A balanced sleep/wakefulness classification method based on actigraphic data in adolescents

G. Orellana, *Student Member, IEEE*, C. M. Held, *Senior Member, IEEE*, P. A. Estevez, *Senior Member, IEEE*, C. A. Perez, *Senior Member, IEEE*, S. Reyes, C. Algarin, P. Peirano.

***Abstract*— Several research groups have developed automated sleep-wakefulness classifiers for night wrist actigraphic (ACT) data. These classifiers tend to be unbalanced, with a tendency to overestimate the detection of sleep, at the expense of poorer detection of wakefulness. The reason for this is that the measure of success in previous works was the maximization of the overall accuracy, disregarding the balance between sensitivity and specificity. The databases were usually sleep recordings, hence the over-representation of sleep samples.**

In this work an Artificial Neural Network (ANN), sleep-wakefulness classifier is presented. ACT data was collected every minute. An 11-min moving window was used as observing frame for data analysis, as applied in previous sleep ACT studies. However, our feature set adds new variables such as the time of the day, the median and the median absolute deviation. Sleep and Wakefulness data were balanced to improve the system training. A comparison with previous studies can still be done, by choosing the point in the ROC curve associated with the corresponding data balance.

Our results are compared with a polysomnogram-based hypnogram as golden standard, rendering an accuracy of 92.8%, a sensitivity of 97.6% and a specificity of 73.4%. Geometric mean between sensitivity and specificity is 84.9%.

I. Introduction

Several studies describe ACT applications. A continuous increase in the number of publication focused in sleep medicine has been reported [1] with promising results. In the specific field that involves this work, Tryon et al. [2] indicate that Polysomnography (PSG) and ACT focus on different steps in the process of falling asleep, which may explain, at least in part, why none of the studies in sleep/wakefulness classification based on ACT data could achieve a perfect match between their markings and those of the PSG in the exact determination of sleep onset. On the other hand, this may be a good opportunity for a learning machine to identify previously unknown relations between PSG and ACT data based on context information.

Different research groups that applied ACT in sleep, such as [1] and [3], pointed to the fact that most of previous studies overestimated the sleep state because of the nature of the data, biased towards sleep data. In a nighttime PSG recording, which serves as the base for ACT state classification, 80%-90% of the epochs are marked as sleep, which means that a classifier focused on sleep state detection will obtain a higher overall accuracy. To offset this situation, Domingues et al. [3] proposed using the geometrical mean (G-mean) between sensitivity and specificity as a classifier performance index, because it has a higher penalization for high sensitivity-low specificity cases than the sum or arithmetic mean.

Cole et al. [4] introduced 1-min frequency data acquisition in ACT and proposed an algorithm that classified looking at an 11-min window that included the five previous samples and the five following samples of the presently observed minute, and linear combinations of these activity values. Sadeh et al. [5] also used the 11-min window and improved previous results extracting several features from it, and selected them using Stepwise Discriminant Analysis. Domingues et al. [3] applied a much higher ACT sample rate, which allowed them to introduce the notion of "purposeless movements", a statistical methodology to find differences between movements in sleep and wakefulness. This is hardly feasible at a 1-min frequency data acquisition rate. Table I shows compared results of these studies. In this paper we compare Sadeh's classifier, which is the best among previous studies, with our system.

Other studies showed the development of sleep/wakefulness classifier algorithms associated to groups with sleep or psychiatric disorders [6], [7] or infants [8], [9].

II. Methods

A. Data

Both ACT and PSG data were recorded simultaneously at the Sleep and Functional Neurobiology Lab, INTA, Universidad de Chile.

ACT data was recorded at one-minute rate. The Actigraph was a Minimitter Actiwatch 64 in crosses-by-zero mode.

TABLE I. Results of Previous ACT Sleep-Wakefulness Studies

Author	Sens %	Spec %	Acc %	G-mean %	Sleep scored by PSG %
Sadeh et al [5]	97.9	74.3	92.6	85.3	77.4
Domingues et al [3]	75.6	81.6	77.8	78.5	80.5
Cole et al [4]	95.2	64.5	88.3	78.4	77.3

Research partially supported by the Department of Electrical Engineering and INTA, Universidad de Chile, and by grants from CONICYT–Chile: FONDECYT 1120319 and 1110513.

G. Orellana, C. M. Held, P. A. Estevez and C. A. Perez are with the Department of Electrical Engineering, Universidad de Chile, Santiago, Chile (gorellanal@ ing.uchile.cl).

S. Reyes, C. Algarín and P. Peirano are with the Sleep and Functional Neurobiology Lab, Institute of Nutrition and Food Technology (INTA), Universidad de Chile, Santiago, Chile.

Actigraph mode was chosen based on literature [5], [10].

The 119 recordings available for this study were obtained from 15-year-old healthy adolescents, totaling 64,102 one-minute epochs. The recording sessions consisted in two consecutive nights of simultaneous PSG and ACT recordings. Some of the recordings (about 5%) were discarded because of heavy artifact presence.

PSG data was recorded at 200Hz, measuring several physiological signals, such as EEG derivations, EOG, EMG, ECG, and others. Based on PSG data, a hypnogram was constructed in 30-seconds epochs by a specialist. The asleep and awake markings of it were used as golden standard for ACT evaluation. The global percentage of sleep in all the considered recordings was 80.5%, not unlike previous studies (see Table I).

B. Data Pre-processing

PSG and ACT data were synchronized using the limb movements signal from the PSG. Synchronization precision between data is limited by the sample frequency of ACT, i.e. one minute. When the two 30-second PSG epochs associated with a single ACT epoch happened to have different classifications (any sleep stage and wakefulness or vice-versa), that one-minute epoch was labeled as wakefulness in the target vector.

Then, preprocessing was performed on ACT data with two goals: 1) filter out low-power noise found in some recordings. This noise consisted in several minutes of constant low-level activity (up to 10 counts per minute, but not fluctuating or only marginally fluctuating). 2) Saturate ACT data, leaving all the epochs with activity over 300 crosses-by-zero, at 300 (figure 1). The saturation value was established using the training data set. It corresponds to a solid wakefulness period, and it helps to standardize a few recordings with high values of ACT.

C. Feature Extraction

An 11-minutes window was applied on the ACT data, as used elsewhere in the literature [4], [5], considering the current sample at the center of the window, and including both the five previous and five following samples.

Thirty four features were calculated for each 11-min window, including raw and logarithm activity level of every sample, and median, median absolute deviation, minimum value and number of minutes with non-zero activity for different combinations of samples. Stepwise Discriminant Analysis [5] was used to discard features with little or no contribution to the classification, The nine characteristics extracted from the window were:

- The natural logarithms of:
  1) The median of the 11-minutes activity counts;
  2) The median of the activity counts of the initial 6 minutes of the 11-minutes window;
  3) The median of the activity counts of the final 6 minutes of the 11-minutes window;
  4) The median absolute deviation of activity counts of the initial 6 minutes of 11-minutes window;
  5) The median absolute deviation of activity counts of the final 6 minutes of 11-minutes window.

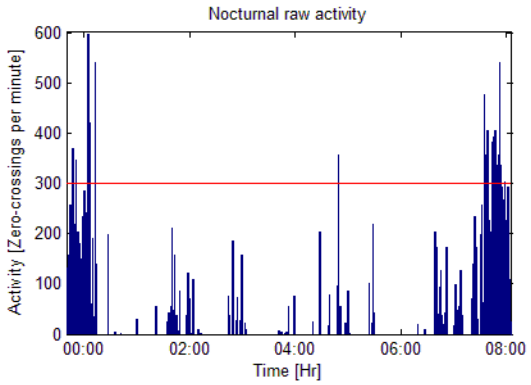

Figure 1. Raw data of night time activity. In red, 300 crosses-by-zero, value in which data was saturated.

All these variables (natural logarithms) were added a unit before serving as inputs to the ANN.

- The number of minutes with modified (because of the pre-processing) activity counts over 0 of:
  6) The initial 6 minutes of the 11-minutes window
  7) The final 6 minutes of 11-minutes window
  8) The activity count of the central minute
  9) The time of the day

The time was expressed as the minute of the day beginning at midnight (values between 1 and 1440). The median was chosen as measure of central tendency and median absolute deviation (MAD) as measure of dispersion because these are robust statistics measures, independent of the distribution of the data.

D. Classification

To carry out a proper classification using ANN, data classified as wakefulness by the PSG was repeated in the training database four times, in order to balance data classes.

Several tests using the training and test sets were made, searching for an optimal ANN architecture. McNemar's Test [11] showed that there were many solutions with two hidden layers and different number of nodes with no significant statistical differences in their results. Finally, a two hidden layer configuration 9:15:14:1 was chosen because it had the highest performance. The ANN was configured using Levenberg-Marquardt as training algorithm, mean square error for performance measurement, and the hyperbolic tangent sigmoid as transfer function in all layers. The selection of this configuration and algorithms was done empirically.

Results were obtained by training and testing the classifier using a 5×2 cross validation scheme [12]: Five iterations of *2-folds* cross validations were performed. In each of the five iterations, the balanced database was divided in two sets, which were used alternatively as training and test set. A portion of the training set was randomly separated and used as validation set (proportion 4:1 between final training and validation sets). The ANN was trained ten times with each configuration; the best performer in the training and validation data sets for each partition was chosen. A total of ten "best performers" were obtained with this method. Once the "best performer" for a balanced dataset partition was

obtained, we searched for the threshold that maximized accuracy in the original unbalanced database. For each ANN obtained, a ROC curve was built by moving the threshold value of the ANN output. We then looked for the point in the ROC curve that had the highest accuracy on the original unbalanced data, considering this accuracy as the weighted mean of sensitivity and specificity, the weight being the proportion of data (0.8047 for sensitivity and 0.1953 for specificity). The threshold found with this method was applied to classify the test data set. This was done ten times, one for each partition, and its average is considered as the system performance.

III. Results

The overall accuracy considering epoch-by-epoch comparison of the output of the ACT system and PSG (as golden standard) was measured using the 5×2 cross validation explained above. Accuracy (index of correctly detected epochs over the total number of epochs) obtained from the ANN trained on the balanced database was 88.1±0.4%, Sensitivity (index of correctly detected sleep epochs over the number of epochs marked as sleep by golden standard) was 92.4±1.5%, and Specificity (index of correctly detected awake epochs over the number of epochs marked as awake by golden standard) was 83.6±1.4%, reaching a G-mean value of 87.9±2.9%. Detailed results are shown in table II.

The ANN output value (between 0 and 1) allows to perform different classifications by modifying the output threshold (default value is 0.5). In this way it can be adapted to improve the score of one state at the expense of the other. Sleep/wakefulness classification of sleep recordings is a prime candidate for this, because the imbalance of data causes that having a higher index of correct sleep classification involves a higher improvement on overall accuracy. The effect of modifying the threshold value can be described in a ROC curve (figure 2).

Considering the database with its original unbalanced proportion of sleep/wakefulness at the optimal threshold on the ROC curve was 0.22±0.01, results showed an accuracy of 92.8±0.4%, a sensitivity of 97.6±0.3%, and a specificity of 73.4±1.4%, reaching a G-mean of 84.6±1.8% (table III).

Comparing with previous studies, our system added new inputs, such as the time of the day and those referring to the initial or end part of 11-min window (features 2-7), relating motor activity with the circadian cycle. In a future work, in order to generalize our method, the use of the time of the day variable would need to be preprocessed in order to correct for daytime sleepers, night workers and others with an altered day-night rhythmicity. For now, we tested how our system would perform without time-related features: time of the day and features 2-7. We replaced features 2-7 (each calculated twice, one taking the initial 6 min of the window and another taking the final 6 min) by the same calculation, but calculated once for the whole window. Using the original dataset with a threshold of 0.21±0.02 calculated with the training data set, results of this classification fell to 91.3±0.2% of accuracy. 5×2 cross validation paired t Test [11] was carried out to compare results of the classification with and without the time-related feature. $\tilde{t}$ Statistic calculated were equal to 5.7027 ($p-\text{value} = 0.0012$). A ROC curve was also constructed by varying the output threshold of this modified ANN classifier (figure 2). Comparisons among algorithms built in this study are presented in table IV.

TABLE II. AVERAGE RESULTS OF THE ANN CLASSIFIER TRAINED ON A 5×2 CROSS VALIDATION TEST USING BALANCED DATA. IN EACH TRIAL, THE NUMBER OF MINUTES IN A STATE IS AN INTEGER.

		PSG		**Total**
		Scored as Sleep	***Scored as Wake***	
		minutes	*Minutes*	
Class. as Sleep	*minutes*	23823±381	4091±352	27914
	% column	92.4±1.5	16.4±1.4	
Class. as Wake	*minutes*	1968±378	20891±350	22859
	% column	7.6±1.5	83.6±1.4	
Total		25791	24982	50773

TABLE III. AVERAGE RESULTS OF AN ANN CLASSIFIER TRAINED USING 5×2 CROSS VALIDATION WITH BALANCED DATA, AND THEN ADAPTING THE THRESHOLD TO OPTIMIZE OUTPUT WITH THE ORIGINAL UNBALANCED DATA.

		PSG		**Total**
		Scored as Sleep	***Scored as Wake***	
		minutes	*Minutes*	
Class. as Sleep	*minute*	25171±80	6653±351	31824
	% column	97.6±0.3	26.2±1.4	
Class. as Wake	*minute*	619±64	18329±331	18948
	% column	2.4±0.3	73.8±1.4	
Total		25790	24982	50772

IV. Discussion

The methodology presented in this study aimed at optimizing classification results, disregarding the data balance between the two states of the database. Our proposition is that the classification can be optimized to the proportion of classes after training, by adjusting the output threshold.

These results allow us to perform comparisons with previous studies because the balanced database does not include any new data, but was enlarged just by repeating wakefulness data. Comparing with previous studies [4]–[9], our selection of characteristics introduced 3 features that were calculated twice (median, number of minutes over 0, and median absolute deviation of ACT), one for the initial segment and another for the final segment in the 11-minutes window. Features calculated twice allowed obtaining tendency data and, along with time of the day, helped to identify the beginning and the end of sleep, reducing false awakening detections in the middle of the night, triggered by unconscious movements.

Besides, use of the median and median absolute deviation as statistical measures of central tendency and dispersion respectively avoids the unproven assumption of normal distribution of the data.

In order to compare the results of our method with previous works, we implemented Sadeh's algorithm for sleep/wakefulness classification [5] and tested it with our original unbalanced dataset. For each iteration of 5×2 *cross*

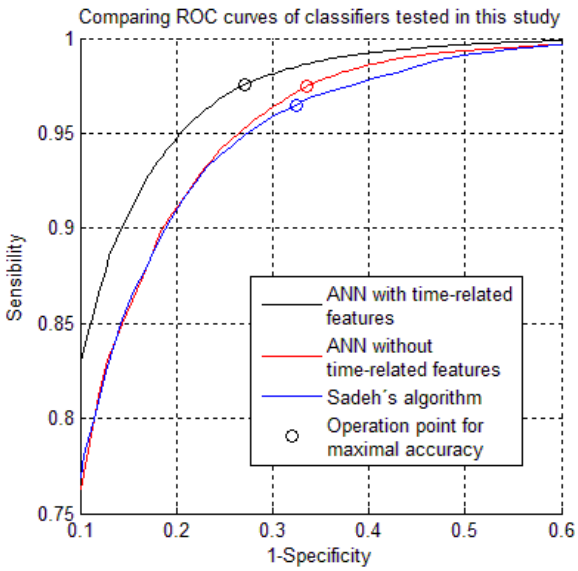

Figure 2. ROC curve of 3 classifiers tested in this study. In black, our ANN classifier using time-related features. In red, our ANN classifier without time-related features. In blue, Sadeh's classifier.

validation, data in the training set was used to adjust the classifier and find an optimal threshold. The classifier was then applied on the test dataset. Sadeh's algorithm was selected for comparison because of its reproducibility and its good published scores. A 5×2 cross validation paired *t* Test [11] was carried out to compare the results of the two algorithms. Calculated $\tilde{t}$ statistics were equal to 5.6341 ($p-\text{value} = 0.0012$). It is also possible to modify the threshold value for Sadeh's algorithm to obtain a ROC curve (figure 2), which gives points with different G-mean values.

We used an ANN as classification tool because of its ability to adjust its performance to any point of the ROC curve. Also, since its output is a continuous value between 0 and 1, it can be interpreted as class membership index for each input. On the other hand, the ANN classifier requires an extensive computing time for its training process to obtain an optimal performance, which is not a significant handicap for this application.

V. Conclusions

In this work a sleep/wakefulness classification method to score 1-minute ACT epochs is presented. Its performance is comparable to a well-known algorithm described in the literature [5]. Our system obtained an accuracy of 92.8±0.4%, a sensitivity of 97.6±0.3% and a specificity of 73.4±1.4%.

TABLE IV. Comparison of the average result of the ACT classification Algorithms studied with our dataset

Algorithm	Sensibility %	Specificity %	Accuracy %	G-mean %
Sadeh et al [5]	96.5±0.4	67.8±1.7	90.9±0.2	80.9±2.4
ANN without time-related features	97.5±0.3	66.8±1.4	91.3±0.2	80.7±1.9
ANN with time-related features	97.6±0.3	73.4±1.4	92.8±0.4	84.6±1.8

The use of an input that links ACT data with the circadian cycle improved the classification results. The time of the day input can't be directly applied in groups of subjects with different sleep schedules. In those cases, we would need further research to learn how to adapt the time of the day variable in order to make it a useful classification input.

The differences between the results obtained by us with Sadeh's algorithm and the results published by Sadeh et al. [5] could be due to difference in actigraph types, size of the database or subjects. Also, as can be seen in figure 2, performance of ANN without time-related features and Sadeh's classifier (table IV) are quite similar. This could be because similarity of features used, and could be independent of the classification method.

Other paths of future work to improve the performance of the classifier would be using the ANN output as an input for a second line classifier, and defining new discriminant features. Another line of research would be to add other data acquisition instruments to record other features simultaneously, such as EOG, ECG, chest volume, etc., which would probably improve the performance and robustness of the classifier.

References

[1] A. Sadeh, "The role and validity of actigraphy in sleep medicine: An update," *Sleep Medicine Reviews*, vol. 15, pp. 259–267, 2011.

[2] W. W. Tryon, "Issues of validity in actigraphic sleep assessment," *Sleep*, vol. 27, pp. 158–165, 2004.

[3] A. Domingues, T. Paiva, and J. Sanches, "Sleep and wakefulness state detection in nocturnal actigraphy based on movement information," *IEEE Trans. Biomed. Eng.*, vol. 61, pp. 426–434, 2013.

[4] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin, "Automatic sleep/wake identification from wrist activity," *Sleep*, vol. 15, pp. 461–469, 1992.

[5] A. Sadeh, K. M. Sharkey, and M. A. Carskadon, "Activity-based sleep-wake identification: an empirical test of methodological issues," *Sleep*, vol. 17, pp. 201–207, 1994.

[6] C. A. Kushida, A. Chang, C. Gadkary, C. Guilleminault, O. Carrillo, and W. C. Dement, "Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients," *Sleep Med.*, vol. 2, pp. 389–396, 2001.

[7] J. Hedner, G. Pillar, S. D. Pittman, D. Zou, L. Grote, and D. P. White, "A novel adaptive wrist actigraphy algorithm for sleep-wake assessment in sleep apnea patients," *Sleep*, vol. 27, pp. 1560–1566, 2004.

[8] E. Sazonov, N. Sazonova, S. Schuckers, and M. Neuman, "Activity-based sleep-wake identification in infants," *Physiological measurement*, 25(5), 1291. 2004.

[9] J. Tilmanne, J. Urbain, M. V Kothare, A. Vande Wouwer, and S. V Kothare, "Algorithms for sleep-wake identification using actigraphy: a comparative study and new results," *J. Sleep Res.*, vol. 18, pp. 85–98, 2009.

[10] A. Sadeh, P. Lavie, A. Scher, E. Tirosh, and R. Epstein, "Actigraphic home-monitoring sleep-disturbed and control infants and young children: a new method for pediatric assessment of sleep-wake patterns," *Pediatrics*, vol. 87, pp. 494–499, 1991.

[11] T. Dietterich, "Approximate Statistical Test For Comparing Supervised Classification Learning Algorithms," *Neural Comput.*, 1998.