Arousal Detection in Obstructive Sleep Apnea Using Physiology-Driven Features

Sandya Subramanian[1], Shubham Chamadia[2], Sourish Chakravarty[1]

[1]Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Massachusetts General Hospital, Boston, MA, USA

Abstract

Obstructive sleep apnea (OSA) is a condition in which a person repeatedly stops breathing during sleep due to closure of the upper airway, leading to a cycle of sleep fragmentation and intermittent hypoxia (oxygen deficiency). Conventional methods for detecting and quantifying OSA are largely based on physiological monitoring during sleep followed by manual labeling of sleep stages and arousals. Here there is scope for computerized methodologies that can efficiently and objectively perform this characterization of sleep.

As part of the CinC/Physionet 2018 challenge to automatically detect arousals in a large, expert-annotated sleep dataset, we extracted 27 spectral and time domain features, chosen for their physiological relevance, from the available training set and implemented two contrasting methods, Generalized Linear Model (GLM) and Random Forest (RF), to classify arousals and non-arousals.

We were able to achieve non-trivial classification accuracy, even in an imbalanced data set with far fewer arousals than non-arousals. This suggests that large machine learning problems can still benefit from physiology-informed feature selection, especially in the biomedical space.

1. Introduction

1.1. Obstructive Sleep Apnea

Obstructive sleep apnea (OSA) is a serious sleep disorder in which patients repeatedly stop breathing during deeper stages of sleep [1]. This is due to upper airway collapse, which leads to rapid desaturation of oxygen. The lack of oxygen (hypoxia) triggers the response of the sympathetic nervous system ("fight or flight" response), a natural response to stress in the body, which triggers recurring arousals to reopen the airway and restart breathing. Most arousals occur without the conscious awakening of the patient at all, including vocalizations, snores, bruxisms, and periodic leg movements. Hypopneas are periods of abnormally slow or shallow breathing, and apneas are periods of no breathing.

While suspicion of OSA can arise from a variety of factors, including co-morbidities such as obesity [2], reported snoring or restless sleep at night or drowsiness during the day, or physical examination findings like narrowed airways, "kissing" tonsils, or abnormal Mallampati score, a diagnosis is usually confirmed in the setting of a sleep laboratory [3]. In such sleep laboratories, behavioral response and physiological signals are recorded from patients while they are asleep. Sleep experts then manually score different sleep stages as well as the presence of different types of arousals in epochs of 30 seconds for the whole night of sleep. In this context, automated detection of arousals could greatly reduce the time and cost required to diagnose OSA, in addition to reducing human error and developing an objective diagnosis scheme.

1.2. The Challenge and Related Dataset

The current work ensued from a competition conducted by Computing in Cardiology (CinC), whose dataset was made available on the Physionet website [4]. The competition aimed at detecting sources of arousal (non-apnea) during sleep using various physiological signals including electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), electrocardiogram (ECG), and blood-oxygen saturation (SpO2) that were each sampled at 200 Hz. This dataset consists of signals from 1983 subjects (994 for training and 989 for testing) that were recorded at Massachusetts General Hospital's (MGH) sleep laboratory dedicated for the diagnosis of sleep disorders. The dataset was also manually annotated for various stages of sleep (30 seconds intervals) as well as the presence of arousals events (hypopneas, snores, vocalization, etc.) by certified sleep technicians at MGH. The aim of the challenge was to correctly detect/classify the target arousal epochs, which included all arousal types except for apneic and hypopneic arousals. Specifically, the scoring was based on how well the vectors of instantaneous probability of arousal predicted for each test subject was able to detect the non-apneic/non-hypopneic arousals. In this paper, however, we also focus on the detection of

 ISSN: 2325-887X DOI: 10.22489/CinC.2018.349

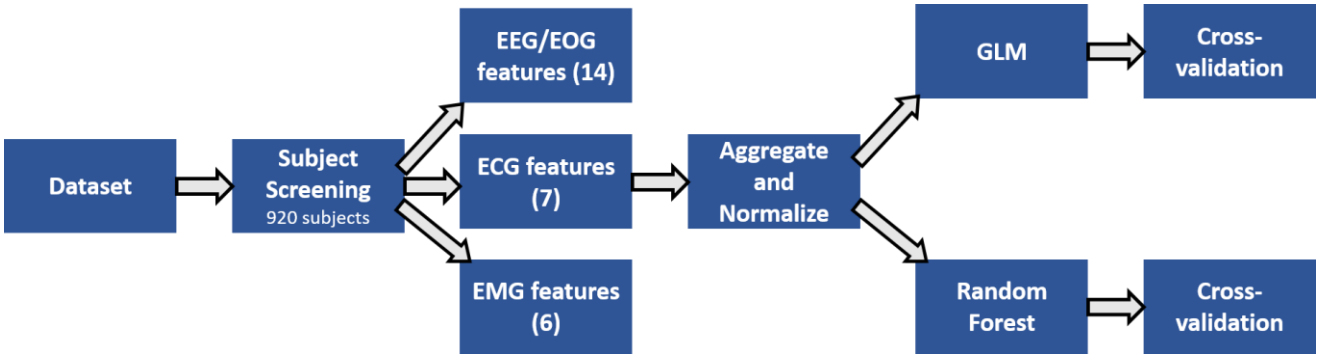

Figure 1 Overall schematic of the methods

arousals including apneic/ hypopneic and non-apneic/non-hypopneic arousals using two classification methods, the Generalized Linear Model (GLM) and the Random Forest (RF) (Fig. 1).

2. Methods & Results

We analyze all signals, sampled at frequency, $f = 200\,Hz$, with a time resolution Δ=5s since the target arousals themselves are often very short in duration (~2 sec). To assign an arousal state to each Δ window (0: non-arousal, +1: non-apneic/non-hypopneic arousal, -1: apneic/hypopneic arousals), we use a "majority vote" polling method to set the arousal state, y_k, based on the mode of the trinary observations within the kth window of size Δ . Since the focus for this paper is detection of all arousals, the -1's are converted to +1's in the arousal state vector prior to model fitting. When missing data scenario is encountered at any instant in any of the signals, the corresponding Δ window is omitted. Physiologically relevant features are chosen from each modality for arousal detection.

2.1. EEG and EOG Features

For every window of size Δ, we extract features from the Multitapered Power Spectral Density (PSD) [5], F_j, for the frequency bin, ω_j (in Hz) with constant width W and using a stationary time window = Δ, time-half-bandwidth product = 2, and number of tapers = 3. We perform the following regression, $log_{10}(F) = b - c\,log_{10}(\omega)$, for the kth window of size Δ, to identify parameters, b_k and c_k, that characterize the background '1/f' decay [6]. Then, using the residuals, $log_{10}\left(R_k(\omega_j)\right) = log_{10}(F_j) - (b_k - c_k log_{10}\omega_j)$ we calculate the following parameters, $\delta_k = W\sum_{0<\omega_j\leq 4} R_k(\omega_j)$, $\theta_k = W\sum_{4<\omega_j\leq 8} R_k(\omega_j)$, $\alpha_k = W\sum_{8<\omega_j\leq 14} R_k(\omega_j)$, $\beta_k = W\sum_{14<\omega_j\leq 30} R_k(\omega_j)$ and $\gamma_k = W\sum_{30<\omega_j\leq 55} R_k(\omega_j)$. Finally, for each subject, the features are rescaled as, $z_k = (x_k - \mu)/\sigma$, where x_k refers to any of the 7 fields, $[b_k, c_k, \delta_k, \theta_k, \alpha_k, \beta_k, \gamma_k]$, extracted for any of the 6 EEG and 1 EOG channels and μ and σ, respectively, correspond to the subject-specific sample mean and standard deviation associated with x_k. Our choice of frequency bands for EEG is inspired by those used for tracking sleep-stages: delta, theta, alpha, beta, and gamma bands [7]. Here our implicit assumption is that the sleep stages can be correlated to arousal states. For computational efficiency, we decided to use 1 EEG channel based on our initial exploratory data analysis of the features across 6 EEG channels.

2.2. ECG Features

The ECG is pre-processed by extracting R peaks using the Pan-Tompkins algorithm [8]. RR intervals that are too long (> 2 seconds) or too short (< 0.3 seconds) are corrected with the addition or removal of R peaks respectively. After the extraction of R peaks, four time-domain ECG features are computed for each window of size Δ as follows: (1) mean of all RR intervals within that window, (2) standard deviation of all RR intervals within that window, (3) root mean square of the difference between consecutive RR intervals within that window, and (4) proportion of differences between consecutive RR intervals greater than 0.05 seconds within that window.

Then the multitapered power spectral density is computed for each stationary time window = Δ, time-half-bandwidth product = 3, and number of tapers = 5. Three more frequency-domain features are computed as follows: (5) total power between 0.04 and 0.15 Hz (low frequency), (6) total power between 0.15 and 0.40 Hz (high frequency), and (7) the ratio of low frequency to high frequency power within that window. All ECG features are heartrate variability measures, which are good indicators of autonomic tone. Since the autonomic nervous system is strongly affected during arousals ("fight-or-flight" response), we hypothesize that these features would be useful.

2.3. EMG Features

The dataset contained three different EMG channels: chin, chest, and abdomen. Chin EMG contained mainly

high frequency activity relating to jaw clench, while chest and abdomen contained much lower frequency sinusoidal activity correlating with the rise and fall of the chest during breathing. Two features are extracted for each channel as follows: (1) total power in each window of size Δ from 0-100 Hz for chin and 0-5 Hz for chest and abdomen, and (2) difference in total power between consecutive windows for each channel. The multitapered spectral estimates are computed using time-half-bandwidth-product = 3 and number of tapers = 5.

2.4. Classification schema

We set up the analysis scheme for both classification methods, GLM and RF, as follows. As mentioned earlier, we group non-target arousals (apneic and hypopneic arousals) with target arousals. Arousal states are down-sampled as previously described to 5-second windows. With such coarse-grained arousal observations and 27 features extracted at the same time resolution, the models can be learnt from a training data set. These predictive models can be implemented on any test case to estimate a vector of instantaneous probabilities of arousal. Once such predictions are computed for every 5 second window, they can be expanded to yield prediction probability vectors at the sampling rate (200 Hz) by simply assigning the prediction computed for each window to all 1000 "samples" within the 5 second window. Feature matrices were also constructed including features from previous and following time windows (Table 1). When tested on datasets where arousal observations are available, metrics such as the area under the receiver operating curve (AUROC) and the area under the precision recall curve (AUPRC) can be computed. Such metrics are useful to analyze the classification abilities of the chosen methods.

Within the training dataset provided by the competition, we identified 74 (out of 994) subjects as outliers in the feature space and discarded them from the model fitting and validation steps. Across the remaining 920 subjects, the mean proportion of arousals compared to non-arousals is found to be 0.201. The proportion of just target arousals is found to be 0.059.

2.5. A naïve Generalized Linear Model

The first method we used for classification was a GLM of the form

$$E\left[log\left(\frac{p}{1-p}\right)\right] = a_0 + \sum_{p=1}^{P} a_p x_p$$

to classify arousals vs. non-arousals. Here, P denotes the number of features used, p is the instantaneous probability of an arousal, $E[\,]$ denotes the expectation operator and x_p indicates the pth feature. To analyze the out-of-sample error from GLM, we use a cross-validation scheme on 920 subjects from the training set by partitioning it into 10 folds with 92 subjects in each. Each fold consisted of around 5000-6000 data points. For every fold, we fit the GLM on the training data from the other 9 folds, and then apply the estimated model to predict the arousal probabilities for each subject from the held-out fold. Using these predicted arousal probabilities and available ground truth arousal data, we calculate the AUROC and AUPRC. The GLM training and testing are conducted using pre-defined MATLAB functions *glmfit()* and *glmval()*.

Table 1 Both GLM and Random Forest models are trained to detect All Arousals including types +1 and -.1. AUPRC and AUROC metrics are calculated for two test scenarios: for predicting All Arousals and for predicting Target Arousals Only (arousals of type +1). The basic Random Forest and GLM were first constructed for zero lag (y_k assumed to depend only on features from the k-th window). Both frameworks were later extended to incorporate additional higher order dependence on m time windows in the past (m back) as well as n time windows in the future (n forward).

Method	Time Lags	Target Arousals Only		All Arousals	
		AUROC	AUPRC	AUROC	AUPRC
Baseline	-	0.5	0.059	0.5	0.201
Random Forest	Zero lag	0.659	0.101	0.694	0.364
Random Forest	1 back	0.675	0.111	0.733	0.425
Random Forest	1 back + 1 forward	0.722	0.139	0.760	0.464
Random Forest	2 back	0.717	0.134	0.756	0.456
Random Forest	2 back + 2 forward	0.766	0.178	0.803	0.541
Random Forest	3 back + 3 forward	0.787	0.204	0.821	0.577
Random Forest	4 back + 4 forward	0.807	0.227	0.838	0.609
Random Forest	5 back + 5 forward	0.812	0.235	0.843	0.622
Random Forest	6 back + 6 forward	0.815	0.238	0.847	0.630
GLM	Zero lag	0.586	0.076	0.601	0.263
GLM	2 back	0.645	0.098	0.651	0.306
GLM	2 back + 2 forward	0.661	0.101	0.664	0.321

2.6. Random Forest

The second method we used for cross-validation is Random Forest. Like the GLM, the training data set is partitioned into the same 10 folds with 92 subjects in each. A random forest of 100 weak learners is trained on each fold, where each decision tree is constrained to having no more than seven splits. The LogitBoost algorithm is used to boost accuracy for all forests. The pre-defined MATLAB function *cfitensemble()* was used to train all forests. Then for each of the 920 subjects, predictions for each window are computed as the average of the predictions yielded by the nine forests not trained on that subject's data (the other nine folds). The AUROC and AUPRC from both GLM and RF are compared against baseline in Table 1.

3. Discussion & Conclusion

There are several points worthy of note from this work. First, we used principled tools to extract physiologically relevant features from 920 subjects. With only 27 such features, we are able to achieve non-trivial classification power. This study, therefore, illustrates the potential of physiology driven feature selection for machine learning problems in biomedical signal processing.

Secondly, we contrast two very different classification methodologies. GLM assumes a parametric distribution for the data, while RF is a fully non-parametric method that makes no such assumption, but relies on a deterministic framework for classification. Note that neither of these methods is black box; in both cases, potentially valuable information regarding which features are important can be extracted post-hoc to help advance our understanding of the disease itself. The significantly better performance of RF over GLM (a popular modeling tool used to describe a simple function of an observed variable of interest by a linear combination of covariates) indicates that algorithms that allow for prediction estimates to be described by nonlinear combination of covariates might lead to better performance for the problem at hand.

Third, inspired by co-participants' in the competition who had observed good performance by incorporating history-dependence, we extended our initial model (that assumed the arousal probabilities to depend only on the features at the same instant) to incorporate additional dependencies on features from few time-windows in the immediate past and future. This extension led to significant improvements in the AUROC and AUPRC scores. Interestingly, some of our features such as the spectral estimates are themselves capturing the temporal correlation in the signals within each 5 second bin. This fact, together with the improvement in the accuracy metrics by leveraging the local temporal structure in the feature estimates across few multiples of 5 seconds, indicates that there may lie a multi-scale temporal correlation pattern co-occuring with arousals. Elucidating this pattern further in greater detail can be a promising direction of research.

Another key challenge that future research can address is with regard to detection of target arousals, specifically, on principled approaches to deal with detection of rare events. Finally, the classification paradigms based on physiological relevant features presented here can also be extended to other biomedical signal processing problems, such as in sleep staging. For example, sleep stages are often distinguished based on their EEG spectral signatures, which we also use as features in this work. Thus, further extensions of this work could also be in sleep stage determination for patients with obstructive sleep apnea.

Acknowledgements

We would like to thank Leon Chlon and Pegah Kahaliardabili for their ideas and assistance. We would also like to acknowledge the resources given to us by the Brown Lab, Akeju Lab, MIT, and MGH.

References

1. Downey III R. Medscape. [Online].; 2018. Available from: emedicine.medscape.com/article/295807-overview.
2. Ho ML, Bass SD. Obstructive sleep apnea. Neurology International. 2011.
3. McNicholas WT. Diagnosis of obstructive sleep apnea in adults. Proceedings of the American Thoracic Society. 2008;: p. 154-160.
4. Ghassemi MM, Moody BE, Lehman LwH, Song C, Li Q, Sun H, et al. You Snooze, You Win: the Physionet/Computing in Cardiology Challenge 2018. Computing in Cardiology. 2018;: p. 1-4.
5. Haller M, Donoghue T, Voytek B. Parametrizing Neural Power Spectra. BioRxiv. 2018.
6. Prerau MJ. Tracking the sleep onset proces: an empirical model of behavioral and physiological dynamics. PLoS Computational Biology. 2014.
7. Pan J, Tompkins WJ. A Real-Time QRS Detection Algorithm. IEEE Transactions on Biomedical Engineering. 1985;: p. 230-236.
8. Thomson DJ. Spectrum estimation and harmonic analysis. Proceedings of the IEEE. 1982;: p. 1055-1096.

Address for correspondence.

Sandya Subramanian, sandya@mit.edu
77 Massachusetts Ave. 46-6057A
Cambridge, MA 02139