Discretize

Discretizes continuous attributes from an input dataset.

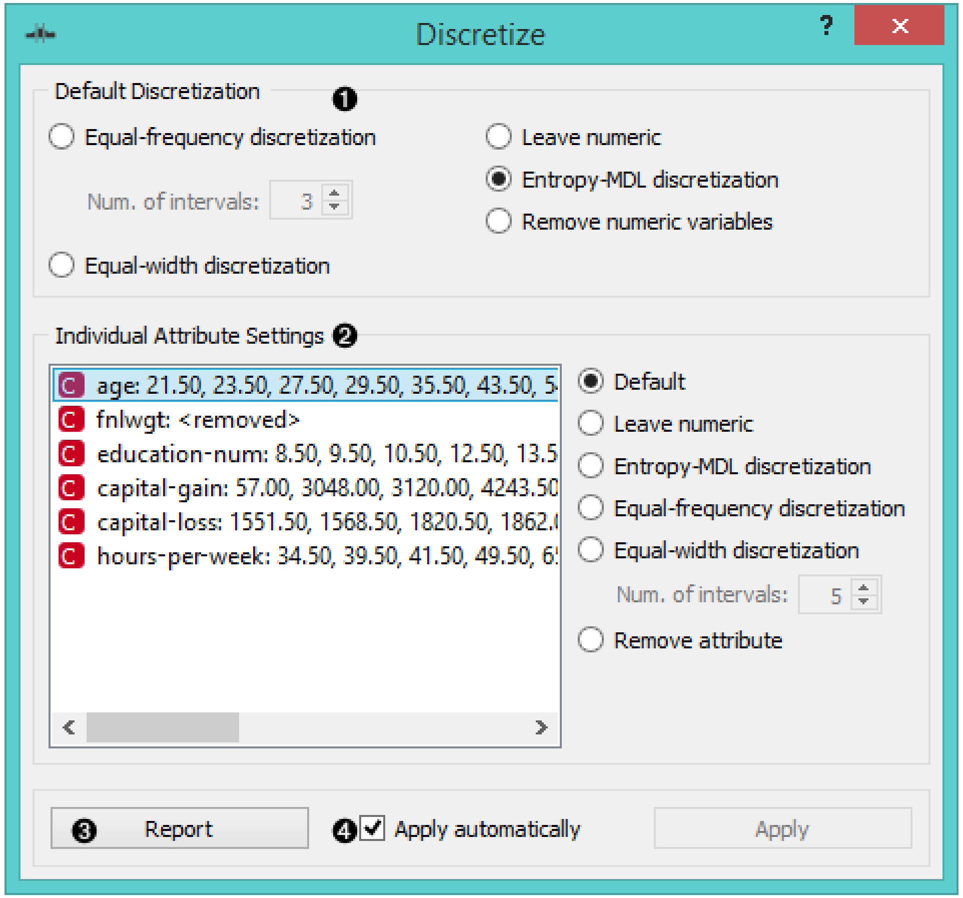
Inputs

- Data: input dataset

Outputs

- Data: dataset with discretized values

The **Discretize** widget discretizes continuous attributes with a selected method.



Discretize

Default Discretization 1

○ Equal-frequency discretization

Num. of intervals: 3

○ Equal-width discretization

○ Leave numeric

● Entropy-MDL discretization

○ Remove numeric variables

Individual Attribute Settings 2

C age: 21.50, 23.50, 27.50, 29.50, 35.50, 43.50, 5

C fnlwgt: <removed>

C education-num: 8.50, 9.50, 10.50, 12.50, 13.5

C capital-gain: 57.00, 3048.00, 3120.00, 4243.50

C capital-loss: 1551.50, 1568.50, 1820.50, 1862.

C hours-per-week: 34.50, 39.50, 41.50, 49.50, 6

● Default

○ Leave numeric

○ Entropy-MDL discretization

○ Equal-frequency discretization

○ Equal-width discretization

Num. of intervals: 5

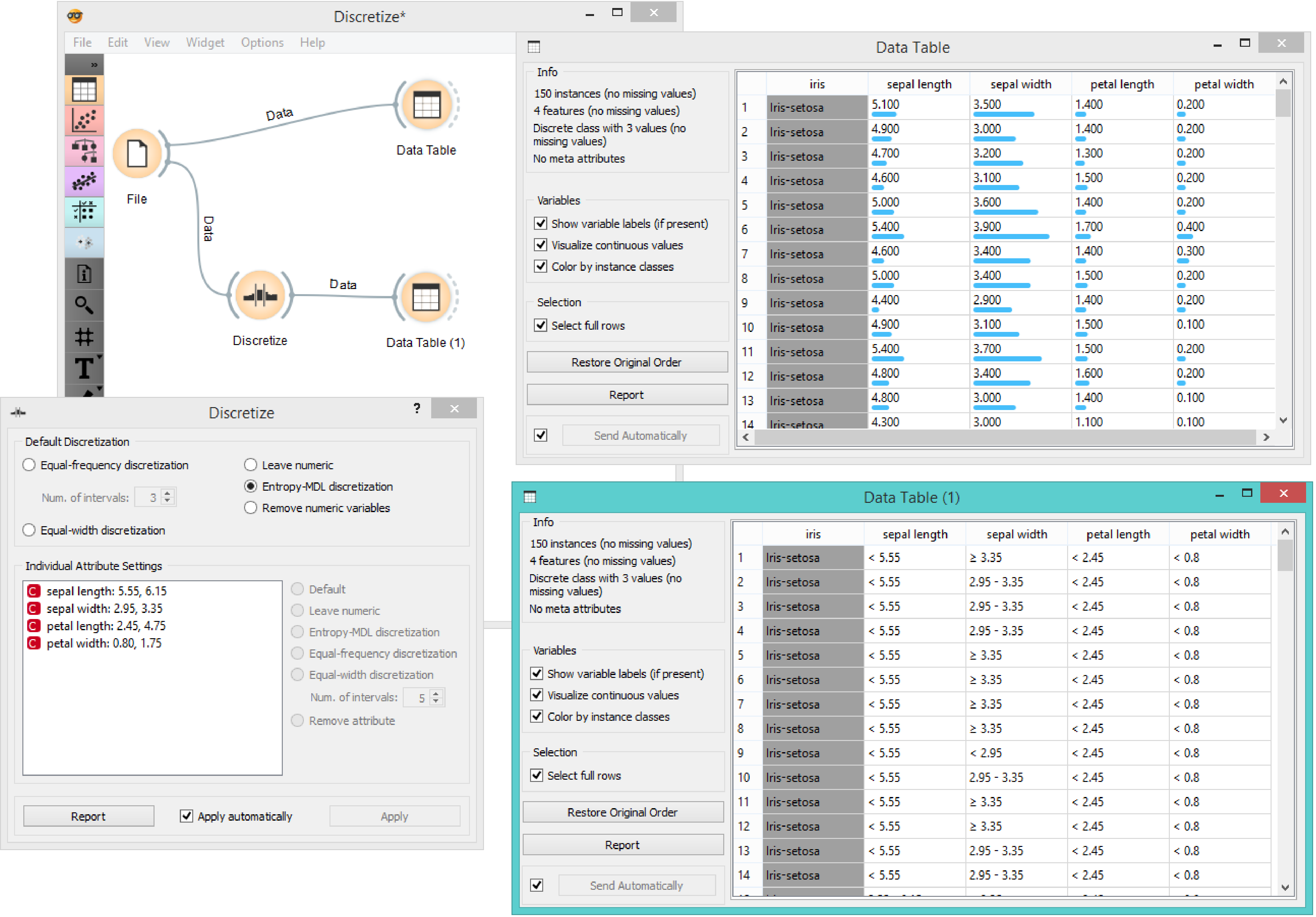
○ Remove attribute

1. The basic version of the widget is rather simple. It allows choosing between three different discretizations.

    - Entropy-MDL, invented by Fayyad and Irani is a top-down discretization, which recursively splits the attribute at a cut maximizing information gain, until the gain is lower than the minimal description length of the cut. This discretization can result in an arbitrary number of intervals, including a single interval, in which case the attribute is discarded as useless (removed).
    - Equal-frequency splits the attribute into a given number of intervals, so that they each contain approximately the same number of instances.
    - Equal-width evenly splits the range between the smallest and the largest observed value. The *Number of intervals* can be set manually.
    - The widget can also be set to leave the attributes continuous or to remove them.
2. To treat attributes individually, go to **Individual Attribute Settings**. They show a specific discretization of each attribute and allow changes. First, the top left list shows the cut-off points for each attribute. In the snapshot, we used the entropy-MDL discretization, which determines the optimal number of intervals automatically; we can see it discretized the age into seven intervals with cut-offs at 21.50, 23.50, 27.50, 35.50, 43.50, 54.50 and 61.50, respectively, while the capital-gain got split into many intervals with several cut-offs. The final weight (fnlwgt), for instance, was left with a single interval and thus removed. On the right, we can select a specific discretization method for each attribute. Attribute *"fnlwgt"* would be removed by the MDL-based discretization, so to prevent its removal, we select the attribute and choose, for instance, **Equal-frequency discretization**. We could also choose to leave the attribute continuous.
3. Produce a report.
4. Tick *Apply automatically* for the widget to automatically commit changes. Alternatively, press *Apply*.

Example

In the schema below, we show the *Iris* dataset with continuous attributes (as in the original data file) and with discretized attributes.



Discretize

Default Discretization

○ Equal-frequency discretization

Num. of intervals: 3

○ Equal-width discretization

○ Leave numeric

● Entropy-MDL discretization

○ Remove numeric variables

Individual Attribute Settings

- C sepal length: 5.55, 6.15
- C sepal width: 2.95, 3.35
- C petal length: 2.45, 4.75
- C petal width: 0.80, 1.75

○ Default

○ Leave numeric

○ Entropy-MDL discretization

○ Equal-frequency discretization

○ Equal-width discretization

Num. of intervals: 5

○ Remove attribute

Report ☑ Apply automatically Apply

Data Table

Info

150 instances (no missing values)
4 features (no missing values)
Discrete class with 3 values (no missing values)
No meta attributes

Variables

☑ Show variable labels (if present)

☑ Visualize continuous values

☑ Color by instance classes

Selection

☑ Select full rows

Restore Original Order

Report

☑ Send Automatically

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.100	3.500	1.400	0.200
2	Iris-setosa	4.900	3.000	1.400	0.200
3	Iris-setosa	4.700	3.200	1.300	0.200
4	Iris-setosa	4.600	3.100	1.500	0.200
5	Iris-setosa	5.000	3.600	1.400	0.200
6	Iris-setosa	5.400	3.900	1.700	0.400
7	Iris-setosa	4.600	3.400	1.400	0.300
8	Iris-setosa	5.000	3.400	1.500	0.200
9	Iris-setosa	4.400	2.900	1.400	0.200
10	Iris-setosa	4.900	3.100	1.500	0.100
11	Iris-setosa	5.400	3.700	1.500	0.200
12	Iris-setosa	4.800	3.400	1.600	0.200
13	Iris-setosa	4.800	3.000	1.400	0.100
14	Iris-setosa	4.300	3.000	1.100	0.100

Data Table (1)

Info

150 instances (no missing values)
4 features (no missing values)
Discrete class with 3 values (no missing values)
No meta attributes

Variables

☑ Show variable labels (if present)

☑ Visualize continuous values

☑ Color by instance classes

Selection

☑ Select full rows

Restore Original Order

Report

☑ Send Automatically

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	< 5.55	≥ 3.35	< 2.45	< 0.8
2	Iris-setosa	< 5.55	2.95 - 3.35	< 2.45	< 0.8
3	Iris-setosa	< 5.55	2.95 - 3.35	< 2.45	< 0.8
4	Iris-setosa	< 5.55	2.95 - 3.35	< 2.45	< 0.8
5	Iris-setosa	< 5.55	≥ 3.35	< 2.45	< 0.8
6	Iris-setosa	< 5.55	≥ 3.35	< 2.45	< 0.8
7	Iris-setosa	< 5.55	≥ 3.35	< 2.45	< 0.8
8	Iris-setosa	< 5.55	≥ 3.35	< 2.45	< 0.8
9	Iris-setosa	< 5.55	< 2.95	< 2.45	< 0.8
10	Iris-setosa	< 5.55	2.95 - 3.35	< 2.45	< 0.8
11	Iris-setosa	< 5.55	≥ 3.35	< 2.45	< 0.8
12	Iris-setosa	< 5.55	≥ 3.35	< 2.45	< 0.8
13	Iris-setosa	< 5.55	2.95 - 3.35	< 2.45	< 0.8
14	Iris-setosa	< 5.55	2.95 - 3.35	< 2.45	< 0.8