第四次作业- 分类 实验报告

1. 对使用的方法的理解

我选择的分类方法是决策树，weka 中的决策树生成使用 C4.5 算法。

决策树是一种简单但是常用的分类和回归模型。在本次实验中使用决策树完成分类问题。

决策树根据训练数据构建一个 if-then 规则的集合，也可以看做是对样本特征空间的划分。决策树的建立过程包括特征选择，决策树的生成，决策树的剪枝三个阶段。

决策树模型建立的三个阶段：

1.1 特征选择

决策树的通过信息增益（比）来进行特征选择，假设 A 为特征 X 的属性集，有 K 类。

数据集 D 的信息熵 $H(D) = -\sum_{i=1}^{K}\frac{|C_k|}{|D|}log_2(\frac{|C_k|}{|D|})$

特征 A_i 的信息增益为 $g(D,A_i) = H(D|A_i) - H(D)$

特征选择每次选择当前还未被分到叶子节点的数据集上信息增益最大的特征。

1.2 决策树的生成（C4.5 算法）

输入：训练数据集 D,特征集 A，阈值 ϵ

输出：决策树 T

1. 如果 D 中所有实例属于同一个类 C_k, 那么建立单节点树。
2. 如果 $A = \phi$, 则 T 为单节点数，节点类是 D 中数量最多的类 C_k
3. 否则，计算信息增益比，选择信息增益比最大的特征 A_g
4. 如果 A_g 的信息增益比小于阈值，则置 T 为单节点树，并将实例数量最多的类作为该节点的类。
5. 否则，对于每个 A_g 可能的取值 a_i,以 D_i 为数据集，$A - A_g$ 为特征集递归地建立子树。

C4.5 算法是 ID3 的改进，使用信息增益比来进行特征选择。

1.3 决策树的剪枝

思路类似于正则化，损失函数是经验熵加上一个对树的复杂度的惩罚的正则项。

决策树模型的优点是简单，速度快，缺点是容易过拟合。

2. 数据集处理思路

2.1 数据集简介

数据集是关于车的数据信息，维度为 7，其中最后一个维度是 class label。

2.2 处理思路

1. 处理缺失值，看了一下数据集并没有缺失值，因此不需要这一步。
2. 建立决策树。

3. 实验结果

```
J48 pruned tree
------------------

safety = low: unacc (576.0)
safety = med
|   persons = 2: unacc (192.0)
|   persons = 4
|   |   buying = vhigh
|   |   |   maint = vhigh: unacc (12.0)
|   |   |   maint = high: unacc (12.0)
|   |   |   maint = med
|   |   |   |   lug_boot = small: unacc (4.0)
|   |   |   |   lug_boot = med: unacc (4.0/2.0)
|   |   |   |   lug_boot = big: acc (4.0)
|   |   |   maint = low
|   |   |   |   lug_boot = small: unacc (4.0)
|   |   |   |   lug_boot = med: unacc (4.0/2.0)
|   |   |   |   lug_boot = big: acc (4.0)
|   |   buying = high
|   |   |   lug_boot = small: unacc (16.0)
|   |   |   lug_boot = med
|   |   |   |   doors = 2: unacc (4.0)
|   |   |   |   doors = 3: unacc (4.0)
|   |   |   |   doors = 4: acc (4.0/1.0)
|   |   |   |   doors = 5more: acc (4.0/1.0)
|   |   |   lug_boot = big
|   |   |   |   maint = vhigh: unacc (4.0)
|   |   |   |   maint = high: acc (4.0)
|   |   |   |   maint = med: acc (4.0)
|   |   |   |   maint = low: acc (4.0)
|   |   buying = med
|   |   |   maint = vhigh
|   |   |   |   lug_boot = small: unacc (4.0)
|   |   |   |   lug_boot = med: unacc (4.0/2.0)
|   |   |   |   lug_boot = big: acc (4.0)
|   |   |   maint = high
|   |   |   |   lug_boot = small: unacc (4.0)
|   |   |   |   lug_boot = med: unacc (4.0/2.0)
|   |   |   |   lug_boot = big: acc (4.0)
|   |   |   maint = med: acc (12.0)
|   |   |   maint = low
|   |   |   |   lug_boot = small: acc (4.0)
|   |   |   |   lug_boot = med: acc (4.0/2.0)
|   |   |   |   lug_boot = big: good (4.0)
|   |   buying = low
|   |   |   maint = vhigh
|   |   |   |   lug_boot = small: unacc (4.0)
|   |   |   |   lug_boot = med: unacc (4.0/2.0)
|   |   |   |   lug_boot = big: acc (4.0)
|   |   |   maint = high: acc (12.0)
|   |   |   maint = med
|   |   |   |   lug_boot = small: acc (4.0)
|   |   |   |   lug_boot = med: acc (4.0/2.0)
|   |   |   |   lug_boot = big: good (4.0)
|   |   |   maint = low
|   |   |   |   lug_boot = small: acc (4.0)
|   |   |   |   lug_boot = med: acc (4.0/2.0)
|   |   |   |   lug_boot = big: good (4.0)
|   persons = more
|   |   lug_boot = small
|   |   |   buying = vhigh: unacc (16.0)
|   |   |   buying = high: unacc (16.0)
|   |   |   buying = med
|   |   |   |   maint = vhigh: unacc (4.0)
|   |   |   |   maint = high: unacc (4.0)
|   |   |   |   maint = med: acc (4.0/1.0)
|   |   |   |   maint = low: acc (4.0/1.0)
|   |   |   buying = low
|   |   |   |   maint = vhigh: unacc (4.0)
|   |   |   |   maint = high: acc (4.0/1.0)
|   |   |   |   maint = med: acc (4.0/1.0)
|   |   |   |   maint = low: acc (4.0/1.0)
|   |   lug_boot = med
|   |   |   buying = vhigh
|   |   |   |   maint = vhigh: unacc (4.0)
|   |   |   |   maint = high: unacc (4.0)
|   |   |   |   maint = med: acc (4.0/1.0)
|   |   |   |   maint = low: acc (4.0/1.0)
|   |   |   buying = high
|   |   |   |   maint = vhigh: unacc (4.0)
|   |   |   |   maint = high: acc (4.0/1.0)
|   |   |   |   maint = med: acc (4.0/1.0)
|   |   |   |   maint = low: acc (4.0/1.0)
|   |   |   buying = med: acc (16.0/5.0)
|   |   |   buying = low
|   |   |   |   maint = vhigh: acc (4.0/1.0)
|   |   |   |   maint = high: acc (4.0)
|   |   |   |   maint = med: good (4.0/1.0)
|   |   |   |   maint = low: good (4.0/1.0)
|   |   lug_boot = big
|   |   |   buying = vhigh
|   |   |   |   maint = vhigh: unacc (4.0)
|   |   |   |   maint = high: unacc (4.0)
|   |   |   |   maint = med: acc (4.0)
|   |   |   |   maint = low: acc (4.0)
|   |   |   buying = high
|   |   |   |   maint = vhigh: unacc (4.0)
|   |   |   |   maint = high: acc (4.0)
|   |   |   |   maint = med: acc (4.0)
|   |   |   |   maint = low: acc (4.0)
|   |   |   buying = med
|   |   |   |   maint = vhigh: acc (4.0)
|   |   |   |   maint = high: acc (4.0)
|   |   |   |   maint = med: acc (4.0)
|   |   |   |   maint = low: good (4.0)
|   |   |   buying = low
|   |   |   |   maint = vhigh: acc (4.0)
|   |   |   |   maint = high: acc (4.0)
|   |   |   |   maint = med: good (4.0)
|   |   |   |   maint = low: good (4.0)
safety = high
|   persons = 2: unacc (192.0)
|   persons = 4
|   |   buying = vhigh
|   |   |   maint = vhigh: unacc (12.0)
|   |   |   maint = high: unacc (12.0)
|   |   |   maint = med: acc (12.0)
|   |   |   maint = low: acc (12.0)
|   |   buying = high
|   |   |   maint = vhigh: unacc (12.0)
|   |   |   maint = high: acc (12.0)
|   |   |   maint = med: acc (12.0)
|   |   |   maint = low: acc (12.0)
|   |   buying = med
|   |   |   maint = vhigh: acc (12.0)
|   |   |   maint = high: acc (12.0)
|   |   |   maint = med
|   |   |   |   lug_boot = small: acc (4.0)
|   |   |   |   lug_boot = med: acc (4.0/2.0)
|   |   |   |   lug_boot = big: vgood (4.0)
|   |   |   maint = low
|   |   |   |   lug_boot = small: good (4.0)
|   |   |   |   lug_boot = med: good (4.0/2.0)
|   |   |   |   lug_boot = big: vgood (4.0)
|   |   buying = low
|   |   |   maint = vhigh: acc (12.0)
|   |   |   maint = high
|   |   |   |   lug_boot = small: acc (4.0)
|   |   |   |   lug_boot = med: acc (4.0/2.0)
|   |   |   |   lug_boot = big: vgood (4.0)
|   |   |   maint = med
|   |   |   |   lug_boot = small: good (4.0)
|   |   |   |   lug_boot = med: good (4.0/2.0)
|   |   |   |   lug_boot = big: vgood (4.0)
|   |   |   maint = low
|   |   |   |   lug_boot = small: good (4.0)
|   |   |   |   lug_boot = med: good (4.0/2.0)
|   |   |   |   lug_boot = big: vgood (4.0)
|   persons = more
|   |   buying = vhigh
|   |   |   maint = vhigh: unacc (12.0)
|   |   |   maint = high: unacc (12.0)
|   |   |   maint = med: acc (12.0/1.0)
|   |   |   maint = low: acc (12.0/1.0)
|   |   buying = high
|   |   |   maint = vhigh: unacc (12.0)
|   |   |   maint = high: acc (12.0/1.0)
|   |   |   maint = med: acc (12.0/1.0)
|   |   |   maint = low: acc (12.0/1.0)
|   |   buying = med
|   |   |   maint = vhigh: acc (12.0/1.0)
|   |   |   maint = high: acc (12.0/1.0)
|   |   |   maint = med
|   |   |   |   lug_boot = small: acc (4.0/1.0)
|   |   |   |   lug_boot = med: vgood (4.0/1.0)
|   |   |   |   lug_boot = big: vgood (4.0)
|   |   |   maint = low
|   |   |   |   lug_boot = small: good (4.0/1.0)
|   |   |   |   lug_boot = med: vgood (4.0/1.0)
|   |   |   |   lug_boot = big: vgood (4.0)
|   |   buying = low
|   |   |   maint = vhigh: acc (12.0/1.0)
|   |   |   maint = high
|   |   |   |   lug_boot = small: acc (4.0/1.0)
|   |   |   |   lug_boot = med: vgood (4.0/1.0)
|   |   |   |   lug_boot = big: vgood (4.0)
|   |   |   maint = med
|   |   |   |   lug_boot = small: good (4.0/1.0)
|   |   |   |   lug_boot = med: vgood (4.0/1.0)
|   |   |   |   lug_boot = big: vgood (4.0)
|   |   |   maint = low
|   |   |   |   lug_boot = small: good (4.0/1.0)
|   |   |   |   lug_boot = med: vgood (4.0/1.0)
|   |   |   |   lug_boot = big: vgood (4.0)

Number of Leaves  :     131

Size of the tree :  182
```

4. 参考资料

1. weka 官方文档
2. 《统计学习方法》. 李航