Density Based Clustering Algorithms in High Dimensional Non-Linear Data Set

R.NANDHAKUMAR
Assistant Professor - Department of Computer Science,
Nallamuthu Gounder Mahalingam College, Pollachi – 642001, India

Dr.ANTONY SELVADOSS THANAMANI
Associate Professor and Head, Department of Computer Science,
Nallamuthu Gounder Mahalingam College, Pollachi-642001, India

Abstract— Clusters that are formed on the basis of density are very helpful and easy to understand. Also, they do not limit to their shapes. Basically, there are two types of density based approaches. First one is density based connectivity which concentrates on Density and Connectivity and another is Density function which is a total mathematical function. In this paper, a study of the three most popular density based clustering algorithms - DBSCAN, DENCLUE, and DBCLASD is presented and finally a comparison is provided between the same.

Keywords— Clustering, Density based clustering, DBSCAN, DENCLUE, DBCLASD.

I. INTRODUCTION

The age of big data has arrived. Big data may be defined as a term for data sets that are so large or complex that traditional data processing applications prove inadequate. Big data exhibits different characteristics like volume, variety, variability, value, velocity and complexity [1] due to which it is very difficult to analyze data and obtain information with traditional data mining techniques.

Data mining is the process of extraction of hidden patterns or characteristics from such large datasets and transform it in an understandable manner. There are many tasks involved in data mining. One of the tasks is clustering where a set of objects is divided into several clusters where the intra-cluster similarity in maximized and the inter-cluster similarity is minimized.

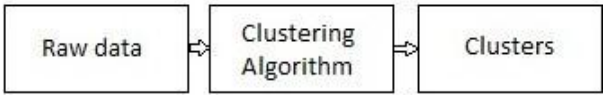

Figure 1: Stages of Clustering [2]

Clustering algorithms may be broadly classified into partition-based algorithms, density-based algorithms, hierarchical-based algorithms, and grid-based algorithms. These algorithms show variation in (i) the procedures used for measuring the similarity (ii) the use of thresholds in developing clusters (iii) the manner of clustering, that is, whether they permit the objects to strictly belong to one cluster or can belong to more clusters in different degrees and the structure of the algorithm [2]. Some of the techniques that fall under partition-based algorithms are PAM, CLARA AND CLARANS. These algorithms take a value k from the user and segment the data into k groups. The popular algorithms under hierarchical category are CHEMELEON and CURE. Grid-based clustering techniques such as CLIQUE, ENCLUS, and WaveCluster aims to divide the space into finite number of cells which make it possible to perform all the clustering operations [3]. The DBSCAN, OPTICS and DENCLUE are some of the most commonly used density-based clustering algorithms.

II. DENSITY BASED CLUSTERING

Density based algorithms find the cluster according to the regions which grow with high density. These algorithms are known as one-scan algorithms. Basically, there are two approaches that may be used in density-based methods. The first approach, called the density-based connectivity clustering, pins density to a training data point. The algorithms that represent this behaviour include DBSCAN and OPTICS. The second approach pins density to a point in the attribute space and is called Density Functions. This behaviour is illustrated by the algorithm DENCLUE.

III. DBSCAN (DENSITY-BASED SPATIAL CLUSTERING OF APPLICATION WITH NOISE)

DBSCAN is one of the most popular clustering algorithms used in data mining. It is a partition type clustering algorithm In this algorithm, a set of points in some space is given and it clusters together points that are closely packed together i.e. points with many nearby neighbors, marking as outliers points that lie away in low-density regions. The data points in DBSCAN may be classified into three categories: (i) Core points i.e. points that are at the interior of a cluster, (ii) Boundary points i.e. non-core points inside a boundary and (iii) Outliers i.e. points that are neither core nor boundary points [4].

ALGORITHM:

The DBSCAN algorithm requires two parameters viz. ε (eps) and the minimum number of points required to build a dense region (MinPts).It selects a random element P from the database. It checks whether P is not a core point, i.e. P has less than MinPts neighbors. If it is true, it will be marked as noise. Else, it will be marked as being in the present cluster and the ExpandCluster function will be called. The objective is to discover all points that are density-reachable from P and are currently being marked as unclassified or

noise. Although it is a recursive function, ExpandCluster is implemented without using recursion. The recursive behavior is accomplished by using a set whose size varies as new density-reachable points are found. The algorithm comes to an end when all points have been properly classified. The pseudo code of DBSCAN algorithm is shown in Figure 2.

```
DBSCAN (Input_Set, ε, MinPts)
foreach p in the Input_Set
   if (p is not in any cluster)
      if (p is a core point)
         generate a new ClusterID
         label p with ClusterID
         ExpandCluster (p, Input_Set, ε, MinPts,
                        ClusterID)
      else
         label(p, NOISE)


ExpandCluster (p, Input_Set, ε, MinPts, ClusterID)
put p in a seed queue
while the queue is not empty
   extract c from the queue
   retrieve the ε-neighborhood of c
   if there are at least MinPts neighbours
      for each neighbour n
         if n is labeled NOISE
            label n with ClusterID
         if n is not labeled
            label n with ClusterID
            put n in the queue
```

Figure 2: Pseudo code of DBSCAN

ADVANTAGES OF DBSCAN:

1. Unlike k-means, DBSCAN does not need one to specify the number of clusters in the data.
2. It can discover arbitrarily shaped clusters. It can also find a cluster completely surrounded by a different cluster.
3. DBSCAN works using two parameters only and it is not affected by the ordering of the points in the database.
4. It is designed for use with databases that are capable of accelerating region queries, e.g. using an R* tree.
5. The algorithm is robust to outliers.

DISADVANTAGES OF DBSCAN:

1. DBSCAN is not completely deterministic.
2. DBSCAN is incapable of clustering data sets well with large differences in densities.
3. In order to choose a meaningful distance threshold ε, the data and scale must be well understood.
4. The quality of DBSCAN depends on the distance measure used in the function regionQuery(P,ε), which ultimately makes it difficult to find an appropriate value for ε.

EXTENSIONS:

In order to overcome these drawbacks, variations of DBSCAN such as VDBSCAN, FDBSCAN and IDBSCAN have been developed. Generalized DBSCAN (GDBSCAN) is a generalization to arbitrary "neighborhood" and "dense" predicates. The MinPts and ε parameters are taken out from the original algorithm and moved to the predicates. Ordering points to identify the clustering structure (OPTICS) is another algorithm whose idea is similar to DBSCAN. But it is capable overcoming one of DBSCAN's major weaknesses. It can detect meaningful clusters in data of varying densities. In order to achieve this, the points of the database are linearly ordered such that points which are spatially closest become neighbors in the ordering. In order to have both points belong to the same cluster, a special distance is stored for each point that represents the density.

IV. DENCLUE (DENSITY BASED CLUSTERING)

DENCLUE [5] (Density based clustering) uses two main concepts i.e. influence and density functions. Influence of each data point can be modeled as mathematical function. The resulting function is called Influence Function. Influence function illustrates the impact of data point within its neighborhood. Second factor is Density function which is sum of influence of all data points. DENCLUE defines two types of clusters i.e. centre defined and multi centre defined clusters. y Є F is an influence function of the data objects. Which is defined in terms of a basic influence function F, F (x) =- F (x, y).

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

DENCLUE is also used to generalize other clustering methods like Density based clustering, partition based clustering, hierarchical clustering. DBSCAN is an example of density based clustering and square wave influence function is used. Multicenter defined clusters here use two parameter σ = Eps, ε = MinPts. In partition based clustering example of k-means clustering is taken where Gaussian Influence function is explained. Here in center defined clusters ε =0 is taken and σ is calculated.

Algorithm:

1. Take Data set in Grid whose each side is of 2σ
2. Find highly dense cells
3. Find out the mean of highly populated cells.
4. If d(mean(cl), mean(c2)) < 4a then two cubes are said to be connected.
5. Now highly populated or cubes that are connected to largely populated cells will be taken into consideration in determining clusters.
6. Find Density Attractors using a Hill Climbing procedure.
7. Randomly pick point r.
8. Compute Local 4 σ density.

9. Pick another point (r+1) close to previous computed density.
10. If den(r) < den(r+1) climb.
11. Put points within (σ /2) of path into cluster.
12. Connect the density attractor based cluster.

ADVANTAGES OF DENCLUE:

1. It has a solid mathematical base and is capable of generalizing various clustering methods like partitioning, hierarchical, and density-based methods.
2. It is a good technique for data sets that contain large amounts of noise.
3. It is faster than DBSCAN.

DISADVANTAGES OF DENCLUE:

1. The density parameter and the noise threshold need to be selected carefully as it significantly affects the quality of results.

V. DBCLASD (DISTRIBUTION BASED CLUSTERING OF LARGE SPATIAL DATABASES)

Basically, DBCLASD [8] is an incremental approach. DBCLASD is based on the assumption that the points inside a cluster are distributed uniformly. DBCLASD dynamically determines the proper number and shape of clusters for a database without needing any input parameters [6]. A random point is assigned to a cluster which is then processed incrementally without considering the cluster.

In DBCLASD, a cluster may be defined by three properties shown below:
1) *Expected Distribution condition:* NNDistSet(C) which is a set of nearest neighbors of cluster C has the expected distribution with required confidence level.
2) *Optimality Condition:* Each point that comes into neighboring of C does not fulfill condition (1).
3) *Connectivity Condition:* Each pair (a,b) are connected to each other through grid cell structure.

Algorithm:
1. Make set of candidates using region query
2. If distance set of C has the expected distribution then point will continue to remain in cluster.
3. Otherwise insert point in list of unsuccessful candidates.
4. In the same way expand cluster and check condition
5. Now list of unsuccessful candidates is checked again through condition.
6. If passes then put in cluster otherwise remain in that list.

Basically, there are two main concepts in DBCLASD. Initial task is generating candidates and candidate generation is done on the basis of region query that shows some radius for circle query to accept candidates. Second task is to test the candidate which is accomplished through chi square testing. Points that lie below the threshold value are considered right candidates while those that lie above threshold remain in unsuccessful candidates' list. In the end, unsuccessful candidate list is checked again and every point goes through the test and points that pass the test are considered in cluster while those left, remain in unsuccessful candidates' list.

ADVANTAGES OF DBCLASD:
1. It is very efficient on large spatial databases.
2. It does not require any input parameters.

DISADVANTAGES OF DBCLASD:
1. The complexity is $3n^2$.
2. It is slower than both DBSCAN and DENCLUE.

VI. COMPARATIVE STUDY OF ALGORITHMS

Table 1. Comparison of three density based algorithms [7]

Name of algorithm	DBSCAN	DENCLUE	DBCLASD
Complexity	$O(n^2)$	O(log |D|)	$O(3n^2)$
Shape of cluster	Arbitrary	Arbitrary	Arbitrary
Varied density type	No	Yes	Yes
Input parameters	Minimum size and radius	Two input parameters	Automatic generation
Cluster testing	No	No	Yes, with 2 features
Type of data	Spatial Data Noise	Large no. of data	Spatial Data with uniformly distributed points
Noise handling	Not Very well	Very well	Good

VII. CONCLUSION

This paper aims to provide an overview of the three most popular Density based clustering algorithms together with their advantages and disadvantages. The algorithms that have been studied are DBSCAN, DENCLUE and

DBCLASD. The paper concludes that each algorithm has a particular use for a particular instance.

VIII. REFERENCES

[1] Avita Katal, Mohammad Wazid, and RH Goudar. Big data: Issues, challenges, tools and good practices. In Contemporary Computing (IC3), 2013 Sixth International Conference on, pages 404-409. IEEE, 2013.

[2] Amandeep Kaur Mann and Navneet Kaur, "Review Paper on Clustering Techniques", Global Journal of Computer Science and Technology, Software and Data Engineering (0975-4350), Volume 13 Issue 5 Version 1.0 Year 2017.

[3] Khan, Kamran, et al. "DBSCAN: Past, present and future." Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on the. IEEE, 2018.

[4] Pavel Berkhin, "Survey of Clustering Data Mining Techniques", Accrue Software, Inc.

[5] Nagpal, Pooja Batra, and Priyanka Ahlawat Mann. "Comparative study of density based clustering algorithms." International Journal of Computer Applications 27.11 (2011): 421-435.

[6] Xu, Xiaowei, et al. "A distribution-based clustering algorithm for mining in large spatial databases." Data Engineering, 1998. Proceedings., 14th International Conference on. IEEE, 2017.

[7] Vivek S Ware, Bharathi H N, "Study of Density based Algorithms", International Journal of Computer Applications (0975 – 8887), Volume 69– No.26, May 2017.

[8] XU, X., ESTER, M., KRIEGEL, H.-P., and SANDER, J.2016. 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 (2016), pp. 264–323.

AUTHORS PROFILE

Name: R.Nandhakumar
Academicians-Department
of Computer Science
Nallamuthu Gounder Mahalingam College,
Pollachi – 642001, India

Name: Dr. Antony Selvadoss Thanamani
Associate Professor and Head
Department of Computer Science,
Nallamuthu Gounder Mahalingam College
College, Pollachi-642001,India