决策树的ID3，C4.5，C5.0，CART算法在原理与应用方面的连续与区别：

ID3：

核心思想：采用各个非叶子结点的信息增益来进行特征选择

具体实例如图1：

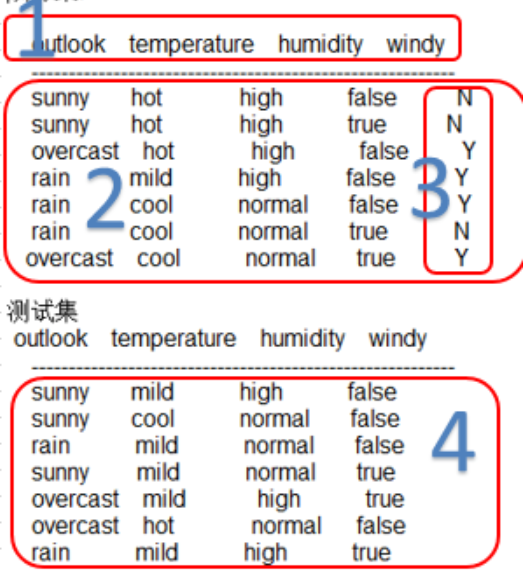


图1 外出天气影响因素与结果

首先计算最终结果result（N,Y）对应的信息熵

H(result) = -P(Y)log2(P(Y)) -P(N)log2(P(N)) P(Y) = number(Y)/number(result)=4/7

如果result有n种时，例如区分颜色（赤橙黄绿青蓝紫....）

H(result) = -P(r1)log2(P(r1)) -P(r2)log2(P(r2)).....-P(rn)log2(P(rn))

P(rn) = number(rn)/number(result)

然后计算每个非叶子结点（特征）的信息增益：

g(result|outlook) = H(result) - H(result|outlook)

如图1所示outlook分为sunny，rain，overcast三个属性，分别是3个，2个，2个

H(result|outlook)具体计算原理如图2所示：

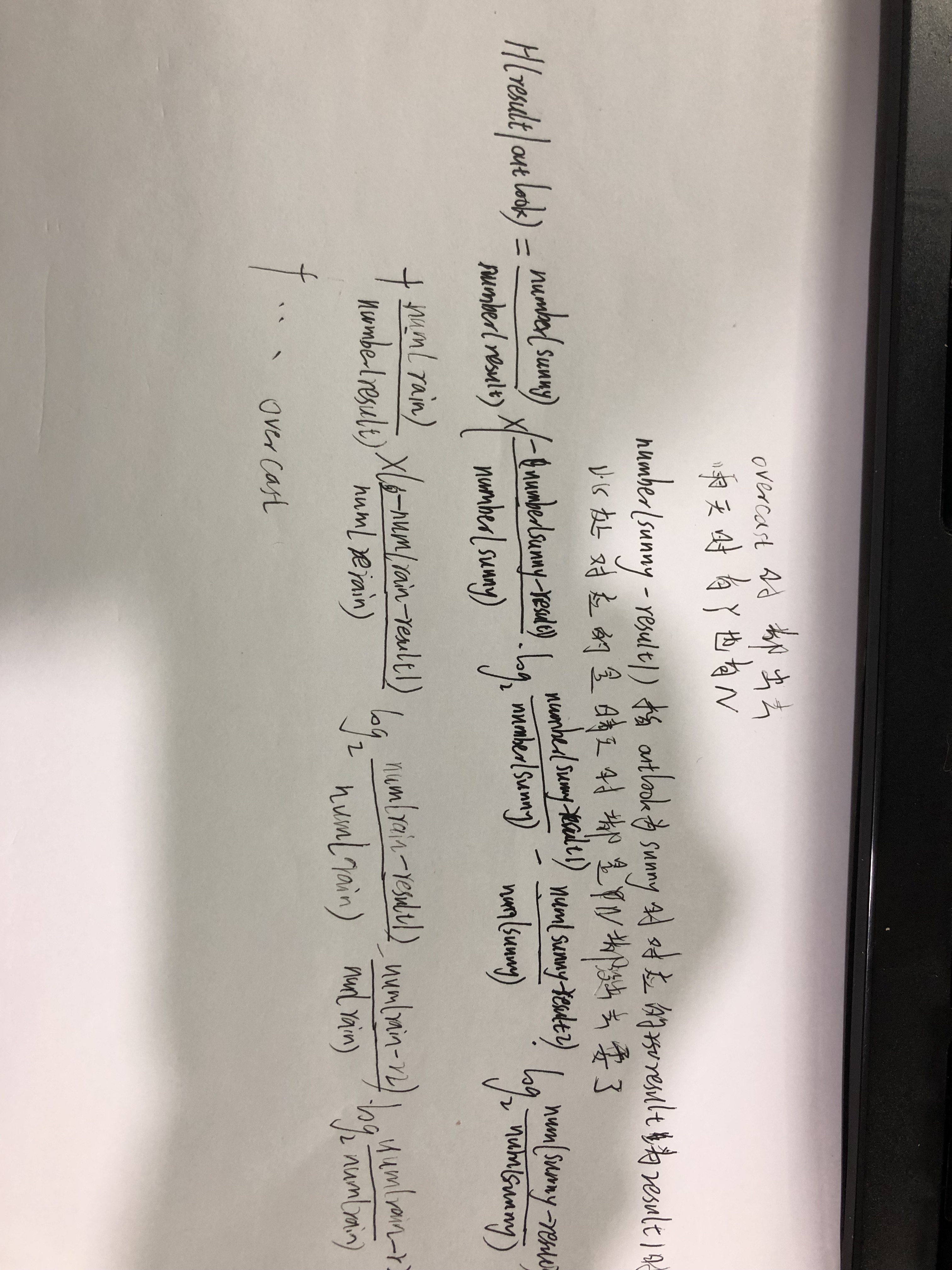


图2 outlook信息增益计算示意

依照此方法依次计算outlook，temperature，humidity，windy特征的信息增益

g(X|A=”outlook”) =0.246  
g(X|A=”temperature”) =0.0289  
g(X|A=” humidity”) =0.1515  
g(X|A=”windy”) =0.048

按照信息增益的计算结果可知，Outlook属性划分获得的信息增益最大，因此我们用这个属性作为我们决策树的根结点，得到下图决策树

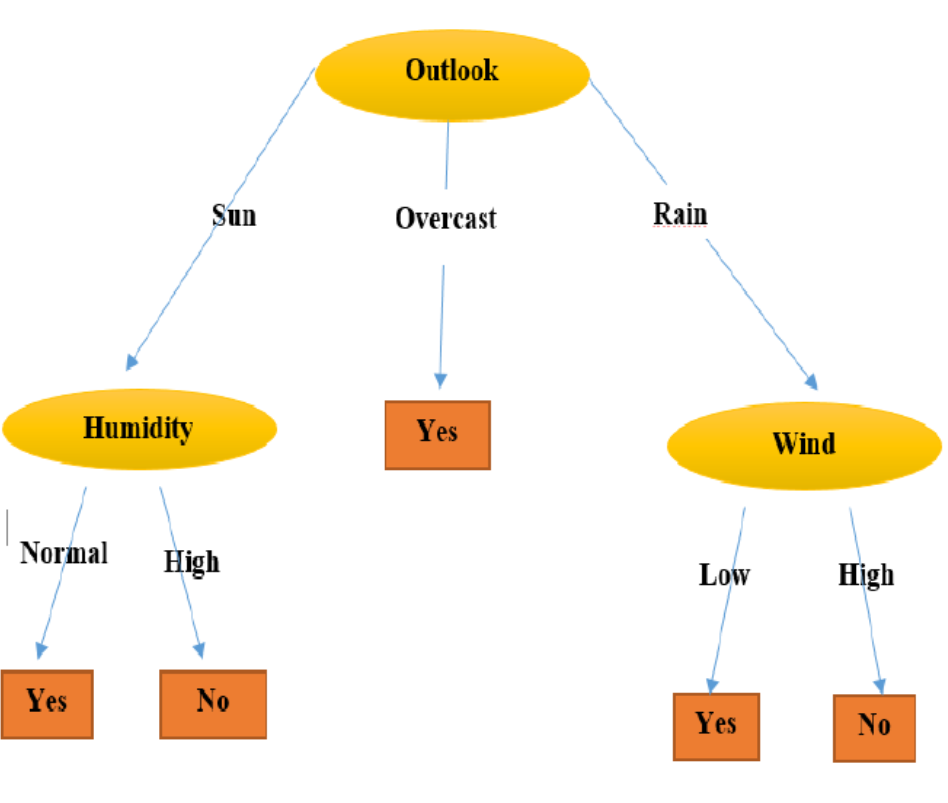


图3

ID3算法有几个缺点：

对于具有很多值的属性它是非常敏感的，例如，如果我们数据集中的某个属性值对不同的样本基本上是不相同的，甚至更极端点，对于每个样本都是唯一的，如果我们用这个属性来划分数据集，它会得到很大的信息增益，但是，这样的结果并不是我们想要的。（也就是说如果如果下雨天出去和不出去的结果都相同，计算出的信息增益固然很大，但是这并不能说明任何问题，反倒下雨没影响）

ID3算法不能处理具有连续值的属性（也就是只能分类，不能回归）

ID3算法不能处理属性具有缺失值的样本（例如，如果outlook中sunny数据丢失，则对应的信息增益无法计算）

由于按照上面的算法会生成很深的树，所有容易产生过拟合现象（只生成树，计算非叶子结点信息增益的过程只是来选择哪个做根节点，没有进行剪枝，所以容易过拟合）

参考博客： https://www.cnblogs.com/chenximcm/p/6285161.html

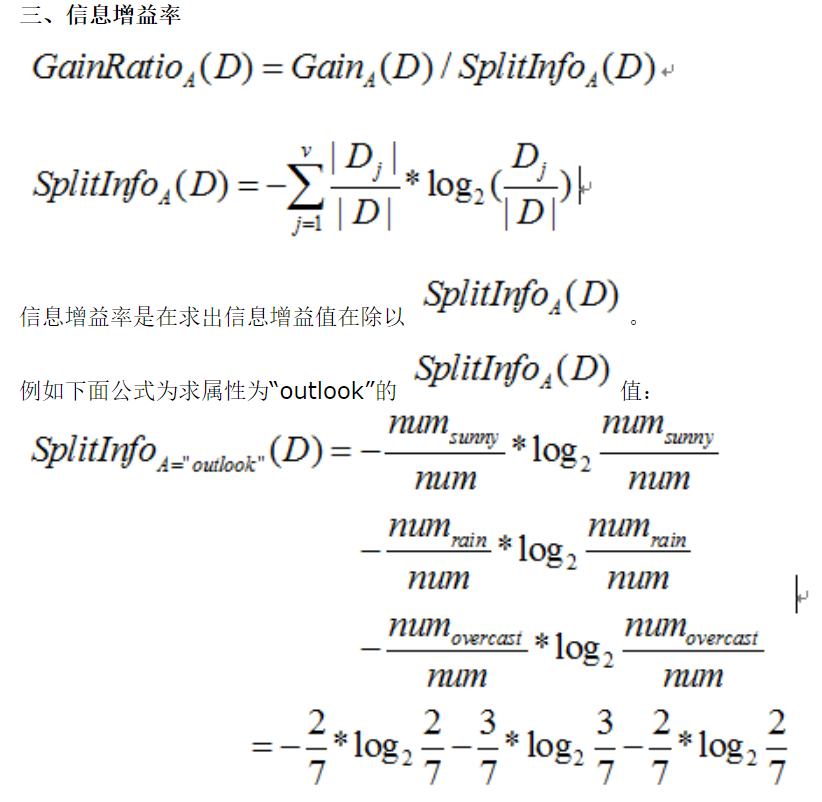
C4.5

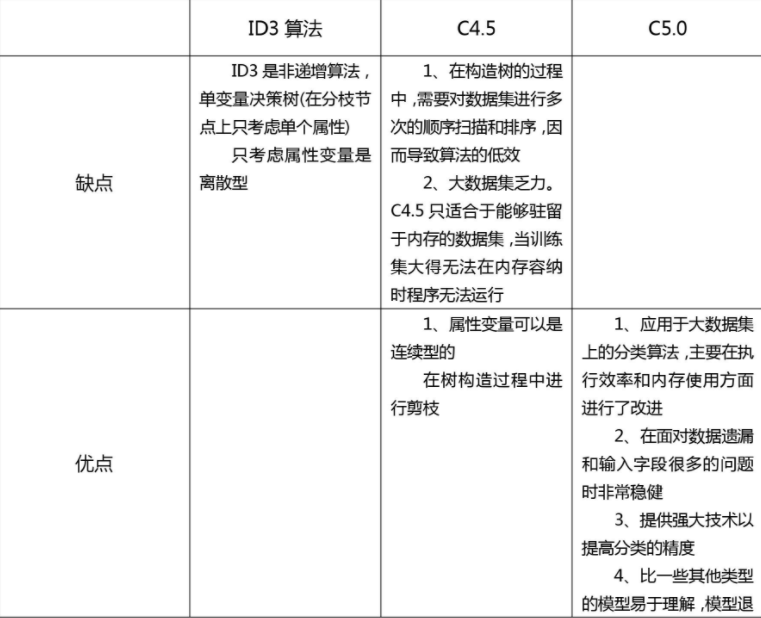
相比于ID3，C4.5选择信息增益率作为选择根节点的标准

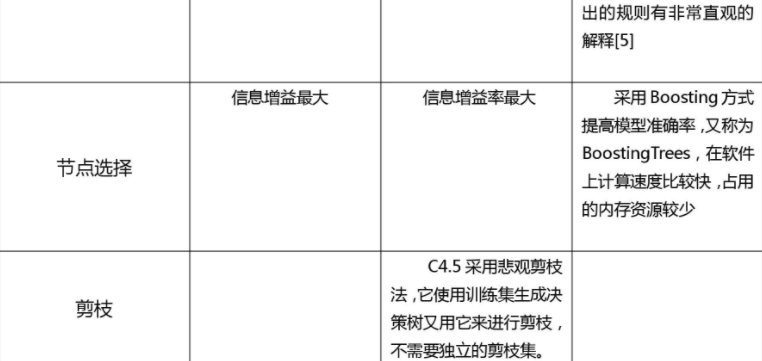
信息增益G(result) = H(result)-H(D|outlook)

信息增益率GR(result) = G(result)/S(outlook)

S(outlook)







参考博客：

https://www.cnblogs.com/chenximcm/p/6285160.html

<https://blog.csdn.net/sinat_26917383/article/details/47617801>

CART：

全称：classification and regression tree

其依据基尼指数进行非叶子结点的选取