IDENTIFICATION OF SLEEP/WAKE STAGES IN ACTIGRAPHY DATA UTILISING GRADIENT BOOSTING ALGORITHM

Marek Mikulec
Doctoral Degree Programme (1st year), FEEC BUT
E-mail: xmikul56@vutbr.cz

Supervised by: Jiří Mekyska
E-mail: mekyska@feec.vutbr.cz

Abstract: Sleep disorders are early markers of various serious diseases that can be treated more effectively when diagnosed in their prodromal stage. Actigraphy is a noninvasive sleep monitoring method for the detection of sleep patterns and determination of sleep parameters that could support the diagnosis of these disorders. This study aims to compare a newly proposed actigraphy-based method of sleep/wake detection with a conventional one in terms of consistency with a polysomnography (PSG) reference. 55 recordings (acquired in 28 subjects) of actigraphy and PSG were modelled by a heuristics-based method and by a new approach utilising a gradient boosting algorithm. In addition, another database (22 subjects, 150 recordings) was used to compare scores of the new method with data reported in sleep diaries. The proposed method achieves 89% accuracy and Mathews correlation coefficient equal to 0.75 when compared to the polysomnography reference. Such results outperform the ones provided by the heuristic technique. The newly proposed method has good consistency with the PSG reference, thus being a good alternative to the golden standard in sleep disorders assessment, especially in decentralised clinical trials.

Keywords: actigraphy, machine learning, polysomnography, sleep, sleep diary

1 INTRODUCTION

Quality of sleep is an essential aspect of a healthy life. Sleep disorders are early indicators of various diseases (e.g. Parkinson’s disease, dementia with Lewy bodies [1]) that can be treated much more effectively if they are diagnosed in the prodromal stage, i.e. the stage wherein early symptoms or signs of the disease are present, but a classic clinical diagnosis is not yet possible. The group of sleep disorders includes, for example, insomnia, restless legs syndrome, idiopathic rapid eye movement sleep behaviour disorder (iRBD), parasomnia, or sleep apnoe [1, 2].

Sleep monitoring via actigraphy is a noninvasive monitoring method that records the occurrence and the degree of limb movement. Actigraph is a bracelet mostly worn on the wrist, ankle, or waist. Despite the movement, some actigraph devices additionally track modalities such as temperature, blood pressure, light intensity, etc. The actigraphy data, expressed in the form of time series, are consequently processed using statistical or machine learning methods for the sake of detection of sleep patterns and determination of sleep parameters. Such measures could be digital biomarkers of the sleep disorders mentioned above [3, 4, 5].

Nowadays, polysomnography (PSG) is considered a gold standard measure of sleep. PSG is a systematic procedure that utilizes electroencephalogram (EEG), electrooculogram (EOG), electrocardiogram (EKG), pulse oximetry, and pulse and respiratory effort [2]. The results of polysomnography are precise and reliable, but they depend on the use of expensive specialized equipment, which is not comfortable for a subject. Moreover, the subject’s sleep might be affected by an unusual laboratory environment [4, 6]. Actigraphy is not a replacement for PSG, but it could be a valuable complement
that finds its place in decentralised clinical trials. Nevertheless, the outcomes of this technology highly depend on correctly detected time windows, where the subject is asleep/awake. For this purpose, several approaches have been proposed [3]. This study aims to compare a newly proposed method of sleep/wake detection with a conventional one in terms of consistency with a PSG reference.

2 MATERIALS AND METHODS

2.1 DATABASE

For this study, we used the Newcastle polysomnography and accelerometer dataset that contains 28 adult patients who were scheduled for a one night PSG assessment while wearing the GENEActiv actigraph (Activinsights Ltd, Kimbolton, UK) with a tri-axial accelerometer (sampling frequency $f_s = 87.5$ Hz) on both wrists [7, 8]. In total, 55 recordings are available (data from the right wrist of one participant are missing).

In addition, the newly proposed method was validated in a database consisting of 22 subjects, where each one wore the actigraph on the non-dominant wrist for 6–7 nights (the database contains 150 recordings). Besides, the participants filled in sleep diaries. They were enrolled at the St. Annes University Hospital Brno. All of them signed the informed consent. The study was approved by the local ethics committee.

2.2 METHOD 1

Method 1 (M1) is a numeric method based on an estimation of the arm angle relative to the horizontal plane. The arm angle was estimated using equation:

$$angle = \left( \tan^{-1} \frac{a_z}{\sqrt{a_x^2 + a_y^2}} \right) \frac{180}{\pi}, \quad (1)$$

where $a_x$, $a_y$, and $a_z$ are the median values of the orthogonally positioned raw acceleration sensors in g-unit derived based on a rolling five-second time window. In the next step, the algorithm identifies periods larger than 5 minutes with the $angle$ change smaller than 10° – these intervals are marked as sleep periods [7]. Results were replicated with 3-minute and 10-minute time windows and with 3° and 5° threshold, respectively.

M1 was precisely validated by V. T. van Hees et al. in [7] on a big Whitehall II study dataset (27981 nights from 4094 participants). In addition, M1 was validated against the Newcastle polysomnography and accelerometer dataset as well [7, 8]. Therefore, we will use M1 as a benchmark for our newly proposed method M2 [7].

2.3 METHOD 2

Method 2 (M2) is our novel method based on a supervised machine learning algorithm. In comparison to M1, besides the accelerometer time series, the proposed method uses data from the temperature sensor as well. M2 consists of several steps. Firstly, the signals are decimated to $f_s = 28.5$ Hz and divided into 30-second segments. Next, the data from the accelerometer are converted into a magnitude representation:

$$magnitude = \sqrt{x^2 + y^2 + z^2}. \quad (2)$$

Consequently, 45 features are extracted from each segment of the acceleration and temperature series, i.e., in total 90 measures per each 30 s time window. To see the whole set of features, we refer to [4, pp. 73–74].
Finally, the features are modelled by XGBoost, which is a state-of-the-art scalable tree boosting classifier. It is a sparsity aware algorithm, theoretically justifying the weighted quantile sketch for approximate learning [9]. The Newcastle polysomnography and accelerometer dataset is significantly smaller than the Whitehall II study database ([8]), therefore the model was trained and its hyper-parameters were optimised employing k-fold cross-validation ($k = 10$). The PSG data were used as targets. In addition, 60% of the total number of data entries were used for training/validation and 40% for testing.

2.4 Evaluation Metrics

Both methods were evaluated in terms of accuracy, sensitivity, specificity, $F_1$ score, and the Mathews correlation coefficient (MCC) (positive cases were used for sleep and negative cases for wake). To evaluate consistency with subject-reported data, the second method was also validated with respect to the information reported in the sleep diaries of the second database.

3 Results

Results based on M1 differ from those reported in [7]. The most balanced results, where $F_1 = 73\%$, were achieved using a 5-minute time window and $10^\circ$ threshold. The highest accuracy (83%) was based on a 3-minute time window and $10^\circ$ threshold1.

During the optimisation phase in M2, we identified the following hyperparameters of the model: the number of gradient boosted trees = 1000, boosting learning rate = 0.1, minimum loss reduction = 0.1, maximum tree depth for base learners = 15, subsample ratio of columns for each level = 0.8, subsample ratio of columns when constructing each tree = 0.5. Some other experiments (there were an attempt to balance the dataset with synthetic data and train the model on a balanced dataset to achieve better results, however, the results were worse [4, pp. 55]) could be found in the original work [4, pp. 49].

As could be seen in figure 1, the learning procedure was stopped after approximately 60 epochs. Figure 1 also displays the ten most important features for the task. The 5th percentile of temperature was identified as the most important measure.

<table>
<thead>
<tr>
<th>Table 1: Results on the test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>accuracy [%]</td>
</tr>
<tr>
<td>sensitivity [%]</td>
</tr>
<tr>
<td>specificity [%]</td>
</tr>
<tr>
<td>$F_1$</td>
</tr>
<tr>
<td>MCC</td>
</tr>
</tbody>
</table>

The overall results for each method are shown in table 1. As can be observed, our newly proposed method significantly outperforms the conventional one. Especially the Mathews correlation coefficient, which measures a trade-off between sensitivity and specificity, reaches much higher values (0.40 vs. 0.75).

Finally, the results where we compared the classification of M2 with subject-reported data are summarised in table 2. Although the method has good sensitivity (73%), it turns out that the specificity is much lower (23%).

1The results and the code of the pipeline are accessible via https://github.com/xsigmu06/actigraphy-processing.
Figure 1: Left part – AUC (area under curve) during the learning process; right part – 10 most important features

<table>
<thead>
<tr>
<th>accuracy [%]</th>
<th>sensitivity [%]</th>
<th>specificity [%]</th>
<th>$F_1$</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>67</td>
<td>73</td>
<td>23</td>
<td>0.34</td>
<td>-0.03</td>
</tr>
</tbody>
</table>

4 DISCUSSION

In comparison to the statistical method M1, which is based on simple heuristics, our newly proposed approach employing machine learning provides better results in terms of all metrics. Nevertheless, M1 was validated on the big Whitehall II dataset, where it proved to be reliable. Thus, to be fair and before claiming strong conclusions, we are going to apply for the database as well and make further comparisons.

Although our method provided a very good performance on the Newcastle polysomnography and accelerometer dataset, it was much worse when comparing its scores with the self-reported data. Vice versa, V.T. van Hess et al. stated in [7], that M1 has a strong correlation with sleep diary outcomes but a weaker correlation with the PSG data. In other words, M2 seems to be complementary to M1, thus its combination (e.g., inputting M1 scores in the model of M2) could probably enhance the results in both scenarios.

The most important features in M2 are mainly temperature-based. Significant correlations between core-body temperature and sleep/wake patterns were mentioned in previous studies [10]. In the first dataset, the information about the temperature is reliable, because it was recorded in a laboratory environment. However, in the case of the second dataset, we measured the temperature of the wrist which could be easily affected by an environment condition, e.g. by covering the wrist with a duvet. This could be the reason we observed a poor correlation with the subject-reported outcomes.
5 CONCLUSION

We proposed a new method of sleep/wake identification, which is based on actigraphy data with consequent modelling by a gradient boosting algorithm. The method achieves 89% accuracy and Mathews correlation coefficient equal to 0.75 when compared to the polysomnography reference. The new approach outperforms the heuristic method proposed by V. T. van Hess et al. [7], nevertheless, it has a lower correlation with subject-reported outcomes.

ACKNOWLEDGEMENTS

This work was supported by grant no. NU20-04-00294 (Diagnostics of Lewy body diseases in prodromal stage based on multimodal data analysis) of the Czech Ministry of Health and by European Regional Development Fund no. CE1581 (Interreg niCE-life – Development of an integrated concept for the deployment of innovative technologies and services allowing independent living of frail elderly) of the European Union.

REFERENCES


