



**选修课课程论文**

**课程名称 机器学习**

**学 院 信息工程**

**年级班别 16级信息工程**

**学 号 3116002423**

**学生姓名 严寿麟**

**任课教师 张良**

**2018年5月**

目录

[1. 目的 3](#_Toc510637261)

[2. 分析方法、过程 3](#_Toc510637262)

[2.1. 流程图与步骤 3](#_Toc510637263)

[3. 结论 3](#_Toc510637264)

[4. 参考文献 3](#_Toc510637265)

## 1. 目的

将之前的20份的案例作为数据集，通过决策树的方法进行二分类，确定是否对患者进行肝癌手术治疗.

## 分析方法、过程

由于数据集数据量过小，每个样本只有十个特征，不宜通过神经网络和支持向量机对数据集，进行分类，而选择通过决策树对数据集进行分类。

### 流程图与步骤

1.分类树两个关键点

（1）将训练样本进行递归地划分自变量空间进行建树。

（2）用验证数据进行剪枝。

**CART分类与回归树**

最佳切分原则对分组进行切分

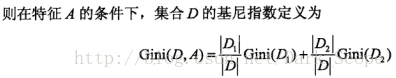
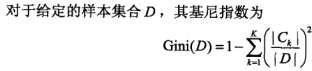
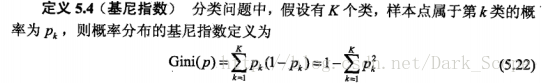
对样本进行两类的分组

满足终止条件

终止二叉树

步骤：

1. 分别取X变量各值的不同组合，将其分到树的左枝或右枝，并对不同组合而产生的树，进行评判，找出最佳组合。如果只有两个取值，好办，直接根据这两个值就可以划分树。取值多于两个的情况就复杂一些了，如变量年纪，其值有“少年”、“中年”、“老年”，则分别生产{少年，中年}和{老年}，{上年、老年}和{中年}，{中年，老年}和{少年}，这三种组合，最后评判对目标区分最佳的组合。因为CART二分的特性，当训练数据具有两个以上的类别，CART需考虑将目标类别合并成两个超类别，这个过程称为双化。这里可以说一个公式,n个属性，可以分出(2^n-2)/2种情况。
2. 二叉树的节点的衡量标准。



通过以上不纯度指标，分别计算每个变量的各种切分/组合情况，找出该变量的最佳值组合/切分点；再比较各个变量的最佳值组合/切分点，最终找出最佳变量和该变量的最佳值组合/切分点。

（3）终止条件：

（1）节点是纯结点，即所有的记录的目标变量值相同。

（2）树的深度达到了预先指定的最大值。

（3）混杂度的最大下降值小于一个预先指定的值。

（4）节点的记录量小于预先指定的最小节点记录量。

（5）一个节点中的所有记录其预测变量值相同。

直观的情况，当节点包含的数据记录都属于同一个类别时就可以终止分裂了。这只是一个特例，更一般的情况我们计算χ2值来判断分类条件和类别的相关程度，当χ2很小时说明分类条件和类别是独立的，即按照该分类条件进行分类是没有道理的，此时节点停止分裂。注意这里的“分类条件”是指按照GINI\_Gain最小原则得到的“分类条件”。

终止条件（3）混杂度的最大下降值小于一个预先指定的值，该枝的分化即停止。所有枝节的分化都停止后，树形模型即成。其实你也可以不使用这个终止条件，让树生长到最大，因为CART有剪枝算法。

## 3. 结论

使用Opencv中的Dtree算法，对16个数据集样本进行训练，使用4个数据集进行测试，参数进行调节可以将训练结果调节得很好，测试结果可以全部正确，但可能出现过拟合问题，不过样本数量太小，并不能说明事件的真实情况。所以还需要更大的数据集进行训练。

## 4. 参考文献