数据挖掘-采样

采样就是按照某种规则从数据集中挑选样本数据。

在Smartbi中支持四种类型的采样：随机采样、加权采样、分层采样、下采样。

概述

类型	说明
	按照随机的原则，从总体数据中随机地抽取部分数据，保证总体样本中每一个样本都有已知的、非零的概率被选入为研究对象，以保证样本的代表性。
	以加权方式生成采样数据。先从总体中，选择用于加权的列，然后按照该列值的大小进行采样，权重值越大，抽取的概率越大。
	数据集分层抽取一定比例或者一定数据的随机样本：先从总体中选择用于分层的列，将总体分成不同的部分，再对每部分进行随机采样。
	分类时，由于训练集中各类别样本数量不均衡，导致模型在测试集上的泛化性不好；下采样通过移除数据量较多类别的部分数据，使样本达到均衡。

输入/输出

输入	只有一个输入端口，用于接收数据集。
输出	只有一个输出端口，用于输出采样结果。

参数配置

随机采样、加权采样、分层采样、下采样的参数设置说明如下：

类型	参数	说明
随机采样	抽样比例	表示样本占总体的比例，范围是[0,1]的数，默认值为0.5。
	抽样种子	作为随机序列的第一个数字，默认值为10。设定抽样种子，可以使随机结果固定，即运行结果在多次运行中保持不变。
加权采样	权重列	选择用于加权的列。 ⚠ 权重列必须为数字类型的字段。
	采样方式值	• 按个数：表示按照需要采样样本的数量值进行抽取，范围为大于等于1的整数。 • 按比例：表示按照采样样本占总体的比例进行抽取，范围为0-1之间的数值。
	随机种子	作为随机序列的第一个数字。设定抽样种子，可以使随机结果固定，即运行结果在多次运行中保持不变。范围是任意整数。
分层采样	分层列	选择用于分层的列（能使数据有较大差异的列）。
	采样方式值	• 按个数：表示按照需要采样样本的数量值进行抽取，范围为大于等于1的整数。 • 按比例：表示按照采样样本占总体的比例进行抽取，范围为0-1之间的数值。
	随机种子	作为随机序列的第一个数字。设定抽样种子，可以使随机结果固定，即运行结果在多次运行中保持不变。
下采样	采样目标列	选择需要采样的列。

	设置各类别的采样方式	类别值	输入需要采样目标列的类别值。 ⚠ 可通过聚合节点（已选字段选择“目标列”，操作选择“Group”）查看目标列具体有哪些类别值，详情请查看 数据挖掘-聚合。
		采样方式值	• 按个数：表示按照需要采样样本的数量值进行抽取，范围为大于等于0的整数。 • 按比例：表示按照采样样本占总体的比例进行抽取，范围为0-1之间的数值。
		采样值/采样比例	• 采样值：采样方式选择按个数，需要输入采样的值。 • 采样比例：采样方式选择按比例，需要输入采样的比例。
		添加	添加一条分类，可满足对多个类别值的采样。
		编辑	修改采样方式、采样值/采样比例。
		删除	删除此条分类。
	随机种子	作为随机序列的第一个数字。设定抽样种子，可以使随机结果固定，即运行结果在多次运行中保持不变。范围是任意整数。	