第一次实验

SA20218084 林睿江

一、实验内容:

编写 ID3 算法进行决策树模型的构造，然后根据模型预测应用的数据。要求使用鸢尾花数据集。

二、实验原理:

分类技术是一种根据输入数据集建立分类模型的系统方法。一般是用一种学习算法确定分类模型，该模型模型不仅要很好拟合输入数据，还要能够正确地预测未知样本的类别。

决策树分类方法是常用的分类方法之一，首先对数据进行处理，利用归纳算法生成可读的规则和决策树，然后使用决策对新数据进行分析。本质上决策树是通过系列规则对数据进行分类的过程。

决策树的基本组成部分：决策结点、分支和叶子。决策树中最上面的结点称为根结点。是整个决策树的开始。每个分支是一个新的决策结点，或者是树的叶子。每个决策结点代表一个问题或者决策。通常对应待分类对象的属性。每个叶结点代表一种可能的分类结果。

在沿着决策树从上到下的遍历过程中，在每个结点都有一个测试。对每个结点上问题的不同测试输出导致不同的分枝，最后会达到一个叶子结点。这一过程就是利用决策树进行分类的过程，利用若干个变量来判断属性的类别。

ID3 算法使用信息增益度选择测试属性。

$$I(a_1,a_2,...,a_n)=\sum_{i=1}^{n}I(a_i)=\sum_{i=1}^{n}p(a_i)\log_2\frac{1}{p(a_i)} \qquad \text{公式1}$$

$$\text{Entropy（S，A）}=\sum\text{（|Sv|/|S|）}*\text{Entropy（Sv）} \quad \text{公式2}$$

$$\text{Gain（S，A）}=\text{Entropy（S）}-\text{Entropy（S，A）} \quad \text{公式3}$$

由这三个公式得到每个属性的信息增益值 Gain（S，A），值越大，说明选择测试属性对分类提供的信息越多。则该属性被加入决策树中。

三、实验步骤

1.先对实验训练数据进行预处理，将 180 列数据转换成对应的 60 个碱基。

即为 getData()函数。

2.再将训练数据集运用递归的方法建立决策树。

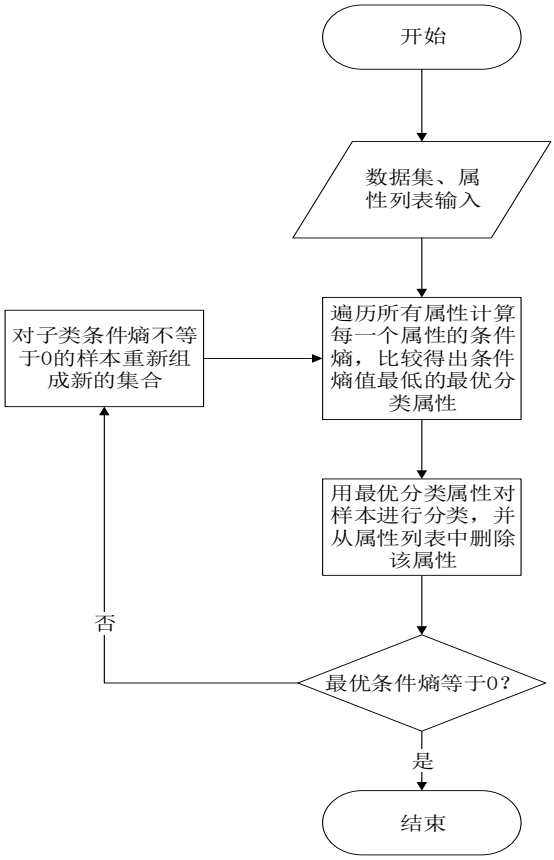
即调用 id3Tree()函数。

3.将所得到的决策树保存，并绘图。

即调用 treeplotter 文件的 createPlot()函数

4.最后调用 Classify()函数，依赖刚刚所得到的决策树来对验证数据集进行分类，将所得分类结果与真实数据结果进行对比，统计分类正确的数值，得出最终准确率。

实验流程图如下：



四、实验结果截图:

```
D:\Users\txt\Desktop\机器学习\作业\SA20218136+仝雪婷\第一次作业>python id3.py
决策树生成
1186
1056
89.04%
```

实验截图

程序根据此实验测试用例，构建决策树模型，输出决策树模型（以前序遍历的顺序）。

五、实验结果分析:

上述的决策树算法增长树的每一个分支的深度，直到恰好能对训练样例比较完美地分类。实际应用中，当数据中有噪声或训练样例的数量太少以至于不能产生目标函数的有代表性的采样时，该策略可能会遇到困难。在以上情况发生时，这个简单的算法产生的树会过渡拟合训练样例（过渡拟合：Over Fitting）.

训练过程应该包含训练样本和验证样本。验证样本用于测试训练后的性能。如果验证结果差，则需要考虑采用不同的结构重新进行训练，例如使用更大的样本集，或者改变从连续值到离散值得数据转换等。

模型过渡拟合的潜在因素：（1）噪声导致的过渡拟合；错误的类别值/类标签，属性值等（2）缺乏代表性样本所导致的过渡拟合：根据少量训练记录作出的分类决策模型容易受过渡拟合的影响。由于训练样本缺乏代表性的样本，在没有多少训练记录的情况下，学习算法仍然继续细化模型就会导致过渡拟合。

解决过度拟合的手段：

1. 及早停止树增长；Occan 法则:具有相同泛化误差的两个模型，较简单的模型比复杂的模型更可取。
2. 后修剪法。