Data Mining

4. Cluster Analysis

4.5 Density-Based Methods

Spring 2010

Instructor: Dr. Masoud Yaghini

Outline

- Introduction
- DBSCAN Algorithm
- OPTICS Algorithm
- DENCLUE Algorithm
- References

Introduction

Introduction

- **Density-based clustering methods**
  - To discover clusters with arbitrary shape, density-based clustering methods have been developed.
  - These typically regard clusters as dense regions of objects in the data space that are separated by regions of low density (representing noise).

Introduction

- **Density-based clustering algorithms**
  - **DBSCAN** grows clusters according to a density-based connectivity analysis.
  - **OPTICS** extends DBSCAN to produce a cluster ordering obtained from a wide range of parameter settings.
  - **DENCLUE** clusters objects based on a set of density distribution functions.

DBSCAN Algorithm

DBSCAN Algorithm

- **DBSCAN Algorithm**
  - stands for **Density-Based Spatial Clustering of Applications with Noise**
  - It is a density-based clustering algorithm.
  - The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise.
  - It defines a cluster as a maximal set of **density-connected points**.

Definitions

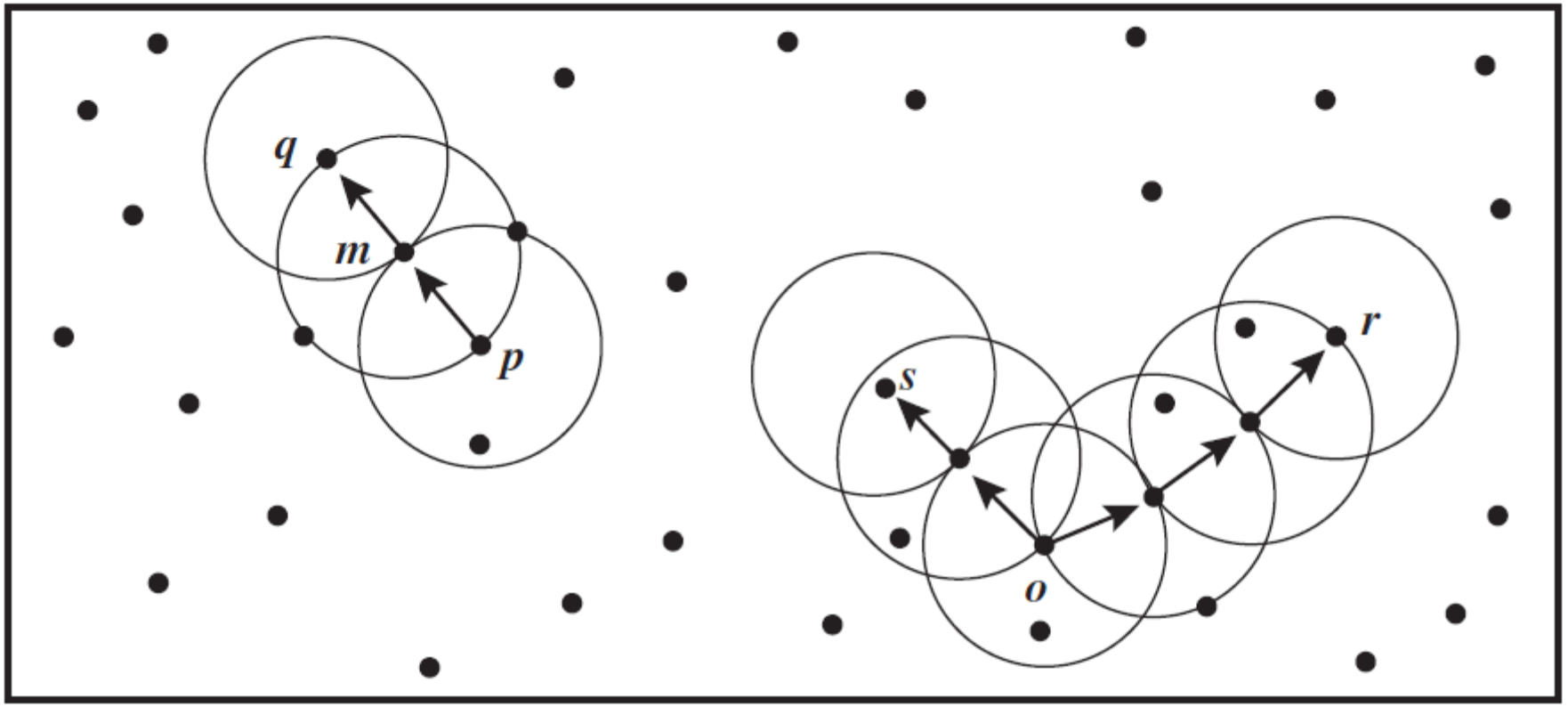
- **ε-Neighborhood of an object**
  - The neighborhood within a radius e of a given object is called the **ε-neighborhood of the object**.
- **Core object**
  - If the ε-neighborhood of an object contains at least a minimum number, *MinPts,* of objects, then the object is called a **core object**.
- **Directly density-reachable objects**
  - Given a set of objects, *D*, we say that an object *p* is **directly density-reachable** from object *q* if *p* is within the ε-neighborhood of *q*, and *q* is a **core object**.

Definitions

- **Indirectly Density-reachable objects**
  - An object p is **indirectly density-reachable** from object q,
  - if there is a chain of objects $p_1, \ldots, p_n$, where $p_1 = q$ and $p_n = p$ such that p_{i+1} is **directly density-reachable** from p_i, for $1 \leq i \leq n$.
- **Indirectly Density-connected objects**
  - An object p is **indirectly density-connected** to object q, if there is an object o such that both p and q are **density-reachable** from o.

Example: Density-reachability and density connectivity

- A given ε represented by the radius of the circles, and, say, let *MinPts* = 3.



Example: Density-reachability and density connectivity

- **Core objects**
  - *m*, *p*, *o*, and *r* are core objects because each is in an ε-neighborhood containing at least three points.
- **Directly density-reachable objects**
  - *q* is directly density-reachable from *m*.
  - *m* is directly density-reachable from *p* and vice versa.

Example: Density-reachability and density connectivity

- **Indirectly density-reachable objects**
  - *q* is indirectly density-reachable from *p* because *q* is directly density-reachable from *m* and *m* is directly density-reachable from *p*.
  - However, *p* is not indirectly density-reachable from *q* because *q* is not a core object.
  - Similarly, *r* and *s* are indirectly density-reachable from *o*, and *o* is indirectly density-reachable from *r*.
- **Indirectly Density-connected objects**
  - *o*, *r*, and *s* are all indirectly density-connected

Definitions

- **A density-based cluster**
  - **A density-based cluster** is a set of density-connected objects that is maximal with respect to density-reachability.
  - Every object not contained in any cluster is considered to be **noise**.

DBSCAN

- DBSCAN searches for clusters by checking the ε-neighborhood of each point in the database.
- If the ε-neighborhood of a point p contains at least *MinPts*, a new cluster with p as a **core object** is 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.

DBSCAN Algorithm

- The **computational complexity** of DBSCAN is $O(n^2)$, where n is the number of database objects.
- With appropriate settings of the user-defined parameters ε and *MinPts*, the algorithm is effective at finding arbitrary-shaped clusters.

OPTICS Algorithm

OPTICS Algorithm

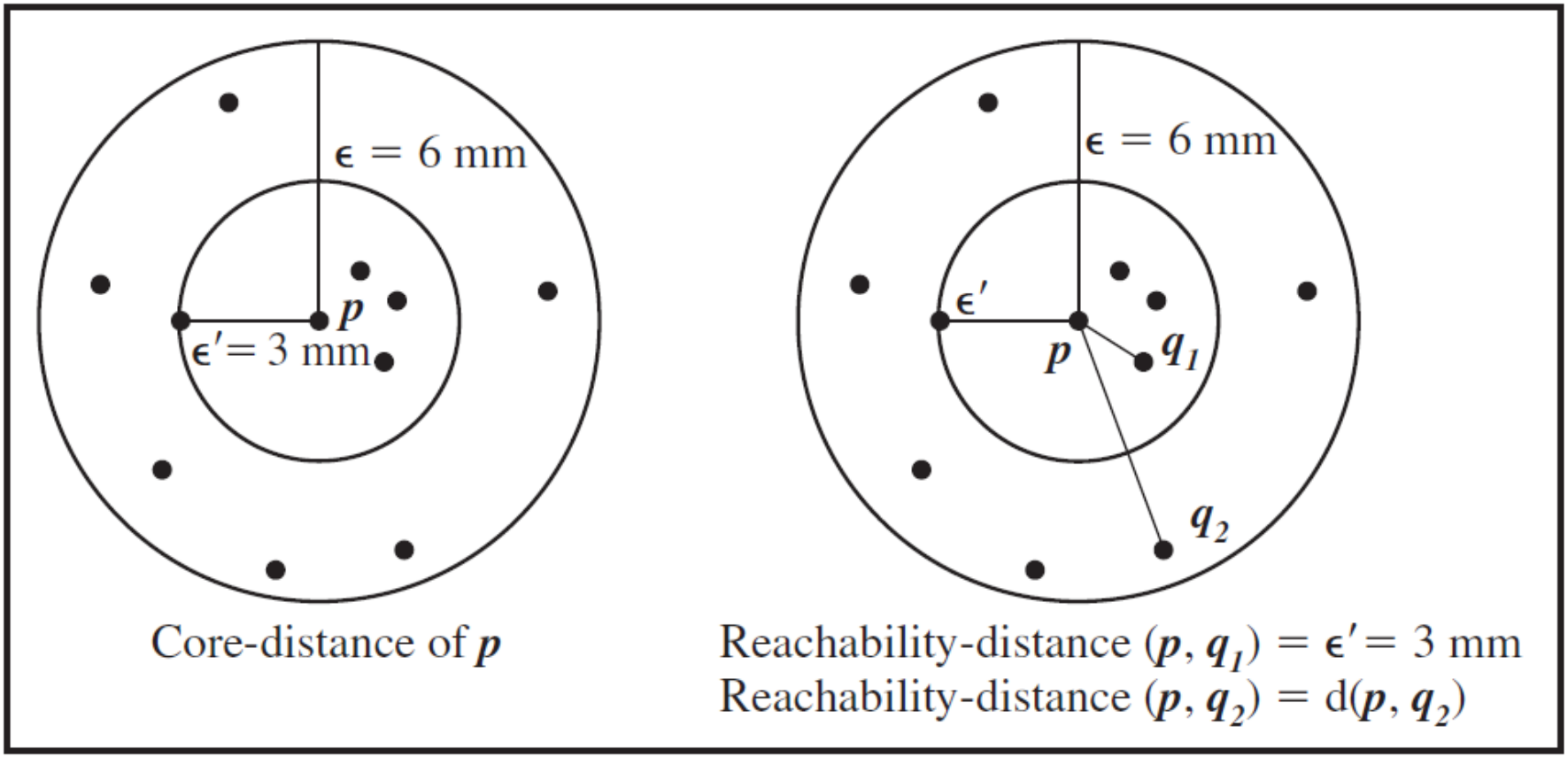
- **OPTICS Algorithm**
  - Stands for **Ordering Points to Identify the Clustering Structure**
  - OPTICS produces a set or ordering of density-based clusters
  - It constructs the different clusterings simultaneously
  - The objects should be processed in a specific order.
  - This order selects an object that is density-reachable with respect to the lowest ε value so that clusters with higher density (lower ε) will be finished first.
  - Based on this idea, two values need to be stored for each object—***core-distance*** and ***reachability-distance***

Definitions

- **Core-distance of an object**
  - The core-distance of an object p is the smallest ε′ value that makes *p* a core object. If *p* is not a core object, the core-distance of *p* is undefined.
- **Reachability-distance of an object**
  - The reachability-distance of an object *q* with respect to another object *p* is the greater value of the core-distance of *p* and the Euclidean distance between *p* and *q*.
  - If *p* is not a core object, the reachability-distance between *p* and *q* is undefined.

OPTICS Algorithm

- OPTICS terminology



OPTICS Algorithm

- The **OPTICS** algorithm creates an ordering of the objects in a database,
- **OPTICS** additionally storing the core-distance and a suitable reachability-distance for each object.
- An algorithm was proposed to extract clusters based on the ordering information produced by OPTICS.
- Such information is sufficient for the extraction of all density-based clusterings with respect to any distance ε′ that is smaller than the distance ε used in generating the order.

DENCLUE Algorithm

DENCLUE Algorithm

- **DENCLUE Algorithm**
  - DENCLUE stands for **DENsity-based CLUstEring**
  - It is a clustering method based on density distribution functions
- **DENCLUE** is built on the following ideas:
  - (1) the influence of each data point can be formally modeled using a mathematical function, called an **influence function**
  - (2) the overall density of the data space are the sum of the influence function applied to all data points
  - (3) clusters can then be determined mathematically by identifying **density attractors**, where density attractors are local maxima of the overall density function.

DENCLUE Algorithm

- **Influence function**
  - Let x and y be objects or points in F^d, a d-dimensional input space.
  - The influence function of data object y on x is a function:

$$f_B^y(x) = f_B(\boldsymbol{x}, \boldsymbol{y})$$

  - It can be used to compute a *square wave influence function,*

$$f_{Square}(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} 0 & if\ d(\boldsymbol{x}, \boldsymbol{y}) > \sigma \\ 1 & otherwise, \end{cases}$$

  - or a *Gaussian influence function,*

$$f_{Gauss}(\boldsymbol{x}, \boldsymbol{y}) = e^{-\frac{d(\boldsymbol{x}, \boldsymbol{y})^2}{2\sigma^2}}$$

  - σ is a threshold parameter

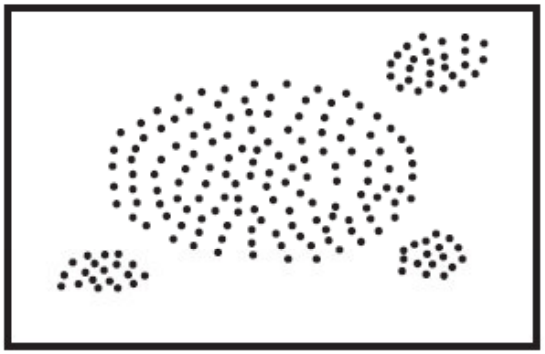
DENCLUE Algorithm

- **Density function**
  - The density function at an object or point x is defined as the sum of influence functions of all data points.
  - That is, it is the total influence on $\boldsymbol{x}$ of all of the data points.
  - Given n data objects, the density function at x is defined as

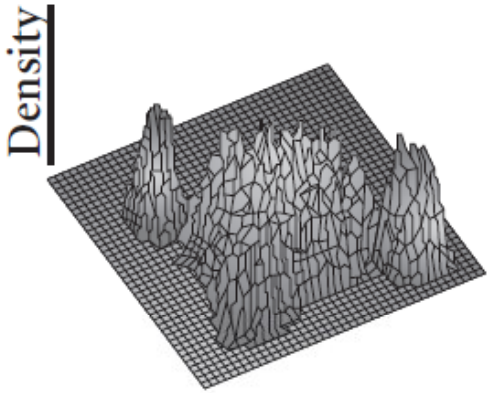
$$f_B^D(\boldsymbol{x}) = \sum_{i=1}^{n} f_B^{x_i}(\boldsymbol{x}) = f_B^{x_1}(\boldsymbol{x}) + f_B^{x_2}(\boldsymbol{x}) + \cdots + f_B^{x_n}(\boldsymbol{x})$$

DENCLUE Algorithm

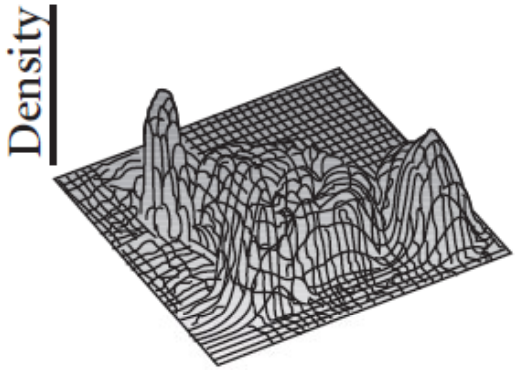
- Possible density functions for a 2-D data set.



(a) Data Set



(b) Square Wave



(c) Gaussian

DENCLUE Algorithm

- From the density function, we can define the **density attractor**, the local maxima of the overall density function.
- A hill-climbing algorithm guided by the gradient can be used to determine the density attractor of a set of data points.

References

References

- J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Elsevier Inc. (2006). (Chapter 7)

The end