**基于水色图像的水质评论报告**

目录

[1 背景介绍 1](#_Toc36758544)

[1.1 行业背景 1](#_Toc36758545)

[1.2 分析目标 1](#_Toc36758546)

[2 数据说明 2](#_Toc36758547)

[3 问题分析 2](#_Toc36758548)

[4 数据预处理 4](#_Toc36758549)

[4.1 图像切割 4](#_Toc36758550)

[4.2 特征提取 5](#_Toc36758551)

[5 分类模型的建立与分析 6](#_Toc36758552)

[5.1 CART决策树 6](#_Toc36758553)

[5.1.1 算法原理 7](#_Toc36758554)

[5.1.2 解决问题的步骤 7](#_Toc36758555)

[5.2 神经网络 8](#_Toc36758556)

[5.2.1 算法原理 8](#_Toc36758557)

[5.2.2 解决问题的步骤 9](#_Toc36758558)

[5.3 SVM支持向量机 10](#_Toc36758559)

[5.3.1 算法原理 10](#_Toc36758560)

[5.3.2 解决问题的步骤 10](#_Toc36758561)

[6 总结 12](#_Toc36758562)

# 背景介绍

## 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。

## 分析目标

使用池塘图片作为数据，专家对池塘图片的水色分为五个等级作为标签，利用图像处理技术，实现水质的自动评价。

# 数据说明

数据来源于某地区的多个罗非鱼池塘的水样数据，包含水产专家按水色判断水质分类的数据（见表 2‑1）以及用数码相机按照标准进行水色采集的数据（见图 2‑1），每个水质图片命名规则为“类别-编号.jpg”，如“1\_1.jpg”说明当前图片属于第1类的样本

表 2‑1水色分类

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 水色 | 浅绿色（清水或浊水） | 灰蓝色 | 黄褐色 | 茶褐色  （姜黄、茶褐、红褐、褐中带绿等） | 绿色（黄绿、油绿、蓝绿、墨绿、绿中带褐等） |
| 水质类别 | 1 | 2 | 3 | 4 | 5 |

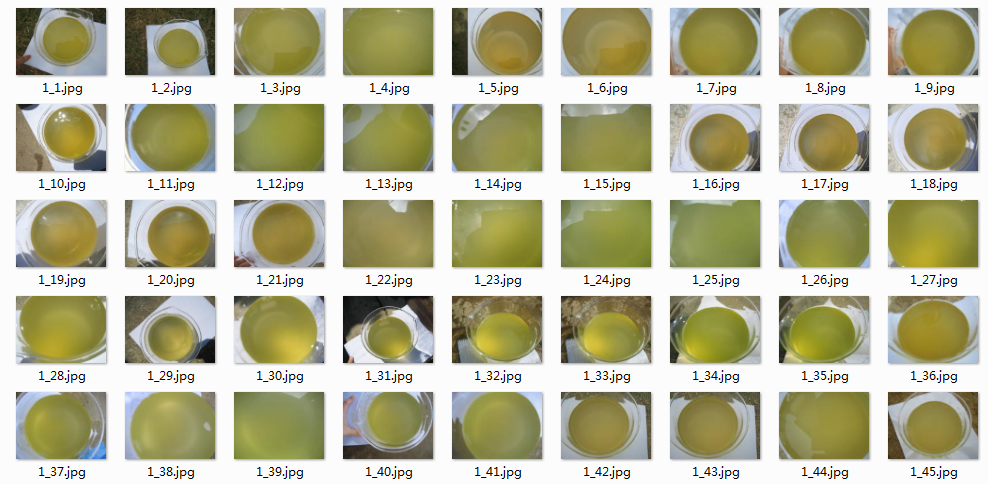


图 2‑1标准条件下拍摄的水样图像

# 问题分析

通过拍摄水样，采集得到水样图像，而图像数据的维度过大，不容易分析，需要从中提取水样图像的特征，提取反映图像本质的一些关键指标，以达到自动进行图像识别或分类的目的。显然，图像特征提取是图像识别或分类的关键步骤，图像特征提取的效果如何直接影响到图像识别和分类的好坏。

图像特征主要包括有颜色特征、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。颜色特征是一种全局特征,描述了图像或图像区域所对应的景物的表面性质。一般颜色特征是基于像素点的特征，所有属于图像或图像区域的像素都有各自的贡献。在利用图像的颜色信息进行图像处理、识别、分类的研究中，在实现方法上已有大量的研究成果，主要采用颜色处理常用的直方图法和颜色矩方法等。

颜色直方图是最基本的颜色特征表示方法，它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。

基于颜色矩提取图像特征的数学基础在于图像中任何的颜色分布均可以用它的矩来表示。根据概率论的理论，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，对于一副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。

颜色直方图产生的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，在本案例中选择采用颜色矩来提取水样图像的特征，即建立水样图像与反映该图像特征的数据信息关系，同时由有经验的专家对水样图像根据经验进行分类，建立水样数据信息与水质类别的专家样本库，进而构建分类模型，得到水样图像与水质类别的映射关系，并经过不断调整系数优化模型，最后利用训练好的分类模型，用户就能方便地通过水样图像，自动判别出该水样的水质类别。图 3‑1为基于水色图像特征提取的水质评价流程，主要包括以下步骤：

1. 从采集到的原始水样图像中进行选择性抽取与实时抽取形成建模数据和增量数据；
2. 对（1）形成的两个数据集进行数据预处理，包括图像切割和颜色矩特征提取；
3. 利用（2）形成的已完成数据预处理的建模数据，由有经验的专家对水样图像根据经验进行分类，构建专家样本；
4. 利用（3）的专家样本构建分类模型；
5. 利用（4）的构建好的分类模型进行水质评价。

图 3‑1 基于水色图像特征提取的水质评价流程

# 数据预处理

## 图像切割

采集到的水样图像包含盛水容器，容器的颜色与水体颜色差异较大，同时水体位于图像中央，为了提取水色的特征，需要提取水样图像中央部分具有代表意义的图像，具体实施方式是提取水样图像中央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，长从第个像素点到第个像素点的子图像。

即把图 4‑1中左边的切割前的水样图像切割并保存到右边的切割后的水样图像。

图 4‑1 切割前水样图像（左）和切割后水样图像（右）

## 特征提取

选择采用颜色矩来提取水样图像的特征，下面给出各阶颜色矩的计算公式：

（1）一阶颜色矩

一阶颜色矩采用一阶原点矩，反映了图像的整体明暗程度，如（式 2‑1）所示。



（式 2‑1）

其中是在第个颜色通道的一阶颜色矩，对于RGB颜色空间的图像，，是第个像素的第个颜色通道的颜色值。

（2）二阶颜色矩

二阶颜色矩采用的是二阶中心距的平方根，反映了图像颜色的分布范围，如（式 2‑2）所示。



（式 2‑2）

其中是在第个颜色通道的二阶颜色矩，是在第个颜色通道的一阶颜色矩。

（3）三阶颜色矩

三阶颜色矩采用的是三阶中心距的立方根，反映了图像颜色分布的对称性，如（式 2‑3）所示。



（式 2‑3）

其中是在第个颜色通道的三阶颜色矩，是在第个颜色通道的一阶颜色矩。

对切割后的图像提取其颜色矩，作为图像的颜色特征。颜色矩的提取，并且提取每个文件名中的类别和序号，同时针对所有的图片都进行同样的操作，最后对203张图片处理后的部分结果（保留4位小数）如表 4‑2所示。

表 2‑1 水色图像特征与相应的水色类别的部分数据

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 水质类别 | R通道一阶矩 | R通道二阶矩 | R通道三阶矩 | G通道一阶矩 | G通道二阶矩 | G通道三阶矩 | B通道一阶矩 | B通道二阶矩 | B通道三阶矩 |
| 1. | 148.6041 | 3.6259 | -3.2171 | 138.6381 | 4.1255 | -4.096 | 64.3694 | 10.485 | -10.562 |
| 1. | 163.6788 | 3.9234 | 2.4431 | 145.5487 | 2.835 | -0.9631 | 54.4581 | 3.479 | -0.9334 |
| 1 | 153.9485 | 2.2088 | -1.2031 | 147.081 | 1.8034 | -0.6695 | 71.9576 | 3.1155 | -2.4168 |
| 1. | 150.3755 | 2.0152 | 0.8614 | 151.3985 | 1.514 | -0.8663 | 64.3118 | 2.6989 | -1.3439 |
| 1. | 150.738 | 1.9029 | -0.3425 | 150.9738 | 1.658 | -0.5528 | 64.6246 | 3.098 | -1.218 |
| 1. | 150.1358 | 1.9201 | -0.9325 | 145.269 | 1.2816 | 0.4698 | 81.3181 | 2.1438 | -1.4855 |
| 1. | 150.0713 | 1.8609 | 0.7906 | 146.1766 | 1.2962 | -0.2269 | 73.5234 | 2.967 | 2.1784 |
| 1. | 157.623 | 2.1804 | -0.6314 | 137.6215 | 1.5738 | 0.9361 | 70.0067 | 3.713 | 3.5668 |
| 1. | 157.7101 | 2.2694 | 0.7963 | 137.5022 | 1.4904 | 0.8447 | 69.7587 | 3.6096 | 3.6955 |
| 1. | 169.3828 | 2.7948 | -1.3436 | 146.971 | 1.9686 | 1.0537 | 74.9072 | 4.8387 | 3.2475 |
| 1. | 178.6782 | 2.1321 | 0.4758 | 153.3402 | 1.8166 | 0.4693 | 74.7462 | 3.367 | 1.3843 |
| 1. | 143.7789 | 2.4835 | -1.3442 | 136.4161 | 1.9944 | 0.7998 | 69.3224 | 3.0158 | -1.3898 |
| 1. | 158.6412 | 2.5609 | -0.5713 | 139.8721 | 1.472 | -0.6618 | 67.8988 | 3.0389 | -2.2942 |
| 1. | 155.9917 | 2.5006 | 1.5021 | 134.7518 | 1.7422 | 1.0453 | 60.818 | 3.7358 | 1.9783 |

# 分类模型的建立与分析

数据挖掘的目的是为了利用水质图像数据，将水质分为5个等级，这显然是一个分类预测的问题，因此，本文将使用决策树、神经网络和SVM这三种模型来解决这个问题。

## CART决策树

### 算法原理

在决策树算法中，分类与回归树 CART（Classifi－cation and Regression Trees）算法[1]是一种十分有效的非参数分类和回归方法，它通过对由测试变量和目标变量构成的训练数据集的循环分析,构建二叉树达到预测的目的。CART算法采用经济学中的基尼系数(Gini Index)作为选择最佳测试变量和分割阈值的准则。基尼系数的定义如下:





式中:是从训练样本集中随机抽取一个样本,当某一测试变量值为时属于第类的概率;为训练样本中测试变量值为时属于第类的样本个数;为训练样本中该测试变量值为的样本个数; 为类别个数。但按照上述过程生成的完整决策树往往会出现“过度拟合”现象,因此有必要对树的结构进行修剪。

CART算法具有如下特点:严格无参数,对输入数据没有任何统计分布的假设要求;能够清楚地指出变量对于分类的重要性,选择与分类相关的变量;方法实现简单,运行速度较快;结构清晰,容易理解等。

### 解决问题的步骤

#### 选取训练集和测试集的数据

选择图片的RGB三个颜色通道的一阶矩，二阶矩，三阶矩这9个特征进行决策树的构建，同时随机选择3/4的数据作为训练集，1/4的数据作为测试集。

#### 决策树的构建与寻优

因为水质分类是一个分类问题，因此，构建一个分类决策树，决策树的最大深度这里设置为1到100，并利用网格搜索法找出一个使模型准确率最高的最大深度。网格搜索法：网格搜索其实是一种暴力搜索参数的方法，它通过我们指定不同的超参列表进行穷举搜索，并计算每一个超参组合对于模型性能的影响，来获取最优的超参组合。利用网格搜索法我们得到模型最优的最大深度为21，对训练数据的准确率为0.858。

#### 模型分析

我们利用最优的模型对测试数据进行预测，预测结果的准确率为0.78，预测结果的混淆矩阵见表 5‑1（注意，由于用随机函数来打乱数据，因此重复试验所得到的结果可能有所不同。）

表 5‑1 决策树模型水质评价的混淆矩阵

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 实际值 预测值 | 1 | 2 | 3 | 4 | 5 |
| 1 | 8 | 2 | 3 | 0 | 0 |
| 2 | 1 | 7 | 1 | 0 | 0 |
| 3 | 1 | 0 | 14 | 0 | 0 |
| 4 | 0 | 0 | 0 | 2 | 1 |
| 5 | 0 | 0 | 0 | 0 | 1 |

## 神经网络

### 算法原理

BP神经网络是一种按误差逆传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一。BP网络能学习和存贮大量的输入-输出模式映射关系，而无需事前揭示描述这种映射关系的数学方程。它的学习规则是使用最速下降法，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小。BP神经网络模型拓扑结构包括输入层（input）、隐层(hide layer)和输出层(output layer)。

神经网络的总体步骤大致如下所示：

（1）权值初始化、，其中表示网络输入层单

元到隐含层单元的连接权值,表示网络隐含层单元到输出层单元的连接权值。

（2）依次输入个学习样本。设当前输入为第个样本。

（3）依次计算各层的输出：、。其中为隐含层上第 个神经元的输出。为输出层上第个神经元的输出。

（4）求得网络各层的反传误差：





并记下各个 、的值

（5）记录已经学习过的样本个数。如果 ,转到步骤（2）继续计算如果,转到步骤（6）。

（6）按权值修正公式修正各层的权值和阈值。

（7）新的权值再计算 、和,若对每个样本和相应的第个输出神经元,都满足同,或达到最大学习次数,则终止学习。否则转到步骤（2）继续新一轮的网络学习。

### 解决问题的步骤

#### 选取训练集和测试集的数据

选择图片的RGB三个颜色通道的一阶矩，二阶矩，三阶矩这9个特征进行决策树的构建，同时随机选择3/4的数据作为训练集，1/4的数据作为测试集。

#### 神经网络的构建与寻优

构建一个神经网络分类器，设置神经网络的参数，一：隐藏层个数及各隐藏层中神经元的个数，二：设置神经网络的迭代次数。隐藏层个数及各隐藏层神经元个数设置三种情况（100,50）、（500,300）与（100,50,20）（注：(100, 50)，表示有两层隐藏层，第一层隐藏层有100个神经元，第二层有50个神经元。）。迭代次数设置为3种情况，100次，500次与1000次。使用网格搜索法对这些参数的不同组合进行训练，最终得到准确率最高的模型参数为：2层隐藏层，隐藏层神经元个数为（500,300），迭代次数为100，该模型对训练集进行预测，预测的准确率为0.951。

#### 模型分析

利用最优的模型对测试数据进行预测，预测结果的准确率为1，预测结果好，预测结果的混淆矩阵见表 5‑2（注意，由于用随机函数来打乱数据，因此重复试验所得到的结果可能有所不同。）

表 5‑2 神经网络模型水质评价的混淆矩阵

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 实际值 预测值 | 1 | 2 | 3 | 4 | 5 |
| 1 | 13 | 0 | 0 | 0 | 0 |
| 2 | 0 | 9 | 0 | 0 | 0 |
| 3 | 0 | 0 | 15 | 0 | 0 |
| 4 | 0 | 0 | 0 | 3 | 0 |
| 5 | 0 | 0 | 0 | 0 | 1 |

## SVM支持向量机

### 算法原理

支持向量机(Support Vector Machine)是一种分类模型。它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。支持向量机方法是建立在统计学习理论的VC 维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度，Accuracy）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折衷，以期获得最好的泛化能力。

算法步骤：

1. 准备数据集，转化为 SVM支持的数据格式
2. 对数据进行简单的标准化操作；
3. 考虑选用核函数（通常选取径函数）；
4. 采用交叉验证（一般采用5折交叉验证），选择最佳参数C与gamma；
5. 用得到的最佳参数C与gamma对整个训练集进行训练得到SVM模型；
6. 用得到的SVM模型对测试样本进行测试。

### 解决问题的步骤

#### 数据标准化

选择图片的RGB三个颜色通道的一阶矩，二阶矩，三阶矩这9个特征进行决策树的构建，发现特征的取值范围差别较大，也就是说如果直接输入SVM模型的话，可能会导致模型精确度下降，因此，在建模之前需要将数据进行标准化，因为颜色矩的像素值在0到255之间，所以将所有数据除以255就可达到标准化效果，从而提高区分度和准确率。对标准化后的样本进行抽样，抽取80%作为训练样本，剩下的20%作为测试样本，用于水质评价检验。标准化后各参数的取值范围如表5‑3

表 5‑3 预测模型输入变量

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **变量名称** | **变量描述** | **取值范围** |
| 1 | R通道一阶矩 | 水样图像在R颜色通道的一阶矩 | 0～1 |
| 2 | R通道二阶矩 | 水样图像在R颜色通道的二阶矩 | 0～1 |
| 3 | R通道三阶矩 | 水样图像在R颜色通道的三阶矩 | -1～1 |
| 4 | G通道一阶矩 | 水样图像在G颜色通道的一阶矩 | 0～1 |
| 5 | G通道二阶矩 | 水样图像在G颜色通道的二阶矩 | 0～1 |
| 6 | G通道三阶矩 | 水样图像在G颜色通道的三阶矩 | -1～1 |
| 7 | B通道一阶矩 | 水样图像在B颜色通道的一阶矩 | 0～1 |
| 8 | B通道二阶矩 | 水样图像在B颜色通道的二阶矩 | 0～1 |
| 9 | B通道三阶矩 | 水样图像在B颜色通道的三阶矩 | -1～1 |
| 10 | 水质类别 | 不同类别能表征水中浮游植物的种类和多少 | 1，2，3，4，5 |

#### svm模型的构建与寻优

构建一个SVM模型，设置SVM模型的核函数为线性核函数时，正则化参数取值为0.0001、0.001、0.01、0.1、1.0、10.0、100.0、1000.0这8种情况，当设置SVM模型的核函数为RBF核函数时，正则化参数取值为0.0001、0.001、0.01、0.1、1.0、10.0、100.0、1000.0这8种情况，gamma参数取值为0.0001、0.001、0.01、0.1、1.0、10.0、100.0、1000.0这8种情况。利用网格搜索法对这几种参数组合进行训练，最终得到准确率最高的模型参数为：RBF为核函数、正则化系数为1000,和gamma系数为10，该模型对训练集进行预测，预测的准确率为0.981。

#### 模型分析

利用最优的模型对测试数据进行预测，预测结果的准确率为0.951，预测结果较好，预测结果的混淆矩阵见表 5‑4（注意，由于用随机函数来打乱数据，因此重复试验所得到的结果可能有所不同。）

表 5‑4 SVM模型水质评价的混淆矩阵

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 实际值 预测值 | 1 | 2 | 3 | 4 | 5 |
| 1 | 12 | 0 | 1 | 0 | 0 |
| 2 | 0 | 8 | 0 | 0 | 1 |
| 3 | 0 | 0 | 15 | 0 | 0 |
| 4 | 0 | 0 | 0 | 3 | 0 |
| 5 | 0 | 0 | 0 | 0 | 1 |

# 总结

最后，对三个模型在对该数据集的预测判断上的优劣进行比较。

**准确性**：模型对测试数据集预测的准确性。

1. **神经网络**

准确率为：0.854

1. **神经网络**

准确率为：1.0

1. **SVM支持向量机**

准确率为：0.951

明显可以看出，就准确率而言，神经网络>SVM支持向量机>决策树。

**时效性**：模型对最优参数的寻找及模型训练的耗时

1. **CART决策树**

耗时为：1.034秒

1. **神经网络**

耗时为：342.688秒

1. **SVM支持向量机**

耗时为：13.106秒

而就时效性而言，决策树>SVM支持向量机>神经网络。

因为决策树模型的准确率较低，故不建议选择该模型进行水质等级分类。当你想要模型运行效率高，但对准确率要求不高时，可以选择SVM支持向量机模型。如果对模型预测的准确率要求高，则选择神经网络模型。综上比较只是简单粗略的比较，没有对所有因素进行严格控制，并且模型参数较多，没有广泛进行调试，存在着许些不足，有待进一步优化。