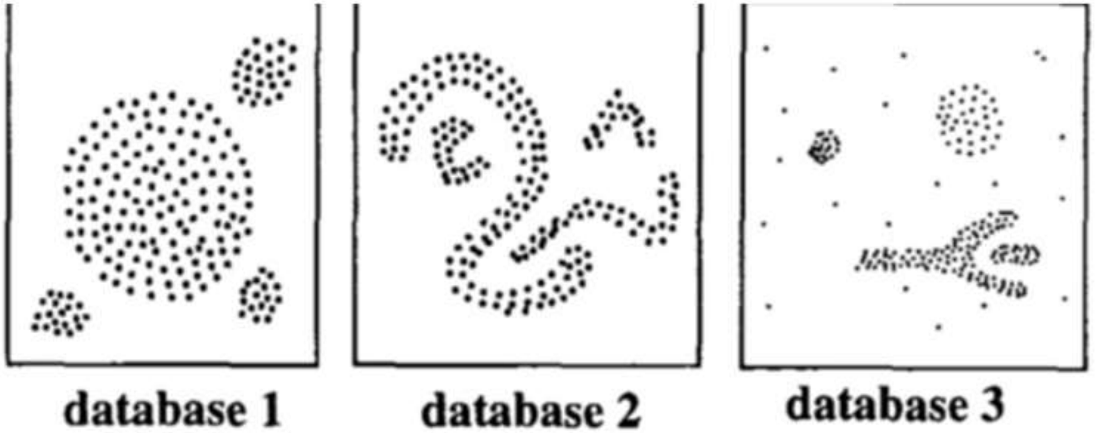
DBSCAN - A Clustering Algorithm

Pınar YAHŞİ

Review

Clustering is to group objects into meaningful subclasses. But there are some difficulties:

- Not to have information about the data to be clustered.
- The separation of clusters in ambiguous/arbitrary shapes.
- Large amount of data.



database 1

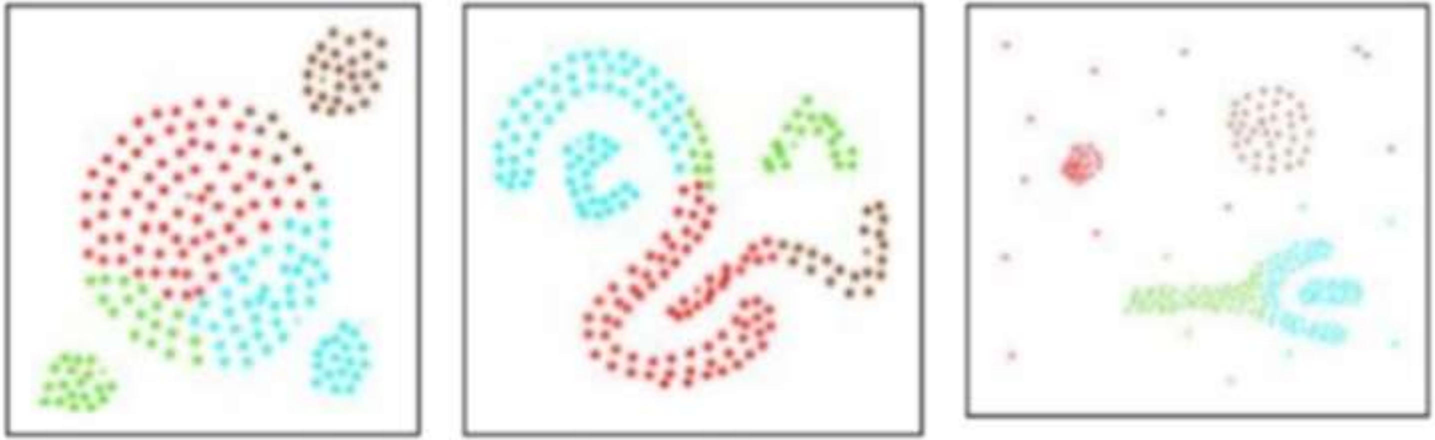
database 2

database 3

Many clustering algorithms clusters according to the distance difference and similarities between data points. Therefore, the result is generally spherical.

similarity = 1 / distance

So these methods fail in concave clusters.



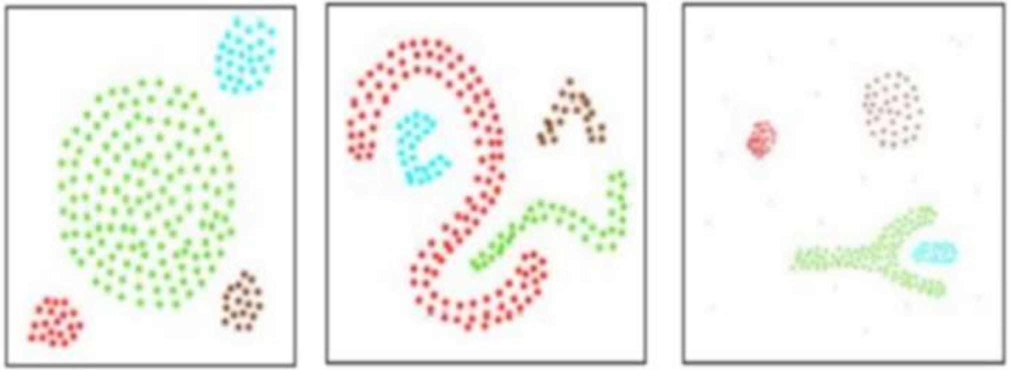
The classification of the CLARANS algorithm.

How does the DBSCAN work?

DBSCAN- Density-Based Spatial Clustering of Applications with Noise.

Clustering is done according to the density of the data. Therefore it is independent of shape and size. So, dbscan is also successful in arbitrary-shaped, large databases and is not affected by the noisy data.

Unlike many clustering algorithms, each point does not have to belong to a cluster.



The classification of the DBSCAN algorithm.

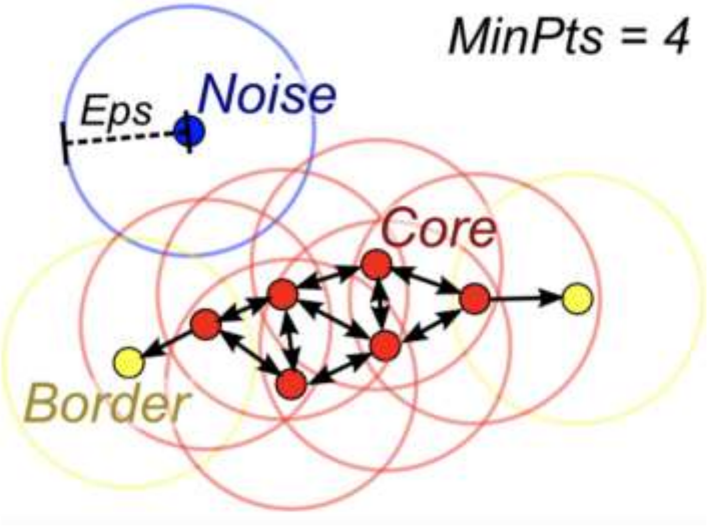
Algorithm marks the lonely points in low density regions and group the points located close together. Two main parameters;

- Ɛ (Epsilon, Eps): largest radius of neighborhood around a point.
- MinPts (minimum points,density): minimum number of points in the neighborhood with radius Ɛ.

Methods such as the distance from Euclidean or Manhattan or other measurement approaches can be used for density measurement.

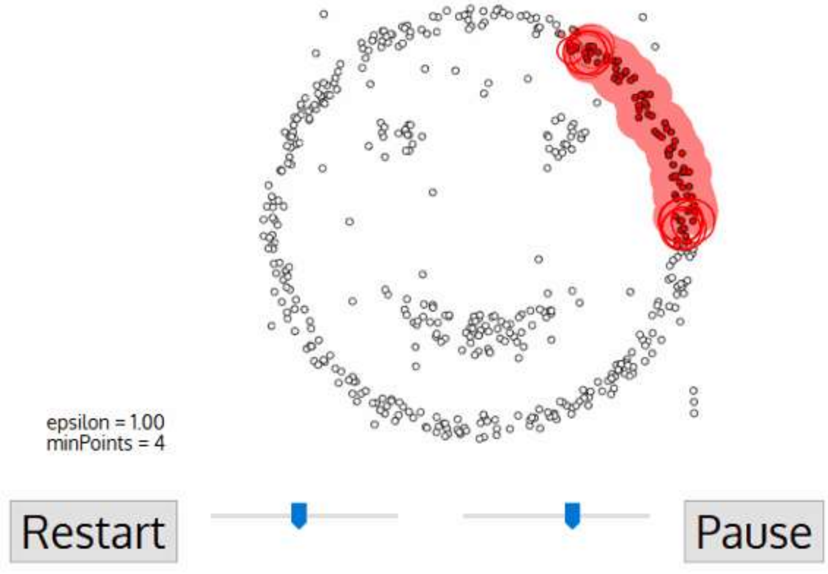
In DBSCAN, the points are labeled in 3 different types:

- Core Point: is a data point that contains greater than or equal to minPts within radius Ɛ.
- Border Point: number of neighbors is less than minPts, but it belongs to the Ɛ-neighborhood of some core point z.
- Noise Point: neither a core nor a border point (outlier).



Algorithm

```
x: data point
D: set of points
for each x ∈ D do
    if x is not yet classified then
       if x is a core point then
           collect all objects density-reachable from o
                   and assign them to a new cluster
       else
           assign x to NOISE
```



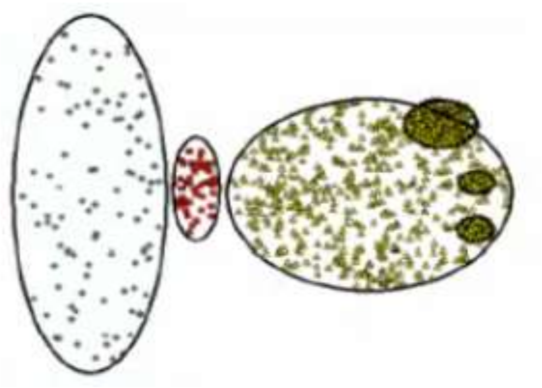
epsilon = 1.00
minPoints = 4

Advantages

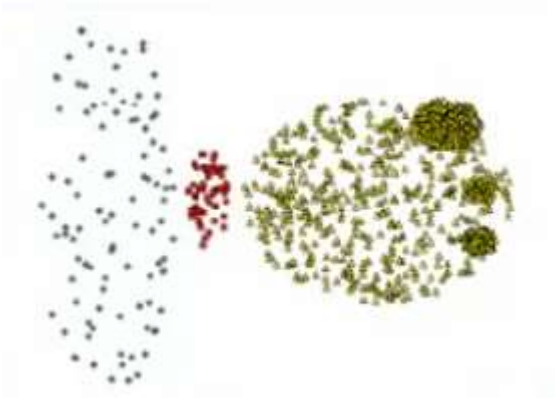
- Can handle clusters different shapes and sizes.
- Resistant to noise

Disadvantages

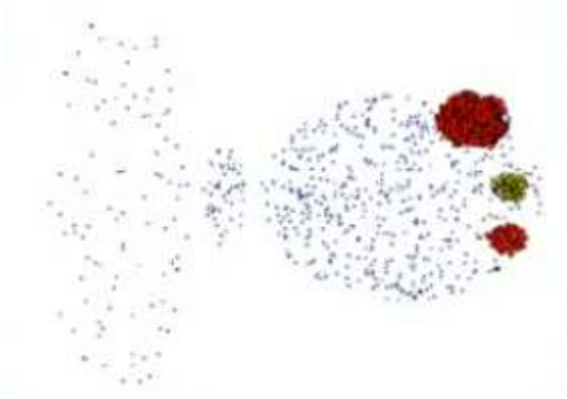
- sensitive in parameter selection.



Original Points

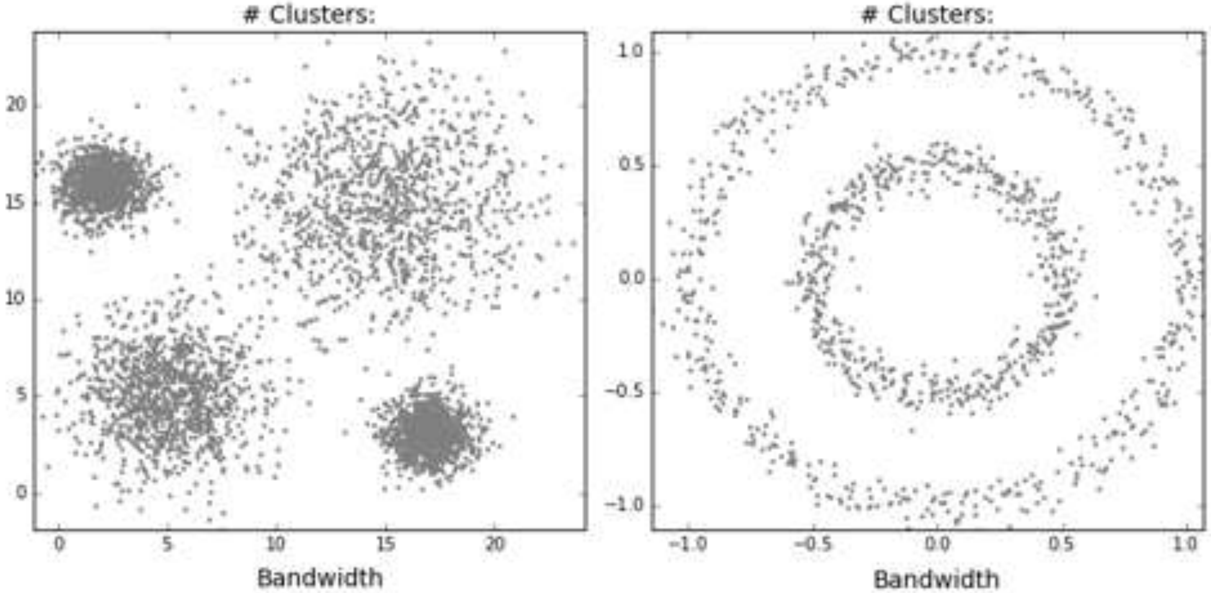


minPts:4, Eps: 9,75



minPts:4, Eps:9,92

Effect of bandwidth value



Thank you for listening…

References

- http://www.sthda.com/english/wiki/wiki.php?id_contents=7940
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html
- https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80
- https://medium.com/@elutins/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818
- https://iq.opengenus.org/dbscan-clustering-algorithm/
- http://ahmetcevahircinar.com.tr/2017/04/17/a-density-based-algorithm-for-discovering-clusters-in-large-spatial-databases-with-noise/
- https://www.youtube.com/watch?v=EtYG-xtU-4g&t=4s
- https://www.youtube.com/watch?v=ktmjTCVmK-s
- http://yarpiz.com/255/ypml110-dbscan-clustering
- https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/ (visualization dbscan algorithm)
- https://www.ahmetcevahircinar.com.tr/wp-content/uploads/2017/04/A-density-based-algorithm-for-discovering-clusters-in-large-spatial-databases-with-noise.pdf (original article)