Kawsar Ahmed
IT-08050

Chapter-5 (Clustering) - I

* what is clustering ?

→ Many definitions for clusters have been proposed :

- Set of like elements. Elements from different clusters are not alike.
- The distance between points in a cluster is less than the distance between a point in the cluster and any point outside it.

Given a database $D = \{t_1, t_2, \ldots t_n\}$ of tuples and an integer value K, the clustering problem is to define a mapping $f: D \to \{1, \ldots K\}$ where each t_i is assigned to one cluster K_j, $1 \leq j \leq K$. A cluster, K_j, contains precisely those tuples mapped to it. That is $K_j = \{t_i \mid f(t_i) = K_j, 1 \leq i \leq n,$ and $t_i \in D\}$.

* what is application of clustering ?

→

1. clustering has been used in many application domains, including biology, medicine, authroplogy, marketing and Economics.

2. clustering applications include plant and animal classification, disease classification, image processing, pattern recognition, and document retrieval.

* Write down the advantages of clustering. or write down the basic features of clustering.

→

1. The (best) number of clusters is not known.
2. There may not be any a prioriti knowledge concerning the clusters.
3. Cluster results are dynamic.

* Write down the disadvantages of clustering.

→

1. Outlier handling is difficult.
2. Dynamic data in the database implies that cluster membership may change over time.
3. Interpreting the semantic meaning of each cluster may be difficult.
4. There is no correct answer to a clustering problem.
5. clustering can be viewed as similar to unsupervised learning.

Diameter: The diameter is the square root of the average mean squared distance between all pairs of points in the cluster.

$$\text{diameter}, = D_m = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(t_{mi}-t_{mj})^2}{(N)(N-1)}}$$

* Discuss different kinds of method to measure the distance between clusters.

→

1. Single link: Smallest distance between an element ~~an elemen~~ in one cluster and an element in the other. We thus have $dis(K_i, K_j) = \min(dis(t_{il}, t_{jm}))\ \forall t_{il} \in K_i \notin K_j$ and $\forall t_{jm} \in K_j \notin K_i$.

2. Complete link: Largest distance between an element in one cluster and an element in the other. We thus have $dis(K_i, K_j) = \max(dis(t_{il}, t_{jm}))\ \forall t_{il} \in K_i \notin K_j$ and $\forall t_{jm} \in K_j \notin K_i$

3. Average: Average distance between an element in one cluster and an element in the other. We thus have $dis(K_i, K_j) = mean(dis(t_{il}, t_{jm}))\ \forall t_{il} \in K_i \notin K_j$ and $\forall t_{jm} \in K_j \notin K_i$.

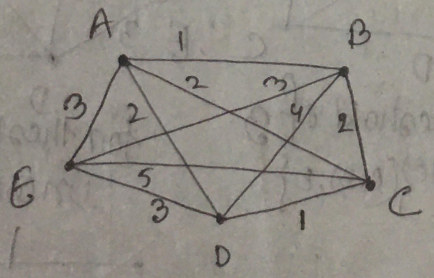
4. Centroid: If clusters have a representative centroid, then the centroid distance is defined as the distance between the centroids. We thus have $dis(K_i, K_j) = dis(C_i, C_j)$, where C_i is the centroid for K_i and similarly for C_j.

5. Medoid: Using a medoid to represent each cluster, the distance between the clusters can be defined by the distance between the medoids: $dis(K_i, K_j) = dis(m_i, m_j)$.

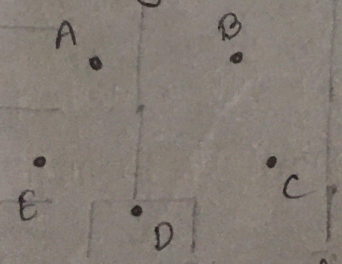
Example 5.3: for the below data draw the dendrogram for Single link, Complete link, Average link.

Item	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

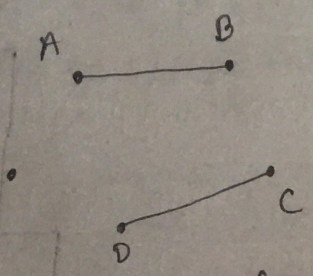
Soln:



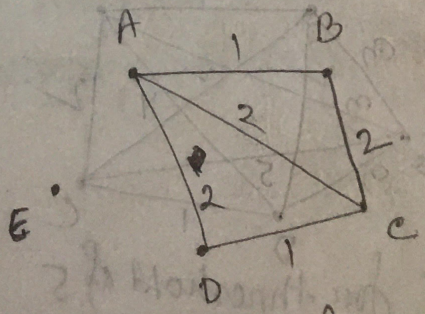
Single link:-



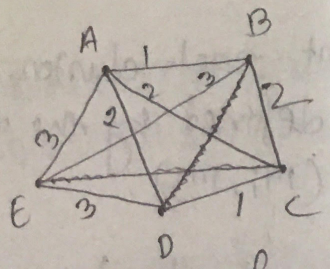
for threshold of 0



for threshold of 1



for threshold of 2



for threshold of 3

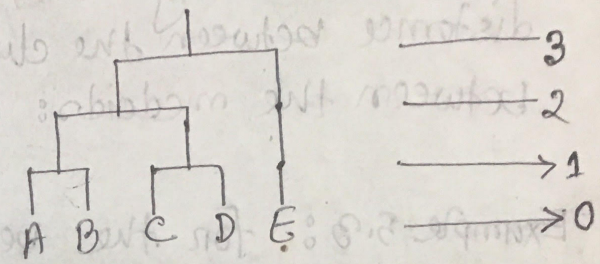
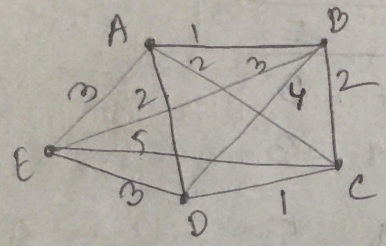
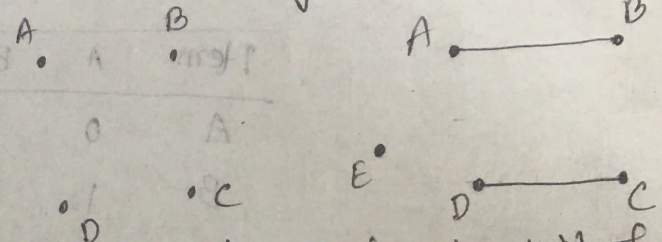


fig: Dendrogram for Single link.

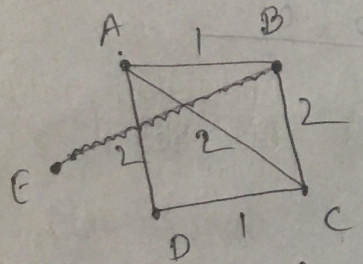
[2] Complete Link:



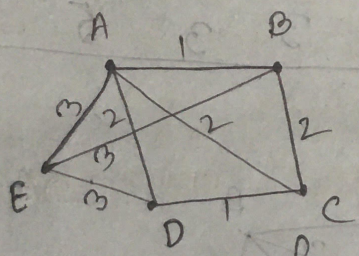


for threshold of 0
(Complete)

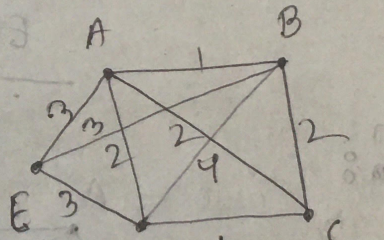
for threshold of 1
(Complete)



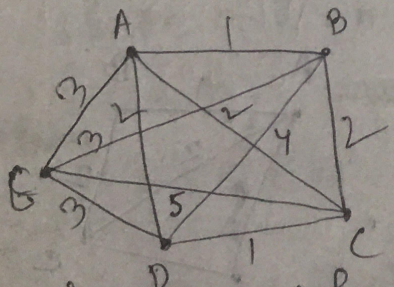
for threshold of 2
(uncomplete)



for threshold of 3
Complete {(A,B), E}



for threshold of 4
uncomplete



for threshold of 5

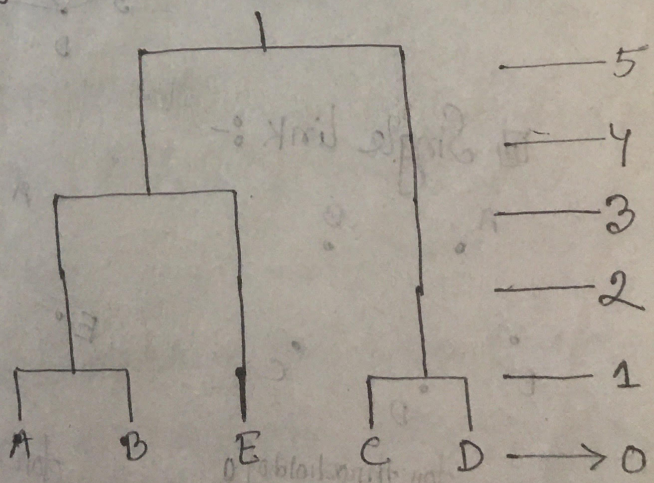
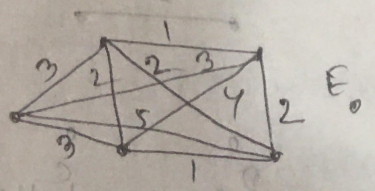
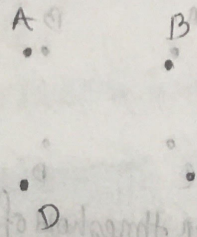


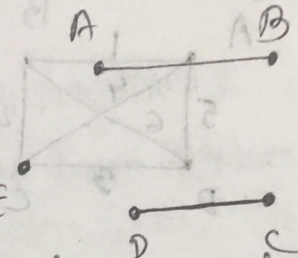
fig: Dendrowgram for Complete link

[iii] Average link:-

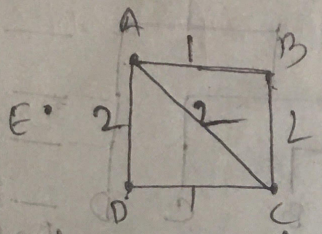




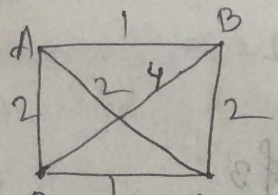
for threshold of 0



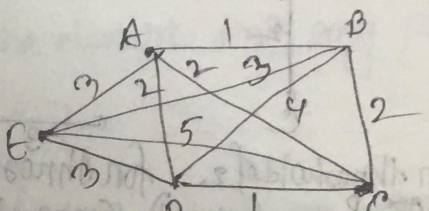
for threshold of 1
taken ∵ average = 0.5 < 1



for threshold of 2



to complete the graph we need the path BD and the average = $\frac{2+2+4+2}{4}$ = 2.5



To Complete the graph we need path BD and CE and the aver is = $\frac{3+3+5+3}{4}$ = 3.5

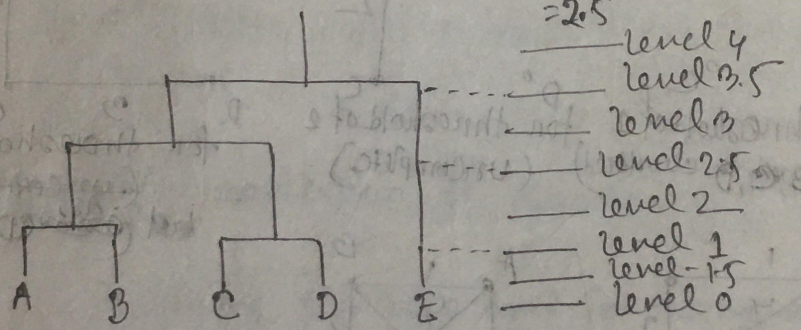


fig. Dendrograph for average link.

* Show the dendrogram created by the single, Complete and avera link clustering algorithms using the following adjacency matrix:

Item	A	B	C	D
A	0	1	4	5
B	1	0	2	6
C	2	2	0	3
D	6	6	3	0

Example 5.4: if $m_1 = 2$, $m_2 = 4$, then the following data items are clustered into two (2) classes.

$$\{2, 4, 10, 12, 3, 20, 30, 11, 25\}$$

Soln: given that,

$m_1 = 2$

$m_2 = 4$

m_1	m_2	K_1	K_2
2	4	{2, 3}	{4, 10, 12, 20, 30, 11, 25}
2.5	16	{2, 3, 4}	{10, 12, 20, 30, 11, 25}
3	18	{2, 3, 4, 10}	{12, 20, 30, 11, 25}
4.75	19.6	{2, 3, 4, 10, 11, 12}	{20, 30, 25}
7	25	{2, 3, 4, 10, 11, 12}	{20, 30, 25}

∴ $K_1 = \{2, 3, 4, 10, 11, 12\}$ and $K_2 = \{20, 30, 25\}$

(Ans).

বি: দ্রঃ- cluster two means K=2 (K_1, K_2). m_1 and m_2 এর মানের সাথে তুলনা করি। ঐ সব মানের নিকটবর্তী মান গুলো K_1 and বাকিগুলো K_2 তে যাবে। এখন K_1 ও K_2 এর মানগুলোর গড় বের করি এবং যথাক্রমে m_1 ও m_2 তে বসাই। এভাবে ততক্ষণ পর্যন্ত করতে হবে যতক্ষণ না পরপর দুইটি step-এ K_1 ও K_2 এর element সমান না হয়।

* clustering the following data items for threshold value 2 (using nearest neighbour Algorithm).

→

Item	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

Solⁿ:

initially, $K_1 = \{A\}$

Now the distance between A to B is $1 \leq 2$ (threshold value)

So B inserted into K_1.

Now $K_1 = \{A, B\}$

∴ The distance between A to C is $= 2 \leq 2$

∴ The distance between B to C is $= 2 \leq 2$

So C inserted into K_1

∴ $K_1 = \{A, B, C\}$

∴ The distance between A to D is $= 2 \leq 2$

∴ The distance between B to D is $= 4$

∴ The distance between C to D is $= 1 \leq 2$

So D inserted into K_1

$\therefore K_1 = \{A, B, C, D\}$

The distance between A to E is = 3
The distance between B to E is = 3
The distance between C to E is = 5
The distance between D to E is = 3
~~The distance between~~ E There is no value below of threshold value.

So E is inserted into new cluster (K_2).

$\therefore K_2 = \{E\}$

$\therefore K_1 = \{A, B, C, D\}$, $K_2 = \{E\}$ Ans.

* What is the usefulness or effectiveness of using large database in clustering algorithm?

→

A clustering algorithm should:

1. require no more than one scan of the database.
2. have the ability to provide status and "best" answer so far during the algorithm execution.
3. be suspendable, stoppable and resumable.
4. be able to update the results incrementally as data are added or removed from the database.