# 数据挖掘-决策树



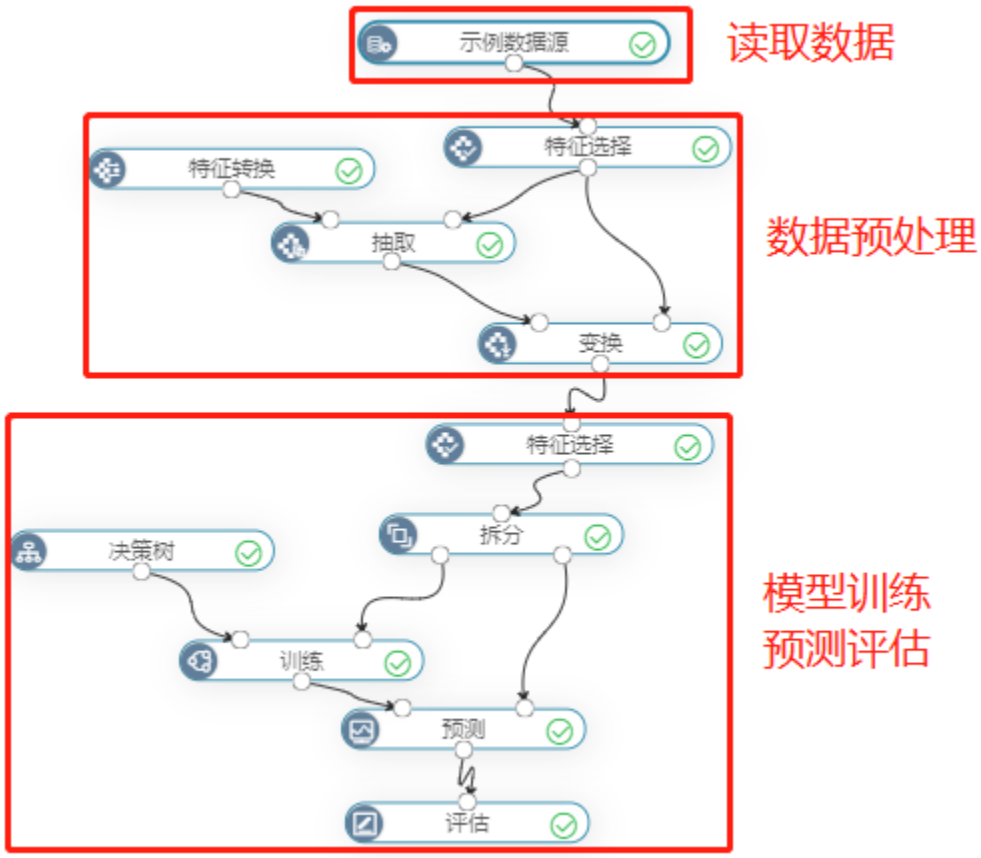
## 概述

决策树是一种常用的分类算法，它是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。根节点到每个叶子节点均形成一条分类的路径规则。而对新的样本进行测试时，只需要从根节点开始，在每个分支节点进行测试，沿着相应的分支递归地进入子树再测试，一直到达叶子节点，该叶子节点所代表的类别即是当前测试样本的预测类别。

优势：可直接查看决策树分析的可视化效果，如下图：


## 示例

使用“糖尿病预测”案例数据，预测是否有糖尿病。




特征转换是为了将各变量中的类别型变量变换成数值型变量，类别型无法进入模型，转换后方便算法模型学习；

## 参数设置

决策树的参数如下：

| 参数名称 | 值 | 说明 |
| --- | --- | --- |
| 自动调参设置 | 系统默认的各项参数值范围。 | 必须结合“启用自动调参”功能使用。系统将对设置指定或范围内的参数值循环调参，匹配出最优的组合。详情请参考 自动调参设置 。 |
| 启用自动调参 | 勾选该项，则系统自动调参数，不需要用户手工设置参数。 | |
| 分裂特征的数量 | 取值范围：>=2的整数； 默认值：32。 | 对连续类型特征进行离散时的分箱数；<br>该值越大，模型会计算更多连续型特征分裂点且会找到更好的分裂点，但同时也会增加模型的计算量； |
| 树的深度 | 取值范围：[1,30]的整数；默认值为4。 | 当模型达到该深度时停止分裂；<br>树的深度越大，模型训练的准确度更高，但同时也会增加模型的计算量且会导致过拟合； |
| 衡量准则 | gini | 裂分标准，Entropy表示熵值，Gini表示基尼指数； |
| | entropy | |
| 子节点最少样本数 | 取值范围：大于0且小于等于1000的整数，默认值：空 | 每次分裂后每个子节点必须拥有的样本数；<br>该值越大，决策树允许分裂的次数就越少。可以防止模型过拟合； |
| 最小分列信息增益 | 取值范围：[0, 10000]，默认值：空 | 每次分裂必须达到的信息增益；<br>该值越大，决策树允许分裂的次数就越少。可以防止模型过拟合； |

### 自动调参设置

系统将对设置指定或范围内的参数值循环调参，匹配出最优的组合。


自动调参的方式分为两种：

- 指定值调参：指定一个固定的值进行自动调参。
- 范围调参：在指定的范围内进行自动调参。

设置项说明如下：

| 设置项 | | 说明 |
| --- | --- | --- |
| 拆分比例 | | 将选择的数据拆分为两部分，一部分部分用于模型的评估，另一部分数据用于训练模型。 |
| 评估标准 | | 用于选择数据的评估指标，包括：f1、precision、recall、accuracy、AUC(二分类)。<br><br>其中，评估标准“AUC(二分类)”仅对二分类问题生效。 |
| 参数 | | 自动调参的参数项。<br><br>• 计算信息增益的方式：支持信息增益和基尼指数；<br>• 树的深度：控制生成决策树的深度，防止数据过拟合；<br>• 分裂特征的数量：连续数据离散分箱的数量；<br>• 最小分裂信息增益；<br>• 子节点最小样本数。<br><br>注意：<br><br>• 计算信息增益的范围支持多选。<br>• 其他参数默认的范围提供了一个推荐值，并不是算法限制的最大值和最小值。<br><br>Spark MLlib对决策树提供了二元以及多label的分类以及回归的支持，支持连续型和离散型的特征变量。为了防止过拟合，需要考虑剪枝。当以下任一情况发生，MLlib的决策树节点就终止划分，形成叶子节点：<br><br>1、树高度达到maxDepth；<br><br>2、minInfoGain，当前节点的所有属性分割带来的信息增益都比这个值要小；<br><br>3、minInstancesPerNode，需要保证节点分割出的左右子节点的最少的样本数量达到这个值。 |
| 指定值调参 | 指定值 | 指定一个固定的值进行自动调参。需要先勾选“是否使用指定值”才能配置。 |
| | 是否使用指定值 | 控制是否使用使用指定值进行调参。 |

| | | |
|---|---|---|
| 范围调参 | 范围 | 设置自动调参参数的范围。<br>若运行速度比较慢，可将参数范围调小一点。 |
| | 步数 | 进行范围调参时，在设置的范围内生成多少个参数值。<br>示例：<br>1）范围为[3,5]，步数为3时，生成的参数值：3，4，5<br>2）范围为[40,100]，步数为4时，生成的参数值：40，60，80，100 |

### 示例

设置自动调参设置如图：


在训练节点查看分析结果如图：