MLPC 2023 Task 4: Challenge

Katharina Hoedt, Paul Primus, Florian Schmid, Jan Schlüter

Institute of Computational Perception
Johannes Kepler University Linz

May 22, 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. In Task 2, you explored the dataset, which included looking for correlations between features (to find features that are redundant) and correlations of features with labels (to find features that are useful). For Task 3, you selected four classifiers, trained and evaluated them on the dataset using cross-validation. The final task now is to optimize your classifier(s) for an application scenario. As in a real machine learning project, the final application deviates slightly from your previous research on the methods, and may benefit from additional tuning.

2 Task Outline

You are provided with a test set of additional recordings, including the same set of audio features that was available for training, but excluding the labels. The task for your team is to provide labels for the test set. As the evaluation will differ a bit from

what you used during your previous experiments, there are opportunities to tune your system to maximize your evaluation score. You may try different solutions on the test set by submitting your predictions to a challenge server that will run the evaluation. You will document not only the final solution, but also the process of getting there, in a slide deck to be submitted by **June 22nd**.

3 Challenge Server

We have set up a challenge server that allows you to download the test set and then submit your predictions for the test files up to twice a day. The web address and login credentials are given on Moodle.

4 Dataset

The training set stays unaltered from the previous tasks. Via the challenge server, we provide a test set of the following characteristics:

- It consists of 16 files of 10 minutes each, so each file has 3000 fragments of 200 ms.
- For each fragment, we provide a feature vector in the same format as for the training data.
- In contrast to the training data, each file can contain vocalizations of different bird species (but not in the same fragment). This was achieved by cutting multiple single-bird recordings and concatenating them. Cutting was only done in positions labeled as 'other', and pieces are not shorter than 6 seconds, so there is still considerable temporal continuity.
- Ground truth labels were formed taking the majority vote of annotators per fragment. Those labels are kept hidden on the server, and not distributed as part of the dataset.

5 Evaluation

The scenario for this task is to find a system that allows biologists to count birds in audio recordings via as little manual corrections of the automatic labelling as possible.

Evaluation is based on a fragment-wise comparison of predictions against ground truth labels. Each combination of the 7 possible predictions and the 7 possible ground truth labels is associated a value in Euros via the gain matrix given in Table **??**, representing how much money the biologists would save each month compared to labelling recordings manually from scratch. The gain matrix encodes the following principles:

- A correct bird prediction saves 1 EUR, a correct background prediction saves 5 cents

		predicted						
		other	comcuc	cowpig1	eucdov	eueowl1	grswoo	tawowl1
ground truth	other	0.05	−0.2	−0.2	−0.2	−0.2	−0.2	−0.2
	comcuc	−0.25	1	−0.3	−0.1	−0.1	−0.1	−0.1
	cowpig1	−0.02	−0.1	1	−0.1	−0.1	−0.1	−0.1
	eucdov	−0.25	−0.1	−0.3	1	−0.1	−0.1	−0.1
	eueowl1	−0.25	−0.1	−0.3	−0.1	1	−0.1	−0.1
	grswoo	−0.25	−0.1	−0.3	−0.1	−0.1	1	−0.1
	tawowl1	−0.25	−0.1	−0.3	−0.1	−0.1	−0.1	1

Table 1: Gain matrix giving the win or loss in EUR/month for predicting a particular fragment label (columns) depending on the true label (rows).

ground truth:	0001110000500000000330330033330000
ignored:	x x x x x x x xx x

Figure 1: Fragments that immediately precede or follow a call are ignored.

- Predicting no bird when there is one costs 25 cents
- Predicting a bird when there is none costs 20 cents (these errors are easier to correct, since they only require checking each predicted call and not all positions without a bird)
- Predicting the wrong bird costs 10 cents (at least a bird was predicted, that already helps)
- Exception: Mistaking any bird for a pigeon (cowpig1) costs 30 cents, missing a pigeon costs 2 cents (pigeons are not important for the biologists to assess habitat quality)

To compute the total savings for an algorithm, the confusion matrix of the algorithm (which counts how often each combination of predicted and true label occurs) is multiplied element-wise with the gain matrix and then summed up.

As the main goal is to count birds, exact call boundaries are not important. Thus, when computing an algorithm's gain, we ignore all fragments that have a ground truth label of 'other' and are immediately preceded or followed by a fragment with a ground truth label of any bird. Figure 1 illustrates this for an example. This entails that short pauses of one or two fragments in a longer call sequence are ignored. For ignored fragments, it is irrelevant whether your algorithm predicts 'other' (as the ground truth would imply) or a bird.

6 Submitting Predictions

Predictions for your team are to be uploaded via the challenge server. You may upload up to two sets of predictions per day, to have plenty of room for experimentation (but not infinitely much room to overfit on the test set).

Each set of predictions for the 16 test recordings must be submitted as a single CSV (comma-separated values) file. It must contain one row per file that starts with the file name without extension, i.e., `test00` up to `test15`, followed by a comma, and then a comma-separated list of 3000 class labels (between `0` and `6`) for the file's 3000 fragments. Any other rows are ignored (so you can have a header or omit it), but additional columns are not allowed. As part of the dataset downloads on the challenge server, we provide an example submission file (`example_submission.csv`) of the correct format that labels all fragments as 'other' (class `0`).

7 Report

Part of the task is to submit a report that documents your process to reach your final solution. This time, there is no predefined set of aspects or questions to address. Instead, you should describe the main ideas you had and hypotheses you tried, as well as their outcome after experimentally verifying or falsifying the hypotheses (via experiments using cross-validation on the training set, or by submitting predictions to the challenge server).

Include visualizations and/or tables to present your results. Compile a slide deck of at most 20 slides (plus a title slide that includes your team name and member names).

Submit your slide deck as a PDF on Moodle by **May 18th**. Only one team member needs to submit on behalf of the team.

8 Grading

Reports for this task are evaluated according to the same criteria as before:

- **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 June 23 will deduct $1/3$ of the points, submitting on June 24 will deduct $2/3$ of the points, and submissions on June 25 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 accounted for 20% of the grade; the classification task and challenge account for 40% each.

9 Task Deliverables

There are two deliverables for this task: (1) your predictions for the test set, to be uploaded to the challenge server as often as you want (up to twice a day) until **June 22nd**, (2) your team's report, to be uploaded as a PDF on Moodle by **June 22nd**.