Group 6: Age Prediction via Electroencephalogram Data

Zihang Wang zwang2547
Jie Sheng jsheng27
Haoran Lu hlu226
Luyang Fang lfang45
Xinran Miao xmiao27

November, 2020

1 Data Set

Name: Electroencephalogram (EEG) Data [1].

URL: https://www.kaggle.com/ayurgo/data-eeg-age-v1

Background:

An electroencephalogram (EEG) is a test used to evaluate the electrical activity in the brain. In short, it attaches many small electrode sensors to different areas of the scalp and each sensor could record brain wave signals. Data from each sensor is called a channel and it looks like wavy lines with peaks and valleys. Since EEG measures activity of brains, it could provide information about a person's brain condition. In practice, this information could be used to diagnose some brain diseases, while our goal is to predict people's age base on their EEG records.

Dataset & Variable Description:

The data set we've chosen consist of 1297 patients' ages and their EEG recordings in .csv format. There is one response variable, age, and there are 32 variables (EEG *-REF) corresponding to 32 channels of EEG data. There are 4 extra variables (PHOTIC-REF, IBI, BURSTS and SUPPR) corresponding to some certain clinical meanings. Since we are mainly interested in infomation within the barin activity patterns, we will only use the response variable and 32 variables of channels and omit the 4 extra variables.

2 Statistical Questions

Question 1: Data mining & dimension reduction

Note that for each patient, there are 32 channels of recordings with a substantial amount of observations (time points). Since having 32 channels is because the EEG measurements use 32 sensors, the 'true dimension' of underlying signals is generally less than 32. Also, we are more interested in the underlying brain activity patterns instead of signals recorded directly by sensors. Thus, it's desired to separate meaningful or useful signals (also called components) from the original EEG recordings using data mining or dimension reduction methods.

Question 2: Prediction

Given results from question 1, it's desired to predict ages from these processed signals for patients. This will give our project a practical usage. On the other hand, since question 1 is an unsupervised task and question 2 is a supervised task by doing prediction, we can evaluate the methods for question 1 by checking prediction performances in question 2.

3 Code

After installing the Kaggle API, we read the file names of the EGG dataset to a .txt file on the command line. Then we download the first one as a pilot run.

```
kaggle datasets files ayurgo/data-eeg-age-v1 | cut -d’ ’ -f1 | tail -n +4 | sed ’$d’ > filenames.txt
test=$(head -n1 filenames.txt)
kaggle datasets download ayurgo/data-eeg-age-v1 -f $test
```

After that, we can use R to read the data from our laptop.

```
library(data.table)
file = list.files(’.’,’*.csv’)
dat1 = readLines(file[1])
#read age data
(age = dat1[1])
#read EEG data
dat = fread(file[1])
dat[1:10,]
```

4 Statistical Methods

For question 1, Independent component analysis (ICA) is our first choice, since it is widely used in processing neuroinformatics data like EEG. ICA is suitable for computing in parallel because it requires to process data of each patient separately and we have data of 1297 patients. Other candidates are principle component analysis (PCA) and factor analysis (FA).

For question 2, we are considering some time series models and machine learning methods for prediction.

5 Computational Tools

For both questions, we plan to use R. We plan to use the following R packages: dplyr and stringr for data pre-processing, fastICA for ICA. We’ll use R in CHTC to pre-process the data and extract useful features in parallel.

References

[1] Alexander Valkovich ayurgo, Enthusiast2020. Eeg for age prediction. https://www.kaggle.com/ayurgo/data-eeg-age-v1, 2020.