# 数据挖掘-OneHot编码



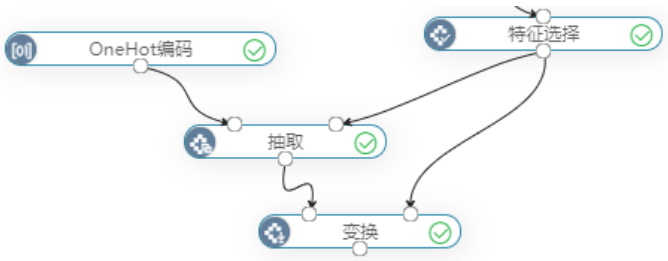
## 概述

OneHot编码是将类别变量转换为机器学习算法易于利用的一种形式的过程。它是分类变量作为二进制向量的表示。这首先要求将分类值映射到整数值。然后，每个整数值被表示为二进制向量，除了整数的索引之外，其它都是零值，它被标记为1。（即：标记位置为1，其他位置为0）我们编码后的结果是一个稀疏向量，稀疏向量就是有特征数量，特征索引和特征值组成。


## 输入/输出

| | |
|---|---|
| **输入** | 没有输入端口。 |
| **输出** | 一个输出端口，用于接入下一个节点，与抽取节点组合使用。 |

## 参数设置

| 参数名称 | 说明 | 备注 |
|---|---|---|
| 新增列后缀 | 用于设置在原字段名后追加后缀生成新的列，默认后缀为：OneHot； | 该节点与抽取、变换节点组合使用：  |
| 非法数据处理策略 | 选择对非法数据进行处理的策略，非法数据指空值或未进模型的类别值。<br>• 自动过滤：在转换时，对在抽取时未出现的类别或空值进行删除；<br>• 特殊编码：在转换时，对在抽取时未出现的类别以一种特殊编码进行转换；<br>• 非法检测：在转换时，对在抽取时未出现的类别进行报错提示。 | |

## 示例

使用“银行零售客户流失数据”，选取性别列转换为OneHot编码，输出结果为数组组形式展示，(2,[1],[1.0])表示为总共有2个类别，索引为1的位置标记为1，其他位置都是0。


当前显示 100 条 / 总共有 100000 条数据

| AUM | # 卡等级 | # 是否个贷 | # 1年内购买理财 | # 下载手机银行 | # 是否领取App权益 | # 是否登录App | # 是否持有信用卡 | # 是否关联还款 | 性别_OneHot |
|---|---|---|---|---|---|---|---|---|---|
| 7876 | 4 | 0 | 0 | 1 | 1 | 32 | 0 | 0 | (2,[1],[1.0]) |
| 80 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | (2,[0],[1.0]) |
| 39 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | (2,[1],[1.0]) |
| 143 | 1 | 0 | 0 | 1 | 1 | 18 | 1 | 0 | (2,[0],[1.0]) |
| 01 | 1 | 1 | 0 | 1 | 1 | 33 | 0 | 0 | (2,[0],[1.0]) |
| 14 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | (2,[1],[1.0]) |
| 59 | 1 | 0 | 0 | 1 | 0 | 44 | 0 | 0 | (2,[1],[1.0]) |
| 52 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | (2,[1],[1.0]) |
| 28 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | (2,[1],[1.0]) |
| 95 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | (2,[1],[1.0]) |
| 706 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | (2,[1],[1.0]) |

表头真名 表头别名 提示：点击单元格可查看超出的内容。注意：表头中 表示特征列，* 表示标签列 下载预览数据