A Survey of Density Based Clustering Algorithms

Rashi Chauhan, Pooja Batra, Sarika Chaudhary
1, 2, 3 Dept. of CSE, Amity University, Haryana, India

Abstract
Clustering means dividing the data into groups (known as clusters) in such a way that objects belonging to a group are similar to each other but they are dissimilar to objects belonging to other groups. This paper is intended to give a survey of density based clustering algorithms in data mining. Various density based clustering algorithms reviewed are: DBSCAN, OPTICS and DENCLUE.

Keywords
Clustering, Outliers, Core, Border, Attributes

I. Introduction
Clustering means dividing the data into groups (known as clusters) in such a way that objects belonging to a group are similar to each other but they are dissimilar to objects belonging to other groups. Clustering, in data mining, is a useful technique for discovering interesting data distributions and patterns in the underlying data [1]. Density-based clustering algorithms cluster the data objects based on density between objects [3].

Fig. 1: Represents Various Density Based Clustering Algorithms

This paper is organized as follows. In section II definitions and notations which are used throughout this paper are presented. Section III presents DBSCAN algorithm along with its procedure, advantages and disadvantages. Section IV presents OPTICS algorithm along with its procedure, advantages and disadvantages. Section V presents DENCLUE algorithm along with its procedure, advantages and disadvantages. Section VI gives an overview of various surveys on clustering techniques. Section VII presents some concluding remarks. Section VIII includes references.

II. Definitions and Notations
This section presents some definitions and notations which are used throughout this paper.

Definition 1:
$\varepsilon$-neighbourhood of an object — neighbourhood within a radius $\varepsilon$ of an object. Mathematically, it can be represented as: $N_\varepsilon(a) = \{b \in D | \text{dist}(a,b) < \varepsilon\}$ [11-12].

Definition 2:
Core object — if $\varepsilon$-neighbourhood of object contains at least minimum number of points (MinPts), then it is called core object. Mathematically, it can be represented as: $|N_\varepsilon(b)| \geq \text{MinPts}$

Definition 3:
Border object — an object whose $\varepsilon$-neighbourhood does not contain at least minimum number of points (MinPts).

Definition 4:
Directly density-reachable — an object $a$ is directly density-reachable from object $b$ if $a$ is within the $\varepsilon$-neighbourhood of $b$, and $b$ is a core object [5, 11-12].

Definition 5:
Density-reachable — A point $a$ is density-reachable from a point $b$ with respect to $\varepsilon$ and MinPts if there is a chain of points $a_1 \ldots a_n$ such that $a_1 = b$, $an = a$ such that $aj+1$ is directly density-reachable from $aj$ [5, 11, 13].

Definition 6:
Density-connected — A point $a$ is density-connected to a point $b$ with respect to $\varepsilon$ and MinPts if there is a point $c$ such that both, $a$ and $b$ are density-reachable from $c$ with respect to $\varepsilon$ and MinPts [5, 11].

Definition 7:
Cluster — If point $a$ is a part of a cluster $C$ and point $b$ is density-reachable from point $p$ with respect to a given distance and a minimum number of points within that distance, then $b$ is also a part of cluster $C$.

- $a, b$: if $a \in C$ and $b$ is density-reachable from $a$ with respect to $\varepsilon$ and MinPts, then $b \in C$.
- $a, b \in C$: $a$ is density-connected to $b$ with respect to $\varepsilon$ and MinPts.

Definition 8:
Noise — Noise is the set of points, in the database, that don’t belong to any of the clusters.

Definition 9:
Core-distance — It is smallest $\varepsilon$ value that makes object $a$ a core object.

Definition 10:
Reachability-distance — The reachability-distance of an object $b$ with respect to another object $a$ is the higher value (of the smallest $\varepsilon$ value that makes object $a$ a core object) of $a$ and the cartesian distance between $a$ and $b$.

III. DBSCAN: Density-Based Spatial Clustering of Applications With Noise
DBSCAN is designed to find non-spherical shaped clusters [3]. It finds the objects that have dense neighbourhoods and connect them with their neighbourhoods to form clusters [1].

Procedure [1]:
- Initially all the objects are marked unvisited.
- Then, randomly select an unvisited object $m$ and mark it as visited.
- Check if its neighbourhood has no minpoints objects, then mark $m$ as noise point.
- Else
  - Form a new cluster $C$ and add $m$ to $C$.
  - Add remaining objects of neighbourhood of $m$ to set $N$.
- Repeat same procedure for each unvisited objects of $N$ until
all objects are visited.
The run time of DBSCAN algorithm is $O(n^2)$ [1].

**Advantages**
- It can detect outliers.
- It can find non-spherical shaped clusters.
- Its processing is fast.
- There is no need to define number of clusters in advance.

**Disadvantages**
- It cannot efficiently find clusters if data has changeable density [14-15].
- It is not suitable for high-dimensional data.
- It requires radius and minimum number of points in neighbourhood to be specified by user in advance.

**IV. OPTICS: Ordering Points To Identify the Clustering Structure [11, 14]**

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

**Procedure:**
- It first creates an ordering of data objects and stores core distance and reachability distance for each data object.
- Then it uses above information to retrieve density based clusters.

The run time of OPTICS algorithm is $O(n\log n)$ [1].

**Advantages**
- It solves the problem of finding good clusters if data has changeable density [14-15].
- It outcomes the objects in a particular ordering.

**Disadvantages**
- It expects some kind of density decline to find cluster borders [3].
- It is less sensitive to erroneous data [3].

**V. DENCLUE: DENsity-based CLUstEring**

It is a clustering algorithm that depends upon density functions [2]. The density function that results from the Gaussian influence function is $
\int \text{D Gauss}(a) = \frac{1}{2\sigma^2} e^{-\frac{(a-a_j)^2}{2\sigma^2}}$

**Procedure [1]:**
- It maintains information about only those grid cells in hierarchical manner that contains data points.
- It can then determine clusters by identifying local maxima (calculated by hill climbing algorithm) of density function.

The run time of DENCLUE algorithm is $O(n^2)$ [1].

**Advantages**
- It detects erroneous data very well.
- It allows a brief description of non-spherical shaped clusters in high-dimensional data sets.
- Its processing is much faster than DBSCAN.

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

**VI. Related Work**

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

B.G. Obula Reddy et al. [3] proposed a comparative analysis of various clustering techniques that enables us to choose best clustering algorithm by explaining each of them with characteristics, examples, positive and negative aspects.

Mariam Rehman et al. [4] provided comparison between DBSCAN and RDBC algorithm by implementing them using iris data set that concluded that RDBC is more efficient algorithm than DBSCAN as it can handle outliers more effectively.

Henrik Bäcklund et al. [5] has first provided description about DBSCAN algorithm with all its relevant terminologies. A comparison has been made between DBSCAN and CLARANS which shows that DBSCAN is more superior to CLARANS in terms of speed and provide good results. Finally they presented various possible applications, positive and negative aspects of DBSCAN algorithm.

Anoop Kumar Jain et al. [8] presented a survey of recent clustering techniques for data mining research that includes centroid-based, connectivity-based, density-based and distributive-based clustering and discussed about k-means, rapid clustering method and hierarchical agglomerative clustering.

Rui Xu et al. [9] explained procedure of cluster analysis using few steps with a feedback loop. They carried out complexity comparison among various clustering algorithms. They made a review on various approaches for clustering large data set, data visualization and sequential data.

**VII. Conclusion**

In this paper, several density-based clustering algorithms proposed for clustering have been discussed. A review has been made on various clustering algorithms which are density based. All are discussed along with their advantages and disadvantages. This survey can be helpful for understanding of several clustering algorithms for choosing appropriate algorithm. The type of algorithm that is to be chosen depends upon type of clusters that are needed to be finded, type of data set and number of attributes.

**References**


Rashi Chauhan has received her B.Tech degree in Computer Science & Engineering from Maharshi Dayand University, Rohtak, India. At present, she is pursuing M.Tech in Computer Science & Engineering from Amity University Haryana, India. Her research interest includes data mining, software testing, software engineering, etc.