SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution

Coimbatore-35

Accredited by NBA – AICTE and Accredited by NAAC – UGC with ‘A++’ Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

19ECT303-ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

III YEAR/ V SEMESTER

UNIT 3 – UNSUPERVISED LEARNING

3.4 Hierarchal clustering

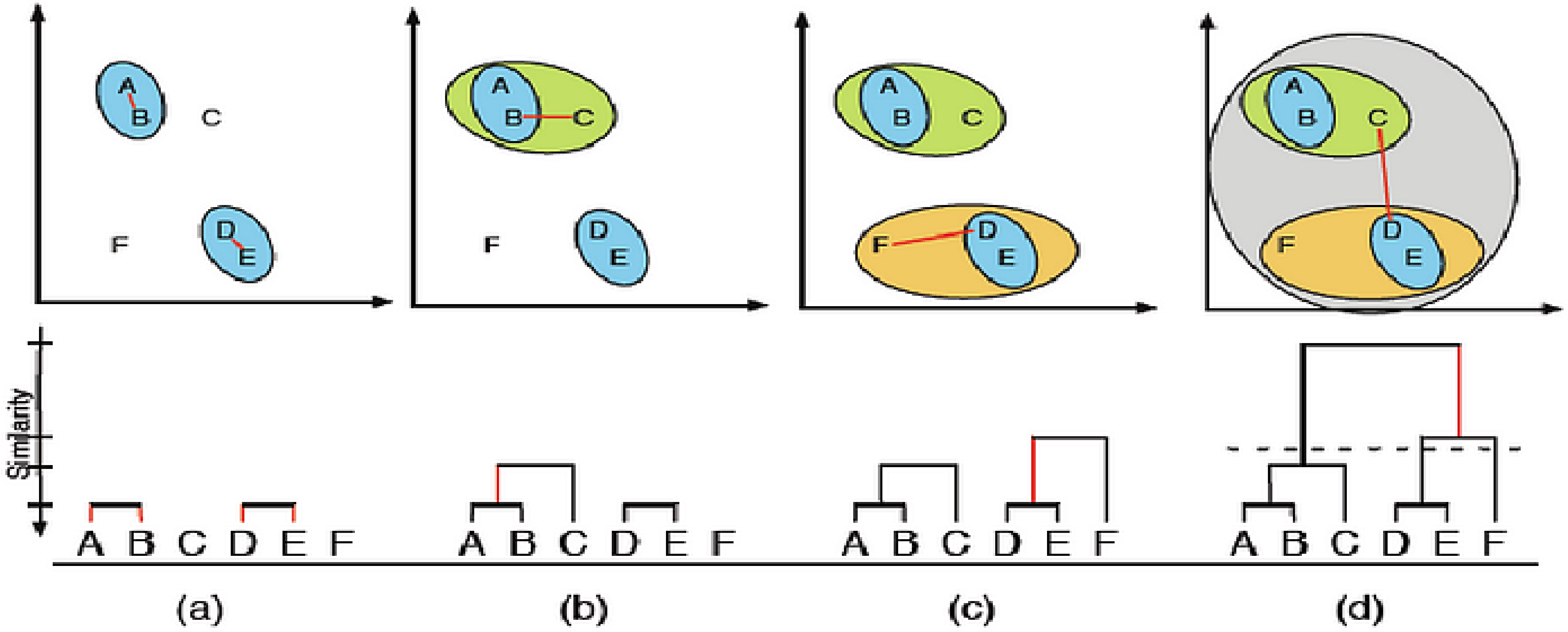
Hierarchical clustering

Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering is this:

1. ***Start by assigning each item to a cluster, N items, →N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.***
2. ***Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.***
3. ***Compute distances (similarities) between the new cluster and each of the old clusters.***
4. ***Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.***

Hierarchical clustering

Example: Hierarchical Agglomerative Clustering



Hierarchical clustering

Problems associated with clustering

•Dealing with large number of dimensions and data items can be problematic because of time complexity;

•The **effectiveness** of the method depends on the definition of **"distance**" (for distance-based clustering). If an *obvious* distance measure doesn't exist we must "define" it, which is not always easy, especially in multidimensional spaces;

•The result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

Hierarchical clustering

Applications

1. *Marketing*: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
2. *Biology*: classification of plants and animals given their features;
3. Insurance: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
4. *Earthquake studies*: clustering observed earthquake epicenters to identify dangerous zones;
5. *World Wide Web*: document classification; clustering weblog data to discover groups of similar access patterns.