三、根据表5.1所给的训练数据集，利用信息增益比(C4.5算法)生成决策树。

解：首先计算数据集D的经验熵 $H(D)$。

$$H(D)=-\frac{9}{15}\log_2\frac{9}{15}-\frac{6}{15}\log_2\frac{6}{15}=0.971$$

然后计算各特征对数据集D的信息增益。分别以 A_1, A_2, A_3, A_4 表示年龄，有工作，有自己的房子和信贷情况4个特征，则

$$\begin{aligned}g(D,A_1)&=H(D)-\left[\frac{5}{15}H(D_1)+\frac{5}{15}H(D_2)+\frac{5}{15}H(D_3)\right]\\&=0.971-\left[\frac{5}{15}\left(-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}\right)+\frac{5}{15}\left(-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5}\right)\right.\\&\quad\left.+\frac{5}{15}\left(-\frac{4}{5}\log_2\frac{4}{5}-\frac{1}{5}\log_2\frac{1}{5}\right)\right]\\&=0.083\end{aligned}$$

$$\begin{aligned}g(D,A_2)&=H(D)-\left[\frac{5}{15}H(D_1)+\frac{10}{15}H(D_2)\right]\\&=0.971-\left[\frac{5}{15}\times 0+\frac{10}{15}\left(-\frac{4}{10}\log_2\frac{4}{10}-\frac{6}{10}\log_2\frac{6}{10}\right)\right]\\&=0.324\end{aligned}$$

$$\begin{aligned}g(D,A_3)&=H(D)-\left[\frac{6}{15}H(D_1)+\frac{9}{15}H(D_2)\right]\\&=0.971-\left[\frac{6}{15}\times 0+\frac{9}{15}\left(-\frac{3}{9}\log_2\frac{3}{9}-\frac{6}{9}\log_2\frac{6}{9}\right)\right]\\&=0.420\end{aligned}$$

$$\begin{aligned}g(D,A_4)&=H(D)-\left[\frac{5}{15}H(D_1)+\frac{6}{15}H(D_2)+\frac{4}{15}H(D_3)\right]\\&=0.971-\left[\left(-\frac{1}{5}\log_2\frac{1}{5}-\frac{4}{5}\log_2\frac{4}{5}\right)\cdot\frac{5}{15}+\frac{6}{15}\left(-\frac{4}{6}\log_2\frac{4}{6}-\frac{2}{6}\log_2\frac{2}{6}\right)\right.\\&\quad\left.+\frac{4}{15}\left(-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4}\right)\right]\\&=0.363\end{aligned}$$

接着计算数据集D关于各个特征值A的值的熵 $H_A(D)$：

$$H_{A_1}(D)=-\frac{5}{15}\log_2\frac{5}{15}-\frac{5}{15}\log_2\frac{5}{15}-\frac{5}{15}\log_2\frac{5}{15}=1.585$$

$$H_{A_2}(D)=-\frac{5}{15}\log_2\frac{5}{15}-\frac{10}{15}\log_2\frac{10}{15}=0.918$$

$$H_{A_3}(D)=-\frac{6}{15}\log_2\frac{6}{15}-\frac{9}{15}\log_2\frac{9}{15}=0.971$$

$$H_{A_4}(D)=-\frac{5}{15}\log_2\frac{5}{15}-\frac{6}{15}\log_2\frac{6}{15}-\frac{4}{15}\log_2\frac{4}{15}=1.566$$

最后计算各个特征的信息增益比.

$$g_R(D, A_1) = \frac{g(D, A_1)}{H_{A_1}(D)} = 0.052$$

$$g_R(D, A_2) = \frac{g(D, A_2)}{H_{A_2}(D)} = 0.353$$

$$g_R(D, A_3) = \frac{g(D, A_3)}{H_{A_3}(D)} = 0.433$$

$$g_R(D, A_4) = \frac{g(D, A_4)}{H_{A_4}(D)} = 0.232$$

比较各个特征的信息增益比。由于特征A_3(有自己的房子)的信息增益比最大,所以,选择特征A_3作为分支的特征条件,把数据集D_1分为两部分D_1(有自己的房子)和D_2(没有自己的房子).再对数据集D_1,D_2建树.D_1,D_2上的特征有A_1,A_2,A_4.数据集D_1中的实例都属于同一类("是"),故建立单节点,类别为"是".

下面针对数据集D_2计算各个特征的信息增益比.

首先计算经验熵$H(D_2)$.

$$H(D_2) = -\frac{3}{9}\log_2\frac{3}{9} - \frac{6}{9}\log_2\frac{6}{9} = 0.918$$

然后计算特征A_1,A_2,A_4对数据集D_2的信息增益.

$$g(D_2, A_1) = H(D_2) - \left[\frac{4}{9}\left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right) + \frac{2}{9}\times 0 + \frac{3}{9}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right)\right]$$
$$= 0.251$$

$$g(D_2, A_2) = H(D_2) - \left[\frac{3}{9}\times 0 + \frac{6}{9}\times 0\right] = 0.918$$

$$g(D_2, A_4) = H(D_2) - \left[\frac{5}{9}\times 0 + \frac{4}{9}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + \frac{1}{9}\times 0\right] = 0.474$$

接着计算数据集D_2关于各个特征的取值的熵$H_A(D_2)$。

$$H_{A_1}(D_2) = -\frac{5}{9}\log_2\frac{5}{9} - \frac{2}{9}\log_2\frac{2}{9} - \frac{2}{9}\log_2\frac{2}{9} = 1.531$$

$$H_{A_2}(D_2) = -\frac{3}{9}\log_2\frac{3}{9} - \frac{6}{9}\log_2\frac{6}{9} = 0.918$$

$$H_{A_4}(D_2) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{4}{9}\log_2\frac{4}{9} - \frac{1}{9}\log_2\frac{1}{9} = 1.392$$

最后计算各个特征的信息增益比.

$$g_R(D_2, A_1) = \frac{g(D_2, A_1)}{H_{A_1}(D_2)} = 0.164$$

$$g_R(D_2, A_2) = \frac{g(D_2, A_2)}{H_{A_2}(D_2)} = 1$$

$$g_R(D_2, A_4) = \frac{g(D_2, A_4)}{H_{A_4}(D_2)} = 0.341$$

比较各个特征的信息增益比。由于特征 A_2(有工作)的信息增益比最大，所以，将特征 A_2 作为分支的条件。再把 D_2 分为两部分 D_3(有工作)和 D_4(无工作)，分别再对 D_3，D_4 建树，由于数据集 D_3 中的所有实例都属于类别"是"，故建立单节点树，类别为"是"。同理，由于数据集 D_4 中的所有实例都属于类别"否"，故建立单节点树，类别为"否"。

从而，由以上信息，我们可以生成决策树为：