

DENCLUE

KHU, M.M.Pedram, pedram@khu.ac.ir

Data Mining, Fall 2011

1

DENCLUE: using density functions

- ❖ DENSity-based CLUstEring: by Hinneburg & Keim (KDD'98)
- ❖ Major features
 - ▶ Solid mathematical foundation
 - ▶ Good for data sets with large amounts of noise
 - ▶ Allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets
 - ▶ Significant faster than existing algorithm (faster than DBSCAN by a factor of up to 45)
 - ▶ But needs a large number of parameters

▶ 2

KHU, M.M.Pedram, pedram@khu.ac.ir

Data Mining, Fall 2011

2

Denclue: Technical Essence

- ❖ Uses grid cells but only keeps information about grid cells that do actually contain data points and manages these cells in a tree-based access structure.
- ❖ DENCLUE is based on the following concepts:
 - ▶ Influence function
 - ▶ Overall density of the data space
 - ▶ Density attractors

▶ 3

KHU, M.M.Pedram, pedram@knu.ac.ir

Data Mining, Fall 2011

3

Definitions

- ❖ The **influence function** $f^y(x)$ for a point $y \in D$ (data space) at point x is a positive function that decays to zero as x “moves away” from y ($d(x,y) \rightarrow \infty$).
- ❖ Influence function describes the impact of a data point within its neighborhood.
- ❖ Typical examples are:

$$f^y(x) = \begin{cases} 1, & \text{if } d(x,y) < \sigma \\ 0, & \text{otherwise} \end{cases}$$

and

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

where σ is a user-defined function/parameter.

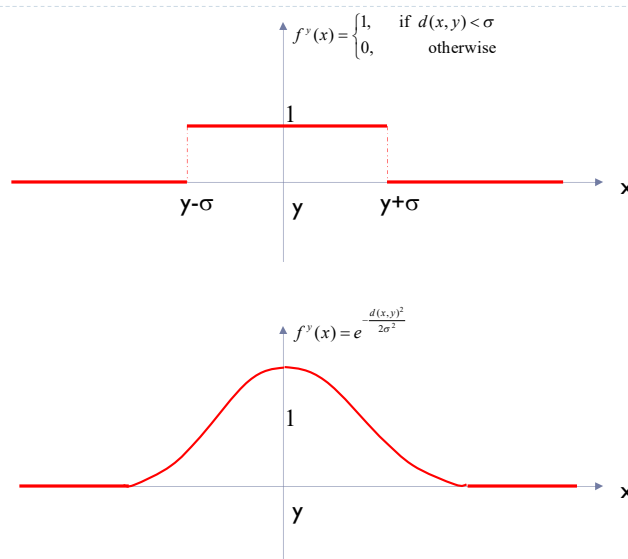
▶ 4

KHU, M.M.Pedram, pedram@knu.ac.ir

Data Mining, Fall 2011

4

Definitions



► 5

KHU, M.M.Pedram, pedram@knu.ac.ir

Data Mining, Fall 2011

5

Definitions

- ❖ The **density function** at x based on a *data space* of N points; i.e. $D = \{x_1, \dots, x_N\}$; is defined as the sum of the influence function of all data points at \underline{x} :

$$f^D(x) = \sum_{i=1}^N f^{x_i}(x)$$

The goal of the definition:

- i. Identify all “**significant**” local maxima, x_j^* , $j=1, \dots, m$ of $f^D(x)$
- ii. Create a cluster C_j for each x_j^* and assign to C_j all points of D that lie within the “**region of attraction**” of x_j^* .

► 6

KHU, M.M.Pedram, pedram@knu.ac.ir

Data Mining, Fall 2011

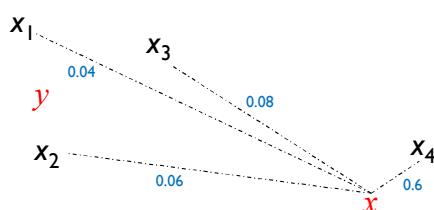
6

Definitions

Example: Density Computation

$$D = \{x_1, x_2, x_3, x_4\}$$

$$f_{\text{Gaussian}}^D(x) = \text{influence}(x_1) + \text{influence}(x_2) + \text{influence}(x_3) + \text{influence}(x_4) \\ = 0.04 + 0.06 + 0.08 + 0.6 = 0.78$$



Remark: the density value of y would be larger than the one for x

► 7

KHU, M.M.Pedram, pedram@khu.ac.ir

Data Mining, Fall 2011

7

Definitions

❖ For a Gaussian influence function:

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

Density function is:

$$f_{\text{Gaussian}}^D(x) = \sum_{i=1}^N e^{-\frac{d(x,x_i)^2}{2\sigma^2}}$$

► 8

KHU, M.M.Pedram, pedram@khu.ac.ir

Data Mining, Fall 2011

8

Definitions

❖ Gradient (The steepness of a slope):

The gradient of function $f^D(x)$ is defined as:

$$\nabla f^D(x, x_i) = \sum_{i=1}^N (x_i - x) \cdot f^{x_i}(x)$$

❖ Why is gradient defined in this way?

❖ Example:

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

$$f^D_{\text{Gaussian}}(x) = \sum_{i=1}^N e^{-\frac{d(x, x_i)^2}{2\sigma^2}}$$

$$\nabla f^D_{\text{Gaussian}}(x, x_i) = \sum_{i=1}^N (x_i - x) \cdot e^{-\frac{d(x, x_i)^2}{2\sigma^2}}$$

▶ 9

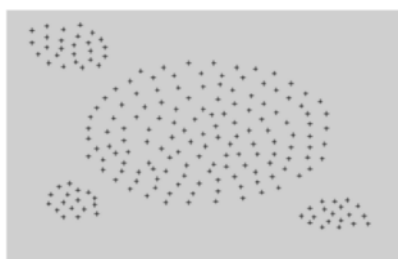
KHU, M.M.Pedram, pedram@knu.ac.ir

Data Mining, Fall 2011

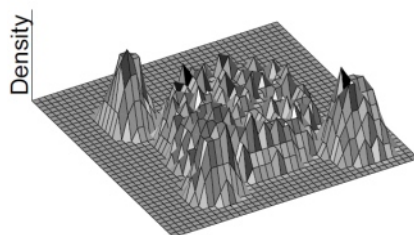
9

Definitions

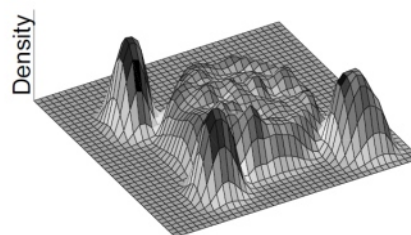
Dataset



Where are the cluster centers?



Based on Square Wave influence function



Based on Gaussian function

▶ 10

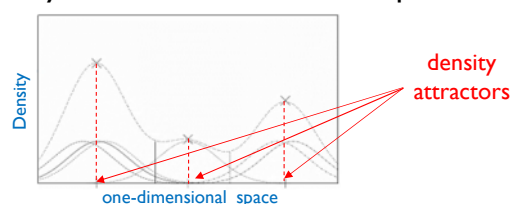
KHU, M.M.Pedram, pedram@knu.ac.ir

Data Mining, Fall 2011

10

Definitions

- ❖ **Density attractors** are local maxima of the overall density function $f^D(x)$.
- ❖ Clusters can then be determined mathematically by identifying density attractors.
- ❖ A hill-climbing algorithm guided by the gradient can be used to determine the density attractor of a set of data points.



▶ 11

KHU, M.M.Pedram, pedram@knu.ac.ir

Data Mining, Fall 2011

11

Definitions

- ❖ A point x is **density-attracted** to a density attractor x^* , if there exists a set of points x_0, x_1, \dots, x_k such that $x_0 = x, x_k = x^*$ and the gradient of x_{i-1} is in the direction of x_i for $0 < i < k$.
or iff $\exists k \in \mathbb{N} : d(x^k, x^*) \leq \varepsilon$ with

$$x^0 = x, x^i = x^{i-1} + \delta \cdot \frac{\nabla f_B^D(x^{i-1})}{\|\nabla f_B^D(x^{i-1})\|}$$

▶ 12

KHU, M.M.Pedram, pedram@knu.ac.ir

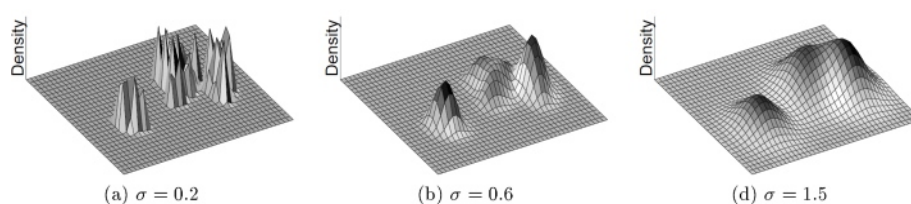
Data Mining, Fall 2011

12

Definitions

Center-Defined Cluster

- ❖ A center-defined cluster (w.r.t. to σ, ξ) for a density attractor x^* is a subset $C \subseteq D$, with $x \in C$ being density-attracted by x^* and $f^D(x) \geq \xi$.
- ❖ **Outlier:** Point $x \in D$ is called outlier if it is density-attracted by a local maximum x_o^* with $f^D(x_o^*) < \xi$.



▶ 13

KHU, M.M.Pedram, pedram@khu.ac.ir

Data Mining, Fall 2011

13

کوه‌ها و ستیغ‌ها



▶ 14

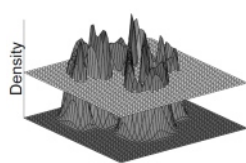
M.M. Pedram, pedram@khu.ac.ir

14

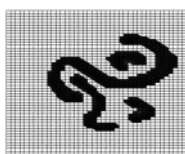
Definitions

Arbitrary-Shape Cluster

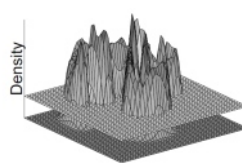
- ❖ An arbitrary-shape cluster (w.r.t. to σ, ξ) for a set of density attractors X is a subset $C \subseteq D$, where
1. $\forall x \in C \quad \exists x^* \in X: f^D(x^*) \geq \xi$, x is density-attracted to x^* , and
 2. $\forall x_1^*, x_2^* \in X: \exists$ a path P from x_1^* to x_2^* with $\forall p \in P: f^D(p) \geq \xi$.



(a) $\xi = 2$



(b) $\xi = 2$



(c) $\xi = 1$



(d) $\xi = 1$

► 15

KHU, M.M.Pedram, pedram@knu.ac.ir

Data Mining, Fall 2011

15

Note 1

- ❖ Note that the number of clusters found by DENCLUE varies depending on σ, ξ .

► 16

KHU, M.M.Pedram, pedram@knu.ac.ir

Data Mining, Fall 2011

16

Two clarifications

1. The **region of attraction** of x_j^* is defined as the set of points in $x \in \mathcal{R}^l$ such that if a “hill-climbing” (such as the steepest ascent) method is applied, initialized by x , it will terminate arbitrarily close to x_j^* .
 2. A **local maximum** is considered as **significant** if $f^D(x_j^*) \geq \xi$ (ξ is a user-defined parameter).
- ❖ Only points of the data set which are close to x actually contribute to the density. This leads to **approximation of $f^D(x)$** , i.e. **local density function**

$$\hat{f}^D(x) = \sum_{x_i \in \text{Near}(x)} f^{x_i}(x)$$

where $\text{Near}(x)$ is the set of points in D that lie “close” to x .

► 17

KHU, M.M.Pedram, pedram@khu.ac.ir

Data Mining, Fall 2011

17

DENCLUE algorithm

- ❖ *Preclustering phase*
- ❖ *Main phase*

► 18

KHU, M.M.Pedram, pedram@khu.ac.ir

Data Mining, Fall 2011

18

DENCLUE algorithm

❖ Preclustering phase (identification of regions dense in points of D)

- A map of the relevant portion of the data space is constructed. The map is used to speed up the calculation of the density function which requires to efficiently access neighboring portions of the data space.

❖ Main phase (clustering)

- The second step is the actual clustering step, in which the algorithm identifies the density-attractors and the corresponding density attracted points.

▶ 19

KHU, M.M.Pedram, pedram@khu.ac.ir

Data Mining, Fall 2011

19

DENCLUE algorithm

❖ Preclustering phase (identification of regions dense in points of D)

- Apply a d -dimensional grid of edge-length 2σ in the \mathbb{R}^d space.
 - The hyper-cubes are numbered depending on their relative position from a given origin. In this way, the populated hyper-cubes (containing d -dimensional data points) can be mapped to one-dimensional keys and stored in tree. The keys of the populated cubes can be efficiently stored in a randomized search-tree or a B⁺-tree.
- Determine the set C_p of the hyper-cubes that contain **at least** one point of D ; (C_p = populated cube)
- Connect neighboring populated cubes :

Two cubes $c_1, c_2 \in C_{sp}$ are connected if

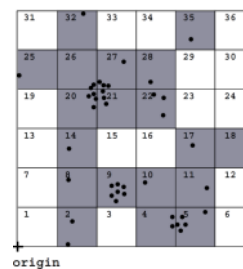
$$d(\text{mean}(c_1), \text{mean}(c_2)) \leq 4\sigma$$

this normally take $O(C_p^2)$ time for all cubes.

To speed up, do as follow:

 - Determine the set $C_{sp} (\subset C_p)$ that contains the “highly populated” cubes of C_p , i.e., cubes that contain at least ζ_c (outlier bound) points of D ; (ζ_c = a second outlier-bound to reduce the time needed for connecting the cubes):

$$C_{sp} = \{c \in C_p \mid N_c > \zeta_c\}$$



▶ 20

KHU, M.M.Pedram, pedram@khu.ac.ir

Data Mining, Fall 2011

20

DENCLUE algorithm

- In general, the number of highly populated cubes C_{sp} is much smaller than C_p , especially in high-dimensional space.
- The time needed for connecting the highly populated cubes with their neighbors is then $O(\|C_{sp}\| \cdot \|C_p\|)$ with $\|C_{sp}\| \ll \|C_p\|$. The cardinality of $\|C_{sp}\|$ depends on ξ_c .
 - A good choice for ξ_c is $\xi_c = \xi/2d$, since in high-dimensional spaces the clusters are usually located on lower-dimensional hyperplanes.

Note:

- ❖ The data structure generated in *Preclustering phase* has the following properties:
 1. The time to access the cubes for an arbitrary point is $O(\log(C_p))$.
 2. The time to access the relevant portion around a given cube (the connected neighboring cubes) is $O(1)$.

► 21

KHU, M.M.Pedram, pedram@knu.ac.ir

Data Mining, Fall 2011

21

DENCLUE algorithm

❖ Main phase

- Determine the set C_r that contains:
 - the highly populated cubes, and
 - the cubes that have at **least** one connection with a highly populated cube.
$$C_r = C_{sp} \cup \{c \in C_p \mid \exists c_s \in C_{sp} \text{ and } \exists \text{connection}(c_s, c)\}$$
- For each point x in a cube $c \in C_r$, determine $Near(x)$ as the set of points that belong to cubes c_j in C_r such that the mean values of c_j s lie at distance **less** than $\lambda\sigma$ from x (typically $\lambda=4$).
- Determine local density-function for each point in the cubes of C_r based on the following approximation. Thus, local gradient can be computed:

$$\hat{f}^D(x) = \sum_{x_i \in Near(x)} e^{-\frac{d(x,y)^2}{2\sigma^2}}$$
- Determine the density-attractors for each point in the cubes of C_r by a hill-climbing procedure based on the local density function and its gradient.
 - Note that after determining the density-attractor x^* for a point x and $f^D(x) \geq \xi$, the point x is classified and attached to the cluster belonging to x^* .

► 22

KHU, M.M.Pedram, pedram@knu.ac.ir

Data Mining, Fall 2011

22

DENCLUE algorithm

For each point x in a cube $c \in C_r$

Apply a hill climbing method starting from x and let x^* be the local maximum to which the method converges.

If x^* is a significant local maximum ($f^D(x^*) \geq \xi$) then

If a cluster C associated with x^* has already been created, then

x is assigned to C

Else

Create a cluster C associated with x^*

Assign x to C

End if

End if

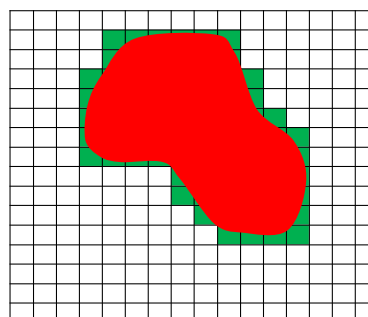
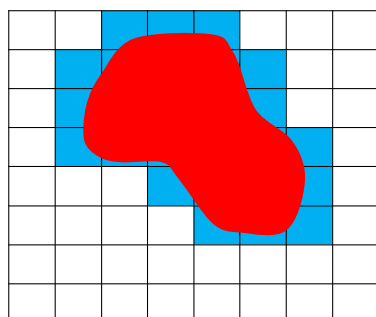
End for

► 23

KHU, M.M.Pedram, pedram@knu.ac.ir

Data Mining, Fall 2011

23



► 24

KHU, M.M.Pedram, pedram@knu.ac.ir

Data Mining, Fall 2011

24

Remarks

- ❖ Shortcuts allow the assignment of points to clusters, without having to apply the hill-climbing procedure.
- ❖ DENCLUE is able to detect **arbitrarily** shaped clusters.
- ❖ The algorithm deals with noise very satisfactory.
- ❖ The **worst-case time complexity** of DENCLUE is $O(N \log_2 N)$.
- ❖ Experimental results indicate that the **average time complexity** is $O(\log_2 N)$.
- ❖ It works efficiently with high-dimensional data.
- ❖ DENCLUE needs at least 3 parameters to be determined, i.e. σ , ξ , ξ_c .