Clustering

M.M. Pedram
pedram@tmu.ac.ir
Tarbiat Moallem University of Tehran
(Fall 2011)

1

Density-Based Clustering Methods

Lecture slides taken/modified from:
- Jiawei Han (http://www-sal.cs.uiuc.edu/~hanj/DM_Book.html)
- Vipin Kumar (http://www-users.cs.umn.edu/~kumar/csci5980/index.html)

2

Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition
- Several interesting studies:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - DENCLUE: Hinneburg & D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)

3

DBSCAN (M. Ester, et al. 1996)

- **DBSCAN**: **D**ensity **B**ased **S**patial **C**lustering of **A**pplications with **N**oise
- Outliers will not effect creation of cluster.
- Have, in principle, the ability to discover clusters of *arbitrarily shaped clusters*.
- One scan
- Need density parameters as termination condition
- Time complexity less than $O(N^2)$.

4

DBSCAN

Nomenclature:

- $ball_\varepsilon(q)$: the hyper-sphere (n-dimensional ball) of radius ε centered at q.
- $N_\varepsilon(q)$: the set of points lying in $ball_\varepsilon(q)$.

  $N_\varepsilon(q) = \{x \mid dist(x,q) \le \varepsilon\}$ or $N_\varepsilon(q) = \{x \mid x \in ball_\varepsilon(q)\}$
- $|N_\varepsilon(q)|$: the number of points lying in $ball_\varepsilon(q)$.

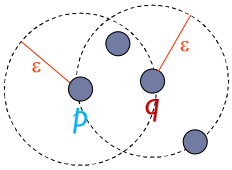
2 input Parameters:

- **ε** : Maximum radius of the neighborhood.
- ***MinPts*** : minimum number of points in cluster (ε-neighborhood of that point).

5

DBSCAN Concepts

- ***ε-neighborhood***: Points within ε distance (radius) of a point.
- ***Density***: number of points within a specified radius (ε).
- ***Core point***: ε-neighborhood dense enough (*MinPts*),
  - or, If the ε-neighborhood contains at least a minimum number of points *Minpts*, then the object is a core object.
  - or, q is core point if $|N_\varepsilon(q)| \ge MinPts$.
  - These are points that are at the interior of a cluster.



ε-neighborhood of q

ε-neighborhood of p

q is a core point (*MinPts* = 4)

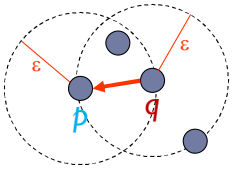
p is not a core point.

6

DBSCAN Concepts

- ***Directly density-reachable***: point p is directly density-reachable from a point q w.r.t. ε, *MinPts* if
  1. p belongs to ε-neighborhood of q,
     - or, the distance is at most ε,
     - or, $p \in ball_{\varepsilon}(q) \equiv p \in N_{\varepsilon}(q)$.
  2. q is a core point,

DDR is an asymmetric relation!



MinPts = 4

p is DDR from q.

q is not DDR from p!

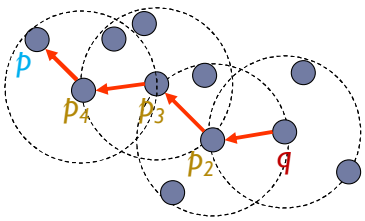
7

DBSCAN Concepts

- ***Density-reachable***: A point p is density-reachable from a point q w.r.t. ε, *MinPts* if there is a chain of points $p_1, \ldots, p_n$, $p_1=q$, $p_n=p$ such that p_{i+1} is directly density-reachable from p_i.

Or, point p is density-reachable form q, if there is a path (chain of points) from p to q consisting of only core points.

DR is an asymmetric relation!



MinPts = 4

p is DR from q.

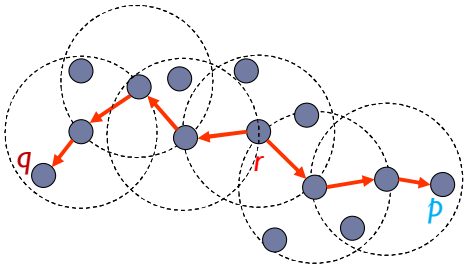
q is not DR from p!

p is not core.

8

DBSCAN Concepts

- ***Density-connectivity***: point *p* is density-connected to point *q* w.r.t. ε, *MinPts* if there is a point *r* such that both, *p* and *q* are density-reachable from *r* w.r.t. ε and *MinPts*.



DC is an symmetric relation!

MinPts = 4

p and *q* are density-connected.

9

DBSCAN Concepts

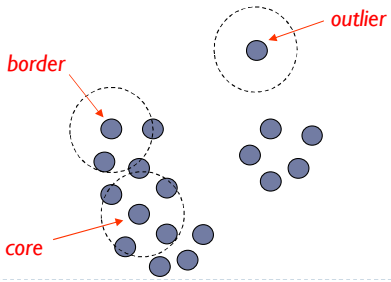
- DBSCAN relies on a *density-based* notion of cluster.
- ***Cluster***: a cluster **C** is a non-empty set of *density-connected* points that is *maximal* w.r.t. density-reachability.
  - Maximality: For all *p*, *q*; if $p \in \mathbf{C}$ and if *q* is density-reachable from *p* w.r.t. ε and *MinPts*, then also $q \in \mathbf{C}$.
- **Border point**: density-reachable from a core.
  - or, has fewer than *MinPts* within ε-neighborhood, but is in the neighborhood of a core point.

10

DBSCAN Density Concepts

- **Noise (outlier) point** : is any point that is not a core point nor a border point.
  - or, not directly density-reachable from at least one core,
  - or, Let $C_1,\ldots,C_m$ be the clusters, every point not contained in any cluster is considered to be *noise*.
- cluster contains *core points* as well as *border points*.

MinPts = 5
ε = circle radius



11

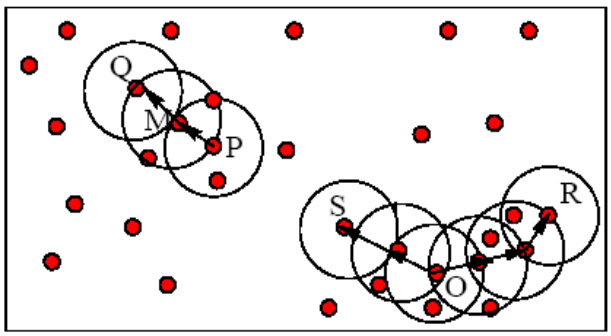
DBSCAN Density Concepts

Example:

- Core objects : M, P, O, R

  (each one is in an ε–neighborhood containing at least 3 points.)

MinPts = 3
ε = circle radius



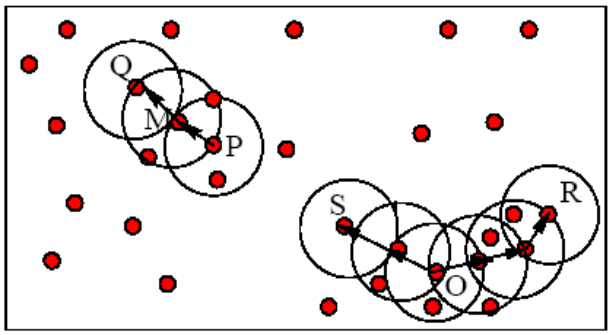
12

DBSCAN Density Concepts

Example:

- Q is directly density reachable from M.
- M is directly density reachable from P and vice versa.

MinPts = 3
ε = circle radius



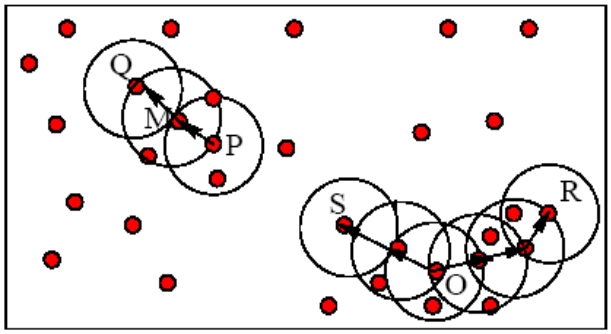
13

DBSCAN Density Concepts

Example:

- Q is (indirectly) density reachable from P since Q is directly density reachable from M and M is directly density reachable from P.
- P is not density reachable from Q since Q is not a core object.

MinPts = 3
ε = circle radius



14

DBSCAN – thoughts behind the algorithm

Note 1:

- If q is a core point and D_q is the set of points that are density reachable from q, then D_q is a cluster.

Note 2:

- If **C** is a cluster and q is a core point in **C**, then **C** equals to the set of the points p that are *density reachable* from q.

⇒ A cluster is uniquely defined by any of its core points.

15

DBSCAN algorithm

- DBSCAN searches for clusters by checking the ε-neighborhood of each point in the database. If the ε-neighborhood of a point p contains more than *MinPts*, a new cluster with p as a core point created.
- DBSCAN then iteratively collects directly density-reachable objects from these core objects, which may involve the merge of a few density-reachable clusters.
- The process terminates when no new point can be added to any cluster.

16

DBSCAN Algorithm

- Arbitrary select a point *p*
- Retrieve all points density-reachable from *p* w.r.t. ε and *MinPts*.
- If *p* is a core point, a cluster is formed.
- If *p* is a border point, no points are density-reachable from *p* and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

DBSCAN Algorithm

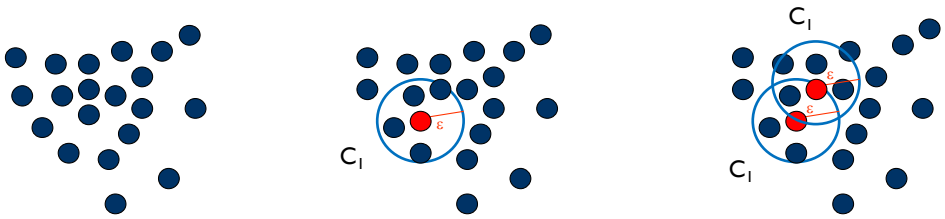
$D_{unprocessed} \leftarrow D$ // D: data-set
$no\_of\_clusters \leftarrow 0$
while $D_{unprocessed} \neq \varnothing$ do
  Arbitrarily select a $p \in D_{unprocessed}$
  if p is a non-core point, then
    - Mark p as noise point
    - $D_{unprocessed} \leftarrow D_{unprocessed} - \{p\}$

  else //p is a core point
    - $no\_of\_clusters$ ++
    - $D_{DR}(p) \leftarrow$ Determine all *Density-Reachable* points in D from p
      //note: The border points that may have been marked as noise, now belong to $D_{DR}(p)$.
    - $Cluster_{no\_of\_clusters} \leftarrow \{p\} + D_{DR}(p)$
    - $D_{unprocessed} \leftarrow D_{unprocessed} - Cluster_{no\_of\_clusters}$

  end-if
end-while

DBSCAN

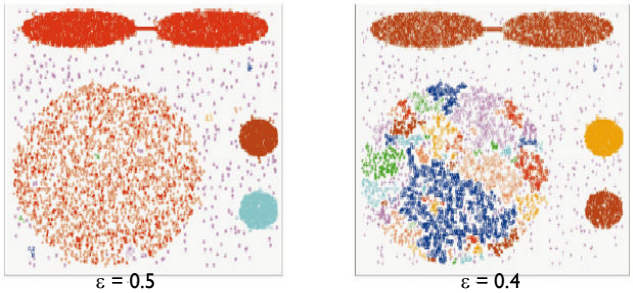
MinPts = 4



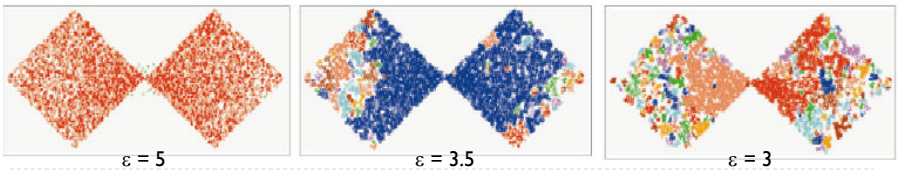
19

DBSCAN

DBSCAN is Sensitive to Parameters. *MinPts* = 4



ε = 0.5 ε = 0.4



ε = 5 ε = 3.5 ε = 3

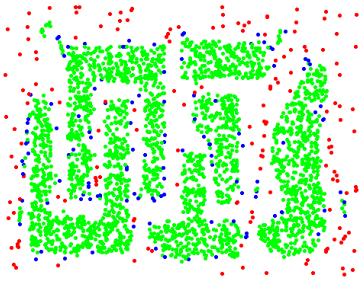
20

DBSCAN

❖ Core, Border and Noise Points:

MinPts = 4, ε = 10

Original Points



Point types: core, border and noise

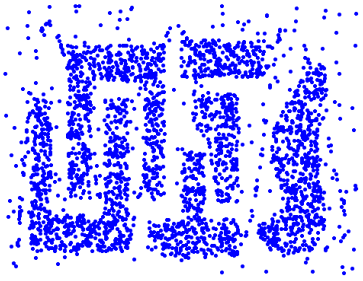
21

DBSCAN

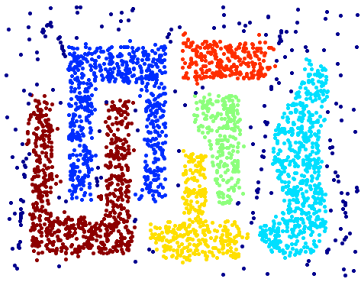
When DBSCAN works well:

❖ Resistant to Noise

❖ Can handle clusters of different shapes and sizes



Original Points

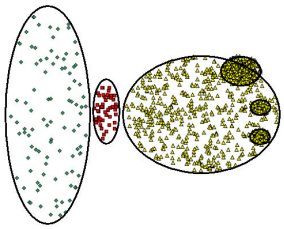


Clusters

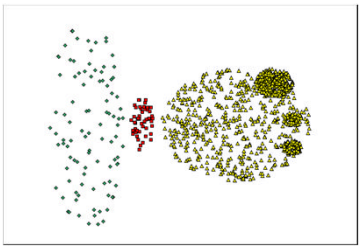
22

DBSCAN

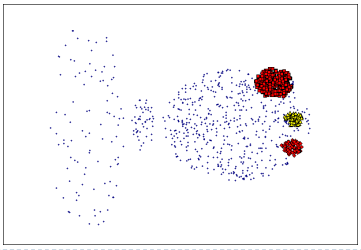
When DBSCAN does not work well:



Original Points



MinPts =4,
ε =9.75



MinPts =4,
ε = ?

- Varying densities
- High-dimensional data

DBSCAN: Determining ε and *MinPts*

- Idea : for points in a cluster, their *k-th* nearest neighbors are at roughly the same distance.
- Noise points have the *k-th* nearest neighbor at farther distance.
- So, plot sorted distance of every point to its *k-th* nearest neighbor.

- *k*-dist : Distance from a point to its *k-th* nearest neighbor.
  - For points that belong to some clusters, the value of *k*-dist will be small if *k* is not larger than cluster size.
  - For points that are not in a cluster such as noise points, the *k*-dist will be relatively large.

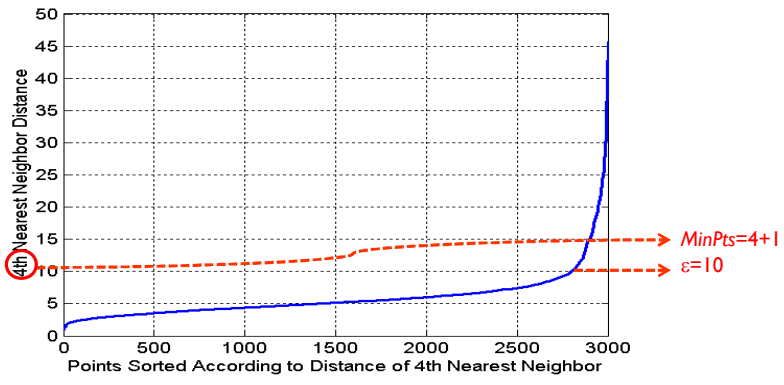
DBSCAN: Determining ε and *MinPts*

Method:

- Compute *k*-dist for all points for some *k*.
- Sort them in increasing order and plot sorted values.
- A sharp change at the value of *k*-dist that corresponds to suitable value of ε and the value of *k* as *MinPts*.
  - Points for which *k*-dist is less than ε will be labeled as core points while other points will be labeled as noise or border points.

25

DBSCAN: Determining ε and *MinPts*



26

DBSCAN: Determining ε and *MinPts*

- If *k* is too small ⇒ Even a small number of closely spaced points that are noise or outliers will be incorrectly labeled as clusters
- If *k* is too large ⇒ small clusters (of size less than k) are likely to be labeled as noise

- If a spatial index (ex, *kd*-tree, R^*-tree) is used, the computational complexity of DBSCAN is $O(n.\log n)$, where n is the number of database objects. Otherwise, it is $O(n^2)$.

Conclusion

- DBSCAN is sensitive to the choice of input parameters.
- DBSCAN is a density-based clustering algorithm which is able to discover clusters of arbitrary shapes. Distance is not the metric unlike the case of hierarchical methods.
- Parameter setting is done empirically:
  - They should be selected such that it would be able to detect the least "dense" cluster, thus experimentation with several values for parameters should be done.
- High dimensional data – more noticeable (pronounced).
- High dimensional data clustering structures are not generally characterized by global density parameters like ε & *MinPts*,
- OPTICS as a solution!