Group 6: Age Prediction via Electroencephalogram Data

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

December, 2020

1 Introduction

An electroencephalogram (EEG) records individual's brain wave signals of difference areas via small electrode sensors in a short period. The recordings of each sensor is called a channel and it looks like wavy lines with peaks and valleys over time. EEG provides information about one's brain condition, which can be used to diagnose brain diseases and make other inference. In this work, we intend to extract structure information of human brain first and try to predict ages based on the extracted signals.

Figure 1 shows the whole pipeline of our analysis. We used Independent Component Analysis (ICA) to reduce the dimension on each patient's EEG data and finished this step in parallel on CHTC. After extracting the mixing matrices which contain the information of brain spatial patterns, we built statistical predictive models such as random forest and k-Nearest Neighbors (kNN) to predict patients' age.

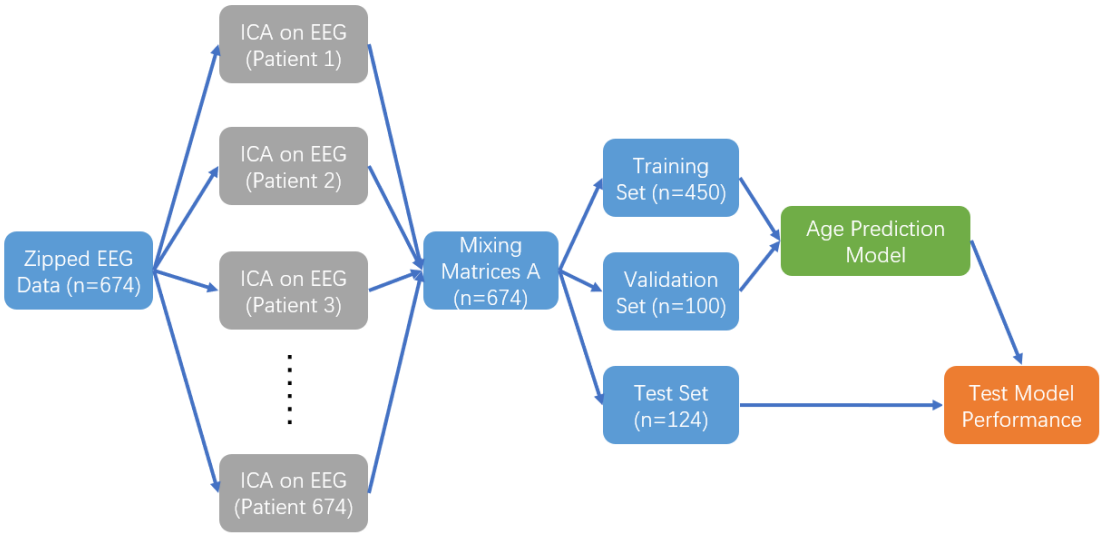

Figure 1: The whole pipeline of our analysis on EEG data.

2 Data Description & Data Cleaning

The EEG data set is from Kaggle [1]. It contains training and test sets with about 1000 patients' recordings on 30 to 34 channels in total, each of which includes signals at more than 300,000 time points. We used the

recordings of 24 channels in common and removed the rest. Figure 2 shows EEG recordings of the first 9 channels for one individual. Some panels look alike, which confirms the necessity and rationale of reducing the dimension to remove redundant information.

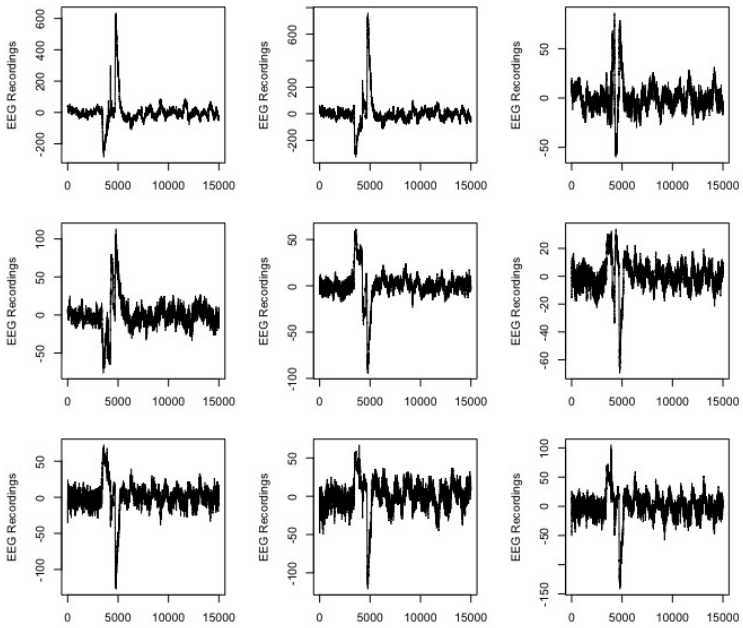

Figure 2: EEG recordings of the first 9 channels of one individual.

3 Statistical Models

3.1 Independent component analysis

Independent component analysis (ICA) is a computational method to find a linear representation of non-Gaussian data so that the components are statistically independent, or as independent as possible. The resulting representation is supposed to capture the essential structure of the data in many applications, including feature extraction and signal separation [2]. Performing ICA on data matrix $\mathbf{X}_{(p\times n)}$, we can get:

$$\mathbf{X}_{p\times n} = \mathbf{S}_{p\times q}\mathbf{A}_{q\times n}$$

where p is the number of time points, n is the number of channels, and q is the number of components we desired to get. $\mathbf{A}$ is the mixing matrix which stores the information related to some underlying spatial patterns of the brain . $\mathbf{S}$ contains the signals we extracted from the original data set.

Our goal is to find a proper way of decomposition such that the signals we extracted are as independent as possible, which can be realized by maximizing the **non-Gaussianity** of those signals [2]. We used **fastICA algorithm** to maximize negative entropy to achieve the maximum non-Gaussianity. And the spatial information included in matrix $\hat{\mathbf{A}}$ could be potentially used to predict a person's age.

3.2 k-Nearest Neighbors Model

3.2.1 Reasons for using mixing matrix A

Next, we only use the mixing matrices $\hat{\mathbf{A}}_i$ form ICA results for age prediction. There were two reasons to use mixing matrices $\hat{\mathbf{A}}_i$ and drop the signal matrices $\hat{\mathbf{S}}_i$. Firstly, rows of $\hat{\mathbf{A}}_i$ contained information related to underlying spatial patterns of the brain, suggested by [3] [4]. From our perspective, we considered that the spatial patterns might be changed through ages, while for $\hat{\mathbf{S}}_i$ we thought it might be hard to extract information about ages. Next, we used kNN on $\hat{\mathbf{A}}_i$ for prediction, but first we need a measure of similarity for $\hat{\mathbf{A}}_i$. Secondly, as we wanted to reduce dimensions for further kNN methods, $\hat{\mathbf{A}}_i$ had a great reduction of dimensions while $\hat{\mathbf{S}}_i$ kept the very large time dimension ($p > 300{,}000$). Therefore, we decided to only include $\hat{\mathbf{A}}_i$ in our following procedure and check whether it could work well in age prediction.

3.2.2 Similarity Measurement and kNN Algorithm

Next step is to build a kNN model for age prediction. In preparation of kNN, we defined a similarity between two people's $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{A}}_2$.

By the property of ICA, denote the j-th row of $\hat{\mathbf{A}}_i$ as $\mathbf{a}_{ij}$. Orders of $\{\mathbf{a}_{ij}\}_j$ in $\hat{\mathbf{A}}_i$ could be of random permutation. So we first aligned $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{A}}_2$ to let their rows match with each other. We simply found m pairs of $\mathbf{a}_{1j}$ and $\mathbf{a}_{2k}$ that have the largest absolute value of correlations and denoted them as $\{\mathbf{a}_{11}, \mathbf{a}_{12}, ..., \mathbf{a}_{1m}\}$ and $\{\mathbf{a}_{21}, \mathbf{a}_{22}, ..., \mathbf{a}_{2m}\}$. Then we defined our aligned similarity of $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{A}}_2$ as (1):

$$\sum_{j=1}^{m} |\text{cor}(\mathbf{a}_{1j}, \mathbf{a}_{2j})| \tag{1}$$

where m is a hyper-parameter regrading how many pairs' correlation we want to involve. Next, we used this similarity to build a kNN model with which to predict age. Note that to use this similarity, we made the assumption that different people's $\hat{\mathbf{A}}_i$ at least has m similar rows. This assumption is not a rigorously defined here, a rigorous assumption of $\hat{\mathbf{A}}_i$ could be found in [3] which is suitable for our method. The two hyper-parameters k in kNN and m in (1) were selected by cross-validation, which is detailed described in Section 5.

3.3 Random Forest

We also tried the random forest method, which is a useful method for almost all type of data. We first aligned spatial patterns corresponding to the same signals in each sample, then used these spatial pattern as variables and age as response to build a random forest model.

Suppose we only want to find one spatial pattern s_1 in every sample, the problem here is that the orders in rows of mixing matrix A were random. We couldn't determine which row is the correct signal we were looking for according to merely the row orders. So we solved this problem by finding a standard reference pattern first. Specifically, we randomly chose 100 samples and used their all q spatial patterns to do cluster analysis. In this way, we obtained the center of clustering and set this vector as the standard reference vector, which represents the pattern of s_1. Under the assumption that the similarity between two spatial patterns corresponding to the same signal was stronger than that between two spatial patterns corresponding to different signals, we could say the row in A that has the largest correlation with this standard vector is

the one corresponding to s_1. Thus, we got one 1×24 vector for each sample, which was used as predictors in the following random forest model. Here we use 10 fold cross-validation to evaluated the performance based on the mean absolute error, and set the number of trees from 40 to 280 to find the optimal tree number.

But the performance of this method is not ideal. The predictions for most samples were ranged from 40 to 55. Thus, although it's mean absolute error is acceptable, we still viewed this as an ill-behaved model for our problem. So we chose to use kNN as out final model, on which and our following analysis will focus.

4 Computations

4.1 Parallel jobs for ICA

Since the independent components could be disintegrated independently on each person's EEG data set, we employed CHTC to run implement ICA in parallel. The method we used are fastICA and the number of components we decided to extract is 20. It took us less than one hour to finishing this computing step. Then the matrices A of all patients were collected.

Specifically, We downloaded the zipped whole EEG dataset of 674 .csv files via Kaggle API to a shared folder on CHTC. Then we extracted about 200 files to our own folders on CHTC server at each time and ran 200 jobs in parallel for 4 times. Each job took less than 5 minutes. We didn't use some alternatives for these two steps because of the following two reasons.

1. For the data downloading part, Kaggle API allowed users to either download a single file or the whole zipped dataset at one time. When we implemented the former, however, some .csv files couldn't be downloaded alone because of "404 not found" errors, while they did get included if we downloaded the whole zipped file. That was why we chose the latter - downloading the whole dataset to a shared folder on CHTC.

2. For the code running part, it's more time-efficient to run all of the 674 parallel jobs at one time compared with what we did. To run 674 jobs, however, we would need to get all .csv files unzipped from the shared folder, which would exceed either the quota limit of the shared CHTC folder or the storage limit of our CHTC location. Thus, we unzipped 200 files to our won CHTC directories at one time for each individual.

4.2 Implement kNN Algorithm

After the dimension was dramatically reduced, following steps could be done on one computer. We separated 674 individuals' ICA processed data into three subsets: 450 for training, 100 for validation and 124 for testing. We wrote our own R codes for aligned similarity and kNN. The two hyper-parameters k and m were selected by cross-validation with mean absolute error. Finally, for each patient, we got his or her age and corresponding predictors, from which we built a kNN model to predict ages on the test set.

5 Results and Discussion

5.1 Results of ICA

Figure 3 and 4 shows four signals of the resulting A matrices for two patients, where each sub panel indicates one extracted signal. The aligning sub panels look similar between patients, which indicates a success of ICA on extracting information.

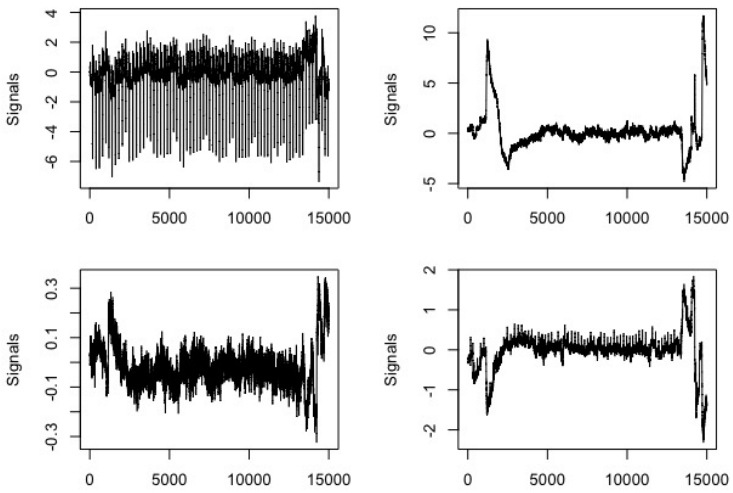

Figure 3: Example one of EEG signals after independent component analysis

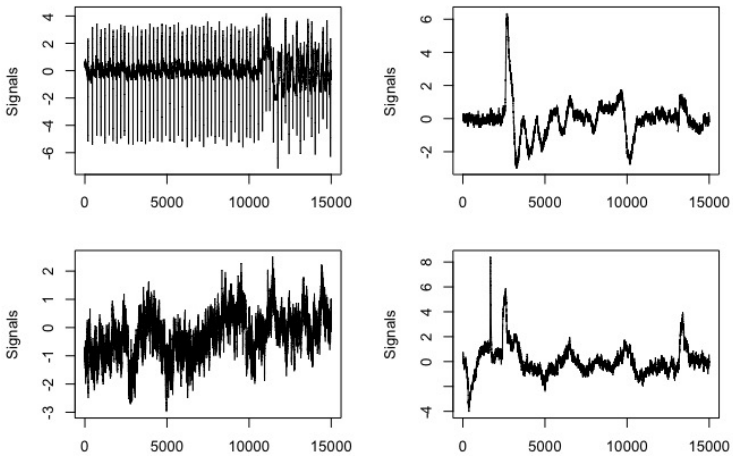

Figure 4: Example two of EEG signals after independent component analysis

5.2 Results of kNN model

We selected hyper-parameters by cross validations evaluated by the mean absolute error. Table 1 shows the result of different settings of k in kNN and m in the aligned similarity. As we can see, $(m, k) = (3, 5)$ gives

the smallest mean absolute error 14.22.

	k=1	k=3	k=5
m=1	18.12	14.67	14.61
m=3	18.03	14.45	14.22
m=5	17.4	15.63	14.88

Table 1: Mean Absolute Errors (k is the parameter of kNN, m is from the aligned similarity)

We chose this setting $(m, k) = (3, 5)$ and made predictions on the test set, with a mean absolute 14.76 on the test set. The result could be interpreted as: our age prediction procedure had an average bias of 14.76 years. We considered this performance satisfying, and it implied the spatial patterns in the mixing matrix A did including useful information about a person's age. The result of prediction performances also verified that our first step, unsupervised ICA, were reliable.

5.3 Discussion & Future Work

We implemented ICA followed by aligned similarity and kNN to predict ages based on EEG signals. The dimension reduction part (ICA) successfully reduced the dimension and extracted information, while the prediction (kNN) outperformed many other methods. In this way, though, we didn't make full use of the whole data set, which can be further explored by the following future directions.

1. For dimension reduction, we avoided using matrix S from ICA. Further methods may be applied to extract information in S.

2. Instead of reducing the dimension before prediction, a direct and computationally heavy prediction method, like deep neural network, can be used.

References

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

[2] Aapo Hyvärinen and Oja Erkki. Independent component analysis: Algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[3] Aapo Hyvärinen. Testing the ica mixing matrix based on inter-subject or inter-session consistency. *NeuroImage*, 58(1):122–136, 2011.

[4] Aapo Hyvärinen and Pavan Ramkumar. Testing independent component patterns by inter-subject or inter-session consistency. *Frontiers in Human Neuroscience*, 7:94, 2013.