

Data Mining

- Clustering II

©University of Queensland, Brisbane Australia
<http://www.itee.uq.edu.au/~dke>

INFS4293 / INFS7203 Data Mining 1

Cluster Analysis

- Unitless Data transformation Formulas
- Partitioning Methods (K-means vs K-Medoids)
- Divisive Analysis Methods
- Density-Based Methods
- Outlier Analysis
- Summary

INFS4293 / INFS7203 Data Mining 2

Measure the Quality of Clustering

- Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, which is typically metric: $d(i, j)$
- There is a separate "quality" function that measures the "goodness" of a cluster.
- The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal and ratio variables.
- Weights should be associated with different variables based on applications and data semantics.
- It is hard to define "similar enough" or "good enough"
 - the answer is typically highly subjective.

INFS4293 / INFS7203 Data Mining 3

Type of data in clustering analysis

- Interval-scaled variables
- Binary variables
- Nominal, ordinal, and ratio variables
- Variables of mixed types

INFS4293 / INFS7203 Data Mining 4

Interval-Scaled Variables

- **Measurement unit** used can affect the clustering analysis.
- Data should be standardized.
- Variables can be weighted.
- Standardization is to convert variable to **unitless**:

$$s_f = \frac{1}{n} (|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

Where $x_{1f} \dots x_{nf}$ are n measurements of variable f , and m_f is the mean of f that is:

$$m_f = \frac{1}{n} (x_{1f} + x_{2f} + \dots + x_{nf})$$

Standardized measurement (z-score):

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

S_f is called the mean absolute deviation.

INFS4293 / INFS7203 Data Mining 5

Once we have standardized the data, how can we calculate the dissimilarity between objects?

- **Euclidean distance:**

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2}$$
- **Manhattan (city block) distance:**

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$
- **Requirements of distance function:**
 - $d(i, j) \geq 0$
 - $d(i, i) = 0$
 - $d(i, j) = d(j, i)$
 - $d(i, j) \leq d(i, h) + d(h, j)$

INFS4293 / INFS7203 Data Mining 6

Generalized Distance Computation

- Minkowski Distance:

$$d(i, j) = (|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)^{1/q}$$

- If each variable is assigned a weight according to its perceived importance (weighted Euclidean Distance):

$$d(i, j) = \sqrt{w_1|x_{i1} - x_{j1}|^2 + w_2|x_{i2} - x_{j2}|^2 + \dots + w_p|x_{ip} - x_{jp}|^2}$$

INFS4293 / INFS7203 Data Mining

7

Binary Variables

Dissimilarity is calculated based on a matrix if **all binary variables have the same weight** (symmetric): given two objects i, j: how do we calculate the similarity between them d(i, j)?

		Object j		
		1	0	Sum
Object i	1	q	r	q+r
	0	s	t	s+t
	Sum	q+s	r+t	p

$$p = q+r+s+t$$

INFS4293 / INFS7203 Data Mining

Binary Variables (cont)

- Dissimilarity between object i and j is the **simple matching coefficient**:

$$d(i, j) = \frac{r+s}{p}$$

- Sometimes the values of binary variables are not treated equally, where number of negative matches is considered unimportant and thus is ignored (**Jaccard coefficient**):

$$d(i, j) = \frac{r+s}{q+r+s}$$

INFS4293 / INFS7203 Data Mining

9

Dissimilarity between Binary Variables

- Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- gender is a **symmetric** attribute
- the remaining attributes are **asymmetric** binary
- let the values **Y** and **P** be set to 1, and the value **N** be set to 0

$$d(\text{jack}, \text{mary}) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(\text{jack}, \text{jim}) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(\text{jim}, \text{mary}) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

INFS4293 / INFS7203 Data Mining

10

Nominal Variables

- A nominal variable is a generalization of binary variable in that it can take on more than two states (e.g., map colour).
- States can be denoted by letters, symbols, or a set of integers, **without specific ordering**.
- Dissimilarity can be computed (simple matching):

$$d(i, j) = \frac{p-m}{p}$$

where m is the number of the matches and p is the total number of variables.

INFS4293 / INFS7203 Data Mining

11

Ordinal Variables

- A discrete ordinal variable resembles a nominal variable, except that the states of the ordinal value are **ordered** in a meaningful sequence (e.g., gold, silver, bronze in a sport).
- The **order is more essential** than the actual values. The **scale of the values is unimportant**.

INFS4293 / INFS7203 Data Mining

12

Ordinal Variables (cont)

- Dissimilarity computation involving variable f has following steps:
 - Replace each object (x_{if})'s f value by its ranking value: r_{if} ($1, \dots, M_f$).
 - Map the ranges of each ordinal variable onto $[0.0, 1.0]$, so that each variable has equal weight:

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$
 - Dissimilarity is computed using any distance measures with z_{if} to represent the f value for the i th object (x_{if}).

INFS4293 / INFS7203 Data Mining 13

Ratio-Scaled Variables

- A ratio-scaled variable makes a positive measurement on a non-linear scale, such as an exponential scale (e.g., the growth of a bacteria population, or decay of a radioactive element):

$$Ae^{Bt} \text{ or } Ae^{-Bt}$$

Where A and B are positive constants.

INFS4293 / INFS7203 Data Mining 14

Ratio-Scaled Variables (cont)

- Three methods to compute the dissimilarity:
 - Treated same as interval-scaled variables (scale may be distorted).
 - Apply logarithmic transformation $y_{if} = \log(x_{if})$. Then y_{if} is treated as interval-scaled values.
 - Treat x_{if} as continuous ordinal data and treat their ranks as interval-scaled values.
- The choice of method is dependent on the given application.

INFS4293 / INFS7203 Data Mining 15

Variables of Mixed Types

- In real databases, objects are described by a mixture of variable types.
- One approach is to perform a separate cluster analysis for each different data type (based on the groups of each kind of variables if cluster analysis derives compatible results).
- Another approach is to combine different variables into a single dissimilarity matrix, bringing all of the meaningful variables onto a common scale of the interval $[0.0, 1.0]$.

INFS4293 / INFS7203 Data Mining 16

Computing of Mixed Types

- Given p variables and a set of objects:

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

Where indicator $\delta_{ij}^{(f)} = 0$ if either (1) x_{if} or x_{jf} is missing, or (2) $x_{if} = x_{jf} = 0$ and f is asymmetric binary; otherwise $\delta_{ij}^{(f)} = 1$.

INFS4293 / INFS7203 Data Mining 17

Computing of Mixed Types (cont)

- The contribution of variable f to the dissimilarity between i and j , $d_{ij}^{(f)}$ is based on the type of f :
 - binary or nominal: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise $d_{ij}^{(f)} = 1$.
 - interval-scaled: $d_{ij}^{(f)} = (|x_{if} - x_{jf}|) / (\max_i x_{if} - \min_i x_{if})$
 - ordinal or ratio-scaled: compute the ranks r_{if} and $z_{if} = (r_{if} - 1) / (M_f - 1)$, then treat z_{if} as interval-scaled.

INFS4293 / INFS7203 Data Mining 18

Partitioning Methods: k-medoids

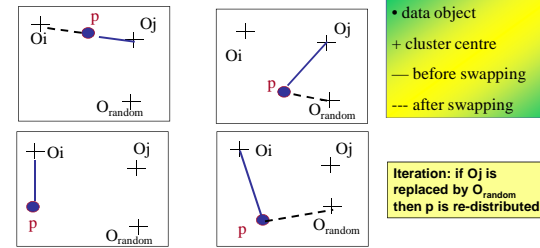
- To find k clusters in n objects, the most centrally located object in a cluster (median object id) is used as a reference point of a cluster.
- Firstly find a representative object (the medoid) for each cluster.
- The remaining objects are distributed to clusters according to the similarity calculations 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).
- Repeat the process until criterion function converges (squared-error criterion):

$$E = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2$$

19

k-medoids (cont)

- To determine whether a non-medoid object (O_{random}) is a good replacement for a current medoid (o_j) each non-medoid p is examined:



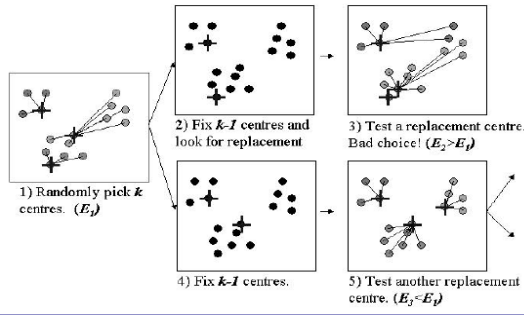
INFS4293 / INFS7203 Data Mining 20

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.

INFS4293 / INFS7203 Data Mining 21

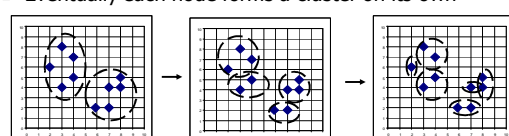
k-medoid example: CLARANS



From: J. Han et al. "Spatial Clustering Methods in Data Mining: A Survey" INFS4293 / INFS7203 Data Mining 22

DIANA (Divisive Analysis)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, e.g., Splus
- Inverse order of AGNES
- Eventually each node forms a cluster on its own

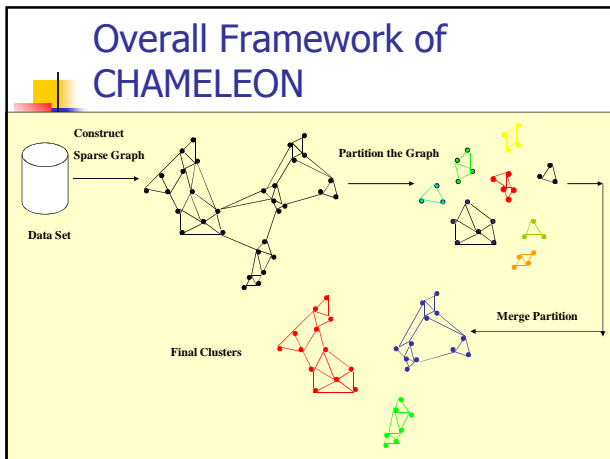


INFS4293 / INFS7203 Data Mining 23

CHAMELEON (Hierarchical clustering using dynamic modeling)

- CHAMELEON: by G. Karypis, E.H. Han, and V. Kumar '99
- Measures the similarity based on a dynamic model
 - Two clusters are merged only if the interconnectivity and closeness (proximity) between two clusters are high relative to the internal interconnectivity of the clusters and closeness of items within the clusters
- A two-phase algorithm
 - Use a graph partitioning algorithm: cluster objects into a large number of relatively small sub-clusters
 - Use an agglomerative hierarchical clustering algorithm: find the genuine clusters by repeatedly combining these sub-clusters

INFS4293 / INFS7203 Data Mining 24



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 .

INFS4293 / INFS7203 Data Mining 26

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
 - p belongs to $N_{Eps}(q)$
 - core point condition: $|N_{Eps}(q)| \geq MinPts$

INFS4293 / INFS7203 Data Mining 27

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 $p_1, \dots, p_n, p_n = p$ 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**.

INFS4293 / INFS7203 Data Mining 28

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

INFS4293 / INFS7203 Data Mining 29

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.
 - Distance-based approach**
 - Deviation-based approach**
 - Statistical approach**

INFS4293 / INFS7203 Data Mining 30

Reading List

- J. Han, M. Kamber, and A. K. H. Tung, “*Spatial Clustering Methods in Data Mining: A Survey*”, H. Miller and J. Han (eds.), Geographic Data Mining and Knowledge Discovery, Taylor and Francis, 2001.