

# 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。

- 概述
- 输入/输出
- 参数设置



## 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

选择字段

说明

用于选择进行异常值处理设置的字段：

⊗异常值处理配置

输入关键字搜索

☐

可选字段

☐ ^入会时间

☐ ^第一次飞行日期

☐ ^性别

☐ ^工作地城市

☐ ^工作地所在省份

☐ ^工作地所在国家

☐ ^观测窗口结束时间

☐ ^末次飞行日期

☐ #会员卡号

☐ #观测窗口季度平均飞行次数

到右边>

<到左边

输入关键字搜索

☐

可选字段

☐ #观测窗口内的飞行次数

检测方法批量处理

请选择

处理策略批量处理

请选择

检测方法	参数设置	处理策略	自定义值充值
四分位距	1.5	上下界	

注意：检测方法为四分位距、标准差时，参数设置：非负数；检测方法为自定义时，参数设置为：下界、上界，上、下界均为数值型且下界小于上界，用英文逗号分隔，例如：1,100

确定

取消

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	<ul style="list-style-type: none"><li>四分位距：四分位距中下界的计算公式为<math>Q1 - p * (Q3 - Q1)</math>；上界的计算公式为<math>Q3 + p * (Q3 - Q1)</math>；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。</li><li>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。</li><li>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</li></ul>																
处理策略	<ul style="list-style-type: none"><li>均值：检测出数据中的异常值后，用均值去替代异常值。</li><li>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</li><li>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。</li><li>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。</li></ul>																
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	<p>将右边选中的行的检测方法全部处理。</p> <div><div><div>⊗异常值处理配置</div><div><div>输入关键字搜索</div><div><div><input type="checkbox"/> 可选字段</div><div><input checked="" type="checkbox"/> #组织机构代码</div><div><input checked="" type="checkbox"/> #企业名称</div><div><input type="checkbox"/> #隶属关系</div><div><input type="checkbox"/> #应收账款</div><div><input type="checkbox"/> #产成品</div><div><input type="checkbox"/> #固定资产合计</div><div><input type="checkbox"/> #固定资产原价</div><div><input type="checkbox"/> #累计折旧</div><div><input type="checkbox"/> #本年折旧</div><div><input type="checkbox"/> #资产总计</div><div><input type="checkbox"/> #流动负债合计</div></div><div>到右边&gt;</div><div>&lt;到左边</div></div></div><div><div>输入关键字搜索</div><div>检测方法批量处理</div><div>标准差法</div><div>处理策略批量处理</div><div>请选择</div><table><tr><th>检测方法</th><th>参数设置</th><th>处理策略</th><th>自定义值</th></tr><tr><td><input checked="" type="checkbox"/> #企业控股情况</td><td>标准差法</td><td>3</td><td>上下界</td></tr><tr><td><input checked="" type="checkbox"/> #流动资产合计</td><td>标准差法</td><td>3</td><td>上下界</td></tr><tr><td><input type="checkbox"/> #存货</td><td>四分位距</td><td>1.5</td><td>上下界</td></tr></table><div>检测方法批量处理选择“标准差法”，则下面右边选中的行的检测方法都为“标准差法”</div></div></div>	检测方法	参数设置	处理策略	自定义值	<input checked="" type="checkbox"/> #企业控股情况	标准差法	3	上下界	<input checked="" type="checkbox"/> #流动资产合计	标准差法	3	上下界	<input type="checkbox"/> #存货	四分位距	1.5	上下界
检测方法	参数设置	处理策略	自定义值														
<input checked="" type="checkbox"/> #企业控股情况	标准差法	3	上下界														
<input checked="" type="checkbox"/> #流动资产合计	标准差法	3	上下界														
<input type="checkbox"/> #存货	四分位距	1.5	上下界														
处理策略批量处理	<p>将右边选中的行的处理策略全部处理。</p> <div><div><div>⊗异常值处理配置</div><div><div>输入关键字搜索</div><div>检测方法批量处理</div><div>请选择</div><div>处理策略批量处理</div><div>上下界</div><div>自定义值</div></div><div><div>输入关键字搜索</div><div>检测方法批量处理</div><div>四分位距</div></div></div></div>	1.5	上下界	上下界													
<input checked="" type="checkbox"/> #流动资产合计	四分位距	1.5	上下界	上下界													
<input type="checkbox"/> #存货	四分位距	1.5	上下界	上下界													