ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Sleep Staging with Deep Learning: A convolutional model

Isaac Fernández-Varela[1], Dimitrios Athanasakis[2], Samuel Parsons[3]
Elena Hernández-Pereira[1], and Vicente Moret-Bonillo[1] *

1- Universidade da Coruña - Departamento de Computación
Facultade de Informática, Campus de Elviña, A Coruña - Spain

2- Data Spartan
60 Ludgate Hill,London EC4M 7AW, United Kingdom

3- University College of London - Department of Computer Science
66-72 Gower Street, London WC1E 6EA, United Kingdom

Abstract. Sleep staging is a crucial task in the context of sleep studies that involves the analysis of multiple signals, thus being a very tedious and complex task. Even for a trained expert, it can take several hours to annotate the signals recorded from a patient's sleep during a single night. To solve this problem several automatic methods have been developed, although most of them rely on hand engineered features. To address the inner problems of this approach, in this work we explore the possibility of solving this problem with a deep learning network that can self-learn the relevant features from the signals. Particularly, we propose a convolutional network, obtaining higher performance than in previous methods, achieving an average precision of 0.91, recall of 0.90, and F-1 score of 0.90.

1 Introduction

Among the main tasks within the medical analysis of the sleep stands out the characterization of the sleep macro structure. Its final goal is the construction of the hypnogram, a graph that helps to interpretate the recorded electrical activities during a polysomnogram (PSG), showing the evolution of the different sleep stages through time.

The construction of the hypnogram was first proposed by Rechtschaffen and Kales (R&K) [1] in 1968 and only recently updated by the American Academy of Sleep Medicine (AASM) [2]. The method establishes a set of rules to assign labels (sleep stages) to time intervals typically lasting 30 s and called epochs. These sleep stages are: wakefulness (W), two stages for drowsy sleep (N1 & N2), one deep sleep (N3), and Rapid Eye Movement (REM).

Sleep staging is a tedious task, very time-consuming because it implies the analysis of multiple signals that record several hours (at least 6), thus there is a need to do it automatically. Several works address this problem with different approaches but they suffer from the problem of using hand engineered features.

*This research was partially funded by the Xunta de Galicia (Grant code GRC2014/035, and ED431G/01) partially supported by the European Union ERDF and from the Xunta de Galicia and the European Union Social Fund ESF.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

The latest review can be found in Penzel and Conradt [3]. Recent works already solve this problem avoiding hand engineered features [4, 5].

This works classifies sleep stages automatically, avoiding the use of hand engineered features using multiple signals at the same time. We also avoid the use of filters of methods to remove artifacts from the signals, feeding the network with the raw signals. The convolutional network that we are proposing is able to learn the relevant features from the signals to classify the sleep stages.

2 Materials

To develop and validate our proposal we have used real PSG recordings from the Sleep Heart Health Scoring database [6]. This database emerged from a multi-center cohort study to determine cardiovascular and other consequences of sleep-disordered breathing. Each recording includes off-line experts annotations following the R&K procedure [1]. The montage includes two EEG derivations (C4A2 and C4A1), rigth and left electrooculograms (EOG), bipolar submental electromyogram (EMG); and other signals which are not relevant to our problem. The EEG, EOG and EMG signals were recorded at 125 Hz and the EOG signals at 50 Hz. All the signals were filtered with a high pass filter set at 0.15 Hz during their acquisition.

A total of 240 recordings from different patients were randomly selected, using 180 for training, 20 for validation and 40 for testing. From each recording from train and validation sets and just to ease the model implementation, only 6 random hours were used, giving a total of 144,000 samples. The test set, which has a total of 49,794 samples, was used completely. No effort was done to select recordings with low noise ratio nor to discard segments with artifacts, as the model should be able to adapt to these situations.

In the train dataset 39.7% of the samples are classified as Awake, 38.3% as Drowsy Sleep, 9.6% as Deep Sleep, and 12.4% as REM. In the validation dataset the class distribution is: 42.0% for W, 37.3% for DS, 9.1% for N3, and 11.6% for REM. Finally, in the test dataset the distribution is: 42.7% for W, 37.3% for DS, 8.8% for N3, and 11.2% for REM.

3 Method

The goal of this work is to classify the different sleep stages of a PSG recording. As our first approximation, we simplify the problem using one label for drowsy sleep which includes both N1 and N2 stages. This was done in previous works [7] given that N1 is the stage for which expert classification presents the lowest inter-agreement [8] and also, the one with lower presence (only 3% of the epochs are classified as N1).

We solve this classification problem using a convolutional network[1]. A convolutional network is a deep feed-forward network that overcomes the limitations of multilayer perceptrons using a shared-weights architecture. The main reason

[1]Code and model are available in https://github.com/bigsasi/deepsleep

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

to use this network is its ability to learn the features that before were hand engineered.

To avoid biasing the model, we use as much data as it is possible to train it. Thus, we use the five available signals, although they are sampled at different rates. To overcome this problem, those signals sampled with a lower rate were padded with zeros. Then, a matrix with a row per signal was created. Obviously, this was done for each 30 seconds window, following the sleep stage definition. This way, each input to the network has a dimension of 3750×5. Although the input is bi-dimensional, our experiments were all done using 1D convolutional networks. With 1D convolutions we avoid imposing some artificial spatial structure between the different signals. Each signal was normalized to zero mean and unit standard deviation using train set as reference.

The convolutional model was selected using the validation set, trying to obtain the smallest network. From one layer models, we kept adding more layers until the performance did not improve. The performance of the model was defined as the average classification recall in the validation set. This experimentation led towards the model represented in Figure 1, which is composed of the following layers: two convolutional layers each with 128 kernels, one pool layer, another convolutional layer with 256 kernel, a max pool layer, and a final fully connected layer.

The filter size was fixed at 20 for every convolutional layer, with padding adjusted to maintain the input dimension. This value was selected after trying values from 3 (recommended value for convolutional networks used in artificial vision) to 65 (which would cover half a second of our signals). We observed that the performance improved with the filter size, but only up to 20, and decaying afterwards. The gradient optimizer was Adam [9] (with learning rate $3e-4$) and the activation functions for all the convolutional layers *relu*, except for a final *softmax* function. We also added a dropout [10] of 0.5 in the final layer as regularization to avoid over-fitting. With this configuration the network had a total of 997,380 trainable parameters. Training was done with batches of size 32 and finished using early stopping over the validation loss with a patience value of 3.

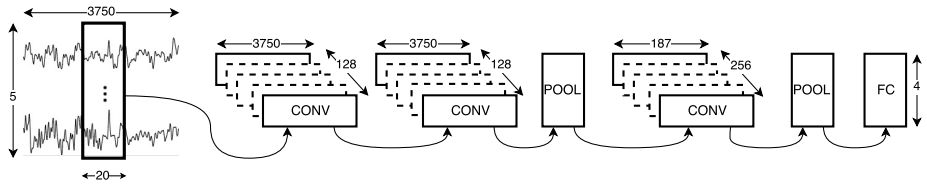

Fig. 1: Outline of the proposed convolutional network.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

4 Results

In order to validate the usefulness of the proposed approach an ensemble of 5 models was trained under the same dataset, but with random different 6 hours selection samples from each recording. To decide the final classification we use the mean value. Table 1 shows the global results obtained with our network and Table 2 the associated confusion matrix. The best precision value is obtained for the Awake class, achieving proximate values for the remaining classes. For the recall measure, it is worth mentioning the drop observed in the Deep Sleep stage.

	Precision	Recall	F-1 Score
Awake (W)	0.96	0.96	0.96
Drowsy Sleep (DS)	0.90	0.91	0.90
Deep Sleep (N3)	0.89	0.82	0.85
REM	0.89	0.90	0.90
Average	0.91	0.90	0.90

Table 1: Precision, recall and F-1 score for each sleep stage

	W	DS	N3	REM
W	20411	712	2	125
DS	741	16917	449	480
N3	1	796	3566	0
REM	152	392	0	5050

Table 2: Confusion matrix

Evaluating each recording individually, the distribution of the different measures is represented in Figure 2. This Figure confirms that the Awake class is the one with the best classification, with F-1 scores over 0.9, and and that the model struggles to classify Deep Sleep, with F-1 values falling to 0.4. Drowsy Sleep and REM classification present similar performance, with values over 0.8 for the F-1 score. Deep Sleep is the class with the highest deviation values, specially regarding the recall measure. Although precision is still greater than 0.8 for most of the records, recall values are worse, with measure values even as low as 0.3, which means that the model tends to underscore this class. Obviously, the fact is also reflected by the F-1 score, showing higher deviation than the other classes. In regards to outliers, the really low values (below 0.2) correspond to those recordings where the number of epochs classified (by the expert) as N3 is limited (lower than 2). Especially, the outliers represented as 0 correspond to a recording with no Deep Sleep epochs.

5 Conclusions

This work presents a method to classify sleep stages in PSG recordings using the common 30 seconds division called epoch. Our solution proposes the use of

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

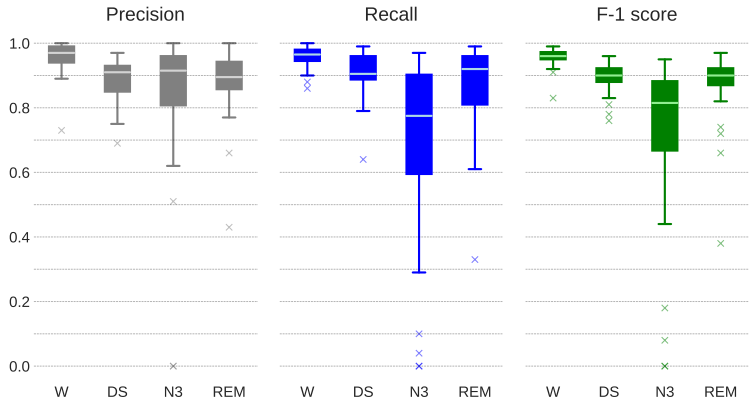

Fig. 2: Distribution of the performance measures for the individual recordings

a convolutional network that is feed with five available signals. The objective of this approach is to avoid human engineered features, which is how most works have solved this problem previously. The raw signals available in our dataset (2 EEG derivations, 1 EMG and 2 EOG) were directly used to feed the network. Thus, we also avoid the use of filters or signal preprocessing methods, apart from those applied within the hardware used to record the data.

Our model was selected using a validation dataset trying to achieve the highest recall with the fewer number of layers. Specifically, the architecture of our final model is composed of three convolutional layers with a pool layer after the second one and a fully connected layer at the end, with a total of 997,380 parameters trained. This yielded an average precision value of 0.91, a recall value of 0.90 and a F-1 score of 0.90.

It is difficult to compare our proposal against previous works due to the lack of benchmarks or clear methodology. Alvarez-Estevez et al. [7] presents a method using fuzzy logic after extracting hand engineered features from the signals. The method is validated on 26 recordings from the SHHS dataset, achieving an average recall value of 0.82, lower than the value obtained by the model described. For each sleep stage our method achieves between 7% (REM class) and 12% (DS class) higher recall. Längkvist et al. [11] avoid the use of hand engineered features with a deep belief network, although they remove noisy segments and select only those epochs with a clear label. Besides, they used a different dataset and classify 5 sleep stages. Assuming that their classification for drowsy sleep would be as good as it is for N2 stage (quite higher than for N1 stage), their average F1-score value is 0.79. Between classes, out method improved between 1% for the N3 class and 23% for the W class. Supratak et al. [4] uses multiples neural networks and a single EEG channel to classify with the Sleep-EDF dataset, achieving lower F-1 values compared to ours. Finally, Sors

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

et al. [5] achieved a similar F-1 for DS and Deep Sleep to ours, although lower in the remaining classes, using a deeper network and the same dataset. The performance measures for the aforementioned works are presented in Table 3

	Alvarez-Estevez et al. [7] (similar dataset, recall)	Längkvist et al. [11] (different dataset, F-1 Score)	Supratak et al. [4] (different dataset, F-1 Score)	Sors et al. [5] (similar dataset, recall)
Awake	0.88	0.78	0.85	0.91
Drowsy Sleep	0.81	0.37 (N1) 0.76 (N2)	0.47 (N1) 0.86 (N2)	0.35 (N1) 0.89 (N2)
Deep Sleep	0.75	0.84	0.85	0.85
REM	0.84	0.78	0.82	0.86
Average	0.82	0.79*	0.85*	0.88*

Table 3: Reported classification performance from previous works. * excluding performance for N1

In the light of the results, there is room for improvement. First of all, future models should include the five sleep classes as it is the standard nowadays. Besides, the easy adjustment to new datasets, which may use different signals or derivations, should be a quality of any model. Finally, actual trends in deep learning are understanding why the models perform the way they do. In this sense, to know which are the learned features or to evaluate the worse performance for the Deep Sleep class would be very valuable.

References

[1] Allan Rechtschaffen and Anthony Kales. A manual of standardized terminology, techniques, and scoring systems for sleep stages of human subjects. 1968.

[2] Richard B Berry et al. *The AASM manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.3*, volume 1. American Academy of Sleep Medicine, Westchester, IL, 2016.

[3] Thomas Penzel and Regina Conradt. Computer based sleep recording and analysis. *Sleep medicine reviews*, 4(2):131–148, 2000.

[4] Akara Supratak et al. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, nov 2017. ISSN 1534-4320. doi: 10.1109/TNSRE.2017.2721116.

[5] Arnaud Sors et al. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomedical Signal Processing and Control*, 42:107–114, apr 2018. ISSN 17468094. doi: 10.1016/j.bspc.2017.12.001.

[6] Stuart F Quan et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

[7] Alvarez-Estevez et al. On the continuous evaluation of the macrostructure of sleep. *Frontiers in Artificial Intelligence and Applications*, 243:189–198, 2012. ISSN 09226389. doi: 10.3233/978-1-61499-105-2-189.

[8] Heidi Danker-hopfe et al. Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard. *Journal of sleep research*, 18(1):74–84, 2009.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL `http://arxiv.org/abs/1412.6980`.

[10] Nitish Srivastava et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[11] Martin Längkvist, Lars Karlsson, and Amy Loutfi. Sleep stage classification using unsupervised feature learning. *Advances in Artificial Neural Systems*, 2012:5, 2012.