Departamento de Electrónica, Telecomunicações e Informática

Exploração de Dados

Data Mining Assignment

Various new paradigms—such as neuroscience, artificial intelligence, cognitive science, and brain-computer interface (BCI) have been developed to understand the brain in more depth [2]. For that controlled experiments are conducted and brain signals (Electroencephalogram, for instance) are recorded while the participants are doing different tasks. Two data sets with features considered relevant in brain cognitive studies are available. And the goal of this assignement is perform the data analysis using supervised machine learning algorithms. With that purpose each group (with two students) must

- Select one data set. And define a strategy for the data analysis.
- Submit in the elearning one page report with your the working plan.*Deadline 7th December.*
- Conduct a machine learning project on the data. Note that the description of the experimental protocols points out to possible studies.
- Produce Jupyter notebook(s) with the project. Note that might be convenient to create new data sets for the project. These data sets versions might be included in the submission to the elearning. *Deadline 30th December.*
- Write four pages with the scope of your study and main results. The report should also include a statement about the contributions of each element of the group to the work. *Deadline 30th December.*

For convenience of the reader the two data sets [1] and the signal acquisition protocols are briefly described.

[1] Both data sets were provided by Ana Rita Teixeira. Any question about the availability of the data out of the scope of this course can be sent to ateixeira@ua.pt

1 The HCT test: looking for biomarkers on the signals

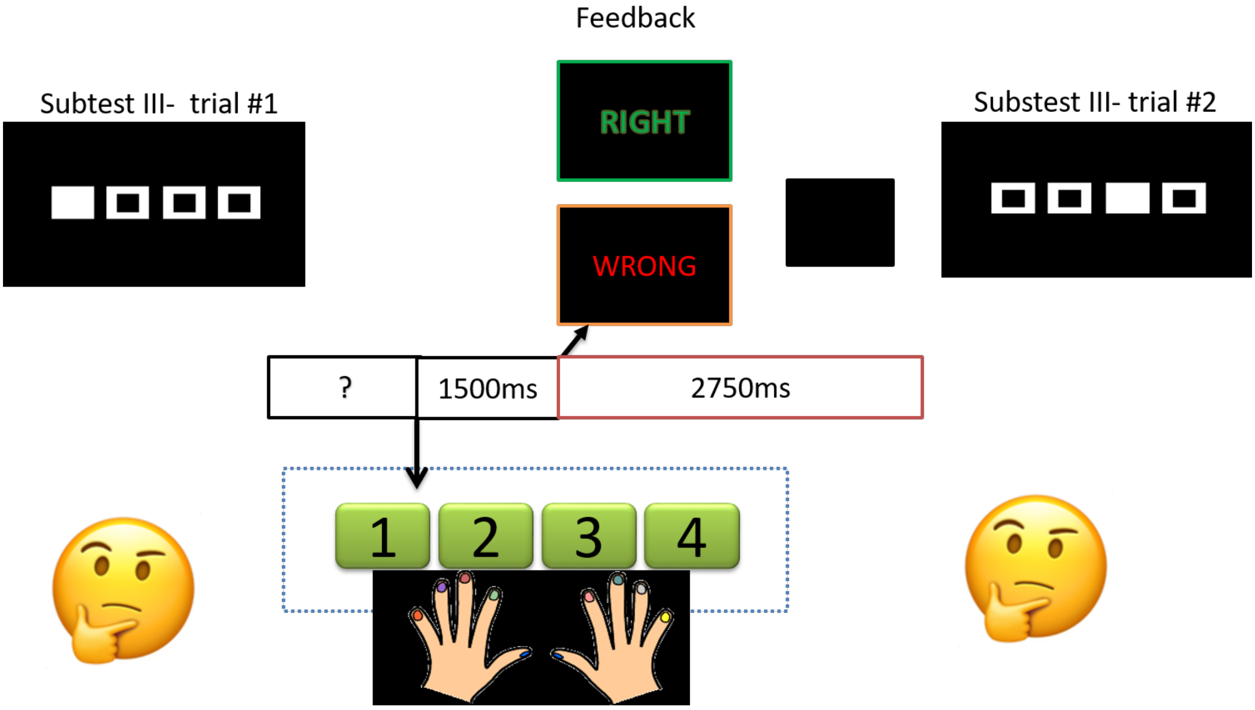

Figure 1: The signal acquisition protocol while participants answer the questions.

The Halstead Category Test (HCT) test, as illustrated in the figure 1, consists of geometric figures or designs (stimulus), and the participant is asked to indicate a number between 1 and 4 as it is suggested by the stimulus. After every response, visual feedback on whether the response was correct or incorrect is provided. The feedback helps participants to adjust their strategy to answer to the next stimulus. The Halstead Category Test is a popular measure of abstraction, concept formation, and logical analysis skills. The test is applied since long clinically and raditionally, the test provides only an overall error score indicative of global frontal brain impairment.

The HCT test was applied to 58 participants and the EEG signals were recorded. The signals were registered according to the $10-20$ system in 26 scalp positions (called channels in the data set). The signals were segmented around the participant's responses as illustrated in the figure 1. For each trial an average signal was calculated with the signals of all participants. Two different strategies to calculate averages were followed:

- A trial template signal as an average of the signals of all participants.
- A template signal of the *wrong* answers. In that case the average of the signals of participants which give incorrect answers.
- A template signal of the *right* answers.In that case the average of the signals of participants which give correct answers.

Note that the average signals are in the first case obtained with 58 signals while the number of signals for *wrong* or *right* templates is variable.It can also happen that *wrong* template is not available.

1.1 Data base of features

The feature extraction block measures different events of the template of the different scalp signals. Two types of events were considered: time and frequency.

1.2 Time events: event related potentials

The detection of characteristic evoked potentials which are peaks that occur in certain defined windows synchronized with

- **response**: ERN (error related negativity) defined in a pre-defined window and it characterized by the latency and the negative peak value.
- **feedback**: FRN (feedback error negativity) defined in a window of and characterized by the latency and the negative peak value. And P500 defined in a window defined by its latency and positive peak value.

The figure 2 illustrates the described events marked on the filtered version of the signal.

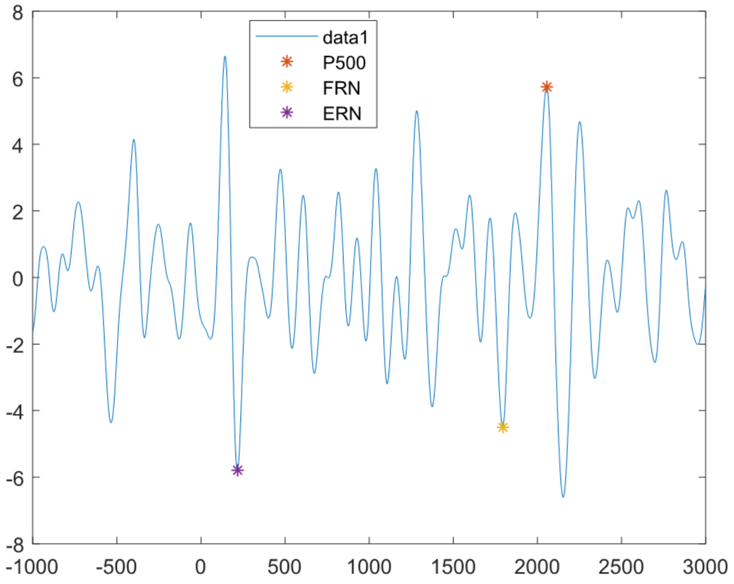

Figure 2: The features stored in database. The first time events is the ERN and occurs after the participant response (t=0) and its time stamp (latency) and amplitude. After the feedback (t=1500)is the FRN and P500

1.3 Frequency events: event related potentials

The energy of the characteristic bands, or EEG rhythms, was calculated into 3 segments of 200*ms* after the response and feedback. Two segments before the response were also included and can be used as possible reference measures. The characteristic bands were: theta, alpha, beta and gamma.

The energy of the reference interval can be used to estimate the increase or decrease of energy of one of the rhythms divided by the energy of reference

$$\Delta = \frac{E_I - E_R}{E_R}$$

where E_I is the energy estimated on I interval and E_R is the energy of the interval. In literature different strategies to define an interval of reference are discussed. The data set provides two intervals of $200ms$ defined before the response. Note that a positive value indicates an increase of the activity relative to the reference interval while the negative value indicates the opposite.

1.4 Machine learning tasks

Different questions can be raised for the data.

- The first question is related with quality of the data and normalization. Therefore the simple univariate analysis of the extracted features can be applied to study the existence of outliers and to know the percentage of non-successful measurements on the trials. Note that *NAN* indicates that the feature is not available.

- Supervised machine learning methods can be applied to study the following problems

  - The HCT test measures the performance of the participants in spatial reasoning and proportional reasoning. The first group is type 3 and 4 and second is type 5 and 6 (see data files). Any evidence on the signal's that allows to classify the trial as proportional or spatial?
  - Any evidence on the signals related with *right* and *wrong* answers.
  - Naturally the participant ignores which strategy(proportional or spatial) should be adopted to answer to the questions but he is informed when the test changes. And notice that it does not change between 3 and 4 but changes between 4 and 5. Any evidence on the signals related with the changes.

2 The Raven test: any difference between groups?

Raven matrices tests have widespread practical use -as a measure of intelligence in the general population for both adults and children, for job applicants as a psychometric test, for applicants to the armed forces, and for assessing clinical populations. In this study the Advanced Progressive Matrices (RAPM) with 48 problems were applied two distinct populations: 21 students of Design and Multimedia and 24 of students Engineering Informatics. EEG signals were registered while the participants perform the tasks of the test. The 48 problems that form the test are divided into two phases: the first 12 where the participant receive a feedback about his answer and the last 36 where no feedback is given. The score of the test is the number (or percentage) of correct answers of the second part. The figure 3 illustrates the time evolution and the relevant marks for each trial

The signals were registered using Enobio 8 EEG recording headset and 8 channels: $F3$, $F4$, $T7$, $C3$, Cz, $C4$, $T8$ and Pz. And the relevant marks for the signal analysis are around

- Problem display;

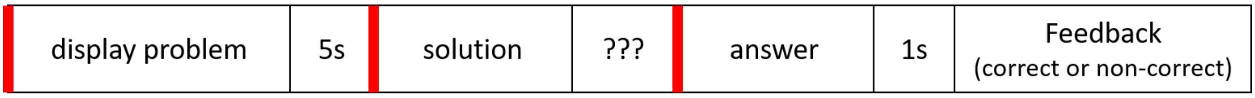

Figure 3: The signal acquisition protocol while participants answer the problems. The first 12 problems include the feedback while the last 36 do not. The red bars represent the time markers stored with the signals

- Possible solutions display;
- Student answer;

And with these time marks the following signal processing windows were considered

- on the problem display $[-75 \quad 500]ms$.
- on the possible solutions display $[-75 \quad 500]ms$.
- on the student answer $[-500 \quad 500]ms$.

3 Feature extraction

The feature extraction was performed using two strategies: in average or in single-trial signals. Therefore the datasets are formed using

- average signals considering the training and the testing phase. The averages might be of all trials or averages of signals when the answer is correct and non-correct.
- single trial signal.

3.1 Time Features

The chosen event-related potentials were $P100$ and $P300$. These were chosen because of attentional characteristics that have sometimes been attributed to the $P100$ potential, as well as the attentional and relationship that is known to exist between the $P300$ potential and cognitive activity. The $P100$ and $P300$ re defined in the first two intervals (display and solution) and its latency and amplitude is stored. Only relevant channels were considered (see the data files).

3.2 Frequency Features

The energy of the characteristic bands are estimated in all defined windows. The energy (E) in the characteristic bands are used to compute other higher level features defined as energy ratios.

- Stress Index [5] : $\frac{E_\beta}{E_\alpha}$
- Mental Fatigue [1], [6]: $\frac{E_\alpha+E_\theta}{E_\beta}$
- Alpha Lateralization [3] :$\frac{E_{\alpha,F3}-E_{\alpha,F4}}{E_{\alpha,F3}+E_{\alpha,F4}}$

- Immersion Index [4]: $\frac{E_\theta}{E_\alpha}$

Several channels were considered (see data set).

3.3 Machine Learning Tasks

- The first question is related with quality of the data and normalization. Therefore the simple univariate analysis of the extracted features can be applied to study the existence of outliers and to know the percentage of non-successful measurements on the trials.
- Supervised machine learning methods can be applied to study the following problems
  - Are the two groups different?
  - What is different in write and wrong answers?
  - And Male versus Female?

References

[1] Sayed Ahmed Alwedaie, Habib Al Khabbaz, Sayed Redha Hadi, and Riyadh Al Hakim. EEG-Based Analysis for Learning through Virtual Reality Environment. *Journal of Biosensors & Bioelectronics*, 09(01):1–6, feb 2018.

[2] Annushree Bablani, Damodar Reddy Edla, Diwakar Tripathi, and Ramalingaswamy Cheruku. Survey on brain-computer interface: An emerging computational intelligence paradigm. *ACM Comput. Surv.*, 52(1):20:1–20:32, February 2019.

[3] Felisa M. Córdova, M. Hernán Díaz, Fernando Cifuentes, Lucio Cañete, and Fredi Palominos. Identifying Problem Solving Strategies for Learning Styles in Engineering Students Subjected to Intelligence Test and EEG Monitoring. *Procedia Computer Science*, 55:18–27, jan 2015.

[4] Yunhan Ga, Taejin Choi, and Gilwon Yoon. Analysis of Game Immersion using EEG signal for Computer Smart Interface. *Journal of Sensor Science and Technology*, 24(6):392–397, nov 2015.

[5] N H A Hamid, N Sulaiman, S A M Aris, Z H Murat, and M N Taib. Evaluation of human stress using EEG Power Spectrum. In *2010 6th International Colloquium on Signal Processing & its Applications*, pages 1–4. IEEE, may 2010.

[6] Budi Thomas Jap, Sara Lal, Peter Fischer, and Evangelos Bekiaris. Using EEG spectral components to assess algorithms for detecting fatigue. *Expert Systems with Applications*, 36(2):2352–2359, mar 2009.