

# CSE5334 DATA MINING

Lecture 13: Clustering (1)

CSE4392/5334 Data Mining, Fall 2009  
Department of Computer Science and Engineering, University of Texas at Arlington  
Chengkai Li (Slides courtesy of Vipin Kumar)

## What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

2

## What is not Cluster Analysis?

- Supervised classification**
  - Have class label information
- Simple segmentation**
  - Dividing students into different registration groups alphabetically, by last name
- Results of a query**
  - Groupings are a result of an external specification
- Graph partitioning**
  - Some mutual relevance and synergy, but areas are not identical

3

## Notion of a Cluster can be Ambiguous

4

## Types of Clusterings

- A **clustering** is a set of clusters
- Important distinction between **hierarchical** and **partitional** sets of clusters
- Partitional Clustering**
  - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering**
  - A set of nested clusters organized as a hierarchical tree

5

## Partitional Clustering

6

### Hierarchical Clustering

Traditional Hierarchical Clustering

Traditional Dendrogram

Non-traditional Hierarchical Clustering

Non-traditional Dendrogram

7

### Other Distinctions Between Sets of Clusters

- **Exclusive versus non-exclusive**
  - In non-exclusive clusterings, points may belong to multiple clusters.
  - Can represent multiple classes or 'border' points
- **Fuzzy versus non-fuzzy**
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  - Weights must sum to 1
  - Probabilistic clustering has similar characteristics
- **Partial versus complete**
  - In some cases, we only want to cluster some of the data
- **Heterogeneous versus homogeneous**
  - Cluster of widely different sizes, shapes, and densities

8

### Types of Clusters

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual
- Described by an Objective Function

9

### Types of Clusters: Well-Separated

- **Well-Separated Clusters:**
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.

3 well-separated clusters

10

### Types of Clusters: Center-Based

- **Center-based**
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster
  - The center of a cluster is often a **centroid**, the average of all the points in the cluster, or a **medoid**, the most "representative" point of a cluster

4 center-based clusters

11

### Types of Clusters: Contiguity-Based

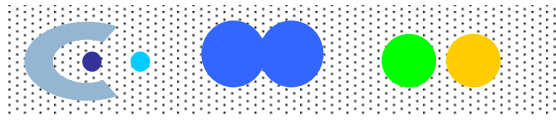
- **Contiguous Cluster (Nearest neighbor or Transitive)**
  - A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

8 contiguous clusters

12

### Types of Clusters: Density-Based

- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.

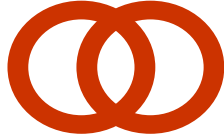


6 density-based clusters

13

### Types of Clusters: Conceptual Clusters

- Shared Property or Conceptual Clusters
  - Finds clusters that share some common property or represent a particular concept.



2 Overlapping Circles

14

## 15 Similarity and Dissimilarity

### Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

16

### Similarity and Dissimilarity

- Similarity and Dissimilarity of Simple Attributes
- Dissimilarity between Objects:
  - Distance
  - Set Difference
  - ...
- Similarity between Objects:
  - Binary Vectors
  - Vectors
  - ...

17

### Similarity/Dissimilarity for Simple Attributes

$p$  and  $q$  are the attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{ p-q }{n-1}$ (values mapped to integers 0 to $n-1$ , where $n$ is the number of values)	$s = 1 - \frac{ p-q }{n-1}$
Interval or Ratio	$d =  p - q $	$s = -d, s = \frac{1}{1+d}$ or $s = 1 - \frac{d - \min d}{\max d - \min d}$

**Table 5.1.** Similarity and dissimilarity for simple attributes

18

### Dissimilarity between Data Objects: Euclidean Distance

- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

Where  $n$  is the number of dimensions (attributes) and  $p_k$  and  $q_k$  are, respectively, the  $k^{th}$  attributes (components) or data objects  $p$  and  $q$ .

- Standardization is necessary, if scales differ.

19

### Euclidean Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

20

### Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left( \sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

Where  $r$  is a parameter,  $n$  is the number of dimensions (attributes) and  $p_k$  and  $q_k$  are, respectively, the  $k^{th}$  attributes (components) or data objects  $p$  and  $q$ .

21

### Minkowski Distance: Examples

- $r = 1$ . City block (Manhattan, taxicab,  $L_1$  norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$ . Euclidean distance
- $r \rightarrow \infty$ . "supremum" ( $L_{max}$  norm,  $L_{\infty}$  norm) distance.
  - This is the maximum difference between any component of the vectors
- Do not confuse  $r$  with  $n$ , i.e., all these distances are defined for all numbers of dimensions.

22

### Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L $_{\infty}$	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

23

### Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.
  - $d(p, q) \geq 0$  for all  $p$  and  $q$  and  $d(p, q) = 0$  only if  $p = q$ . (Positive definiteness)
  - $d(p, q) = d(q, p)$  for all  $p$  and  $q$ . (Symmetry)
  - $d(p, r) \leq d(p, q) + d(q, r)$  for all points  $p, q$ , and  $r$ . (Triangle Inequality)

where  $d(p, q)$  is the distance (dissimilarity) between points (data objects),  $p$  and  $q$ .
- A distance that satisfies these properties is a **metric**

24

### Common Properties of a Similarity

- Similarities, also have some well known properties.
  - $s(p, q) = 1$  (or maximum similarity) only if  $p = q$ .
  - $s(p, q) = s(q, p)$  for all  $p$  and  $q$ . (Symmetry)

where  $s(p, q)$  is the similarity between points (data objects),  $p$  and  $q$ .

25

### Similarity Between Binary Vectors

- Common situation is that objects,  $p$  and  $q$ , have only binary attributes
- Compute similarities using the following quantities
  - $M_{01}$  = the number of attributes where  $p$  was 0 and  $q$  was 1
  - $M_{10}$  = the number of attributes where  $p$  was 1 and  $q$  was 0
  - $M_{00}$  = the number of attributes where  $p$  was 0 and  $q$  was 0
  - $M_{11}$  = the number of attributes where  $p$  was 1 and  $q$  was 1
- Simple Matching and Jaccard Coefficients**
  - SMC = number of matches / number of attributes  
 $= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$
  - J = number of 11 matches / number of not-both-zero attributes values  
 $= (M_{11}) / (M_{01} + M_{10} + M_{11})$

26

### SMC versus Jaccard: Example

$p = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$   
 $q = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$

$M_{01} = 2$  (the number of attributes where  $p$  was 0 and  $q$  was 1)  
 $M_{10} = 1$  (the number of attributes where  $p$  was 1 and  $q$  was 0)  
 $M_{00} = 7$  (the number of attributes where  $p$  was 0 and  $q$  was 0)  
 $M_{11} = 0$  (the number of attributes where  $p$  was 1 and  $q$  was 1)

$SMC = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$

$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$

27

### Cosine Similarity

- If  $d_1$  and  $d_2$  are two document vectors, then  
 $\cos(d_1, d_2) = (d_1 \cdot d_2) / (||d_1|| ||d_2||)$ ,  
 where  $\cdot$  indicates vector dot product and  $||d||$  is the length of vector  $d$ .
- Example:
  - $d_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$
  - $d_2 = 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$

$d_1 \cdot d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$   
 $||d_1|| = (3^2+2^2+0^2+5^2+0^2+0^2+0^2+2^2+0^2+0^2)^{0.5} = (42)^{0.5} = 6.481$   
 $||d_2|| = (1^2+0^2+0^2+0^2+0^2+0^2+1^2+0^2+2^2)^{0.5} = (6)^{0.5} = 2.245$

$\cos(d_1, d_2) = .3150$

28

### General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.

- For the  $k^{th}$  attribute, compute a similarity,  $s_k$ , in the range  $[0, 1]$ .
- Define an indicator variable,  $\delta_k$ , for the  $k_{th}$  attribute as follows:
 
$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$
- Compute the overall similarity between the two objects using the following formula:
 
$$similarity(p, q) = \frac{\sum_{k=1}^n \delta_k s_k}{\sum_{k=1}^n \delta_k}$$

29

### Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use weights  $w_k$  which are between 0 and 1 and sum to 1.

$$similarity(p, q) = \frac{\sum_{k=1}^n w_k \delta_k s_k}{\sum_{k=1}^n \delta_k}$$

$$distance(p, q) = \left( \sum_{k=1}^n w_k |p_k - q_k|^r \right)^{1/r}$$

30