International journal of Computer Science and its Applications [ISSN 2250 - 3765]

# Survey of Density Based Clustering Algorithms

Pooja Batra Nagpal, Priyanka Ahlawat Mann

Computer Science Department
NIT Kurukshetra
Kurukshetra,India
poojabatra9@gmail.com

*Abstract*— **This paper presents a survey of Density based Clustering Algorithms. The paper focus not only on renowned density based algorithms such as DBSCAN, DBCLASD, OPTICS, GDBSCAN, DENCLUE but also algorithms like UDBSCAN, PDBSCAN, DENCLUE2.0 which have better efficiency and cluster clarity even in varying conditions and densities. Here not only advantages and disadvantages of the most common algorithms are discussed but also complexity and run time is given.** *(Abstract)*

*Keywords*— **Density based Clustering, Clustering algorithm, Clustering in the Presence of Noise.** *(Keywords)*

## I. Introduction

Clustering is the technique which is used in almost every place in our real word and different clustering techniques are used to implement clustering .To give operational definition is always tough rather than to give functional definition. So according to Everitt "Clusters may be described as connected regions of a multi-dimensional space containing a relatively high density of points, separated from other such region containing a relatively low density of points."[1] Clustering's main objective is to distribute objects, events etc in groups such that degree of association should be strong enough among members of same cluster and should be weak among members of different clusters.

Generally clustering is classified into two categories i.e. non-exclusive (overlapping) and exclusive (non-overlapping). Exclusive clustering is further divided into two categories i.e. extrinsic (supervised) and intrinsic (unsupervised). Now intrinsic clustering is further divided into hierarchical and partitional methods. [10]. Our main concern, density based algorithms belong to partitional methods. We will present a brief introduction of both methods.

Hierarchical clustering ,as its name suggests is a sequence of partitions in which each partition is nestled into the next partition in the sequence i.e. Maintaining hierarchy. Hierarchical clustering is depicted by binary tree or dendrogram [2] Whole data set is represented by root node and rest of the leaf node is represented as data object. At any level cutting a dendrogram defines clustering and identifies new clusters.

Partitional clustering is of non hierarchical type. It generates a single partition of the data in order to achieve natural groups present in the data. Density based algorithm belongs to partitional clustering.

## II. Density Based clustering

In Density based clustering [3] there is partition of two regions i.e. low density region to high density region .A cluster is defined as a connected dense component that grows in any direction where a density leads. This is the reason that density based algorithms are capable of discovering clusters of arbitrary shapes and provides natural protection to outliers. Basically, density based clustering is divided into two categories i.e. density based connectivity and density function [5].while discussing about density based connectivity, density and connectivity are two main concept comes under this and both measured in terms of local distribution of nearest neighbours. Density based connectivity includes DBSCAN [6], GDBSCAN [7], OPTICS [8] and DBCLASD [9] algorithms and density function includes DENCLUE [12] algorithm.

### A. *DBSCAN*

In DBSCAN (Density-Based Spatial Clustering of Applications with Noise) it relies on a density based notion of cluster and discovers the clusters and noise in a database and based on fact that a cluster of arbitrary shape with noise.

Now two basic concepts or definitions are discussed here which will be used in many of density based algorithms and in DBSCAN as well.

Density reachable-A point p is density reachable from a point q with respect to Eps, Minpts if there is a chain of points $p_1$…..$p_n$ such that $p_{i+1}$ is directly density reachable from $p_i$ .

Density connected- A point p is density connected to a point q with respect to Eps, Minpts if there is a point m such that both p and q are density reachable from $p_i$

Selection of cluster is done but validation of cluster is still remaining. Following lemmas are used for this purpose i.e. clusters satisfying following lemmas will be considered as validates cluster and lemmas are as follows:

**Lemma 1:** Let p be a point in D and $|N_{Eps}(p)| \geq$ MinPts. Then the set M = {m | m $\in$ D and m is density-reachable from p wrt. Eps and MinPts} is a cluster wrt. Eps and MinPts.

It is not obvious that a cluster C wrt. Eps and MinPts is uniquely determined by any of its core points. However, each point in C is density-reachable from any of the core points of

C and, therefore, a cluster C contains exactly the points which are density-reachable from an arbitrary core point of C.

**Lemma 2:** Let C be a cluster wrt. Eps and MinPts and let p be any point in C with $|N_{Eps}(p)| \geq$ MinPts Then C equals to the set M = {m| m is density-reachable from p wrt. Eps and MinPts}.

The runtime of the algorithm is of the order O (n log n) if region queries are efficiently supported by spatial index structures, i.e. at least in moderately dimensional spaces.

*(1) Advantages*

a) DBSCAN does not require you to know the number of clusters in the data a priori, as opposed to k-means.

b) DBSCAN can find arbitrarily shaped clusters. It can even find clusters completely surrounded by (but not connected to) a different cluster. Due to the MinPts parameter, the so-called single-link effect (different clusters being connected by a thin line of points) is reduced.

c) DBSCAN has a notion of noise.

d) DBSCAN requires just two parameters and is mostly insensitive to the ordering of the points in the database.

*(2) Disadvantages*

a) Need to specify global parameters Eps, MinPts in advance from user, which is very difficult.

b) DBSCAN does not respond well to data sets with varying densities (called hierarchical data sets)

### *B UDBSCAN*

UDBSCAN (Density based clustering for uncertain objects) [11] Existing traditional clustering algorithm [12] were designed to handle static objects. The UDBSCAN which is an extension of DBSCAN algorithm uses sample based on the object's uncertain model.

In UDBSCAN extends the existing DBSCAN algorithm to make use of their derived vector deviation function which defines deviation in each direction from the expected representative. Vector deviation, VD= ($vd_1$,…...,$vd_{2^m}$), is a set of $2^m$ vectors which initiate from the expected representative. .In this a new metric is also defined to measure the quality of cluster of density based clustering.

*(1) Advantage*

a) Overcome the drawback of DBSCAN and used for uncertain objects.

*(2) Disadvantage*

a) Problem with this algorithm is that it finds very difficult to extend to high dimensional spaces.

### *C.GDBSCAN*

GDBSCAN (Generalized Density-Based Spatial Clustering of Applications with Noise).It is a generalized version of DBSCAN. It can cluster point objects as well as polygon objects using spatial and non-spatial attributes. GDSCAN generalized DBSCAN in two ways; first if symmetric and reflexive are the two properties of neighbourhood then we can use any notion of the neighbourhood of an object. The two properties that are symmetric and reflexive are termed as binary predicate. Second by calculating the non spatial attributes by defining cardinality of the neighbourhood. GDBSCAN has five important applications. In the first application we cluster a spectral space (5D points) created from satellite images in different spectral channels which is a common task in remote sensing image analysis. The second application comes from molecular biology. The points on a protein surface (3D points) are clustered to extract regions with special properties. To find such regions is a subtask for the problem of protein-protein docking. The third application uses astronomical image data (2D points) showing the intensity on the sky at different radio wavelengths. The task of clustering is to detect celestial sources from these images. The last application is the detection of spatial trends in a geographic information system. GDBSCAN is used to cluster 2D polygons creating so-called influence regions which are used as input for trend detection.

Spatial index structures such as R-trees may be used with GDBSCAN to improve upon its memory and runtime requirements and when not using such a structure the overall complexity is O (n log n).

### *D. P-DBSCAN*

P-DBSCAN [4], a new density-based clustering algorithm based on DBSCAN for analysis of places and events using a collection of geo-tagged photos. Here two new concepts are introduced: (1) density threshold, which is defined according to the number of people in the neighbourhood, and
(2) adaptive density, which is used for fast convergence towards high density regions.

### *E. OPTICS*

OPTICS (**O**rdering **P**oints **T**o **I**dentify the **C**lustering **S**tructure).The basic idea of OPTICS and DBSCAN algorithm is same but there is one drawback of DBSCAN i.e. while densities are varying it is difficult to detect meaningful clusters which was overcome in OPTICS algorithm. It not only stores the core distance but also a suitable reachability distance for each object and creates a proper ordering of database. It requires two parameters ε and MinP. E describes radius and MinP tells minimum no of points that are required to form a cluster. Point p is a core point if at least MinP are found within its ε neighbourhood Nε(p).

$$Core\_dist\varepsilon{,}MinP(p)=\begin{cases} Undefined \quad if\ (|N\varepsilon\ '\ (P)|>=MinP\ ) \\ Distance\ to\ the\ MinPth\ point{,}otherwise \end{cases}$$

$$Reachability\_dist_{\varepsilon\ ,MinP(p,o})\ = \begin{cases} undefined\ if\ (|N\varepsilon\ '\ (o)|>=MinP\ ) \\ Max(core\_\ dist_{\varepsilon\ ,MinP(o),}\ dist_{(o,p)}\ ,othrw \end{cases}$$

*(1) Advantage*

a) Clustering ordering can be used to extract basic clustering information.

*(2) Disadvantage*

b) Wide range of parameter setting is required.

Complexity is O($kN^2$) where k is no of dimensions  
Run time is *O(n log n)*

*F. DBCLASD.*

*DBCLASD* (Distribution Based Clustering of Large spatial Databases).The main idea behind DBCLASD is the assumption that points within a given cluster are uniformly distributed i.e points of a cluster are distributed in such a manner that it should be like a homogeneous i.e Poisson point process which is controlled to certain part of the data space just like a rain fall .So according *to* DBCLASD definition of cluster which is based on the distribution of nearest neighbour distance set (*NNDistSet (K)*) is mentioned as follows:

Definition Let R be set of points .A cluster K is a nonempty subset of R with following properties:

- *NNDistSet(K*)(Nearest neighbour distance set of a set of points)has the expected distribution with a required satisfying level.
- Maximality condition: K is maximal i.e. extension done by neighbouring points of K does not fulfill condition (1).
- Connectivity condition: There is connectivity or path for each pair of points (l,m) of the cluster i.e K is connected.

  Basically two steps are followed in DBCLASD which are as follows:
- Generating Candidates

  First step is to generate candidates i.e. for each new member p of cluster K ,we retrieve new candidate using a region query (i.e. circle query) by selected radius which is choosen such that for no point of cluster a larger distance to the nearest neighbour is expected. So a necessary condition for m(radius) is

  $$N{*}P(NNDistK(p)>m)<1$$
- Testing Candidates

  Now there may be possibility that selected candidates may not fulfil the above conditions for cluster so testing of candidates is required which is done through $\chi^2$ test [15]. Some times few candidates does not satisfy the criteria and considered to be unsuccessful but these are not discarded but again tested later. There is also a possibility that pointsassigned to a particular cluster may switch to another cluster.

*(1) Advantages*

a) Better run time than CLARANS[14]

b) DBCLASD requires no user input

*(2) Disadvantages*

a) It is slower than DBSCAN

The run time of DBCLASD is roughly three times the run time of DBSCAN

*G. DENCLUE*

DENCLUE (DENsity based CLUstEring).In this algorithm concept of influence and density function is used .Here according to authors influence of each data point can be modelled formally using a mathematical function and that is called an influence function. Influence function describe the impact of data point within its neighbourhood. Now the next concept comes of density function which is sum of influences of all data points. According to DENCLUE two types of clusters are defined i.e. centre defined and multi centre defined clusters .In centre defined cluster a density attractor x* ( $f_B^D(x^*) > \xi$ ) is the subset of the database which is density attracted by x* and in multicenter defined cluster it consist of a set of center defined clusters which are linked by a path with significance ξ and ξ is noise threshold.

The influence function of a data object y Є $F^d$ is a function $f^Y : F^d \rightarrow R^+_0$ which is defined in terms of a basic influence function Fb $f_B\ (x) =- f_B$ (x, y).

$$f^y_B\ (x) =- f_B\ (x, y).$$

The density function is defined as the sum of the influence functions of all data points.

$$\hat{f}_B^D(x) = \sum_{x_i \in near(x)} f_B^{x_i}(x)\ .$$

DENCLUE also generalizes other clustering methods such as density based clustering; partition based clustering, hierarchical clustering. In density based clustering DBSCAN is the example and square wave influence function is used and multicenter defined clusters are here which uses two parameter σ = Eps, ξ = MinPts. In partition based clustering example of k-means clustering is taken where Gaussian influence function is discussed. Here in center defined clusters ξ=0 is taken and σ is determined. In hierarchical clustering center defined clusters hierarchy is formed for different value of σ.

Faster than DBSCAN by a factor of up to 45.

*(1) Advantages*

a) It has a firm mathematical basis.

b) It has good clustering properties in data sets with large amounts of noise.

c) It allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets.

d) it is significantly faster than existing algorithms

*(2) Disadvantages*

a) Data points are assigned to clusters by hill climbing, the used hill climbing may make unnecessary small steps in the beginning and never converges exactly to the maximum, it just comes close.

b) Needs a large no of input parameters

### *H. DENCLUE.*

DENCLUE2.0 (Fast Clustering based on Kernel Density Estimation) [13].In both DENCLUE1.0 and DENCLUE2.0 hill climbing procedure is used. Here the main aim of this hill climbing procedure is to maximize the density i.e. ^p(x). Now in gradient based hill climbing first derivative of ˆp(x)is set to be zero and solve for x. Resulting equation is

$$X = \frac{\sum_{t=1}^{N} K(x-x_t/h)x_t}{\sum_{t=1}^{N} K(x-x_t/h)}$$

Where $x_t$ ЄX $R^d$ ,d Є N, t = 1….N and K = Gaussian Kernel Since x influences the right hand side only through the kernel, the idea is to compute the kernel for some fixed x and update the vector on the Left hand side according to above formula. This gives a new iterative procedure with the update formula

$$X= \frac{\sum_{t=1}^{N} K((x^{(l)} - x^{(t)})/h)x_t}{\sum_{t=1}^{N} K(x^{(l)} - x^{(t)})/h}$$

Where h = quantity

## III. Conclusion

*(1) Summary*

This paper presents an up-to-date survey on density based clustering algorithms. It tries to focus not only the renowned algorithms such as DBSCAN, GDBSCAN, DBCLASD, OPTICS ,DENCLUE but also on algorithms such as UDBSCAN, PDBSCAN, DENCLUE2.0 that having improvements in efficiency and runtime than existing algorithms. In addition advantages and disadvantages of the most common algorithms with run time and complexities are discussed.

*(2) Future Work*

All the density based clustering algorithm are efficient but the question arises what are the scenarios in which these algorithms performs better among themselves .More generally we can say that which algorithm is more computationally efficient when compared with others.

So in our next paper we will present a comparative study of density based clustering algorithms.

## Acknowledgment

My sincere thanks to Mrs. Priyanka Ahlawat Mann for her guidance in formulating the paper.

## References

[1] B.*S. Everitt Cluster Analysis*, John Wiley &sons, Inc., NewYork,1974

[2] R. Xu and D.Wunsch, II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005

[3] P.N. Tan, M. Steinbach,V. Kumar, *Introduction to Dat*a *Mining*,Pearson , *Addison Wesley,2006*

[4] S.Kisilevich, F. Mansmann, D. Keim, "P-DBSCAN: a density based Clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos" in Proc of COM,Geo ' 10 of the 1st international conference and exhibition on computing for geospatial Research & Application, doi>10.1145/1823854.1823897

[5] Pavel Berkhin, Survey of Clustering Data Mining Techniques , Wyswilson,2002

[6] Ester M., Kriegel H.-P., Xu X.: "Knowledge Discovery in Large Spatial Databases: Focusing Techniques for efficient Class Identification", Proc. 4th Int. Symp. on large Spatial Databases, Portland, ME, 1995, in: Lecture Notes In Computer Science, Vol. 951, Springer, 1995, pp. 67-82

[7] SANDER, J., ESTER, M., KRIEGEL, H.-P., and XU, X. 1998. Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. In Data Mining and Knowledge Discovery, 2, 2, 169-194

[8] Ankerst, M., Breunig, M., Kreigel, H.-P., and Sander, J. 1999. OPTICS:Ordering points to identify clustering structure. In Proceedings of the ACM SIGMOD Conference, 49-60, Philadelphia, PA.

[9] XU, X., ESTER, M., KRIEGEL, H.-P., and SANDER, J. 1998. A distribution-based clustering algorithm for mining in large spatial databases. In Proceedings of the 14th ICDE, 324-331, Orlando, FL.[10]A. K. Jain, M. N. Murty and P. J. Flynn, Data clustering: a review, CM, 31 (1999), pp. 264–323.

[10] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988

.[11] A. Hinneburg and D. Keim, "An efficient approach to clustering Large multimedia databases with noise," in *Proc. 4th Int. Conf. KnowledgeDiscovery and Data Mining (KDD'98)*, 1998, pp. 58–65.

[12] A. K. Jain, M. N. Murty and P. J. Flynn, Data clustering: a review, ACM, 31 (1999), pp. 264–323

[13] Alexander Hinneburg , Hans Henning Gabriel , "DENCLUE 2.0:Fast Clustering Based on Kernel Density Estimation",Advances in Intelligent Data Analysis vii,vol 4723,2007

[14] Ng R. T., Han J.: "Eficient and Effective Clustering Methods for Soatial Data Mining". Proc. 20th Int. Conf. on Very Large Data Bases, Santiag;, Chile, Morgan Kaufmann

Publishers, San Francisco, CA, 1994, pp. 144-155.

[15] Devore J. L.: ‘*Probability and Statistics for Engineering and the Sciences*’, Duxbury Press, 1991.