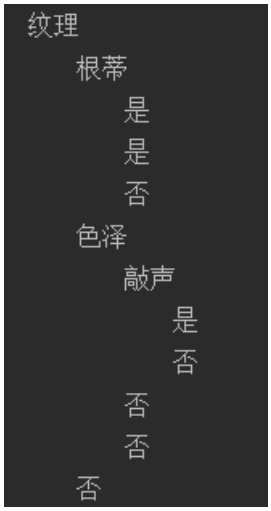
实验六 决策树

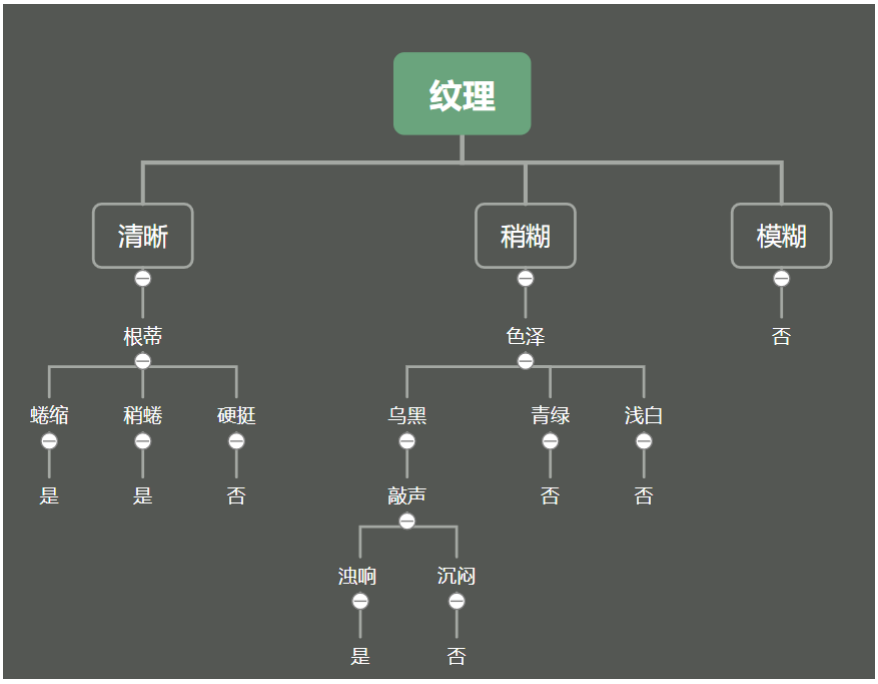
杨航 1811451 计算机科学与技术

基本要求

（1）基于 Watermelon-train1数据集（只有离散属性），构造ID3决策树；

- ID3是基于信息熵增益大小来决定分裂节点的算法，熵增最大的属性，作为此次分叉的结点。
- 伪代码思路：
    1. 输入此次需要分类的训练样本X。
    2. 查看是否满足递归终止条件（如所有样本都属于同一类别）,如果是：return退出，否进入三。
    3. 计算样本原始H（Y）
    4. 计算如果使用各个属性xi进行分裂时的熵H（Y|xi），值最小的就是熵增最大的。
    5. 通过4中找到的xi分裂训练样本X，把分裂的样本分别当做训练样本开始下一轮递归。
- 最后决策树的结构输出如下：
- 

- 好像不太直观，我手画一下：
-



（2）基于构造的 ID3 决策树，对数据集 Watermelon-test1进行预测，输出 分类精度；

测试结果：

```
[1, '浅白', '蜷缩', '浊响', '清晰', '是'] is correctly predicted
[2, '乌黑', '稍蜷', '沉闷', '清晰', '是'] is correctly predicted
[3, '乌黑', '蜷缩', '沉闷', '清晰', '是'] is correctly predicted
[5, '浅白', '蜷缩', '浊响', '清晰', '是'] is correctly predicted
[8, '青绿', '稍蜷', '浊响', '模糊', '否'] is correctly predicted
[9, '乌黑', '稍蜷', '沉闷', '稍糊', '否'] is correctly predicted
[10, '青绿', '硬挺', '清脆', '模糊', '否'] is correctly predicted
accuracy is 0.7
```

准确率70%

中级要求：

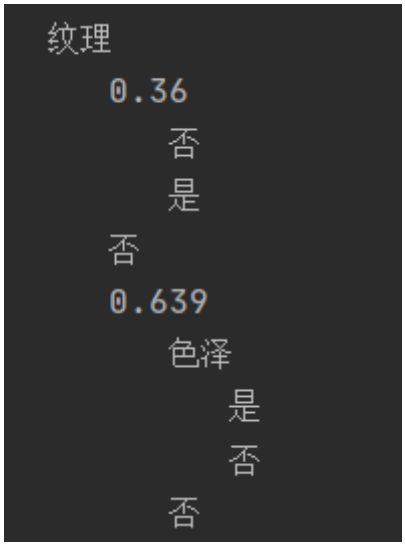
（1）对数据集Watermelon-train2，构造C4.5或者CART决策树，要求可以处 理连续型属性；

- 我构造了C4.5决策树，与ID3所不同的是使用了熵的增益率来决定。用分裂信息度量来考虑某种属性进行分裂时分支的数量信息和尺寸信息，我们把这些信息称为属性的内在信息（instrisic information）。信息增益率用信息增益 / 内在信息，会导致属性的重要性随着内在信息的增大而减小（也就是说，如果这个属性本身不确定性就很大，那我就越不倾向于选取它），这样算是对单纯用信息增益有所补偿。
- 思路基本和ID3一致，这里强调一下连续变量的处理方式：
  - 连续变量可以看作是无限个离散变量，而我们的数据量是有限的，所以可以离散化处理。
  1. 首先将样本按照连续变量排序（我是由小到大）
  2. 尝试每一个样本的值作为分裂阈值把样本分为大于这个阈值的一部分和小于等于这个样本的一部分，例如有三个样本【0.5 ，1，1.5】，当用0.5为阈值时，分成：

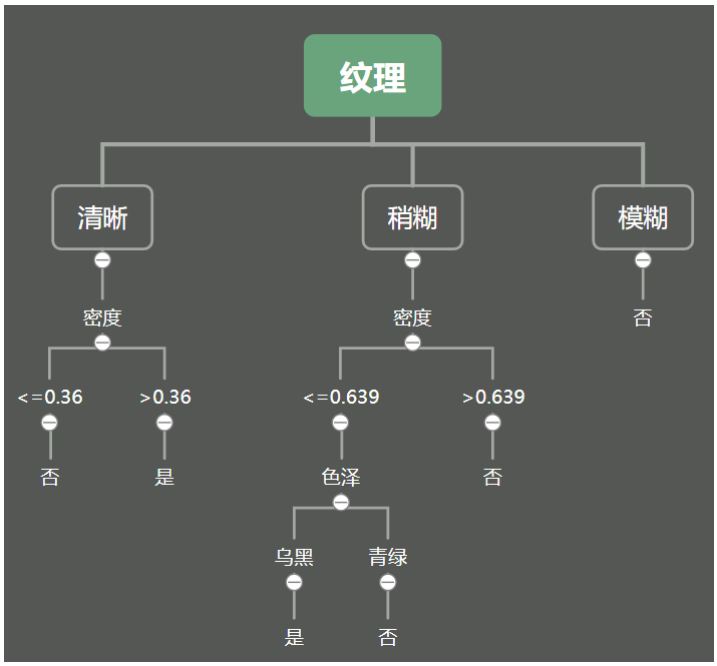
【0.5】【1 , 1.5】。当1为阈值时，分成【0.5 , 1】【1.5】。当1.5 为阈值时，分为【0.5 ,1 ,1.5】

3. 分别计算2中不同阈值条件下的信息增益率，找到最大的那个maxR
4. 用3中得到的maxR 和其他离散特征的信息增益率作比较，找到最大的增益率特征。用这个特征作为分裂节点。

- 最后决策树的结构如下：
-



-



（2）对测试集Watermelon-test2进行预测，输出分类精度；

测试结果：

```
[1, '乌黑', '稍蜷', '浊响', '清晰', 0.40299999999999997, '是'] is correctly predicted
[4, '乌黑', '稍蜷', '沉闷', '稍糊', 0.6659999999999999, '否'] is correctly predicted
[5, '青绿', '硬挺', '清脆', '清晰', 0.243, '否'] is correctly predicted
accuracy is 0.6
```

准确率：60%

高级要求：

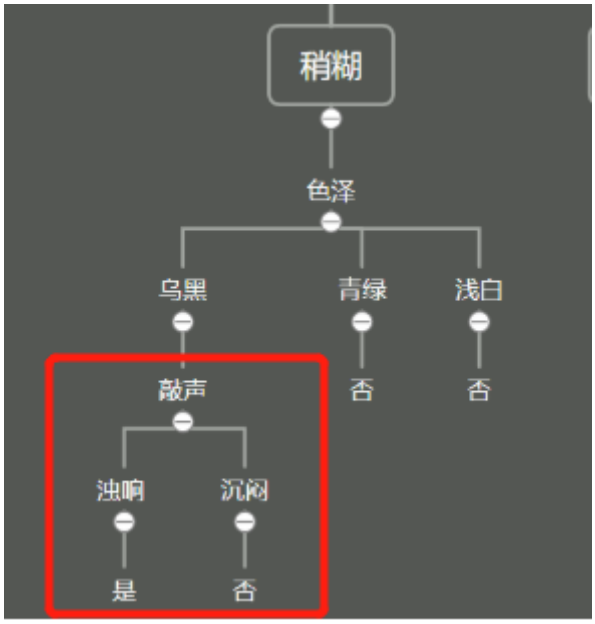
使用任意的剪枝算法对构造的决策树（基本要求和中级要求构造 的树）进行剪枝，观察测试集合的分类精度是否有提升，给出分析过程。

我使用了预剪枝的两种策略进行剪枝：

1. 通过限制节点最小样本数剪枝

上述在基础和中级要求中的每个节点的最小样本数都是1。这样很容易把孤例样本也形成一个分支，出现过拟合的问题。

- 当每个节点最小样本数设为3的时候，ID3算法的精度有上升10%
- ```
  [1, '浅白', '蜷缩', '浊响', '清晰', '是'] is correctly predicted
  [2, '乌黑', '稍蜷', '沉闷', '清晰', '是'] is correctly predicted
  [3, '乌黑', '蜷缩', '沉闷', '清晰', '是'] is correctly predicted
  [5, '浅白', '蜷缩', '浊响', '清晰', '是'] is correctly predicted
  [7, '乌黑', '稍蜷', '浊响', '稍糊', '否'] is correctly predicted
  [8, '青绿', '稍蜷', '浊响', '模糊', '否'] is correctly predicted
  [9, '乌黑', '稍蜷', '沉闷', '稍糊', '否'] is correctly predicted
  [10, '青绿', '硬挺', '清脆', '模糊', '否'] is correctly predicted
  accuracy is 0.8
  ```
- 上升的原因是，敲声这个地方的过拟合部分被消除。



- 再继续把最小样本数的条件升高到4,5,6，结果依旧没有变化，精度为80%

- 但是在c4.5算法中，这个剪枝算法没有精度提升。
  ```

2.通过树的深度限制进行剪枝

- 在树的限制的算法中，依然只是对ID3的精度有提升，当深度限制为2层的时候，和限制节点最小样本数取得了同样的效果。

```
[1, '浅白', '蜷缩', '浊响', '清晰', '是'] is correctly predicted
[2, '乌黑', '稍蜷', '沉闷', '清晰', '是'] is correctly predicted
[3, '乌黑', '蜷缩', '沉闷', '清晰', '是'] is correctly predicted
[5, '浅白', '蜷缩', '浊响', '清晰', '是'] is correctly predicted
[7, '乌黑', '稍蜷', '浊响', '稍糊', '否'] is correctly predicted
[8, '青绿', '稍蜷', '浊响', '模糊', '否'] is correctly predicted
[9, '乌黑', '稍蜷', '沉闷', '稍糊', '否'] is correctly predicted
[10, '青绿', '硬挺', '清脆', '模糊', '否'] is correctly predicted
accuracy is 0.8
```

- 但是依旧对c4.5没有提升效果，很有可能是因为测试集样本数太少了，只有5个，不是很能说明问题。