人工智能导论

机器学习方法实现树叶枯竭情况的量化

实验报告

20191620310024 孟 浩

20191620310046 张沈晖

20191620310149 陈时彦

**目录**

[1 问题背景 3](#_Toc11186)

[2 问题分析 3](#_Toc10193)

[3 数据集获取及标注 4](#_Toc24033)

[4 特征工程 5](#_Toc19096)

[5 模型选取与训练 8](#_Toc27952)

[6 模型拟合 9](#_Toc19172)

[7 下一步工作 9](#_Toc8773)

[8 实验总结 10](#_Toc18392)

# 1 问题背景

当前国内的林业保护工作，主要以人工进行为主。通过森林巡护监察人员对林区树木进行肉眼观察等传统的检查，来登记可疑病死木、枯死木。检查周期长且耗费人力物力。急需一种智能化的检测手段来降低这种人力物力的损耗。

同时，由于对树木的检查主要是以肉眼观察为主，有一些树木生长过程中的病症难以在初期得到很好地排查（例如维束管病症，患病树木往往只有部分枝条枯萎而其他枝条发育正常），为树木之后的健康生长埋下了隐患。需要智能化自动化程度高的病症检测方案来降低人工病症检测的失误率提高准确率。为林业健康发展保驾护航。

国内目前针对树叶枯竭量化的研究尚少，针对本课题的研究可以为推动相关工作的发展。

# 2 问题分析

## 2.1 原理介绍

机器学习模型是图像识别分类的主要研究方法之一，本文将其应用于树叶枯竭情况的量化研究。所使用方法主要包括图像预处理、RGB特征提取、决策树回归等。

## 2.2 图像预处理

为了去除原始数据集中的图像中的无用噪声等信息，增加感兴趣区域的信息密度，提高特征提取的质量，我们需要对原始图像库进行预处理。本方法做的工作主要有最大轮廓矩阵获取、超绿色法处理与OTSU二值化。具体效果如图1所示。

## 2.3 特征选取

根据对题目的研判，考虑到常绿落叶阔叶林树木的枯竭程度应主要从两个方面进行判断，一是树木的颜色，二是树木枝干是否茂密。因此树木枯竭程度的主要特征为树叶新鲜程度和树叶稀疏程度。

# 3 数据集获取及标注

## 3.1 数据集来源

本方法所使用的树木、树叶图像来源于网络上公开的图像数据集、谷歌图片等。为了使模型对常绿落叶阔叶林具有一定普适性，我们选取的训练集图像都是常绿落叶阔叶林树木。如图2所示。



图1 常绿落叶阔叶树木示例

## 3.2 数据集标注

数据标记是使用标签或标签手动注释内容的过程。我们将添加这些标签的人称为贴标签者。在计算机视觉领域，标签标识图像中的元素。然后将注释数据用于监督学习。根据用例，有不同类型的数据标记。在对象检测任务中，我们不仅要知道目标对象的类别，还要知道它们的位置。因此，我们在图像中的目标对象周围绘制边界框。还有图像分类、语义分割和实例分割任务。我们用类、分割图和实例分割图进行标记，如下图所示。

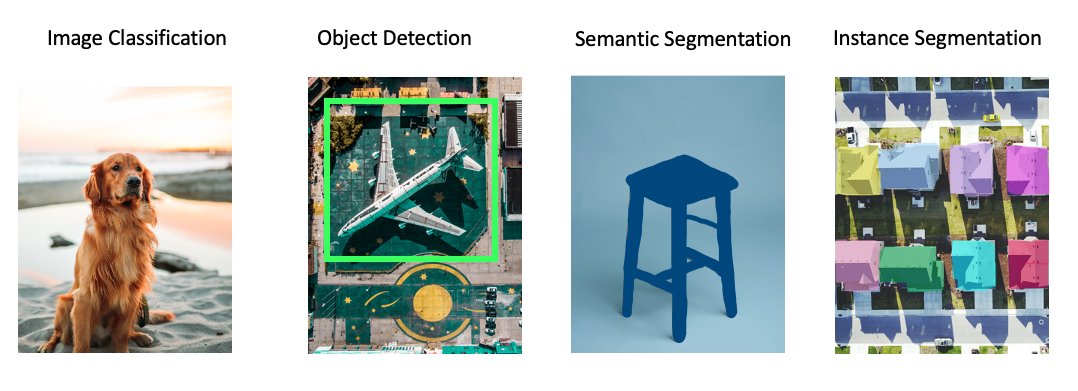


图2 计算机视觉任务的常见标签类型

本方法中，我们对训练集中每张图片依据事先定好的较为详细的标准进行认为评估。评分标准如下表。

表1 枯竭程度视觉标准



表2 枯竭程度评分标准



# 4 特征工程

通过问题分析，我们可以将树叶枯竭程度的特征提取分为两个方面：树叶新鲜程度和树叶稀疏程度。树叶的新鲜程度主要表现在树叶轮廓的颜色方面，树叶的稀疏程度在业界并没有明确的研究，通过相关论文的参考与分析，我们将树叶稀疏程度做如下的定义：一副图像中树叶与树干占整颗树或是整条树枝的面积比。便从两方面入手进行特征工程的开展。

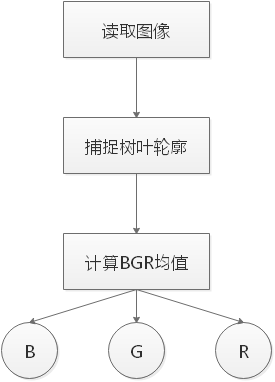


图3 树叶新鲜程度的特征提取流程

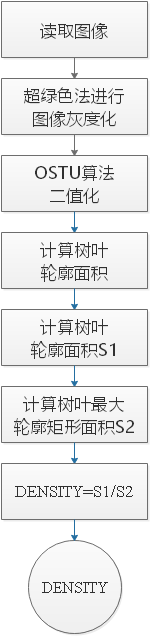


图4 树叶稀疏程度的特征提取流程

使用Python和OpenCV按照上述流程进行实现，最终实现效果如下：

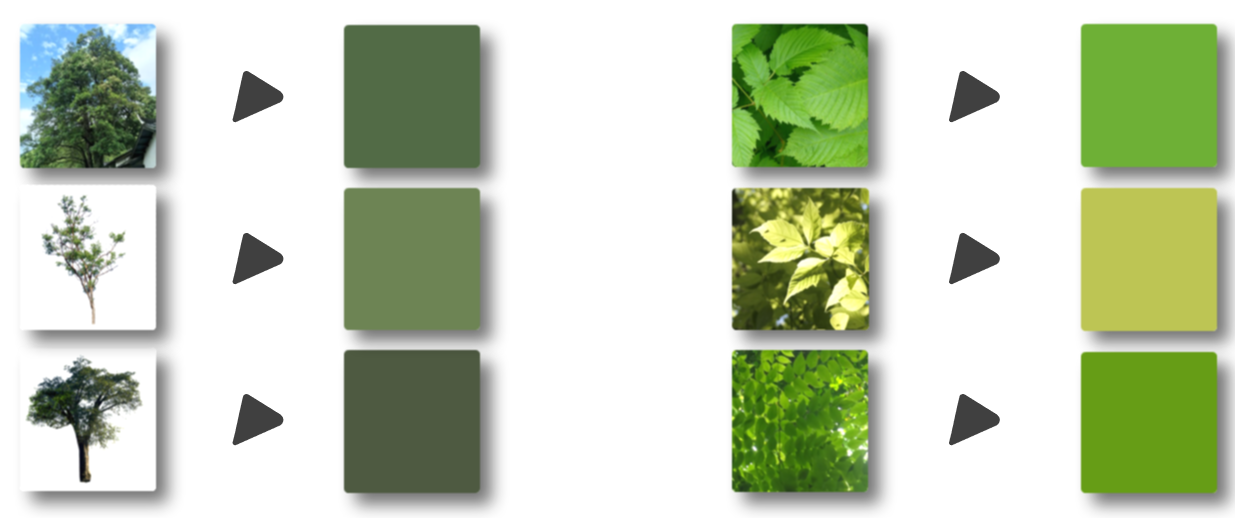


图5 从树叶图片中提取RGB颜色信息





图6 树叶图片灰度处理提取轮廓

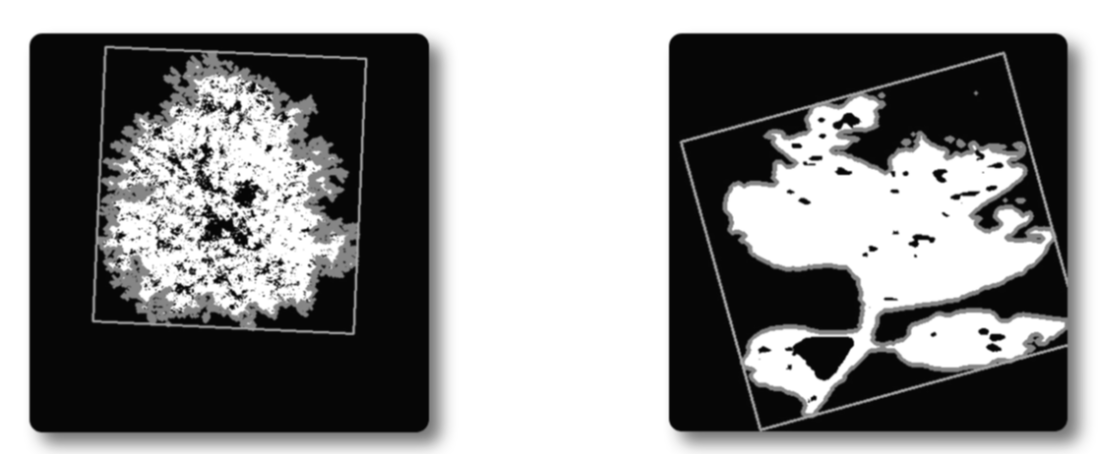


图7 获取到的树叶最大轮廓矩形

表3 特征提取得到的数据

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **JPG\_ID** | **B** | **G** | **R** | **DENSITY** | **RESULT** |
| 1 | 69.96375911 | 151.5825274 | 93.19723534 | 0.824978094 | 4 |
| 2 | 53.5056871 | 180.2731527 | 109.084952 | 0.801192036 | 5 |
| 4 | 57.4802039 | 163.2877425 | 87.61880892 | 0.586446654 | 4 |
| 5 | 20.89874334 | 173.5697266 | 140.3937032 | 0.235607001 | 3 |
| 6 | 36.92194579 | 173.275419 | 140.5898323 | 0.930413044 | 3 |
| 7 | 54.97282531 | 160.8509379 | 92.29514021 | 0.786621485 | 3 |
| 9 | 8.727287052 | 165.9701125 | 136.9945645 | 0.438902792 | 3 |

通过实验代码获取每一张树叶图片的特征数值，将其存入数据表当中，供模型训练使用，部分数据列举如下表：

将特征数据进行异常值处理（如存在缺失值、稀疏程度比率大于1等）和数据标准化，处理过的数据如下表：

表4 特征数据处理后所得数据

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **JPG\_ID** | **B** | **G** | **R** | **DENSITY** | **RESULT** |
| 1 | 0.3346 | 0.6246 | 0.2834 | 0.8231 | 4 |
| 2 | 0.2541 | 0.7848 | 0.3637 | 0.7991 | 5 |
| 4 | 0.2736 | 0.6900 | 0.2553 | 0.5820 | 4 |
| 5 | 0.0947 | 0.7474 | 0.5218 | 0.2274 | 3 |
| 6 | 0.1731 | 0.7457 | 0.5228 | 0.9297 | 3 |
| 7 | 0.2613 | 0.6764 | 0.2789 | 0.7843 | 3 |
| 9 | 0.0352 | 0.7050 | 0.5046 | 0.4329 | 3 |

# 5 模型选取与训练

我们将量化问题看作是一个回归问题，对于几个特征数值的变化，评分越高代表枯竭程度越低，作为分类问题是不合适的，原因是五档评分之间每一个评分之间是有阶梯关系而不是毫无关系。

对于回归问题，在机器学习算法中常见的解决模型是：

表5 常见机器学习回归方法模型列举

|  |  |
| --- | --- |
| **线性回归模型** | 线性回归模型是一种利用数理统计中回归分析来确定变量之间的相关关系的一种数学回归模型。 |
| **lasso回归模型** | 该方法是一种压缩估计。它通过构造一个惩罚函数得到一个较为精炼的模型，使得它压缩一些回归系数，即强制系数绝对值之和小于某个固定值；同时设定一些回归系数为零。 |
| **决策树回归模型** | 决策树模型是一种简单易用的非参数分类器，当决策树用于分类时称为分类树，用于回归时称为回归树。 |
| **SVM回归模型** | SVM，又称支持向量机，是一类按[监督学习](https://baike.baidu.com/item/%E7%9B%91%E7%9D%A3%E5%AD%A6%E4%B9%A0/9820109" \t "https://baike.baidu.com/item/%E6%94%AF%E6%8C%81%E5%90%91%E9%87%8F%E6%9C%BA/_blank)方式对数据进行[二元分类](https://baike.baidu.com/item/%E4%BA%8C%E5%85%83%E5%88%86%E7%B1%BB/15635322" \t "https://baike.baidu.com/item/%E6%94%AF%E6%8C%81%E5%90%91%E9%87%8F%E6%9C%BA/_blank)的广义线性分类器，其[决策边界](https://baike.baidu.com/item/%E5%86%B3%E7%AD%96%E8%BE%B9%E7%95%8C/22778546" \t "https://baike.baidu.com/item/%E6%94%AF%E6%8C%81%E5%90%91%E9%87%8F%E6%9C%BA/_blank)是对学习样本求解的最大边距超平面。 |
| **KNN回归模型** | KNN法即K最邻近法，如果一个样本在特征空间中的K个最相似的样本中的大多数属于某一个类别，则该样本也属于这个类别。使用K近邻进行回归时，预测结果为K个邻居的平均值。 |
| **随机森林回归模型** | 随机森林属于Bagging类算法，建立多个决策树，当需要对某个样本进行预测时，统计森林中的每棵树对该样本的预测结果，然后通过投票法从这些预测结果中选出最后的结果。 |

使用python当中的scikit-learn机器学习框架对模型进行训练，以0.85:0.15的比例将原始数据集分割成训练集和测试集，分别采取六种模型在相同的预处理和分割情况下做训练，得到的模型拟合优度如下表所示。

表6 各模型的训练拟合优度

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **模型** | 线性回归 | lasso回归 | 决策树 | SVM | KNN | 随机森林 |
| **拟合优度** | 0.47 | 0.35 | 0.63 | 0.32 | 0.44 | 0.55 |

# 6 模型拟合

由几种模型的训练结果可知，随机森林回归模型的效果最好，即将其作为我们最终的模型进行新的树叶图片的枯竭程度拟合，具体拟合样例如下：



图8 对于新的树叶图片预测枯竭程度

# 7 下一步工作

本方法基于机器视觉和图像处理技术，使用机器学习方法建模实现对树叶枯竭情况的评估与量化，实现了计算机视觉自动化评估枯竭程度，但仍然存在一些问题。实验中，机器学习模型对树叶图像拟合的正确率未达到预期，精度仍要继续完善。

首先，要继续完善的是数据集，本方法使用的数据集含有较多干扰，质量较低。其次，对于数据集的标注工作也需要指定一个更加规范的标准，避免人为因素对模型的干扰。第三，可以基于大数据库应用深度学习模型，例如卷积神经网络(convolutional neural network, CNN)在预测人类感知方面表现出色，下一步可以使用CNN神经网络实现对图片进行分类与回归。

# 8 实验总结

我们的以量化树叶枯竭程度为目标，采用机器学习的方法，通过爬虫爬取、人力搜集、人工打标的方式清洗、整理出了1000张树叶的数据集。采用超绿色图像灰度法、OSTU算法进行树叶新鲜程度和疏密程度两种主要特征的提取。选取了包括线性回归、随机森林回归、KNN回归在内的6种不同的模型，对这些模型训练的效果以拟合优度为指标做了横向对比。最后选取了决策树模型训练的参数。

从效果来看，我们的实验并不是特别理想。主要有一下两个原因：1、数据集质量不过关，且是人工打标，存在主观性。2、数据集数量较少，导致模型训练过程中信息素不足且离散程度较大。

基于以上两点，我们下一步工作会继续进行数据集的收集，收集100000张图片做为数据集。并采用编写自动化脚本的方式进行打标，去除人工打标的主观性。同时采取CNN神经网络的深度学习算法，优化模型效果。