

图 82

sepal_width	petal_length	species	sepal_length	petal_width
3.5	1.4	setosa	5.1	0.2
3	1.4	setosa	4.9	0.2
3.2	1.3	setosa	4.7	0.2
3.1	1.5	setosa	4.6	0.2
3.6	1.4	setosa	5	0.2
3.9	1.7	setosa	5.4	0.4
3.4	1.4	setosa	4.6	0.3
3.4	1.5	setosa	5	0.2

共 149 条 25 条/页 < 1 2 3 4 5 6 > 前往 1 页

图 83

3.4.2.9 数据离散化

图标:  数据离散化

描述: 某些模型算法，特别是某些分类算法如 ID3 决策树算法和 Apriori 算法等，要求数据是离散的，此时就需要将连续型特征（数值型）变换成离散型特征（类别型），即连续特征离散化。常用的离散化方法主要有三种：等宽法，等频法和通过聚类分析离散化（一维）。

字段属性

待离散化数据: 必选。请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，

下个组件可能无法获取所有列。勾选多列时，自动对每一列数据进行离散化如图 84 所示。



字段	类型	取值范围
a	数值	-
b	数值	-

* 待离散化数据

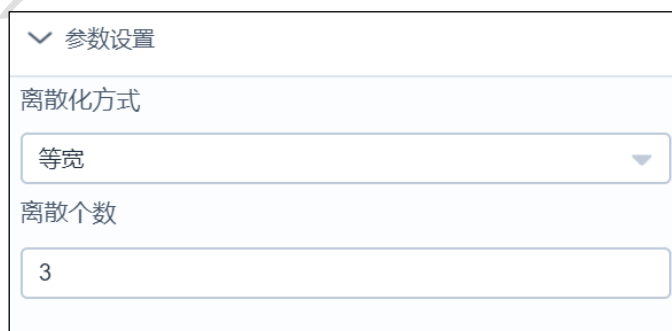
a × b ×

图 84

参数设置

离散化方式：选取要使用的离散方式，支持等宽、等频、聚类离散化，默认等宽。

离散个数：离散的个数，默认 2，如图 85 所示。



参数设置

离散化方式

等宽

离散个数

3

图 85

输出

表结果：对勾选的每一列进行离散化后的结果。

报告：无。

示例

下面对数据进行离散化。原数据如图 86 所示。

预览数据	
a	b
1	2
2	3
3	4
4	5
5	6

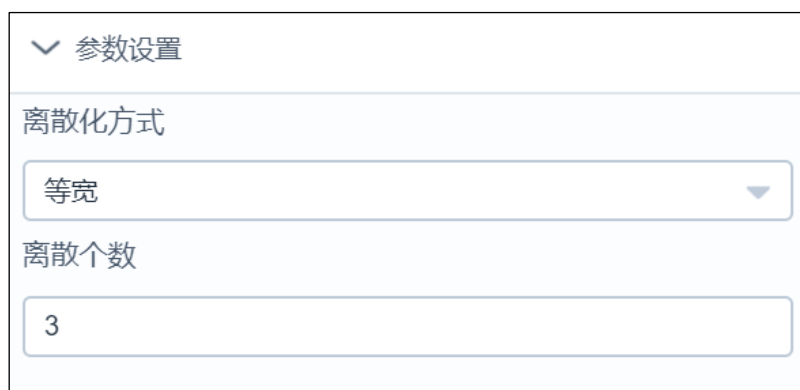
图 86

- 勾选需要进行离散化的数据，如图 87 所示。
- 选择离散化方式为等宽，离散个数为 2。如图 88 所示。
- 运行该组件，右击选择查看数据，如图 89 所示。

The screenshot shows a data processing workflow in a software interface. On the left, a canvas displays two components: '输入源' (Input Source) and '数据离散化' (Data Discretization), connected by a flow line. The right-hand panel is titled '字段属性' (Field Properties) and contains a table of field information. Below the table, there is a section for '待离散化数据' (Data to be Discretized) with a dropdown menu showing 'a' and 'b' selected. At the bottom of the panel, there are expandable sections for '参数设置' (Parameter Settings) and '组件描述' (Component Description).

字段	类型	取值范围
a	数值	-
b	数值	-

图 87



参数设置

离散化方式

等宽

离散个数

3

图 88



预览数据	
a	b
(0.995, 2.333]	(1.995, 3.333]
(0.995, 2.333]	(1.995, 3.333]
(2.333, 3.667]	(3.333, 4.667]
(3.667, 5.0]	(4.667, 6.0]
(3.667, 5.0]	(4.667, 6.0]

图 89

3.4.2.10 排序

图标: 

描述: 根据某一列的顺序将所有数据重新排序。

字段属性

特征列: 勾选的列必须包含关键字段。勾选的列将传入下一个组件。如图 90 所示。

关键字: 必选，由该列值的顺序将待排序的数据重新排序。如图 91 所示。