimilarity, Fuzzy c-Means, Possibilistic c-Means, Noise Clustering.*

## 1. INTRODUCTION

Remotely sensed image data is extensively applied in a range of oceanographic, atmospheric and other socio-economic applications, such as resource utilization, environmental modelling, monitoring urban development and land-cover mapping. Thematic maps have a wide application among the end products of remote sensing. In the digital domain, thematic maps are created by assigning labels to each pixel in an image and, this process is called as Digital Image Classification [8]. However, classifying any remotely sensed data into a thematic map remains a challenge due to many factors, such as, complexity of landscape, selection of remotely sensed data, image-processing and classification approaches, which may affect the success of a classification [3].

Conventional classification technique presumes that, each pixel in an image contains a single class. However, a pixel may contain more than one class, and such pixels are known as mixed pixels [8]. Occurrence of mixed pixels may be a problem in mapping and monitoring land cover, and in particular, their effect is most severe in mapping heterogeneous landscape from coarse spatial resolution images [5]. The concept of fuzzy set was found suitable for addressing the mixed pixel problem, so that pixels may have multiple or partial class membership [6]. In Fuzzy case, a measure of the strength of membership for each class is output by the classifier, resulting in a soft classification technique [15]. In recent scenario, it has been observed that conventional hard classification techniques, which allocate each pixel to a specific class, are often inappropriate for applications where mixed pixels are present in ample amount [4].

Various fuzzy based classifiers like Fuzzy *c*-Mean (FCM), FCM with Entropy (FCME), Possibilistic *c*-Means (PCM), Noise Clustering (NC) and NC with Entropy (NCE), can handle mixed pixels using fuzzy set theory. Generally these classifiers have been applied with Euclidean, Mahalonobis and Diagonal Mahalonobis norms only [3, 14]. These classifiers have not been used with various similarity and dissimilarity measures. Various other similarity and dissimilarity measures can also be incorporated in these classifiers. In this paper, similarity and dissimilarity measures considered were Normalized-Squared-Euclidean, Cosine, Braycurtis, Correlation, Manhattan, Canberra, Chessboard, Median-Absolute-Difference, and Mean-Absolute-Difference in single or composite mode with weighted component.

Commercial image processing software e.g. ERDAS Imagine contains Minimum distance classifier, Parallelepiped and Linear Mixture Model, ENVI also contains the same as well as Decision Tree, Artificial Neural Network and SVM classifiers, PCI Geomatica consist of FCM, GRASS and ILWIS incorporate Minimum distance, Parallelepiped or called Box Classifiers. Many fuzzy based soft classifiers like FCM with Entropy, PCM, Noisy Clustering, and Noisy Clustering with Entropy are not implemented yet in commercial image processing software that to with large number of similarity and dissimilarity norms. This paper presents the capabilities of SMIC [10] (Sub Pixel Multispectral Image Classifier) Tool with fuzzy classifiers incorporating various similarity and dissimilarity measure in single or composite mode with weighted component.

## 2. SOFT CLASSIFICATION METHOD

One of the major approaches to generate land cover information from remotely sensed images is classification. Numerous classification algorithms have been developed. Among the most popular are the maximum likelihood classifier (MLC), neural network classifiers, decision tree classifiers and support vector machine (SVM). Maximum likelihood classification (MLC) is a supervised statistical approach for thematic mapping using pixel based information. MLC follow Gaussian rule requires high degree of computation when large number of classes is to be classified and become unreliable [7], if sample size for each class is not large enough. ANN is a computational model inspired by the biological neural network. It is a non parametric approach and does not follow Gaussian rule. An advantage of neural network lies in the high computation rate achieved by their massive parallelism. At the other hand, ANN can be very complex as the learning time of a neural network can be very long for high dimensional data [1]. Generalization requires large training data, as data dimensionality increases its structure becomes more complex.

Decision tree uses a hierarchical mechanism. It breaks a complex classification problem into multiple stages of simple decision making processes [12]. The main benefit of using a hierarchical mechanism is that the tree structure can be viewed as white box, so that easier to interpret as compare to ANN [13]. Another benefit it requires less complicated training compared to the ANN but for decision tree, decision rules must be framed. It become complex when decision rules are large in number. SVM is a statistical learning classification technique. It was originally linear binary classifier which allocates the labels [13]. The core operation of SVM is to construct a separating hyperplane on the basis of the properties of the training samples. SVM provides higher accuracy compared to other methods, such as MLC, ANN and decision trees [9]. On the other hands, SVM is time consuming and costly when labelled samples are collected [2]. Therefore, these classifiers are having some limitations like, ANN requires large training data, MLC follow Gaussian rule, SVM is time consuming as well as costly and decision rules must be framed in Decision tree. Thus to overcome these limitations a different category of classifiers needed. Fuzzy concept is also a valuable approach for dealing with classification problem. This paper provides the details of fuzzy set theory based algorithms with various similarity and dissimilarity measures.

## 3. SMIC GUI DETAILS

Input for SMIC system can be any multispectral remote sensing image. This system includes five different fuzzy classification algorithms (FCM, PCM, NC, NCE, and FCME) incorporating nine similarity and dissimilarity measures in a supervised mode. Fig. 1 shows the classifier GUI of SMIC in which user can select any similarity and dissimilarity measures including Euclidean, Mahalonobis and Diagonal Mahalonobis norms for FCM, FCME, NC, NCE and PCM classifier with or without MRF. There is the provision kept in this system that composite distance can be generated will using any of the two distances in combination by choosing a weighting component. In developed tool, user has a choice to select either Type 1 or Type 2 fuzzy outputs. Type 2 fuzzy set is the fuzziness in a fuzzy set. In Type 2 fuzzy clustering algorithm [11], the membership value of each pattern in the image is extended as Type 2 fuzzy memberships by assigning membership grades (triangular membership function) to Type 1 fuzzy membership. The membership values for the Type 2 have been obtained in equation (1).

$$a_{ij} = \mu_{ij} - \frac{1-\mu_{ij}}{2} \tag{1}$$

Where $a_{ij}$ and $\mu_{ij}$ are the Type 2 and Type 1 fuzzy membership respectively.

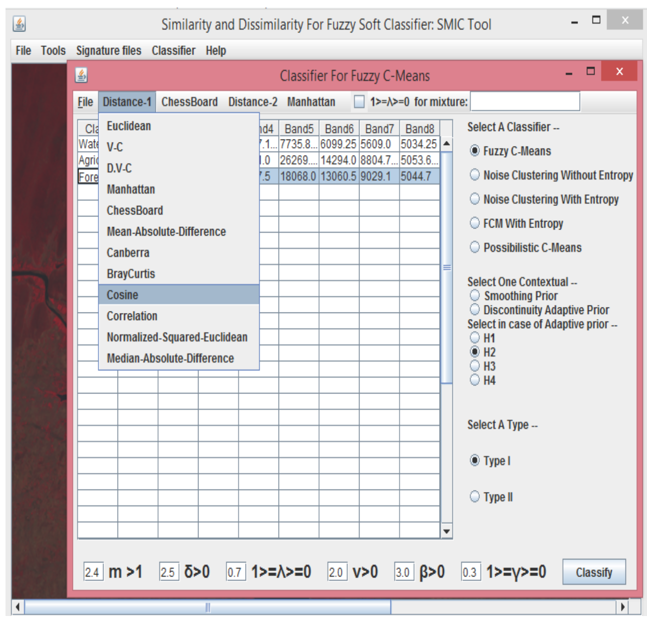



**Fig. 1: Graphical user interface of SMIC**

## 4. FUZZY CLASSIFIERS ADOPTED

In the following subsections fuzzy classifiers adopted as well as similarity and dissimilarity measures applied have been describes from 4.1 to 4.7.

### 4.1 Fuzzy *c*-Means (FCM)

Fuzzy *c*-Means (FCM) method is a partitioning algorithm and is widely used in pattern recognition, medical imaging and remote sensing. It calculates the membership values, which gives the degree of sharing of a single pixel to different land cover classes. Objective function for the FCM classifier [3] has been mentioned in equation (2).

$$J_m(U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c} \mu_{ij}{}^{m} D(X_j, V_i) \quad (2)$$

Where, n is the total number of the pixels, c is the number of classes, $\mu_{ij}$ the fuzzy membership value of the ith pixel for class j, m is the weighing exponent 1<m<∞, $X_j$ is the vector pixel value, $V_i$ is the mean vector of a class and $D(X_j, V_i)$ is a similarity and dissimilarity measures as described in equation (8) to equation (16).

Classification results of the popular FCM classifier was found inaccurate in the presence of noise and outliers. To solve this issue associated with FCM, a Possibilistic *c*-Means algorithm (PCM) was developed by slightly modifying the objective function of FCM and relaxing the membership restriction which exists for FCM.

### 4.2 Possibilistic *c*-Means (PCM)

PCM was developed to address the drawback of FCM. It assigns a pixel to more than one cluster in the form of membership value and this membership value does not follow the constraint in FCM. It was developed by adding an additional term called regularizing term to the objective function of FCM. Objective function for the PCM classifier [3] has been mentioned in equation (3).

$$J_m(U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c} \mu_{ij}{}^{m} D(X_j, V_i) + \sum_{i=1}^{c}\eta_i \sum_{j=1}^{n}(1-\mu_{ij}{}^{m}) \quad (3)$$

Where $\eta_j$ is a parameter that depends on the distribution of pixels in the cluster, "j". However, the basic problem associated with PCM is proper accounting of noise pixels.

### 4.3 Noise Clustering (NC)

The concept of Noise Cluster was introduced such that noisy data points may be assigned to noisy class. This method is fundamentally based on FCM, where an additional cluster is introduced such that is supposedly contains all outliers. Objective function of the NC classifier [3] has been mentioned in equation (4).

$$J_m(U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c} \mu_{ij}{}^{m} D(X_j, V_i) + \sum_{j=1}^{n} \mu_{j,c+1}{}^{m}\delta \quad (4)$$

$$D_{j,c+1} = \delta \quad (5)$$

Where U= n×c+1matrix, the noise information class has no centre and the dissimilarity $D_{j,c+1}$ between $X_j$ and this noise information class can be expressed as in equation (5). $\delta > 0$ is a fixed parameter, also known as resolution parameter.

### 4.4 Noise Clustering with Entropy (NCE)

Another group of classifiers considers hybridization, which has its origin in Entropy theory. The term entropy was first introduced by Rudolf Clausius to state the second law of thermodynamics. A typical regularization is done by adding a regularization function. The objective function of NCE [3] has been mentioned in equation (6).

$$J_m(U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c} \mu_{ij}{}^{m} D(X_j, V_i) + \sum_{j=1}^{n} \mu_{j,c+1}{}^{m}\delta + \nu \sum_{i=1}^{c+1}\sum_{j=1}^{n} \mu_{ij}\log\mu_{ij} \quad (6)$$

Where $\nu$ is regularizing parameter, $\nu$ >0.

### 4.5 Fuzzy c-Means with Entropy (FCME)

FCME is a hybridization approach of a classification where the emphasis is to integrate entropy based regularization method with FCM. The objective function for FCME approach [3] has been mentioned in equation (7).

$$J_m(U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c} \mu_{ij}{}^{m} D(X_j, V_i) + \nu \sum_{i=1}^{c+1}\sum_{j=1}^{n} \mu_{ij}\log\mu_{ij} \quad (7)$$

Where $\nu$ is regularizing parameter, $\nu$ >0. Detail of different measures of similarity and dissimilarity $D(X_j, V_i)$ has been described in equation (8) to equation (16).

### 4.6 Dissimilarity Measures

Dissimilarity between two sequences of measurement is a measure that quantifies the independency between the sequences. A dissimilarity measure D is considered a metric if it produces a higher value as corresponding values in sequences become less dependent. Dissimilarity measures applied with fuzzy classifiers have been mentioned in equation (8) to equation (14).

Manhattan: Manhattan norm or sum of absolute intensity differences is one of the oldest dissimilarity measures used to compare images. It has been given as equation (8).

$$D(X_j, V_i) = \sum Abs(X_j, V_i) \quad (8)$$

Chessboard: It is also called chebyshev distance, maximum metric, it is a metric defined on a vector space. It has been mentioned in equation (9).

$$D(X_j, V_i) = Max[Abs(X_j, V_i)] \quad (9)$$

Braycurtis: It is directly related to Sorenson similarity index. It is used to quantify the compositional dissimilarity. Braycurtis has been given as equation (10).

$$D(X_j, V_i) = \frac{\sum Abs[X_j - V_i]}{\sum Abs[X_j + V_i]} \quad (10)$$

Canberra: It is a numerical measure of the distance between pairs of points in a vector space. It is weighted version of Manhattan distance. Formula of Canberra has been given as equation (11).

$$D(X_j, V_i) = \sum \frac{Abs[X_j - V_i]}{Abs[X_j] + Abs[V_i]} \quad (11)$$

Mean-Absolute-Difference: It is also known as Gini mean absolute difference equation. It is a measure of statistical dispersion. It has been mentioned in equation (12).

$$D(X_j, V_i) = \frac{1}{b}[\sum Abs(X_j - V_i)] \quad (12)$$

Median-Absolute-Difference: To reduce the effect of impulse noise on the calculated dissimilarity measure, instead of the average of absolute differences, the median of absolute differences (MAD) may be used to measure the dissimilarity between two images. MAD has been mentioned in equation (13).

$$D(X_j, V_i) = Median[Abs(X_j - V_i)] \quad (13)$$

Normalized-Squared-Euclidean: In this, normalizes the measure with respect to image contrast. Normalized-Squared-Euclidean has been mentioned in equation (14).

In the calculation of correlation coefficient, scale normalization is performed once after calculating the inner product of the normalized intensities.

$$D(X_j, V_i) = \frac{\sum \begin{matrix} Abs[X_j + \frac{1}{b}(\sum - X_j) \\ -V_i + \frac{1}{b}(\sum V_i)^2] \end{matrix}}{\begin{matrix} 2[\sum Abs\{X_J + \frac{1}{b}(\sum - X_j)^2\} \\ + \sum Abs\{V_i + \frac{1}{b}(\sum - V_i)^2\}] \end{matrix}} \quad (14)$$

### 4.7 Similarity Measures

Similarity between two sequences of measurement is a measure that quantifies the dependency between the sequences. Two similarity measures (cosine, correlation) applied with fuzzy classifiers have been mentioned in equation (15) and equation (16).

Cosine: It is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. It is most commonly used in high dimensional positive spaces like information retrieval and text mining.

$$D(X_j, V_i) = 1 - \frac{\sum X_j V_i}{\sqrt{\sum Abs[X_j]^2} \sqrt{\sum Abs[V_i]}^2} \quad (15)$$

Correlation: In this case similarity between two items is measured by computing the Pearson-r correlation. It is presented as a normalized form of covariance.

$$D(X_j, V_i) = 1 - \frac{\begin{matrix} [\{X_j + \frac{1}{b}(\sum - X_j)\} * \\ \{V_i + \frac{1}{b}(\sum - V_i)\}] \end{matrix}}{\begin{matrix} \sqrt{\sum Abs[X_j + \frac{1}{b}(\sum - X_j)^2]} * \\ \sqrt{\sum Abs[V_i + \frac{1}{b}(\sum - V_i)^2]} \end{matrix}} \quad (16)$$

Where, b denotes the number of bands in image.

## 5. DATA USED AND STUDY AREA

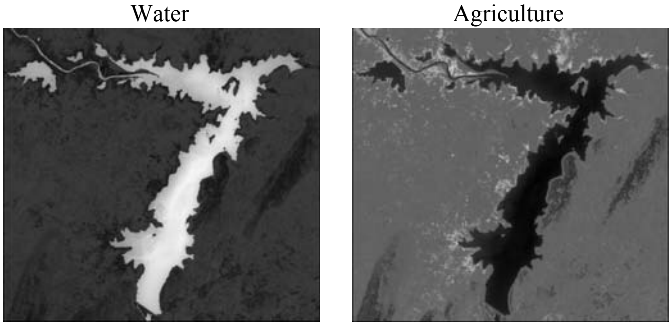
Salem district (11° 39' 0" North, 78° 10' 0" East) is a district of Tamil Nadu state in southern India, has been selected as the study area for testing SMIC Tool with similarity and dissimilarity measures. It has a reservoir namely Stanley also called Mettur Dam. The maximum percentage of water requirements for irrigation in Tamil Nadu depends on the Mettur Dam.

Landsat 8, Operational Land Imager (OLI) sensor data has been used to test the implemented similarity and dissimilarity measures. Landsat 8 OLI bands 1 to 7 and 9 with a spatial resolution of 30 meters and Thermal Infrared Sensor (TIRS) sensor consist of ten spectral bands. Band 1 (ultra-blue) is useful for coastal and aerosol studies. Band 9 is useful for cirrus cloud detection. The resolution for Band 8 (panchromatic) is 15 meters. Thermal bands 10 and 11 are useful in providing more accurate surface temperatures and are collected at 100 meters. From Landsat 8 data, Bands 1 to 7 and 9 has been used only for testing the algorithm.

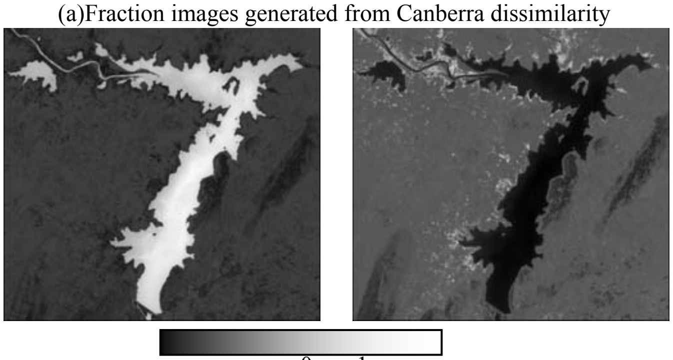
## 6. DESCRIPTION OF OUTPUTS

SMIC system generates the output in the form of fraction images in such a way that certain land cover classes are clearly represented in the resulting image. For an example fraction images generated using Fuzzy c-Means algorithm with dissimilarity measures (Canberra and Manhattan) and similarity measures (Cosine and Correlation) had been shown in Fig. 2 and Fig. 3 at m=2.4 and output pixel value were represented between 0 and 1 membership values. In this study two land cover classes water and agriculture were extracted.

Water Agriculture



(a)Fraction images generated from Canberra dissimilarity

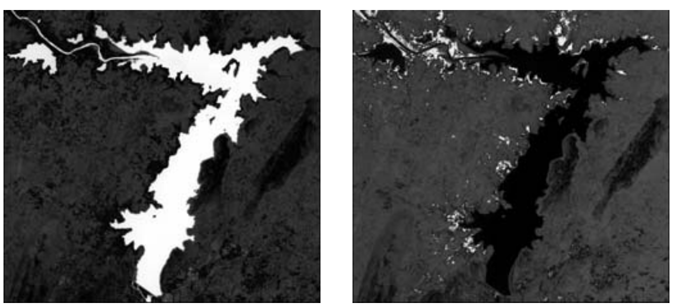


0 μ 1

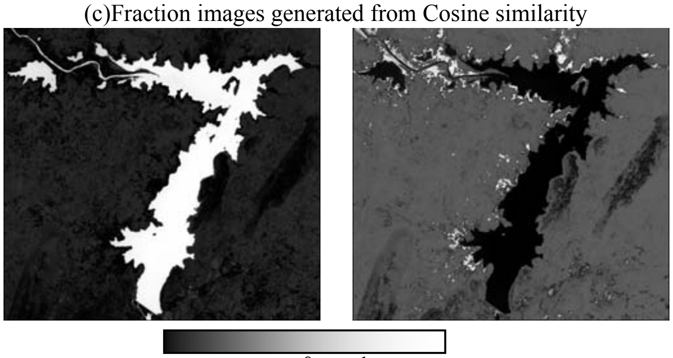
μ = Membership Value

(b)Fraction Images generated from Manhattan similarity

**Fig. 2: Poor results generated from (a) Canberra and (b) Manhattan dissimilarity measures.**



(c)Fraction images generated from Cosine similarity



0 μ 1

μ = Membership Value

(d)Fraction Images generated from Correlation similarity

**Fig. 3: Best results generated from (c) Cosine and (d) Correlation similarity measures.**

## 7. CONCLUSION

Performance of similarity and dissimilarity with fuzzy based classifier has been tested on Landsat 8, OLI sensor data and two classes water and agriculture has been extracted. Water bodies in input images, were homogeneous. Fraction images (output image) of water bodies generated from Canberra and Manhattan dissimilarity measures shows less homogeneity (high variance) while Cosine and Correlation similarity measures provides fraction images with more homogeneity (less variance) which is closer to ground information about water body. Also, Canberra and Manhattan has assigned less membership values to favourable class while Cosine and Correlation has assigned high membership values to favourable class. Thus, with respect to membership values and homogeneity, among all dissimilarity and similarity measures which have been defined in this system, Canberra and Manhattan provides poor results and Cosine and Correlation generates best results. Developed SMIC tool with similarity and dissimilarity measures in fuzzy classifiers is in JAVA programming language so it is platform independent. It has a graphical user friendly environment so that any resources management professionals can easily use this system and can study various similarity and dissimilarity measures for fuzzy classifiers in single or composite mode.

## 8. ACKNOWLEDGEMENTS

The first author is thankful to IIRS ISRO for allowing to doing her six month M.Tech project on study of similarity and dissimilarity measures with fuzzy based classifiers for which the tool has been developed and supported throughout the research process.

## REFERENCES

[1] Ablin R, and Sulochna C.H, "A Survey of Hyperspectral Image Classification in Remote Sensing", *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 8., 2013.

[2] Patra S. and Bruzzone L, "A Fast Cluster Assumption Based Active-Learning Technique for Classification of Remote Sensing Images", *Geoscience and Remote Sensing, IEEE Transactions on,* vol 49, 2011, pp. 1617-1626.

[3] Dwivedi R.K, *Fuzzy Hybrid Approaches for Soft Classification of Satellite Images*, Indian Institute of Technology Roorkee, Roorkee, India, unpublished Doctoral Thesis, 2014.

[4] Foody, G.M., "Approaches for the Production and Evaluation of Fuzzy Land Cover Classifications from Remotely Sensed Data", *International Journal of Remote Sensing*, vol. 17, no. 7, 1996, pp. 1317-1340.

[5] Foody, G.M., "The Role of Soft Classification Techniques in Refinement of Estimates of Ground Control Points Location", *Photogrammetric Engineering and Remote Sensing,* 68, 2002, pp. 897-903.

[6] Foody, G.M., Lucas,R.M., Curran, P.J. and Honzak M., "Non-Linear Mixture Modelling without End-Members using an Artificial Neural Network", *International Journal of Remote Sensing,* vol. 18, no. 4, 1997, pp. 937-953.

[7] Gopinath R.A., "Maximum Likelihood Modeling with Gaussian Distributions for Classification", in *Proceedings International Conference on Speech Signal Processing,* Seattle, WA, 1998, pp. 661-664.

[8] Harikumar A., *Effect of Discontinuity Adaptive MRF Models with Noise Classifier*, University of Twente, Enschede, The Netherlands & Indian Institute of Remote Sensing, ISRO, Dehradhun, India, published M.Sc Thesis, 2014.

[9] Huang. C, Davis L.S. and Townshend J.R.G., "An Assessment of Support Vector Machines for Land Cover Classification", *International Journal of Remote Sensing*, vol 23, no. 4., 2002, pp. 725-749.

[10] Kumar A., Gosh S.K. and Dadhwal V.K., "Sub-Pixel Land Cover Mapping: SMIC System", *ISPRS International Symposium on Geospatial Databases for Sustainable Development, Goa*, India, September 27-30, 2006.

[11] Rhee F.C.H. and Hwang C., "A Type-2 Fuzzy c Means Clustering Algorithm", in *Proceedings in Joint 2$^{th}$ IFSA world congress and 20$^{th}$ NAFIPS International conference*, 2001, pp.1226-1222.

[12] Safavian S.R., and Landgrebe D., "A Survey of Decision Tree Classifier Methodology". *IEEE Transactions on Systems, Man and Cybernetics,* 21, 1991, pp. 660-674.

[13] Tso B., and Mather P.M., *Classification Methods for Remotely Sensed Data,* CRC Press, Boca Raton, 2009, pp. 125-183.

[14] Upadhyay P., Ghosh S.K. and Kumar A., "A Brief Review of Fuzzy Soft Classification and Assessment of Accuracy Methods for Identification of Single Land Cover", *Studies in Surveying and Mapping Science (SSMS),* vol. 2, 2014.

[15] Yannis S.A., and Stefanos D.K., "Fuzzy Image Classification using Multi- Resolution Neural with Applications to Remote Sensing" *Electral Engineering Department, National Technical University of Athens*, Zographou 15773, Greece, 1999.