# NSGA DBSCAN: An Efficient Clustering Technique

*Thesis submitted in partial fulfillment of the requirements for the award of degree of*

**Master of Engineering**

In

**Computer Science and Engineering**

*Submitted By*

**NITIKA**

**Roll No. 801632032**

Under the supervision of:

**Dr. V.P. Singh**
Associate Professor, CSED

**Dr. Vinay Gautam**
Lecturer, CSED


COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY
PATIALA – 147004

**June 2018**

# CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*NSGA DBSCAN: An Efficient Clustering Technique*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. V.P. Singh* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Nitika

Signature:

(Nitika)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

19/12/18

(Dr. Vinay Gautam)

Lecturer,CSED

(Dr.V.P. Singh)

Associate Professor,CSED

# ACKNOWLEDGEMENT

The successful completion of any task would be incomplete without acknowledging the people who made it possible and whose constant guidance and encouragement secured the success.

First of all, I wish to acknowledge the benevolence of omnipotent God who gave me strength and courage to overcome all obstacles and showed me the silver lining in the dark clouds with the profound sense of gratitude and heartiest regard. I express my sincere feelings of indebtedness to **Dr. V.P. Singh and Dr. Vinay Gautam** for their positive attitude, excellent guidance, constant encouragement, keen interest, invaluable co-operation, generous attitude and above all their blessings. She has been a source of inspiration for me.

I am grateful to **Dr. Maninder Singh**, Head of Department and **Dr. Ashutosh Mishra**, P.G. Coordinator, Computer Science and Engineering Department, Thapar Institute of Engineering and Technology for the motivation and inspiration for the completion of this thesis. I will be failing in my duty if I don't express my gratitude to **Dr. S.S. Bhatia**, Senior Professor and Dean of Academics Affairs in the institute for making provisions of infrastructure such as Library facilities, Computer Lab equipped with internet facility immensely useful for the learners to equip themselves with latest in the field.

Last but not the least I would like to express my heartfelt thanks to my parents and friends who helped in doing this thesis.

Nitika

**-Nitika**

**801632032**

## ABSTRACT

Clustering is one of the significant streams useful for determining groups and identifying significant distributions in the underlying data. Remote sensing images are utilized to automatically detect the high-resolution images. However, it demands accurate descriptions of the characteristics of the objects.

Density-based spatial clustering of applications with noise (DBSCAN) evaluates clusters of arbitrary shape relying on a density-based notion of clusters. Additional implementation contains KD-Trees to save the data that allow efficient retrieval of data and bring down the time complexity from ($n^2$) to (n log n). Therefore, improving the computational speed of the DBSCAN is the main motivation behind this research work. Because majority of existing clustering techniques used for remote sensing images suffer from noise issue and parameter tuning issue, which may degrade the performance of remote sensing vision systems.

Therefore, to overcome this issue in this research work, a novel adaptive density-based clustering (NSGA DBSCAN) technique with noise is designed which is used to tune the parameters of DBSCAN based clustering technique for remote sensing images. Initially, random population is generated. Thereafter, for each random solution DBSCAN is implemented for clustering process. The solution that has accurate cluster with lesser noise is selected as non-dominated solutions. Thereafter, selection, mutation and crossover operator are used to explore the proposed technique further. After, getting the termination condition, tuned parameters for DBSCAN are obtained. Extensive experiments are carried out by considering benchmark remote sensing images (i.e., obtained from satellite sensors such as QUICKBIRD, IKONOS, MODIS, SPOT etc.). From visual and quantitative analysis, it is found that the proposed technique outperforms existing techniques in terms of Accuracy and Root mean square error. Therefore, the proposed technique is more applicable to real-time imaging systems.



KEYWORDS – DBSCAN, NSGA-DBSCAN, K-Means.

# Table of Content

# List of Figures

# List of Tables

## LIST OF ABBREVIATIONS

| ABBREVIATIONS | DETAILS |
|---|---|
| DBSCAN | Density Based Spatial Clustering of Applications with Noise |
| NSGA DBSCAN | Non-Dominate Sorting Genetic Algorithm DBSCAN |
| V-DBSCAN | Varied DBSCAN |
| P-DBSCAN | Photo DBSCAN |
| ST-DBSCAN | Spatio-Temporal DBSCAN |
| DSets-DBSCAN | Dominant DBSCAN |
| PAM | Partitioning around Medoids |
| CLARANS | Clustering Large Applications |
| AGNES | Agglomerative Clustering |
| DIANA | Divisive Clustering |
| GA | Genetic Algorithm |
| SOM | Self-Organizing Map |
| DT | Decision Trees |
| ANNs | Artificial Neural Networks |
| NF | Neuro-Fuzzy |
| MARS | Multivariate Adaptive Regression Splines |
| DENCLUE | Density Based Clustering |
| OPTICS | Ordering Points To Identify Clustering Structure |
| STING | Statistical Information Grid-based methods |
| CLIQUE | Clustering in Quest. |
| TI-DBSCAN | Triangle Inequality DBSCAN |
| T-DBSCAN | Temporal-DBSCAN |
| N-Cuts | Normalized Cuts |
| FCM | Fuzzy C-Means |
| FP Growth | Frequent Pattern Growth |

# CHAPTER 1
# INTRODUCTION

The concept of clustering introduces new emerging platform for computing human mobility via GPS locations using geospatial data. In today's era of computing, vast amount of geospatial data is generated in almost all the fields which can be obtained from satellite images, remote sensing, GIS and technologies which allow collection and distribution of huge amount of user generated data[1]. Spatial data clustering known as one of the popular mining clustering techniques that are very useful in cases of spatial data when the data size is sufficiently large. The analysis on this dataset is carried out with this technique that results in non-trivial useful patterns. DBSCAN is one of them which discovers arbitrary shaped clusters and handle outliers effectively. Hence it is the most commonly used data clustering method in literature.

This user generated data is used for wide range of applications i.e. business, science, government, research and machine learning. So enormous amount of data can be obtained from these sources which is almost incomprehensible. These type of research domains give the attractive way to cluster the information based upon their parameters.

Clustering known as the significant streams useful for determining groups and identifying significant distributions in the underlying data. In clustering, remote sensing image data are segmented into various objects and thereafter clustered based upon object characteristics[2]. Clustering in remote sensing images are utilized to automatically detect the high-resolution images. However, it demands accurate descriptions of the characteristics of objects.

Clustering algorithms are beneficial to group the user generated data in a way such that group of similar data points referred to one cluster and group of dissimilar data points referred to another cluster. Clustering analysis is a major tool which is used in many research areas covering image analysis, data compression, pattern recognition, computer graphics, bioinformatics and information retrieval. As clustering analysis is an unsupervised learning technique that is applied if no knowledge about the dataset available. Clustering helps to determine the intrinsic grouping from unlabelled training data to predict the unlabelled data from the labelled set of training points.

As Clustering is a useful data mining technique applied in all research domains where no priori labelled data is available and cluster formation must be inferred from the data alone. Looking at this from a machine learning perspective, resulting clustering structure denotes a data concept. Thus, clustering can be seen as the unsupervised learning of a hidden data concept. Figure 1.1 shows basic concept of clustering.

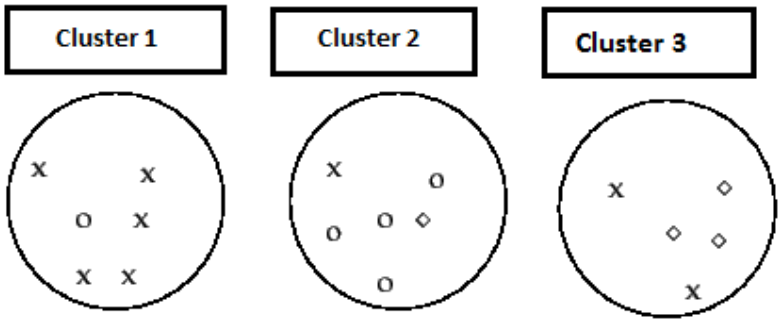


Figure 1.1: The basic concept of clustering

## 1.1 Clustering Techniques

The clustering techniques are broadly classified in their different types as shown in figure 1.2.

- Partitioning Clustering
- Hierarchical Clustering
- Density-Based Clustering
- Grid based Clustering algorithms

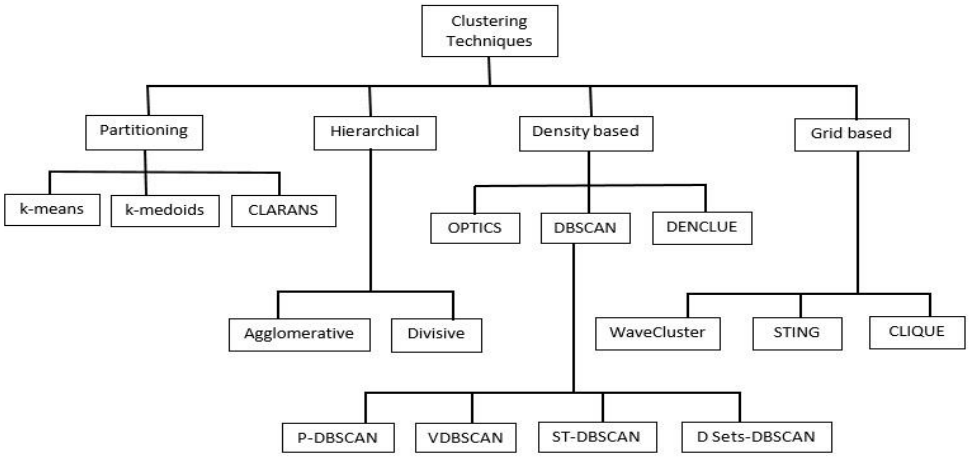


Figure 1.2: Clustering Techniques

### 1.1.1 Partitioning Algorithm

Partitioning Algorithms are the most popular and fundamental version of cluster analysis which is also known as iterative relocation algorithms or (combinatorial optimization algorithms). It constructs $k$ partition of $n$ documents from the given dataset D, where each partition identifies the pair of cluster ($k \leq n$). It starts with initial $k$ random centers and uses the following technique to improve the partitioning of objects by moving each object from one group to other. Figure 1.3 shows different types of partitioning algorithm include K-Means Clustering, K-Mediod Clustering, PAM, CLARANS and FCM clustering[2] [4] [5].

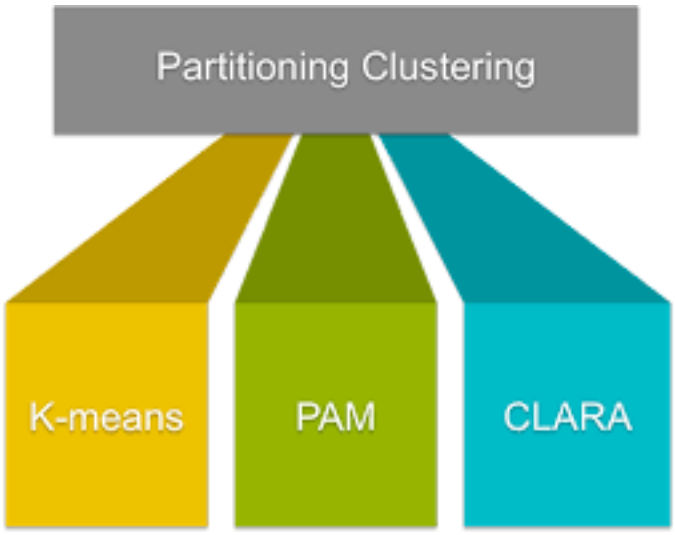


Figure 1.3: Partitioning based Clustering

**1. K-Means Algorithm**

K-Means algorithm[4] [5] is one of the widely used partitioning clustering approaches which are applied on unlabelled data. K-Means algorithm partitions $n$ number of objects into $k$ corresponding groups known as clusters. $K$ is a fixed priori number in which clusters are obtained according to fixed points called cluster centroids. K-Means follows iterative approach in which each object intends to partition into n clusters and defines $k$ centroids one for each cluster with the nearest mean. Figure 1.4 shows illustration of K-Means Algorithm.

**Steps of Algorithm includes**:

Let data points be $x_1, x_2, x_3 \ldots\ldots x_n$ and centroid clusters be $c_1, c_2, c_3 \ldots\ldots c_n$.

- Partition $n$ number of data points into $k$ clusters where k is a fixed priori element.
- Select randomly k cluster centers.
- Compute the Euclidean distance function and assign the data points and centroids to their closest center of cluster.
- Compute mean or centroid in each value of cluster.
- Repeat all steps until the condition is met and points are assigned as the cluster.

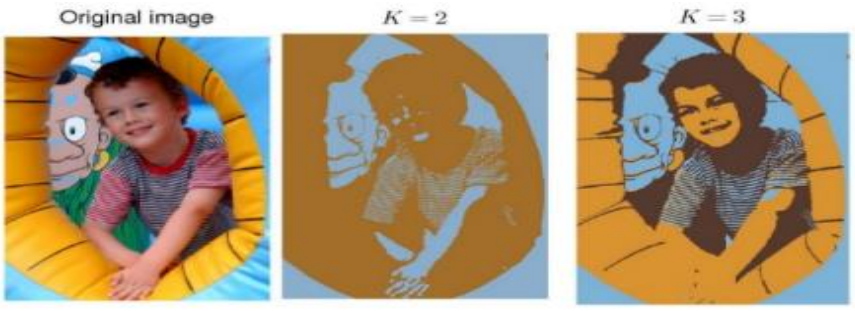


Figure 1.4: K-Means Clustering

**2. K-Mediod Clustering**

K-Mediod algorithm is a partitioning clustering algorithm which is an extension of K-Means algorithm as given in[6]. K-Mediod makes use of mediods rather than mean/centroid to find cluster. K-Mediod clustering attempts to minimize the average dissimilarity between the points. K-Mediod algorithm is an object representative technique which overcomes the drawbacks of handling noise and outliers in K-Means. It starts with randomly selected n objects as mediods and represents the clusters which have its mediods place near to them.

**3. PAM**

PAM abbreviated as partitioning around mediods which is realisation of K-mediod clustering[7]. Cluster Formation of PAM based upon Gaussian mixture. It starts from the given set of mediods and step by step one mediod is replaced by other mediod which helps to improve the distance of resulting clusters then all the resulted pairs are analysed for replacement. The biggest issue of PAM model is that it is effective for small data set regions and not applicable for large datasets because this algorithm uses greedy approach which not finds optimal solution but it is faster than K-mediod

algorithm. As compared to K-means, PAM algorithm is less sensitive to noise or outlier regions.

**4. CLARA**

Clustering LARge ApplicatioNS (CLARA) is a randomized approach for clustering large number of objects in a dataset[8]. CLARA is a first spatial data mining clustering technique which identifies spatial structures. This algorithm handles points and polygon objects efficiently. CLARA takes two input parameters: (maxNeighbor) as maximum neighbor and (numLocal) as local minima obtained. It starts with randomly selected k data points to make set of clusters and then minimize the summation distance to make new set of clusters. Run-time complexity of CLARA is $O(n_2)$.

**5. FCM**

Fuzzy C-Means clustering is a clustering technique which identifies one piece of data to many clusters[2]. The FCM algorithm assigns membership to each data object identical to each cluster with respect to distance between both data object and cluster centre. Figure 1.5 shows FCM clustering image analysis.

**Steps of Algorithm includes:**

Let data points be $\{x_1, x_2, x_3, \ldots\ldots x_n\}$ and centroid clusters be $\{v_1, v_2, \ldots\ldots, v_c\}$.

1. Select randomly 'c' cluster centers.
2. Fuzzy membership function$'\mu_{ij}'$ calculated as

$$\mu\dot{ij}=\frac{1}{\sum_{k=1}^{c}(d\imath\dot{\jmath}/d\imath\dot{k})\left[\frac{2}{m-1}\right)}$$

3. Fuzzy centers$'v_j'$can be calculated as

$$V_J = (\sum_{\imath=1}^{n}(\mu\imath\dot{j})^m \, xi/\sum_{\imath=1}^{n}(\mu\imath\dot{j})^m), \forall J = 1,\dot{2} \ldots\ldots c$$

4. Repeat both the steps until the value 'J' is achieved.

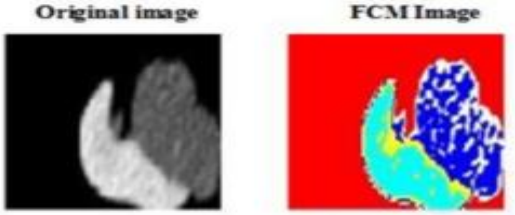


Figure 1.5: FCM Clustering

### 1.1.2 Hierarchical Algorithm

Hierarchical clustering provides a series of hierarchical decomposition of the given objects[3] [9]. Hierarchical algorithms follow recursive process which can be divided into two approaches: top-down (or divisive approach) and bottom-up (or agglomerative approach). In Agglomerative, it starts with the point as an individual cluster and merges the group of objects that are close to each other and keeps on merging until the termination condition holds. In Divisive, it starts with group of objects in the same cluster and a cluster is split up into a number of clusters. Hierarchical algorithms are also called "nested set of partitions" which can be represented in the form of tree structure called dendogram. Figure 1.6 shows various types of hierarchical algorithm including agglomerative clustering and divisive clustering.

1. **Agglomerative clustering or AGNES:** Agglomerative is a bottom-up clustering method in which observation starts in its own individual cluster (leaf) and the group of clusters are merged as according to the approach in which clusters moved up according to the hierarchy[9].
2. **Divisive clustering or DIANA:** Divisive clustering algorithm is an inverse of the agglomerative clustering algorithm. Divisive is a top-down clustering method in which observation starts into single cluster (root) and recursively perform splitting in which clusters move down according to the hierarchy[9].

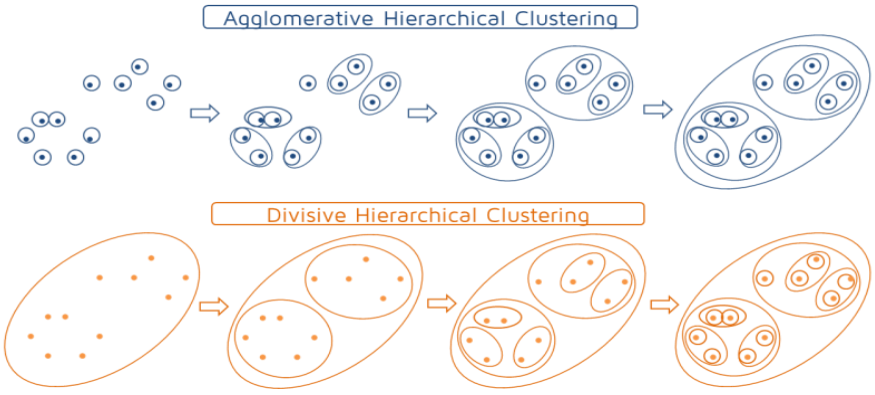


Figure: 1.6: Hierarchical based Clustering

### 1.1.3 Density-Based Algorithm

Density-based method introduces concept of algorithm as notion of density. Density-based clustering algorithm discovers arbitrary shaped clusters which is one of the

cardinal methods for clustering in data mining[9]. These algorithms discover clusters based upon the density of regions in a data set. The cluster formation based upon density does not limit itself to the shapes of cluster. The density of data in that region will be formulated as the ratio of number of object to the volume of object. Density based clustering algorithm can be categorized as DBSCAN, OPTICS and DENCLUE.

As there are number of Density-based clustering algorithms such as DBSCAN, OPTICS and DENCLUE. I have focussed on DBSCAN because of its fascinating properties rather than other clustering algorithms:

- In DBSCAN, there is no need to define number of clusters as in case of K-Means clustering algorithm.
- DBSCAN can find clusters of non-spherical shape and can detect outlier effectively.
- Processing is fast in DBSCAN which works well with spatial database.

1. **DBSCAN**: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a dynamic density-based clustering algorithm[10]. In these algorithms, data points in some region is provided, which is then grouped together and mark them as a packed cluster (points with many nearby neighborhood data points), and mark outlier points as noise which are not able to make clusters (data points whose neighbors are too far). As it describes the notion of clusters surrounded by a given radius in which neighborhood of a given radius has to minimum number of data objects. In other words, two points are density-connected if and only if they are enough similar and at least the surrounded area is dense. The object will be marked as NOISE, if the sum of the neighborhood region is below the specified threshold limit. In other respects, by finding the maximal group of density connected objects in the neighborhood region the new cluster will be formed inside the cluster from the core object. Two parameters are used as input inside the cluster i.e. radius of the cluster ($\epsilon$) and minimum points (Minpts). In this algorithm, first of all the number of objects surrounded around the neighborhood region ($\epsilon$) is to be calculated i.e. threshold distance or ($\epsilon$): two points are considered to be neighbors if the distance between the two points is less than ($\epsilon$) and second, threshold density or MinPts: two points are considered to be as neighbors if the distance between the

two points is less than ($\epsilon$) and at least one of them has MinPts surrounded around the region. Figure 1.7 shows arbitrary shaped clusters.

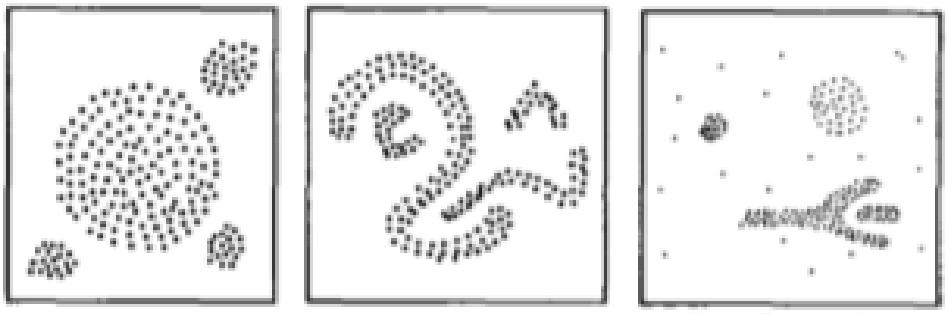

Figure 1.7: Arbitrary shaped clusters [10]

Following formal definitions and related notions of density based clustering need to be understood before detail of DBSCAN algorithm:

**Definition 1**: **Eps Neighborhood of point p:** The neighborhood of a point $x$ is defined as the point which lies in the radius of point $p$. NEps($p$) is defined as set of points in Dataset D which have distance from p in the range of $\epsilon$.

$$NEps(p) = \{q \in D \mid dist(p,q) \Leftarrow Eps\}$$

**Definition 2**: **Core Point:** A point $p$ satisfies the core point condition, if points are at the interior of a cluster which contains at least MinPts points within $\epsilon$.

**Definition 3**: **Border Point:** A point $p$ satisfies the border point condition, if points lie on the border of the cluster which has fewer MinPts within $\epsilon$, but is in the neighborhood of a core point.

**Definition 4**: **Noise Point:** A point $p$ satisfies the noise point/outlier condition, if point is neither a core point nor a border point comes under the category of noise point.

Figure 1.8 shows core, border and noise points.

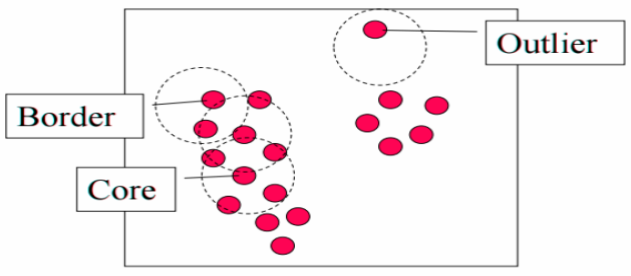


Figure 1.8: Core, Border and Noise point

**Definition 5**: **Directly Density-Reachable:** A point $p$ satisfies directly density-reachable condition from another point $q$ w.r.t radius and MinPts, if it follows two conditions:

1. $p \in NEps(q)$

2. $|NEps(q)| >= MinObjs(core\ object\ condition)$

**Definition 6**: **Density-Reachable:** A point $p$ satisfies density-reachable condition from another point $q$ w.r.t radius and MinPts, if there exists a sequence of chain forming points $q_1, q_2, q_3, \ldots\ldots q_n$ with $q_1 = q$ and $q_n = p$ such that $q_i + 1$ is directly-reachable from $q_1$.

**Definition 7**: **Density-Connectivity:** A point p satisfies density-connectivity condition from another point $q$ w.r.t radius and MinPts, if both points $p$ and $q$ are density reachable from point O. Figure 1.9 shows density-reachability and density-connectivity clusters.

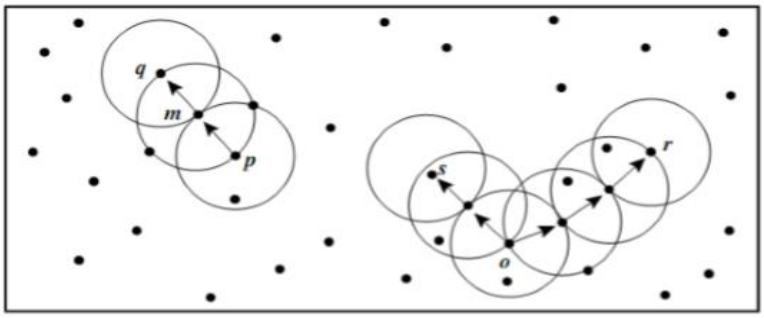


Figure 1.9: Density-reachability and Density-connectivity

**Definition 8**: **Cluster Formation:** Let D is a subgroup of database D of points which satisfies the following conditions:

1. $\forall\, p, q$: w.r.t radius and Minpts, if $p$ belongs to cluster C and point $q$ is density-reachable from point $p$ then $q$ belongs to cluster C (Maximality).

2. $\forall\, p, q$: w.r.t radius and MinPts, point $q$ is density-connected to point $p$ (Connectivity).

## DBSCAN Variants

DBSCAN addresses some of the weaknesses which are hard to be applicable on all types of datasets. So, as to address these weaknesses, some of the variants of DBSCAN algorithm are as follows:

**P-DBSCAN**

P-DBSCAN is a variation of DBSCAN which captures geo-tagged collection of photos of interesting places and events held[1]. It also works on the concept of DBSCAN with respect of new definition of density based on the number of people (owner of photos). P stands for photo that has been captured by a user and uploaded with location coordinates. A photo through its point can be reachable from object 1 to object 2 is called neighborhood. The distance is measured from the Point (Latitude, Longitude). If friends have multiple photos reachable, it is a region.

Adaptive density defines the ratio of previously recorded number of photos to the currently recorded number of photos that are in the region and reachable with respect to their points.

**VDBSCAN**

VDBSCAN is varied density based spatial clustering of application with noise, which addresses varied-density datasets analysis[17]. VDBSCAN discovers some of the important weakness of DBSCAN i.e. DBSCAN cannot discover varying density clusters while VDBSCAN can identify varied density clusters. The idea of VDBSCAN clustering is same as of DBSCAN but additionally difference is it addresses varied-density datasets analysis.

VDBSCAN can follow the various steps, first choose $\epsilon$ and second cluster formation based on varied-density. Also it uses $k$-dist graph to find out the $\epsilon$ values. $k$-dist graph is plotted for the different density level datasets for finding parameters. VDBSCAN calculates and stores the $k$-dist value and then number of densities is given by $k$-dist graph and automatically it will choose $\epsilon$ value for corresponding density value. Then final cluster will be formed with varied densities. Figure shows the sample of $k$-dist graph. The complexity of VDBSCAN is same as that of DBSCAN.

**ST-DBSCAN**

Spatio-Temporal DBSCAN is a marginal extension to DBSCAN[18]. In relation to existing density clustering algorithms, Spatio-Temporal DBSCAN has the capability to discover clusters with non-spatial, spatial and temporal object values. ST-DBSCAN needs four input parameters from users i.e. $\varepsilon 1$, $\epsilon 2$ , MinPts, $\Delta\in$. $\epsilon 1$ is a parameter which

is used as a distance for spatial latitude and longitude attributes. $\epsilon 2$ is a parameter which is used as a distance for non-spatial attributes. MinPts as minimum points within $\epsilon 1$ and $\epsilon 2$ distance. The working of whole algorithm is similar to DBSCAN with respect to $\varepsilon 1$ and $\epsilon 2$.

ST-DBSCAN can cluster the spatial-temporal type of data which has the capability to discover clusters with non-spatial, spatial and temporal object values. As DBSCAN algorithms fails to identify noise point regions containing varied density but ST-DBSCAN algorithm overcomes the problem with the density factor. ST-DBSCAN also clusters the points having different densities.

**D-Sets DBSCAN**

D-Sets DBSCAN is a parameter-free algorithm which is a combination of D-Sets (dominant sets) and DBSCAN algorithm[19]. In the D-Sets DBSCAN algorithm, histogram equalization is applied to find the pairwise similarity matrix to make D-Sets clusters which is not depend upon the input parameters. Then, clusters can be formed with D-Sets DBSCAN and input parameters are determined automatically. D-Sets DBSCAN algorithm can be effective in both clustering algorithms and image segmentation. D-Sets DBSCAN algorithm is similar to D-Sets and DBSCAN, as it extracts the sequential clusters.

**2. Ordering Points to Identify the Clustering Structure** (**OPTICS**): A density based clustering algorithm in which clusters are generated using "cluster-ordering" of points[11]. Its works on the principle of DBSCAN, but it helps to identify one of the DBSCAN algorithm major weaknesses of detecting clusters in varying density. Similar to DBSCAN. It requires two input parameters: maximum distance of the cluster ($\epsilon$) and minimum number of points to form a cluster (Minpts).

**3**. **DENsity CLUstering** (**DENCLUE)** works differently from both DBSCAN and OPTICS which uses kernel density estimation for finding clusters in data)[12]. DENCLUE uses two concepts: In influence functions, data points can be influenced as mathematical function known as resulting function and density functions is determined using sum of influence of all the data objects. Two types of cluster are formed in DENCLUE, multi centre defined clusters and centre defined clusters. Clusters are

identified using density attracters which are local maximal of the overall density function.

### 1.1.4 Grid-Based Algorithm

Grid based algorithm together with the object to form grid. Grid structure formation depends upon the quantizing object space into finite number of cells and then performs the operation on quantized space[3]. Grid based algorithms are totally different from conventional clustering algorithm but it concerned with space partitioning. Types of grid based clustering include STING, WaveCluster and CLIQUE.

1. **STatistical Information Grid** (**STING)** method is used to cluster spatial database represented in the form of hierarchical structure**)[**13] [16]. These algorithms facilitate region oriented queries by dividing the spatial data into rectangle cells and stores the cell in hierarchical structure. It starts with the root level of the hierarchy and at next level, number of cells can be obtained by partitioning each cell into 4 cells and parent cell corresponds to union of all of its children. This algorithm is applicable to only two-dimensional space in which each cell is divided into adjacent 4 cells with each cell corresponding to quadrant of parent cell. Figure 1.10 shows STING clustering.

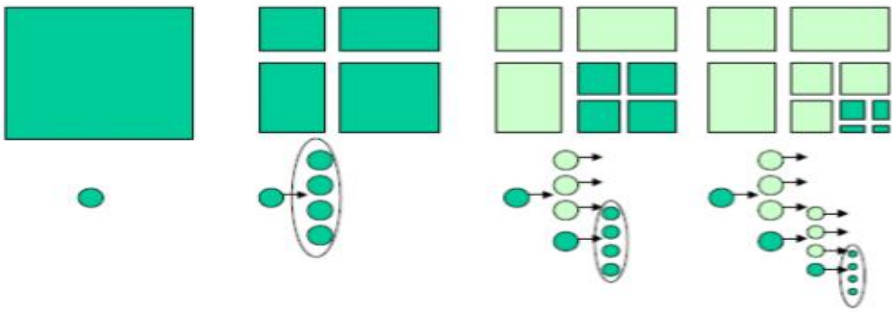

Figure: 1.10: Grid-based Clustering

2. **WaveCluster:** WaveCluster is a grid-based approach, which specially used for large databases[14]. WaveCluster is multi-resolution approach which uses wavelet transforms and applies it to data points and use the data to generate clusters. WaveCluster is a spatial data clustering algorithm which has the capability to find arbitrary shaped cluster formation at different scales. The complexity of algorithm is $O(N)$.

3. **CLustering In QUEst** (**CLIQUE)** is defined as a subspace algorithm which makes static grid with the help of apriori approach which minimizes the chances to reduce search space[15] [16]. CLIQUE is Grid-density based clustering algorithm which finds cluster using input parameters by taking number of grids and by taking density threshold value.

   Clustering has a major role on remote sensing images. Remote Sensing is helpful to identify the high density regions and low density regions with noise and time complexity.

## 1.2 Remote Sensing

Data obtained from satellite images, remote sensing, GIS and internet technologies which allow collection and distribution of huge amount of user generated data. The concept of geospatial data clustering introduces new emerging platform for computing human mobility via GPS locations[20]. Remote Sensing is a field of science or art which provides the method to sense the valuable information about an object or phenomenon from a distance without having any direct physical relationship with the object. Remote Sensing is a part of Earth Science disciplines which involves wide range of applications in various fields including natural resource management, geography, intelligence, planning and humanitarian applications.

Remote Sensing helps to identify and collect information of extremely large areas at a distance from the targeted areas. Remote Sensing helps in spatial data clustering which helps to detect the inaccessible information and observe the broader area at a given period of time. Remote Sensing technologies capture a single image which can be analyzed and processed for various use and applications[21]. It helps to locate the highly density regions which makes is easier to obtain the information about the targeted areas. Spatial data clustering known as the popular data mining clustering techniques that are very useful in cases of spatial data when the data size is sufficiently large. Figure 1.11 showing areas where remote sensing is applicable to make density based notion of clusters.

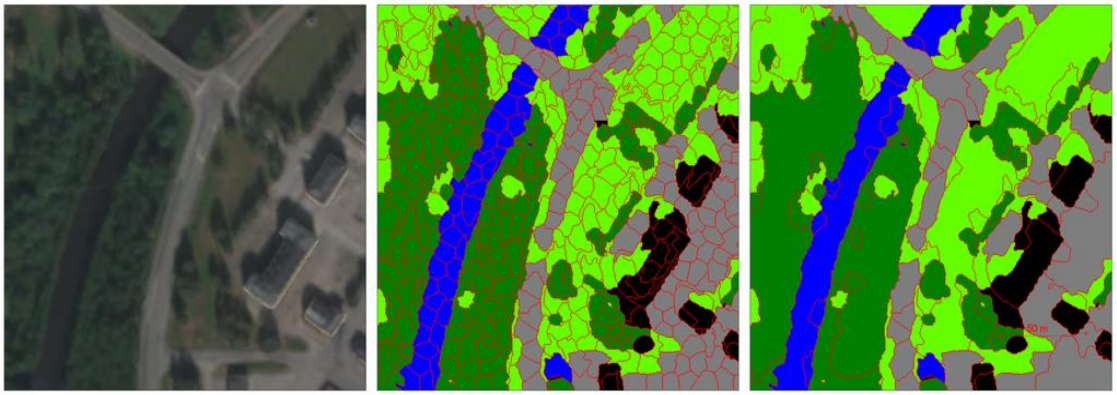

Figure 1.11: Remote Sensing

### 1.2.1 Remote Sensing Major Application Area

Remote Sensing Satellites is an application which processes the remote sensing data. Just like the graphics software, remote sensing applications allow us to generate word mapping, sensor image data, GIS and GPS etc. Collection of such big satellite systems caused a worldwide buzz in real-world scenario which allows us to record information without being physically there. The areas where Remote Sensing is highly applicable are:

1. **Tracking urban growth:** Tracking urban growth will demand a high density of world's population that is expected to live in the high density regions which will easily the spatial history about the urban people living in the nearby areas. It includes industrial developments, residential areas and urban areas from where data sources can be collected using geographic information retrieval systems.
2. **Collecting earth's pictures from space**: Remote Sensing systems will use different sensors to collect complementary information regarding oceans, atmosphere and lands. It will be helpful to cluster the information where high density regions can be found.
3. **Geology of Earth's surface:** Remote Sensing helps in understanding the Earth and its surface features through which we can easily identify the regions that come from geological observation of earth's surface. Geology of Earth's surface will remain constant in our lives which are used in remote sensing as structural mapping.

### 1.2.2 Remote Sensing Exploration using Machine Learning Techniques

Machine Learning Algorithms are known as "universal approximators", as machine learning in remote sensing gathers the data using satellites or drones. These algorithms learn the behaviour of the data from the set of training data. One of the interesting features of machine learning is that they train the data themselves as they do not need knowledge about the data. Machine Learning algorithms are also used in regression and classification models to train the systems. Machine Learning algorithms can be categorized as follows:

1. **Artificial Neural Networks (ANN):** Most of the geosciences problems deal with the Artificial Neural Network. Artificial Neural Networks is a neural network model which is based upon input and output model. The formulation of neural network is based on the simple hidden layer network. Artificial Neural Networks are also called connectionist systems.
2. **Self-organizing Map (SOM):** Self-organizing Map is an unsupervised learning technique which comes under artificial neural network. Self-organizing feature Map represent the visual data as a hexagonal or rectangular grid. Self-organizing Map apply competitive learning technique on the high-dimensional datasets.
3. **Decision Trees (DT):** Decision Trees make decision about the item's target value which is represented in the form of leaves by observing an item which is represented in the form of branches. Decision Trees can be categorized as classification trees (which can take discrete values) and regression trees (which can take continuous values). Decision Trees can be used in data mining techniques as tree-like model decision making.
4. **Genetic Algorithm (GA):** Genetic Algorithms solve the problem using principle of Darwinian natural selection. These algorithms are widely used in geosciences and remote sensing domains. The mechanism of genetic algorithms based on genetics and selection. Genetic Algorithm is a part of artificial intelligence which solves the problem and finds the potential solution.

5. **Neuro-Fuzzy (NF):** Neuro-Fuzzy is a part of artificial intelligence, which is a mixture of ANN and fuzzy logic. These are combined to overcome the limitations of each isolated paradigm referred to as Neuro-Fuzzy Systems**.**
6. **Multivariate Adaptive Regression Splines (MARS):** Multivariate Adaptive Regression Splines (MARS) is a part of nonparametric regression analysis which is used to predict the values of a continuous variable from a set of variables. Multivariate Adaptive Regression Splines is a popular data mining area because it does not assume any particular type of relationship.
7. **Random Forest and Random Tree**

   These two concept are very similar. Random Forests is a collection of decision trees. A Random Forest is a supervised learning based algorithm for classification and regression based tasks. At training time, random forest works as multitude of decision trees. Random Forests make multiple decision trees and then combines them together in order to get more accurate and reliable prediction. While expanding the tree we obtain some additional randomness to the model. It splits the node and searches the best feature among all the random features.

   Random tree is an assembling of trees predicators**.** Random tree is compared with basic and simple classifiers such as perceptron. Perceptron means trees originate from different type of family models.
8. **Regression**

   Regression analysis is a supervised learning technique which is used for prediction and forecasting to predict and analyze the relationship among variables. Regression helps you understand the characteristic value based upon the other values in data. It is also called “predictive power”. For e.g. time series prediction of stock market.

   In Regression predictive modeling, the task is to approximate a map function from input to continuous output variables. But it may predict a discrete value, which is in the form of integer quantity. In this analysis one dependent variable and more than one independent variable is given, focuses on how the value of dependent variable changes with respect to independent variables in varied region remaining other values fixed.

9. **Prediction**

   Prediction is a supervised learning technique, which tells us about a future event. In predictive modeling, the task is to approximate a map function from input to output variables. Prediction is useful in wide range of areas including machine learning, artificial intelligence, data mining and modeling etc. Prediction makes use of statistics and machine learning algorithms to predict outcomes.

## 1.3 Structure of the thesis

The rest of the thesis part is organized in the following order:

**Chapter 2 Literature Review -** This section defines different density based clustering techniques and their variants.

**Chapter 3 Research Problem –** This section of thesis describes the need for designing the algorithm and research gaps. It discusses the circumstances which led to formulate the problem and set the objectives to achieve problem stated. It also specifies the methodology needed to solve the research problem**.**

**Chapter 4 Proposed Methodology** – In this section, NSGA DBSCAN is proposed and implemented in the suitable environment then, the design of proposed work is explained in detail.

**Chapter 5 Experiments and Result –** This section includes the implementation setup, experiments performed on MATLAB tool and comparison of the results obtained.

**Chapter 6 Conclusion and Future Scope** – This section gives the conclusion of the thesis, summary of the contributions and the future scope.

# CHAPTER 2

# LITEARTURE SURVEY

This chapter covers different clustering techniques which are used to define the problem. An excellent focus into the studies of clustering, in this many literature surveys were analysed as follows:

**Atrayee Dhua1 et al. (2018) [25]** introduces density based algorithms which forms the cluster images using DBSCAN algorithm and Mean shift algorithm into different homogenous structures. Cluster formulation is based two parameters: minimum number of points within a cluster and the maximum distance between the points. Minimum number of points within a cluster is selected by the user. As in this algorithm cluster formation deals with the images with larger pixel values, accuracy of the algorithm will be improved if the values with larger radius are taken. These algorithms help to identify the cluster formation with respect to segmented images.

**Maguelonne Teisseire et al. (2018) [30]** produced object-oriented satellite image time series analysis using a graph-based representation. The proposed work applied on spatial-temporal values which are hard to understand. As method works by detecting spatio-temporal values then it creates their evolutions of graph representation by using mean and formes spatio-temporal clusters. As previous work has used pixel-intensity values and proposed work uses object-oriented images where segmentation is used for the formation of clusters.

**Simon Schmidt et al. (2018) [31]** proposed spatio-temporal mapping approach to cluster erosion prone regions that identifies high density regions. The analysis has been carried out by different datasets containing satellite data, aerial data and MODIS remote sensor. The proposed Swiss C-factor, a model that identifies grasslands regions situated at different areas.

**Kaiguang Zhao et al. (2018) [32]** have originated multi temporal lidar for monitoring of forest and carbon. For the following work ranging from 2002 to 2012, airborne lidar surveys have to conducted to describe robust multi temporal lidar analysis. Lidar monitoring is the best technology for tracking forest areas into grid level analysis. At lower and higher sampling rates, lidar failed to identify some components in lower

regions which contains more noise. In higher density regions, it improves the efforts applied in remote sensing images. At the end, time series analysis has been carried out to differentiate between lower and higher sampling rates.

**Ariel E. Baya et al. (2017) [28]** proposed clustering stability for automated color image segmentation. The proposed work adopted a cluster validation method which automatically segment images. Segmentation process includes optimal number of partitions based upon the color, texture, shape, energy and dissimilarity parameters. The adapted clustering stability detects feature extraction automatically and results are carried out on natural images, MRI data and color-texture images.

**Noel Gorelick et al. (2017) [29]** proposed planetary-scale geospatial data analysis over google earth engine. It will bring massive google earth engine computations to produce high-intensity density areas including forest, environment protection, disaster management areas etc. Results can be carried out over geospatial dataset where user can produce data products.

**Jianbing Shen et al. (2016) [26]** have described DBSCAN clustering algorithm to the generation of real-time super pixel segmentation. As it also follows the same concept as DBSCAN, set of points in some region are provided and cluster formation is based upon the closely packed points having points with many nearby neighbors, low-density regions are known as outliers whose neighbors are too far. The proposed super pixel segmentation algorithm first performs DBSCAN clustering algorithm to produce the super pixel clusters with color-similarity and then the super pixel clusters will be combined with their nearest neighbors by taking consideration of color and spatial information.

**Jian Hou et al. (2016) [21]** defined parameter-independent data clustering and image segmentation. In this paper, the experiments that have been carried out using various algorithms i.e. D-Sets, N-Cuts K-Means and DBSCAN algorithms which gives optimal results while testing datasets. Dominant set takes the input as pairwise data similarity matrix and cluster formation depends upon the similarity matrix. Different regulation parameters are applied on remote sensing images which results into segmented clustered results. Histogram equalization is also proposed to enhance images by maximizing the overall intensity in image.

**Tahereh Kamali et al. (2016) [27]** have discussed about automated segmentation density based clustering algorithm of white matter fiber bundles using diffusion tensor imaging data. The result demonstrated in three views such as axial view, sagittal view and coronal view. A new density based clustering algorithm i.e. neighborhood distance is proposed which uses local and global densities information to form natural clusters. Further proposed algorithm compared with other clustering algorithms i.e. DBSCAN, K-Means, Chameleon and soon. Among all clustering algorithms, neighborhood distance entropy consistency produces highest density clustering results.

**W. Chen et al. (2014) [24]** have described T-DBSCAN. T-DBSCAN stands for Trajectory DBSCAN, which is an extended and modified version of DBSCAN based upon the GPS based trajectories. T-DBSCAN considers the trajectories along with the GPS points by considering time-sequential characteristics. It will decompose the trajectory into sequence of stops and moves and defines that cluster formation is based upon the points surrounding a stop condition. Two parameters are used as input inside the cluster to define density i.e. search radius ($\epsilon$) and minimum points (MinPts).

**Yujie Li1 et al. (2012) [23]** have proposed a method for segmented image which will compute their similarity coefficient in RGB color then apply TI-DBSCAN. First the method will reduce the noise and then will get the segmented numbers. Secondly, it is not required to change the RGB color space. Thirdly, TI-DBSCAN i.e. Triangle Inequality property is used to reduce the neighborhood color space.

**Kisilevich et al. (2010) [1]** proposed P-DBSCAN which is a variation of DBSCAN which captures geo-tagged collection of photos of interesting places and events held. It also works on the concept of DBSCAN with respect of new definition of density based on the number of people (owner of photos). P stands for photo that has been captured by a user and uploaded with location coordinates. A photo through its point can be reachable from point 1 to point 2 is called neighborhood. The distance is measured from the Point (Latitude, Longitude). If point have multiple points reachable, it is a region. Adaptive density is the ratio of previously recorded number of photos to the currently recorded number of photos that are in the region and reachable with respect to their points.

**T. Blaschke (2010) [22]** formulated image analysis based upon the object for the respective remote sensing images dataset. Image segmentation and image processing, both concepts are used here for the formulation of clusters. Segmentation analysis of image partitions the image into number of layers, spectral and contextual information which can be further used for cluster formation.

**Ester et al. (1996) [10]** have given DBSCAN Algorithm which mines the clusters according to the density of the object. In DBSCAN, two points are density-connected if and only if they are enough similar and at least the surrounded area is dense. The object will be marked as NOISE, if the sum of the neighborhood region is below the specified threshold limit. In other respects, by finding the maximal group of density connected objects in the neighborhood region the new cluster will be formed inside the cluster from the core object. Two parameters are used as input inside the cluster i.e. radius of the cluster (Eps) and minimum points (Minpts).

# CHAPTER 3

# PROBLEM FORMULATION

As brought out above in chapter 2, following research gaps are formulated based upon the existing results as given below.

## 3.1 Research Gaps

After conducting the survey of existing techniques, followings gaps are formulated:

1. Majority of existing clustering techniques used for remote sensing images suffers from noise issue, which may degrade the performance of remote sensing vision systems.
2. The DBSCAN and majority of its variants suffer from the parameter tuning issue.
3. The use of Non-dominate sorting genetic algorithm has been neglected by majority of researchers to optimize the existing clustering techniques.

## 3.2 Problem Definition

Clustering techniques usually segments a remote sensing image into various objects and thereafter objects are clustered based on their characteristics. Basically these clusters are used to automatically detect the high-resolution images. However, it demands accurate descriptions of the characteristics of objects. As brought out above that the existing technique suffers from two problems such as parameter tuning and noise issue. Therefore, to overcome these issues, in this paper, a novel adaptive NSGA density-based clustering technique is proposed.

NSGA DBSCAN improves noise issues and smoothening by using morphological operations and tuning parameter issue by using NSGA Algorithm. Morphology is a broad set of clustering operations that process image based on shape. For smoothening images and to remove imperfection in images, three morphological operations are used here i.e. imerode and imreconstruct.

- Imerode function erodes the greyscale, binary or packed binary image. Imerode function set output pixel to 0, if any of the function set to 0. For binary images imerode function takes value 1 and for greyscale images imerode function takes value for uint8 is 255.

- Imreconstruct conceptually known as marker image which processes repeated dilation of the segmented image. It reconstructs image marker under the image mask. Marker and mask are the two intensity images or two binary images with the same size.

NSGA is a well-known meta-heuristic technique which is used to tune the parameters of DBSCAN based clustering technique for remote sensing images. Initially, random population is generated. Thereafter, for each random solution DBSCAN is implemented for clustering process. The solution that has accurate cluster with lesser noise are selected as non-dominated solutions. Thereafter three operators are used to explore the proposed technique further i.e. selection, crossover and mutation.

- Selection: Selection operator selects values for the next population.
- Crossover: Crossover produces fraction between 0 and 1 which combines two values to produce next values. Crossover is also known as Recombination.
- Mutation: After the random population is generated, GA makes small changes in random population to create their mutant children.

After, getting the termination condition, tuned parameters for DBSCAN are obtained. Figure 3.1 shows step by step working of selection, crossover and mutation operator.

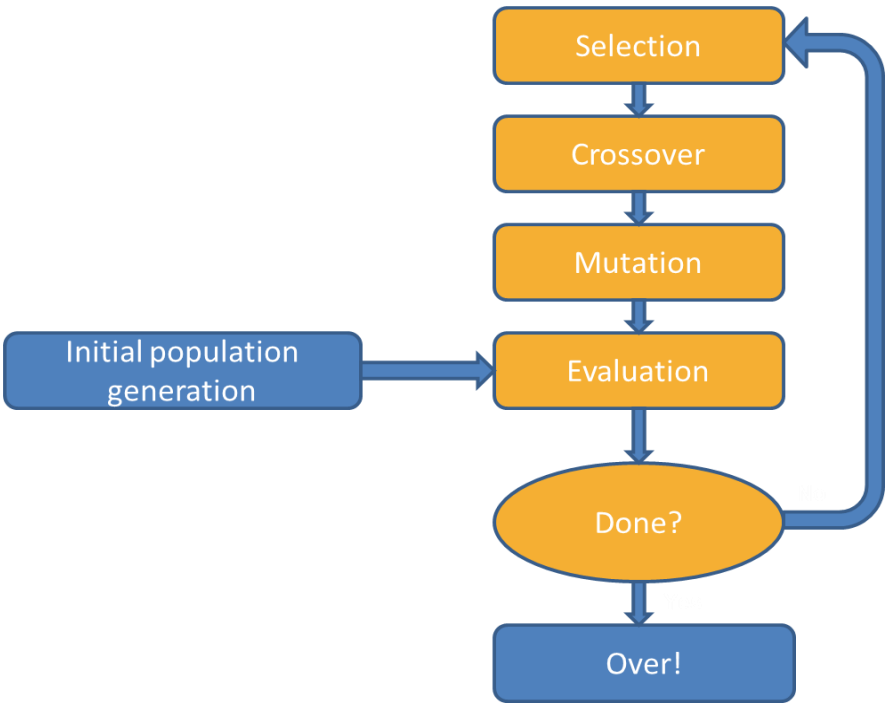


Figure 3.1: Selection, Crossover and Mutation

## 3.3 Objectives

To overcome these gaps following objectives are formulated:

1. To study and analyses the performance of existing DBSCAN technique to cluster remote sensing images.
2. To propose NSGA based adaptive DBSCAN technique to improve the clustering of remote sensing images.
3. To compare the proposed techniques with existing DBSCAN by considering the well-known quality metrics.

# CHAPTER 4

# PROPOSED METHODOLOGY

In the below section, an adaptive NSGA DBSCAN have proposed to formulate the problem that have been discussed above in chapter 3.

## 4.1 Flowchart

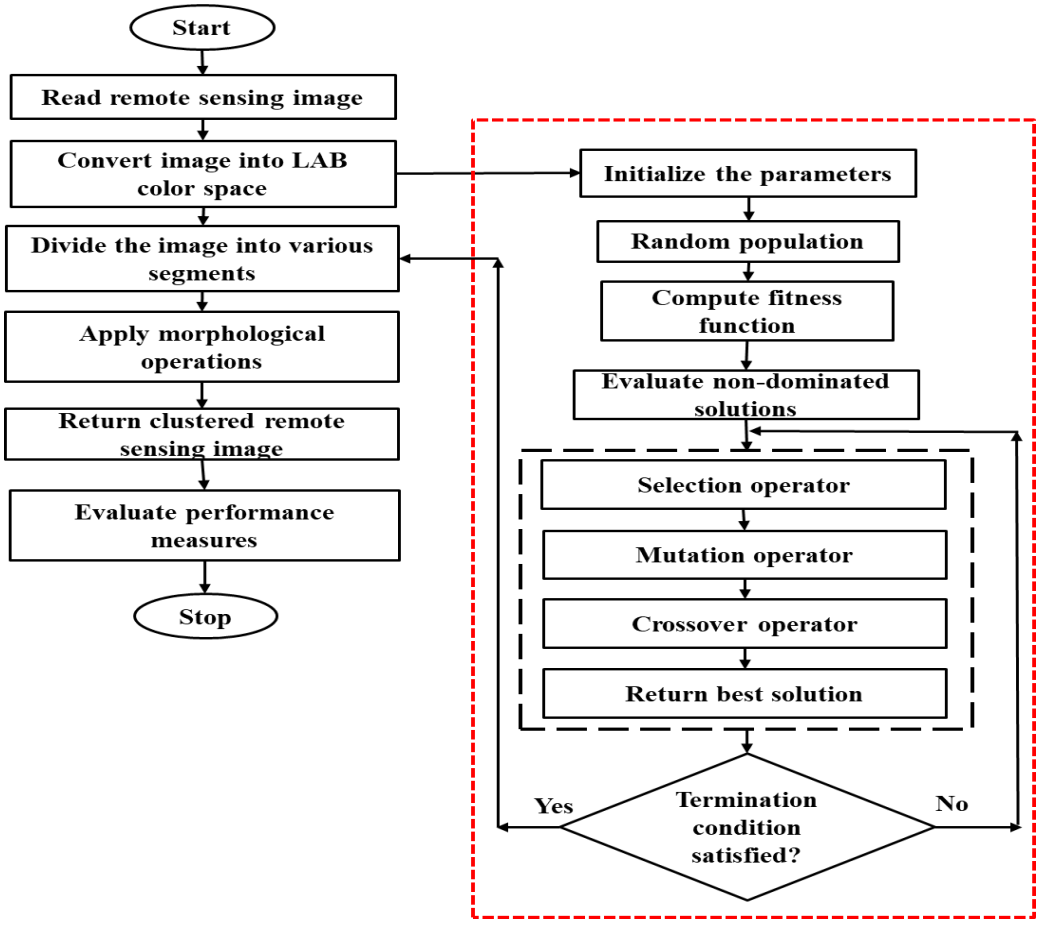


Figure 4.1: Graphical representation of proposed methodology

## 4.2 Pseudocode

**Step 1** – Read remote sensing dataset.

**Step 2** – Divide the remote sensing dataset into training and testing sets by using Naïve Bayes model.

**Step 3** – Read the image.

**Step 4** – Then convert the resulted image into LAB color space.

**Step 5** – Initialize the parameters which is obtained using clustered images with cluster as classification index. Eight features are there in training data i.e., mean, variance, texture, diameter, shape, energy, dissimilarity, entropy etc.

**Step 6** – Obtain the random population size of the dataset.

**Step 7** – Evaluate Euclidean distance measurement and then sort it.

**Step 8** – Apply DBSCAN on distance matrix with the two input parameters i.e. Epsilion and Minimum number of points.

**Step 9** – To calculate the optimal solution of the images compute the fitness function which is based on root mean square error.

**Step 10** – Apply selection, Mutation and recombination operations to get the desired results.

**Step 11** – Returns the best solution and check if the terminated condition is satisfied, if not then apply the same operations on the image to get the optimal solution.

**Step 12-** Divide the image into various segments using N-cuts matrix and apply morphological operations which will first smooth the image and then remove noise from the image.

**Step 13** – Apply K-Means on Ncut matrix to obtain final segmented image in LAB space.

**Step 14** – Convert the image into RGB colorspace.

**Step 15** – Evaluate classification index of matrix weighted.

**Step 16** – Return classification image based on index.

**Step 17** –Return segmented image of DBSCAN based N-cuts and segmentation analysis of DBSCAN based N-cuts.

---

**4.3 Algorithm:** Efficient clustering technique using NSGA DBSCAN.

---

NSGA DBSCAN (Layers, Window size (w_size), Structure Element (se), image (n), path (p), crossnum (c), Population Size (pop), VarHigh, VarLow, Treebagger (b2), predicted value (b1))

1. Load the remote sensing dataset.
2. Divide the training and testing dataset using holdout.
3. Select the image from the file
    a. read the input image n and provide the input path p

    b. Convert the input image using double datatype.
4. Convert the input image using double datatype.
5. Set the layers, window size, structure element of the image.
6. If mod(c!=0)
    a. c=c+1;
7. if i=1 to c
    a. select i
    b. else i+1
8. if pop>VarHigh
    a. select VarHigh
9. if pop<VarLow
    a. select VarLow
10. implement Treebagger model b2
11. if b1= b2
    a. best accuracy=accuracy
12. else
13. calculate Minmat and Meanmat

# CHAPTER 5

# EXPERIMENTS AND RESULTS ANALYSIS

## 5.1 Experiments analysis

The proposed technique is evaluated using MATLAB tool. Results are carried out on different sensor images and each image having different features such as mean, variance, texture, diameter, shape, energy, dissimilarity, entropy etc. Given features are divided based upon their training and testing datasets and comparison shows that results generated with Non-Dominate Based Sorting Genetic Algorithm is better than existing DBSCAN with N-Cuts Algorithm.

The results can be carried out based upon analysis of segmented image which describes the various parameters to test the dataset in terms of their performance measures i.e. Accuracy, Error rate and RMSE value.

1. In experiment analysis, results have been carried out by DBSCAN based N-Cuts produces 82% accurate results while results have been carried out by NSGA-DBSCAN produces 99% accurate results of remote sensing image while testing on different datasets.

   a. Calculate correct accuracy classification from confusion matrix as 1-c.

   b. Apply Structure Element TP/(TP+FN) as SE = cm (2,2)/sum (cm (2, :)) TN/(TN+FP) SP=cm (1,1)/sum (cm (1, :)) to a binary image.

2. The inaccurate predicated values are called as error. If the targeted values are absolute, then it is called as error rate. In experiment analysis, results have been carried out by DBSCAN based N-Cuts produces 0.1795 inaccurate results while results have been carried out by NSGA DBSCAN produces 0.0088 inaccurate results of remote sensing image while testing on different datasets.

$$\text{Error Rate} = \left[1 - \frac{Correct\ Predictions}{Total\ Predictions}\right] * 100$$

3. **Root Mean Square Error-** It is also known as Root Mean Square Deviation. It uses fitness function value as an objective or cost function to compute better results than DBSCAN Based N-Cuts by adapting density value automatically. In

experiment analysis, results have been carried out by DBSCAN based N-Cuts produces 0.4237 RMSE value while results have been carried out by NSGA DBSCAN produces 0.0941 RMSE value of remote sensing image while testing on different datasets.

RMSE $(x)$ returns the root mean square value of the objective function.

N=input

X_rms=sqrt(1/N.*(sum(x(ii).^2)));

## 5.2 Results Analysis

First, datasets need to be uploaded as CSV file. This task is done in the following steps: first, read image and divide the dataset in training and testing sets. Morphology operation is also used to remove noise and for smoothening of image. Next, analysis is carried out to compare the DBSCAN based N-Cuts segment results and NSGA-Based DBSCAN segment results.

### 5.2.1 Visual Analysis of Quick Bird Dataset

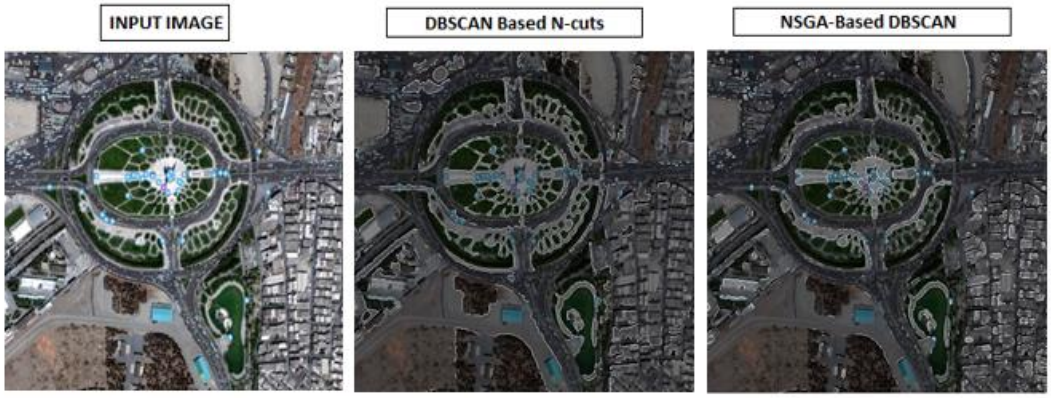


Figure 5.1: Quick Bird Analysis

Analysis carried out on input image with respect to DBSCAN Based N-Cuts and NSGA-Based DBSCAN shows clustering result as shown in figure 5.1.

Following values are used to determine the accuracy, error rate and RMSE value of images as shown in table 5.1.1.

Table 5.1.1: Quick Bird Results

| Quick Bird Analysis | DBSCAN Based N-Cuts | NSGA-Based DBSCAN |
|---|---|---|
| Accuracy | 0.8205 | 0.9912 |
| Error Rate | 0.1795 | 0.0088 |
| RMSE | 0.4237 | 0.0941 |

### 5.2.2 Visual Analysis of IKONOS Dataset

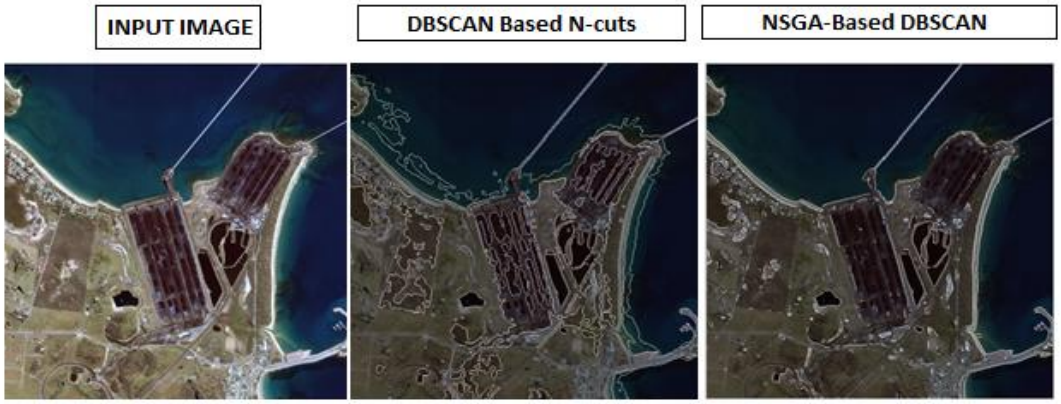


Figure 5.2: IKONOS Analysis

Analysis carried out on input image with respect to DBSCAN Based N-Cuts and NSGA-Based DBSCAN shows clustering result as shown in figure 5.2.

Following values are used to determine the accuracy, error rate and RMSE value of images as shown in table 5.1.2.

Table 5.1.2: IKONOS Results

| IKONOS Analysis | DBSCAN Based N-Cuts | NSGA-Based DBSCAN |
|---|---|---|
| Accuracy | 0.8500 | 0.9652 |
| Error Rate | 0.1500 | 0.0348 |
| RMSE | 0.3873 | 0.1865 |

### 5.2.3 Visual Analysis of MODIS Dataset

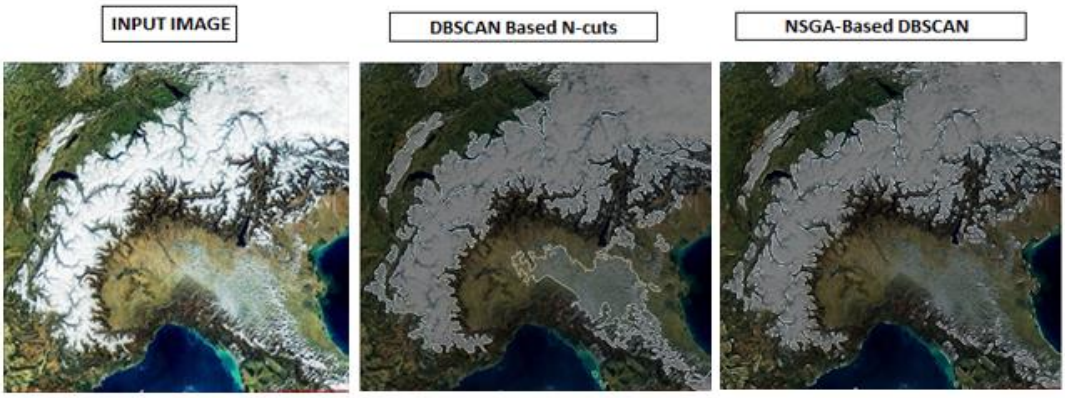


Figure 5.3: MODIS Analysis

Analysis carried out on input image with respect to DBSCAN Based N-Cuts and NSGA-Based DBSCAN shows clustering result as shown in figure 5.3.

Following values are used to determine the accuracy of images as shown in table 5.1.3.

Table 5.1.3: MODIS Results

| MODIS Analysis | DBSCAN Based N-Cuts | NSGA-Based DBSCAN |
|---|---|---|
| Accuracy | 0.8407 | 0.9652 |
| Error Rate | 0.1593 | 0.0348 |
| RMSE | 0.3991 | 0.1865 |

### 5.2.4 Visual Analysis of SPOT Dataset

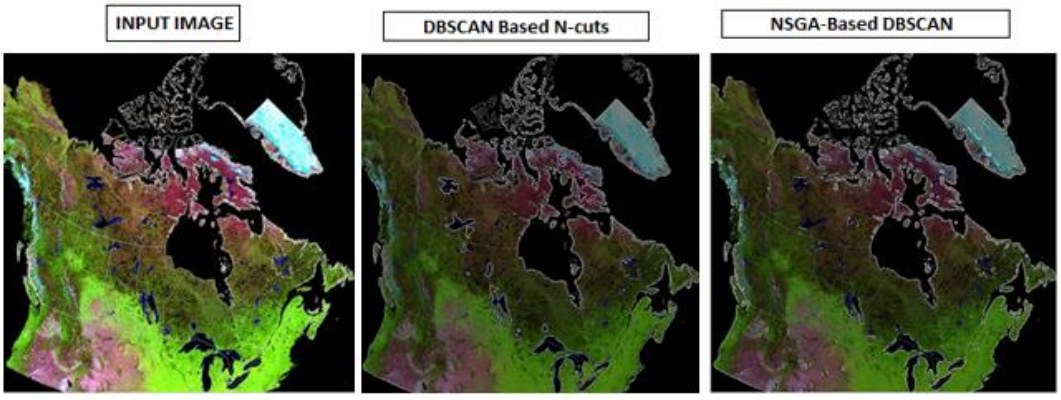


Figure 5.4: SPOT Analysis

Analysis carried out on input image with respect to DBSCAN Based N-Cuts and NSGA-Based DBSCAN shows clustering result as shown in figure 5.4.

Following values are used to determine the accuracy, error rate and RMSE value of images as shown in table 5.1.4.

Table 5.1.4: SPOT Results

| SPOT Analysis | DBSCAN Based N-Cuts | NSGA-Based DBSCAN |
|---|---|---|
| Accuracy | 0.8345 | 0.9407 |
| Error Rate | 0.1565 | 0.0593 |
| RMSE | 0.3956 | 0.2436 |

## 5.3 Performance Analysis

### 5.3.1 Result of Quick Bird Dataset

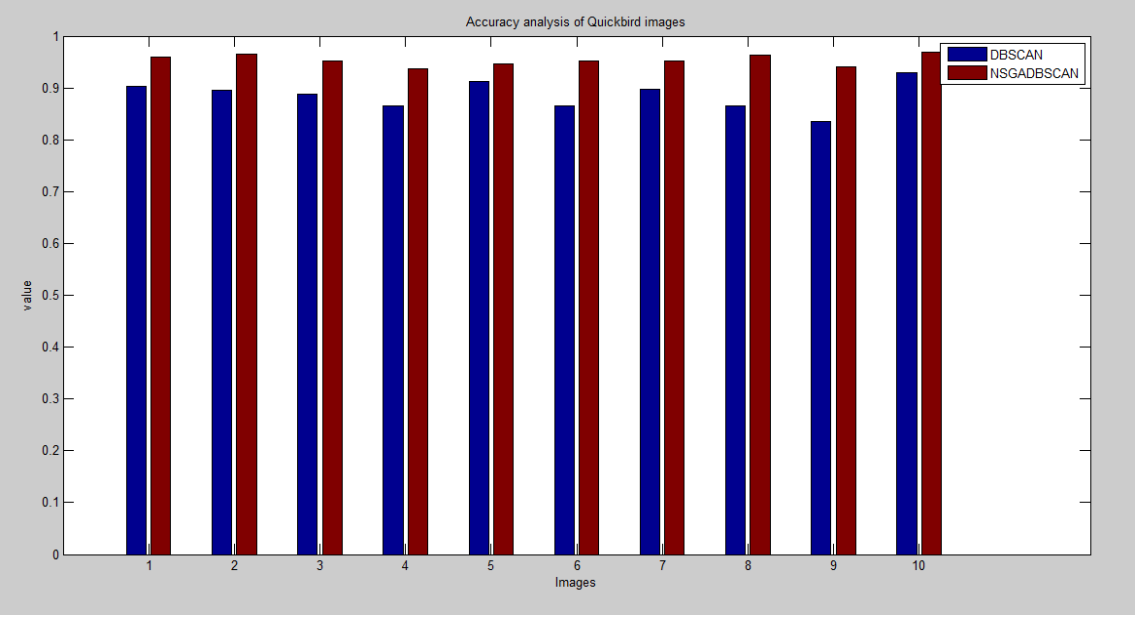


Figure 5.5: Quick Bird Analysis

### 5.3.2 Result of IKONOS Dataset

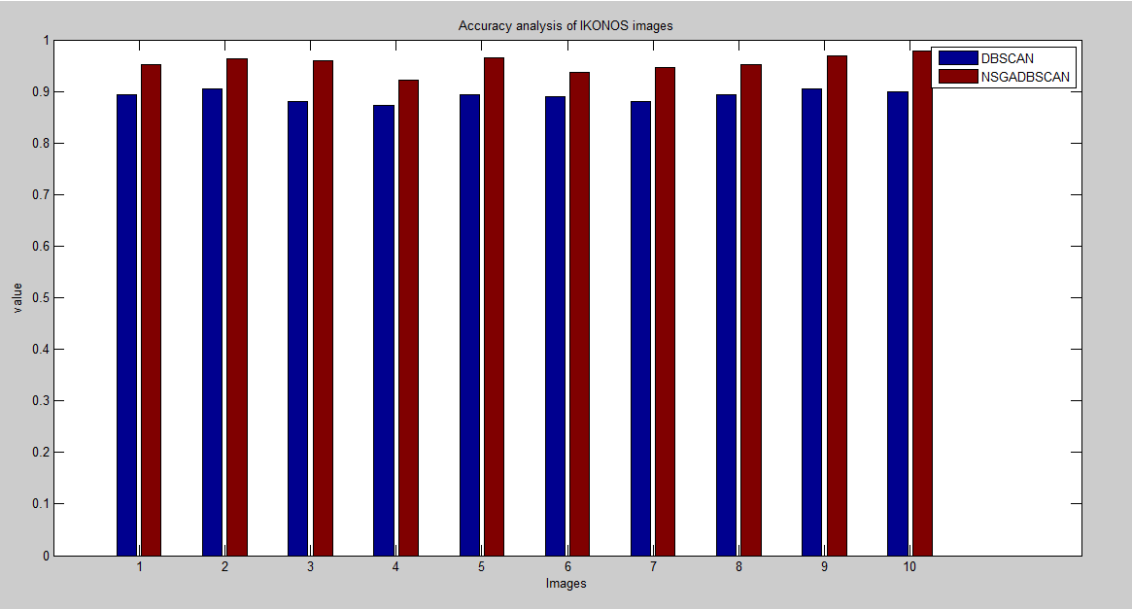


Figure 5.6: IKONOS Analysis

### 5.3.3 Result of MODIS Dataset

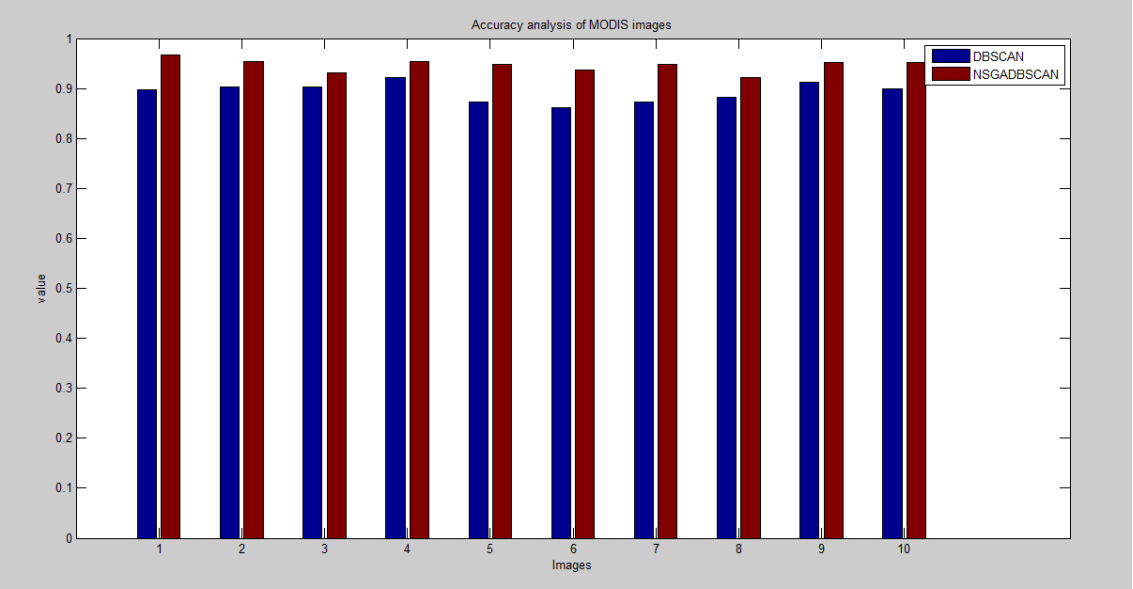


Figure 5.7: MODIS Analysis

### 5.3.4 Result of SPOT Dataset

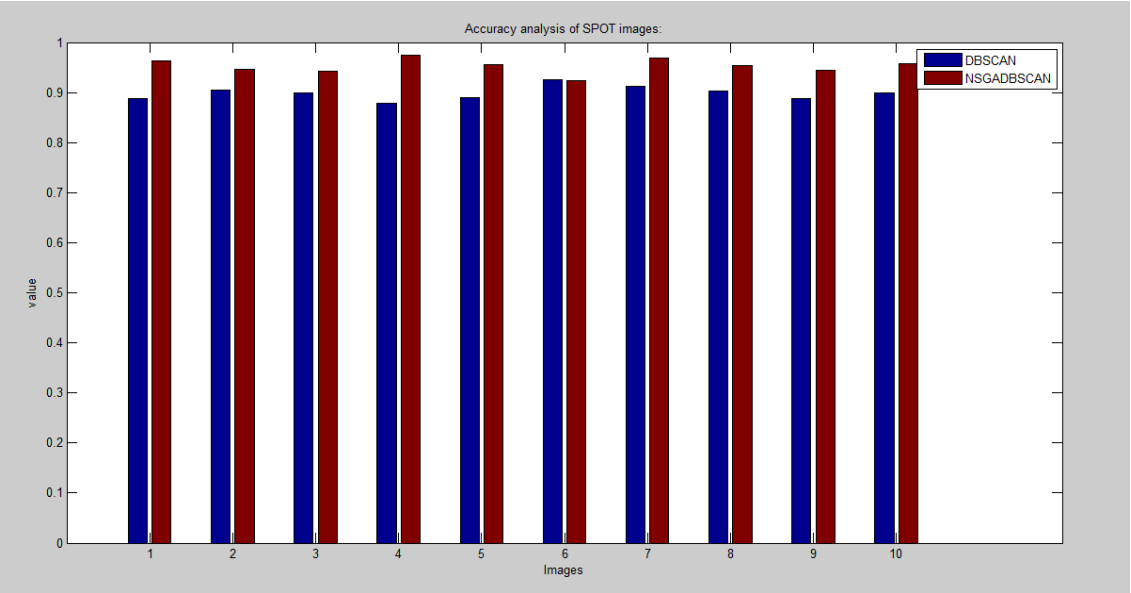


Figure 5.8: SPOT Analysis

# CHAPTER 6

# CONCLUSION AND FUTURE WORK

In the research work, density-based spatial clustering discovers arbitrary shape clusters which follows density-based notion of clusters. Given Epsilion ($\epsilon$) as input attribute, unlike k-means clustering, it explores all the possible clusters by categorizing core point, border point or outlier. However, it can be computationally expensive as computation of nearest neighbors demands to compute all pair wise proximities. Additional result of implementation contains KD-Trees to save the data that allow efficient retrieval of data and bring down the time complexity from $O(m^2)$ to $O(m \log m)$. Therefore, improving the computational speed of the DBSCAN is the main motivation behind this research work. Because majority of existing clustering techniques used for remote sensing images suffer from noise issue, which may degrade the performance of remote sensing vision systems.

Therefore, to overcome this issue in this research work, a novel adaptive density-based clustering (NSGA DBSCAN) technique with noise is designed. Initially, DBSCAN identify the core point, border point or noise point. Thereafter, it minimizes noise regions and put an edge among all core points that are sharing Epsilion ($\epsilon$) values of each other. Thereafter, DBSCAN make each group of connected core points into an individual cluster. In the end, it assigns each border point as clusters of their associated core points.

Additionally, a well-known meta-heuristic technique (i.e., non-dominated sorting based genetic algorithm (NSGA) is used to tune the parameters of DBSCAN based clustering technique for remote sensing images. Initially, random population is generated. Thereafter, for each random solution DBSCAN is implemented for clustering process. The solution that has accurate cluster with lesser noise are selected as non-dominated solutions. Thereafter, selection, mutation and crossover operator are used to explore the proposed technique further. After, getting the termination condition, tuned parameters for DBSCAN are obtained. Extensive experiments are carried out by considering benchmark remote sensing images (i.e., obtained from satellite sensors such as QUICKBIRD, IKONOS, MODIS, SPOT etc.). From visual and quantitative analysis, it is found that the proposed technique outperforms existing techniques in terms of

Accuracy and Root mean square error. Therefore, the proposed technique is more applicable to real-time imaging systems.

**Future directions for the proposed work:**

1. In this work, on the tuning of the DBSCAN has been done, therefore in future, one may use the NSGA to optimize other clustering techniques such as K-means, Fuzzy-c-means, etc.
2. In near future proposed technique will be applied to the fields such as biomedical processing, image machine learning etc. to evaluate the performance of the proposed technique for other applications.
3. Also, in near future a novel multi-objective fitness function will also be considered to enhance the results further.

# PUBLICATION

Nitika, Dr.V.P. Singh, Dr.Vinay Gautam,"NSGA-DBSCAN: An Efficient Clustering Technique" (communicated to International Journal of Engineering Science and Computing)

## REFERNCES


[1] S.Kisilevich, F.Mansmann, and k.Daniel "P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos." *Proceedings of the 1st international conference and exhibition on computing for geospatial research & application*. ACM, 2010.

[2] P.Berkhin "A survey of clustering data mining techniques." *Grouping multidimensional data*. Springer, Berlin, Heidelberg, 2006. 25-71.3.

[3] Mann, K.Amandeep and K.Navneet "Survey paper on clustering techniques." *International Journal of Science, Engineering and Technology Research* 2.4 (2013): pp-0803.

[4] Yadav, Jyoti and S.Monika. "A Review of K-mean Algorithm." *International Journal of Engineering Trends and Technology (IJETT)–Volume* 4 (2013).

[5] Raval, R.Unnati and J.Chaita "Implementing and Improvisation of K-means Clustering." *International Journal of Computer Science and Mobile Computing* 4.11 (2015): 72-76.

[6] T.Velmurugan "Efficiency of k-means and k-medoids algorithms for clustering arbitrary data points." *Int. J. Computer Technology & Applications* 3.5 (2012): 1758-1764.7.

[7] FBAl.Abid "A Novel Approach for PAM Clustering Method." *International Journal of Computer Applications* 86.17 (2014).

[8] RT.Ng and J.Han.''CLARANS: A method for clustering objects for spatial data mining." *IEEE transactions on knowledge and data engineering* 14.5 (2002): 1003-1016.

[9] JS.Saket. and S.Pandya. "An Overview of Partitioning Algorithms in Clustering Techniques."

[10] M.Ester,HP.Kriegel,J.Sander,X.XU. "A density-based algorithm for discovering clusters in large spatial databases with noise." *Kdd*. Vol. 96. No. 34. 1996.

[11] M.Ankerst,MM.Breuing,HP.Kriegel and J.Sander. "OPTICS: ordering points to identify the clustering structure." *ACM Sigmod record*. Vol. 28. No. 2. ACM, 1999.

[12] H.Rehioui,A.Idrissi,M.Abourezq and F.Zegrari . "DENCLUE-IM: A new approach for big data clustering." *Procedia Computer Science* 83 (2016): 560-567.

[13] W.Wang, J.Yang, and R. Muntz. "STING: A statistical information grid approach to spatial data mining." *VLDB*. Vol. 97. 1997.

[14] G.Sheikholeslami, S.Chatterjee, and A.Zhang. "WaveCluster: a wavelet-based clustering approach for spatial data in very large databases." *The VLDB Journal—The International Journal on Very Large Data Bases* 8.3-4 (2000): 289-304.

[15] R.Agrawal,JE.Gehrke,D.Gunopulos and P.Raghavan . "Automatic subspace clustering of high dimensional data for data mining applications." U.S. Patent No. 6,003,029. 14 Dec. 1999.

[16] S.Saini, and P.Rani. "A Survey on STING and CLIQUE Grid Based Clustering Methods." *International Journal of Advanced Research in Computer Science* 8.5 (2017).

[17] P.Liu, D.Zhou, and N.Wu. "VDBSCAN: varied density based spatial clustering of applications with noise." *Service Systems and Service Management, 2007 International Conference on*. IEEE, 2007.

[18] D.Birant, and A.Kut. "ST-DBSCAN: An algorithm for clustering spatial–temporal data." *Data & Knowledge Engineering* 60.1 (2007): 208-221.

[19] J.Hou, H.Gao, and X.Li. "DSets-DBSCAN: a parameter-free clustering algorithm." *IEEE Transactions on Image Processing* 25.7 (2016): 3182-3193.

[20] PH.Swain., and SM.Davis. "Remote sensing: the quantitative approach." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 6 (1981): 713-714.

[21] J.Hou, W.Liu, E.Xu and H.Cui. "Towards parameter-independent data clustering and image segmentation." *Pattern Recognition* 60 (2016): 25-36.

[22] T.Blaschke, S.Lang, E.Lorup, J.Strobl and P.Zeil . "Object-oriented image processing in an integrated GIS/remote sensing environment and perspectives for environmental applications." *Environmental information for planning, politics and the public* 2 (2000): 555-570.

[23] Y.Li, H.Lu, L.Zhang, S.Yang and S.Serikawa. "Color image segmentation using fast density-based clustering method." *Future Communication, Computing, Control and Management*. Springer, Berlin, Heidelberg, 2012. 593-598.

[24] Y.Li, H.Lu, L.Zhang, S.Yang and S.Serikawa. " Density based color segmentation using fast density-based clustering method." *Future Communication, Computing, Control and Management*. Springer, Berlin, Heidelberg, 2012. 593-598.

[25] O.Sudana, D.Putra, M.Sudarma, RS.Hartati and A.Wirdiani. "Image clustering of complex balinese character with DBSCAN algorithm." *Journal of Engineering Technology* 6.1 (2018): 548-558.

[26] J.Shen, X.Hao, Z.Liang, Y.Liu, W.Wang and L.Shao. "Real-time superpixel segmentation by DBSCAN clustering algorithm." *IEEE Transactions on Image Processing* 25.12 (2016): 5933-5942.

[27] D.Wassermann, L.Bloy, E.Kanterakis, R.Verma and R.Deriche. "Unsupervised white matter fiber clustering and tract probability map generation: Applications of a Gaussian process framework for white matter fibers." *NeuroImage* 51.1 (2010): 228-241.

[28] AE.Bayá., MG. Larese, and R.Namías. "Clustering stability for automated color image segmentation." *Expert Systems with Applications* 86 (2017): 258-273.

[29] RT.Moore and MC. Hansen. "Google Earth Engine: a new cloud-computing platform for global-scale earth observation data and analysis." *AGU Fall Meeting Abstracts*. 2011.

[30] L.Khiali, D.Lenco, and M.Teisseire. "Object-oriented satellite image time series analysis using a graph-based representation." *Ecological Informatics* 43 (2018): 52-64.

[31] KA.Borden., and SL.Cutter. "Spatial patterns of natural hazards mortality in the United States." *International journal of health geographics* 7.1 (2008): 64.

[32] HK.Gibbs, S.Brown, JO.Niles and JA.Foley. "Monitoring and estimating tropical forest carbon stocks: making REDD a reality." *Environmental Research Letters* 2.4 (2007): 045023.

ORIGINALITY REPORT

| 9% | 5% | 7% | 4% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Marzena Kryszkiewicz. "Faster Clustering with DBSCAN", Advances in Soft Computing, 2005<br>Publication | 1% |
| 2 | Tripathy, Animesh, Sumit Kumar Maji, and Prashanta Kumar Patra. "FDCA: A fast density based clustering algorithm for spatial database system", 2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011), 2011.<br>Publication | <1% |
| 3 | Submitted to Caledonian College of Engineering<br>Student Paper | <1% |
| 4 | Hou, Jian, Huijun Gao, and Xuelong Li. "DSets-DBSCAN: A Parameter-Free Clustering Algorithm", IEEE Transactions on Image Processing, 2016.<br>Publication | <1% |
| 5 | OZKOK, Fatma Ozge and CELİK, Mete. "A New Approach to Determine Eps Parameter of | <1% |

DBSCAN Algorithm", Selçuk Üniversitesi Teknoloji Fak., 2017.
Publication

| | | |
|---|---|---|
| 6 | rss.onlinelibrary.wiley.com<br>Internet Source | <1% |
| 7 | Submitted to The Glasgow School of Art<br>Student Paper | <1% |
| 8 | ijarcsse.com<br>Internet Source | <1% |
| 9 | Submitted to Thapar University, Patiala<br>Student Paper | <1% |
| 10 | Chih-Wei Liu. "FICA: A New Data Clustering Technique Based on Partitional Approach for Data Mining", 2007 International Conference on Machine Learning and Cybernetics, 08/2007<br>Publication | <1% |
| 11 | Deepti Joshi, Ashok K. Samal, Leen-Kiat Soh. "Density-based clustering of polygons", 2009 IEEE Symposium on Computational Intelligence and Data Mining, 2009<br>Publication | <1% |
| 12 | Mukesh Kumar, Santanu Kumar Rath. "Classification of microarray using MapReduce based proximal support vector machine classifier", Knowledge-Based Systems, 2015<br>Publication | <1% |

| | | |
|---|---|---|
| 13 | Submitted to University of Central Lancashire<br>Student Paper | <1% |
| 14 | Xiaowei Xu. "A Fast Parallel Clustering Algorithm for Large Spatial Databases", High Performance Data Mining, 2002<br>Publication | <1% |
| 15 | Submitted to Pathfinder Enterprises<br>Student Paper | <1% |
| 16 | advanceddataanalytics.net<br>Internet Source | <1% |
| 17 | Submitted to Birla Institute of Technology and Science Pilani<br>Student Paper | <1% |
| 18 | "Medical Image Understanding and Analysis", Springer Nature, 2017<br>Publication | <1% |
| 19 | Harmanpreet Kaur, Navleen Kaur. "Efficient image fusion using visual salient features and cross-contrast with edge weakening guided image filter", 2017 Fourth International Conference on Image Information Processing (ICIIP), 2017<br>Publication | <1% |
| 20 | Submitted to Indian Institute of Technology, Kanpur<br>Student Paper | <1% |

| | | |
|---|---|---|
| 21 | Marzena Kryszkiewicz. "TI-DBSCAN: Clustering with DBSCAN by Means of the Triangle Inequality", Lecture Notes in Computer Science, 2010<br>Publication | <1% |
| 22 | Submitted to KIIT University<br>Student Paper | <1% |
| 23 | Dua, R.L., and N. Gupta. "Fast color image quantization based on bacterial foraging optimization", Fourth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom2012), 2012.<br>Publication | <1% |
| 24 | Singh, Sweta, Sanya Ambegaokar, Kiran Singh Champawat, Animesh Gupta, and Shirish Sharma. "Time series analysis of clustering high dimensional data in precision agriculture", 2015 International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS), 2015.<br>Publication | <1% |
| 25 | calidad.uniovi.es<br>Internet Source | <1% |
| 26 | Submitted to Victoria University<br>Student Paper | <1% |

| | | |
|---|---|---|
| 27 | www.doh.dot.state.nc.us<br>Internet Source | <1% |
| 28 | www.docstoc.com<br>Internet Source | <1% |
| 29 | repository.um.edu.my<br>Internet Source | <1% |
| 30 | Goran Šimić. "chapter 10 E-Government Documents and Data Clustering", IGI Global, 2015<br>Publication | <1% |
| 31 | www.slideshare.net<br>Internet Source | <1% |
| 32 | Hong-Jiang Zhang. "On clustering and retrieval of video shots", Proceedings of the ninth ACM international conference on Multimedia - MULTIMEDIA 01 MULTIMEDIA 01, 2001<br>Publication | <1% |
| 33 | tutcris.tut.fi<br>Internet Source | <1% |
| 34 | dspace.thapar.edu:8080<br>Internet Source | <1% |
| 35 | www.onr.navy.mil<br>Internet Source | <1% |
| 36 | link.springer.com<br>Internet Source | <1% |

| | | |
|---|---|---|
| 37 | repository.tudelft.nl<br>Internet Source | <1% |
| 38 | cfsites1.uts.edu.au<br>Internet Source | <1% |
| 39 | Feng Guo. "A Local Density Based Spatial Clustering Algorithm with Noise", 2006 IEEE International Conference on Systems Man and Cybernetics, 10/2006<br>Publication | <1% |
| 40 | Submitted to Visvesvaraya Technological University<br>Student Paper | <1% |
| 41 | perun.im.ns.ac.yu<br>Internet Source | <1% |
| 42 | en.cs.ustc.edu.cn<br>Internet Source | <1% |
| 43 | Peng Liu, Dong Zhou, Naijun Wu. "VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise", 2007 International Conference on Service Systems and Service Management, 2007<br>Publication | <1% |
| 4 | Gabriela Serban, Grigoreta Moldovan. "A New k-means Based Clustering Algorithm in Aspect Mining", 2006 Eighth International Symposium | <1% |

on Symbolic and Numeric Algorithms for Scientific Computing, 2006
Publication

| | | |
|---|---|---|
| 45 | Bing Liu. "A Fast Density-Based Clustering Algorithm for Large Databases", 2006 International Conference on Machine Learning and Cybernetics, 2006<br>Publication | <1% |
| 46 | Lecture Notes in Computer Science, 2006.<br>Publication | <1% |
| 47 | Lecture Notes in Computer Science, 2009.<br>Publication | <1% |

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | < 5 words |
| Exclude bibliography | On | | |