End semester report on R & D Project (NU 302)
Academic Year- 2019-20

On

INTELLIGENT ANALYSIS ON SATELLITE IMAGERY

A dissertation

Submitted in partial fulfillment of the requirements for the award of the degree

Bachelor of Technology

by

Anjali Mishra	BT17GCS157	CSE
Arundhati Das	BT17GCS016	CSE
Manaal Soni	BT17GCS052	CSE
Yash Kejriwal	BT17GCS123	CSE
Himanshu Nikhare	BT17GCS036	CSE

Under supervision of
Dr. Sudip Sanyal

CERTIFICATE BY SUPERVISOR(S)

This is to certify that the present R&D project entitled Intelligent Analysis on Satellite Imagery being submitted to NIIT University, Neemrana, in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology, in the area of BT/CSE/ECE/GIS, embodies faithful record of original research carried out by ----------. They have worked under my guidance and supervision and that this work has not been submitted, in part or full, for any other degree or diploma of NIIT or any other University.

Place:

Name of the Supervisor(s) with signature

Date:

ACKNOWLEDGEMENT

We are grateful to NIIT University for providing us an opportunity to undertake this project on "Intelligent Analysis on Satellite Imagery".We would like to express our sincere gratitude to our R&D mentor Dr.Sudip Sanyal, Mr.Prashant Srivastava & Mr.Thota Sivasankar for guiding us immensely through the course of this project. Their constructive advice, encouragement & constant motivation have been responsible for the completion of this project.

Place:Neemrana,Rajasthan

Date:15-07-2020

1)	**Anjali Mishra**	**BT17GCS157**	**CSE**
2)	**Arundhati Das**	**BT17GCS016**	**CSE**
3)	**Manaal Soni**	**BT17GCS052**	**CSE**
4)	**Himanshu Nikhare**	**BT17GCS036**	**CSE**
5)	**Yash Kejriwal**	**BT17GCS123**	**CSE**

NU
NIIT UNIVERSITY
THE UNIVERSITY OF THE FUTURE
Industry-linked | Technology-based | Research-driven | Seamless

DECLARATION BY STUDENT(S)

We hereby declare that the project report entitled Intelligent Analysis on Satellite Imagery which is being submitted for the partial fulfilment of the Degree of Bachelor of Technology, at NIIT University, Neemrana, is an authentic record of our original work under the guidance of Dr.Sudip Sanyal. Due acknowledgements have been given in the project report to all other related work used. This has previously not formed the basis for the award of any degree, diploma, associate/fellowship or any other similar title or recognition in NIIT University or elsewhere.

Place:Neemrana,Rajasthan

Date:15-07-2020

Anjali Mishra	**BT17GCS157**	**CSE**	
Arundhati Das	**BT17GCS016**	**CSE**	
Manaal Soni	**BT17GCS052**	**CSE**	
Himanshu Nikhare	**BT17GCS036**	**CSE**	
Yash Kejriwal	**BT17GCS123**	**CSE**	

Introduction:

Nowadays, satellite imagery is widely used in remote sensing and other research areas for various applications. Intelligent analysis of satellite images is a process of collecting and interpreting the satellite images from multiple satellites such as Sentinel-2B, Landsat-8, etc.The aim of the study is high resolution satellite imagery land cover classification with higher accuracy. Land cover classification is a well-studied problem in the domain of remote sensing.In this study we will compare different algorithms on the basis of their accuracy and find the most efficient one.The study includes comparison of different single class algorithms for single class classification and predicts the higher accurate results implementing the best defined approach.We'll begin by comparing different unsupervised learning techniques & then accordingly choosing the best technique for the classification. Future scope of the study is to extend the optimization criteria for multiple classes. We will also try to introduce several methods that will make the system more robust so that the same model can be used for a larger region.

Problem statement:

Analyzing the unsupervised learning techniques to identify clusters of an image & then comparing the algorithms to check which algorithms obtain the best results.

Literature review:

A.W.Abbas et. al 2016, this paper analyzes the two unsupervised classification algorithms ISODATA & K-means in remote sensing for the city of Abbottabad, Pakistan. These two algorithms evaluate statistically by iterative techniques to automatically group similar spectral features into unique clusters.The test region of Abbottabad is divide into five bands i.e. NDVI (Normalized Difference Vegetation Index), green, near infrared, far infrared, and green.The ROIs (regions of interest) selected for classification of land cover data comprises five different types of classes namely water bodies, agriculture,settled area, forest and barren land.In this paper the analysis started from data acquisition, pre-processing, then unsupervised classification by

K-means and ISODATA method was carried out, with final processing in post-classification and accuracy assessment has been done using SAGA GIS .

Lasri 2016, the study combines various papers and their conclusion to build one of its own. Self Organizing Maps can be widely used for clustering and classification. However, there are some loopholes when using this model alone. The paper suggests using SOM along with other techniques to understand whether the loopholes can be fixed or not. The paper performs two tests in which Self Organising Maps are used with different parameters. These tests show the faults in the SOM algorithm. The first test is performed on a 5x5 SOM with inputs(4x2 Matrix form) that have linear regularities. Based on this experimentation, SOM is incapable of distinguishing the objects that possess liner regularity. In cases of inputs having linear regularities the SOM generates two BMUs also called winner neurons. In the second test a SOM of dimension 8x3 and it was processed on inputs having linear dependencies. The results showed that SOM was again incapable of distinguishing inputs with linear dependencies however, it worked perfectly for other inputs.

Hemant Kumar Aggarwal & Sonajharia Minz 2015,this paper presents the three types of unsupervised learning techniques used for the change detection in water, vegetation and built-up land cover classes of Delhi region in India. Eight images were taken from satellite Landsat TM and Landsat ETM+ from the year 1998-2011 for preprocessing. Three features namely Soil Adjusted Vegetation Index(SAVI), Modified Normalized Difference Water Index (MNDWI), and Built-up from Normalized Difference Built-up Index (NDBI) were extracted at the preprocessing stage. These features corresponds to vegetation, water, and built-up classes respectively.The three clustering algorithms k means, fuzzy c mean and expectation maximization were selected to represent the partition based, fuzzy, and probability based technique respectively.The three algorithms were implemented to cluster the pixels of all the eight images using the above three features namely SAVI, MNDWI and NDBI.Those features enabled a 50% reduction in the dimensionality of data & the three algorithms also yielded good results.Based on the results obtained from the features & algorithms, it had been seen that vegetation has decreased every year whereas urban area has increased. Based on the measure of

silhouette coefficients,partition based clustering algorithm is more effective in comparison to probabilistic and fuzzy based clustering techniques and thus for change detection.

Usman 2013,the study mainly focuses on the use of K-means clustering algorithm to classify satellite imagery into specific objects within it.Segmentation and classification of high resolution satellite imagery is a challenging problem due to the fact that it is no longer meaningful to carry out this task on a pixel-by-pixel basis.K-means clustering algorithm is a better method of classifying high resolution satellite imagery. with the use of minimum distance decision rule the extracted regions are classified.The procedure significantly reduces the mixed pixel problem suffered by most pixel based methods and helps in getting better classification accuracy with an overall accuracy of 88.889%.

Balasubramanian, Sowmya & Sheelarani, B. 2011,this paper talks about a new reformed fuzzy c means technique(RFCM) for land cover classification. By comparing it with other techniques such as FCM , PFCM(Possibilistic fuzzy C means) on the basis of quality measures such as Kappa Coefficient,Peak signal to time ratio,compression ratio, execution time it has been found out that RFCM performs better land cover classification than other FCM techniques.

Gonçalves et. al 2011, this particular paper is on Land-Cover Classification Using Self-Organizing Maps Clustered with Spectral and Spatial Information. Using unsupervised learning is a good approach when there is very less prior information about the data. Unsupervised learning uses clustering methods for classifications. These methods examine the unknown pixels in an image and cluster them into different classes. The basis of selecting a method depends on image size and feature dimension. Self Organising Maps (SOM) is an unsupervised and competitive learning Artificial Neural Network model. SOM is a method which is used (i) When data to be clustered is unlabelled i.e., the number of classes are unknown and (ii) For dimension reduction of the data. This paper presents a 2 level SOM based clustering approach. The first level is SOM Training map original patterns of image to a reduced set of prototypes arranged in a 2D rectangular grid. Here two factors are kept in mind: Image sampling process and determination of the SOM training parameters. The second level is SOM segmenting the SOM output map using an additional clustering method. This approach reduces the computational load of the classification process. After training the SOM the second level will

automatically, without user interference, reduce the dimensions as per the need. Only two parameters that need to be defined by the user: size of samples and number of SOM neurons.

Shi Na et. al 2010,this paper discusses the standard k-means clustering algorithm and analyzes the shortcomings of standard k-means algorithm.One of the shortcoming of this algorithm is that the k-means clustering algorithm has to calculate the distance between each data object and all cluster centers in each iteration, which makes the reduces the efficiency of the algorithm.This paper proposes an improved k-means algorithm in order to solve the shortcomings of K-means algorithm.This paper presents a simple and efficient way for assigning data points to clusters. The proposed method in this paper ensures that the entire process of clustering is in O(nk) time without sacrificing the accuracy of clusters. Experimental results show that the improved algorithm can improve the execution time of the k-means algorithm. So the proposed k-means method is feasible.

Dehuri et. al 2006 the study focuses on the choice of the best clustering algorithm among three algorithms like K-means, Self Organizing Map (SOM) and Density Based Clustering for Applications with Noise (DBSCAN) .The experimental results showed that if the clusters are of arbitrary shape, a density based clustering algorithm (DBSCAN) is the most preferred, where as if the clusters are of convex shape ,hyper spherical and well-separated then the SOM or K-means is most preferred.

Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." *Kdd*. Vol. 96. No. 34. 1996. The algorithm groups together different meaning classes in the database. The author compares algorithms like CLARANS(Clustering Large Applications based on RANdomizedSearch) algorithm, Hierarchical algorithms and explores Jain's density based clustering approach to identify clusters in k-dimensional point sets. The proposed algorithm was found to be more effective than CLARANS by a factor of at least 100 in terms of efficiency, the evaluation of this is done on both synthetic and real data of the SEQUOIA 2000 benchmark.

Sa'ada, Nailus & Harsono, Tri & Basuki, Ahmad. (2019). Improvement of Segmentation Performance for Feature Extraction on Whirlwind Cloud-based Satellite Image using DBSCAN

Clustering Algorithm. EMITTER International Journal of Engineering Technology. 7. 10.24003/emitter.v7i1.372. In this study, the author used satellite images as input and segment it using DBSCAN to get useful information from it. And also compares K-means from DBSCAN in which DBSCAN seems to outperform meng hee heng's K-means in terms of finding centroid points of the cloud, it also requires 0.05 second to finish clustering whereas Kmeans need more than a second.

Proposed methodology:

A high resolution satellite imagery illustrating various types of land use and land cover will be used as the test image for classification.Initially the supplied image will be extracted from its compressed format of around 10 m resolution for the bands 5, 4, 3, and 2 of NIR, Red, Green and Blue and then the pixels of the image is clipped out and saved. A composite of Bands 4,3,2 will then be performed that will give an output and be saved as TIFF/GEOTIFF format. Now,first an unsupervised classification will be performed on the image using different algorithms to classify the image into the desired classes . A raster layer will be generated using the unsupervised algorithm while running QGIS/SAGA GIS. The pixels will be identified for each of the categories and then they will be grouped into land cover categories.

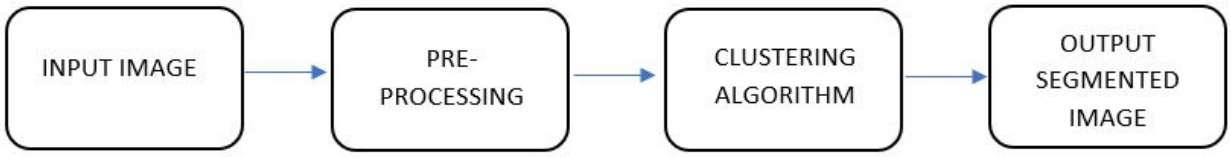


Figure 1: Proposed Methodology

Expansion/ Modification of the project:

We realised with research that it is difficult to find the training data set according to the workflow requirements. To deal with this the team along with guidance has planned to perform consistent labelling using unsupervised learning. Previously we planned to perform land cover classification using supervised learning on a labelled data set found online. However, now we

shall label our own data using unsupervised learning. We will indulge into various clustering algorithms and study them to identify the algorithm that gives most consistent labelling. Therefore, now before classification we will first find an image data set which may not be labelled. Till then the following are the added steps to the process:

To the the imaged:

1. Run clustering algorithms on the images and get labels
2. Study various clustering algorithms
3. Compare and check the consistency of every algorithm.
4. Apply fuzzy c-means on the data that doesn't get clustered. We assume this happens because an image might identify itself under more than one cluster.
5. Analyze the results and make the data as consistent as possible using clustering techniques.

After these steps we will have a data set ready to train our supervised learning model of classification.

Revised Deliverables:

1. Consistent labeling using unsupervised learning
2. Land Cover classification model.

- Workflow (detailed description of each module/item)

- Collecting Data
- Data Preprocessing
- Classification of Data(unsupervised)
- Post classification(error reduction)
- Accuracy assessment
- Result ,analysis & conclusion

- Technology

1) **Software Used** :
   i) QGIS 3.12.1
   ii) SAGAGIS
2) **Language:**
   i) Python 3.6
3) **Hardware Used:**

i) Laptop

4) **Dataset :**

i) UCI Machine learning landsat satellite Dataset

Algorithms Studied:

We have analyzed some of the unsupervised learning algorithms like K-means,ISODATA,Fuzzy c-means, DBSCAN, SOM.

A.] K-means Clustering:

It is one of the most common unsupervised learning clustering algorithm used to get an intuition about the structure of the data. It's an iterative algorithm that clusters, or partitions the given data into K-clusters or parts based on the K-centroids. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The less variation we have within clusters, the more similar the data points are within the same cluster.

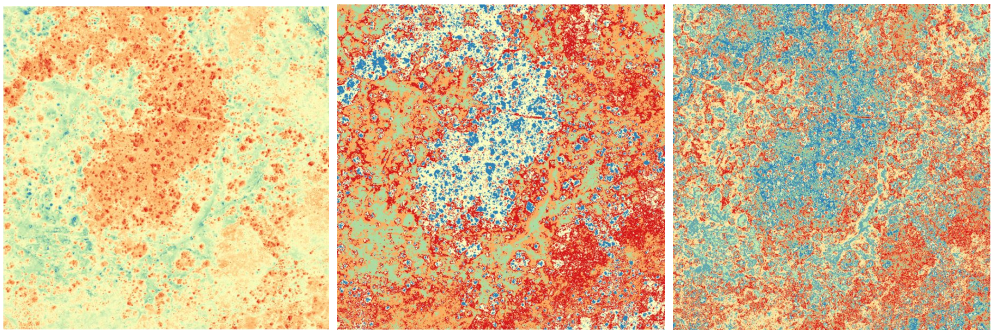
Algorithm:

1. Select the number of clusters K.
2. Choose randomly K points, the centroids.
3. Assign each data point such that it is assigned to the closest centroid
4. Compute the new centroid of each cluster and place it accordingly.
5. Reassign each data point to the new nearest centroid.
6. If any further reassignment of data point took place, go to step 4, otherwise, the model is ready.

Analysis:

The original image is being displayed below. On that, we performed K-means clustering by first using 5 clusters & then with 10 clusters. A significant change has been observed after increasing the cluster to 10. Increasing the number of clusters actually helps us to optimize the model in terms of how good it actually separates from other clusters.

Output:



Original image **Cluster=5** **Cluster=10**

Figure 2: K- means clustering

B.] ISODATA

The ISODATA classifier is a modified form of the K-means classifier, with the ability to split classes with too much variance and merge classes that are too similar between each iteration.

ISODATA Algorithm:

1) ISODATA computes class means consistently circulated in the data space before iteratively clusters the continuing pixels utilizing least distance approaches.
2) Every iteration recalculates means as well as reclassifies pixels through respect to the new means.
3) It may turn out later that more or fewer clusters would fit the data better.

Output:

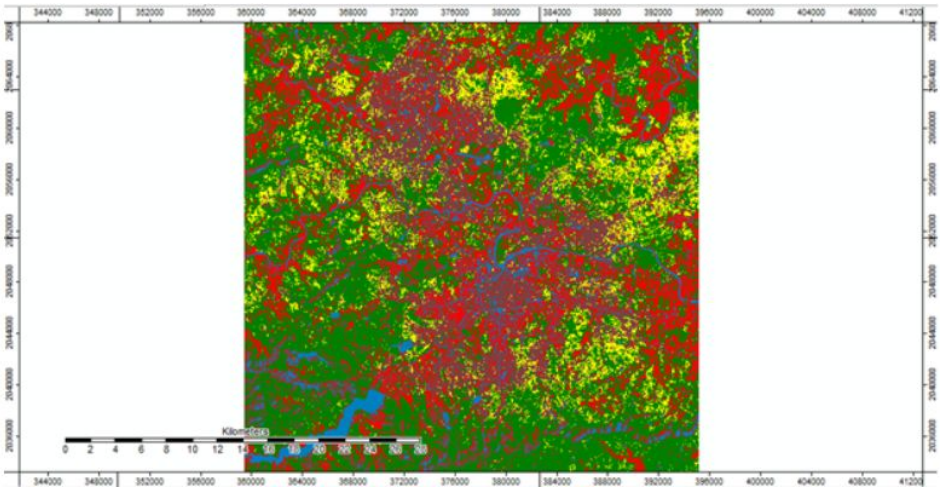


Figure 3: ISODATA Classification (Cluster =5)

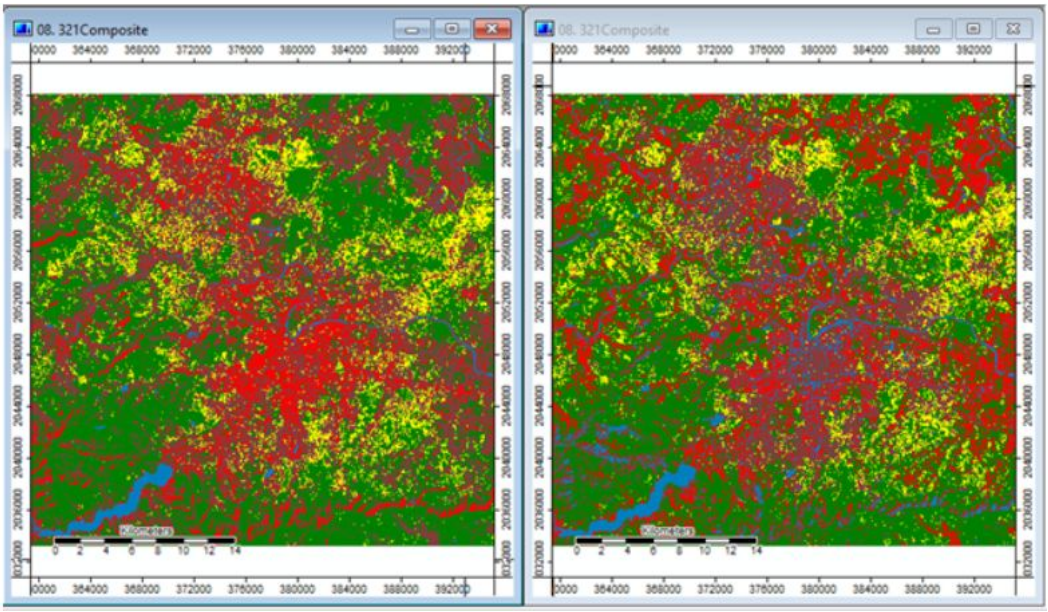


Figure 4:Comparison of K-Means & ISODATA Classification (Cluster=5)

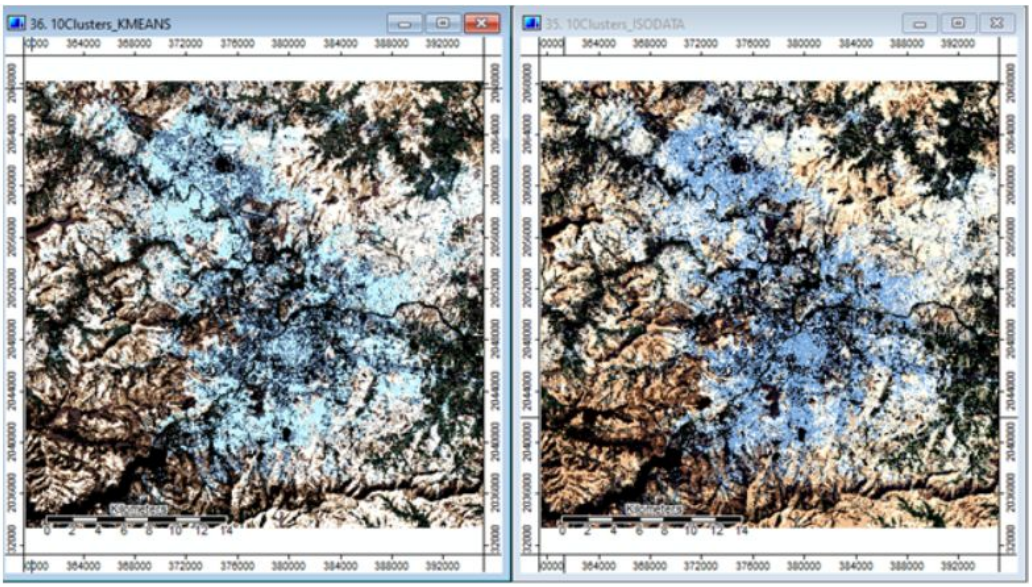


Figure 5:Comparison of K-Means & ISODATA Classification (Iterations=1000, 452 Composite)

	CLUSTER	ELEMENTS	MEANDIST	MEAN.Band1	STDV.Band1	MEAN.Band 2	STDV.Band 2	MEAN.Band 3	STDV.Band 3	MEAN.Band 4	STDV.Band 4	MEAN.Band 5
1	1	133838	1.018910	79.189348	2.531597	36.950709	1.648922	40.263991	2.667045	41.732154	3.409328	62.647821
2	2	65312	2.033308	67.239160	4.379753	28.930319	2.649862	28.231810	4.373313	21.309907	9.042953	39.280362
3	3	101260	1.141925	69.519050	2.280676	32.171252	1.331924	33.438722	2.488557	48.748621	7.772038	68.210369
4	4	52913	1.058667	73.251999	2.247781	35.424924	1.513008	39.533952	2.630184	57.339463	5.445938	82.149377
5	5	160553	1.114220	75.760901	2.411511	37.632153	1.418177	46.790331	2.821501	56.207508	4.157744	105.580855
6	6	99072	1.018440	70.835080	2.118965	33.486111	1.439983	39.552749	2.553885	50.723958	3.434474	92.292585
7	7	69805	0.879026	75.565418	2.079684	35.676470	1.308999	39.525507	2.231537	48.693647	3.093902	71.952797
8	8	104257	0.978488	78.428844	2.391938	37.932916	1.283630	44.405038	1.948270	46.948800	3.487820	79.777991
9	9	49259	0.920902	80.798880	1.842763	39.852433	1.241863	46.545890	2.175836	49.770052	3.869395	81.857772
10	10	96288	0.921232	73.562573	1.779603	34.869776	1.237870	40.120088	2.217071	47.237891	3.376087	80.416636
11	11	142828	1.014671	70.366203	2.812821	31.702320	1.418190	34.177976	2.138489	43.029021	3.351011	64.881263
12	12	82053	0.968945	79.081789	1.565763	39.873838	1.188354	49.773402	2.433386	54.176983	3.733822	100.539944
13	13	89855	1.049175	68.597585	2.951047	30.605542	1.517193	30.481097	2.335214	49.835557	4.940503	57.853920
14	14	39506	1.242859	86.886321	3.397978	40.025201	2.595388	49.472029	3.873407	48.520630	4.659101	77.657343
15	15	36714	1.681171	68.967233	3.714837	30.637196	2.500833	33.447595	3.968319	36.426132	6.947053	62.207332
16	16	61620	1.531857	80.233827	5.249441	43.284015	3.637329	53.450487	4.839551	55.241042	5.414834	99.049854
17	17	132965632	0.000000	67.239160	0.000000	28.930319	0.000000	28.231810	0.000000	39.395813	0.000000	39.280362
18	18	0	0.000000	69.519050	0.000000	32.171252	0.000000	33.438722	0.000000	64.292698	0.000000	68.210369
19	19	132966272	0.000000	75.760901	0.000000	37.632153	0.000000	46.790331	0.000000	56.207508	0.000000	105.580855
20	20	0	0.000000	86.886321	0.000000	45.215978	0.000000	49.472029	0.000000	48.520630	0.000000	77.657343
21	21	132966912	0.000000	68.967233	0.000000	30.637196	0.000000	33.447595	0.000000	36.426132	0.000000	62.207332
22	22	0	0.000000	90.732710	0.000000	43.284015	0.000000	53.450487	0.000000	55.241042	0.000000	99.049854

Table 1:ISODATA cluster Statistics

	CLUSTER	ELEMENTS	MEANDIST	MEAN.Band1	STDV.Band1	MEAN.Band 2	STDV.Band 2	MEAN.Band 3	STDV.Band 3	MEAN.Band 4	STDV.Band 4	MEAN.Band 5
1	1	9993	0.826544	79.121785	0	36.864605	0	40.101871	0	42.004303	0	64.793656
2	2	573083	1.748039	70.816726	0	32.056861	0	34.062186	0	43.664139	0	62.630778
3	3	81792	1.365965	70.558319	0	33.164014	0	35.574139	0	57.626981	0	74.394122
4	4	10915	1.427431	72.328081	0	35.246083	0	39.206322	0	59.239762	0	77.376913
5	5	273189	1.435152	75.129478	0	37.072543	0	45.246218	0	56.030667	0	100.347386
6	6	10587	0.821480	71.593180	0	33.555587	0	39.842259	0	48.240767	0	92.265136
7	7	148	0.715169	77.175676	0	36.270270	0	40.540541	0	50.168919	0	70.425676
8	8	92523	1.264542	77.910574	0	38.161884	0	45.215946	0	46.660701	0	76.258692
9	9	332702	1.677291	81.408308	0	39.838910	0	46.972035	0	49.416899	0	86.089437
10	10	201	0.723681	72.900498	0	35.761194	0	39.587065	0	46.885572	0	80.134328

Table 2:K-Means Cluster Statistics

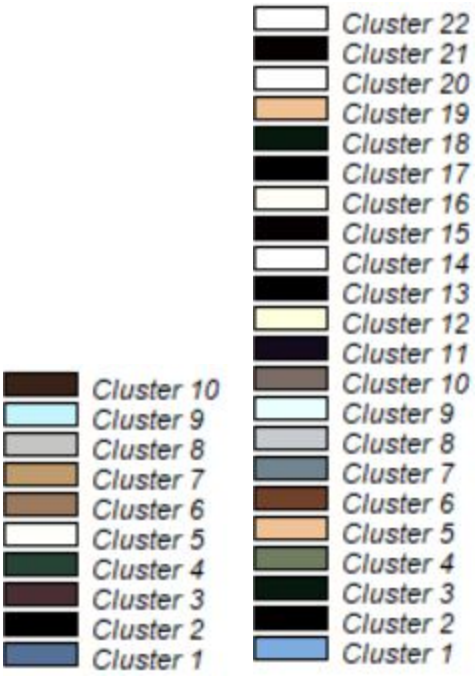


Figure 6:Legends Detail (i)K-Means (1000 iterations) (ii)ISODATA (1000 iterations)

Analysis:

ISODATA Algorithm, allows the number of clusters to be adjusted automatically during the iteration by merging similar clusters and splitting clusters with large standard deviations, unlike K-Means Algorithm and works better than K-means as the results are almost comparable for both the clustering algorithm for comparison we have considered 5 cluster(here fixed) information of both the algorithms and also drew a comparison based on no. of clusters identified for same no. of iterations (here 1000) statistics details which computes mean pairwise distance(MPD).

Gaps/Challenges :

Often times in ISODATA clustering algorithm it becomes difficult to converge.

How can we overcome this Gap?

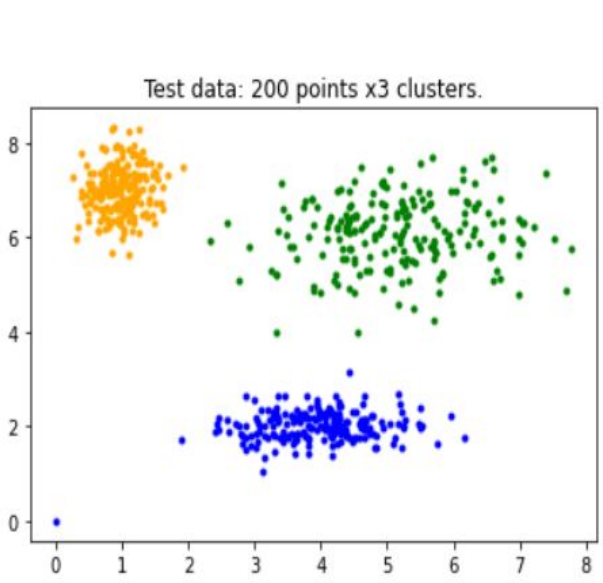
We can start the ISODATA process with an overestimate of the number of classes and then simply merges (i.e. there is no splitting of classes).

C.] Fuzzy C-means clustering

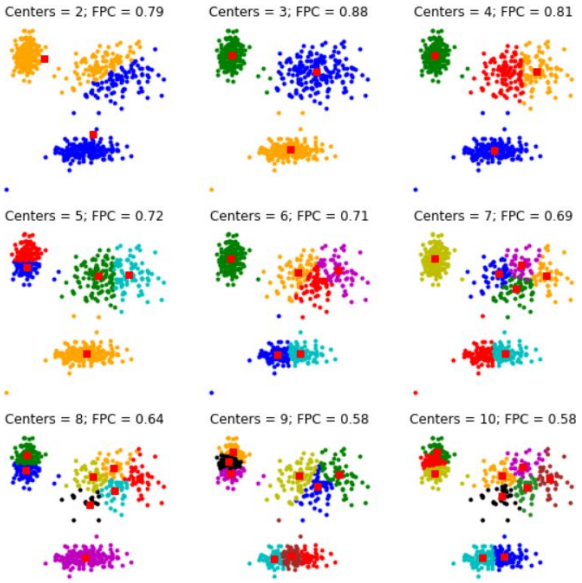
It is a clustering algorithm which allows a single data input to identify itself in more than one cluster. Here the data point arranges itself at a distance from a particular cluster property in such a way that the distance between signifies the belongingness of that data point to the cluster.

For inconsistent labelling of any data we will run fuzzy c-means to identify if there is overlapping in the data that creates inconsistency. All the algorithms mentioned above, a data point is always grouped to only one cluster. Fuzzy C-means will help us have maximum consistency in labelling.

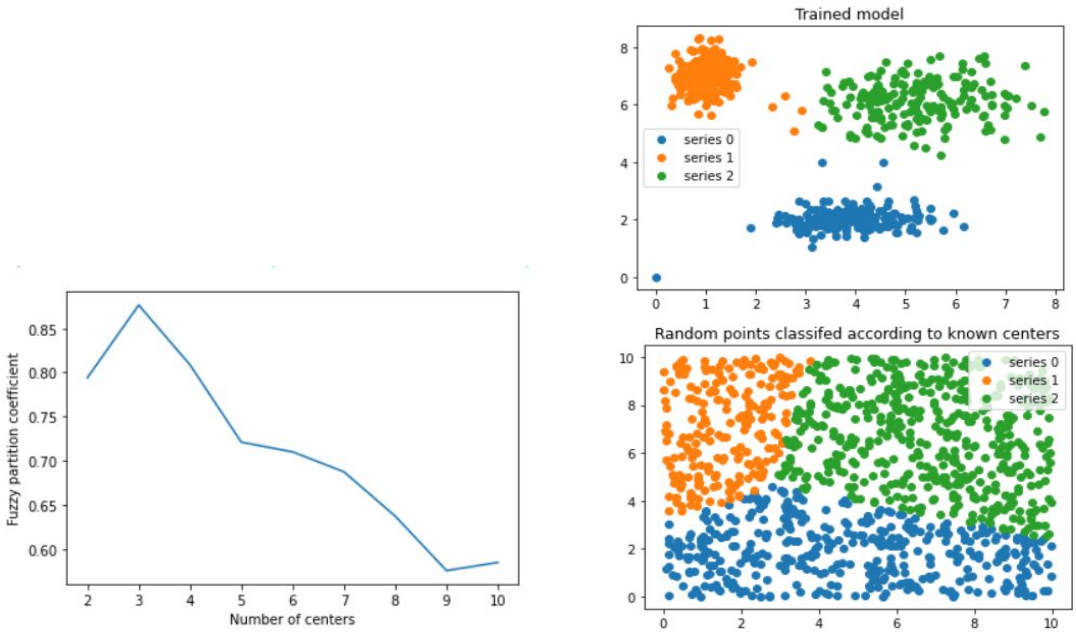
Output:



a)



b)



c) **d)**

Figure 7: Clockwise from top:(a)Test data , b)Centers of each fuzzy clusters , c)Number of centers , d)Trained Model & Fuzzy Result)Fuzzy C-Means classification

Analysis:

From fig 6(c) it is clearly evident that with increased no. of centers frequency partition coefficient decreases(FPC) in a non-linear fashion, and also it is observed that different initializations results in different evolutions of the algorithm. In fact it converges to the same result but probably with a different number of iteration steps.

D.] DBSCAN:

DBSCAN - Density-based spatial clustering of applications with noise, is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu in 1996.DBSCAN is one of the most common clustering algorithms.

Why DBSCAN?

Partitioning methods like K-means or PAM clustering and hierarchical clustering work for finding spherical-shaped clusters or convex clusters. In other words, we can say that they are suitable only for compact and well-separated clusters. Also, these algorithms are very much affected because of the presence of noise and outliers in the data.

Real-life data may contain irregularities, such as -

i) Clusters can be of different shapes such as those shown in the figure below.

ii) Data might be noisy.

The DBScan Algorithm tried to overcome many of these drawbacks of algorithms like K-Means. It identifies the dense region by grouping together data points that are closed to each other based on distance measurement.

Abstract Algorithm:

DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a dense region MinPts.

1. Find all the neighbour points within eps and identify the core points or visited with more than MinPts neighbours.
2. For each core point if it is not already assigned to a cluster, create a new cluster.
3. Find recursively all its density connected points and assign them to the same cluster as the core point.

   A point a and b are said to be density connected if there exists a point c that has a sufficient number of points in its neighbours and both the points a and b are within the eps distance. This is a chaining process. So, if b is neighbour of c, c is neighbour of d, d is neighbour of e, which in turn is neighbour of a implies that b is neighbour of a.
4. Iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are noise.

A simple implementation of this requires storing the neighbourhoods in 1, and hence requires substantial memory.

We used Canada Weather data for the year 2014 to cluster weather stations

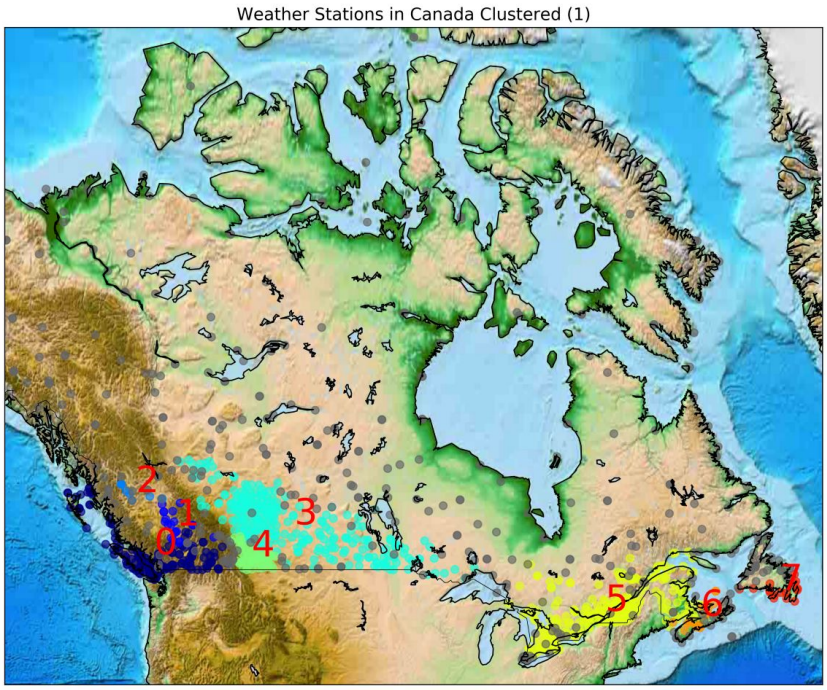
We tried to implement the original DBScan algorithm steps in our implementation as mentioned in the research paper by Martin Ester et.al. [1]. We used Basemap Toolkit to help us plot 2D data for visualizing maps in our collab project (python) and the result is as follows:

Output:

The output shows 8 different clusters in Canada Based on our selected features from dataset.

$\epsilon = 0.3$

MinPts = 10



The second output is for caparison of K-means and DBScan done using randomly generated samples using SKLearn dataset.

Figure 8:DBSCAN Classification

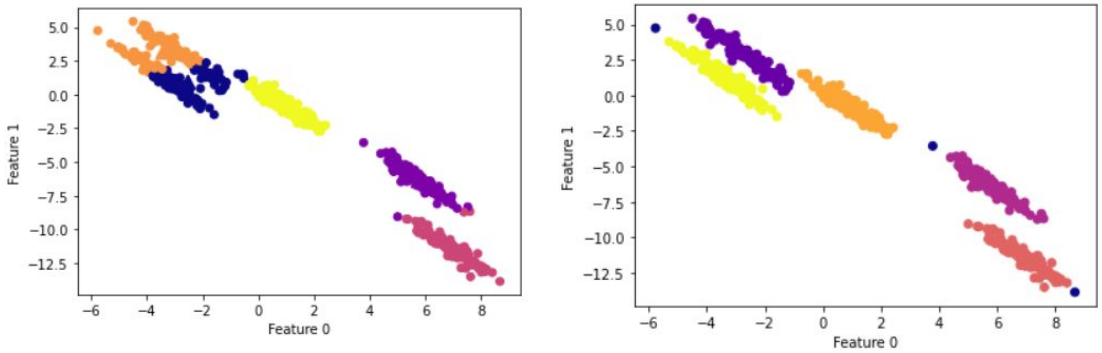


Figure 9:Comparison of K-means clustering and DBSCAN (i) K-means (ii) DBSCAN

Complexity:

- O(n log n)
- a non-matrix based implementation of DBSCAN only needs O(n) memory.

Analysis:

- Identifies dense region by clustering together the data points close to each other.
- Able to identify noise
- By changing the value of ϵ we were able to identify the clustering problem if lack of domain knowledge is there.

E.] Self Organising Map:

Self Organising Maps **(SOM)** is an unsupervised and competitive learning Artificial Neural Network model. SOM is a method that is used **(i)** When data to be clustered is unlabelled i.e., the number of classes is unknown and **(ii)** For dimension reduction of the data.

Algorithm:

1. Initialize the SOM.
2. Get an input pattern in vector x.
3. Initialize all the neurons i and a weight vector w_i associated to a neuron i.
4. Repeat steps 5 and 6 for all nodes.
5. Calculate the Euclidean distance between the input vector and the weight vector.
6. Find best matching unit (BMU) i.e., the node with the minimum euclidean distance
7. Calculate the learning rate and neighbourhood function
8. Calculate neighbourhood distance weight matrix
9. Modify SOM weight matrix
10. Repeat from step 4 until the maximum number of iterations is reached.

SOM v/s K Means and ISODATA clustering algorithms

1. To begin with K Means and ISODATA, the number of classes needs to be known beforehand.
2. Users need to manually define many parameters to which K Means and ISODATA algorithms are very sensitive.
3. The high computational cost for K Means and ISODATA, when data to be analyzed is very large.

Output:

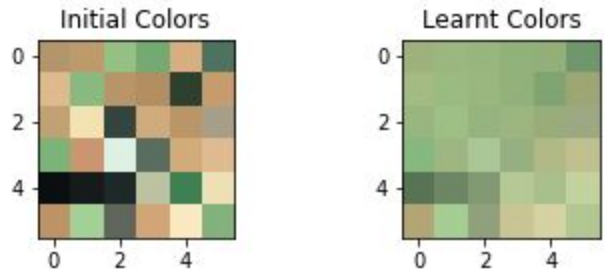


Figure 10: Dimensionality Reduction

Analysis:

1. SOM algorithm reduces noise
2. It reduces the dimensionality of the images and yet successfully clusters them. This reduces the computational complexity and cost.
3. The learnt colors can be utilised to further classify.

Concluding remarks:

Choosing a particular clustering algorithm depends on the type of data to be clustered and the purpose of the clustering application. The fuzzy c-Means is suitable for clustering for noise related problems or mixed media information, and human interactions, and can provide approximate solutions faster . Isodata is suitable for recovering compact clusters.Our study indicates that DBSCAN is better than K-means ,Fuzzy c-Means,ISODATA and SOM in discovering non-convex clusters & also it is a better approach than K-means,ISODATA and SOM in extracting convex clusters. It can easily detect noise. However, the DBSCAN algorithm also takes much more CPU time for large data sets. When the clusters are of arbitrary shape DBSCAN is more preferable. On the other hand, when the clusters are of hyperspherical or convex shape and well separated and the data set is large, then SOM or K-means,ISODATA,Fuzzy c-means may be preferable as they are faster.

References:

[1]A. W.Abbas, N. Minallh,N.Ahmad, S.A.R.Abid, M.A.A.Khan(2016) K-Means and ISODATA Clustering Algorithms for Landcover Classification Using Remote Sensing

[2] Rafik LASRI, 2016, Clustering and Classification Using a Self-Organizing MAP, The Main Flaw and The Improvement Perspectives

[3]Hemant Kumar Aggarwal & Sonajharia Minz (2015) Change Detection Using Unsupervised Learning Algorithms for Delhi, India

[4]Usman Babawuro(2013) Satellite Imagery Land Cover Classification using K-Means Clustering Algorithm: Computer Vision for Environmental Information Extraction, Elixir Comp. Sci. & Engg. 63 (2013) 18671-18675

[5]Balasubramanian, Sowmya & Sheelarani, B. (2011) Land cover classification using reformed fuzzy C-means. c Indian Academy of Sciences. 36. 153-165. 10.1007/s12046-011-0018-4.

[6] M. L. Gonçalves, J. A. F. Costa and M. L. A. Netto (2011) Land-Cover Classification Using Self-Organizing Maps Clustered with Spectral and Spatial Information

[7]Shi Na, Liu Xumin ,Guan yong (2010) Research on k-means Clustering Algorithm An Improved k-means Clustering Algorithm

[8]Satchidanandan Dehuri, Chinmay Mohapatra, Ashish Ghosh and Rajib Mall, 2006. A Comparative Study of Clustering Algorithms. *Information Technology Journal, 5: 551-559.*

[9]*"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise"*; Martin Ester et.al. KDD-96 Proceedings.

[10] Density Based Clustering Methods; Gao,J., Associate Professor Buffalo University.

Annexure:

We had tried to implement the algorithms like SOM, DBSCAN, Fuzzy C-means in google colab, you can view our implementation on:

https://colab.research.google.com/drive/1FcA5cUKTGVHdwY_aE5cGXnljDnfqZCrH?usp=sharing#scrollTo=z6i6pxmWBw74

https://colab.research.google.com/drive/1e8_dGLceuTqX_w65OvT_22_B1KM430x-?usp=sharing

https://colab.research.google.com/drive/13caj9umpNNG_3DM70OXG2qzJzUsJ6Ael?usp=sharing

K-Means and ISODATA Algorithms are implemented in QGIS AND SAGA GIS Software.