Hierarchical Clustering

Groups items using a hierarchical clustering algorithm.

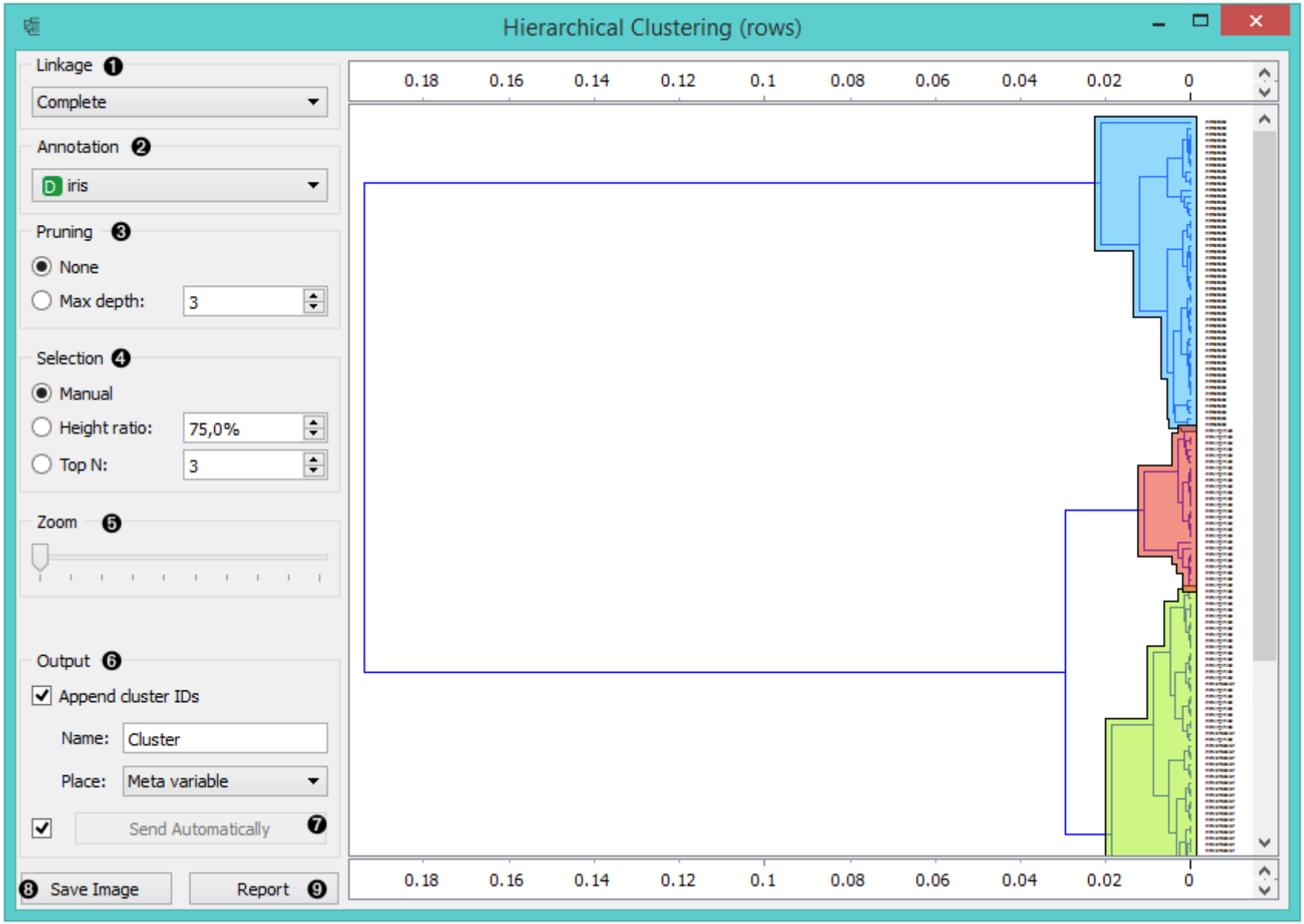
Inputs

- Distances: distance matrix

Outputs

- Selected Data: instances selected from the plot
- Data: data with an additional column showing whether an instance is selected

The widget computes hierarchical clustering of arbitrary types of objects from a matrix of distances and shows a corresponding dendrogram.

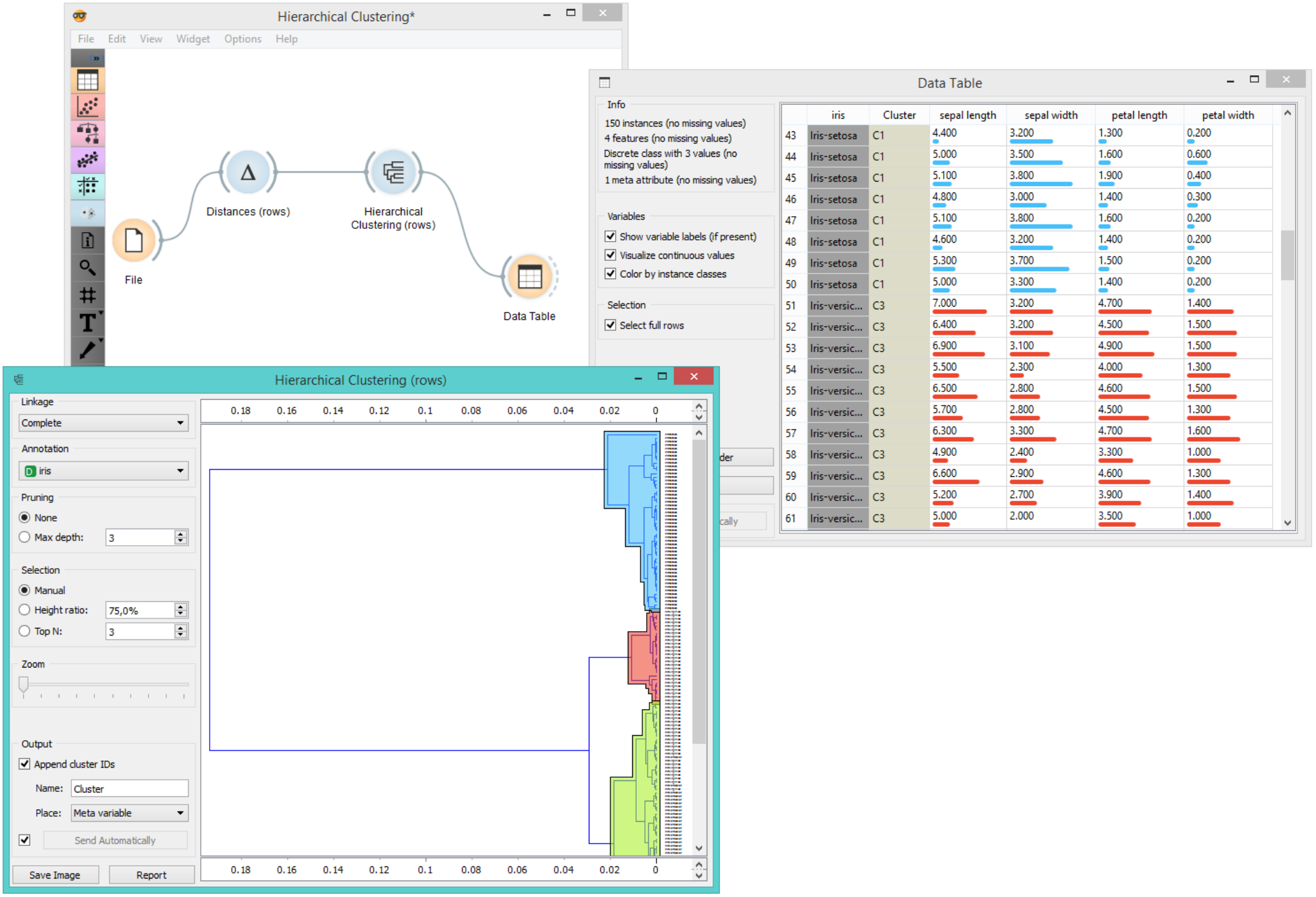


1. The widget supports four ways of measuring distances between clusters:
   - **Single linkage** computes the distance between the closest elements of the two clusters
   - **Average linkage** computes the average distance between elements of the two clusters
   - **Weighted linkage** uses the WPGMA method
   - **Complete linkage** computes the distance between the clusters' most distant elements

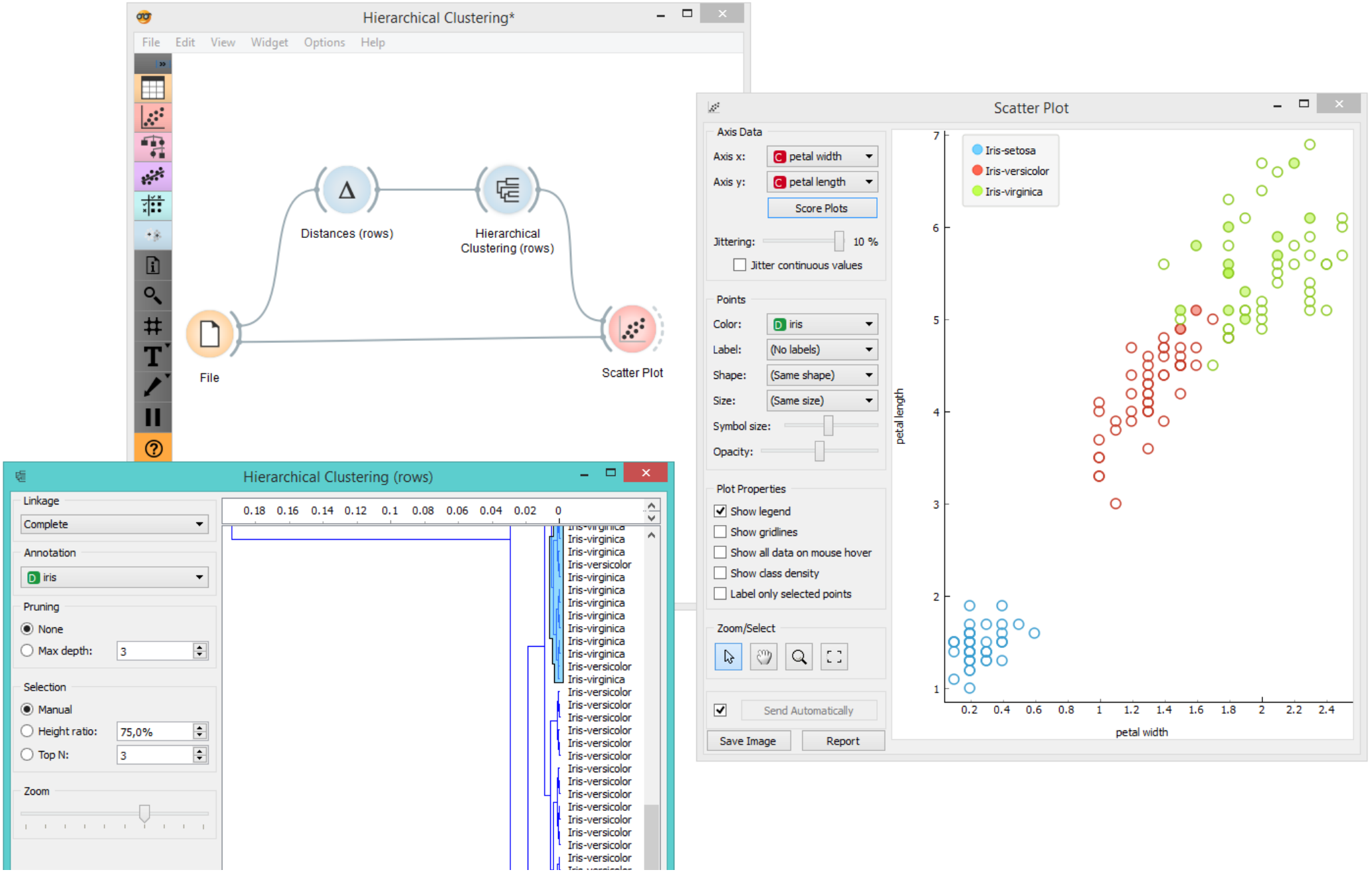
2. Labels of nodes in the dendrogram can be chosen in the **Annotation** box.
3. Huge dendrograms can be pruned in the *Pruning* box by selecting the maximum depth of the dendrogram. This only affects the display, not the actual clustering.
4. The widget offers three different selection methods:
    - **Manual** (Clicking inside the dendrogram will select a cluster. Multiple clusters can be selected by holding Ctrl/Cmd. Each selected cluster is shown in a different color and is treated as a separate cluster in the output.)
    - **Height ratio** (Clicking on the bottom or top ruler of the dendrogram places a cutoff line in the graph. Items to the right of the line are selected.)
    - **Top N** (Selects the number of top nodes.)
5. Use *Zoom* and scroll to zoom in or out.
6. If the items being clustered are instances, they can be added a cluster index (*Append cluster IDs*). The ID can appear as an ordinary **Attribute**, **Class attribute** or a **Meta attribute**. In the second case, if the data already has a class attribute, the original class is placed among meta attributes.
7. The data can be automatically output on any change (*Auto send is on*) or, if the box isn't ticked, by pushing *Send Data*.
8. Clicking this button produces an image that can be saved.
9. Produce a report.

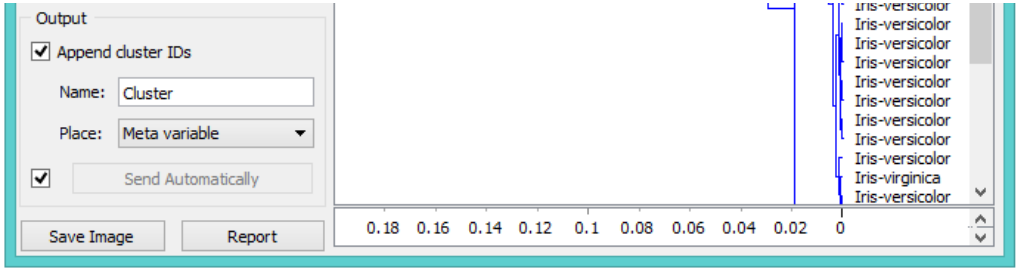
Examples

The workflow below shows the output of **Hierarchical Clustering** for the *Iris* dataset in Data Table widget. We see that if we choose *Append cluster IDs* in hierarchical clustering, we can see an additional column in the **Data Table** named *Cluster*. This is a way to check how hierarchical clustering clustered individual instances.


Hierarchical Clustering*
File Edit View Widget Options Help
Distances (rows)
Hierarchical Clustering (rows)
File
Data Table
Data Table
Info
150 instances (no missing values)
4 features (no missing values)
Discrete class with 3 values (no missing values)
1 meta attribute (no missing values)
Variables
Show variable labels (if present)
Visualize continuous values
Color by instance classes
Selection
Select full rows
iris Cluster sepal length sepal width petal length petal width
43 Iris-setosa C1 4.400 3.200 1.300 0.200
44 Iris-setosa C1 5.000 3.500 1.600 0.600
45 Iris-setosa C1 5.100 3.800 1.900 0.400
46 Iris-setosa C1 4.800 3.000 1.400 0.300
47 Iris-setosa C1 5.100 3.800 1.600 0.200
48 Iris-setosa C1 4.600 3.200 1.400 0.200
49 Iris-setosa C1 5.300 3.700 1.500 0.200
50 Iris-setosa C1 5.000 3.300 1.400 0.200
51 Iris-versic... C3 7.000 3.200 4.700 1.400
52 Iris-versic... C3 6.400 3.200 4.500 1.500
53 Iris-versic... C3 6.900 3.100 4.900 1.500
54 Iris-versic... C3 5.500 2.300 4.000 1.300
55 Iris-versic... C3 6.500 2.800 4.600 1.500
56 Iris-versic... C3 5.700 2.800 4.500 1.300
57 Iris-versic... C3 6.300 3.300 4.700 1.600
58 Iris-versic... C3 4.900 2.400 3.300 1.000
59 Iris-versic... C3 6.600 2.900 4.600 1.300
60 Iris-versic... C3 5.200 2.700 3.900 1.400
61 Iris-versic... C3 5.000 2.000 3.500 1.000
Hierarchical Clustering (rows)
Linkage
Complete
Annotation
iris
Pruning
None
Max depth: 3
Selection
Manual
Height ratio: 75,0%
Top N: 3
Zoom
Output
Append cluster IDs
Name: Cluster
Place: Meta variable
Send Automatically
Save Image
Report
0.18 0.16 0.14 0.12 0.1 0.08 0.06 0.04 0.02 0

In the second example, we loaded the *Iris* dataset again, but this time we added the Scatter Plot, showing all the instances from the File widget, while at the same time receiving the selected instances signal from **Hierarchical Clustering**. This way we can observe the position of the selected cluster(s) in the projection.




Output
Append cluster IDs
Name: Cluster
Place: Meta variable
Send Automatically
Save Image
Report
Iris-versicolor
Iris-versicolor
Iris-versicolor
Iris-versicolor
Iris-versicolor
Iris-versicolor
Iris-versicolor
Iris-versicolor
Iris-virginica
Iris-versicolor
0.18 0.16 0.14 0.12 0.1 0.08 0.06 0.04 0.02 0