EXPERIMENT NO. 8

Aim: To implement Hierarchical Clustering method.

Requirements: Windows O.S and Weka tool.

Problem Statement: To implement Hierarchical Clustering (Single Link, Complete Link and Average Link) on iris data set.

Theory:

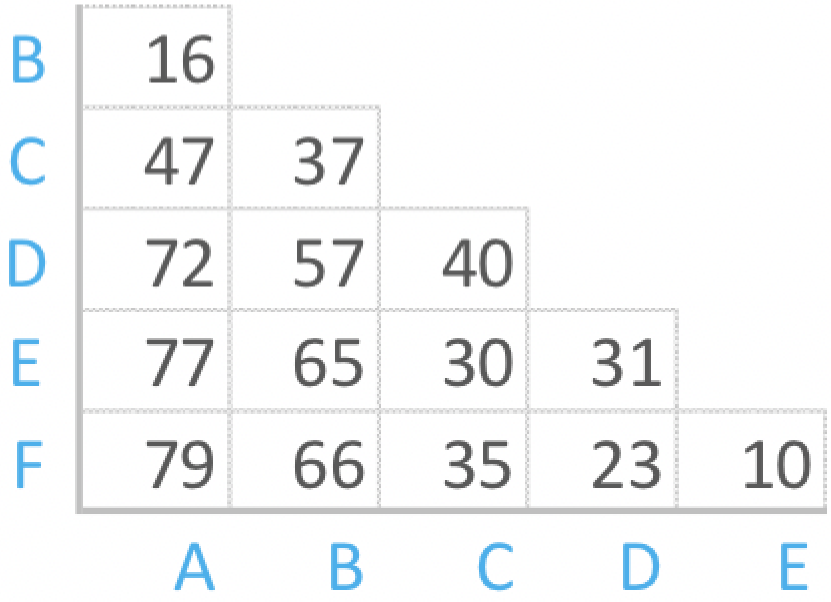
Hierarchical Clustering:

Hierarchical clustering, also known as *hierarchical cluster analysis*, is an algorithm that groups similar objects into groups called *clusters*. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

If you want to do your own hierarchical cluster analysis, use the template below - just add your data!

Required data

Hierarchical clustering can be performed with either a *distance matrix* or *raw data*. When raw data is provided, the software will automatically compute a distance matrix in the background. The distance matrix below shows the distance between six objects.



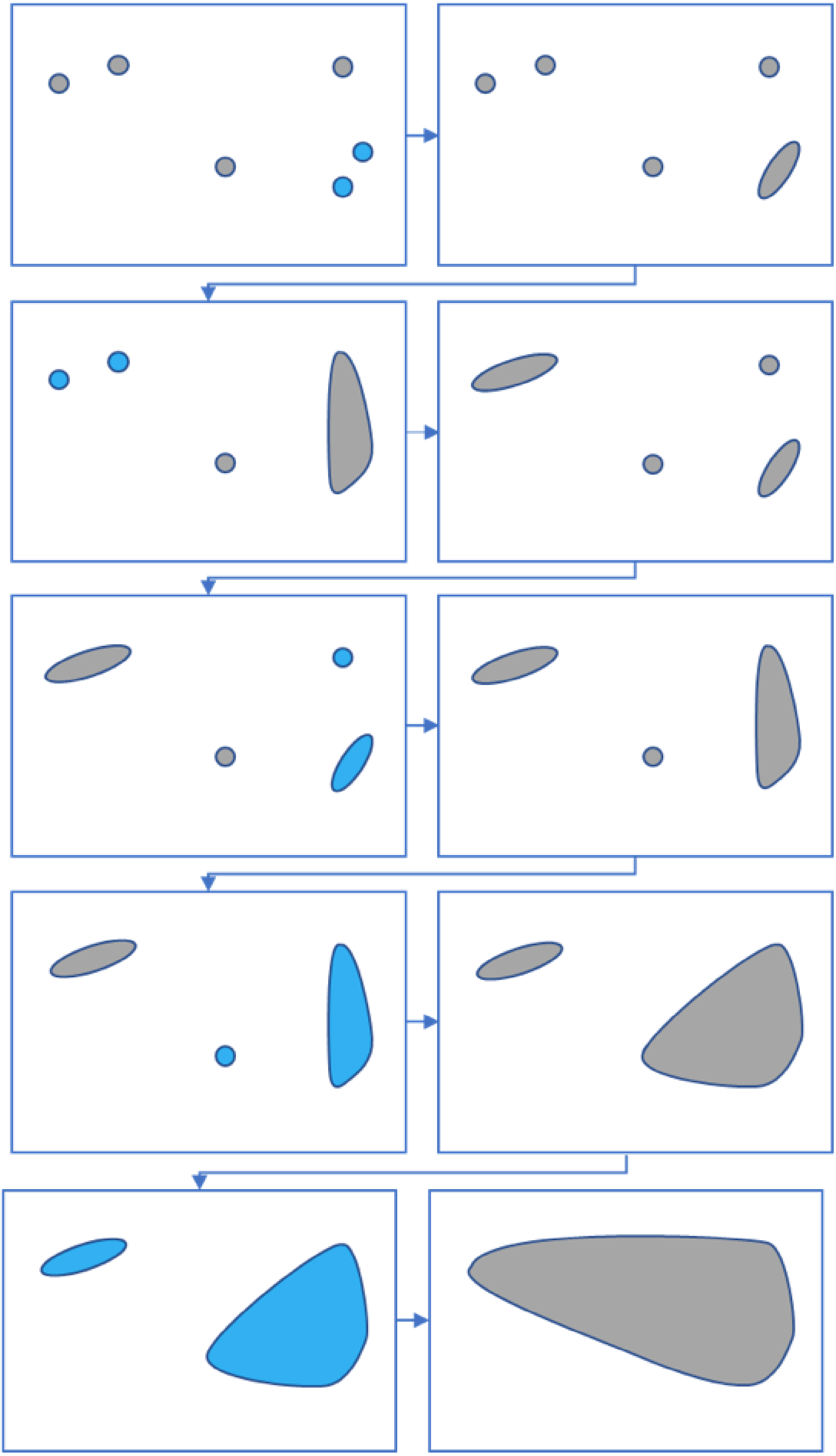
B	16				
C	47	37			
D	72	57	40		
E	77	65	30	31	
F	79	66	35	23	10
	A	B	C	D	E

How hierarchical clustering works

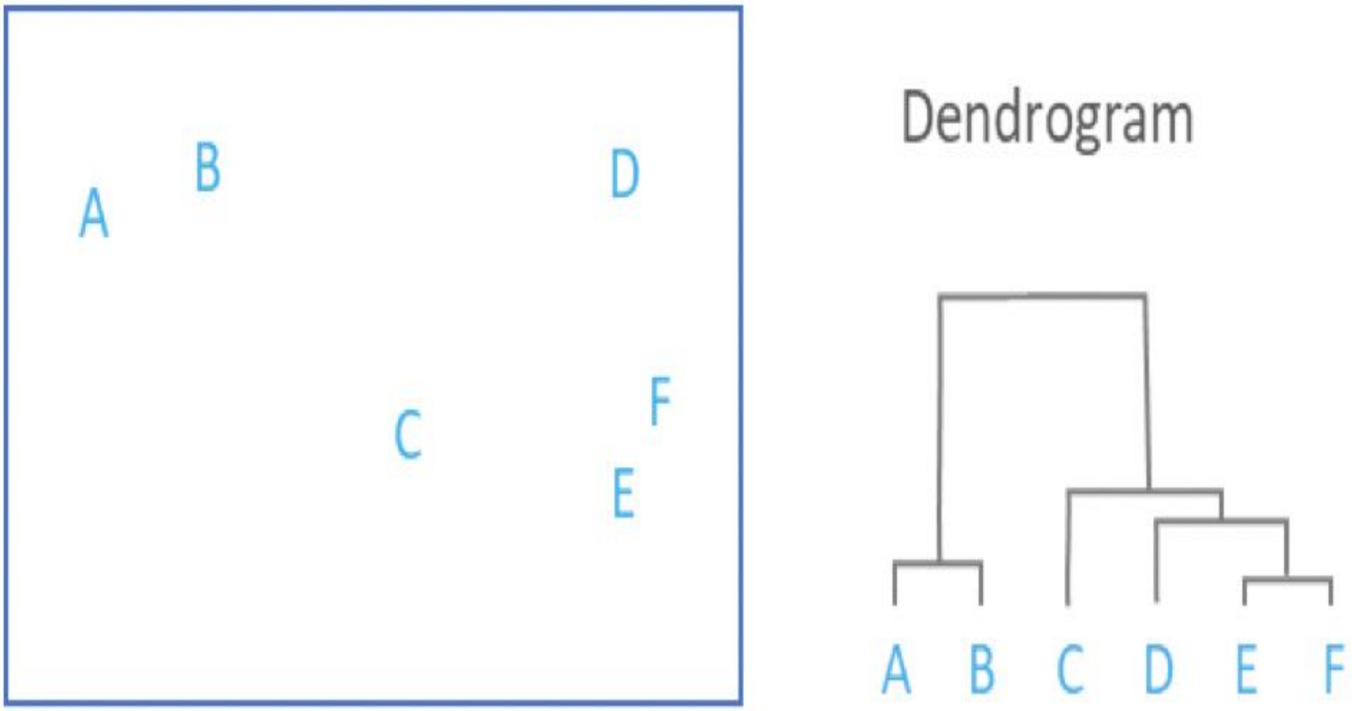
Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps: (1) identify the two clusters that are closest together, and (2) merge the two most similar clusters. This iterative process continues until all the clusters are merged together. This is illustrated in the diagrams below.

Identify the two clusters that are closest together

Merge the two most similar clusters



The main output of Hierarchical Clustering is a *dendrogram*, which shows the hierarchical relationship between the clusters:



Measures of distance (similarity)

In the example above, the *distance* between two clusters has been computed based on the length of the straight line drawn from one cluster to another. This is commonly referred to as the *Euclidean distance.* Many other *distance metrics* have been developed.

The choice of distance metric should be made based on theoretical concerns from the domain of study. That is, a distance metric needs to define similarity in a way that is sensible for the field of study. For example, if clustering crime sites in a city, city block distance may be appropriate. Or, better yet, the time taken to travel between each location. Where there is no theoretical justification for an alternative, the Euclidean should generally be preferred, as it is usually the appropriate measure of distance in the physical world.

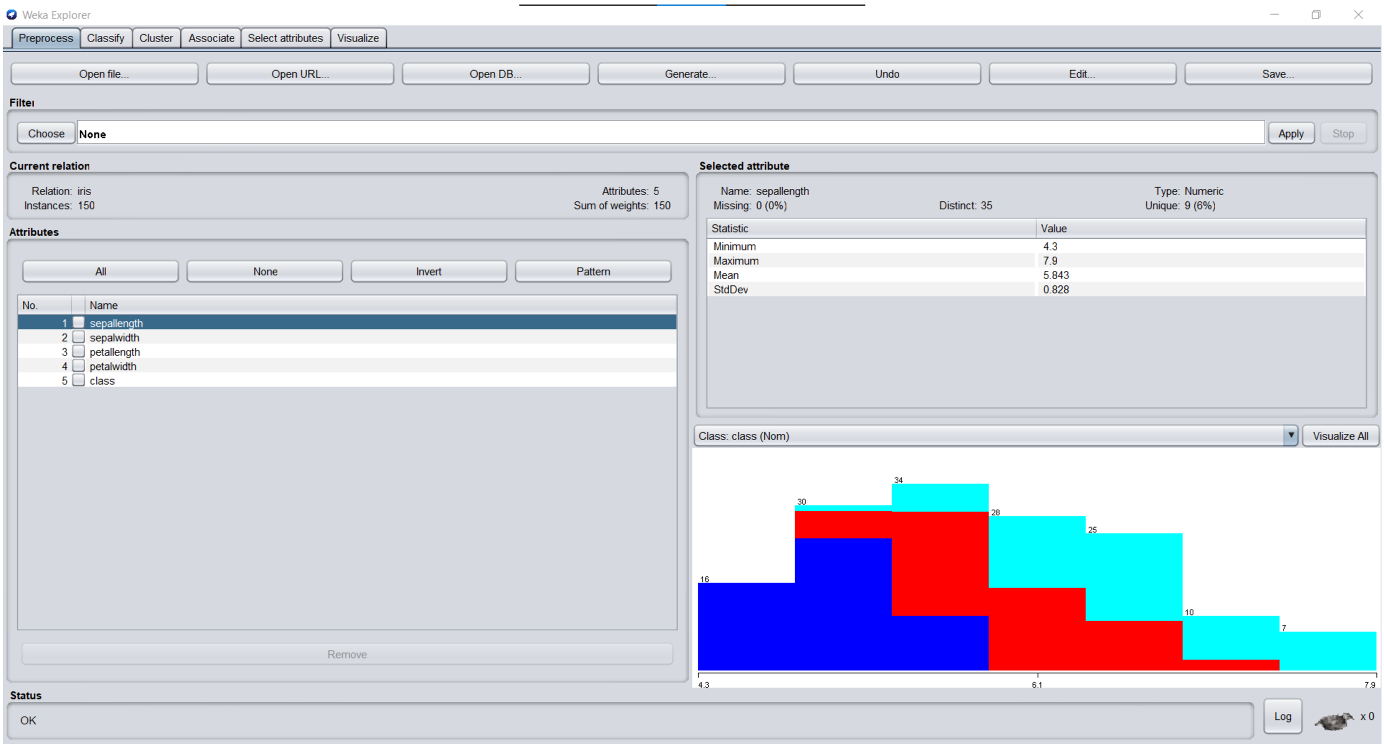
Linkage Criteria

After selecting a distance metric, it is necessary to determine from where distance is computed. For example, it can be computed between the two most similar parts of a cluster (*single-linkage*), the two least similar bits of a cluster (*complete-linkage*), the center of the clusters (*mean* or *average-linkage*), or some other criterion. Many linkage criteria have been developed.

As with *distance metrics,* the choice of linkage criteria should be made based on theoretical considerations from the domain of application. A key theoretical issue is what causes variation. For example, in archeology, we expect variation to occur through innovation and natural resources, so working out if two groups of artefacts are similar may make sense based on identifying the most similar members of the cluster.

Output:

Iris Data set:



Filter: None

Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150

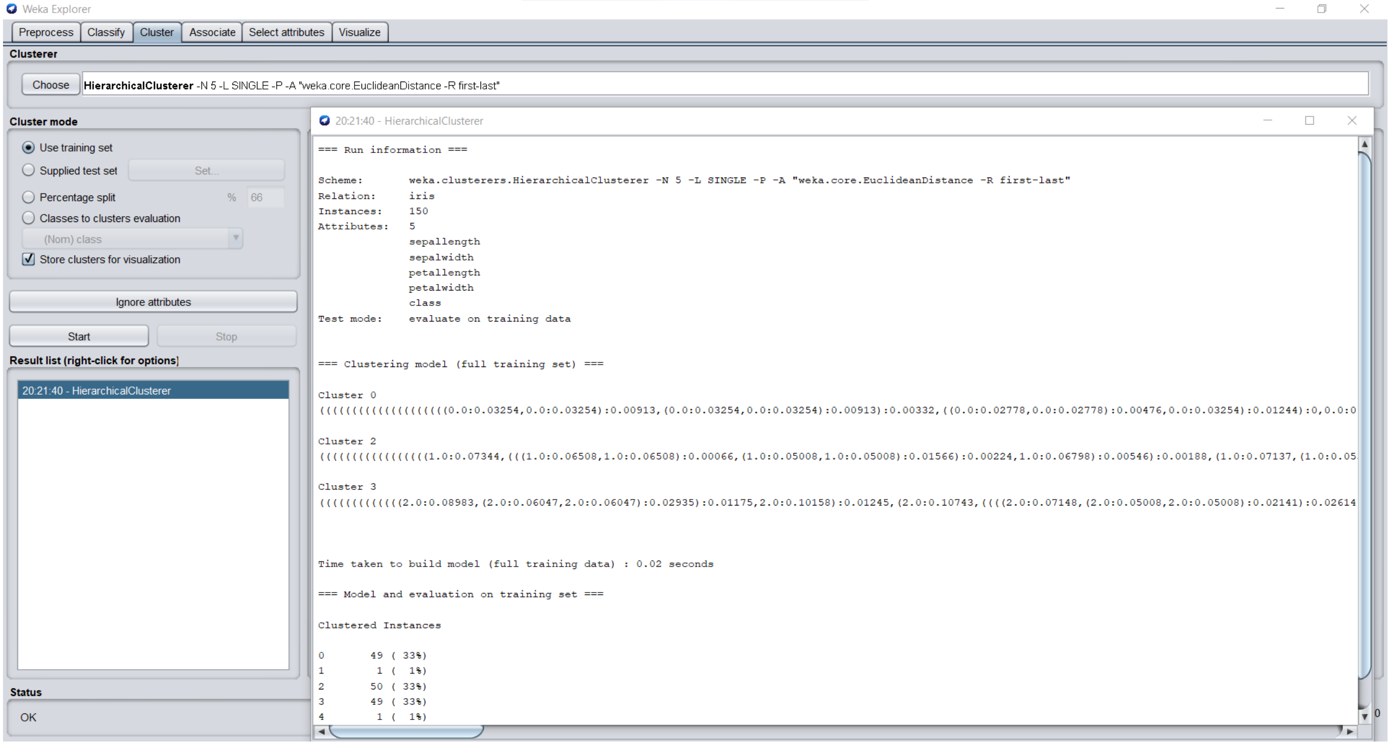
No.		Name
1	□	sepallength
2	□	sepalwidth
3	□	petallength
4	□	petalwidth
5	□	class

Name: sepallength
Missing: 0 (0%)
Distinct: 35
Type: Numeric
Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Status: OK

Single Link Clustering using 5 clusters and Euclidian Distance:

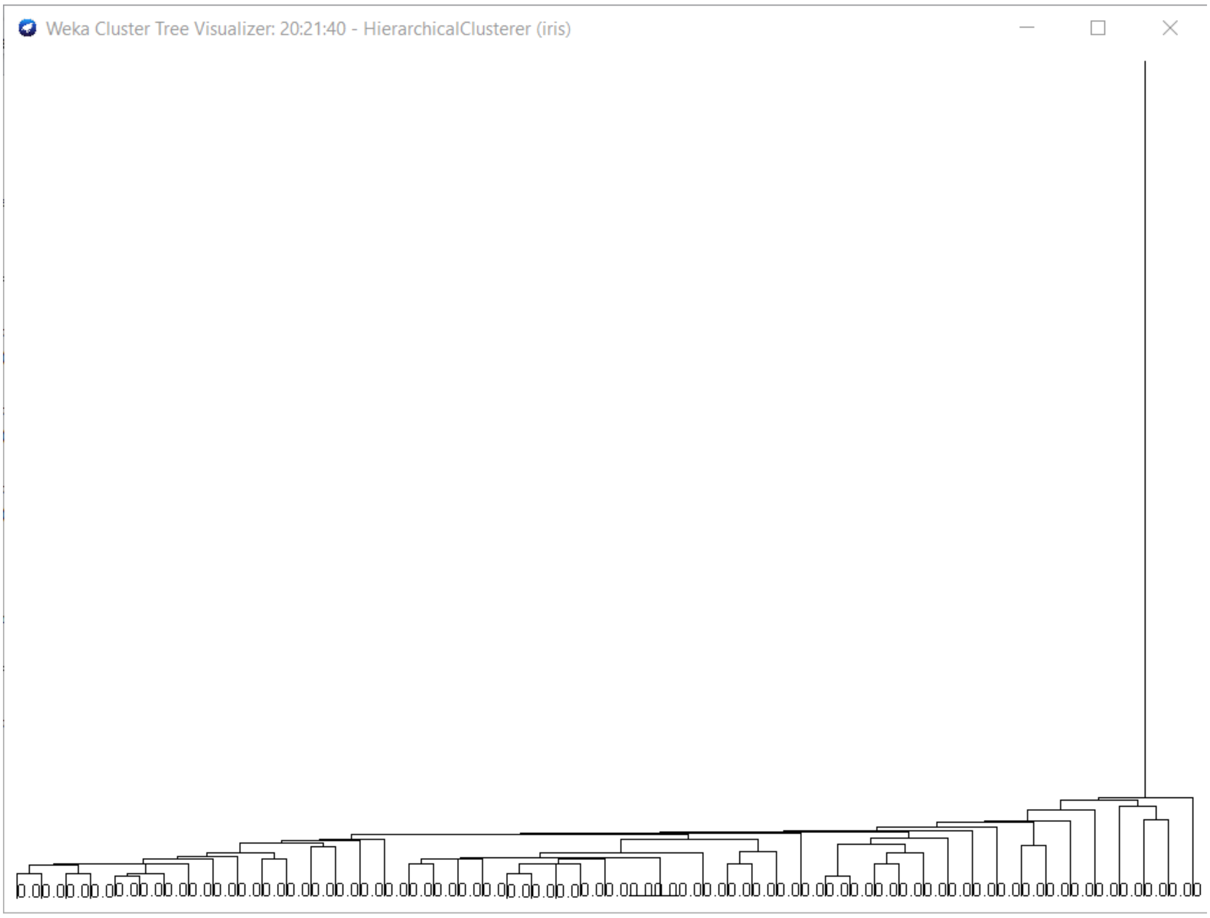


HierarchicalClusterer -N 5 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

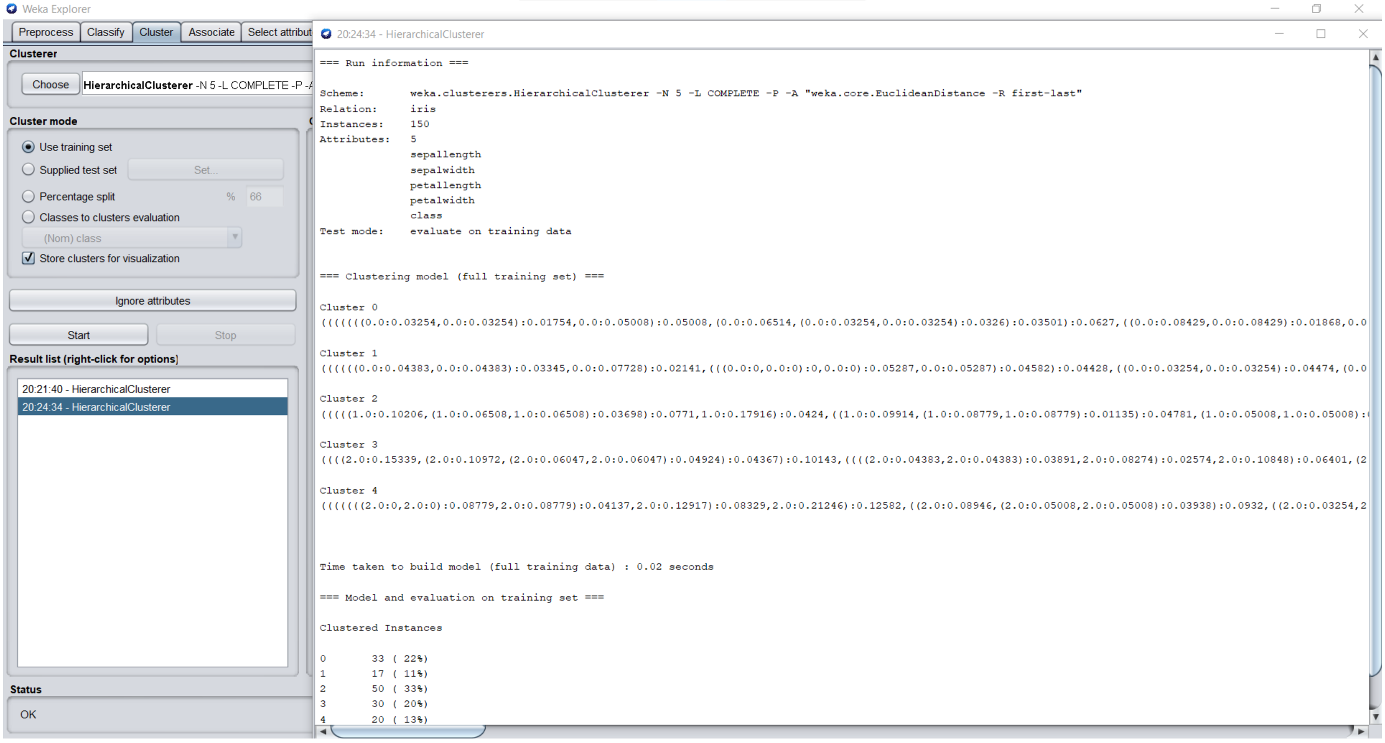
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) class
☑ Store clusters for visualization

```
=== Run information ===

Scheme:       weka.clusterers.HierarchicalClusterer -N 5 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    evaluate on training data


=== Clustering model (full training set) ===

Cluster 0
((((((((((((((((((((0.0:0.03254,0.0:0.03254):0.00913,(0.0:0.03254,0.0:0.03254):0.00913):0.00332,((0.0:0.02778,0.0:0.02778):0.00476,0.0:0.03254):0.01244):0,0.0:0

Cluster 2
(((((((((((((((((1.0:0.07344,(((1.0:0.06508,1.0:0.06508):0.00066,(1.0:0.05008,1.0:0.05008):0.01566):0.00224,1.0:0.06798):0.00546):0.00188,(1.0:0.07137,(1.0:0.05

Cluster 3
(((((((((((((2.0:0.08983,(2.0:0.06047,2.0:0.06047):0.02935):0.01175,2.0:0.10158):0.01245,(2.0:0.10743,((((2.0:0.07148,(2.0:0.05008,2.0:0.05008):0.02141):0.02614



Time taken to build model (full training data) : 0.02 seconds

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

Clustered Instances

0       49 ( 33%)
1        1 (  1%)
2       50 ( 33%)
3       49 ( 33%)
4        1 (  1%)
```

Single Link Dendrogram:



Complete Link Clustering using 5 clusters and Euclidian Distance:



Clusterer

Choose **HierarchicalClusterer** -N 5 -L COMPLETE -P -A

Cluster mode

● Use training set

○ Supplied test set Set...

○ Percentage split % 66

○ Classes to clusters evaluation

(Nom) class

☑ Store clusters for visualization

Result list (right-click for options)

20:21:40 - HierarchicalClusterer

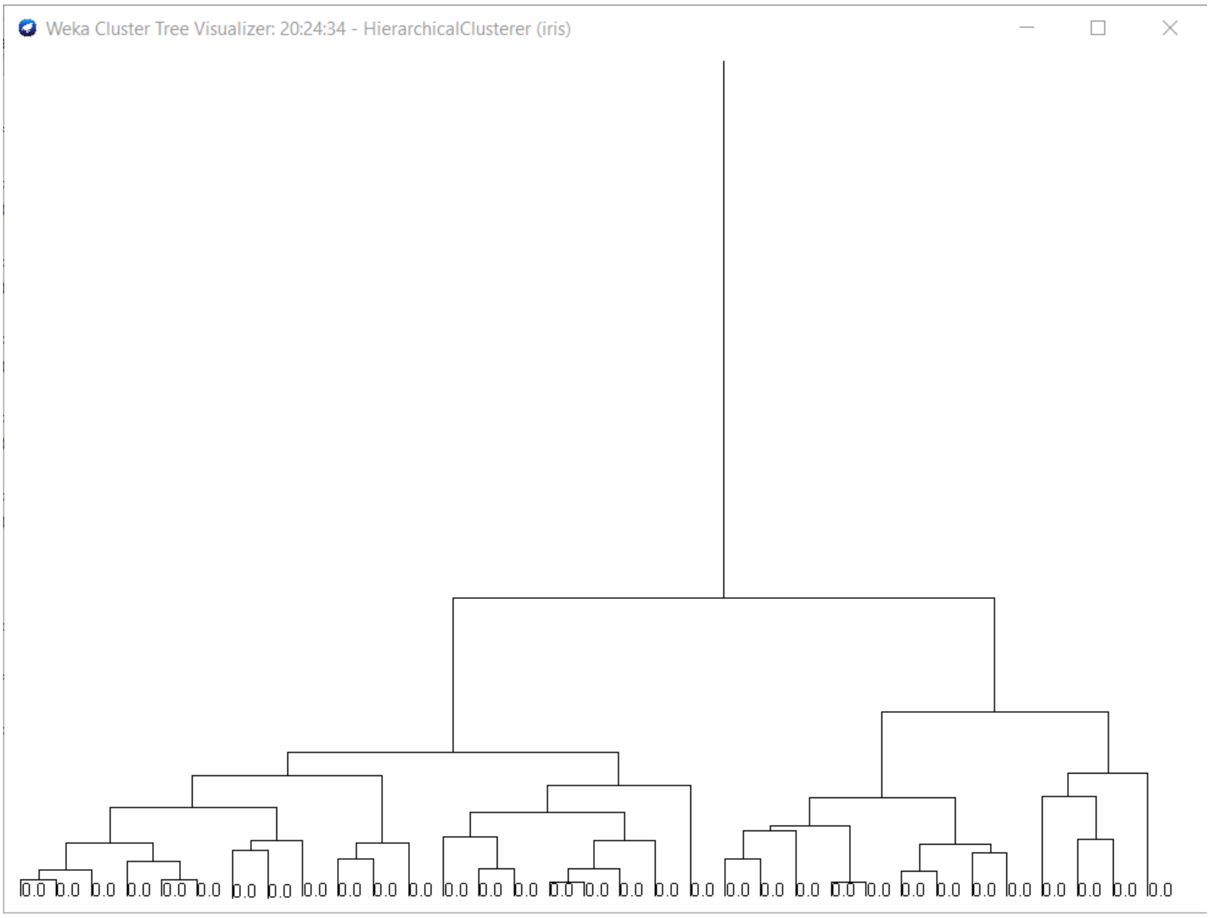
20:24:34 - HierarchicalClusterer

Status

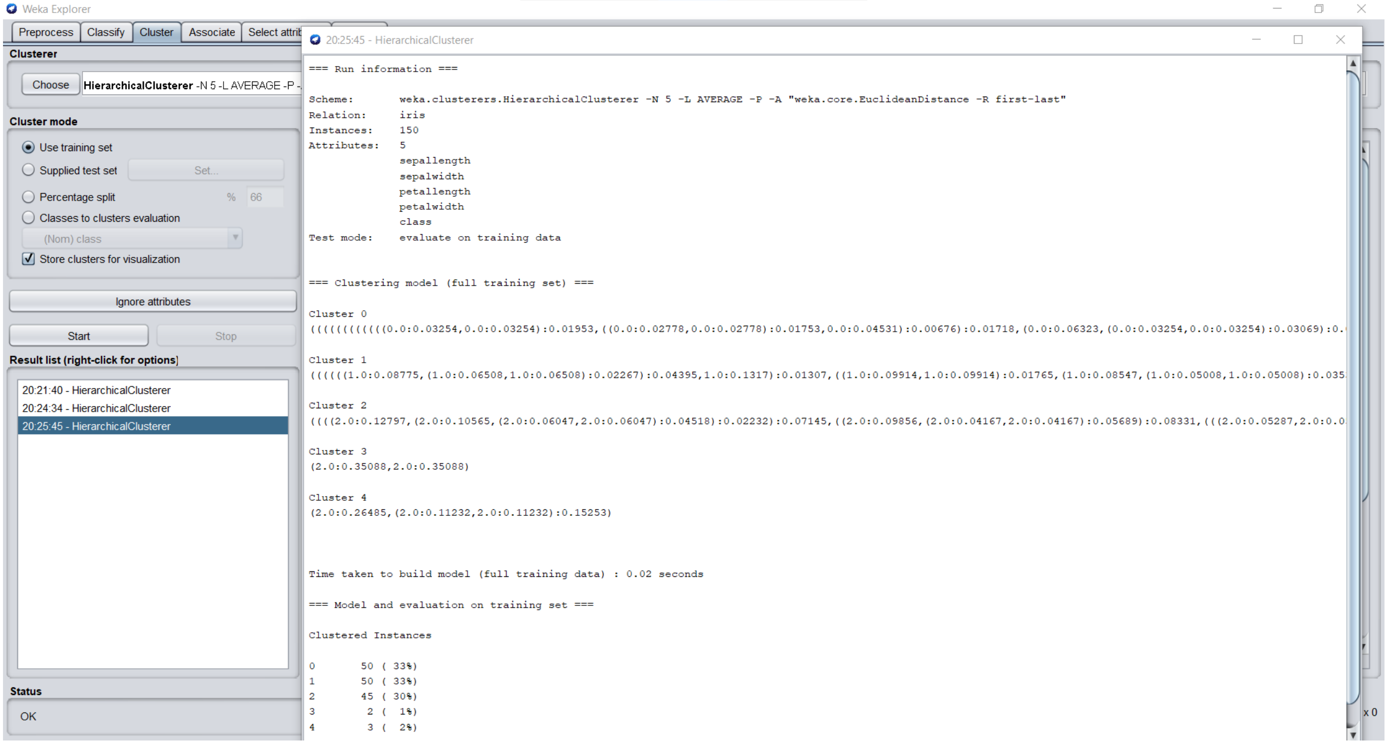
OK

```
=== Run information ===

Scheme:       weka.clusterers.HierarchicalClusterer -N 5 -L COMPLETE -P -A "weka.core.EuclideanDistance -R first-last"
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    evaluate on training data


=== Clustering model (full training set) ===

Cluster 0
((((((0.0:0.03254,0.0:0.03254):0.01754,0.0:0.05008):0.05008,(0.0:0.06514,(0.0:0.03254,0.0:0.03254):0.0326):0.03501):0.0627,((0.0:0.08429,0.0:0.08429):0.01868,0.0

Cluster 1
(((((0.0:0.04383,0.0:0.04383):0.03345,0.0:0.07728):0.02141,(((0.0:0,0.0:0):0,0.0:0):0.05287,0.0:0.05287):0.04582):0.04428,((0.0:0.03254,0.0:0.03254):0.04474,(0.0

Cluster 2
((((1.0:0.10206,(1.0:0.06508,1.0:0.06508):0.03698):0.0771,1.0:0.17916):0.0424,((1.0:0.09914,(1.0:0.08779,1.0:0.08779):0.01135):0.04781,(1.0:0.05008,1.0:0.05008):

Cluster 3
((((2.0:0.15339,(2.0:0.10972,(2.0:0.06047,2.0:0.06047):0.04924):0.04367):0.10143,((((2.0:0.04383,2.0:0.04383):0.03891,2.0:0.08274):0.02574,2.0:0.10848):0.06401,(2

Cluster 4
((((((2.0:0,2.0:0):0.08779,2.0:0.08779):0.04137,2.0:0.12917):0.08329,2.0:0.21246):0.12582,((2.0:0.08946,(2.0:0.05008,2.0:0.05008):0.03938):0.0932,((2.0:0.03254,2



Time taken to build model (full training data) : 0.02 seconds

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

Clustered Instances

0       33 ( 22%)
1       17 ( 11%)
2       50 ( 33%)
3       30 ( 20%)
4       20 ( 13%)
```

Complete Link Dendrogram:



Average Link Clustering using 5 clusters and Euclidian Distance:



HierarchicalClusterer -N 5 -L AVERAGE -P -

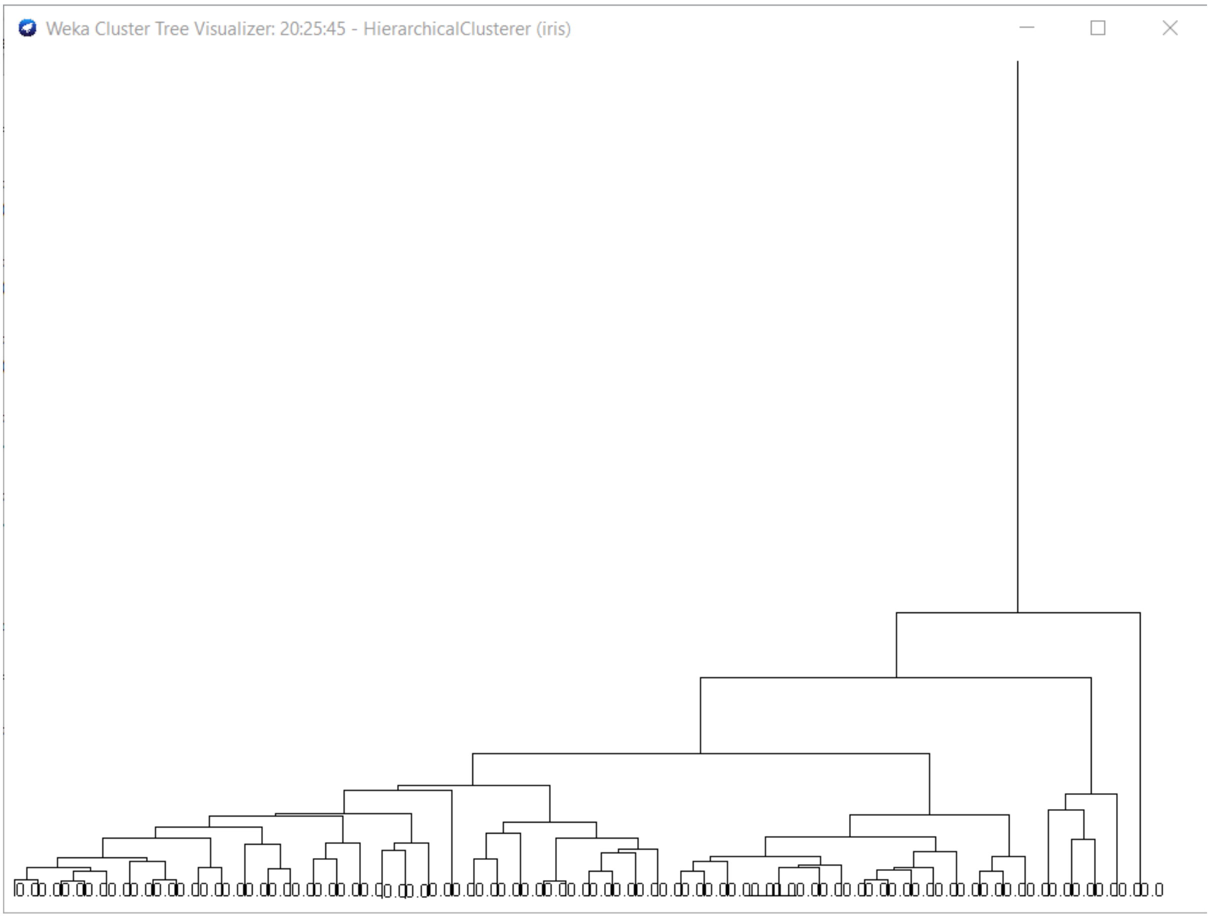
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) class
☑ Store clusters for visualization

20:21:40 - HierarchicalClusterer
20:24:34 - HierarchicalClusterer
20:25:45 - HierarchicalClusterer

OK

```
=== Run information ===

Scheme:       weka.clusterers.HierarchicalClusterer -N 5 -L AVERAGE -P -A "weka.core.EuclideanDistance -R first-last"
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    evaluate on training data


=== Clustering model (full training set) ===

Cluster 0
((((((((((((0.0:0.03254,0.0:0.03254):0.01953,((0.0:0.02778,0.0:0.02778):0.01753,0.0:0.04531):0.00676):0.01718,(0.0:0.06323,(0.0:0.03254,0.0:0.03254):0.03069):0.

Cluster 1
((((((1.0:0.08775,(1.0:0.06508,1.0:0.06508):0.02267):0.04395,1.0:0.1317):0.01307,((1.0:0.09914,1.0:0.09914):0.01765,(1.0:0.08547,(1.0:0.05008,1.0:0.05008):0.035

Cluster 2
((((2.0:0.12797,(2.0:0.10565,(2.0:0.06047,2.0:0.06047):0.04518):0.02232):0.07145,((2.0:0.09856,(2.0:0.04167,2.0:0.04167):0.05689):0.08331,(((2.0:0.05287,2.0:0.0

Cluster 3
(2.0:0.35088,2.0:0.35088)

Cluster 4
(2.0:0.26485,(2.0:0.11232,2.0:0.11232):0.15253)



Time taken to build model (full training data) : 0.02 seconds

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

Clustered Instances

0       50 ( 33%)
1       50 ( 33%)
2       45 ( 30%)
3        2 (  1%)
4        3 (  2%)
```

Average Link Dendrogram:


Weka Cluster Tree Visualizer: 20:25:45 - HierarchicalClusterer (iris)

Conclusions: We have successfully implemented Hierarchical Clustering (Single Link, Complete Link and Average Link) on iris data set using Weka Tool.