

# 数据挖掘-自动特征组合

## 概述

自由特征组合是将现有特征按照一定方式进行组合，形成新的特征，为后续的挖掘工作提供基础。它是通过将单独的特征进行组合（相乘）而形成的合成特征，特征组合有助于表示非线性关系。

例如：在电商推荐场景中，预测一个用户是否购买一个东西的场景下，性别（男女）是一个特征，时间维度是另一个特征（节假日，工作日，周末等待），特征组合就是假设性别特征是女，时间维度双11，会发现这两个特征组合起来一定是非常强对用户预测是否买东西的组合特征。女性在双十一购买商品的概率非常高，这就是交叉后的非线性的组合特征，非线性的组合特征对数据挖掘非常关键。

## 输入/输出

|    |                         |
|----|-------------------------|
| 输入 | 一个输入端口，用于接收前置节点传下来的数据集。 |
| 输出 | 一个输出端口，用于输出特征组合过后的数据集。  |

## 参数设置

| 参数名称   | 说明                                                                                                                                                                                       | 备注                                                  |
|--------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------|
| 选择列    | 选择需要进行组合的特征列，必须是数值列。                                                                                                                                                                     | 必填                                                  |
| 选择标签列  | 用于选择作为标签列的字段                                                                                                                                                                             | 类别数需小于100                                           |
| 选择方式   | 用于确定自由组合的方式： <ul style="list-style-type: none"><li>全局组合：将所有所选特征按照派生度进行组合，该方法所组合出的特征数最得多，但性能消耗也随着派生度的增大随之大幅增长。</li><li>搜索组合：按照所选特征进行搜索找出局部最优的组合特征，其产生特征数较少，但性能相对较快、同时给出可解释性也较好。</li></ul> | 必填组合方式                                              |
| 派生度    | 派生度越多，所产生的特征越多，全局组合成指数增长，搜索组合成倍数增长                                                                                                                                                       |                                                     |
| 输出的特征数 | 用于设置从自由组合的特征列集合中需要输出特征列的数量。                                                                                                                                                              | 必填，从组合出的特征中选择重要性最高的特征数，当所填值超过产生的特征数时，输出特殊数为全部组合特征数。 |

## 示例

使用“鸢尾花数据”，选择4个特征列和1个标签列，选择搜索组合方式，设置派生度为2，输出特征数为5。输出的结果特征数为5个特征，如下图：

⌂ 当前显示 100 条 / 总共有 150 条数据 提示:点击单元格可查看超出的内容

| c-1863743831         | c-1706677299        | c2035154531         | c-1128853625         | c-1086510588       | id | Sepal_Length |
|----------------------|---------------------|---------------------|----------------------|--------------------|----|--------------|
| 0.015323478335036694 | 0.1485152557275677  | 0.10726434834525683 | 0.01776841022191236  | 0.3907486975434356 | 1  | 5.1          |
| 0.015323478335036694 | 0.1485152557275677  | 0.10726434834525683 | 0.01776841022191236  | 0.375425219208399  | 2  | 4.9          |
| 0.015323478335036694 | 0.1379070231755986  | 0.0996026091777385  | 0.01776841022191236  | 0.3601017408733623 | 3  | 4.7          |
| 0.015323478335036694 | 0.15912348827953687 | 0.11492608751277521 | 0.01776841022191236  | 0.3524400017058439 | 4  | 4.6          |
| 0.015323478335036694 | 0.1485152557275677  | 0.10726434834525683 | 0.01776841022191236  | 0.3830869583759173 | 5  | 5.0          |
| 0.03064695667007339  | 0.1803399533834751  | 0.1302495658478119  | 0.03553682044382472  | 0.4137339150459907 | 6  | 5.4          |
| 0.022985217502555035 | 0.1485152557275677  | 0.10726434834525683 | 0.026652615332868532 | 0.3524400017058439 | 7  | 4.6          |
| 0.015323478335036694 | 0.15912348827953687 | 0.11492608751277521 | 0.01776841022191236  | 0.3830869583759173 | 8  | 5.0          |
| 0.015323478335036694 | 0.1485152557275677  | 0.10726434834525683 | 0.01776841022191236  | 0.3371165233708073 | 9  | 4.4          |
| 0.007661739167518347 | 0.15912348827953687 | 0.11492608751277521 | 0.00888420511095618  | 0.375425219208399  | 10 | 4.9          |
| 0.015323478335036694 | 0.15912348827953687 | 0.11492608751277521 | 0.01776841022191236  | 0.4137339150459907 | 11 | 5.4          |

点击鼠标右键查看分析效果：

| # | 特征名称         | 重要程度                | 组合特征                                                                             |
|---|--------------|---------------------|----------------------------------------------------------------------------------|
| 1 | c-1863743831 | 0.3267472175571302  | $\text{Length}^1 \cdot \text{Petal}^4 \cdot \text{Width}^3$                      |
| 2 | c-1706677299 | 0.2714870522635013  | $\text{Length}^3 \cdot \text{Petal}^4 \cdot \text{Width}^1$                      |
| 3 | c2035154531  | 0.21800694984358607 | $\text{Length}^2 \cdot \text{Petal}^4 \cdot \text{Width}^2$                      |
| 4 | c-1128853625 | 0.1247132825660405  | $\text{Petal}^4 \cdot \text{Width}^4$                                            |
| 5 | c-1086510588 | 0.03760271767722775 | $\text{Length}^2 \cdot \text{Sepal}^1 \cdot \text{Petal}^3 \cdot \text{Width}^2$ |