数据挖掘-异常值处理

概述

异常值检测和处理是数据挖掘中常用的数据处理方法，添加异常值检测节点，满足数据处理、欺诈行为检测等应用场景。

用户可以针对异常值选择相应的填充策略进行异常值的替换。

输入/输出

输入	一个输入端口，用于接收数据集。
输出	一个输出端口，用于输出异常值处理的结果。

参数设置

设置异常值处理的参数：

设置说明如下：

参数	说明
选择字段	用于选择进行异常值处理设置的字段： 
检测方法	• 四分位距：将数据按数值从小到大分成四等分，分隔点为Q1、Q2、Q3，四分位距则为上四分位值Q3与下四分位值Q1两者之差。 • 标准差法：假定数据是服从正态分布的，计算数据的标准差，对偏离标准差的数据进行处理如用均值、上下界数值、指定值替换。 • 自定义检测：可以自定义上下界，对异常值进行处理。

参数设置	• 四分位距：四分位距中下界的计算公式为Q1 - p * (Q3 - Q1)；上界的计算公式为Q3 + p * (Q3 - Q1)；公式中的p就是参数设置中的值，用户可以根据需求调整公式中的p，p需为非负数。 • 标准差法：标准差中下界的计算公式： 均值 - 系数 * 标准差；上界的计算公式： 均值 + 系数 * 标准差；公式中的系数就是参数设置中的值，同样的，系数的取值也需要为非负数。 • 自定义检测：自定义检测的参数为使用英文逗号分隔的上下界。如参数设置为“1,2”，则代表下界为1，上界为2。注意：上下界的值需要为数值型
处理策略	• 均值：检测出数据中的异常值后，用均值去替代异常值。 • 指定值：检测出数据中的异常值后，用指定值去替代异常值。当处理策略为指定值时，需要输入数值型数据 • 上下界：检测出数据中的异常值后，当异常值超出上界，用上界替换异常值；超出下界，用下界替换异常值。下界必须不大于上界，且均为数值型数据。 • 异常值处理：检测出数据中的异常值后，直接删除异常值所在的行。
自定义填充值	只有当处理策略选择“指定值”时，自定义填充值才允许编辑，用户可以自定义填充的指定值。
检测方法批量处理	将右边选中的行的检测方法全部处理。 
处理策略批量处理	将右边选中的行的处理策略全部处理。