## National University of Computer and Emerging Sciences

# Lab Manual 04
# Fundamentals of Big Data Lab

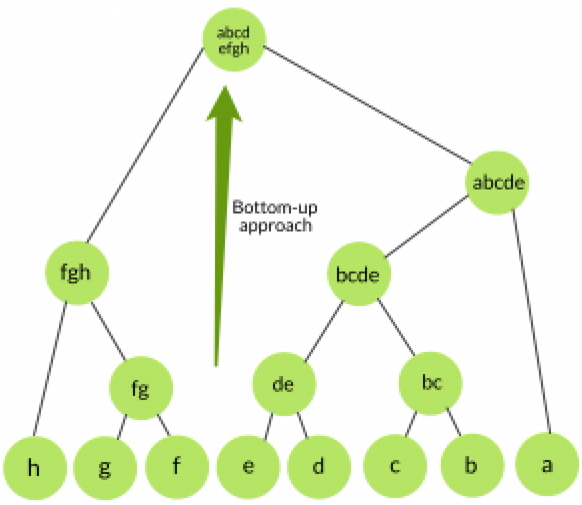
| | |
|---|---|
| Course Instructor | Dr. Iqra Safdar |
| Lab Instructor (s) | Aiss Shahid,<br>Muhammad Mazarib |
| Section | |
| Semester | Spring 2022 |

In data mining and statistics, hierarchical clustering analysis is a method of cluster analysis that seeks to build a hierarchy of clusters i.e. tree-type structure based on the hierarchy.

**Agglomerative Clustering:** Also known as bottom-up approach or hierarchical agglomerative clustering (HAC). A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data.

**Algorithm :**

```
given a dataset ($d_1$, $d_2$, $d_3$, ....$d_N$) of size N
# compute the distance matrix
for i=1 to N:
   # as the distance matrix is symmetric about
   # the primary diagonal so we compute only lower
   # part of the primary diagonal
   for j=1 to i:
      dis_mat[i][j] = distance[$d_i$, $d_j$]
each data point is a singleton cluster
repeat
   merge the two cluster having minimum distance
   update the distance matrix
until only a single cluster remains
```



Hierarchical agglomerative clustering

**Python implementation of the above algorithm using the scikit-learn library:**

- **Python3**

```
from sklearn.cluster import
AgglomerativeClustering
import numpy as np

# randomly chosen dataset
X = np.array([[1, 2], [1, 4], [1, 0],
              [4, 2], [4, 4], [4, 0]])

# here we need to mention the number of clusters
# otherwise the result will be a single cluster
# containing all the data
clustering = AgglomerativeClustering(n_clusters =
2).fit(X)

# print the class labels
print(clustering.labels_)
```

**Output :**

```
[1, 1, 1, 0, 0, 0]
```

**Computing Distance Matrix: While merging two clusters we check the distance between two every pair of clusters and merge the pair with least distance/most similarity. But the question is how is that distance determined. There are different ways of defining Inter Cluster distance/similarity. Some of them are:**

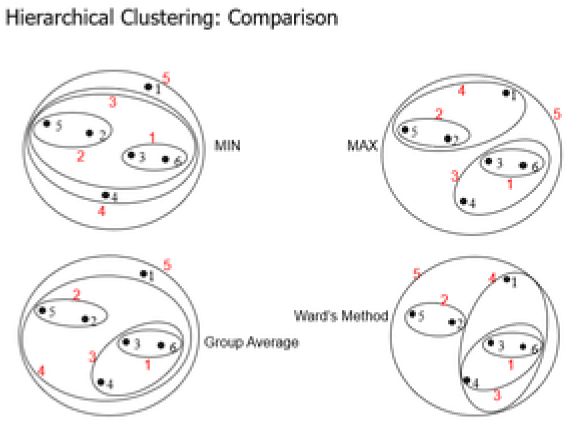
**1. Min Distance: Find minimum distance between any two points of the cluster.**

**2. Max Distance: Find maximum distance between any two points of the cluster.**

**3. Group Average: Find average of distance between every two points of the clusters.**

**4. Ward’s Method: Similarity of two clusters is based on the increase in squared error when two clusters are merged.**

**For example, if we group a given data using different method, we may get different results:**



**Task A:**

1. Use built in Agglomerative to work on given data, use annual income and spending score.

**Task B:**

1. Design your own Agglomerative clustering algorithm, use annual income and spending score.