STAT 605 Project

Xiaoyu Liu[1],Hongyi Liu[1], Tianhang Li[1],Xiaofeng Wang[1], Rui Huang[1]

[1]Department of Statistics – University of Wisconsin-Madison

xliu969,hliu557,tli425,xwang2443,rhuang95

1. Introduction

The general purpose of our task is to predict each participants' age group (adult, middle-aged adult, and senior adult) based on the brainwave data received. In this project, we perform an analysis of the data set of EEG signals which contains brainwave data and age of 308 participants. We preprocess the raw data in parallel jobs by using the fast Fourier transform method to convert the time series data into frequency domain data, and then the power spectrum function is further used to obtain the power spectrum components, and the entropy of the δ wave, θ wave, α wave, and β wave are obtained by integral calculation. KNN model and Random Forest model are used for the classification of age group. The accuracy reaches nearly 70%.

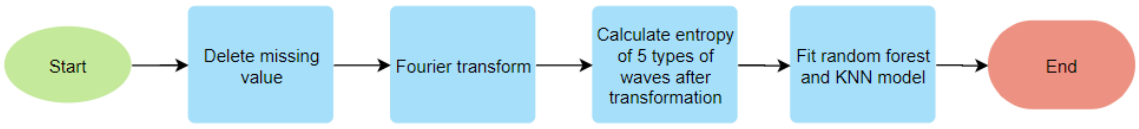


Figure 1. Project flowchart

2. Data

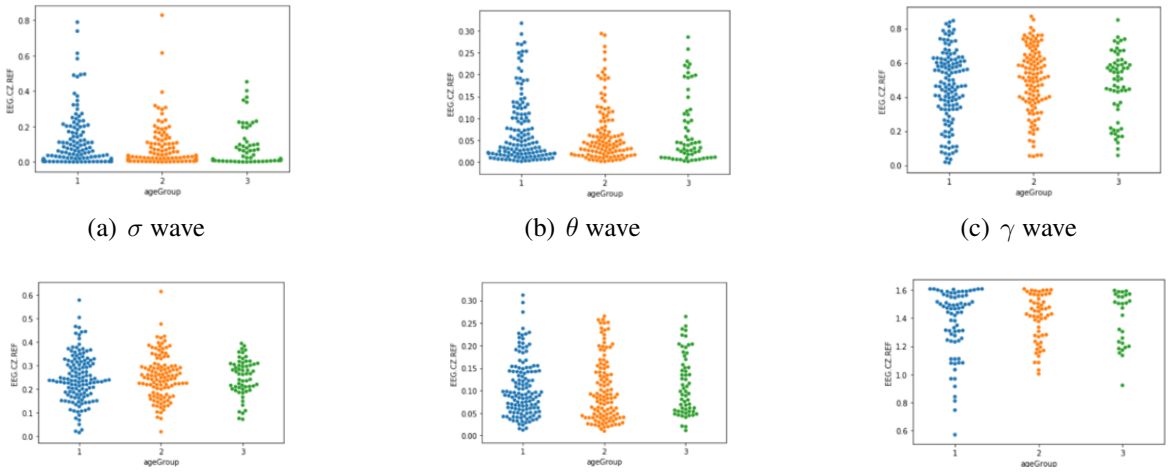
The data we use is EEG for Age Prediction from Kaggle[1]. The data set contains brainwave data and age of 1297 participants and we randomly pick 308 files from the entire data set for our project. The size of each file is between 70MB to 100MB. The first line of each csv file is the age of the participant, the second line is the names of brainwave signals and follows are the data for signals. For every participant, we have 36 kinds of brainwave signals and each signal contains over 300,000 lines of data.

3. Extracting feature

We first use CHTC to do data pre-processing. We treat the processing of each csv file as a job and in each job we do the data cleaning and extract the features in data set. For data cleaning, we firstly read the csv file and save the age value in the first line. Then we delete the last three columns and NA values in the data set because according to research, these three kinds of signals are not useful for our project. To find the feature of each participant, we read several papers about EEG signals. According to Al Zoubi et al.[2], EEG signals could be transformed into five types of waves. On the other hand, percentage of waves or degree of chaos of waves are related with age. Therefore, We use FFT to get the coefficient of discrete Fourier transformationthen we get the power propotion of each wave and use information entropy to describe the degree of chaos.

We classify the ages into three groups, people below 40 years old are classified into adult, those who are between 40 and 60 are classified into middle age adult and those above 60 are regarded as senior adult.

δ(Hz)	θ(Hz)	α(Hz)	β(Hz)	γ(Hz)
1-3	4-7	8-13	14-25	26-30



(a) σ wave (b) θ wave (c) γ wave

(d) α wave (e) β wave (f) entropy

Figure 2. Distribution

After observing the relationship between age and different types of waves and entropy, we think the distribution of entropy, alpha-wave and sigma-wave are more related with age. We will do more research next time to find an accurate type of wave from these five.

However, we can also observe that samples in age group 3 is obviously smaller than samples in group 1 and group 2, which might cause bias in our prediction. We will include more samples to fix this problem next time.

For this time, we decide to use information entropy as the feature for each brainwave signal. We firstly do the discrete Fourier transformation (DFT) to each signal. After the DFT, we transform data from time domain to frequency domain. After that we calculate the power spectral density by Fourier component. After normalizing according to the total spectral power, we then calculate the density of each type of wave in every signal. Then we calculate the information entropy for each wave.

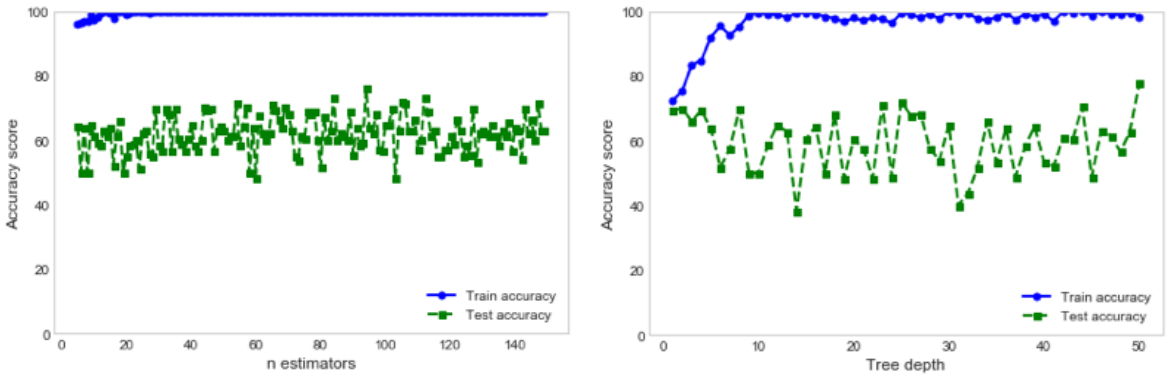
After that, we sum them up to get the total information entropy for each signal. After the transformations above, we get 33 numbers as total information entropy for each participant. Combining them with the age information as mentioned above, each job outputs a csv file as the features of the participant. We write the sub file and the .sh file to make CHTC run. After we get the output files, we write the merge file to combine 308 participants' information as one csv file were the first line contains the first column contains the age and last 33 columns contains the information entropy of each signal.

4. Statistical Models

After pre-processing the original data sets, we combine them into a merged set including each observation's age and the entropy in each of the EEG signal channels. We include 308 observations this time to test parameters in our models.

We try random forest model to predict the age group of a participant this time. To obtain optimal hyperparameter, we use grid search method to calculate accuracy of different combination of hyper-parameters on split sets. Hyperparameters that we are interested in are number of trees, maximum of trees and criterion method of choosing parameters.

According to the plots above, we chose the optimal hyperparameter (shown in following table1). We also notice that predict accuracy of training set is much higher than reasonable value, which indicates that the model might not be suitable for our data. On the other hand, this might also imply that



(a) n_estimator (b) tree depth

Figure 3. Accuracy of random forest model

Model	Hyperparameters	Train Accuracy	Validation Accuracy
Random Forest	Max_depth=6 Number of tree = 60	78.06%	67.02/%
KNN	K=14	52.09%	48.62/%

number of samples we include this time is not larger enough, which causes the problem of overfitting. Therefore, we also try K-Nearest-Neighbors model (denote as KNN) to compare.

We use similar method to obtain optimal hyperparameter in KNN. The hyperparameter we focus on this time is number of k.

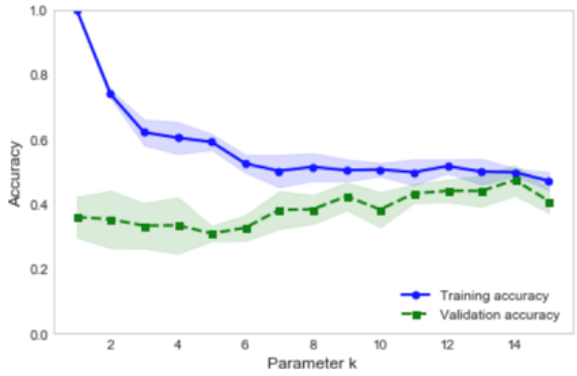


Figure 4. Accuracy of KNN model

As shown in Figure 4, accuracy of training set is decreasing when parameter k is increasing, which is reasonable because when k goes larger, train set is less likely to be overfitted. However, when parameter k is equal to 1, we also obtain 100% of accuracy, which is not reasonable. Therefore, comparing this result with the one we calculate using random forest model, we think the reason for this problem is the small number of samples we include.

5. Review and plan

After revisiting our question of predict age group, we found that our method is not very reliable. As we mentioned above, our data set is not large enough this time and the problem of overfitting is obvious, we will include 800 participants next time to fix this.

Besides that, we have plan for future work related with more methods of extracting feature and fitting models. Except for entropy of brainwaves, we found out that percentage of a certain type

of brainwave is also related with age group. On the other hand, we also notice that Support Vector Machine or XGBoost model might also be better when fitting our data. We will try finding out which type of wave is more related with age and extract percentage of it from original data. We will also try assembling different models to reduce bias and variance.

6. Source

[1] EEG data, Kaggle, https://www.kaggle.com/ayurgo/data-eeg-age-v1

[2] O.Al Zoubi, C. Ki Wong, R. T. Kuplicki, H.-w. Yeh, A. Mayeli, H. Refai, M. Paulus, and J. Bo-durka, "Predicting age from brain eeg signals—a machine learning approach,"Frontiers in Aging-Neuroscience, vol. 10, p. 184, 2018.

[3] S. Kanokwan, W. Pramkamol, K. Wipatcharee, W. Warissara, R. Siwarit, S. Sompiya, B. Onuma, andS. Mitra, "Age-related differences in brain activity during physical and imagined sit-to-stand inhealthy young and older adults,"Journal of Physical Therapy Science, vol. 31, no. 5, pp. 440–448,2019.