数据挖掘-拆分

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

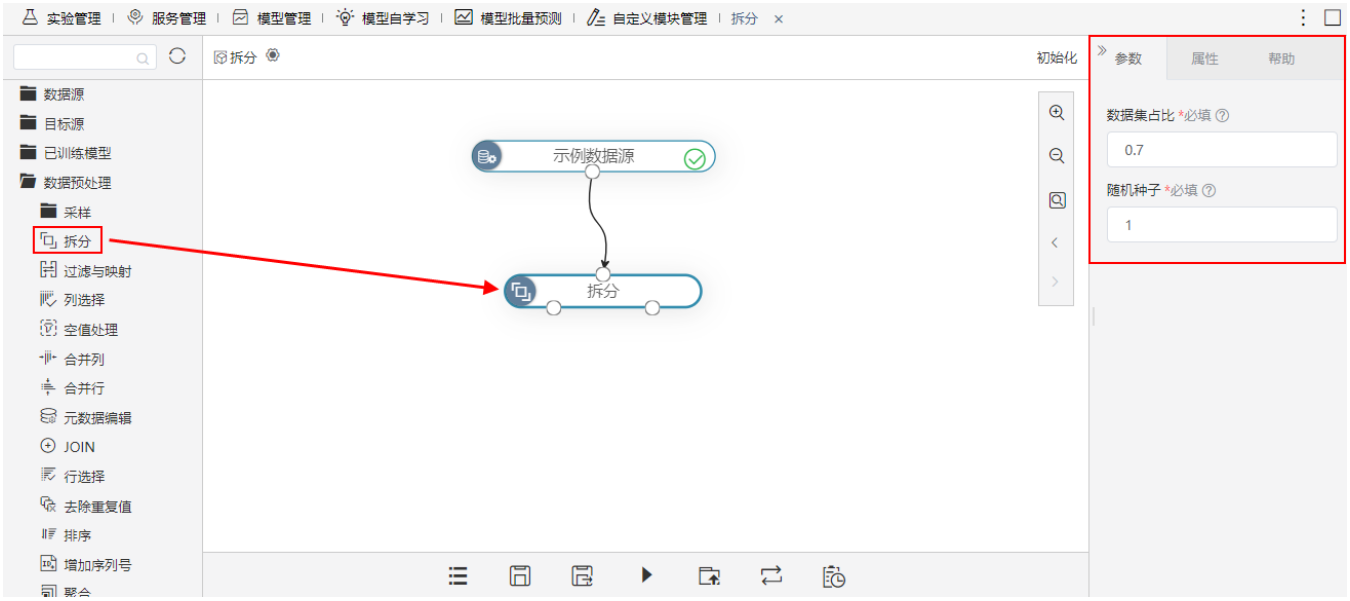
设置拆分的参数：

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：

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。