基于水色图像的水质评价

目录

案例背景

案例背景

- 水产养殖的关键因素之一是水质
- 养殖水体生态系统的平衡状况可通过水质颜色体现
- 传统水质监控的关键：行家

专家判断的局限性

- 对个人经验要求高
- 存在主观性引起的观察性偏差
- 观察结果的可比性、可重复性不高，不易推广应用

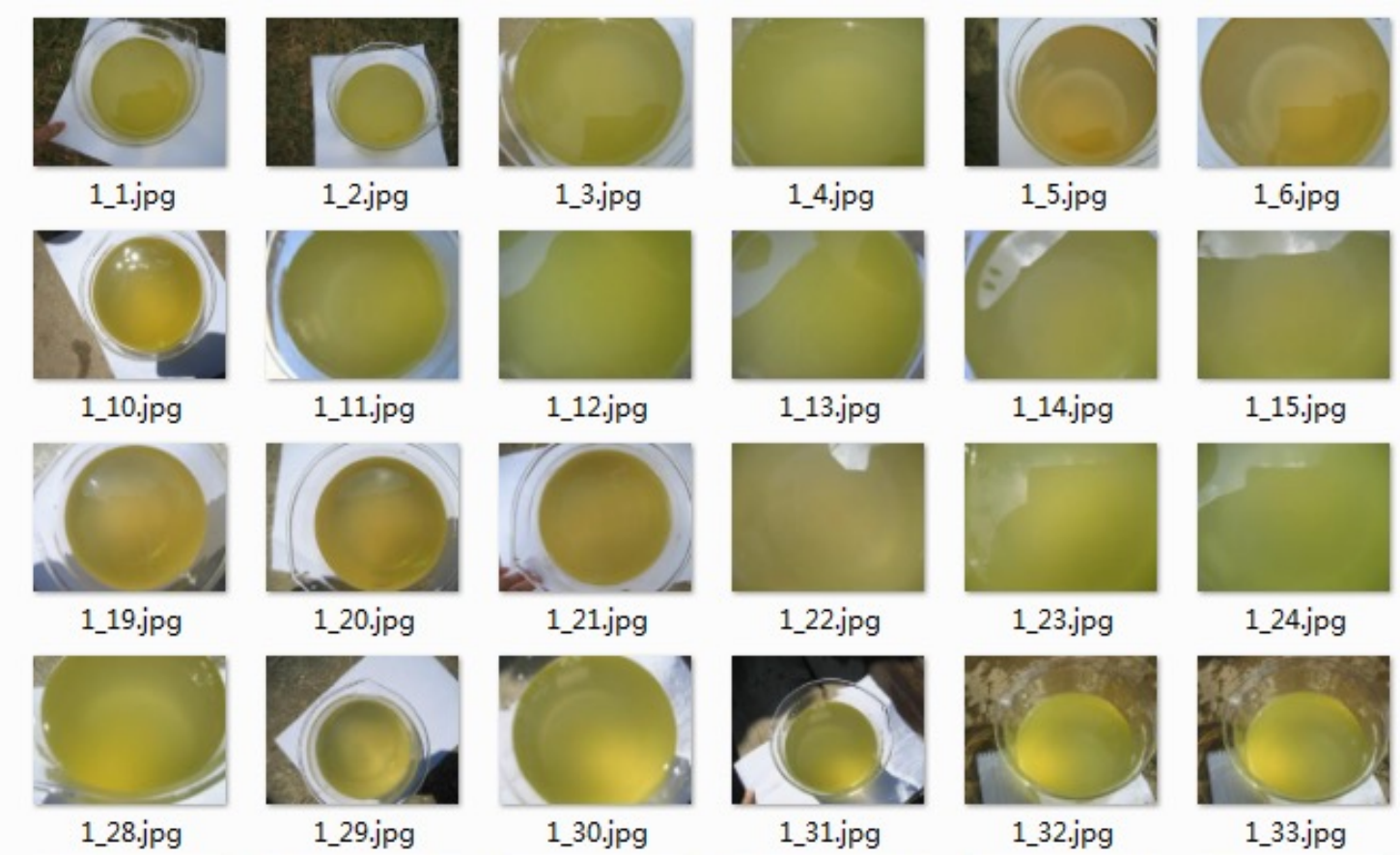
案例背景

在线水质监测

- 计算机视觉
- 数字图像处理技术
- 专家经验（专家数据）
- 机器学习算法

案例背景

原始数据



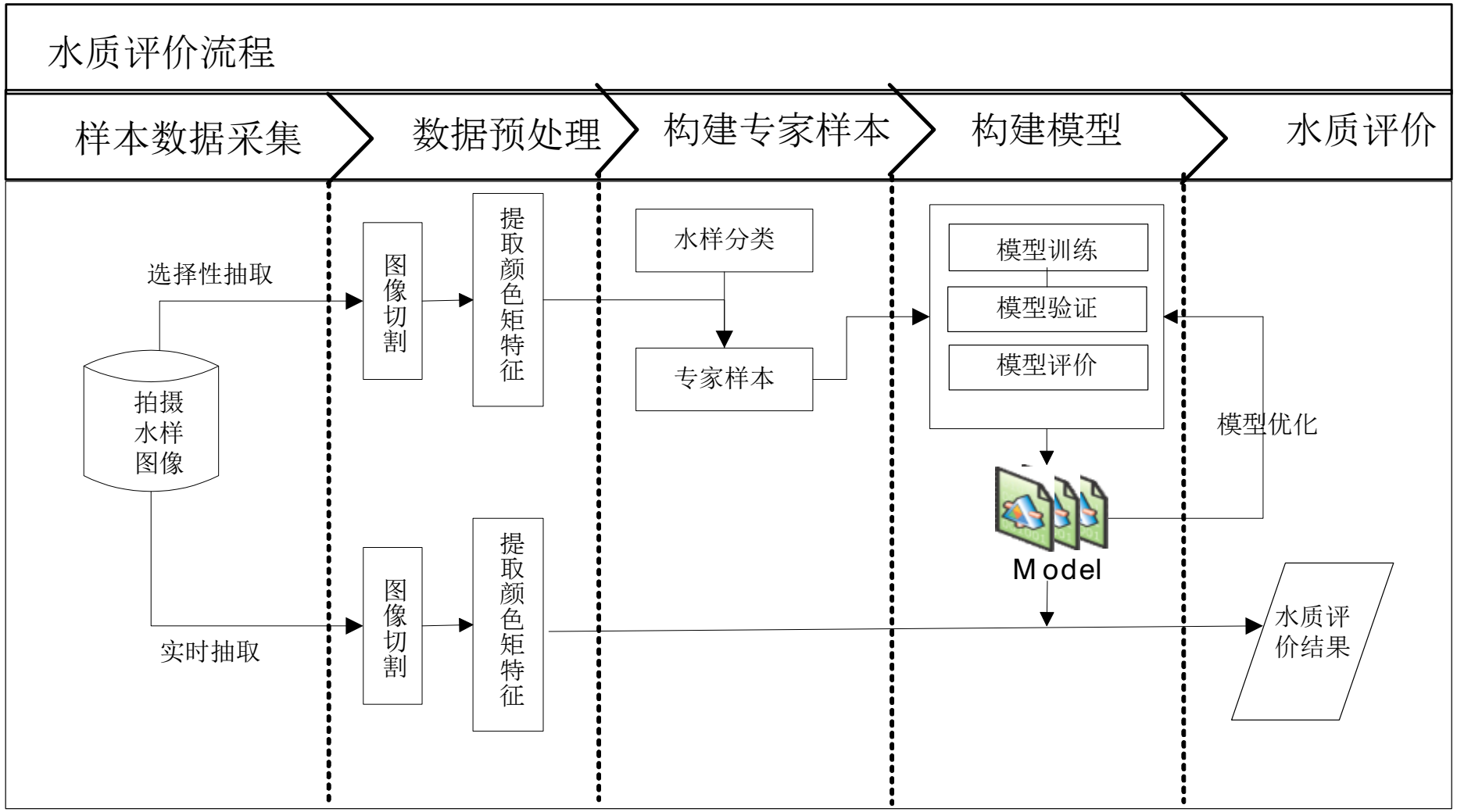
水质分类标准

水色	水质类别
浅绿色	1
灰蓝色	2
黄褐色	3
茶褐色/姜黄	4

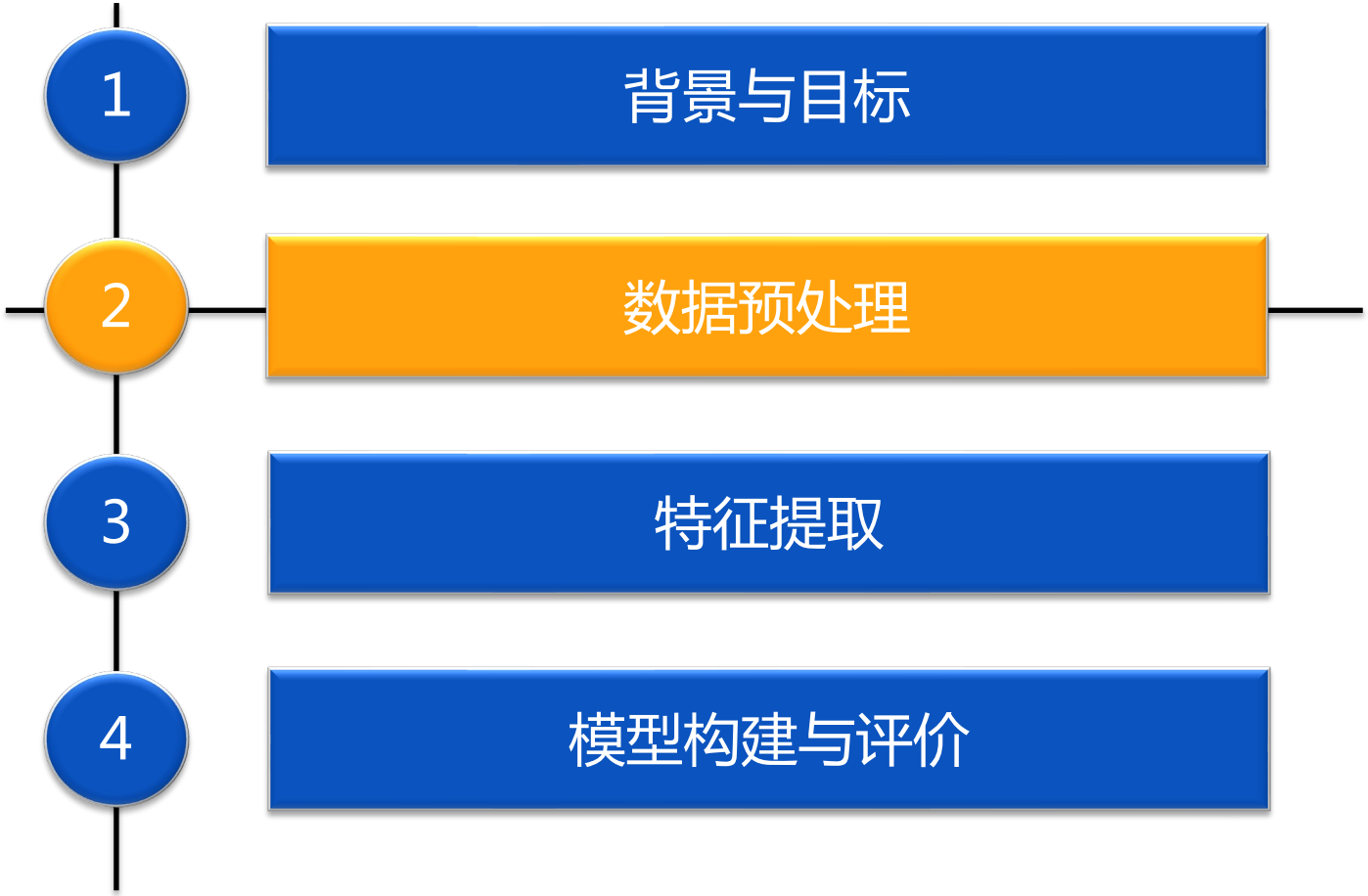
挖掘目标

请根据水质图片，利用图像处理技术和相应模型，实现水质的自动评价。

挖掘目标


水质评价流程
样本数据采集
数据预处理
构建专家样本
构建模型
水质评价
选择性抽取
拍摄水样图像
实时抽取
图像切割
图像切割
提取颜色矩特征
提取颜色矩特征
水样分类
专家样本
模型训练
模型验证
模型评价
模型优化
Model
水质评价结果

目录



采集水样图像

数据转化（Python）

- 图片转像素值矩阵：PIL Image.open()
- r,g,b = im.split() *#分成3个颜色通道*
- r_d = np.asarray(r) *#取出各通道像素值*

采集水样图像

图像切割

提取水样图像中央101*101像素的图像。设原始图像的大小是 $M \times N$，则截取宽从第 $fix(\frac{M}{2}) - 50$ 个像素点到第 $fix(\frac{M}{2}) + 50$ 个像素点，高从第 $fix(\frac{N}{2}) - 50$ 个像素点到第 $fix(\frac{N}{2}) + 50$ 个像素点的子图像。

目录

特征提取

- 特征提取

- 图像特征主要包括：颜色特征、纹理特征、形状特征、空间关系特征等。
- 与几何特征相比，颜色特征更为稳健，对于物体的大小和方向均不敏感，表现出较强的鲁棒性。
- 本案例中由于水色图像是均匀的，故主要关注颜色特征。

特征提取

颜色特征

- 颜色直方图：反映的是图像中颜色的组成分布，即出现了哪些颜色以及各种颜色出现的概率。
- 颜色矩：图像中任何的颜色分布均可以用它的矩来表示。颜色矩包含各个颜色通道的一阶距-（平均值）、二阶矩（方差）、三阶矩阵（偏度）和四阶矩（峰度），对于一幅RGB颜色空间的图像，具有R、G和B三个颜色通道，则有12个分量。

特征提取

特征提取：各阶颜色矩

- 一阶颜色矩-均值：反映了图像的整体明暗程度。

$$E = \frac{1}{N}\sum pix$$

- 二阶颜色矩-方差：表示图像颜色分布的离散程度 。

$$\sigma = \frac{1}{n}\sum_{j=1}^{n}(p_{i,j} - \mu_i)^2$$

- 三阶颜色矩-偏度：表示图像颜色分布(正态分布)的偏斜程度 。

$$\text{Skew(x)} = E[(\frac{x-\mu}{\sqrt{\sigma}})^3]$$

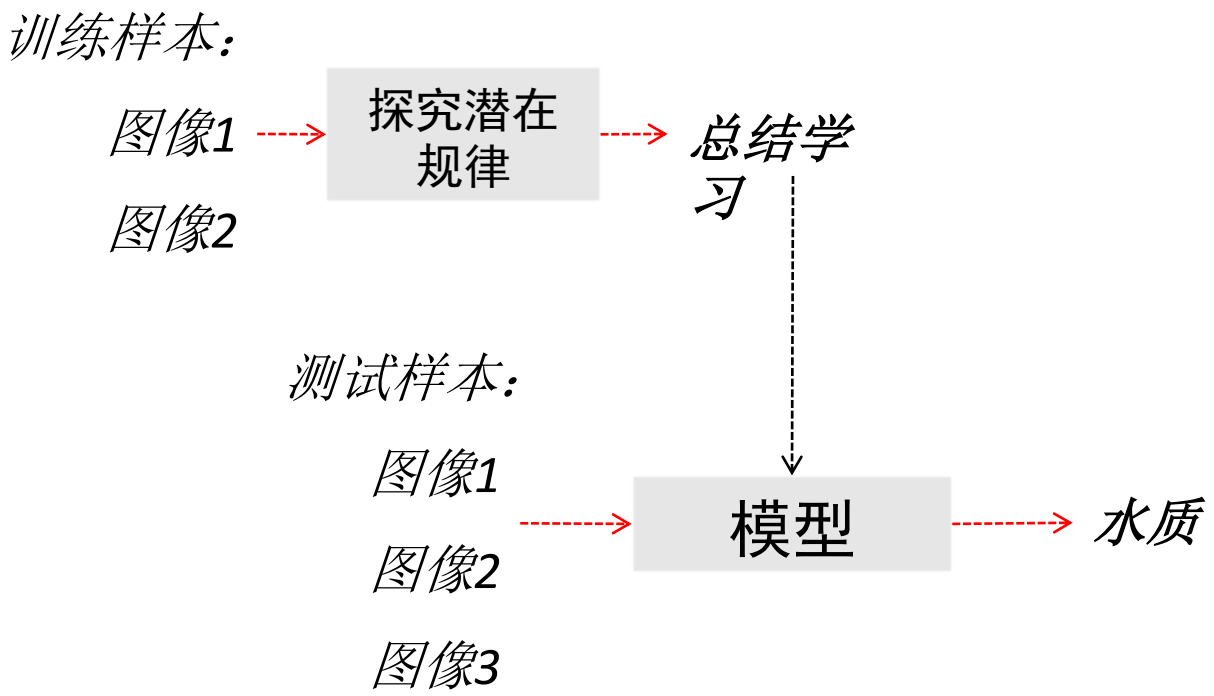
- 三阶颜色矩-峰度：即分布的尖锐程度。包括正态分布（峰度值=3），厚尾（峰度值>3），瘦尾（峰度值<3）

$$\text{Kurt(x)} = E[(\frac{x-\mu}{\sqrt{\sigma}})^4]$$

目录

模型构建与评价

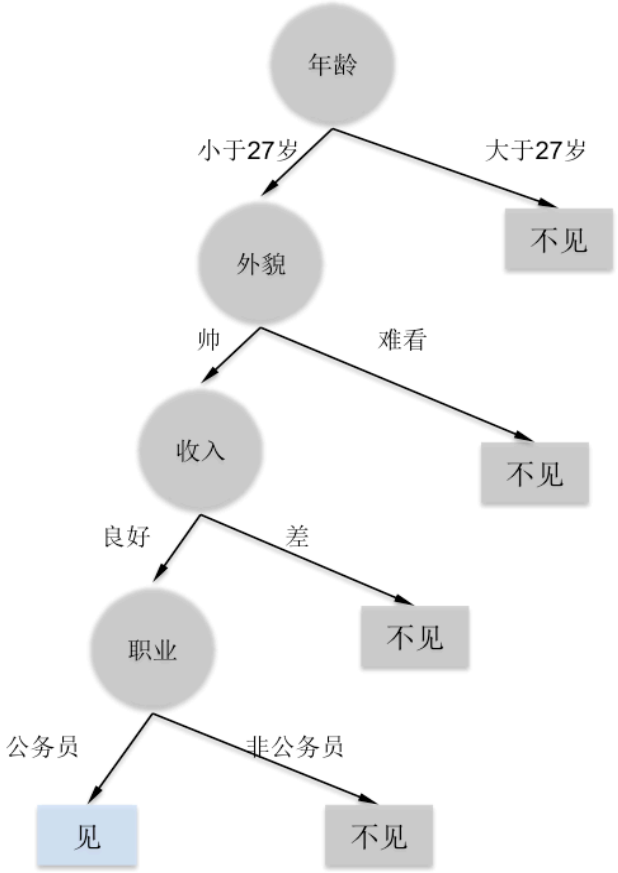
1. 抽取80%作为训练样本，剩下的20%作为测试样本。
2. 用训练集样本对模型进行训练。
3. 用测试集样本对模型性能进行评价。



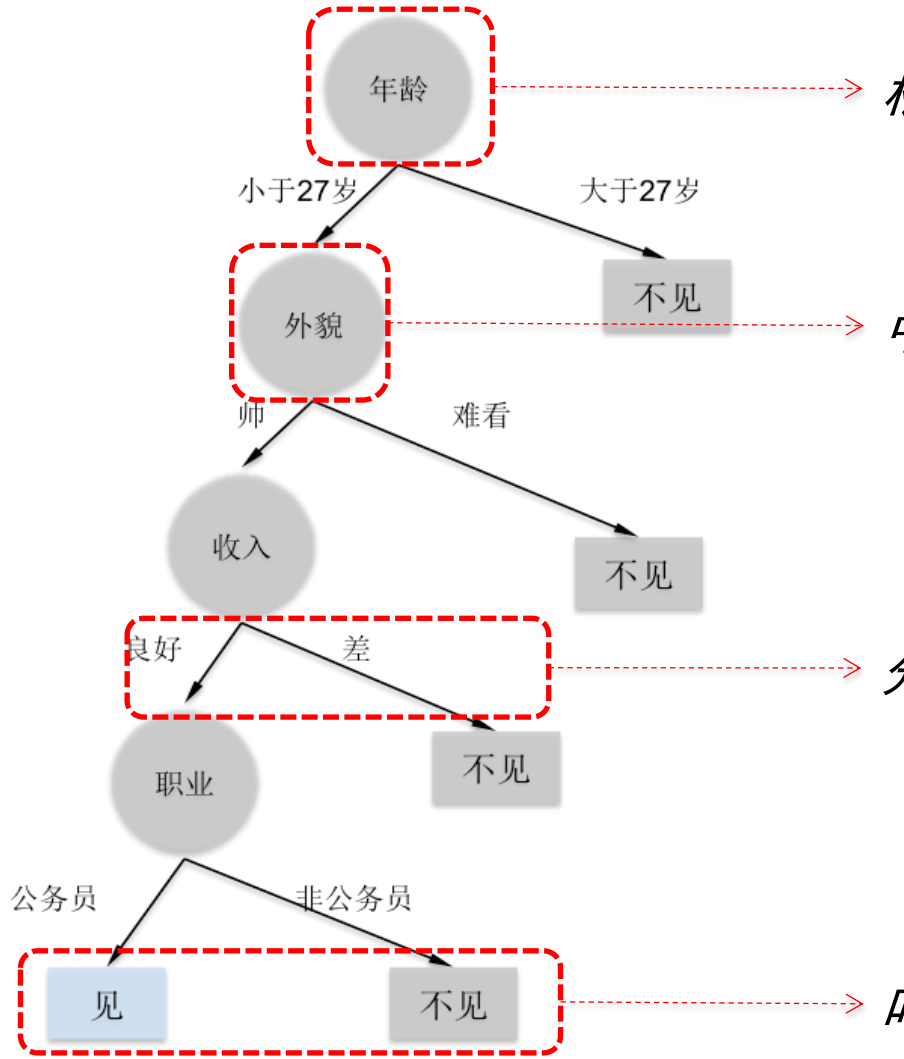
决策树

现想象一个女孩的母亲要给这个女孩介绍男朋友，于是有了下面的对话：

- 女儿：多大年纪了？
- 母亲：26
- 女儿：长的帅不帅？
- 母亲：挺帅的
- 女儿：收入高不？
- 母亲：不算很高，中等情况。
- 女儿：是公务员不？
- 母亲：是，在税务局上班呢。
- 女儿：那好，我去见见。



决策树



根部节点(root node)

中间节点(non-leaf node)(代表测试的条件)

分支(branches)(代表测试的结果)

叶节点(leaf node)(代表分类后所获得的分类标记)

决策树

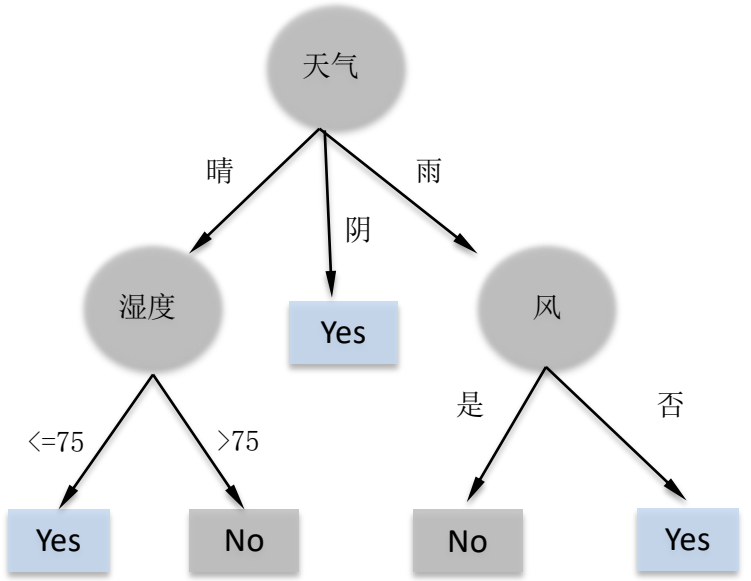
天气情况对是否打高尔夫球的影响

日期	天气	温度(华氏度)	湿度	起风	打球?
1	晴	85	85	F	No
2	晴	80	90	T	No
3	阴	83	78	F	Yes
4	雨	70	96	F	Yes
5	雨	68	80	F	Yes
6	雨	65	70	T	No
7	阴	64	65	T	Yes
8	晴	72	95	F	No
9	晴	69	70	F	Yes
10	雨	75	80	F	Yes
11	晴	75	70	T	Yes
12	阴	72	90	T	Yes
13	阴	81	75	F	Yes
14	雨	71	80	T	No
15	阴	85	90	F	?
16	雨	80	79	F	?
17	晴	78	70	T	?

决策树

- 天气情况对是否打高尔夫球的影响

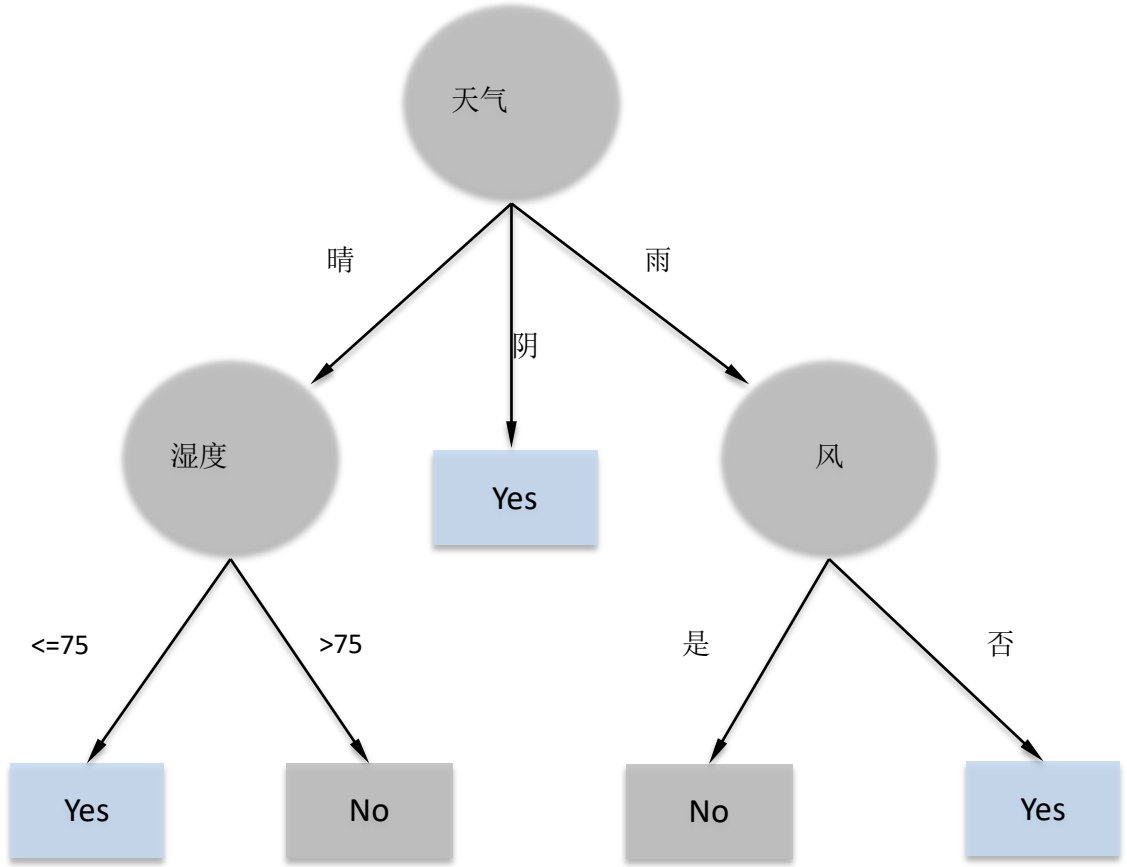
日期	天气	温度(华氏度)	湿度	起风	打球?
1	Sunny	85	85	F	No
2	Sunny	80	90	T	No
8	Sunny	72	95	F	No



决策树

决策树关键词

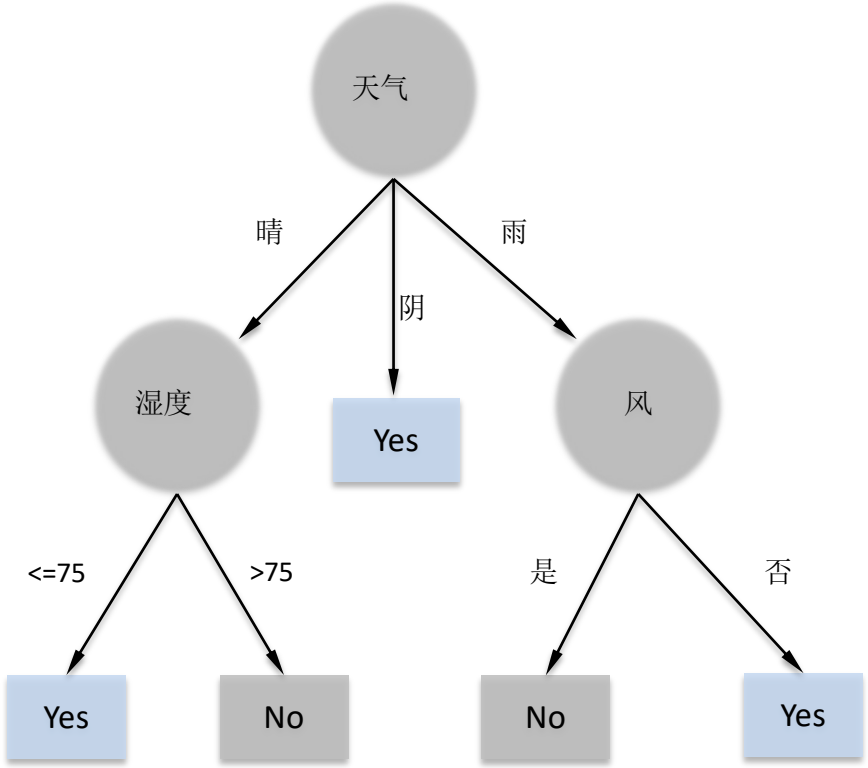
- 属性选择的先后顺序
- 熵值
- 信息增益
- 信息增益率



决策树

问题：对于给定样本集，如何判断应该在哪个属性上进行拆分

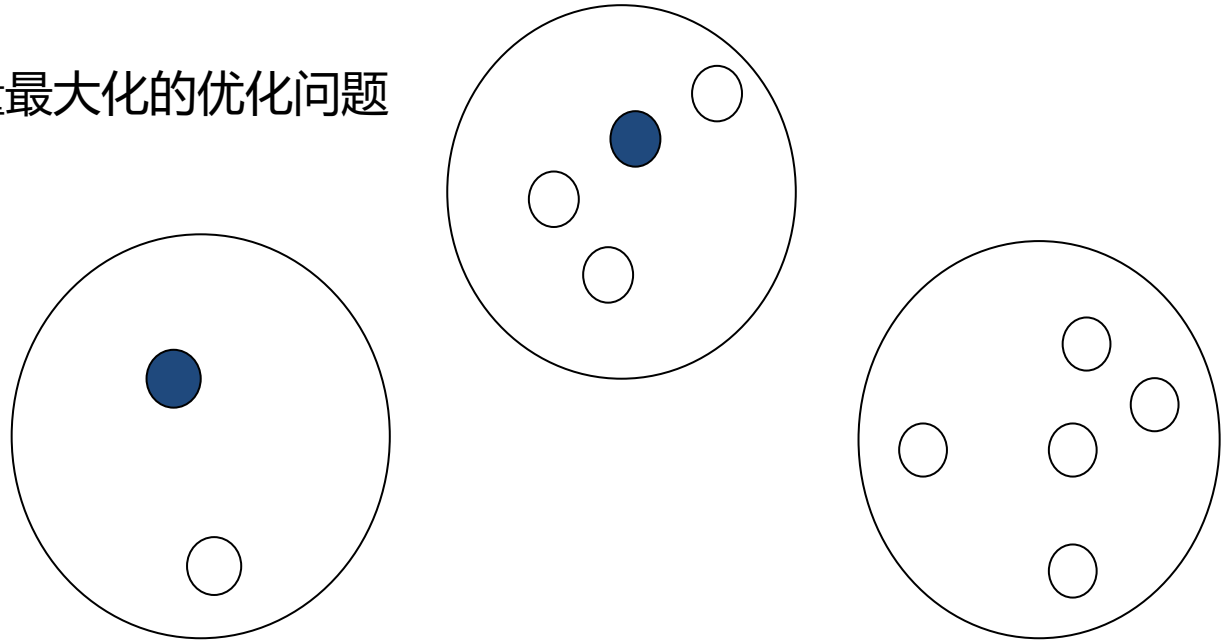
- 理想情况：在拆分过程中，当叶节点只拥有单一类别时，将不必继续拆分。
- 目标是寻找较小的树，希望递归过程尽早停止，得到较小的树？
- 当前最好的拆分属性产生的拆分中目标类的分布应该尽可能地单一（单纯，分支少），多数类占优。
- 决策树算法通常按照纯度的增加来选择拆分属性。



决策树

纯度的概念

- 纯度度量
  - 当样本中没有两项属于同一类：0
  - 当样本中所有项都属于同一类：1
- 最佳拆分可以转化为选择拆分属性使纯度度量最大化的优化问题



决策树

纯度的度量

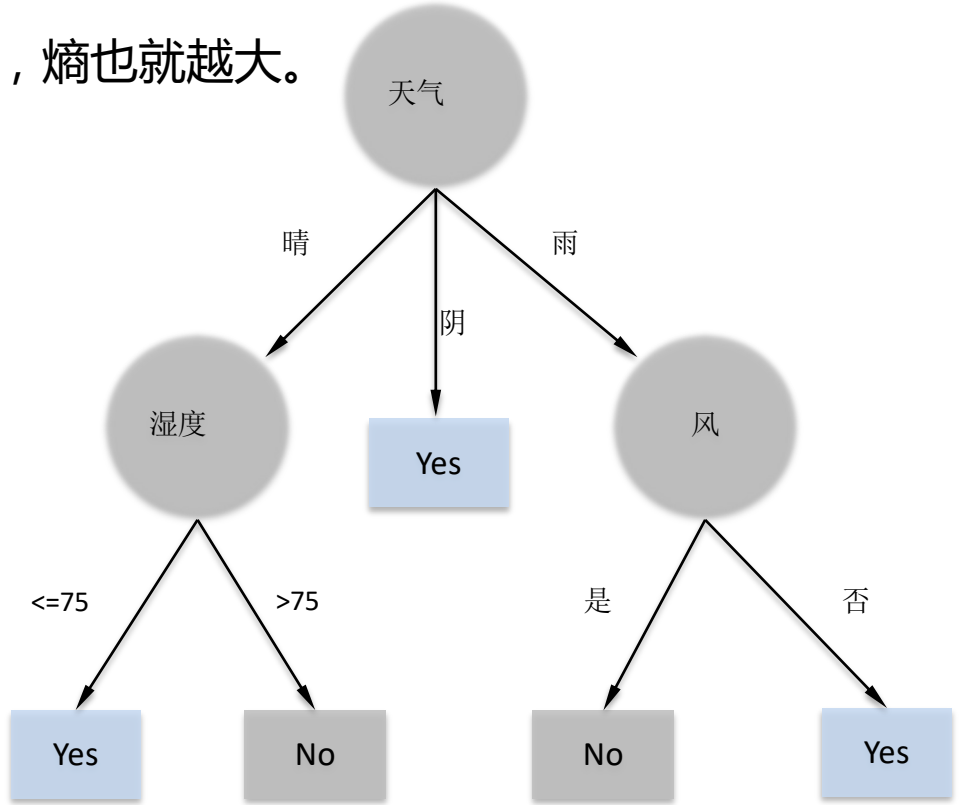
- 拆分增加了纯度，但如何将这种增加量化呢，或者如何与其他拆分进行比较呢？
- 用于评价拆分分类目标变量的纯度度量包括
  - 熵(entropy，信息量)
  - 信息增益(Gain) ID3
  - 信息增益率　C4.5，C5.0
  - 基尼(Gini，总体发散性) CART
- 改变拆分准则（splitting criteria）导致树的外观互不相同

决策树

熵(entropy)

- 信息论中的熵：是信息的度量单位，是一种 对属性“不确定性的度量”。
- 属性的不确定性越大，把它搞清楚所需要的信息量也就越大，熵也就越大。
- 如果一个数据集D有N个类别，则该数据集的熵为：

$$Ent(D) = -\sum_{i=1}^{N} p_i \log_2 p_i$$



决策树

打球与否？

日期	天气	温度(华氏度)	湿度	起风	打球?
1	晴	85	85	F	No
2	晴	80	90	T	No
3	阴	83	78	F	Yes
4	雨	70	96	F	Yes
5	雨	68	80	F	Yes
6	雨	65	70	T	No
7	阴	64	65	T	Yes
8	晴	72	95	F	No
9	晴	69	70	F	Yes
10	雨	75	80	F	Yes
11	晴	75	70	T	Yes
12	阴	72	90	T	Yes
13	阴	81	75	F	Yes
14	雨	71	80	T	No
15	阴	85	90	F	?
16	雨	80	79	F	?
17	晴	78	70	T	?

$$Ent(D) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

打球数据集的熵为：

$$-(9/14)\log_2(9/14) - (5/14)\log_2(5/14) = 0.940$$

决策树

信息增益(gain)：对纯度提升的程度，即节点的纯度提升越多，信息增益越大，提供的信息量就越多，那么信息的不确定性较小幅度越大。

- 若离散属性a有V个取值，则其信息增益为：

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

决策树

信息增益(gain)

天气属性的信息增益

- 晴：打球记录2条，不打球记录为3条

$$Ent(D^1) = -(2/5)\log_2(2/5) - (3/5)\log_2(3/5) = 0.97$$

- 阴：打球记录4条，不打球记录0条

$$Ent(D^2) = -(4/4)\log_2(4/4) - (0/4)\log_2(0/4) = 0$$

- 雨：打球记录3条，不打球记录2条

$$Ent(D^3) = -(3/5)\log_2(3/5) - (2/5)\log_2(2/5) = 0.97$$

决策树

信息增益(gain)

天气属性的信息增益： $Gain(D,a)=Ent(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}Ent(D^v)$

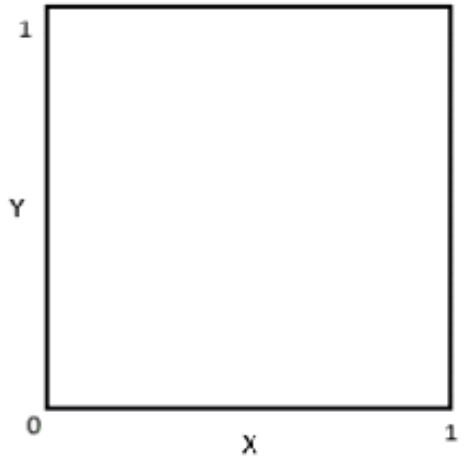
$$=0.940-\frac{5}{14}\times 0.97-\frac{4}{14}\times 0-\frac{5}{14}\times 0.97$$

$$=0.247$$

起风属性的信息增益： 0. 048

决策树

- 模型构建示意图



For more tutorials: annalyzin.wordpress.com

决策树算法分类

- 常用的决策树算法见下表：

决策树算法	算法描述
ID3算法	其核心是在决策树的各级节点上，使用信息增益作为属性的选择标准，来帮助确定每个节点所应采用的合适属性。
C4.5算法	C4.5决策树生成算法相对于ID3算法的重要改进是使用信息增益率来选择节点属性。C4.5算法既能够处理离散的描述属性，也可以处理连续的描述属性。
C5.0算法	C5.0是C4.5算法的修订版，适用于处理大数据集，采用Boosting方式提高模型准确率，根据能够带来的最大信息增益的字段拆分样本。
CART算法	CART决策树是一种十分有效的非参数分类和回归方法，通过构建树、修剪树、评估树来构建一个二叉树。当终结点是连续变量时，该树为回归树；当终结点是分类变量，该树为分类树。

Thank you!