Automatic Detection of Target Regions of Respiratory Effort-Related Arousals Using Recurrent Neural Networks

Heiðar Már Þráinsson*, Hanna Ragnarsdóttir*, Guðni Fannar Kristjánsson, Bragi Marinósson, Eysteinn Finnsson, Eysteinn Gunnlaugsson, Sigurður Ægir Jónsson, Jón Skírnir Ágústsson, Halla Helgadóttir

Nox Research, Nox Medical ehf, Reykjavík, Iceland

*These authors contributed equally

Abstract

We present a method for classifying target sleep arousal regions of polysomnographies. Time- and frequency-domain features of clinical and statistical origins were derived from the polysomnography signals and the features fed into a Bidirectional Recurrent Neural Network, using Long Short-Term Memory units (BRNN-LSTM). The predictions of five recurrent neural networks, trained using different features and training sets, were averaged for each sample, to yield a more robust classifier. The proposed method was developed and validated on the PhysioNet Challenge dataset which consisted of a training set of 994 subjects and a hidden test set of 989 subjects. Five-fold cross-validation on the training set resulted in an area under precision-recall curve (AUPRC) score of 0.452, an area under receiver operating characteristic curve (AUROC) score of 0.901 and intraclass correlation ICC(2,1) of 0.59. The classifier was further validated on the PhysioNet Challenge test set, resulting in an AUPRC score of 0.45.

1. Introduction

In the scoring manual by the American Academy of Sleep Medicine (AASM), arousals are defined as abrupt shifts of electroencephalography (EEG) frequency that last at least 3 seconds, with at least 10 seconds of previous stable sleep [1]. Arousals can occur spontaneously or as a result of sleep-disordered breathing or other sleep disorders [2]. Respiratory Event Related Arousals (RERA) are arousals that are caused by sequences of breaths lasting more than 10 seconds characterized by increasing respiratory effort [1]. The identification of arousals is important for the evaluation of sleep continuity and for diagnosis of various sleep disorders [3].

Manual scoring of these events is costly due to the huge amount of data recorded per night, and difficult due to variance across patients and technicians experience [4] [5]. Automation of the detection procedure is therefore important and different works have explored different ways of automating the process. Alvarez-Estevez and Moret-Bonillo [6] developed a method for the automatic detection of EEG arousals using two EEG channels and electromyography (EMG). Experiments conducted on 20 patients reported a sensitivity and specificity respectively of 0.86 and 0.76. Behera et al. [7] followed the study, adding more features to the input of an artificial neural network, and combining different models. Experiments conducted on 26 patients reported a sensitivity of 0.81 and a specificity of 0.88 with an error of 0.13. More recently, Isaac Fernández-Varela et al. [8] combined various signal analysis solutions to identify relevant arousal patterns with special emphasis on robustness and artifact tolerance. Experiments conducted on 22 patients reported precision of 0.86 and F1 score of 0.79. However, all of these methods were developed on datasets containing relatively few subjects and may not generalize well across different populations.

We propose a recurrent neural network-based approach for classifying target sleep arousal regions, using full polysomnography recordings. The algorithm was trained and tested on the PhysioNet Challenge 2018 database, which includes 1985 subjects [9]. The results are thus based on a larger dataset than previous methods. The method was implemented using Keras 2.1.5, using Tensorflow 1.8.0 backend. The code was submitted for the Open-Source Challenge call of the PhysioNet Challenge 2018 [9].

2. Methods

We employed a three layer neural network. The first hidden layer is a BRNN and the second hidden layer is a dense neural network. Time- and frequency domain features were derived from multiple available signal including

 ISSN: 2325-887X DOI: 10.22489/CinC.2018.126

the EEG, ECG and respiratory signals. The features were fed into the neural network, which after training outputs the probability that a given segment is a target arousal region.

2.1. Feature extraction

For each subject a variety of biometric signals, relevant to sleep studies, were recorded. EEG recordings were made in the following configurations: F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, O2-M1. Additionally the left eye electrooculogram was recorded using the E1-M2 configuration. Signals relating to cardiorespiratory activity were recorded and were as follows: EMG recordings made at the chin, chest and abdomen; oxygen saturation (SaO2), airflow, and electrocardiogram (ECG). Features were extracted from all relevant signals with different signals requiring unique processing methods. All features were calculated over a 10 second sliding window with 50% overlap unless otherwise specified.

2.1.1. EEG features

For each EEG signal, various frequency and time domain features were extracted. The signals were decomposed into sub-bands using the wavelet packet decomposition (WPD). The Daubechies 4 wavelet has been shown to perform well in EEG feature extraction [10] and was used to decompose the signal down to the 4th level, resulting in sub-bands of 6.25 Hz resolution. For each sub-band, statistical features were calculated, as well as sub-band energy. Additionally, the Hjorth parameters were calculated for the signals. These parameters are Hjorth activity, mobility, and complexity and represent signal power, mean frequency, and change in frequency respectively [11].

2.1.2. Respiratory features

Features were extracted from the respiratory signals which could indicate respiratory disturbance. Statistical features calculated from the SaO2 signal indicate changes in oxygen saturation which correlate with apnea [12]. Since the characteristics of the airflow, chest, and abdomen signals vary between individuals, the statistical features give information about changes in respiratory activity. Correlation between abdomen and thorax signals was calculated to detect when the two signals go out of phase, which is an indicator of obstructive apnea [13].

2.1.3. ECG features

For the ECG signal we derived various features relating to the heart rate. The QRS complexes of the ECG signals were detected using a robust R-peak finder [14]. From the locations of the R-peaks, the heart rate and Heart Rate Variability (HRV) signals were calculated. Statistical features were calculated from the heart rate while more complex frequency domain features were derived from the HRV. The power spectrum of the HRV is an important indicator of the function of the nervous system and has been shown to be a good indicator of apneas [15]. The HRV signal was interpolated using cubic spline interpolation to get a signal of constant sampling frequency. A spectrogram of the HRV was then calculated using Welch's method with a sliding Hamming window. This was done using windows of 5 minute and 30 minute duration to capture the short and long term dynamics of the HRV. For both windows a stride of 5 seconds was used. The frequency bands of interest are the very low frequency (VLF) 0.003 - 0.04 Hz, low frequency (LF) 0.04 - 0.15 Hz and high frequency (HF) 0.15 - 0.4 Hz [16]. For each band we calculated the normalized total energy, peak energy and peak frequency as well as the ratio of LF and HF power.

2.2. Classification

Recurrent neural networks, using LSTM hidden units, are powerful models for learning from sequential data since they are capable of remembering information for a long period of time. Bidirectional recurrent neural networks can further learn from both past and future states, which is important when context of the input is needed, such as when detecting sleep arousals [17]. We thus considered LSTM-based BRNN model for the sleep arousal detection.

2.2.1. Data preparation

After feature extraction, the data was reshaped for the BRNN-LSTM layer into a three-dimensional array, where the three dimensions are:

- Number of training sequences, N
- Sequence length (number of time-steps), W
- Number of features of each sequence, F

By experimenting with different values for the sequence length and different positions of the label, we found $W = 20$ to result in the best performance, positioning the label at time step 11. The time-steps were composed of features extracted over a 10 second window with a 5 second overlap. Thus, each sequence considered by the classifier was $20 \cdot (10 \cdot 0.5) = 100$ seconds long, with the neural network looking 50 seconds in the past and 40 seconds in the future.

Regions in the training dataset labeled neither as normal regions nor target arousal regions were ignored, as those regions are not considered in the PhysioNet Challenge. The remaining training dataset is unbalanced, with 7% of

the data being arousal regions and 93% being normal sleep regions. To achieve a more balanced training dataset we randomly removed 90% of the normal sleep regions.

2.2.2. Sequence Classifiers

We experimented with several model structures and hyperparameters. The model that performed the best was a three-layer model, where the first hidden layer is a LSTM-based BRNN layer consisting of 50 LSTM blocks, and the second hidden layer is a dense layer consisting of 50 nodes. We use a softmax activation function on the output layer to extract probabilities for classification. To combat overfitting, dropout is applied to the output of both hidden layers, and all layers have l2-kernel regulizer of strength 0.01 to further combat overfitting [18]. The neural network was trained using a batch size of 200, learning rate was reduced on plateau and early stopping was used. The loss function used was binary cross-entropy and the optimizer was Adam.

2.2.3. Ensemble Classifier

Five classifiers of the same structure as described above were trained on different subsets of the training data and using different sets of features. The predictions of these classifiers were then averaged per sample, to create a more robust classifier and to reduce variance arising from the random initialization of the weights and the random split between train and validation set. The final ensemble classifier thus consisted of five classifiers, each trained on all the respiratory features, but with different set of two to three EEG and ECG features.

3. Results and Discussion

In this section we evaluate the performance of our method for classifying target arousal regions, and furthermore compare the importance of different feature groups.

3.1. Model Validation

To evaluate the performance of our method, we performed a 5-fold cross-validation on the training dataset. During each cross validation fold, 20% of the available training data was set aside for final testing. The other 80% were used for the training and validation of the models during the cross validation. For each model the validation set was randomly selected containing 10% of the training and validation data.

According to the PhysioNet Challenge scoring system, results are reported as gross AUPRC score and AUROC score [9], however only the gross AUPRC score is used to rank competitors in the PhysioNet Challenge. The cross validated scores of the individual classifiers, as well as the ensemble classifier, are shown in table 1. The ensemble classifier gave the best results, its performance was higher than any of the individual classifiers. The method was verified using the hidden test set of the PhysioNet Challenge. It achieved an AUPRC score of 0.45, which places it second in the competition.

Table 1. Cross validated AUPRC and AUROC scores of the individual models as well the final ensemble

Model	AUPRC		AUROC	
	Mean	STD	Mean	STD
Model 1	0.432	0.037	0.893	0.0026
Model 2	0.429	0.035	0.893	0.0027
Model 3	0.426	0.038	0.891	0.0030
Model 4	0.430	0.040	0.893	0.0020
Model 5	0.428	0.032	0.895	0.0030
Ensemble model	0.452	0.038	0.901	0.0030

The performance of the classifier was further analyzed by comparing the arousal indices of the manual and automatic annotations. The arousal index for the automatic classifier was calculated by setting the prediction threshold as 0.82 and removing all predicted arousals in unscored regions. The intraclass correlation ICC(2,1) [19] was calculated for the arousal indices and resulted in a cross-validated score of 0.59. This value is within the reported range of intraclass correlation between human scorers (0.50-0.85), but lower than the reported average (0.68) [5].

3.2. Feature Importance

Calculating feature importance with recurrent neural networks is not straight forward, as standard feature importance calculations do not take into account the temporal attribute of the RNNs. To get an estimation of the importance of the features, we trained our classifier using three different groups of features and compared the cross validated AUPRC scores. The feature groups compared are features derived from EEG signals, features derived from ECG signals and features derived from respiratory signals. The respiratory features performed best with an AUPRC score of 0.41, suggesting they are most important in detecting arousals. The EEG signals resulted in an AUPRC score of 0.27 and the ECG signals performed the worst with an AUPRC score of 0.20. This was to be expected since the majority of the scored arousals were RERAs.

4. Conclusion

The problem of automatically detecting sleep arousals is not a trivial one and more work remains to be done.

However, being able to effectively score arousals automatically is important, as manual scoring of arousals is time consuming and difficult. In this paper we have proposed a method for classifying target sleep arousal regions, using a BRNN-LSTM ensemble model. The method was validated on PhysioNet Challenge 2018 dataset and the results are encouraging, suggesting that the automatic classification of arousals is an achievable task. We intend to further develop our method for clinical application, by generalizing it for a different dataset and improving efficiency with feature selection and code optimization.

Acknowledgements

This work was supported by the Icelandic Centre for Research under the Icelandic Student Innovation Fund and the Horizon 2020 SME Instrument, project number 733461.

References

[1] Berry RB, Albertario CL, Harding SM, Lloyd RM, Plante DT, Quan SF, Troester MM, Vaughn BV. The AASM Manual for the Scoring of Sleep and Associated Events; Rules, Terminology and Technical Specifications, version 2.5. American Academy of Sleep Medicine, 2018.

[2] Bonnet M, Carley D, Carskadon M, Easton P, Guilleminault C, Harper R, Hayes B, Hirshkowitz M, Ktonas P, Keenan S, Pressman M, Roehrs T, Smith J, Walsh J, Weber S, Westbrook P, Jordan B. EEG arousals: Scoring rules and examples. a preliminary report from the Sleep Disorders Atlas Task Force of the American Sleep Disorder Association Jan 1992;15:173–184.

[3] American Academy of Sleep Medicine. International Classification of Sleep Disorders. American Academy of Sleep Medicine, 2014.

[4] Bonnet MH, Doghramji K, Roehrs T, Stepanski EJ, Sheldon SH, Walters AS, Wise M, Chesson AL. The scoring of arousal in sleep: Reliability, validity, and alternatives. Journal of Clinical Sleep Medicine 2007;03(2):133–145.

[5] Magalang UJ, Chen NH, Cistulli PA, Fedson AC, Gíslason T, Hillman D, Penzel T, Tamisier R, Tufik S, Phillips G, Pack AI, for the SAGIC Investigators. Agreement in the scoring of respiratory events and sleep among international sleep centers. Sleep 2013;36(4):591–596.

[6] Álvarez Estévez D, Moret-Bonillo V. Identification of electroencephalographic arousals in multichannel sleep recordings. IEEE Transactions on Biomedical Engineering Jan 2011;58(1):54–63.

[7] Behera CK, Reddy TK, Behera L, Bhattacarya B. Artificial neural network based arousal detection from sleep electroencephalogram data. 2014 International Conference on Computer Communications and Control Technology I4CT Sept 2014;458–462.

[8] Moret-Bonillo V, Fernández-Varela I, Hernández-Pereira E, Alvarez-Estévez D, Perlitz V. On The Automation of Medical Knowledge and Medical Decision Support Systems. Computers in biology and medicine 2018;87:77–86.

[9] Ghassemi MM, Moody BE, Lehman LH, Song C, Li Q, Sun H, Mark RG, Westover MB, Clifford GD. You Snooze, You Win: the PhysioNet/Computing in Cardiology Challenge 2018. In Computing in Cardiology, volume 45. Maastricht, Netherlands, 2018; 1–4.

[10] Faust O, Acharya UR, Adeli H, Adeli A. Wavelet-based EEG processing for computer-aided seizure detection and epilepsy diagnosis. Seizure 2015;26:56–64.

[11] Hjorth B. EEG analysis based on time domain properties. Electroencephalography and Clinical Neurophysiology 1970;29(3):306–310.

[12] Alvarez D, Hornero R, Marcos JV, del Campo F. Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis. IEEE Transactions on Biomedical Engineering Dec 2010;57(12):2816–2824.

[13] Berry RB, Budhiraja R, Gottlieb DJ, Gozal D, Iber C, Kapur VK, Marcus CL, Mehra R, Parthasarathy S, Quan SF, Redline S, Strohl KP, Ward SLD, Tangredi MM. Rules for scoring respiratory events in sleep: Update of the 2007 AASM manual for the scoring of sleep and associated events: Deliberations of the sleep apnea definitions task force of the american academy of sleep medicine. J Clin Sleep Med Oct 2012;8(5):597–619.

[14] Kathirvel P, Sabarimalai Manikandan M, Prasanna SRM, Soman KP. An efficient R-peak detection based on new nonlinear transformation and first-order gaussian differentiator. Cardiovascular Engineering and Technology Dec 2011;2(4):408–425.

[15] Penzel T, McNames J, de Chazal P, Raymond B, Murray A, Moody G. Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings. Medical and Biological Engineering and Computing Jul 2002;40(4):402–407.

[16] Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Frontiers in Public Health 2017;5:258.

[17] Lipton ZC. A critical review of recurrent neural networks for sequence learning. CoRR 2015;abs/1506.00019.

[18] Cogswell M, Ahmed F, Girshick RB, Zitnick L, Batra D. Reducing overfitting in deep networks by decorrelating representations. CoRR 2015;abs/1511.06068.

[19] E. Shrout P, L. Fleiss J. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin Apr 1979; 86:420–8.

Address for correspondence:

Nox Research, Nox Medical
Katrínartún 2, 105 Reykjavík
halla@noxmedical.com