International Journal of Civil Engineering and Technology (IJCIET)
Volume 8, Issue 12, December 2017, pp. 763–767, Article ID: IJCIET_08_12_083
Available online at http://http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=8&IType=12
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

COMPARATIVE STUDY OF DENSITY-BASED CLUSTERING ALGORITHMS

Y. Vijay Bhaskar Reddy
Research Scholar, Rayalaseema University, Kurnool, AP., India

Dr. L.S.S. Reddy
Vice Chancellor, KL University, Vaddeswaram. Guntur, India

Dr. S.Sai Satya Naryana Reddy
Principal, Vardhaman College of Engineering, Hyderabad, India

ABSTRACT

Clustering is an unsupervised learning. It will divide clusters without assigning labels. The process of partitioning the data into groups known as clusters in such a way that the intraclass similarity is high and interclass similarity is low. This paper is proposed to give a comparative study of various density-based clustering algorithms of data mining. The following are different density-based clustering algorithms which will be reviewed in this paper: DBSCAN, OPTICS, and DENCLUE.

Key words Clustering, Outliers, Core Point, Border Point.

Cite this Article: Y. Vijay Bhaskar Reddy, Dr. L.S.S. Reddy, Dr. S. Sai Satya Naryana Reddy. Comparative Study of Density-Based Clustering Algorithms. *International Journal of Civil Engineering and Technology*, 8(12), 2017, pp. 763-767. http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=12

1. INTRODUCTION

Clustering is an unsupervised learning .It will divide clusters without assigning labels. The process of partitioning the data into groups known as clusters in such a way that the intraclass similarity is high and interclass similarity is low. Clustering, in data mining, is a useful technique for discovering structures and patterns in the underlying data [1]. Density-based clustering algorithms cluster the data objects based on density between objects [3].

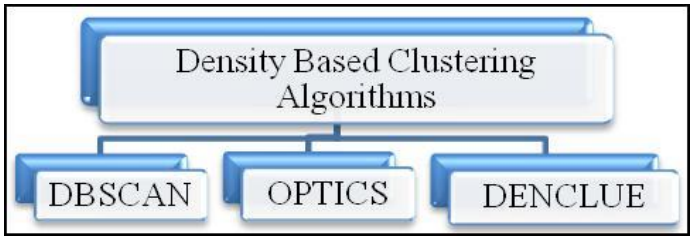

Figure 1 Represents Various Density Based Clustering Algorithms.

This paper is planned as follows. In the section II we will discuss different terminologies and notations which are used throughout this paper are presented. The sections III, IV and V, will explain different Density Based Clustering algorithms such as DBSCAN, OPTICS and DENCLUE. The algorithm and procedures along their pros and cons are listed out. Section VI gives an overview of different surveys on clustering techniques. Section VII presents some concluding remarks. Section VIII includes references.

2. TERMINOLOGIES AND NOTATIONS

This section presents some Terminologies and notations which are used throughout this paper.

Definition 1:

Density: The total Number of Points within a specified radius r (ε)

Definition: 2:

Epsilon (ε): within a specified radius r

Definition: 3:

Min Points: Minimum number of points within a specified radius r.

Definition: 4:

ε -neighborhood of p and q — neighborhood within a radius ε .Mathematically, it can be represented as:

$N\ \varepsilon(a) = \{b \in D | dist(a,b) < \varepsilon\}$ [11-12].

Definition 5:

Core Point: A point is a core point if it has more than a specified number of points (Min points) within – ε. These are points that are present inside of a cluster.

Mathematically, it can be represented as:

$|N\ \varepsilon(q)| \geq$ Min points.

Definition 6:

Border Point: A border point has less than Min points within Epsilon(ε) , but is in the neighborhood of a core point.

Definition 7:

Directly Density-reachable: A point x is directly density-reachable from point y if x is within the ε neighborhood of y, and y is a core point [5].

Definition 8:

Density-reachable: A point x is density-reachable from a point y with respect to ε and Min Points if there is a series of points x1,x2,x3…. xn,

Definition 9:

Density-Connected: A point x is density-connected to a point y with respect to ε and Min Points if there is a point z such that both, x and y are density-reachable from z with respect to ε and Min points [5].

Definition 10:

Cluster Forming: If point 'p' is a part of a cluster C1 and point 'q' is density-reachable from point 'p' with respect to a given distance and a minimum number of points within that distance, then 'q' is also a part of cluster C1.

- p, q: if P ϵ C1 and q is density-reachable from a with respect to ε and Min Points, then q ϵ C1.
- p, q ϵ C1: 'p' is density-connected to 'q' with respect to ε and Min Points.

Definition 11:

Noise Point: Noise point is a point or the set of points, that don't belong to any of the clusters i.e. a noise point, is any point that is not a core point and not a border point.

3. DBSCAN: DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE

DBSCAN is designed to find arbitrary shaped clusters [3]. It finds the points that have dense neighborhoods and connect them with their neighborhoods to form clusters [1].

Algorithm

for each point ***p*** ∈ D (data set) do
if ***p*** is not yet classified then
if ***p*** is a core -object then
Collect all objects density-reachable from ***p***
And assign them to a new cluster.
Else
Assign ***p*** to NOISE

Repeat above procedure for each unvisited objects of N until all objects are visited.
The runtime of DBSCAN algorithm is $O(n^2)$[1].

Advantages

- It can detect outliers.
- It can find arbitrary shaped clusters.
- Scan entire dataset within a single iteration.
- No need to define the number of clusters in advance like k-Means.

Disadvantages

- It can't generate good clusters if data has varying densities [14-15].
- It is not suitable for high-dimensional data like neck type data.
- It requires radius such as epsilon and the Min points in neighborhood to be specified by the user in progress.

4. OPTICS

It is a one of the density-based clustering technique which identifies the implicit clustering in a given dataset [3]. It extends DBSCAN algorithm. Here, priority will be given to high-density clusters over lower density clusters thereby maintaining the order of each data objects that are processed [3]. Thus OPTICS produces an ordering of the given data set that leads to good property clusters [1, 3].

Procedure:

- It initially creates an ordering of data points and also stores the each point core distance and reachability distance.
- It uses the above procedure to retrieve density -based clusters.

The runtime of OPTICS algorithm is O(nlogn) ..

Advantages

- It can handle clusters efficiently, If data has varying densities.
- It retrieves objects in particular order using ordering mechanism.

Disadvantages

- It is less sensitive to erroneous data

5. DENCLUE: DENSITY-BASED CLUSTERING

This algorithm will generate clusters based upon density functions.[2]. The density function internally uses the Gaussian influence function then produce results..

Procedure [1]: It maintains information about It can then determine clusters by identifying local maxima. It means hill climbing technique of the density function.

The estimated runtime of this algorithm is $O(n^2)$

Advantages

- It can handle erroneous data very well.
- It allows a brief description of clusters in high dimensional data sets of non spherical shape.
- It can process clusters much faster than DBSCAN.

Disadvantages

- It needs many constants [3].
- It is less sensitive to outliers.

6. RELATED WORK

Alexander Hinneburg et.al.[2] proposed a new algorithm for clustering in databases which consists of large multimedia information i.e. called DENCLUE that can handle noise. In this approach, they are able to find clusters using local density function. They evaluated the performance of DBSCAN with DENCLUE shows that DENCLUE is having high performance than DBSCAN.

B.G. Obula Reddy et. al.[3] proposed a comparative study of different clustering techniques that enables us to choose the best and efficient clustering algorithm by explaining each of them with their functionality, advantages , and disadvantages.

Mariam Rehman et. al.[4] provided the comparative study of DBSCAN and RDBC(Recursive Density - Based Clustering) by applying these two on iris sample data set and concluded that RDBC is more efficient algorithm than DBSCAN. This can handle outliers more effectively.

Henrik Bäcklund et.al.[5] has provided a brief study of DBSCAN algorithm and terminologies used in DBSCAN. A comparison has been made between DBSCAN and CLARANS (Clustering Large Applications based on Randomized Search) which shows that DBSCAN is more superior to CLARANS in terms of speed and provide good results.

7. CONCLUSIONS

In this paper, the different density- based clustering algorithms were presented. A review has been made on various clustering algorithms with their advantages and disadvantages. The main challenge of Density- based clustering algorithm is handling High dimensional data and Handling of data with varying densities.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Jian Pei,"Data Mining Concepts and Techniques", third edition, 2012.

[2] Alexander Hinneburg, Daniel A.Keim (1998),"An Efficient Approach to Clustering in Large Multimedia Databases with Noise [Online] Available: http://www.aaai.org

[3] B.G. Obula Reddy, Dr. Maligela Ussenaiah, "Literature Survey On Clustering Techniques", IOSR Journal of Computer Engineering, Vol. 3, Issue 1, July 2012.

[4] Mariam Rehman, Syed Atif Mehdi, "Comparison of Density Based Clustering Algorithms", research work, Lahore College for Women University, Lahore, Pakistan.

[5] Henrik Bäcklund, Anders Hedblom, Niklas

[6] Neijman, "DBSCAN A Density-Based Spatial

[7] Clustering of Application with Noise",2011.

[8] R. Elankavi, R. Kalaiprasath and R. Udayakumar, A Fast Clustering Algorithm for High-Dimensional Data. International Journal of Civil Engineering and Technology, 8(5), 2017, pp. 1220–1227

[9] S. Radha Rammohan Anomaly Detection in Mobile Adhoc Networks (MANET) using C4.5 Clustering Algorithm. International Journal of Information Technology & Management Information System (IJITMIS), 7 (1), 2015, pp. 01 - 10 .

[10] Mrs. Meghana P. Lokhande, Mrs. Namrata Gawande, Mrs. Shweta Koprade and Mrs .M. S. Bewoor. Text Summarization Using Hieararchical Clustering Algorithm and Expectation Maximization Clustering Algorithm. International Journal of Computer Engineering and Technology, 6 (10), 2015, pp. 58 - 6 5 .

[11] Neeti Arora, Dr. Mahesh Motwani, A Distance Based C lustering Algorithm, International Journal of Computer Engineering and Technology, Volume 5, Issue 5, May (2014), pp. 109-119.