Probabilistic Machine Learning
2024 BUPT Summer School

Coursework Guideline

July 3, 2024

1 Overview

Mosquitoes are one of the deadliest animals by the number of human deaths they cause, as some of the mosquito species are the major vectors of diseases such as malaria, dengue, and yellow fever. We would like to develop a machine learning approach to detect those dangerous mosquito species by sound.

Creating your own sound datasets is impractical. So, we use an open dataset (`https://zenodo.org/records/4904800`), which provides a large multi-species sound samples and their ground truth labels i.e. species. In this coursework, you will develop suitable classification techniques to evaluate the results. You will create a poster to introduce the problem, your solutions, and results. You will need to submit a Jupyter notebook by 12noon on Friday 19th July 2024. Any of your experiment figures and/or tables included in your poster need to be reproducible in the Jupyter notebook that you submit.

2 Datasets

You are provided with the datasets in the repository (`https://zenodo.org/records/4904800`), which includes a multi-part zip of audio files (around 4GB in total) and a CSV file "neurips_2021_zenodo_0_0_1.csv" that contains the meta data.

Essential task: Detect mosquito presence as indicated in Column F of the CSV file under "sound_type" (mosquito, audio, background).

Optional task: For those seeking additional challenges, you are encouraged to explore species classification using the information in Column G "species" of the CSV file. For this coursework, we consider classifying only the these six mosquito species (An. arabiensis, Culex pipiens, Ae. aegypti, An. funestus ss, An. squamosus, An. coustani).

Notes: You are provided here all the data including audio samples and labels. Present your results in a way that demonstrates how well your model is likely to perform on unseen samples. .

More information about the dataset can be found in [1].

3 Group allocation

You should form a group of up to 4 people to complete the coursework. You need to finalise your group choice by 5pm on Thursday the 11th July. Each person will need to explain in the final presentation the individual contribution he/she makes to the completion of the coursework.

4 Deliverables

You should submit the following two files:

1. A poster that describes the problem, literature review, your approach and evaluation results.

2. A Jupyter notebook containing code to train and evaluate your developed algorithms with the data available. The submitted Jupyter notebook needs to include the figures for evaluating performance which you included in your poster.

There will be a group presentation of 3 minutes with the poster for each group, in which the evaluation is based on. A Chinese presentation is fine, although an English presentation is encouraged with a 5 point bonus added to the coursework score. (**Starting from 6:30pm on Friday 19th July**)

5 Coursework poster

The poster should be in A0 size, in digital format (No hard print is needed). It is important to demonstrate your understanding of the subject area, your approach to the coursework and present the outcome. There are many useful guidelines on how to design an academic poster, for examples: `https://www.makesigns.com/tutorials/`. You are also provided with a sample poster as a research presentation poster example. You are free to choose either the landscape or portrait orientation.

You should consider including sufficient evidence of your work through the poster, covering the following areas in the marking criteria.

6 Marking criteria

You will be evaluated based on the following criteria with your **presentation** and **poster**.

- **Introduction to the project**: Provide appropriate background information, and detail the objectives of the project. Literature review should be carried out on relevant topics and how they may relate to your work. You are expected to research and discuss other sources of information, but must show their origins by referencing all sources used. The reference list should be included in the poster. (30/100)
- **Analysis of the problems**: Research and technical issues, challenges, and description of the approach you have developed with justification. Effective graphic illustration will be appreciated. Multiple classification techniques can be designed and implemented. It could be also a multiple versions of the same technique, with different hyperparameters or different ways of training approaches, so you can compare the performance across different models in the poster. (30/100)
- **Implementation and evaluation of the methods**: You need to discuss the system results, present the data you used, how many of them for training and validation in your model, and what are suitable performance metrics etc. (hint: you may want to include not only accuracy here as your performance metrics). Figures and/or tables can be useful to present such information, especially when multiple approaches are developed and comparison of various approaches/choices are conducted. Figures included in the poster need to be reproducible as demonstrated in your Jupyter notebook (otherwise they will not be counted in the scores). (40/100)
- An English presentation is encouraged with a 5 point bonus score. (bonus: 5/100)

References

[1] Ivan Kiskin et al. HumBugDB: a large-scale acoustic mosquito dataset. In *Proc. Adv. Neur. Inf. Process. Sys. (NeurIPS 2021) Track on Datasets and Benchmarks.*, 2021.