姓名和学号：李胜志 2210180232

章节名称：C4.5算法

知识目标：掌握C4.5算法的基本原理和流程。

能力目标：能够使用C4.5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C4.5算法构建决策树的流程。

知识难点：C4.5在ID3算法的的改进。

C4.5算法：

C4.5算法是在ID3算法的基础上开发的，用于生成决策树。主要用于监督学习，即给定一个数据集，每个数据元组由一组属性值描述，并属于某个互斥类别。C4.5的目标是通过学习找到从属性值到类别的映射关系，这个映射可以用来对未知类别的实体进行分类。

C4.5算法使用信息熵的概念来构建决策树，类似于ID3算法，C4.5算法的改进是用信息增益率取代信息增益作为衡量特征属性的标准，对决策树进行剪枝以避免过拟合，以及处理非离散数据和缺失数据的能力。

C4.5算法的核心思想是使用决策树进行分类。C4.5算法通过递归地选择最优划分属性来构建决策树，直到满足停止条件（如所有样本属于同一类别、样本集合为空等）。

在C4.5算法中，选择最优划分属性的标准是信息增益率。与ID3算法使用的信息增益不同，信息增益率通过引入“分裂信息”(split\_information)来规范化信息增益，从而克服了信息增益对具有大量值属性的偏好问题。其数学表达式为：

其中，S表示训练样本集，A表示特征属性，表示信息增益，表示分裂信息，分裂信息的数学表达式为：

其中，S表示训练样本集，表示含有第i个属性值的样本集。

例如：使用C4.5算法对【例6-1】进行决策树分类。

使用C4.5算法构建决策树并对新样本进行预测的步骤如下：

1. 分别计算各个特征属性的信息增益，计算方法于ID3算法相同。
2. 计算各个特征属性的分裂信息。特征属性“天气”中属性值晴、阴和雨的概率分别为，则“天气”的分裂信息和信息增益率为：

同理可得：

1. 计算各个特征属性的信息增益率

可见，“天气”的信息增益率最大，因此选择“天气”作为决策树的根节点。信息增益率越高，表示使用该属性进行划分所获得的纯度提升越大，因此被选为最优划分属性。

“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. 分别计算“天气=晴”和“天气=雨”两个分支子数据集中各个特征属性的信息增益率，步骤与(1)(2)(3)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率：
2. 计算“天气=晴”条件下的信息增益率。

表 1“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 |

可见，“湿度”的信息增益率最大，因此选择“湿度”作为决策树的根节点。

1. 计算“天气=雨”条件下的信息增益率。

可见，“风速”的信息增益率最大，因此选择“湿度”作为决策树的根节点。此时，每个叶子节点中所有的样本都属于同一个类别，C4.5算法结束。

1. 构建最终决策树。如图 1所示。

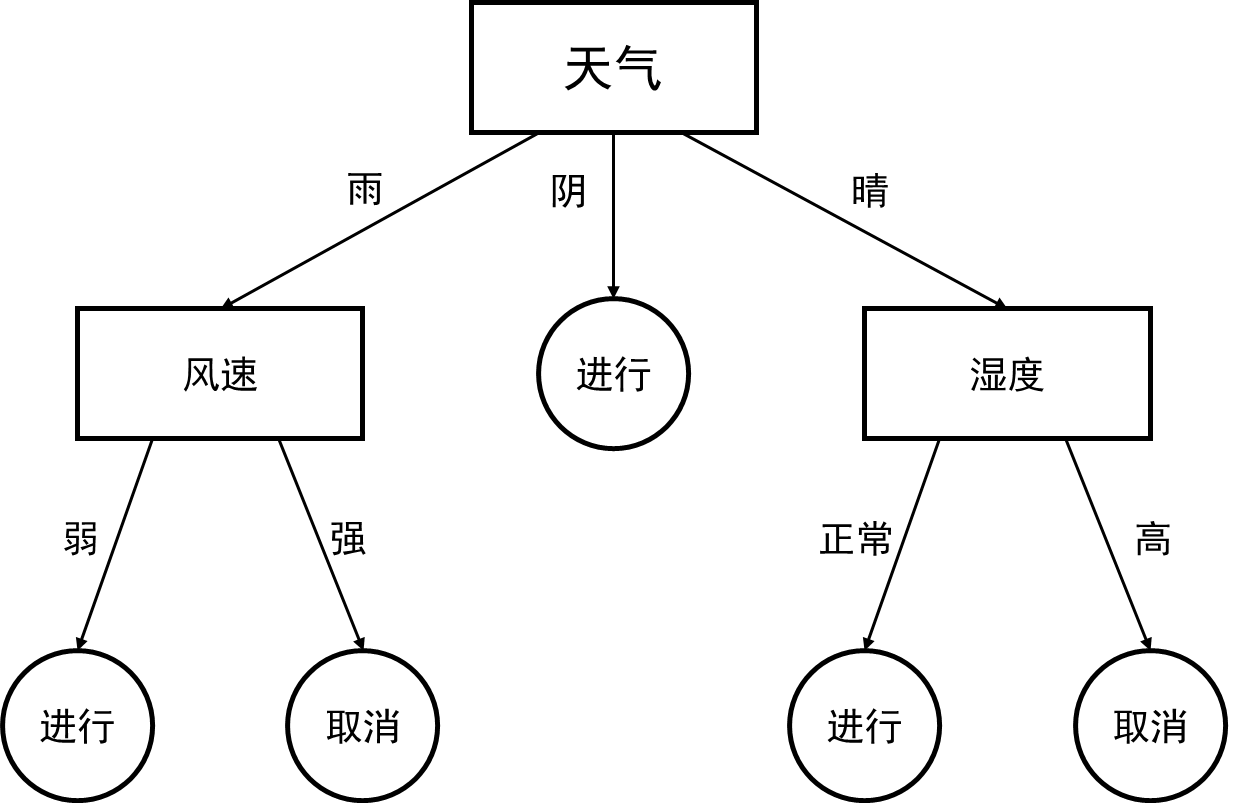


图 1最终决策树

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：

改进与优势

相较于ID3算法，C4.5算法在一下两个方面进行了改进：

1. C4.5算法可以避免偏向取值较多的特征属性。
2. C4.5算法能够处理连续性数值数据。

缺点：

1. 算法计算效率低。
2. 容易过拟合。