Write a short note on Hierarchical Clustering

written 6.0 years ago by aartisahitya ♦ 140

modified 2.0 years ago by prashantsaini ♦ 0

0

8.5k views

data warehouse and mining

1 Answer

written 6.0 years ago by aartisahitya ♦ 140

0

- A hierarchical clustering method works by grouping data objects into a tree of clusters.
- It uses distance (similarity) matrix as clustering criteria.
- Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-

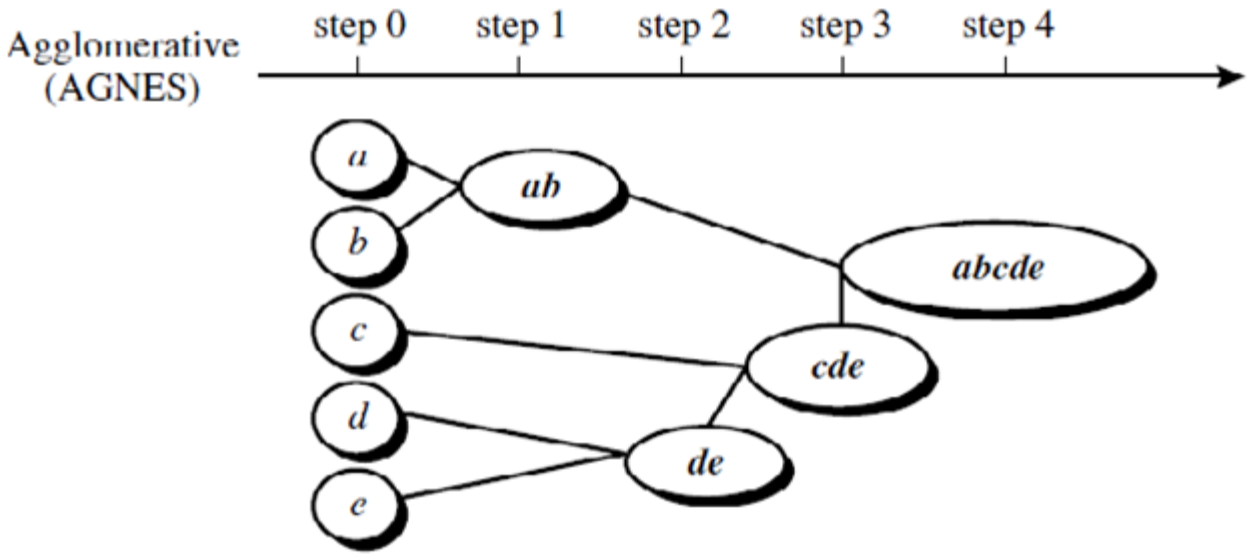
merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied.

- Divisive hierarchical clustering:

- This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster.
- It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, such as a desired number of clusters is obtained or the diameter of each cluster is within a certain threshold.

AGGLOMERATIVE HIERARCHICAL CLUSTERING: - Figure shows the application of AGNES (AGglomerativeNESting), an agglomerative hierarchical clustering method to a data set of five objects(a, b, c, d, e).

- Initially, AGNES places each object into a cluster of its own.
- The clusters are then merged step-by-step according to some criterion.



Agglomerative Algorithm: (AGNES)

Given

-a set of N objects to be clustered

-an N*N distance matrix ,

The basic process of clustering id this:

Step1: Assign each object to a cluster so that for N objects we have N clusters each containing just one Object.

Step2: Let the distances between the clusters be the same as the distances between the objects they contain.

Step3: Find the most similar pair of clusters and merge them into a single cluster so that we

now have one cluster less.

Step4: Compute distances between the new cluster and each of the old clusters.

Step5: Repeat steps 3 and 4 until all items are clustered into a single cluster of size N.

- Step 4 can be done in different ways and this distinguishes single and complete linkage.

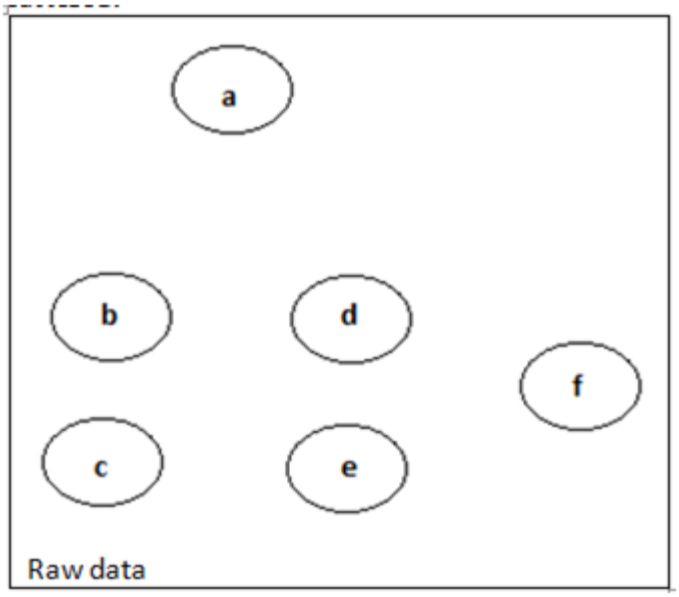
  -> For complete-linkage algorithm:

  - clustering process is terminated when the maximum distance between nearest clusters exceeds an arbitrary threshold.

  -> For single-linkage algorithm:

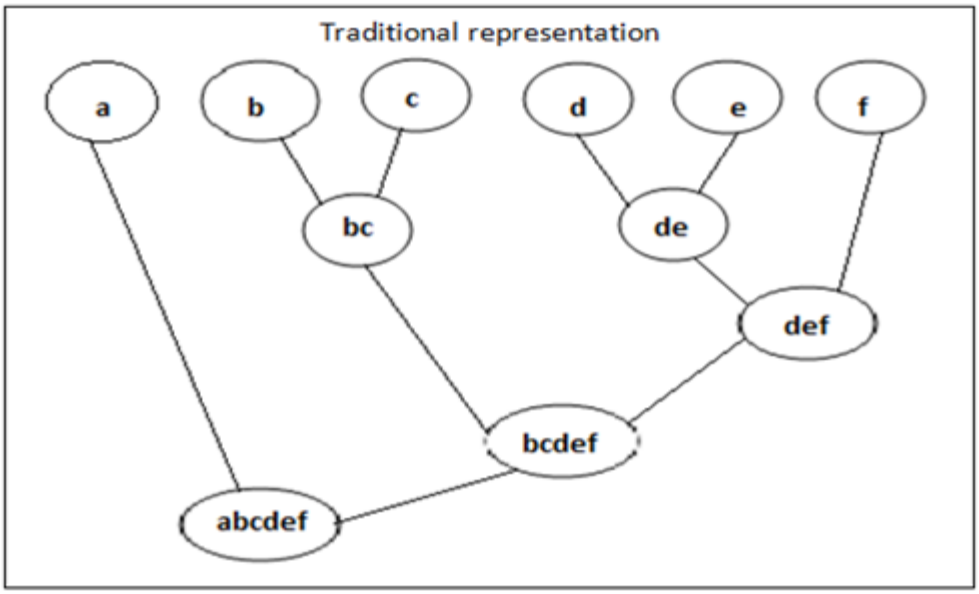
  - clustering process is terminated when the minimum distance between nearest clusters exceeds an arbitrary threshold. **EXAMPLE:**

Suppose this data is to be clustered.



- In this example, cutting the tree after the second row of the dendrogram will yield clusters {a} {b c} {d e} {f}.
- Cutting the tree after the third row will yield clusters {a} {b c} {d e f}, which is a coarser clustering, with a smaller number but larger clusters.

The hierarchical clustering dendrogram would be as such:



In our example, we have six elements {a} {b} {c} {d} {e} and {f}.

The first step is to determine which elements to merge in a cluster.

Usually, we take the two closest elements, according to the chosen distance.

Then, as clustering progresses, rows and columns are merged as the clusters are merged and the distances updated. Suppose we have merged the two closest elements b and c, we now have the following clusters {a}, {b, c}, {d}, {e} and {f}, and want to merge them further.

To do that, we need to take the distance between {a} and {b c}, and therefore define the distance between two clusters. Usually the distance between two clusters *A* and *B* is one of the following:

- The maximum distance between elements of each cluster (also called complete-linkage clustering):

  max {d(x,y):x∈A,y∈B}

- The minimum distance between elements of each cluster (also called single-linkage clustering):

  min {d(x,y):x∈A,y∈B}

- The mean distance between elements of each cluster (also called average linkage clustering, used e.g. in UPGMA):

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y)$$

- Each agglomeration occurs at a greater distance between clusters than the previous agglomeration, and one can decide to stop clustering either when the clusters are too far apart to be merged (distance criterion) or when there is a sufficiently small number

of clusters (number criterion).

ADD COMMENT SHARE EDIT

Please log in to add an answer.