

数据挖掘-拆分

- 概述
- 输入/输出
- 参数设置
- 示例

概述

拆分是将原始样本集按照训练集和测试集的方式拆分为两个子集。拆分后各个子集的比例总和小于等于100%。

数据拆分经常作为回归或者分类算法节点的前置节点。



输入/输出

输入	一个输入端口，用于接收数据集。
输出	两个输出端口，用于输出不同的拆分结果。

参数设置

设置拆分的参数：

>> 参数

属性帮助

数据集占比 *必填 ?

0.7

随机种子 *必填 ?

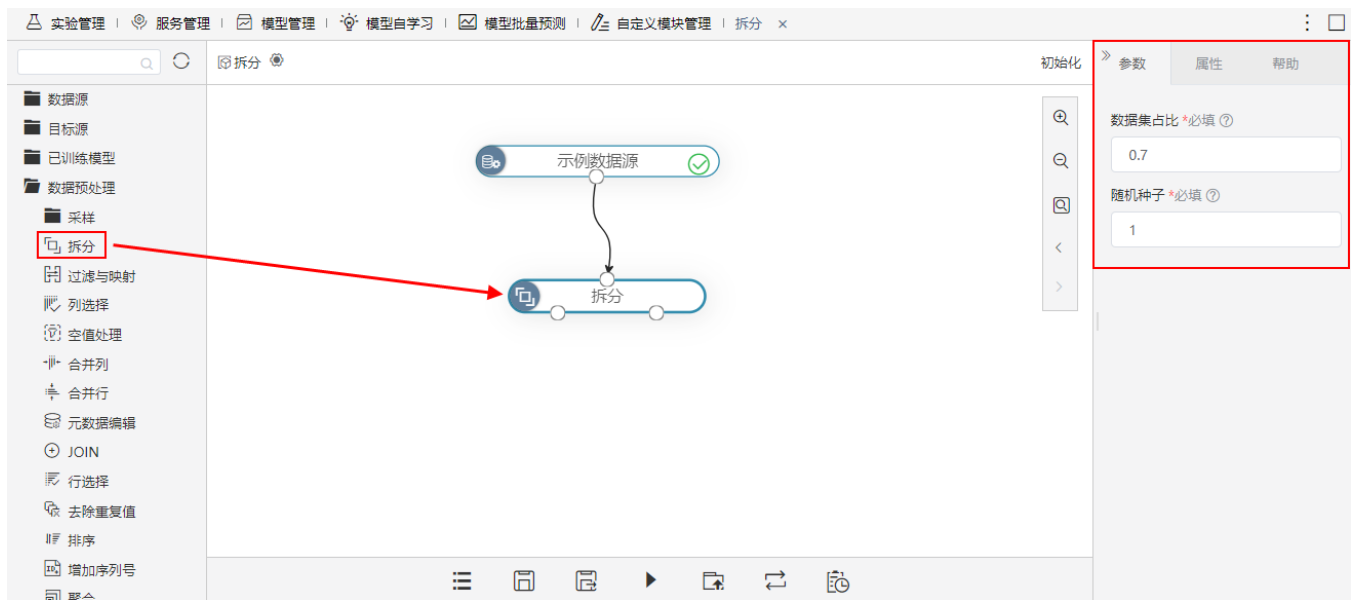
1

设置说明如下：

参数	说明
数据集占比	表示用于算法模型训练的数据集占总体数据的比例，范围是[0, 1]的数值，默认是0.7。
随机种子	作为随机序列的第一个数字，默认值为1，取值为整数。 设定随机种子，可以生成规律的随机数。

示例

1、原先示例数据源的输出结果有150条数据，对其进行拆分，设置数据集占比为0.7，即用于算法模型训练的数据集占总体数据的0.7，基于算法模型进行测试的数据集占总体数据的0.3。设置随机种子为1。



2、拆分后的结果：

示例数据源的输出结果有150条数据

id	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa

拆分

进行拆分后的输出结果1有107条数据

id	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa

进行拆分后的输出结果2有43条数据

id	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Species
5	5.0	3.6	1.4	0.2	setosa
11	5.4	3.7	1.5	0.2	setosa
17	5.4	3.9	1.3	0.4	setosa
19	5.7	3.8	1.7	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
25	4.8	3.4	1.9	0.2	setosa
29	5.2	3.4	1.4	0.2	setosa
33	5.2	4.1	1.5	0.1	setosa
37	5.5	3.5	1.3	0.2	setosa
39	4.4	3.0	1.3	0.2	setosa
40	5.1	3.4	1.5	0.2	setosa

拆分后有两个输出结果，左侧输出端口为训练集即输出结果1，有107条数据；右侧输出端口为测试集即输出结果2，有43条数据。