Data Mining:

Concepts and Techniques

— Chapter 7 —

Jiawei Han

Department of Computer Science

University of Illinois at Urbana-Champaign

www.cs.uiuc.edu/~hanj

Cluster Analysis

1. What is Cluster Analysis?
2. Types of Data in Cluster Analysis
3. A Categorization of Major Clustering Methods
4. Partitioning Methods
5. Hierarchical Methods
6. Density-Based Methods

What is Cluster Analysis?

- Cluster: a collection of data objects
  - *Similar* to one another within the same cluster
  - *Dissimilar* to the objects in other clusters
  - → distance (or similarity) measures
- Cluster analysis
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- Unsupervised learning: no predefined classes
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

Requirements of Clustering in Data Mining

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

Explanations on whiteboard

Explanations on whiteboard

Cluster Analysis

1. What is Cluster Analysis?
2. Types of Data in Cluster Analysis
3. A Categorization of Major Clustering Methods
4. Partitioning Methods
5. Hierarchical Methods
6. Density-Based Methods

Data Structures

- Data matrix
  - *n* objects, *p* attributes
  - (two modes)
  - One row represents one object

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix
  - Distance table
  - (one mode)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

Distances between objects

<u>Distances</u> are normally used to measure the <u>similarity</u> or <u>dissimilarity</u> between two data objects

- Properties
  - $d(i,j) \geq 0$
  - $d(i,i) = 0$
  - $d(i,j) = d(j,i)$
  - $d(i,j) \leq d(i,k) + d(k,j)$

Example on whiteboard

Is the following a distance measure?

$$d(i,j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{otherwise} \end{cases}$$

Type of data in clustering analysis

- Interval-scaled variables
  - Continuous measurements (weight, temperature, …)
- Binary variables
  - Variables with 2 states (on/off, yes/no)
- Nominal variables
  - A generalization of the binary variable in that it can take more than 2 states (color/red,yellow,blue,green)
- Ordinal
  - ranking is important (e.g. medals(gold,silver,bronze))
- Ratio variables
  - a positive measurement on a nonlinear scale (growth)
- Variables of mixed types

Interval-valued variables

- Sometimes we need to standardize the data
  - Calculate the mean absolute deviation:

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

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

  - Calculate the standardized measurement (*z-score*)

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

Example on whiteboard

Data set: { 1, 2, 6 } → n = 3

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

mf = 1/3 (1 + 2 + 6) = 3

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

sf = 1/3 (| 1 - 3 | + | 2 - 3 | + | 6 - 3 |) = 2

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

z1 = (1 – 3) / 2 = -1

z2 = (2 – 3) / 2 = -0.5

z3 = (6 – 3) / 2 = 1.5

Distances between objects

- Euclidean distance:

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+\ldots+|x_{i_p}-x_{j_p}|^2)}$$

- Manhattan distance:

$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+\ldots+|x_{i_p}-x_{j_p}|$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects,

Distances between objects

- *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + ... + |x_{i_p} - x_{j_p}|^q)}$$

  q is a positive integer

- If *q* = *1*, *d* is Manhattan distance
- If *q* = 2, *d* is Euclidean distance

Example on whiteboard

Binary Variables

- symmetric binary variables: both states are equally important; 0/1
- asymmetric binary variables: one state is more important than the other (e.g. outcome of disease test); 1 is the important state, 0 the other

Contingency tables for Binary Variables

		Object j		
		1	0	*sum*
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

a: number of attributes having 1 for object i and 1 for object j
b: number of attributes having 1 for object i and 0 for object j
c: number of attributes having 0 for object i and 1 for object j
d: number of attributes having 0 for object i and 0 for object j
p = a+b+c+d

Distance measure for *symmetric* binary variables

		Object j		
		1	0	sum
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

Example on whiteboard

	Fever	Cough	test1	test2	test3	test4
Jack	p	n	p	n	n	n
Mary	p	n	p	n	p	n
Jane	p	p	n	n	n	n

p → 1; n → 0

Example on whiteboard

	Fever	Cough	test1	test2	test3	test4
Jack	p	n	p	n	n	n
Mary	p	n	p	n	p	n
Jane	p	p	n	n	n	n

		Mary	
		1	0
Jack	1		
	0		

		Jane	
		1	0
Jack	1		
	0		

		Jane	
		1	0
Mary	1		
	0		

d(Jack, Mary) =

d(Jack, Jane) =

d(Mary, Jane) =

Distance measure for *asymmetric* binary variables

		Object j		
		1	0	sum
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

$$d(i,j) = \frac{b+c}{a+b+c}$$

Jaccard coefficient = 1- d(i,j) = $sim_{Jaccard}(i,j) = \frac{a}{a+b+c}$

Example on whiteboard

	Fever	Cough	test1	test2	test3	test4
Jack	p	n	p	n	n	n
Mary	p	n	p	n	p	n
Jane	p	p	n	n	n	n

		Mary	
		1	0
Jack	1		
	0		

		Jane	
		1	0
Jack	1		
	0		

		Jane	
		1	0
Mary	1		
	0		

d(Jack, Mary) =

d(Jack, Jane) =

d(Mary, Jane) =

Nominal or Categorical Variables

- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables

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

- Method 2: use a large number of binary variables
  - creating a new asymmetric binary variable for each of the *M* nominal states *(Homework)*

Example on whiteboard

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - replace x_{if} by their rank $r_{if} \in \{1, \ldots, M_f\}$
  - map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - compute the dissimilarity using methods for interval-scaled variables

Example on whiteboard

Ratio-Scaled Variables

- Ratio-scaled variable: a positive measurement on a nonlinear scale, approximately at exponential scale, such as Ae^{Bt} or Ae^{-Bt}
- Methods:
  - treat them like interval-scaled variables—*not a good choice!* (why?—the scale can be distorted)
  - apply logarithmic transformation

$$y_{if} = log(x_{if})$$

  - treat them as continuous ordinal data, treat their rank as interval-scaled

Example on whiteboard

Variables of Mixed Types

- A database may contain all the six types of variables
  - symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio
- One may use a weighted formula to combine their effects:

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

  - *f* is binary or nominal:

    $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
  - *f* is interval-based: use the (normalized) distance
  - *f* is ordinal or ratio-scaled
    - compute ranks r_{if} and
    - and treat z_{if} as interval-scaled

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - delta(i,j) = 0 iff (i) x-value is missing or (ii) x-values are 0 and f asymmetric binary attribute

Example on whiteboard

A,E: interval-based variable, Euclidean distance
B: symmetric binary variable
C,D: asymmetric binary variables

	A	B	C	D	E
I	1	Y	Y	N	5
J	2	Y	N	N	No-value

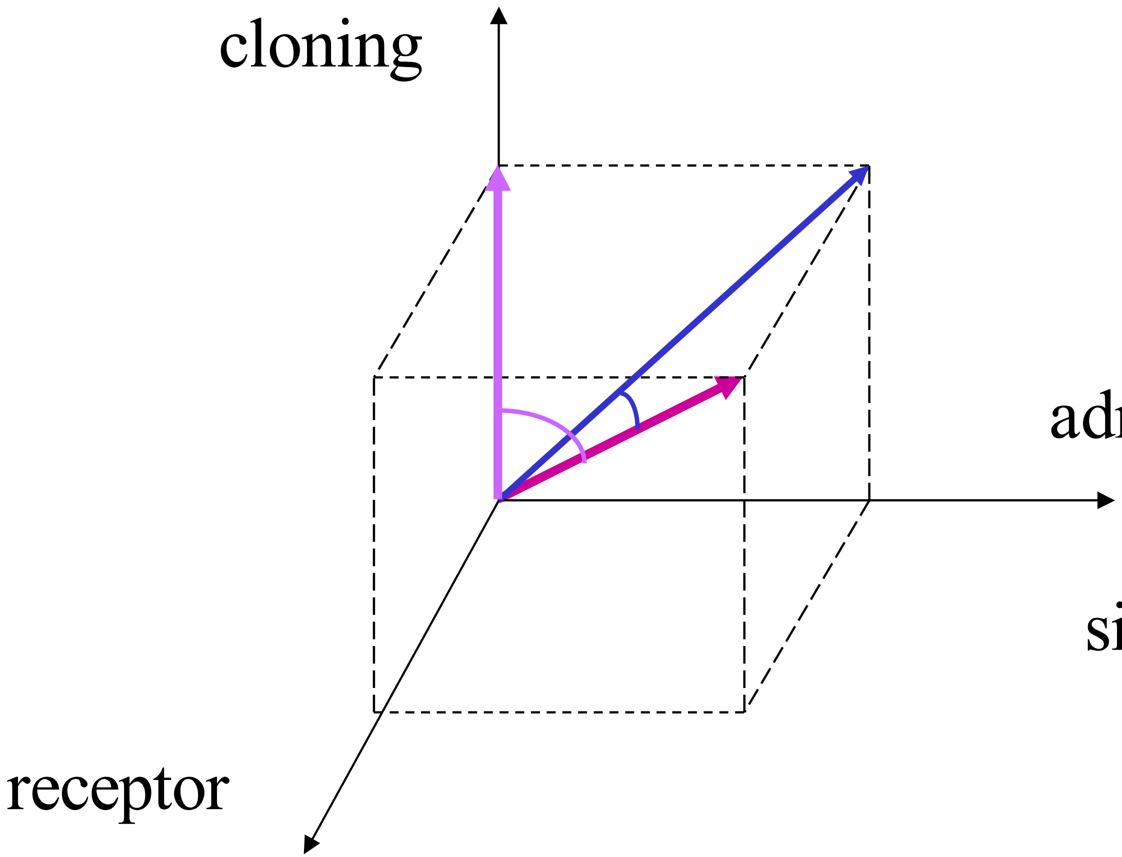
Vector Objects

- Vector objects: keywords in documents, gene features in micro-arrays, etc.
- Broad applications: information retrieval, biologic taxonomy, etc.
- Cosine measure

$$s(\vec{X}, \vec{Y}) = \frac{\vec{X}^t \cdot \vec{Y}}{|\vec{X}||\vec{Y}|},$$

$\vec{X}^t$ is a transposition of vector $\vec{X}$, $|\vec{X}|$ is the Euclidean normal of vector $\vec{X}$,

Vector model for information retrieval (simplified)



Doc1 (1,1,0)
Doc2 (0,1,0)

Q (1,1,1)

$$\text{sim}(d,q) = \frac{d \cdot q}{|d| \times |q|}$$

Typical Alternatives to Calculate the Distance between Clusters

- Single link: smallest distance between an element in one cluster and an element in the other, i.e., $d(K_i, K_j) = \min d(t_{ip}, t_{jq})$
- Complete link: largest distance between an element in one cluster and an element in the other, i.e., $d(K_i, K_j) = \max d(t_{ip}, t_{jq})$
- Average: avg distance between an element in one cluster and an element in the other, i.e., $d(K_i, K_j) = \text{avg } d(t_{ip}, t_{jq})$
- Centroid: distance between the centroids of two clusters, i.e., $d(K_i, K_j) = d(C_i, C_j)$
- Medoid: distance between the medoids of two clusters, i.e., $d(K_i, K_j) = d(M_i, M_j)$
  - Medoid: one chosen, centrally located object in the cluster

Example on whiteboard

Centroid, Radius and Diameter of a Cluster (for numerical data sets)

- Centroid: the "middle" of a cluster

$$C_m = \frac{\sum_{i=1}^{N} (t_{ip})}{N}$$

- Radius: square root of average distance from any point of the cluster to its centroid

$$R_m = \sqrt{\frac{\sum_{i=1}^{N} (t_{ip} - c_m)^2}{N}}$$

- Diameter: square root of average mean squared distance between all pairs of points in the cluster

$$D_m = \sqrt{\frac{\sum_{i=1}^{N} \sum_{i=1}^{N} (t_{ip} - t_{iq})^2}{N(N-1)}}$$