Machine Learning and Brain States

Alencar M. Gabriel Henrique

INSA de Rouen

May 28, 2019

Summary

The subject

Context

Being able to classify the brain state

Final Goal

Study different classification methods and find tools that improve the performance of the selected models.

General framework of the project

1. Create a python code with all the procedures for the PAO of the INSA Rouen ;
2. Develop a short article on the results of the project for the Federal University of Ceará.

Theory

There are 5 types of brain waves that work almost like musical notes. Some act at low frequency, others at a higher frequency.

The brain is an electrochemical organ. All this electrical activity is responsible for different types of brain waves. Throughout the day, the brain keeps active the 5 types of brainwaves. Depending on what we do at any given moment, some waves will show greater activity in certain areas of the brain, and others will work less intensely in other regions. None of them will be, in other words, "disconnected."

Wave types

1. Delta waves (1 to 3 Hz) : Delta waves are those that have a higher wavelength and are related to deep sleep (but without dreams) and bodily activities that we are not aware of normally;
2. Theta waves (3.5 to 8 Hz) : is related, above all, to the imaginative capacities, the reflection and the sleep;
3. Alpha waves (8 to 13 Hz) : Appear in the midnight dusk where there is calm, but not sleep, where there is relaxation and a favorable state to meditate. A high level of Alpha waves would prevent us from focusing attention;
4. Beta waves (12 to 33 Hz) : Accented by states that are related to the daily activities in which we focus all the attention, when we are alert and we need, at the same time, to be attentive to numerous stimuli;
5. Gamma waves (25 to 100 Hz) : This wave is related to tasks of high cognitive processing. They have to do with our learning style, with the ability to record new information, and also with our senses and perceptions, states of happiness and the phases of REM sleep.

Analysis of the theory

Knowing the different types of brain waves allows us to understand the mental processes, emotions, activities and dynamics that generate a type of "energy" in the brain.

The Alpha, Beta and Gamma waves are very interesting for recognition of brain states linked to actions, feeling and concentration.

But we have more frequencies that can be explored through pretreatment.

Synchronized Brainwave Dataset

[1] The MIDS class at the UC Berkeley School of Information is sharing a dataset collected using consumer-grade brainwave-sensing headsets, along with the software code and visual stimulus used to collect the data (the dataset includes all subjects' readings during the stimulus presentation).

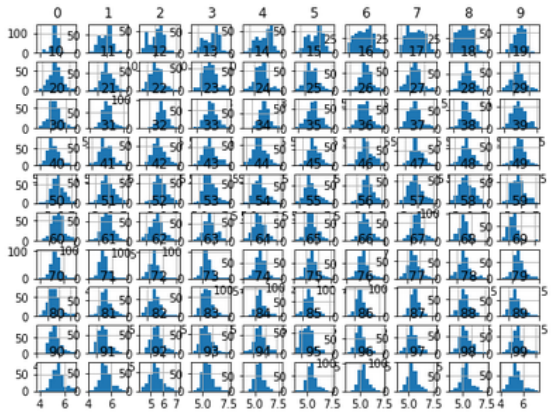
To simplify our work, we will focus only on the first experiment that is an experiment that measures the level of concentration of a patient when he makes mental calculations of different equations

Description of the type of methods

We will be working with a binary classification algorithm, since we want to know if the patient is relaxed or concentrated.

Analyzing the data - Histogram

```
# Math data
plots = math_data.hist()
```



```
# Relax data
plots = relax_data.hist()
```

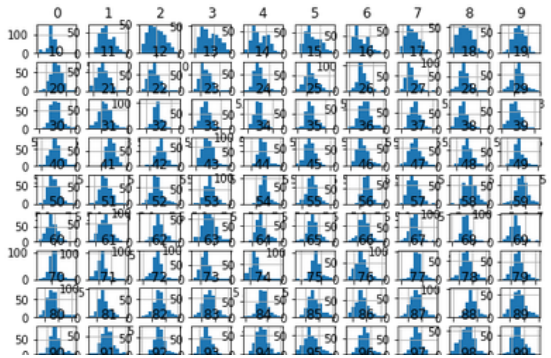


Figure: Histogram.

Analyzing the data - Covariance

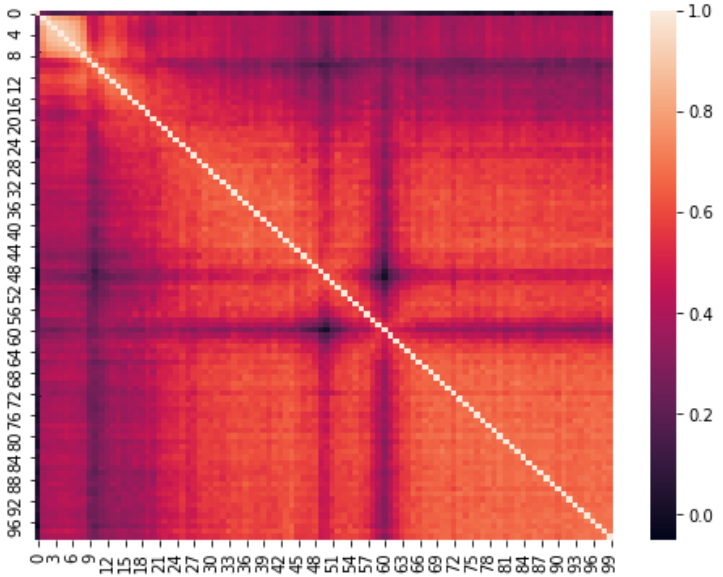


Figure: Covariance matrix.

Analyzing the data - PCA

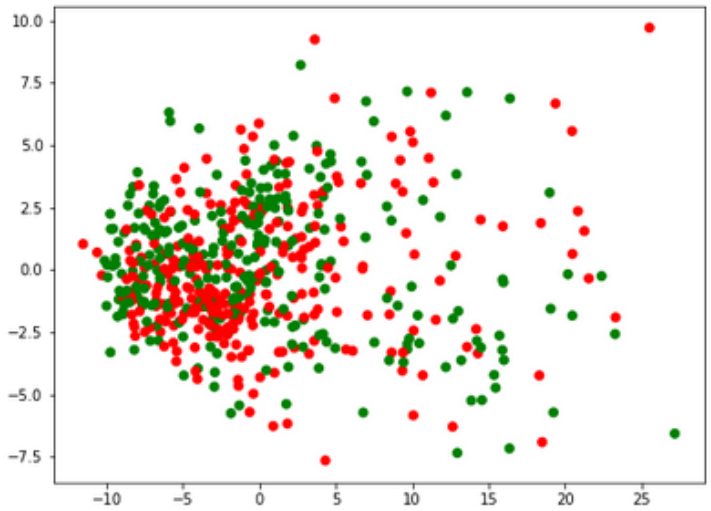


Figure: PCA.

Other results for this data set

On the site Kaggle - Synchronized Brainwave Dataset[1], several people have posted their codes with their results, but only two posts are really interesting for us:

1. Classifying relaxation versus doing math[2]
2. math/relax prediction with different classifiers[3]

Classifying relaxation versus doing math

Each group of 512 raw values produced by the device was taken, and a fast fourier transformation (FFT) was performed to produce a spectrum of energy. Then groups of 3 power spectra are picked up, averaged and the result is logarithmically grouped to produce feature vectors of 100 values.

Problem: Used one person as a training set and another as a test

Math/relax prediction with different classifiers

8 possible brain waves with common frequencies as attributes were used. Got good precision with SVC (Support Vector Classification), but showed that in the end using the preprocessing of the data set from the previous solution is even better.

Tested Methods

Linear Methods:

1. LDA - Linear discriminant analysis
2. The logistic regression
3. Linear SVM

Nonlinear Methods:

1. Naive bayes classifier
2. QDA - Quadratic Discriminant Analysis
3. Gaussian Mixture Model - GMM
4. K-nearest neighbors - KNN
5. Random Forest and AdaBoost
6. Kernelized SVM

Hyper parameter management

It is not clear if the first parameters used in the tests will result in the best possible model. For this, there are techniques for finding the best parameters within a set, such as Random Subsampling, Leave-one-out Cross-Validation, Bootstrap, and K-Fold Cross-Validation.

Pipeline

A machine learning pipeline is a set of processes where the result of one process is input from another, and each process acts together with others to generate better results.

Automated Machine Learning - AutoML

It is a family of methods (classifiers, features preprocessing, data preprocessing, ...) that work together to obtain an improved classification model using transfer learning.

Methodology

We do the same job of **Classifying relaxation versus doing math** for preprocessing, but splitting the dataset into 2 (training and test without shuffle)

Method handlers

Basically, experiments were carried out with 2 classes in python:

1. BCIBinaryClassifierModel : to have and manage different classification methods, as well as their performances in a given data set ;
2. BCIBinaryClassifierModelCrossValidation : perform a cross-validation to find out the best parameter for each method.

AutoML

AutoML uses 15 classifiers and 14 preprocesses of features. Therefore, at least 210 different models must be taken into account.

For 3 days, that is, 259200 seconds, to test at least 210 models, each model must have at least 1235 seconds to run.

Results - 1/4

```
bci_model.showResults()
```

	Naive bayes classifier	LDA	Logistic regression	Mixture model	Mixture model Bayes	k nearest neighbors	k nearest neighbors Cross Validation	k nearest neighbors Grid
Precision on a test set	54.455446	53.960396	55.940594	50.0	49.50495	58.910891	52.475248	60.39604

	Pipeline VS_knn	Pipeline + HP_tuning	Linear SVM Linear	Linear SVM Grid	Random Forest	Ada Boost	Kernel SVM
Precision on a test set	53.960396	61.386139	51.980198	55.940594	49.009901	50.990099	50.49505

Figure: Methods without cross validation.

Results - 2/4

```
bci_model_cross_validation.showResults()
```

	Naive bayes classifier	Logistic regression	Mixture model	Mixture model Bayes	k nearest neighbors	k nearest neighbors Cross Validation	k nearest neighbors Grid
Precision on a test set	54.455446	58.415842	66.831683	66.831683	58.910891	58.910891	55.445545

	Pipeline VS_knn	Pipeline + HP_tuning	Linear SVM Linear	Linear SVM Grid	Random Forest	Kernel SVM	AdaBoost
Precision on a test set	61.386139	58.415842	58.910891	53.960396	67.326733	62.871287	63.861386

	Auto ML
Precision on a test set	58.415842

Figure: Methods with cross validation.

Results - 3/4

The last value was not bigger than 61 percent accuracy. Therefore, different situations can be taken into account: perhaps overfitting has occurred, perhaps training has taken into account sets of methods that do not provide good estimates for a long time execution, among other factors.

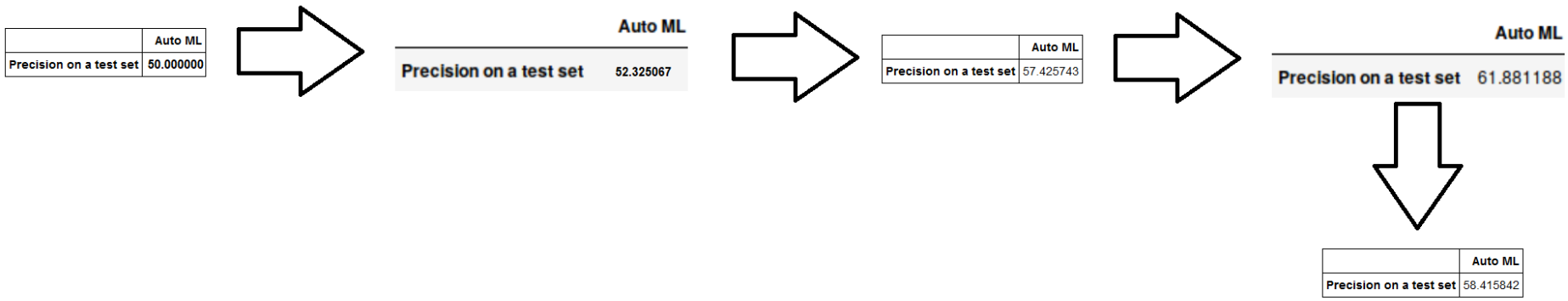


Figure: AutoML progress.

Results - 3/4

However, we can see that AutoML still has a lot of potential. You then only need to discover the set configurations that do not provide good performance and remove them, or increase the number of data, or change training times, and so on.

Results - 4/4

Best Parameters

```
bci_model_cross_validation.showBestParams()
```

	Logistic regression	Mixture model	Mixture model Bayes	k nearest neighbors Cross Validation
Best params of test set	{'lr_penaltys': 'l2', 'lr_solvers': 'saga'}	{'gmm_n_components': 3, 'gmm_tol': 0.888888888...	{'gmm_bayes_n_components': 3, 'gmm_bayes_tol':...	{'knntest_size': 3, 'knn_random': 0.44, 'rando...

	Pipeline VS_knn	Pipeline + HP_tuning	Linear SVM Linear	Random Forest	Kernel SVM
Best params of test set	{'anova k': 1, 'svcc': 1}	{'anova k': 4, 'n_neighbors': 1}	{'svmC': 0.001}	{'maxdep': 200, 'estimators': 2000, 'mfeatures...	{'sgamma': 0.1, 'sC': 1, 'skernel': 'rbf'}

Figure: Best parameters.

Achievements

What were the achievements of the PAO?

EEG Headset - Unicorn

Figure: Unicorn The brain interface.

Code

As a result of this experiment, a python code was created that not only is able to import, analyze and pre-treat data related to mental states, but also to test different methods and see which are the best results for the most classical methods and AutoML.

Article

An article that has the information about the work done, talking about the information of this presentation in more depth.

Conclusion - 1/2

It is easy to see that normally the performances would not be very good, but that with a tunning of hyperparametres one can find better performances than those found in the old results.

There are numerous ways to increase performance, one of them being the possibility to increase the number of people in the database or use better sensors.

Conclusion - 2/2

What is important to note are the methods that gave above 60 percent and especially those that gave above 65 percent because they have shown themselves to be capable of solving problems involving BCI data, and AutoML is a powerful tool that if handled well can result in performance better than all the performances shown here and hence better than the old results.

Références

Synchronized Brainwave Dataset :

https://www.kaggle.com/berkeley-biosense/synchronized-brainwave-dataset

Classifying relaxation versus doing math

https://www.kaggle.com/seaneuron/math-relax-prediction-with-different-classifiers

math/relax prediction with different classifiers

https://www.kaggle.com/elsehow/classifying-relaxation-versus-doing-math

Merci

Merci beaucoup !