INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 5)
Available online at: www.ijariit.com

Hybrid clustering algorithm using Ad-density-based spatial clustering of applications with noise

Deepak Kumar Sharma
dks7990@gmail.com
Panipat Institute of Engineering and Technology, Panipat, Haryana

Sachin Dhawan
ersachind.ece@piet.co.in
Panipat Institute of Engineering and Technology, Panipat, Haryana

ABSTRACT

With a pace of time clustering algorithms are stunningly utilized in spatial databases for classification. Now a day's databases are of different class and variable in length, therefore, some basic essential parameters are needed for a clustering algorithm for example efficiency must be a pinnacle for a larger database, finding of clusters with an unpredictable shape for larger database and must have the knowledge of input parameters so that cluster can be formed. Now a day's data clustering is so much vital data mining innovation which plays supreme par in diverse scientific exercise. But real problem is that with the time size of data set increasing exponentially and to process huge data set one of the tedious tasks. Clustering technology in data mining is well defined whose main focus is to provide orthodox, manifest shape with help of different data set which is collected for the desired goal. An adroit clustering technique must be systematic and able to recognize clusters of erratic shapes. In our research work, there is a comparative analysis executed between DBSCAN and Ad-DBSCAN where six data sets considered of different quantity of 3-D nature. In base paper 1-D data set considered and after analysis between the base and proposed technique, in proposed technique accuracy is up to 99.99% and process for forming a cluster is very less as compared to base technique. There are different DBSCAN algorithms which can be implemented to execute diverse function so that the formation of the cluster would be dynamic. In this research, article summarization shows that formation of cluster executed fastly and with accuracy almost 100%.

Keywords— *Cluster, Data mining, EPS, Radius, Ad-DBSCAN, 3-D, Database*

1. INTRODUCTION

A cluster is a collection of data points that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data points can be treated collectively as one group and so may be considered as a form of data compression. Clustering is also called data segmentation in some applications because of clustering partitions large data sets into groups according to their similarity.

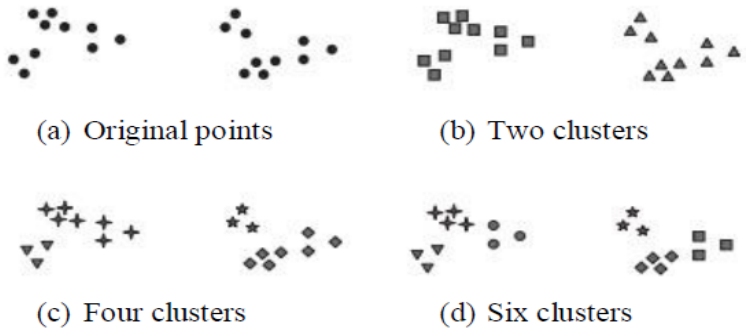

Fig. 1: Different ways of dividing points into clusters

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge which can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration [3]. Density-Based Spatial Clustering of Applications with Noise DBSCAN is a typical density-based clustering algorithm. DBSCAN can discover clusters of arbitrary shape. But it is sensitive to the input parameters, especially when the density of data is non-uniform [4]. The DBSCAN clustering algorithms usually can be classified into the following different categories:

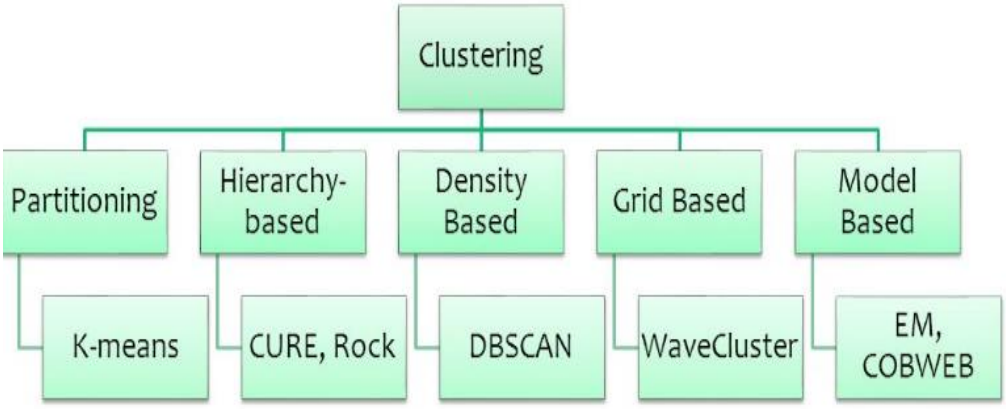

Fig. 2: Clustering methods

Density-based algorithms like DBSCAN and OPTICS find the core objects at first and they grow the clusters based on these cores and search for objects that are in a neighborhood within a radius of a given object. The advantage of these types of algorithms is that they can detect arbitrary forms of clusters and they can filter out the noise [7].

Grid-based algorithms quantize the object space into a finite number of cells (hyper-rectangles) and then perform the required operations on the quantized space. The advantage of this approach is the fast processing time that is in general independent of the number of data points. The popular grid-based algorithms are STING, Wave Cluster, and ClIQUE.

Model-based algorithms find good approximations of model parameters that best fit the data. They can be either partition or hierarchical, depending on the structure or model they hypothesize about the data set and the way they refine this model to identify partitioning. They are closer to density-based algorithms, in that they grow particular clusters so that the preconceived model is improved [9]. However, they sometimes start with a fixed number of clusters and they do not use the same concept of density. The most popular model-based clustering methods are EM.

Fuzzy algorithms suppose that no hard clusters exist on the set of objects, but one object can be assigned to more than one cluster. The best known fuzzy clustering algorithm is FCM.

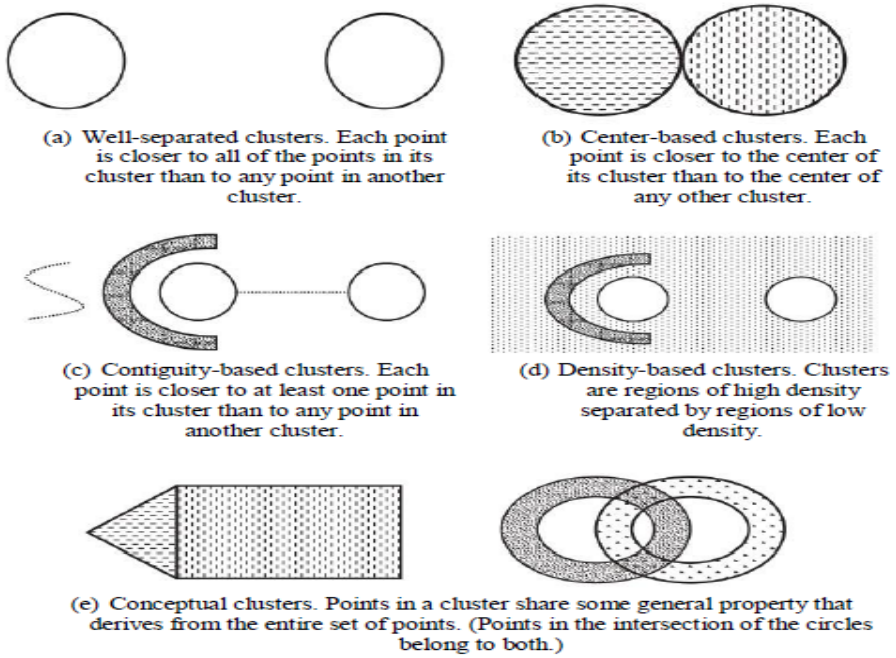

Fig. 3: Different types of clusters

2. TYPES OF CLUSTER

(1) **Well-Separated:** In this type of cluster, each object is closer (or more similar) to every other object in the cluster than to any object not in the cluster Figure 3 (a).

(2) **Prototype-Based:** It is a cluster in which each object is closer to the prototype that defines the cluster than the prototype of any other cluster. We commonly refer to prototype-based clusters as center-based clusters Figure 3 (b).

(3) **Graph-Based:** A cluster can be defined as a connected component; i.e., a group of objects that are connected to one another, but that have no connection to objects outside the group. An important example of graph-based clusters is contiguity-based clusters, where two objects are connected only if they are within a specified distance of each other. This implies that each object in a continuity-based cluster is closer to some other object in the cluster than to any point in a different cluster Figure 3 (c)

(4) **Density-Based:** A cluster is a dense region of objects that is surrounded by a region of low density. Objects in these sparse areas- that are required to separate clusters - are usually considered to be noise and border points Figure 3 (d).

3. LITERATURE SURVEY

Data mining is the process of identifying hidden and interesting patterns from the large data set, which can further be used in decision making and future prediction. Clustering is an important technique of class identification in spatial databases. The objective of the clustering is to maximize the intracluster similarity and minimizing the inter-cluster similarity. Clustering is used to find useful patterns in unlabeled data. Clustering has been extensively studied for over 40 years and across many disciplines due to its broad applications.

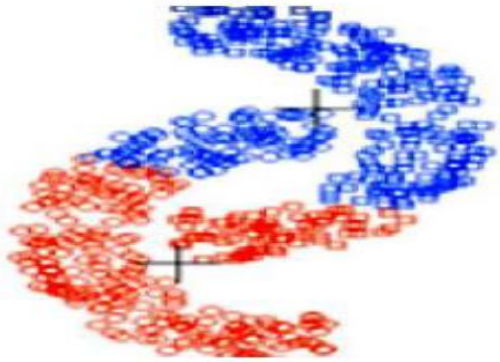

Fig. 4: Clustering with k-means and PAM algorithms

K-means is a widely used clustering algorithm in the field of data mining across different disciplines in the past fifty years. However, *k*-means heavily depends on the position of initial centers, and the chosen starting centers randomly may lead to poor quality of clustering. Motivated by this, this paper proposes an optimized *k*-means clustering method along with three optimization principles named $k*$-means. First, we propose a hierarchical optimization principle initialized by $k*$ cluster centers ($k* > k$) to reduce the risk of randomly seeds selection, and then utilize proposed top-*n* method to merge the nearest clusters associated with the shortest *n* edges in each round until the number of clusters reaches at k [1]. Different clustering algorithms highlight the characteristics of big data. A brief overview of various clustering algorithms which are grouped under partitioning, hierarchical, density, grid-based and model-based are discussed [2]. In data mining, Clustering is the most popular, powerful and commonly used unsupervised learning technique. It is a way of locating similar data objects into clusters based on some similarity. Clustering algorithms can be categorized into seven groups, namely Hierarchical clustering algorithm, a Density-based clustering algorithm, Partitioning clustering algorithm, Graph-based algorithm, Grid-based algorithm, a Model-based clustering algorithm, and Combinational clustering algorithm. Clustering is a technique in which a given data set is divided into groups called clusters in such a manner that the data points that are similar lie together in one cluster. Clustering plays an important role in the field of data mining due to a large number of data sets. This paper reviews the various clustering algorithms available for data mining and provides a comparative analysis of the various clustering algorithms like DBSCAN, CLARA, CURE, CLARANS, K-Means etc. [6].

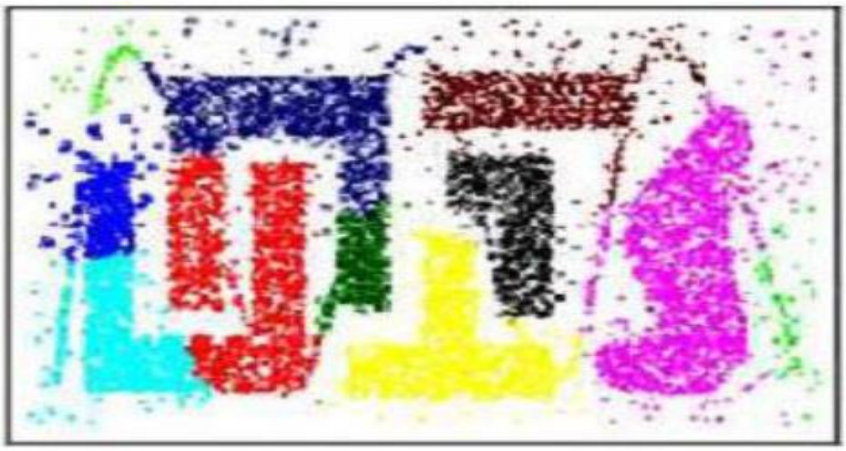

Fig. 5: Clustering an artificial dataset with CURE algorithm

4. PLANNING OF WORK/METHODOLOGY

DBSCAN, a well-known density-based clustering algorithm and the main feature of this algorithm are that each point of a cluster must contain at least a certain number of points (Min Pts) within its ϵ- neighborhood. In other words, the density in the ϵ-neighbourhood of a point belonging to a cluster has to be greater or equal to a predefined threshold. The clustering process in DBSCAN is based on the following concepts of relations between points: directly density reaches ability and density reachability. Moreover, DBSCAN discerns three types of points: core points, border points, and noise points. These concepts and types of points are defined in the next section. A cluster in the context of the DBSCAN algorithm is a region of high density. Regions of low density constitute noise. A point in space is considered a member of a cluster if there are a sufficient number of points within a given distance from it. Definitions and notions related to the DBSCAN algorithm are given below.

Core point

p is a core point with respect to ε if its ε-neighbourhood contains at least *MinPts*; that is if $|\varepsilon NN(p)| \geq MinPts$.

Directly density reachable points

Point p is directly density-reachable from point q with respect to ε *and MinPts* if the following two conditions are satisfied:

(a) $p \in \varepsilon NN(q)$

(b) q is a core point.

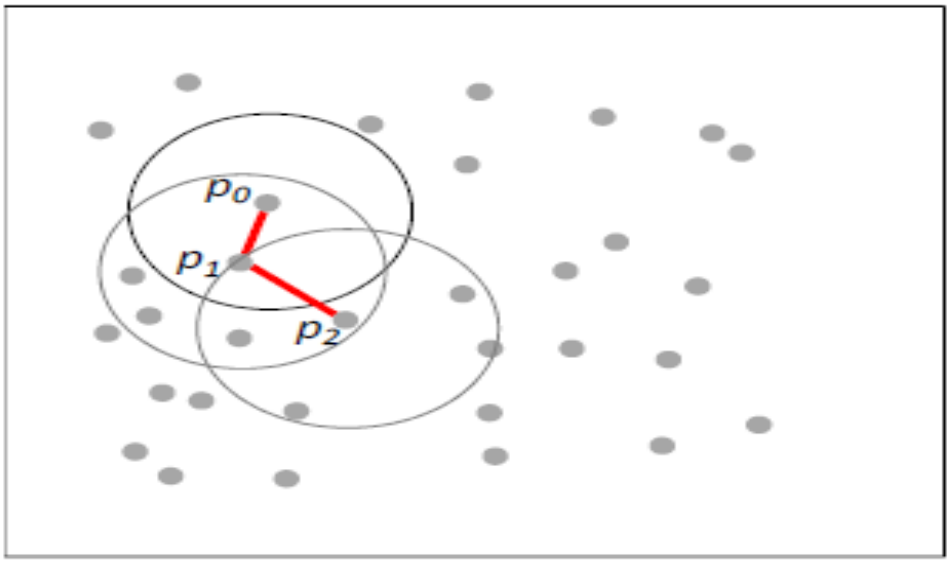

Fig. 6: Po is directly density-reachable from core point Pi; P0 is density-reachable from

Ad-DBSCAN is a more representative density-based clustering algorithm. Unlike the partitioning and stratified clustering methods, it defines the cluster as a density phase. The largest set of connected points can divide a region with a sufficiently high density into clusters and can be in a spatial database of noise. In our research work, six different 3-D data sets considered for clustering purpose of different length for example 210, 266 and 1000 and many more data set can be considered. There are two important parameters which we have to focus that is accuracy and processing time. In our approach, FDBSCAN and K mean clustering algorithm is utilized. Different data sets are considered for experimental work and our main emphasis on the radius of the cluster, processing time, number of cluster formation which depend on the values of data set and its length. If there are a number of mean value came into existence then the number of the cluster is formed. FDBSCAN algorithm is very fast when a comparative analysis made with respect to base paper and also accuracy lies in the range 99.6 % to 100%. The key idea of density-based c1ustering is that for each object of a c1uster the neighborhood of a given radius (Eps) has to contain at 1east a minimum number of objects (MinPts), i.e. the cardina1ity of the neighborhood have to exceed some thresho1d. We wi1l first give a short introduction to DBSCAN inc1uding the definitions which are required for incremental c1ustering.

```
Algorithm DBSCAN (D, Eps, MinPts)
// Precondition: All objects in D are unclassified.
  FORALL objects o in D DO:
    IF o is unclassified
      call function expand_cluster to construct a cluster wrt.
       Eps and MinPts containing o.

FUNCTION expand_cluster (o, D, Eps, MinPts):
  retrieve the Eps-neighborhood N_Eps(o) of o;
  IF | N_Eps(o) | < MinPts    // i.e. o is not a core object
     mark o as noise and RETURN;
  ELSE // i.e. o is a core object
     select a new cluster-id and mark all objects in N_Eps(o)
     with this current cluster-id;
     push all objects from N_Eps(o)\{o} onto the stack seeds;
     WHILE NOT seeds.empty() DO
        currentObject := seeds.top();
        retrieve the Eps-neighborhood N_Eps(currentObject)
        of currentObject;
        IF | N_Eps(currentObject) | ≥ MinPts
           select all objects in N_Eps(currentObject) not yet
           classified or are marked as noise,
           push the unclassified objects onto seeds
            and mark all of these objects with current
            cluster-id;
        seeds.pop();
  RETURN
```

Fig. 7: Ad- DBSCAN algorithm

5. SOFTWARE USED AND SIMULATION RESULT

5.1 Software

Proposed scheme have developed in MATLAB 2015a tool. It is powerful software that provides an environment for numerical computation as well as a graphical display of outputs. In Matlab, the data input is in the ASCII format as well as binary format. It is a high-performance language for technical computing integrates computation, visualization, and programming in a simple way where problems and solutions are expressed in familiar mathematical notation. In our dissertation work six 3-D data set of variable length are used to know the actual performance of the implemented algorithm for various parameters. Density-based Ad-DBSCAN technique used to measure accuracy, processing time, the radius of the cluster, no of the cluster formed.

Table 1: Specification of system hardware

Processor	Intel ® Core ™ i5
RAM	4GB
Operating system	Window 10 Pro, 64 Bit
Software	MATLAB 2016a

5.2 Experimental Result for Data Set

Data Set 1 Result **Data Set 2 Result**

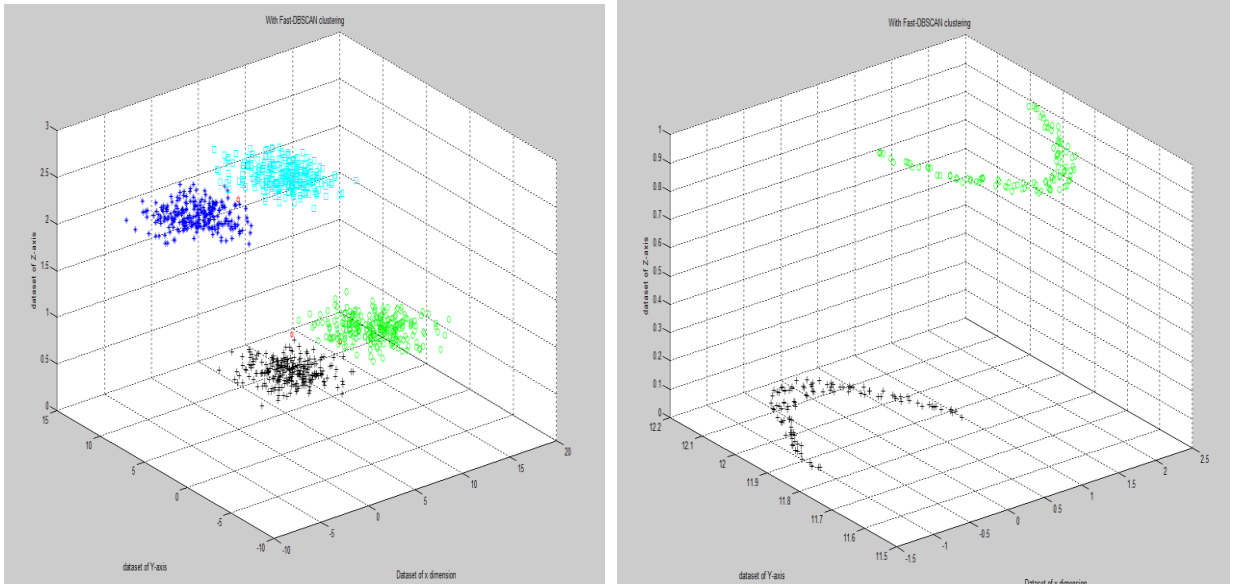

Fig. 6: Experimental result for Ad-DBSCAN for data set 1 and 2 forming 4 clusters and 2 respectively

Length of data set 1=1000
No of data fall into cluster 1=249
No of data fall into cluster 2=249
No of data fall into cluster 3=250
No of data fall into cluster 4=249
Accuracy=997/1000= 99.7

Length of data set 2=210
No of data fall into cluster 1=105
No of data fall into cluster 2=105
Accuracy=210/210= 100

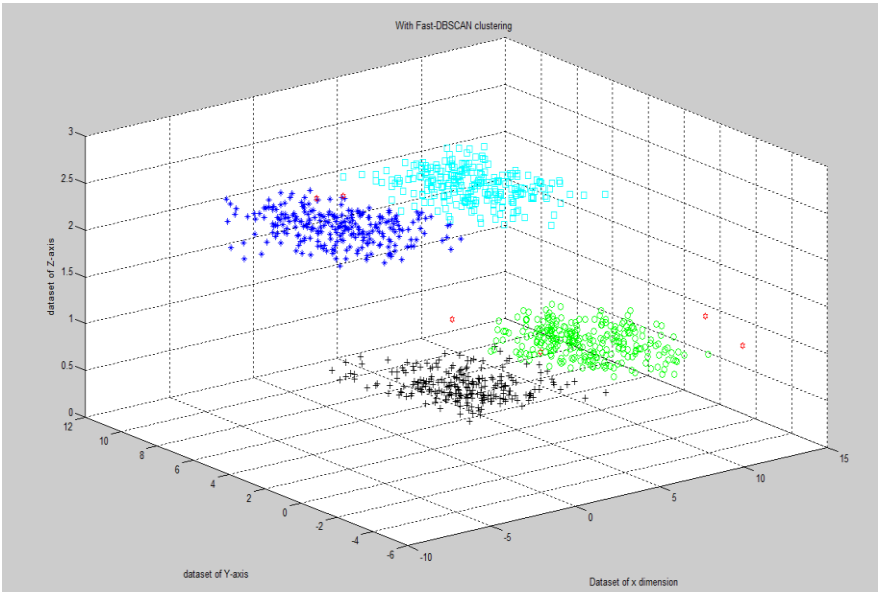

Fig. 7: Experimental Result for Ad-DBSCAN for data set 3 forming 4 clusters

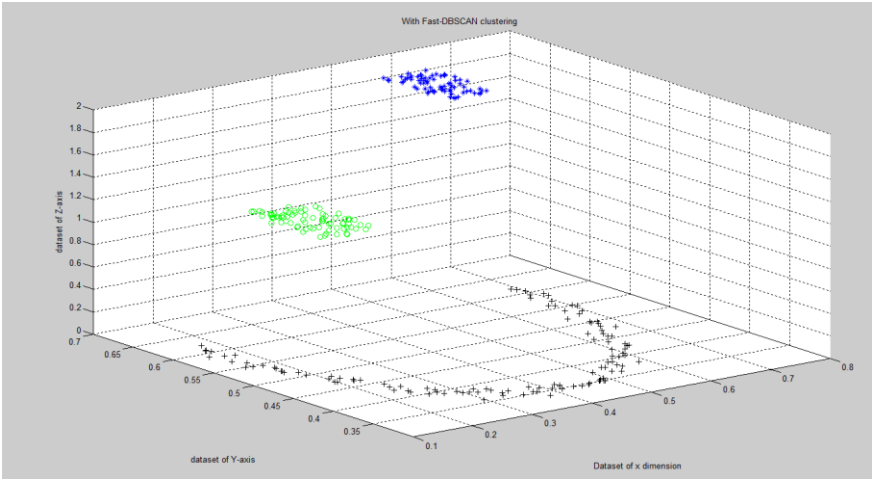

Fig. 8: Experimental Result for Ad-DBSCAN for data set 4 forming 3 clusters

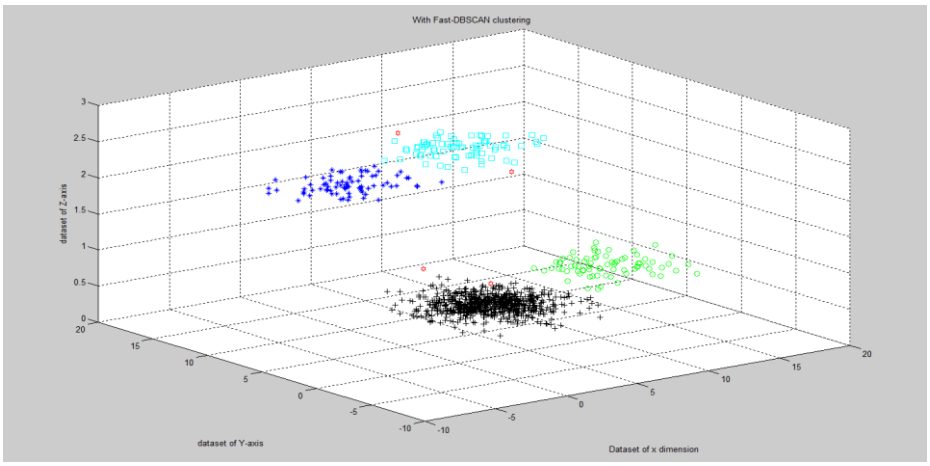

Fig. 9: Experimental Result for Ad-DBSCAN for data set 5 forming 4 clusters

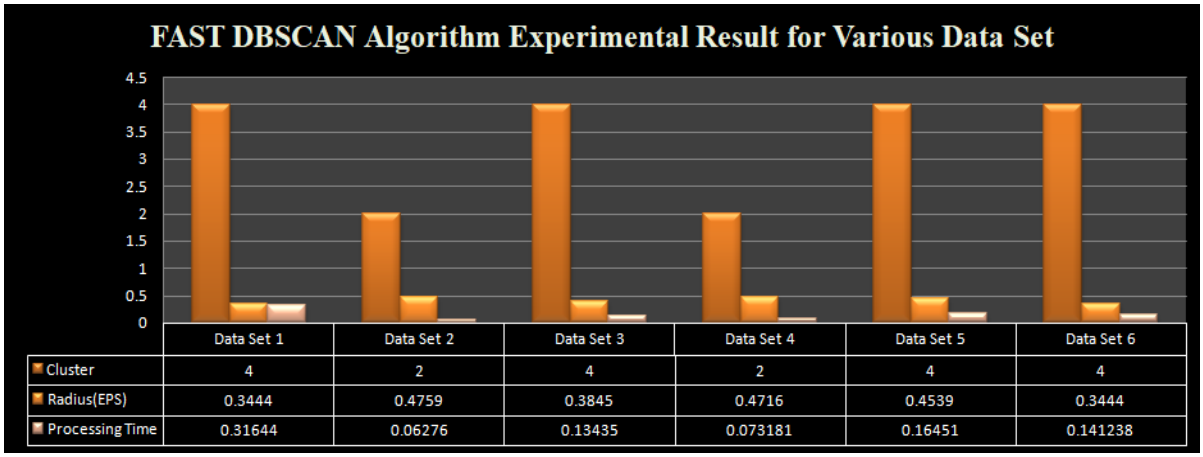

Fig. 10: Graphical representation of research work for 3 parameters

Table 2: Parameters values for different data set

Parameters	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Data Set 5	Data Set 6
Cluster	4	2	4	2	4	4
Radius(EPS)	0.3444	0.4759	0.3845	0.4716	0.4539	0.3444
Processing Time	0.31644	0.06276	0.13435	0.073181	0.16451	0.141238

6. CONCLUSION

With the pace of time huge development going on, in the area of clustering with the latest technologies so that better result could be achieved. An advanced technique to boost up DBSCAN has been proposed which is known as Ad-DBSCAN and this technique are based upon K-means clustering algorithm. In this research paper, the different data set is used during the investigation of various parameters of variable length and length varies from 200 to 1000. There are two crucial parameters which must be improved that is accuracy and processing time as compared to reference work. How many numbers of the cluster will be formed depends upon data set values? Clustering algorithm helps out to extract noise (unwanted data) from data because noises having a high frequency as compared to data or information. According to our experimental result which is executed in MATLAB

2015a, a diverse number of the cluster formed depending upon data set values of variable length. In our research work, 3-D data set values used whereas in reference work 1-D data used. In our research both the parameters, accuracy and processing time (PT) to form a cluster are achieved in exponential form as compared to base paper. In our research work accuracy is almost 100% and processing time is also fast. Further, this technique can be implemented with IoT, artificial intelligence (AI) and suitable languages for these techniques are SCALA and SPARK which are capable of handling huge databases having fast response

7. REFERENCES

[1] Jianpeng Qi, Yanwei Yu*, Lihong Wang, and Jinglei Liu, "K*-Means: An Effective and Efficient K-means Clustering Algorithm", 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking

[2] T. Sajana, C. M. Sheela Rani and K. V. Narayana, " A Survey on Clustering Techniques for Big Data Mining", Indian Journal of Science and Technology, Vol 9(3), DOI: 10.17485/ijst/2016/v9i3/75971, January 2016

[3] KM Archana Patel and Prateek Thakral, "The Best Clustering Algorithms in Data Mining", International Conference on Communication and Signal Processing, April 6-8, 2016, India

[4] Ajin V W, Lekshmy D Kumar, "Big Data and Clustering Algorithms", International Conference on Research Advances in Integrated Navigation Systems (RAINS - 2016), April 06-07, 2016, R. L. Jalappa Institute of Technology, Doddaballapur, Bangalore, India

[5] Dr. Anu Saini, Jagrit Minocha, Jaypriya Ubriani, and Dhruv Sharma, " New Approach for Clustering of Big Data: DisK-Means", International Conference on Computing, Communication, and Automation (ICCCA2016)

[6] Garima, Hina Gulati and P.K Singh, "Clustering techniques in data mining: A comparison", 2nd International Conference on Computing for Sustainable Global Development (INDIA Com), March 2015

[7] Mythili S, Madhiya E, "An Analysis of Clustering Algorithms in Data Mining", IJCSMC, Vol. 3, Issue. 1, January 2014, pg.334 – 340

[8] Chintan Shah1 and Anjali Jivani, "Comparison of Data Mining Clustering algorithms", 2013 Nirma University International Conference on Engineering (NUiCONE)

[9] Hassan A Kingravi M Emre Celebi and Patricio A Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm, Expert Systems with Applications, 40(1):200–210, 2013.

[10] Huang Darong, Wang Peng "Grid-based DBSCAN Algorithm with Referential Parameters" 2012 International Conference on Applied Physics and Industrial Engineering

[11] Naresh Kumar Nagwani and Ashok Bhansali, "An Object-Oriented Email Clustering Model Using Weighted Similarities between Emails Attributes", International Journal of Research and Reviews in Computer science (IJRRCS), Vol. 1, No. 2, June 2010

[12] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Michael Wimmer, Xiaowei Xu, "Incremental clustering for mining in a data warehousing", University of Munich Oettingenstr. 67, D-80538 München, Germany

[13] **S**auravjyoti Sarmah, Dhruba K. Bhattacharyya, "An Effective Technique for Clustering Incremental Gene Expression data", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 3, May 2010.

[14] Xiaoke Su, Yang Lan, Renxia Wan, and Yuming "A Fast Incremental Clustering Algorithm", International Symposium on Information Processing (ISIP'09), Huangshan, P.R.China, August-21-23, pp:175-178,2009

[15] Zhipeng Cai, Randy Goebel, Mohammad R Salavatipour, Yi Shi, Lizhe Xu, and Guohui Lin. Selecting genes with dissimilar discrimination strength for sample class prediction. In APBC, pages 81–90, 2007.

