

Probabilistic Distance Clustering: A Spatial Clustering Approach for Remote Sensing Imagery

Y. Jayababu
Associate professor,
Department of CSE
Pragati Engineering College,
Andhra Pradesh, India

G.P.S. Varma
Head, Department of
Information Technology,
S R K R Engineering College,
Andhra Pradesh, India

A. Govardhan
Director, School of Information
Technology
Jawaharlal Nehru
Technological University
Hyderabad, India

ABSTRACT

Spatial clustering has been widely applied in various applications, especially in remote sensing technology. Clustering the geographical nature of the remote sensing imagery is challenging due to its wide and dense spatial distribution. Renowned clustering algorithms such as k-means and other probabilistic clustering algorithms have been reported in the literature. However, they are not robust to handle such peculiar data distribution. This paper employs probabilistic d – clustering algorithm to cluster the geographical information of the remote sensing imagery. The methodology considers diverse neighborhood connectivity and degree of connectivity to investigate the performance of probabilistic d – clustering algorithm. Experimental investigation demonstrates that probabilistic d – clustering algorithm is better than k – means clustering algorithm in handling remote sensing imagery.

Keywords

Spatial clustering; probabilistic d – clustering; remote sensing; geographic clustering

1. INTRODUCTION

Clustering is the significant task in an image segmentation method [1], which refers to a process of grouping or partitioning an image into non-overlapping segments with the following constraints: (i) each segment should be homogenous and (ii) union of two adjacent segments should be heterogeneous [2]. Since image segmentation plays a crucial role in image classification [3] [4], the significance of clustering remains high [5]. However, the traditional clustering algorithms consider each pixel as individual data element and so these methods are vulnerable under noisy environment. Hence, spatial clustering does a main job to overcome this problem, because it exploits the spatial relationship between the pixels [1].

Spatial clustering finds numerous applications in remote sensing [6], biomedical engineering [7] and many other fields. Advancements in GPS and smart mobile devices have increased the volume of data with spatial attributes because they utilize location sensitive information. However, clustering such spatial data is not a simple task because of arbitrarily distributed nature, noisy environment and large volume of hidden information.

Researchers pay great attention towards clustering such spatially distributed data in an efficient manner [8]. Few spatial clustering algorithms are combined with smoothing filters to make the clustering robust against noise [9] [10]

[11], while texture and spectral analysis have also been introduced in other works [12] [13].

Since such preprocessing methods may degrade the significant information, achieving a remarkable level of clustering efficiency remains a challenging task [1]. Hence, dedicated spatial clustering methods have been reported in the literature [14] [15] [16] [17].

2. RELATED WORKS

Remote sensing is one of the key areas, where spatial clustering methods can be essentially adopted [6]. Numerous spatial clustering algorithms have been reported in the literature. Few of them are briefly reviewed here.

Yong-Qiang Zhao *et al* [6] have proposed a band-subset-based clustering to improve the accuracy of hyperspectral image classification. In their method, they have segmented the hyperspectral data into numerous uncorrelated subsets, followed by determining the confidence of each subset using an eigenvalue – based approach. Segments with arbitrary shapes in both the spatial and the spectral domain have been extracted using a nonparametric technique. The hyperspectral classification has been performed based on the spectral reference of each cluster. Experimental investigation on Hyperspectral Digital Imagery Collection Experiment (HYDICE) has demonstrated the improved performance of their method over the spatially constrained fuzzy c-means clustering method. Moreover, their method has been proved to offer more robustness against noise than the supervised K-Nearest Neighbor (KNN) classifier.

Vahid Akbari *et al* [18] have unified the statistical distribution with the spatial contextual information to introduce an unsupervised clustering algorithm for multilook polarimetric synthetic aperture radar (PolSAR). This algorithm was based on the Markov random field (MRF) model that is an integrated version of K-Wishart distribution and Potts model for the PolSAR data statistics and spatial context. Moreover, the algorithm has utilized stochastic expectation maximization (SEM) algorithm to solve the clustering problem, in addition to estimating the parameters of MRF and K-Wishart distribution model. The performance of the algorithm was assessed through experiments on real as well as simulated PolSAR data.

Shuyuan Yang *et al* [19] have introduced an assumption for relaxed clustering and a laplace regularizer to work in the spatial domain. Based on these, they have performed the semi-supervised hyperspectral image classification. Since they have worked on both the noisy and mixed hyperspectral

image pixels, the clustering assumptions have been relaxed and so, similar hyperspectral vectors are allowed to share similar labels rather than the same labels. A spatial regularizer has also been constructed by considering the spatial homogeneity assumption. AVIRIS data has been used to investigate the performance of their method and it is found that the performance has improved than that of the state-of-the-art methods, just for a sample volume of training data.

Yanfei Zhong *et al* [20] have proposed an adaptive fuzzy clustering algorithm with spatial information (AFCM_S1), in which an objective function with adaptive spatial information weight has been introduced based on the entropy. Further, they have introduced an adaptive memetic fuzzy clustering algorithm with spatial information (AMASFC) that solves the clustering problem as an optimization problem. The memetic algorithm has been formulated as a combination of differential evolution algorithms and Gaussian local search (GLS) method. Experimental investigations have demonstrated the performance of their algorithm over the conventional algorithms.

Jian Ji and Ke-Lu Wang [21] have addressed the challenges on using the conventional Fuzzy c-means (FCM) algorithm to segment the synthetic aperture radar (SAR) images. They have overcome the problem by introducing a nonlocal fuzzy clustering algorithm with between-cluster separation measure (NS_FCM), in which an adaptive binary weighted distance measure and the adaptive filtering degree parameters have been used along with a fuzzy between-cluster variation term. By minimizing the NS_FCM objective function, they have simultaneously maximized the within-cluster compactness measure and the between-cluster separation measure of the segments. This supports the distance between cluster centers to be adjusted flexibly, when the fuzzy between-cluster variation term has been regulated. They have proved the performance of their algorithm on both the real and the synthetic SAR images.

3. PRELIMINARIES

3.1 Problem formulation

Since spatial clustering poses great challenge to the researchers, the scope of working on it remains wide. K-means clustering is the popular clustering methods that have been widely used for this purpose [22]. However, there are few significant challenges reside with the K-means clustering [23]. Probabilistic clustering addresses these challenges and overcomes the problem well [23]. Since, probabilistic clustering has outperformed, numerous variants have been reported in the literature.

3.2 Contribution

Recently, probabilistic distance clustering has been reported in the literature. Probabilistic distance clustering works based on joint distance function, rather than traditional probabilistic models [24]. Since it has been introduced recently, the algorithm has been found in limited applications. Hence it has not been applied much to cluster remote sensing imagery.

4. GEOGRAPHICAL CLUSTERING METHODOLOGY

4.1 Preprocessing

The proposed geographical clustering, as given in Figure 1, is comprised of three stages namely, preprocessing, constructing clustering data and clustering geographical information. Let us represent the input remote sensing image as

$$I_{mn} : 0 \leq m \leq M - 1, 0 \leq n \leq N - 1.$$

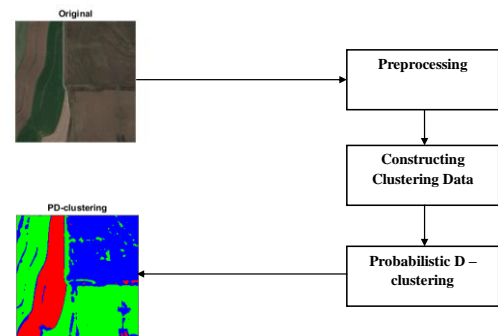


Fig 1: Proposed architecture of geographical clustering

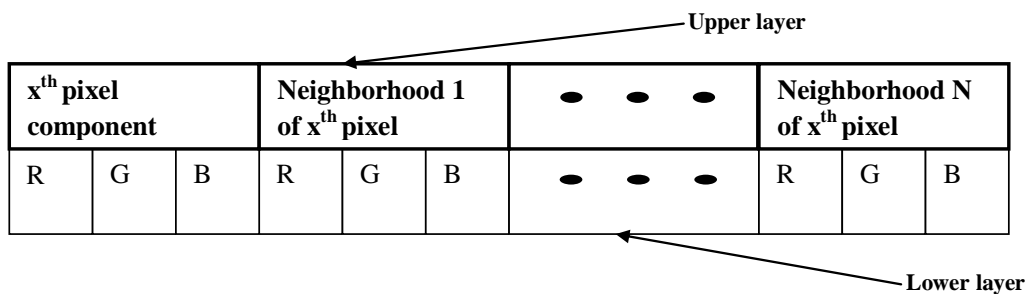


Fig 2: Anatomy of N-neighborhood clustering data

Input	C_d
Output	C_L
Process	Geographic clustering using Probabilistic D - Clustering
Step 1	Initialize $\{c\}_k : 0 \leq k \leq N_c - 1$
Step 2	Calculate $P_k(C_d)$ using equation (5)
Step 3	Calculate C_L using equation (6)
Step 4	Determine $\{c\}_k^+$ using equation (7)
Step 5	Go to Step 2, if $\{c\}_k^+$ differs from $\{c\}_k$
Step 6	Terminate the process

Fig 3: Pseudo code of the probabilistic D – clustering for remote sensing imagery

I_{mn} is a matrix of pixels in three dimensions that can be represented as I_{mn}^R, I_{mn}^G and I_{mn}^B , where, M and N refers to the height and width of the image, respectively. Individual pixel component can be referred here as $i(m,n):(0,255)$. The acquired image is in any of the size as per the setting of the acquisition device. However, to process the image in a common platform, all the images should have similar height and width, i.e., $M_1 = M_2 = \dots = M_H$ and $N_1 = N_2 = \dots = N_H$, where H is the number of images subjected to geographical clustering. Since the images can be corrupted by noise, median filtering is applied over the resized images under 3×3 neighborhood to obtain $[I_1]_{M \times N}$.

4.2 Constructing Clustering data

Clustering data refers to the actual data to be clustered in which the data to be clustered is constituted by multiple attributes. Since the primary objective is to assign clustering label to every pixel, each pixel is defined by attributes such as its intensity as well as its neighborhood intensity. Let the clustering data be represented as $\{C_D\}_{xy} : 0 \leq x \leq C_M, 0 \leq y \leq C_N$, where C_M and C_N are the number of records and attributes of the clustering data, respectively. Given the clustering image, $[I_1]_{M \times N}$, C_M and C_N can be determined as follows.

$$C_M = MN - 2(M + N) + 4 \quad (1)$$

$$C_N = N_c(N + 1) \quad (2)$$

In equation (2), N_c refers to the number of color spaces, which is 3 here, because the clustering image is in RGB color space. N given in equation (1) is different from N given in equation (2), which defines the neighborhood of the pixel, represented by m and n . Here, m and n can be referred as $f(x, M)$ and $f(x, N)$ as follows

$$m = \Gamma\left(\frac{x}{M}\right) \quad (3)$$

$$n = \Gamma\left(\frac{x}{N}\right) \quad (4)$$

In equation (3) and (4), $\Gamma(\bullet)$ represents floor function. Using these functions, corresponding pixel and the neighborhood information are extracted as attributes and hence, the clustering data is constructed. The anatomy of clustering data is depicted in fig. 2 in which the upper layers illustrate the number of neighborhoods considered in the clustering data C_D and the lower layer illustrates the color space components considered for every pixel.

4.3 Clustering geographical information

The geography of the subjected remote sensing information is clustered by clustering C_D using probabilistic D – clustering [25]. The probabilistic D – clustering intends to identify the likely probability of a data point to be associated with a particular cluster, which is defined by the probability function given in equation (5).

$$P_k(C_d(x)) = \frac{\prod_{j \neq k} D(C_d(x), c_j)}{\sum_{l=1}^{N_c} \prod_{j \neq l} D(C_d(x), c_j)}; 1 \leq k \leq N_c \quad (5)$$

where, N_c represents number of clusters, $P_k(C_d(x))$ refers to the likely probability of the x^{th} component to be associated with the k^{th} cluster, $D(A, B)$ represents distance between A and B and c_j represents centroid of the j^{th} cluster. The pseudo code of the probabilistic D – clustering algorithm is illustrated in fig. 3.

The $C_L \in (0, N_c - 1) : |C_L| = C_M \times 1$ in fig. 3 refers to the cluster labels for the data points that can be calculated as

$$C_L(x) = \arg \max_k P_k(C_d(x)) \quad (6)$$

The centroid updating formulation can be given as

$$\{c\}_k^+ = \sum_{x=1}^{C_M} \left(\frac{\mu_k(C_d(x))}{\Psi_k} \right) C_d(x) \quad (7)$$

$$\mu_k(C_d(x)) = \frac{P_k(C_d(x))^2}{D(C_d(x), c_k)} \quad (8)$$

$$\Psi_k = \sum_{x=1}^{C_M} \mu_k(C_d(x)) \quad (9)$$

where, $\mu_k(\bullet)$ and $\Psi_k(\bullet)$ given in equations (8) and (9) are membership and factorization functions, respectively. The termination criterion for the algorithm is given as The subjected images are indexed based on the obtained C_L to visualize the geographical nature of the image.

5. EXPERIMENTAL RESULTS

Simulation of the proposed method has been done in MATLAB and the performance investigation is carried out. The investigation has considered K-means, which is a renowned clustering algorithm, to ensure the competency of

the proposed methodology with probabilistic D – clustering algorithm.

This paper consider three types of neighborhood, namely 4-connected neighborhood, 4-connected diagonal neighborhood and 8-connected neighborhood. 4-connected neighborhood considers top, bottom, left and right pixels of the center pixel to construct the clustering data, whereas 4-connected diagonal neighborhood considers left bottom, left top, right bottom and right top pixels. The 8-connected neighborhood considers all the first degree neighbors of the center pixels.

The experimentation is carried out on ten images acquired from “Google Maps” at arbitrary altitude and they are subjected to geographical clustering. The clustering outcome of sample images is produced in fig. 4. Since it is complex to acquire the ground truth results, Davis-Bouldin (DB) index [26], which is an internal evaluation scheme, is considered here to study the performance of both the algorithms. The metric values obtained for the ten images are tabulated in Table I and its statistical study are given in Table II.

5.1 Discussion

From fig. 4, one can observe that clustering variations remain minor though the neighborhood connectivity varies. However, the DB index shows considerable variation between different neighborhood connectivity. For instance, for image 1, the DB – index of probabilistic D – clustering under 4-connected neighborhood is 1.08, whereas it has increased to 1.13 under 4-connected diagonal neighborhood. It is a 4% increase from the 4-connected neighborhood.

Similarly, the DB-index of the 4-connected diagonal neighborhood exhibited a 2% increase over the 8-connected neighborhood with respect to it. The relative performance between the probabilistic D – clustering and the K-means clustering given in Table I shows that the probabilistic D – clustering outperforms K-means clustering. This has also been ensured by Table II through the statistical analysis in which mean, median, best, worst and standard deviation metrics have been given. The mean clustering performance exhibited by probabilistic D – clustering is 13.8%

$$= \left(\frac{0.81 - 0.69}{0.81} \times 100 \right)$$
 more than the K-means clustering algorithm under 4-connected neighborhood.

Table 1. Probabilistic D- clustering versus K-means clustering based on DB-index obtained for different neighborhood connectivity

Images	4 – connected neighborhood		4 – connected diagonal neighborhood		8 – connected neighborhood	
	Probabilistic D – clustering	K – means clustering	Probabilistic D – clustering	K – means clustering	Probabilistic D – clustering	K – means clustering
1	1.0816	0.7064	1.1289	0.7567	1.1152	0.7421
2	0.7366	0.6209	0.7634	0.6543	0.7585	0.5716
3	0.6198	0.5665	0.6429	0.5945	0.6366	0.4912
4	1.2007	0.8962	1.2833	0.9726	1.2627	0.9516
5	0.8019	0.5658	0.8802	0.8941	0.8602	0.6044
6	0.7059	0.6839	0.7595	0.7336	0.7444	0.7196
7	0.6876	0.6139	0.8800	0.6822	0.8601	0.6000
8	0.8235	0.6538	0.9044	0.8976	0.8824	0.6869
9	0.8255	0.8251	0.9167	0.7149	0.8918	0.6929
10	0.6369	0.8668	1.0269	0.9881	1.0008	0.9574

Table 2. Probabilistic D- clustering versus K-means clustering based on statistical analysis of DB-index

Statistical measures	4 – connected neighborhood		4 – connected diagonal neighborhood		8 – connected neighborhood	
	Probabilistic D – clustering	K – means clustering	Probabilistic D – clustering	K – means clustering	Probabilistic D – clustering	K – means clustering
Mean	0.8120	0.6999	0.9186	0.7889	0.9013	0.7018
Median	0.7693	0.6688	0.8923	0.7452	0.8713	0.6899
Best	1.2007	0.8962	1.2833	0.9881	1.2627	0.9574
Worst	0.6199	0.5658	0.6429	0.5945	0.6366	0.4912
Standard deviation	0.1898	0.1220	0.1878	0.1387	0.1841	0.1532


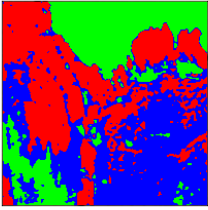
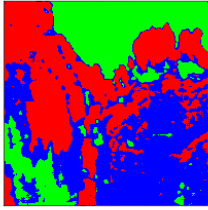
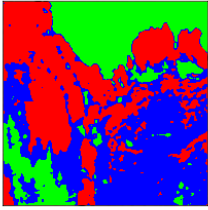
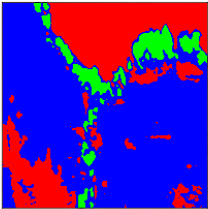
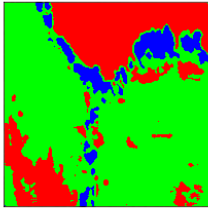
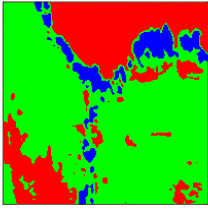
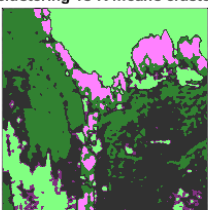
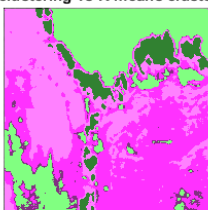
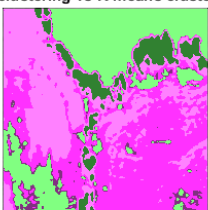
Original Image	<div style="text-align: center;">Original</div> 		
Neighborhood	4-neighborhood	4-neighborhood diagonal	8-neighborhood
Outcome from probabilistic D – clustering	<div style="text-align: center;">PD-clustering</div> 	<div style="text-align: center;">PD-clustering</div> 	<div style="text-align: center;">PD-clustering</div> 
Outcome from K-means clustering	<div style="text-align: center;">K-means clustering</div> 	<div style="text-align: center;">K-means clustering</div> 	<div style="text-align: center;">K-means clustering</div> 
Comparison between K-means and Probabilistic D – clustering	<div style="text-align: center;">PD-clustering vs K-means clustering</div> 	<div style="text-align: center;">PD-clustering vs K-means clustering</div> 	<div style="text-align: center;">PD-clustering vs K-means clustering</div> 

Fig 4: Clustering results from probabilistic D – clustering and K-means clustering under three neighborhood conditions

Under 4-connected diagonal neighborhood and 8-connected neighborhood, the probabilistic D – clustering is 14% and 22% better than K-means clustering, respectively. Similarly, the other statistical metrics like median, best and worst cases have also demonstrate a considerable performance dominance by probabilistic D – clustering. However, the standard deviation is relatively higher in probabilistic D – clustering, which is a drawback. Nevertheless, the other metrics find rather sensitivity and hence the deviation can be neglected here.

6. CONCLUSION

This paper proposed a probabilistic D – clustering based spatial clustering method for remote sensing imagery. The methodology was developed in MATLAB and experimental investigation was carried out with ten images acquired from “Google Maps”. DB index was used to quantify the clustering performance and comparison was made with the renowned K-means clustering algorithm. The obtained results have

demonstrated that probabilistic D – clustering algorithm have exhibited more than 10% higher clustering performance than K-means clustering algorithm. This has also been ensure through first order statistical functions such as mean, median, best and worst cases, despite the standard deviation is relative higher than K-means clustering. The obtained results are encouraging to apply probabilistic D – clustering in other spatial clustering applications, where data density is found to be high.

7. REFERENCES

- [1] Li, N., Huo, H., Zhao, Y.M., Chen, X. and Fang, T. 2013. A Spatial Clustering Method with Edge Weighting for Image Segmentation. IEEE Geoscience and Remote Sensing Letters. 10 (Sep. 2013), 1124-1128.
- [2] Pal, N.R. and Pal, S.K. 1993. A review on image segmentation techniques. Pattern Recognit. 26 (Sep. 1993), 1277-1294.

- [3] Cloude, S.R. and Pottier, E. 1997. An entropy based classification scheme for land applications of polarimetric SAR. *IEEE Trans. Geosci. Remote Sens.* 35 (Jan. 1997), 68-78.
- [4] Van, Z. 1989. Unsupervised classification of scattering mechanisms using radar polarimetry data. *IEEE Trans. Geosci. Remote Sens.* 27 (1989), 36-45.
- [5] Lee, J.S., Grunes, M. and Kwok, R. 1994. Classification of multi-look polarimetric SAR imagery based on the complex Wishart distribution. *Int. J. Remote Sens.* 15 (1994), 2299-2311.
- [6] Zhao, Y.Q., Zhang, D. and Kong, S.G. 2011. Band-Subset-Based Clustering and Fusion for Hyperspectral Imagery Classification. *IEEE Trans. Geosci. Remote Sens.* 49 (Feb. 2011), 747-756.
- [7] Yu, D.J., Hu, J., Yang, J., Shen, H.B., Tang, J. and Yang, J.Y. 2013. Designing Template-Free Predictor for Targeting Protein-Ligand Binding Sites with Classifier Ensemble and Spatial Clustering. *IEEE/ACM Trans. Computational Biology and Bioinformatics.* 10 (Jul.-Aug. 2013), 994-1008.
- [8] Packer, E., Bak, P., Nikkila, M., Polishchuk, V. and Ship, H.J. 2013. Visual Analytics for Spatial Clustering: Using a Heuristic Approach for Guided Exploration. *IEEE Trans. Visualization and Computer Graphics.* 19 (Dec. 2013), 2179-2188.
- [9] Comaniciu, D. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 24 (May 2002), 603-619.
- [10] Ilea, D.E. and Whelan, P.F. 2008. CTex—An adaptive unsupervised segmentation algorithm based on color–texture coherence. *IEEE Trans. Image Process.* 17 (Oct. 2008), 1926-1939.
- [11] Kuwahara, K.H., Ehiu, S. and Kinoshita, M. 1976. Processing of riangiocardiographic images. In *Proceedings of the Digital Processing of Biomedical Images*. New York: Plenum, 187-203.
- [12] Dong, G. and Xie, M. 2005. Color clustering and learning for image segmentation based on neural networks. *IEEE Trans. Neural Netw.* 16 (Jul. 2005), 925-936.
- [13] Xiangrong, Z., Jiao, L., Liu, F., Bo, L. and Gong, M. 2008. Spectral clustering ensemble applied to SAR image segmentation. *IEEE Trans. Geosci. Remote Sens.* 46 (Jul. 2008), 2126-2136.
- [14] Tarabalka, Y., Benediktsson, J.A. and Chanussot, J. 2009. Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques. *IEEE Trans. Geosci. Remote Sens.* 47 (Aug. 2009), 2973-2987.
- [15] Liew, A.W., Yan, H. and Law, N.F. 2005. Image segmentation based on adaptive cluster prototype estimation. *IEEE Trans. Fuzzy Syst.* 13 (Aug. 2005), 444-453.
- [16] Dulyakarn, P. and Rangsanteri, Y. 2001. Fuzzy C-means clustering using spatial information with application to remote sensing. In *Proceedings of the 22nd Asian Conference on Remote Sens.*, Singapore.
- [17] Ahmed, M.N., Yamany, S. M., Mohamed, N., Farag, A.A. and Moriarty, T. 2002. A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data. *IEEE Trans. Med. Imag.* 21 (Mar. 2002), 193-199.
- [18] Akbari, V., Doulgeris, A.P., Moser, G., Eltoft, T., Anfinson, S.N. and Serpico, S.B. 2013. A Textural–Contextual Model for Unsupervised Segmentation of Multipolarization Synthetic Aperture Radar Images. *IEEE Trans. Geosci. Remote Sens.* 51 (Apr. 2013), 2442-2453.
- [19] Yang, S., Qiao, Y., Yang, L., Jin, P.L. and Jiao, L. 2014. Hyperspectral Image Classification Based on Relaxed Clustering Assumption and Spatial Laplace Regularizer. *IEEE Geoscience and Remote Sensing Letters.* 11 (May 2014), 901-905.
- [20] Zhong, Y., Ma, A. and Zhang, L. 2014. An Adaptive Memetic Fuzzy Clustering Algorithm with Spatial Information for Remote Sensing Imagery. *IEEE J. Selected Topics in Applied Earth Observations and Remote Sensing.* 7 (Apr. 2014), 1235-1248.
- [21] Ji, J. and Wang, K.L. 2014. A Robust Nonlocal Fuzzy Clustering Algorithm with Between-Cluster Separation Measure for SAR Image Segmentation. *IEEE J. Selected Topics in Applied Earth Observations and Remote Sensing.* 7 (Dec. 2014), 4929- 4936.
- [22] Wang, J. and Su, X. “An improved K-Means clustering algorithm”, 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN), Page(s): 44 – 46, 2011
- [23] Bacher, J. 2000. A Probabilistic Clustering Model for Variables of Mixed Type. *Quality and Quantity.* 34 (Aug. 2000), 223-235.
- [24] Iyigun, C., Ben-Israel, A. 2009. Semi-supervised Probabilistic Distance Clustering and the Uncertainty of Classification. *Advances in Data Analysis, Data Handling and Business Intelligence, Studies in Classification, Data Analysis, and Knowledge Organization*, 3-20.
- [25] Ben-Israel, A., Iyigun, C. 2008. Probabilistic D-Clustering. *J. Classification.* 25 (Jun. 2008), 5-26.
- [26] Davies, D.L., Bouldin, D.W. 1979. A Cluster Separation Measure. *IEEE Trans. Pattern Analysis and Machine Intelligence.* PAMI-1(Apr. 1979), 224-227.