ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.9

NSGA DBSCAN: An Efficient Clustering Technique

Nitika[1], V.P. Singh[2] , Vinay Gautam[3]
Research Scholar[1], Associate Professor[2], Lecturer[3]
Department of Computer Science & Engineering
Thapar Institute of Engineering & Technology, Patiala, Punjab, India

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. The accurate descriptions of the characteristics of the objects are highly demanded. The majority of existing clustering techniques are suffered from noise and parameter tuning issues, which may degrade the performance of remote sensing vision systems. The techniques used in improving the computational speed of the clustering technique have been considered in this research work. An algorithm based on novel adaptive density-based clustering technique with noise has been proposed to tune the parameters of DBSCAN for remote sensing images. The results are based on 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, Clustering Technique, K-Means.

INTRODUCTION

With the emerging technologies and the development of smart devices massive amount of information gathered via IOT Technologies, recommendation systems, business intelligence, GPS systems, bioinformatics, chemo informatics and social networking domains caused a worldwide buzz in real-world scenario. The amount of data generated from such sources provides an efficient way of querying big data clusters. Researchers have gained considerable attention in data clustering which resulted in the development of wide variety of successful clustering algorithm. The concept of geospatial data clustering introduce new emerging platform for computing human mobility via GPS locations[1]. Spatial data clustering is one of the popular data mining clustering techniques that are very useful in cases of spatial data when the data size is sufficiently large.

Clustering algorithms 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. Clustering in remote sensing images is utilized to automatically detect the high-resolution images. However, it demands accurate descriptions of the characteristics of objects.

CLUSTERING TECHNIQUES

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 is available. Clustering algorithms helps to determine the intrinsic grouping from unlabeled training data to predict the unlabelled data from the labeled set of training points. The clustering techniques are broadly classified in their different types as shown in figure1.

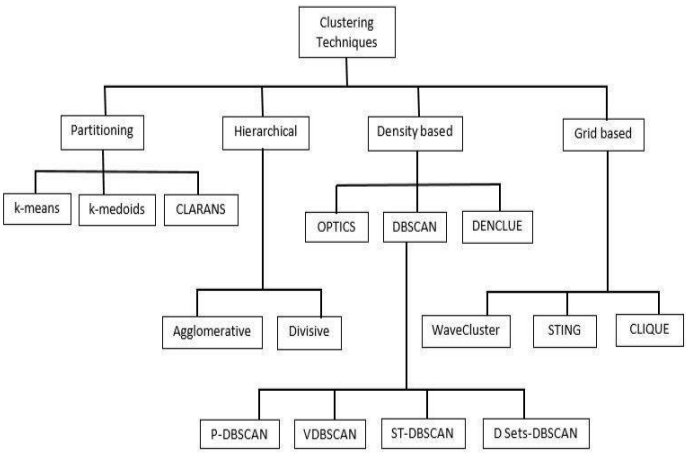

Figure1: Clustering Techniques

Partitioning Clustering

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

Hierarchical Clustering

Hierarchical clustering provides a series of hierarchical decomposition of the given objects. Hierarchical algorithms follow recursive process which can be divided into two approaches: top-down (or divisive approach) and bottom-up (or agglomerative approach)[3]. 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[6]. 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.

Density-based Clustering

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[6]. 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.

- **Density-Based Spatial Clustering of Applications with Noise (DBSCAN)** is a dynamic density-based clustering algorithm[7]. 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. DBSCAN has number of variants which 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.

  VDBSCAN: VDBSCAN is varied density based spatial clustering of application with noise, which addresses varied-density datasets analysis[8]. 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.

  ST-DBSCAN: 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[9]. 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[10]. 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.

- **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 and minimum number of points to form a cluster (Minpts).

- **DENsity CLUstering (DENCLUE)** works differently from both DBSCAN an**d** 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.

Grid-based Clustering

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.

- **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 store 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.

- **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 (N).

- **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].

LITERATURE REVIEW

Remote sensing technologies produce huge amount of satellite images that can useful to monitor geographical areas. [17] have discussed about the object-oriented satellite image time

series analysis using graph-based representation. A new technique SITS is proposed to detect spatio-temporal data and produce statio-temporal clusters having similar behavior. In [18] author have presented image segmentation of noisy image between two density based clustering algorithm i.e. DBSCAN and Mean-shift. [19] have discussed density based cluster extension method to overcome over segmentation. Histogram Equalization is also proposed to enhance images by maximizing the overall intensity in images. In [1] author proposed P-DBSCAN which captures geo-tagged collection of photos of interesting places and events held where P stands for photo that has been captured by a user and uploaded with location coordinates. 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.

The earlier work focused on the analysis of K-means algorithm, DBSCAN algorithm, NCuts and soon over remote sensing images. Proposed work includes analysis of these algorithms including additional factor NSGA technique to overcome the issues highlighting in earlier work to make algorithm adaptive so as to improve noise and parameter tuning issue. We have designed NSGA-DBSCAN which is used to tune the parameters of DBCSCAN based clustering technique for remote sensing images.

PROPOSED APPROACH

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, two 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 known as the 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 here 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 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.

Algorithm: Efficient Clustering technique using NSGA DBSCAN.

Require: remote sensing image dataset, No. of layers, Clustering information.

Ensure: clustering of remote sensing images, compute the clusters, optimize the clusters.

Load the remote sensing dataset.
Divide the training and testing dataset using holdout.
Select the image from the file
    Read the input image n and provide the input path p
    Convert the input image using double datatype.
Convert the input image using double datatype.
Set the layers, window size, and structure element of the image.
If mod(c!=0)
    c=c+1;
if i=1 to c
    select i
    else i+1
if pop>VarHigh
    select VarHigh
if pop<VarLow
    select VarLow
implement Treebagger model b2
if b1= b2
    best accuracy=accuracy
else
calculate Minmat and Meanmat

EXPERIMENTAL 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.

Results Analysis

The following steps are used here: first, read image and divide the dataset in training and testing sets. Morphology operation is 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. From visual and performance analysis, it is found that the proposed technique outperforms existing techniques in terms of accuracy, root mean square error and error rate and found that existing technique producing approximately 83% accurate clustered images and proposed technique producing approximately 99% accurate clustered images.

Visual Analysis of Quick Bird Dataset

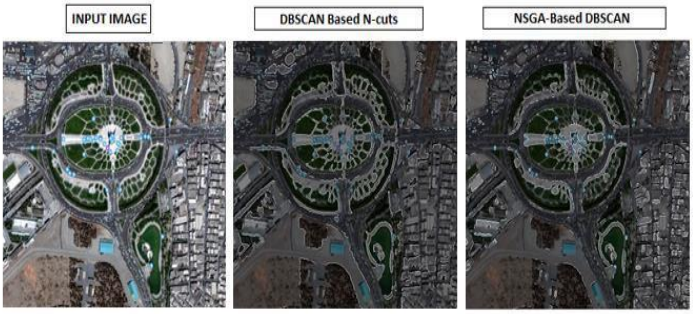

Figure 2: Visual Analysis of Quick Bird

Table 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

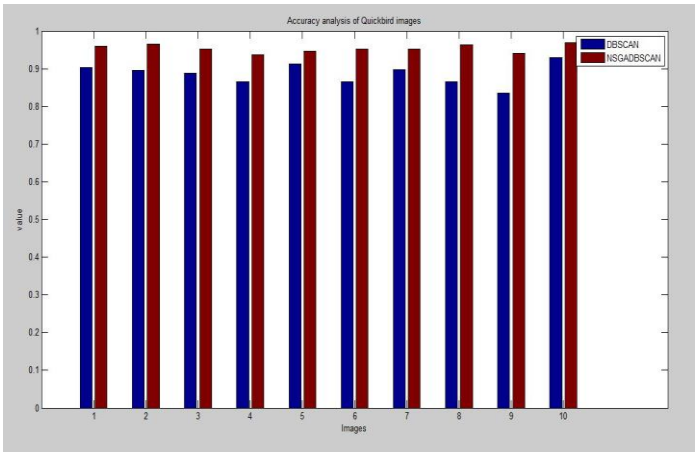

Figure 3: Quick Bird Performance Analysis

Visual Analysis of IKONOS Dataset

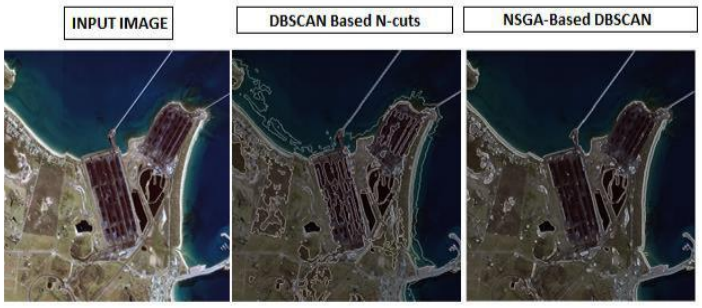

Figure 4: Visual Analysis of IKONOS

Table 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

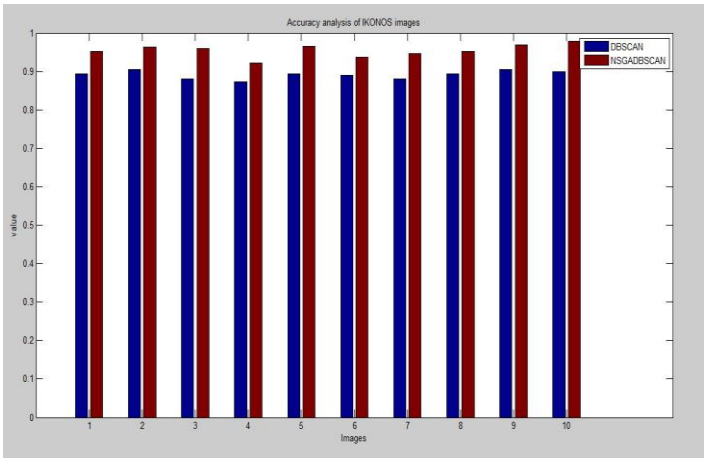

Figure 5: IKONOS Performance Analysis

Visual Analysis of MODIS Dataset

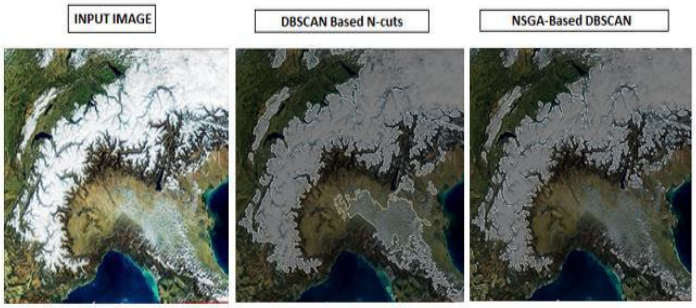

Figure 6: Visual Analysis of MODIS Analysis

Table 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

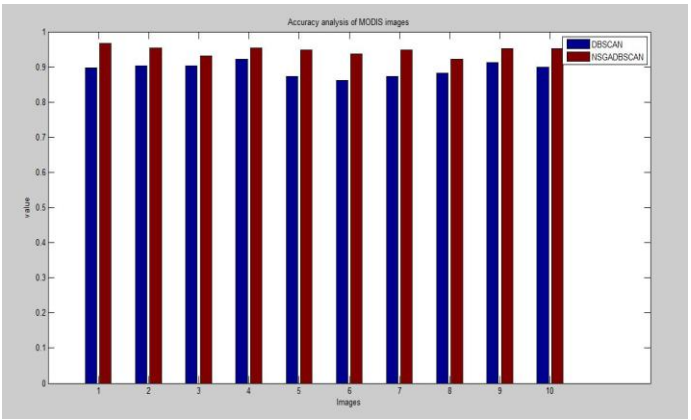

Figure 7: MODIS Performance Analysis

Visual Analysis of SPOT Dataset

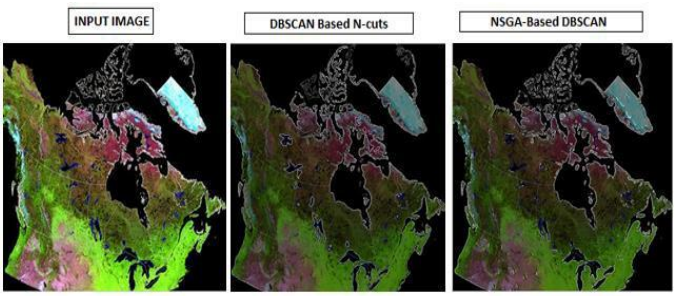

Figure 8: Visual Analysis of SPOT Analysis

Table 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

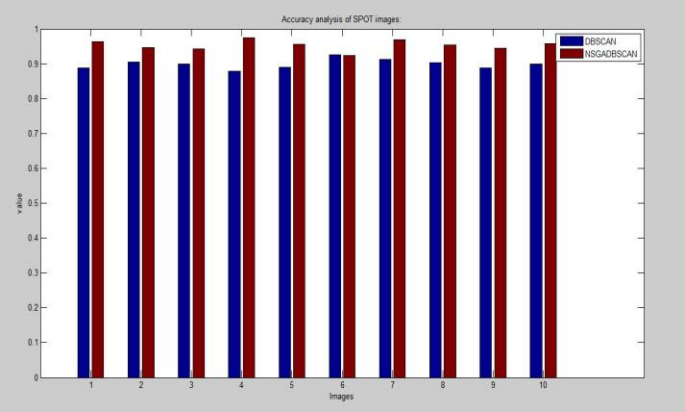

Figure 9: SPOT Performance Analysis

CONCLUSION

In this paper, the density-based spatial clustering discovers arbitrary shape clusters which follows density-based notion of clusters. 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. 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. 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 and found that existing technique producing approximately 83% accurate images and proposed technique producing approximately 99% accurate images. Therefore, the proposed technique is more applicable to real-time imaging systems.

REFERENCES

[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 geotagged 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] JS.Saket. and S.Pandya. "An Overview of Partitioning Algorithms in Clustering Techniques." [7] FBAl.Abid "A Novel Approach for PAM Clustering Method." International Journal of Computer Applications 86.17 (2014).

[7] 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.

[8] 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.

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

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

[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 . q"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] 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.

[18] 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.

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