姓名和学号：李胜志 2210180232

章节名称：ID3算法

知识目标：掌握ID3算法的基本原理和流程。

能力目标：能够使用ID3算法构建决策树。

素质目标：养成分析问题、事前规划的良好习惯。

知识重点：使用ID3算法构建决策树的流程。

知识难点：信息熵和条件熵的计算。

ID3算法：

ID3算法的基本思想是：利用信息增益来选择最优的属性作为决策树的节点，算法通过计算每个属性的信息增益，选择信息增益最大的属性作为节点，从而逐步构建决策树。

信息增益是衡量一个属性区分数据的能力，即得知特征属性A的信息而使得类别属性Y的取值不确定性减少的程度，可表示为：

其中为属性类别的无条件熵，为已知特征属性的值后类别属性的无条件熵。表示已知特征属性的值后导致类别属性熵的减小值（即类别属性的取值不确定性减少的程度），的值越大，说明特征属性提供的信息越多。

熵：

熵表示随机变量的不确定性。用符号表示。对于一个离散随机变量，其熵的定义是：

1. 是随机变量取值的概率。
2. 以2为底数时，单位是比特（bit）；当使用e为对数底的时候，单位为奈特（nit）。
3. 求和是对所有可能的取值进行的。

条件熵：

条件熵是指在已知信源发出的信息后，信宿仍然存在的不确定性。在给定的条件下，的条件熵为，公式为：

ID3算法的流程：

1. 确定决策树的根节点：首先，计算给定数据集中类别属性的信息熵，然后，计算数据集中每个特征属性的条件熵，最后，计算各个特征属性对应类别属性的信息增益，选择信息增益最大的特征属性作为决策树（子树）的根节点。
2. 划分数据集：根据选定的划分属性，将数据集划分为若干个子集。
3. 递归构建决策树：重复以上步骤，直至子集包含单一特征属性或节点的样本个数小于预定阈值。

【例6-1】根据给出的数据集（表 1），使用ID3算法构建决策树，判定活动（天气：阴，寒冷，适度高，风速弱）是否能进行活动。

表 1天气情况和是否进行活动的数据集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 天气 | 温度 | 湿度 | 风速 | 活动 |
| 1 | 晴 | 炎热 | 高 | 弱 | 取消 |
| 2 | 晴 | 炎热 | 高 | 强 | 取消 |
| 3 | 阴 | 炎热 | 高 | 弱 | 进行 |
| 4 | 雨 | 适中 | 高 | 弱 | 进行 |
| 5 | 雨 | 寒冷 | 正常 | 弱 | 进行 |
| 6 | 雨 | 寒冷 | 正常 | 强 | 取消 |
| 7 | 阴 | 寒冷 | 正常 | 强 | 进行 |
| 8 | 晴 | 适中 | 高 | 弱 | 取消 |
| 9 | 晴 | 寒冷 | 正常 | 弱 | 进行 |
| 10 | 雨 | 适中 | 正常 | 弱 | 进行 |
| 11 | 晴 | 适中 | 正常 | 强 | 进行 |
| 12 | 阴 | 适中 | 高 | 强 | 进行 |
| 13 | 阴 | 炎热 | 正常 | 弱 | 进行 |
| 14 | 雨 | 适中 | 高 | 强 | 取消 |

使用ID3算法构建决策树并对新样本进行预测的步骤如下：

1. 确定根节点。该数据集共有4个特征属性，因此需分别计算各个特征属性的信息增益：

因此，需分别计算出信息熵，条件熵

、、。

1. 计算出信息熵。在活动列中，类别“取消”出现了5次，“进行”出现了9次，因此取消活动的概率为，进行活动的概率为，信息熵为：
2. 天气列中共有3个属性值：晴、阴和雨，其出现的概率分别为，，。当天气为晴时，进行活动的概率为，取消活动的概率为；当天气为阴时，进行活动的概率为，取消活动的概率为；当天气为雨时，进行活动的概率为，取消活动的概率为。因此：

因此已知天气的情况下，计算类别属性“活动”的条件熵：

1. 同理，可计算出条件熵
2. 根据以上计算的结果可以计算出各个特征属性对类别属性的信息增益：

其中天气的信息增益最大，所以选择天气作为决策树的根节点。如图 1所示：

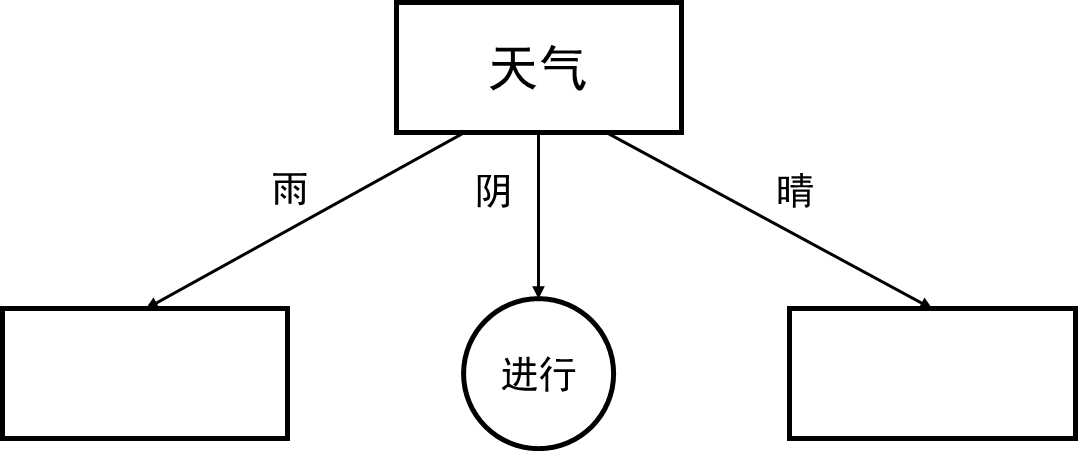


图 1

其中，“天气=阴”条件下所有样本的类别均为一个类别“进行”。因此，只需确定“天气=晴”和“天气=雨”分支的根节点即可。

1. 确定“天气=雨”分支的根节点。找到“天气=雨”条件下的所有样本，如表 2所示。

表 2“天气=雨”条件下的样本

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 天气 | 温度 | 湿度 | 风速 | 活动 |
| 4 | 雨 | 适中 | 高 | 弱 | 进行 |
| 5 | 雨 | 寒冷 | 正常 | 弱 | 进行 |
| 6 | 雨 | 寒冷 | 正常 | 强 | 取消 |
| 10 | 雨 | 适中 | 正常 | 弱 | 进行 |
| 14 | 雨 | 适中 | 高 | 强 | 取消 |

根据表 2的数据确定本分支的根节点，步骤与(1)的步骤相同，需分别计算温度、湿度和风速3个特征属性的信息增益：

其中“风速”的信息增益最大，所以选择“风速”作为“天气=雨”分支节点的根节点。

1. 同理，确定“天气=晴”分支的根节点，步骤和(2)相同，找到“天气=晴”条件下的所有样本，如表 3所示。

表 3“天气=晴”条件下的样本

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 天气 | 温度 | 湿度 | 风速 | 活动 |
| 1 | 晴 | 炎热 | 高 | 弱 | 取消 |
| 2 | 晴 | 炎热 | 高 | 强 | 取消 |
| 8 | 晴 | 适中 | 高 | 弱 | 取消 |
| 9 | 晴 | 寒冷 | 正常 | 弱 | 进行 |
| 11 | 晴 | 适中 | 正常 | 强 | 进行 |

分别计算温度、湿度和风速3个特征属性的信息增益。

通过计算可知，“湿度”的信息增益最大，因此选择“湿度”作为“天气=晴”分支节点的根节点。

此时，每个叶子节点中所有的样本都属于同一个类别，ID3算法结束，分类完成，最终可得到以下决策树图 2：

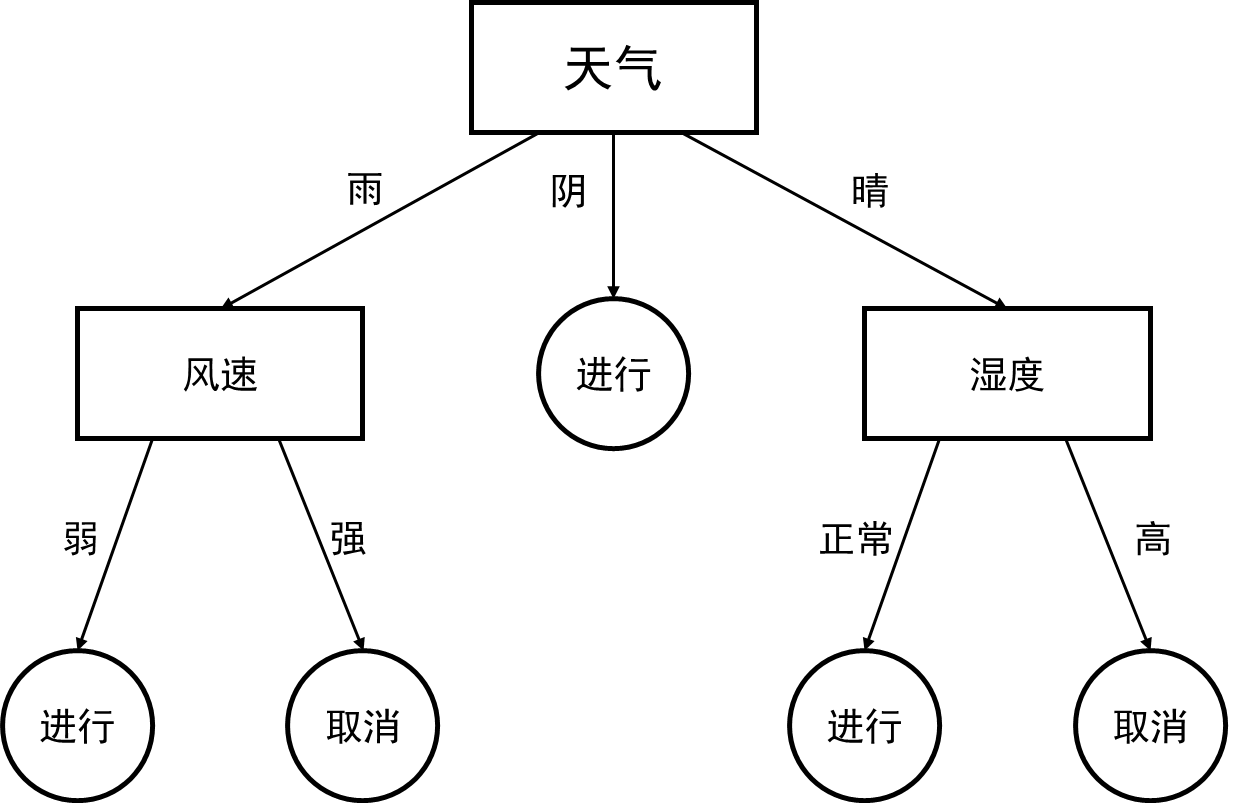


图 2最终决策树

1. 根据决策树写出决策规则。从决策树的根节点到叶子节点的每条路径对应一条合取规则：
2. 使用决策规则对新样本进行预测。对于新样本（阴，寒冷，湿度高，风速弱），应划分为“进行”类别。

ID3算法的优缺点：

优点：

1. 易于理解和解释。ID3算法生成的决策树结构简单直观，便于理解和使用。
2. 处理离散数据。算法能够很好地处理具有离散属性的数据集。
3. 多分类问题。能够处理具有多个类别的分类问题。

缺点：

1. 信息增益的计算倾向于选择具有较多值的属性。ID3算法在选择属性时可能会偏向于那些具有更多值的属性，这可能导致生成的决策树不够准确。
2. 可能产生过度拟合。由于ID3算法会尽可能地将数据集划分得非常细致，这可能导致决策树对训练数据的过度拟合，从而在新的数据上表现不佳。
3. 抗噪性差。ID3算法对噪声数据和异常值比较敏感，这可能会影响决策树的质量。
4. 只能用于离散值属性，不能直接用于连续值属性。
5. 计算效率低。在处理大型数据集时，ID3算法可能会因为递归计算所有可能的分割点而变得效率低下。