DonsApps
STUDENT/ALUMNI

Sunny Zeng <szeng8@dons.usfca.edu>

[Reminder] EEG data needed and following meet-up schedules

William Bosl <wjbosl@usfca.edu> Sat, Jul 16, 2016 at 10:32 AM
To: Ellen Lin <rlin11@dons.usfca.edu>
Cc: Sunny Zeng <szeng8@dons.usfca.edu>

Hi Ellen and Sunny,

I am going to send two files to you using wetransfer.com, since the files are too big to send by email. I will also put them in our google drive folder later. The files are in Excel format. You can open in Excel and save as either .csv or .txt (tab-separated), as you wish.

The first file has the raw computed EEG features. Some of the data is clearly corrupted (zeroes for Sample entropy is the clue). You are welcome to play with this. Later today I will process this and create a clean data set with imputed data filling in for the corrupt data.

The data consists of subject IDs by row. Each subject has one or more measurements at the ages given in months. These are:
3, 6, 9, 12, 18, 24, and 36 months. Very few subjects have all measurements. Most have at least 3. The challenge is to build trajectories using this data.

There are 3 classes represented here:
asd - confirmed diagnosis of autism spectrum disorder after 3 years
typ - confirmed to be typically developing after 3 years (that is, no autism)

You may want to just work with these two classes for now.

The final class, labeled hra for “high risk autism” are children who have an older sibling with autism, but were determined at 3 years to not have autism. Nevertheless, they may exhibit some autistic symptoms. These group is expected to be more heterogenous. It’s not known if they’ll follow a third trajectory or fall into one of the other two (more likely).

I’ve also included gender. It’s not known if that can be used for distinct subgroups or not.

If you are able to build trajectories for this data, and we can distinctly separate the groups, this will be a highly visible and publishable result. I will help to write it up and find the right journal to submit.

Finally, the features are a little complicated and are made up of 3 dimensions:

sensor or channel (19): channels = ["C3", "C4", "O1", "O2", "Cz", "F3", "F4", "F7", "F8", "Fz", "Fp1", "Fp2", "P3", "P4", "Pz", "T7", "T8", "P7", "P8"]

features (9): ['Power', 'SampE', 'RR', 'DET', 'LAM', 'L_entr', 'L_max', 'L_mean', 'TT']

Frequency band or scale (6): [0,1,2,3,4,5]

Each feature is named with each of these dimensions, separated by a colon, like
F7:RR:s2
for channel Fp1 (left frontal), nonlinear signal value RR (recurrence rate), and scale 2 (it’s a wavelet scale).
Thus, there are 19*9*6 = 1026 total features.

In my work so far, it appears that the most informative nonlinear values are RR, SampE, and Power; all channels; scales 0 to 2.

Let me know when you have questions!

Bill

On Jul 13, 2016, at 11:04 PM, Ellen Lin <rlin11@dons.usfca.edu> wrote:

[Quoted text hidden]