HOMEWORK 6

梁敬聪　　18307110286

2022 年 4 月 25 日

设 A_1, A_2, A_3 代表资历、教育程度和有无经验三个特征。首先对根节点计算经验熵:

$$H(D_0) = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10} = 0.5 + 0.5 = 1$$

接下来对三个特征计算经验条件熵和固有值:

$$\begin{aligned}
H(D_0|A_1) &= \frac{5}{10}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{3}{10}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) \\
&\quad + \frac{2}{10}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) \\
&= 0.4855 + 0.2755 + 0.2 = 0.9610 \\
H_{A_1}(D_0) &= -\frac{5}{10}\log_2\frac{5}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{2}{10}\log_2\frac{2}{10} = 0.5 + 0.5211 + 0.4644 = 1.485 \\
H(D_0|A_2) &= \frac{3}{10}\times 0 + \frac{7}{10}\left(-\frac{2}{7}\log_2\frac{2}{7} - \frac{5}{7}\log_2\frac{5}{7}\right) = 0.6042 \\
H_{A_2}(D_0) &= -\frac{3}{10}\log_2\frac{3}{10} - \frac{7}{10}\log_2\frac{7}{10} = 0.5211 + 0.3602 = 0.8813 \\
H(D_0|A_3) &= \frac{4}{10}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{6}{10}\left(-\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}\right) \\
&= 0.3245 + 0.5510 = 0.8755 \\
H_{A_3}(D_0) &= -\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10} = 0.5288 + 0.4422 = 0.9710
\end{aligned}$$

由此我们可以得到三个特征的信息增益比:

$$\begin{aligned}
g_R(D_0, A_1) &= \frac{H(D_0) - H(D_0|A_1)}{H_{A_1}(D_0)} = \frac{1 - 0.9610}{1.485} = 0.02628 \\
g_R(D_0, A_2) &= \frac{H(D_0) - H(D_0|A_2)}{H_{A_2}(D_0)} = \frac{1 - 0.6042}{0.8813} = 0.4491 \\
g_R(D_0, A_3) &= \frac{H(D_0) - H(D_0|A_3)}{H_{A_3}(D_0)} = \frac{1 - 0.8755}{0.9710} = 0.1282
\end{aligned}$$

其中特征 A_2（教育程度）的信息增益比最大，因此选择 A_2 作为根节点的特征，将 D_0 划分为 D_1（A_2 为“本科”）和 D_2（A_2 为“硕士”）。由于 D_1 中的类别均为“普通”，因此对应子节点成为叶节点，类标记为“普通”。对于 D_2 对应的子节点，我们继续从 A_1, A_3 中选择特征。对该节点计算经验熵:

$$H(D_2) = -\frac{2}{7}\log_2\frac{2}{7} - \frac{5}{7}\log_2\frac{5}{7} = 0.5164 + 0.3467 = 0.8631$$

剩余特征的经验条件熵和固有值为:

$$
\begin{aligned}
H(D_2|A_1) &= \frac{4}{7}\left(-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4}\right)+\frac{1}{7}\times 0+\frac{2}{7}\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)\\
&= 0.4636+0.2857=0.7493\\
H_{A_1}(D_2) &= -\frac{4}{7}\log_2\frac{4}{7}-\frac{1}{7}\log_2\frac{1}{7}-\frac{2}{7}\log_2\frac{2}{7}=0.4613+0.4011+0.5164=1.379\\
H(D_2|A_3) &= \frac{2}{7}\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)+\frac{5}{7}\left(-\frac{1}{5}\log_2\frac{1}{5}-\frac{4}{5}\log_2\frac{4}{5}\right)\\
&= 0.2857+0.5157=0.8014\\
H_{A_3}(D_2) &= -\frac{2}{7}\log_2\frac{2}{7}-\frac{5}{7}\log_2\frac{5}{7}=0.5164+0.3467=0.8631
\end{aligned}
$$

进而它们的信息增益比为:

$$
\begin{aligned}
g_R(D_2,A_1) &= \frac{H(D_2)-H(D_2|A_1)}{H_{A_1}(D_2)}=\frac{0.8631-0.7493}{1.379}=0.08255\\
g_R(D_2,A_3) &= \frac{H(D_2)-H(D_2|A_3)}{H_{A_3}(D_2)}=\frac{0.8631-0.8014}{0.8631}=0.07154
\end{aligned}
$$

其中特征 A_1（资历）的信息增益比最大，因此选择 A_1 作为该节点的特征，将 D_2 划分为 D_3（A_1 为“3 年以下”）、D_4（A_1 为“3 年至 5 年”）和 D_5（A_1 为“5 年以上”）。由于 D_4 中的类别均为“优秀”，因此对应的子节点成为叶节点，类标记为“优秀”；而 D_3 和 D_5 对应的子节点还需要再生成子节点，但由于仅剩特征 A_3（有无经验），因此它们只能选择 A_3 作为其特征。其中，D_3 划分为 D_6（A_3 为“无经验”）和 D_7（A_3 为“有经验”），均为子节点；D_6 中的类别均为“优秀”，因此类标记为“优秀”；D_7 中有 2 个“优秀”和一个“普通”，按实例数最大的类也标记为“优秀”。同理，D_5 划分为 D_8（A_3 为“无经验”）和 D_9（A_3 为“有经验”），类标记分别为“普通”和“优秀”。由此便生成了一棵有四个内部节点的决策树（尽管有一些冗余节点），如图 1 所示。

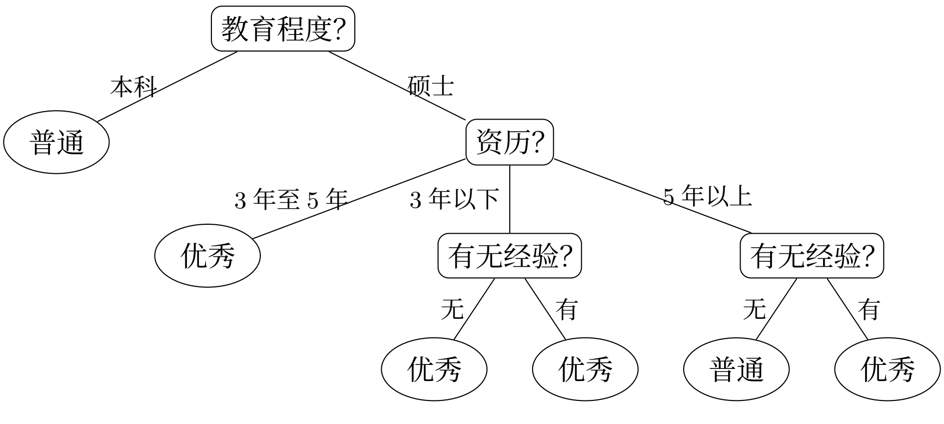

图 1: C4.5 算法得到的决策树