目录

[一、背景介绍 2](#_Toc37880061)

[1、行业背景 2](#_Toc37880062)

[2、问题分析 2](#_Toc37880063)

[二、数据说明 3](#_Toc37880064)

[1、数据来源 3](#_Toc37880065)

[2、水色图片获取方法（控制变量进行拍照） 3](#_Toc37880066)

[3、数据量 4](#_Toc37880067)

[三、数据探索与数据预处理 5](#_Toc37880068)

[1、切分图片 5](#_Toc37880069)

[2、特征提取 5](#_Toc37880070)

[四、模型训练 6](#_Toc37880071)

[1、决策树 7](#_Toc37880072)

[2 、k近邻 8](#_Toc37880073)

[3、朴素贝叶斯 9](#_Toc37880074)

[4、支持向量机 9](#_Toc37880075)

[5、神经网络 10](#_Toc37880076)

[五、小结 12](#_Toc37880077)

# 一、背景介绍

1、行业背景

从事渔业生产有经验的从业者可通过观察水色变化调控水质，以维持养殖水体生态系统中浮游植物、微生物类、浮游动物等合理的动态平衡。由于这些多是通过经验和肉眼观察进行判断，存在主观性引起的观察性偏倚，使观察结果的可比性、可重复性降低，不易推广应用。当前，数字图像处理技术为计算机监控技术在水产养殖业的应用提供更大的空间。在水质在线监测方面，数字图像处理技术是基于计算机视觉，以专家经验为基础，对池塘水色进行优劣分级，达到对池塘水色的准确快速判别。

2、问题分析

由于图片数据由专业人士获取，因此我们的分析主要从数据采集完成后如下五个步骤着手分析问题：

**①、数据转换**：为了将图像数据转换为计算机能够识别的形式，通过PIL库对图片进行读取，并通过numpy进行数值计算。

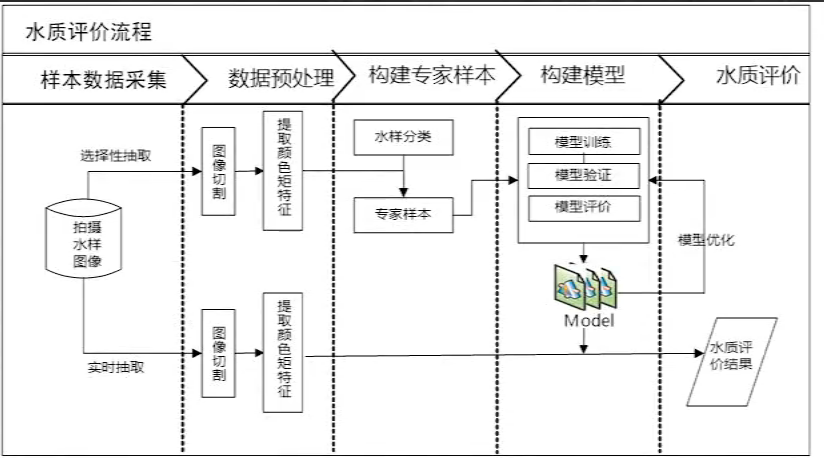
**②、数据清洗**：为了去除每张图片噪声数据（外部环境，非水色图片部分）的影响，通过PIL库截取每张图片中心100\*100像素区域的水色图片作为我们的分析数据。

**③、特征提取**：将图片的R、G、B颜色通道分离出来。使用概率论的基础知识，分别对每张图片的R、G、B三阶颜色矩阵进行计算，作为每张图片的特征。

**④、数据整理**：将存储在本地的图片数据进行批量读取，转换为我们的特征数据，并将特征数据与数据标签保存，为模型训练做准备。

**⑤、模型构建与评价**：通过上一步得到的特征数据与数据标签，划分训练集与测试集。通过scikit-learn构建决策树、k近邻、朴素贝叶斯、支持向量机、神经网络等分类模型并进行GridSearchCV参数搜索，最后使用模型分类准确率和混淆矩阵对模型进行评估。

整个分析过程的流程图如下：



# 二、数据说明

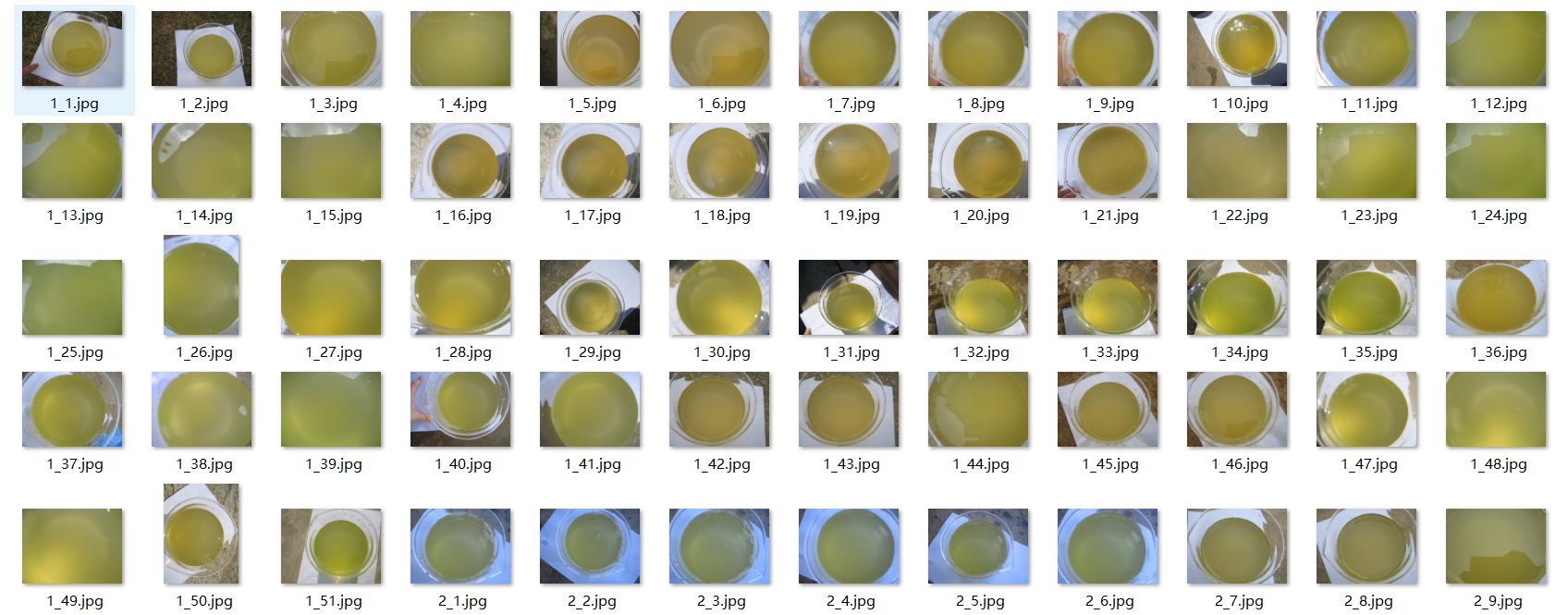
1、数据来源

水样由专业人士获取。

2、水色图片获取方法（控制变量进行拍照）

将所有采样的水，全部用一个桶装，拍照时将其放入一个预先搭设的“小黑屋”,排除周围光源的影响，而后将获得的水源放置在固定的位置，保证每次接受到的光源都是一样的。在确保水面平静，没有波纹后，对不同的水源进行拍照。

通过上述方法，一共获取到了203张图片。基本保证了每一张图片采集时，外部环境一致。部分图片如下





共计五类水色图片，命名规则为 水色类别\_编号。

3、数据量

图片数量：203张图片(每张图片约为400万像素)

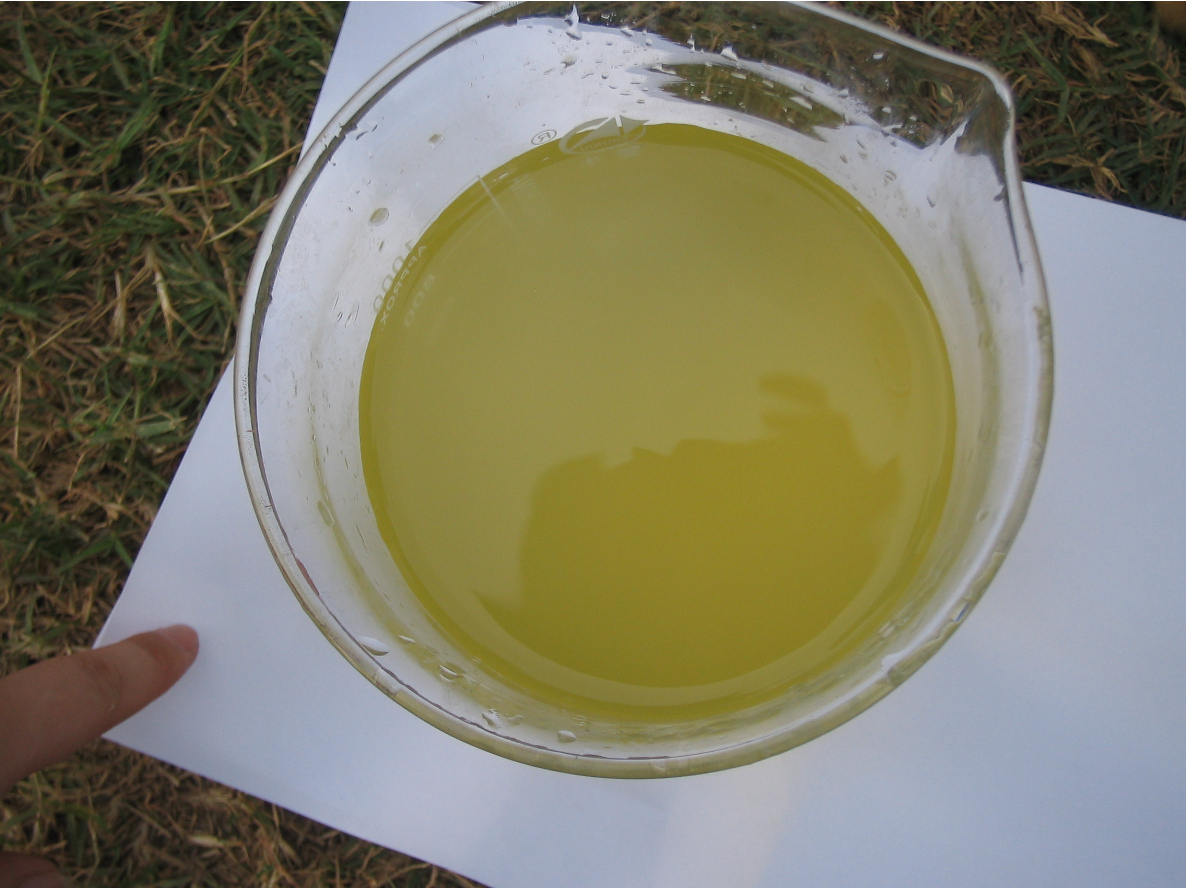
图片类别：5类图片

# 三、数据探索与数据预处理

1、切分图片

由于是通过图片的水色来对图像进行判别，为了消除图片数据中非水样的部分的影响，我们通过分割图像（首先通过os库读取本地的图片文件夹，再通过PIL库读取图片属性，通过PIL.crop()进行分割，截取图片中心区域），取每张图片的中心区域(100\*100)。示例如下

*原图*

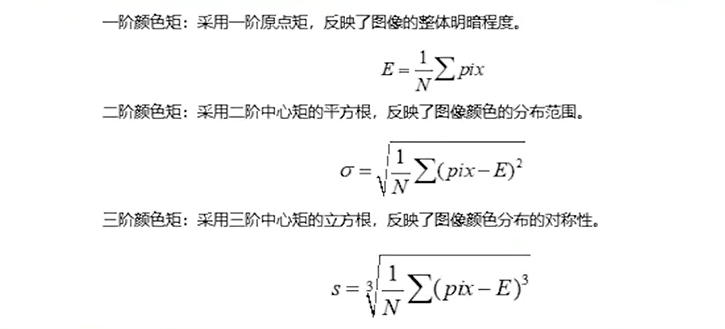


*分割得到的中心区域图片*



2、特征提取

为了获取图像的特征，我们将每张图片R、G、B三种颜色通道分离出来，以矩阵的形式存放，即颜色矩阵。分别计算每张图片的R、G、B通道的一、二、三阶矩阵。



计算结果如下所示:



图a

上*图a*中的水质类别作为分类标签label，剩下的数据作为data进行训练。

# 四、模型训练

我们对数据进行划分，将*图a*中的水质类别作为分类标签label，剩下的数据作为data进行训练。

将测试集与训练集按2:8进行划分、分别代入决策树、K近邻、朴素贝叶斯、神经网络与支持向量机等分类模型中进行训练，并通过分类的准确率与混淆矩阵对模型进行评估。

在训练模型时，为了找到最好的参数，我们通过sklearn中的GridSearchCV对模型的参数进行网格搜索，使得模型的效果尽可能地好。相关搜索参数以及模型训练结果如下：

1、决策树

**①搜索参数**

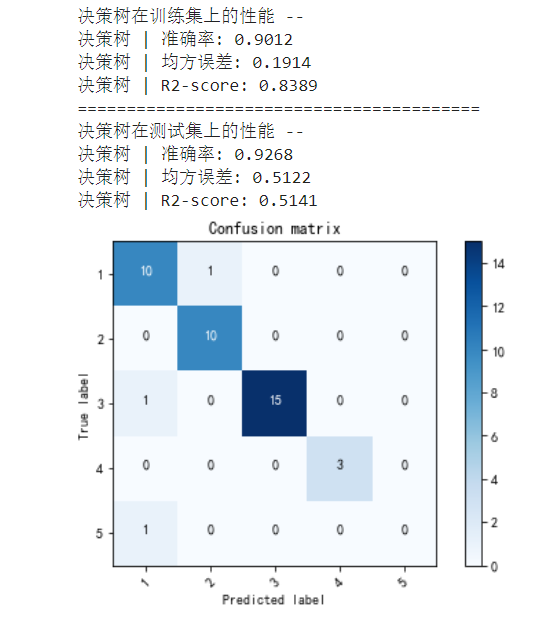
**param = {' criterion':['gini','entropy'],**

**'max\_depth':[i for i in range (1,50)],**

**'min\_samples\_leaf':[2,3,5,10],**

**'min\_impurity\_decrease':[0.1,0.2,0.5]}**

**②模型评估(准确率 92.68%)**



2 、k近邻

**①搜索参数**

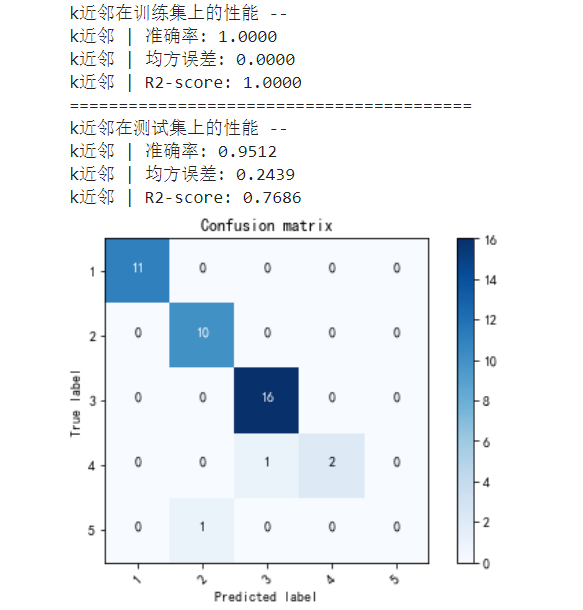
**param={'weights':['uniform','distance'],**

**'n\_neighbors':[i for i in range(1,11)],**

**'p':[i for i in range(1,6)]**

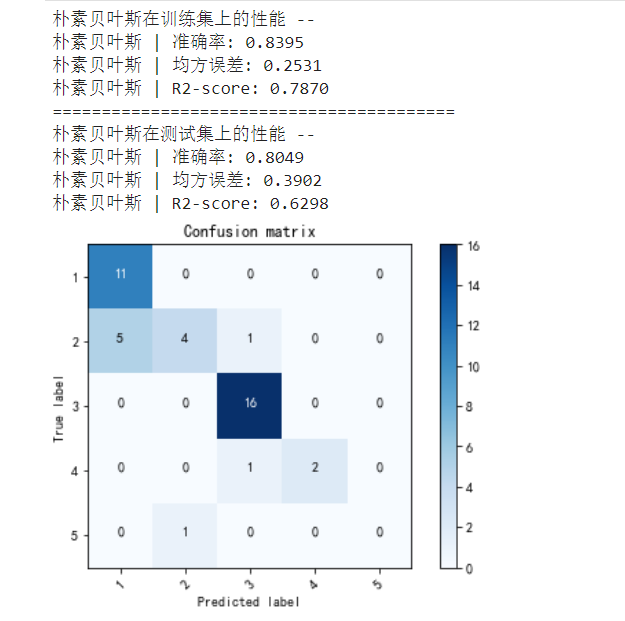
**}**

**②模型评估(准确率 95.12%)**



3、朴素贝叶斯

**模型评估（准确率80.49%）**



4、支持向量机

**①搜索参数**

**param={ 'cache\_size' : [100, 200, 400],**

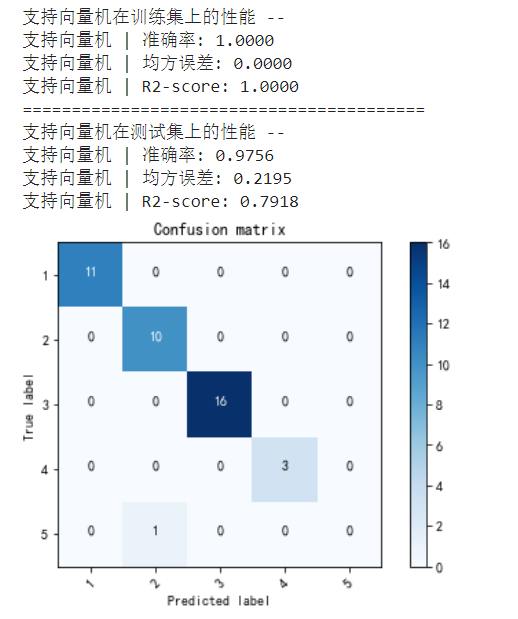
**'C': [1, 10, 100],**

**'kernel' : ['rbf', 'linear'],**

**'degree' : [1, 2, 3, 4],**

**}**

**②模型评估（准确率 97.56%）**



5、神经网络

**①搜索参数**

**param={**

**'activation':['identity', 'logistic', 'tanh', 'relu'],**

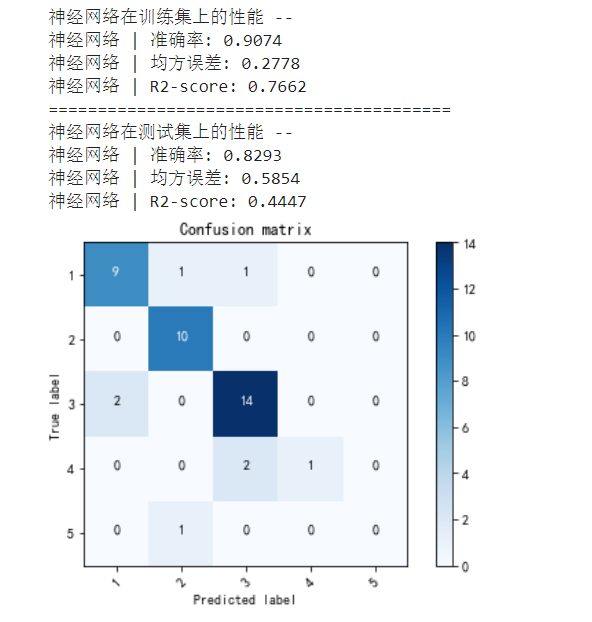
**'alpha':[10.0 \*\* ~ np.arange(1,7).any()],**

**'hidden\_layer\_sizes': [(100,),(50,)],**

**'max\_iter': [1000,500,200,100]**

**}**

**②模型评估（准确率 82.93%）**



# 五、小结

通过对上述模型的对比，我们发现使用支持向量机对水色图片进行分类能够达到更理想的效果，准确率为97.56%。