Artificial Intelligence
——Hierarchical Clustering

Yanghui Rao
Assistant Prof., Ph.D
School of Data and Computer Science,
Sun Yat-sen University
raoyangh@mail.sysu.edu.cn

Hierarchical Clustering

- A hierarchical clustering is a set of nested clusters that are organized as a tree

- There are 2 basic approaches for generating a hierarchical clustering
  - Agglomerative (凝聚式)
  - Divisive (分裂式)

Hierarchical Clustering

- In ***agglomerative*** hierarchical clustering, we start with the points as individual clusters
- At each step, we merge the closest pair of clusters
- This requires defining a notion of cluster distance (如何计算簇与簇之间的距离?)

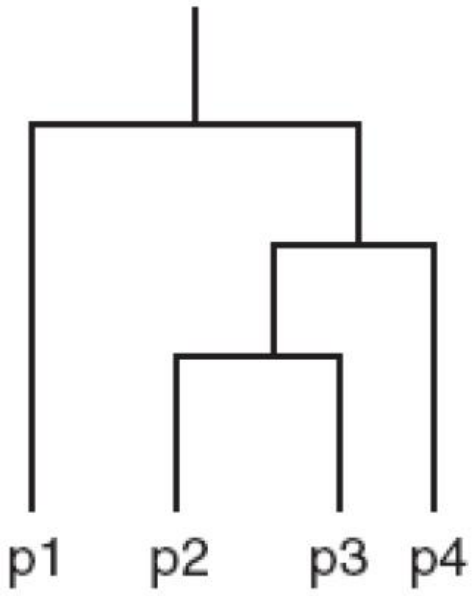
Hierarchical Clustering

- In ***divisive*** hierarchical clustering, we start with one, all-inclusive cluster.
- At each step, we split a cluster.
- This process continues until only singleton clusters of individual points remain (每个簇只包含一个样本/对象).
- In this case, we need to decide
  - Which cluster to split at each step and
  - How to do the splitting.

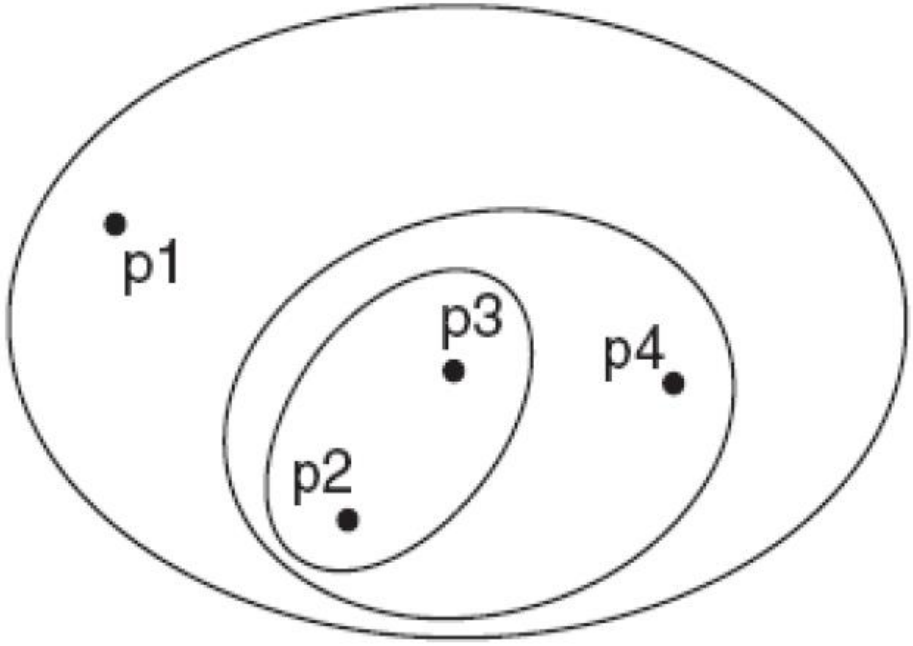
Hierarchical Clustering

- A hierarchical clustering is often displayed graphically using a tree-like diagram called the dendrogram (树状图).
- The dendrogram displays both
  - the cluster-subcluster relationships and
  - the order in which the clusters are merged (agglomerative) or split (divisive).
- For sets of 2-D points, a hierarchical clustering can also be graphically represented using a nested cluster diagram.

Hierarchical Clustering



(a) Dendrogram.



(b) Nested cluster diagram.

Hierarchical Clustering

- The general agglomerative hierarchical clustering is summarized as follows:
- Compute the distance matrix.
- Repeat
  - Merge the closest two clusters
  - Update the distance matrix to reflect the distance between the new cluster and the original clusters.
- Until only one cluster remains

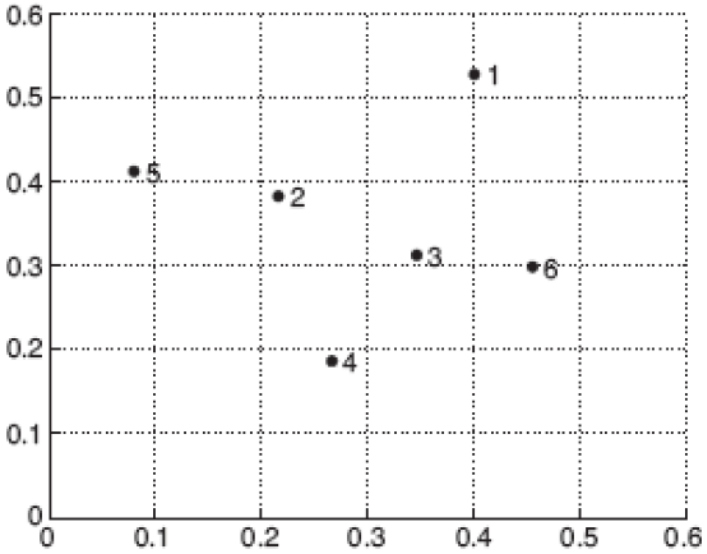
Hierarchical Clustering

- Different definitions of cluster distance leads to different versions of hierarchical clustering.

- These versions include
  - Single link (单连接) or MIN
  - Complete link (全连接) or MAX
  - Group average (组平均)

Hierarchical Clustering

- We consider the following set of points.
- The Euclidean distance matrix for these data points is shown in the following slide.



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

Hierarchical Clustering

	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

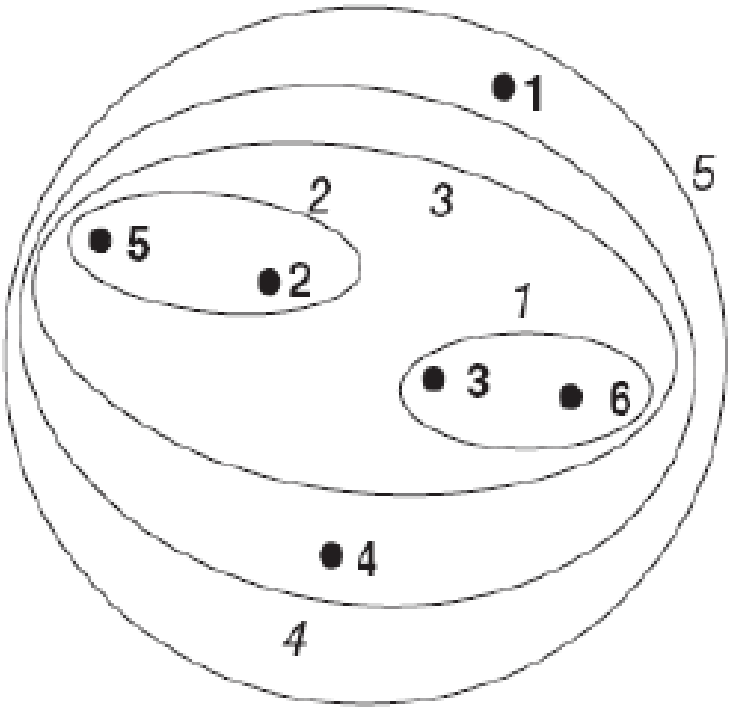
Single Link

- We now consider the single link or MIN version of hierarchical clustering.
- In this case, the distance of two clusters is defined as the minimum of the distance between any two points in the two different clusters.
- This technique is good at handling non-elliptical (非球状的) shapes.
- However, it is sensitive to noise and outliers.

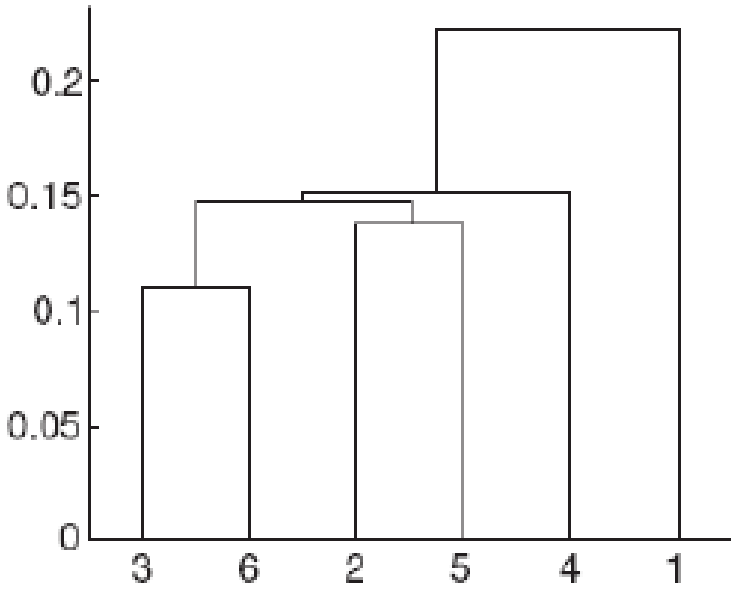
Single Link

- The following figure shows the result of applying the single link technique to our example data.
- The left figure shows the nested clusters as a sequence of nested ellipses.
- The numbers associated with the ellipses indicate the order of the clustering.
- The right figure shows the same information in the form of a dendrogram.
- The height at which two clusters are merged in the dendrogram reflects the distance of the two clusters (树状图的高度即距离值).

Single Link



(a) Single link clustering.



(b) Single link dendrogram.

Single Link

- For example, we see that the distance between points 3 and 6 is 0.11.
- That is the height at which they are joined into one cluster in the dendrogram.
- As another example, why is the distance between clusters {3,6} and {2,5} being 0.15?

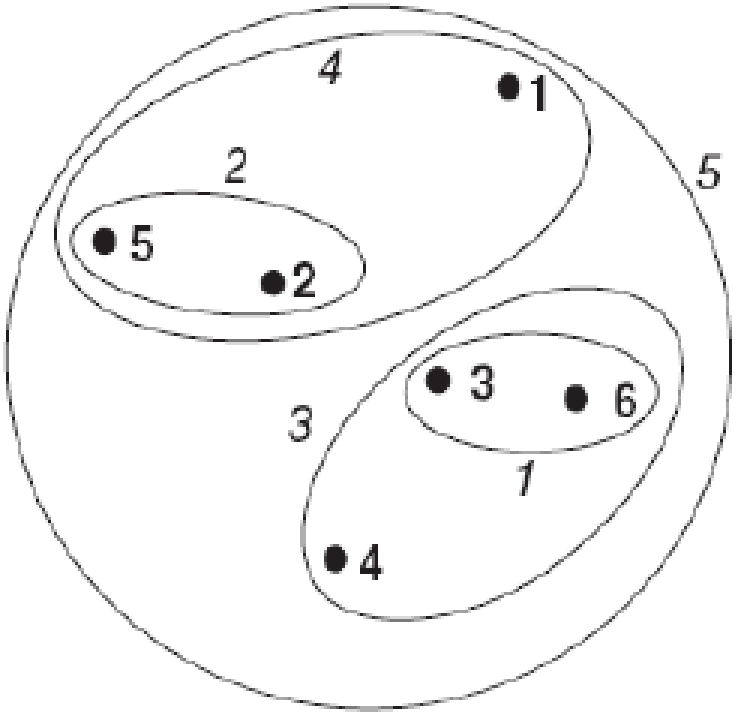
Complete Link

- We now consider the complete link or MAX version of hierarchical clustering.
- In this case, the distance 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 (不敏感) to noise and outliers, but it tends to produce clusters with globular (球状) shapes.

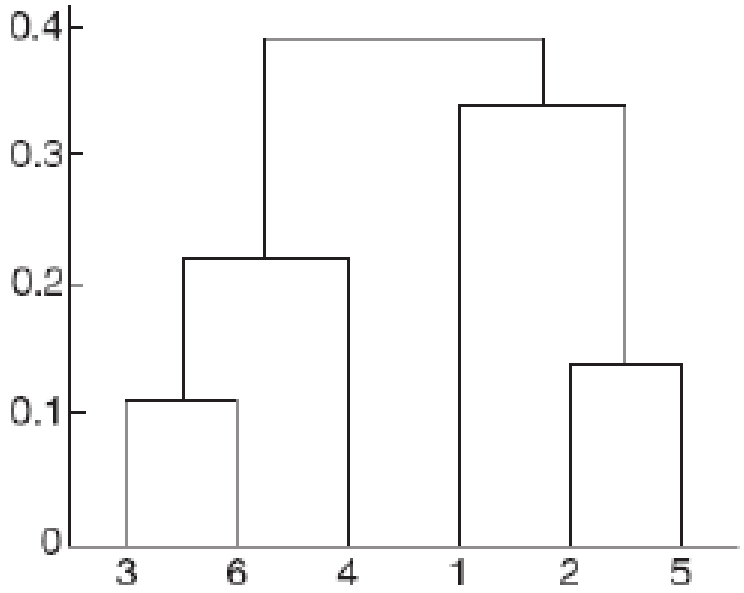
Complete Link

- The following figure shows the results of applying the complete link approach to our sample data points.
- As with single link (or MIN), points 3 and 6 are merged first.
- Points 2 and 5 are then merged.
- After that, {3,6} is merged with {4}.

Complete Link



(a) Complete link clustering.



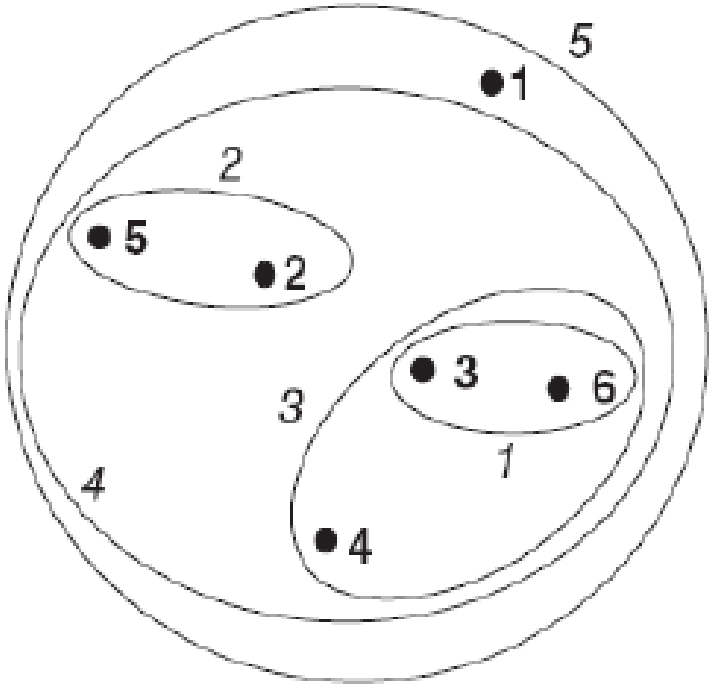
(b) Complete link dendrogram.

Group Average

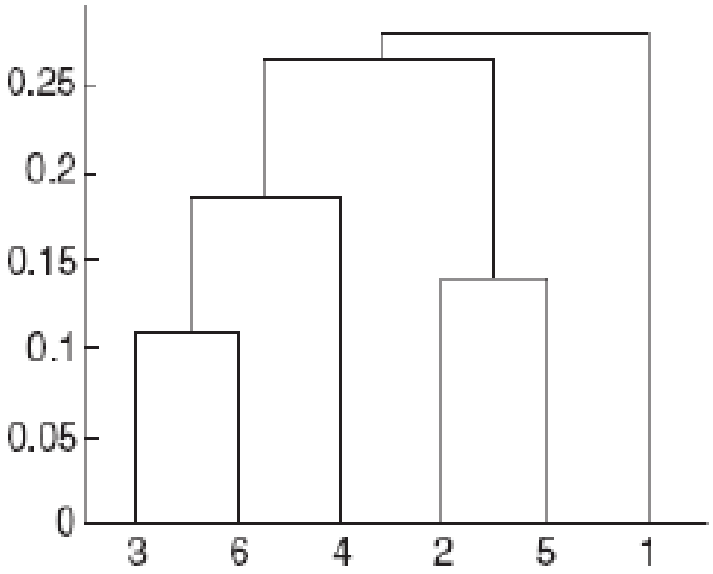
- We now consider the group average version of hierarchical clustering.
- In this case, the distance of two clusters is defined as the average pairwise distance among all pairs of points in the different clusters.
- This is an intermediate approach between the single and complete link approaches.

Group Average

- The following figure shows the results of applying the group average to our sample data.



(a) Group average clustering.



(b) Group average dendrogram.

Key Issues

- Hierarchical clustering is effective when the underlying application requires the creation of a multi-level structure.
- However, they are expensive in terms of their computational and storage requirements.
- In addition, once a decision is made to merge two clusters, it cannot be undone at a later time.