Survey of Density Based Clustering Algorithms and its Variants

Pradeep Singh
Department of Computer Science & Engineering
National Institute of Technology, Raipur, India
psingh.cs@nitrr.ac.in

Prateek A. Meshram
Department of Information Technology
National Institute of Technology, Raipur, India
prateekmeshram100@gmail.com

***Abstract*— Clustering technique is a unsupervised machine learning technique in the domain of data mining. Many of the clustering techniques are inherently sensitive to the input parameters. Different clustering techniques works differently for different types of the input datasets. Among all different varieties of clustering techniques, DBSCAN is one of the most important clustering technique whose working principle based on the density estimation while forming the clusters of the input dataset points which is basically used for spatial datasets of random shapes and sizes. It also eliminates the noise during the clustering formation process with a worst case run-time complexity of O(n^2). DBSCAN technique also produces a bad result for varied density datasets. In this paper we have discussed about different density based clustering techniques along with DBSCAN, its variants and some of its modified algorithms with respect to their input parameters and running time complexities. Also we have presented the comparison analysis of all the different variants of DBSCAN algorithms over different benchmark datasets for computing various measures.**

Keywords—DBSCAN, Fast-DBSCAN

I. INTRODUCTION

Clustering mechanism divides the input datasets in to number of groups of subclasses or clusters in such a way that the objects those are in one cluster are very similar to each other and dissimilar to the objects present in another cluster. The main emerging applications of clustering techniques are in the field of medical and image analysis [1] , pattern recognition[2], knowledge discovery[3] which is being posed to find clusters with random shapes and sizes, by determining the input parameters of algorithms with a minimum domain knowledge. Different clustering algorithms follow different approaches to the same input dataset. One of the methods of clustering approach is the partitional clustering which uses the K-means[4] algorithm to discover clusters of spherical shapes and requires the number of clusters to be formed as an initial input parameter. Kernel K-means another clustering technique forms random clusters by transforming in to kernel functions which is having a time complexity of $O(n^2)$ and for this reason it is not efficient for large data sets. Hierarchical clustering is another clustering approach which partitions the data sets into hierarchical structures discovered by combining the subsets using minimum distance criteria at various stages[5].The hierarchical clustering method has a running time complexity of $O(n^2)$.To derive the cluster this approach follows a stopping criterion on splitting and merging of partitions.BIRCH is another hierarchical clustering approach which uses a tree like structure representation in order to decrease the running time complexity to find the clusters for spherical datasets. In this BIRCH algorithm the order of input dataset can affect the whole clustering output result[6]. Currently, semi-supervised[7] and multi-view based[8][11] methods have shown effective improvement in the accuracy of clustering.

Clustering algorithms splits the input dataset into a specific number of clusters C. In most of the research papers cluster is being defined by considering into the internal consistency and the external disassociation [9],[10] that is the paradigms in the same clusters must be identical to each other and that in the different clusters must not. The mathematical expressions for the representation of few clustering algorithms based on the explanations described in[9].

Given a set of parameters $X = \{X_1, X_2, X_3, \dots, X_i, \dots X_n\}$

where, $X_j = \{x_{i1}, x_{i2}, \dots \dots . . x_{id}\}^T \in R^d$ where x_{ij} is called as an attribute.

(Hard)Partitional clustering strives to find n-partitions of X, $C = \{ C_1, C_2, \dots \dots . . C_n\} \,\forall\, n \leq k)$ such that

1. $C_i \neq \Phi, where\ i = 1,2, \dots, n$
2. $\cup_{i=1}^{n} C_i = X$;
3. $C_i \cap C_j = \emptyset, i, j = 1, \dots ., n\ and\ i \neq j$

Density based spatial clustering applications with noise DBSCAN[11] is the initial algorithms which discovers clusters random shapes and handles noise outliers efficiently.

In this paper we have shown various density based clustering algorithms including DBSCAN and its variants along with some of the modified density based approaches with their runtime complexities.

II. LITERATURE WORK

DBSCAN [12] Density Based Spatial Clustering Algorithm with Noise uses density based approach to form spherical shaped clusters in order to cluster the input datasets of varied density. E. Bicici [13] proposed a new algorithm LSDBC which impulse local scaling by guessing the threshold based on the locally available data in the density clustering Here K-nearest neighbors are used for finding potential centers of the clusters discovering local maxima. The value of the cluster is increased until its density falls below the specified density parameters, resulting in the clustering of noise and random shaped data sets on density gradients.

STDBSCAN [14] proposed by Birant et. Al is an another algorithm, having the same runtime complexity as that of DBSCAN which uses two input parameters $Eps1$ and $Eps2$,one is the spatial dimension and the other is the temporal dimension which is being used to handle the varied density points by defining the density factors[14]. Minpts with a density threshold value is used to give the outputs for the datasets with two different dimensions.

Borah and Bhattacharya [15] proposed DDSC a new technique with a time complexity of $O(n * logn)$ having radius Eps, Minpts and the density threshold value K as the input parameters for the algorithm which mainly focuses on partitioning the datasets such that the adjoining regions significantly get differs in the density by the use of homogeneity testing for detecting the variations in the density.

Enhanced DBSCAN proposed by A.fahim et.al [16] whose complexity was equal to the complexity of the DBSCAN. The algorithm also takes the input parameter as the number of nearest neighbors and a new parameter with the limitation of the greatest density Minpts[17]. The proposed algorithm is based on the concept of local density functions and to finds its value at each points which will be an approximation of overall density function.

Darong and Peng [18] proposed a new technique GRPDBSCAN and this technique was also having the same complexity as DBSCAN the input parameters which were used as the number of grid units N[18]. The main goal of this algorithm is that it combines the grid partition technology with the multi-density based clustering. Hence this technique has also the same limitation that it could also not handle the datasets with varied density.

ISDBSCAN[19] which is used to enhance the density-based clustering , having the same complexity of the DBSCAN where the number of nearest neighbours –K as the input parameter to the algorithm and used the main concept of space stratification based on both the INFLO function and KNN distance[20].

Y.Kim et.al [21] proposed a new algorithm DEBCURE which was basically using the technology to parallelize the density based algorithms by using the map-reduced approach and the algorithms with varying densities was also used to take the outputs from the datasets with varying densities successfully[21]. Specialty of this algorithm was that it was 9.5 times faster than the conventional DBSCAN algorithm.

Andrade and et. al [22] proposed G-DBSCAN a GPU accelerated technique which was also used to design or implement the parallel working of the algorithms which were based on density. This algorithm claimed to be different from the other algorithms which were being implemented at that time by its simplicity in which the datasets are indexed with the graphs allowing the others with various parallel opportunities. According to the authors GPU accelerated algorithm could be 100 x times faster than compared to its currently working sequential versions using GPU.

G.Panda and et.al [23] proposed a new technique Computationally efficient density- based clustering algorithm.This paper was proposed to reduce the computational complexity related to the conventional DBSCAN by implementing the merge functions at the start of the evaluation of the cluster. Though these DBS algorithms are not computationally efficient but still they are effective on giving the outputs. Hence in the proposed algorithm they have mearged the new clustering technique which lead to produce computationally effective outputs.

M.Kue and et.al [11] proposed a technique with a modified density-based scanning algorithm with noise for spatial travel pattern analysis from smart card AFC data. This paper focuses the application of the DBSCAN algorithm by using smart AFC data while proposing an algorithm named weighted-DBSCAN where they have tested the passenger dataset and examined the given dataset with their proposed algorithm. The out coming result have been compared with traditional DBSCAN and proved that the proposed algorithm worked more efficiently than the conventional DBSCAN but this was also failing to implement the high dimensional dataset. This algorithm could run 45% faster than conventional DBSCAN.

J. Zhang and et.al [24] proposed a new technique based on density based spatial clustering named Adaptive Density Based Model for Extracting Surface Returns from Photon-Counting Laser Altimeter Data. They proposed an application of density based algorithm by computing them against the datasets taken from the real time such as The Ice cloud and the land Elevation satellite2 to smooth the rounding edged of the algorithm and has produced efficient output by adaptive based model.

III. DENSITY BASED CLUSTERING ALGORITHM AND ITS VARIOUS TYPES

Density based algorithms play a very important role while performing clustering over various nonlinear datasets. DBSCAN is the most widely used algorithm for the formation of clusters of spherical shapes based on the density approach. In this section various density based algorithms nave been discussed along with their variants.

A. DBSCAN:

A cluster is defined as a region of connected points as a dense region collected from the datasets and those regions are separated by sparse regions. Euclidean distance is used as a similarity measure for the connected points in the dense regions.

- eps-neighborhood: for a point $i \in P$, the Eps-neighborhood symbolizes the set of points whose remoteness from i is less than or equal to Eps. The cardinality of eps-neighborhood defines the threshold density of i
- eps-connected: For a pair of points i and $y \in P$, if $||i - y|| \leq eps$, then i and y are eps-connected

points. In this techniques the regions of data points are mostly classified as core points and border points. Border point can also be a density connected point or a noise point.

- Core points are the points where the threshold density is greater than or equal to the defined Minpts.
- Border points are the points where the threshold density is fewer than the Minpts.
- A point d is termed as a noise point if all points in the eps-neighborhood of d are border points and the threshold density of d is fewer than Minpts.
- A border point with minimum one core point in its eps-neighborhood is the Density-connected point.

Algorithm randomly starts with choosing an unvisited point and discovering its eps-neighborhood. If the number of points in the eps-neighborhood is less than the Minpts then it is identifiable as a noise point or an outlier, otherwise that point will be considered as a dense point and after that a density connected cluster is created. The same procedure is carried out in an iterative manner, if no new points can be found to add in to the cluster then we can conclude that partition with a new cluster as a whole and no points will be further added to the cluster in succeeding iterations. Hence to form a new cluster from the given initial data points the same procedure will be carried out repeatedly until a new cluster is found or modeled. The clustering process stops when all the data points are assigned to some cluster or the points are identified as noise points. Within a cluster every data point must be eps-connected with atleast one data point to which it belongs and not eps-connected with any other points in the remaining clusters. There may be cases in which a border point could be associated to the border points of other clusters.

Some of the important applications of the DBSCAN are:

Satellites images: Data is being received from satellites all over the place in the world which has to be translated into useful information, for example, categorizing areas of satellite-taken images with admiration to forest, mountains and water.

X-ray crystallography: A real-world application that discovers all the atoms in the interior of a crystal, which consequences in a large quantity of data. To discover and categorize the different types and amount of atoms in the data which are taken out by the crystal can use the DBSCAN algorithm.

Anomaly Detection in Temperature Data: Application which emphasizes basically on design anomalies in the data, which is significant in various cases, e.g. credit fraud, health condition etc. Measures alterations in temperatures, which is applicable due to the ecological changes (global warming). It can also determine the errors happening in the equipment. These unusual designs need to be noticed and examined to get control.

B. DENCLUE Algorithm:

DENCLUE[8] describes the concept of influence function and the density function. Influence function is called as the data point which could be modelled as the resulting function and computative function. The impact of data point within its neighborhood is described by the influence function and the sum of influence of all the data points is called as density function. DENCLUE defines two types of clusters i.e. one is the center defined cluster and the other one is the multi-center defined cluster. The density function is defined as the sum of the influence functions of all data points. The DENCLUE is a two-step algorithm, the first step is the pre-clustering step, in this step we make of a specific location of the dataset. To speed up the calculation of the density function it requires the efficient access of the neighboring parts of the relevant data space that is used for mapping. The second step under this algorithm is the actual clustering process. In this step the density-attractors and the corresponding density points are identified.

C. DBCLASD

DBCLASD (Application Based Clustering Algorithms for Mining in Large Spatial Databases)[25] is another clustering algorithm in which a point given to the cluster which is processed incrementally without taking into consideration of the cluster. In DBCLASD a Cluster is defined by the following three properties:

1) NNDistSet(C): set of nearest neighbors of cluster C are called as Expected Distribution sets.

2) Points that come inside the cluster C but do not fulfill the first condition, called as maximal conditions.

3) A pair (a,b) which is always connected by grid cells ,called as connectivity conditions.

D. OPTICS(Generalized Density based Approach)

OpticsXi[26]can also be called as generalized Density Based clustering technique by producing an collection of the points that allows the abstraction of clusters with random values for ε .The generating-distance ε is the major distance reflected for clusters. Clusters can be removed for all ε_isuch that $0 \leq ε\_i \leq ε$. The core-distance is the least distance ε' among p and an object in its ε-neighborhood such that p would be a core piece. The reachability-distance of p is the least distance such that p is density-reachable from a core point O.

E. BRIDGE

BRIDGE [27] is another clustering process in which the whole data set points are divided in to a set of k different clusters by hybrid approach. This approach is used to create density clusters by K-means from the initial observed data points and then the process is followed by one of the density-based clustering mechanism. At the last step the formed K-means clusters are refined by removing the noise in the density based clustering approach. In this hybrid approach K-means technique is used as it is faster than the density-based clustering as it is computationally more cost effective for finding out the neighborhoods in density based clustering.

DBSCAN can be applied on over these clusters to find the dense clusters.

F. CUBIN

The basic key idea behind this algorithm is to find variable size and non-spherical clusters. The runtime complexity of this algorithm is O(n) where n is the size of the dataset. The algorithm consists of the following steps:

Algorithm: DBSCAN

Input:

- D: Dataset containing n objects
- ε: radius parmeter
- $Minpts$: neighborhood density threshold

Output: A set of density-based clusters

Procedure:

mark all objects as unvisited
$uVisited = \{|D|\}$;// $uVisited$ is a list contains all the objects of the dataset;//$|D|$= n(Number of objects)
do {
  Randomly select an unvisited object p i.e. $p \in uVisited$
  Mark p as visited
  $Z \leftarrow \varepsilon - neighborhood(p, \varepsilon, Minpts)$ // Z is the neighborhood list of the visited point p
  if $|Z| \geq Minpts$ //if the neighborhood list contains at least $Minpts$ objects
    $clusterID = \{\emptyset\}$
    $clusterId = clusterId \cup \{p\}$// add p to the new cluster
    for each point p' in $|Z|$ i.e. for all points in the ε-neighborhood of p
      if p' is unvisited
        mark p' is visited
        $W \leftarrow \varepsilon - neighborhood\ (p', \varepsilon, Minpts)$
        if $|W| \geq Minpts$ i.e. if the $\varepsilon - neighborhood$ of p' has atleast $Minpts$ points
          then $Z \leftarrow Z \cup |W|$ // add all points to the $\varepsilon - neighborhood$
      if $p' \notin$ any $clusterId$
        then $clusterId = clusterId \cup \{p'\}$
    end for
    Clusters= $clusterId$
  else
    Mark p as noise point
until $uVisited$ list is empty

- The first step includes the application of erosion for finding of boundary points.
- Next step applies the nearest neighboring strategy on the identified boundary points.
- The last step describes the formation of the cluster by using the above method

G. ST-DBSCAN

ST-DBSCAN[13] as compared to the conventional DBSCAN algorithm, requires two more additional parameters $Eps1$,$Eps2$, $MinPts$, and $\Delta\in$, where $Eps1$ is the spatial attribute distance parameter and $Eps2$ non-spatial attribute distance parameter.Euclidean, Manhattan or Minkowski distance measure techniques are used for the calculation of the values of $Eps1$, $Eps2$. The minimum number of points within $Eps1$ and $Eps2$ are called as Minpts. If the density of a region is high, then it contains more number of points than the MinPts threshold value.

H. LSDBC: (Locally Scaled Density Based Clustering Approach)

LSDBC[28] Locally scaled density based clustering system works by manufacturing cluster points by connecting solid regions of space until compactness falls below the threshold value determined by the epicenter of the cluster. LSDBC clustering technique takes two input constraints, one is the k an order of adjacent neighbor consideration for each argument in the dataset for density control and α determines the boundary of the present cluster growth based on its density. The LSDBC algorithm first analyses the ε values for each point grounded on their kNN distances. ε allows us to order points grounded on their density. Smaller ε values resemble to solid regions in dataset. The function kNN Distance takes a point and an amount k and returns the distance to its kth adjacent neighbor ε, as well as the set of its k adjacent neighbors. LocalMax function guarantees that the designated point is the densest point locally in its neighborhood.

I. VDBSCAN: Varied Density Spatial Clustering for Apllications with Noise

This algorithm is used to find the clusters for varied density automatically unlike DBSCAN which is incapable of finding clusters for varied densities. To find out meaningful clusters from the given datasets with varied densities is the sole purpose of this algorithm. The input parameter Eps is generated automatically from the dataset.VDBSCAN has the running time complexity as equal to DBSCAN[29].

J. CUDA-DClust

CUDA-Dclust [30] can produce multiple clusters assigning p an undecided single object simultaneously to the GPU, as instead of DBSCAN could form clusters for only undecided object p (∈D) at time. Each cluster created is called as a chain. To improve the speed of computing ∈ - neighbor N_∈ (p) of p in each block, CUDA-DClust can perform multiple distance operations from p to another object in each thread simultaneously. Collision matrix is a dissimilar metric which can verify whether there is a collision occurred or not ,in case of collision all the chains will merged into a single cluster. Bohm et al [30] inferred that CUDA-DClust outperforms the conventional DBSCAN by 15 times by carrying number of experiments. One of the major disadvantage of this approach is that it is not always necessary to compute the distance for the undecided data objects. The off-chip device memory of GPU stores the collision matrix, which is very high in cost for memory access.

K. DVBSCAN

DVBSCAN [29] algorithm is able to detect clusters from various shapes and sizes but is unable to produce clusters for varied density data. This algorithm also capable to handle within cluster density deviations locally. The input parameters are minimum objects(μ),radius, threshold values(α,λ).The mean and variance of the cluster densities are calculated to expand the consideration of the eps-neighborhood The core object is only allowed to expand if its density variance is less than or equal to its threshold value which will fulfill the similarity index of cluster. To limit the amount allowed local density variations inside the cluster parameters α and λ are used.

L. UDBSCAN

Normally all the traditional clustering algorithms are designed for handling static data objects. But UDBSCAN is an extension to the conventional DBSCAN algorithm to make use of the derivational vector regression function [31]

M. G-DBSCAN :- A Grid Based Density Clustering Approach

G-DBSCAN[32] is the enhanced version or model of the conventional DBSCAN algorithm. The main goal of this algorithm is to decrease the number of the demand objects which are used as preliminary points in the DBSCAN algorithm and to organize the data objects in layout of grids with the middle point of the data into lattice. As a result the request object will be condensed exponentially, which will go on to the better amplified efficiency of the algorithm, and results in the lessening of the memory footprints. But while resolving the G-DBSCAN the following things has to be considered:

The size of the lattice which is to be occupied by taking the data as an input at the genuine performance time.

The importance of the noise threshold charge so that the data which lies less the value of the threshold value will come in noise. The center of the data into the grid. The distance can be calculated as:

$$D = \sqrt{\sum_{i=0}^{n}(x_i - x)}$$

N. GRP-DBSCAN

Grid-Based DBSCAN[18] algorithm is the combination of both grid based approach and density based approach. This algorithm performs better than conventional DBSCAN algorithm in efficiency.

O. DBCURE

DBCURE [21] is a density based clustering algorithm which can find variable density clusters. It is also called as generalized form of DBSCAN algorithm which uses ellipsoidal neighborhoods.

P. GDBSCAN: Generalized Density Based Spatial Clustering of Applications with noise

GDBSCAN [33]This algorithm is the generalized form of the conventional DBSCAN. This algorithm clusters the data using spatial and non-spatial attributes and can also cluster point and polygon objects respectively. GDSCAN generalizes DBSCAN in the following ways, we could use any of the notations of the object neighborhood when reflexive and symmetric are the properties. The properties symmetric and reflexive are binary predicate. We could also define the cardinality of the neighbor with the help of non-spatial attributes. For improving the runtime necessities and the memory we can use R-tress with G-DBSCAN. The Complexity of this algorithm is O (n log n).GDBSCAN has an application in image analysis, remote sensing, molecular biology, geographical information systems.

Q. Fast-DBSCAN

A unique Graph-based indexed structure method group that is able to accelerate the neighbor hunt operations and also accessible for high dimensional datasets[34]. The performance of this algorithm on benchmark datasets with an increase speed of 1.5-2.2 than that of DBSCAN algorithm. The efficiency of the DBSCAN algorithm degraded due to noise points due to the extra work for the computation of the distance while the Fast- DBSCAN method is more efficient to deal with noise in an early stages and reducing the cost of computing the distance calculations.

R. DENDIS

DENDIS[35] is a hybrid DENsity and DIStance based algorithm whose main objective is to give efficient output. This algorithm operates with a concept of keep on adding new representatives at each iterations in order to satisfy the basic objectives. The second objective focuses on comparable space used to fit cluster structure. To pledge with the density requisite the new representatives are chosen in the most crowded set of similar patterns. The algorithm works in two steps:

Step1:- This step considers the distance notations while looking for space density.

Step2:- This step is also known as post processing step and the main purpose is to protect from selecting the noise points or outliners as the representatives.

In this paper we have discussed all the different categories of density based clustering algorithms. All the observations of

Table 2:- Density Based Clustering Algorithms and Techniques.

Algorithms And Techniques	Complexity	I/P Parameters	Author and Year of Publication
DBSCAN	$O(n^2)$ Using R-tree index- $O(n * \log n)$	Two Parameters Eps, Minpts.	Ester et. Al.,1996
DENCLUE	$O(\log \lvert n \rvert)$	Two Parameters σ, ξ	Hinneburg ad Kiem, 1998
DBCLASD	$O(3n^2)$	None	Xu et.Al, 1998
OPTICS	Without index- $O(n^2)$ With Spatial index- $O(n \log n)$	Two Parameters Eps, Minpts.	Mihael Ankrest, H Kriegel,1999
BRIDGE	$O(n \log n)$	Dataset D, and K(Number of cluster)	Dash et. Al, 2001
CUBIN	$O(n)$	Boundary values	Wang and Wang 2003
ST-DBSCAN	$O(n * \log n)$	Eps1(Distance parameter of spatial attribute),Eps2(Distance Parameter of non-spatial attribute),Minpts and $\Delta\in$	Derya Birant , Alp Kut ,2007
LSDBC	$O(n\ log\ n)$	K and α	Ergun Bicici ,2006
VDBSCAN	$O(\mathrm{n})$	Automatically generated	Guilherme Priólli Daniel,2007
CUDA-DClust	$O(n * log\ n)$	Same as DBSCAN	Christian Bohm ,2009
DVDBSCAN	$O(n^2)$	Radius , 2 threshold parameters	M.Parimala ,2011
UDBSCAN	Greater than DBSCAN	O_i an object in consideration	Apinya Tepwankul,2010
Grp-DBSCAN	n^2	Eps,MinPts	Huang Darong,2012
G-DBSCAN	$O(n * log\ n)$	Grid size, noise pts	Li Ma, Lie Gu, 2014
GDBSCAN	$O(n * log\ n)$	Eps,Minpts	Martin Ester, 1998
Fast-DBSCAN	$O(n + nd)$	5*Eps	K.Mahesh Kumar ,2016
DENDIS	$O(n * s)$	Xi,I	Fredric Ros, 2016

Table 3: Comparative analysis of various density based algorithms.

Author/Year	Algorithm	Dataset	Running Times	Measures	
K.Mahesh Kumar A.Rama Mohan Reddy[34],2016	Fast-DBSCAN	Pen-Digits	2.23(s)	Rand index	1
H.Rehioui, A Idrissi[8],2016	DENCLUE DENCLUE-IM	Page-Blocks	71.028(s) 5.749(s)	Dunn Dunn	0.712 0.693
F.Ros, S. Guillaume [35],2016	DENDIS	R#6	662(s)	Rand Index	0.98
S. Nanda , G. Panda [23],2015	DBSCAN	Iris Soyabean	0.1760(s) 0.1650(s)	Accu Accu	80.60 78.72
X. Zhang , H.Liu [36], 2012	FDBSCAN FOPTICS PDBSCAN POPTICS	ECOLI		Efficiency	0.6820 0.6850 0.7156 0.8991
A.Tepwankul,[31] , 2010	UDBSCAN	Artificial	687(ms)	D_{intra} D_{inter}	0.0094 0.0045
P.Vishwanath VS.Babu [37], 2009	Rough DBSCAN	Banana Dataset	0.079(s)	Rand Index	0.98
M. Dash [27],2001	BRIDGE	DS1	78(s)	Efficiency	Better than DBSCAN
A.Hinneburg, D. Keim [3],2001	DENCLUE	Real Time	120(S)	Efficiency	0.95
X.Xu, M.Ester[25] 1998	DBCLASD CLARANS	Synhtetic	77(s) 5013(s)	Efficiency	DBCLASD outperforms CLARANS By Factor of 60

the algorithms with respect to the runtime complexities ,input parameters and the proposed author has been listed in Table-2. In this paper we have analysed the performance of various density based algorithms on specific benchmark datasets and the measures has been listed out in Table-3.

IV. CONCLUSION

In this paper we have described the working principle of density based DBSCAN algorithm and its variants. In this paper the working of DBSCAN algorithm and different density based algorithms is are analyzed and compared. with respect to its performance and input parameters. there are some density based algorithms which are improved and effective in functioning when compared with the DBSCAN algorithm. Some algorithms outperformed the conventional DBSCAN such as Fast-DBSCAN, DENDIS, GDBSCAN. These algorithms worked efficiently than DBSCAN. Still there is need of new density based algorithms which can handle clusters of different densities, free from parameter sensitivity and requires less computational time and space.

References

[1] J. Kutarnia and P. Pedersen, "A Markov random field approach to group-wise registration/mosaicing with application to ultrasound," Med. Image Anal., vol. 24, no. 1, pp. 106–124, 2015.

[2] D. Huang, J. Lai, and C. D. Wang, "Ensemble clustering using factor graph," Pattern Recognit., vol. 50, no. February, pp. 131–142, 2016.

[3] A. Hinneburg and D. Keim, "An Efficient Approach to Clustering in Large Multimedia Databases with Noise," Proc. 4th Int. Conf. Knowl. Discov. Data Min. (KDD 98), no. October 2001, pp. 58–65, 1998.

[4] G. Hamerly, "Making k -means even faster," 2010 SIAM Int. Conf. data Min. (SDM 2010), pp. 130–140, 2010.

[5] X. Tang and P. Zhu, "Hierarchical Clustering Problems and Analysis of Fuzzy Proximity Relation on Granular Space," IEEE Trans. Fuzzy Syst., vol. 21, no. 5, pp. 814–824, 2013.

[6] C. Liabilities and P. Accounting, Chapter 11. 2002.

[7] H. Zhang and J. Lu, "Knowledge-Based Systems Semi-supervised fuzzy clustering : A kernel-based approach," Knowledge-Based Syst., vol. 22, no. 6, pp. 477–481, 2009.

[8] H. Rehioui, A. Idrissi, M. Abourezq, and F. Zegrari, "DENCLUE-IM : A New Approach for Big Data Clustering," Procedia Comput. Sci., vol. 83, pp. 560–567, 2016.

[9] P. Hansen and B. Jaumard, "Cluster analysis and mathematical programming," Math. Program., vol. 79, pp. 191–215, 1997.

[10] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data," Prentice Hall, p. 320, 1988.

[11] L. M. Kieu, A. Bhaskar, and E. Chung, "A modified Density-Based Scanning Algorithm with Noise for spatial travel pattern analysis from Smart Card AFC data," Transp. Res. Part C Emerg. Technol., vol. 58, pp. 193–207, 2015.

[12] M. Ester, H. Kriegel, X. Xu, and D.- Miinchen, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," Kdd, vol. 96, no. 34, pp. 226–231, 1996.

[13] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial-temporal data," Data Knowl. Eng., vol. 60, no. 1, pp. 208–221, 2007.

[14] D. Birant and A. Kut, "ST-DBSCAN : An algorithm for clustering spatial – temporal data," Data Knowl. Eng., vol. 60, no. 1, pp. 208–221, 2007.

[15] B. Borah and D. K. Bhattacharyya, "DDSC: A density differentiated spatial clustering technique," J. Comput., vol. 3, no. 2, pp. 72–79, 2008.

[16] [16] T. Ali, S. Asghar, N. A. Sajid, and M. Ali, "Critical Analysis of DBSCAN Variations," Inf. Emerg. Technol., pp. 1–6, 2010.

[17] A. Ram, A. Sharma, A. S. Jalal, R. Singh, and A. Agrawal, "An enhanced density based spatial clustering of applications with noise," 2009 IEEE Int. Adv. Comput. Conf. IACC 2009, no. November, pp. 1475–1478, 2009.

[18] H. Darong and W. Peng, "Grid-based DBSCAN algorithm with referential parameters," Phys. Procedia, vol. 24, pp. 1166–1170, 2012.

[19] C. Cassisi, A. Ferro, R. Giugno, G. Pigola, and A. Pulvirenti, "Enhancing density-based clustering: Parameter reduction and outlier detection," Inf. Syst., vol. 38, no. 3, pp. 317–330, 2013.

[20] A. Amini, H. Saboohi, T. Herawan, and T. Y. Wah, "MuDi-Stream: A multi density clustering algorithm for evolving data stream," J. Netw. Comput. Appl., vol. 59, pp. 370–385, 2016.

[21] J. S. Kim, Y., Shim, K., Kim, M. S., & Lee, "DBCURE-MR: an efficient density-based clustering algorithm for large data using MapReduce," Inf. Syst., vol. 42, pp. 15–35, 2014.

[22] G. Andrade, G. Ramos, and D. Madeira, "G-DBSCAN : A GPU Accelerated Algorithm for Density-based Clustering," Procedia Comput. Sci., vol. 18, pp. 369–378, 2013.

[23] S. J. Nanda and G. Panda, "Design of computationally efficient density-based clustering algorithms," Data Knowl. Eng., vol. 95, pp. 23–38, 2015.

[24] J. Zhang and J. Kerekes, "An adaptive density-based model for extracting surface returns from photon-counting laser altimeter data," IEEE Geosci. Remote Sens. Lett., vol. 12, no. 4, pp. 726–730, 2015.

[25] X. Xu, M. Ester, H. Kriegel, and J. Sander, "A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases," 14th Int. Conf. Data Eng. (ICDE ' 98), 1998.

[26] M. Ankerst, M. M. Breunig, P. Kriegel, and J. Sander, "OPTICS : Ordering Points To Identify the Clustering Structure," ACM Sigmod Rec., vol. 28, no. 2, pp. 49–60, 2004.

[27] H. Liu, M. Dash, X. Xu, and C. Technology, "'1+ 1 > 2':Merging Distance and Density Based Clustering," Database Syst. Adv. Appl. Seventh Int. Conf. on. IEEE, pp. 32–39, 2001.

[28] E. Bic and D. Yuret, "Locally Scaled Density Based Clustering," Int. Conf. Adapt. Nat. Comput. Algorithms. Springer, Berlin, Heidelb., pp. 739–748, 2007.

[29] M. Parimala, D. Lopez, and N. C. Senthilkumar, "A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases," Int. J. Adv. Sci. Technol., vol. 31, pp. 59–66, 2011.

[30] C. Böhm, C. Plant, B. Wackersreuther, and R. Noll, "Density-based Clustering using Graphics Processors," Proc. 18th ACM Conf. Inf. Knowl. Manag. ACM, pp. 661–670, 2009.

[31] A. Tepwankul and S. Maneewongwattana, "U-DBSCAN : A Density-Based Clustering Algorithm for Uncertain Objects," Data Eng. Work., pp. 136–143, 2010.

[32] L. Ma, L. Gu, B. Li, S. Qiao, and J. Wang, "G-DBSCAN : An Improved DBSCAN Clustering Method Based On Grid," Adv Sci Technol Lett, vol. 74, no. Asea, pp. 23–28, 2014.

[33] X. Xu, "Density-Based Clustering in Spatial Databases : The Algorithm GDBSCAN and Its Applications," Data Min. Knowl. Discov., vol. 2, no. 2, pp. 169–194, 1998.

[34] K. M. Kumar and A. R. M. Reddy, A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method, vol. 58. Elsevier, 2016.

[35] F. Ros and S. Guillaume, "DENDIS: A new density-based sampling for clustering algorithm," Expert Syst. Appl., vol. 56, pp. 349–359, 2016.

[36] X. Zhang, H. Liu, and X. Zhang, "Novel density-based and hierarchical density-based clustering algorithms for uncertain data," Neural Networks, vol. 93, pp. 240–255, 2017.

[37] P. Viswanath and V. S. Babu, "Rough-DBSCAN : A fast hybrid density based clustering method for large data sets," Pattern Recognit. Lett., vol. 30, no. 16, pp. 1477–1488, 2009.