Experiment: 6

Aim: To demonstrate hierarchical clustering on given data set.

Theory:

Clustering is one of the most common exploratory data analysis techniques used to get an intuition about thestructureof the data. It can be defined as the task of identifying subgroups in the data such that data points inthe same subgroup (cluster) are very similar while data points in different clusters are very different. In otherwords, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similaritymeasure to use is application-specific.

Hierarchical clustering:

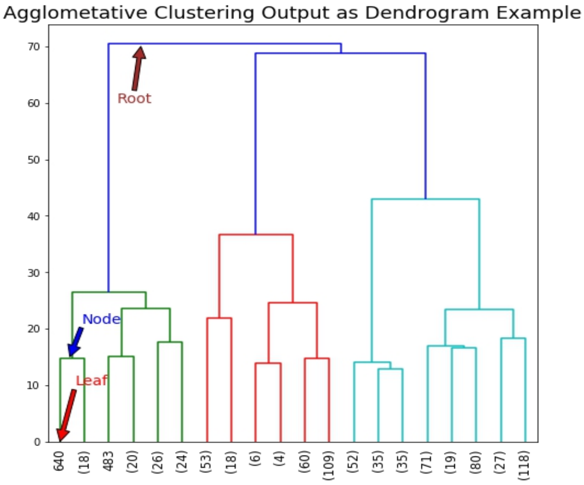
Hierarchical clustering, also known as Hierarchical cluster analysis or HCA, is an unsupervised clustering approach that includes forming groups with a top-to-bottom order.

The hierarchical clustering technique has two approaches:

1. **Agglomerative:** Agglomerative is a **bottom-up** approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
2. **Divisive:** Divisive algorithm is the reverse of the agglomerative algorithm as it is a **top-down approach.**

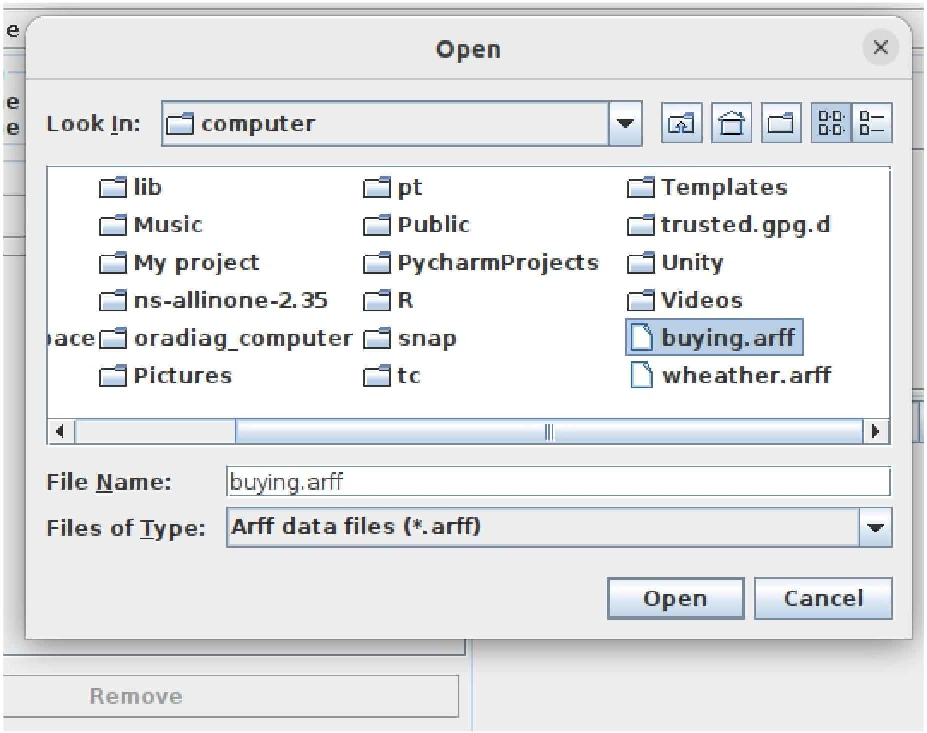
Agglomerative Hierarchical clustering:

Agglomerative Clustering or bottom-up clustering essentially started from an **individual cluster** (each data point is considered as an individual cluster, also called **leaf**), then every cluster calculates their **distance** with each other. The two clusters with the shortest distance with each other would **merge** creating what we called **node**. Newly formed clusters once again calculating the member of their cluster distance with another cluster outside of their cluster. The process is repeated until all the data points assigned to one cluster called **root**. The result is a tree-based representation of the objects called **dendrogram**.



Steps to be followed:

Step 1: Open the Weka explorer in the preprocessing interface and import the appropriate dataset;



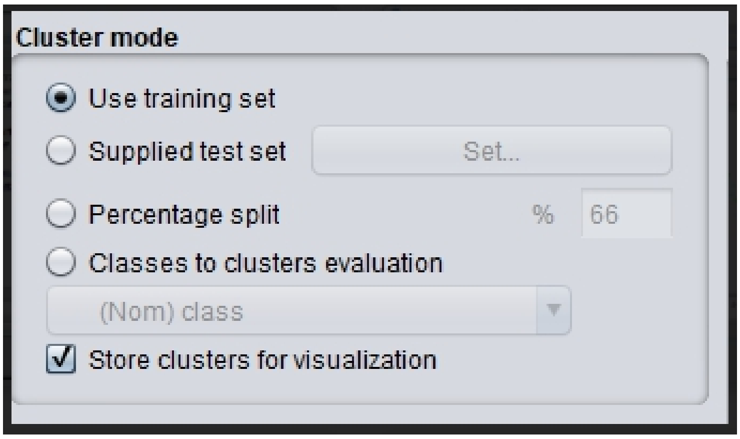
File Name: buying.arff

Files of Type: Arff data files (*.arff)

Step 2: To perform clustering, go to the explorer's 'cluster' tab and select the select button. As a result of this step, a dropdown list of available clustering algorithms displays; pick the Hierarchical algorithm.

Step 3: Then press the text button to the right of the pick icon to bring up the popup window seen in the screenshots. In this window, we input three for the number of clusters and leave the seed value alone. The seed value is used to generate a random number that is used to allocate cluster instances to each other internally.

Step 4: One of the options has been selected. Before we execute the clustering method, we need to make sure they're in the 'cluster mode' panel. The option to employ a training set is chosen, after which the 'start' button is hit. The process and the resulting window are depicted in the screenshots below.



Cluster mode

● Use training set

○ Supplied test set Set...

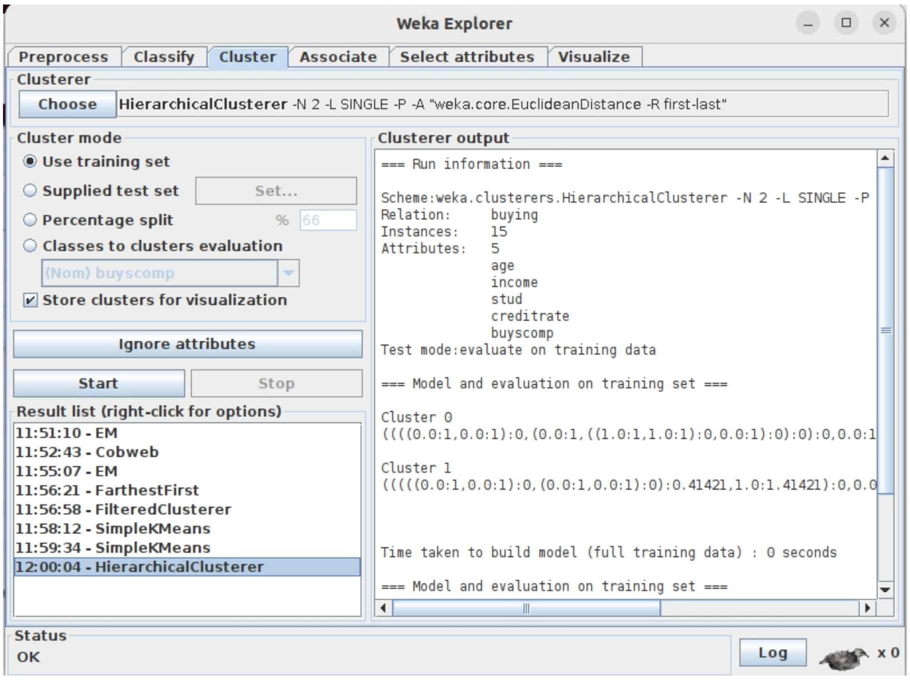
○ Percentage split % 66

○ Classes to clusters evaluation

(Nom) class

☑ Store clusters for visualization

Step 5: The resulting window displays the centroid of each cluster, as well as data on the number and proportion of instances assigned to each cluster. A mean vector is used to represent each cluster centroid. A cluster can be described using this cluster.



Clusterer: HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

Cluster mode

● Use training set

○ Supplied test set Set...

○ Percentage split % 66

○ Classes to clusters evaluation

(Nom) buyscomp

☑ Store clusters for visualization

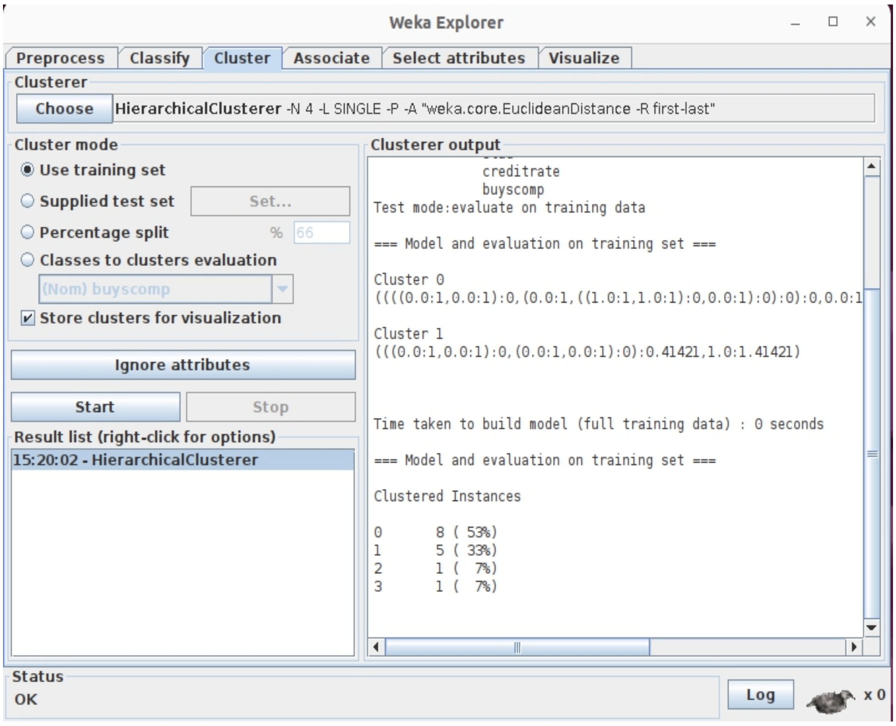
Result list (right-click for options)

11:51:10 - EM
11:52:43 - Cobweb
11:55:07 - EM
11:56:21 - FarthestFirst
11:56:58 - FilteredClusterer
11:58:12 - SimpleKMeans
11:59:34 - SimpleKMeans
12:00:04 - HierarchicalClusterer

Clusterer output

```
=== Run information ===

Scheme:weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P
Relation:     buying
Instances:    15
Attributes:   5
              age
              income
              stud
              creditrate
              buyscomp
Test mode:evaluate on training data

=== Model and evaluation on training set ===

Cluster 0
((((0.0:1,0.0:1):0,(0.0:1,((1.0:1,1.0:1):0,0.0:1):0):0):0,0.0:1

Cluster 1
(((((0.0:1,0.0:1):0,(0.0:1,0.0:1):0):0.41421,1.0:1.41421):0,0.0



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===
```

Status
OK



Clusterer: Choose HierarchicalClusterer -N 4 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

● Use training set

☑ Store clusters for visualization

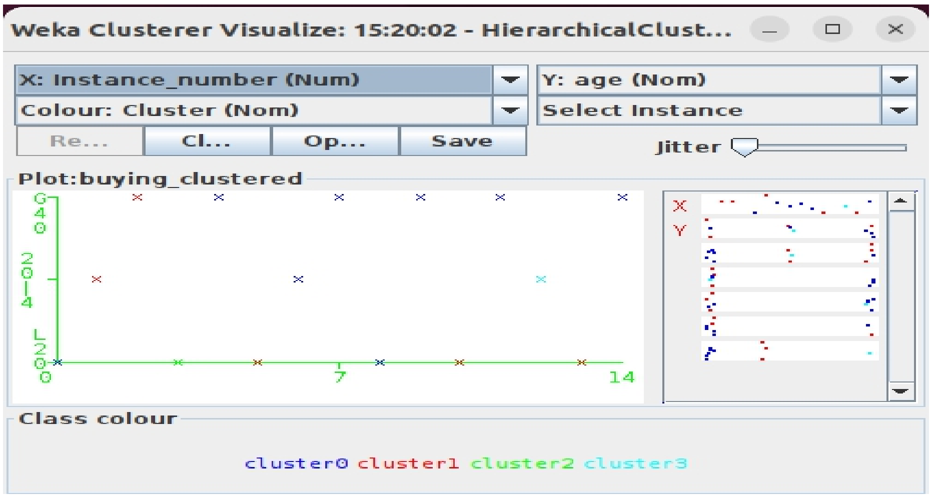
Result list (right-click for options): 15:20:02 - HierarchicalClusterer

Clusterer output

```
                    creditrate
                    buyscomp
Test mode:evaluate on training data

=== Model and evaluation on training set ===

Cluster 0
((((0.0:1,0.0:1):0,(0.0:1,((1.0:1,1.0:1):0,0.0:1):0):0):0,0.0:1

Cluster 1
(((0.0:1,0.0:1):0,(0.0:1,0.0:1):0):0.41421,1.0:1.41421)



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       8 ( 53%)
1       5 ( 33%)
2       1 (  7%)
3       1 (  7%)
```

Status: OK

Step 6: Visualizing the qualities of each cluster is another approach to grasp them. Right-click the result set on the result to do so. To visualize cluster assignments are selected from the list column.



Weka Clusterer Visualize: 15:20:02 - HierarchicalClust...
Re...
Cl...
Op...
Save
Jitter
Select Instance
G
40
20-4
L200
0
7
14
X
Y
Class colour
cluster0 cluster1 cluster2 cluster3

X: Instance_number (Num)

Y: age (Nom)

Colour: Cluster (Nom)

Plot:buying_clustered

Conclusion: Thus we have implemented hierarchical algorithm for clustering successfully.

Viva Voce

Q.1 What is a Hierarchical Clustering Algorithm?

Q.2 What are the various types of Hierarchical Clustering?

Q.3 What is a dendrogram in Hierarchical Clustering Algorithm?

Q.4 Explain the different linkage methods used in the Hierarchical Clustering Algorithm.