Hierarchical Clustering

Lecture 9

Marina Santini

Acknowledgements
Slides borrowed and adapted from:
Data Mining by I. H. Witten, E. Frank and M. A. Hall

Lecture 9: Required Reading

Witten et al. (2011: 273-284)

Outline

- Dendogram
- Agglomerative clustering
- Distance measures
- Cut-off

Hierarchical clustering

- Recursively splitting clusters produces a hierarchy that can be represented as a *dendogram*
  - Could also be represented as a Venn diagram of sets and subsets (without intersections)
  - Height of each node in the dendogram can be made proportional to the dissimilarity between its children

Agglomerative clustering

- Bottom-up approach
- Simple algorithm
  - Requires a distance/similarity measure
  - Start by considering each instance to be a cluster
  - Find the two closest clusters and merge them
  - Continue merging until only one cluster is left
  - The record of mergings forms a hierarchical clustering structure – a *binary dendogram*

Distance measures

- *Single-linkage*
  - Minimum distance between the two clusters
  - Distance between the clusters closest two members
  - Can be sensitive to outliers
- *Complete-linkage*
  - Maximum distance between the two clusters
  - Two clusters are considered close only if all instances in their union are relatively similar
  - Also sensitive to outliers
  - Seeks compact clusters

Distance measures cont.

- Compromise between the extremes of minimum and maximum distance
- Represent clusters by their centroid, and use distance between centroids – *centroid linkage*
  - Works well for instances in multidimensional Euclidean space
  - Not so good if all we have is pairwise similarity between instances
- Calculate average distance between each pair of members of the two clusters – *average-linkage*
- Technical deficiency of both: results depend on the numerical scale on which distances are measured

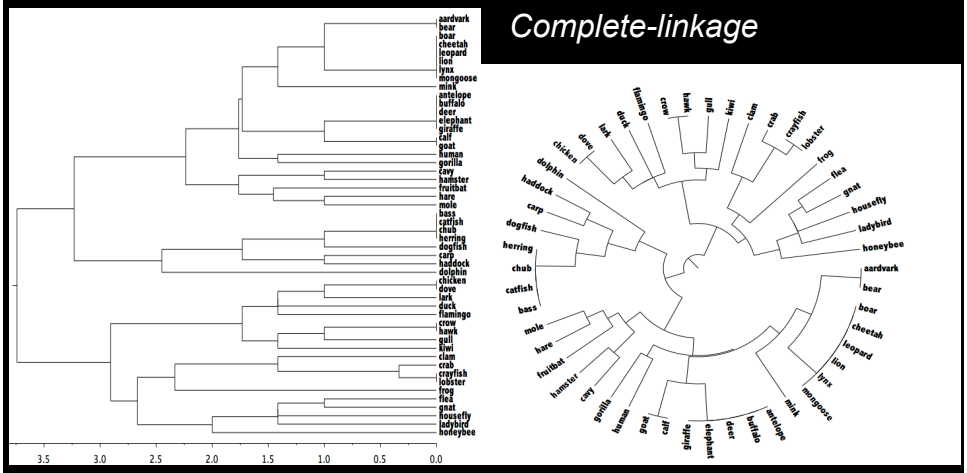
More distance measures

- *Group-average* clustering
  - Uses the average distance between all members of the merged cluster
  - Differs from average-linkage because it includes pairs from the same original cluster
- *Ward's* clustering method
  - Calculates the increase in the sum of squares of the distances of the instances from the centroid before and after fusing two clusters
  - Minimize the increase in this squared distance at each clustering step
- **All** measures will produce the same result if the clusters are compact and well separated

Example hierarchical clustering

- 50 examples of different creatures from the zoo data

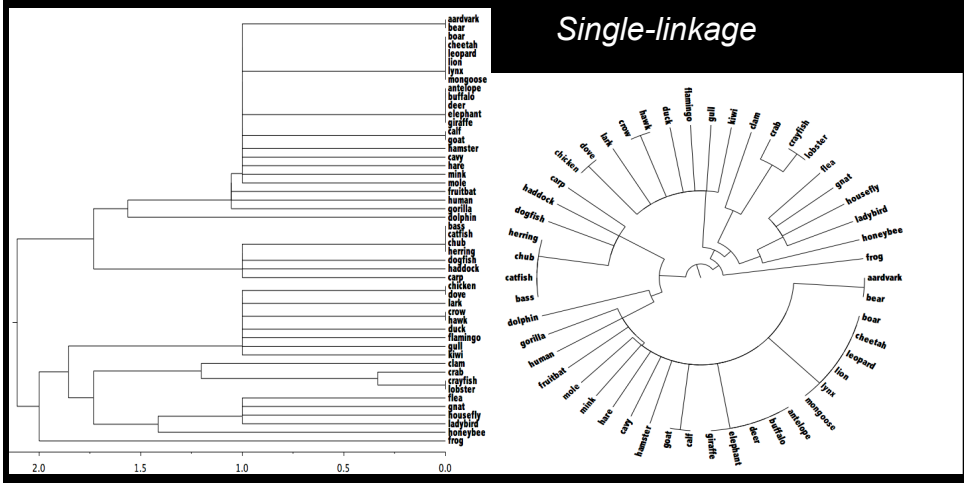
Complete-linkage



Dendogram Polar plot

Example hierarchical clustering 2

Single-linkage



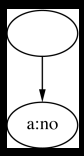
Incremental clustering

- Heuristic approach (COBWEB/CLASSIT)
- Form a hierarchy of clusters incrementally
- Start:
  - tree consists of empty root node
- Then:
  - add instances one by one
  - update tree appropriately at each stage
  - to update, find the right leaf for an instance
  - May involve restructuring the tree
- Base update decisions on *category utility*

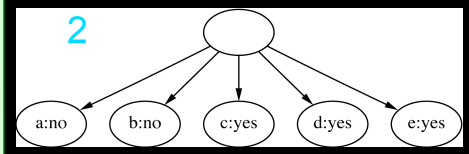
Clustering weather data

ID	Outlook	Temp.	Humidity	Windy
A	Sunny	Hot	High	False
B	Sunny	Hot	High	True
C	Overcast	Hot	High	False
D	Rainy	Mild	High	False
E	Rainy	Cool	Normal	False
F	Rainy	Cool	Normal	True
G	Overcast	Cool	Normal	True
H	Sunny	Mild	High	False
I	Sunny	Cool	Normal	False
J	Rainy	Mild	Normal	False
K	Sunny	Mild	Normal	True
L	Overcast	Mild	High	True
M	Overcast	Hot	Normal	False
N	Rainy	Mild	High	True

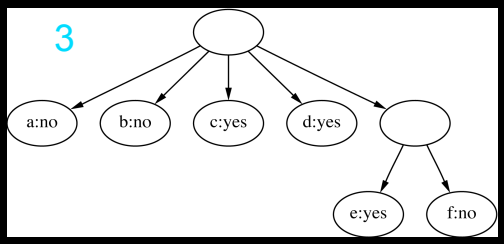
1



2



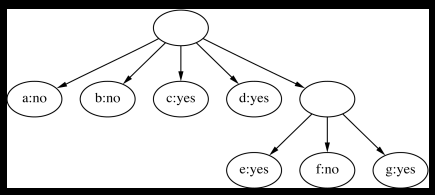
3



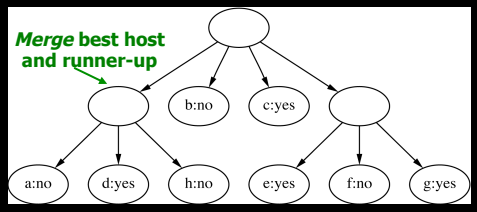
Clustering weather data

ID	Outlook	Temp.	Humidity	Windy
A	Sunny	Hot	High	False
B	Sunny	Hot	High	True
C	Overcast	Hot	High	False
D	Rainy	Mild	High	False
E	Rainy	Cool	Normal	False
F	Rainy	Cool	Normal	True
G	Overcast	Cool	Normal	True
H	Sunny	Mild	High	False
I	Sunny	Cool	Normal	False
J	Rainy	Mild	Normal	False
K	Sunny	Mild	Normal	True
L	Overcast	Mild	High	True
M	Overcast	Hot	Normal	False
N	Rainy	Mild	High	True

4

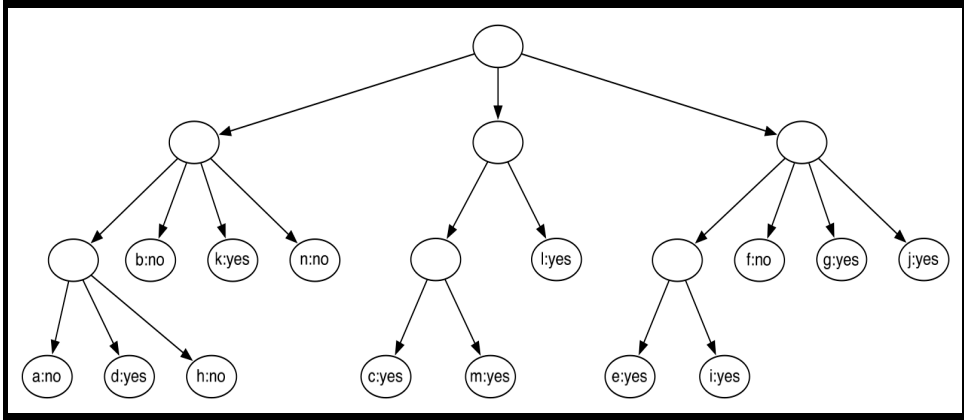


5

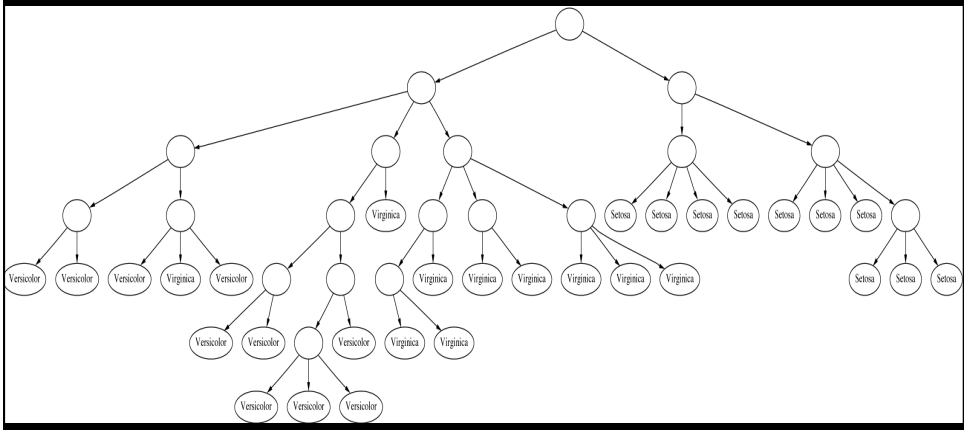


Consider *splitting* the best host if merging doesn't help

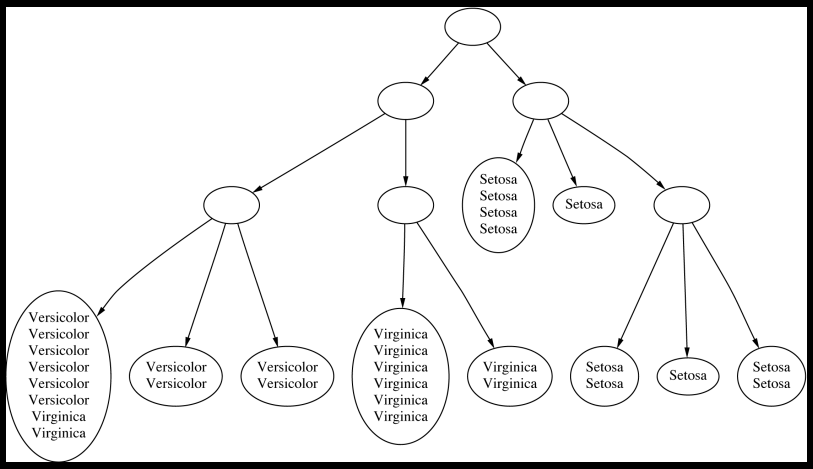
Final hierarchy



Example: the iris data (subset)



Clustering with cutoff



The end