Identifying Sleeping Difficulty from PolySomnoGraphic (PSG) recordings

Final Project of CSE6250 – Fall 2019

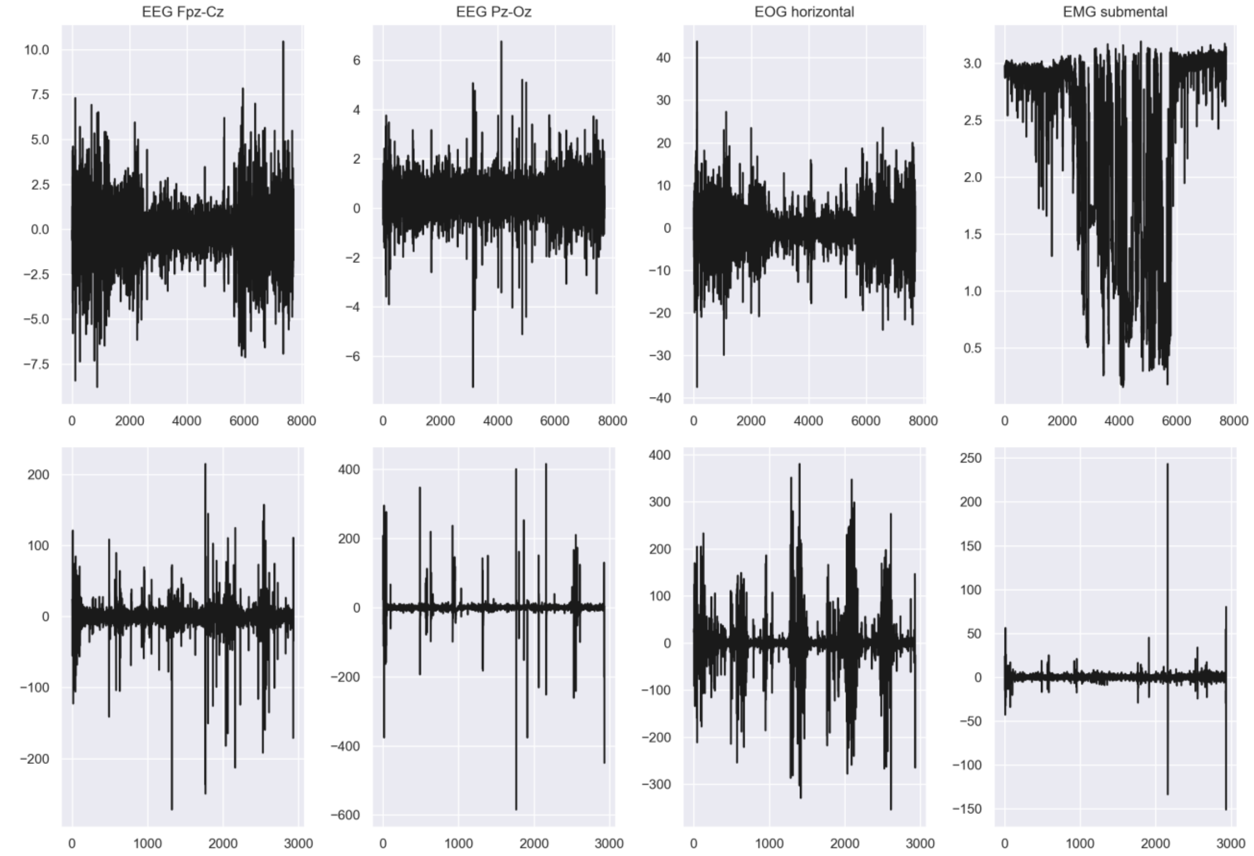
Chuan Wang, Oriah Ulrich, Desmond Chan, Hanlin Chen

Abstract

One common health issue in people's daily life is sleep difficulty. One of the common methods to monitor people's sleeping stage and identify potential sleep problems is to use electroencephalogram (EEG) which tracks and records the electric activity of human brains, electrooculography (EOG) which measures the corneo-retinal standing potential that exists between the front and the back of the human eye. Electromyogram (EMG) which tracks and records the electric activity of skeletal muscle. A medical practitioner would be able to identify sleeping difficulty based on long time of practice. However, for a non-professional, EEG records are extremely hard to interpret. In this paper, we are using machine learning model to provide a simple yet strong classifier to tell whether an EEG record indicate potential sleep difficulty. In this work, we have used various strategies/methods such as Fourier transformation, data augmentation, etc. to improve the classification accuracy. We explore multiple model options from the basic logistic regression to more advanced ones such as random forest, gradient boosting machine, neural networks to improve the classifier's prediction performance.

Introduction

The data being used in this study are the 8.1GB PhysioNet dataset from sleep-edf database. The sleep-edf database contains 197 whole-night PolySomnoGraphic (PSG) sleep recordings, containing EEG, EOG, chin EMG, and event markers. There are two groups of test subjects in the collected data. One group is obtained in a 1987-1991 study of age effects on sleep in healthy Caucasians aged 25-101 containing 153 PSG sleep recordings. The other group is obtained in a 1994 study of temazepam effects on sleep in 22 Caucasian males and females without other medication containing 44 PSG sleep recordings. Subjects in this group had mild difficulty falling asleep but were otherwise healthy (https://www.physionet.org/content/sleep-edfx/1.0.0/). Since the first group of test subjects generally do not have sleep difficulty while the second group contain test subjects who have mild sleep difficulty. We utilize this as a binary response that labels whether a PSG recording indicates sleep difficulty. The figure below shows two sets of sample PSG recordings from two groups. The top row of graphs shows the recordings of EEG, EOG and EMG from a test subject that has no sleep difficulty while the bottom shows those recordings of a subject with sleep difficulty. Note that original readings were recorded at every milisecond, for simplicity of visualization without affecting the general patterns, we have taken the average of each 10 seconds, that is, every 1000 readings.



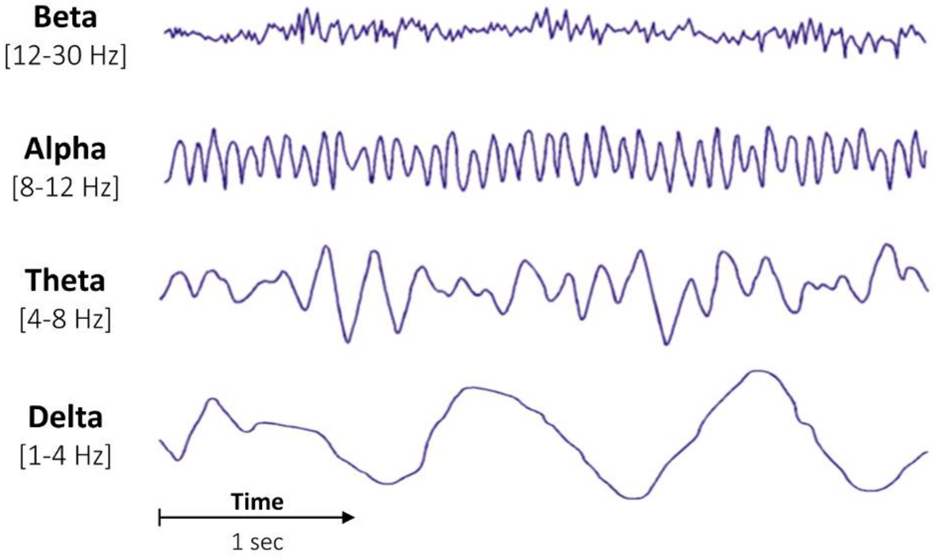
Methods

Data augmentation

The original database contains only 197 whole-night (with time ranging from 5 hours to 20 hours) PSG sleep recordings which is a very small dataset for the purpose of machine learning modeling. In order to reach a relatively larger data size that help us build a more robust model, we used a data augmentation strategy by separating the whole-night recording into one-hour buckets. By utilizing such an approach, we augmented the 197 observations into 3,747 observations. Then we will try to predict whether a person has sleep difficulty based on only one hour of PSG recording.

Feature engineering with band power

Even after separating the complete PSG recordings into one-hour buckets, each one-hour recording is consisted of 360,000 (3600*100) readings. It is impractical to directly put all those readings as features into the model for two major reasons. First it brings too much computational cost, second it leads to serious overfitting. In previous work related to machine learning/deep learning with PSG sleep recordings such Biswal et al., some summary statistics such as the mean, median, max/min or quantile values are often utilized as features at the aggregate level. In this work, we decided to use Fourier transformation over the sleep recordings, decompose the EEG, EOG and EMG signals into frequency component and then calculate the average power of a signal in different frequency ranges (see figure below for the specific frequency ranges being considered in this study). We used 3 channels of waves (EEG Fpz-Cz, EEG Pz-Oz and EOG) and 4 frequency ranges which resulted in 12 (3x4) features in the final modeling dataset.



Models and Evaluation Metrics

Given that the problem we are addressing in this study is essentially a binary classification problem and we have generated 12 numerical features in the previous step, we have started with some basic classification models such as logistic regression, random forest and gradient boosting machine (GBM). We randomly split the 3,747 observations into 70% vs 30% for training and test respectively.

In the original dataset, there are more recordings for test subjects without sleeping difficulty than test subjects with sleep difficulty (153 vs 44). On the other hand, it appears that test subjects with sleep difficulty generally have less sleeping time than those without. After data augmentation, the unbalancedness in response is further aggravated (3394 vs 353), the negative cases (sleep difficulty) only takes 8.3% of population. In order to avoid overly skewed predictions under the data skewness, we used 4 different evaluation metrics to evaluate the model predictions. Aside the common metrics for binary classification such as accuracy and area under curve (AUC) of receiver operating characteristics (ROC) curve, we also utilized metrics AUC under precision-recall (PR) curve (Davis and Goadrich) and log loss which are more robust against data unbalancedness.

Initial Results

In table below we have shown the initial results of our classification models.

	Accuracy	**ROC AUC**	**PR AUC**	**Log Loss**
baseline	0.917	0.500	0.083	0.693
Logistic Regression	0.940	0.957	0.710	0.151
Random Forest	0.956	0.954	0.721	0.231
GBM	0.972	0.986	0.903	0.128

Note that the first row “baseline” uses naïve predictions (e.g., in case of accuracy, simply always predict the more frequent case while in other cases, simply use a predicted probability of 0.5). From the results we can see that the band power features we have generated are effective in that even the simplest model logistic regression was able to significantly improve the metrics over the baseline. Secondly across all the four evaluation metrics, GBM is outperforming the other models by a significant margin. The table below shows the confusion matrix resulting from GBM model on test data. Again, we can see the classifier is making reasonably good predictions.

	negative	positive
negative	1017	15
positive	16	77

Next Step

For next steps, we would like proceed with our exploration in three directions: more advanced models especially recurrent neural network (RNN) based models, automatic feature engineering based on autoencoding of PSG waveform, and including more data points from other databases such as Sleep Heart Healthy Study (https://sleepdata.org/datasets/shhs).

RNN

The PSG recordings are in their essence sequence data and there have already been lots of research related to utilizing RNN based networks to identity sleep state, problems, etc. (Zhang and Fabbri). We will try to improve our classifier using those ideas and frameworks.

Autoencoding

Currently our model used manually generated features based on a combination of Fourier transformation and band power calculation. One approach to improve the current pipeline is to utilize autoencoding techniques to automatically generate features which could potentially reduce information loss during the process of manually generating features.

More data

Even with data augmentation, our current dataset is relatively small and we would like to see if the classifier performance are consistent over other sources of dataset.

Reference

O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou. Automatic sleep stage scoring with single-channel eeg using convolutional neural networks. arXiv preprint arXiv:1610.01683, 2016.

S. Biswal, J. Kulas, H. Sun, B. Goparaju, M. Brandon Westover, M. T. Bianchi, and J. Sun. SLEEPNET: Automated sleep staging system via deep learning. 26 July 2017.

M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In International Conference on Machine Learning, pages 4100–4109, 2017.

R. U. D. K. Linda Zhang, Daniel Fabbri. Automated sleep stage scoring of the sleep heart health study using deep neural networks. 2019.

Dean DA 2nd, Goldberger AL, Mueller R, Kim M, Rueschman M, Mobley D, Sahoo SS, Jayapandian CP, Cui L, Morrical MG, Surovec S, Zhang GQ, Redline S. Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. Sleep. 2016 May 1;39(5):1151-64. doi: 10.5665/sleep.5774. Review. PubMed PMID: 27070134; PubMed Central PMCID: PMC4835314.

Zhang GQ, Cui L, Mueller R, Tao S, Kim M, Rueschman M, Mariani S, Mobley D, Redline S. The National Sleep Research Resource: towards a sleep data commons. J Am Med Inform Assoc. 2018 May 31. doi: 10.1093/jamia/ocy064. [Epub ahead of print] PubMed PMID: 29860441.

Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, Rapoport DM, Redline S, Robbins J, Samet JM, Wahl PW. The Sleep Heart Health Study: design, rationale, and methods. Sleep. 1997 Dec;20(12):1077-85. PubMed PMID: 9493915.

Redline S, Sanders MH, Lind BK, Quan SF, Iber C, Gottlieb DJ, Bonekat WH, Rapoport DM, Smith PL, Kiley JP. Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. Sleep Heart Health Research Group. Sleep. 1998 Nov 1;21(7):759-67. PubMed PMID: 11300121.

B Kemp, AH Zwinderman, B Tuk, HAC Kamphuisen, JJL Oberyé. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. IEEE-BME 47(9):1185-1194 (2000).

Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals (2003). Circulation. 101(23):e215-e220.

Davis, Jesse and Goadrich, Mark, The Relationship Between Precision-Recall and ROC Curves (2006). International Conference on Machine Learning