实验四 决策树

**一、实验目的**

编程实现决策树算法C4.5；理解算法原理。

**二、实验原理**

**（1）ID3算法**

ID3算法的核心思想就是以信息增益度量属性选择，选择分裂后信息增益最大的属性进行分裂。下面先定义几个要用到的概念。设D为用类别对训练元组进行的划分，则D的[熵](http://en.wikipedia.org/wiki/Entropy)（entropy）表示为：



其中pi表示第i个类别在整个训练元组中出现的概率，可以用属于此类别元素的数量除以训练元组元素总数量作为估计。熵的实际意义表示是D中元组的类标号所需要的平均信息量。现在我们假设将训练元组D按属性A进行划分，则A对D划分的期望信息为：



      而信息增益即为两者的差值：

http://latex.codecogs.com/gif.latex?gain(A)=info(D)-info_A(D)

ID3算法就是在每次需要分裂时，计算每个属性的增益率，然后选择增益率最大的属性进行分裂。

对于特征属性为连续值，可以如此使用ID3算法：先将D中元素按照特征属性排序，则每两个相邻元素的中间点可以看做潜在分裂点，从第一个潜在分裂点开始，分裂D并计算两个集合的期望信息，具有最小期望信息的点称为这个属性的最佳分裂点，其信息期望作为此属性的信息期望。

**（2）C4.5算法**

    ID3算法存在一个问题，就是偏向于多值属性，例如，如果存在唯一标识属性ID，则ID3会选择它作为分裂属性，这样虽然使得划分充分纯净，但这种划分对分类几乎毫无用处。ID3的后继算法C4.5使用[增益率](http://en.wikipedia.org/wiki/Information_gain_ratio)（gain ratio）的信息增益扩充，试图克服这个偏倚。

      C4.5算法首先定义了“分裂信息”，其定义可以表示成：



      其中各符号意义与ID3算法相同，然后，增益率被定义为：



      C4.5选择具有最大增益率的属性作为分裂属性，其具体操作与ID3类似。

**三、实验内容**

利用traindata.txt的数据（75\*5，第5列为标签）进行训练，构造决策树；利用构造好的决策树对testdata.txt的数据进行分类，并输出分类准确率。

**四、实验要求**

（1）提交源代码及可执行文件。

（2）提交实验报告，内容包括：对代码的简单说明、运行结果的截图及说明等。

（3）需画出决策树，指明每个分支所对应的特征/属性，以及分裂值。

**提交时间：**11月12日。

**提交方式：**

源代码、可执行文件及实验报告打包（姓名+学号+实验三）发送给班长，收齐后交给实验老师。