

IEEE Bigdata Cup 2018 FEMH Challenge Report

1st Soumya Ray

*Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, Ohio, United States
sray@case.edu*

2nd Mingxuan Ju

*Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, Ohio, United States
mxj255@case.edu*

3rd Zhengkai Jiang

*Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, Ohio, United States
zxj89@case.edu*

4th Yufan Chen

*Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, Ohio, United States
yxc775@case.edu*

I. INTRODUCTION

A. Background:

Vocal classification is perceived as one of the challenging tasks in medical field. Many researchers did meaningful work on designing optimal classifiers which can help diagnose vocal diseases from patient's voice record. Some vocal diseases such as neoplasm are notoriously hard to distinguish solely by listening due to noise and subtlety of symptom.

B. Related work:

Vahid Majidnezhad tried artificial neural network with Mel-Frequency-Cepstral-Coefficients as feature vectors to achieve optimal result on vocal pathology classification [5]. P. Kukharchik also used wavelet transform with support vector machine to optimize classifiers' performance [4]. However, there are not any researcher try multiple ensemble-methods which divide mul-class classification into multiple binary classification in vocal pathology classification.

II. METHODOLOGY

A. Data Preprocessing:

- Silence and Noise Removal

Doing a spot check on both training and testing set, we discovered that there exists silence or noise at the beginning of most audio files. We removed those silence or noise parts by calculating the average loudness of each audio file, and recursively comparing the value of 30% of average loudness to first $\frac{3}{4}$ of the audio file.

- Equalizing Loudness:

We also figured out that ,between training set and testing sets, there is a loudness difference not representative of class label. So, we linearly equalized average loudness of each file by the average loudness of all audio files.

B. Feature Extraction:

We utilized general audio features: Zero Crossing Rae, Energy, Entropy of Energy, Spectral Centroid, Spectral Entropy, Spectral Spread, Spectral Entropy, Spectral Flux, Spectral Rolloff, MFCCs, Chroma Vector and Chroma Deviation. The library we used for those feature extractions are pyAudioAnalysis [1].

After analyzing the dataset, we also noticed that number of local minimum amplitude peaks is one representative feature and we added that to feature list.

C. Model Selection:

Our whole model-building infrastructure is based on scikit-learn [2] [3].

We tried K-Nearest Neighbor, Support Vector Machine, Boosting, Random Forest, Extratrees, Multiple Instance Learning, Label Propagation. After few experiments, it seemed like tree algorithms performed poorly on this task and we finalized our focus on SVM, Label Propagation, SVC and MILR with pipeline. This classifier undercalled Normal and Neoplasm patients, which we think is resulted from different class distributions between training and testing set. So, we converted this relatively complex learning task into three less complicated tasks: Normal vs. Pathological, Vocal vs. Rest of Diseases, and Phonotrauma vs. Neoplasm.[Fig. 1.] The reason why we design the pipeline this way is based on the difficulty of classification. (From easiest Normal vs. Pathological to hardest Phonotrauma vs. Neoplasm)

D. Hyper Parameter Tuning:

Since the size of training set is getting smaller and smaller as we propagate through the pipeline, in order to get trustworthy accuracy to tune parameter for SVC, for all three ensembles, we use the same parameter tuned in the first ensemble. The process is pretty straightforward; we simply set up two loops, one of which for C and one of which for gamma. The tuned parameters are 10 for C, and 0.01 for gamma. Also due to the

