# Design and Implementation of DBSCAN Algorithm for Generating Clusters in Dense Region of Complex Datasets

Deepak Jain[1], Manoj Singh[2], Dr.Arvind K Sharma[3]
[1]M.Tech, Dept. of CSE, Gurukul College of Engineering & Technology, Kota
[2]Dept. of CSE, Gurukul College of Engineering & Technology, Kota
[3]Dept. of CSE, University of Kota, Rajasthan

***Abstract-*** At present, data clustering is a most popular and frequently used analytical approach of data mining. There are various types of partitions and hierarchal based algorithms applied in clustering and the clusters which formed based on the density are easy to understand and do not limit itself to certain shapes of clusters. DBSCAN is a traditional data clustering algorithm and widely used in network security and data mining. It has best ability to detect clusters with different shapes and sizes. This algorithm clusters the dataset based on two received parameters from the user one is min-points and another is radius. This paper presents complete implementation of DBSCAN algorithm by using MATLAB.

***Keywords –*** *Data Mining, DBSCAN, MATLAB*

## I. INTRODUCTION

In the present scenario large amount of data to be uploaded on different websites and required to be classified.Data mining is the process to extract useful information from databases by using various techniques that shows the different type of mining and various data recovery operation. Several kinds of approaches like classification, prediction, time series analysis, clustering, association, summarization techniques are used to perform various operations to generate results. These techniques can be implemented frequently on existing hardware and software platform to enhance the value and performance of existing information system. Data mining algorithm also have a lot of potential to continually mine large amount of data. Data mining contributes important benefits to the various sectors like IT industries sector, banking sector, blood bank sector, etc. It can be a fundamental tool to analyses the data through the system. Basically it is a process of the knowledge discovery in databases(KDD) which helps to convert raw data into knowledge and informative data.

### 1.1 Data Mining Techniques

*Refer Fig.1*

### 1.2 Data Mining Issues

Data mining applications have been evolving ever since due to the wide applications involved. There are various issues however, which are identified in the data mining systems. There are some issues which are enlisted below.

### *A. Security and Social Issues*

Any type of data which is collected to be shared or can be used for strategic decision-making requirements a proper secure measure in all possible ways. Along with the collection of data for customer profiling, the correlation of personal data, user behavior understanding, etc. there is further sensitive and private data related to organizations or individuals which is collected and stored. The confidential nature of the data is sometimes revealed that causes problems within the system. There might be some activities performed by the data mining, which would disclose the implicit knowledge of groups or individuals. This might cause the violation of privacy policies of system especially if the information is very crucial. The proper manner in which the data mining method is to be used is another major concern here. There might be some information here, which can be controlled and not shared with the other system and the other unimportant information which is not much of hiding, can be shared openly.

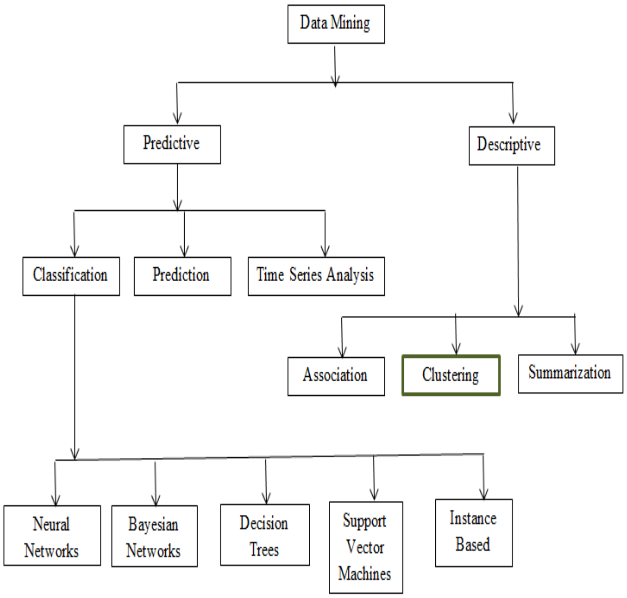


Fig.1. Classification of Data Mining Techniques

### *B. User Interface Issues*

When the knowledge discovered by data mining tools is interesting and understandable, the user finds it to be beneficial. The data mining results are to be better understood with the help of good data visualization. This also helps the user to understand better the applications.

### *C. Mining Methodology Issues*

The data mining approaches are applied and pertained which are related to these issues. The data mining methodology choices could be interrupted through dimensionality domain, assessment of knowledge discovered, the exploitation of background knowledge and metadata, the control and handling of noise in data etc. [1].

### *D. Performance Issues*

For the purpose of data analysis and interpretation, various artificial intelligence and statistical methods are present which are however, not available for huge datasets. The common size of datasets currently is terabytes. The scalability and efficiency are the major issues of data mining techniques of datasets currently which are being raised here when the processing of large data is considered.

### *E. Data Source Issues*

There are various issues which are concerned with the data sources. Some of them are involved with practical applications such as diversity of data types. There are philosophical issues such as the data glut issue. There is a lot of data present in the system which cannot be handled. It is gradually increasing at high rate at each instant. There is more data harvesting to be required where there is an increase in the collection of information through the database management systems. The gathering and processing of data at the instant or to process it later is the current issue. There are some issues which are to be taken care of them. The collection of right information in exact amount, the knowledge of what tasks to be performed using it, and the identification of important or useless data are some of them.

## II. DBSCAN ALGORITHM

There are many types of clustering techniques used in data mining which involve partitioning, hierarchical, density, grid, model and constraint based clustering. On the density based parameters, the density based clustering algorithm is applicable. The regions which are different from a thin region are formed as thick regions or areas. Until the density in the neighbors rises above certain threshold, the identified cluster is increased here. DBSCAN stands for (Density Based Spatial Clustering of Applications with Noise). This algorithm identifies the arbitrary shaped clusters which also includes separating the noise from large spatial databases. There are two parameters which are accepted by DBSCAN algorithm i.e. Eps (radius) and minPts (minimum points-a threshold). The numbers of points within a specific radius Eps are counted for the purpose of estimating the density at a specific point of the data set. This is known as a center-based approach which is applied here. There are many points which are classified in this approach in the categories such as core point, border point and noise points. The important step here is to provide minimum number of points (minPts) for the neighborhood of a given radius (Eps) of each point of a cluster. In other words, the density of the neighbor can be more than the predefined threshold value. There are three input parameters of this algorithm as under:

- K, the neighbor list size.
- Eps, the radius that delimitate the neighborhood area of a point (Eps neighborhood).
- minPts, the minimum number of points that must exist in the Eps-neighborhood.

There are many parameters on which the clustering process depends. The classification of points in the dataset as core points, border points and noise points as well as the usage of density relations between the points are involved [2].

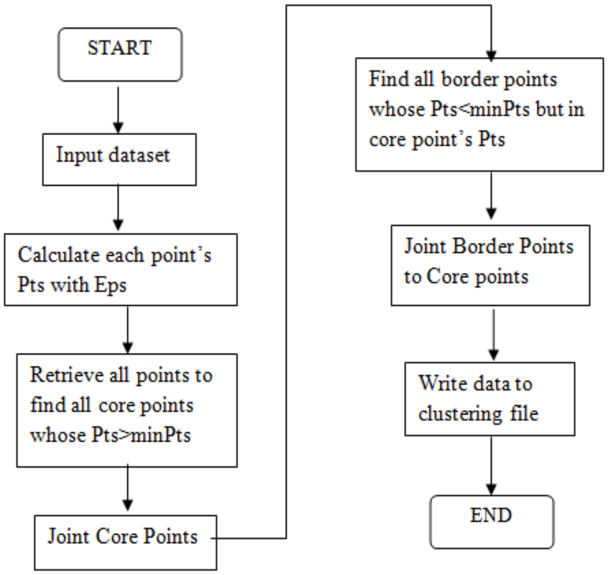

START
Input dataset
Calculate each point's Pts with Eps
Retrieve all points to find all core points whose Pts>minPts
Joint Core Points
Find all border points whose Pts<minPts but in core point's Pts
Joint Border Points to Core points
Write data to clustering file
END

Fig.2. Flowchart of DBSCAN Algorithm

### *COMPLEXITY OF DBSCAN*

Every point of a database is visited numerous times by the DBSCAN algorithm (i.e. Candidates to various clusters). The time complexity of the algorithm is calculated by using the number of region Query invocations as per the practical algorithms are involved. For each point, one query is executed by the DBSCAN. A neighborhood query is executed as O(log n) when the indexing structure is utilized. An overall average runtime complexity of O(nlogn) is obtained. The worst case run time complexity is $O(n^2)$ that occurs when the accelerating index structure is not utilized or when the degenerated data is involved [3].

### *2.1 Advantages*

a. The number of clusters needs not to be specified in the data apriori in the case of DBSCAN algorithm which is different from that of K-means algorithm.
b. The arbitrarily shaped clusters are identified by the DBSCAN algorithm. A cluster which is completely surrounded by other cluster is also identified here. The single-link effect is reduced with the help of minPts parameter.
c. The notion of noise is involved in the DBSCAN algorithm. It is also robust to outliers present in the data set.
d. There are only two parameters required in the DBSCAN. The ordering of points within a database is insensitive in nature.
e. The databases which could accelerate the region queries are designed with the help of DBSCAN algorithm.
f. If the data is well understood, the domain expert can be used for setting the minPts and Eps parameters [4].

### *2.2 Disadvantages*

a. The DBSCAN algorithm is not deterministic completely. On the basis of the order in which the data is processed, the border points which are near to the numerable clusters are involved as a part of any one cluster. However, not all the situations involve this type of issue to arise. Here, the DBSCAN is deterministic. The border points are considered as noise and the completely deterministic result is achieved through this method.
b. The distance measure which is used in the regionQuery(P,ε) function is utilized for calculating the quality of DBSCAN. Euclidean distance is the most common distance metric involved.
c. As there are huge difference is densities, the DBSCAN is not able to cluster data. The combination of minPts-ε cannot be selected in an appropriate manner for all the clusters present.
**d.** The selection of a meaningful distance threshold ε becomes tough when the data and scale are not understood properly [5].

## III. PROPOSED WORK

The density based technique is the type of algorithm in which density of the whole dataset is calculated and most dense region is calculated to find similarity between the elements of the dataset. In the existing work, technique of density based clustering is applied in which density of whole dataset is calculated and dense region is calculated. On the dense region EPS value is calculated to analyze similarity between the elements. The Euclidian distance is applied to analyze similarity between the elements. The EPS is calculated in the dynamic order to achieve maximum accuracy. The Euclidian distance is calculated in the static manner due to which accuracy is not achieved at the maximum point. In this proposed work, enhancement in DBSCAN algorithm has been proposed that calculate Euclidian distance in the iterative manner to increase accuracy of clustering in data mining process.

### *3.1 Objectives of Proposed Work*

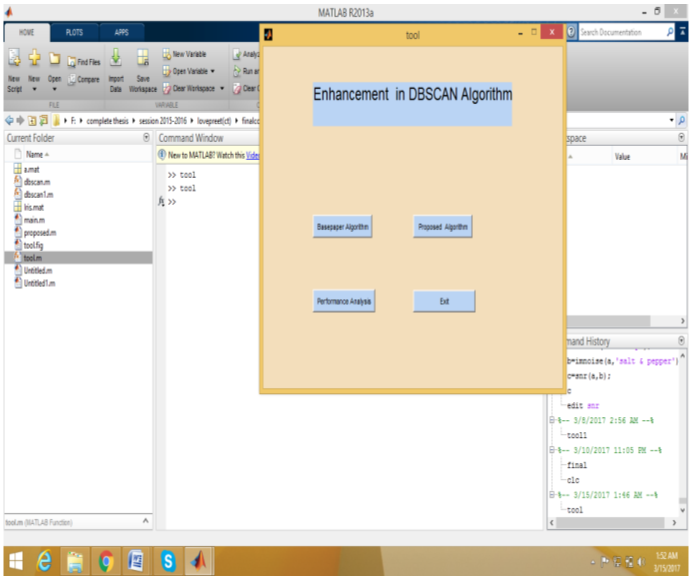
a. To study and analyze various densities based clustering algorithm for data clustering.
b. To propose enhancement in DBSCAN algorithm to increase accuracy of clustering
c. The proposed enhancement is based on back propagation algorithm to calculate Euclidian distance in the dynamic manner
d. To implement proposed and existing techniques and compare results in terms of accuracy and execution time.

### *3.2 Proposed Methodology*

In DBSCAN algorithm the most dense region is calculated from the datasets. The central point is calculated from the most dense region which is the called EPS value of the datasets. To calculate similarity between the data points of the data Euclidian distance is calculated from central point to all other points. In the base paper, to improve accuracy of clustering EPS values is calculated in the dynamic manner which leads to the clustering of the points which are remained uncluttered. The basic DBSCAN algorithm is the static algorithm which the EPS value is given by the user. The EPS value is the radius value which can be the covered in the dataset to part the larger cluster. The Euclidian distance is calculated which cluster the similar and dissimilar values from the defined radium. The incremental DBSCAN algorithm is the further enhancement in the DBSCAN algorithm in which the EPS value is not given statically by the user by it the calculated according to the input datasets. To calculate EPS value according to the dataset, the most dense region is calculated from the dataset and arithmetic mean of the dense region is calculated which is the central point of the datasets. The EPS value defines the class of the input dataset. The Euclidian distance is calculated from the EPS point which clusters similar and dissimilar values from the datasets.

## IV. EXPERIMENTAL EVALUATION

The implementation of the proposal is done using MATLAB, version 2012B with 32-bit windows 7 operating system.

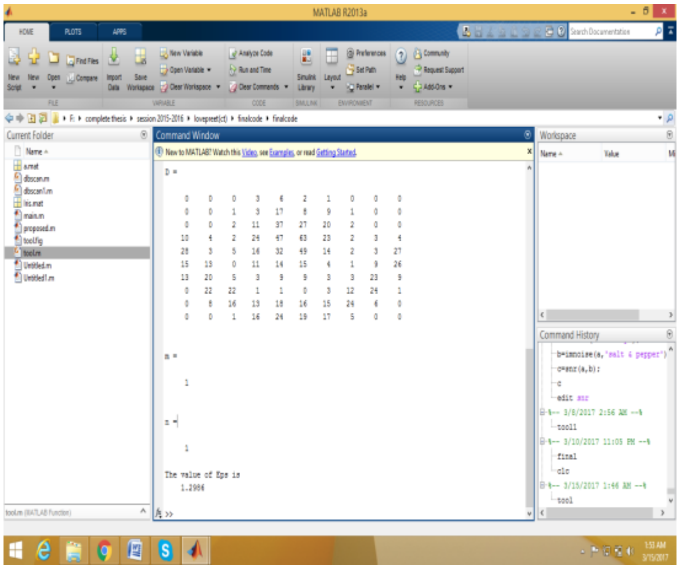


Enhancement in DBSCAN Algorithm

```
>> tool
>> tool
>>
```

Fig.3. Default Interface of Model

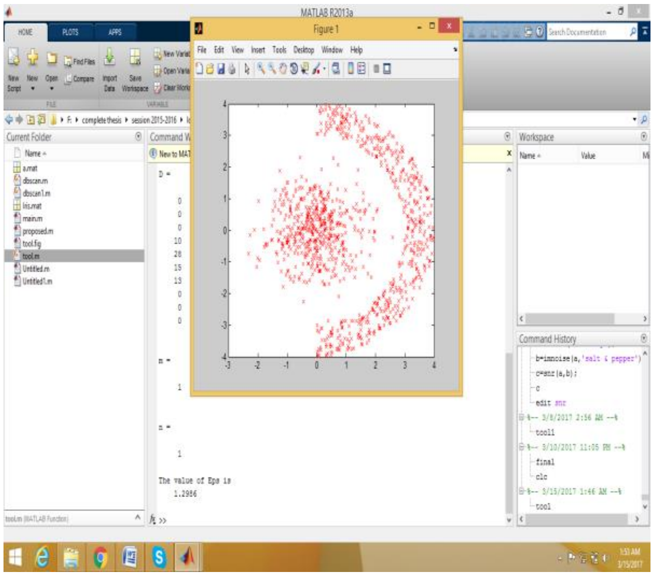
As shown in fig.3, the model is designed in the MATLAB which shows the execution of incremental DBSCAN algorithm and proposed DBSCAN algorithm.



```
D =

     0     0     0     3     6     2     1     0     0     0
     0     0     1     3    17     8     9     1     0     0
     0     0     2    11    37    27    20     2     0     0
    10     4     2    24    47    63    23     2     3     4
    28     3     5    16    32    49    14     2     3    27
    15    13     0    11    14    15     4     1     9    26
    13    20     5     3     9     9     3     3    23     9
     0    22    22     1     1     0     3    12    24     1
     0     8    16    13    18    16    15    24     6     0
     0     0     1    16    24    19    17     5     0     0


m =

     1


n =

     1

The value of Eps is
    1.2986
```

Fig.4. Calculation of Dense Region

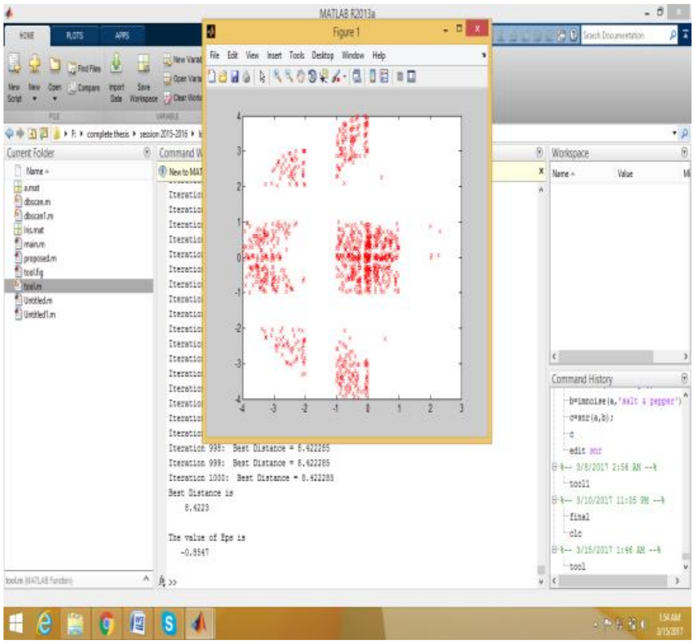
As shown in fig.4, the incremental DBSCAN algorithm has been implemented in which the most dense region has been calculated. The calculated dense region is shown in the snapshot. According to the most dense region the EPS value is calculated which defines radius of the cluster.



```
D =

     0
     0
     0
    10
    28
    15
    13
     0
     0
     0


m =

     1


n =

     1

The value of Eps is
    1.2986
```

Fig.5. Generation of Clusters

As shown in fig.5, the EPS value is defined according to the input dataset. The EPS value defines the Class of the dataset and Euclidian distance from the central point is calculated according to that similar and dissimilar values are clustered.



```
Iteration 998:  Best Distance = 8.422285
Iteration 999:  Best Distance = 8.422285
Iteration 1000:  Best Distance = 8.422285
Best Distance is
    8.4223

The value of Eps is
   -0.9547
```

Fig.6. Generation of Final Clusters

As shown in fig.6, the final clusters are generated according to Euclidian distance value and results shows that generated clusters are different from the existing clusters.

*Comparison Results*

In this section, the performance enhancement of DBSCAN algorithm and Enhancement DBSCAN (E-DBSCAN) algorithm has been compared together in terms of accuracy, time, EPS value and distance.

## V. CONCLUSION AND FUTURE SCOPE

The clustering is the most popular datamining technique in which similar and dissimilar types of data can be clustered together to analyze complex data. The technique of density based clustering is applied which could cluster the similar and dissimilar type of data according to the data density in the input dataset. In the density based clustering the most dense region is calculated from which similar and dissimilar type of data is calculated using similarity techniques. In the DBSCAN algorithm which is applied in this work, the EPS value has been calculated which will be the central of the dataset. The EPS value is calculated dynamically to achieve maximum accuracy. The technique of Euclidian distance is applied to calculate similarity between the data points in the datasets.

## VI. REFERENCES

[1]. Zhe Zhang et al., "*Improved K-means clustering algorithm*", 2008, Congress on Image and Signal Processing CISP, vol. 5, May pp. 169–172.

[2]. D. Widyantoro, T. Ioerger, J. Yen, "*An incremental approach to building a cluster hierarchy*", 2002, ICDM Proceedings IEEE International Conference on DataMining, pp. 705–708.

[3]. S.A.L. Mary, K.R.S. Kumar, "*A density based dynamic data clustering algorithm based on incremental dataset*", 2012, J. Computer Sci. 8 (5) 656–664.

[4]. K.M. Hammouda, M.S. Kamel, "*Incremental document clustering using cluster similarity histograms*", 2003, IEEE/WIC Proceedings International Conference on Web Intelligence, pp. 597–601.

[5]. M. Ester et al., "*Incremental clustering for mining in a data warehousing environment*", 1998, Proceedings of the 24th VLDB Conference, Institute for Computer Science, University of Munich, Germany, New York, USA.