*International Journal of Engineering Science Invention (IJESI)*
*ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726*
*www.ijesi.org ||Volume 8 Issue 06 Series. II || June 2019 || PP 49-55*

# Reduction of DBSCAN Time Complexity for Data Mining Using Parallel Computing Techniques.

Anupam Kumari, Prof. Vishal Shrivastava, Dr. Akhil Pandey
*M. Tech.Scholar Dept.of CSE Arya college of Engg.& I.T. mishra.*
*.Tech co-ordinator Dept.of CSE Arya college of Engg.& I.T.*
*HOD Dept.of CSE Arya college of Engg.& I.T.*
*Corresponding Author: Anupam Kumari*

***Abstract****: Data mining has become the buzz word of computational research due to its enormous -economical & social significance. Clustering is subset of data mining operations & is a type of learning using observation techniques. Clustering uses on unsupervised learning model & does not require any training data to generate the clustering model. Clustering enables grouping of similar & dissimilar type of data in separate groups. DBSCAN is on emerging & power full clustering algorithm & has gained wide spread popularity in recent times. This work is aimed at reducing the time complexity of DBSCAN algorithm by parallel computing technology. Parallel computing techniques have been implemented using multiple cores with & without code information.*

***Keywords****-Data mining, Clustering, DBSCAN, Parallel Computing*

---

Date of Submission: 22-06-2019 Date of acceptance: 08-07-2019

---

## I. Introduction

In which age We are is often speak of to the information age. In this age, as we believe that information points to power and achievement, and thanks to sophisticated technologies such as computers, satellites, etc., we have been collecting incredible amounts of information. Primarily, with the introduction of computers and means for mass digital storage, we had started gathering and putting away all sorts of data, including on the power of computers to help sort through this combination of information. Inappropriately, these huge collections of data stored on unrelated structures very quickly became irresistible. This early disorder has led to the creation of structured databases and database management systems (DBMS). We have been collecting a

numberless data, from simple numerical amounts and text documents, to more difficult information such as spatial data, multimedia channels, and hypertext documents. Data Mining is well-known as Knowledge Discovery in Databases (KDD. It is nontrivial concept of formerly unknown and possibly useful figures from data in databases. Even though data mining and knowledge discovery in databases (or KDD) both are synonyms, data mining is basically part of the knowledge discovery procedure.

The technique of finding similar groups of data in a set is called clustering. In this technique, Objects in every group are relatively more comparable to objects of that group than those of the other groups. In this paper, I will be taking through the different clustering types, different algorithms and evaluation between two of the most commonly used clustering methods.

Generally speaking, clustering can be divided into two subcategories:-

1.**Hard Clustering**
2. **Soft Clustering**

Now I will go through two of the most standard clustering systems– K Means clustering and Hierarchical clustering.

- K means algorithm is an iterative clustering algorithm that aims to find local maxima in each iteration.
- Hierarchical clustering is those who build hierarchy of clusters. It starts with all the data points allocated to their own cluster. After that two nearest clusters are merged into the one cluster. At last, this algorithm let go when there is only one cluster left.

Density based algorithm has played a vigorous role in finding nonlinear shapes structure based on the density. Density-Based Latitudinal Clustering applications with Noise (DBSCAN) is most widely used density based algorithm. Here we used the conceptionof reachability and connectivity of density.

➢ **DBSCAN Algorithm**

In this algorithm, here first I have to explain some points like -

▪ Epsilon neighbourhood: where all points sets within a distance.

▪ Base point : A point which has least number of 'min Point's neighbour..

▪ Direct Density Reachable (DDR) : The Point q is straight density reachable from a point p if p is base point and q $\in N\varepsilon$ .

▪ Density Reachable (DR): Two points which are DR if there is a sequence of DDR points connected with these two points.

▪ Boundary Point: Point that are Direct Density Reachable but not a base point.

▪ Noise: Points that do not belong to any neighbour point's.

.Parallel Computing Toolbox™ using this software I solved computationally and data-intensive problems with the help of multicore processors, and computer clusters. High-level paradigms—parallel for-loops, special array types, and parallelized numerical algorithms—after that I parallelize MATLAB® without CUDA or MPI programming. we can use the toolbox with Simulink® to run multiple simulations of a model parallely.

## II. Literature Review

This paper presents a comprehensive study of DBSCAN algorithm and the enhanced version of DBSCAN algorithm with its implementation using mat lab. It is concluded that the density based cluster is that the economical kind of cluster throughout that clusters unit printed on the density of the input dataset. [1]

This work is based on the density based clustering which is applied to calculate similarity from the densest region which can define clusters on the basis of similar and dissimilar type of data. [2]

The clustering is one of the most popular data mining algorithms in which the similar and dissimilar type of data could be clustered together to analyse complex data. Moreover, the algorithm 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. [3]

In this review work the most commonly used k-means clustering methods of data mining is analyzed. The work shows that there are more than a few methods to improve the clustering with different methodologies. Various clustering techniques are go through which improve the existing algorithm with different perspective. [4]

In this paper ,a new technique is presented , in that the data is separated into clusters based on the density that determined by applying some limitation .In this process ,the data has been separated perfectly , then new sampling technique will applied in order to reduce the density of dense data and obtain data with only one density distribution (sparse data), the results of sampling were very effective . [5]

Here we use mat lab to implementing thesesuggested clustering and outlier detection system and tested with these data unnaturally created by Gaussian distribution function. These data will be in the form of circular or spherical clusters. Our Future works may report some issues involved in applying the algorithm in a particular application area.[6]

These algorithms is been applied within the context of the LOCAL project [Local2005] as the part of a task, its aim to create models of the geographic space (Space Models) and to be used in context-aware mobile systems. Here, the character of the clustering algorithms is to identify clusters of Points of Interest (POIs) and then use the clusters to automatically characterize geographic regions. [7]

In this paper, we have suggested exploiting user log information (or user document clicks) as a supplement. A new clustering principle is proposed: if two queries give rise to the same document clicks, they are similar. [8]

This algorithm can dynamically discover nested clusters, which is very useful for highly irregular data distributions. Experimental results presented in this paper also showed the efficiency and effectiveness of the proposed algorithm compared to the grid-based methods using single uniform meshes. [9]

First of all we present data-sensitive analysis of the algorithm's running time, which shows that the algorithm runs faster as the segregation between clusters increases. After that a number of experiential studies on unnaturally created data and on real data sets from applications in color quantization, data compression, and image segmentation.[10]

## III. Methodology

This work Is performed on the three core of the system:-

1.**Single Core Setup (Basic DBSCAN)**

Data clustering algorithm is commonly used in data mining and machine learning. based on a set of points , DBSCAN groups together points that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions

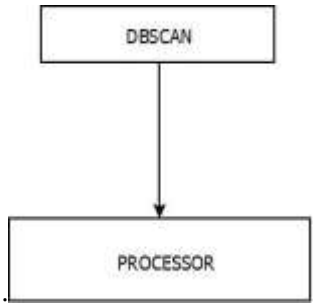


**Fig:3.1** Single Core Setup (Basic DBSCAN)

**2 Multicore Setup (Basic DBSCAN)**

In multi core setup, DBSCAN work on two or more core of processer in this diagram we can see two core of processer in used in DBSCAN. And here we can see automatic job sequencer which is used for split load between two core of processer automatically when we run matlab code here not require any command for split the load between two processer.

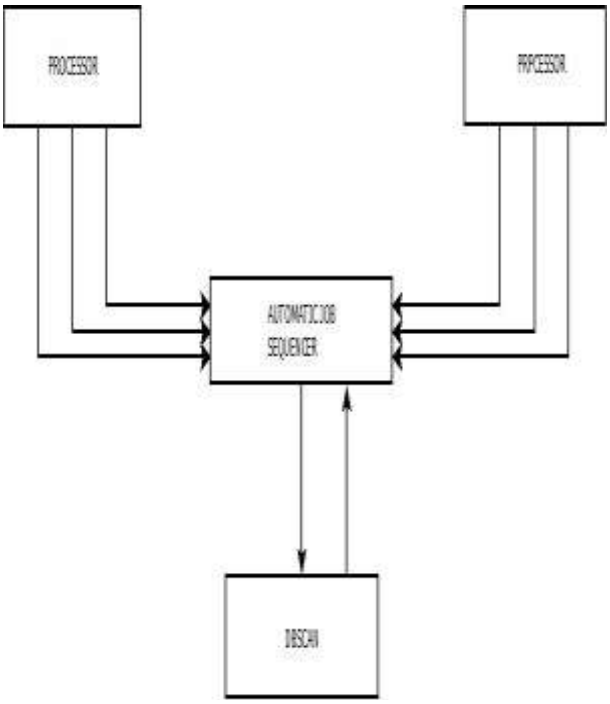


**Fig:3.2 Multicore Setup (Basic DBSCAN)**

**3 Multicore Setup- Modified DBSCAN**

In this type of DBSCAN operation which is multicore setup-modified DBSCAN we use multi core processer but here job sequencer method is change which is user specified. It mean user can be specified work load on the core.

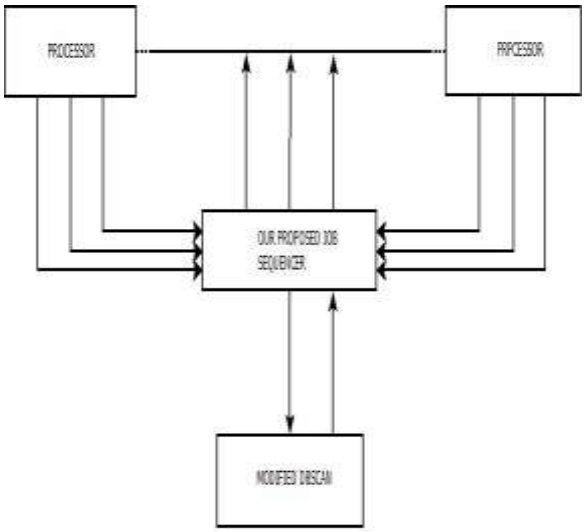


**Fig:3.3 Multicore Setup- Modified DBSCAN**

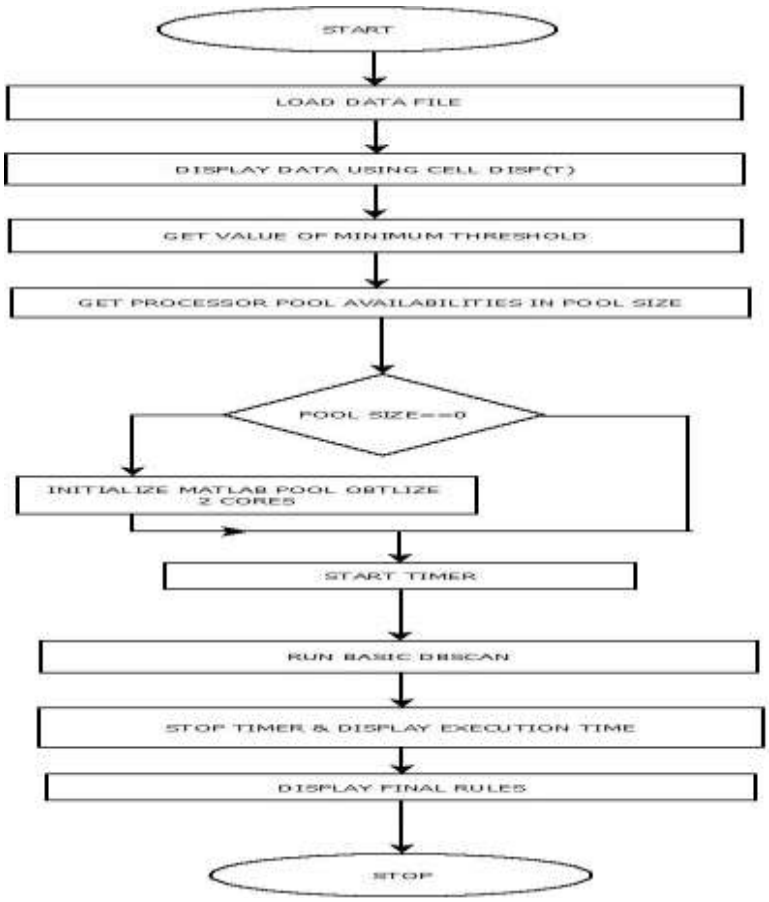


**Fig.3.4-flowchart for Multicore Simple DBSCAN**

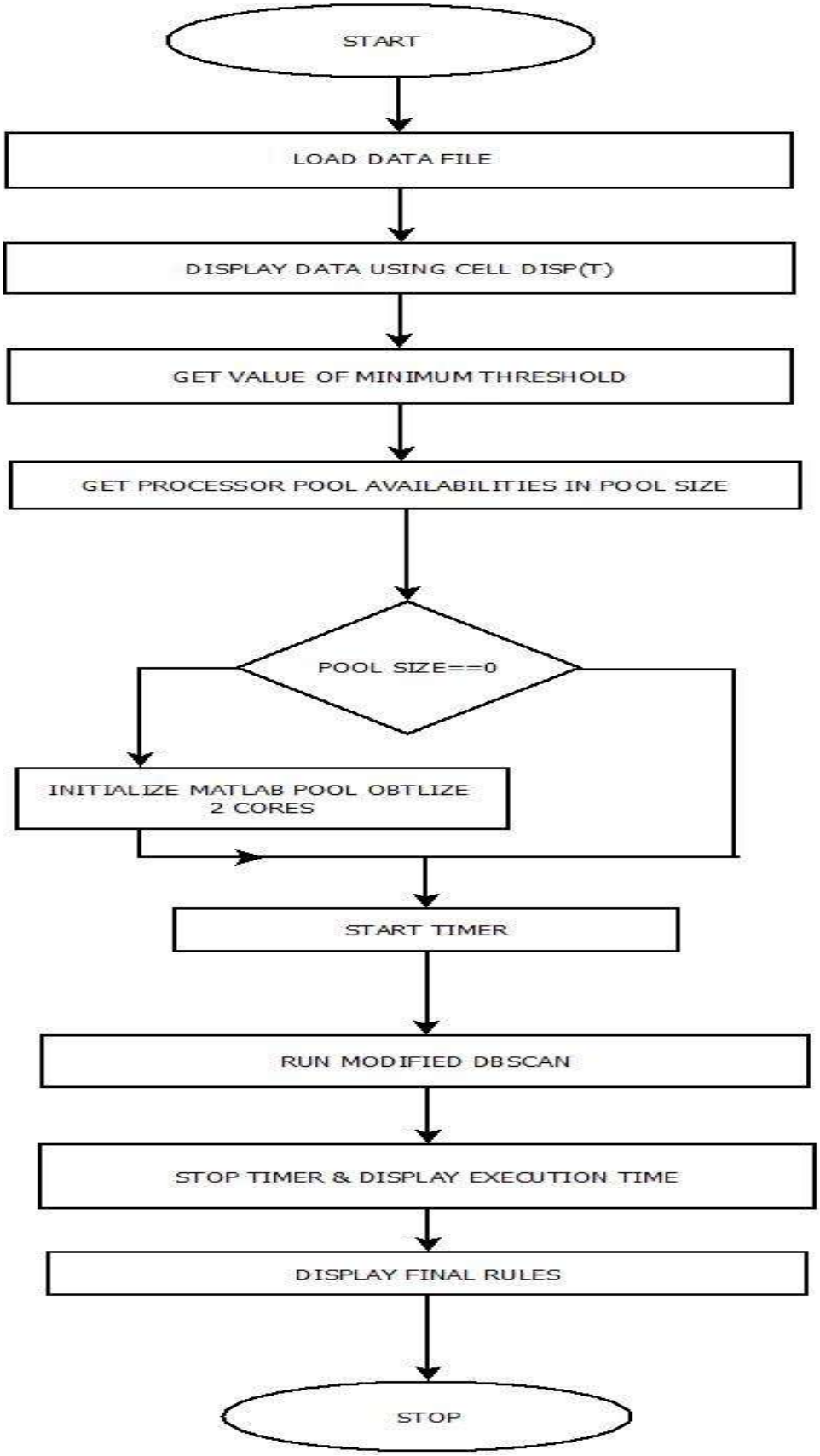


**Fig 3.5-Flowchart for Multi Core Modified DBSCAN**

## IV. Result

In this window we can see the DBSCAN clustering on epsilon 0.5 and minimum points 10 and we can see noise that is denoted by circle sign and cluster#1 and cluster#2 that is denoted by different colour cross sign.

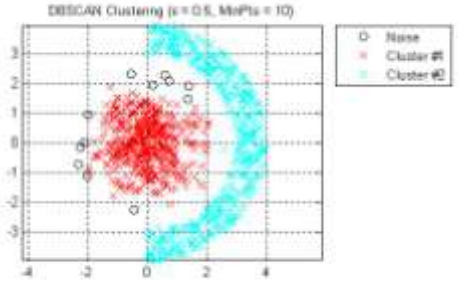


**Fig 4.1- Cluster in single core**

In this window we can see figure of DBSCAN clustering at epsilon 0.5 and minimum points is 10. And we can also see noise and cluster#1 andcluster#2

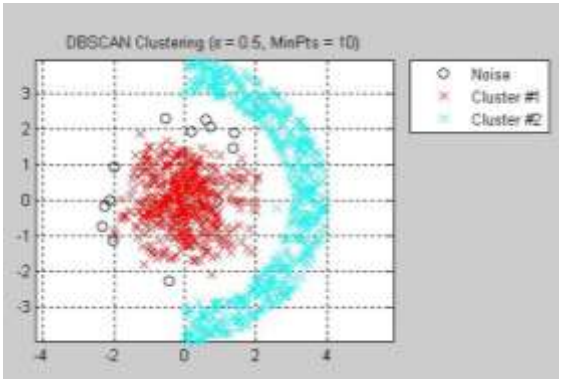


**Fig 4.2- Cluster using parallel processing without code intervention**

In this window we can see the figure of DBSCAN clustering at the value of epsillon is 0.5 and minimum points is 10. And we can also see noise , cluster#1 and cluster#2 that is denoted by different sign.

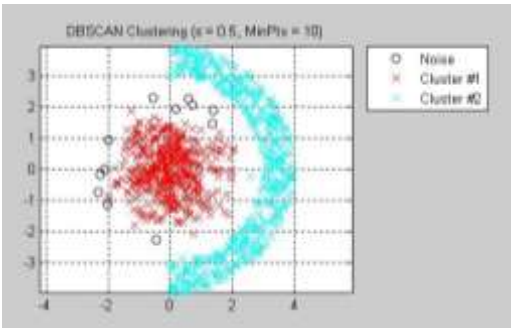


**Fig 4.3- Cluster using parallel processing with code intervention**

## V. Conclusion

This work has aimed at reduction of time complexity of DBSCAN clustering algorithm. The some is achieved by employing MATLAB parallel computing toolbox, & various execution methods of DBSCAN clustering have been tested on a reference dataset vizsingle core execution (without parallel computing) multi core execution without code intervention that when processers cores are employed but no intervention for tasks sharing is provided/amended in code & multi core execution with code intervention that when multi processors

we employed & these are Specific instruction.As observed by the results above the proposed method of employing multiple cores with code intervention yields the best result with lowest execution time.

## VI. Future Scope

A demonstrated by the results above a updated & advanced DBSCAN algorithm is presented. That improves the system performance multiple times, especially for huge datasets, bus as data mining and parallel computing are rapidly advanced techniques, the proposed system must aim to meet the future demands & trends,

## References

[1]. Deepak Jain, Manoj Singh, Dr.Arvind K Sharma "Performance Enhancement of DBSCAN Density based Clustering Algorithm in Data Mining" ICECDS-2017
[2]. Mandeep Kaur1, Rupinderpal Singh2 "Efficient Incremental Density-Based Algorithm for Clustering Large Datasets" IJESE 2017.
[3]. 1Deepak Jain, 2Manoj Singh, 3Dr. Arvind K Sharma "Comparative Study of Density Based Clustering Algorithms for Data Mining" IJCST 2017.
[4]. 1*AnshulYadav, 2Sakshi Dhingra "A REVIEW ON K-MEANS CLUSTERING TECHNIQUE " International Journal of Latest Research in Science and Technology 2016.
[5]. Safaa O. Al-mamory, IsraaSalehKamil "Enhancing of DBSCAN based on Sampling and Densitybased Separation" NSDS 2015.
[6]. K. Mumtaz1 and Dr. K. Duraiswamy 2, "A Novel Density based improved k-means Clustering Algorithm – Dbkmeans" IJCSE 2010.
[7]. Adriano Moreira, Maribel Y. Santos and Sofia Carneiro "Density-based clustering algorithms – DBSCAN and SNN" DBSCAN & SNN 2005.
[8]. Ji-Rong Wen, Jian-yunnie, Hong-jingzhang "Clustering User Queries of a Search Engine" IJSC 2005.
[9]. Wei-kengLiao , Ying Liu , Alok Choudhary "A Grid-based Clustering Algorithm using Adaptive Mesh Refinement" Mining Scientific and Engineering Datasets 2004.
[10]. Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, SeniorMember, IEEE "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" IEEE 2002.

Anupam Kumari" Reduction of DBSCAN Time Complexity for Data Mining Using Parallel Computing Techniques." International Journal of Engineering Science Invention (IJESI), Vol. 08, No. 06, 2019, PP 49-55