COMP 345: Data Mining
More on Clustering

Slides Adapted From : Jiawei Han, Micheline Kamber & Jian Pei
Data Mining: Concepts and Techniques, 3rd ed.

Rhodes College

Announcements

- Assignment 5 has been assigned
  - Due Wed. Oct. 24th/Thurs. Oct. 25th at the beginning of class
- Extra Credit Opportunity (2 points):
  - Attend Dr. Lee Giles talk on Deep Learning
  - **Friday, October 19th at 3p.m. in FJ-B**
  - Turn in a 1 paragraph summary of what you learned and what you found interesting about the talk by beginning of class on Mon. Oct. 22nd/Tues. Oct. 23rd
- Future Extra Credit Opportunity (2 points)
  - Attend Dr. Stanley Pounds talk on his Biostatistics Research at St. Jude
  - **Thurs. Nov. 1st at 4pm in Spence Wilson Room**
  - Turn in a 1 paragraph summary of what you learned and what you found interesting about the talk by beginning of class on Mon. Nov. 5th /Tues. Nov. 6th

2

What is Cluster Analysis?

- Cluster: A collection of data objects
  - similar (or related) to one another within the same group
  - dissimilar (or unrelated) to the objects in other groups
- Cluster analysis (or *clustering*, *data segmentation, …*)
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- Unsupervised learning: no predefined classes (i.e., *learning by observations* vs. learning by examples: supervised)
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

3

Density-Based Clustering Methods

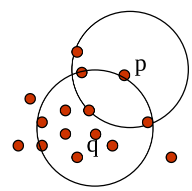
- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition
- Several interesting studies:
  - DBSCAN: Ester, et al. (KDD'96)
  -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

4

Density-Based Clustering: Basic Concepts

- Two parameters:
  - *Eps*: Maximum radius of the neighborhood
  - *MinPts*: Minimum number of points in an Eps-neighborhood of that point
- $N_{Eps}(q)$: {p belongs to D | dist(p,q) ≤ Eps}
- Directly density-reachable: A point *p* is directly density-reachable from a point *q* w.r.t. *Eps*, *MinPts* if
  - *p* belongs to $N_{Eps}(q)$
  - core point condition:

    $|N_{Eps}(q)| \geq MinPts$



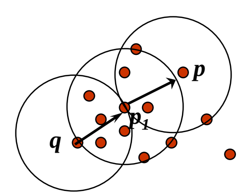
MinPts = 5

Eps = 1 cm

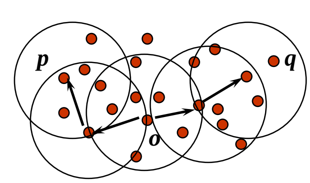
5

Density-Reachable and Density-Connected

- Density-reachable:
  - A point *p* is density-reachable from a point *q* w.r.t. *Eps*, *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



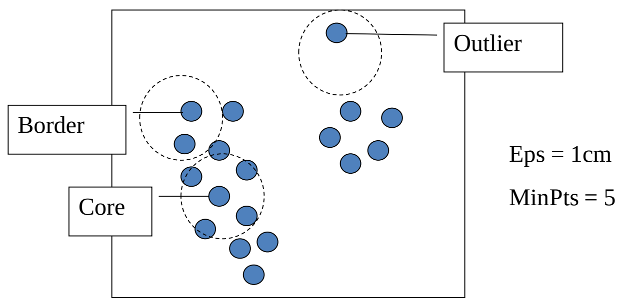
- Density-connected
  - A point *p* is density-connected to a point *q* w.r.t. *Eps*, *MinPts* if there is a point *o* such that both, *p* and *q* are density-reachable from *o* w.r.t. *Eps* and *MinPts*



6

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

- Relies on a *density-based* notion of cluster: A *cluster* is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise



7

DBSCAN: The Algorithm

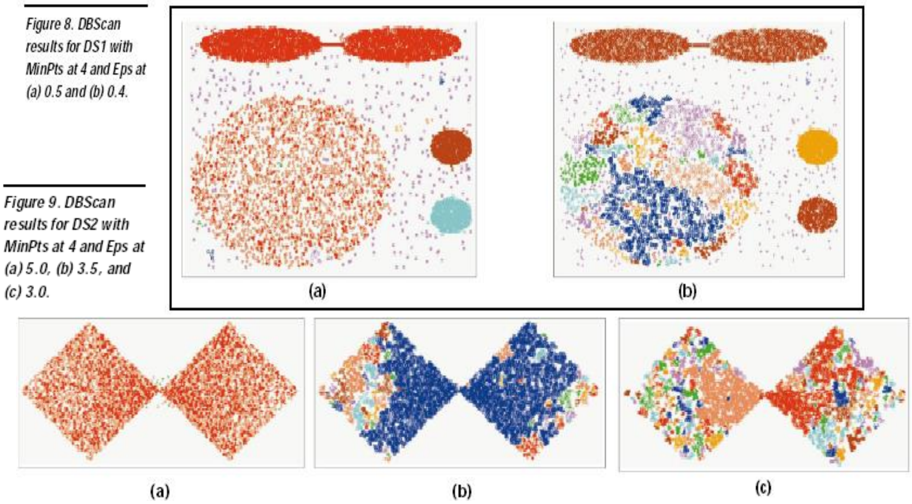
- Arbitrary select a point *p*
- Retrieve all points density-reachable from *p* w.r.t. *Eps* and *MinPts*
- If *p* is a core point, a cluster is formed
- If *p* is a border point, no points are density-reachable from *p* and DBSCAN visits the next point of the database
- Continue the process until all of the points have been processed
- *If a spatial index is used, the computational complexity of DBSCAN is O(n log n), where n is the number of database objects. Otherwise, the complexity is* $O(n^2)$

8

DBSCAN: Sensitive to Parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



DBSCAN online Demo:

https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

9

OPTICS: A Cluster-Ordering Method (1999)

- OPTICS: Ordering Points To Identify the Clustering Structure
  - Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99)
  - Produces a special order of the database wrt its density-based clustering structure
  - This cluster-ordering contains info equivalent to the density-based clusterings corresponding to a broad range of parameter settings
  - Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
  - Can be represented graphically or using visualization techniques

10

OPTICS: Some Extension from DBSCAN

- Index-based: k = # of dimensions, N: # of points
  - Complexity: O(N*log N)
- Core Distance of an object p: the smallest value ε such that the ε-neighborhood of p has at least MinPts objects

  Let $N_\varepsilon(p)$: ε-neighborhood of p, ε is a distance value

  $\text{Core-distance}_{\varepsilon,\, MinPts}(p)$ = Undefined if $card(N_\varepsilon(p))$ < MinPts

  MinPts-distance(p), otherwise
- Reachability Distance of object p from core object q is the min radius value that makes p density-reachable from q

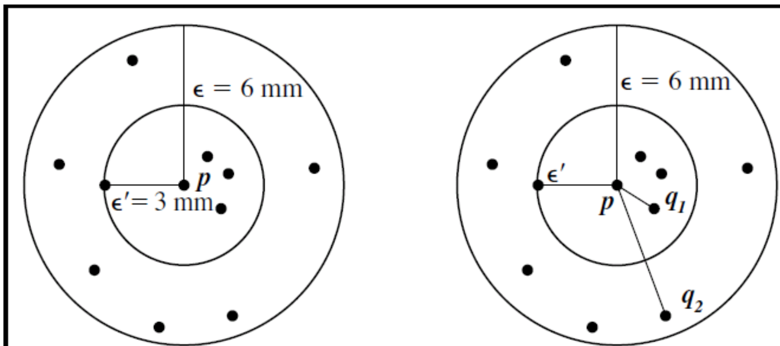
  $\text{Reachability-distance}_{\varepsilon,\, MinPts}(p, q)$ =

  Undefined if q is not a core object

  max(core-distance(q), distance (q, p)), otherwise

11

Core Distance & Reachability Distance



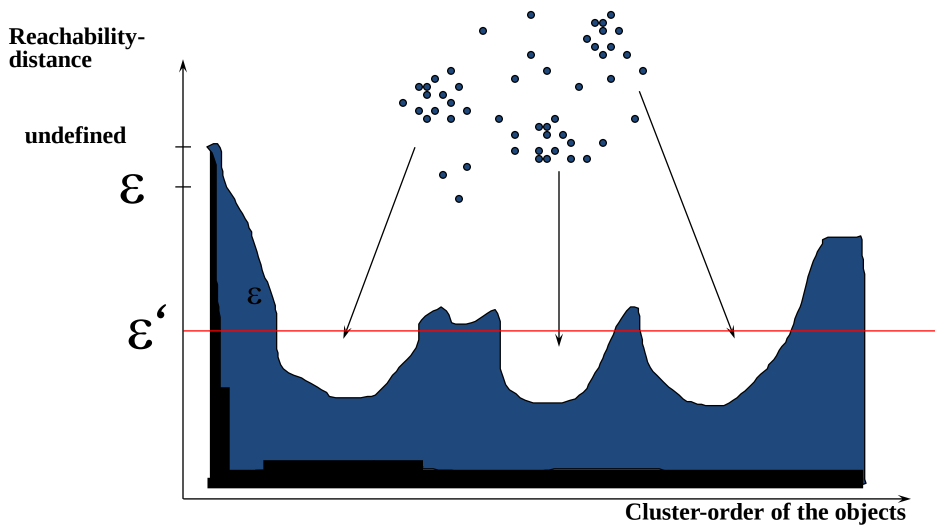
Core-distance of p

Reachability-distance $(p, q_1) = \epsilon' = 3$ mm

Reachability-distance $(p, q_2) = d(p, q_2)$

Figure 10.16: OPTICS terminology. Based on [ABKS99].

12



Density-Based Clustering: OPTICS & Applications

demo: http://www.dbs.informatik.uni-muenchen.de/Forschung/KDD/Clustering/OPTICS/Demo