Decision Tree Algorithm 实验报告

2012011307 黄必胜

算法设计：

1、算法概况

实现了 ID3 决策树在给定数据集上的训练和测试。利用信息增益作为属性分类训练数据的能力度量标准。在每一个分类结点上，选取信息增益最大的属性进行展开，建立子结点。直至结点中的数据标签相同或者每一个属性都已经被展开。

2、缺失值处理

训练数据中有可能会有缺失的属性值。对于每一个训练的数据点，若有缺失的属性值，则根据该数据点的类别（＞50k 或≤50K），选择该类别中该缺失属性的最常见值。

3、连续值处理

原来是按照教材中合并连续值属性的办法，但是由于某些属性的阈值过多，要评价候选阈值需要进行的计算量太大，最后采取的策略是：将连续属性 A 排序样例，然后确定目标分类不同的相邻事例，产生一组候选阈值。而后取这组阈值的平均值作为最终阈值。

4、剪枝策略

先进行决策树的初建立，允许过拟合的发生；然后对于每一个结点，考察删除该结点为根的子树后对整体正确率的影响；将正确率最大的结点子树删除，用结点中最常见的类别标签来定义该新叶子结点；重新考虑每一个结点删除后的影响，直至树中结点的删除不能带来正确率的提高。

实验结果：

按照实验要求，进行以下实验：
未剪枝，随机选取 5%，50%，100%作为训练集，
剪枝，随机选取 5%，50%，100%中的 60%作为训练集，其余 40%作为剪枝的验证集
在测试集数据上进行测试，对于每一次实验项目分别进行五次测试，
结果如下，

训练集百分比	最大值	最小值	平均值
5%	78.4%	78.2%	78.3%
5%(post-pruning)	81.0%	80.6%	80.7%
50%	80.9%	80.7%	80.8%
50%(post-prunning)	83.5%	83.0%	83.4%
100%	81.1%	81.0%	81.0%
100%(post-prunning)	83.4%	83.1%	83.2%

实验分析：

1、 训练集的大小

从实验结果中可以看出，训练集的大小会对分类的效果产生一定的影响。总体而言训练集较小时，正确率较低；训练集的增加，会带来正确率的提升；但是 50%和 100%未剪枝的测试结果相差不大，可能是 100%的训练集反而导致了过拟合。

2、 剪枝策略的效果

在不同的样本量上，剪枝都带来了正确率的提升，提升量在 2%左右，因而剪枝是有一定的效果的。剪枝之前，树的结点数在三万左右，而剪枝算法在减掉 30 个结点左右就收敛了。也许进一步地探寻剪枝空间可以带来更大的进步。

3、 剪枝策略的加速

在算法的实现过程中，遇到了许多困难，例如数据的预处理较为繁杂，算法计算时间过长等，其中值得一提的是剪枝算法的实现。

剪枝算法需要考虑删除结点之后对于整体正确率的影响。

一开始我的做法是对每一个结点进行修改之后，都要对验证集的所有数据点进行重新分类，导致计算过程非常地慢。后来我仔细思考，其实对于每一个结点而言，删除该结点之后只会影响那些在分类时经过该结点的数据。所以只要在数据点分类的时候，在经过的每一个结点上进行计数，记录该数据点的分类结果和实际的类别标签。

利用这些信息，再加上总结点数和当前的判断正确数，就可以计算出删除该结点的子树后的正确率变化。这样在对验证集进行一轮分类后，即可直接选出正确率最高的结点，将其删除，使得算法的运行时间进入可以接受的范围。