LECTURE 34

Dr.Vani V

Source : Textbook 1 , Chapter 6 , Selected Exercises

Exercise

Given a data set of five objects characterized by a single continuous feature:

	a	b	C	d	e
Feature	1	2	4	5	6

Apply the agglomerative algorithm with single-link, complete-link and averaging cluster distance measures to produce three dendrogram trees, respectively.

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0

Summary

- Hierarchical algorithm is a sequential clustering algorithm
  - *Use distance matrix to construct a tree of clusters (dendrogram)*
  - *Hierarchical representation without the need of knowing # of clusters (can set termination condition with known # of clusters)*
- Major weakness of agglomerative clustering methods
  - *Can never undo what was done previously*
  - *Sensitive to cluster distance measures and noise/outliers*
  - *Less efficient: O (n^2 logn), where n is the number of total objects*
- There are several variants to overcome its weaknesses
  - *BIRCH: scalable to a large data set*
  - *ROCK: clustering categorical data*
  - *CHAMELEON: hierarchical clustering using dynamic modelling*

Exercise 1:

Draw a contingency table for each of the following rules using the transactions shown in Table

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

Rules: {b} → {c}, {a} → {d}, {b} → {d}, {e} → {c}, {c} → {a}.

Solution 1:

Contingency tables for each of the

Rules: {b} → {c}, {a} → {d}, {b} → {d}, {e} → {c}, {c} → {a}.

using the transactions given

Answer:

	c	$\overline{c}$
b	3	4
$\overline{b}$	2	1

	d	$\overline{d}$
a	4	1
$\overline{a}$	5	0

	d	$\overline{d}$
b	6	1
$\overline{b}$	3	0

	c	$\overline{c}$
e	2	4
$\overline{e}$	3	1

	a	$\overline{a}$
c	2	3
$\overline{c}$	3	2

Exercise 2 & Solution 2:...

Use the contingency tables in part (a) to compute and rank the rules in decreasing order according to the following measures.

i. Support.

Rules	Support	Rank
$b \longrightarrow c$	0.3	3
$a \longrightarrow d$	0.4	2
$b \longrightarrow d$	0.6	1
$e \longrightarrow c$	0.2	4
$c \longrightarrow a$	0.2	4

Exercise 2 & Solution 2:...

Use the contingency tables in part (a) to compute and rank the rules in decreasing order according to the following measures.

ii. Confidence.

Rules	Confidence	Rank
$b \longrightarrow c$	3/7	3
$a \longrightarrow d$	4/5	2
$b \longrightarrow d$	6/7	1
$e \longrightarrow c$	2/6	5
$c \longrightarrow a$	2/5	4

Exercise 2 & Solution 2:...

Use the contingency tables in part (a) to compute and rank the rules in decreasing order according to the following measures.

iii. Interest(X $\rightarrow$ Y) = $\frac{P(X,Y)}{P(X)}P(Y)$

Rules	Interest	Rank
$b \longrightarrow c$	0.214	3
$a \longrightarrow d$	0.72	2
$b \longrightarrow d$	0.771	1
$e \longrightarrow c$	0.167	5
$c \longrightarrow a$	0.2	4

Note: Please use support and not support count in interest formula.

Exercise 2 & Solution 2:...

Use the contingency tables in part (a) to compute and rank the rules in decreasing order according to the following measures.

iii. IS(X → Y) = $\frac{P(X,Y)}{\sqrt{P(X)P(Y)}}$

Rules	IS	Rank
$b \longrightarrow c$	0.507	3
$a \longrightarrow d$	0.596	2
$b \longrightarrow d$	0.756	1
$e \longrightarrow c$	0.365	5
$c \longrightarrow a$	0.4	4

Exercise 3:

Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are

A1(2, 10),A2(2, 5),A3(8, 4),B1(5, 8),B2(7, 5),B3(6, 4),C1(1, 2),C2(4, 9).

The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only

(a) the three cluster centers after the first round of execution.

(b) the final three clusters.

Solution 3:

(a) The three cluster centers after the first round of execution.

Answer:

After the first round, the three new clusters are: (1) {A1}, (2) {B1,A3,B2,B3,C2}, (3) {C1,A2}, and their centers are (1) (2, 10), (2) (6, 6), (3) (1.5, 3.5).

(b) The final three clusters.

Answer:

The final three clusters are: (1) {A1,C2,B1}, (2) {A3,B2,B3}, (3) {C1,A2}.