Hierarchical clustering

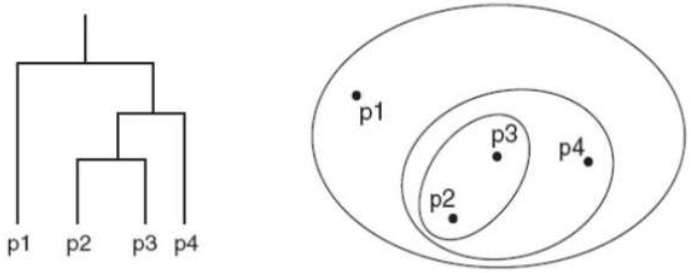
In this type of clustering, the set of clusters are nested clusters that are organized in the form of a tree known as dendrogram. Each node in the tree is the union of its children and root of the tree is the cluster containing all the objects.

There are two types of hierarchical clustering. They are:

1. Agglomerative hierarchical clustering
2. Divisive hierarchical clustering

Agglomerative hierarchical clustering

Start with the points as individual clusters and at each step, merge the closest pair of clusters.

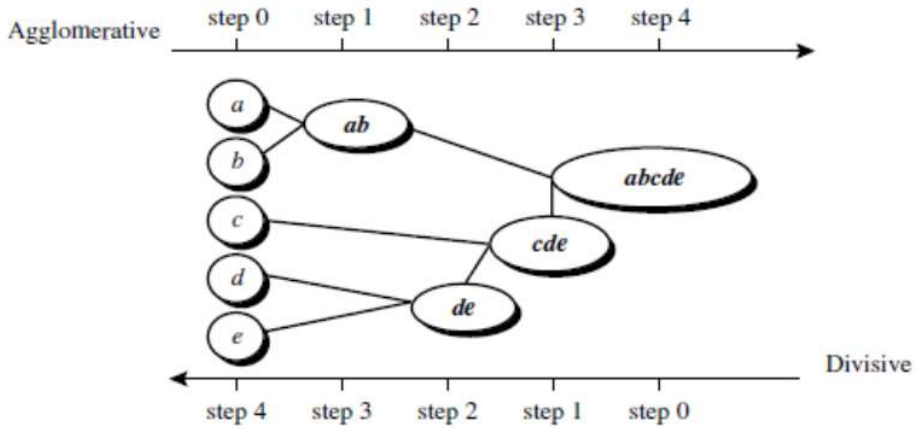


(a) Dendrogram. (b) Nested cluster diagram.

A hierarchical clustering of four points shown as a dendrogram and as nested clusters.

Divisive hierarchical clustering

Start with one i.e.; group all data objects into a single cluster, at each step, split a cluster until only single cluster of individual points remains.



Basic agglomerative hierarchical clustering algorithm

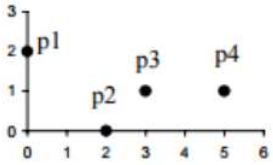
Starting with initial points as clusters, successively merge the two closest clusters until only one cluster remains.

Algorithm 8.3 Basic agglomerative hierarchical clustering algorithm.

1: Compute the proximity matrix, if necessary.
2: **repeat**
3: Merge the closest two clusters.
4: Update the proximity matrix to reflect the proximity between the new cluster and the original clusters.
5: **until** Only one cluster remains.

cluster proximity

Cluster proximity is defined as similarity or dissimilarity between elements or clusters. They are generally defined by proximity matrix.



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0.000	2.828	3.162	5.099
p2	2.828	0.000	1.414	3.162
p3	3.162	1.414	0.000	2.000
p4	5.099	3.162	2.000	0.000

Four points and their corresponding data and proximity (distance) matrices.

Proximity matrix is a matrix containing distance between elements.

Methods for defining the proximity between the clusters

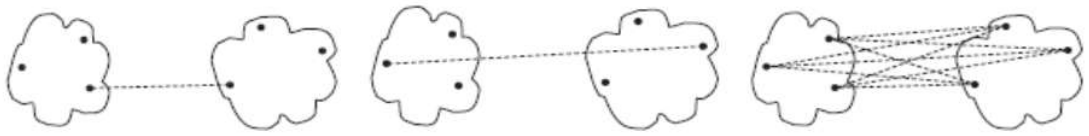
Proximity between clusters can be defined in three ways. They are:

1. Max
2. Min
3. Group average

Min defines cluster proximity between the closest two points that are in different clusters. This type of technique is also known as **single link technique.**

Max defines cluster proximity between the farthest two points that are in different clusters. This type of technique is also known as **complete link technique.**

Group average defines cluster proximity to be the average proximities of all pairs of points from different clusters.



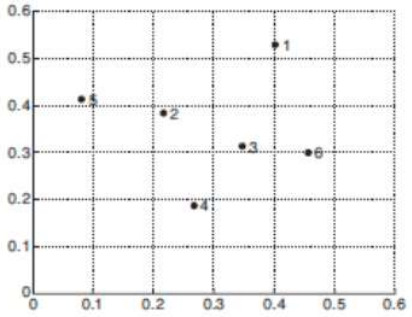
(a) MIN (single link.) (b) MAX (complete link.) (c) Group average.

Graph-based definitions of cluster proximity

Proximity between two clusters using MIN technique (Or)

Single- Link hierarchical clustering

Let us consider a data set with 6 data points.



Set of 6 two-dimensional points.

Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

xy coordinates of 6 points.

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Euclidean distance matrix for 6 points.

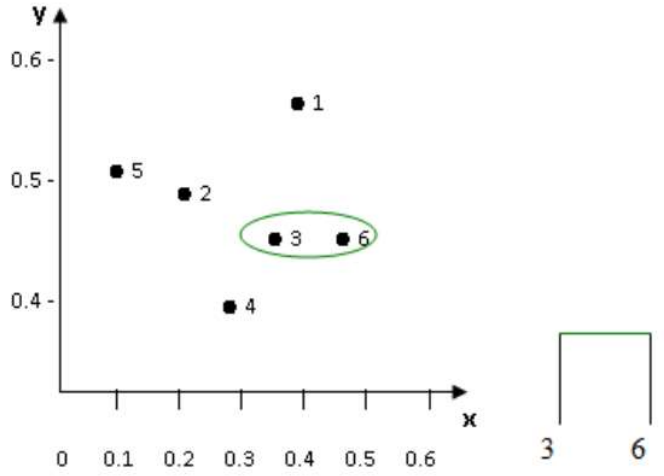
Min or single link technique

The proximity of two clusters is defined as the minimum of the distance between any two points I the two different clusters. The single link technique is good at handling non-elliptical shapes, but is sensitive to noise and outliers.

	P1	P2	P3	P4	P5	P6
P1	0	0.24	0.22	0.37	0.34	0.23
P2	0.24	0	0.15	0.20	0.14	0.25
P3	0.22	0.15	0	0.15	0.28	0.11
P4	0.37	0.20	0.15	0	0.29	0.22
P5	0.34	0.14	0.28	0.29	0	0.39
P6	0.23	0.25	0.11	0.22	0.39	0

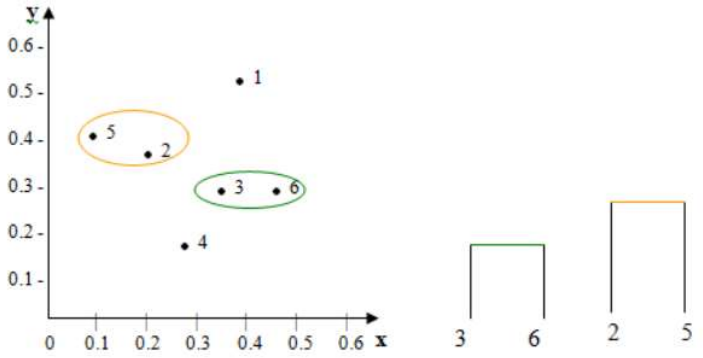
In the above table, p3, p6 is having the lowest distance, so merge the data points into a single cluster.

	P1	P2	P3P6	P4	P5
P1	0	0.24	0.22	0.37	0.34
P2	0.24	0	0.15	0.20	0.14
P3P6	0.22	0.15	0	0.15	0.28
P4	0.37	0.20	0.15	0	0.29
P5	0.34	0.14	0.28	0.29	0



The next lowest distance is for P2P5, so merge those two data points into a single cluster. The matrix obtains as follows:

	P1	P2P5	P3P6	P4
P1	0	0.24	0.22	0.37
P2P5	0.24	0	0.15	0.20
P3P6	0.22	0.15	0	0.15
P4	0.37	0.20	0.15	0



The distance between (p3, p6) and (p2, p5) would be calculated as follows:

dist((p3, p6), (p2, p5)) =MIN (*dist*(p3, p2) , *dist*(p6, p2), *dist*(p3, p5), *dist*(p6, p5))

=MIN (0.15, 0.25, 0.28, 0.39)

=0.15

dist((p3, p6), (p1)) =MIN (*dist*(p3, p1) , *dist*(p6, p1)

=MIN (0.22, 0.23)

=0.22

dist((p3, p6), (p4)) =MIN (*dist*(p3, p4) , *dist*(p6, p4)

=MIN (0.15, 0.22)

=0.15

dist((p2, p5), (p1)) =MIN (*dist*(p2, p1) , *dist*(p5, p1)

=MIN (0.24, 0.34)

=0.24

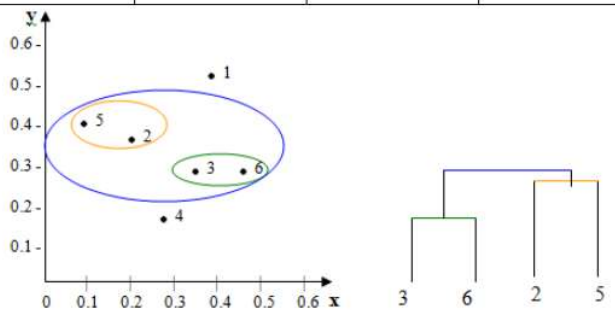
dist((p2, p5), (p4)) =MIN (*dist*(p2, p4) , *dist*(p5, p4)

=MIN (0.20, 0.29)

=0.20

So, looking at the last distance matrix above, we see that (p2, p5) and (p3, p6) have the smallest distance from all - 0.15. We also notice that p4 and (p3, p6) have the same distance - 0.15. In that case, we can pick either one. We choose (p2, p5) and (p3, p6). So, we merge those two in a single cluster, and re-compute the distance matrix.

	P1	P2P5P3P6	P4
P1	0	0.22	0.37
P2P5P3P6	0.22	0	0.15
P4	0.37	0.20	0



The distance between (P2, P5, P3, P6) and P1 would be calculated as follows:

dist((P2, P5, P3, P6), (p1)) =MIN (*dist*(p2, p1) , *dist*(p5, p1), *dist*(p3, p1) , *dist*(p6, p1))

=MIN (0.24, 0.34, 0.22, 0.23)

=0.22

The distance between (P2, P5, P3, P6) and P4 would be calculated as follows:

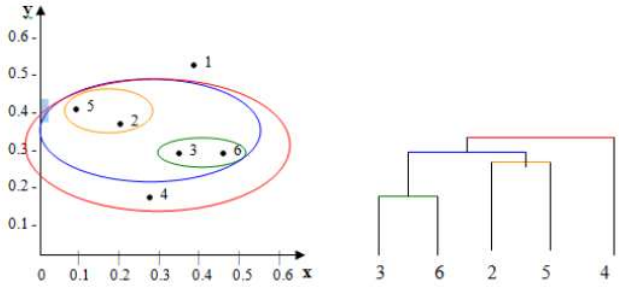
dist((P2, P5, P3, P6), (p4)) =MIN (*dist*(p2, p4) , *dist*(p5, p4), *dist*(p3, p4) , *dist*(p6, p4))

=MIN (0.20, 0.29, 0.15, 0.22)

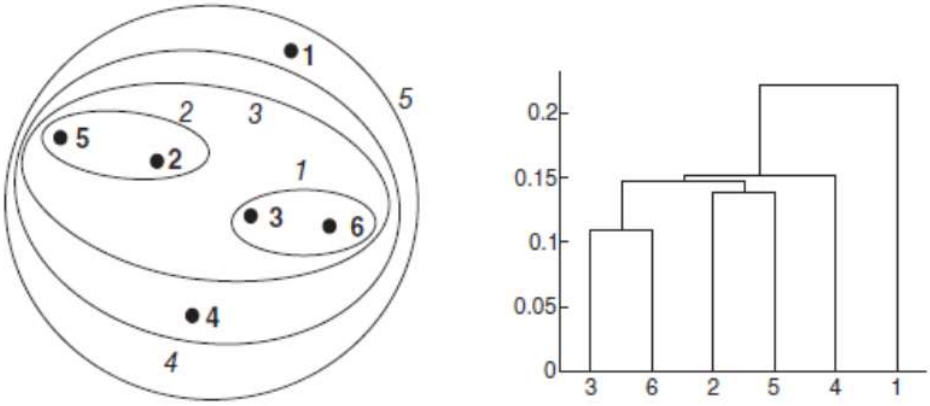
=0.15

So, looking at the last distance matrix above, we see that (p2, p5, p3, p6) and p4 have the smallest distance from all - 0.15. So, we merge those two in a single cluster, and re-compute the distance matrix.

	P1	P2P5P3P6P4
P1	0	0.22
P2P5P3P6P4	0.22	0



Finally merge clusters (P2, P5, P3, P6, P4) and P1. The clusters and dendrogram are formed as follows:



(a) Single link clustering. (b) Single link dendrogram.

Fig: single link clustering and the dendrogram of the 6 data points

Proximity between two clusters using MAX technique (Or)

Complete Link hierarchical clustering (Or)

Clique technique for clustering.

The proximity of two clusters is defined as the maximum of the distance between any two points in the two different clusters. Complete link is less susceptible to noise and outliers, but it can break large clusters and it favors globular shapes.

The procedure is same as Min but the difference is max distance is considered while combing to closest clusters.

For Example,

After the clusters (P2, P5) and (P3, P6) are formed, the distances are calculated as follows:

dist((p3, p6), (p2, p5)) =MAX (*dist*(p3, p2) , *dist*(p6, p2), *dist*(p3, p5), *dist*(p6, p5))

=MAX (0.15, 0.25, 0.28, 0.39)

=0.39

dist((p3, p6), (p1)) =MAX (*dist*(p3, p1) , *dist*(p6, p1)

=MAX (0.22, 0.23)

=0.23

dist((p3, p6), (p4)) =MAX (*dist*(p3, p4) , *dist*(p6, p4)

=MAX (0.15, 0.22)

=0.22

dist((p2, p5), (p1)) =MAX (*dist*(p2, p1) , *dist*(p5, p1)

=MAX (0.24, 0.34)

=0.34

dist((p2, p5), (p4)) =MAX (*dist*(p2, p4) , *dist*(p5, p4)

=MAX (0.20, 0.29)

=0.29

Here the proximity is defined as maximum of distance but minimum of similarity. The lowest among all the distances is 0.22. So, merge clusters (P3, P6) and P4.

The distance between (P3, P6, P4) and remaining points are calculated as follows:

dist((p3, p6, p4), (p2, p5)) =MAX (*dist*(p3, p2) , *dist*(p6, p2), *dist*(p4, p2), *dist*(p3, p5) , *dist*(p6, p5), *dist*(p4, p5))

=MAX (0.22, 0.23, 0.37, 0.28, 0.39, 0.29)

=0.39

dist((p3, p6, p4), (p1)) =MAX (*dist*(p3, p1) , *dist*(p6, p1), *dist*(p4, p1))

=MAX (0.22, 0.23, 0.37)

=0.37

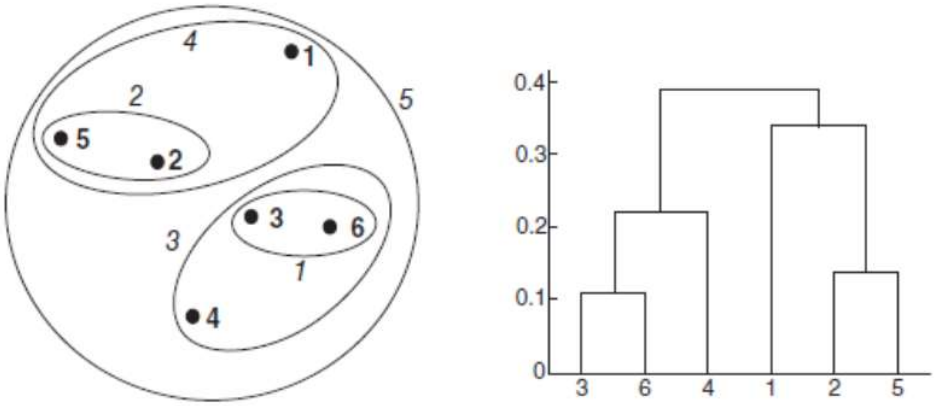
dist((p2, p5), (p1)) =MAX (*dist*(p2, p1) , *dist*(p5, p1)

=MAX (0.24, 0.34)

=0.34

The lowest among all the distances is 0.34. So, merge clusters (P2, P5) and P1.

Now merge clusters (P2, P5, P1) and (P3, P6, P4). The final clusters and dendrogram are formed as follows:



(a) Complete link clustering. (b) Complete link dendrogram.

Fig: Complete link clustering and the dendrogram of the 6 data points