MLPC Report - Task 3: Classification Experiments

Team OBSERVE

Johannes Grafinger **Jonas Gantar** **Leonhard Markus Spanring** **Reinhard Josef Pötscher**

Contributions

Johannes, Jonas and Reinhard were responsible for tasks 1.) Labeling Function, 2.) Data Split, 3.) Audio Features and 4.) Evaluation. Leonhard was responsible for task 5.) Experiments. All of us together were responsible for task 6.) Analysing Predictions. We prepared this report in the same constellation. We all worked together on the presentation. We held regular meetings at which each member presented their results and the others critically evaluated the work. Everyone communicated via a dedicated Discord server and the editing was done over a Github repository.

1.) Labeling Function : For your analysis, you may focus on a subset of the 58 classes provided

For this task, we mainly focused on the following, randomly chosen subset of classes of size 7: ['Airplane', 'Bird Chirp', 'Power Drill', 'Cowbell', 'Siren', 'Trumpet', 'Rooster Crow'].

a.) Assess how accurately the applied labeling functions capture the intended classes.

a.).1 Do the mapped classes correspond well to the free-text annotations?

a.).2 Are the labeled events clearly audible within the indicated time regions?

b.) Which audio features appear most useful for distinguishing between the classes of interest?

In order to answer this question, we first of all randomly collected 1000 frames from the dataset (representing a collection of all frames from all files) per corresponding class, resulting in a total of around 7000 frames (a bit less, since some frames contain several of the classes). For each frame, we additionally include the previous and next 2 frames as context length by simply concatenating them with the features from the central frame (in temporal order). Each class therefore has 1000 vectors of shape 5 * 942 = 4710, where 942 represents the summed number of dimensions for all features.

For each class, we then estimated the Mutual Information between the label vector and all features dimensions. For each feature, we then calculated the average Mutual Information over all feature dimension corresponding to it. This gives us a measure for how informative the feature generally is for predicting the label correctly, which also captures nonlinear relationships. We then normalized the values per label, resulting in the first plot of (Figure 1). However, since this approach cannot capture effects features only have in combination with one another, we also tried fitting a simple Random Forest Classifier on the data and extracted its feature importances, resulting in the second plot of (Figure 1).

The features 'embeddings', 'melspectrogram', 'mfcc' and 'contrast' all seem to be pretty useful according to our Mutual Information graphic, with 'embeddings' being the most important one. However, the fact that the Random Forest almost exclusively relies on the embeddings might suggest heavy correlations between them and the other features. Which does make sense, considering many of the other features are usually used for creating such audio embeddings.

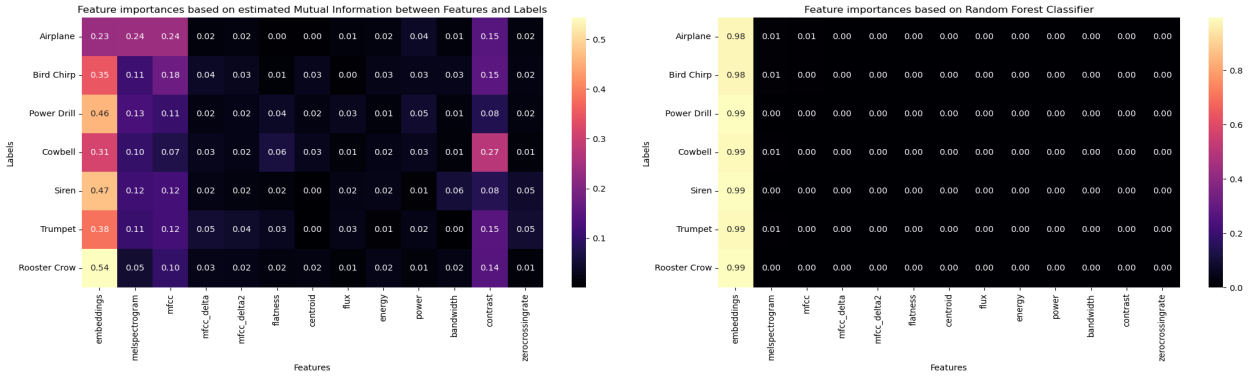

Figure 1: Estimated feature importances as by Mutual Inforamation and Random Forest methods.

c.) How well do the chosen audio features group according to the discretized class labels? Do samples of the same class form tight clusters?

Using the 4 features mentioned above, we reduced the high-dimensional data to 2 dimension using T-SNE and plotted the results in (Figure 2). Note that for examples corresponding to more than 1 of the classes, we randomly chose 1 for coloring.

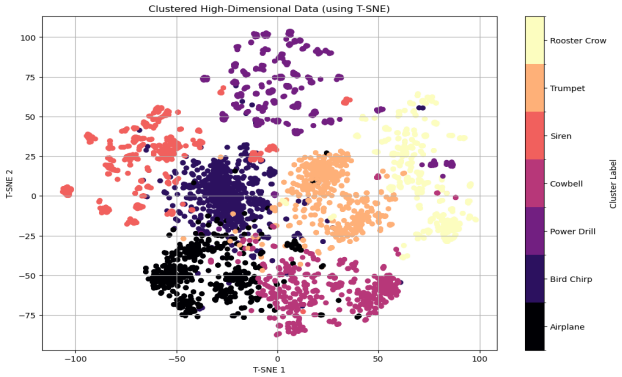

Figure 2: Downprojected examples using selected features and their class correspondences.

Overall, the results look very reasonable. Clusters are identifiable and examples within mostly have the same attributed class. Some overlaps are of course inevitable, considering the relatively high complexity of the dataset.

2.) Data Split

a.) Describe how you split the data for model selection and performance evaluation.

b.) Are there any potential factors that could cause information leakage across the data splits if they are not carefully designed?

b.).1 If yes, how did you address these risks?

c.) Describe how you obtained unbiased final performance estimates for your models.

3.) Audio Features

a.) Which subset of audio features did you select for your final classifier? Describe the selection process and the criteria you used to make your choice.

After reflecting on the results from section 1.), b.), we have decided to run our experiments only on the 'embeddings' feature, as there was also was no significant drop in performance for tests we ran on a small scale. Our reason for only using that feature stays the same as discussed in the last paragraph of that section: Seemingly heavy correlations and possible redundancy of the other three features considering the possible creation process of these audio embeddings.

b.) Did you apply any preprocessing to the audio features? If so, explain which techniques you used and why they were necessary.

Preprocessing is an important step for ensuring consistency in the data and making the lives of models like Support-Vector-Machines much easier as the features are confined in a smaller space. We used the global mean and standard deviation from the training data and normalized our training, validation and test sets using these global statistics (to avoid data leakage). We then additionally normalized the features frame-wise to have unit mean and variance for each example.

4.) Evaluation

a.) Which evaluation criterion did you choose to compare hyperparameter settings and algorithms, and why?

b.) What is the baseline performance? What could be the best possible performance?

5.) Experiments

a.) For at least three different classifiers, systematically vary the most important hyperparameters and answer the following questions for each of them:

a.).1 How does classification performance change with varying hyperparameter values? Visualize the change in performance.

a.).2 (To what extent) Does overfitting or underfitting occur, and what does it depend on?

b.) After selecting appropriate hyperparameters, compare the final performance estimate of the three classifiers.

6.) Analysing Predictions

Find two interesting audio files that have not been used for training and qualitatively evaluate your classifier's predictions.

a.) Use the spectrogram and the sequence of predictions to visualize the classifier output.

b.) Listen to the audios and inspect the corresponding predictions of the classifier.

b.).1 How well does the classifier recognize the classes?

c.) What are particular problematic conditions that cause the classifier to mispredict classes?

c.).1 Can you think of simple postprocessing steps that might help improve the predictions?