Introduction:

Due to a low number of data samples, audio domain, and scarce sources open to the public in environmental sound classification; devising accurate classification comes as a challenge. So, in this project, depending on the open source dataset ESC-50 and the dataset owner's baseline methods, different supervised learning methods are applied for sound classification. Dataset owner's GitHub link containing more detailed material and other people's solutions is provided in the appendix.

Problem statement: Using a dataset for ESC (environmental sound classification) reaching maximum accurate classification by machine and deep learning methods.

The Dataset: The main dataset comprises 2000 samples of environmental sounds each with 5 seconds length in .wav format. There are 50 classes each with 40 samples. The 50 classes make up the ESC-50 dataset. Some of these classes are also labeled as ESC-10 which constitutes the simplified 10 classes as a benchmark. Sound clips were sampled at 44.1kHz frequency and a single channel. The 50 classes are loosely categorized into 5 groups: animal sounds, natural soundscapes and water sounds, human (non-speech) sounds, interior/ domestic sounds, and exterior/urban noises.

Materials and methods:

In this project, Anaconda 3, python 3 and Jupyter notebook are used for coding.

Main packages used for:

Liborsa : feature extraction(mfcc, mel, zero crossing rate) and signal loading.

Pandas : reading csv files.

NumPy : matrix manipulations.

Scikit-Learn: fitting data, classification algorithms

PyTorch : feedforward neural network

Feature extraction methods:

From each clip two types of the feature were extracted: zero-crossing rate and mel-frequency cepstral coefficients. Discarding the 0^{th} coefficient, the first 12 MFCCs were used concatenating with zero crossing rate each clip with their mean and standard deviation across frames. So that the feature dimension for one clip was 26. As another method, Mel filter banks were used with 80 filter banks by using Librosa.

Supervised learning methods applied:

k-NN = k-Nearest Neighbors

SVM = Support Vector Machine

RF = Random Forest

FNN = Feedforward Neural Network

Discussion and Results:

- With ESC-50 dataset and MFCC, zero crossing rate feature extraction as explained in detail in the materials and methods section; following accuracy rates were reached for the respected

algorithms in defaults and 5-fold cross-validated: k-NN = 30 %, Support Vector Machine = 33.8 %, Random Forest = 40.8 %

- With ESC-10 dataset and MFCC, zero crossing rate feature extraction as explained in detail in the materials and methods section; the following accuracy rates were reached for the respected algorithms in defaults and 5-fold cross-validated: k-NN = 64.8 %, Support Vector Machine = 67.5%, Random Forest = 72.3 %

Results were close to that of dataset owners as they did the same process.

- Since each clip is fixed to five seconds and some of them contain silence. This may affect MFCCs as we took its average over time. So as a preprocessing voice activity detection was applied to delete the silence. And by the same feature extraction technique accuracy improved respectfully for ESC-50 dataset: k-NN = 30.4 %, Support Vector Machine = 36.3%, Random Forest = 41.8 %
- As an alternative way of feature extraction Mel filter banks were used by using Librosa with 80 mels concatenated with zero crossing rate like in the MFCC. For ESC-50 dataset following accuracy rates were reached for the respected algorithms in defaults and 5-fold cross-validated: k-NN = 30.4 %, Support Vector Machine = 36.4%, Random Forest = 42%

Mel filter banks performed slightly better than MFCC.

- Feedforward Neural networks were applied using PyTorch. Features were extracted with MFCC and zero-crossing rate as in previous cases. One hidden layer and ReLU activation function applied with 5 -fold cross validation %33 accuracy rate was obtained.

Doubling the hidden layer and learning rate increased the accuracy to %36.6 with the same features.

Dropout methods after each hidden layer gave poor results.

Feedforward with Mel filter bank feature extraction also gave poor results compare to MFCC. (%27 accuracy)

To sum up, Random forest gave the best accuracy. Voice activity detection increased the performance slightly. Feedforward neural network performed close to other algorithms. Since the dataset size is low and we averaged time components of the clips much information was lost. So in the end, the project didn't achieve better results than the baselines (proposed by the dataset owners). Further analysis can be done by CNN (convolutional neural network) as it turns out CNN gives better results than baseline methods which can be checked from the link provided in the appendix with other methods.

Appendix:

GitHub - karolpiczak/ESC-50: ESC-50: Dataset for Environmental Sound Classification

(Dataset owner's GitHub page)

furkanatak/Supervised-Learning (github.com)

(project's code)

Furkan Atak 25.07.2022