Rapport projet Dreem

Classification supervisée des phases du sommeil par signaux EEG, cardiaques et d'accélérométrie

Barthel Cyprien et Doyen Baptiste

Janvier 2019

1 Introduction

Le sommeil comporte plusieurs type de phases : éveillé (non-sommeil) ; léger ; profond et paradoxal. Ces phases correspondent toutes à des fonctions physiologiques différentes et contribuent ainsi à la qualité du sommeil de l'individu. Il apparaît donc comme important de pouvoir déterminer dans quelle phase se situe un individu au cours de son sommeil afin d'obtenir un meilleur diagnostic de ses éventuels troubles du sommeil.

C'est l'objet de ce projet : à partir de différents signaux physiologiques du type electroencephalogramme, electrocardiogramme et enfin des signaux liés à la position dans l'espace, déterminer dans quelle phase de sommeil se situe l'individu.

Figure 1: L'appareil de mesure, le bandeau Dreem

2 Données d'apprentissage

2.1 Mesures

Les données utilisées dans ce projet proviennent d'enregistrements réalisés à partir de bandeaux Dreem tel que celui pris en photo plus haut.
Plus précisément, nous disposons de 38289 observations chacune labellisée par une phase de sommeil et collectées sur plusieurs individus.
Chacun de ces observations regroupe 10 types de signaux observés chacun pendant une fenêtre temporelle de 30s.
Ces 10 signaux sont les suivants : "accelerometer-x","accelerometer-y","accelerometer-z","eeg-1", "eeg-2", "eeg-3", "eeg-4", "eeg-5", "eeg-6", "eeg-7" et "pulse-oximeter-infrared". "eeg" fait ici référence aux signaux du type encéphalogramme (mesure d'une activité électrique du cerveau à partir d'une électrode positionnée à un certain endroit du crâne); "accelerometer" mesure l'accélération linéaire du bandeau sur la tête et "pulse" le pouls cardiaque.
La fréquence d'échantillonage des signaux du type "eeg" est de 50Hz, celle des signaux du type "accelerometer" et du signal "pulse-oximeter-infrared" est de 10Hz. Pour une observation fixée et pour chaque signal du type "eeg", nous disposons donc de $30 \times 50 = 1500$ points de données et pour les signaux du type "accelerometer" et "pulse-oximeter-infrared" de 300 points de données.
Enfin, il s'agit de classifier ces signaux en 5 classes précisément : Wake (*éveil*), N1, N2, N3 et REM (*paradoxal*)

On peut visualiser ces signaux comme suit :

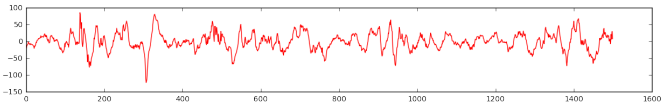

Figure 2: Une observation du signal "eeg-1" pendant $30s$

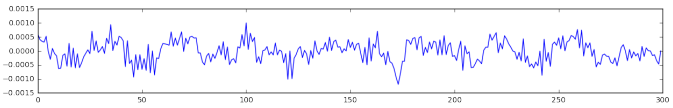

Figure 3: Une observation du signal "accelerometer-y" pendant $30s$

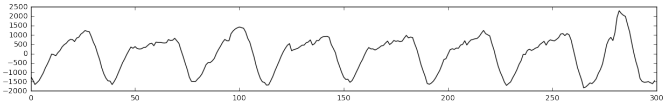

Figure 4: Une observation du signal "pulse-oximeter-infrared" pendant $30s$

2.2 Modélisation mathématiques

Concernant les signaux du type EEG, on peut les modéliser comme des signaux non-stationnaires. En effet, les signaux ne semblent pas posséder de formes particulière.
D'un point de vue neurologique, cela s'explique par le fait que le signal enregistré change de source au cours du temps : passant d'une structure du cerveau à une autre. Ces changements de sources induisent ainsi une non-stationnarité au signal. Quantitativement, ces changements interviennent tous les $0,25s$[1]. Afin d'observer des signaux quasi-stationnaires, il conviendrait ainsi de segmenter le signal selon des bandes de fréquence de longueur $\frac{1}{0,25} = 4\text{Hz}$.
Cela explique ainsi la longueur des bandes θ $(0-4\text{Hz})$, α $(4-8\text{Hz})$ et β $(8-13\text{Hz})$ présentées dans l'article de Khald Ali I. Aboalayon et al[2].

Ainsi, certaines features sont calculées par rapport à chacune de ces bandes séparément au lieu d'être calculée sur toute la gamme de fréquence du signal. Cela nous permet ainsi de nous ramener à des signaux quasi-stationnaires. Dans le cas contraire, le calcul de ces features perdrait alors de son sens et on risquerait un certain effet de confusion.

3 Pre-processing

3.1 Filtration

Conformément à ce qui a été présenté plus haut, nous avons filtré le signal d'entrée de sorte à ne sélectionner que les bandes de fréquences souhaitées. Pour cela nous avons utilisé 3 filtres passe-bande du type Butterworth pour les bandes θ, α et β et un filtre passe-bas pour la bande δ.
La bande γ n'a pas été sélectionnée car il s'agit d'un signal de fréquence supérieure à 30Hz et les signaux dont nous disposons ont été échantillonnés avec une fréquence de 50Hz. Le spectre de ces signaux sera donc sans pertes d'information pour les fréquences inférieures à 25Hz selon le théorème d'échantillonnage de Nyquist-Shannon.

Visuellement, voici une illustration avant et après filtration passe-bande pour la bande θ sur une acquisition du signal EEG-1 :

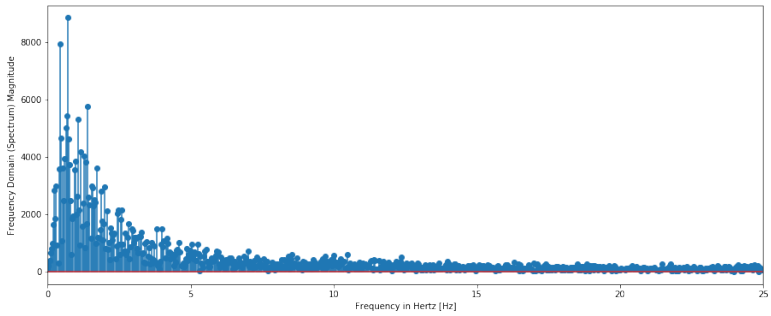

Figure 5: Avant filtration

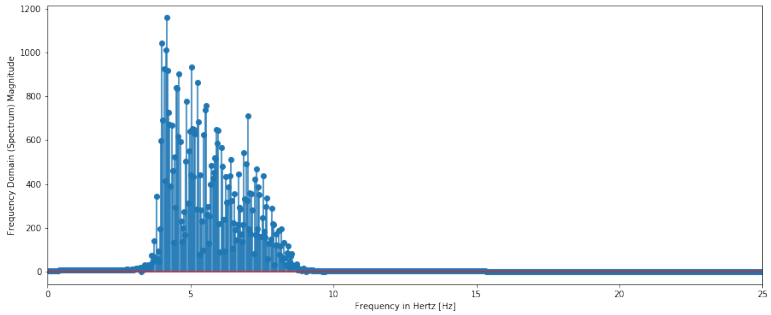

Figure 6: Après filtration

La filtration est donc conforme à ce que l'on pourrait attendre : les fréquences hors-bande sont coupées et le gain sur la bande passante est unitaire.

3.2 Extraction de features

La section qui suit présente toutes les features qui ont été testées pour réaliser la classification automatique. On les a regroupés selon 3 catégories. Pour la plupart de ces features, il était plus pertinent de calculer ces grandeurs sur des signaux stationnaires et nous les avons donc calculées pour chacun des bandes de fréquence mentionnées plus haut. À l'exception des coefficients d'autorégression, qui eux ont été calculés sans filtration au préalable.

3.2.1 Features statistiques empiriques usuelles (domaine temporel)

Soient X la variable aléatoire correspondant aux réalisations d'un signal donné, n le nombre d'observations étudiées et x_i une réalisation de X.

- **Moyenne (moment d'ordre 1)** : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Variance (moment d'ordre 2)** : $Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2$
- **Assymétrie (skewness, ordre 3)** : $Skew(X) = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^3}{\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2\right)^{3/2}}$
- **Aplatissement (kurtosis, ordre 4)** : $Kurt(X) = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2\right)^{2}} - 3$

3.2.2 Autres features calculées dans le domaine temporel

On pose $p_i := \frac{|x_i|}{\sum_{i=1}^{n} |x_i|}$ de sorte à ce que (p_i) forme une distribution au sens des probabilités.
De plus, on note $X^{'}$ la dérivée discrète de X, c'est-à-dire : $X'_t := X_t - X_{t-1}$

- **Entropie de Shannon** : $H(X) = -\sum_{i=1}^{n} p_i log(p_i)$
- **Mobilité (Hjorth)** : $Mobilité(X) = \sqrt{\frac{Var(X')}{Var(X)}}$
- **Complexité (Hjorth)** : $Complexité(X) = \frac{Mobilité(X')}{Mobilité(X)}$
- **Coefficients d'Auto-Regression** : On cherche à modéliser le signal d'entrée à l'aide de coefficients (a_k) tels que $x_n = \sum_{k=1}^{n-1} a_k x_k + e(n)$.

L'entropie de Shannon est une mesure de la quantité d'information portée par le signal.

Les features *Mobilité* et *Complexité* sont elles liées à la dynamique temporelle du signal (d'où la présence d'une "dérivée") et peuvent être aussi interprétées dans le domaine fréquentiel : la mobilité est une estimée de la fréquence moyenne du signal et la complexité est une mesure de ressemblance du signal à un signal sinusoïdal pur (une seule fréquence).

3.2.3 Énergie du Signal (domaine fréquentiel)

On note $X(f_i)$ la transformée de Fourier du signal (x_i).

- **Énergie** : $E = \frac{1}{n} \sum_{i=1}^{n} |X(f_i)|^2$

Et d'après le théorème de Parseval (version discrète) : $E = \sum_{i=1}^{n} x_i^2$.
En pratique, c'est la somme dans le domaine temporel qui a été utilisée pour calculer E car aucune transformée de Fourier est alors à calculer.

4 Modèle d'apprentissage

4.1 Descriptif du modèle

Un modèle de type *Random Forest* a été choisi pour la tâche de classification. Ce modèle consiste à réaliser un ensemble d'arbres de décisions participant par "vote" à la prédiction finale du modèle.
Chaque arbre est formé à partir d'un sous-ensemble choisi aléatoirement sur les données (de taille fixée) et pour un sous-ensemble aussi aléatoire de features
Cette méthode de construction d'arbres de décision à partir de sous-ensembles assez réduits et formées aléatoirement permet ainsi d'éviter un effet de sur-apprentissage sur les données. En effet, cela permet de diminuer la corrélation entre arbres de décision et donc l'erreur de généralisation.
De manière générale, le meilleur choix résulte d'un compromis entre la capacité prédicitive à l'échelle individuelle des arbre de décision et la capacité prédicitve de la forêt dans son ensemble.
Empiriquement, la partie entière de la racine carrée du nombre total de features du dataset est choisi comme nombre de noeuds pour chaque arbre de décision. La racine carré ou bien aussi le choix du logarithme permettent ainsi de limiter la profondeur des arbres quand le nombre de features devient trop important.

4.2 Entraînement du modèle

Les classes sont fortement non-équilibrées dans la distribution du dataset d'entraînement comme le montre l'histogramme suivant :

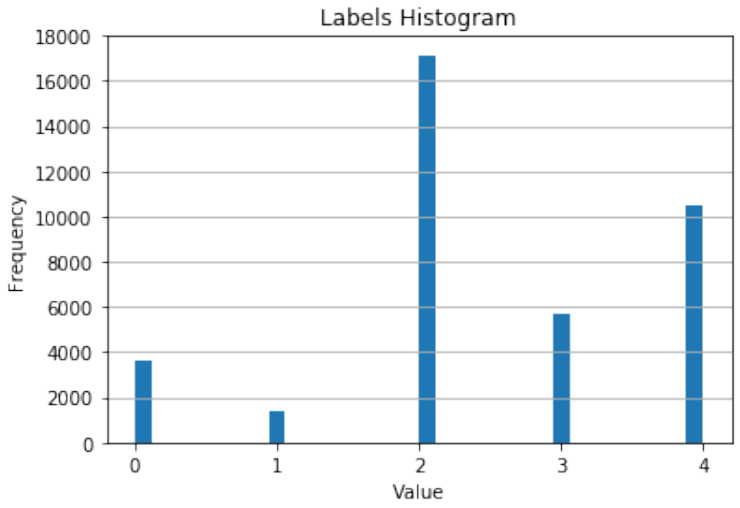

Figure 7: Distribution des classes sur le dataset d'entraînement

Le modèle final étant une méthode par vote, si l'on ne prend pas en compte ce déséquilibre de proportion des classes, il peut en résulter un choix systématiquement biaisé envers les classes majoritaires : dans notre cas, se tromper sur la classe 1 peut être en effet considéré comme bien moins coûteux que se tromper sur la classe 2 (la classe 2 compte plus de 10 fois plus d'observations que la classe 1). À l'inverse, si l'on équilibre strictement les classes, nous perdons de l'information sur les classes majoritaires (on réduit le nombre d'observations) et aussi sur la distribution des classes en général.
Si cette distribution est la même pour le dataset de test, il serait alors plus pertinent de ne pas équilibrer les classes. Après essai avec et sans équilibrage, il apparaît comme plus pertinent de ne pas équilibrer les classes. Pour un même modèle entraîné, le score $F1$ avec équilibrage est de l'ordre de $0,58$ et sans équilibrage de $0,64$.

4.3 Validation croisée

Pour la méthode de validation croisée, nous avons choisi une méthode du type $K - Folds$: après un $shuffle$ des observations, le dataset est splité 5 fois en un dataset d'entraînement et un dataset de validation selon un rapport 80/20. Un score $F1$ est calculé pour chaque split et enfin moyenné comme résultat final.

4.4 Importance des features

L'index de Gini nous donne une mesure du poids relatif de chaque variable dans la prédiction finale. L'histogramme qui suit donne ainsi les 20 premières features pour le modèle testé (qui en compte 260 au total).

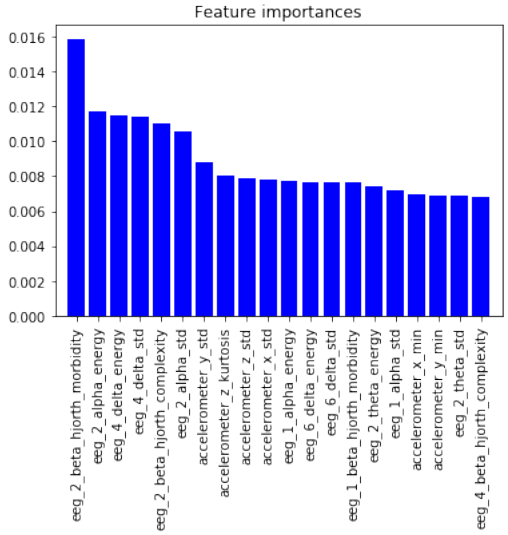

Figure 8: Importance des 20 premières features du modèle global

Sur ces 20 features, tous les signaux EEG (à l'exception de $eeg-5$) ainsi que les signaux d'accélérométrie sont présents. Pour les signaux EEG, les features statistiques de variance et celles liées aux paramètres de Hjorth ains que l'énergie du signal sont les plus pertinentes.
Pour les signaux d'accélérométrie, les features statistiques kurtosis (qui signale la présence ou non d'un pic dans la distribution du signal), variance et minimum sont celles qui comptent le plus.

4.5 Résultats du modèle

De manière générale, le modèle choisi (sans équilibrage) performe raisonnablement bien sur les classes 0, 2, 3 et 4. La classe 1 est la classe où la performance de la classification est insuffisante comme l'atteste la figure suivante :

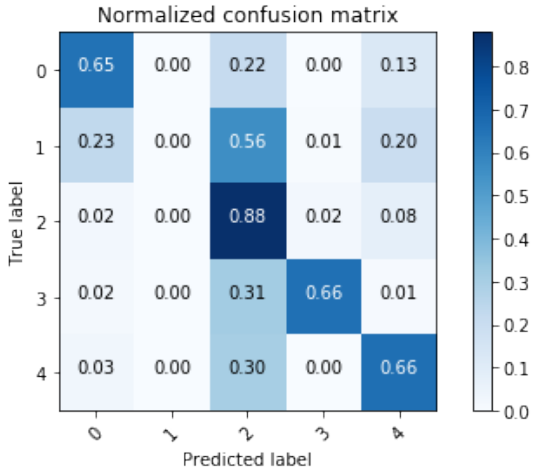

Figure 9: Matrice de confusion normalisée (classes non-équilibrées)

Cette matrice de confusion a été obtenue après un test sur un dataset de validation splité selon le principe 80/20.
Elle permet de visualiser rapidement où le modèle performe le mieux et où il se trompe le plus.
De manière prévisible, la classe 2 (fortement majoritaire) est correctement prédite dans 88% pourcent des cas.
En revanche, la classe 1 n'est jamais prédite et le modèle prédit près de 56% de ses observations dans la classe 2.
Pour un compte rendu plus précis des performances pour chaque classe C compte tenu de sa proportion dans le dataset d'entraînement, on peut aussi calculer le ratio $R_{up-lift} = \frac{\mathbb{P}_{modèle}(pred \in C | obs \in C)}{\mathbb{P}_{data}(obs \in C)}$:

Classe	$R_{up-lift}$
0	6,9
1	0
2	1,97
3	4,41
4	2,4

Figure 10: Performance du modèle par rapport à un choix aléatoire selon une distribution uniforme

Dans l'ordre, les classes les mieux identifiées par le modèle sont donc les suivantes : 0 (Wake), 3 (deep sleep), 4 (paradoxical sleep), 2 (light sleep 2) et 1 (light sleep 1).

À titre de comparaison, on retrouve globalement cet ordre de comparaison avec la matrice de confusion normalisée du modèle sur classes équilibrées :

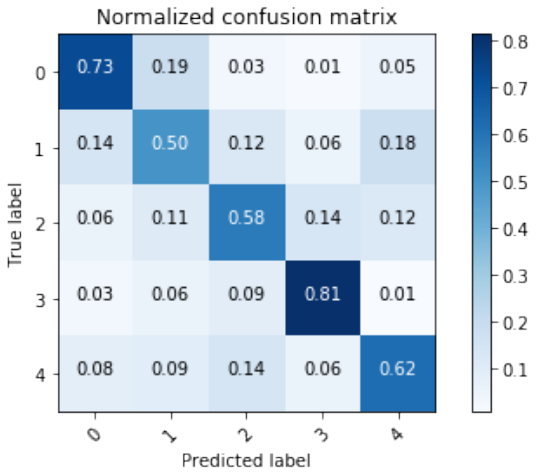

Figure 11: Matrice de confusion normalisée (classes équilibrées)

5 Conclusion

Comparativement aux autres modèles proposés lors du challenge interne, le modèle décrit plus haut se situe au mieux à la 9ème position pour les soumissions en équipe (avec un score final de 0.63967) et le meilleur modèle est $2,2\%$ au-dessus..
Afin d'améliorer ce score, on peut envisager des features plus spécifiques à la classe 1, classe que le modèle ne parvient pas à prédire de manière satisfaisante. De nouvelles features liées aux pulsations cardiaques plus spécifiquement pourraient également être intéressantes à étudier : en effet, ce signal ne figure pas parmi les variables les plus discriminantes du modèle étudié. Enfin, l'ensemble du code est aussi disponible à l'adresse suivante :
https://github.com/bdoyen/Challenge$_D reem/blob/master/code_d reem.ipynb$

References

[1] *Everything you wanted to ask about EEG but were afraid to get the right answer.* Wlodzimierz Klonowski. Nonlinear Biomedical Physics 2009.

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