Clustering: Unsupervised Learning

IF-3270 Pembelajaran Mesin

Teknik Informatika ITB

Modul 7: Clustering

04 Hierarchical Clustering

IF3270 - Pembelajaran Mesin
(Machine Learning)

Fariska Z. Ruskanda, S.T., M.T.
(fariska@informatika.org)

KK IF -Teknik Informatika - STEI ITB

Outline

Hierarchical Clustering

Agglomerative Clustering

Divisive Clustering

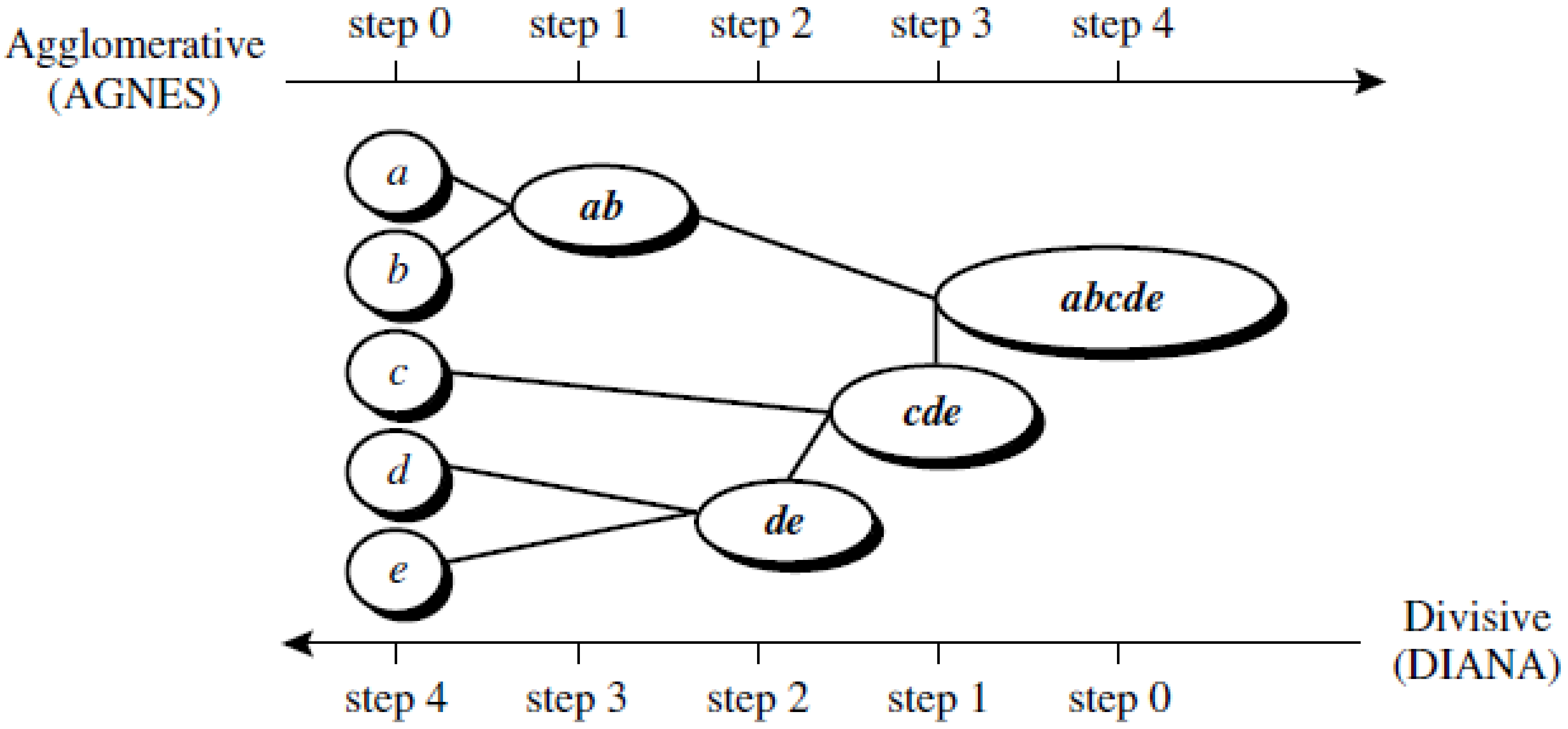
Hierarchical Clustering

A hierarchical clustering method works by grouping data objects into a hierarchy or “tree” of clusters.

a. Agglomerative hierarchical clustering method uses a bottom-up strategy.
b. Divisive hierarchical clustering method employs a top-down strategy.

Han & Kamber (2006)

Hierarchical Clustering



Han & Kamber (2006)

Agglomerative Clustering Algorithm

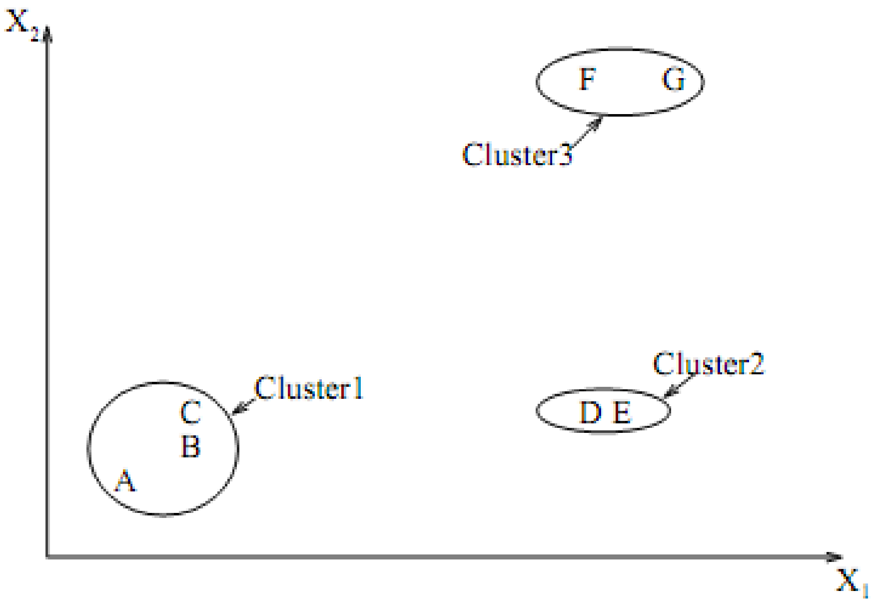
1) Start with N singleton clusters. Calculate the proximity matrix for the N clusters.
2) Search the minimal distance

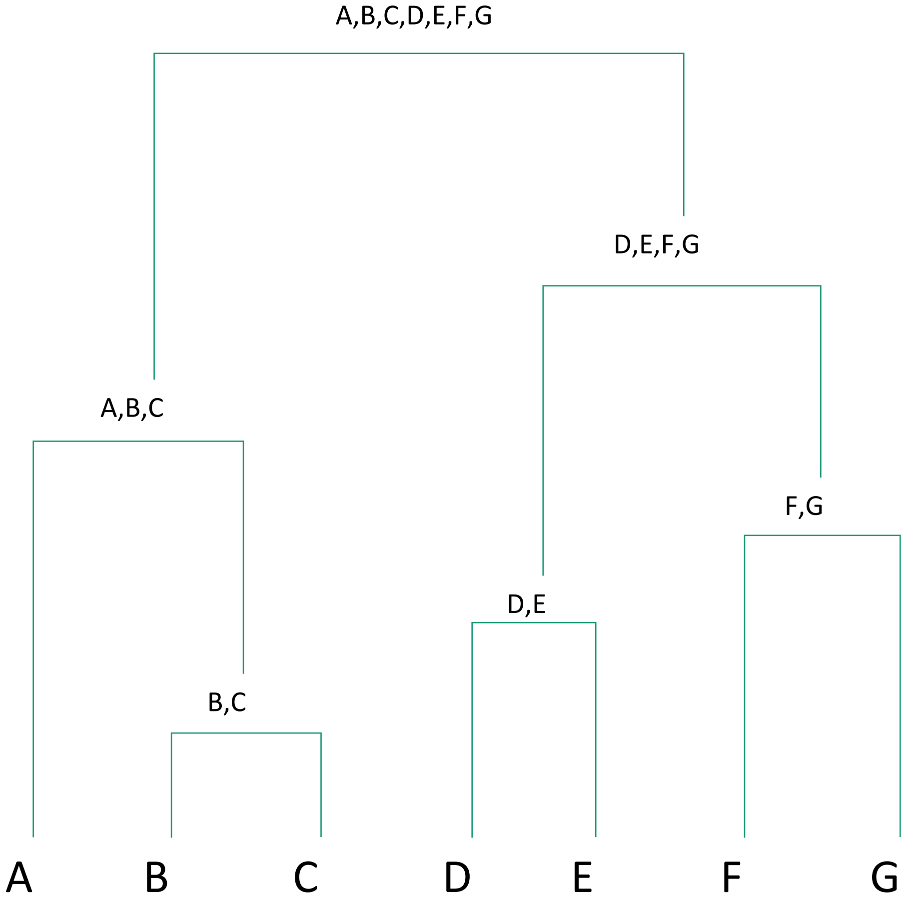
$$D(C_i, C_j) = \min_{\substack{1 \le m,l \le N \\ m \ne l}} D(C_m, C_l)$$

where $D(*, *)$ is the distance function discussed before, in the proximity matrix, and combine cluster C_i and C_j to form a new cluster.
3) Update the proximity matrix by computing the distances between the new cluster and the other clusters.
4) Repeat steps 2)–3) until all objects are in the same cluster.

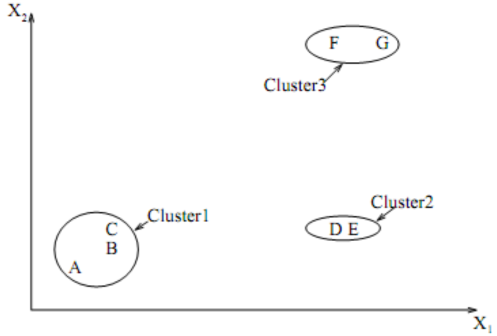
Sumber: Xu & Wunsch (2005); Jain dkk (1999)

Ilustrasi Agglomerative HC





Ilustrasi Agglomerative HC (lanj)



	A	B	C	D	E	F	G
A							
B							
C							
D							
E							
F							
G							

	A	B,C	D	E	F	G
A						
B,C						
D						
E						
F						
G						

	A	B,C	D,E	F	G
A					
B,C					
D,E					
F					
G					

	A	B,C	D,E	F,G
A				
B,C				
D,E				
F,G				

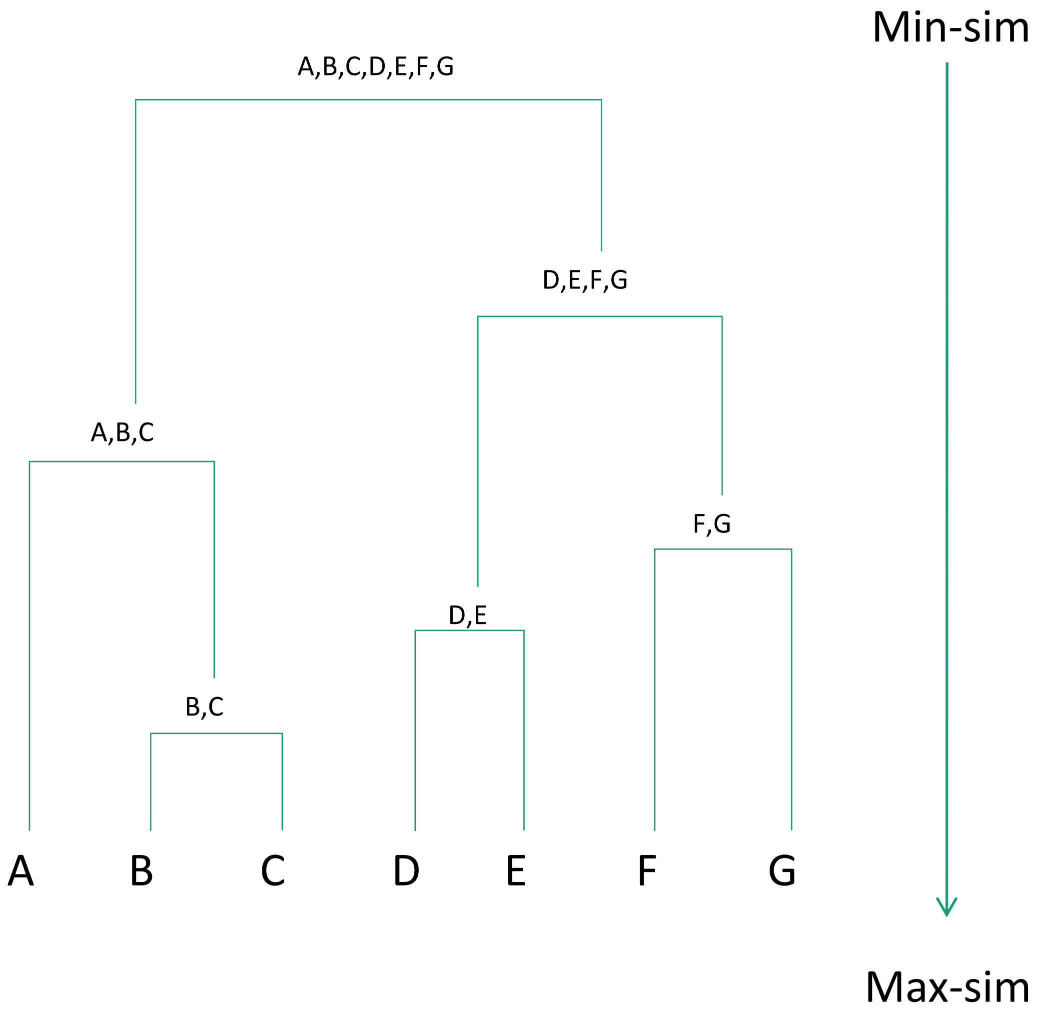
- Iterasi 0: (A),(B),(C),(D),(E),(F),(G)
- Iterasi 1 : (A),(B,C),(D),(E),(F),(G)
- Iterasi 2 : (A),(B,C),(D,E),(F),(G)
- Iterasi 3 : (A),(B,C),(D,E),(F,G)
- Iterasi 4 : (A,(B,C)),(D,E),(F,G)
- Iterasi 5 : (A,(B,C)),((D,E),(F,G))
- Iterasi 6 : ((A,(B,C)),((D,E),(F,G)))

	A, (B,C)	D,E	F,G
A, (B,C)			
D,E			
F,G			

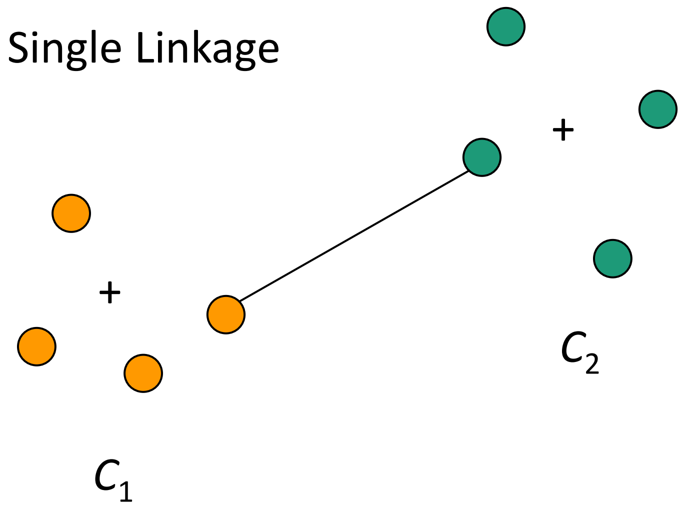
	A, (B,C)	(D,E), (F,G)
A, (B,C)		
(D,E), (F,G)		

Dendogram

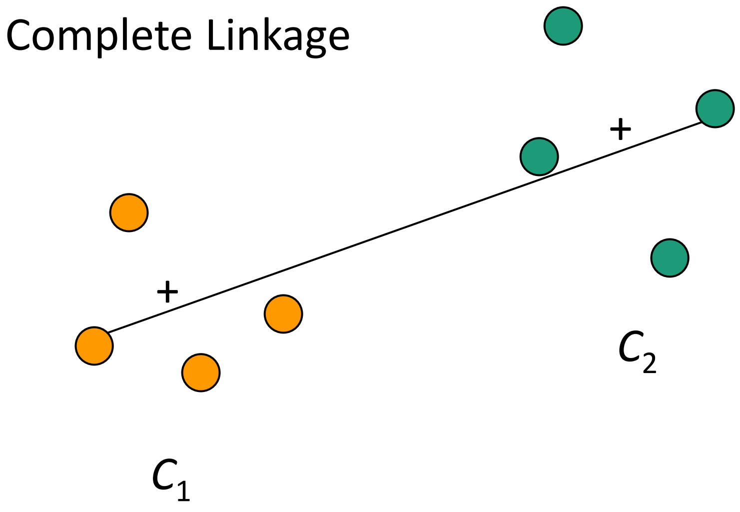
Iterasi 0: (A),(B),(C),(D),(E),(F),(G)
Iterasi 1 : (A),(B,C),(D),(E),(F),(G)
Iterasi 2 : (A),(B,C),(D,E),(F),(G)
Iterasi 3 : (A),(B,C),(D,E),(F,G)
Iterasi 4 : (A,(B,C)),(D,E),(F,G)
Iterasi 5 : (A,(B,C)),((D,E),(F,G))
Iterasi 6 : ((A,(B,C)),((D,E),(F,G)))



Linkage: Single, Complete, Average, Average Group



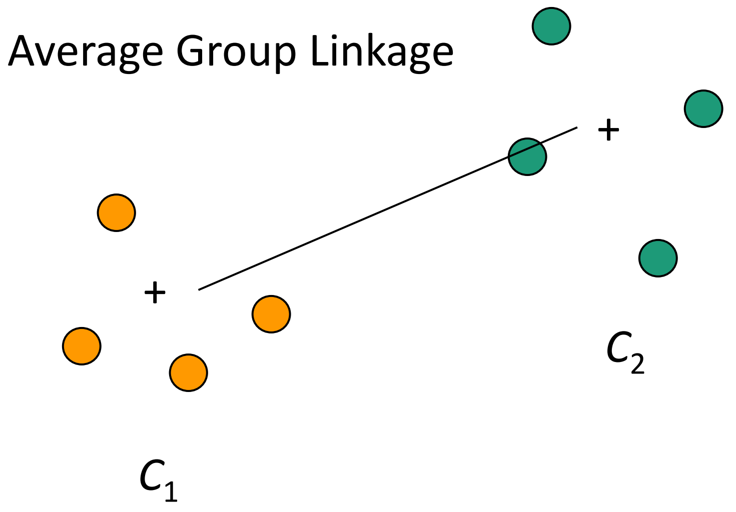
Dissimilarity between two clusters = Minimum dissimilarity between the members of two clusters



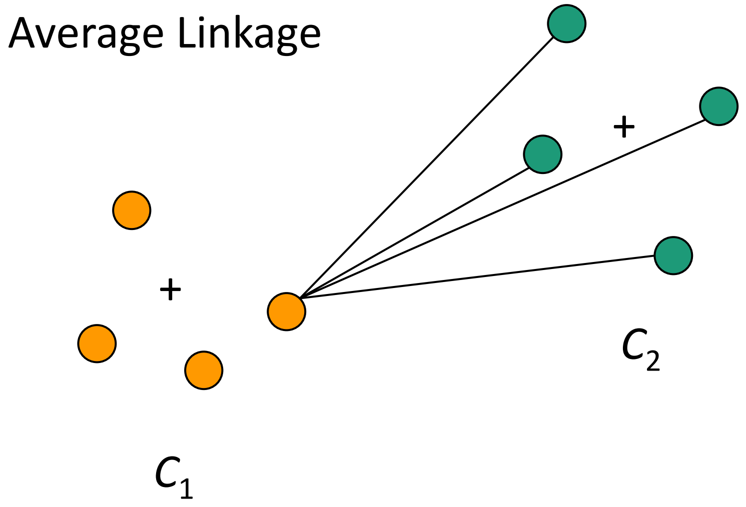
Dissimilarity between two clusters = Maximum dissimilarity between the members of two clusters

Sumber: Hong (2005)

Linkage: Single, Complete, Average, Average Group (lanjutan)



Dissimilarity between two clusters = Distance between two cluster means.



Dissimilarity between two clusters = Averaged distances of all pairs of objects (one from each cluster).

Sumber: Hong (2005)

Single vs Complete-Link

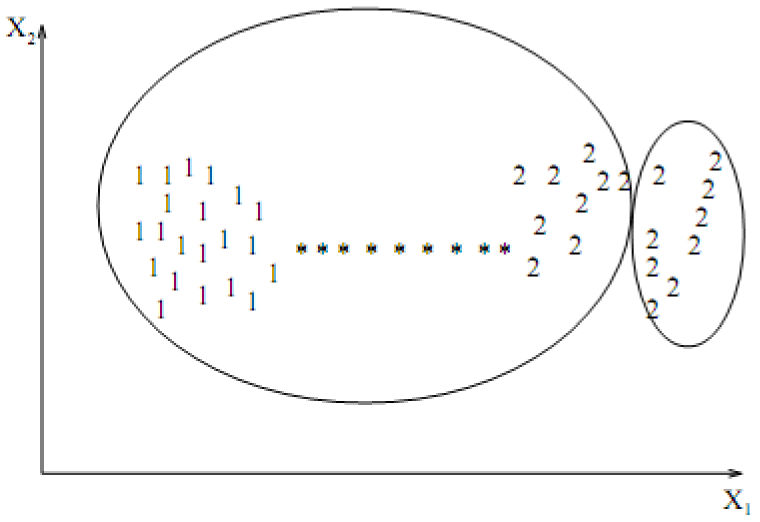


Figure 12. A single-link clustering of a pattern set containing two classes (1 and 2) connected by a chain of noisy patterns (*).

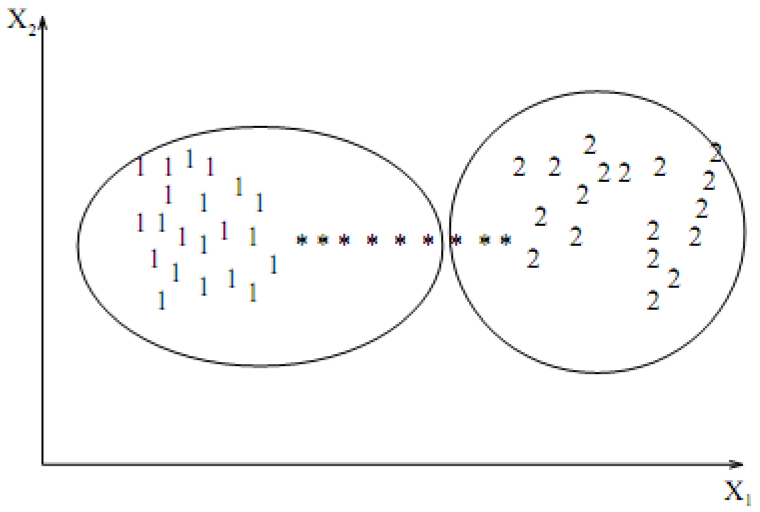


Figure 13. A complete-link clustering of a pattern set containing two classes (1 and 2) connected by a chain of noisy patterns (*).

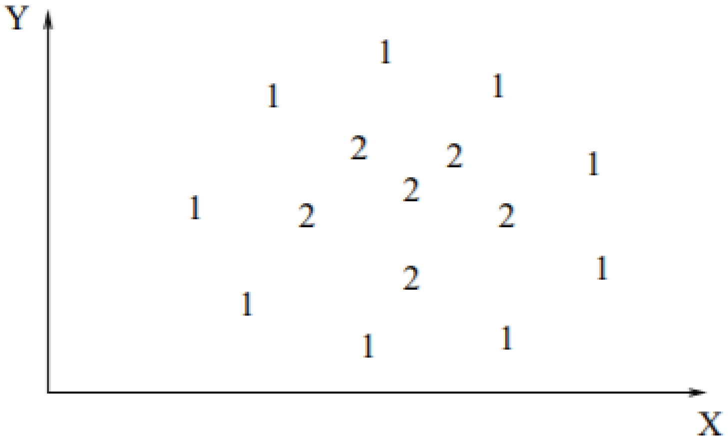
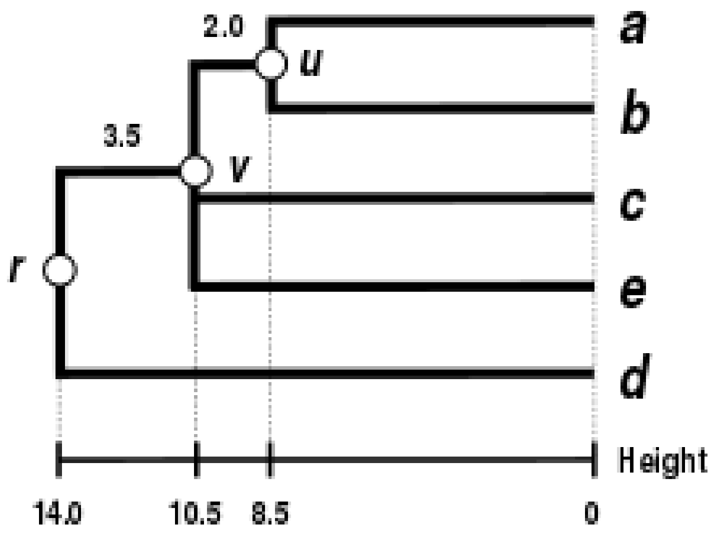


Figure 11. Two concentric clusters.

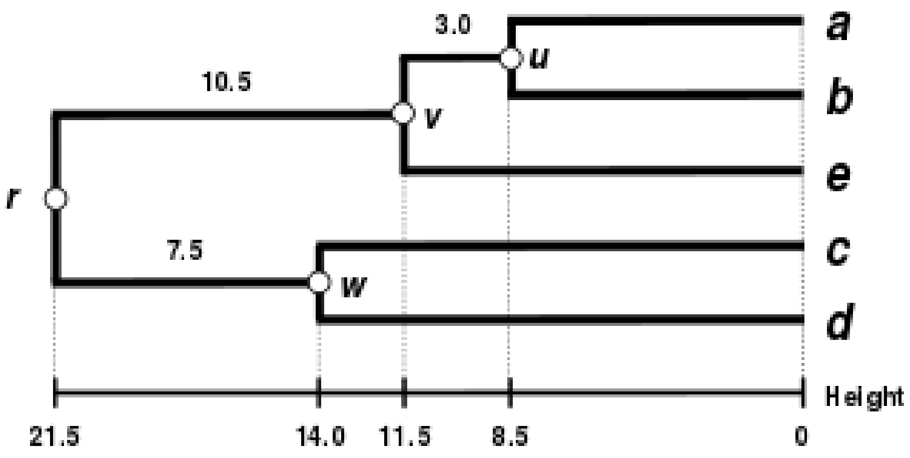
- Single link clustering suffers from a chaining effect.
- From a pragmatic viewpoint, it has been observed that the complete-link algorithm produces more useful hierarchies in many applications that single-link alg.

Sumber: Jain dkk (1999)

Single vs Complete-Link



Single-link



Complete-link

Divisive Clustering

- In the beginning, the entire data set belongs to a cluster and a procedure successively divides it until all clusters are singleton clusters.
- Divisive clustering is not commonly used in practice:
  - For a cluster with N objects, there are $2^{N-1}-1$ possible two-subset divisions, which is very expensive in computation (Xu & Wunsch, 2005).

05 Cluster Evaluation

IF3270 Pembelajaran Mesin