3.4.7 聚类

3.4.7.1 K-Means

图标： K-Means

描述：K-Means 是 Mac Queen 提出的一种非监督实时聚类算法，在最小化误差函数的基础上将数据划分为预定的类数 K。

字段属性

特征列：需要进行聚类的列，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。如图 351 所示。

图 351

参数设置

聚类个数：聚类的个数，默认 3。

最大迭代次数：迭代的次数。

如图 352 所示。

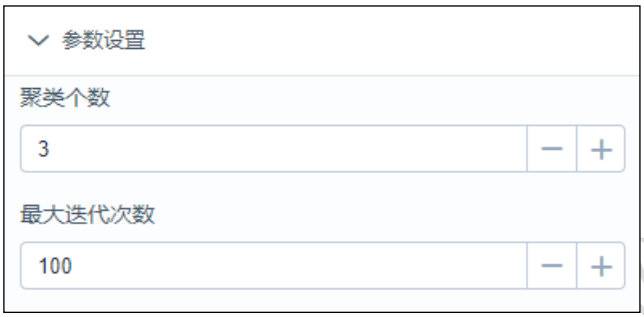

图 352

输出

表结果：包含聚类结果的数据表。

报告：聚类中心、饼图。

示例

下面对某数据进行 K-Means 聚类。

- 选择待聚类的序列，数据必须为数值型。如图 353 所示。
- 点击参数设置，聚类个数设置为 3，最大迭代次数设置为 100。如图 354 所示。
- 运行该组件，对组件右击，选择查看数据与报告，结果如图 355 与图 356 所示。

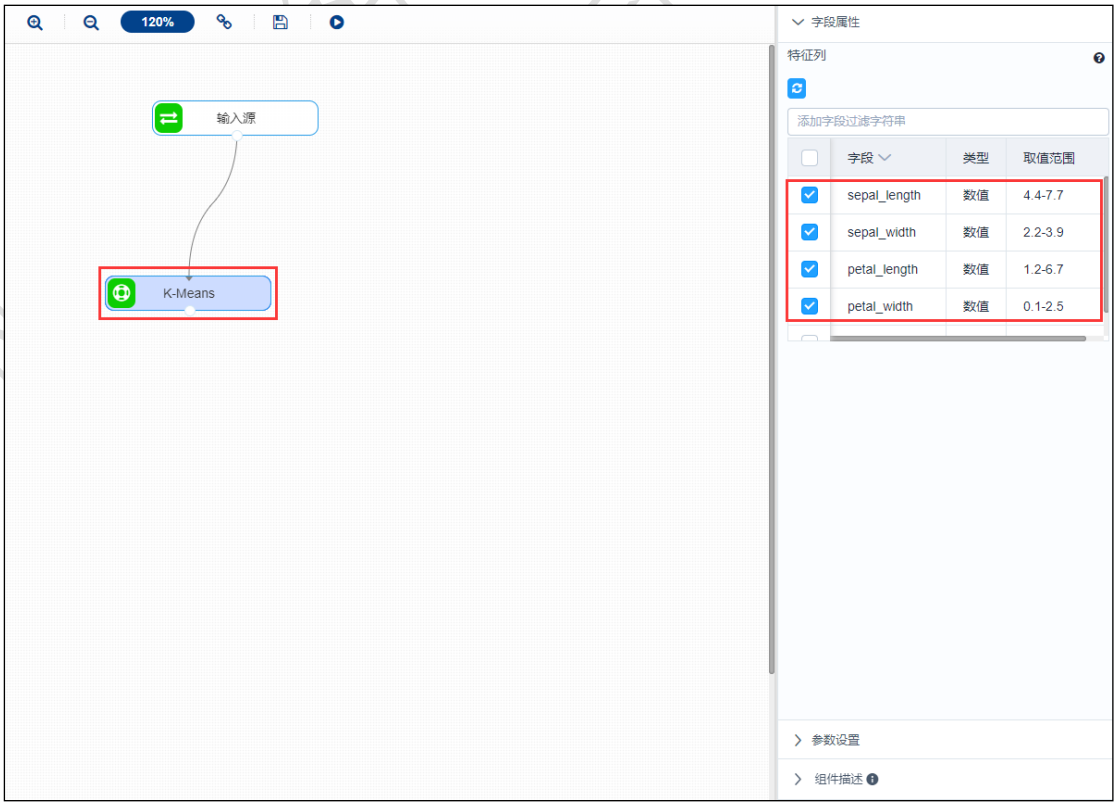

图 353

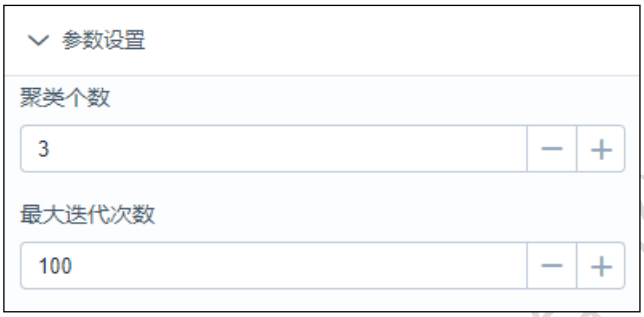

图 354

sepal_length	sepal_width	petal_length	petal_width	cluster_id
5.1	3.5	1.4	0.2	2
4.9	3	1.4	0.2	2
4.7	3.2	1.3	0.2	2
4.6	3.1	1.5	0.2	2
5	3.6	1.4	0.2	2
5.4	3.9	1.7	0.4	2
4.6	3.4	1.4	0.3	2
5	3.4	1.5	0.2	2

共 150 条 25 条/页 1 2 3 4 5 6 前往 1 页

图 355

聚类中心

cluster_id	sepal_length	sepal_width	petal_length	petal_width
1	5.901612903225806	2.7483870967741937	4.393548387096774	1.4338709677419355
2	5.006	3.428	1.4619999999999997	0.24600000000000033
3	6.85	3.0736842105263156	5.742105263157894	2.0710526315789473

饼图

图 356

3.4.7.2 GMM(高斯混合模型)

图标： GMM

描述：GMM（高斯混合模型）是用高斯概率密度函数精确地量化事物，将一个事物分解为

若干的基于高斯概率密度函数形成的模型。

字段属性

特征列：需要进行聚类的列，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。如图 357 所示。

图 357

参数设置

聚类个数：聚类的个数，默认 2。

指定协方差类型：包括球状型、结点型、对角型、全型，其中球状型：所有的分模型的协方差矩阵都是一个标量值；结点型：所有的分模型都共享一个协方差矩阵；对角型：每个分模型的协方差矩阵都是对角矩阵；全型：每个分模型都有自己的协方差矩阵。

指定初始化次数：默认为 1。

指定初始化权重的策略：默认为 kmeans。

如图 358 所示。

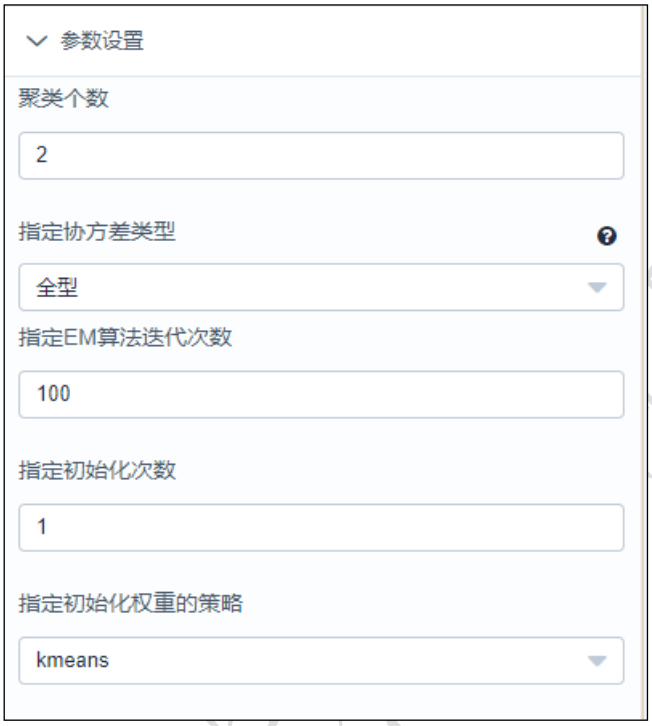

图 358

输出

表结果：包含聚类结果的数据表。

报告：无。

示例

下面对某数据进行 GMM 聚类。

- 选择待聚类的序列，数据必须为数值型。如图 359 所示。
- 运行该组件，对组件右击，选择查看数据，结果如图 360 所示。

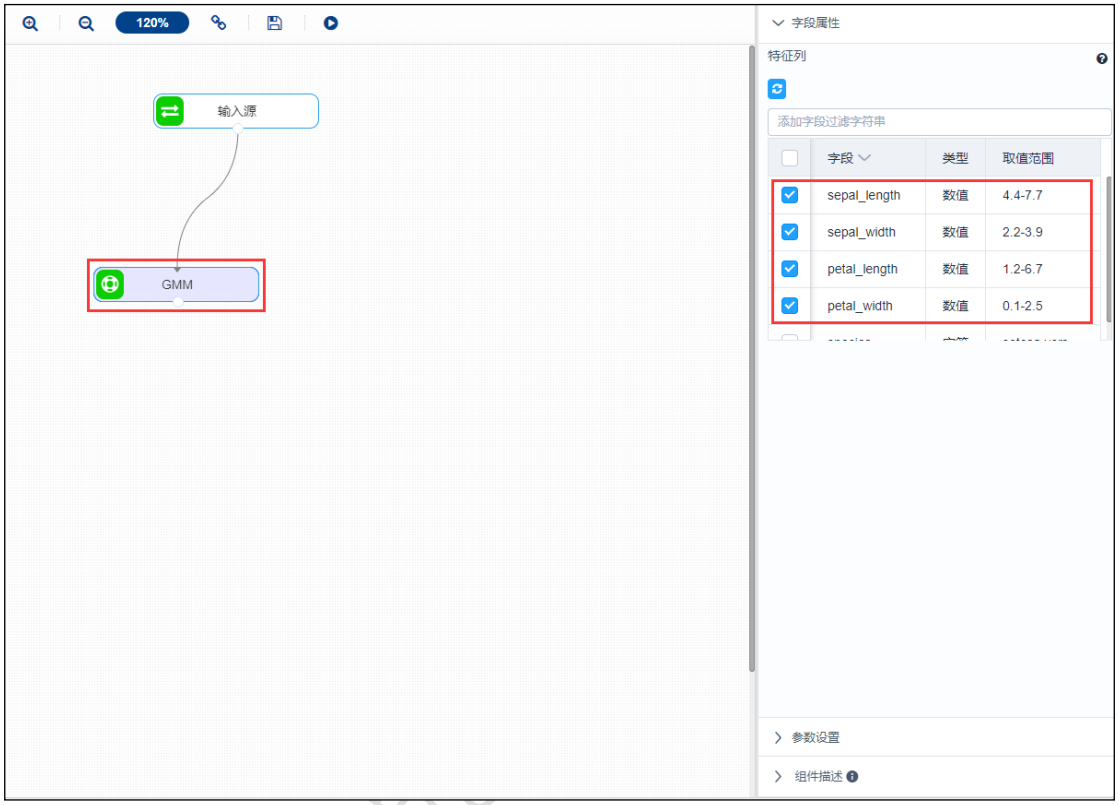

图 359

sepal_length	sepal_width	petal_length	petal_width	cluster_id
5.1	3.5	1.4	0.2	1
4.9	3	1.4	0.2	1
4.7	3.2	1.3	0.2	1
4.6	3.1	1.5	0.2	1
5	3.6	1.4	0.2	1
5.4	3.9	1.7	0.4	1
4.6	3.4	1.4	0.3	1
5	3.4	1.5	0.2	1

共 150 条 25 条/页 1 2 3 4 5 6 前往 1 页

图 360

3.4.7.3 密度聚类

图标： 

描述：密度聚类的核心思想是从某个核心点出发，不断向密度可达的区域扩张，从而得

到一个包含核心点和边界点的最大化区域，区域中任意两点密度相连。对于噪声样本，其簇标记为-1。

字段属性

特征列：需要进行聚类的列，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。如图 361 所示。

图 361

参数设置

邻域半径：设置某个半径，默认 0.5。

邻域内最小数目：设置邻域内最小点的个数，默认 5。

如图 362 所示。

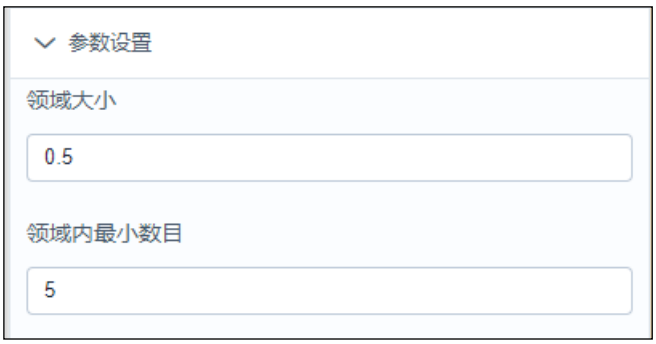

图 362

输出

表结果：包含聚类结果的数据表。

报告：无。

示例

下面对某数据进行密度聚类。

- 选择待聚类的序列，数据必须为数值型，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。。如图 363 所示。
- 运行该组件，对组件右击，选择查看数据，结果如图 364 所示。

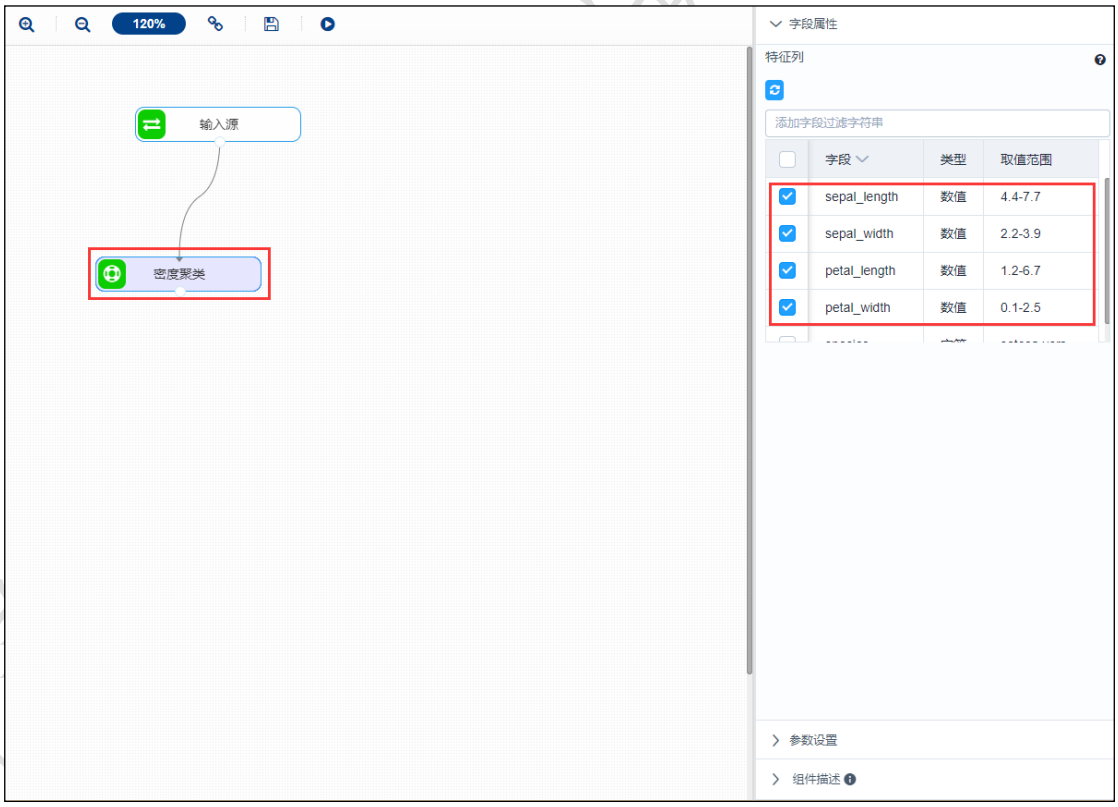

图 363

sepal_length	sepal_width	petal_length	petal_width	cluster_id
5.1	3.5	1.4	0.2	0
4.9	3	1.4	0.2	0
4.7	3.2	1.3	0.2	0
4.6	3.1	1.5	0.2	0
5	3.6	1.4	0.2	0
5.4	3.9	1.7	0.4	0
4.6	3.4	1.4	0.3	0
5	3.4	1.5	0.2	0

共 150 条 25 条/页 1 2 3 4 5 6 前往 1 页

图 364

3.4.7.4 K-Medoids

图标： K-Medoids

描述：K-Medioids 是 Kmeans 算法的改进，减轻了 Kmeans 算法对孤立点的敏感性，选用簇中离平均值最近的对象作为簇中心。

字段属性

特征列：需要进行聚类的列，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。如图 365 所示。

图 365

参数设置

聚类个数：聚类的个数，默认 3。

最大迭代次数：迭代的次数。

如图 366 所示。

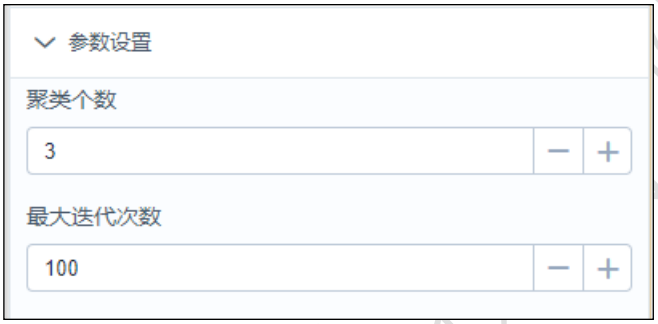

图 366

输出

表结果：包含聚类结果的数据表。

报告：无。

示例

下面对某数据进行 K-Medoids。

- 选择待聚类的序列，数据必须为数值型。如图 367 所示。
- 运行该组件，对组件右击，选择查看数据和报告，结果如图 368 与图 369 所示。

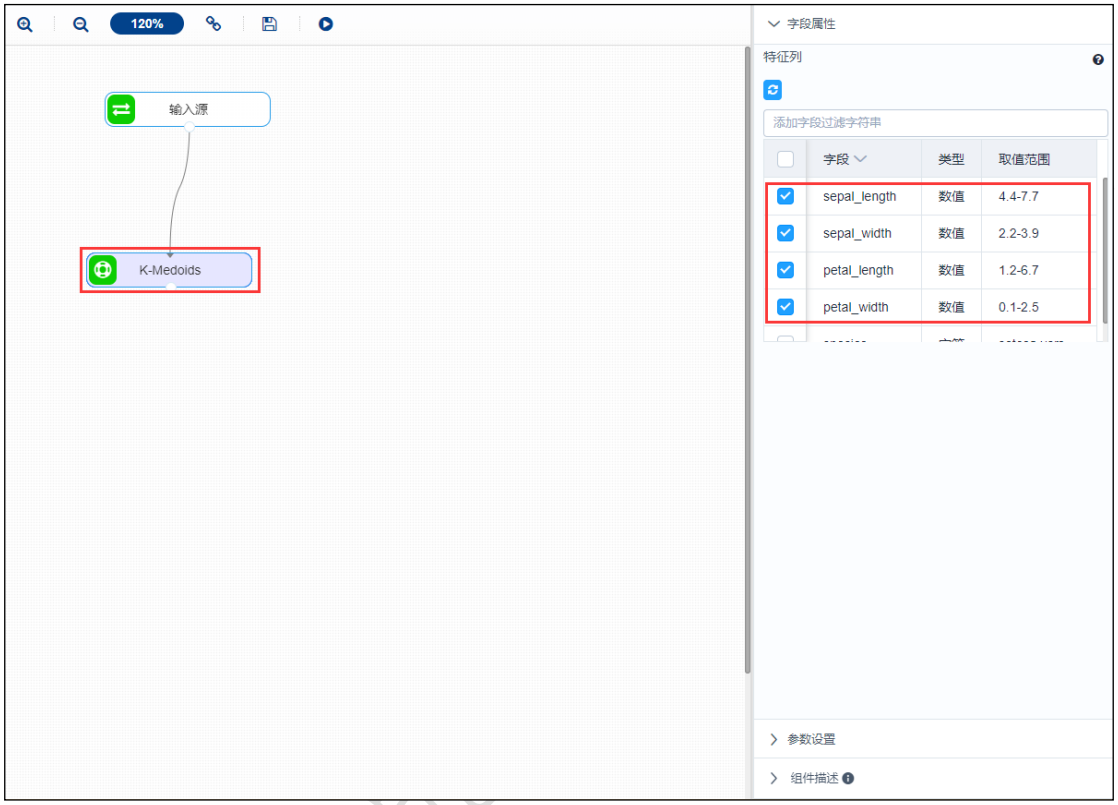

图 367

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

共 150 条 25 条/页 1 2 3 4 5 6 前往 1 页

图 368

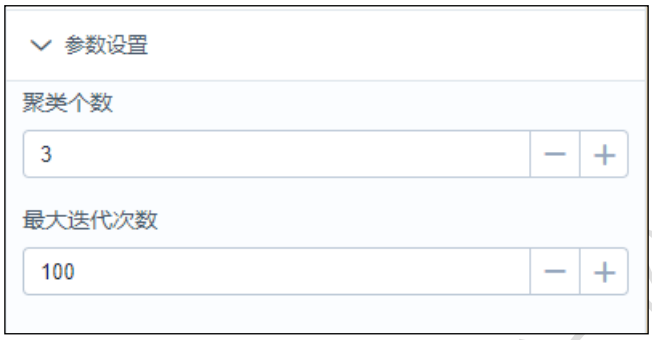

图 369

3.4.7.5 层次聚类

图标： 层次聚类

描述：层次聚类也叫系统聚类，分类单位所处的位置越低，其所包含的个体越少，但这些个体间的共同特征越多。

字段属性

特征列：需要进行聚类的列，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。如图 370 所示。

图 370

参数设置

输出聚类数：默认显示 2 类。如图 371 所示。

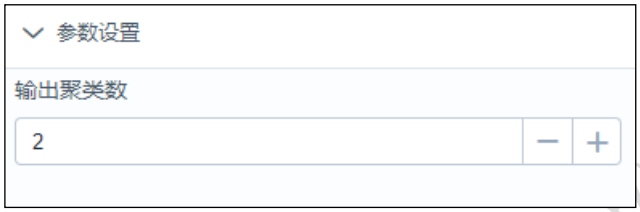

图 371

输出

表结果：包含聚类结果的数据表。

报告：无。

示例

下面对某数据进行层次聚类。

- 选择待聚类的序列，数据必须为数值型。如图 372 所示。
- 运行该组件，对组件右击，选择查看数据，结果如图 373 所示。

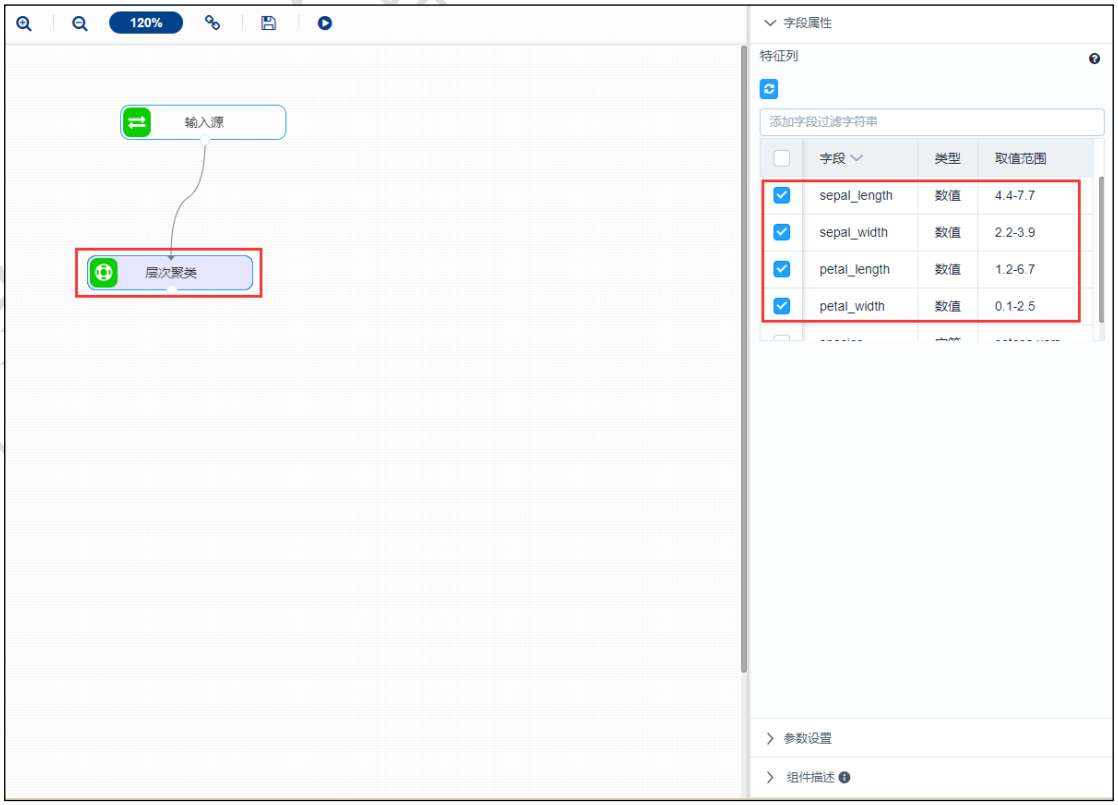

图 372

sepal_length	sepal_width	petal_length	petal_width	cluster_id
5.1	3.5	1.4	0.2	2
4.9	3	1.4	0.2	2
4.7	3.2	1.3	0.2	2
4.6	3.1	1.5	0.2	2
5	3.6	1.4	0.2	2
5.4	3.9	1.7	0.4	2
4.6	3.4	1.4	0.3	2
5	3.4	1.5	0.2	2

图 373