# 机器学习期末报告

组别：A2

组长：王茂华

组员：钟宗幸、陈远涛、侯君凌、刘煜龙

## 前言

我国不仅是一个严重缺水的国家，而且水资源的总量十分有限，水污染也很严重，甚至到了触目惊心的地步。据统计，我国淡水域普遍受到污染，全国3/4的江河呈现不同程度的富营养化；大部分城市水体污染严重，对我国128个南北分布城市的地下水进行调查显示，125个城市的地下水受到了不同程度的污染。

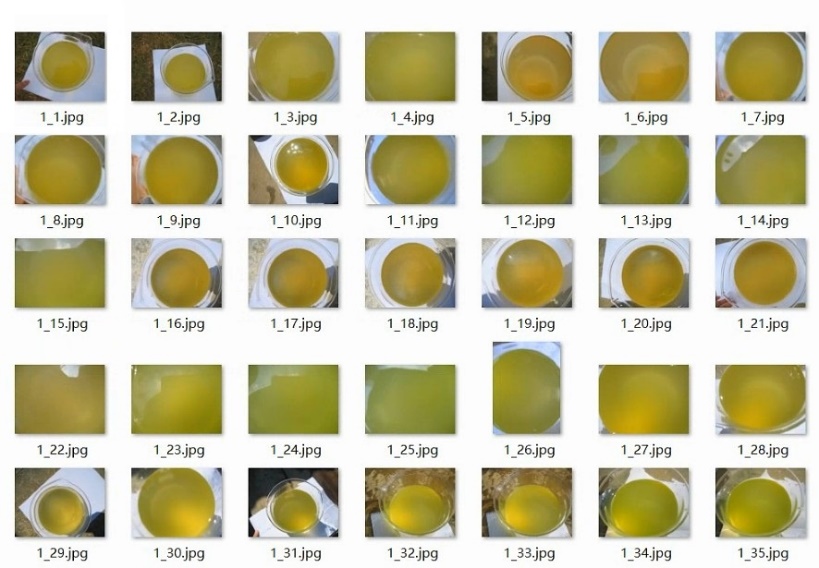
水是生命之源。水体（质）的优劣与人类健康密切相关，水质检测则是确保水环境的必备工作之一。传统的水质检测技术主要是采用化学方法对水质进行检测，一般是依靠人工现场采取水样，利用实验室仪器对其进行测试分析。实验室中进行水质检测，虽然所需仪器俱全、工作环境稳定，但是实验室检测存在周期长、监测频率低、采样误差大以及造成二次污染等诸多缺点。

由此，我小组决定以基于水色图像的水质评价为题，完成本次期末实践。

## 分析方法与过程

### 1.原始数据与分类标准

我们选择了197张不同水质等级、在近似条件下拍摄的水样照片作为原始数据，并将其按【水质等级】\_【图片序号】的方式命名。



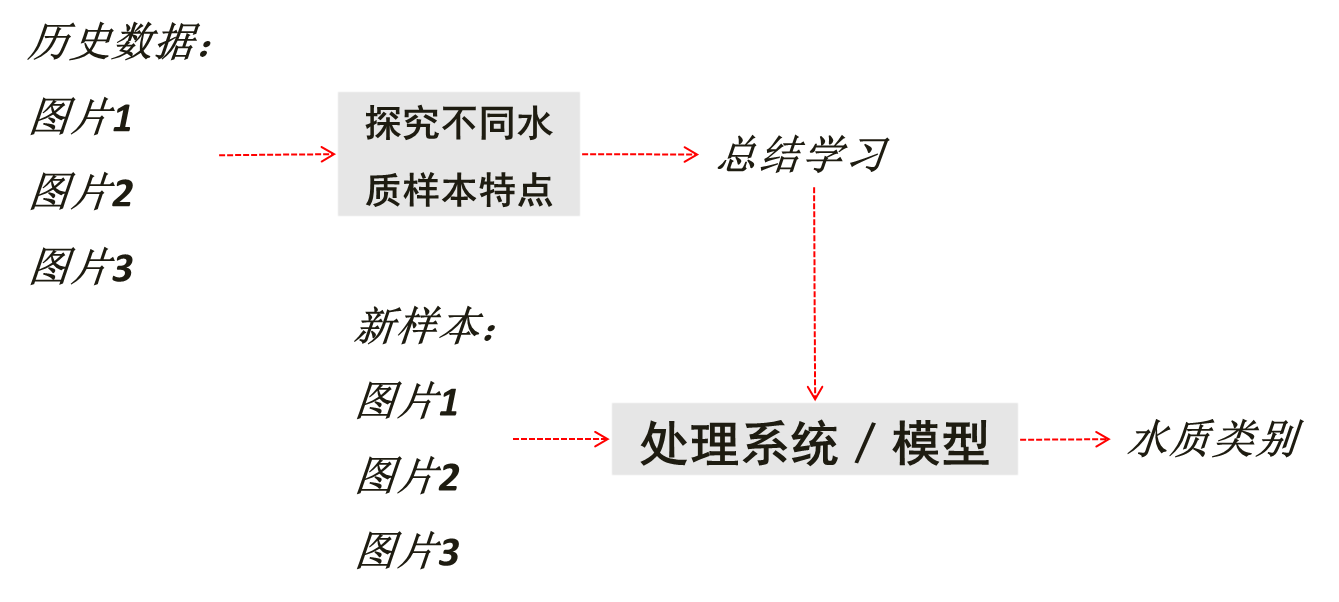
其水质等级分类方法如下：

|  |  |
| --- | --- |
| 水样颜色 | 水质级别 |
| 浅绿 | 1 |
| 灰蓝 | 2 |
| 黄褐 | 3 |
| 茶褐/姜黄 | 4 |
| 绿/黄绿/油绿 | 5 |

### 2.实践目标

我们希望利用图像处理技术和对应的模型，利用水体样本图片，实现水质的自动评价。

通过对历史图像样本的总结学习，探究不同级别水体样本图片的特征，以此建立水样图片自动识别、评价与水质分类的模型。



### 3.对于样本图像的预处理与特征提取

#### 预处理：

##### 数据转化：

图片转像素值矩阵：PIL

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

##### 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

则截取宽从第个像素点到第个像素点，高从第个像素点到第个像素点的子图像。

#### 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。所以我们选择分析原始数据图像的颜色特征作为本次实践中对样本的分析方式。

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。其优点在于它能简单描述一幅图像中颜色的全局分布，即不同色彩在整幅图像中所占的比例，特别适用于描述那些难以自动分割的图像和不需要考虑物体空间位置的图像。其缺点在于它无法描述图像中颜色的局部分布及每种色彩所处的空间位置，即无法描述图像中的某一具体的对象或物体。

颜色矩：图像中任何的颜色分布均可以用它的矩来表示。根据概率论，随机变量的概率分布可以由其各阶矩唯一的表示和描述。一副图像的色彩分布也可认为是一种概率分布，那么图像可以由其各阶矩来描述。颜色矩包含各个颜色通道的一阶距、二阶矩和三阶矩，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9个分量。

颜色直方图产生特征维数一般大于颜色矩的特征维数，为了避免过多变量影响后续的分类效果，在本案例采用颜色矩来提取水样图像的特征。

以下为颜色矩计算公式与对样本图片的处理结果：

* **一阶颜色矩**

采用一阶原点矩反映图像的整体明暗程度

* **二阶颜色矩**

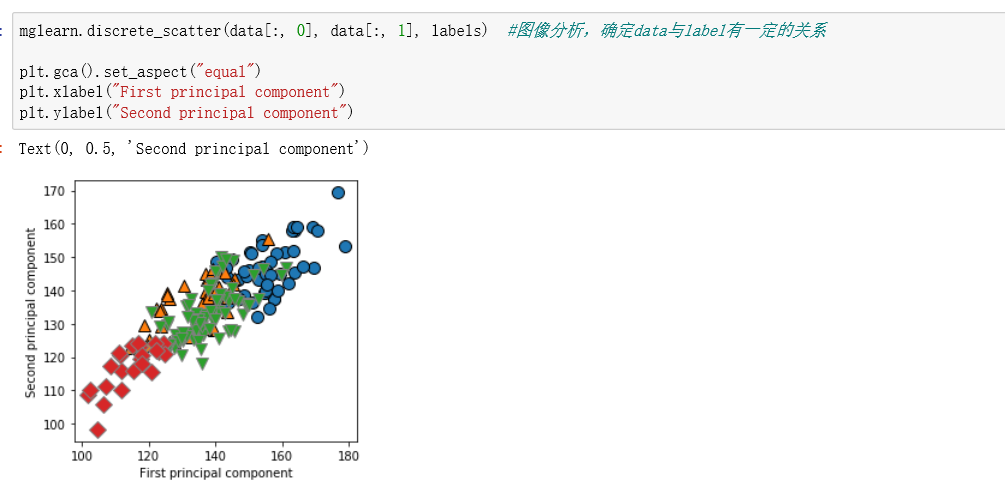
采用二阶中心矩的平方根，反映图像颜色的分布范围

* **三阶颜色矩**

采用三阶中心矩的立方根，反映了图像颜色分布的对称性

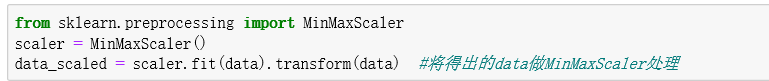


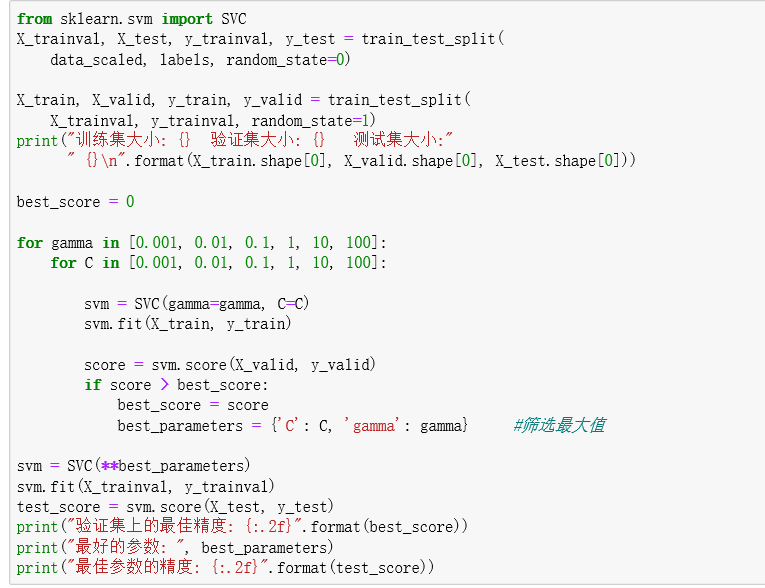
对数据进行图像分析，确认data与label有一定关系



### 4.模型优化与交叉验证

使用MinMaxScaler对得到的数据data进行处理





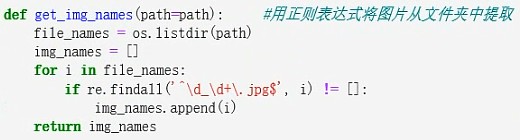
## 源代码



path为样本图片所在位置



求颜色通道三阶颜色矩



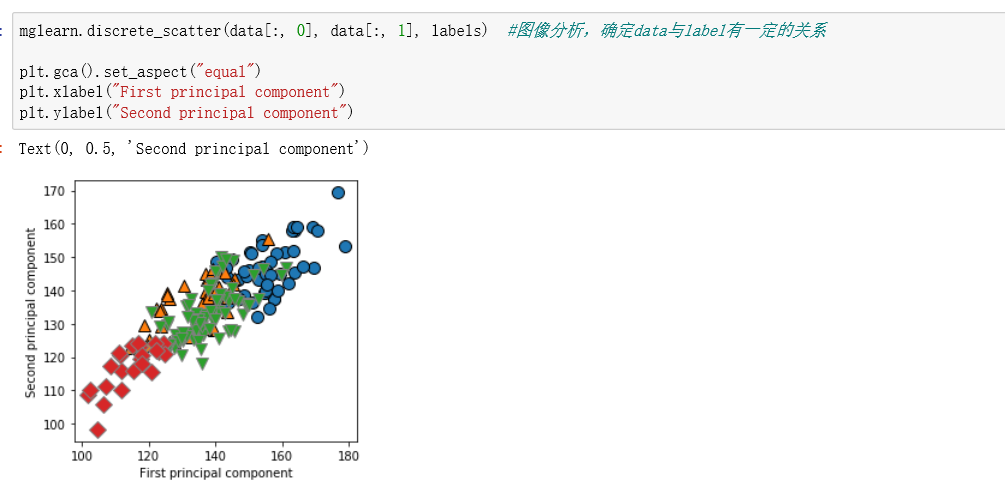
利用正则表达式提取样本图片



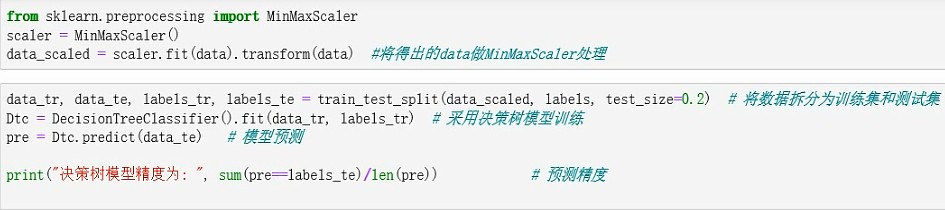
利用此函数处理图片数据



得到data与label



分析图象，确定data、label关系



利用MinMaxScaler进行处理，并用决策树模型训练，最后模型预测得出模型精度为0.875



使用网格交叉验证得出最佳精度、最好参数与最好参数的精度