决策树的实现

周国兵 MF1333062 zgbdsg@gmail.com 15250986966

(南京大学 计算机科学与技术系, 南京 210093)

# 实现细节

（阐述清楚所实现的方法即可，例如如何处理缺失属性，如何寻找分割属性等，此部分不需超过此页）

分类与回归树是由Breiman等在1984年提出，是一种决策树算法，可以同时处理分类问题和回归问题。本次试验的代码实现主要依据CART理论来实现。

CART的生成是通过递归的构建二叉决策树，对回归树使用平方误差最小化准则，对分类树用基尼指数最小化准则，进行特征选择，生成二叉树。

1. 最小二乘回归树生成算法
2. 选择最优切分属性j，以及属性切分点s

遍历变量j，对固定的切分变量j扫描切分点s，选择使上式大道最小值的(j,s)

1. 用选定的（j，s）对数据集进行划分，并决定相应的输出值
2. 继续对左右子空间进行划分
3. 输出决策树
4. 分类树生成算法
5. 对于数据集D，计算对于每一个特征，对于其可能的每个取值，将D分成D1和D2，计算此时的基尼指数。
6. 在所有的可能特征中以及所有可能的切分点a中，选择基尼指数最小的特征及其对应的切分点，生成两个子节点，将数据集依据特征分配到子节点中
7. 生成CART

对于缺失属性，在训练决策树过程中丢弃该属性，在预测过程中忽略该属性的作用，直接使用归属的最近的决策树点的属性作为预测值。分割属性在连续型变量中，先将该属性所有值排序，去相邻值的重点作为可能分割点，n个数据有n-1个分割点，判断规则为与该值大小的比较。离散属性，统计所有的非重复值作为分割点，判断规则为与该值是否相等。

# 结果

## 实验设置

可对叶子节点的元素个数进行控制，小于4个数据点将不再分裂。

## 试验结果

|  |  |  |
| --- | --- | --- |
| Classification | breast-cancer | segment |
| Accuracy(mean±std) | 0.645±0.003 | timeout |
| Regression | housing | meta |
| RMSE(mean±std) | 6.055±1.37 | fault |