

作业二

1. 采用信息增益准则，基于表 4.2 中编号为 1、2、3、6、7、9、10、14、15、16、17 的 11 个样本的色泽、根蒂、敲声、文理属性构建决策树。（本次作业可以用笔算，鼓励编程实现，但都需要列出主要步骤，其中

$\log_2(3)=1.585, \log_2(5)=2.322, \log_2(6)=2.585, \log_2(7)=2.807, \log_2(9)=3.17, \log_2(10)=3.322, \log_2(11)=3.459$ ）（40 分）

注：此题也可用敲声做第一次划分属性，结果见后面

（10 月 22 日上课前提交纸质版）

1、采用信息增益准则，基于表 4.2 中编号为 1、2、3、6、7、9、10、14、15、16、17 的 11 个样本的色泽、根蒂、敲声、文理属性构建决策树。（本次作业可以用笔算，鼓励编程实现，但都需要列出主要步骤，其中

$\log_2(3)=1.585, \log_2(5)=2.322, \log_2(6)=2.585, \log_2(7)=2.807, \log_2(9)=3.17, \log_2(10)=3.322, \log_2(11)=3.459$ ）

为了利用已知数据构建一棵决策树，首先从根节点开始选择一个划分属性对该节点进行划分，划分属性的选取由属性的信息增益所决定，信息增益的计算公式为：

$$Gain(D, a) = Ent(D) - \sum_{v=1}^V \left(\frac{|D^v|}{|D|} Ent(D^v) \right)$$

其中 $Ent(D)$ 是指样本集合 D 中的信息熵， V 是指离散属性 a 中的 V 个不同取值。

信息熵的计算公式为

$$Ent(D) = - \sum_{k=1}^y p_k \log_2 p_k$$

其中 p_k 是指集合 D 中第 k 类样本所占比例。

根据公式算出根节点的信息熵为

$$Ent(D) = - \sum_{k=1}^y p_k \log_2 p_k = - \left(\frac{6}{11} \log_2 \frac{6}{11} + \frac{5}{11} \log_2 \frac{5}{11} \right) = 0.994$$

以西瓜的色泽为例计算它的信息增益，色泽共有三种取值：{青绿，乌黑，浅白}，使用该属性对样本 D 进行划分，可得到三个子集： D^1, D^2, D^3 ，其中 D^1 中的样本包括 {1, 6, 10, 17}，正例有 2 个， $p_1 = 2/4$ ， D^2 中的样本包括 {2, 3, 7, 9, 15}，正例有 3 个， $p_2 = 3/5$ ， D^3 中的样本包括 {14, 16}，正例有 0 个 $p_3 = 0/2$ ，可算得 3 个分支的信息熵分别是

$$Ent(D^1) = - \left(\frac{2}{4} \log_2 \frac{2}{4} + \frac{2}{4} \log_2 \frac{2}{4} \right) = 1$$

$$Ent(D^2) = - \left(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5} \right) = 0.9710$$

$$Ent(D^3) = - \left(\frac{0}{2} \log_2 \frac{0}{2} + \frac{2}{2} \log_2 \frac{2}{2} \right) = 0$$

可算得“色泽”属性的信息增益是

$$Gain(D, 色泽) = 0.994 - \left(\frac{4}{11} \times 1 + \frac{5}{11} \times 0.9710 + \frac{2}{11} \times 0 \right) = 0.1890$$

类似的其它属性的信息增益为

$$Gain(D, 根蒂) = 0.1113$$

$$Gain(D, 敲声) = 0.1981$$

$$Gain(D, 纹理) = 0.1981$$

显然属性信息增益最大值有两个，随机选取一个作为划分属性，这里选取纹理。图 1.1 给出基于脐部对根节点的划分结果，各分支节点所包含的样例子集显示在各节点中。



图 1.1 基于纹理属性对节点进行划分

然后利用决策树算法继续对每个分支节点进行划分，以图 1.1 中第一个分支节点为例，该节点包含编号为{1,2,3,6,10,15}这 6 个节点，可用属性是{色泽，根蒂，敲声}。以此节点为根节点选择一个属性进行再次划分。

根据公式算出根节点的信息熵为

$$Ent(D) = - \sum_{k=1}^y p_k \log_2 p_k = - \left(\frac{4}{6} \log_2 \frac{4}{6} + \frac{2}{6} \log_2 \frac{2}{6} \right) = 0.9183$$

基于这个新的 D 计算出各个属性的信息增益

$$Gain(D, 色泽) = 0$$

$$Gain(D, 根蒂) = 0.5850$$

$$Gain(D, 敲声) = 0.3774$$

明显根蒂这个属性的信息增益最大，这里我们选取根蒂作为划分属性。图 1.2 所示为划分结果



图 1.2 基于根蒂属性对节点进行划分

对纹理模糊的分支进行划分

根据公式算出根节点的信息熵为

$$Ent(D) = - \sum_{k=1}^y p_k \log_2 p_k = - \left(\frac{1}{4} \log_2 \frac{1}{4} + \frac{3}{4} \log_2 \frac{3}{4} \right) = 0.8113$$

$\text{Gain}(D, \text{色泽}) = 0.3113$
 $\text{Gain}(D, \text{根蒂}) = 0.1226$
 $\text{Gain}(D, \text{敲声}) = 0.8113$

这里选取敲声作为划分属性
划分结果如图 1.3 所示

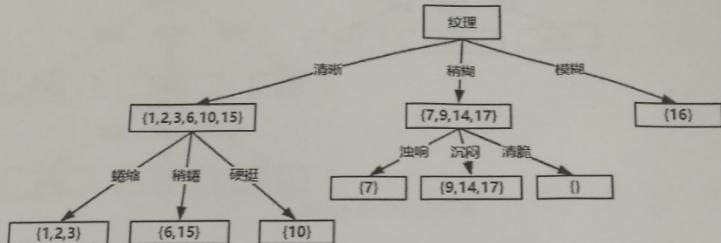


图 1.3 基于敲声属性对节点进行划分

以此类推对其他节点进行划分，划分终点是划分类别是一种类别，若划分有空节点，则选取整个样本最多的类，即将其标记为坏瓜。

最终的划分结果如图 1.4 所示

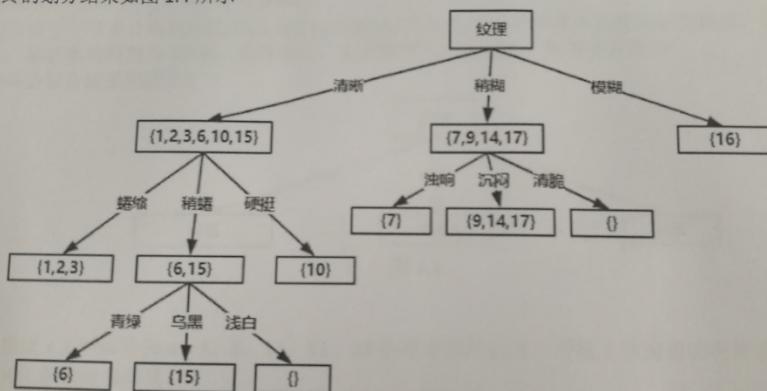


图 1.4 最终划分结果

抽象出来如图 1.5 所示

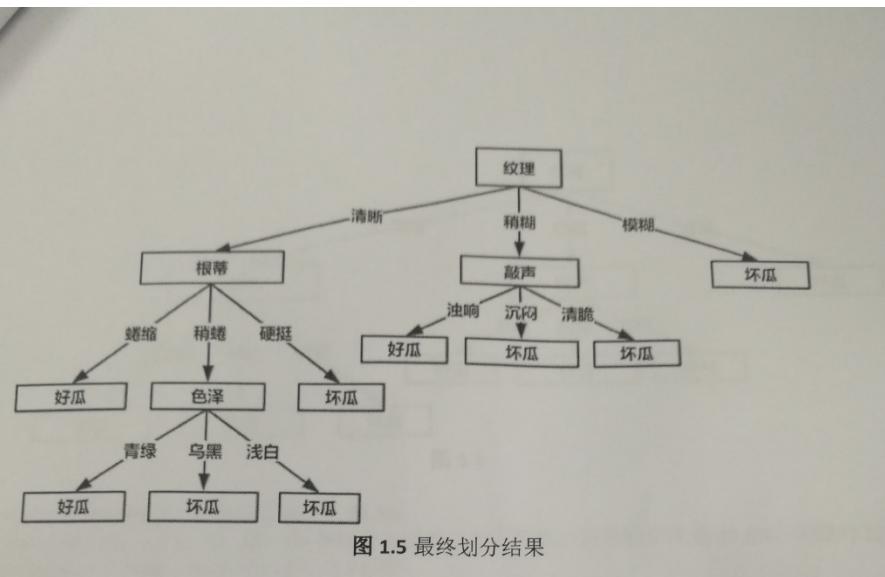


图 1.5 最终划分结果

敲声 > 处理 > 色泽 > 根蒂

① 划分敲声
根据信息增益准则

敲声 = ?
响亮 混乱 清脆

$$Ent(D) = \left(\frac{4}{6} \log_2 \frac{4}{6} + \frac{2}{6} \log_2 \frac{2}{6} \right) = 0.918$$

一、色泽 {清绿, 乌黑, 淡白}
 $D^1 \{1, 6\}$
 $D^2 \{3, 7, 15\}$
 $D^3 \{16\}$

正 反
0 $\frac{1}{3}$
 $\frac{2}{3}$ 0
0 1

$$Ent(D^1) = 0$$

$$Ent(D^2) = 0.918$$

$$Ent(D^3) = 0$$

$$Gain(D^1, \text{色泽}) = 0.459$$

二、根蒂 {蜷缩, 稍蜷}
 $D^1 \{1, 3, 16\}$
 $D^2 \{6, 7, 15\}$

正 反
 $\frac{2}{3}$ $\frac{1}{3}$
 $\frac{1}{3}$ 1

$$Ent(D^1) = 0.918$$

$$Ent(D^2) = 0.918$$

$$Gain(D^1, \text{根蒂}) = 0$$

三、纹理：{清晰, 模糊, 模糊}

$D^1 \{1, 3, 6, 15\}$ 正 反
 $D^2 \{7\}$ 1 0
 $D^3 \{16\}$ 0 1

$$Ent(D^1) = 0.811$$

$$Ent(D^2) = 0$$

$$Ent(D^3) = 0$$

$$Gain(D^1, \text{纹理}) = 0.377$$

② 选色泽



电子科技大学

University of Electronic Science and Technology of China

2. 用表 4.2 中编号为 4、5、8、11、12、13 的样本做测试集，对上题的训练数据采用预剪枝策略构建决策树，并汇报验证集精度。(35 分)

2、用表 4.2 中编号为 4、5、8、11、12、13 的样本做测试集，对上题的训练数据采用预剪枝策略构建决策树，并汇报验证集精度。

首先对根节点进行判断，在划分前，其类别标记为训练样例中最多的类别，于是将这个类别标记为坏瓜，精度为 $\frac{3}{6} \times 100\% = 50\%$

划分后三个节点分别包括 {1,2,3,6,10,15}, {7,9,14,17}, {16}，这三个节点分别被标记为好瓜，坏瓜，坏瓜，验证集的精度为 100%，进行划分，由于精度无法再增大。所有停止划分

最终的划分结果如图 2.1

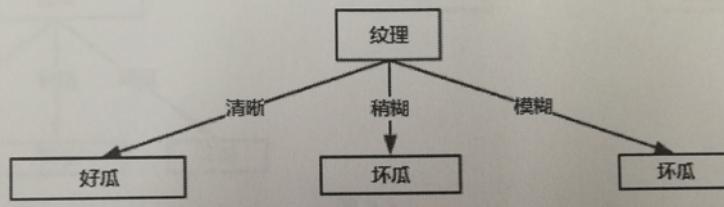


图 2.1

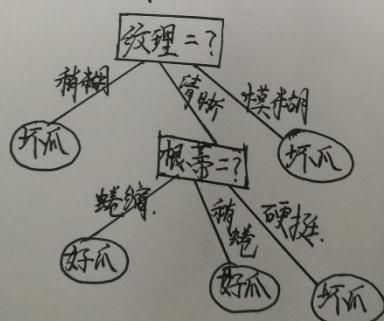
3. 用表 4.2 中编号为 4、5、8、11、12、13 的样本做测试集，对题 1 所构建的决策树进行后剪枝，并汇报验证集精度。(25 分)

对于右图决策树，基于表 4.2 中编号为 4、5、8、11、12、13 为样本的测试集易知，验证集精度为 66.7%。

考察结点⑤，若将该结点不平衡的分支剪除，相当于把⑤替换为叶结点，该叶结点类别为“坏瓜”，此时验证集精度为 83.3%，于是后剪枝策略决定剪枝。

考察结点②，将其标记为叶结点，类别为“坏瓜”此时验证集精度为 100%，于是后剪枝策略决定剪枝。

因精度已达 100%，故无需进行继续剪枝，最终决策树如下：



选择属性“敲声”

电子科技大学
University of Electronic Science and Technology of China

作业二

1.	端号	色泽	根蒂	敲声	纹理	脐部	触感	是否
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑		是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑		是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑		是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘		是
7	乌黑	稍蜷	浊响	稍模糊	稍凹	软粘		是
9	乌黑	稍蜷	沉闷	稍模糊	稍凹	硬滑		否
10	青绿	硬挺	清脆	清晰	平坦	软粘		否
14	浅白	稍蜷	沉闷	稍模糊	凹陷	硬滑		否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘		否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑		否
17	青绿	蜷缩	沉闷	稍模糊	稍凹	硬滑		否

$|y|=2$ 正例 $P_1 = \frac{5}{11}$ 反例 $P_2 = \frac{6}{11}$

信息熵 $Ent(D) = -\sum_{k=1}^2 P_k (\log_2 P_k) = -(\frac{5}{11} \log_2 \frac{5}{11} + \frac{6}{11} \log_2 \frac{6}{11}) = 0.994$

一、色泽：{青绿, 乌黑, 浅白}

D¹ D² D³

地址：成都市建设北路二段四号
高新区西源大道2006号

查询电话：83201114

邮政编码：610054（沙河校区）
611731（清水河校区）



电子科技大学

University of Electronic Science and Technology of China

三、敲声：{法响、沉闷、清脆}

	D'	D*	D''	正	反
D'	{1, 3, 6, 7, 15, 16}			$\frac{4}{6}$	$\frac{2}{6}$
D*	{2, 9, 14, 17}			$\frac{1}{4}$	$\frac{3}{4}$
D''	{10}			0	1
$Ent(D')$	= 0.918			$\frac{6}{11}$	
$Ent(D^*)$	= 0.811			$\frac{4}{11}$	
$Ent(D'')$	= 0			$\frac{1}{11}$	
$Gain(D, 敲声)$	= 0.198				

四、纹理：{清晰、稍糊、模糊}

	D'	D*	D''	正	反
D'	{1, 2, 3, 6, 10, 15}			$\frac{4}{6}$	$\frac{2}{6}$
D*	{7, 9, 14, 17}			$\frac{1}{4}$	$\frac{3}{4}$
D''	{16}			0	1
$Ent(D')$	= 0.918			$\frac{6}{11}$	
$Ent(D^*)$	= 0.811			$\frac{4}{11}$	
$Ent(D'')$	= 0			$\frac{1}{11}$	
$Gain(D, 纹理)$	= 0.198				

地址：成都市建设北路二段四号
高新区西源大道2006号

查询电话：83201114

邮政编码：610054（沙河校区）
611731（清水河校区）

对于敲声=法响 色泽=乌黑
 $Ent(D^2) = 0.918$ 正 反
 1. 根蒂 { 稍皱, 稍皱 } $D' \{ 3 \}$ 1 0
 $D^2 \{ 7, 15 \}$ $\frac{1}{2}$ $\frac{1}{2}$
 $Ent(D') = 0$
 $Ent(D^2) = 1$
 $Gain(D^2, 根蒂) = 0.25$

2. 纹理 { 清晰, 稍糊 } $D' \{ 3, 15 \}$ 正 反
 $D^2 \{ 7 \}$ 1 0
 $Ent(D') = 1$
 $Ent(D^2) = 0$
 $Gain(D^2, 纹理) = 0.25$
 (1) 根蒂

对于 敲声=法响 色泽=乌黑 根蒂=稍皱
 只剩一个(5) 纹理 { 清晰, 稍糊 } $D' \{ 15 \}$ 正 反
 $D^2 \{ 7 \}$ 0 1
 $Ent(D') = 0$
 $Ent(D^2) = 0$
 $Gain(D, 纹理) = 1$
 其实也不用算 直接用上

