

**模式识别大作业**

题 目 决策树的算法原理与分析

学 院 信息科学与工程

专 业 信息与通信工程

组 员 李雪晴

指导教师 赵海涛

**完成日期： 2017 年 10 月16日**

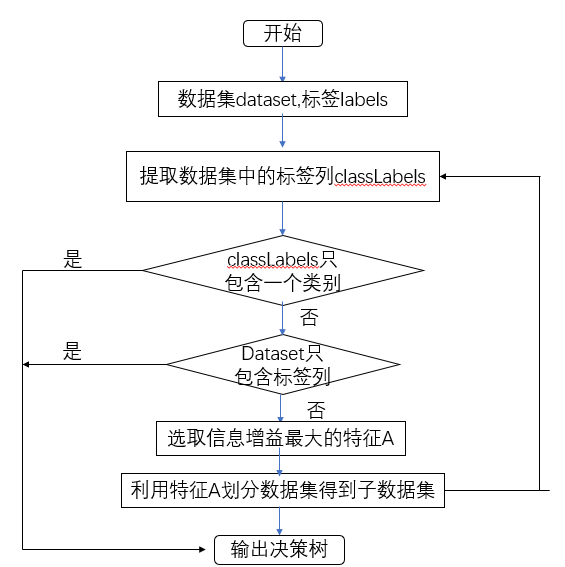
模式识别作业课程报告—决策树的算法原理与分析

组员：李雪晴

初次接触模式识别这门课，其实有点不太适应其繁杂冗长的推导过程，但是经过赵海涛老师的耐心讲解与辛勤指导下，我对该课程有了一定的了解。本次作业我选择了决策树的算法原理与分析，并通过python编程实现来巩固课堂所学的内容。下面我将从以下几个方面一一阐述。

1. 决策树的介绍
2. 决策树的算法分析
3. 测试与总结
4. **决策树的介绍**

决策树（Decision Tree,DT）是一种比较常用的分类方法，属于监督学习中的一种，适用于解决各种分类问题。决策树的原理大致为，首先构建了样本的特征属性及其类别的映射关系，然后依据已经构建的映射关系来形成预测模型，最后便可通过训练完毕的预测模型来对待分类集进行分类。其流程图如下图所示。



决策树模型是一种用于对实例进行分类的树形结构，由节点和有向边组成，结点分为内部节点和叶子结点，每一个节点都代表一个对象，每一个分支都代表一个属性值。每一个叶子结点代表最后的分类类别，每个叶子结点具有从根节点到该叶子结点所经过的路径的所有属性值。常用的算法有CART（Classification And Regressin Tree）,ID3,C4.5,随机森林等。本文采用的是ID3算法。

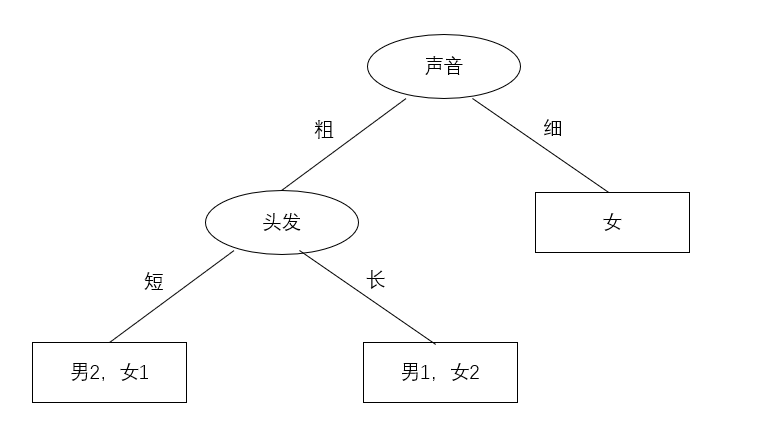
ID3是一种经典的决策树算法，从根节点开始，对该根节点所赋予的属性的每个取值都生成相应的分支，又在每个分支上生成新的节点。ID3采用基于信息熵定义的信息增益来选择内节点的测试属性。

1. **本例中的决策树的算法分析**

首先介绍一下本文所选的数据集：根据头发和声音判断一个同学的性别。下表是简单统计的七位同学的相关特征数据：

| **头发** | **声音** | **性别** |
| --- | --- | --- |
| 长 | 粗 | 男 |
| 短 | 粗 | 男 |
| 短 | 粗 | 男 |
| 长 | 细 | 女 |
| 短 | 细 | 女 |
| 短 | 粗 | 女 |
| 长 | 粗 | 女 |
| 长 | 粗 | 女 |

根据上述数据可以画出一个简单的数据集的决策树，如下图所示，首先判断声音，声音细是女生，声音粗是男生，然后判断头发，头发短是男生，头发长是女生。



划分数据集的工作的原则本质上都是将无序的数据变得更加有序，对于按不同特征分类后的数据的复杂度，若按某一特征分类后复杂度减少的多，那么该特征极为最佳分类特征。

代码如下：



对于分类的原则，Claude Shannon给出了两个定义：信息熵（entropy）和信息增益（Information Gain）。

在信息论中，熵是表示随机变量不确定性的度量，刻画了任意样本集的属性。熵的取值越大，随机变量的不确定性也越大。

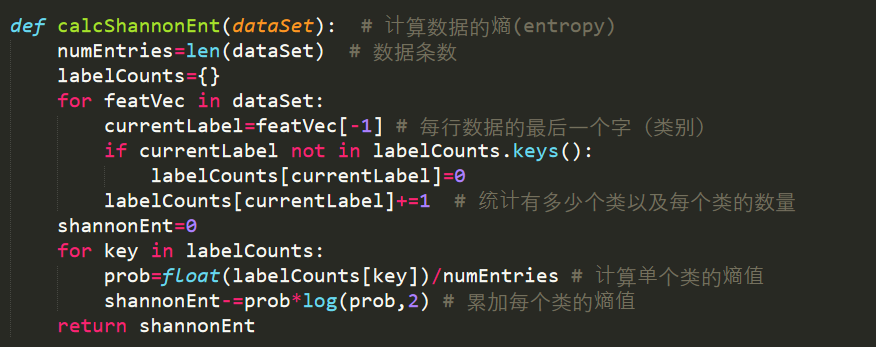
设S是n个数据样本的集合，将样本集划分为c个不同的类,每个类含有的样本数目为，则S划分c个类的信息熵或期望信息，有

其中为S中的样本属于第i类的概率，即。

假设属性A的所有不同值的集合为,是S中属性A的值为v的样本子集，即，在选择属性A后的每一个分支节点上，对该节点的样本集分类的熵为。选择A导致的期望熵定义为每个子集的熵的加权和，权值为属于的样本占原始样本S的比例，即期望熵为，其中是将中的样本划分到c个类的信息熵。属性A相对样本集合S的信息增益定义为

其中是指因知道属性A的值后导致的熵的期望压缩，越大，说明选择测试属性A对分类提供的信息越多，ID3算法就是将每个节点选择信息增益最大的属性作为测试属性。

用python代码实现计算上述例题的熵：

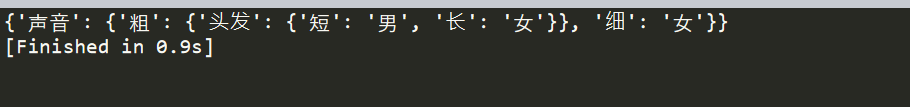


1. **测试与总结**
2. **测试**

输入数据集



其输出结果为;



1. **总结**

由于初次做有关决策树的问题，加之自己理解不够深刻，选择了一个简单的实例。我会在接下来的时间里，认真研究其深层内涵，做出一个比较有含量的题目，感谢老师平时的指导。