**Apriori节点使用说明**

作者：张青松

目录

[1. 建立工作流 2](#_Toc423025692)

[1.1. 数据处理 2](#_Toc423025693)

[1.1.1. 内容组成项集 2](#_Toc423025694)

[1.1.2. 表头组成项集 3](#_Toc423025695)

[2. 设置Apriori模型参数并执行 6](#_Toc423025696)

[3. Apriori执行结果解释 7](#_Toc423025697)

# 建立工作流

将Apriori节点直接与数据源连接，输入数据。



图1 Apriori工作流

## 数据处理

数据处理的目的是，将每一个项集装换成BitVector或者CollectionColumn类型。根据实际情况数据可以分为两大类：数据表的表头组成项集、数据表内容组成项集。下面以关联规则常用的购物车为例，描述数据的处理过程。打开Apriori的设置界面。

### 内容组成项集



表1 CollectionColumn类型输入数据

如上表所示，每一行表示一个购物车中商品名称。这种类型的输入数据，将生成一个CollectionColumn列用来计算频繁项集。

针对这种类型输入数据，在设置面板的设置页签中的项集生成方式的下拉菜单中，选择“内容”，并根据需要选择是否忽略空值。在下方的表中选择用来生成CollectionColumn列的数据列。



图2 Apriori设置界面

### 表头组成项集



表2 bitvector类型输入数据

如上表，表头为所有商品。每一行表示个购物车，表中的元素用于标示对应的购物车中是否购买了该商品。这里，用1表示购买，0表示未购买。这种输入数据将生成一个BitVector类型的列来计算频繁项集。

针对表头为项集中的项的输入数据，选择项集生成方式为“列头”。并根据表中的标示信息进一步设置。



图3 Apriori对于bitvector输入数据的设置

针对表头为项集组成部分的数据，转换的目的是将标示信息转换成只含有0和1的bit表示。根据输入数据中表示信息的类型分为以下两种处理。

1. 标示信息为字符串

选择字符串列。为了将字符串转化成bit，选择正则表达式或者通配符匹配，并设置匹配中的设置为1，反之为0；或者匹配中的为0，反之为1。以及大小写是否敏感等等。

在下方的表中只显示类型为字符串的列，选择用来生成BitVector的列。



图4 字符串类型生成bitvector设置

1. 标示信息为数字类型

选择数字列，并设置阈值。阈值的作用是如果标示信息的数字不小于设定的阈值，用1表示，反之则为0.

阈值的设置分为使用固定值和该列的均值的百分比。



图5 数值类型生成bitvector设置

# 设置Apriori模型参数并执行

根据实际情况设置Apriori的最小项集数量，最小支持度，最小置信度和前项是否排序等信息。



图5 Apriori模型参数设置

保存并执行Apriori节点

# Apriori执行结果解释

根据输入数据设置适当的参数并执行，得到的结果如表，



表3 Apriori执行结果举例

表中的表头依次表示：支持后项、支持前项、置信度、支持度和提升度。每一行表示一个规则。