SURVEY ON DENSITY – BASED CLUSTERING METHODS IN DATA MINING

J.Margaret Sangeetha
Assistant Professor
Department of Computer Science
St.Xavier's College (Autonomous)
Palayamkottai.

INTRODUCTION

Partitioning and Hierarchical Methods are designed to find spherical shaped clusters. They have difficulty in finding clusters of arbitrary shape such as the "S" shaped and oval clusters. To find clusters of arbitrary shape, we can model clusters as dense regions in the data space separated by space regions. This is the main strategy behead density-based clustering methods, which can discover clusters of non-spherical shape. Density Based Clustering have proven to be very effective for analyzing large amounts of heterogeneous, complex data for clustering of complex agents.

Density Based Methods

Density Based clustering is to discover clusters of arbitrary shape in spatial databases with noise. It forms clusters based on maximal set of density, connected points. The case part in Density Based clustering is density – reach ability and density connectivity. Also it requires two input parameters i.e) Eps which is known as radius and Min pts is the minimum number of points required to form a cluster. It starts with an arbitrary sharing point that has not varied one. Then the E neighborhood is retrieved, and if it contains sufficiently many points than a cluster is started. Otherwise the point is labeled as noise.

Density Based Clustering Algorithms mainly include three method

Methods	Descriptions	Primary Input	Dataset
DBSCAN	Which grows cluster accordingly to a density based connectivity analysis	Cluster radius, min. of objects	High dimensional data
OPTICS	Extends DBSCAN to produce a cluster ordering obtained from a wide range of parameter settings	Density threshold	High dimensional data
DENCLUE	Cluster objects based on a set of density distribution functions	Radius	High dimensional data

Among many types of clustering algorithms, density based algorithms is more efficient is detecting the cluster with varied density.

INTRODUCTION OF DBSCAN ALGORITHM

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is the most well-known density-based clustering algorithm, first introduced in 1996 by Ester et.al. Due to its importance in both theory and applications, this algorithm is one of three algorithms awarded the Test of Time Award at SIGKDD 2014.

For the DBSCAN algorithm following terms is used in consideration of Database D, Core point (q), Border point (p), Minimum no. of points in cluster (Minpts) and Radius (Eps).

Definition 1: Minimum number of points

Minpts are use to determine whether a neighborhood is denser or not. Minpts specify the density threshold of the denser regions.

Definition 2: Distance of point p within given Eps
Neighborhood point p within Eps value that is referred NEps (p),
Here **NEps (p) ={q€D │ dist (p,q) >= Minpts}**

Definition 3: Core point q condition
Number of NEps (p) is greater than equal to Minpts
i.e. │NEps (p)│>=Minpts

Definition 4: Directly density reachable points
Directly density reachable points are core point q and border point p.
Core point: Minimum numbers of points are needed within Eps-neighborhood.

│NEps (q)│>= Minpts

Border Point: Eps-neighborhood of border point has less point than the Eps of core point.

p€ NEps (q)
q=core point
p=Border point

Definition 5: Density reachable points
Point p is referred as density reachable from another point q in order to Eps and Minpts. If there is a connected chain of point P l to pi, pl=q, pi=p such as pn+1 is directly reachable from pn.

Definition 6: Noise

Any point that is neither core point nor border point and as well as not belongs to any of the cluster is called noise point.

Following are the Algorithm steps for DBSCAN: | 1 |

```
DBSCAN (D, Eps, Minpts)
    C=0
    For each unvisited point p in the dataset D
    Mark p as visited
    N=regionQuery (p, Eps)
    If sizeof(N)<Minpts
        Mark p as Noise
    Else
        Put p into new cluster C
    Enlargecluster(Eps, Minpts, C, p, N)

Enlargecluster(Eps, Minpts, C, p, N)
    Add p to cluster C
    For each point p' in N
    If p' is unvisited
```

```
            Mark p’ as visited
            N’=regionQuery(p’Eps)
    If sizeof(N’)>=Minpts
            N=N’ combine to N
            If p’ is note in any cluster than add p’ to cluster C
```

<u>Advantages of DBSCAN</u>

1. DBSCAN can find arbitrary shape cluster
2. DBSCAN can remove noise from the dataset.
3. It is requires only two parameter which are mostly insensitive ordering of the point in the database

<u>Disadvantages of DBSCAN</u>

1. Multi density dataset are note complete by DBSCAN
2. Run time complexity is high.
3. DBSCAN cannot cluster data sets well with large differences in densities.

VARIOUS METHODS OF DBSCAN

The following are the existing DENSITY BASED algorithms based on the essential parameters needed for a good clustering algorithm.

Algorithm	**Input Parameter**	**Advantages**	**Disadvantages**
DBSCAN	❖ Radius of the cluster (Eps) ❖ Minimum points required inside the cluster (min Pts)	❖ Discovers cluster of arbitrary shape ❖ Holds good for large spatial databases	❖ Only considers point subjects ❖ Fails to detect clusters with varied density
VDBSCAN (Varied Density Based Spatial Clustering of	❖ Automatically select Eps values for different densities	❖ Find clusters with respect to widely varied	❖ Time complexity is high on same as DBSCAN

Applications with Noise)	❖ The parameter K is automatically generated based on the characteristics of the datasets	densities ❖ Automatic generation of i/p parameters depending upon the datasets	❖ The consequence of the magnitude of parameter K for a particular databases is one of the interesting challenges
DVBSCAN (Density Based Algorithm for discovering Density Varied Clusters in Large Spatial Density	❖ Minimum objects ❖ Radius ❖ Threshold values (α & λ)	❖ Able to handle the density variations that exist within the cluster ❖ Finds the cluster that represent relatively ❖ Finds the cluster that represent relatively uniform regions without being	❖ Time complexity is high ❖ i/p parameter can be determined automatically for better clustering

		separated by sparse regions	
DBCLASD (Distribution Based Algorithm for Mining Large Spatial Databases)	❖ no i/p parameters	❖ very efficient & good clustering ❖ suitable for uniform distributions of points	❖ not suitable for non-uniform distribution of points

The **DBSCAN** clustering algorithm usually can be classified into the following categories

Incremental DBSCAN algorithm

Incremental DBSCAN algorithm is capable of adding points in to bulk to existing set of clusters. In this algorithm data points are added to the first cluster using DBSCAN algorithm and after that new clusters are merged with the existing clusters to come up with the modified set of clusters. In this algorithm Clusters are added incrementally rather than adding points incrementally.

Algorithm Steps for the incremental DBSCAN

Advantages of Incremental DBSCAN

1. Allow to see the clustering pattern of the new data along with existing cluster pattern
2. This algorithm clusters can be merged.
3. Incremental clustering approach is more suitable to use in a large multidimensional Dynamic database

Partition Based DBSCAN algorithm

The steps of PDBSCAN are as follows:

1. Partitioning the initial database into N partitions
2. For each partition, building local R/-tree, analyzing and selecting local

Eps and MinPts, and then clustering it with DBSCAN.

3. Merging the partial clusters.

Advantages of PDBSCAN

The initial database is partitioned into N partitions to reduce the time cost.

1. Partitioning database can also alleviate the burden of memory and find more precise parameter. Eps for every partition.

Disadvantages of PDBSCAN:

1. In the first step of PDBSCAN, some articles partition database over the data dimensions. This method will lead to many problems.

C.Boundary Detection Algorithm for each cluster based on DBSCAN (BDAEC)

Firstly, accordingly to the core point percent and the density value of each data subject, all the core points are extracted by this algorithm from the data set. Then, may connected undirected graphs will be constituted by these core points. And the cluster numbers of the data set can be known by those connected undirected graphs for each one of them represents a cluster. Finally, Eps field will be divided into two fields: the positive field and the negative field. And the boundary of each cluster or the whole data set can be detected by the distribution characteristics of the data objects which are located in the positive field and negative field of the given data object.

Advantage:

- BDAEC can obtain the numbers and the boundaries of the clusters with different size or shapes effectively
- BDAEC can extract the boundary of each cluster and the whole data set with the function of avoiding the interference of noises in the dataset.

Disadvantage:

- BDAEC algorithm solves the boundary detection problems only on the numerical datasets and not in categorical datasets and the mixed datasets

D.ST-BDSCAN(Spatial Temporal Density Based Clustering)

ST-DBSCAN algorithm is constructed by modifying DBSCAN [7] algorithm. In contract to existing density based clustering algorithm, ST-DBSCAN [12] algorithm has the ability of discovering clusters with respect to non-spatial, spatial and temporal values of the objects. The three modifications done in DBSCAN algorithm are as follows,

(i) ST-DBSCAN algorithm can cluster spatial-temporal data according to non-spatial. Spatial and temporal attributes.

(ii) DBSCAN does not detect noise points when it is of varied density but this algorithm overcomes this problem by assigning density factor to each cluster.

(iii) In order to solve the conflicts in border objects it compare the average value of a cluster with new coming value.

Advantages:

- Has the ability to discover cluster with respect to non-spatial, spatial and temporal values of the objects
- Used in GIS, medical imaging & weather forecasting

Disadvantage:

- Performance has to be improved.
- Input parameter has to be automatically generated.

CONCLUSION

This paper gives an idea about the density based clustering algorithms and various DBSCAN algorithm based on the essential parameters need for a good clustering. The actual DBSCAN approach is not suitable for a large multidimensional database which is frequently updated. When some records are added to existing data, then it deals with the problem of scanning the whole database again. The time complexity is very high due to rescanning the whole database. It requires more effort as well. The Existing system is not efficient with respect to time and effort. That’s why new system with incremental clustering approach is more suitable to use in a large multidimensional dynamic database. In that case, the incremental clustering approach in much better.

REFERENCES

J.M.Ester, H.P.Kriegel, J.Sandar, X.Xu. **A density-based algorithms for discovering clusters in large spatial databases with noise**. in: Proceedings of the Second International Conference on knowledge. Discovery and Data Mining, Portland, Oregon, 1996, pp, 226-231.

https://blog.dominodatatlab.com/topology-and-density-based-clustering/ on 28/08/2016.

[1] Yaminee S.Patil, M.B.Vaidya " **A technical survey on clustering analysis in data mining"** International Journal of Emerging Technology and Advanced Engineering.

[2] Pradeep Rai, Shubha Singh" **A Survey of Clustering Techniques"** Internationial Journal of Computer Applications.

[3] Sanjay Chakraborty, Prof. N.K.Nagwani " **Analysis and Study of Incremental DBSCAN Clustering Algorithm**" International journal of Enterprise computing and business system.

[4] Martin Ester, Hans-Peter Kriegel Jorg Sander, Michael Wimmer, Xiaowei
Xu, " **Incremental Clustering for mining in a data ware housing",** Univeristy of Munich Oettingenstr. 67, D-80538 Miinchen,Germany.

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

[13] Derya Birant*, Alp Kut " **ST-DBSCAN: An Algorithm for clustering spatial-temporal data**" www.elsevier.com/locate/datak.

Temu Verma, Dr. Deepti Gaur ITM University, Gurgaon, India. **" A survey on Study of Enchanced DBSCAN Algorithm**" International Journal of Engineering Research & Technology (UERT) Vol.2 Issue 11, November – 2013.

J. Data Mining Concepts and Techniques, Kaufman, 2006.

Glory H.Shah, C.K. Bhensdadia, Amit P.Ganatra" An Empirical Evaluation of Density-Based Clustering Techniques".

M.Perimala, D.Lopex, N.C.Senthilkumar, " A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases.". International Journal of Advanced Science and Technology, Vol,31, June 2011.

Data Mining Concepts and Techniques by Jiawei Han and Kamber.