Machine Learning Approach to Classifying Insomnia with EEG

ESE 400 Independent Study

<https://github.com/4Jamintaek/Insomnia>

**Group Members:**

**Andy Dohoon Kim**

Bachelor of Science in Computer Science

[dohoon.kim@wustl.edu](mailto:dohoon.kim@wustl.edu)

**Mintaek Lim**

Bachelor of Science in Electrical Engineering

[mintaeklim@wustl.edu](mailto:mintaeklim@wustl.edu)

**Project Advisor:**

**Jr-Shin Li**

Das Family Career Development Distinguished Associate Professor in the Preston M. Green Department of Electrical and Systems

[jsli@ese.wustl.edu](mailto:jsli@ese.wustl.edu)

Time period: Jan 14-May 9

**Table of Contents**

**1. Introduction**

***1.1 Background information and problem statement***

***1.2 Aims or objectives of the project***

**2. Data**

**3. Preprocessing**

**4. Data processing**

**5. Dimensionality Reduction**

**6. Machine Learning model selection**

**7. Results/ Discussion**

**8. Conclusion and Implications**

**9. References**

**1. Introduction**

***1.1 Background information and problem statement***

According to American Sleep Association (ASA), 50-70 million adults in the United States have a sleep disorder. Insomnia is the most common specific sleep order with sleep apnea, with short term issues reported by about 30% of adults and chronic insomnia by 10%. 25 Million US obstructive sleep apnea. 37% of 20-39 year-olds report short sleep duration, and 40% of 40-59 year-olds report short sleep duration.

Approximately 9% to 18% of adults in the United States meet criteria for chronic insomnia disorder, and up to one-third of a general population presents at least one of the insomnia criteria as defined by the DSM-IV [1][2]. Chronic insomnia is characterized by persistent difficulties initiating or maintaining sleep or non-restorative sleep accompanied by significant daytime impairments [3].

The exact mechanism that causes symptoms still remains obscure. According to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV-TR), the criteria for insomnia are difficulty initiating or maintaining sleep, that the disturbance does not occur during the course of breathing-related sleep disorder and another mental disorder, and the disturbance not is due to the direct physiological effects of a substance.

The exact mechanism that causes insomnia still needs elucidation, and so are the results revealed by sleep researchers. It was revealed that 14.7-30 H activity dropped more slowly on the sleep-onset period, and observed less 4 Hz activity in the persistent primary insomnia (PPI) group [4]. In a subsequent study, the same group of researchers compared EEG frequency for the first 4 REM and NREM periods in 119 PPI and 20 control subjects. Activity for all frequencies below 14.75Hz decreased and 14.75-30 Hz activity in NREM sleep in the PPI group increased. The results are further supported by another group of researchers, who compared 9 PPI subjects with 9 controls and 9 with major depression [5]. Data from the first 3 NREM periods were analyzed and showed evidence for greater relative power (power within a frequency band divided by the sum of power in all bands) of high-frequency activity (14-45 Hz) in the PPI group than in the other two groups.

Nonetheless, association of specific EEG frequency activity with insomnia has been challenged. Greater high-frequency activity (18-30 Hz) during waking and stage 1 from patients with PPI was found, but any differences from normals in Stage 2 was not clearly shown [7]. Nofzinger stated that there was no difference in high-frequency NREM EEG activity between 15 PPI subjects and 15 controls, but found greater 0.5-8 Hz activity in the PPI group. Furthermore, in 2013, the absence of significant differences between groups in waking or NREM power was revealed [8].

***1.2 Aims or objectives of the project***

***Broad objectives or goals of the project***

Our ultimate goal of the project is to develop a method to classify insomnia group individuals from controls, given a patient’s EEG dataset. We would like to use comprehensive data from the corresponding EEG channels, which include Fp2-F4, F4-C4, C4-P4, P4-O2, and C4-A1.

***Specific objectives or aims of the project***

Our specific objective for the project is to implement machine learning techniques to increase the accuracy of our classification program. Due to the lack of available online dataset, we thoroughly analyzed the EEG dataset with caution and created hypothetical sets to enable machine learning.

**2. Data**

Credibility for each dataset was thoroughly evaluated. The majority of the datasets are from PhysioNet, which is supported by the National Institute of General Medical Sciences (NIGMS) and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number 2R01GM104987-09.

We used the dataset from h[ttps://physionet.org/pn6/capslpdb/](https://physionet.org/pn6/capslpdb/)

The CAP Sleep Database is a collection of 108 polysomnographic recordings registered at the Sleep Disorders Center of the Ospedale Maggiore of Parma, Italy. The waveforms include at least 3 EEG channels, 2 EOG channels, EMG of the submentalis muscle, bilateral anterior tibial EMG, respiration signals, and EKG.  
  
 The 16 healthy subjects included in the study did not present any neurological disorders and were free of drugs affecting the central nervous system. The 92 pathological recordings include 40 recordings of patients diagnosed with NFLE, 22 affected by RBD, 10 with PLM, 9 insomniac, 5 narcoleptic, 4 affected by SDB and 2 by bruxism [9].

As our primary concern is accurate classification of insomnia, we selected EEG datasets from individuals that share certain EEG channels: Fp2-F4, F4-C4, C4-P4, P4-O2, and C4-A1. Thus, ins1, ins2, ins3, ins4, ins5, ins6, ins7, ins8, ins9, n2, n3, n4, n5, n11 were initially chosen.

Ins4, ins7, and n16 were dropped due to bad Independent Component Analysis (ICA) performance; we invested more than two hours separating the EEG signals, but the dataset did not satisfy our threshold value for Pearson correlation value of 0.8, which was used for ICA.

N4 was dropped as the dataset did not include ROC-LOC (EOG channel), and EMG1-EMG2 (EMG channel).

**3. Preprocessing**

Channel selection was done with EEGLAB and EDFBrowser. We assumed the five channels: Fp2-F4, F4-C4, C4-P4, P4-O2, and C4-A1 to be located on top of each channel list, and removed other irrelevant EEG channels. No modification other than channel selection was made. We ran the main script in python to run the signal process, and change inputs to fit data.   
 ***Signal preprocess package workflow in python:***

1. Process annotation and header files  
2. Process the record to contain the channels of interest:

Fp2-F4, F4-C4, C4-P4, P4-O2, and C4-A1

3. Filter the record using notch filter and low pass filter  
4. Remove artifact by ICA and export the processed data to MATLAB

***Feature extraction package workflow in MATLAB and python:***

1. Run Time frequency analysis: Fast Fourier Transform (FFT)

2. Extract minimum, maximum, and mean values of FFT for five different EEG frequency bands: delta [0.5 3.5], low [0.5 14], alpha [8.5 12], sigma [12.5 16], and beta [16 30] [10].

3. One to one mapping: add sleep stage label based on hypnogram to the output of feature extraction step 2.

4. Take average of values that belong to same sleep stage, such as beta wave that belongs to NREM sleep stages.

***Filtering and artifact removal:***

Channels that were not EEG, ECG, and EOG, and power-line noise was removed with notch filtering: American AC power at 60Hz, and European at 50Hz.

The matrix was transposed for FastICA. EOG, ECG, and EMG channels were determined with Pearson correlation. With inverse of mixing matrix, the artifact channels were zeroed-out. Corrected data were added to the original raw signal. ICA was repeated until the r values we set were not satisfied. Our threshold for r value was 0.8.

**4. Data Processing**

Through the investigation on the feature values, data points were observed to be not equally scaled for the same feature for various patients. Hence, we decided to scale them in same units to prevent the high variance between the values. After normalization of the data and scaling on the feature vectors to set in same units, we used existing data to build multivariate distribution using the means and covariances within the data to generate simulated dataset due to the limitation of our available real dataset. The generated data points from the multivariate distribution contribute to the hypothetical dataset we use further for training and test for our machine learning algorithm. Hypothetical set was generated with 1000 insomnia patients and 1000 normal patients. The function we used for generating hypothetical set was “FakeData.R” in OpenMx for R programmers.

**5. Dimensionality Reduction**

Principal component analysis method is an orthogonal transformation method on observations that are possibly correlated to convert into linearly uncorrelated feature vectors [15]. Hence, the principal components attained from this method enables dimensionality reduction in high dimensional feature space that helps prevent overfitting and generates independence among the newly generated feature components. Each component has the highest variance, which is called the explained variance and through the elbow’s method, we figure the components where the marginal increment in explained variance diminishes.



Figure 1. Cumulative explained variance for number of components

The graph indicates that about 99.7% of the explained variance is achieved from 4 components.

**6. Machine Learning model selection**

***Machine learning model selection:***

The labels or outputs of are binary in that the patient is either normal or insomnia. Hence, we decided to implement general and popular machine learning techniques adaptive to binary classification problem following:

1. **Binary Logistic Regression:**

Binary logistic regression is well suited probabilistic model for dichotomous outputs in that it is well applied to classify binary outputs, for we are unsure of the linearity between independent and dependent variables. Through PCA, we made the feature space to be linearly independent, thus it is valid to implement the binary logistic regression.

1. **Linear SVM:**

SVMs are generalized linear classifiers and treated as advanced version of perceptron algorithm [13]. With support vectors and margins, SVM is applied to more flexibly incorporate the data points, hence we decided to implement SVM on our dataset with thought on SVM will perform well applicable to binary classification. For our dataset, among the various kernels, such as polynomial and RBF, we chose to use the linear SVM since we figured our PCA version feature space is not high dimensional and there is no need for nonlinear transformation to deal with high dimensional feature space. Thus, we chose linear SVM for binary classification model. For the model parameters, we applied with the l2 penalty and hinge loss because it’s a classification model. Also, tolerance is determined to 0.0001.

1. **Adaboost:**

Adaboost is an advanced ensemble technique that learns an accurate strong classifier by combining an ensemble of inaccurate weak learners. Weak learners would be thought as the weak hypotheses that classify the data points, and they are eventually built up as a strong hypothesis. The ensemble of weak learners is combined into the strong classifier through the weighted majority vote and the algorithm selects the most pivotal features and give different weights for them. Among the ensemble methods, we selected Adaboost method for its robustness to overfitting and advanced generalization property with important feature selection. Since we focused on binary classification on our route options from relatively independent features we have, iterations over weak classifiers and updated weights on features to build the strong classifier in Adaboost method is thought to be reasonable boosting method.

**10-fold cross validation:**

In 10-fold cross validations, the data is partitioned into 10 equally sized subsamples. For the 10 iterations, each subsample becomes a validation set for testing the model and rest data are used as the training data. After the 10 iterations, cross validation errors for 10 iterations are averaged to single estimation of error and it has less bias and more accurate on testing the model than single iteration with testing and training set.

**7. Results/ Discussion**

After dimensionality reduction with PCA and 10 fold cross validation, we obtained the mean cross validation accuracy using Python Scikit-Learn library.

The first model was binary logistic regression, which produces below results:

1. Dataset with alpha wave

* PCA and 10 cross validation result:

|  |  |
| --- | --- |
| Accuracy | 0.50 (+/- 0.00) |

1. Dataset without alpha wave

* PCA and 10 cross validation result:

|  |  |
| --- | --- |
| Accuracy | 0.97 (+/- 0.02) |

To check the performance of our model, Receiver Operating Characteristic(ROC) curve and Area Under the Curve(AUC) were measured:



Figure 2. ROC curve and AUC value for Logistic Regression

The next model was Linear SVM, which produces below results:

1. Dataset with alpha wave

* PCA and 10 cross validation result:

|  |  |
| --- | --- |
| Accuracy | 0.50 (+/- 0.00) |

1. Dataset without alpha wave

* PCA and 10 cross validation result:

|  |  |
| --- | --- |
| Accuracy | 0.98 (+/- 0.02) |

To check the performance of our model, Receiver Operating Characteristic(ROC) curve and Area Under the Curve(AUC) were measured:



Figure 3. ROC curve and AUC value for Linear SVM

Our final model was Adaboost, which produces below results,

1. Dataset with alpha wave
   * PCA and 10 cross validation result:

|  |  |
| --- | --- |
| Accuracy | 0.50 (+/- 0.00) |

1. Dataset without alpha wave
   * PCA and 10 cross validation result:

|  |  |
| --- | --- |
| Accuracy | 0.85 (+/- 0.04) |

To check the performance of our model, Receiver Operating Characteristic(ROC) curve and Area Under the Curve(AUC) were measured:



Figure 4. ROC curve and AUC value for Adaboost

From the observations on the results from three different models and ROC curves, logistic regression and linear SVM show strength in our dataset in binary classification. Adaboost also accurately classify our data points but since it is more aware of the outliers that being sensitive to some noise date could make it overfit in test set, thus it is reasonable that it is less efficient than two other models. With the independent features from PCA were well applicable for logistic regression and the accuracy was favorable with almost 99 percent. Linear SVM also flexibly classify our data points in that the classification result was outstanding.

Also, as we compare the datasets with Alpha wave and without alpha, it was figured that there are significant effect from absence of alpha values that increase the overall accuracy and existence of alpha values decreases the overall accuracy; thus, determining factors that affect classification on insomnia patients are based on beta and low wave values in features.

**8. Conclusion and Implications**

Our 97% classification accuracy agrees with the past literature that activity for all frequencies below 14.75Hz decreased and 14.75-30 Hz activity in NREM sleep in the PPI group increased and that high frequency activity (14-45 Hz) in the PPI group from the first 3 NREM periods showed greater relative power [4][5].

Our results that demonstrate that alpha brain wave that belongs to NREM 2 sleep stage indeed are surprising as it lowered the overall classification accuracy to 50%. Our set range for alpha brain wave was [8.5 12], and the frequency band [10 12] has been known to associate with bursts of neural oscillatory activity, sleep spindle.

Sleep spindles generated in the thalamus have been shown to aid sleeping in the presence of disruptive external sounds. A correlation has been found between the amount of brainwave activity in the thalamus and a sleeper's ability to maintain tranquility [12].

How alpha brain wave that belongs to NREM 2 sleep stage associated with sleep spindles lowered the overall classification accuracy still needs further elucidation. It is obscure whether the strong correlation between low [0.5 14] and beta [8.5 12] brain waves from the first 3 NREM sleep stages eliminated alpha brain wave in NREM 2 sleep stage, or whether the alpha brain wave that includes sleep spindle range dilutes the overall classification accuracy.

**References**

[1] Leger D, Scheuermaier K, Raffray T, Metlaine A, Choudat D, Guilleminault C. HD-16: a new quality of life instrument specifically designed for insomnia. Sleep Med 2005;6:191-8.

[2] Ohayon MM. Epidemiology of insomnia: What we know and what we still need to learn. Sleep Med Rev 2002;6:97-111.

[3] J.D. Edinger, M.H. Bonnet, R.R. Bootzin, K. Doghramji, C.M. Dorsey, C.A. Espie, A.O.Jamieson, W.V. McCall, C.M. Morin, E.J. Stepanski: Derivation of research diagnostic criteria for insomnia: report of an American Academy of Sleep Medicine Work Group

Sleep, 27 (2004), pp. 1567-1596

[4] Merica, H., Blois, R. and Gaillard, J. M. Spectral characteristics of sleep EEG in chronic insomnia. Eur. J. Neurosci., 1998, 10: 1826–1834.

[5] Perlis ML, Smith MT, Andrew PJ, Orff H, Giles DE. Beta/gamma EEG activity in patients with primary and secondary insomnia and good sleeper controls. Sleep. 2001;24:110-117.

[6] Kai Spiegelhalder, Wolfram Regen, Bernd Feige, Johannes Holz, Hannah Piosczyk, Chiara Baglioni, Dieter Riemann, Christoph Nissen: Increased EEG sigma and beta power during NREM sleep in primary insomnia

Biological Psychology, Volume 91, Issue 3, 2012, Pages 329-333, ISSN 0301-0511,

[7] Freedman, R. R. EEG power spectra in sleep-onset insomnia. Electroencephalogr. Clin. Neurophysiol., 1986, 63: 408–413.

[8] Wu YM, Pietrone R, Cashmere JD, et al. EEG Power During Waking and NREM Sleep in Primary Insomnia. *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*. 2013;9(10):1031-1037. doi:10.5664/jcsm.3076.

[9] MG Terzano, L Parrino, A Sherieri, R Chervin, S Chokroverty, C Guilleminault, M Hirshkowitz, M Mahowald, H Moldofsky, A Rosa, R Thomas, A Walters. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. Sleep Med 2001 Nov; 2(6):537-553.

[10] 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. Circulation 101(23):e215-e220 [Circulation Electronic Pages; <http://circ.ahajournals.org/cgi/content/full/101/23/e215>]; 2000 (June 13).

[11] Krystal, A. D., Edinger, J. D., Wohlgemuth, W. K., & Marsh, G. R. (2002). NREM sleep EEG frequency spectral correlates of sleep complaints in primary insomnia subtypes. *Sleep: Journal of Sleep and Sleep Disorders Research, 25*(6), 626-636.

[12] Thien Thanh Dang-Vu, Scott M. McKinney, Orfeu M. Buxton, Jo M. Solet, Jeffrey M. Ellenbogen. Spontaneous brain rhythms predict sleep stability in the face of noise. Current Biology - 10 August 2010 (Vol. 20, Issue 15, pp. R626-R627)

[13] S.B. Kotsiantis. (2007) “Supervised machine learning: a review of classification techniques”

[14] Divya Ramani, Harshita Kanani, Chirag Pandya. (2013) “Ensemble of classifiers based on association rule mining”

[15] [Hervé Abdi](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorStored=Abdi%2C+Herv%C3%A9), Lynne J.Williams. (2010) “Principal Component Analysis”