Cluster Analysis

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.

Applications

→ market research → clustering can help marketers discover distinct groups in their customers based on their purchasing patterns.
→ pattern recognition
→ data analysis
→ image processing
→ in biology → categorize genes with similar functionality

clustering is also called **data segmentation.** because clustering partitions large data sets into groups according to their similarity.

→ Clustering can also be used for **outlier detection**. Applications of outlier detection include the detection of credit card fraud and monitoring of criminal activities in e-commerce.

For example, exceptional cases in credit card transactions, such as very expensive and frequent purchases, may be of interest as possible fraudulent activity.

The following are typical requirements of clustering in data mining:

(1) Scalability → Many clustering algorithms work well on small data sets, however, clustering on a sample of a given large data set may lead to biased results. Highly scalable clustering algorithms are needed.

(2) Ability to deal with different types of attributes →

Many algorithms are designed to cluster interval based (numerical) data. However applications may require clustering other types of data, such as binary, categorical (nominal) and ordinal data.

(3) Discovery of clusters with arbitrary shape.

Many clustering algorithms determine clusters based on Euclidean distance. Algorithms based on such distance tend to find spherical clusters with similar size and density. It is important to develop algorithms that can detect clusters of arbitrary shape.

(4) Minimal requirements for domain knowledge to determine input parameters. :- Many clustering algorithms require users to input certain parameters in cluster analysis (such as the no. of desired clusters).

The clustering results can be quite sensitive to input parameters. parameters are often difficult to determine for high-dimensional objects.

(5) Ability to deal with noisy data

Most real world databases contain outliers or missing, unknown or erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.

(6) Incremental clustering and insensitivity to the order of input records :-

Some clustering algorithms cannot incorporate newly inserted data into existing clustering structures. and determine a new clustering from scratch.

Some clustering algorithms are sensitive to the order of input data.

It is important to develop incremental clustering algorithms that are insensitive to the order of input

(7) High dimensionality :-

A database or a data warehouse can contain several dimensions or attributes.

many clustering algorithms are good at handling two or three dimensions.

Finding clusters of data objects in high dimensional space is challenging.

(8) constraint based clustering :-

Real world applications may need to perform clustering under various kinds of constraints.

A challenging task is to find groups of data with good clustering behavior that satisfy specified constraints.

Interpretability and usability.

users expect clustering results to be interpretable comprehensible and usable.

II. Types of Data in cluster analysis

Data matrix → (object-by-variable structure)

This represents n objects, with p variables.
The structure is ~~is~~ n X p matrix. Also called two-mode matrix

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

Dissimilarity matrix (object by object structure)

Represented by n X n table.

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

where $d(i,j)$ is the measured difference between objects i and j. Also called one-mode matrix.

$d(i,j) = d(j,i)$ and $d(i,i) = 0$.

Interval scaled Variables → continuous measurements of a roughly linear scale.

Ex- weight and height, weather temp. etc.

The measurement unit used can affect the clustering analysis. For example, changing measurements units from meters to inches for height or from kg to pounds for weight, may lead to different clustering structures. So we have to standardized the data.

(1) calculate the mean absolute deviation.

$$s_f = \frac{1}{n}\left[|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|\right]$$

m_f → mean value of f. More robust to outliers.

(2) calculate the Z-score.

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

(3) Euclidean distance.

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{in} - x_{jn})^2}$$

Manhattan or city block distance.

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{in} - x_{jn}|$$

Both the distances satisfy the following.

1. $d(i,j) \geq 0$ = distance is nonnegative no.
2. $d(i,j) = 0$, the distance of an object to itself.
3. $d(i,j) = d(j,i)$, distance is symmetric.
4. ~~$d(i,j) \leq d(i,k) + d(k,j)$~~

Minkowski distance.

$$d(i,j) = \left[|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \ldots + |x_{in} - x_{jn}|^p\right]^{1/p}$$

p = +ve integer.

Binary variables → it has only two states: 0 or 1

How can we compute the dissimilarity between two binary variables?

Ans → By computing a dissimilarity matrix from the given binary data.

		object j		
		1	0	sum
object i	1	q	r	q+r
	0	s	t	s+t
	sum	q+s	r+t	p

where q = no. of variables that equal 1 for both i & j objects.
r = " " that equal 1 for i and 0 for j objects.
s = " " that equal 0 for i and 1 for j =
t = " " that equal 0 for i and 0 for j objects.
p = total no. of variables = q+r+s+t.

Binary variable is symmetric → if both of its states are equally valuable and carry the same weight. Ex: gender - male or female.

symmetric binary dissimilarity (or distance) can be calculated between objects i and j by

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

A binary variable is asymmetric; if the outcomes of the states are not equally important.
Ex → +ve or -ve outcomes of a disease test.

The asymmetric binary dissimilarity, where the no. of -ve matches, t is considered unimportant and thus is ignored is calculated as

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

Ex suppose that a patient record table contains the attributes name, gender, fever, cough, test-1, test-2, test-3 and test-4, where name is an object identifier, gender is a symmetric attribute and remaining attributes are asymmetric binary.

Let the values Y(yes) and P(+ve) set to 1. and the value N(no or -ve) set to 0.

suppose that the distance between objects is computed based only on the asymmetric variables. The distance between each pair of three patients is.

$$d(\underset{i}{\text{jack}}, \underset{j}{\text{mary}}) = \frac{r+s}{q+r+s} = \frac{0+1}{2+0+1} = 0.33$$

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

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

name	gender	fever	cough	t-1	t-2	t-3	t-4
(i) Jack	M	Y(1)	N(0)	P(1)	N(0)	N(0)	N(0)
(j) Mary	F	Y(1)	N(0)	P(1)	N(0)	P(1)	N(0)
Jim	M	Y(1)	Y(1)	N(0)	N(0)	N(0)	N(0)

q = no. of variables that equal 1 for both i, j
r = " " that equal 1 for i and 0 for j
s = " " that equal 0 for i and 1 for j

Categorical variables

A categorical variable is a generalization of the binary variable that can take more than two states.

Ex: map_color is a categorical variable that have, five states, red, yellow, green, pink and blue.

Let the no. of states of a categorical variable be M. The states can be denoted by letters, symbols or a set of integers.

The dissimilarity between two objects i and j can be computed based on the ratio of mismatches.

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

where m = no. of matches (i.e the no. of variables for which i and j are in the same state)
p = total no. of variables.

Ex

object identifier	test-1
1	code-A
2	code-B
3	code-C
4	code-A

Dissimilarity matrix =

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Here, we have one categorical variable, test-1, so, p = 1.

d(i,j) = 0 if objects i and j match and 1 if the objects differ.

Ordinal variables

A discrete ordinal variable, resembles a categorical variable, except that the M-states of the ordinal value are ordered in a meaningful sequence.

Ex = Assistant, associate and professor.

continuous ordinal variables

A set of continuous data for an unknown scale, i,e. the relative ordering of the values is essential but their actual magnitude is not.

Ex Relative ranking in sports (i,e Gold, Silver, bronze)

~~Ordinal variables may~~

Suppose, f is a variable from a set of ordinal variables describing n objects.

The dissimilarity computation involves the following:

1. The value of f for the ith object is x_{if} and f has M_f ordered states, representing the ranking $1, \ldots M_f$. Replace each x_{if} by its corresponding rank, $r_{if} \in \{1, \ldots M_f\}$
2. since each ordinal variable can have a different no. of states, it is often necessary to map the range of each variable onto $\{0, 1\}$. This can be done by replacing the rank r_{if} of the ith object in the fth variable by

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

3. Dissimilarity can then be computed using any of the distance measures, using Z_{if} to represent the f value for the ith object.

Ex

object identifier	test-2	rank	Normalized value
1	Excellent	3	1
2	Fair	1	0
3	Good	2	0.5
4	Excellent	3	1

$M_f = 3$, three states, Ex, Fair, good.

① Replace each value for test-2 by its rank.

Ex = 3; Fair = 1, Good = 2

② Normalizing the ranking by mapping rank 1 to 0.0, rank 2 to 0.5 and rank 3 to 1.0.

③ use Euclidean distance, which results the following dissimilarity matrix.

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix}$$

Ratio scaled variables

A ratio scaled variable makes a +ve measurement on a nonlinear scale, such as an exponential scale.

Ae^{Bt} or Ae^{-Bt}, where A and B are +ve constants, t is time.

Ex → growth of a bacteria population.

There are three methods to handle ratio-scaled variables:-

(1) Treat ratio-scaled variables like interval scaled variables. But it is not a good choice.

(2) Apply logarithmic transformation:- A ratio scaled variable f having value x_{if} for object i, ~~by~~ $y_{if} = \log(x_{if})$. Then treat the y of value them as interval-valued.

(3) Treat it as continuous ordinal data and treat their ranks as interval-valued.

Ex

Object	te^{Bt+3}	$\log(te^{t-3})$
	445	2.65
1	22	1.34
2	164	2.21
3	1210	3.08
4		

Then use the Euclidean distance on the transformed value, we get

The dissimilarity matrix =

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & & & \\ 1.31 & 0 & & \\ 0.44 & 0.87 & 0 & \\ 0.43 & 1.74 & 0.87 & 0 \end{bmatrix}$$

2.65
1.34
1.31

Variables of mixed types

Suppose that the data set contains p variables of mixed type. The dissimilarity $d(i,j)$ between objects i and j is defined as

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

where the indicator $\delta_{ij}^{(f)} = 0$ if either x_{if} or x_{jf} is missing (or)

(2) $x_{if} = x_{jf} = 0$ and variable f is asymmetric binary, otherwise $\delta_{ij}^{(f)} = 1$

The dissimilarity between i and j is computed as follows -

① If f is interval-based: $d_{ij}^{f} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$

where h runs over all nonmissing objects for variable f.

② If f is binary or categorical $d_{ij}^{f} = 1$ otherwise $d_{ij}^{f} = 0$, if $x_{if} = x_{jf}$

(3) If f is ordinal, compute the ranks r_{if} and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$ and treat z_{if} as interval scaled.

(4) If f is ratio scaled: either perform logarithmic transformation and treat the transformed data as interval-scaled.

(or) treat f as continuous ordinal data, compute r_{if} and z_{if}, and then treat z_{if} as interval scaled.

Ex

object	test-1 (categorical)	test-2 (ordinal)	test-3 (ratio-scaled)
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1210

The procedure we followed for test-1 and test--2 are the same.

Let us consider for test-3. After applying logarithmic transformation we get.

2.65
1.34
2.21
3.08.

Let $\max_h x_h = 3.08$

$\min_h x_h = 1.34$

Then divide the values in the dissimilarity matrix by $3.08 - 1.34 = 1.74$.

$$\begin{bmatrix} 0 & & \\ 1.31/1.74 & 0 & \\ 0.44/1.74 & 0.87/1.74 & 0 \\ 0.43/1.74 & 1.74/1.74 & 0.87/1.74 \quad 0 \end{bmatrix}$$

$$d_3 = \begin{bmatrix} 0 & & & \\ 0.75 & 0 & & \\ 0.25 & 0.50 & 0 & \\ 0.25 & 1 & 0.5 & 0 \end{bmatrix}$$

Now use the dissimilarity matrices for the 3-variables using eqn (1)

$$d_1 = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Now. $d(2,1) = \frac{1(1)+1(1)+1(0.75)}{3}$

$= 0.92$

$d(3,1) = \frac{1(1)+1(0.5)+1(0.25)}{3}$

$= 0.58$

$$d_2 = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.50 & \end{bmatrix}$$

The final result is

$$\begin{bmatrix} 0 & & \\ 0.92 & 0 & \\ 0.58 & .67 & 0 \\ 0.09 & 1 & .670 \end{bmatrix}$$

Vector objects

To measure distance between complex objects.

Similarity function, $s(x, y)$, to compare two vectors x and y. calculate cosine measure

$$s(x, y) = \frac{x^t \cdot y}{\|x\| \|y\|}$$

x^t = transposition of vector x

$\|x\|$ = Euclidean norm of vector x

$\|y\|$ = " " y of $\|y\|$

s = cosine angle between vectors nearly.

Euclidean norm of vector $x = \sqrt{x_1^2 + x_2^2 + \cdots x_p^2}$

Ex Given $x = (1, 1, 0, 0)$

$y = (0, 1, 1, 0)$

$$s(x, y) = \frac{0+1+0+0}{\sqrt{2}\sqrt{2}} = 0.5$$

Tanimoto coefficient = $s(x, y) = \frac{x^t \cdot y}{x^t \cdot x + y^t \cdot y + - x^t \cdot y}$

y distance

Types of clustering Method

partitioning methods - K-mean, K-medoids.

Hierarchical methods → agglomerative or divisive

Density based methods

Grid based methods

Model based methods