

復旦大學

本科毕业论文



论文题目：基于北美东部鸟鸣数据的分类实证研究

姓 名：戴濡羽 学 号：17307110321

院 系：复旦大学管理学院

专 业：统计学

指导教师：沈娟 职 称：副教授

单 位：复旦大学管理学院

完成日期：2022 年 5 月 22 日

---

摘要.....	2
Abstract .....	3
1. 引言 .....	4
1.1. 研究背景.....	4
1.2. 文献综述.....	4
2. 信号特征.....	6
2.1. 时频特征 .....	7
2.1.1. 过零率.....	7
2.1.2. 频谱质心.....	7
2.1.3. 带宽.....	7
2.1.4. 衰减截止频率.....	7
2.1.5. 频率范围.....	8
2.2. 倒频域特征 .....	8
2.2.1. 梅尔频率倒谱系数 MFCC.....	8
3. 数据集 .....	9
3.1. 数据来源 .....	9
3.2. 描述性统计 .....	9
3.2.1. 以音频文件为单位.....	9
3.2.2. 以鸟类为单位.....	10
3.2.3. 采样日期与出现鸟类之间的关系 .....	11
3.3. 数据处理 .....	11
3.3.1. 过滤低频信号.....	11
3.3.2. 过滤强重叠片段.....	12
3.4. 信号特征的描述性统计.....	13
4. 分类模型.....	15
4.1. 隐式马尔可夫模型.....	15
4.2. 随机森林 .....	16
4.3. XGBoost .....	16
4.4. 卷积神经网络模型 .....	17
5. 结果与讨论 .....	18
5.1. 对于 HMM 模型的讨论 .....	18
5.2. 不同模型对音频重叠比的敏感程度 .....	19
6. 总结 .....	22
参考文献 .....	23
致谢 .....	25
附录 .....	26

---

## 摘要

本文所使用的是一份由 Dryad 发布的、采集于美国宾夕法尼亚州的鸟鸣数据集，包含鸟鸣文件(.wav)与注解文件(.txt)两个部分。与其他鸟鸣数据集不同的是，它包含了由专家校验过的鸟鸣片段起讫时间。这有助于更多研究者聚焦于基于鸟鸣的分类器的研究。

在前期的数据处理工作中，本文将注解文件中的片段起讫时间视为真值，重新切割音频文件生成鸟鸣片段，并且对每一个鸟鸣片段生成“收集日期”、“音频重叠比”这两个变量，这对后续分类器的表现都有所影响。

本文主要尝试了四种分类器，分别是在鸟鸣识别领域中较为经典的隐式马尔可夫模型(HMM, Hidden Markov Model)，在目前的鸟鸣识别赛事中非常流行的卷积神经网络模型(CNN, Convolutional Neural Network)，以及较为常用的两种基于决策树的模型，即随机森林和 XGBOOST。

从结果来看，对于不同重叠比、包含收集日期与否的数据集，随机森林与 XGBOOST 的表现都非常优越，准确率在 99% 左右，CNN 的准确率在 90%-95% 之间，较为经典的 HMM 模型准确率最高仅在 77% 左右。对 HMM 模型进一步分析，可以发现它在不包含收集日期时，重叠比越高，准确率越低，准确率的波动越大。当加入收集日期后，模型有更好的稳健性，即重叠比上升，准确率波动变小。

对于表现较好的 CNN 模型和 XGBOOST 模型，本文进一步探讨了它们在去除注解文件中其他变量，而仅使用根据鸟鸣片段提取的音频特征时的表现，发现两者都几乎不受重叠比影响，而当数据量增大，CNN 的分类效果明显好于 XGBOOST。

**【关键词】** 鸟鸣识别，隐式马尔可夫模型，随机森林，XGBOOST，卷积神经网络

---

## Abstract

The dataset used in this article is from Dryad. It's a bird-audio dataset collected in Pennsylvania, USA, including two parts: birds singing file (.wav) and annotation file (.txt). Different from other bird-audio datasets, it contains the start time and end time of the birds' call or singing, which have been verified by experts. This helps more researchers focus on the study of the classifiers.

In the early data-processing work, this article considers the clips in the annotation document as true value, re-cuts audio files to generate audio fragments, and generate "collection date" and "overlap ratio" for each audio fragment. These two variables affect the performance of the following classifiers.

These four classifiers are the focus of this article: Hidden Markov Model (HMM), which is a classic model in the field of bird recognition, Convolutional Neural Network (CNN), which is more prevalent in the current bird identification data contest, and the other two commonly used tree-based models, namely Random Forest and XGBOOST.

From the results, the performance of random forests and XGBOOST is outstanding for all data sets with different overlapping ratios and with or without date variables. The accuracy rate is about 99%. CNN comes second with an accuracy between 90%-95%. HMM with the poorest performance has a 77% accuracy. For further analysis of the HMM model, we discovered that when it does not include the collection date variable, the higher the overlapping ratio, the lower the accuracy, and the greater the fluctuation of classification accuracy. When the collection date is added, the model is more robust, that is, the overlapping ratio rises and the accuracy fluctuations become smaller.

For the well-performed CNN model and XGBOOST model, this article further discusses their performance in removing other variables in the annotation file, and only uses the performance of the audio features extracted based on audio fragments. When the amount of data increases, the classification effect of CNN is significantly better than XGBOOST.

[Keyword] bird recognition, Hidden Markov Model, Random Forest, XGBOOST, Convolutional Neural Network

# 1. 引言

## 1.1. 研究背景

鸟类的数目和种类对生态监测的重要性也是不言而喻的：一个良好的生态环境会吸引更多的鸟类生活其中，或者当气候、环境发生变化，鸟类会比人更敏锐地察觉（Virkkala et al., 2014）。因此，对不同鸟类的行为模式或整体数量的记录，可以作为生态监测的一个好的指标。

但长久以来，鸟类识别就不是一项轻松的工作。它需要观测者眼观四路，耳通八方，脑内还要存有一个足够的数据库，以调用观测地点所存在的多种可能的鸟种类及其习性，从鸟的外形、声音、行为等多方面，判断鸟的种类、性别与成熟度。因此，观鸟者通常一手拿着望远镜，一手拿着纸质的鸟类指南，完成观鸟活动。但是，随着机器学习等技术逐渐深入到鸟类识别这一领域，刚入门的观鸟者也可以运用如“懂鸟”、“Warblr”等商用程序，通过录入鸟鸣或者拍摄照片等方式，快速了解所观测的鸟。

## 1.2. 文献综述

本文所关注的鸟鸣识别技术，它的基本研究路径包含三个部分：收集鸟鸣数据，从音频数据中提取特征，根据特征进行分类。在此基础上，为了说明分类模型的准确度，会把数据集分为训练集与测试集。训练集用来训练不同鸟类对应的模型的参数，用这个模型来拟合测试集，用拟合最好的模型对应的鸟类作为分类结果。通过预测结果与测试集的真值之比得到准确度。

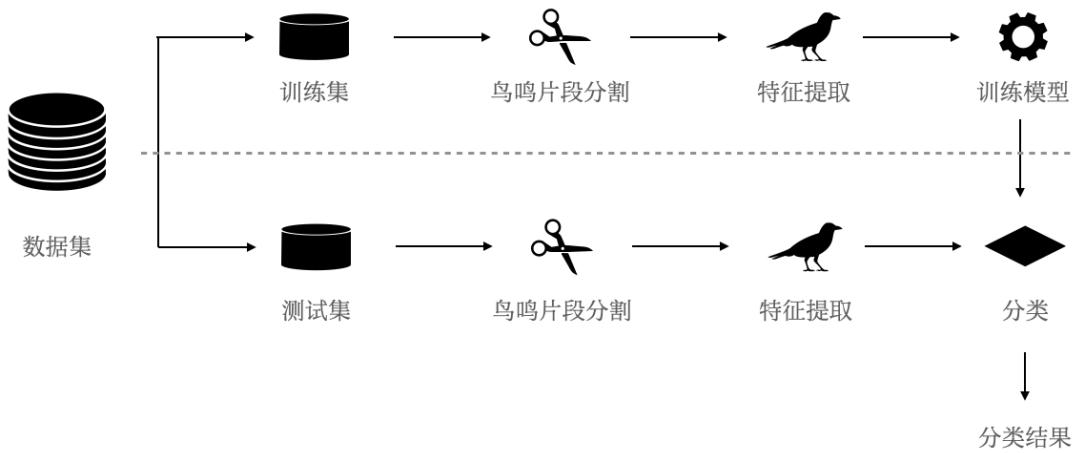


图 1.1 鸟鸣识别的基本研究路径

鸟鸣片段分割主要是提取一段音频中只包含单一鸟鸣的片段，而去除较长的空白片段以及其他噪音。分割的准确度对后续的分类至关重要。通常对较小的数据集，采取人工分割是可行的(Anderson, 1996; Kogan, 1998)。但对于大的数据集，采取自动分

---

割的方式是必要的。通常来讲，有以下两种路径：

1. 基于能量包络(energy envelope)的切割（Fagerlund, 2004; Chen, 2006）。包络的意思是对信号震荡的某一方面特征进行平滑后得到的曲线(Johnson 2011)。具体来说，就是以某个能量包络的阀阈值为起点，向后搜索信号。若出现高于这个阀域值的信号区间，则标注鸟鸣片段的起讫点。同时，根据不同时间出现的噪声，不断更新阀阈值。这一方法的优点是简洁明了，易于操作。一旦音频质量不稳定，比如出现较大音量的噪声，就会出现分类失误。此外，该方法也不适用于噪声能量大于鸟鸣的情况。

2. 无监督学习的聚类方法(Jancovic, 2011; Graciarena, 2011; Zakeri, 2017)。随着机器学习的发展，鸟鸣片段的识别问题也可以转变成为某种模式识别的问题。Javcovic 和 Zakeri 将用正弦信号拟合鸟鸣，结合傅立叶快速变换，聚类得到一系列鸟鸣频率向量(frequency tracks)。这一方法平均正确率为 86%。但是这一方法同样容易受到变化的噪声影响，因此他们引入了另一个噪声数据库辅助分类。此外，Graciarena 借鉴了语言识别中的方法，将鸟鸣最小的音节类比做语言识别中的音素(phoneme)。从而，找出鸟鸣的模式，就是找出概率意义上音素最大可能出现的排列方式。

经过切割后，下一步是对鸟鸣片段进行特征提取。目前比较常用的特征是倒频谱特征(MFCC) (Somervuo, 2006; Graciarena, 2011；徐淑正, 2018)。还有一些常见的频谱特征 (Fagerlund, 2004)，比如频谱质心、带宽、衰减截止频率等等的均值和方差。具体可以参考本文第二部分。

最后，将提取到的音频特征投入不同的分类器。一般常见的分类器可以分为三种 (Montgomery, 2020): 1. 神经网络模型，如时延神经网络模型(TDNN, Time Delay Neural Network)(Cai, 2007)；2. 深度学习模型，如卷积神经网络(CNN, Convolutional Neural Network)(Xie et al., 2019)、深度神经网络(DNN, Deep Neural Network)(Kahl, 2021)，3. 概率模型，如隐式马尔可夫模型(HMM, Hidden Markov Model)(Trifa, 2008)。

在数据集比较纯净的时候，以 HMM 为代表的概率模型表现不错。Zakeri(2017)探讨了将 HMM 与动态时间规整(DTW, Dynamic Time Warping)相结合的情况，即从总的同种鸟鸣片段中提取出某种固定的特征。再与每个鸟鸣片段同时输入 HMM，可以达到 94% 的准确率。但该方法的缺点是对噪声敏感。

目前在各类大型鸟鸣识别的机器学习竞赛中表现较为优异的模型一般都属于深度学习模型(Kahl, 2021)。该算法的特征就是通过大量数据自主学习其中的规律性，产生分类。这与鸟鸣识别的主要难点是匹配的。

## 2. 信号特征

本文所处理的音频信号，本质上是声波。最基础的声波单元是正弦波，从而对于现实世界中的声音，也可以通过叠加一系列正弦波的形式去还原。具体地说，就是通过傅里叶变换，使得一段时间的音频示波图(Oscillogram)可以用一系列正弦波的振幅和频率表示为频谱(Spectrum)。

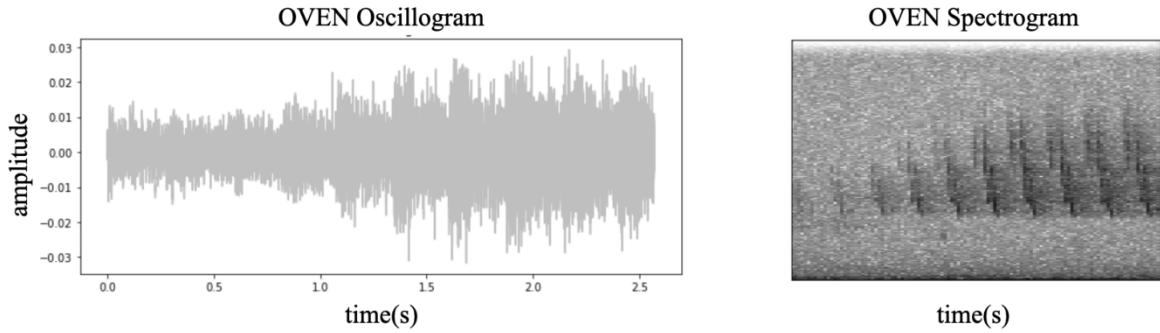


图 2.1 橙顶杜鹃 OVEN 的音频示波图和频谱

但是，通过傅里叶变换会损失掉音频的时域信息。这也就是说，傅里叶变化把一段时间的音频信号当作瞬时发生的，将其分解为正弦波，并得到每个频率的能量强度。为了同时保有时域、频域信息，需要对音频进行短时傅里叶变换，从而得到了时频图。

此外，对时频图进行傅里叶逆变换，可以得到倒频谱(cepstrum)，相比于频谱呈现了更多有关音调的信息(Noll 1964)。由此得到的梅尔倒频谱系数(MFCC)，在鸟鸣识别领域中也应用广泛(Kogan, 1998; Fagerlund, 2007; Cheng, 2010)

后文将继续从这时频域、倒频域这两个方面具体说明可以提取的音频特征。需要注意的是，在这个过程中，所有过程都是基于帧的。因此必然涉及到三个重要的参数的选择：1.帧长(frame size)，2. 相邻两帧之间的间隔(hop size)，3.加窗(window)。本文默认使用 Hanning 窗，帧长为 512，间隔为 256。

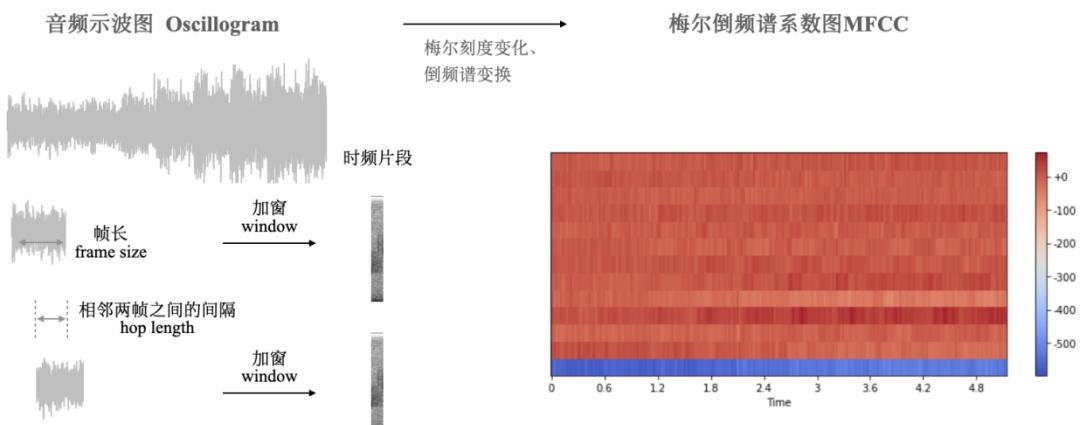


图 2.2 快速傅立叶变换与 MFCC 示意图

## 2.1. 时频特征

### 2.1.1. 过零率

由示波图可以看出，一段音频的强度(magnitude)被正则化之后，会表现出在零附近震荡。因此，对每一帧计算信号经过振幅为 0 的次数可以得到过零率(ZCR, zero crossing rate)。具体来说：

$$ZCR = \sum_{n=0}^{M-1} |sgn(x(n)) - sgn(x(n+1))| \quad (2.1)$$

其中  $x(n)$  为一帧信号， $n$  为一帧信号里的信号点(sample point)， $M$  为一帧信号中所含的信号点总数，示性函数  $sgn$  定义如下：

$$sgn(x(n)) = \begin{cases} 1, & x(n) \geq 0, \\ -1, & x(n) < 0. \end{cases}$$

### 2.1.2. 频谱质心

由频谱可以看出，我们可以算得频谱质心(SC, spectral centroid)。这一特质在感知上与声音的亮度相关，声音越亮，频谱质心越高。具体来说：

$$SC = \frac{\sum_{n=0}^M n |X(n)|^2}{\sum_{n=0}^M |X(n)|^2},$$

其中  $X(n)$  是对一帧进行离散傅里叶变换(DFT)， $M$  是 DFT 得到频率范围的一半。这是因为 DFT 结果是对半共轭的，因此只需要一半的数据。 $|X(n)|^2$  指信号的强度。

### 2.1.3. 带宽

基于频谱质心，我们可以得到各个频率段相对于频谱质心的加权“距离”，这就是带宽(BW, bandwidth)，其中权重即该频率段的强度。具体来说：

$$BW = \sqrt{\frac{\sum_{n=0}^M (n - SC)^2 |X(n)|^2}{\sum_{n=0}^M |X(n)|^2}}.$$

### 2.1.4. 衰减截止频率

同样基于频谱分布，我们可以得到特定能量分位点的频率，即衰减截止频率(SRF, spectral roll-off frequency)。这一特点用于区分高频和低频，类似于频谱的“斜度”描述。具体地说：

$$SRF = \max \left( K \left| \sum_{n=0}^K |X(n)|^2 \right| < TH \sum_{n=0}^M |X(n)|^2 \right),$$

其中， $TH$  为 0-1 之间的值。本文取常用值 0.95.

---

### 2.1.5. 频率范围

本数据集中，每个鸟鸣片段有对应的鸟鸣频率范围。需要注意的是，这个频率范围并不是以帧为基础的，而是反应了一小段时间内的鸟鸣频率特征。

## 2.2. 倒频域特征

### 2.2.1. 梅尔频率倒谱系数 MFCC

获得梅尔频率倒谱系数的过程如下：1. 对波形图做离散傅里叶变换 2. 对得到的频率做梅尔刻度变换(Mel-scaling)。这一步的目的在于使得数量级能更符合人耳对频率的感知。3. 对此时的频谱做离散对数变换，得到倒频谱系数。本文取 MFCC 的常用值，即前 12 个 MFCC。

### 3. 数据集

#### 3.1. 数据来源

数据集来源于 Dryad.(2021), 包含鸟鸣音频 (.wav 格式) 与相应注解 (.txt 文件)。其中, 每份音频文件时长为 5 分钟, 总计 385 分钟, 采样频率为 32kHz。音频数据收集于 2018 年 4 月至 6 月, 美国宾夕法尼亚州的 Powdermill 自然保护区。随后, 研究人员通过 Raven Pro 软件标注每份音频中的鸟鸣起讫时间(Begin Time (s)、End Time (s))、鸟鸣的频率范围(Low Freq (Hz)、High Freq (Hz)), 以及相应的鸟类(species), 并对最后结果进行人为复查。

由于鸟鸣分割经过专业软件与专家双重校验, 因此本文将注释文件中的鸟鸣起始时间当作真实值, 开展后续的特征提取与分类工作。

表 2.1 Recording\_03\_Segment\_01 对应的注解文件示意

selection	View	Channel	Begin Time (s)	End Time (s)	Low Freq (Hz)	High Freq (Hz)	species
1	Spectrogram 1	1	0.24588	0.69425	1950.0	9600.0	NOCA
2	Spectrogram 1	1	0.25070	1.37404	5512.5	8887.5	BWWA
3	Spectrogram 1	1	0.59783	0.90638	3112.5	4968.7	EATO
4	Spectrogram 1	1	0.88710	1.12334	2700.0	3768.7	EATO
5	Spectrogram 1	1	1.34029	2.63237	2400.0	9281.2	EATO

#### 3.2. 描述性统计

本节从三个方面介绍了本数据集。以音频文件为最小单位的描述性统计, 便于理解数据的收集过程。以鸟类为单位的描述性统计, 是建立在读取了所有音频文件的基础上, 对数据重新组合, 以分类结果为导向建立的描述性统计。最后, 对音频文件收集日期和鸟类交叉分析, 将“收集日期”纳入自变量考虑范围。

##### 3.2.1. 以音频文件为单位

图 1.1 展示了数据收集的过程以及相应的结果。具体来说, 数据集按照采样日期分为 4 个文件夹, 每个文件夹中含有若干个 5 分钟的音频文件(.wav)与对应的注解(.txt), 每份注解文件中所含的鸟类数目、注解数目是不同的。在图 1.1 中以文件夹为

整体，以音频文件为最小统计单位，给出了音频时长、注解数、鸟类的汇总。

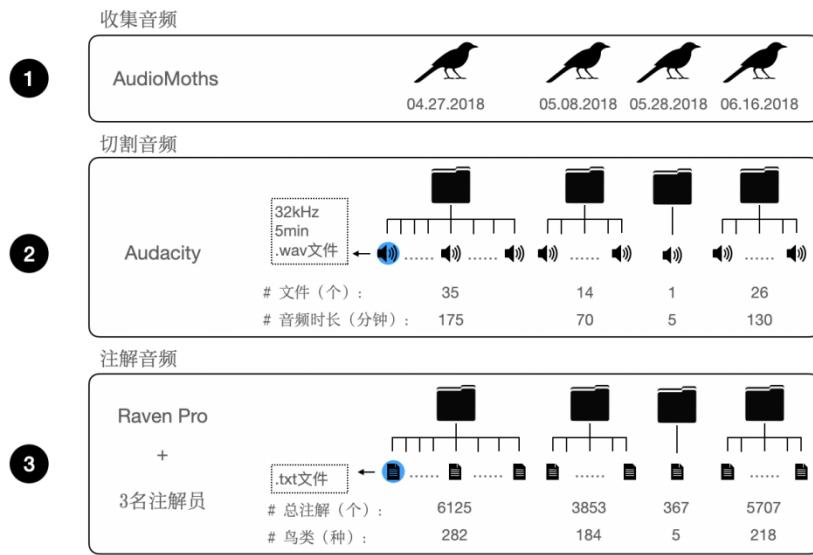


图 3.1 Chronister(2021)数据集内容和收集过程

### 3.2.2. 以鸟类为单位

由于音频文件中有一部分声音是没有鸟鸣的，这对后续的分类相当于“无效数据”。我们仅需要在每个 5 分钟的音频文件中提取含有鸟鸣注解的有效部分，即鸟鸣片段。鸟鸣片段的提取依赖于注解文件中的起讫时间列。从这个角度，可以统计数据集中各类鸟的鸟鸣片段数与总时长，这有助于了解分类器所处理数据的稀疏程度。

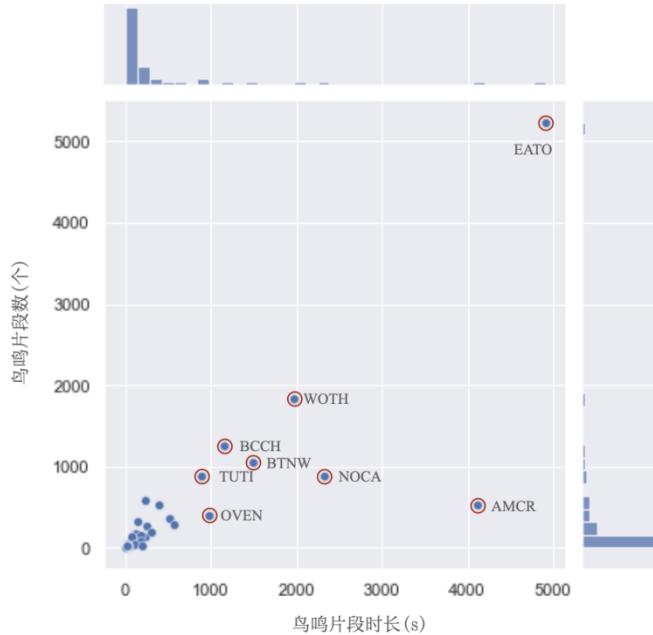


图 3.2 鸟鸣片段时长与个数分布图

从图 2.2 中可以看出，大部分鸟鸣片段个数在 0-500 之间，鸟鸣片段总时长也在

500s 以内，说明不同鸟类数据的稀疏差别较大。鸟鸣片段总时长短、鸟鸣片段个数少的鸟类，会削弱分类器的准确度。

### 3.2.3. 采样日期与出现鸟类之间的关系

一般在野外观鸟的时候，观鸟的日期和地点是非常重要的线索。但是在机器学习的模型中，这一点往往因为不属于音频特征的范围而被排除在外。

按照采样日期和鸟类，对音频片段的数目进行统计，可以发现在 6/16 日采集的音频中，多出了 7 种鸟类。经过查阅资料，确认这 7 种鸟皆属于候鸟，即在冬日会迁徙往美国南部过冬。

因此，本文将采集日期 date 纳入自变量的范围。这一变量对分类结果影响的讨论，具体见本文第四部分。

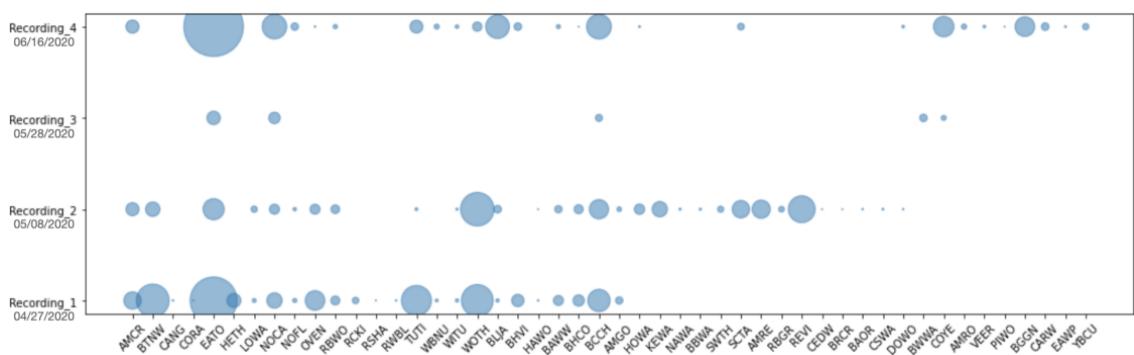


图 3.3 鸟鸣片段与收集日期之间的关系

## 3.3. 数据处理

出于数据量与数据质量两方面的考虑，本文需要对音频文件做以下处理工作。首先，本文根据注解文件筛去一定频率以下的信息，即过滤低频信号。其次，由于音频采集的过程是完全自然的，因此存在大量的鸟鸣重叠文件。音频的重叠程度会影响后续的分类结果，因此本文对每条鸟鸣记录计算重叠比。最后，由于经过重叠比筛选的鸟类所含的鸟鸣记录容量不一致，因此本文仅选取不同重叠比下数据量最多的 8 种鸟类进行分类。

### 3.3.1. 过滤低频信号

由于不同音频文件的录制日期不同，因此噪音的频率和强度也不同。又因为在注解文件中包含了鸟鸣的频率区间，因此一个自然的想法是，以所有鸟鸣中最低的频率为标准，筛去比该频率更低的信号，由此获得更为纯净的鸟鸣信号。

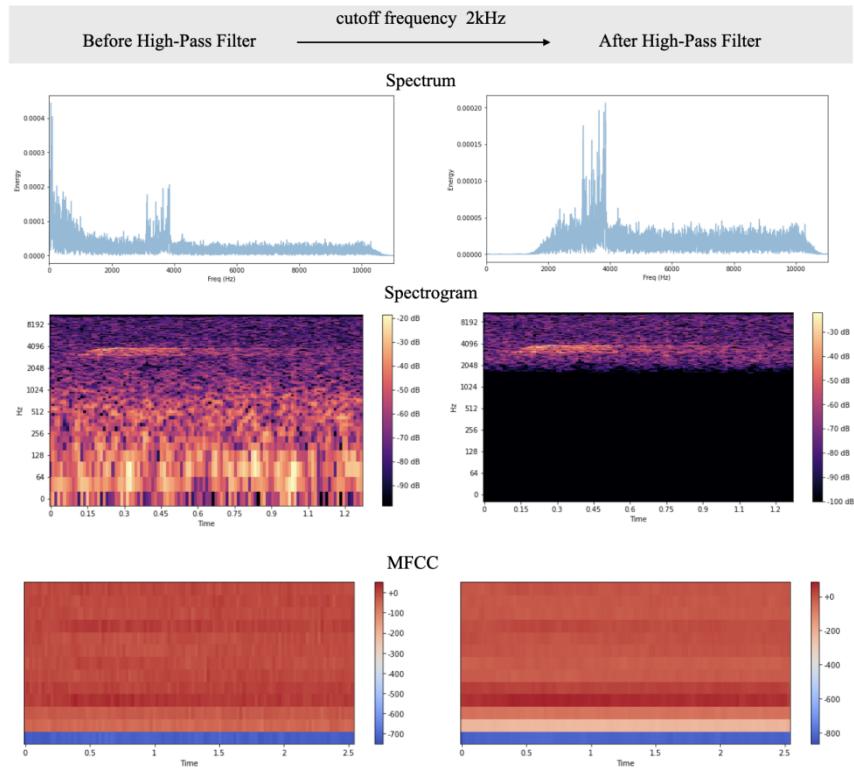


图 3.4 以某个 EATO 片段为例，经过高频过滤后的频谱、时频与 MFCC 图

可以发现，经过过滤后的频谱和时频图将低于 2kHz 的信号能量设为 0，从而仅留下了我们关心的鸟鸣部分频率，从图中也可以清晰地辨别出鸟鸣能量是高于其余高频噪音的。对于 MFCC 图，经过过滤后的图更为“平滑”了，说明特定种类的鸟鸣片段更容易呈现某种清晰的模式，有利于后续的分类。

### 3.3.2. 过滤强重叠片段

在分类器的搭建阶段，笔者发现分类器极度依赖注解文件中的频率范围，这与本文的探究目的相违背：分类器需要尽少依赖注解，完成较准确的分类工作。经过分析，这是由于本文中所提取的音频特征，不能很好地区分同时包含多种鸟类的音频。而这种复杂的音频处理也暂不在本文的研究范围之内，因此需要进一步过滤出弱重叠的音频片段。

根据注解文件中的鸟鸣片段起讫时间列，本文计算音频重叠的方法如下：

步骤 1：（计算与前一个鸟鸣片段的重叠）对于特定鸟鸣片段，向上比较至多 5 条鸟鸣片段的结束时间。前重叠时长= $\max\{[\text{前 5 条鸟鸣片段中的结束时间}-\text{本条鸟鸣片段开头}], 0\}$ ；

步骤 2：（计算与后一个鸟鸣片段的重叠）对于特定鸟鸣片段，向下比较至多 5 条鸟鸣片段的开始时间。后重叠时长= $\max\{[\text{后 5 条中的开始时间}-\text{本条鸟鸣片段结尾}], 0\}$ ；

步骤 3：（计算重叠比）对于特定鸟鸣片段，总重叠时长=前重叠时长+后重叠时

长，从而得到重叠比=总重叠时长/片段时长。

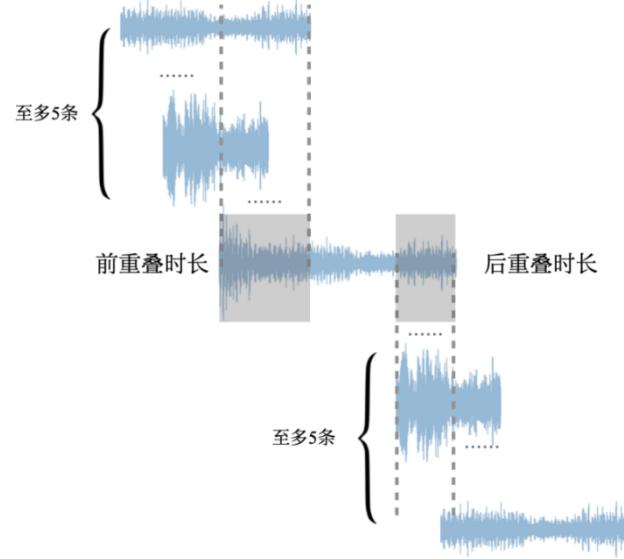


图 3.5 对于特定鸟鸣片段，计算重叠比的过程图示

当设置重叠比小于等于 0.1 的鸟鸣片段时，筛选得到的各类鸟的数据量如图 3.6 所示。此时符合标准的鸟类从原数据集的 48 种下降到 36 种，其中仅有 17 种鸟类所含片段数大于 30。考虑到其中有些鸟鸣片段过短，不利于后续隐式马尔可夫模型的分类，因此本文最后选取的数据集仅包含鸟鸣片段总数前 10 的鸟类，并筛选鸣叫片段应大于 0.4s。

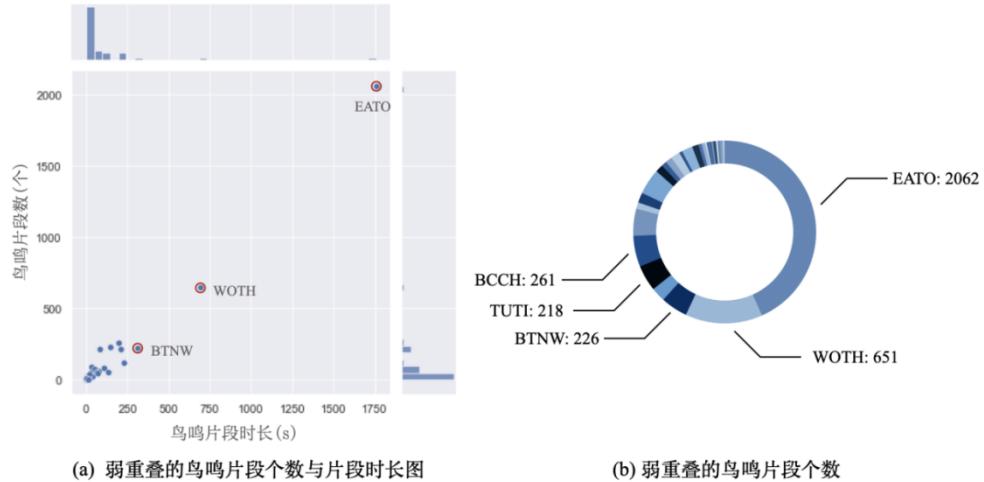


图 3.6 弱重叠鸟鸣片段的描述性统计

### 3.4. 信号特征的描述性统计

在不同重叠比下，得到各类鸟鸣时频、倒频特征分布如附录 2 所示。区分度一直很好的特征分别是来自于注解文件中的最低频率、最高频率。主要原因在于频率区间

是通过人工校验，仅包含鸟鸣信息。其余从音频中提取的特征，容易受噪声影响。

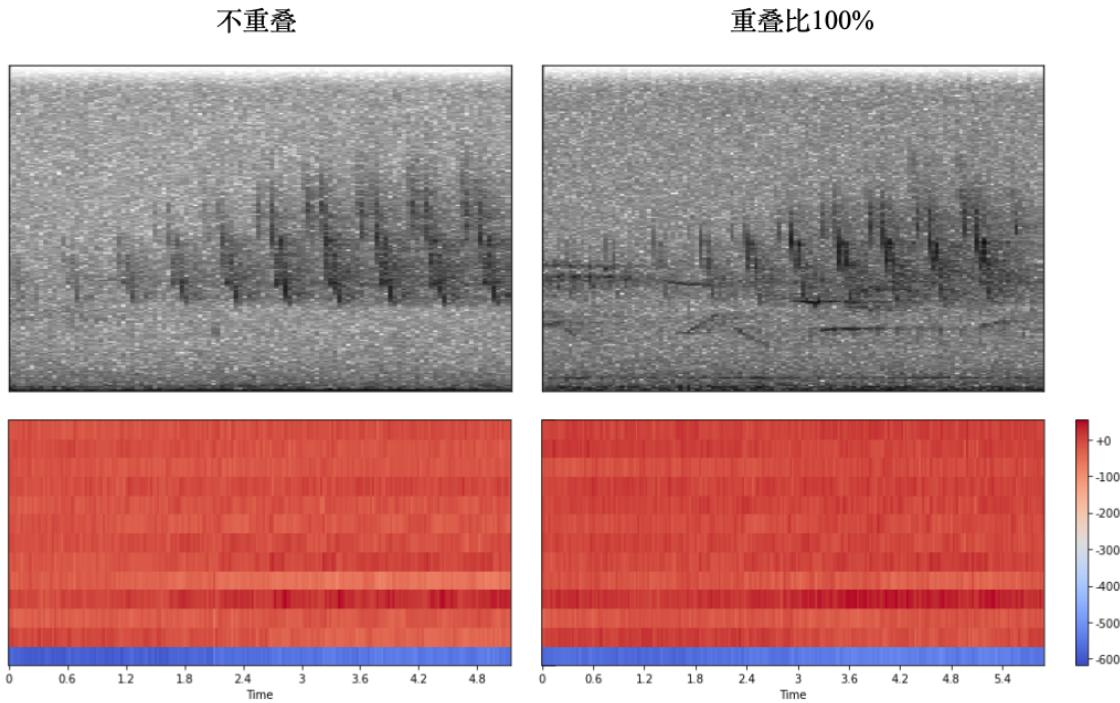


图 3.7 不同重叠比 OVEN 信号的频谱图、MFCC 图对比

MFCC 的特征中，靠前的系数更具有区分度，如 `mfcc_0` 到 `mfcc_3`。这与 MFCC 本身的性质有关：Mermelstein(1976)提出梅尔倒频谱系数，作为语音识别提取的特征，更靠前的系数意味着含有更多的“有用”信息，而靠后的系数，则包含着语句、发音之间的停顿、环境的噪声等“无用”信息。因此通常在语音识别领域常用到前 12-13 个 MFCC，对于鸟类来说，它们发音频率更高，每声鸟鸣较单词而言更短，因此有用的 MFCC 可能会比人声识别领域常用的数量更少一点。

相较而言，其余的时频特征区分度并不高，尤其是带宽和衰减截止频率，除了中位数最低的那种鸟类的均值、方差显著区别于其余 7 种，剩下的鸟类该特征的分布几乎一致。

## 4. 分类模型

### 4.1. 隐式马尔可夫模型

Trifa(2008)提出可以用隐式马尔可夫模型(HMM, Hidden Markov Model)来处理鸟鸣识别的问题。具体地说，假设每一种鸟鸣可以根据鸟鸣时长分割成若干个阶段，在每一个阶段，我们可以通过观测音频特征推测它所处的状态。

如图所示，记鸟鸣中的含有 n 个状态，状态 i 有观测  $O_k$  的概率为  $b_i(O_k)$ 。在本文中，所有音频特征都是连续的，因此采用高斯混合模型对多元向量建模，即假设该向量服从高斯混合分布。记 M 为混合的高斯分布总数， $c_{dim}$  为某一种高斯模型在混合分布中所占的权重， $N(O_k; \mu_{dim}, \Sigma_{dim})$  为观测  $O_k$  服从的均值为  $\mu_{dim}$ ，方差为  $\Sigma_{dim}$  的多元高斯分布，则有：

$$b_i(O_k) = \sum_{m=1}^M c_{dim} N(O_k; \mu_{dim}, \Sigma_{dim}).$$

直观地说，HMM 中的状态可以理解为某种鸟鸣模式。以橙顶杜鹃 OVEN 为例，将鸣叫按照时间分成不同状态，可以描述鸟鸣之间的间隔、短促鸟鸣音节与更长的鸟鸣音节组成的鸟鸣模式。每种发音模式以概率  $b_i(O_k)$  出现观测向量  $O_k$ ，这个概率是根据高斯混合模型拟合得到的。最终根据观测到的各特征的概率，推测该鸟鸣模式所对应的鸟类。因此，对于 HMM 来说，鸟鸣片段的准确切割是非常重要的。

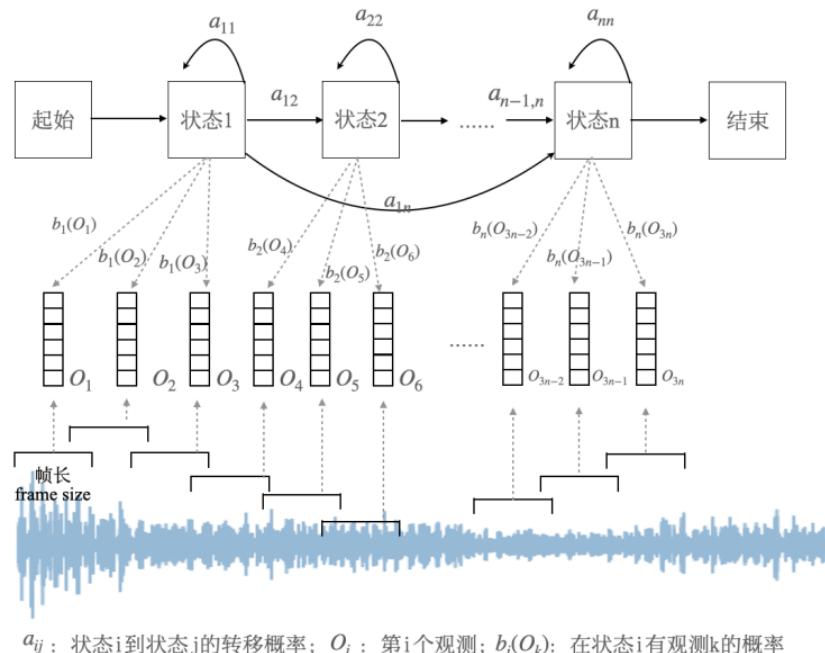


图 4.1 HMM 算法示意图（改自 Trifa, 2008）

---

## 4. 2. 随机森林

这一算法由 Breiman(2001)提出，是一种基于决策树的算法。随机森林会从完整的数据集中，用有放回抽样(SRS)抽取部分观测和部分特征，用决策树进行分类。对所有决策树的分类结果，按多数结果输出为最终结果。这也就是说，每枝决策树只处理有限随机样本的分类问题，通过最后对所有树结果的平均，使得结果更为精确，并能够避免过拟合的问题。这种对数据集抽样，再归拢自数据集的结果的结果被称为袋装算法(Bagging Algorithm)。

这一算法主要的优点有：1. 由于在建树的时候只随机选取了部分特征，因此随机森林能够规避决策树算法经常遇到的过拟合问题。2. 决策树的节点的构造是极为不稳定的，观测某一特征微小的变化就会导致整棵树的变化(Li, 2002)，从而导致分类结果的变化。因此在分枝时引入更多的随机性，有利于选取更好的特征，从而得到更稳定的分类结果。

## 4. 3. XGBoost

XGBoost(Chen, 2014)属于 Gradient Boosting 算法，这一算法会把上一层决策树的预测误差加入下一层的决策树进行分类，以减少模型的误差，从而使得弱分类器实现强效果。这也是 Boosting 算法与 Bagging 算法最大的不同。

具体地说，记第  $n$  步预测值为  $\hat{y}_n$ ，此时对应模型的伪残差为  $\hat{e}_n$ （伪残差的意思即用该步预测值与观测值之差，与真残差之别在于此时的预测并不是最终的），步长参数为  $\eta$ ，则有：

$$\hat{y}_n = \hat{y}_{n-1} + \eta \hat{e}_{n-1},$$

这一算法的目标函数为：

$$Obj(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k),$$

其中  $l$  为 Logit 损失函数， $n$  为观测数量， $K$  为树的总数， $\Omega$  为正则化函数， $f_k$  为第  $k$  棵树。

XGBoost 相对于其他 Gradient Boosting 算法的不同之处在于，在训练模型阶段使用了基于柱状图的算法(histogram-based algorithm)，从而使得模型表现更好，建模更快。

## 4.4. 卷积神经网络模型

卷积神经网络模型(CNN, Convolutional Neural Network)属于深度学习算法的一种。它包含三个结构：卷积层(Convolutional Layer)、池化层(Pooling Layer)以及完全连接层(Fully Connected Layer)。

对于输入的图像数据，卷积层的功能是，获得卷积核(kernel)与每个点所对应的空间定位(spatial location)之间的点积。所谓卷积核，本质是一个矩阵，用来提取空间定位的特征。需要注意的是，卷积核与空间定位之间的点积方式，是通过以特定步长移动卷积核，直至其遍历过整个图像。

在池化层，数据中的信息被进一步浓缩。具体地说，在卷积层输出的数据被进一步划分成若干个矩阵，对每一个子矩阵，输出其中的最大值。

最后，在完全连接层，所有特征都被激活，将前两层学到的特征模式，通过卷积映射回样本，得到最后的分类结果。

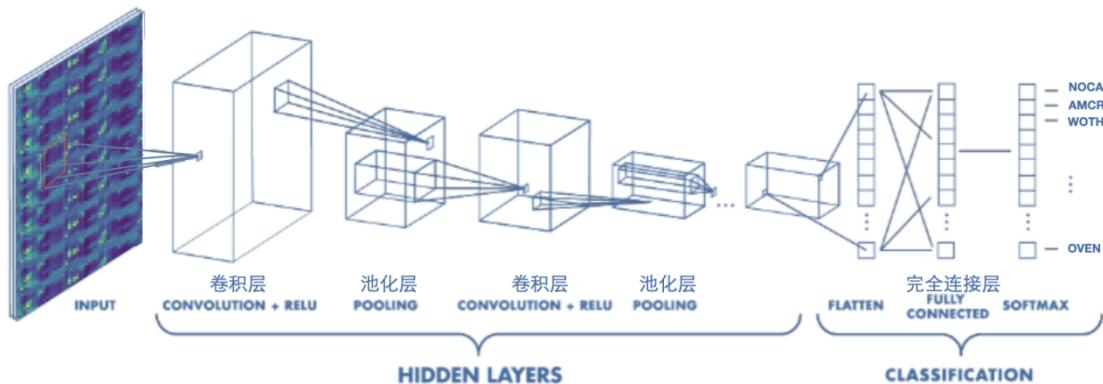


图 4.3 CNN 算法结构说明（改自电子书 Introduction Deep Learning with MATLAB 插图）

---

## 5. 结果与讨论

本文报告了四种模型对筛选过的数据集的分类表现，以及讨论了采集日期变量 date 和数据集的重叠比对不同分类器表现的影响。分类表现以准确率的形势呈现，其中 number of sample 是机器学习中测试集的样本量，number of true prediction 是经过训练得到的模型对测试集正确分类的数量：

$$\text{accuracy} = \frac{\text{number of true prediction}}{\text{number of sample}}.$$

对于 HMM 模型，笔者在实验中发现其预测的准确性波动性较大，并受到上述所提的两种变量的影响。考虑到数据量较大，模型运行较慢，因此对每种状态数的 HMM 模型反复运行 5 次，得到准确率的标准差 std：

$$std = \sqrt{\sum_{i=1}^5 (\text{accuracy}_i - \bar{\text{accuracy}})^2}, \quad \bar{\text{accuracy}} = \frac{1}{5} \sum_{i=1}^5 \text{accuracy}_i.$$

对于其余三种模型，笔者进而研究了不同模型对注解文件中频率范围的依赖程度。并且对表现比较好 XGBoost 和 CNN，尝试了仅包含音频特征的数据集进行分类。

### 5.1. 对于 HMM 模型的讨论

尽管 HMM 模型的方法论对鸟鸣片段有一定的解释程度（见 4.1），但此时状态数的定义并不明确，从而使得 HMM 状态数的参数选择成为一种仅基于特定数据集的调试艺术。在许多鸟鸣识别的文献中，提议取 12-13 个状态数为佳。因此针对本数据集，笔者也实验了不同状态数下的 HMM 模型表现情况，并对每种状态数反复计算 5 次，计算其标准差，结果如图所示。

对于不包含日期的数据集，当重叠比上升时，HMM 的分类准确度出现大幅波动更为剧烈，并且即使是最合适的状态数，分类准确度仍然不如重叠比较低时的情形。对于重叠比在 1%-10% 的数据集，预测的波动性比重叠比 0%-1% 的数据集所对应的预测波动小 0.01 左右，这可能是由于随机性带来的。

对于包含日期的数据集，上述规律仅体现在 50%-70% 重叠比的数据集与 70%-100% 重叠比的数据集的对比中。可能是因为增加了 date 变量使得分类更稳健了。但是总的来说，最佳准确率对应标准差最低的仍然是重叠比最低(1%)的数据集。因此，增加 HMM 预测准确度和稳健性的重要条件，是保证鸟鸣片段的单一性。

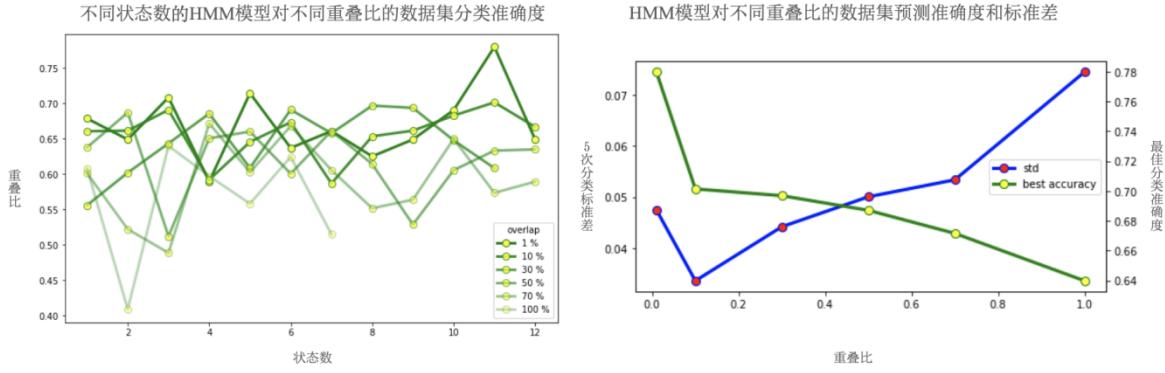


图 5.4 不包含 date, HMM 模型对不同重叠比数据集的分类结果分析

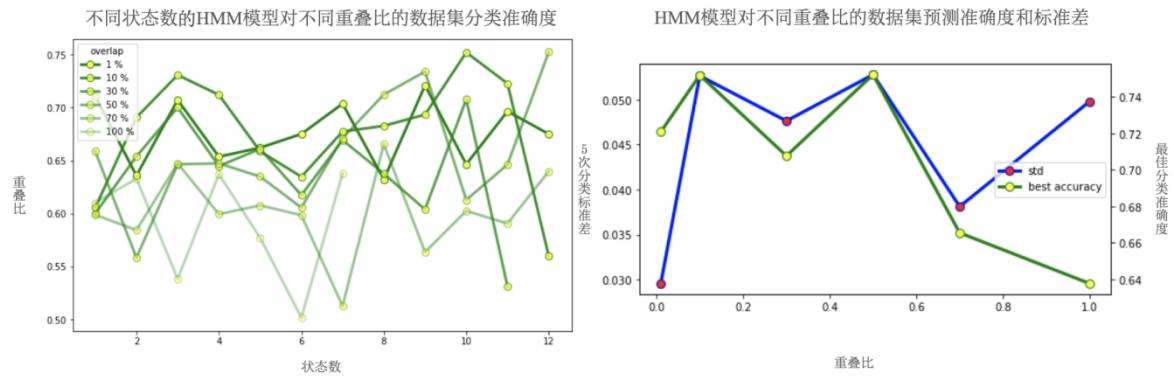


图 5.5 包含 date, HMM 模型对不同重叠比数据集的分类结果

## 5.2. 不同模型对音频重叠比的敏感程度

针对不同重叠比、包含与不包含数据采集日期 date 的数据，本文测试了 HMM、随机森林、XG Boost 和 CNN 这四种分类器的准确度。其中，HMM 的准确度取最佳状态数的准确度（见 5.2 的讨论），CNN 的准确度为 epoch=100 内最佳。可以发现，不论是否含有 date 这一变量，两种基于树的分类器都比较稳健，一直稳定在 98% 以上。重叠比对 CNN 的影响影响并不显著，其准确率基本在 90% 以上。但是 HMM 的准确度明显受到重叠比的影响，尤其是不包含 date 的数据集。

表 5.1 不包含 date, 不同重叠比数据集的准确度

模型 \ 重叠比	1%	10%	30%	50%	70%	100%
HMM*	77.98%	70.13%	69.68%	68.70%	67.15%	63.97%
随机森林	99.52%	99.58%	99.53%	99.27%	98.66%	98.35%
XG Boost	99.39%	99.19%	99.40%	99.22%	98.98%	98.69%
CNN*	89.47%	90.20%	98.57%	95.83%	98.13%	94.30%

表 5.2 包含 date, 不同重叠比数据集的准确度

模型	重叠比	1%	10%	30%	50%	70%	100%
HMM*	72.09%	75.20%	70.77%	75.23%	66.54%	63.76%	
随机森林	99.88%	99.88%	99.98%	99.81%	99.84%	99.86%	
XG Boost	100.00%	100.00%	100.00%	99.98%	100.00%	99.99%	
CNN*	97.36%	92.16%	98.58%	90.27%	97.19%	97.93%	

从各变量对分类结果的贡献度来看（图 5.2），基于树的两种模型表现如此优异的原因或许可以得到解答：这两种分类器始终高度依赖于 low\_freq 与 high\_freq。因此，受重叠比影响的其他信号特征并不会较大程度地影响分类结果。但是对于 HMM 算法来说，并不会赋予较大预测效果的变量更大的权重。因此，当信号特征受重叠污染，HMM 算法就不能稳定地分类。

具体地说，就 XGBoost 与随机森林而言，随机森林对 low\_freq 和 high\_freq 的总依赖程度一直在 60%-70% 左右，且随着重叠比显著增加。而 XGBoost 对频率范围的依赖与重叠比并没有明确的正向关系。对于重叠比在 30%-50% 的数据集，随机森林分类器对频率范围的依赖度为 66% 左右，而 XGBoost 对应的依赖度在 62% 左右，甚至低于重叠比在 1%-10% 的数据集。因此作者认为，在明确鸟鸣片段起讫时间的情况下，对于不同重叠比的数据集，XGBoost 有更大的区分能力，因此是一个比较好的算法。

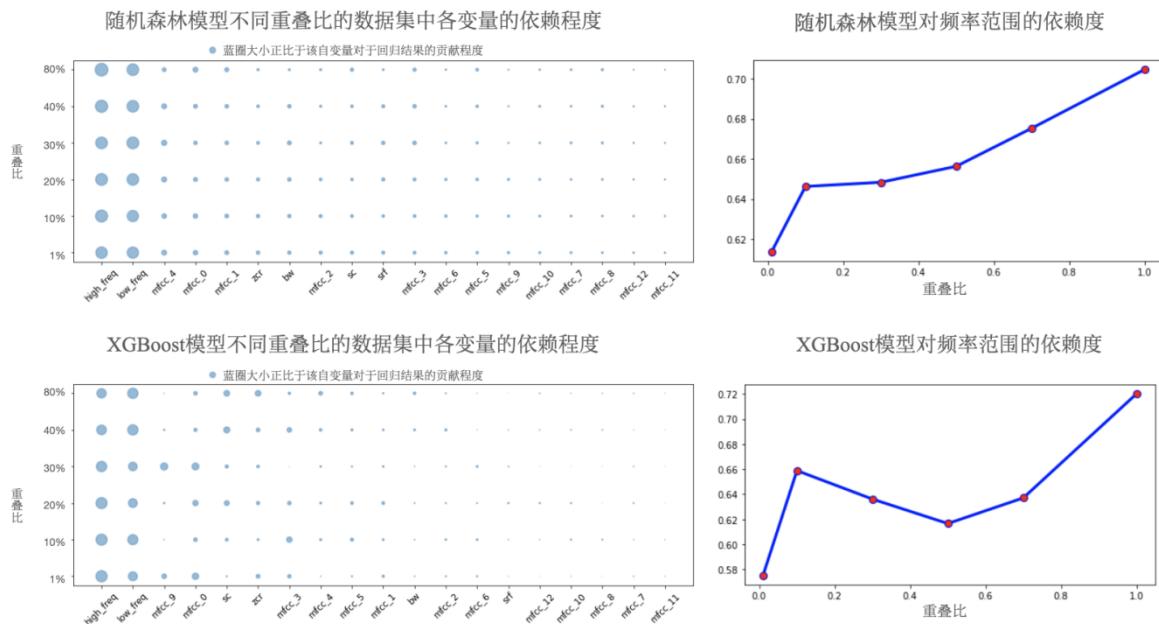


图 5.2 不含 date, 两种基于决策树的模型的变量重要程度分析

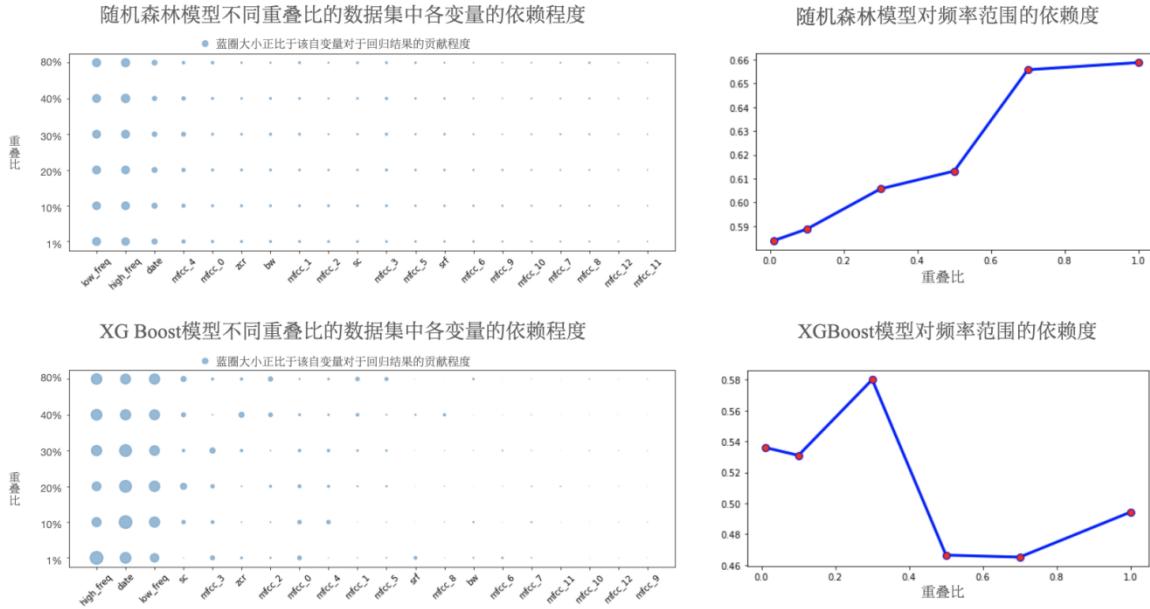


图 5.3 包含 date，两种基于决策树的模型的变量重要程度分析

对于 CNN 来说，分类准确率更重要的影响因素是数据量。对于重叠比在 0-1% 的数据，仅含有 190 条鸟鸣片段，因此分类效果最差。对于重叠比在 10%-30% 的数据，含有 350 条鸟鸣片段，准确率达到 98%。重叠比在 30%-50% 的数据，含有 356 条鸟鸣片段，与重叠比在 10%-30% 的数据量相仿。因此我们可以看到，此时分类的准确率有所降低，大概在 95% 左右。

因此，笔者推测，如果从数据集中去掉注解文件中的信息 (low\_freq, high\_freq, date)，CNN 的分类效果也很好。对此，笔者主要对比了 XGBoost 和 CNN 的表现效果，如表 5.3 所示。可以发现对于数据重叠比大于 30% 的数据集，CNN 的分类效果要显著好于 XGBoost。并且 CNN 的表现基本与数据量呈正比，即重叠比越高，数据量越大，分类效果越好。

表 5.3 CNN 和 XGBoost 对不含注解文件信息的数据集的分类结果比较

模型 \ 重叠比	1%	10%	30%	50%	70%	100%
CNN*	73.68%	62.74%	84.29%	90.28%	91.59%	92.22%
XGBoost	78.31%	78.79%	79.68%	81.41%	77.82%	81.32%

---

## 6. 总结

针对含有起讫时间标注的鸟鸣数据集，本文在数据整理阶段，加入了鸟鸣片段的重叠比，以及鸟鸣片段的收集时间。对于不同重叠比、是否包含收集时间的数据集，都做了用 HMM 模型、随机森林模型、XGBoost 模型与 CNN 模型的分类。

对于 HMM 模型，准确率与音频重叠比关系较大。在较纯净的音频下，准确度为 77%。随着重叠比增加，不仅分类准确度下降，且分类的波动性提升。当增加收集时间变量，分类的稳健性略有提升。

对于基于决策树的两种模型，准确度非常好。对不同重叠比的数据，若包含收集时间，则准确度基本在 98% 以上；若包含收集时间，则准确度在 99% 左右。本文进而探讨了分类结果对注解文件中的频率范围的依赖程度，发现随机森林对这两个变量的依赖性更大一点。

相较而言，CNN 虽然准确度在 90%-95% 左右，但对于不含注解文件信息的数据集，在数据量大的时候准确率仍然可以达到 90%，优于 XGBoost 的 80% 准确率。

最后，本文在以下方面仍然存在不足：

1. 本文缺失切割(segmentation)这一步骤，直接将注解文件中的鸟鸣起讫时间作为单音节鸟鸣(syllable)的真值。这一步骤的必要性在于，在实际的数据探索中，作者发现部分数据仍然存在鸟鸣前后有留白，以及一个片段中含有多声鸟鸣的情况，即非单音节。
2. 对于原数据集中的噪音，本文采用了直接截取大于特定频率以上的部分，作为降噪的手段。这种一刀切的方法，仅对于含有频率注解的数据是有效的。更好的办法是现对信号的高频部分进行加强，再进行高频过滤，并对过滤后的数据处理噪音。但限于笔者对信号处理背景知识的了解，暂时无法实现更好地信号过滤方法。

---

## 参考文献

1. Virkkala, R., & Lehikoinen, A. (2014). Patterns of climate-induced density shifts of species: poleward shifts faster in northern boreal birds than in southern birds. *Global Change Biology*, 20(10), 2995 – 3003.
2. 乔玉,钱昆, & 赵子平. (2020). 基于机器听觉的鸟声识别的中文研究综述. 复旦学报(自然科学版).59. 375-380.
3. Kogan, J. A., & Margoliash, D. (1998). Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study. *The Journal of the Acoustical Society of America*, 103(4), 2185 – 2196.
4. Anderson SE, Dave AS, Margoliash D.(1996) Template-based automatic recognition of birdsong syllables from continuous recordings. *The Journal of the Acoustical Society of America*, 100(2):1209 – 1219.
5. 徐淑正,孙忆南,皇甫丽英 & 方玮骐.(2018).基于 MFCC 和时频图等多种特征的综合鸟声识别分类器设计. 实验室研究与探索(09),81-86,91.
6. Jr, R. J. C., Sethares, W. A., & Klein, A. G. (2011). *Software Receiver Design: Build your Own Digital Communication System in Five Easy Steps* (1st ed.). Cambridge University Press.
7. Fagerlund, S. (2004, November). Automatic recognition of Bird Species by Their Sound (Master's dissertation).
8. Chen, Z., & Maher, R. C. (2006). Semi-automatic classification of bird vocalizations using spectral peak tracks. *The Journal of the Acoustical Society of America*, 120(5), 2974 – 2984.
9. Somervuo, P., Harma, A., & Fagerlund, S. (2006). Parametric Representations of Bird Sounds for Automatic Species Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6), 2252 – 2263.
10. Fagerlund, S. (2007). Bird Species Recognition Using Support Vector Machines. *EURASIP Journal on Advances in Signal Processing*, 2007(1).
11. Cai, Jinhai et al. (2007). Sensor Network for the Monitoring of Ecosystem: Bird Species Recognition. 2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information. 293-298.
12. Xie, Jie, et al.( 2019). Investigation of Different CNN-Based Models for Improved Bird Sound Classification. *IEEE Access*, vol. 7, pp. 175353 – 61.
13. Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning

- 
- solution for avian diversity monitoring. *Ecological Informatics*, 61, 101236.
14. Trifa, V. M., Kirschel, A. N. G., Taylor, C. E., & Vallejo, E. E. (2008). Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. *The Journal of the Acoustical Society of America*, 123(4), 2424 – 2431.
  15. Kahl, Stefan et al. (2021). Overview of BirdCLEF 2021: Bird call identification in soundscape recordings. *CLEF*.
  16. Dryad. (2021, April 6). An annotated set of audio recordings of Eastern North American birds containing frequency, time, and species information [Dataset]. Dryad. <https://doi.org/10.5061/dryad.d2547d81z>
  17. Noll AM, Schroeder MR.(1964). Short-time ‘Cepstrum’ pitch detection. *The Journal of the Acoustical Society of America*, 36(5). 1030 – 1030.

---

## 致谢

本文的选题和本人的两只鹦鹉分不开，分别是小鸟和木鸭。在倒数第三次汇报的时候，两只鸟飞走了。之后一天，它们在小区内邻居的帮助下回了家。正如本篇论文和我的大学生活，都有惊无险地完成了。

此外，我的家人和朋友，尤其母亲顾女士，一如既往地给予了我很多精神和物质上的支持，使我能够保持相对平稳的心态，完成困难的选题。

最后，我想感谢指导教师沈娟老师。每次汇报时，她都用认真负责、耐心细致的态度，督促着我不断思考论文可以改进的地方。这种学术态度，我相信会令我受益终身。

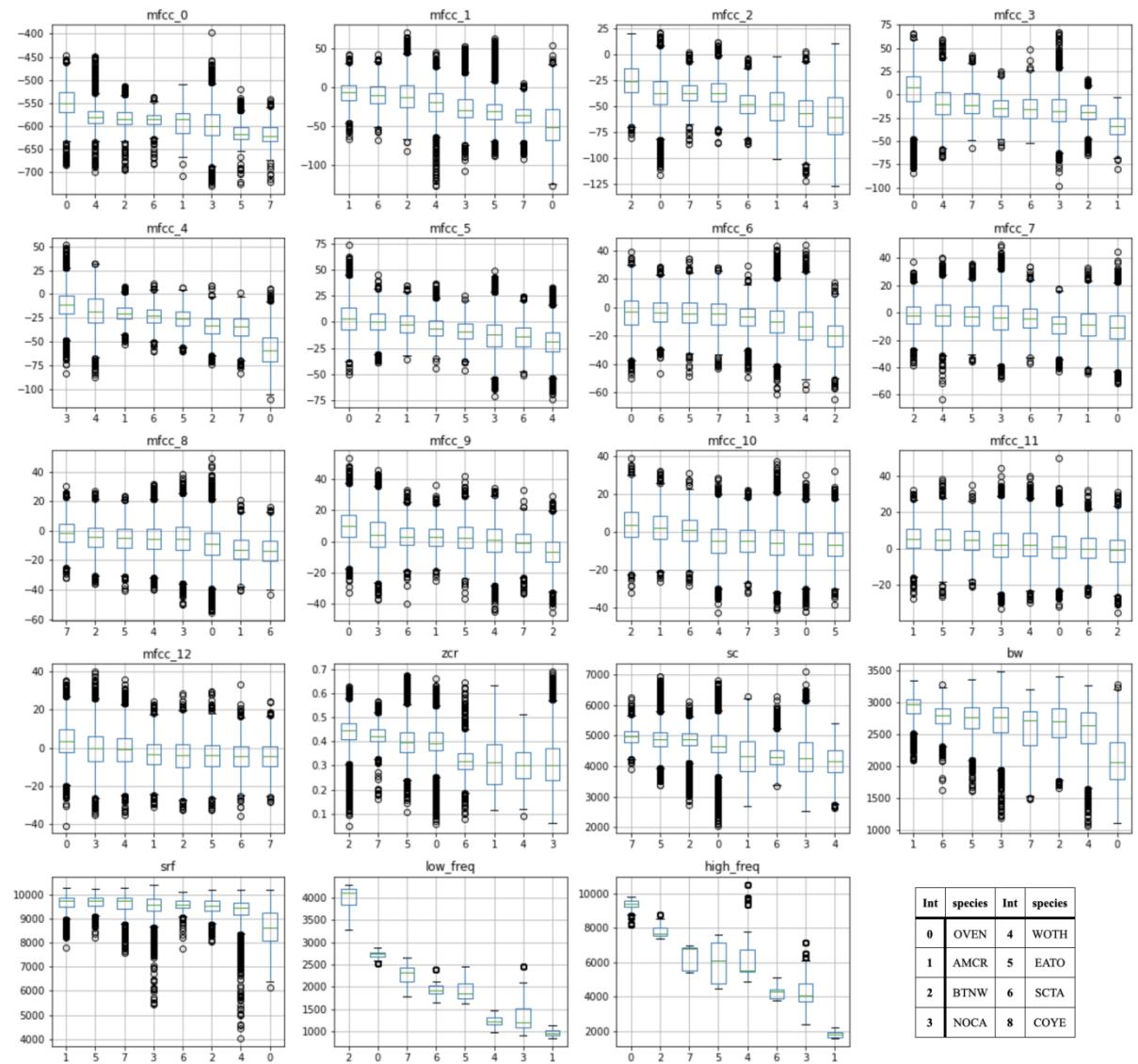
## 附录

附录 1：鸟鸣简写对应的鸟类常见名、拉丁学名

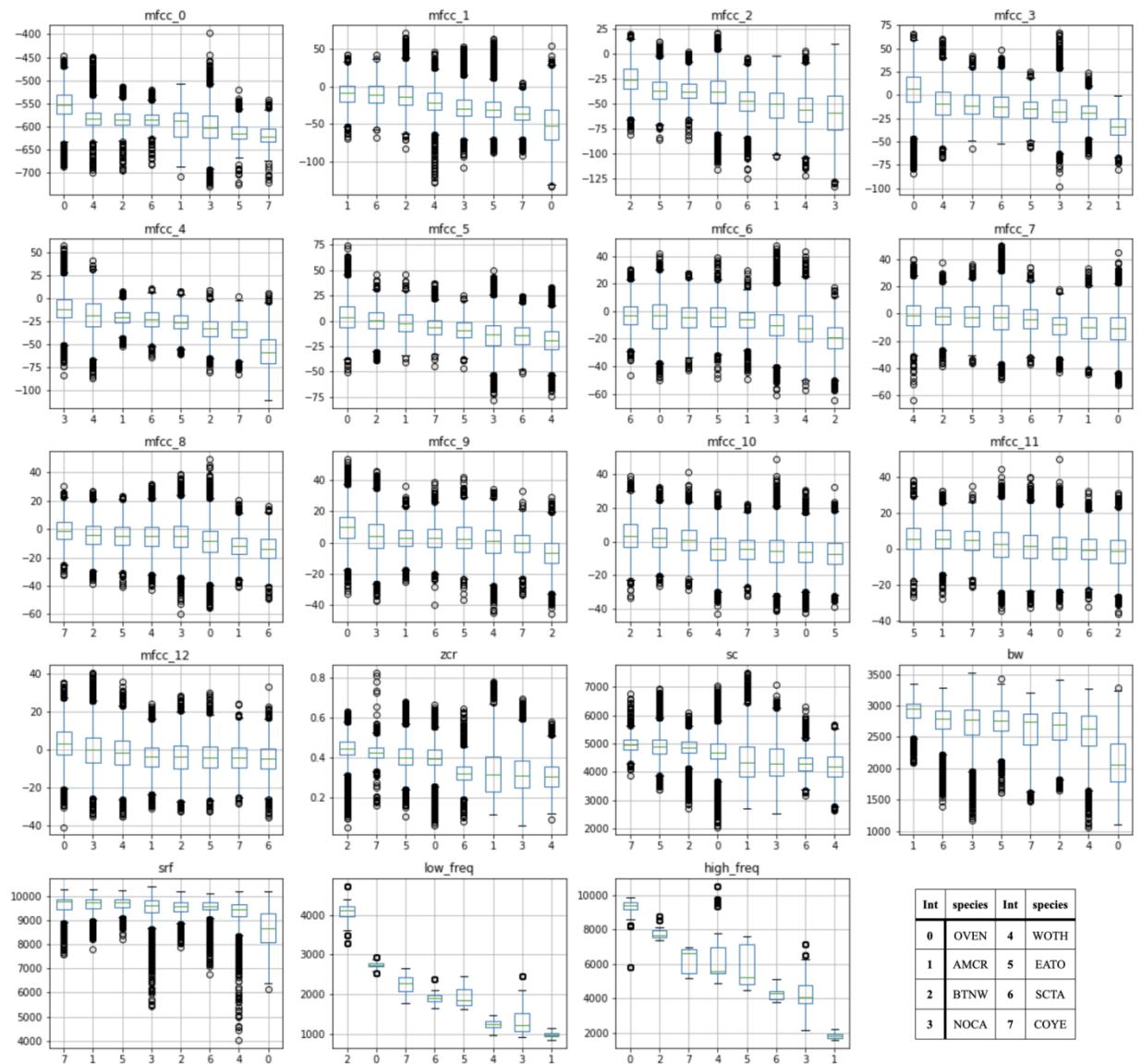
Common Name	Alpha Code	Latin Name
American crow	AMCR	<i>Corvus brachyrhynchos</i>
American goldfinch	AMGO	<i>Spinus tristis</i>
American redstart	AMRE	<i>Setophaga ruticilla</i>
American robin	AMRO	<i>Turdus migratorius</i>
Baltimore oriole	BAOR	<i>Icterus galbula</i>
Black-and-white warbler	BAWW	<i>Mniotilla varia</i>
Bay-breasted warbler	BBWA	<i>Setophaga castanea</i>
Black-capped chickadee	BCCH	<i>Poecile atricapillus</i>
Blue-gray gnatcatcher	BGGN	<i>Polioptila caerulea</i>
Brown-headed cowbird	BHCO	<i>Molothrus ater</i>
Blue-headed vireo	BHVI	<i>Vireo solitarius</i>
Blue jay	BLJA	<i>Cyanocitta cristata</i>
Brown Creeper	BRCR	<i>Certhia americana</i>
Black-throated green warbler	BTNW	<i>Setophaga virens</i>
Blue-winged warbler	BWWA	<i>Vermivora cyanoptera</i>
Canada goose	CANG	<i>Branta canadensis</i>
Carolina wren	CARW	<i>Thryothorus ludovicianus</i>
Cedar waxwing	CEDW	<i>Bombycilla cedrorum</i>
Common raven	CORA	<i>Corvus corax</i>
Common yellowthroat	COYE	<i>Geothlypis trichas</i>
Chestnut-sided warbler	CSWA	<i>Setophaga pensylvanica</i>
Downy woodpecker	DOWO	<i>Picoides pubescens</i>
Eastern towhee	EATO	<i>Pipilo erythrrophthalmus</i>
Eastern wood peewee	EAWP	<i>Contopus virens</i>
Hairy woodpecker	HAWO	<i>Leuconotopicus villosus</i>

Hermit thrush	HETH	<i>Catharus guttatus</i>
Hooded warbler	HOWA	<i>Setophaga citrina</i>
Kentucky warbler	KEWA	<i>Geothlypis formosa</i>
Louisiana waterthrush	LOWA	<i>Parkesia motacilla</i>
Nashville warbler	NAWA	<i>Leiothlypis ruficapilla</i>
Northern cardinal	NOCA	<i>Cardinalis cardinalis</i>
Northern flicker	NOFL	<i>Colaptes auratus</i>
Ovenbird	OVEN	<i>Seiurus aurocapilla</i>
Pileated woodpecker	PIWO	<i>Dryocopus pileatus</i>
Rose-breasted grosbeak	RBGR	<i>Pheucticus ludovicianus</i>
Red-bellied woodpecker	RBWO	<i>Melanerpes carolinus</i>
Ruby-crowned kinglet	RCKI	<i>Regulus calendula</i>
Red-eyed vireo	REVI	<i>Vireo olivaceus</i>
Red-shouldered hawk	RSHA	<i>Buteo lineatus</i>
Red-winged blackbird	RWBL	<i>Agelaius phoeniceus</i>
Scarlet tanager	SCTA	<i>Piranga olivacea</i>
Swainson's thrush	SWTH	<i>Catharus ustulatus</i>
Tufted titmouse	TUTI	<i>Baeolophus bicolor</i>
Veery	VEER	<i>Catharus fuscescens</i>
White-breasted nuthatch	WBNU	<i>Sitta carolinensis</i>
Wild turkey	WITU	<i>Meleagris gallopavo</i>
Wood thrush	WOTH	<i>Hylocichla mustelina</i>
Yellow-billed cuckoo	YBCU	<i>Coccyzus americanus</i>

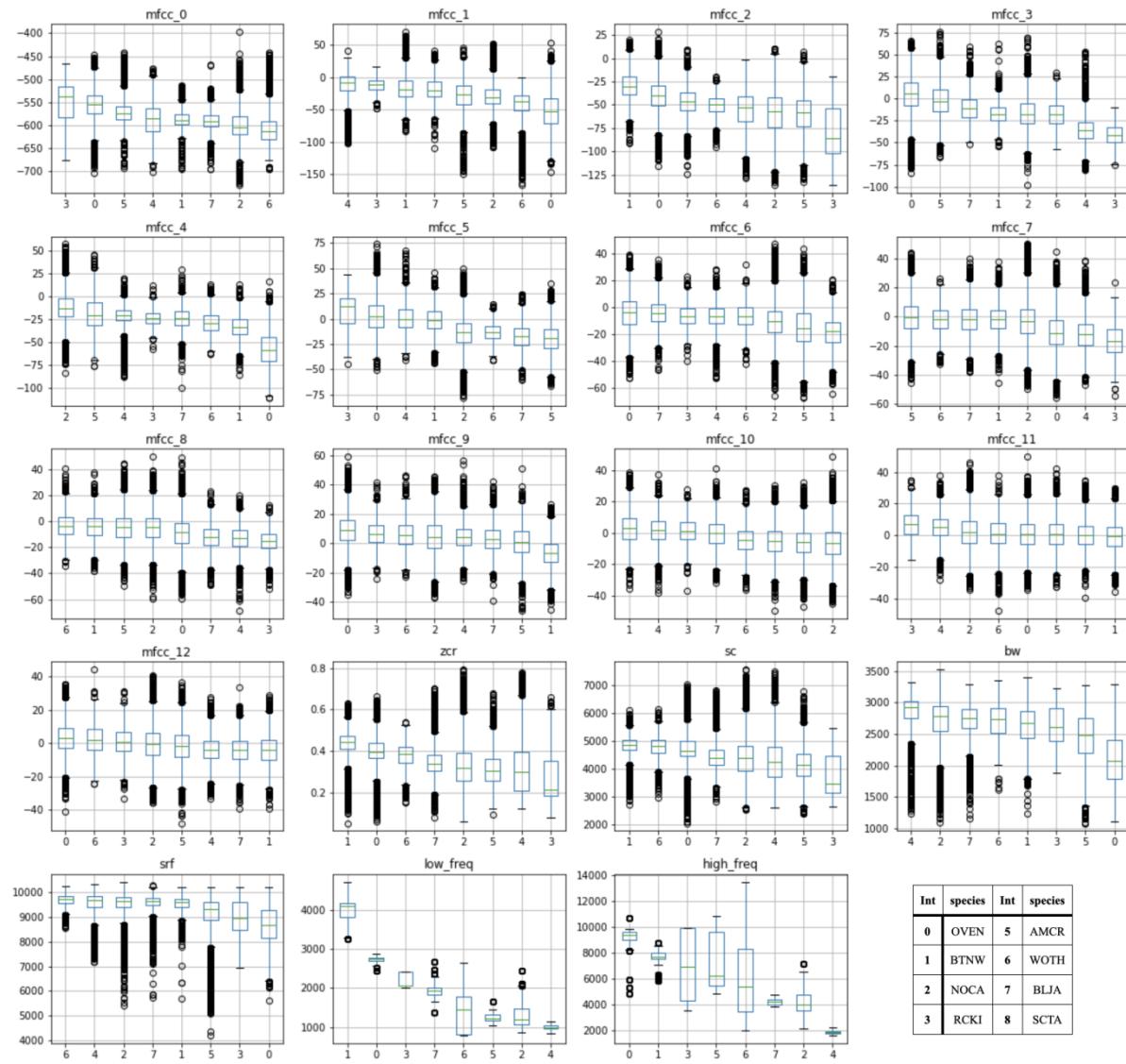
## 附录 2：不同重叠比数据，信号特征的描述性统计



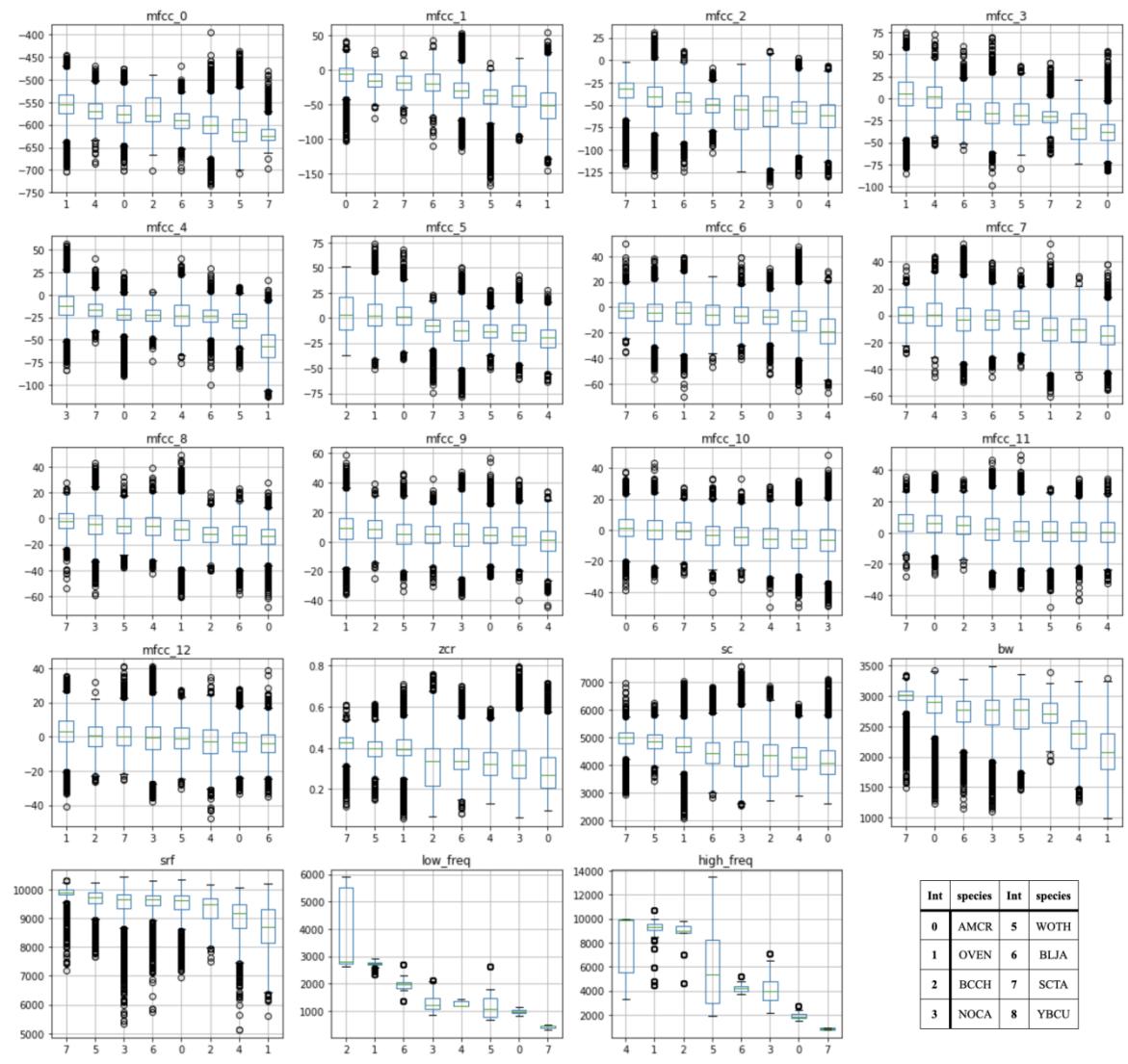
重叠比为 0% - 1% 的数据集各项特征分布



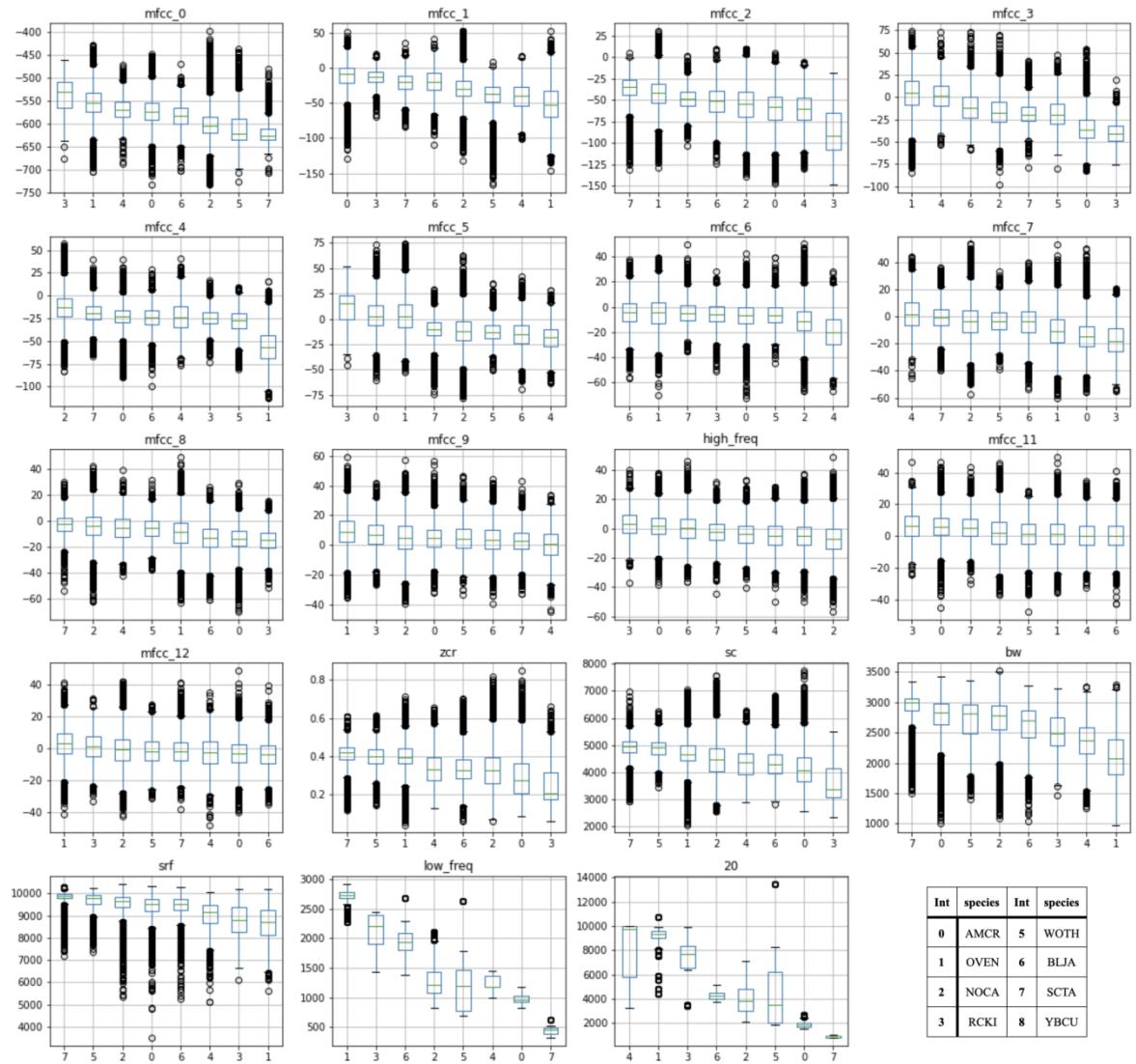
重叠比为 1% - 10% 的数据集各项特征分布



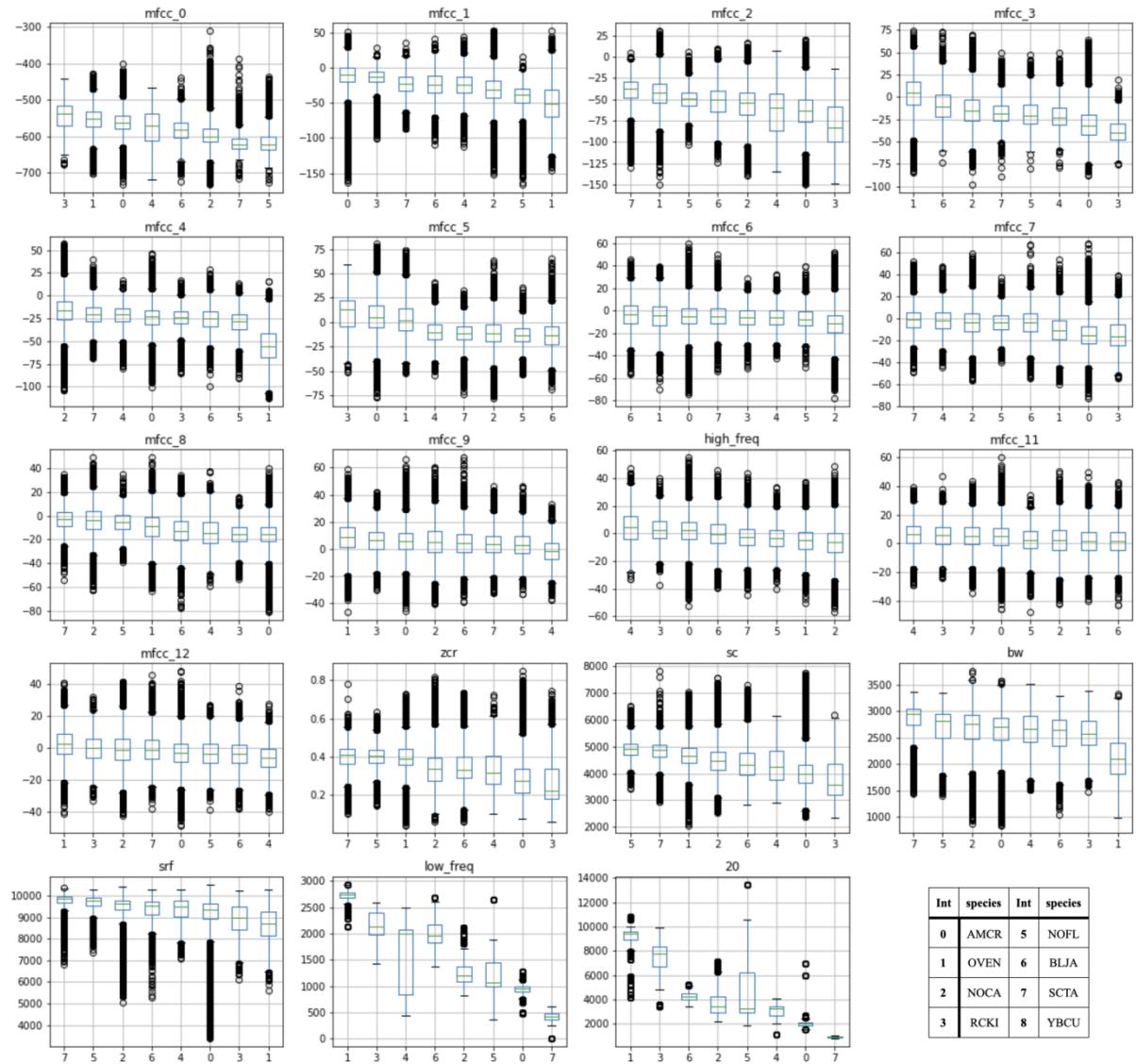
重叠比为 10% - 30%的数据集各项特征分布



重叠比为 30% - 50% 的数据集各项特征分布



重叠比为 50% - 70% 的数据集各项特征分布

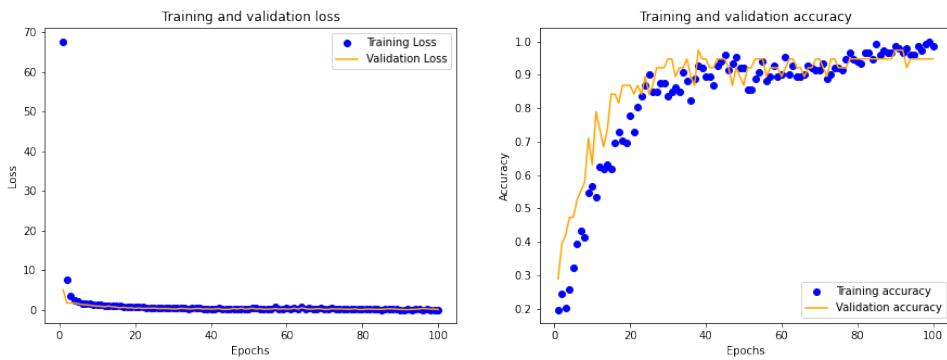


重叠比为 70% - 100%的数据集各项特征分布

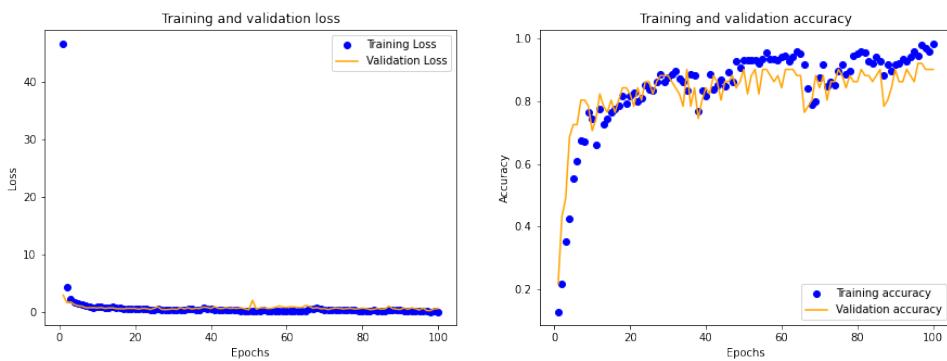
---

### 附录 3: CNN 的分类结果

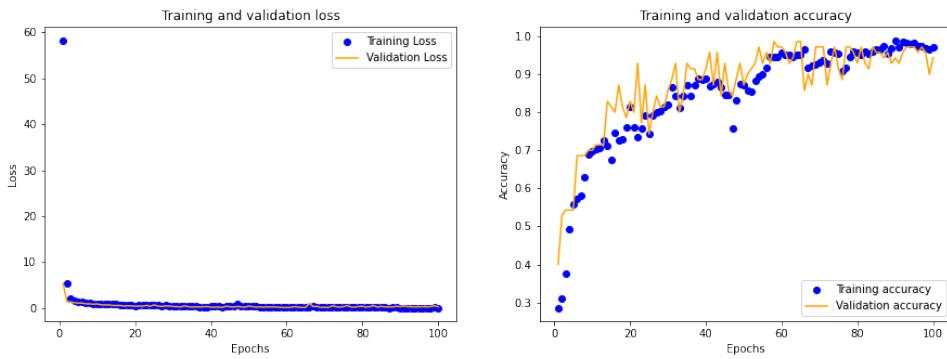
#### 【1】包含 Date 变量的分类结果



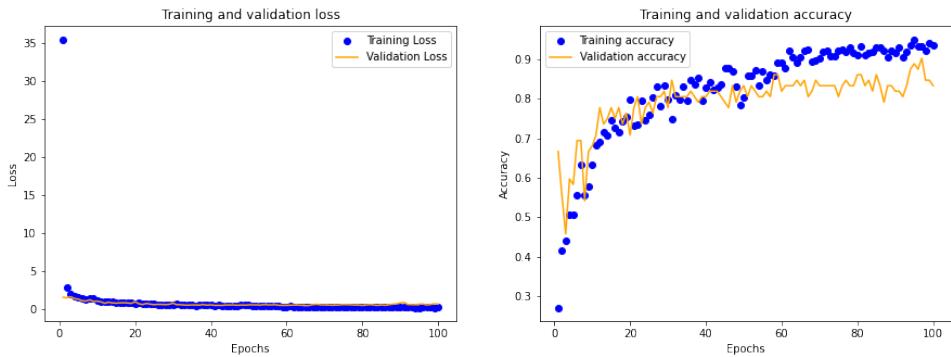
重叠比 0%-1%



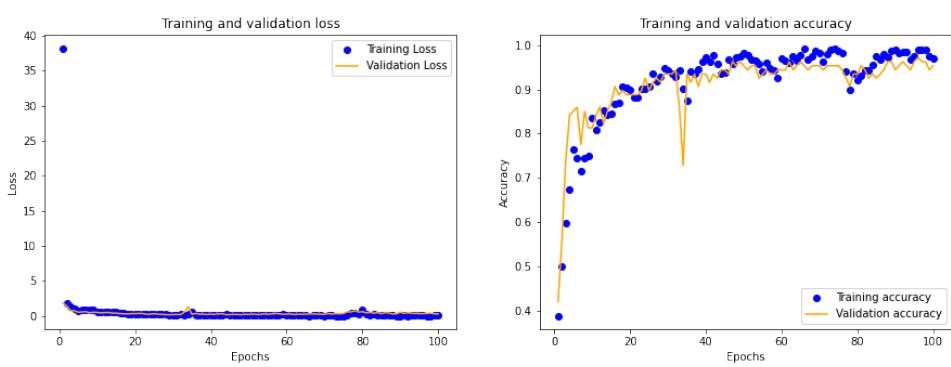
重叠比 1%-10%



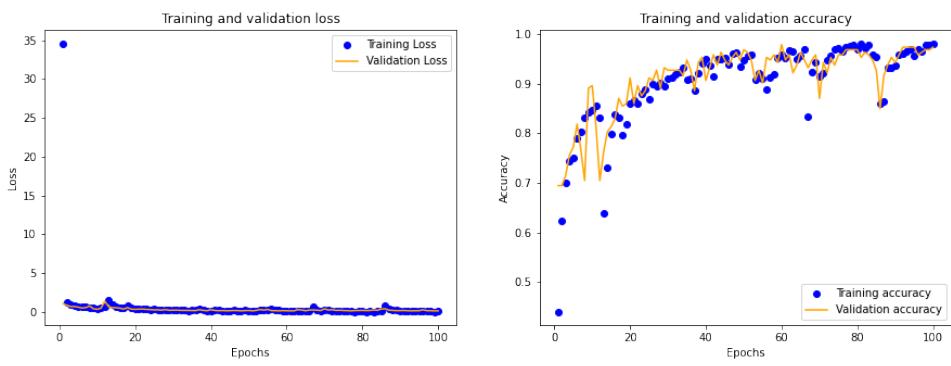
重叠比 10%-30%



重叠比 30%-50%

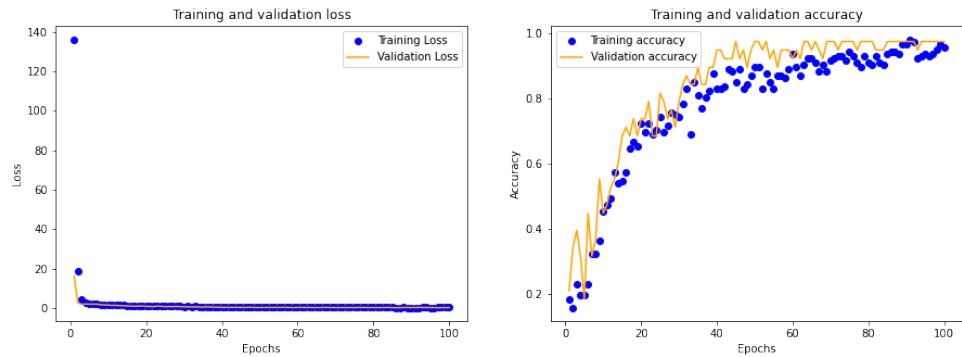


重叠比 50%-70%

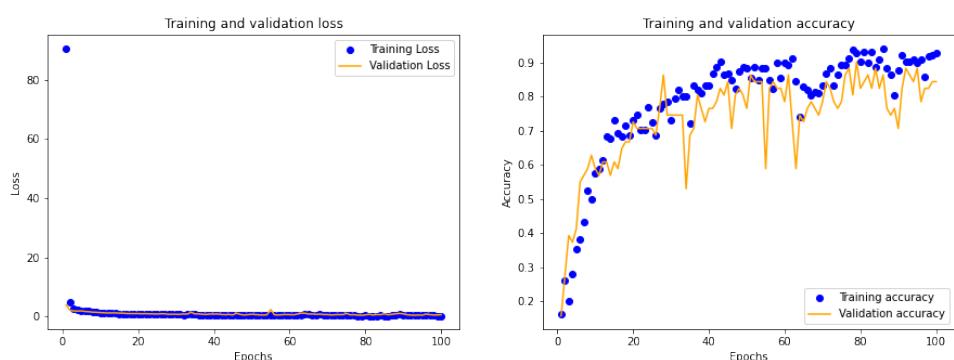


重叠比 70%-100%

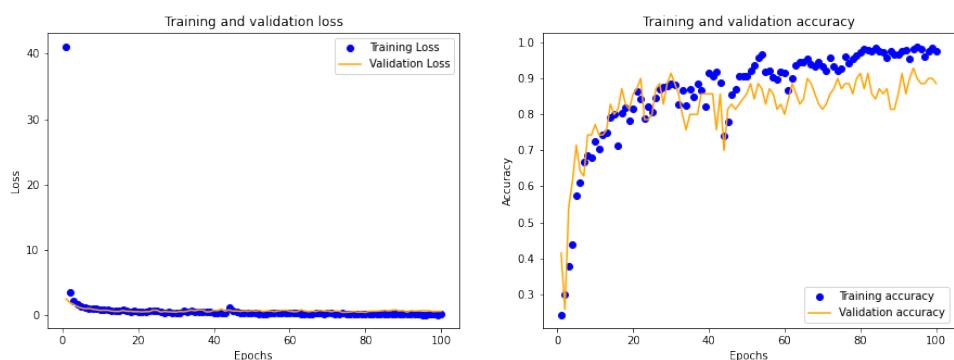
## 【2】不包含 Date 变量的分类结果



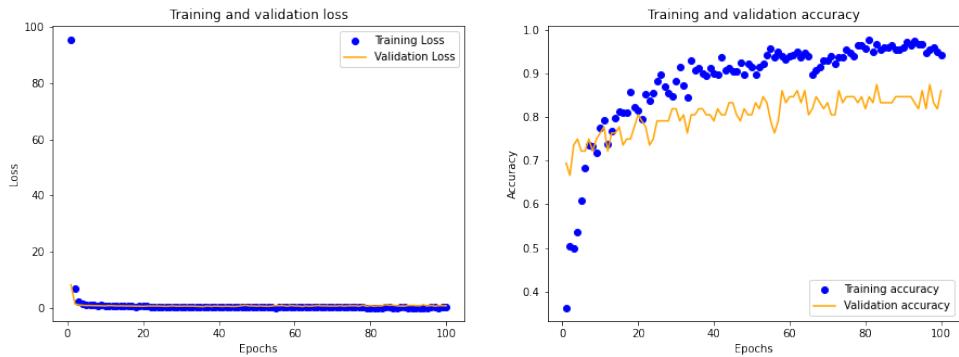
重叠比 0%-1%



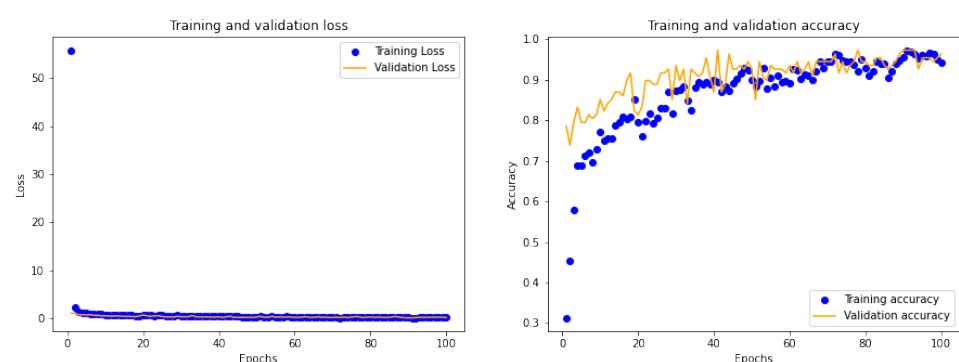
重叠比 1%-10%



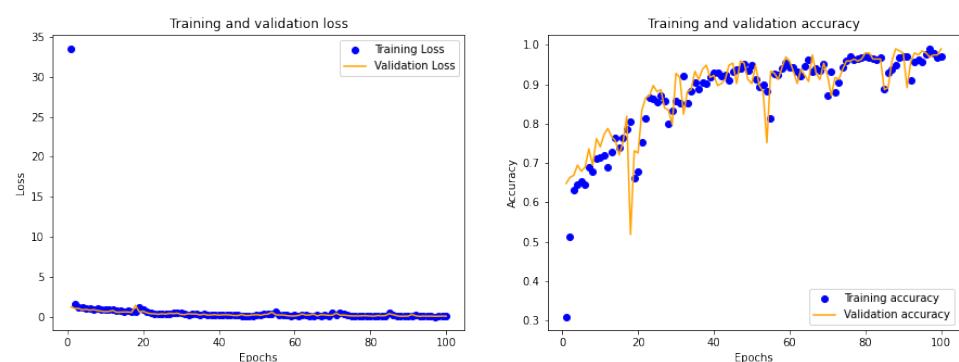
重叠比 10%-30%



重叠比 30%-50%

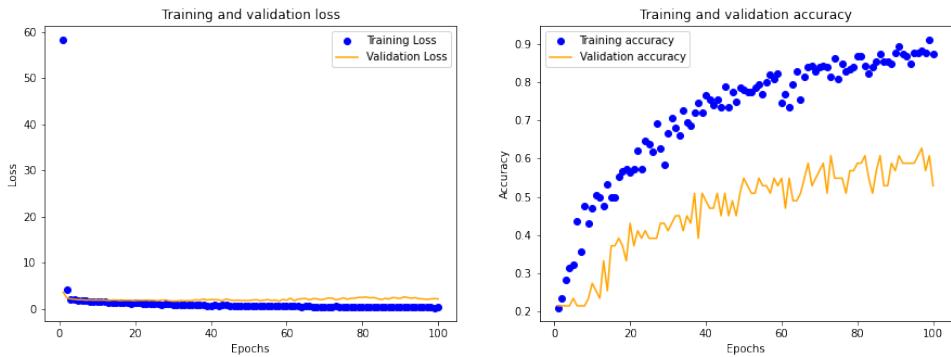


重叠比 50%-70%

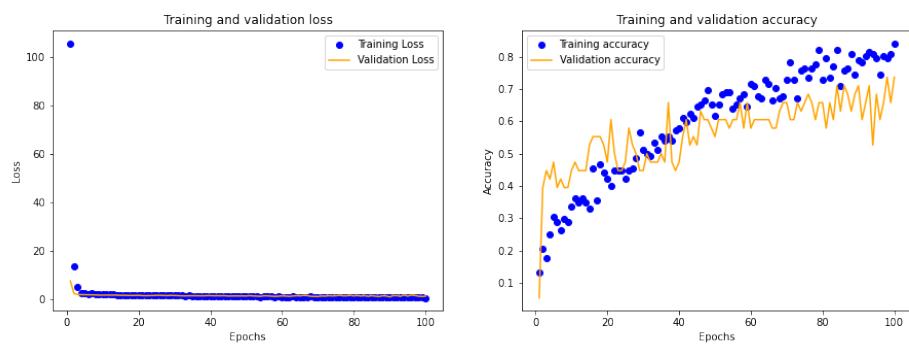


重叠比 70%-100%

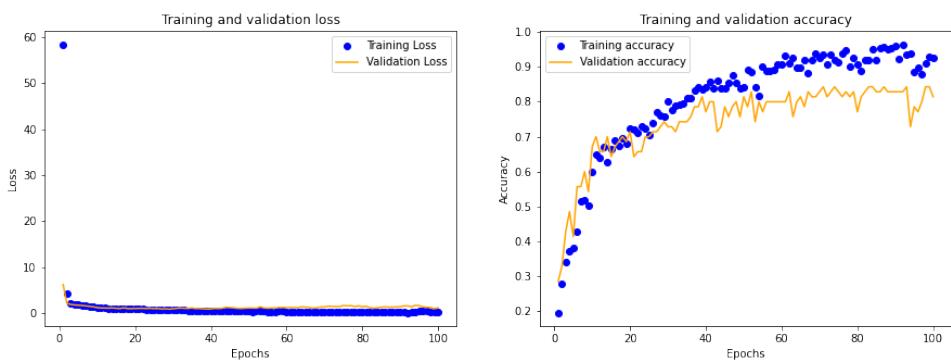
**【3】** 不包含 Date 以及频率范围(low\_freq, high\_freq)的分类结果



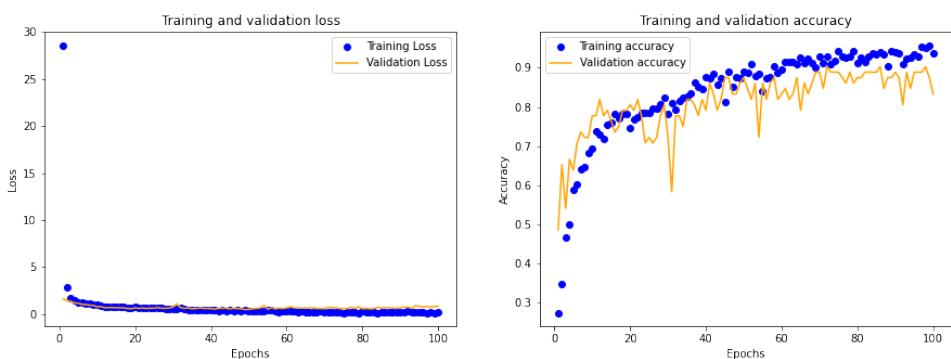
重叠比 0%-1%



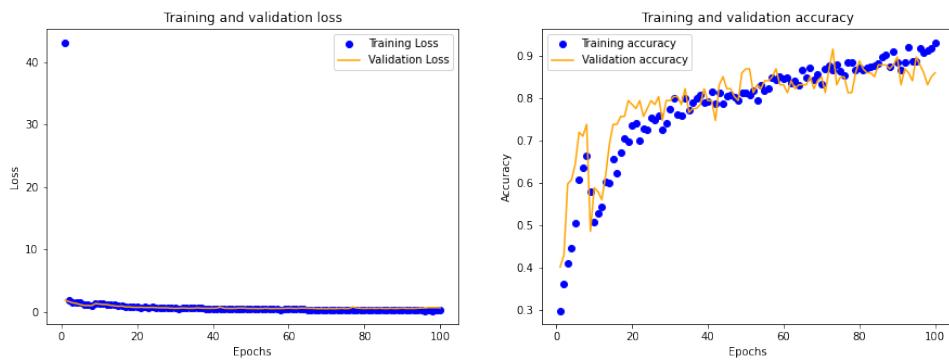
重叠比 1%-10%



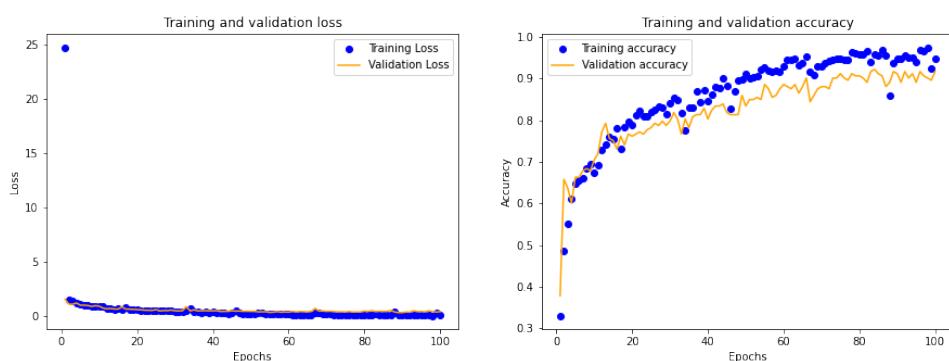
重叠比 10%-30%



重叠比 30%-50%



重叠比 50%-70%



重叠比 70%-100%