

WEKA – Clustering

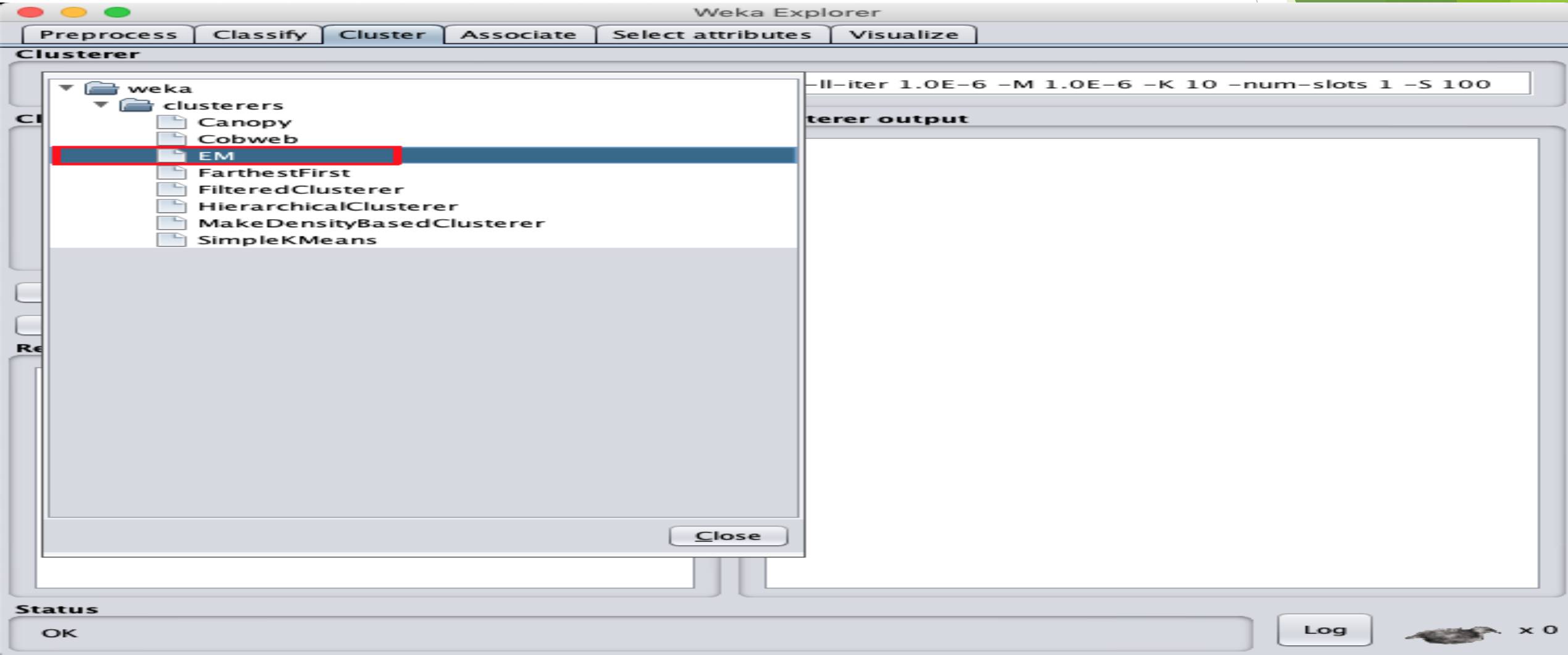
You can observe that there are 150 instances and 5 attributes. The names of attributes are listed as **sepalength**, **sepalwidth**, **petallength**, **petalwidth** and **class**.

The first four attributes are of numeric type while the **class** is a nominal type with 3 distinct values.

Examine each attribute to understand the features of the database. We will not do any preprocessing on this data and straight-away proceed to model building.

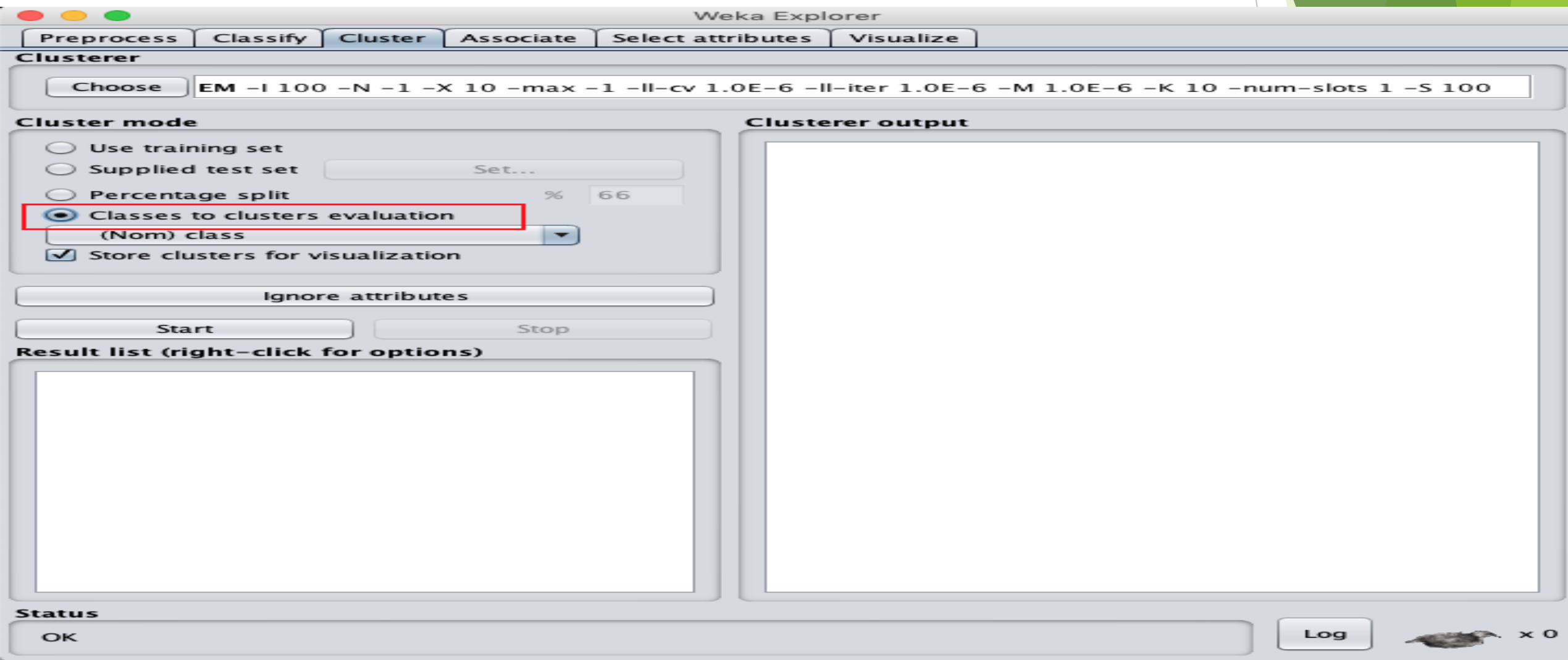
WEKA – Clustering

Clustering - Click on the **Cluster** TAB to apply the clustering algorithms to our loaded data. Click on the **Choose** button. You will see the following screen:



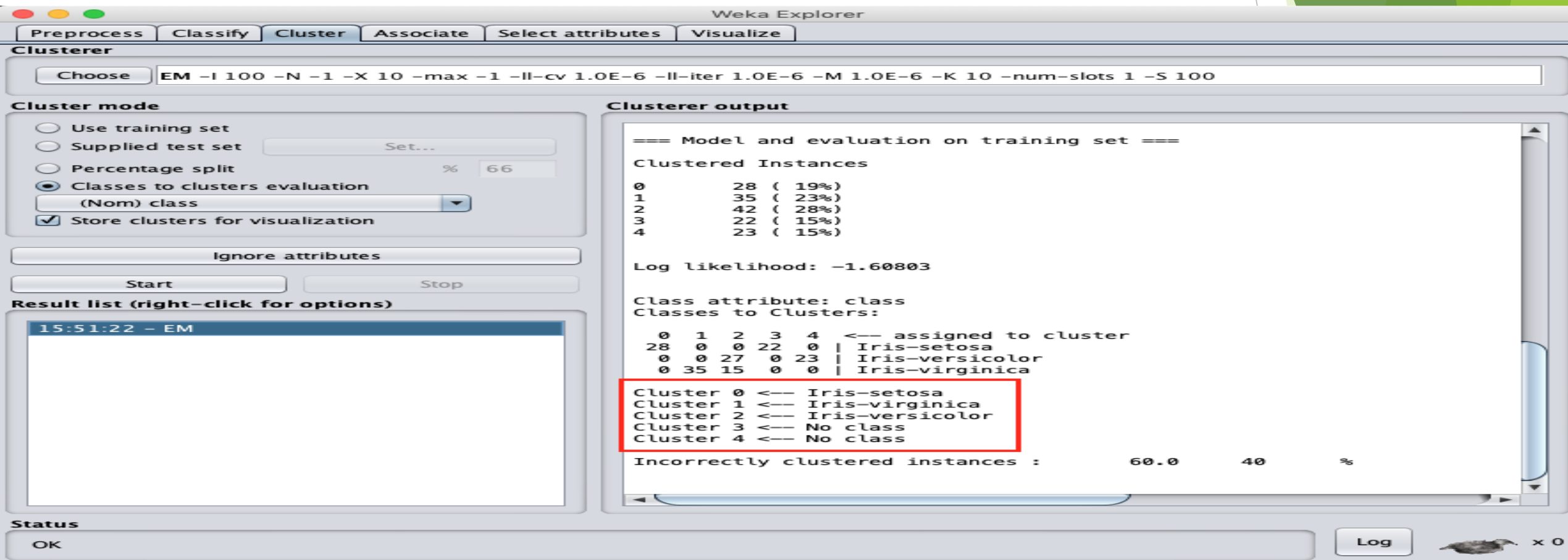
WEKA – Clustering

Now, select **EM** as the clustering algorithm. In the **Cluster mode** sub window, select the **Classes to clusters evaluation** option as shown in the screenshot below:



WEKA – Clustering

Click on the **Start** button to process the data. After a while, the results will be presented on the screen. Next, let us study the results. The output of the data processing is shown in the screen below:



From the output screen, you can observe that: □ There are 5 clustered instances detected in the database.

□ The **Cluster 0** represents setosa, **Cluster 1** represents virginica, **Cluster 2** represents versicolor, while the last two clusters do not have any class associated with them.

WEKA – Clustering

If you scroll up the output window, you will also see some statistics that gives the mean and standard deviation for each of the attributes in the various detected clusters. This is shown in the screenshot given below:

The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. The 'Clusterer' dropdown is set to 'EM' with the following command: `EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100`. The 'Cluster mode' section has 'Classes to clusters evaluation' selected with a '(Nom) class' dropdown and 'Store clusters for visualization' checked. The 'Clusterer output' window displays the following statistics:

```
=== Clustering model (full training set) ===  
  
EM  
===  
Number of clusters selected by cross validation: 5  
Number of iterations performed: 16
```

Attribute	Cluster 0 (0.18)	Cluster 1 (0.23)	Cluster 2 (0.28)	Cluster 3 (0.15)	Cluster 4 (0.15)
sepal.length					
mean	4.7748	6.8585	6.1613	5.2823	5.5432
std. dev.	0.2405	0.5228	0.4138	0.2407	0.3159
sepal.width					
mean	3.1789	3.0862	2.8547	3.7037	2.5786
std. dev.	0.2599	0.2891	0.2687	0.2857	0.2512
petal.length					
mean	1.4194	5.7859	4.7484	1.5173	3.863
std. dev.	0.1692	0.4745	0.3193	0.1592	0.3516
petal.width					
mean	0.1948	2.1327	1.5757	0.3028	1.1696
std. dev.	0.0557	0.2359	0.2196	0.1212	0.1351

The 'Status' bar at the bottom shows 'OK' and a 'Log' button.

Next, we will look at the visual representation of the clusters.

WEKA – Clustering

To visualize the clusters, right click on the **EM** result in the **Result list**. You will see the following options:

The screenshot shows the Weka Explorer interface. The top menu bar includes Preprocess, Classify, Cluster, Associate, Select attributes, and Visualize. The 'Cluster' tab is active. In the 'Clusterer' section, the 'Choose' button is selected, and the 'EM' algorithm is chosen with the following parameters: -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100. The 'Cluster mode' section has 'Classes to clusters evaluation' selected, with a dropdown menu showing '(Nom) class' and a checked box for 'Store clusters for visualization'. The 'Ignore attributes' button is also visible. The 'Result list (right-click for options)' shows a list of results, with '15:51:22 - EM' selected. A right-click context menu is open, showing options like 'View in main window', 'View in separate window', 'Save result buffer', 'Delete result buffer(s)', 'Load model', 'Save model', 'Re-evaluate model on current test set', 'Re-apply this model's configuration', 'Visualize cluster assignments' (highlighted with a red box), and 'Visualize tree'. The 'Clusterer output' section displays the clustering model results, including the number of clusters selected by cross validation (5) and the number of iterations performed (16). Below this, a table shows the mean and standard deviation for each attribute across five clusters.

Clusterer output

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

EM
==
Number of clusters selected by cross validation: 5
Number of iterations performed: 16
```

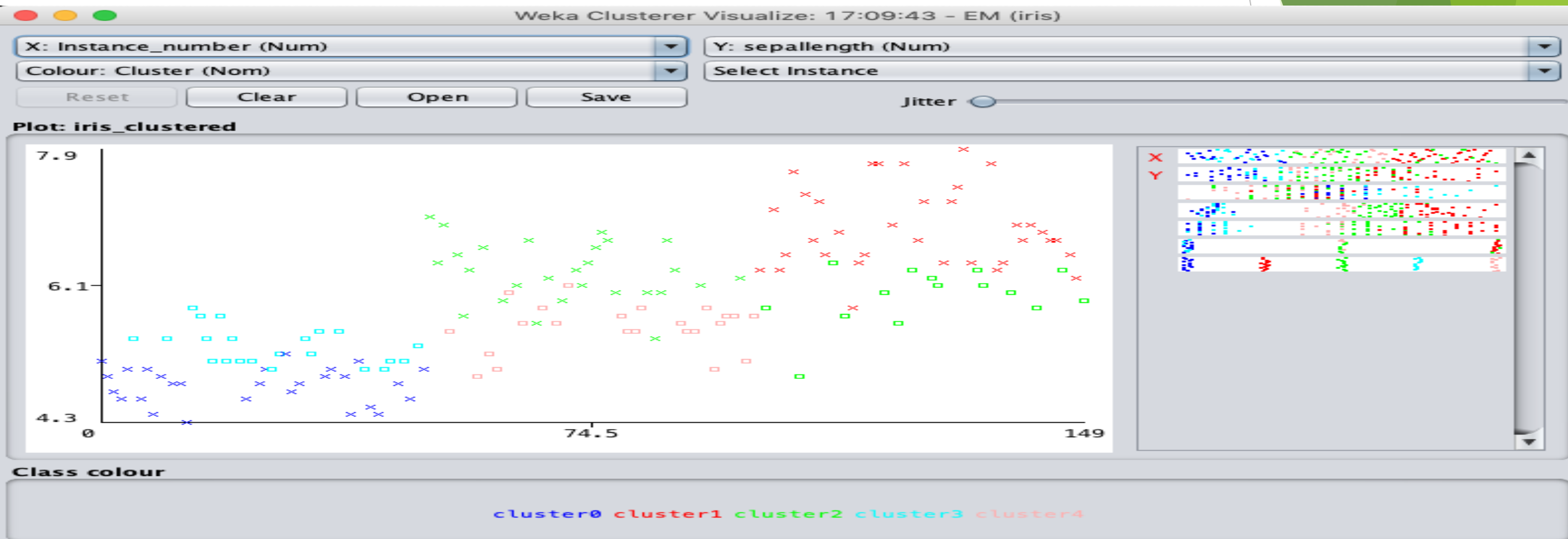
Attribute	Cluster 0 (0.18)	1 (0.23)	2 (0.28)	3 (0.15)	4 (0.15)
sepal length					
mean	4.7748	6.8585	6.1613	5.2823	5.5432
std. dev.	0.2405	0.5228	0.4138	0.2407	0.3159
sepal width					
mean	3.1789	3.0862	2.8547	3.7037	2.5786
std. dev.	0.2599	0.2891	0.2687	0.2857	0.2512
petal length					
mean	1.4194	5.7859	4.7484	1.5173	3.863
std. dev.	0.1692	0.4745	0.3193	0.1592	0.3516
petal width					
mean	0.1948	2.1327	1.5757	0.3028	1.1696
std. dev.	0.0557	0.2359	0.2196	0.1212	0.1351

Status

OK Log x 0

WEKA – Clustering

Select **Visualize cluster assignments**. You will see the following output:

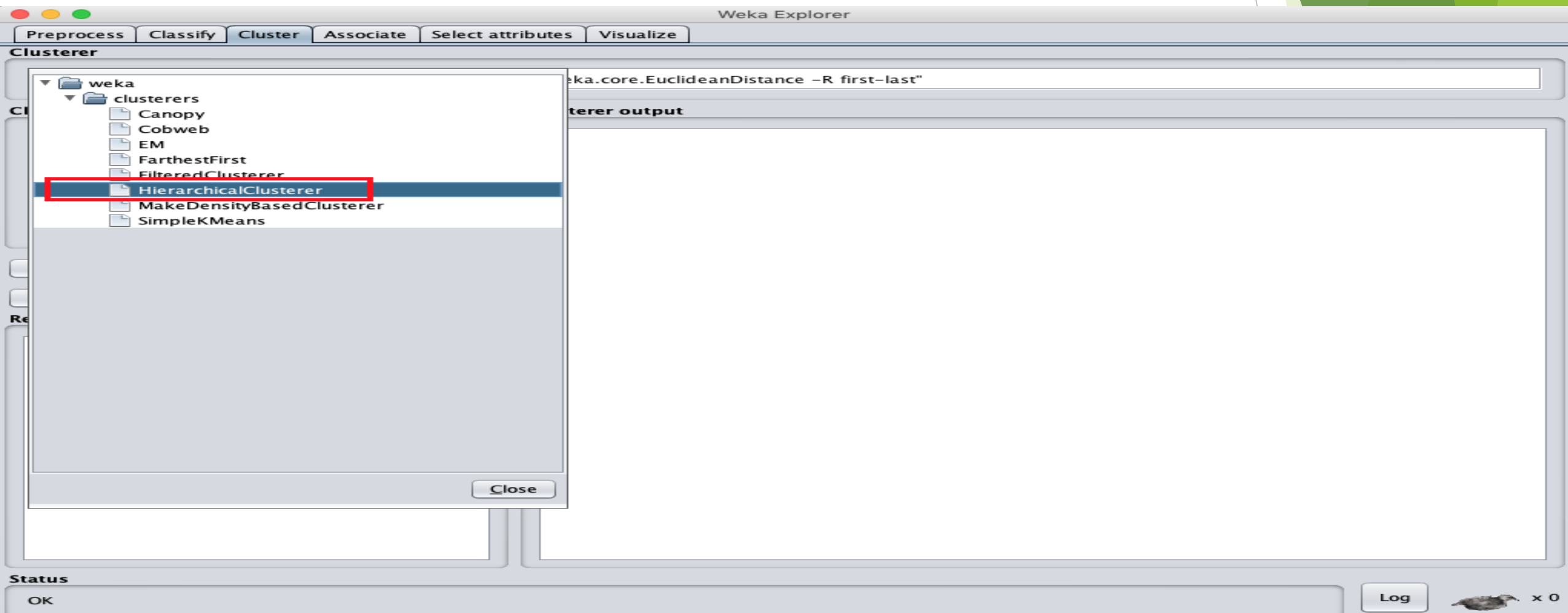


As in the case of classification, you will notice the distinction between the correctly and incorrectly identified instances. You can play around by changing the X and Y axes to analyze the results. You may use jittering as in the case of classification to find out the concentration of correctly identified instances. The operations in visualization plot are similar to the one you studied in the case of classification.

WEKA – Clustering

Applying Hierarchical Clusterer

To demonstrate the power of WEKA, let us now look into an application of another clustering algorithm. In the WEKA explorer, select the **HierarchicalClusterer** as your ML algorithm as shown in the screenshot shown below:



rt button. You will see the

The screenshot shows the Weka Explorer interface. At the top are tabs for Preprocess, Classify, Cluster, Associate, Select attributes, and Visualize. The "Cluster" tab is active. Below it, the "Clusterer" section has a dropdown menu set to "HierarchicalClusterer -N 2 -L SINGLE -P -A \"weka.core.EuclideanDistance -R first-last\"". Under "Cluster mode", three options are listed: "Use training set", "Supplied test set", and "Percentage split", all unselected. A red rectangle highlights the selected option, "Classes to clusters evaluation", which also includes "(Nom) class" and a percentage value of "66". There is a checked checkbox for "Store clusters for visualization". Buttons for "Ignore attributes", "Start", and "Stop" are present. On the left, a "Result list (right-click for options)" pane shows two entries: "17:23:50 - EM" and "17:24:11 - HierarchicalClusterer", with the latter being highlighted. The main area displays the output of the clustering process:

```
=== Clustering model (full training set) ===  
  
Cluster 0  
(((((((.....((0.2:0.03254,0.2:0.03254):0.00913,(0.3:0.03254,0.3:0.03254):0.00913):0.  
  
Cluster 1  
(((((((.....((1.4:0.07344,(((1.5:0.06508,1.5:0.06508):0.00066,(1.4:0.05008,  
  
Time taken to build model (full training data) : 0.03 seconds  
  
== Model and evaluation on training set ==  
  
Clustered Instances  
  
0      50 ( 33%)  
1     100 ( 67%)  
  
Class attribute: class  
Classes to Clusters:  
  
   0    1 <-- assigned to cluster  
50    0 | Iris-setosa  
   0    50 | Iris-versicolor  
   0    50 | Iris-virginica  
  
Cluster 0 <-- Iris-setosa  
Cluster 1 <-- Iris-versicolor  
  
Incorrectly clustered instances :           50.0       33.3333 %
```

The bottom status bar contains an "OK" button and a "Log" button next to a small animal icon.

Notice that in the **Result list**, there are two results listed: the first one is the EM result and the second one is the current Hierarchical. Likewise, you can apply multiple ML algorithms to the same dataset and quickly compare their results.

WEKA – Clustering

If you examine the tree produced by this algorithm, you will see the following output:

