|  |
| --- |
| **机器学习** |
| **XX案例名** |
| 主 研 人：  参 研 人： |
| 审 核 人：  方 向：工业算法案例研究  版 本 号：A |
|  |

**版本更新**

|  |  |  |  |
| --- | --- | --- | --- |
| **日期** | **版本** | **更新描述** | **作者** |
| 2019/3/1 | A | 初稿 | 董乐诗 |

**选题表格**

|  |  |  |  |
| --- | --- | --- | --- |
| 时间 | 竞赛名 | 竞赛背景描述（50字以内） | 类型（分类/回归） |
| 2019/2/26 | Marinexpl-ore和康奈尔大学鲸鱼检测挑战赛 | 在导航技术计算最佳运输路线实现商业可持续性的同时，如何减少海洋运输累积噪音“脚印”对鲸鱼的影响值得关注。 | 分类 |

目录

[1. 背景描述](#_Toc6117_WPSOffice_Level1) [4](#_Toc6117_WPSOffice_Level1)

[1.1 竞赛赛题描述](#_Toc18630_WPSOffice_Level2) [4](#_Toc18630_WPSOffice_Level2)

[1.2 评估指标描述](#_Toc16215_WPSOffice_Level2) [4](#_Toc16215_WPSOffice_Level2)

[2. 数据来源及描述性统计分析](#_Toc18630_WPSOffice_Level1) [5](#_Toc18630_WPSOffice_Level1)

[2.1 大赛数据来源](#_Toc26629_WPSOffice_Level2) [5](#_Toc26629_WPSOffice_Level2)

[2.2 数据的描述性统计](#_Toc25354_WPSOffice_Level2) [5](#_Toc25354_WPSOffice_Level2)

[2.2.1 数据基本情况描述：](#_Toc18630_WPSOffice_Level3) [5](#_Toc18630_WPSOffice_Level3)

[2.2.2 数据字段介绍：](#_Toc16215_WPSOffice_Level3) [6](#_Toc16215_WPSOffice_Level3)

[2.2.3 数据描述性统计](#_Toc26629_WPSOffice_Level3) [6](#_Toc26629_WPSOffice_Level3)

[第一名得主计算了每个频率切片的 ：质心，宽度，倾斜和总变化。](#_Toc1088_WPSOffice_Level2) [6](#_Toc1088_WPSOffice_Level2)

[3. 优秀算法思路](#_Toc16215_WPSOffice_Level1) [6](#_Toc16215_WPSOffice_Level1)

[3.1 方案一](#_Toc21881_WPSOffice_Level2) [6](#_Toc21881_WPSOffice_Level2)

[3.1.1 方案一数据预处理及特征工程部分方案](#_Toc25354_WPSOffice_Level3) [6](#_Toc25354_WPSOffice_Level3)

[3.1.2 方案一模型设计、建立部分方案](#_Toc1088_WPSOffice_Level3) [7](#_Toc1088_WPSOffice_Level3)

[3.1.3 方案一结果、排名等](#_Toc21881_WPSOffice_Level3) [7](#_Toc21881_WPSOffice_Level3)

[3.1.4 方案一算法流程图](#_Toc16433_WPSOffice_Level3) [7](#_Toc16433_WPSOffice_Level3)

[4. 算法比较](#_Toc26629_WPSOffice_Level1) [8](#_Toc26629_WPSOffice_Level1)

[表4-1 算法比较](#_Toc16433_WPSOffice_Level2) [9](#_Toc16433_WPSOffice_Level2)

[5. 总结与展望](#_Toc25354_WPSOffice_Level1) [9](#_Toc25354_WPSOffice_Level1)

[5.1 总结](#_Toc11799_WPSOffice_Level2) [9](#_Toc11799_WPSOffice_Level2)

[5.2 建模思路](#_Toc16041_WPSOffice_Level2) [9](#_Toc16041_WPSOffice_Level2)

[第一名得主的思路梳理如下图所示：](#_Toc4307_WPSOffice_Level2) [9](#_Toc4307_WPSOffice_Level2)

1. 背景描述

我们依靠导航技术结合准确的位置和环境数据规划最佳运输路线，以此实现不间断的海洋运输能力。如何考虑减少商业航运对海洋环境的影响的同时，实现商业可持续性，具有越来越重要的意义，特别是涉及各种累积海洋噪音“足迹”对大鲸鱼生存状况的影响。

Marinexplore团队和领先的海洋专业人士社区一起组建了Planet海洋数据库。其中一个重要的数据集包括可用于探测栖息在全球海洋中的物种的声学记录。有心人士想要将这些关于动物位置的知识应用于工业操作。

康奈尔大学的生物声学研究项目在识别濒临灭绝的鲸鱼物种方面拥有丰富的经验，并且已经部署了一个24/7浮标网络来避免船只与世界上最后的400只北大西洋露脊鲸发生碰撞。

露脊鲸可以发出六种类型的声音，但其中一种叫做“向上呼叫”（“up-call”）的特征性呼叫 是由自动探测浮标识别的声音。“向上呼叫”由于高度的可识别性和露脊鲸的高频率利用而变得很有用。这种“向上呼叫”有点像露脊鲸之间简短的谈话，目的是让附近的同类可以获取自己的行踪。在每段剪辑的录音中，“向上呼叫”很容易听到 ，是一个大约持续一秒钟的上升的“呐喊”声。

**1.1 竞赛赛题描述**

赛方要求参赛者创建一种算法，通过检测录音中是否有北大西洋露脊鲸的“向上呼叫”（“up-call”），来防止航运流量与露脊鲸发生冲突。

Marinexplore和康奈尔大学的研究人员向你提出挑战，要求你击败现有的鲸鱼检测算法来识别正确的露脊鲸的“向上呼叫”。这将推进该地区的船舶航线决策。

挑战是匹配专家分析师标签，以确定特定的2秒声音片段是否包含正确的露脊鲸呼叫片段。

**1.2 评估指标描述**

提交的结果应该是一个大小为54503 x 1的实数列，并且与测试集中的剪辑顺序相同（test1，test2，...）。 行标题不是必需的。 根据ROC曲线区域判断提交的内容：

在Matlab中（使用统计工具箱）：

[~, ~, ~, auc ] = perfcurve(true\_labels, predictions, 1);

在R中（使用ROCR包）：

pred = prediction(predictions, true\_labels);

auc.tmp = performance(pred,"auc");

auc = as.numeric(auc.tmp@y.values);

在python中（使用scikit-learn的metrics模块）：

fpr, tpr, thresholds = metrics.roc\_curve(true\_labels, predictions, pos\_label=1)

auc = metrics.auc(fpr,tpr)

Ps:本文档整理的方案运用的是python语言

1. 数据来源及描述性统计分析

**2.1 大赛数据来源**

数据来源：Marinexplore团队和领先的海洋专业人士社区一起组织的Planet海洋数据库（marinexplore.com提供了各种与海洋相关的测量数据）

本次比赛官方提供的数据如下（全部数据都在此压缩包中）：

<whale-detection-challenge.zip>

官方数据下载地址如下：

<https://www.kaggle.com/c/whale-detection-challenge/data>

**2.2 数据的描述性统计**

**2.2.1 数据基本情况描述：**

数据库包括30,000个训练样本和54,503个测试样本。 每一段都是一个2秒的.aiff声音片段，采样率为2 kHz。文件“train.csv”（“<whale-detection-challenge.zip>”的子文件）给出了训练组的标签。 将分类结果用1和0表示，即正确的“露脊鲸的叫声片段”（目标片段）的标签= 1，否则标签= 0。

所有的训练和测试数据都已经打包在了 "whale\_data.zip"（“<whale-detection-challenge.zip>”的子文件）中，如果要在下载完整数据文件之前浏览数据格式，“small\_data\_sample\_revised.zip”（“<whale-detection-challenge.zip>”的子文件）包含10个示例剪辑。

建议：这些剪辑可能是露脊鲸，非生物噪音和其他种类鲸鱼的叫声的混合音频。

通过FFT将时间序列变换到频域可能会对于数据的研究有帮助。Cornell的RavenLite软件可能会对于快速浏览剪辑很有用。

**2.2.2 数据字段介绍：**

**<templateReduced.csv>（**‘moby-master’的子文件**）**

**2.2.3 数据描述性统计**

第一名得主计算了每个频率切片的 ：质心，宽度，倾斜和总变化。

绘图代码在‘plotting.py’（‘moby-master’的子文件）中。（由于python版本的问题无话画出）

3. 优秀算法思路

**3.1 方案一**

**3.1.1** 方案一数据预处理及特征工程部分方案

# 第一名得主对于典型的音频处理功能mfcc(语音特征参数)和 rolloff没有过多的使用，更多的是在使用模板匹配和图像处理，由于没有很多独特的露脊鲸的呼叫声，所以可以对于模板进行适度的删减。第一名得主的全部代码放到了GitHub上【https://github.com/nmkridler/moby2 】

# 所有操作运用的都是python语言，并使用opencv【开源发行的计算机视觉库】的matchTemplate【图形模板匹配算法】进行模板匹配。得主从平均spectogram

【https://github.com/flyingfishes/Spectrogram 这个网站介绍频谱图的原理】开始，但随后开始从数据中删除示例，根据看到的哪些类型的调用我得到/缺失。得主清理了模板，使它们成为二进制掩码。

最大的收获来自于试图增强图像的对比度。得主使用了一个简单的滑动装置，中间的几个箱子被移除了。我首先尝试了一个盒式过滤器，但如果我在两个图像上进行模板匹配，效果会更好：一个在频率维度中有贬值，另一个在时间维度上有贬值。它们非常相关，但是对于非露脊鲸来说，最大值出现的位置似乎不同。得主总共使用了大约25个模板。使用的其他指标基于时间维度中的切片：质心，宽度，倾斜和总变化。得主从排序中得到的特征只是一个简单的中心北移除的滑动平均值。得主发现使用64个箱子和128个箱子，得到了类似的结果，但是它们有点不相关，所以得主取了平均值，将分数提高了一点。

**3.1.2** 方案一模型设计、建立部分方案

生成功能后，将它们提供给sklearn的GradientBoostingClassifier进行分类。这种方法更多的是在关注出错的实例。（出错的概率也很低，大约为0.2或者更低），得主更多的是在尝试创建指标来解决它们。

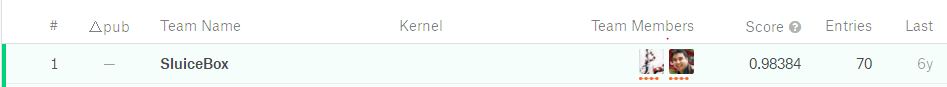
对于训练，得主只是进行了交叉验证，以便最后对所有样本进行预测。使用的是10折交叉验证。

下面是作者对于这次比赛的[主要方法的描述]:对于那些不熟悉marinexplore竞赛的人，我使用了多模板匹配方法。我使用滑动平均值来增强对比度，然后使用openCV的matchTemplate计算最大归一化互相关系数。我仅在频率维度上进行滑动平均移除，然后仅在时间维度上进行滑动平均移除。这产生了相当相关的值，但最大值的位置略有不同，给出了很好的提升。除了模板匹配之外，我还计算了每个频率切片的一些统计数据。我发现了从频率区0到60的质心，带宽，偏斜和总变化。在计算这些统计数据时，我没有做任何对比度增强。

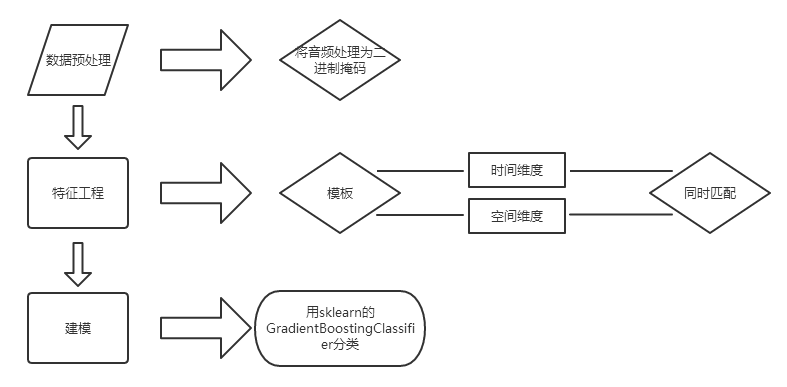
**3.1.3** 方案一结果、排名等

结果：得分0.98384

排名：最终排名top1



**3.1.4** 方案一算法流程图



4. 算法比较

【赛方总结讨论https://www.kaggle.com/c/whale-detection-challenge/discussion/4472】方法：

前两位获胜者都采用了一种方法，该方法定义了一个频率 - 时间“紧密盒子”，它限制了谱图中右鲸调用的发生，然后为每个紧密盒子提取一组定制的特征。第一名获奖团队采用多模板匹配方法，而第二名获胜团队使用维特比算法查找频率上扫描的精确轨迹。紧密的盒子使功能更加一致和稳健，因此更加频率不变和/或时间不变。

两种获胜方法还从不同的角度设计了几个特征向量，以结合来自频谱，呼叫频率调制的时间动态，甚至标记的时间顺序（正或负）的信息。最后一个变量，即时间排序，来自识别数据集中调用的文件和标签的排序和编号。结果，许多积极的分类事件连续出现。此数据集中的此时间聚类功能可能不是可靠的，我们可以在更新的自动检测系统中使用。然而，这个特征可能有助于区分正确的鲸鱼上行呼叫（几乎总是作为单独的瞬变发生）和座头鲸频率调制的上行呼叫，它们是歌曲中的音符或者是作为一系列呼叫产生的。

许多参与者应用深度学习方法（特别是卷积网络）并获得高分（例如排名为＃3，＃4和＃6的参赛者）。在我们对其深度学习方法的理解中，右鲸呼叫的频谱图被视为与手写数字大致相同的图像。

许多参赛者使用Python作为首选编程语言，反映出Python的模块，如Sci-Kit-learn，Sci-py，Num-py，已经成为数据分析领域的标准。因此，参与者优选几种分类器，例如梯度增强和随机森林。

**表4-1 算法比较**

|  | **评估指标** | **特征工程** | **基础算法** | **基本库** |
| --- | --- | --- | --- | --- |
| **算法1** | **ROC曲线** | **多模板匹配** | **opencv【开源发行的计算机视觉库】的matchTemplate【图形模板匹配算法】** | **Numpy/Scipy**  **Pandas**  **OpenCV**  **Sci-kit Learn** |
| **算法2** |  |  |  |  |

5. 总结与展望

**5.1 总结**

本项目的主要目标是对于音频中的特征片段进行识别，主要运用的方法是机器视觉的多模板匹配方法，和GradientBoostingClassifier分类方法。

**5.2 建模思路**

第一名得主的思路梳理如下图所示：

