Chapter9.4

决策树

分类概述

什么是分类?

- 分类就是根据以往的数据和结果对另一部分数据进行结果的预测。
- 模型的学习在被告知每个训练样本属于哪个类的“指导” 下进行新数据使用训练数据集中得到的规则进行分类

分类的基本过程:

- 学习阶段：建立一个分类模型，描述预定数据类或概念集。

  评估模型的预测准确率

  如果准确率可以接受，那么使用该模型来分类标签为未知的样本。
- 分类阶段：即使用分类模型，对将来的或未知的对象进行分类。

数据集：训练集、测试集、预测数据集

分类概述

分类与预测

- 不同点
  - 分类是预测类对象的分类标号（或离散值），根据训练数据集和类标号属性，构建模型来分类现有数据，并用来分类新数据。
  - 预测是建立连续函数值模型评估无标号样本类，或评估给定样本可能具有的属性值或值区间，即用来估计连续值或量化属性值，比如预测空缺值。
- 相同点
  - 分类和预测的共同点是两者都需要构建模型，都用模型来估计未知值。预测中主要的估计方法是回归分析。

ID3算法

- **ID3算法**最早是由罗斯昆（J. Ross Quinlan）于1975年在悉尼大学提出的一种分类预测算法，算法的核心是**“信息熵”**。ID3算法通过计算每个属性的**信息增益**，认为信息增益高的是好属性，每次划分选取信息增益最高的属性为划分标准，重复这个过程，直至生成一个能完美分类训练样例的决策树。
- 决策树是对数据进行分类，以此达到预测的目的。该决策树方法先根据训练集数据形成决策树，如果该树不能对所有对象给出正确的分类，那么选择一些**例外**加入到训练集数据中，重复该过程一直到形成正确的决策集。决策树代表着决策集的树形结构。
- 该算法是以信息论为基础，以信息熵和信息增益度为衡量标准，从而实现对数据的归纳分类。

ID3

例 使用ID3算法进行分类预测

表2和表3为训练数据和测试数据，其中“患病与否”是类标记，使用ID3算法构建决策树然后进行分类预测。

表2 某疾病患病情况的训练数据

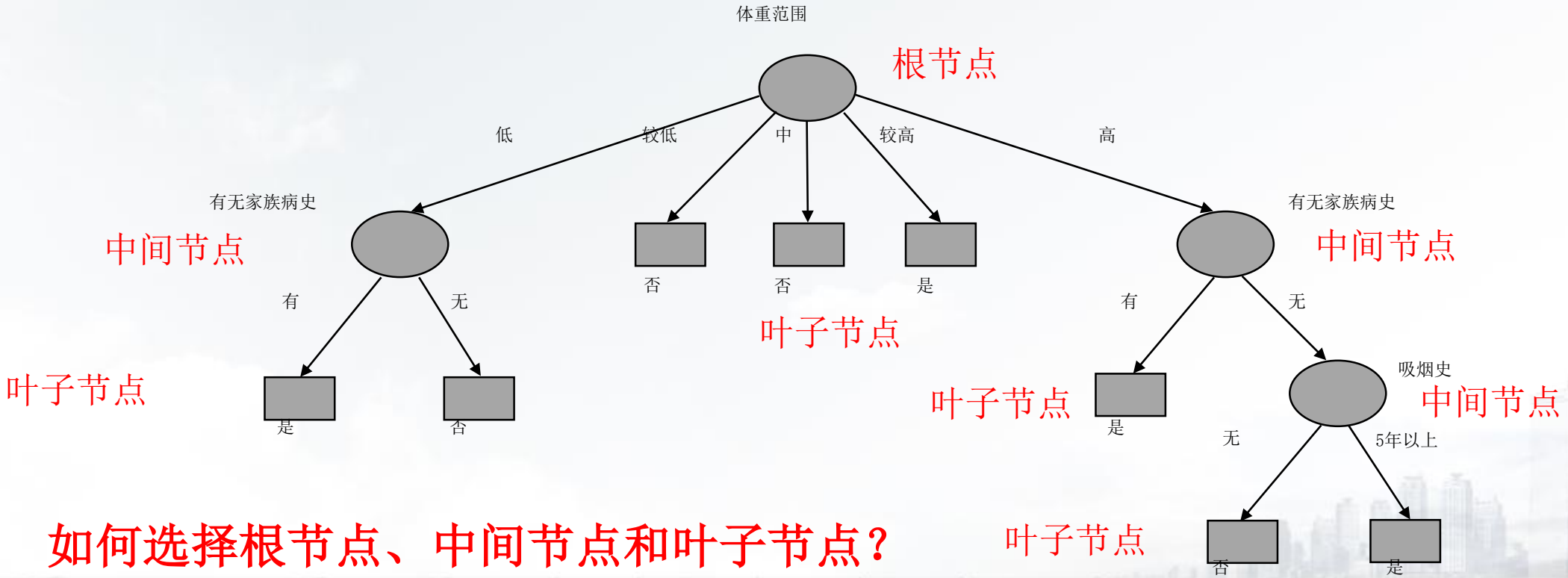
ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
1	23	无	无	较低	否
2	25	无	无	中	否
3	27	0-5年	无	中	否
4	30	0-5年	有	低	是
5	39	无	无	较低	否
6	41	无	无	低	否
7	43	无	无	高	否
8	45	5年以上	有	高	是
9	46	无	有	高	是
10	47	无	有	高	是
11	62	无	有	较高	是
12	63	无	有	高	是
13	66	5年以上	无	高	是
14	66	5年以上	无	较高	是
15	68	0-5年	无	中	否

表3 某疾病患病情况的测试数据

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
1	25	无	无	较低	?
2	42	无	无	高	?
3	67	5年以上	无	较高	?

ID3

利用ID3算法构建的决策树如图5所示。



如何选择根节点、中间节点和叶子节点？

图5 ID3算法构造的决策树

分类的相关知识：

– 1、信息熵

信息熵用来衡量事件的不确定性的大小，计算公式如下：

$$Infor(x) = -p(x) \times log_2 p(x)$$

信息熵具有可加性，即多个期望信息，计算公式如下：

$$Infor(X) = -\sum_{i=1}^{m} p(x_i) \times log_2 p(x_i)$$

分类的相关知识：

– **2、信息增益**

信息增益表示某一特征的信息对类标签的**不确定性减少的程度**。

$$g(D|A) = Infor(D) - Infor(D|A)$$

其中$Infor(D|A)$是在特征*A*给定条件下对数据集合*D*进行划分所需要的期望信息，它的值越小表示分区(分类)的**纯度越高**，计算公式如下。

$$Infor(D|A) = \sum_{j=1}^{n} \frac{|D_j|}{|D|} \times Info(D_j) \qquad (7\text{-}4)$$

其中n是数据分区数，$|D_j|$表示第j个数据分区的长度，$\frac{|D_j|}{|D|}$表示第*j*个数据分区的权重。

- **例1 信息增益的计算**
- 表1是带有标记类的训练集D，训练集的列是一些特征，表中最后一列的类标号为是否提供贷款，有两个不同的取值，计算按照每个特征进行划分的信息增益。

表1 贷款申请的训练集

ID	学历	婚否	是否有车	收入水平	类别
1	专科	否	否	中	否
2	专科	否	否	高	否
3	专科	是	否	高	是
4	专科	是	是	中	是
5	专科	否	否	中	否
6	本科	否	否	中	否
7	本科	否	否	高	否
8	本科	是	是	高	是
9	本科	否	是	很高	是
10	本科	否	是	很高	是
11	研究生	否	是	很高	是
12	研究生	否	是	高	是
13	研究生	是	否	高	是
14	研究生	是	否	很高	是
15	研究生	否	否	中	否

①根据公式计算信息熵 $Infor(D)$。

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

②计算按照每个特征进行划分的期望信息，A代表特征“学历”，B代表特征“婚否”，C代表特征“是否有车”，E代表特征“收入水平”。

$$Infor(D|A) = \frac{5}{15} \times \left(-\frac{2}{5} log_2 \frac{2}{5} - \frac{3}{5} log_2 \frac{3}{5}\right) + \frac{5}{15} \times \left(-\frac{3}{5} log_2 \frac{3}{5} - \frac{2}{5} log_2 \frac{2}{5}\right)$$
$$+ \frac{5}{15} \times (-\frac{4}{5} log_2 \frac{4}{5} - \frac{1}{5} log_2 \frac{1}{5}) = 0.888$$

$$Infor(D|B) = Infor(D|B) = \frac{10}{15} \times \left(-\frac{6}{10} log_2 \frac{6}{10} - \frac{4}{10} log_2 \frac{4}{10}\right) + \frac{5}{15} \times \left(-\frac{5}{5} log_2 \frac{5}{5}\right) = 0.647$$

$$Infor(D|C) = \frac{9}{15} \times \left(-\frac{6}{9} log_2 \frac{6}{9} - \frac{3}{9} log_2 \frac{3}{9}\right) + \frac{6}{15} \times \left(-\frac{6}{6} log_2 \frac{6}{6}\right) = 0.951$$

$$Infor(D|E) = \frac{5}{15} \times \left(-\frac{4}{5} log_2 \frac{4}{5} - \frac{1}{5} log_2 \frac{1}{5}\right) + \frac{6}{15} \times \left(-\frac{2}{6} log_2 \frac{2}{6} - \frac{4}{6} log_2 \frac{4}{6}\right) + \frac{4}{15} \times (-\frac{4}{4} log_2 \frac{4}{4}) = 0.608$$

③计算信息增益

$$g(D|A) = Infor(D) - Infor(D|A) = 0.083$$
$$g(D|B) = Infor(D) - Infor(D|B) = 0.324$$
$$g(D|C) = Infor(D) - Infor(D|C) = 0.020$$
$$g(D|E) = Infor(D) - Infor(D|E) = 0.363$$

分类的相关知识：

– 3、信息增益率

信息增益率是指按照某一特征进行划分的信息增益与训练集关于这个特征的信息熵的比值：

$$g_r(D,A)=\frac{g(D|A)}{SplitInfor_A(D)}$$

其中：

$$SplitInfor_A(D)=-\sum_{i=1}^{n}\frac{|D_j|}{|D|}\times log_2(\frac{|D_j|}{|D|})$$

例 2 信息增益率的计算

基于例1的数据，计算按照每个特征进行划分的**信息增益率**。

解：①根据例1计算出的按照每个特征划分的信息增益，*A*代表特征“学历”，*B*代表特征“婚否”，*C*代表特征“是否有车”，*E*代表特征“收入水平”，计算$SplitInfor_A(D)$。

$$SplitInfor_A(D) = -\frac{5}{15} \times log_2 \frac{5}{15} - \frac{5}{15} \times log_2 \frac{5}{15} - \frac{5}{15} \times log_2 \frac{5}{15} = 1.585$$

$$SplitInfor_B(D) = -\frac{10}{15} \times log_2 \frac{10}{15} - \frac{5}{15} \times log_2 \frac{5}{15} = 0.918$$

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

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

②按照公式计算**信息增益率**。

$$g_r(D,A) = \frac{0.083}{1.585} = 0.052$$

$$g_r(D,B) = \frac{0.324}{0.918} = 0.331$$

$$g_r(D,C) = \frac{0.020}{0.971} = 0.021$$

$$g_r(D,E) = \frac{0.363}{1.566} = 0.232$$

ID3

例4 使用ID3算法进行分类预测

表2和表3为训练数据和测试数据，其中“患病与否”是类标记，使用ID3算法构建决策树然后进行分类预测。

表2 某疾病患病情况的训练数据

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
1	23	无	无	较低	否
2	25	无	无	中	否
3	27	0-5年	无	中	否
4	30	0-5年	有	低	是
5	39	无	无	较低	否
6	41	无	无	低	否
7	43	无	无	高	否
8	45	5年以上	有	高	是
9	46	无	有	高	是
10	47	无	有	高	是
11	62	无	有	较高	是
12	63	无	有	高	是
13	66	5年以上	无	高	是
14	66	5年以上	无	较高	是
15	68	0-5年	无	中	否

表3 某疾病患病情况的测试数据

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
1	25	无	无	较低	?
2	42	无	无	高	?
3	67	5年以上	无	较高	?

年龄数据有何不同?
是否可以直接应用?

ID3

解：①连续型数据的离散化**(分段)**。ID3算法不能直接处理连续型数据，只有通过离散化将连续型数据转化成离散型数据再进行处理。

此例采用等宽分箱法对连续型特征“年龄”离散化：

设定区域范围（设箱子数为3，箱子宽度为（68-23）/3=15），分箱结果为：

箱1：23 25 27 30

箱2：39 41 43 45 46 47

箱3：62 63 66 66 68

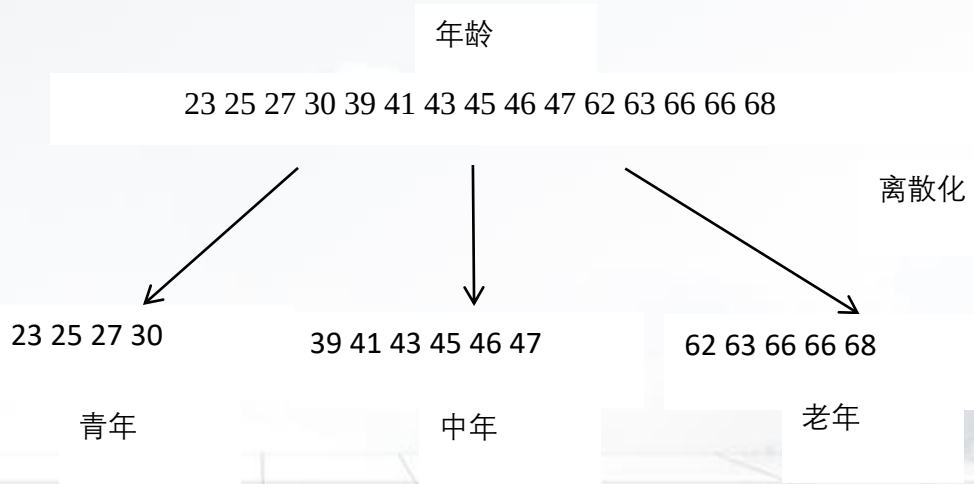


图4 对特征“年龄”分箱

ID3

离散后训练数据集如表4所示。

表4 某疾病患病情况的训练数据（离散化后）

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
1	青年	无	无	较低	否
2	青年	无	无	中	否
3	青年	0-5年	无	中	否
4	青年	0-5年	有	低	是
5	中年	无	无	较低	否
6	中年	无	无	低	否
7	中年	无	无	高	否
8	中年	5年以上	有	高	是
9	中年	无	有	高	是
10	中年	无	有	高	是
11	老年	无	有	较高	是
12	老年	无	有	高	是
13	老年	5年以上	无	高	是
14	老年	5年以上	无	较高	是
15	老年	0-5年	无	中	否

ID3

②根据训练数据构造ID3算法的决策树，其中Z代表训练集，A、B、C、D分别代表特征“年龄”、“吸烟史”、“有无家族病史”、“体重范围”，按照**每个特征计算其分裂的信息增益。**

$$Infor(Z) = -\frac{8}{15} \times \log_2 \frac{8}{15} - \frac{7}{15} \times \log_2 \frac{7}{15} = 0.997$$

$$Infor(Z|A)$$
$$= \frac{4}{15} \times \left(-\frac{3}{4} \times \log_2 \frac{3}{4} - \frac{1}{4} \times \log_2 \frac{1}{4}\right) + \frac{6}{15} \times \left(-\frac{3}{6} \times \log_2 \frac{3}{6} - \frac{3}{6} \times \log_2 \frac{3}{6}\right) + \frac{5}{15}$$
$$\times \left(-\frac{4}{5} \times \log_2 \frac{4}{5} - \frac{1}{5} \times \log_2 \frac{1}{5}\right) = 0.857$$

$$Infor(Z|B)$$
$$= \frac{9}{15} \times \left(-\frac{5}{9} \times \log_2 \frac{5}{9} - \frac{4}{9} \times \log_2 \frac{4}{9}\right) + \frac{3}{15} \times \left(-\frac{2}{3} \times \log_2 \frac{2}{3} - \frac{1}{3} \times \log_2 \frac{1}{3}\right) + \frac{3}{15} \times \left(-\frac{3}{3} \times \log_2 \frac{3}{3}\right)$$
$$= 0.778$$

ID3

②根据训练数据构造ID3算法的决策树，其中Z代表训练集，A、B、C、D分别代表特征“年龄”、“吸烟史”、“有无家族病史”、“体重范围”，按照每个特征计算其分裂的信息增益。

$$Infor(Z|C)=\frac{9}{15}\times\left(-\frac{7}{9}\times\log_2\frac{7}{9}-\frac{2}{9}\times\log_2\frac{2}{9}\right)+\frac{6}{15}\times\left(-\frac{6}{6}\times\log_2\frac{6}{6}\right)=0.459$$

$$Infor(Z|D)$$

$$=\frac{2}{15}\times\left(-\frac{2}{2}\times\log_2\frac{2}{2}\right)+\frac{2}{15}\times\left(-\frac{1}{2}\times\log_2\frac{1}{2}-\frac{1}{2}\times\log_2\frac{1}{2}\right)+\frac{3}{15}\times\left(-\frac{3}{3}\times\log_2\frac{3}{3}\right)+\frac{6}{15}$$

$$\times\left(-\frac{5}{6}\times\log_2\frac{5}{6}-\frac{1}{6}\times\log_2\frac{1}{6}\right)+\frac{2}{15}\times\left(-\frac{2}{2}\times\log_2\frac{2}{2}\right)=0.393$$

ID3

②根据训练数据构造ID3算法的决策树，其中Z代表训练集，A、B、C、D分别代表特征“年龄”、“吸烟史”、“有无家族病史”、“体重范围”，按照每个特征计算其分裂的信息增益。

$$g(Z|A) = Infor(Z) - Infor(Z|A) = 0.997 - 0.857 = 0.140$$

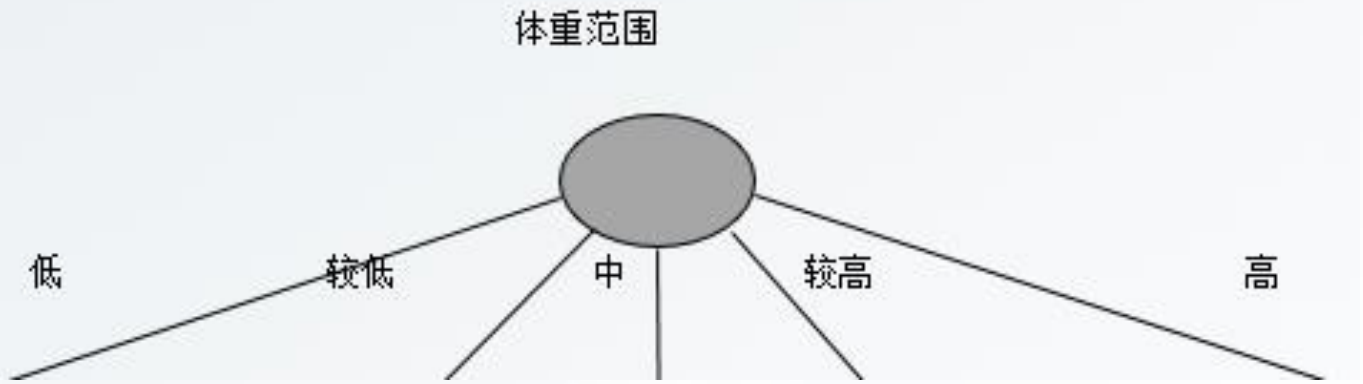
$$g(Z|B) = Infor(Z) - Infor(Z|B) = 0.997 - 0.778 = 0.219$$

$$g(Z|C) = Infor(Z) - Infor(Z|C) = 0.997 - 0.459 = 0.538$$

$$g(Z|D) = Infor(Z) - Infor(Z|D) = 0.997 - 0.393 = 0.604$$

选择信息增益最大特征“体重范围”作为根结点的分裂属性，将训练集Z划分为5个子集Z_1、Z_2、Z_3、Z_4和Z_5，对应的“体重范围”取值分别为“低”、“较低”、“中”、“较高”、“高”。由于Z_2、Z_3和Z_4只有一类数据，所以它们各自成为一个叶结点，三个结点的类标签分别为“否”、“否”、“是”。

根据信息增益最大特征“体重范围”作为根结点的分裂属性



ID3

表 Z_1训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
4	青年	0-5年	有	低	是
6	中年	无	无	低	否

表 Z_2训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
1	青年	无	无	较低	否
5	中年	无	无	较低	否

表 Z_3训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
2	青年	无	无	中	否
3	青年	0-5年	无	中	否
15	老年	0-5年	无	中	否

表 Z_4训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
11	老年	无	有	较高	是
14	老年	5年以上	无	较高	是

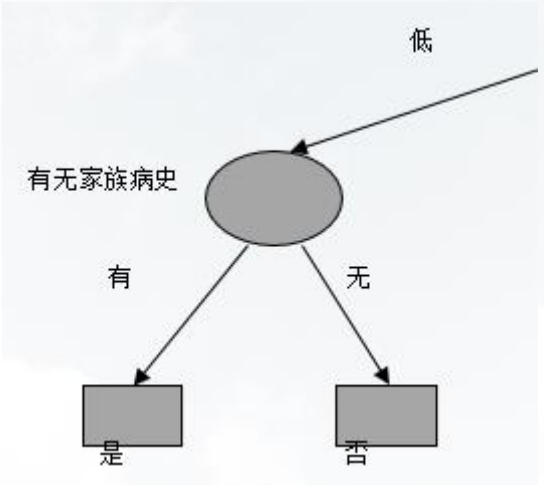
表 Z_5训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
7	中年	无	无	高	否
8	中年	5年以上	有	高	是
9	中年	无	有	高	是
10	中年	无	有	高	是
12	老年	无	有	高	是
13	老年	5年以上	无	高	是

ID3

③对于Z_1继续进行分裂，选择剩余特征中信息增益最大的作为分裂属性。

$$Infor(Z_1) = -\frac{1}{2} \times \log_2 \frac{1}{2} - \frac{1}{2} \times \log_2 \frac{1}{2} = 1$$



$$Infor(Z_1|A) = \frac{1}{2} \times \left(-\frac{1}{1} \times \log_2 \frac{1}{1}\right) + \frac{1}{2} \times \left(-\frac{1}{1} \times \log_2 \frac{1}{1}\right) = 0$$

$$Infor(Z_1|B) = \frac{1}{2} \times \left(-\frac{1}{1} \times \log_2 \frac{1}{1}\right) + \frac{1}{2} \times \left(-\frac{1}{1} \times \log_2 \frac{1}{1}\right) = 0$$

$$Infor(Z_1|C) = \frac{1}{2} \times \left(-\frac{1}{1} \times \log_2 \frac{1}{1}\right) + \frac{1}{2} \times \left(-\frac{1}{1} \times \log_2 \frac{1}{1}\right) = 0$$

ID3

③对于Z_1继续进行分裂，选择剩余特征中信息增益最大的作为分裂属性。

$$g(Z_1|A) = Infor(Z_1) - Infor(Z_1|A) = 1 - 0 = 1$$

$$g(Z_1|B) = Infor(Z_1) - Infor(Z_1|B) = 1 - 0 = 1$$

$$g(Z_1|C) = Infor(Z_1) - Infor(Z_1|C) = 1 - 0 = 1$$

选择信息增益最大特征作为Z_1结点的分裂属性，由于三个属性的信息增益相同，**随机挑选**一个作为分裂属性，此处选取“**有无家族病史**”作为分裂属性，它将数据集Z_1分成2个子集Z_{21}和Z_{22}，对应的“有无家族病史”的取值分别为“有”和“无”，在这2个子集中的数据都各自属于同一类，于是就不需要再继续分裂。

ID3

表 Z_{21}训练数据集

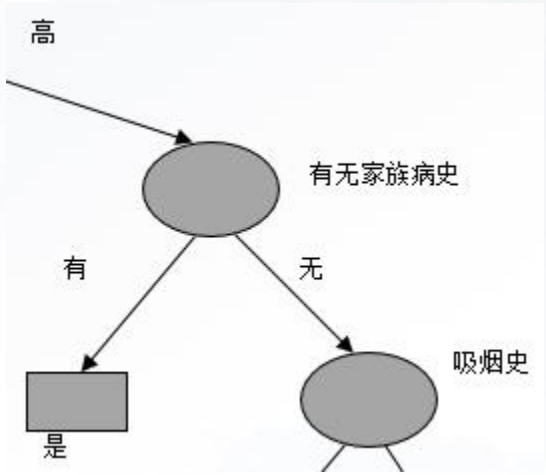
ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
6	中年	无	无	低	否

表 Z_{22}训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
4	青年	0-5年	有	低	是

ID3

④对于Z_5继续进行分裂，选择剩余特征中信息增益最大的作为分裂属性。



$$Infor(Z_5) = -\frac{1}{6}\times\log_2\frac{1}{6} - \frac{5}{6}\times\log_2\frac{5}{6} = 0.650$$

$$Infor(Z_5|A) = \frac{4}{6}\times\left(-\frac{1}{4}\times\log_2\frac{1}{4} - \frac{3}{4}\times\log_2\frac{3}{4}\right) + \frac{2}{6}\times\left(-\frac{2}{2}\times\log_2\frac{2}{2}\right) = 0.541$$

$$Infor(Z_5|B) = \frac{4}{6}\times\left(-\frac{3}{4}\times\log_2\frac{3}{4} - \frac{1}{4}\times\log_2\frac{1}{4}\right) + \frac{2}{6}\times\left(-\frac{2}{2}\times\log_2\frac{2}{2}\right) = 0.541$$

$$Infor(Z_5|C) = \frac{4}{6}\times\left(-\frac{4}{4}\times\log_2\frac{4}{4}\right) + \frac{2}{6}\times\left(-\frac{1}{2}\times\log_2\frac{1}{2} - \frac{1}{2}\times\log_2\frac{1}{2}\right) = 0.333$$

ID3

④对于Z_5继续进行分裂，选择剩余特征中信息增益最大的作为分裂属性。

$$g(Z_5|A) = Infor(Z_5) - Infor(Z_5|A) = 0.650 - 0.541 = 0.109$$

$$g(Z_5|B) = Infor(Z_5) - Infor(Z_5|B) = 0.650 - 0.541 = 0.109$$

$$g(Z_5|C) = Infor(Z_5) - Infor(Z_5|C) = 0.650 - 0.333 = 0.317$$

选择信息增益最大特征“**有无家族病史**”作为Z_5结点的分裂属性，将数据集Z_5分成2个子集Z_{51}和Z_{52}，对应的“有无家族病史”的取值分别为“有”和“无”，“有”对应的Z_{51}中的数据都属于同一类，不需要再继续分裂，“无”对应的Z_{52}中的数据属于不同类，需要对其进行分裂。

ID3

表 Z_{51}训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
8	中年	5年以上	有	高	是
9	中年	无	有	高	是
10	中年	无	有	高	是
12	老年	无	有	高	是

表 Z_{52}训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
7	中年	无	无	高	否
13	老年	5年以上	无	高	是

ID3

⑤对于Z_{52}继续进行分裂，选择剩余特征中信息增益最大的作为分裂属性。

$$Infor(Z_{52}) = -\frac{1}{2} \times \log_2 \frac{1}{2} - \frac{1}{2} \times \log_2 \frac{1}{2} = 1$$

$$Infor(Z_{52}|A) = \frac{1}{2} \times \left(-\frac{1}{1} \times \log_2 \frac{1}{1}\right) + \frac{1}{2} \times \left(-\frac{1}{1} \times \log_2 \frac{1}{1}\right) = 0$$

$$Infor(Z_{52}|B) == \frac{1}{2} \times \left(-\frac{1}{1} \times \log_2 \frac{1}{1}\right) + \frac{1}{2} \times \left(-\frac{1}{1} \times \log_2 \frac{1}{1}\right) = 0$$

$$g(Z_{52}|A) = Infor(Z_{51}) - Infor(Z_{51}|A) = 1 - 0 = 1$$

$$g(Z_{52}|B) = Infor(Z_{51}) - Infor(Z_{51}|B) = 1 - 0 = 1$$

由于两个属性的信息增益相同，随机挑选一个作为分裂属性，此处选取“**吸烟史**”作为Z_{52}结点的分裂属性，将数据集Z_{52}分成2个子集Z_{521}和Z_{522}，对应的“吸烟史”的取值分别为“无”和“5年以上”，在这两个子集中的数据都各自属于同一类，于是就不需要再继续分裂，决策树构造完毕。

ID3

表 Z_{521}训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
7	中年	无	无	高	否

表 Z_{522}训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
13	老年	5年以上	无	高	是

ID3

利用ID3算法构建的决策树如图5所示。

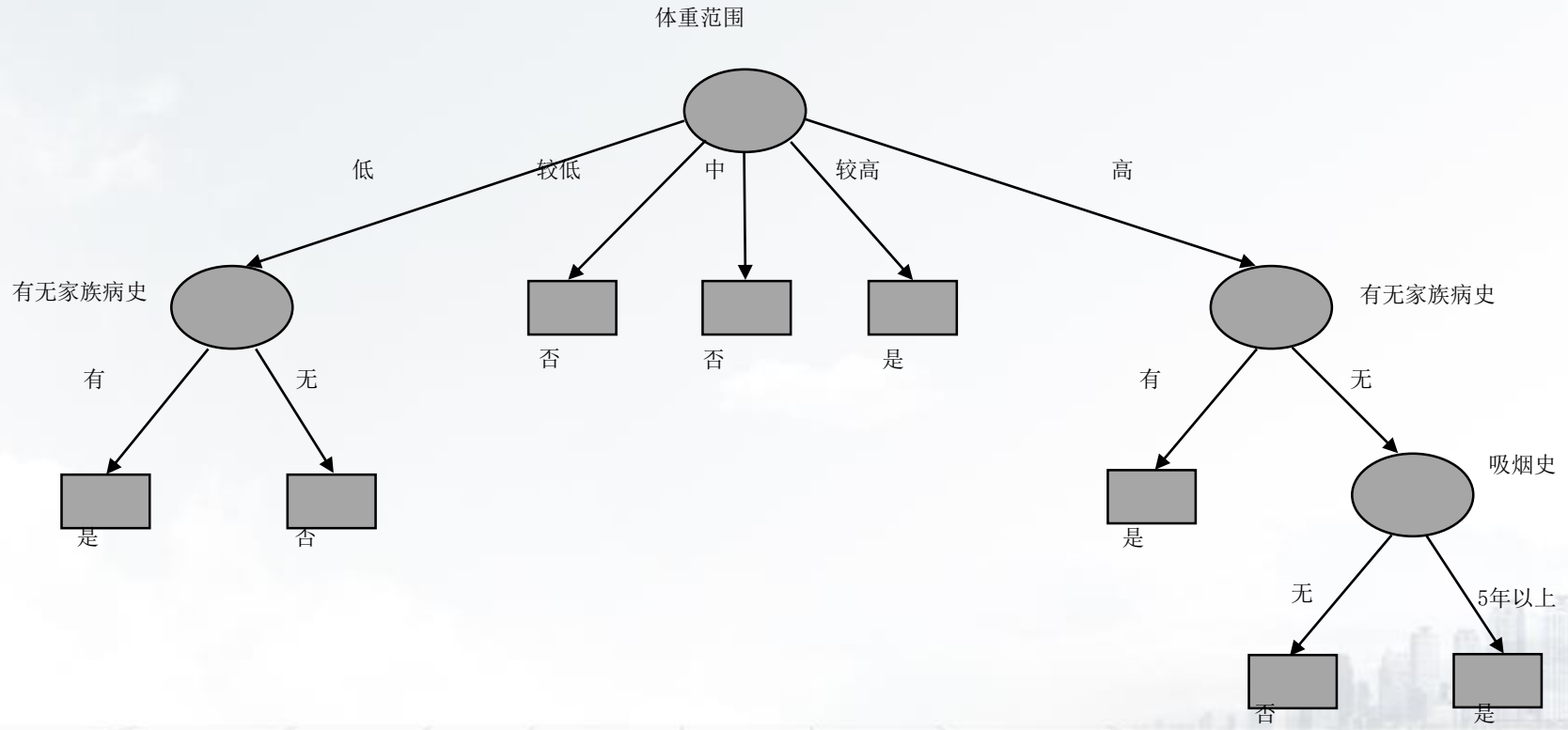


图5 ID3算法构造的决策树

ID3

根据构建的决策树，可以提取分类规则如下。

①IF “体重范围”=“低” AND “有无家族病史”=“有” THEN 患病

②IF “体重范围”=“低” AND “有无家族病史”=“无” THEN 没有患病

③IF “体重范围”=“较低” THEN 没有患病

④IF “体重范围”=“中” THEN 没有患病

⑤IF “体重范围”=“较高” THEN 患病

⑥IF “体重范围”=“高” AND “有无家族病史”=“有” THEN 患病

⑦IF “体重范围”=“高” AND “有无家族病史”=“无” AND “吸烟史”=“无” THEN 没有患病

⑧IF “体重范围”=“高” AND “有无家族病史”=“无” AND “吸烟史”=“5年以上” THEN 患病

决策树算法过程

（1）构造决策树

决策树构造过程如下。

①输入数据，主要包括训练集的特征和类标号。

②选取一个属性作为根节点的分裂属性进行分裂。

③对于分裂的每个分支，如果已经属于同一类就不再分了，如果不是同一类，依次选取不同的特征作为分裂属性进行分裂，同时删除已经选过的分列属性。

④不断的重复③，直到到达叶子节点，也就是决策树的最后一层，这时这个节点下的数据都是一类了。

⑤最后得到每个叶子节点对应的类标签以及到达这个叶子节点的路径。

（2）决策树的预测

得到由训练数据构造的决策树以后就可以进行预测了，当待预测的数据输入决策树的时候，根据分裂属性以及分裂规则进行分裂，最后即可确定所属的类别。

ID3

ID3算法的构建方法和决策树的构建基本是一致的，不同的是分裂节点的特征选择的标准。该算法在分裂节点处将信息增益作为分裂准则进行特征选择，递归的构建决策树。

ID3算法的步骤如下：

① 输入数据，主要包括训练集的特征和类标号。

② 如果 所有实例都属于一个类别，则决策树是一个单结点树，否则执行③。

③ 计算训练数据中每个特征的信息增益。

④ 从根节点开始选择最大信息增益的特征进行分裂。依次类推，从上向下构建决策树，每次选择具有最大信息增益的特征进行分裂，选过的特征后面就不能继续进行选择使用了。

⑤ 不断的构建决策树，至没有特征可以选择或者分裂后的所有元组属于同一类别时候停止构建。

⑥ 决策树构建完成，进行预测。

ID3算法的问题

1.倾向于选择数值较多的特征

例如考虑唯一标识符的属性*product_ID*，因其每个值只有一个元组，在*product_ID*的划分会产生元组总数个分区，且每个分区都是纯的，即该分区的元组属于同一个类，基于该划分对数据集*D*分类的*Info(D|product_ID)*=0，则*g(D|product_ID)*最大，选择*product_ID*作为分裂属性不是一个好的策略，属性取值最多的属性并不一定最优。

再如例4中使用ID3算法构造的决策树第一个分裂属性选择的是“体重范围”，很大程度是因为其数值较多，但在实际情况中，“体重范围”不一定是决定性属性。

2.对于每个属性均为离散值属性，如果是连续值属性需先离散化。