

IE5331 Engineering Data Analytics

Final Project,

Due May 04, 2021

Parkinson's disease is a brain disorder that leads to shaking, stiffness, and difficulty with walking, balance, and coordination. Approximately 60,000 Americans are diagnosed with Parkinson's disease each year. More than 10 million people worldwide are living with Parkinson's disease. Parkinson's disease can be diagnosed based on medical history, a review of your signs and symptoms, and a neurological and physical examination. Additionally, speech data are also useful for non-invasive diagnosis. This project aims to analyze speech data from 40 subjects and develop machine learning models to predict Parkinson's disease.

Data: The dataset attached consists of training and testing files. The training data belongs to 20 subjects with Parkinson (6 female, 14 male) and 20 healthy individuals (10 female, 10 male). From all subjects, multiple types of sound recordings (26 voice samples including sustained vowels, numbers, words, and short sentences) are taken. A group of 26 linear and time frequency-based features are extracted from each voice sample. The testing data set is collected from 28 Parkinson's disease patients. The patients are asked to say only the sustained vowels 'a' and 'o' three times respectively which makes a total of 168 recordings. The same 26 features are extracted from voice samples of this dataset.

Training Data File:

Each subject has 26 voice samples including sustained vowels, numbers, words and short sentences. The voice samples in the training data file are given in the following order:

Sample # - corresponding voice samples

1: sustained vowel (aaa), 2: sustained vowel (ooo), 3: sustained vowel (uuu), 4-13: numbers from 1 to 10, 4-17: short sentences, 18-26: words.

Test Data File:

28 PD patients are asked to say only the sustained vowels 'a' and 'o' three times respectively which makes a total of 168 recordings (each subject has 6 voice samples) The voice samples in the test data file are given in the following order:

Sample# - corresponding voice samples

1-3: sustained vowel (aaa), 4-6: sustained vowel (ooo)

Feature Information:

Training Data File:

Column 1: Subject id, Column 2-27: features

Features 1-5: Jitter (local), Jitter (local, absolute), Jitter (rap), Jitter (ppq5), Jitter (ddp),

Features 6-11: Shimmer (local), Shimmer (local, dB), Shimmer (apq3), Shimmer (apq5), Shimmer (apq11), Shimmer (dda),

Features 12-14: AC, NTH, HTN (measures of the ratio of the total noise component in the voice)

features 15-19: Median pitch, Mean pitch, Standard deviation, Minimum pitch, Maximum pitch,

Features 20-23: Number of pulses, Number of periods, Mean period, Standard deviation of period, features

24-26: Fraction of locally unvoiced frames, Number of voice breaks, Degree of voice breaks

Column 28: class information

Test Data File:

Column 1: Subject id, Column 2-27: features

Features 1-5: Jitter (local), Jitter (local, absolute), Jitter (rap), Jitter (ppq5), Jitter (ddp),

Features 6-11: Shimmer (local), Shimmer (local, dB), Shimmer (apq3), Shimmer (apq5), Shimmer (apq11), Shimmer (dda),

Features 12-14: AC, NTH, HTN,

Features 15-19: Median pitch, Mean pitch, Standard deviation, Minimum pitch, Maximum pitch,

Features 20-23: Number of pulses, Number of periods, Mean period, Standard deviation of period,

Features 24-26: Fraction of locally unvoiced frames, Number of voice breaks, Degree of voice breaks

Column 28: class information

Analysis: (use train_data.csv for step 1-3 and test_data.csv for step 4)

1. Leave-One-Subject-Out classification
 - a. From the training dataset, select one subject (26 samples) as the testing set; use the rest subjects' data to train a classification model.
 - b. Use the model to compute predictions for testing set. Take the majority vote of the predictions as the class of the testing subject.
 - c. Repeat step a and b for all subjects; compare the predicted and true classes to evaluate the model performance.
 - d. Repeat step a-c for different tuning parameters to identify the best model.
 - e. Try Logistic Regression, SVM, and Random Forest, and compare their performance.
2. Feature Extraction
 - a. Calculate central tendency and dispersion features:
 - (1) Central tendency features: mean, median, trimmed mean of the 26 voice samples of each subject for different attributes.
 - (2) Dispersion features: Standard deviation, interquartile range, mean absolute deviation of the 26 voice samples of each subject for different attributes.
 - ✓ Trimmed mean is the average of the samples after removing a 25% of the largest and smallest values.
 - ✓ Mean absolute deviation is
$$\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|, \text{ where } \bar{x} \text{ is the mean of } x$$
3. Leave-One-Out classification using the extracted features from 2.
 - a. From the training dataset, select one subject as the testing set; use the rest subjects' data to train a classification model.
 - b. Use the model to compute the prediction for the testing set.
 - c. Repeat step a and b for all subjects; compare the predicted and true classes to evaluate the model performance.
 - d. Repeat step a-c for different tuning parameters to identify the best model.
 - e. Try Logistic Regression, SVM, and Random Forest, and compare their performance.
4. Testing using the independent testing dataset (test_data.csv).
 - a. Use the classification models (Logistic Regression, SVM, and Random Forest) from step 1 and 3 to predict subjects' status in the testing dataset; evaluate the model performance.

Report: Write a final project report using the template provided.

Reference:

Erdogdu Sakar, B., Isenkul, M., Sakar, C.O., Sertbas, A., Gurgun, F., Delil, S., Apaydin, H., Kursun, O., 'Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of Sound Recordings', IEEE Journal of Biomedical and Health Informatics, vol. 17(4), pp. 828-834, 2013