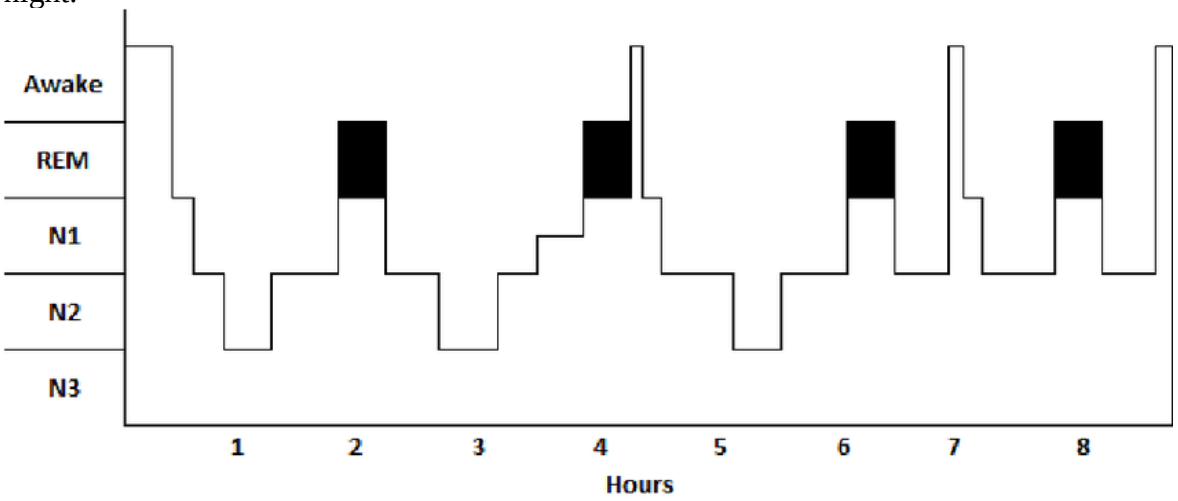
Sleep Stage Classification with EEG signal analysis using Pyspark

Thibault Leblanc, Gabriel Olympie, Antoine de Mathelin

I. Motivation and Problem Definition:

Sleep plays a vital role in an individual's health and well-being. Sleep progresses in cycles that involve multiple sleep stages : wake, light sleep, deep sleep, rem sleep. Different sleep stages are associated to different physiological functions. Monitoring sleep stage is beneficial for diagnosing sleep disorders. The gold standard to monitor sleep stages relies on a polysomnography study conducted in a hospital or a sleep lab. Different physiological signals are recorded such as electroencephalogram (EEG signals), electrocardiogram etc.

Sleep stage scoring is then performed visually by an expert on epochs of 30 seconds of signals recording. The resulting graph is called a hypnogram. He provides a compact description of the night.



The idea of the project is to develop an algorithm of sleep staging able to differentiate between Wake, N1, N2, N3 and REM on windows of 30 seconds of raw data. We focused on the feature extraction part which is time consuming in the process and tried to use pyspark to speed it up

II. Data and methodology

II.1 Data

The data come from the Kaggle challenge of the Machine Learning course. The ‘Dreem’ company provides them.

It includes 7 EEGs channels in frontal and occipital position, 1 pulse oximeter infrared channel, and 3 accelerometers channels (x, y and z). Each row from the data is separated between the different signals recorded by the Dreem headband. Signal is shapes in windows of 30 seconds.

Electroencephalogram is measured at 7 different locations of the head. The sampling frequency is 50 Hz. Wich makes 1.500 values per signals channel. An exemple of EEG signal is represented on Figure 1.

Pulse oximeter and accelerometer channels are sampled at 10Hz. Wich makes 300 values per signals channel.

There are two sets of data: a “train” and a “test” one, they are respectively composed of nearly 40.000 rows. The total amount of data for each sets is **3.2 Go** what could be considered huge enough to use distributed and parallelized methods for analyzing them.

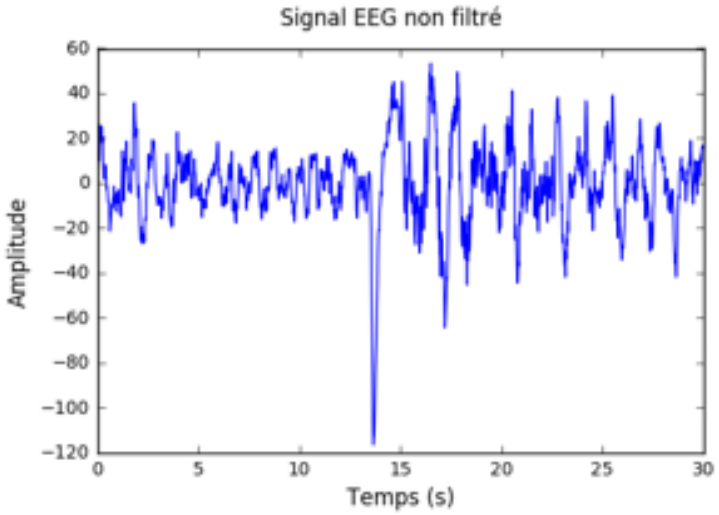
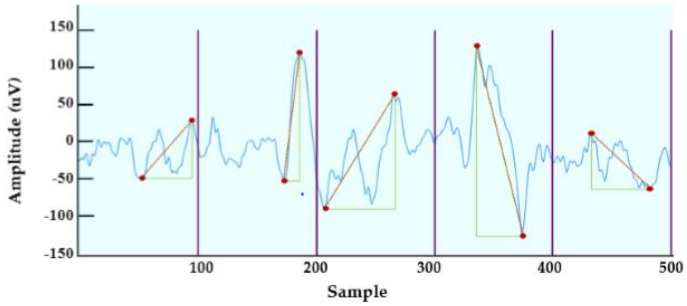


Figure I

II.2 Methodology

In order to use classifiers on this data we first need to preprocess data and to extract features. Preprocessing consist in extracting for each EEGs signals 4 different features: it’s mean,

standard deviation, its minimum maximum distance calculated on subwindows of the signal:



Minimum maximum distance

And its entropy

The entropy is defined by sampling the signal into 40 sub window between -200 and 200, then calculating the probability Pk that a point belong to one window, and finally applying the entropy formula:

$$S = \sum -p_k \log(p_k)$$

Finally, we applied a pipeline of Standard scaling, PCA with 4 components and svm with gaussian kernel on sklearn

The large amount of data and the complexity of the preprocessing cause the features extraction to take a lot of time with numpy. Therefore, we considered this problem to be a good opportunity to take advantage of pyspark's distributed framework

Finally, we applied a pipeline of Standard scaling, PCA with 4 components and svm with gaussian kernel using sklearn to be able to classify the sleep stages accordingly.

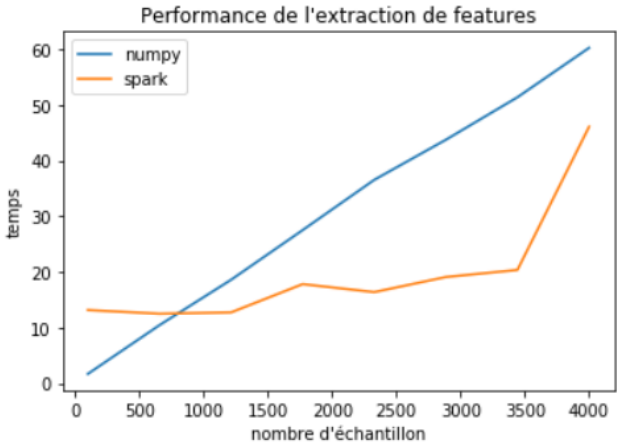
III. Results

We compared the time taken to extract the features as a function of the number of samples.

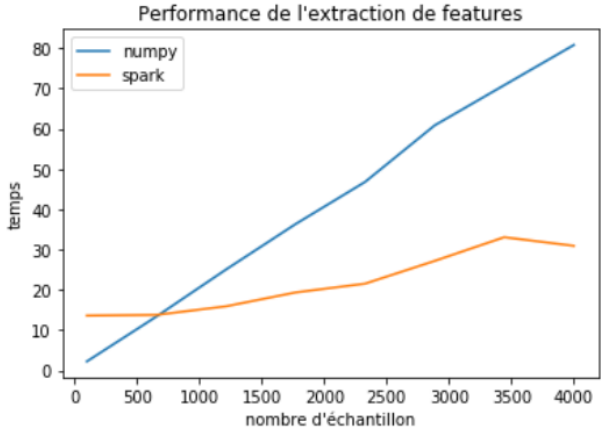
We applied the extraction process with numpy on the one hand, and with a pyspark dataframe on the other.

The configuration we used for the spark context was 8 cores, 8Gb of executors memory and 8 Gb of dirvers memory.

We could notice that the time taken to extract the features with pyspark was not linear as it is with numpy. Pyspark was faster when the number of samples exceeded 500.



However, we also ran it on a configuration with a wider RAM (16 go) and noticed that the saturation for large amount of data didn't append anymore:



References

[1] A Comprehensive Survey and New Investigation Khald Ali I. Aboalayon, Miad Faezipour, Wafaa S. Almuhammadi and Saeid Moslehpour, Sleep Stage Classification Using EEG Signal Analysis, 23 August 2016

[2] Mohammed Diykh, Yan Li, Complex Networks Approach for EEG Signal Sleep Stages Classification July 2016