Journal of Theoretical and Applied Information Technology
31st October 2016. Vol.92. No.2
© 2005 - 2016 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

K-NEAREST NEIGHBOR BASED DBSCAN CLUSTERING ALGORITHM FOR IMAGE SEGMENTATION

SURESH KURUMALLA[1], P SRINIVASA RAO[2]

[1]Research Scholar in CSE Department, JNTUK Kakinada
[2]Professor, CSE Department, Andhra University, Visakhapatnam, AP, India
E-mail id: kurumallasuresh@gmail.com

ABSTRACT

Clustering is a primary and vital part in data mining. Density based clustering approach is one of the important technique in data mining. The groups that are designed depending on the density are flexible to understand and do not restrict itself to the outlines of clusters. DBSCAN Algorithm is one of the density grounded clustering approach which is employed in this paper. The author addressed two drawbacks of DBSCAN algorithm i.e. determination of Epsilon value and Minimum number of points and further proposed a novel efficient DBSCAN algorithm as to overcome this drawback. This proposed approach is introduced mainly for the applications on images as to segment the images very efficiently depending on the clustering algorithm. The experimental results of the suggested approached showed that the noise is highly reduced from the image and segmentations of the images are also improved better compared to the existing image segmentation approaches.

Keywords: *Data Mining, Clustering, Density Based Clustering, DBSCAN, K-Nearest Neighbor, Image Segmentation.*

1. INTRODUCTION

Emerging of current methods for logical information gathering had ensued in huge scale accretion of information relating to dissimilar arenas. Data mining is the encouraging methodology that arrives at the conclusion in the domain of computer science where it mine the vital or beneficial knowledge from enormous data samples or huge amount of data. It is a stage in the Knowledge Discovery in Databases (KDD) procedure comprising of the application for data investigation and detection of procedures under adequate computational efficacy restrictions, generates a specific enumeration of trends above the data [8]. It employs sophisticated arithmetical study and modeling methods to reveal patterns and interactions concealed in administrative data samples. Clustering plays a significant part in the data mining. The approach of identifying similarities amongst information rendering to the features obtained in the data and merging analogous data entities into groups is known as clustering. It is an unsupervised categorization of distinguishing set of identical entities in outsized data samples without having definite clusters through unambiguous features. Clustering Methods are advantageous in numerous domains like excavating skin lesion images, pattern analysis, machine learning circumstances and numerous other domains.

Clustering is categorized into five kinds: Partitioned dependent, Hierarchical dependent, Density dependent, Model dependent and Grid dependent. Density dependent approaches are grounded on an approximation of the density of information. The universal perception of these approaches is that the clusters to construct are designed of a group of points of greatly density surrounded by a lower density values. In this Method, most segregating procedures group entities depending on the distance amongst entities. Such procedures could discover random designed groups. The universal perception is to endure developing the specified cluster as elongated as the density or number of entities or data values in the neighborhood beats certain threshold. These procedures could be employed to filter the noise or outliers. Density dependent clustering approach is one of the crucial procedures for grouping in data mining. The clusters that are designed depending on the density are flexible to understand and do not restrict itself to the outline of clusters.

Almost all of the famous clustering procedures necessitate input parameters that are rigid to define however have a substantial effect on the clustering outcome. Amongst numerous kinds of clustering procedures density dependent approach is further effective in distinguishing the clusters with wide-ranging density. There has been done numerous study on clustering approaches for years however the application to huge spatial data samples presents the subsequent necessities:

- *Minimal number of input factors*: Because for huge spatial data samples it is very difficult to identify the early factors such as count of clusters, shape and density priori.
- *Discovery of groups with random figure*: Since the outline of clusters might be in any arbitrary shape.
- *Good efficiency* ought to be accomplished in very huge data samples.

DBSCAN (Density Based Spatial Clustering of Application with Noise) is a prominent density-dependent clustering approach that can determine the groups with random shapes and does not require to know the number of clusters primarily in its procedure [9, 10]. Though DBSCAN Approach is the most popular clustering approach due to various benefits, it still suffers from certain limitations in the methodology. One of them is, the determination of initial values of Epsilon and minimum number of points are done by the user depending upon experience of the user, which is not always acceptable and these two parameters need to be adjusted accordingly. Several existing survey has been done on this issue on different application as to overcome the problem to certain extent.

On the same lines, this paper also addressed the above issue and proposed a novel efficient DBSCAN clustering algorithm which is mainly concentrating on the image applications and shown that an improved segmentation of an image with size MXN is attained. It is to be strictly noted that this proposed approach can only be applied on images of size MXN. In this approach two different techniques are employed to determine two parameters i.e. the minimum number of points and epsilon values.

1.1 Organization of the paper

In this paper, a brief discussion on data mining, clustering and motivation for the suggested methodology is defined in this section. The section 2 provide a detailed explanation on the existing approaches of diverse DBSCAN based clustering algorithms on different application domain. The existing Density Based Spatial Clustering of Applications with Noise (DBSCAN) technique is briefly explained in section 3. The suggested efficient DBSCAN approach for the image applications is briefly given in section 4. The experimental results and its analysis for the suggested method is given in section 5. The conclusion and referenced of the recommended methodology is specified in section 6 and 7 respectively.

2. LITERATURE SURVEY

There are different types of clustering methods have been developed namely partitioning, hierarchical, density, grid, model, and constraint dependent. Amongst the given, the density grounded method works depending on the idea of density. Some of the recent issues in density based clustering approaches are given in this section. In 2013, the G-DBSCAN algorithm presented a parallel implementation of DBSCAN on GPU. While there are other parallel versions of this algorithm, the G-DBSCAN is distinguished by simplicity of indexing the data through graphs and G-DBSCAN is up to 100 times faster than DBSCAN [1].

The SNN methodology [2], similar to DBSCAN, is a density grounded clustering approach. The foremost alteration amongst this approach and DBSCAN is that it states the likeness amongst data points by observing at the numerous adjacent neighbors that two points can share. By means of this similarity function in the SNN method, the density is determined as the summation of the matches of the adjacent neighbors of a point. Points having higher density is a core points, whereas points having lower density specifies noisy points. All the rest, which are intensely identical to a definite core points will signify a novel clusters.

MR-DBSCAN is an ascendable MapReduce dependent DBSCAN approach for deeply twisted information. It is an efficient ascendable DBSCAN procedure by means of MapReduce is presented in [3] with an aim of load balancing in large-scale datasets and effective speed-up and scale-up for twisted huge data. It works under three level namely data segregating, local clustering, and global merging. In the first level, dataset is divided into smaller partitions based on spatial proximity. During second level, each partition is clustered independently. Then at the final level, the partial clustering results are combined to produce the global clusters.

MDBSCAN is a Modified DBSCAN Using MST depending value for ε in DBSCAN given in [4] an adapted form of DBSCAN approach for clustering datasets using minimum spanning tree (MST) based objective function and to discover natural grouping. A threshold depending on MST of data values of every cluster therefore obtained is employed to eliminate noise from the final clustering. DMDBSCAN is a Dynamic Method for Discovering Density Varied Clusters introduced in [5] have given this new algorithm for the intention of varied density data samples investigation. The main notion is that it uses energetic approach to identify appropriate value of Eps for every density level of the dataset.

PACA-DBSCAN is an Enhancing Clustering Procedure depending on Segregating DBSCAN and Ant Clustering approach depending on partitioning based DBSCAN and Ant clustering is proposed in [6]. It applies one of the two partitioning methods namely PD grounded segregation and PACA segregation depends on dimension of datasets. That is, if the dataset is 2D, it employs PD grounded segregation approach to divide the data, otherwise if dataset is multiple dimension, then it uses PACA method. And for every division, this methodology constructs R*-tree, plots k-dist graph and executes DBSCAN. Finally, the fractional group will be combined depending on pre-specified rules.

DBSCAN-DLP is a Multi-density DBSCAN Algorithm Based on Density Levels Partitioning in [7] to generalize the typical DBSCAN to spontaneously determine clusters of dissimilar densities through the concept of density level partitioning. The essential idea of this algorithm is that partitioning the data samples into diverse density stages sets by examining numerical features of its density disparity, formerly evaluates Eps for every density stage set, and lastly accepts DBSCAN on every density stage set with equivalent Eps to obtain resulting groups.

3. DENSITY BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE (DBSCAN) APPROACH

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a basic density grounded clustering approach suggested in [9] to find out random shaped clusters besides distinguishing noise from large spatial databases [11]. This technique employed two significant input constraints i.e. Epsilon *(ϵ)* and minimal number of point and further employed no. of cluster, unclustered samples, improper samples along with time and noise ratio. DBSCAN generates clusters pertaining to the density aided interconnection technique. This technique could define clusters in huge three-dimensional data samples by observing at the native density of database components, by means of only two input constraints. Moreover, the individual obtains an idea where the constraint value will be appropriate. Consequently, negligible understanding of the area is essential.

The DBSCAN similarly define what data need to be categorized as noise or as outliers. Despite this, its operating procedure is speed and measures better with the dimension of the data set almost linearly. By employing the density dissemination of nodes in the data sample, DBSCAN can classify these nodes into distinct groups that describe dissimilar groups. DBSCAN could discover clusters of random form. Nevertheless, clusters that exists nearby to one another tend to pertain to the similar category. This approach produces areas with adequately higher density into clusters and detects clusters of random shape in spatial data samples with noise. The algorithm employs certain notions for the description of clusters, to well understand how it functions.

➢ Core point: A point p is known as core point if the neighborhood of p outstrips certain specified threshold value like Minpts in the interior to the Epsilon range.
➢ Border point: This has the lesser neighborhood compared to the Minpts in the interior to the Epsilon, nevertheless present in the neighborhood of a core point.

- Noise point (outlier): This is any point which is neither a core point nor a border point.
- Density reachability: An entity q is directly density-reachable from entity p if q is interior to the Epsilon neighborhood of p where p is a core object.
- Density connectivity: Entity p is density-connected to entity q pertaining to Epsilon and Minimum points if there is an entity o in such a way that p and q are density-reachable from o pertaining to Epsilon and Minimum points.

Figure 3 showed core point, border point and outlier for Minimum points=4 and Epsilon = 1unit. DBSCAN determine a cluster C pertaining to Epsilon and Minimum points in a data sample D as a non-empty group in D sustaining the circumstance of maximality and interconnectivity:

- Maximality: if p pertains to C and q is density reachable from p relating to Epsilon and Minimum points formerly q also pertain to C
- Interconnectivity: p and q pertain to C formerly p is density-connected from q relating to Epsilon and Minimum points in DDBSCAN [9] is a significant and extensively employed procedure for class detection in spatial data sets.

The procedure for the DBSCAN Approach is given below:

1. Choose an random point p.
2. Explore whole points that are density reachable from p rendering to Epsilon and Minimum points.
3. If p is a core point, formerly a cluster is generated.
4. If p is a border point, no point is accessible by density from p and DBSCAN move to the succeeding point of the database.
5. Repeat the procedure until whole points are visited.

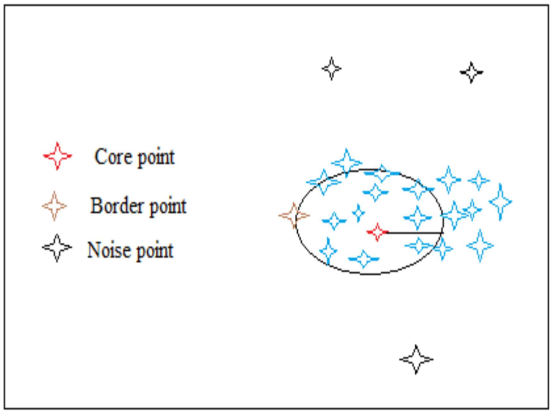

Fig 1: Core Point, Border Point And Outlier For DBSCAN

Benefits of DBSCAN: Maximum clustering approaches employ distance as a measurement amongst two different clusters that fails in identifying haphazard shaped clusters. The DBSCAN could identify random shaped clusters, which is the key characteristic of this approach in recognizing clusters. DBSCAN do not necessitate to know the number of clusters in the data previously, as different from k-means.

- This algorithm could discover random form of clusters and could determine clusters entirely encircled with a diverse cluster. Owing to the Minimum Points constraint, the so called single-related influence (diverse clusters getting inter-related through single line of points) is minimized.
- This technique have the perception of noise.
- This technique needs merely two constraints and is frequently unresponsive to the arrangement of the points in the data set.

Drawbacks of DBSCAN: As the initial density-based clustering approach that identifies clusters with random shape and outliers, DBSCAN has some restrictions that are given as:

- It is not flexible to define accurate preliminary values of Epsilon and Minimum points. Even though, exists in the similar data samples, whenever the number of instances are altered, the two constraints need to be altered consequently.

- The computational complexity of DBSCAN deprived of any distinctive format is O(n), here n is the number of data elements. If any spatial indexing is employed, the complexity could be minimized to O (nlogn). Nevertheless, the job of constructing a spatial indexing is time taking and fewer appropriate to higher dimensional data samples.

4. PROPOSED EFFICIENT DBSCAN CLUSTERING APPROACH

In this section, a novel efficient DBSCAN clustering approach is proposed especially for the image database. The two parameters that is minimum number of points and epsilon value is determined using two different techniques in the proposed approach. Initially consider an RGB Color image of size MXN and convert it into Grey image. The algorithm for the proposed approach is briefly given below and the block diagram for this approach is represented in the Figure 2.

Determination of Minimum Number of Points:

The minimum number of points for the proposed approach purely depends on the size of an image. Consider that the size of the image is MXN, then the minimum number of points are estimated as given:

$$minpts = \frac{M * N}{256}$$

Where M, N are the size of the pixel image, 256 is the grey level value.

Determination of Epsilon (ϵ)

The Epsilon (ϵ) is determined based on the minimum number of points and k-nearest neighbor algorithm. In this methodology, the traditional k-nearest neighbor approach is performed on the pixels of the grey image where the k value depends on the minimum number of points. A graph known as k-dist plot is computed for the obtained k-nearest neighbor distance of the every pixel pixels. The k-dist plot is defined as to determine the behaviors of the distance from a point to its k^{th} nearest neighbor. The k-dist are estimated for entire data points of certain k (minimal number of points), arranged in non-descending order, and formerly plotted by means of the arranged values, as an outcome, a sharp modification is probable to perceive. The sharp modification at the value of k-dist resembles to a appropriate value of Epsilon. For instance consider the Figure 1 with 3 nearest neighbor distances with K-NN approach where the dashed lines given horizontally determine the corresponding Epsilon value ate the sharp change in the plot.

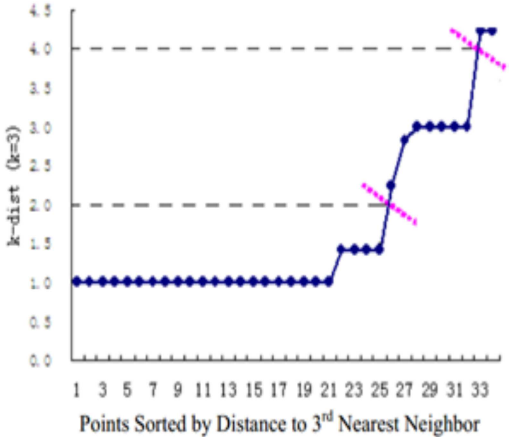

Fig 1: Example Chat For K-Dist Plot

Efficient DBSCAN clustering for image segmentation Algorithm:

1. Initially consider an RGB color image of size MXN and transform into grey scale image.
2. The noise is removed from the converted image using the filtering techniques of image processing.
3. Determine the values of the parameters i.e. minimum number of points and Epsilon value as per the approach evaluated above.
4. Consider these parameters as the initial values and execute the traditional DBSCAN Clustering algorithm on the grey image relating to Epsilon and minimal number of points.

5. The resulted output the segmented image obtained from the clustering technique

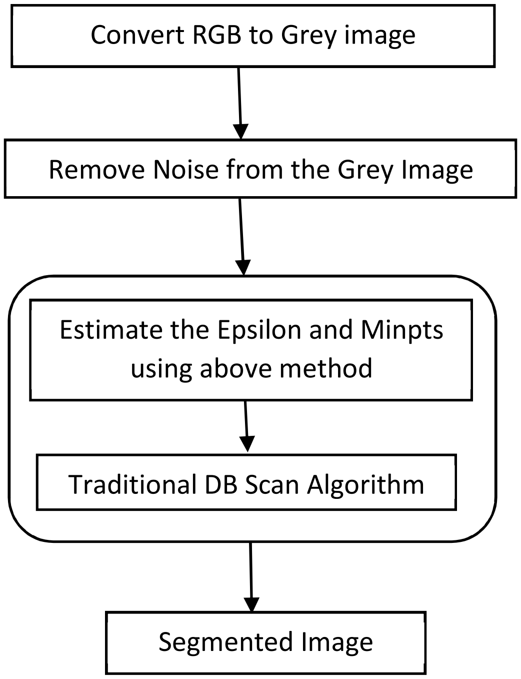
Convert RGB to Grey image
Remove Noise from the Grey Image
Estimate the Epsilon and Minpts using above method
Traditional DB Scan Algorithm
Segmented Image

Fig 2: Block Diagram Of The Efficient DBSCAN For Clustering The Image

5. EXPERIMENTAL RESULTS AND ITS ANLYSIS

In this section a brief analysis of the proposed approach compared to existing approaches that normally employed in the for image segmentation. The experimental result is carried out on two different images. They are an old monument and the ….image. The input are used by adding 5% of salt and pepper noise in the image. The existing approaches that are used for the comparison are the k-means algorithm with 4 clusters, moving k-means algorithm. The Fig.3 shows the original monument image with the segmentation by means of k-means and moving k-means. The Fig.4 shows the original picture image with the segmentation by means of k-means and moving k-means.

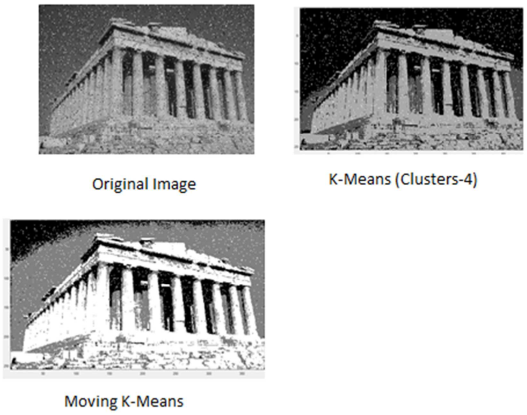

Fig 3: Original Monument Image With K-Means And Moving K-Means Segmentation

The two input images are given to the proposed DBSCAN Approach and the results are shown in the figures below. The analysis for this approach is performed with three different epsilon values obtained from the k-nearest neighboring approach which is employed in the DBSCAN. The k-plot distance of the k-nearest neighboring approach attained three different epsilon values such as 1, 1.5 and 2 respectively for the minimum values 40 depending on the size of the image. The dimension of the image considered in this experiment is 100X100 where from the evaluation method, the obtained minimum point is 40.

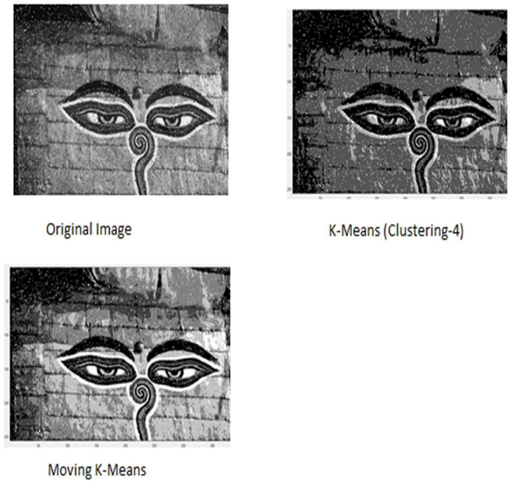

Fig 4: Original Picture Image With K-Means And Moving K-Means Segmentation

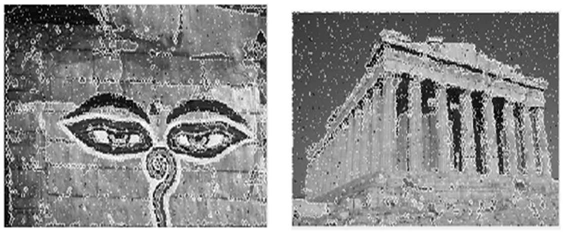

Fig 5: Segmented Image With Epsilon=1 And Minimum Points=40

The Fig. 5, Fig. 6 and Fig. 7 represents the segmented image of the original two input images with epsilon values 1, 1.5 and 2 respectively and with the minimum points as 40. From the outcomes, it is perceived that the noisy is highly reduced when matched with existing segmentation approaches and the images are also improved. As the epsilon values is increasing the segmentation of the image is improving gradually even with small patches in the image. But the only limitation is that the response time for the suggested approach is more compared to the existing systems that is other approaches took 1 minute to execute whereas Proposed DB scan took 1 minute 30 seconds to execute. However, the results of the propose approach showed that the segmentation is performed very well even with the small parts of the image and can be observed from the figures.

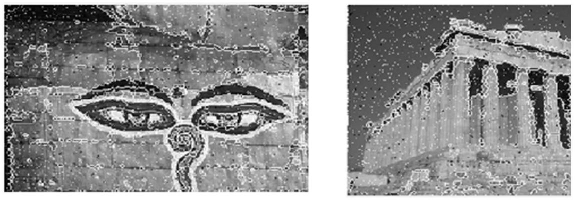

Fig 6: Segmented Image With Epsilon=1.5 And Minimum Points=40

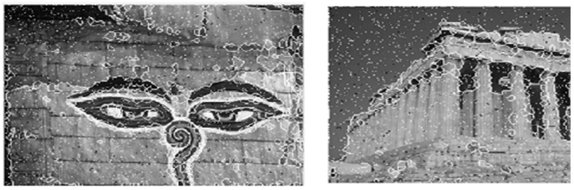

Fig 7: Segmented Image With Epsilon=2 And Minimum Points=40

6. CONCLUSION

The Proposed Efficient DBSCAN clustering approach improved the performance of the image with better segmentation and also reduced the salt and pepper noise compared to the existing segmentation approaches in the image. This DBSCAN approach functions only on the image database where the improvement of this approach is shown by modifying the existing DBSCAN algorithm. In this approach instead to considering random minimum points and epsilon values, these values are asses by the size of the image and evaluating the k-dist plot for the image. The performance of the obtained showed shown an improved compared to the existing segmentation approaches in the image and also observed that the noise is also highly reduced from existing system. As the epsilon value is increasing, better segmented image is obtained even with small patches in the image. This approach can be highly applicable to the old monuments or old pictures where the damage in the image can be segmented better. One of the drawback of this approach is it takes more execution time compared to existing system.

REFERENCES

[1] G. Andrade, G. Ramos, D. Madeira, R. Sachetto, R. Ferreira, and L. Rocha, "G-DBSCAN: A GPU Accelerated Algorithm for Densitybased Clustering," Procedia Comput. Sci., vol. 18, pp. 369–378, Jan. 2013.

[2] Levent Ertoz, Michael Steinback, Vipin Kumar, "Finding Clusters of Different Sizes, Shapes, and Density in Noisy, High Dimensional Data", Second SIAM International Conference on Data Mining, San Francisco, CA, USA, 2003.

[3] Yaobin HE, Haoyu TAN, Wuman LUO, Shengzhong FENG, and Jianping FAN, "MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data," Research Article, Front. Computer Science, vol. 8, no. 1, pp. 83-99, 2014.

[4] Nirmalya Chowdhury and Preetha Bhattacharjee, "Using an MST-based Value for ε in DBSCAN Algorithm for Obtaining Better Result," Int. Journal of Information Technology and Computer Science, vol. 6, pp. 55-60, 2014.

[5] Mohammed T. H. Elbatta and Wesam M. Ashour, "A Dynamic Method for Discovering Density Varied Clusters," Int. Journal of Signal Processing, Image Processing, and Pattern Recognition, vol. 6, no. 1, pp. 123-134, 2013.

[6] G. Chaudhari Chaitali, "Optimizing Clustering Technique based on Partitioning DBSCAN and Ant Clustering Algorithm," Int. Journal of Engineering and Advanced Technology (IJEAT), vol. 2, no. 2, pp. 212-215, 2012.

[7] Zhongyang Xiong, Ruotian Chen, Yufang Zhang, and Xuan Zhang, "Multi-density DBSCAN Algorithm Based on Density Levels Partitioning," Journal of Information & Computational Science (JOICS), pp. 2739-2749, 2012.

[8] Fayyad U., Piatetsky-Shapiro G., and Smyth P. 1996. "Knowledge Discovery and Data Mining: Towards a Unifying Framework". Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, pp. 82-88.

[9] M Ester, H-P. Kriegel. J. Sander, and X, Xu. 1996. "A density-based algorithm for discovering clusters in large spatial databases". KDD'96.

[10] M.Parimala, Daphne Lopaz, N.C. Senthilkumar, "Survey on Density based Clustering Algorithm for mining large spatial databases", IJAST 2011.

[11] L. Duan, L. Xu, F. Guo, J. Lee and B. Yan, "A local-density based spatial clustering algorithm with noise," Information Systems, vol. 32, pp. 978-986, 2007