戴申：ID3 算法可视化解读

网络上有许多关于 ID3 算法的讲解，但是看来看去总是觉得一头雾水，每到想要关注细节的地方，文章便嘎然而止，一句计算方法同理便把需要关心的内容一带而过，不得不感叹，这个世界聪明的人太多，像我这样笨人太少。说句实话，ID3 的计算不是太复杂，也就是三两的公式，但是涉及递归计算，绕来绕去就把自己给饶糊涂了，由于本人对数字不敏感，但对图形还是可以的，其实大家对图像对很敏感。于是就有了用图形形象地把 ID3 的计算过程记载下来的想法，以便帮助同我一样笨的人学起来不再痛苦。在网上找了些文章，拼凑了一下就产生了本文。

属性选择

ID3 怎样决定那一个属性最好？使用一个叫做信息增益的统计特性。选择增益最大的那一个（信息对分类最有用）。为了定义增益，首先要借助信息理论的一个概念——熵。熵可以测量属性的信息量。

已知有 C 个结果的训练集 S

$$Entropy(S) = \Sigma\ -p(I)\ log2\ p(I)$$ ------（公式 1）

这里 p(I) 是属于类 I 的 S 的比例。Σ 是对 C 求和。Log2 以 2 为底的自然对数。

如果所有 S 属于相同的类，熵为 0（数据分类完毕）。熵的范围是 0（分类完毕）到 1（完全随机）。

注意： S 不但是属性而且也是整个样本集（这一点刚开始可能有点混淆）。

$$Entropy(S, A) = \Sigma\ (|S_v| / |S|) * Entropy(S_v)$$ -----（公式 2）

这里:

Σ 是属性 A 的所有可能的值 v

S_v = 属性 A 有 v 值的 S 子集

$|S_v|$ = S_v 中元素个数

|S| = S 中元素个数

Gain(S, A) 是属性 A 在集 S 上的信息增益，定义为：

$$Gain(S, A) = Entropy(S) - Entropy(S, A)$$ -----（公式 3）

Gain(S, A)是指已知属性 A 的值后导致熵的减少。Gain(S, A)越大，说明选择测试属性 A 对分类提供的信息越多。

ID3 的例子

假设我们希望用 ID3 决定天气是否适合打垒球。在过去的两周中，收集了 14 天的数据帮助 ID3 建立决策树。

目标分类是“我们可以去打垒球吗？”，它有两种选择，可以或不可以。

天气可以用四个属性来刻画，**户外**，**温度**，**湿度**和**风速**。它们的属性值分别为:

户外 = { 晴天, 阴天, 雨天 }

温度 = {炎热, 温柔, 凉爽 }

湿度 = { 高, 正常 }

风速 = {弱, 强 }

根节点的选择：

根节点的选择标准就是看哪一个属性的增益最大。下面是计算四个属性的增益：

我天生懒惰，对一些数来数去的工作感觉不能胜任，头发昏，于是就用软件来代替查数的工作，这样便选择了用 SPSS 完成计算过程的演示工作，繁琐的事情都由机器代劳了，免去我的烦恼。

使用 SPSS 进行计算过程的演示：

首先导入数据：

```
DATA LIST LIST /天数(A8) 户外(A8) 温度(A8) 湿度(A8) 风速(A8) 活动(A8).
BEGIN DATA
D1  晴天    炎热    高    弱    取消
D2  晴天    炎热    高    强    取消
D3  阴天    炎热    高    弱    进行
D4  雨天    温柔    高    弱    进行
D5  雨天    凉爽    正常  弱    进行
D6  雨天    凉爽    正常  强    取消
D7  阴天    凉爽    正常  强    进行
D8  晴天    温柔    高    弱    取消
D9  晴天    凉爽    正常  弱    进行
D10 雨天    温柔    正常  弱    进行
D11 晴天    温柔    正常  强    进行
```

```
D12 阴天    温柔    高    强    进行
D13 阴天    炎热    正常  弱    进行
D14 雨天    温柔    高    强    取消
END DATA.
EXE.
```

第一步 ：计算决策属性的熵

决策属性**活动**有 14 个记录，其中 9 个记录活动可以进行，5 个记录不适合活动，

那么使用公式 1 计算熵。

$$Entropy(活动) = -(9/14)\ Log2\ (9/14) - (5/14)\ Log2\ (5/14) = 0.940$$

声明一下，由于在 SPSS 中没有找到以 2 为底的对数函数，所以这部分计算熵的工作不得不自己动手完成，尽管十分不情愿。

如果直接计算上式，相信多数人不会有太多的印象，要是对熵的概念不是很熟悉，还真有点不好记，还是利用人类适合记忆图像的特点，将计算过程图示化。

按变量**活动**对数据进行排序：

```
SORT CASES BY  活动 .
EXE.
```

观察变量**活动**那一列，9 个适合活动的记录，5 个不适合活动的记录，比较醒目。

天数	户外	温度	湿度	风速	活动
D7	阴天	凉爽	正常	强	进行
D11	晴天	温柔	正常	强	进行
D12	阴天	温柔	高	强	进行
D3	阴天	炎热	高	弱	进行
D4	雨天	温柔	高	弱	进行
D5	雨天	凉爽	正常	弱	进行
D9	晴天	凉爽	正常	弱	进行
D10	雨天	温柔	正常	弱	进行
D13	阴天	炎热	正常	弱	进行
D2	晴天	炎热	高	强	取消
D6	雨天	凉爽	正常	强	取消
D14	雨天	温柔	高	强	取消
D1	晴天	炎热	高	弱	取消
D8	晴天	温柔	高	弱	取消

公式的形象记忆：
用红框的长度（9）除以黑框的长度（14）再乘上这个数的以2为底的对数；
用绿框长度（5）除以黑框长度（14）再乘上这个数的以2为底的对数；
把这两个数相加再取负号。

第二步：计算条件属性的熵

样本集共有四个条件属性，**户外**，**温度**，**湿度**和**风速**。使用公式 2 计算条件属性的熵。

■ **风速**的熵：

计算分两个过程，首先使用公式 1 计算属性值的熵，即风速强和风速弱。

$Entropy(S_{弱}) = -(6/8)*\log2(6/8) - (2/8)*\log2(2/8) = 0.811$

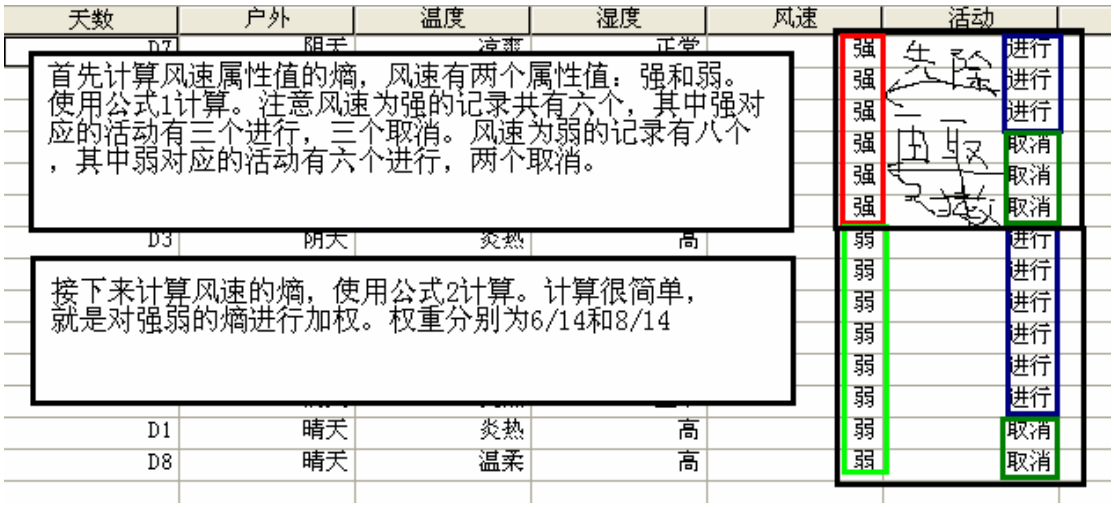
$Entropy(S_{强}) = -(3/6)*\log2(3/6) - (3/6)*\log2(3/6) = 1.00$

然后使用公式 2 计算属性的熵。

$Entropy(S, 风速) = (8/14)*Entropy(S_{弱}) + (6/14)*Entropy(S_{强})$

$= (8/14)*0.811 + (6/14)*1.00$

$= 0.892$



■ **户外**的熵：

户外有三个属性值，晴天，阴天和雨天。其熵分别为：

$Entropy(S_{晴天}) = -(2/5)*\log2(2/5) - (3/5)*\log2(3/5) = 0.971$

$Entropy(S_{阴天}) = -(4/4)*\log2(4/4) = 0$ （熵为 0 表示这一支比较纯，没有分下去的必要）

$Entropy(S_{雨天}) = -(3/5)*\log2(3/5) - (2/5)*\log2(2/5) = 0.971$

户外的熵：

Entropy (S,户外)=(5/14)*Entropy($S_{晴天}$)+(4/14)*Entropy($S_{阴天}$)

+5/14) * Entropy($S_{雨天}$)= (5/14)*0.971 + (4/14)*0 +（5/14）*0.971

= 0.693

为了便于比较，需要将各个变量的位置进行轮换。实现的程序如下：

```
SORT CASES BY 户外 活动.
MATCH FILES FILE=*
 /KEEP=天数 温度 湿度 风速 户外  活动.
EXE.
```

天数	温度	湿度	风速	户外	活动
D11	温柔	正常	强	晴天	进行
D9	凉爽	正常	弱	晴天	进行
D2	炎热	高	强	晴天	取消
D1	炎热	高	弱	晴天	取消
D8	温柔	高	弱	晴天	取消
D7	凉爽	正常	强	阴天	进行
D12	温柔	高	强	阴天	进行
D3	炎热	高	弱	阴天	进行
D13	炎热	正常	弱	阴天	进行
D4	温柔	高	弱	雨天	进行
D5	凉爽	正常	弱	雨天	进行
D10	温柔	正常	弱	雨天	进行
D6	凉爽	正常	强	雨天	取消
D14	温柔	高	强	雨天	取消

■ **温度**的熵：

温度有三个属性值，凉爽，温柔和炎热。它们的熵分别为

Entropy($S_{凉爽}$) = - (3/4)*log2(3/4) - (1/4)*log2(1/4) = 0.811

Entropy($S_{温柔}$) = - (4/6)*log2(4/6) - (2/6)*log2(2/6) = 0.918

Entropy($S_{炎热}$) = - (2/4)*log2(2/4)- (2/4)*log2(2/4) = 1

温度的熵：

Entropy (S,温度)=(4/14)*Entropy($S_{凉爽}$)+(6/14)*Entropy($S_{温柔}$)

+（4/14）* Entropy($S_{炎热}$)

= (4/14)*0.811 + (6/14)*0.918+（4/14）*1

= 0.911

位置轮换的程序为：

```
SORT CASES BY 温度 活动.
MATCH FILES FILE=*
 /KEEP=天数 湿度 风速 户外 温度  活动.
EXE.
```

天数	湿度	风速	户外	温度	活动
D9	正常	弱	晴天	凉爽	进行
D7	正常	强	阴天	凉爽	进行
D5	正常	弱	雨天	凉爽	进行
D6	正常	强	雨天	凉爽	取消
D11	正常	强	晴天	温柔	进行
D12	高	强	阴天	温柔	进行
D4	高	弱	雨天	温柔	进行
D10	正常	弱	雨天	温柔	进行
D8	高	弱	晴天	温柔	取消
D14	高	强	雨天	温柔	取消
D3	高	弱	阴天	炎热	进行
D13	正常	弱	阴天	炎热	进行
D2	高	强	晴天	炎热	取消
D1	高	弱	晴天	炎热	取消

■ **湿度的熵：**

变量**湿度**有两个属性值，湿度正常和湿度高，它们的熵分别为

Entropy($S_{高}$) = - (3/7)*log2(3/7) - (4/7)*log2(4/7) = 0.985

Entropy($S_{正常}$) = - (6/7)*log2(6/7) - (1/7)*log2(1/7) = 0.591

湿度的熵：
Entropy (S,湿度)=(7/14)*Entropy($S_{高}$)+(7/14)*Entropy($S_{正常}$)

= (7/14)*0.985 + (7/14)*0.591

= 0.789

位置轮换的程序为：

```
SORT CASES BY 湿度 活动.
MATCH FILES FILE=*
 /KEEP=天数 风速 户外 温度 湿度  活动.
EXE.
```

天数	风速	户外	温度	湿度	活动
D12	强	阴天	温柔	高	进行
D4	弱	雨天	温柔	高	进行
D3	弱	阴天	炎热	高	进行
D8	弱	晴天	温柔	高	取消
D14	强	雨天	温柔	高	取消
D2	强	晴天	炎热	高	取消
D1	弱	晴天	炎热	高	取消
D9	弱	晴天	凉爽	正常	进行
D7	强	阴天	凉爽	正常	进行
D5	弱	雨天	凉爽	正常	进行
D11	强	晴天	温柔	正常	进行
D10	弱	雨天	温柔	正常	进行
D13	弱	阴天	炎热	正常	进行
D6	强	雨天	凉爽	正常	取消

第三步：计算条件属性的增益

使用公式 3 计算条件属性的增益为：

Gain(S, 户外) = Entropy(活动) - Entropy (S,户外) = 0.94- 0.693 = 0.246

Gain(S, 温度) = Entropy(活动) - Entropy (S,温度) = 0.94- 0.911 = 0.029

Gain(S, 湿度) = Entropy(活动) - Entropy (S,湿度) = 0.94- 0.789 = 0.151

Gain(S, 风速) = Entropy(活动) - Entropy (S,风速) = 0.94- 0.892 = 0.048

条件属性**户外**有最大的增益，所以它用于决策树的根节点。

如果你连数都懒得数，可以使用下列程序直接生成频数表，在里面自己挑选对应的数字。

生成计算熵的频数表程序：

```
CTABLES
 /VLABELS VARIABLES=活动 DISPLAY=none
 /TABLE (户外 + 温度 + 湿度 + 风速) > 活动
 /SLABELS  VISIBLE=NO
 /CLABELS ROWLABELS=OPPOSITE.
```

		进行	取消
户外	晴天	2	3
	阴天	4	0
	雨天	3	2
温度	凉爽	3	1
	温柔	4	2
	炎热	2	2
湿度	高	3	4
	正常	6	1
风速	强	3	3
	弱	6	2

支节点的选择：

因为**户外**有三种类型，根节点就有三个分支（晴天，阴天，雨天）。由于阴天的熵为 0，就不用考虑它了。下面考虑晴天和雨天。

晴天支节点的选择：

接下来的问题是“在晴天支节点处应该检验什么属性？”。 因为已经使用**户外**为根节点，只能用剩余三个变量：温度，湿度或风速。

户外为晴天的记录有 5 个，$S_{晴天}$ = {D1, D2, D8, D9, D11}

第一步:计算户外为晴天的熵，前面已经计算完成，即 Entropy($S_{晴天}$) = 0.970

第二步:计算户外为晴天的条件下各属性的熵

■ 温度的熵

温度有三个属性值，凉爽，温柔和炎热。它们的熵分别为

Entropy($S_{凉爽}$) = - (1/1)*log2(1/1) = 0 （纯洁了）

Entropy($S_{温柔}$) = - (1/2)*log2(1/2) - (1/2)*log2(1/2) = 1

Entropy($S_{炎热}$) = - (2/2)*log2(2/2) = 0 （纯洁了）

温度的熵：

Entropy ($S_{晴天}$, 温度)=(1/5)*Entropy($S_{凉爽}$)+(2/5)*Entropy($S_{温柔}$)

+ (2/5) * Entropy($S_{炎热}$)

= (1/5)*0 + (2/5)*1+（2/5）*0

= 0.4

位置轮换的程序为：

```
SORT CASES BY 户外 温度 活动.
MATCH FILES FILE=*
 /KEEP=天数 风速 湿度 户外 温度  活动.
EXE.
```

天数	风速	湿度	户外	温度	活动
D9	弱	正常	晴天	凉爽	进行
D11	强	正常	晴天	温柔	进行
D8	弱	高	晴天	温柔	取消
D2	强	高	晴天	炎热	取消
D1	弱	高	晴天	炎热	取消
D7	强	正常	阴天	凉爽	进行
D12	强	高	阴天	温柔	进行
D3	弱	高	阴天	炎热	进行
D13	弱	正常	阴天	炎热	进行
D5	弱	正常	雨天	凉爽	进行
D6	强	正常	雨天	凉爽	取消
D4	弱	高	雨天	温柔	进行
D10	弱	正常	雨天	温柔	进行
D14	强	高	雨天	温柔	取消

■ 湿度的熵

湿度有两个属性值，湿度正常和湿度高，它们的熵分别为

Entropy($S_{高}$) = - (3/3)*log2(3/3) = 0

Entropy($S_{正常}$) = - (2/2)*log2(2/2) = 0

湿度的熵：

Entropy ($S_{晴天}$,湿度)=(3/5)*Entropy($S_{高}$)+(2/5)*Entropy($S_{正常}$)

= (3/5)*0 + (2/5)*0

= 0 （纯洁了）

位置轮换的程序为：

```
SORT CASES BY 户外 湿度 活动.
MATCH FILES FILE=*
 /KEEP=天数 风速 温度 户外 湿度  活动.
EXE.
```

天数	风速	温度	户外	湿度	活动
D8	弱	温柔	晴天	高	取消
D2	强	炎热	晴天	高	取消
D1	弱	炎热	晴天	高	取消
D9	弱	凉爽	晴天	正常	进行
D11	强	温柔	晴天	正常	进行
D12	强	温柔	阴天	高	进行
D3	弱	炎热	阴天	高	进行
D7	强	凉爽	阴天	正常	进行
D13	弱	炎热	阴天	正常	进行
D4	弱	温柔	雨天	高	进行
D14	强	温柔	雨天	高	取消
D5	弱	凉爽	雨天	正常	进行
D10	弱	温柔	雨天	正常	进行
D6	强	凉爽	雨天	正常	取消

■ 风速的熵

风速的属性值强和正常的熵分别为

$Entropy(S_{弱}) = -(1/3)*\log2(1/3) - (2/3)*\log2(2/3) = 0.918$

$Entropy(S_{强}) = -(1/2)*\log2(1/2) - (1/2)*\log2(1/2) = 1.00$

风速的熵

$Entropy(S_{晴天}, 风速) = (3/5)*Entropy(S_{弱}) + (2/5)*Entropy(S_{强})$

$= (3/5)*0.918 + (2/5)*1.00$

$= 0.9508$

位置轮换的程序为：

```
SORT CASES BY 户外 风速 活动.
MATCH FILES FILE=*
 /KEEP=天数 温度 湿度 户外 风速 活动.
EXE.
```

天数	温度	湿度	户外	风速	活动
D11	温柔	正常	晴天	强	进行
D2	炎热	高	晴天	强	取消
D9	凉爽	正常	晴天	弱	进行
D8	温柔	高	晴天	弱	取消
D1	炎热	高	晴天	弱	取消
D12	温柔	高	阴天	强	进行
D7	凉爽	正常	阴天	强	进行
D3	炎热	高	阴天	弱	进行
D13	炎热	正常	阴天	弱	进行
D14	温柔	高	雨天	强	取消
D6	凉爽	正常	雨天	强	取消
D4	温柔	高	雨天	弱	进行
D5	凉爽	正常	雨天	弱	进行
D10	温柔	正常	雨天	弱	进行

第三步：计算属性的增益

在户外为晴天的记录中，三个变量的增益分别为：

Gain($S_{晴天}$, 温度) = Entropy($S_{晴天}$) - Entropy ($S_{晴天}$, 温度) = 0.970 - 0.4 = 0.570

Gain($S_{晴天}$, 湿度) = Entropy($S_{晴天}$) - Entropy ($S_{晴天}$, 湿度) = 0.970 - 0 = 0.970

Gain($S_{晴天}$, 风速) = Entropy($S_{晴天}$) - Entropy ($S_{晴天}$, 风速) = 0.970-0.951= 0.019

湿度有最大增益；所以它用作晴天的支节点。因为 Entropy ($S_{晴天}$, 湿度) = 0，所以这一支的分类结束。

雨天支节点的选择：

需要了解的问题是“在雨天支节点处应该检验什么属性？”。

户外为雨天的记录有 5 个，$S_{雨天}$ = {D4, D5, D6, D10, D14}

第一步:计算户外为雨天的熵，前面已经计算完成，即 Entropy($S_{雨天}$) = 0.970

第二步:计算户外为雨天的条件下各属性的熵

■ 温度的熵

在嵌套计算中，**温度**有三个属性，凉爽，温柔和炎热。它们的熵分别为

Entropy($S_{凉爽}$) = - (1/2)*log2(1/2) - (1/2)*log2(1/2) = 1

Entropy($S_{温柔}$) = - (2/3)*log2(2/3) - (1/3)*log2(1/3) = 0.918

Entropy($S_{炎热}$) = 0 （纯洁了）

温度的熵：

Entropy (S 雨天, 温度)=(2/5)*Entropy(S 凉爽)+(3/5)*Entropy(S 温柔)

= (2/5)*1 + (2/5)*0.918

= 0.767

位置轮换的程序为：

```
SORT CASES BY 户外 温度 活动.
MATCH FILES FILE=*
 /KEEP=天数 风速 湿度 户外 温度 活动.
EXE.
```

天数	风速	湿度	户外	温度	活动
D9	弱	正常	晴天	凉爽	进行
D11	强	正常	晴天	温柔	进行
D8	弱	高	晴天	温柔	取消
D2	强	高	晴天	炎热	取消
D1	弱	高	晴天	炎热	取消
D7	强	正常	阴天	凉爽	进行
D12	强	高	阴天	温柔	进行
D3	弱	高	阴天	炎热	进行
D13	弱	正常	阴天	炎热	进行
D5	弱	正常	雨天	凉爽	进行
D6	强	正常	雨天	凉爽	取消
D4	弱	高	雨天	温柔	进行
D10	弱	正常	雨天	温柔	进行
D14	强	高	雨天	温柔	取消

■ **湿度**的熵

在嵌套计算中，变量**湿度**有两个属性，湿度正常和湿度高，它们的熵分别为

Entropy(S 高) = - (1/2)*log2(1/2) - (1/2)*log2(1/2) = 1

Entropy(S 正常) = - (2/3)*log2(2/3) - (1/3)*log2(1/3) = 0.918

湿度的熵：

Entropy (S 雨天, 湿度)=(3/5)*Entropy(S 高)+(2/5)*Entropy(S 正常)

= (3/5)*1 + (2/5)*0.918

= 0.967

位置轮换的程序为：

```
SORT CASES BY 户外 湿度 活动.
MATCH FILES FILE=*
 /KEEP=天数 风速 温度 户外 湿度 活动.
EXE.
```

天数	风速	温度	户外	湿度	活动
D8	弱	温柔	晴天	高	取消
D2	强	炎热	晴天	高	取消
D1	弱	炎热	晴天	高	取消
D9	弱	凉爽	晴天	正常	进行
D11	强	温柔	晴天	正常	进行
D12	强	温柔	阴天	高	进行
D3	弱	炎热	阴天	高	进行
D7	强	凉爽	阴天	正常	进行
D13	弱	炎热	阴天	正常	进行
D4	弱	温柔	雨天	高	进行
D14	强	温柔	雨天	高	取消
D5	弱	凉爽	雨天	正常	进行
D10	弱	温柔	雨天	正常	进行
D6	强	凉爽	雨天	正常	取消

■ **风速**的熵

在嵌套计算中，风速的属性值强和正常的熵分别为

Entropy($S_{弱}$) = - (3/3)*log2(3/3) = 0 （纯洁了）

Entropy($S_{强}$) = - (2/2)*log2(2/2) = 0 （纯洁了）

风速的熵：

Entropy ($S_{雨天}$,风速)=(3/5)*Entropy($S_{弱}$)+(2/5)*Entropy($S_{强}$)

= (3/5)*0 + (2/5)*0

= 0 （纯洁了）

位置轮换的程序为：

```
SORT CASES BY 户外 风速 活动.
MATCH FILES FILE=*
 /KEEP=天数 温度 湿度 户外 风速 活动.
EXE.
```

天数	温度	湿度	户外	风速	活动
D11	温柔	正常	晴天	强	进行
D2	炎热	高	晴天	强	取消
D9	凉爽	正常	晴天	弱	进行
D8	温柔	高	晴天	弱	取消
D1	炎热	高	晴天	弱	取消
D12	温柔	高	阴天	强	进行
D7	凉爽	正常	阴天	强	进行
D3	炎热	高	阴天	弱	进行
D13	炎热	正常	阴天	弱	进行
D14	温柔	高	雨天	强	取消
D6	凉爽	正常	雨天	强	取消
D4	温柔	高	雨天	弱	进行
D5	凉爽	正常	雨天	弱	进行
D10	温柔	正常	雨天	弱	进行

第三步：计算属性的增益

在户外为雨天的记录中，三个变量的增益分别为：

$Gain(S_{雨天}, 温度) = Entropy(S_{雨天}) - Entropy(S_{雨天}, 温度) = 0.970 - 0.767 = 0.203$

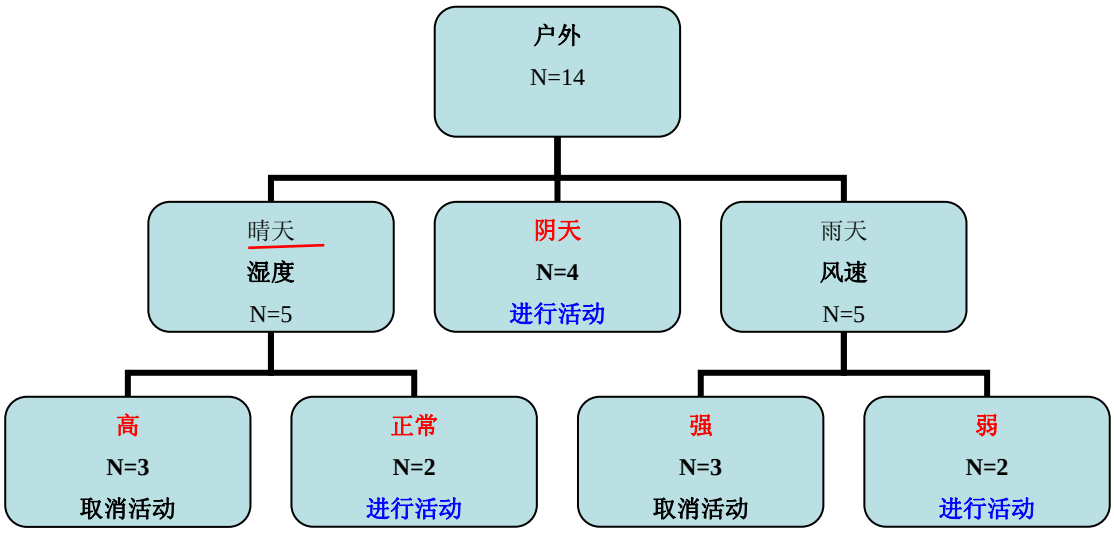
$Gain(S_{雨天}, 湿度) = Entropy(S_{雨天}) - Entropy(S_{雨天}, 湿度) = 0.970 - 0.967 = 0.003$

$Gain(S_{雨天}, 风速) = Entropy(S_{雨天}) - Entropy(S_{雨天}, 风速) = 0.970 - 0 = 0.970$

风速有最大增益；所以它用作雨天的支节点。

因为晴天的湿度高和正常的熵为 0，所以这一支划分结束。雨天的风速强和弱的熵为 0，所以这一支也划分结束。

最后形成的分类树大致是这个样子：



决策树也能用规则公式表示：

如果户外为**晴天**并且湿度**高**，那么活动**取消**

如果户外为**晴天**并且湿度**正常**，那么活动**进行**

如果户外为**阴天**，那么活动**进行**

如果户外为**雨天**并且风**大**，那么活动**取消**

如果户外为**雨天**并且风**弱**，那么活动**进行**

至此，ID3 的算法就演示完毕，尽管这个算法比较土，但它是分类树的根，还是了解一下比较好。

作者联系方式：itellin@163.com