University of Reading

CS3DS19 - Data Science Algorithms and Tools

Clustering (2 of 2)

Module convenor
Dr. Carmen Lam
carmen.lam@reading.ac.uk
Department of Computer Science

Lecture notes and videos created by
Prof. Giuseppe Di Fatta

Clustering

- Cluster analysis, unsupervised classification
- Proximity measure
- Types of Clusters
- Clustering Approaches
  - Partitioning
  - Hierarchical
  - Density-based
  - Grid-based
  - Model-based
- Cluster Validity

University of Reading

CS3DS19 - Data Science Algorithms and Tools

Hierarchical Clustering

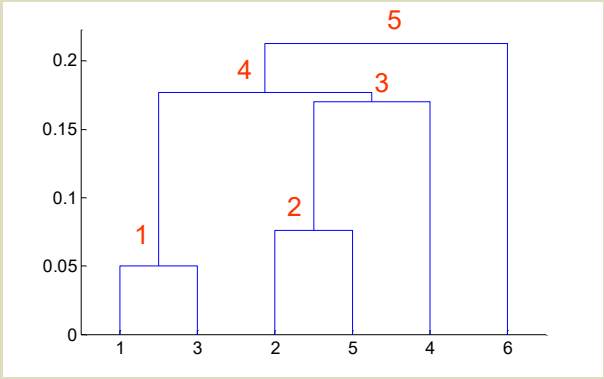
Module convenor
Dr. Carmen Lam
carmen.lam@reading.ac.uk
Department of Computer Science

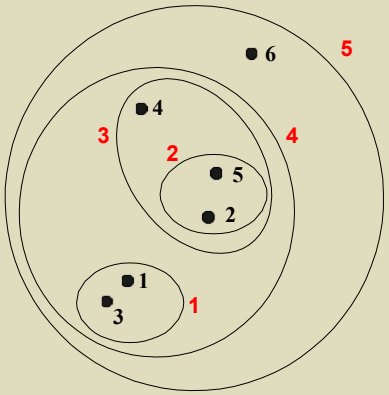
Lecture notes and videos created by
Prof. Giuseppe Di Fatta

Hierarchical Clustering

➢ Hierarchical clustering approach

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits

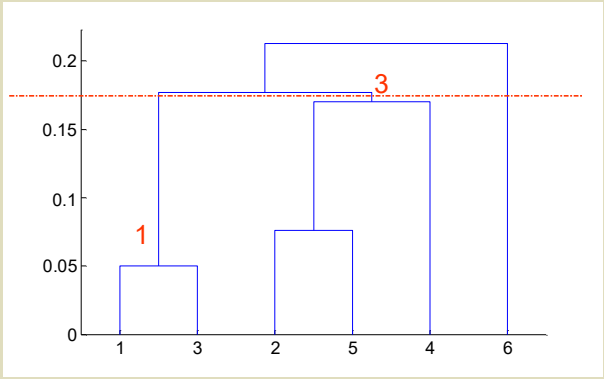


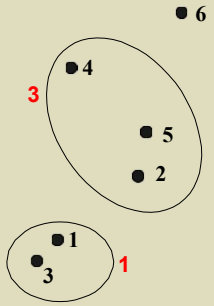


Strengths of Hierarchical Clustering

- Do not have to assume any particular number of clusters
  - Any desired number of clusters can be obtained by ‘cutting’ the dendogram at the proper level

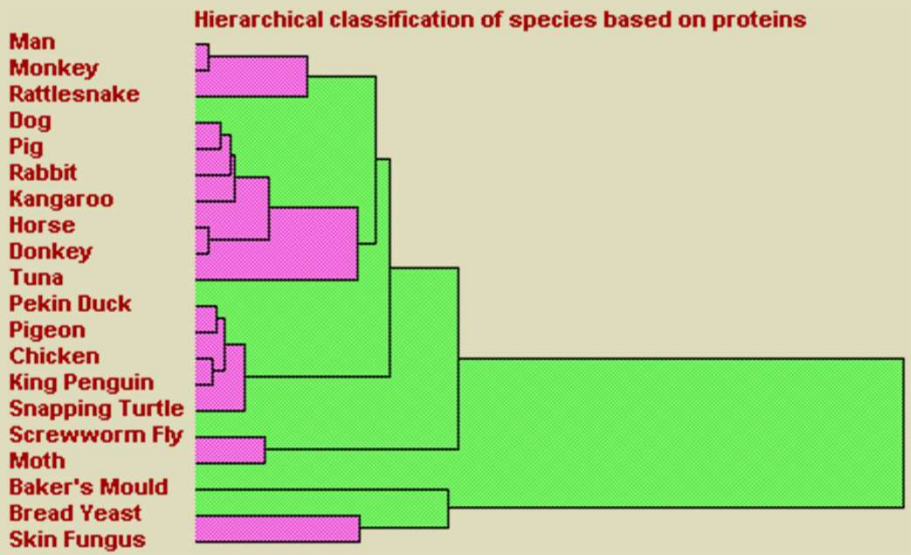
- They may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)





Hierarchical Classification of Species

- Hierarchical clusters may correspond to meaningful taxonomies

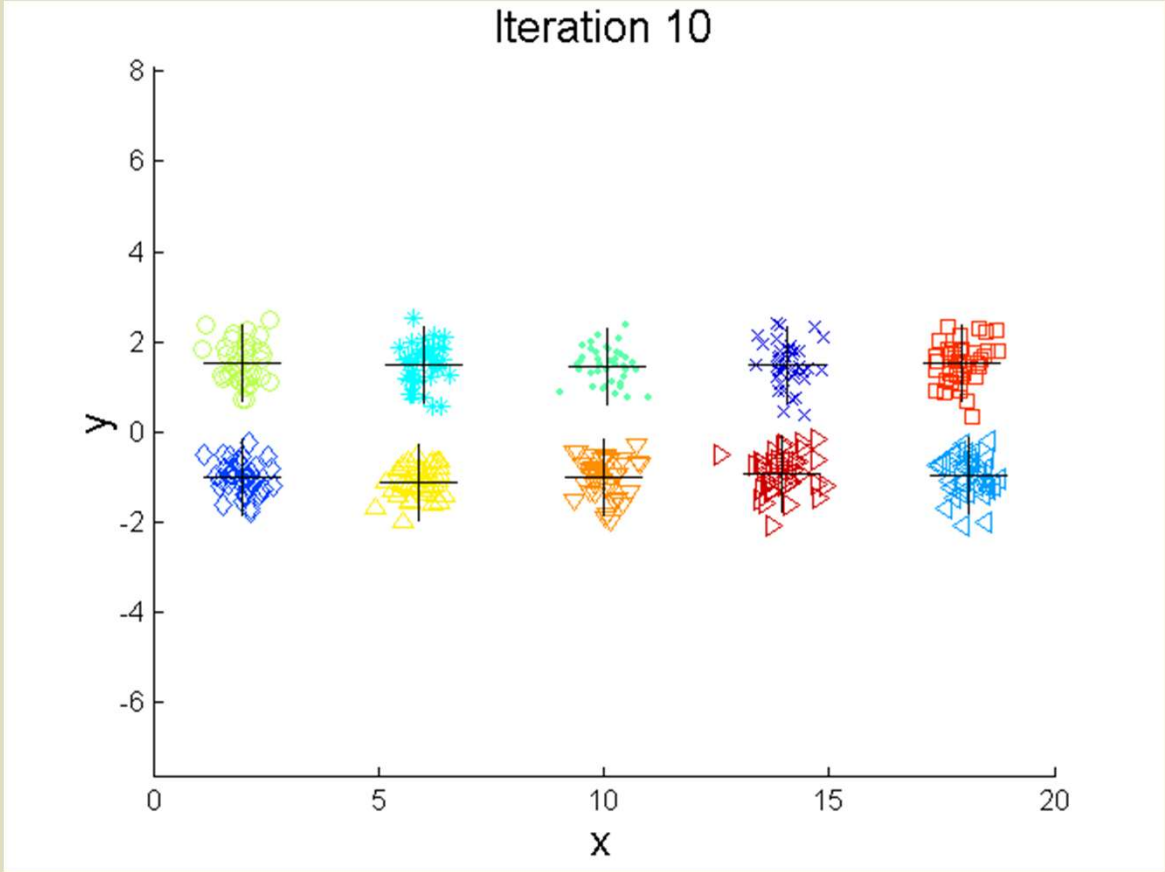


Bisecting K-means

- Bisecting K-means algorithm
  - Variant of K-means that can produce a partitional or a hierarchical clustering

1: Initialize the list of clusters to contain the cluster containing all points.
2: **repeat**
3: Select a cluster from the list of clusters
4: **for** $i = 1$ to *number_of_iterations* **do**
5: Bisect the selected cluster using basic K-means
6: **end for**
7: Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: **until** Until the list of clusters contains K clusters

Bisecting K-means Example



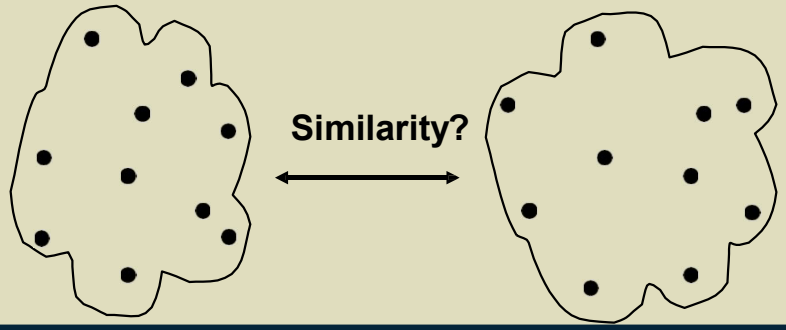
Hierarchical Clustering

- Two main types of hierarchical clustering:
  - **Agglomerative**:
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
  - **Divisive**:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are k clusters)
- Traditional hierarchical algorithms use a similarity (proximity) or distance matrix
  - Merge or split one cluster at a time

Agglomerative Clustering Algorithm

- Basic algorithm:
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. **Repeat**
  4. Merge the two closest clusters
  5. Update the proximity matrix
  6. **Until** only a single cluster remains

- Key operation is the computation of the proximity of two clusters
  - different definitions of the inter-cluster similarity/distance distinguish the different algorithms: e.g., min, max, group average, distance between centroids.



Next video lecture:

- Cluster Validity

University of Reading

CS3DS19 - Data Science Algorithms and Tools

Cluster Validity

Module convenor
Dr. Carmen Lam
carmen.lam@reading.ac.uk
Department of Computer Science

Lecture notes and videos created by
Prof. Giuseppe Di Fatta

Cluster Validity

Clustering is an unsupervised learning task

- If there is no ground truth, how to evaluate the "goodness" of the resulting clusters?
- Clusters are in the eye of the beholder!

❑ Why do we want to evaluate them?

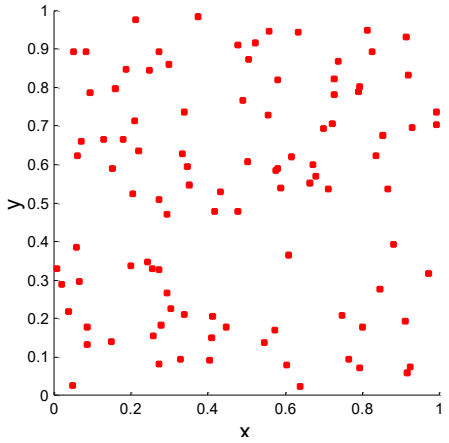
- To avoid finding patterns in noise
- To compare clustering algorithms
- To compare two clusters
- To compare two sets of clusters

❑ How to evaluate them?

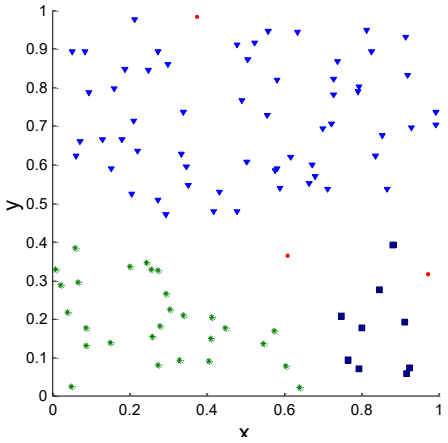
- in an unsupervised setting: internal indices
- in a supervised setting: external indices

Clusters found in Random Data

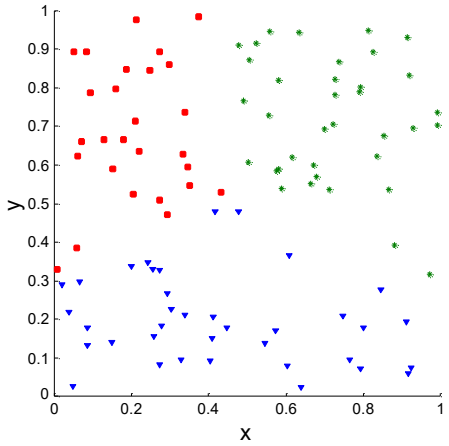
Random Points



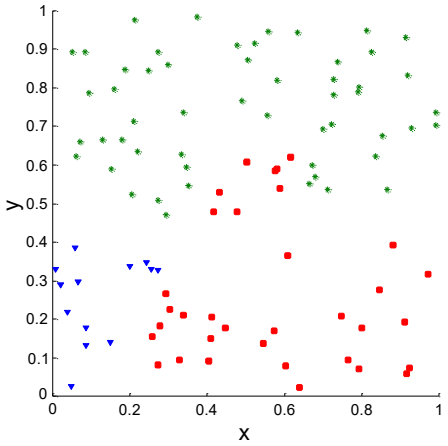
DBSCAN



K-means



Complete Link



Measures of Cluster Validity

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy
  - Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    - Sum of Squared Error (SSE)
  - Relative Index: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy
- Sometimes these are referred to as criteria instead of indices
  - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

Internal Measures: Cohesion and Separation

- Cluster Cohesion: Measures how closely related are objects in a cluster
  - Example: SSE
- Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters
- Example: Squared Error
  - Cohesion is measured by the **W**ithin cluster **S**um of **S**quares (SSE)

$$WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

  - Separation is measured by the **B**etween cluster **S**um of **S**quares

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

    - Where $|C_i|$ is the size of cluster i, m_i is the cluster mean, m is the overall mean.

External Measures of Cluster Validity: Entropy and Purity

Entropy: $H = -\sum_i p_i \log(p_i)$

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = -\sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

Final Comment on Cluster Validity

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a **black art** accessible only to those true believers who have experience and great courage."

From "Algorithms for Clustering Data", Jain and Dubes

Next:

- P05: practical on Clustering in KNIME

Next week:

- Classification