MLPC 2023 Task 2: Data Exploration

Katharina Hoedt, Paul Primus, Florian Schmid, Jan Schlüter

Institute of Computational Perception
Johannes Kepler University Linz

March 20, 2023

1 Context

Our overall goal for this year's project is to train a system that can tell, for any instant in an audio recording, which bird species is audible (if any). Turning this into a machine learning problem entails:

1. splitting up the audio into small fragments, which will become the training examples,
2. computing a set of audio features for each fragment,
3. assigning a label to each fragment,
4. training a classifier to predict the label from the audio features.

Using your annotations from Task 1, we have prepared a dataset of 120,000 labeled audio fragments along with audio features. As a preparation for the classification task, you will now have a detailed look at the dataset.

2 Task Outline

Instead of blindly throwing that data at machine learning algorithms, a good data scientist will first do some exploratory data analysis. In your team, we would like you to analyze the data with respect to the following aspects:

1. **Annotator agreement**: How consistent are the annotations? Do different annotators agree in their labels for the same fragment?
2. **Label characteristics:** How are the class labels distributed? Are the classes unbalanced, and how much? What is the average duration of a species' calls (or drumming)? Are there large inter-/intra-class variations?

3. **Feature characteristics:** How are the features distributed? Are there any pairs or subsets of features that seem highly correlated or redundant?
4. **Feature/Label agreement:** Which features seem useful for classification? Which ones are correlated with the labels?
5. **Consequences:** Any conclusions you can draw from your analysis for doing classification?

For this, use any kind of statistical computation or visualization that you find enlightening. Compile your results into a report, in the form of a slide deck, with at most 7 slides (plus a title slide that includes your team name and member names). Make sure to address all five aspects in your report.

Submit your slide deck as a PDF on Moodle by **April 14th**. Only one team member needs to submit on behalf of the team.

3 Dataset

The dataset download links are available on Moodle, and the format and content of the dataset is described in detail in the slide deck for Meeting 2 (March 20). Please refer to that slide deck for information on the audio features, the derivation of the labels, and the file formats.

4 Grading

Reports for this task are evaluated according to the following criteria, for each of the five aspects given in the task outline:

- **Thoroughness and correctness:** Have you seriously thought about the problem? Are the proposed procedures and experiments sound, correct?
- **Presentation, Completeness, Clarity:** Are the ideas, features, algorithms, and results described clearly? Does the report contain all the information needed for the reader to reproduce the results (e.g., exactly which features were used, what were the parameter settings, ...)? Are results presented in a structured way (e.g., tables, graphics) that permits the reader to easily grasp them?
- **Punctuality:** The reports must be submitted in time. Any delay will result in reduced grades. Specifically, submitting on April 15 will deduct 1/3 of the points, submitting on April 16 will deduct 2/3 of the points, and submissions on April 17 or later will be rejected.
- **Additional bonus: Creativity:** Have you tried to explore alternative paths? Have you tried to come up with novel, creative ways of addressing the problem?

The data exploration task accounts for 20% of the grade; the upcoming classification task and challenge will account for 40% each.

5 Task Deliverables

The only deliverable for this task is your team's report. Upload it as a PDF on Moodle by **April 14th**.