Application of Random Forest to classify EEG data of mTBI patients and control adults obtained during a Visuospatial Working Memory Task

Cruz, W.,[1] Cavanagh, J.F.,[2] Lin, C.Y.[1]

[1]National ChengKung University, NCKU [2]University of New Mexico, UNM

Introduction

The combination of electroencephalographic (EEG) recording and cognitive experimental tasks provides an excellent tool for studying human neural dynamics. EEG provides high temporal resolution time series data sampled across multiple scalp locations that produces large amounts of data, and some of these datasets are freely available in repositories on the internet (Fig. 1).

Figure 1. Open-science neuroinformatics database repository.

The analysis of EEG data requires the implementation of many signal extraction methods that are further used to characterize the psycho physiological phenomena been studied. Sometimes the large amount of data collected for one experiment is not fully used because most of the researchers prefer to take a confirmatory and deductive approach, thus focusing their attention in a rather narrow number of features but unintentionally overlooking other important latent features. The utility of the random forest algorithm was investigated for extracting the most relevant features from an EEG data set obtained from mild Traumatic Brain Injury (mTBI) patients and Healthy Controls (HC) during a Visuospatial Working Memory (VSWM) Task (?).

Random forest is a supervised ML method used for classification purposes; the algorithm generates a set number of decision trees, each of which is made based on different subsets of data extracted from the training set. These subsets are selected following a random sampling approach; this iterative process and the number of decision trees computed are further used to reach a classification consensus, and the most common output is selected as the most relevant model which is usually the one that contains the most relevant features for the purpose of classifying the cases or instances considered (?). Therefore it is possible to classify EEG signal recording using this data analysis framework to characterize the neural dynamics and predict the performance and diagnosis (?).

Advantages of Analyzing EEG data with Random Forest

Machine Learning techniques include different computational frameworks such as SVM, Logistic Regression and Random Forest, that are capable of mining large datasets, some advantages of using these frameworks and in particular Random Forest when analyzing EEG datasets are:

- **Identify** relevant questions concerning EEG data.
- **Discover** new knowledge through pattern recognition and mathematical modeling.
- **Applicability** in both medical and social sciences' datasets.
- **Predict** the performance in a task.
- **Diagnose** patients based on their neural activity patterns.

Methods

Subjects

An EEG data set obtained during a Visuospatial Working Memory (VSWM) task was downloaded from OpenNeuro (?); the subjects that conformed the final groups for the present analysis, mild Traumatic Brain Injury patients (mTBI, *n* = 27) and Healthy Controls (HC, *n* =27), were matched using demographic variables and their scores in the task, thus ensuring that both groups did not differ significantly by age ($p = 0.67$), sex ($p = 0.58$), nor by hit ratio ($p= 0.97$).

Procedure

The VSWM task (see Figure 2) was deployed using MATLAB, participants had to perform a yes-no recognition task and were asked to respond whether a location defined by a square containing a question mark (i.e. probe) had been occupied by a red dot in the preceding visual array. Participants were told to ignore the yellow dots. Altogether there were three conditions, either showing three targets or three red dots (*Condition 3*), showing three targets and two distractors (*Condition 3 + 2*), or showing 5 targets (*Condition 5*).

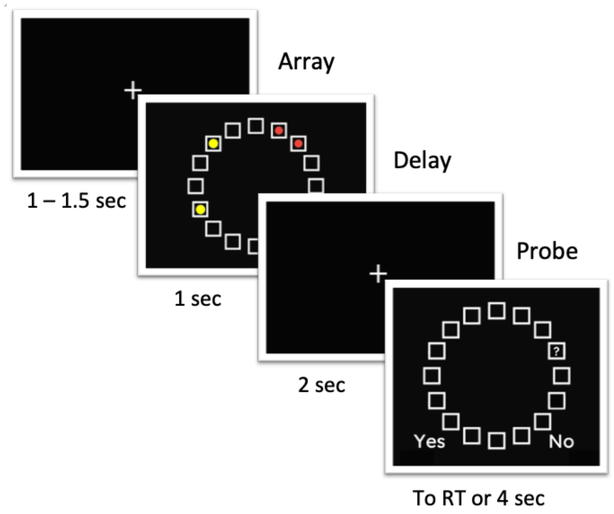

Figure 2. Visuospatial Working Memory (VSWM) task.

EEG Data

EEG activity was recording using a 64 channel cap and the electrodes were located according to the 10-20 system. Continuous EEG was monitored and reference to the *Fz* channel and acquired at a 1000 Hz sampling rate. Total EEG set up time was in average 30 minutes, and the VSWM task included a practice block and an experimental block with 150 trials. The data of each subject that comprised the final sample was first loaded into MATLAB using EEGLab (?), then the EKG and VEOG were removed leaving only the channels that recorded brain activity.

Following, each subject's data was decomposed by ICA using the picard algorithm (?) and the artifact components were removed with the SASICA module (?). Lastly, the data was epoched including a period from -4000 to 0 ms (i.e. 0 was the time in which the probe appeared on the screen), thus covering three memory phases (i.e. Baseline, Encoding, Retention). This data was analyzed to extract 5 frequency components from each of the scalp sites. The EEG data was labelled by group and separated as either correct or incorrect for classification purposes independently of the condition, and was exported into plain format for further processing with R (?). The Random Forest Algorithm was trained with 60% of the data to build EEG classifiers of VSWM trial accuracy and diagnosis.

Results

The first model classified Healthy Controls trials with an accuracy of 86%; occipital α at encoding provided one of the highest importance values as well as parietal θ at baseline and parietal δ at encoding. The second model correctly had an accuracy of 78% when cross-validated in mTBI data and included θ and β bands of several channels.

Model 3 identified central-parietal β at retention and posterior occipital γ and β at encoding as primary classifiers of diagnosis, providing a 98% classification accuracy. Model 4 using the incorrect trials only, identified Central γ at retention and baseline as primary classifiers of group belongingness. Table 1 presents a summary of the models along with the some of the variables with the larger importance in the construction of EEG classifiers for the given data set, the larger the value the more significant is for the outcome. For a complete list of the variables that resulted for each model along with the R script visit my Github page by scanning the QR code.

Table 1. Summary of Random Forest Models with some of the variables with larger importance values

Channel	Location	Stage	Frequency	Importance
Model 1 - *Healthy controls Only* - Accuracy 0.859				
P4	Parietal	Baseline	θ	2.02
P4	Parietal	Encode	δ	1.35
O2	Occipital	Encoding	α	1.13
Model 2 - *mTBI Only* - Accuracy 0.785				
CP2	Central-Parietal	Retention	θ	2.32
TP10	Temporo-Parietal	Retention	θ	2.31
TP10	Temporo-Parietal	Encoding	α	1.72
Model 3 - *Correct trials Only*- Accuracy 0.982				
CP6	Central-Parietal	Retention	β	5.24
POz	Posterior-Occipital	Encoding	γ	4.75
POz	Posterior-Occipital	Encoding	β	4.66
Model 4 - *Incorrect trials Only* - Accuracy 0.947				
Cz	Central	Retention	γ	17.4
Cz	Central	Baseline	γ	16.98
CP6	Central-Parietal	Retention	β	15.16

Discussion

These analyses indicate that the performance in the VSMW task as well as diagnosis of the subjects based on a trial based classification approach using the Random Forest is useful for classification purposes given the high accuracy of the models; in addition, the resulting set of scalp sites and band frequencies coincide with previous findings in the working memory literature.

For example, some experiments suggest that α varies as a function of memory performance, being smaller during the encoding (?). In addition, θ band activity has been associated with higher memory load (?). γ band has been associated with the integration of multi-modal sensory processes in memory tasks, and as θ, it increases with higher cognitive load. Importantly γ has been found to support short-term maintenance of information and indicates the recruitment of cognitive resources to match the demands of the cognitive task (?).