目录

[一、背景介绍 2](#_Toc37225779)

[1、项目背景 2](#_Toc37225780)

[2、项目目标 2](#_Toc37225781)

[二、数据探索与预处理 2](#_Toc37225782)

[1、截取图像的有效区域 2](#_Toc37225783)

[2、计算三个颜色通道的颜色矩 2](#_Toc37225784)

[3、建立数据集 3](#_Toc37225785)

[三、模型训练与预测 3](#_Toc37225786)

[1、决策树模型简介 3](#_Toc37225787)

[2、利用决策树训练模型与预测 3](#_Toc37225788)

[四、模型评估 3](#_Toc37225789)

# 一、背景介绍

## 1、项目背景

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、项目目标

利用203张不同罗非鱼池塘水样图像数据，建立合适的模型对水色图像进行识别。

# 二、数据探索与预处理

## 1、截取图像的有效区域

文件中给出的图片不仅仅包含水质信息，还包含了拍摄水样时的周围环境，为了防止图片中周围环境对结果的影响，获取图像中反应仅水质的有效区域，我们用python查看图片，截取图像的中间100行、100列。截取效果如下图所示：



图 1 images文件夹中图片1\_1.jpg



图 2 截取出的图片1\_1.jpg的有效区域

我们可以看出，截取出的图片的有效区域仅包含水质信息，排出了周围环境对结果的影响。

## 2、计算三个颜色通道的颜色矩

颜色矩是反映水质图像特征的数矩，颜色矩的含义及计算公式为：

一阶颜色矩：采用一阶原点矩，反映了图像的整体明暗程度。计算公式为：



二阶颜色矩：采用二阶中心矩的平方，反映了图像颜色的分布范围。计算公式为：



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。计算公式为：



首先、我们把截取出的有效图片数据划分为RGB三个颜色通道，分别将三个颜色通道的图片数据转换为像素值矩阵。其次，我们在Python中自定义计算三阶颜色矩的函数，分别计算三个颜色通道的一阶颜色矩、二阶颜色矩、三阶颜色矩。这样，我们就得到了每张图片三个颜色通道每个颜色通道三个颜色矩（共九个数据）的特征数据。

## 3、建立数据集

首先，为了正确获取指定路径中的所有图片，我们自定义函数，获取images文件夹中所有以数字1-5开头，以’.jpg’结尾的图片名称。数据信息如下表所示：

|  |  |
| --- | --- |
| 水质标签 | 图片数量 |
| 1 | 51 |
| 2 | 44 |
| 3 | 78 |
| 4 | 24 |
| 5 | 6 |
| 总共 | 203 |

其次，利用循环语句，截取所有图片的中间100行、100列。部分截取结果如下所示（数据详见：/result/images\_1）：

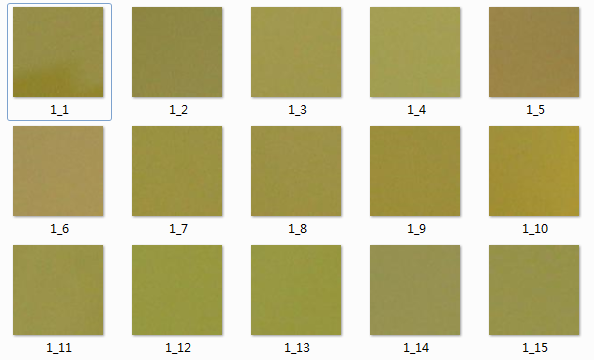


图3 部分截取结果显示

查看所有的截取结果，所有的截取结果只包含水质信息，无周围环境信息。

最后，利用循环语句，计算所有图片的9个特征数据，同时以图片名称的第一个数字作为图片标签，建立204张图片特征数据与标签的数据集（数据详见：/result/result.json），用于模型的训练与预测。

# 三、模型训练与预测

## 1、决策树模型简介

分类决策树模型是一种描述对实例进行分类的树形结构。决策树由节点和有向边组成。节点有两种类型：内部节点和叶节点。内部节点表示一个特征或属性，叶节点表示一个类。用决策树分类，从根节点开始，对实例的某一特征进行测试，根据测试结果，将实例分配到其子节点，这时，每个子节点对应着该特征的一个取值。如此递归地对实例进行测试并分配，直到达到叶节点，将实例分到叶节点的类中。

## 2、利用决策树训练模型与预测

我们将第一步得到的数据集拆分为训练集与测试集，其中，训练集占总数据的80%，测试集占总数据的20%。我们使用决策树分类模型，利用训练集训练模型，然后利用模型预测测试数据标签。

# 四、模型评估

分类指标的文本报告如下表所示：

表一 分类指标的文本报告

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | precision | recall | | f1-score | | support | | |
| 1 | 0.57 | | 0.67 | | 0.62 | | 6 | | |
| 2 | 0.83 | | 0.91 | | 0.87 | | 11 | | |
| 3 | 0.94 | | 0.89 | | 0.91 | | 18 | | |
| 4 | 1.00 | | 0.80 | | 0.89 | | 5 | | |
| 5 | 1.00 | | 1.00 | | 1.00 | | 1 | | |
| avg/total | 0.87 | | 0.85 | | 0.86 | | 41 |

从各类的准确率、召回率、F1 score可以看出，模型的分类效果较好。

混淆矩阵为：



从混淆矩阵可以看出，第一类水质图片中有一张被分到第二类，一张被分到第三类，、；第二类水质图片中有一张被分到第一类；第三类水质图片中有两张被分到第一类；第四类水质图片中有一张被分到第二类。