CS598 Project Proposal: Automatic Sleep Stage Classification

Pradeep Dwivedi
University of Illinois Urbana-Champaign
901 West Illinois Street, Urbana, IL 61801
pkd3@illinois.edu

Michael Renteria
University of Illinois Urbana-Champaign
901 West Illinois Street, Urbana, IL 61801
mrente5@illinois.edu

Jason Kolter
University of Illinois Urbana-Champaign
901 West Illinois Street, Urbana, IL 61801
jkolter2@illinois.edu

Zichun Xu
University of Illinois Urbana-Champaign
901 West Illinois Street, Urbana, IL 61801
zichunx2@illinois.edu

1. MOTIVATION

Sleep disorders are widespread and can have a large impact on a person's health and quality of life [1]. Sleep stages are critical for diagnosing sleep disorders[2]. There are four stages of sleep that various studies analyze in order to find non regular sleeping patterns, and furthermore pinpoint various sleep related disorders. The four stages are broken down into two categories: rapid eye movement (REM) sleep and non-REM sleep. non-REM sleep is composed of stages one, two, and three. REM sleep is a stage and category by itself. Since, it typically takes 90 minutes to hit the fourth stage of sleep in a normal patient [3], it is needless to say that you are often working with many hours of data on each specific patient. The hours of data that you are working with is then multiplied by each of the various sensors you choose to use to help you predict the specific stages of sleep.

It takes a human expert hours to label the polysomnogram (PSG) for a single patient's one night sleep. It is time and labor intensive because sleeping stages are scored based on a collection of rules given by American Academy of Sleep Medicine (AASM)[4]. They include (but are not limited to) signal intensity, patterns, duration, sequence or co-occurrence of signals and so on. Deep learning methods have shown great promise in generating sleep stage labels [5-8] and some models have achieved comparable accuracy to human experts. It can greatly aid the process of diagnosing sleep disorder. Therefore we propose to research deep learning models for automatic sleep stage classification.

2. LITERATURE SURVEY

Researchers have devoted significant effort in developing accurate and reliable deep learning algorithms for automated sleep stage classification.

In SLEEPNET [5], the authors experimented with convolutional neural network (CNN), multi-layered recurrent neural network (RNN) with Long short-term memory (LSTM) network, and RCNN where the final output of CNN is fed into a 2-layer RNN for classification. For feature representations, they experimented with 3 options: raw waveforms of Electroencephalography (EEG) data from 6 different channels, spectrogram (transformed by Fast Fourier Transformation), and expert defined features from time and frequency domain. The dataset consisted of 80M EEG data from 10K patients. RNN with expert defined features showed the best performance with an accuracy of 0.86 and Cohen's Kappa of 0.79. It was able to perform well on a broad cohort of 10K patients and thus better addressed the variability between subjects, compared with other studies with less PSG recordings.

Paper [6] presented a deep learning model using a combination of CNN and LSTM. The convolutional layers evaluate the co-occurrence of signal patterns and LSTM layers discover temporal relationships in signals. The model is trained on PSG data from 5213 patients, sums to about 42,560 hours of sleep data. Compared with previous paper which only used EEG data, this paper used signals from EEG, electromyogram (EMG), and electrooculogram (EOG). Similarly, this paper experimented using raw data as input, as well as spectrograms from applying short-time Fourier transforms. The CNN+LSTM model achieved a high F1-score of 0.87 and Cohen's Kappa score of 0.82 when using spectrogram as input. The model using raw data as input had a lower F1-score of 0.846 but might be more preferable since it's more robust to noise. The accuracy in N1 scoring was very low compared with the other categories because of large class imbalance and PSG scoring rules.

In paper [7], single channel EEG data from 20 healthy young adults, each lasting about 20 hours, were used to train a classifier. Raw EEG signal was used without pre-processing to train a CNN model. The output of CNN was fed into a softmax layer to produce labels. Since they had a smaller amount of data, they used 20-fold cross validation. Mean F1-score was 0.79 and mean accuracy was 0.80 across sleep stages.

Paper [8] illustrates the details of performing sleep stage identification on EEG signals by means of ensemble empirical mode decomposition and random undersampling boosting. The results of using the RUSBoost algorithm are better in most of the cases and comparable in other state-of-the-art methods proposed for sleep stage identification. The paper goes on to highlight the importance of undersampling in this process, which not only removes the class imbalance in this problem, arised due to the inherent nature of human sleep but also requires much lesser computing resources. The RUSBoost is based on a decision tree algorithm and thus much more interpretative then the current deep learning mechanisms.

SPINDLE [3] (Sleep Phase Identification with Neural networks for Domain-invariant Learning) was developed to increase accuracy of detecting sleep stages in lab rats. SPINDLE utilizes two CNN networks and a hidden Markov model. Hidden Markov is a generative model that utilizes hidden states. The research of this paper eventually led to the development of a python package (deepsleep) that has the SPINDLE implementation, along with some extra utilities to process EEG data (EDF file format). The package also includes the functionality to preprocess the data utilizing the technique that was defined in the paper. The layers

utilized in the CNN are convolution, max-pooling, dense, and softmax.

3. DATASET

For this project, we intend to use sleep recording data from the Sleep-EDF Expanded database available at PhysioNet[9] as our primary data source.

This database contains 197 samples of whole-night PolySomnoGraphic recordings suitable for general purpose sleep analysis. All records contain readings from EEG, EOG, chin EMG and event markers with some samples containing respiration and temperature data. The recordings are available in EDF/EDF+[10] format which is the de facto standard for PSG data. The data from these samples will be used to build the features that will be used to train our scoring models. Along with the sample data, there also is a set of expert-scored hypnograms available that contain ground truth annotations corresponding to the PSG recordings.

These recordings were obtained from two separate studies [11, 12]. The first contains 153 healthy Caucasian subjects not using sleep related medication. Each of these subjects has 2 PSG recordings of approximately 20 hours each. The EOG and EEG samples were taken at 100Hz, while the event markers were sampled at 1Hz. The second study contains data from 22 subjects who suffer from sleep issues. There are again 2 samples for each subject however this group contains one each of medicated and placebo dosed sleep assistance.

We have also found the ISRUC-Sleep dataset [13] which is a publicly available sleep dataset also suitable for general purpose analysis that contains an additional 118 samples from 3 different groups, also scored by two human experts. Initial experimentation with the data demonstrated that the Sleep-EDF database was easier to process, so while that will be our primary data source, this ISRUC-Sleep dataset is available for additional data as needed.

Since each PSG file is formatted using the standard EDF/EDF+ format we have several tools available to us for processing. Each PSG and Hypnogram recording will be converted to a delimited text format for use programmatically in our implementation. We believe this dataset gives a good representative sample of different sleep patterns and a mix of healthy and unhealthy subjects that we can use for robustly training the models as discussed in the approach section.

4. APPROACH

4.1 Initial Data Analysis

After reading the data from the input files, we will analyze the sleep data for each field and get the basic statistics of the data e.g., mean, standard deviation, field summary etc. We'll check for the correlation among fields and will remove or transform the fields if they are highly correlated with other fields, present in the input dataset. We'll plot the graph of each field to inspect its distribution. Our endeavor will be to make sure that the data is normally distributed and is devoid of any class imbalances. Additionally, we will perform below steps as well, as part of the data wrangling and data analysis:

- Find the zero-level classifier i.e., finding the maximum occurring class and setting it as zero level classifier. Our endeavor will be to make sure that other algorithms result in significant learning over the zero-level base classifier.
- Analyze the moments of the data i.e., mean, standard deviation, skewness and the kurtosis as part of the exploratory data analysis
- Perform under-sampling and/or oversampling, as needed, to ensure that the input data is representative, produces unbiased results and the class imbalances are taken care of. Additionally, we might be oversampling to avoid overfitting the data.

4.2 Machine Learning Algorithms

We want to explore both traditional machine learning and deep learning algorithms as part of our project work to auto classify the different sleep stages. Based on the exploratory data analysis performed earlier, we plan to use at least one traditional machine learning algorithm and a minimum of one deep learning algorithm.

4.2.1 Traditional Machine Learning Algorithms

The classification metric obtained using this algorithm will be used for comparison with the results obtained using advanced deep learning framework(s). We plan to first set the zero-level classifier as the majority class of the training dataset to ensure that the algorithms used later would result in significant learning on top of the baseline classifier. Depending on the results obtained from the data analysis done before, we'll implement logistic regression (LR) or linear discriminant analysis (LDA) or the quadratic discriminant analysis (QDA). In some of the previous studies of the sleep data it has been shown that the LDA performs better than even the deep learning frameworks on the sleep data [8]. We're open to try ensemble learning using decision tree stumps if time permits [8].

4.2.2 Deep Learning Algorithms

We plan to use CNN to extract the spatial features and RNN to extract the temporal dependencies for the sleep data. The final algorithm used will be a combination of CNN and RNN layers, including transfer learning frameworks as well, as applicable. We're yet to decide the number of layers to use for CNN, RNN and FF networks. The number of layers will be optimized based on the computing requirement to train the model.

We plan to make our model as interpretable as possible. Therefore, if the deep learning framework gives better classification AUC-ROC, the non-deep learning framework with similar accuracy, can be used as the interpretable model.

4.3 Success Metrics

We'll measure classification performance of our models using the AUC-ROC, F1 score and Cohen's Kappa score. Cohen's Kappa score measures inter-rater reliability. It allows us to tell to what extent the labels produced by our model are in agreement with that from a human expert. The overall score from eight European centers is 0.76 [14], and we will be comparing ours with this human level performance.

5. EXPERIMENTAL SETUP

We could either use a local machine (40 virtual CPU cores, 128GB RAM) or an AWS cluster, depending on what we find from the exploratory data analysis step. We'll be working in a python3 environment using pytorch.

6. TIMELINE

We propose the following tentative timeline for our project:

- Get access to the sleep data and read it using standard python libraries – week of April 3rd
- Perform exploratory data analysis – week of April 10th
- Work on the non-deep learning models – week of April 17th
- Work on deep learning models – week of April 17th
- Compare the metrics of different models used and draw conclusions – week of April 24th
- Complete the project documentation, presentation etc. – week of May 1st

7. REFERENCES

[1] Raymond C. Rosen, Mark Rosekind, Craig Rosevear, Wesley E. Cole, William C. Dement, Physician Education in Sleep and Sleep Disorders: A National Survey of U.S. Medical Schools, Sleep, Volume 16, Issue 3, May 1993, Pages 249–254, https://doi.org/10.1093/sleep/16.3.249

[2] Abad, V. C., & Guilleminault, C. (2003). Diagnosis and treatment of sleep disorders: a brief review for clinicians. Dialogues in clinical neuroscience, 5(4), 371–388.

[3] Miladinović Đ, Muheim C, Bauer S, Spinnler A, Noain D, Bandarabadi M, et al. (2019) SPINDLE: End-to-end learning from EEG/EMG to extrapolate animal sleep scoring across experimental settings, labs and species. PLoS Comput Biol 15(4):e1006968. https://doi.org/10.1371/journal.pcbi.1006968

[4] Danker-Hopfe H, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res. 2009;18(1):74–84.

[5] 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.

[6] L. Zhang, D. Fabbri, R. Upender, and D. Kent. Automated sleep stage scoring of the sleep heart health study using deep neural networks. 2019.

[7] 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.

[8] Ahnaf RashikHassan, Mohammed Imamul HassanBhuiyan, Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka 1205, Bangladesh, Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting https://doi.org/10.1016/j.cmpb.2016.12.015

[9] Goldberger, A., et al. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220." (2000).

[10] B Kemp, J Olivan. European data format 'plus' (EDF+), an EDF alike standard format for the exchange of physiological data. Clinical Neurophysiology 114:1755–1761 (2002).

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

[12] MS Mourtazaev, B Kemp, AH Zwinderman, HAC Kamphuisen. Age and gender affect different characteristics of slow waves in the sleep EEG. Sleep 18(7):557–564 (1995).

[13] Khalighi Sirvan, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. “ISRUC-Sleep: A comprehensive public dataset for sleep researchers.“Computer methods and programs in biomedicine 124 (2016): 180-192.

[14] Danker-Hopfe H, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res. 2009;18(1):74–84.