Machine Learning and Pattern Classification

Team Report

Assignment 2: Data Exploration

Team Aberrant

Ábel Boros (k11944603)
Henry-Willy Lechner (k12007370)
Luiz Henrique Madjarof (k12141871)

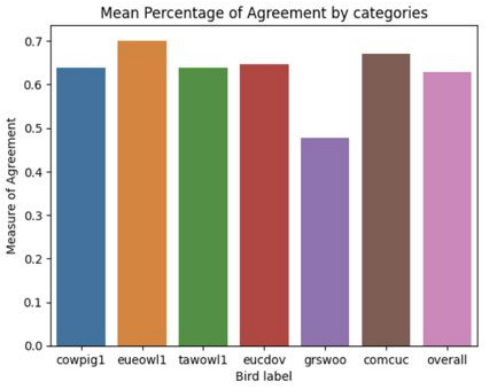
1. Annotator Agreement:

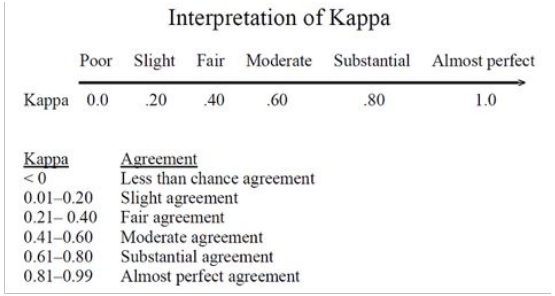
Method

Here, our objective was to assess the consistency of annotations by multiple annotators. To quantify it, we utilized a measure known as Inter-Annotator Agreement (IAA), which is commonly used to evaluate the level of agreement among annotators for a given category. Based on our preconditions, which included having at least 3 annotators for each fragment and binary labeling of birds as either the bird category or "other", we decided to use the Fleiss Kappa method. We began by iterating through the different bird directories and loading the label files one by one. We removed the majority vote column from each file, retaining only the annotations made by the individual annotators. We then preprocessed the data to create a matrix where each value in a row represented the number of annotators who voted for the corresponding label. Once the data was preprocessed, we used the fleiss_kappa method on the data and appended the results to a list. We repeated this process for each bird category in our dataset. Finally, we calculated the mean of the list to obtain the average agreement between annotators for each bird category. The results were interpreted according to established guidelines, as shown in the second image.

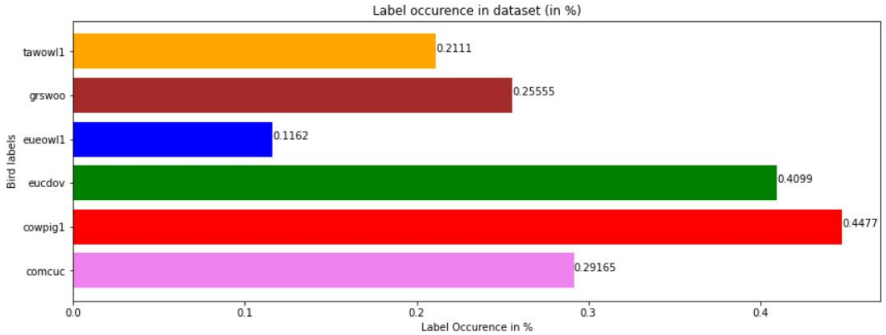
Findings:

Based on our findings, we determined that there was substantial agreement among annotators for most of the bird categories, with the exception of the "grswoo" category where the agreement was moderate. Overall, our results indicate that annotators exhibited substantial agreement in labeling the birds.





2. Label Characteristics



This data was created, by counting all the nonzero elements of the majority vote in the .labels.npy file, and then dividing them by the files length.

How are the class labels distributed?

Here we see the percentage of manually annotated bird calls in the .label.npy file.

This means, that in all the “comcuc” label files, for example, about 29 % of the file contains a annotation of the “comcuc”-bird call - the same scheme follows for the other bird annotation files.

Are the classes unbalanced, and how much?

Yes, we can see from the class label distribution, that they are really unbalanced for every bird.

comcuc	609.421264
cowpig1	1318.090873
eucdov	1154.670639
eueowl1	676.097402
grswoo	510.917666
tawowl1	623.820285

What is the average duration of a species’ calls?

The table below shows the mean duration of a bird call, derived from the annotation files (.label.npy) in ms.

This table was created, by using the majority vote of the .labels.npy file, and counting how many of the array inputs are equal to the assigned integer of the bird (e.g. “comcuc” has the assigned integer equal to “1”).

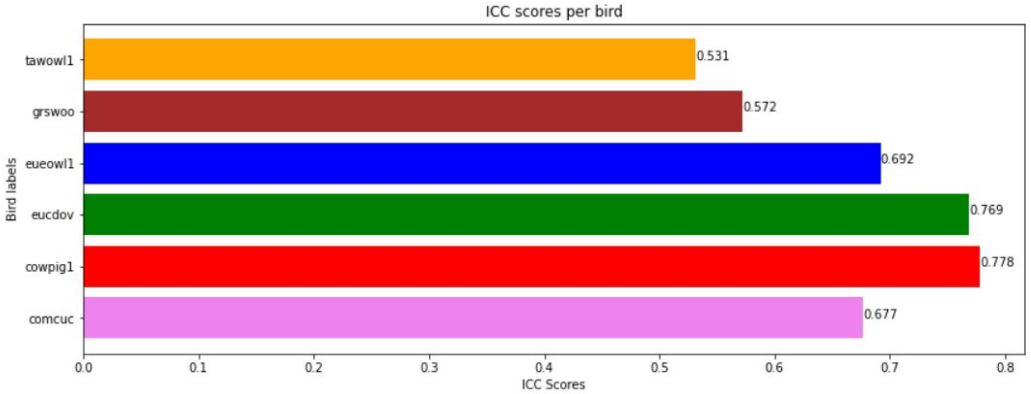
2. Label Characteristics

Are there large inter-/intra-class variations?

Inter-class variation:

From the table for the above question of average duration of a species' call, we can conclude, that there are two groups, the "comcuc", "eueowl1", "grswoo", "tawowl1" have an average call of about 510 to 680 ms, while the "cowpig1" and "eucdov" have very long average calls of about 1150 to 1320 ms.

Intra-class variations:



- **Less than 0.50:** Poor reliability
- **Between 0.5 and 0.75:** Moderate reliability
- **Between 0.75 and 0.9:** Good reliability
- **Greater than 0.9:** Excellent reliability

These thresholds tell us the reliability of the icc scores from above.

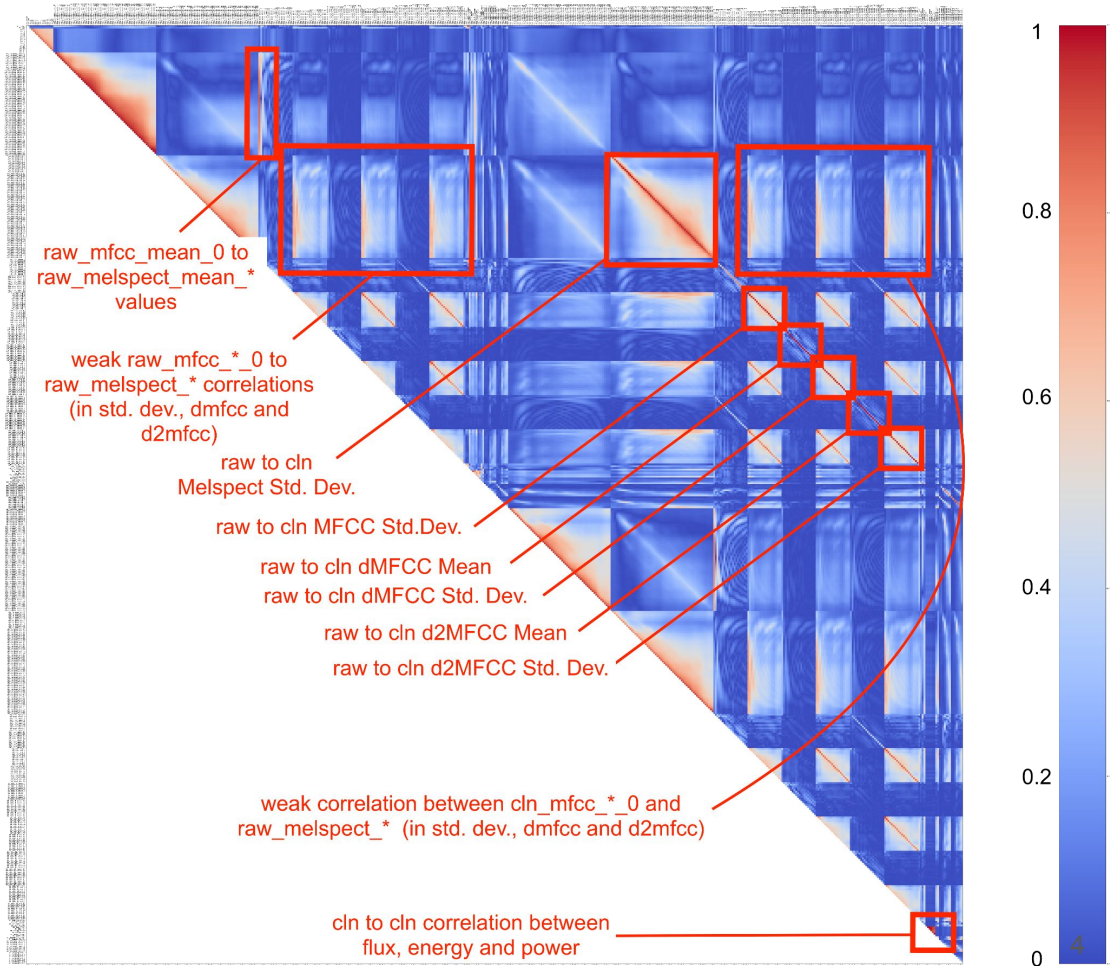
So we have "moderate reliability" for most of the intra-class annotations in our dataset.

3. Label Characteristics

Here, we explore our feature data and look into patterns of correlation between features themselves. The goal was to test if there were pairs of features with very high correlation that might make their use in the analysis of bird call data redundant. This redundancy could result from features that are just mirror images of each other, such that they always return the same values.

Method

Due to the large number of features, it was simply not feasible to look for feature correlation in a correlation matrix which would have a size of 548 rows and 548 columns. Instead, we created a heat map in which features with high correlation appear in a tone of the color red and low correlation appear in a variation of the color blue. That way, we can navigate a bit more easily through correlation values.



3. Label Characteristics

Findings:

We find a few redundancies and a few interesting unexpected correlations. To begin with redundancies, our correlation map shows that raw and cln features correlate at over 90% with their respective counterpart in their standard deviation, and overall d and d2 values, such that using these measures as a pair to analyse bird calls might not be very informative. A similar pattern is observed as well for the contrast features with a slight correlation in their mean and stronger in their standard deviation.

This correlation may be strong, but it is not very surprising since they come from the same measurements which are just applied to different variations of the same data. An interesting pattern of high correlation occurs between the MFCC_0 feature and all mel features. This correlation reaches over 80% and is observed both for the raw and cln data, although as the table shows, in the raw data, this correlation is generally stronger.

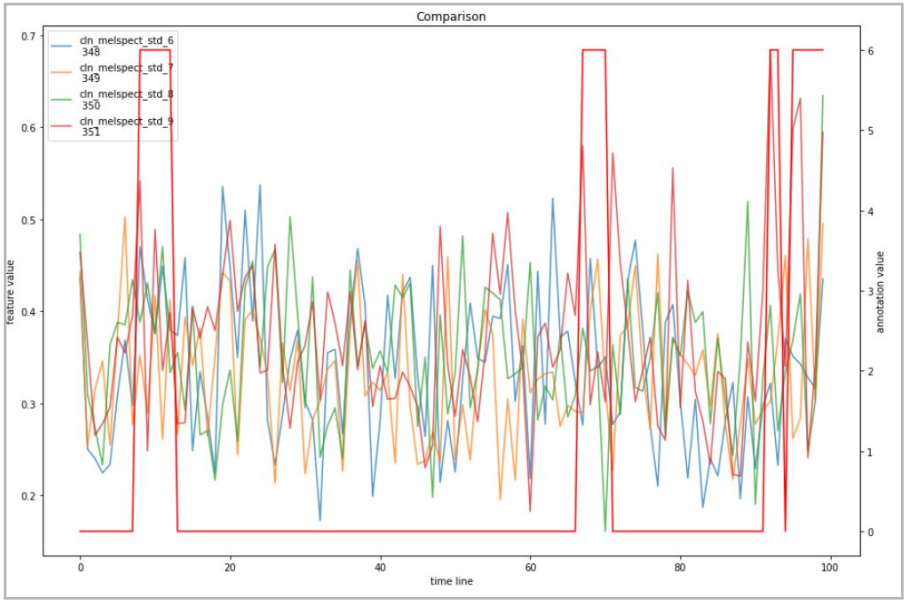
Another surprising correlation happens between the raw centroid mean and standard deviation and the ZCR mean and standard deviation. This occurs only in the raw data.

Finally, flux, energy and power features were shown to very strongly correlate (over 90% correlation) among themselves, but only in the cln data. There was a bit of correlation for those features in the raw data as well, but all below 80%.

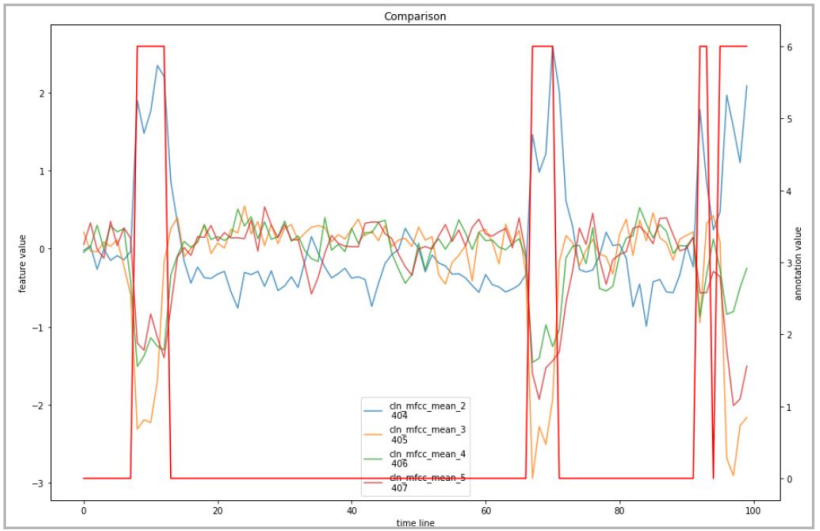
	Raw MFCC Mean 0		Cln MFCC Mean 0
raw_mfcc_mean_0	1.0	**cln_mfcc_mean_0**	1.0
raw_melspect_mean_26	0.8919	**cln_melspect_mean_21**	0.8190
raw_melspect_mean_27	0.8892	**cln_melspect_mean_23**	0.8188
raw_melspect_mean_25	0.8835	**cln_melspect_mean_22**	0.8172
raw_melspect_mean_28	0.8821	**cln_melspect_mean_20**	0.8166
raw_melspect_mean_29	0.8748	**cln_melspect_mean_24**	0.8149
raw_melspect_mean_24	0.8696	**cln_melspect_mean_19**	0.8082
raw_melspect_mean_30	0.8695	**cln_melspect_mean_25**	0.8062
raw_melspect_mean_31	0.8673	**cln_melspect_mean_26**	0.8027
raw_melspect_mean_32	0.8658	**cln_melspect_mean_18**	0.8002

4. Feature/Label Agreement

Some features are showing linear relationships with the annotation. Those could be useful for the classification, but also looking at the mean features could prove fruitful. As shown on the right plot



This plot shows the “eucdov” majority annotation (right scale), compared to some melspect std features (right scale).



This plot shows the “eucdov” majority annotation (right scale), compared to some mfcc mean features (right scale).

Some however, are just too noisy, or could potentially be combined with other features for usefulness, but are not inherently useful on their own (see left plot).

5. Consequences

To address the consequences of our observations for the purpose of our project, we will go through each of our findings:

1. Data annotators were in strong agreement with each other in their annotation. Therefore, the labeling data is consistent for training purposes in later steps of our machine learning project.

2. Having clear patterns in the label data that make the identification of a certain bird type possible would be another prerequisite for the training of our machine learning algorithm. In our dataset, as shown previously, we found that there was a very clear call length pattern for each bird type, such that bird identification would be possible by patterns of bird call length for instance.

3. We tested correlation between pairs of features in our feature data for all bird types. Our goal by testing correlation was to see if there were pairs of features that showed extremely high correlation with each other and which combined use in the training of our algorithm, therefore, might be simply redundant. For the majority of features in our data, such a strong correlation was not observed although there was strong inter-feature correlation for many features. The consequence of these findings is that the selection of features for training in different phases has to be done carefully as to avoid the pairing of features which use could be redundant.

4. Lastly, we looked for patterns of correlation between labels and features which could make the identification of bird types even more precise. We found that there were features that clearly oscillated with the labeling of bird calls, such that whenever a specific bird was identified by our experts there was a large movement in the feature data for some of the feature variables.

These observations go to show that identifying the bird types by their calls using the data we have at hand is feasible and precise. There are very clear and strong patterns in our data which can be used to train a machine learning algorithm.