

Data Mining

- Clustering II

INFS4293 / INFS7203 Data Mining

1

How to avoid becoming a statistic in the ITEE MISCONDUCT REPORT

"Last semester, ITEE investigated 60 students for possible misconduct (involving cheating, plagiarising or copying work). You need to know about what is unacceptable so you don't become a statistic, so please pay close attention to the advice below and take the time to check the web links for more information."

"As a necessary part of this undertaking, we are obliged to identify and investigate cases of possible academic misconduct within the School. A report on our activities in this regard last semester, which gives you an indication of our diligence in ensuring students' work is their own, is provided below. **This is done to raise your awareness of the implications of copying or plagiarising work, and to take the opportunity to refer you to information about academic misconduct** (which includes plagiarising, copying or sharing work, not referencing your work, or colluding on your work)."

INFS4293 / INFS7203 Data Mining

2

Clustering Methods

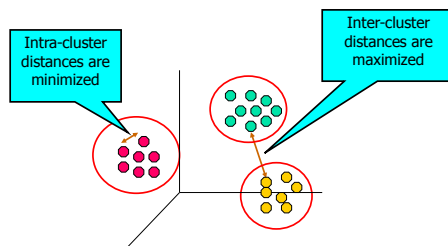
- K-means
- Agglomerative Hierarchical Clustering
- Graph-based Clustering
- K-medoids
- Divisive Hierarchical Clustering
- Density-based Clustering

INFS4293 / INFS7203 Data Mining

3

Distance Measures

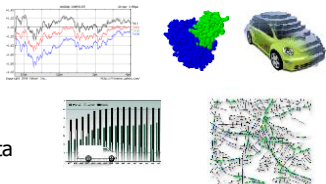
- Each clustering problem is based on some kind of "distance" between points.



4

Complex Data Types

- Complex data
 - Text Data
 - Temporal data
 - Spatial data
 - Spatial-temporal data
 - Multimedia data
- Not always so simple!
 - E.g., Cluster these numbers into 3 groups..
18, 22, 25, 42, 27, 43, 33, 35, 56, 28



5

Euclidean Vs. Non-Euclidean

- The **Euclidean space** has some number of real-valued dimensions and "dense" points.
 - The Euclidean distance is based on the **locations** of points in such a space.
- A **Non-Euclidean distance** is based on **properties** of points, but not their "location" in a space.

6

Distance Measure

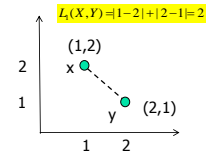
- d is a **distance measure** if it is a function from pairs of points to real values such that:
 - $d(x,y) \geq 0$. (**non negativity**)
 - $d(x,y) = 0$ iff $x = y$.
 - $d(x,y) = d(y,x)$. (**symmetry**)
 - $d(x,y) \leq d(x,z) + d(z,y)$ (**triangle inequality**).

7

Distance Measures

- L_1 (1-norm)

$$L_1(X, Y) = \sum_{i=1}^{\dim} |X_i - Y_i|$$



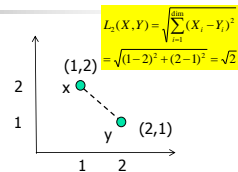
- sum of the differences in each dimension.
- Manhattan distance, city block distance**

8

Distance Measures

- L_2 (2-norm)

$$L_2(X, Y) = \sqrt{\sum_{i=1}^{\dim} (X_i - Y_i)^2}$$



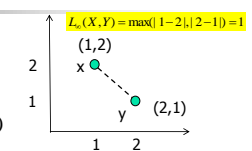
- square root of the sum of the squares of the differences between x and y in each dimension.
- The most common notion of "distance."
- Euclidean Distance**

9

Distance Measures

- L_∞ norm

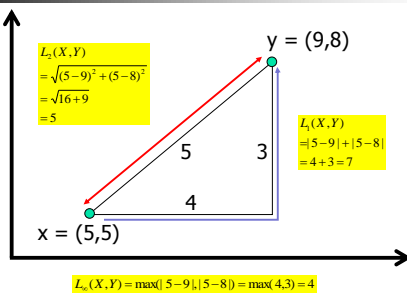
$$L_\infty(X, Y) = \max_{i=1}^{\dim} (|X_i - Y_i|)$$



- the maximum of the differences between x and y in any dimension.

10

An Example



11

Quiz!

- $X=(1,0,5), Y=(2,4,9)$
- $L_1(X,Y)=?$
- $L_2(X,Y)=?$
- $L_\infty(X,Y)=?$



$X=(1,0,5), Y=(2,4,9)$

- $L_1(X,Y)=|1-2|+|0-4|+|5-9|=9$
- $L_2(X,Y)=\text{Sqrt}((1-2)^2+(0-4)^2+(5-9)^2)$
 $=\text{Sqrt}(33)=5.74$
- $L_\infty(X,Y)=\max(|1-2|,|0-4|,|5-9|)=4$

Non-Euclidean Distances

- *Jaccard distance*
 - Binary vectors
- *Cosine distance*
 - angle between vectors from the origin to the points in question.
- *Edit distance*
 - number of edit operations to change one string into another.

Jaccard Distance

- Distance between binary vectors
- Jaccard Measure (JM)
 - ratio of sizes of intersection and union

$$JM(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

- Jaccard Distance: $JD(x,y) = 1 - JM$

X	1	0	1	1	1	JM=3/4 JD=1-3/4=1/4
Y	1	0	0	1	1	
Intersection	1			1	1	← 3
Union	1	1	1	1	1	← 4

Why J.D. Is a Distance Measure

- $JD(x,x) = 0$
 - $x \cap x = x \cup x$.
- $JD(x,y) = JD(y,x)$
 - Union and intersection are symmetric.
- $JD(x,y) \geq 0$
 - $|x \cap y| \leq |x \cup y|$.
- $JD(x,y) \leq JD(x,z) + JD(z,y)$
 - Trickier – optional homework!

Quiz!

- $X=(1,0,0,0,1,0,1)$
- $Y=(0,0,0,1,1,1,1)$
- $JM(X,Y)=?$
- $JC(X,Y)=?$



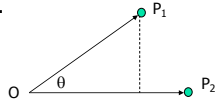
$X=(1,0,0,0,1,0,1), Y=(0,0,0,1,1,1,1)$

X	1	0	0	0	1	0	1
Y	0	0	0	1	1	1	1
Intersection					1	1	
Union	1			1	1	1	1

JM=2/5;
 JC = 1-2/5 = 3/5

Cosine Distance

- Measure the distance between two vectors
 - Think of a point as a vector from the origin (0,0,...,0) to its location.
 - Two points' vectors make an angle, whose cosine is the normalized dot-product of the vectors: $\frac{P_1 \cdot P_2}{|P_2| |P_1|}$.

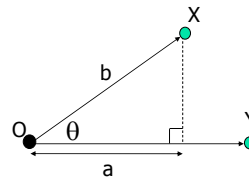


- Compare the documents in text mining
 - Will be discussed in text mining ...

19

Cosine Distance

dot product $X_1 \times Y_1 + X_2 \times Y_2 + \dots$



$$\text{CosineD}(X, Y) = \theta = \arccos\left(\frac{X \cdot Y}{\|X\| \times \|Y\|}\right)$$

Magnitude, L_2

dot product	
X	1 0 3 2
Y	0 1 2 1
	0 0 6 2
	Σ
	8

$$\text{Cosine}(\theta) = \frac{a}{b} = \frac{X \cdot Y / \|Y\|}{\|X\|} = \frac{X \cdot Y}{\|X\| \times \|Y\|}$$

[Ref: http://en.wikipedia.org/wiki/Dot_product]

20

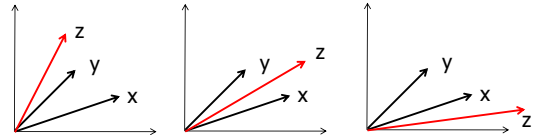
Why C.D. Is a Distance Measure

- $d(x,x) = 0$ because $\arccos(1) = 0$.
- $d(x,y) = d(y,x)$ by symmetry.
- $d(x,y) \geq 0$ because angles are chosen to be in the range 0 to 180 degrees.

21

Why C.D. Is a Distance Measure

- Triangle inequality: $d(x,y) \leq d(x,z) + d(y,z)$
 - physical reasoning. If I rotate an angle from x to z and then from z to y, I can't rotate less than from x to y.



CASE 1) $C.D.(X,Y) \leq C.D.(X,Z) \leq C.D.(X,Z) + C.D.(Y,Z)$
 CASE 2) $C.D.(X,Y) = C.D.(X,Z) + C.D.(Y,Z)$
 CASE 2) $C.D.(X,Y) \leq C.D.(Y,Z) \leq C.D.(X,Z) + C.D.(Y,Z)$

22

Quiz!

- $X=(1,0,0,0,1,0,1)$
- $Y=(0,0,0,1,1,1,1)$
- $\text{CosineD}(X,Y)=?$

Note that, it can be any real number not only 0/1.

$$\text{Cosine}(\theta) = \frac{a}{b} = \frac{X \cdot Y / \|Y\|}{\|X\|} = \frac{X \cdot Y}{\|X\| \times \|Y\|}$$

dot product $X_1 \times Y_1 + X_2 \times Y_2 + \dots$

INFS4293 / INFS7203 Data Mining

23

$$X=(1,0,0,0,1,0,1), Y=(0,0,0,1,1,1,1)$$

X	1	0	0	0	1	0	1
Y	0	0	0	1	1	1	1
	0	0	0	0	1	0	1

$$X \cdot Y = 0+0+0+0+1+0+1=2$$

$$\|X\| = \text{Sqrt}(1^2+0^2+0^2+0^2+1^2+0^2+1^2) = \text{Sqrt}(3) = 1.732$$

$$\|Y\| = \text{Sqrt}(0^2+0^2+0^2+1^2+1^2+1^2+1^2) = \text{Sqrt}(4) = 2$$

$$\text{CosineD} = \arccos(2 / (1.732 \times 2)) = \arccos(0.58) = 54.5^\circ$$

Not required

INFS4293 / INFS7203 Data Mining

24

Edit Distance

- Extensively used to measure the distance between strings
- Edit Distance between two strings, X and Y, is defined as the **minimum** number of operations needed to transfer X to Y.
 - Insertion
 - Deletion
 - Substitution

ant ant ant
a(n)t an act
ED=1

Edit Distance Calculation

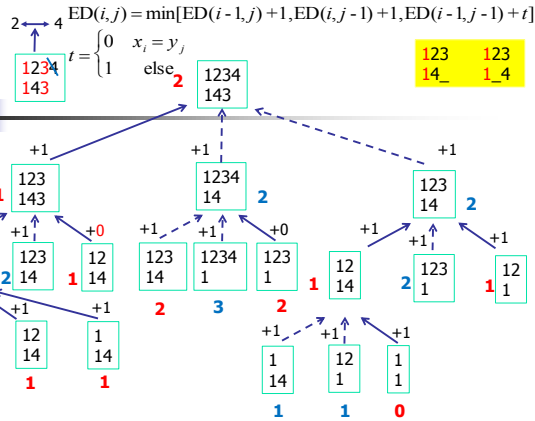
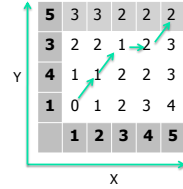
distance between first *i* letters of X and first *j* letters of string Y

$$ED(i, j) = \min[ED(i-1, j)+1, ED(i, j-1)+1, ED(i-1, j-1)+t]$$

$$t = \begin{cases} 0 & x_i = y_j \\ 1 & \text{ignore } X(i) \quad \text{ignore } Y(j) \quad \text{substitute } X(i) \& Y(j) \end{cases}$$

X: 12345
Y: 1435

EditDistance(X,Y)=2
12345
143 5



Why E.D. Is a Distance Measure

- $d(x,x) = 0$ because 0 edits suffice.
- $d(x,y) = d(y,x)$ because insert/delete are inverses of each other.
- $d(x,y) \geq 0$: no notion of negative edits.
- Triangle inequality:**
 - $ED(X,Z) + ED(Y,Z): X \rightarrow Z \rightarrow Y$
 - changing x to z and then to y is one way to change x to y .

Quiz!

- $x = abcde$
- $y = bcduve$
- $ED(x,y) = ?$

e					
v					
u					
d					
c					
b					
	a	b	c	d	e

Answer,

e	6	6	5	4	3
v	5	5	4	3	3
u	4	4	3	2	2
d	3	3	2	1	2
c	2	2	1	2	3
b	1	1	2	3	4
	a	b	c	d	e

ED=3

abcde
bcduve

Clustering Methods

- K-means
- Agglomerative Hierarchical Clustering
- Graph-based Clustering
- K-medoids
- Divisive Hierarchical Clustering
- Density-based Clustering

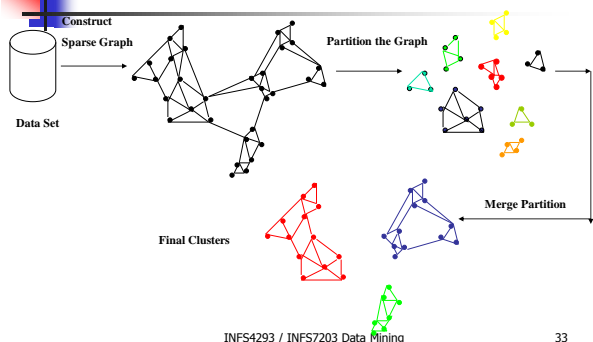
INFS4293 / INFS7203 Data Mining

31

Graph-Based Clustering

- CHAMELEON (Hierarchical clustering using dynamic modeling)
- Graph-Based clustering uses the proximity graph
- In the simplest case, clusters are connected components in the graph.

Overall Framework of CHAMELEON



INFS4293 / INFS7203 Data Mining

33

Chameleon: Steps

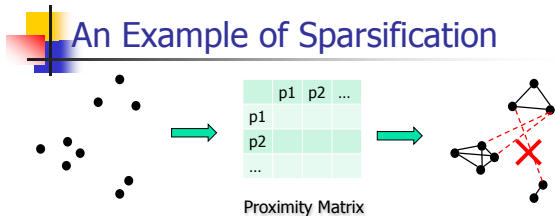
- **Preprocessing Step:**
Represent the Data by a Graph
 - Given a set of points, construct the k-nearest-neighbor (k-NN) graph to capture the relationship between a point and its k nearest neighbors
 - Concept of neighborhood is captured dynamically (even if region is sparse)
- **Phase 1:** Use a multilevel graph partitioning algorithm on the graph to find a large number of clusters of well-connected vertices
 - Each cluster should contain mostly points from one "true" cluster, i.e., is a sub-cluster of a "real" cluster

Chameleon: Steps ...

- **Phase 2:** Use Hierarchical Agglomerative Clustering to merge sub-clusters
 - Two clusters are combined if the resulting cluster shares certain properties with the **constituent** clusters
 - Two key properties used to model cluster similarity:
 - **Relative Interconnectivity:** Absolute interconnectivity of two clusters normalized by the internal connectivity of the clusters
 - **Relative Closeness:** Absolute closeness of two clusters normalized by the internal closeness of the clusters

Sparsification

- From proximity Matrix to a graph
 - Each node is connected to all others
 - The weight of the edges between any pair of nodes reflects their pairwise proximity
- K-nearest neighbour (KNN)
 - keep the connections to the KNN
 - Break all links that have a distance greater than a specified threshold.



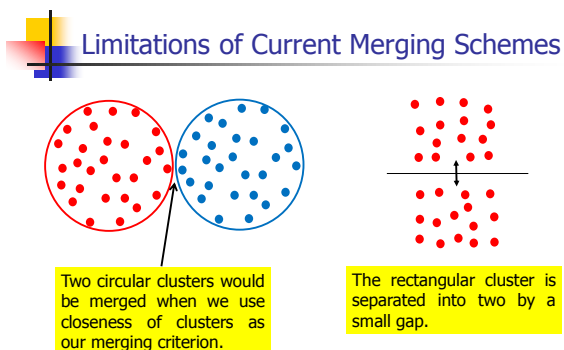
- ### Sparsification
- The amount of data that needs to be processed is drastically reduced
 - Sparsification can eliminate more than 99% of the entries in a proximity matrix
 - The amount of time required to cluster the data is drastically reduced
 - The size of the problems that can be handled is increased

- ### Sparsification
- Clustering may work better
 - Sparsification techniques keep the connections to the most similar (nearest) neighbors of a point while breaking the connections to less similar points.
 - The nearest neighbors of a point tend to belong to the same class as the point itself.
 - This reduces the impact of noise and outliers and sharpens the distinction between clusters.
 - Sparsification facilitates the use of graph partitioning algorithms
 - Further partition is required
 - Graph partitioning algorithms

- ### Merging
- Agglomerative Hierarchical clustering
 - Bottom-up
 - Which Clusters to Merge?
 - Existing merging schemes in hierarchical clustering algorithms are static in nature
 - MIN: merge two clusters based on their *closeness* (or minimum distance)
 - GROUP-AVERAGE: merge two clusters based on their average *connectivity*

INFS4293 / INFS7203 Data Mining

40



- ### Merging
- Adapt to the characteristics of the data set to find the natural clusters
 - Use a dynamic model to measure the similarity between clusters
 - Main properties are the *relative closeness* and *relative inter-connectivity* of the cluster
 - Two clusters are combined if the resulting cluster shares certain *properties* with the constituent clusters
 - The merging scheme preserves *self-similarity*

Relative Closeness

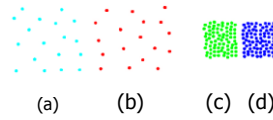
- Closeness:
 - minimum distance between two clusters
- Relative Closeness:
 - The absolute closeness of two clusters normalized by the internal closeness of the clusters.

$$RC = \frac{\bar{S}_{EC}(C_i, C_j)}{\frac{m_i}{m_i + m_j} \bar{S}_{EC}(C_i) + \frac{m_j}{m_i + m_j} \bar{S}_{EC}(C_j)}$$

Number of edges in C_i → m_i
 Average weight of the edges of cluster C_i → $\bar{S}_{EC}(C_i)$
 Average weight of the edges of cluster C_j → $\bar{S}_{EC}(C_j)$
 Average weight of the edges that connect cluster C_i and C_j → $\bar{S}_{EC}(C_i, C_j)$

Illustration of Relative Closeness

Which two clusters should be merged?



INFS4293 / INFS7203 Data Mining

44

Relative Interconnectivity

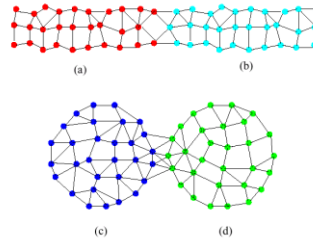
- Interconnectivity:
 - Average connectivity between two clusters
- Relative Interconnectivity:
 - The absolute interconnectivity of two clusters normalized by the internal interconnectivity of the clusters.

$$RC = \frac{EC(C_i, C_j)}{\frac{1}{2}(EC(C_i) + EC(C_j))}$$

Sum of the edges in cluster C_i → $EC(C_i)$
 Sum of the edges in cluster C_j → $EC(C_j)$
 The sum of the edges that connect cluster C_i and C_j → $EC(C_i, C_j)$

Illustration of Relative Interconnectedness

Which two clusters should be merged?

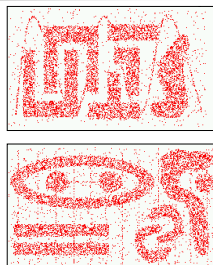


INFS4293 / INFS7203 Data Mining

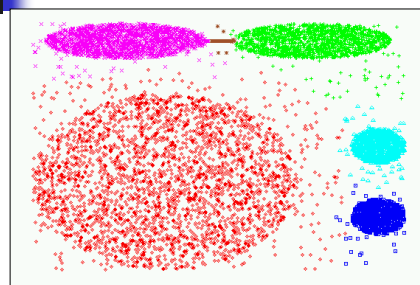
46

Chameleon

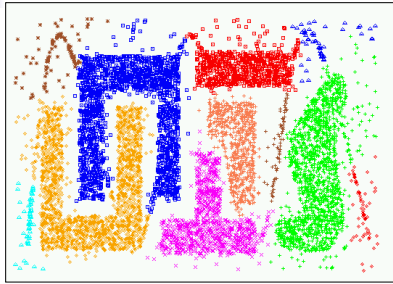
- One of the areas of application is **spatial data!**
- Characteristics of Spatial Data
 - Clusters are defined as densely populated regions of the space
 - Clusters have arbitrary shapes, orientation, and non-uniform sizes
 - Difference in densities across clusters and variation in density within clusters
 - Existence of special artifacts (*streaks*) and noise



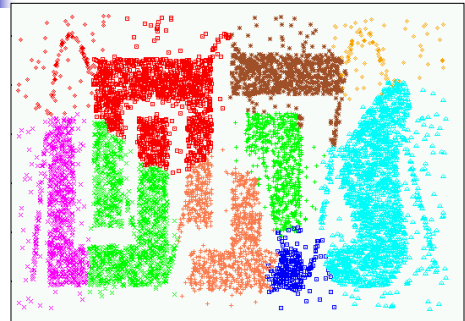
Experimental Results: CHAMELEON



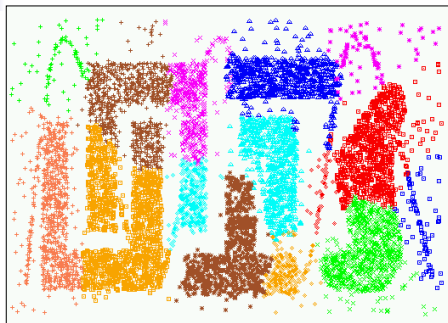
Experimental Results: **CHAMELEON**



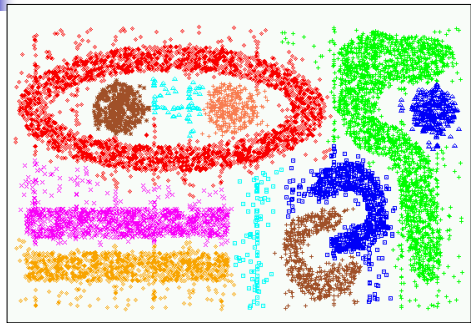
Experimental Results: **CURE (10 clusters)**



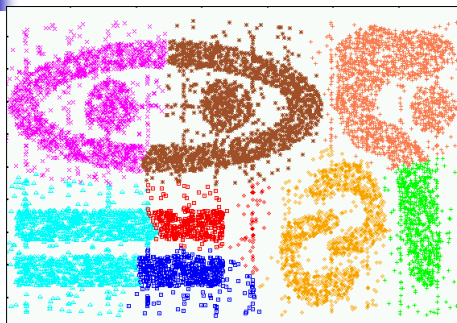
Experimental Results:
CURE (15 clusters)



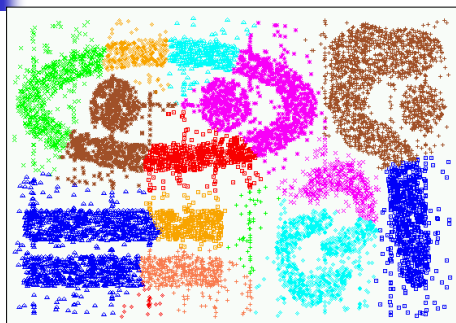
Experimental Results: **CHAMELEON**



Experimental Results: **CURE (9 clusters)**



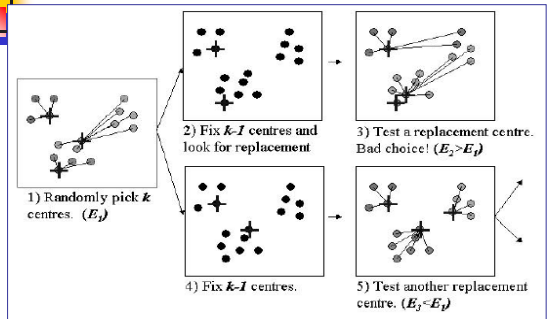
Experimental Results: **CURE (15 clusters)**



Partitioning Methods: k-medoids

- To find k clusters in n objects, the most centrally located object in a cluster (median object id) is used as a reference point of a cluster.
- Firstly find a representative object (the medoid) for each cluster.
- The remaining objects are distributed to clusters according to the similarity calculations with the medoids.
- The process then iteratively replaces one of the medoids by one of the non-medoids as long as the quality of the resulting clustering is improved (a cost function is used to measure the average dissimilarity an object and the medoid of its cluster).

k-medoid example: CLARANS



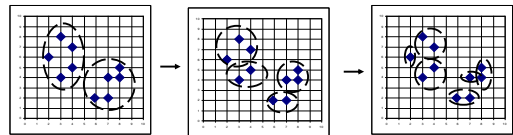
From: j. Han etc "Spatial Clustering Methods in Data Mining: A Survey"

k-means vs k-medoids

- k-medoids method uses the most centrally located objects (medoids) in a cluster to be the cluster centre, so it is less sensitive to noise and outliers.
- k-medoids method result in a higher running time.
- Both need to determine k and use the same criterion function (squared-error function) to converge the computation.
- K-medoids method extends the k-means paradigm to cluster categorical data by replacing the means of clusters with medoids.

DIANA (Divisive Analysis)

- Introduced in Kaufmann and Rousseeuw (1990)
- Inverse order of Agglomerative Clustering
- Eventually each node forms a cluster on its own

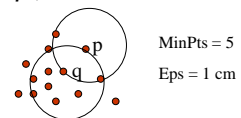


Density-Based Methods (DBSCAN)

- To discover clusters with arbitrary shape.
- The neighbourhood with radius ϵ of a given object is defined as ϵ -neighbourhood of the object.
- If the ϵ -neighbourhood of an object contains at least a minimum number, $MinPts$, of objects, then the object is called a core object.
- If q is a core object, p is within the ϵ -neighbourhood of q , then p is directly density-reachable.
- A chain of directly density-reachable defines the density-reachable objects.
- Two objects, p and q , are density-connected if both of them are density-reachable from an object o .

Density-Based Clustering: Background

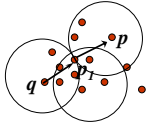
- Two parameters:
 - Eps : Maximum radius of the neighbourhood
 - $MinPts$: Minimum number of points in an Eps -neighbourhood of that point
- $N_{Eps}(p)$: $\{q \text{ belongs to } D \mid dist(p,q) \leq Eps\}$
- Directly density-reachable: A point p is directly density-reachable from a point q wrt. $Eps, MinPts$ if
 - 1) p belongs to $N_{Eps}(q)$
 - 2) core point condition: $|N_{Eps}(q)| \geq MinPts$



Density-Based Clustering: Examples:

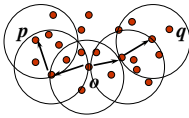
■ Density-reachable:

- A point p is density-reachable from a point q wrt. $Eps, MinPts$ if there is a chain of points $p_1, \dots, p_n, p_1 = q, p_n = p$ such that p_{i+1} is directly density-reachable from p_i



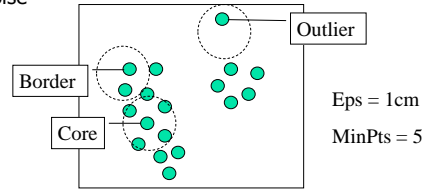
■ Density-connected

- A point p is density-connected to a point q wrt. $Eps, MinPts$ if there is a point o such that both, p and q are density-reachable from o wrt. Eps and $MinPts$.



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



Outlier Analysis

- Data objects that do not comply with the general behaviour or model of the data (grossly different from or inconsistent with the remaining set of data).
- Many data mining algorithms try to minimize the influence of outliers or eliminate them all together. This however, could result in the loss of information since *one person's noise could be another's signal*.
- **Outlier mining** can be used for fraud detection, early warning sign detection, exception handling, etc.

Web Mining

Dr Heng Tao Shen