PERFORMANCE ANALYSIS OF CLASSIFICATION ON ENVIRONMENTAL SOUND CLASSIFICATION (ESC_50) DATASET

Mie Mie Oo[1], *Lwin Lwin Oo*[2]

[1]*University of Computer Studies, Kyaing Tong, 095, Myanmar*
[2]*University of Computer Studies, Mandalay, 095, Myanmar*

Abstract

The classification of audio dataset is intended to distinguish between the different source of audio such as indoor, outdoor and environmental sounds. The environmental sound classification (ESC-50) dataset is composed with a labeled set of 2000 environmental recordings. The spectral centroid method is applied to extract audio features from ESC-50 dataset with waveform audio file (WAV) format.* The decision tree is easy to implement and fast for fitting and prediction therefore this proposed system is utilized the coarse tree and medium tree as a classifier. *Then fivefold cross-validation is also applied to evaluate the performance of classifier. The proposed system is implemented by using Matlab programming. The classification accuracy of coarse tree is 63.8% whereas the medium tree is 58.6% on ESC-50 dataset.

Keyword: audio, ESC, classification, features

1. INTRODUCTION

The audio classification is useful in various audio processing applications. The relevant tasks in classification of audio signals are source identification, labeling / classification / tagging, music/ speech / environmental sound segmentation and so on [1]. The environmental sound of ESC-50 is consists of various sounds such as dog bark, rain, sea waves, baby cry, clock tick, person sneeze, helicopter, chainsaw, rooster and fire crackling [3]. The feature extraction of audio is important to get the appropriate features for successful classification. The audio feature extraction can perform with various feature extraction techniques and machine learning algorithm. The physical and perceptual features from a sound are extracted and use these features to identify the labeled classes. The decision tree model can be used in both classifications as well as regression problems solving. It is a very powerful to achieve high accuracy in many tasks of classification. The knowledge is learned as training data by decision tree as a hierarchical structure. In this proposed system, the hamming window is used to analyze the frequency content in dataset and also to segment a short time of a longer audio signal. For the audio features extraction, the spectral centroid is applied. These features are the characterization off short-time spectrum. The decision trees such as coarse and medium tree use as classifiers and then analyzed the performance of classification accuracy in order to know which classifier can give better performance on ESC-50 dataset.

2. LITERATURE REVIEW

In [2] the author was evaluated the short audio clips of environmental sounds with convolutional neural network. The max-pooling and 2 fully connected layers were applied in training for audio data. The accuracy analyzed on three datasets (ESC-10, ESC-50 and UrbanSound 8K) of environmental and urban recording. In paper, the fivefold and tenfold cross-validation used as a validation set. The log-scaled mel-spectrograms extracted from all audio recordings with 1024 window size and 60 mel-bands. The first layer, ReLU with 80 filters of rectangular shape (57*6, 1*1) and the second layer, ReLU with 80 filters (1*3, 1*1) was applied in this paper. The model was based on convolutional neural network that have the better performance than the manually-engineered features.

The authors were considered the classification accuracy with Spectrogram, MFCC and CRP image representations methods. For the evaluation of accuracy was implemented by using convolutional deep neural networks (AlexNet and GoogLeNet) over three datasets: ESC-10, ESC-50 and UrbanSound8K. The authors experimented on Ubuntu 14.04 LTS and used Anaconda Python and the deep learning frameworks Caffee, TensorFlow and NVIDIA Deep Learning GPU training system (DIGITS). The GoogLeNet could give the better possible classification accuracy than AlexNet [3].

3. METHODOLOGY

Firstly, the environmental sound classification (ESC-50) dataset with WAV file format is downloaded from the Github. Then read this audio dataset and calculate the centroid of the power spectrum for audio frames over time for 50 ms hamming windows of data with 25 ms overlap. After getting the audio features train these with coarse tree model for classification by using fivefold cross validation. Finally the classification of these two decision tree models is analyzed. The proposed system design is described in the following figure.

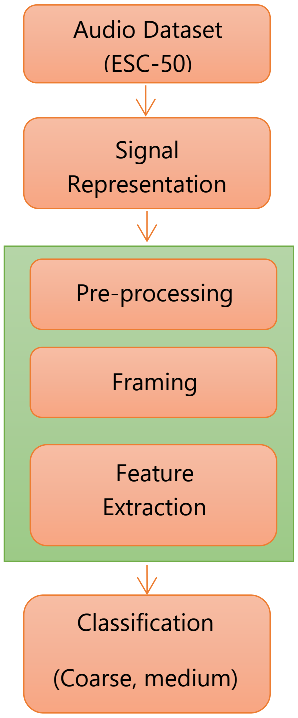

Figure 1. System Design of Proposed System

3.1. Environmental Sound Classification (ESC-50) Dataset

The environmental sound classification (ESC-50) dataset is a collection of recordings 5 seconds environmental sound with the frequency range of 44.1 kHz. The ESC-50 dataset has composed with 2000 labeled set that is 50 classes and 40 clips per class. There are included five categories of sound in this ESC-50 dataset such as animal sound, indoor/family sound, outdoor/urban noise sound, natural sound and human sound.

3.2. ESC-50 Dataset Preprocessing

The windowing method is used to reduce the discontinuities at the boundaries of each finite audio sequence. The hamming window could be satisfied the good frequency resolution and reduced spectral leakage.

A Hamming Window

The windowing is the mathematical function of zero-valued outside of some interval [wiki window function]. It is used to view a short time segment of a signal and then analyze the content of frequency range. The hamming window was proposed by Richard W. Hamming and it could smooth the truncated autocovariance function in time domain. The equation of hamming window is

$$w(n)= 0.54 - 0.46 \cos \left(\frac{2\pi n}{N\text{-}1}\right)$$

B. Spectral Centroid

The spectral centroid is used to measure the characterization of a spectrum and it is shown that the center of mass of the spectrum. The calculation of spectral centroid is as the weighted mean of the frequencies range in the audio signal and it is determind by using a fourier transform. The spectral centroid is the good predictor of the brightness of a sound and mostly used In digital audio and music processing. Each frame of magnitude spectrum is normaized as a distribution over frequency bins from the mean (centroid) is extracted per frame. The spectral centroid of the audio signal can be calculated by using the following formula:

$$\text{centroid}= \frac{\sum_{k=b_1}^{b_2} f_k S_k}{\sum_{k=b_1}^{b_2} S_k}$$

where, fk is the frequency in Hz corresponding to bin k and sk is the spectral value at bin k. b1 and b2 are the band edges[4].

3.3. Decision Trees: Coarse Tree

The coarse tree has a few large leaves for a response function that is the maximum number of splitting is 4 and the minimum leaf size is 36. The optimizaton of coarse tree is based on the Gini's diversity index. The Gini index is the measurement of degree or probability in wrong classification. The value of Gini index is varied between 0 and 1, the value 0 is denoted that all elements are belonging to the certain class or if there is in only one class. The Gini index 1 is that the elements are randomly distributed across the various classes. The middle value of Gini (0.5) is equally distributed elements into some classes. The formula for the calculation of Gini index is as follows [5]:

$$\text{Gini}= 1-\sum_{i=1}^{n} \rho_i^{2}$$

Where, ρ_i is the probability of classification to a particular class.

3.3. Decision Trees: Medium Tree

In the medium tree, there have a medium sized leaves for response function with the maximum number of splitting is 20 and the minimum leaf size is 12. The Gini's diversity index is also used for optimization.

4.RESULTS

The classification of environmental sound classification (ESC-50) dataset is performed by using Matlab programming. The spectral centroid feature extraction technique is applied to extract audio features on a frame by frame. The cross-validation used in this classification task is fivefold.

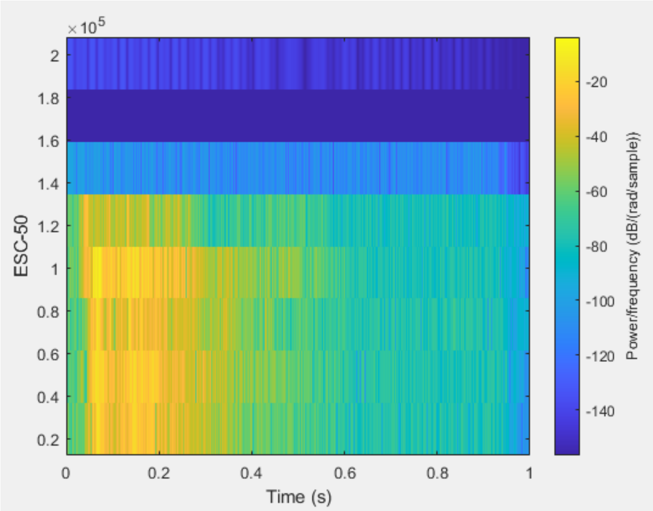

Figure 2. Spectrogram for ESC-50 Dataset

The spectral centroid is the normalization of frequency-weighted sum by the unweighted sum. The spectral centroid features for the ESC-50 dataset is shown in figure 3 [4].

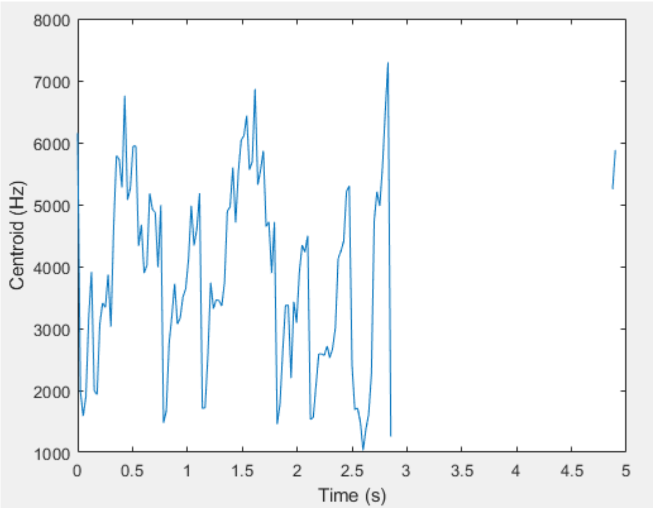

Figure 3. Spectral Centroid Features

There are four possible outcomes in classification: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The true positive rate (TPR) is the proportion of positive condition for positive test result and false negative rate (FNR) is also the false condition for negative test result. The TPR and FNR for model 1; coarse tree classifier is illustrated in figure 4.

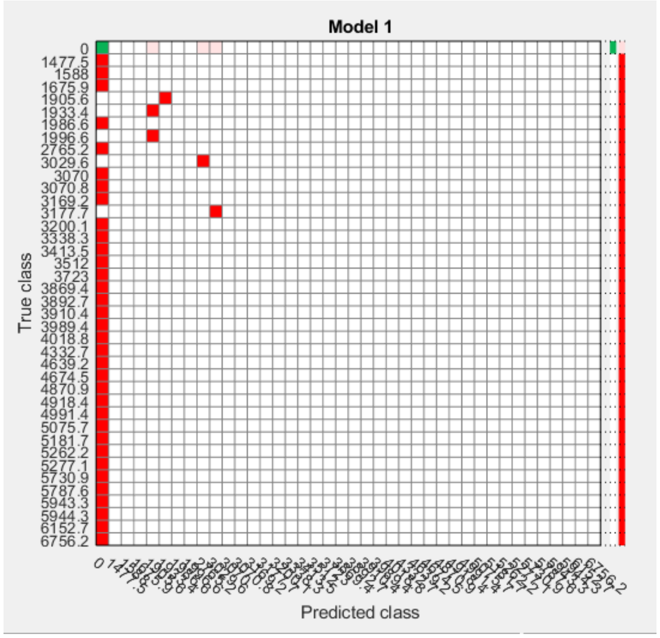

Figure 4. TPR and FNR for Model 1

The medium tree classifier's TPR and FNR is described in the figure 5.

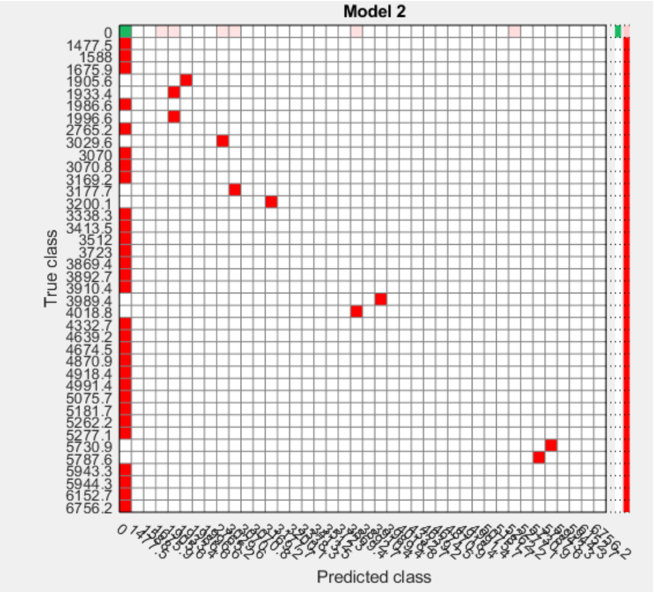

Figure 5. TPR and FNR for Model 2

In the following ROC curve for model 1 (coarse tree) model, the marker value 0.87 indicates the classifier assigns 87% of the observations incorrectly to the positive class and 96% assigns for true positive class or correctly positive class.

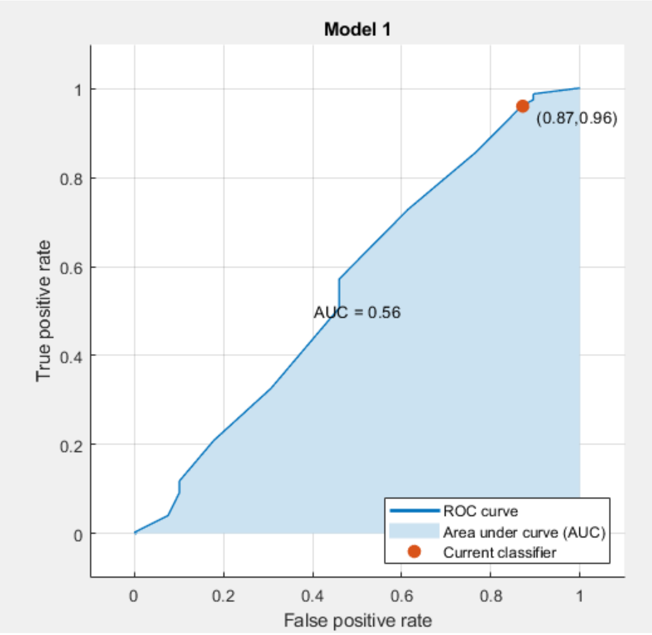

Figure 6. Spectral Centroid Features

Whereas, the model 2 (medium tree model) is assigned 74% of the observations incorrectly to the positive class and 88% assigned for true positive class or correctly positive class.

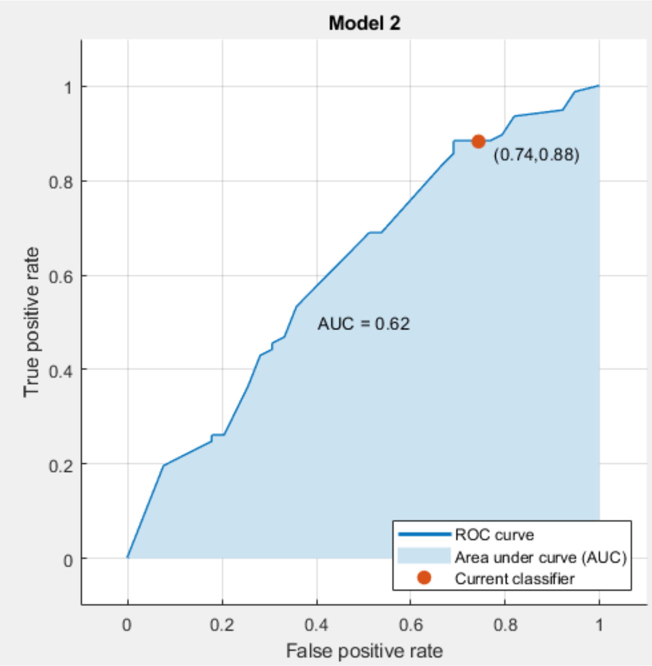

Figure 7. Spectral Centroid Features

5.CONCLUSION

The classification of ESC-50 dataset is implemented and analyzed the performance of two of decision trees classifiers in this paper. The spectral centroid features of ESC-50 audio dataset are trained with classification trees to predict the response of data. The decisions in the decision tree is followed from root node and down to a leaf node. The model validation of medium tree is no better than the coarse tree because the classification accuracy of medium tree is 58.6% whereas 63.8% of coarse tree. Therefore, the medium tree is less accurate than the coarse tree in classification of ESC-50 dataset.

REFERENCES

[1] Juan Pablo Bello, "Sound Classification", EL9173 Selected Topics in Signal Processing: Audio Content Analysis NYU Poly.

[2] Karol J. Piczak, "Environmental Sound Classification with Convolutional Neural Networks",2015 IEEE International Workshop on Machine Learning for Signal Processing, SEPT.17-20, 2015, BOSTON, USA, 2015.

[3] Venkatesh Boddapati, Andrej Petef, Jim Rasmusson et al., "Classifying environmental sounds using image recognition networks", International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille France, 2017.

[4] Spectral Descriptors, http://www.mathworks.com

[5] Gini Index For Decision Trees- QuantInsti's Blog, https://www.blog.quantinsti.com