第四章

王照国

2019 年 4 月 1 日

1 熵和信息增益

1.1 样本集合D的信息熵定义为：

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

熵的值越小表示数据D越纯

假设离散属性a有V个可能的属性值a_1, a_2, a_3, a_v，如果使用a属性来对数据集D进行划分，产生V个分支节点

1.2 信息增益：

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

一般而言，信息增益越大，意味着利用属性a来进行划分得到的“纯度提升越大”,ID3决策树算法就是根据信息增益来计算的,但是信息增益对于可能取值较多的属性有所偏好

1.3 增益率：

$$Gain\_ratio(D, a) = \frac{Gain(D, a)}{IV(a)}$$

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{D} log_2 \frac{|D^v|}{|D|}$$

采用增益率的方法消除了采用信息增益对于多类别的偏好，属性a的类别越多，IV值就会越高，但是相应的，增益率对于类别较少的属性有偏好，所有在实际应用中，C4.5算法不是直接选择增益率最大的元素，而是首先选出信息增益在平均水平之上的属性，然后在其中选择一个增益率最高的属性

1.4 基尼系数

$$Gini(D) = 1 - \sum_{k=1}^{|y|} p_k^2$$

$$Gini\_index(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v)$$

采用基尼系数的CART决策树，在候选属性集中选择使得基尼系数最小的属性进行划分

2 剪枝

预剪枝是指在决策树生成过程中，对每个结点在划分之前先进行估计，若当前结点的划分不能带来决策树泛化性能的提升，则停止划分并将当前结点标记为叶节点

后剪枝是从训练集中生成一棵完整的决策树，然后自底向上地对非叶节点进行考察，若将该叶节点对应的子树替换为叶节点能带来决策树泛化性能的提升，则将该子树替换为叶节点

3 缺失值处理

现实任务中经常出现不完整样本，即某些样本的某些属性缺失，这时，我们有两个问题需要考虑：

1.如何在属性值缺失的情况下进行划分属性选择？

2.给定划分属性，如果样本在当前属性上相应的值缺失，如何对样本进行划分？

$$Gain(D,a)=\rho\times Gain(\tilde{D},a)=\rho\times(Ent(\tilde{D})-\sum_{v=1}^{V}\tilde{r_v}Ent(\tilde{D^v}))$$

问题1的解决方法：其中，ρ表示对于属性a无缺失样本所占的比例，$\tilde{D}$表示对属性a没有缺失的样本数目，$\tilde{r_v}$表示无缺失样本中在属性a上取值a^v的样本所占的比例

问题2的解决方法：首先我们将所有的样本分配一个权值w_x，如果样本x在划分属性a上的取值已知，则将x划入与其值对应的子节点，且样本权值在子节点中保持为w_x。如果样本x在划分属性a上的取值未知，则将x划入到所有子节点，并且样本权值在与属性值a^v对应的子节点中调整为$\tilde{r_v}*w_x$