Methodology Overview:

The proposed method to cluster channels comprises of three stages. In the first stage, rolling window FFT (Fast Fourier Transform) spectral analysis has been used to extract important frequency features from EEG signal, and after this stage, those features were then combined for each athelets. In the second stage, auto-Kmeans clustering has been used to cluster channels on each individual subject using the features produced by first stage, and a [cluster ring] was produced from the auto kmeans. In the third stage, Jaccard Distance measure has been used to determine the distance between two channels across the whole athletes.

Rolling window fft spectral analysis:

1. Data acquisition:
2. Windowlization
3. fft
4. standardization

As a feature extraction process, we have used rolling window fft analysis method to transform EEG signals from time domain to frequency domain.

Data Acquisition:

Our data came from 3 sources, EDF, incomplete EEG and complete EEG files. Complete EEG files come with 3 files for each record, EEG, vhdr file, ahdr file. Incomplete EEG files were transformed manually using wineeg software to ASCII text files. EDF files were transformed to ASCII text files as well.

All data from the three sources were then inserted into mysql database. The table schema is: table name = filename, table content= time, power, channel, primary key = time, channel. Time is not the real time, but 1 data. Channel is one of the sensors we put on athletes’ head when we collected the EEG.

Windowlization and fft:

We split the data into overlapping trunks on the time plane, using 60 as interval and 10 as break points. Therefore e.g. .if we have 1-120 sample points, then we have 0-60, 10-70,20-80,30-90,40-100,50-110,60-120 windows. Then we do FFT transformation on each of the window to find characteristics in the frequency domain, and combining them to get charactertics along the whole period for one channel.

Standardization:

Then we standardize each window across all channels. E.g. the window 1 at channel 1 is standardized with window1 at channel1,2,3,4,5,6 .. channel\_max, window2 at channel 1 is standardized with window1 at channel1,2,3,4,5…. Channel\_max etc.

From the above process, the data we got is dictionary d, for all athelets. D[atheletFileName][channelNum][startTime] will give us 1 particular fft transformed, standardized window array for that atheletFileName. This dictionary d then was passed to intra athlete clustering step.

Intra Athlete clustering on channels:

Clustering alforithms are used widely to collectsiimlar or dissimilar data. The most used clustering algorithms are k-means clustering, and we built on that to automate the k selection process.

Auto kmeans & sscluster threshold

After we obtain the dictionary d for all athelets, we combined all the windows together so then D[atheletFileName][channelNum] is going to return 1 array for that channel. This array is pretty large as it has all the information for that channel across the whole records.

Then we apply k-means, from 1 then up. After each iteration, we check if the information gained is more than 10% of what it has changed from the last iteration. If it is, then we continue grow on K. If it did not get us the extra 10%, then we quit the kmeans and say the last K is the best.

To check the information gained, we used sscluster threshold. The ssCluster is give by: pF = (distortion/(clusterNumber-1))/((ssBetween)/(n-clusterNumber))

Distortion is the distortion of each data points from centroids, ssBetween is the distance between each centroids, n is the data length.

The output of this step is a dictionary of athelet names with their clustering Rings. The clustering is something like: If we have channels 1-9, then the clustering for 1 athelet would be e,g, [[1,2,3],[4,5,6],[7,8,9]].

Inter Athelet Clustering on channels:

We used Jaccard distance to get the distance measure between all pairs of channels, and then we use this distance matrix to generate the dendrograms for all channels.

We also have different interval/space settings, namely not using 60 as the interval and 10 as the space at the windowlization step, and follow alone the path to here. The difference is minor, which indicates setting different interval/space would not interfere with our findings.

[new]

Inter Athelet clustering on athelets:

This step we use Mutual information gain for calculating how similar different athelets are, and then we generate dendrograms for those athelets, hoping to find out same type of athelet got grouped together. The data we use is after the Intra\_athelet channel clustering, namely, using the Cluster Ring.