Available online at www.sciencedirect.com
ScienceDirect
Biomedical Signal Processing and Control 2 (2007) 171–179
Biomedical Signal Processing and Control
www.elsevier.com/locate/bspc

Feature selection for sleep/wake stages classification using data driven methods

Lukáš Zoubek [a,b], Sylvie Charbonnier [b,*], Suzanne Lesecq [b], Alain Buguet [e], Florian Chapotot [c,d]

[a] VSB-Technical University of Ostrava, Department of Measurement and Control, 17. listopadu 15/2172, Ostrava-Poruba 708 33, Czech Republic
[b] GIPSA-Lab, Control Systems Department, BP 46, 38402 Saint Martin d'Hères Cedex, France
[c] Sleep, Chronobiology and Neuroendocrinology Laboratory, The University of Chicago, 5841 South Maryland Avenue, AMB M374 (MC 1027) Chicago, IL, USA
[d] PhiTools, 1 rue du Général de Castelnau, Strasbourg, France[1]
[e] Neurobiologie des états de vigilance, EA 3734, Université Claude-Bernard Lyon 1, Lyon, France

Received 27 February 2007; received in revised form 11 May 2007; accepted 18 May 2007
Available online 2 July 2007

Abstract

This paper focuses on the problem of selecting relevant features extracted from human polysomnographic (PSG) signals to perform accurate sleep/wake stages classification. Extraction of various features from the electroencephalogram (EEG), the electro-oculogram (EOG) and the electromyogram (EMG) processed in the frequency and time domains was achieved using a database of 47 night sleep recordings obtained from healthy adults in laboratory settings. Multiple iterative feature selection and supervised classification methods were applied together with a systematic statistical assessment of the classification performances. Our results show that using a simple set of features such as relative EEG powers in five frequency bands yields an agreement of 71% with the whole database classification of two human experts. These performances are within the range of existing classification systems. The addition of features extracted from the EOG and EMG signals makes it possible to reach about 80% of agreement with the expert classification. The most significant improvement on classification accuracy is obtained on NREM sleep stage I, a stage of transition between sleep and wakefulness.

Keywords: Decision making; Diagnosis; Medical applications; Pattern recognition; Signal processing

1. Introduction

Polysomnography (PSG) consists in study of sleep and wakefulness from the concurrent recording of multiple bio-electric signals including the electroencephalogram (EEG), electro-oculogram (EOG) and electromyogram (EMG). A system of standardized rules established in the conventional Rechtschaffen and Kales (R&K) human sleep/wake stage scoring manual [1] enables the visual recognition by medical and technical experts of up to six different vigilance stages: wakefulness, non-rapid eye-movement (NREM) sleep stages I, II, III and IV, and REM or paradoxical sleep (PS). NREM stages III and IV represent the slow wave sleep (SWS). The successive visual interpretation, by 20 or 30 s epochs, of 8–24 h PSG recordings leads to the representation of the temporal distribution of sleep/wake stages called a hypnogram, an example of which is presented in Fig. 1. A hypnogram reveals the internal architecture of sleep and the alternation of NREM and REM sleep phases, which makes the discrimination between normal and abnormal sleep much simpler. PSG is thus a powerful tool in the diagnosis of sleep disorders, which are rather common with about 5% of the general population affected [2].

Since 1970 and the development of computerized methods, automated systems have emerged in order to automatically score PSG recordings, so as to avoid the expert to spend too much time to this tedious and time-consuming work. The visual interpretation of PSG recordings is a typical pattern recognition task. Physicians look at the signals and classify successive epochs from the shape of their traces. Two problems must be

* Corresponding author. Tel.: +33 476826415; fax: +33 476826388.
E-mail address: Sylvie.Charbonnier@inpg.fr (S. Charbonnier).
[1] www.phitools.com.

1746-8094/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bspc.2007.05.005

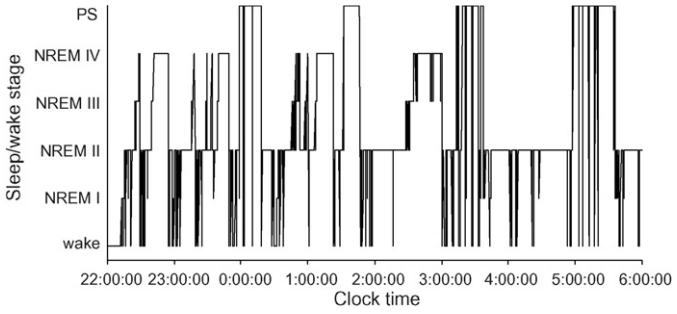

Fig. 1. An example of hypnogram obtained from a night-time PSG recording.

solved to obtain automatic classifiers able to obtain results similar to human experts: to choose the classification function that will give the best results on this problem and to process the signals with adequate techniques so as to obtain inputs to the classifier (features) which are the most similar to the visual information used by the expert.

An important number of publications can be found in the literature on automatic sleep/wake stages analysis. Many of these papers focus on the choice of an adequate type of classifier to achieve accurate classification. The authors use either classical algorithms or artificial intelligence methods, such as neural networks [3–7]. Features used as inputs to the classification systems are extracted from PSG signals at constant intervals (epoch) using various signal processing techniques operating in the time domain and/or in the frequency domain. Though several propositions were made to process the PSG signals, only few studies have been performed to determine the optimal set of features achieving an accurate sleep/wake stage classification [8–14].

Thanks to the development of computerized methods and in parallel to automated systems, a research field has emerged, known as data mining or knowledge discovery. This research field proposes methods that enable the extraction of knowledge from large sets of examples [15]. The aim of the study presented in this paper was to apply data mining methods to extract knowledge about sleep/wake stages classification. Knowledge extraction was performed from a large database composed of 47 night sleep recordings from 41 healthy subjects. Feature selection algorithms and systematic statistical assessment were performed to determine which signals and processing methods are the most relevant and accurate for sleep/wake stage automated classification.

The outline of the paper is the following. The whole database and the techniques used to process the signals and extract the features are presented in Section 2. The features selection methods used are described in detail in Section 3. The results are presented in Section 4 and discussed in Section 5.

2. Materials

2.1. Presentation of the PSG recording database

In this study, a large database of PSG recordings was used. The full database contains 47 night-time PSG recordings obtained from 41 healthy adult subjects (19–47 years old, 39 males and 2 females). Recordings were made continuously during the night (8 h between 22:00 h and 06:00 h). Four EEG channels (C3-A2, P3-A2, C4-A1, and P4-A1), one transversal EOG and one chin EMG were registered and digitized at a sampling frequency f_s = 128 Hz. The EEG leads were attached onto the scalp according to the International 10-20 EEG System of Electrodes Placement [16].

All the 47 PSG recordings were visually interpreted by two independent sleep physicians. Visual sleep/wake stage scoring was performed with constant epoch duration of 20 s according to the conventional rules of the R&K manual [1]. Each epoch was thus classified into one of five different stages: wakefulness, NREM sleep stage I, NREM sleep stage II, slow wave sleep (SWS or NREM stages III and IV), and paradoxical sleep. To avoid the introduction of expert inaccuracies in the database, only the epochs classified in the same stage by both experts were considered in this project. They represent 84% of the original PSG recording database and only that subset was used to form our study database. The total number of epochs included was 63,254. As it can be seen in Table 1, the number of epochs classified in each sleep/wake stage is different. NREM stage II lasts a long time, whereas NREM stage I is rather short. To avoid classification errors related to differences in the sample size of each class, the database was further reduced to a smaller one where each class is composed of about the same number of epochs. The numbers of epochs classified in each sleep stage for the database reduced are presented in the second row of Table 1.

The database used in this study thus consisted of 10,000 randomly selected epochs classified into one of the five sleep/wake stages. The set S of 10,000 epochs was split in 10 subsets $S = \{S_1, S_2, \ldots, S_{10}\}$, each subset S_k containing 1000 epochs, equally distributed in the five classes. The size of the subsets (1000 epochs) was chosen from a previous study whose goal was to analyze the effect of the number of examples on the classification error [17]. Its main conclusion was that a minimal number of 500 examples was required to train and validate a classifier on a sleep/wake classification problem and get an unbiased evaluation of the classification accuracy.

2.2. Features extracted from the PSG recordings

Each epoch stored in the database consists of a 20 s recording of six signals (four EEG, one EOG and one EMG) [18]. Since the PSG recordings were sampled at 128 Hz, each time series contains 2560 samples. Various features describing different signal characteristics were extracted from each signal using multiple processing techniques. The PRANA software for PSG analysis (PhiTools, Strasbourg, France) was used to

Table 1
Description of the database used in this study (number of epochs in the sleep/wake stages)

	Wake	NREM I	NREM II	SWS	PS
Full database	5232	1989	32966	7701	15,366
Reduced database	1914	1879	2206	1902	2,099

visually interpret the recording database and to perform feature extraction.

EEG is traditionally analyzed in the frequency domain, since each sleep stage is characterized by a specific pattern of frequency contents, but some useful information can be added from temporal analysis. EOG and EMG are most often analyzed in the time domain, because these signals do not exhibit obvious frequency patterns.

2.2.1. EEG features

- A set of five features was used to describe the spectral activity of the EEG in traditional frequency bands [19,20]. They were calculated using Fourier transformation. Relative powers were computed in five frequency bands by dividing absolute powers in each frequency range by the sum of powers in the 0.5–32.5 Hz frequency range:
  - $P_{\text{rel}}(\text{EEG}, \delta_{\text{FT}})$ with $\delta_{\text{FT}} = [0.5; 4.5]$ Hz;
  - $P_{\text{rel}}(\text{EEG}, \theta_{\text{FT}})$ with $\theta_{\text{FT}} = [4.5; 8.5]$ Hz;
  - $P_{\text{rel}}(\text{EEG}, \alpha_{\text{FT}})$ with $\alpha_{\text{FT}} = [8.5; 11.5]$ Hz;
  - $P_{\text{rel}}(\text{EEG}, \sigma_{\text{FT}})$ with $\sigma_{\text{FT}} = [11.5; 15.5]$ Hz;
  - $P_{\text{rel}}(\text{EEG}, \beta_{\text{FT}})$ with $\beta_{\text{FT}} = [15.5; 32.5]$ Hz.
- Another set of five features was used to characterize the EEG signal. They were computed from the Wavelet coefficients generated by discrete Wavelet transform. A four-level Wavelet packet decomposition (with Daubechies3 Wavelet) was used to compute the features in the five frequency bands considered:
  - $\delta_{\text{WT}} = [0; 4]$ Hz;
  - $\theta_{\text{WT}} = [4; 8]$ Hz;
  - $\alpha_{\text{WT}} = [8; 12]$ Hz;
  - $\sigma_{\text{WT}} = [12; 16]$ Hz;
  - $\beta_{\text{WT}} = [16; 32]$ Hz.

The information contained in the selected arrays of Wavelet coefficients is characterized by the quadratic mean value (root mean square value, RMS) of the coefficients:

$$\text{RMS}_{\text{FB}} = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m} c_{\text{FB}}(i)^2} \tag{1}$$

where m is the number of Wavelet coefficients $c_{\text{FB}}(i)$ in each frequency band FB and $\text{FB} \in \{\delta_{\text{WT}}, \theta_{\text{WT}}, \alpha_{\text{WT}}, \sigma_{\text{WT}}, \beta_{\text{WT}}\}$. The features are then expressed as relative values of RMS_{FB} computed over these five frequency bands and are labeled as $\{\text{RMS}_{\text{rel}}\,\delta, \text{RMS}_{\text{rel}}\,\theta, \text{RMS}_{\text{rel}}\,\alpha, \text{RMS}_{\text{rel}}\,\sigma, \text{RMS}_{\text{rel}}\,\beta\}$.

- Five features were used to describe the signal in the time domain, namely, the signal entropy, the 75th percentile of the signal distribution, the standard deviation, the skewness and kurtosis numbers.

The entropy, entr_{EEG} [21], is computed from a histogram of the signal during one epoch:

$$\text{entr}_{\text{EEG}} = -\sum_{j=1}^{N} \frac{n_j}{n} \ln \frac{n_j}{n} \tag{2}$$

where n is the number of samples $y(i)$ of the measured signal y in the epoch, N the number of bins used for the calculation of the histogram and n_j is the number of samples $y(i)$ which values are within the jth bin. In this study, N is chosen as the largest integer inferior to n squared root, it is the same for each epoch.

The 75th percentile of the signal distribution, $\text{prctile75}_{\text{EEG}}$, is defined as

$$\text{card}\{y(i)/y(i) < \text{prctile75}_{\text{EEG}}\} = \frac{75n}{100} \tag{3}$$

where n is the number of samples $y(i)$ of the measured signal y in the epoch and card stands for the number of elements in the set.

The standard deviation, std_{EEG}, is defined as

$$\text{std}_{\text{EEG}} = \left[\frac{1}{n-1}\sum_{i=1}^{n}(y(i) - \bar{y})^2\right]^{1/2} \tag{4}$$

where n is the number of samples $y(i)$ of the measured signal y in the epoch and $\bar{y}$ represents the mean value (5) of the signal y:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y(i) \tag{5}$$

The skewness, skew_{EEG}, is defined as

$$\text{skew}_{\text{EEG}} = \frac{M_3}{M_2\sqrt{M_2}} \tag{6}$$

with

$$M_k = \frac{1}{n}\sum_{i=1}^{n}(y(i) - \bar{y})^k \tag{7}$$

The kurtosis, kurt_{EEG}, is defined as

$$\text{kurt}_{\text{EEG}} = \frac{M_4}{M_2 M_2} \tag{8}$$

2.2.2. EMG and EOG features

EMG and EOG signals were processed in the time domain. Both signals are characterized by their entropy $\{\text{entr}_{\text{EMG}}, \text{entr}_{\text{EOG}}\}$, their 75th percentile $\{\text{prctile75}_{\text{EMG}}, \text{prctile75}_{\text{EOG}}\}$, their standard deviation $\{\text{std}_{\text{EMG}}, \text{std}_{\text{EOG}}\}$, their skewness $\{\text{skew}_{\text{EMG}}, \text{skew}_{\text{EOG}}\}$ and their kurtosis $\{\text{kurt}_{\text{EMG}}, \text{kurt}_{\text{EOG}}\}$ as defined in Eqs. (2)–(8).

The EMG signal was also processed in the frequency domain. The relative power of the EMG signal in a high frequency band [12.5; 32] Hz, $P_{\text{rel}}(\text{EMG, high})$, was calculated as

$$P_{\text{rel}}(\text{EMG, high}) = \frac{P(\text{EMG}, [12.5{-}32\,\text{Hz}])}{P(\text{EMG}, [8{-}32\,\text{Hz}])} \tag{9}$$

2.3. Transformation of the features

In order to reduce the influence of extreme values that are often observed on physiological variables, each feature of the

Table 2
Transformations toward normal distribution

Feature	Transformation
$P_{rel}\,\delta$, $P_{rel}\,\theta$, $RMS_{rel}\,\delta$, $RMS_{rel}\,\theta$	$\arcsin(\sqrt{x})$
$P_{rel}\,\alpha$, $P_{rel}\,\sigma$, $P_{rel}\,\beta$, P_{rel}(EMG, high), $RMS_{rel}\,\alpha$, $RMS_{rel}\,\sigma$, $RMS_{rel}\,\beta$	$\log\left(\frac{x}{1-x}\right)$
$entr_{EEG}$, $entr_{EMG}$, $entr_{EOG}$, $prctile75_{EEG}$, $prctile75_{EMG}$, $prctile75_{EOG}$, std_{EEG}, std_{EMG}, std_{EOG}, $kurt_{EEG}$, $kurt_{EMG}$, $kurt_{EOG}$	$\log(1+x)$
$skew_{EEG}$, $skew_{EMG}$, $skew_{EOG}$	–

database was transformed using a non-linear transformation [22].

Each night recording was processed as follows. Firstly, the features were extracted from the recording using signal processing techniques and transformed using an appropriate function. The list of transformations that were applied to each feature is presented in Table 2. They were chosen from [17]. After this transformation, each feature x was normalised in a new variable z, using a z-score normalisation:

$$z = \frac{x - \mu}{\sigma} \tag{10}$$

where μ is the night recording mean value of the transformed feature x and σ is its standard deviation.

Each epoch is represented by a set of 26 features, which are summarized in Table 3.

3. Feature selection methods

In this section, the methods used to select the most relevant features are presented. Sequential methods were implemented, increasing or decreasing the number of features to be used according to the value of a criterion J. Though these methods are not optimal, they were used because the results they provide are easy to analyze.

Let $f_1, f_2, \ldots, f_n$ be a set of n features to select. Let F be a subset of these n features and $\bar{F}$ be the subset of features that are not in F:

$$F \cup \bar{F} = \{f_1, f_2, \ldots, f_n\}, \qquad F \cap \bar{F} = \varnothing$$

Let J be a criterion to be maximised and $J(F)$, the criterion J that is calculated with the features contained in the subset F. The sequential selection is an iterative technique which selects at each step i the subset F_i of features that maximises J.

Table 3
The set of features used in the study to characterize an epoch

EEG signal	$P_{rel}\,\delta$, $P_{rel}\,\theta$, $P_{rel}\,\alpha$, $P_{rel}\,\sigma$, $P_{rel}\,\beta$ $RMS_{rel}\,\delta$, $RMS_{rel}\,\theta$, $RMS_{rel}\,\alpha$, $RMS_{rel}\,\sigma$, $RMS_{rel}\,\beta$ $entr_{EEG}$, $prctile75_{EEG}$, std_{EEG}, $skew_{EEG}$, $kurt_{EEG}$
EMG signal	$entr_{EMG}$, $prctile75_{EMG}$, std_{EMG}, $skew_{EMG}$, $kurt_{EMG}$ P_{rel} high
EOG signal	$entr_{EOG}$, $prctile75_{EOG}$, std_{EOG}, $skew_{EOG}$, $kurt_{EOG}$

3.1. Sequential forward selection (SFS)

The method consists in increasing at each step i the number of features contained in F_{i-1} by one. Let F_{i-1} be the subset of features selected at step $i-1$, that maximises $J(F_{i-1})$. F_{i-1} contains $i-1$ features, which were previously selected. $\bar{F}_{i-1}$ contains the $n-i+1$ features still to be selected. At step i, a new feature f_i is selected out of $\bar{F}_{i-1}$ as $J(F_{i-1} \oplus f_i) = \max(J(F_{i-1} \oplus f_k))$ with $f_k \in \bar{F}_{i-1}$.

The first subset is initialised to the empty set $F_0 = \{\varnothing\}$.

3.2. Sequential backward selection (SBS)

It consists in decreasing at each step i the number of features contained in F_{i-1} by one. Let F_{i-1} be the subset of features selected at step $i-1$, that maximises $J(F_{i-1})$. F_{i-1} contains $n-i+1$ features, which were previously selected. $\bar{F}_{i-1}$ contains the $i-1$ features that were rejected. At step i, a new feature f_i is rejected out of F_{i-1} as $J(F_{i-1} - f_i) = \max(J(F_{i-1} - f_k))$ with $f_k \in F_{i-1}$. The first subset is initialised to the subset containing all the features, $F_0 = \{f_1, f_2, \ldots, f_n\}$.

3.3. Criterion

In this study, the criterion J to be maximised is a function of the percentage of epochs correctly classified by a classifier C.

As presented in Section 2, the database S was split into 10 subsets, $S = \{S_1, S_2, \ldots, S_{10}\}$. Each subset S_k contains 1000 epochs. Each of the five classes to recognise is equally represented in S_k. A classifier C is trained on one subset S_k and validated on the nine other subsets $S_{\bar{k}}$, $S_{\bar{k}} \in \bar{S}_k$ with $\bar{S}_k = S - S_k$.

An accuracy function is calculated on each of the nine subsets $S_{\bar{k}}$ as

$$\mathrm{Acc}(k, \bar{k}) = \frac{\mathrm{card}[\{\mathrm{ep}(i) \in S_{\bar{k}} / C(\mathrm{ep}(i)) - E(\mathrm{ep}(i)) = 0\}]}{\mathrm{card}[S_{\bar{k}}]} \tag{11}$$

where $\mathrm{ep}(i)$ is an epoch belonging to $S_{\bar{k}}$, $C(\mathrm{ep}(i))$ the class assigned to epoch(i) by the classifier C, trained on the subset k. $E(\mathrm{ep}(i))$ is the class assigned by the experts to $\mathrm{ep}(i)$.

A circular permutation is performed on the 10 subsets S_k. The classifier is trained 10 times using the different data sets S_k. Thus, 90 values of $\mathrm{Acc}(k, \bar{k})$ are obtained. The criterion J used to select the features is

$$J = \frac{1}{10} \sum_{k=1}^{10} \left(\frac{1}{9} \sum_{\substack{j=1 \\ j \neq k}}^{10} \mathrm{Acc}(k, j) \right) \tag{12}$$

$J(F_i)$ is the value of criterion J defined by (11) and (12) using the features contained in the feature subset F_i. In Eq. (12), the term in brackets corresponds to the mean accuracy obtained on the nine validation sets, when the classifier C is trained on one training set. J corresponds to the mean accuracy obtained on the

validation sets, when the classifier C is trained 10 times with 10 different training sets. Computing J this way ensures that the accuracy obtained is insensitive to the training set used. The standard deviation of the accuracy Acc obtained using classifier C is computed with:

$$\mathrm{std}_{\mathrm{Acc}} = \left[\frac{1}{89} \sum_{k=1}^{10} \left(\sum_{\substack{j=1 \\ j \neq k}}^{10} (\mathrm{Acc}(k,j) - J)^2 \right) \right]^{1/2} \quad (13)$$

Actually, $\mathrm{std}_{\mathrm{Acc}}$ is an indicator of the dispersion of the accuracies. It can be used to determine if the accuracies obtained using different features are statistically different.

3.4. Classifiers

To ensure that the results obtained are independent of the classifier used, the features selection methods were processed with three different classifiers, each of them calculating the frontiers of each class in a different way [23,24]:

- Two Bayes rule-based classifiers:
  - A parametric one, the quadratic classifier, where the probability density function of each class is assumed to be a multidimensional Gaussian model, the mean and covariance matrix being estimated for each class from the training set.
  - A non-parametric one (no prior assumptions are made on the probability density functions), the k-nearest neighbours classifier, where the probability density function is estimated with the volume occupied by a fixed number of neighbours. Ten nearest neighbours were used.
- A multi-layer perceptron (MLP), where the frontiers of each class are directly calculated from the training set. A neural network with three layers was implemented as an automatic classifier. The architecture of the neural network was chosen from [17], the transfer functions being adjusted using a trial and error method. The number of neurons in the first layer is defined by the number of input features extracted from the epoch to be processed. The transfer function of the neurons in this layer is a hyperbolic tangent function. The hidden layer of the network contains six neurons; the transfer function is a logarithmic sigmoid function. The output layer of the network consists of five neurons; the transfer function of each neuron is a hyperbolic tangent. The number of neurons in the output layer is determined by the number of target sleep/wake stages to be classified. The neural network is trained using feedforward backpropagation gradient algorithm. The weights representing connections between the neurons were randomly initiated at the beginning of the learning phase. The network was trained 10 times with 10 different random initialisation sets and the best network was kept, so as to avoid being trapped in a local minimum during the training phase and not reach the global minimum.

4. Results

The results obtained by the data mining methods are presented in this section. Only the C3-A2 EEG channel was used to obtain the results that are presented below. Actually, tests were performed using each of the four EEG channels but the results showed that no EEG channel outperforms the others. The mean accuracies obtained using criterion J, defined in Eq. (12), were not statically different, whatever the channel used.

4.1. Comparison of Fourier and Wavelet transform

The ability of Wavelet transform compared to Fourier analysis to process EEG signals was first analyzed. To do so, only the features describing EEG activity in different frequency band, using the Fourier transform or the Wavelet transform were used to train the classifiers. The classification accuracy (12), and its standard deviation (13), obtained using only the features $\{P_{\mathrm{rel}}\,\delta, P_{\mathrm{rel}}\,\theta, P_{\mathrm{rel}}\,\alpha, P_{\mathrm{rel}}\,\sigma, P_{\mathrm{rel}}\,\beta\}$ extracted from the Fourier transform (defined in Section 2.2) or using only the features obtained by means of the Wavelet transform $\{\mathrm{RMS}_{\mathrm{rel}}\,\delta, \mathrm{RMS}_{\mathrm{rel}}\,\theta, \mathrm{RMS}_{\mathrm{rel}}\,\alpha, \mathrm{RMS}_{\mathrm{rel}}\,\sigma, \mathrm{RMS}_{\mathrm{rel}}\,\beta\}$ (defined in Section 2.2) are presented in Table 4.

The results are quite similar. Actually, the Wavelet transform performs a decomposition of the signal over different frequency bands and provides about the same information as the Fourier transform. Nonetheless, whatever the classifier, Table 4 shows that the accuracy is significantly higher (t-test; $p < 0.01$) when the relative EEG powers are calculated using the Fourier transform, the best results being obtained with a neural network classifier ($71.56 \pm 1.46\%$). As they are also longer to process, Wavelet transform features were eliminated from further analysis.

4.2. Selection of the most relevant features

The SFS and the SBS methods were both applied to the set of features presented in Table 3. The features corresponding to EEG signal processed by Wavelet transform ($\mathrm{RMS}_{\mathrm{rel}}$ EEG) were removed from the set. The subset of features representing relative power of EEG in the frequency bands obtained with the Fourier transform was considered as a single feature (P_{rel} EEG). The selecting feature method could select P_{rel} EEG only, meaning that all the energies in the five bands were selected.

Table 4
Classification accuracies obtained with Wavelet and Fourier transform—mean values and standard deviations

%	Classification accuracy	
	Fourier transform	Wavelet transform
Quadratic	69.91 ± 1.73	67.58 ± 1.34
Neural network	71.56 ± 1.46	68.85 ± 1.23
k-NN	67.83 ± 1.57	63.49 ± 1.44

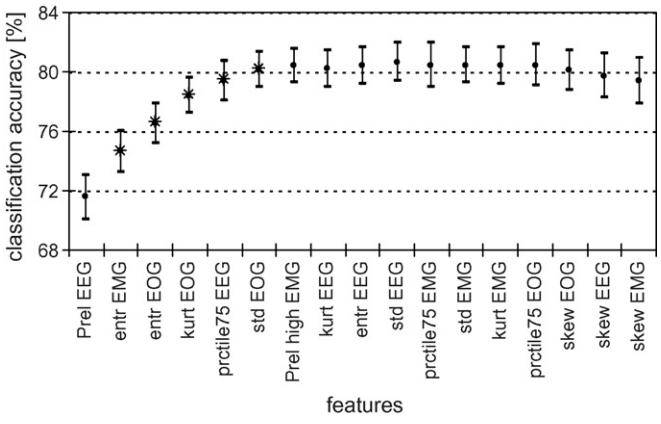

Fig. 2. Selection of features by SFS performed by the neural network classifier.

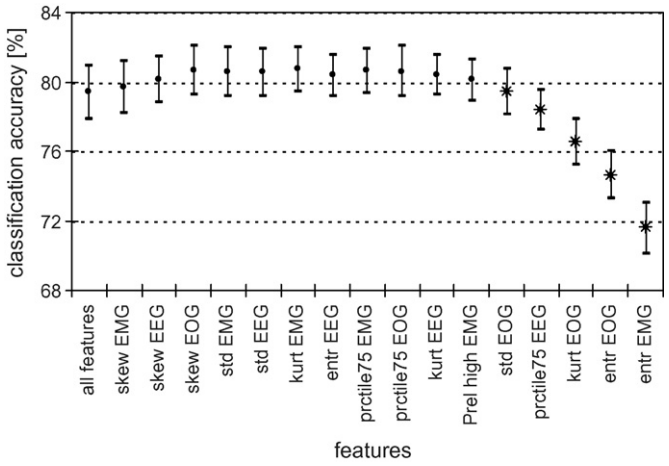

Fig. 3. Selection of features by SBS performed by the neural network classifier.

The results obtained using the SFS method with the neural network classifier are shown in Fig. 2. The dots show the classification accuracy (12), obtained at each step of the feature selection process, the bars express the corresponding standard deviation (13). The axis of abscissas shows the features selected at each step. Steps where the addition of a feature generated a significant increase in the percentage accuracy (t-test; $p < 0.01$) are represented by a star. The most relevant feature is the set expressing the EEG relative power in the five frequency bands, which is able to correctly classify 71% of the epochs. The accuracy is significantly increased (t-test; $p < 0.01$) when the entropy of EMG, the entropy of EOG, the kurtosis number of EOG, the 75th percentile of EEG and the standard deviation of EOG are added (increase from $71.56 \pm 1.46\%$ to $80.11 \pm 1.17\%$). Adding other features does not significantly improve the classification accuracy. The addition of some features, such as skewness numbers, can even diminish it. The optimal set of features is then $\{(P_{\text{rel}}\,\delta, P_{\text{rel}}\,\theta, P_{\text{rel}}\,\alpha, P_{\text{rel}}\,\sigma, P_{\text{rel}}\,\beta)$, entr_{EMG}, entr_{EOG}, kurt_{EOG}, $\text{prctile75}_{\text{EEG}}$ and $\text{std}_{\text{EOG}}\}$. The same set of optimal features was obtained when J was calculated using the quadratic classifier or the k-nearest neighbours classifier. Adding entr_{EMG}, entr_{EOG}, kurt_{EOG}, $\text{prctile75}_{\text{EEG}}$ and std_{EOG} increases the global classification accuracy approximately of about 9% for each classifier.

When applying the SBS method, the same features were detected as relevant for the classification. The selection procedure is shown in Fig. 3. It displays the value of the criterion J at each step of the feature selection process, when one feature is removed from the set. Steps where the removal of a feature shows a significant accuracy decrease are represented by a star. The removal of the first 11 features, from skew_{EMG} to P_{rel} high EMG, does not significantly decrease the value of J. Then, a significant decrease in the classification accuracy is observed when P_{rel} EEG, entr_{EMG}, entr_{EOG}, kurt_{EOG}, $\text{prctile75}_{\text{EEG}}$ and std_{EOG} are removed from the features set.

Table 5 presents the confusion matrix obtained when the optimal set of features is used. The columns represent the stages classified by the neural network classifier and the rows represent the stages determined by the experts. Each case (i, j) corresponds to the number of examples classified as i by both experts and j by the classifier, expressed as a percentage of the examples classified as i by the experts.

Table 5 shows that except for NREM stage I and PS, the incorrect classifications occur between adjacent phases. For example, errors on wakefulness classification are mainly due to a wrong attribution to NREM stage I, which is the phase normally succeeding wakefulness in a non-pathognomonic sleep episode. This can be explained by the fuzzy boundaries between two succeeding sleep stages due to the fact that sleep is a dynamical process. Transitions between two successive stage phases may occur during an epoch or may last longer than 20 s period during which it is difficult for the expert to be certain of his decision. Using data which expertise is the same by two experts diminish the uncertainty area but does not fully eliminate it.

On the contrary, NREM stage I and PS are difficult to be discriminated, though they are not adjacent phases. Yet, Fig. 4 shows that their accuracies increased when adding features extracted from EMG and EOG signals. Indeed, Fig. 4 displays the percentage of correct classification for sleep/wake stages (case(i, i) of the confusion matrix) obtained at each step of the optimal feature selection process (SFS). It can be seen that wakefulness, NREM stage II and SWS stages are correctly classified using EEG spectral information (the accuracy is at least 80%). The addition of new information processed from the two other signals, EMG and EOG, improves the percentage accuracy of these three phases by a few digits only. The increase in the global percentage accuracy (from 71% to 80%) is mainly due to the increased ability of the classifier to discriminate

Table 5
Confusion matrix obtained with the neural network classifier using the optimal set of features

%	Wake	NREM I	NREM II	SWS	PS
Wake	**84.57**	8.13	2.36	1.99	2.95
NREM I	8.47	**64.56**	6.74	0.50	19.73
NREM II	0.79	4.23	**85.55**	7.05	2.38
SWS	0.37	0.06	6.62	**92.90**	0.05
PS	2.33	22.30	2.38	0.18	**72.81**

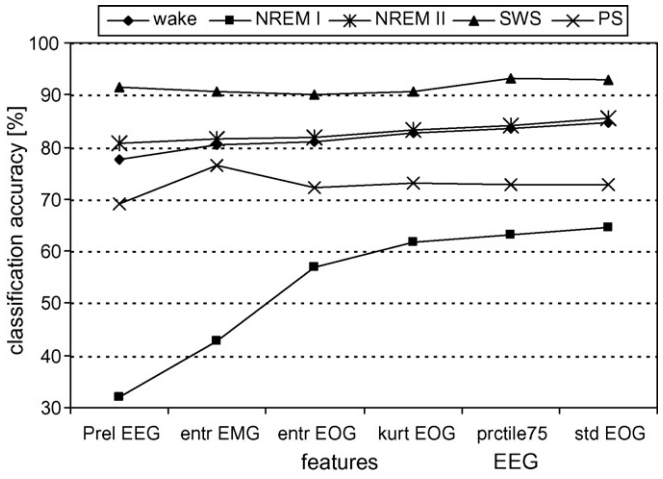

Fig. 4. Classification accuracy of each sleep/wake stage obtained at each step of SFS.

NREM stage I from PS. These two phases are hardly distinguishable by the EEG signal analyzed in the frequency domain. Indeed, when only the EEG spectral information is used, 45% of the NREM stage I epochs are classified into PS. This percentage drops to 20% when the optimal features, processed with EMG and EOG, are used while, at the same time, the percentage of PS wrongly classified as NREM stage I remains the same (about 20%). The percentage of NREM stage I epochs that are correctly classified increases by 32%. The NREM stage I classification accuracy is especially improved by $entr_{EMG}$, $entr_{EOG}$ and $kurt_{EOG}$. The parameter $entr_{EMG}$ improves the accuracy of the PS stage.

5. Discussion

The use of data mining methods enabled knowledge to be extracted from the available database. The methods were implemented in a particular way (different classifiers and several training sets were used) to make it sure that the results obtained were insensitive to the classification method used or to the training set chosen.

Data were selected so that each class was equally represented in the database. The classification accuracy J is then a fair compromise between each class. The selection of features does not favor one class to another one. This is a key point in this study. Indeed, the feature selection procedure showed that wakefulness, NREM stage II and SWS could be correctly classified using EEG spectral information and that the improvement obtained by adding the EOG and EMG signals was not so significant for these stages. Table 1 shows the repartition of the data in the initial database, which is similar to the time spent by a patient in the different sleep/wake stages. About 50% of the night are spent in NREM stage II, while only 3% are spent in NREM stage I. Using a repartition of the data similar to a night sleep would have put a very large weight on NREM stage II and a very small one on NREM stage I, leading to the conclusion that EEG spectral analysis is sufficient to correctly classify a night-time PSG recording. Yet, since PSG is used to diagnose sleep disorders, all sleep/wake stages are to be classified with equal accuracy.

The large amount of data used in this study, recorded on an important number of healthy adults, ensures that the results are not specific to one subject and that they could be extrapolated to new subjects. These methods are data driven; no prior knowledge is introduced in the decision process. Results can then be confronted to the neurophysiologist's point of view.

Results have shown that the EEG relative power spectrum is the most discriminating feature to classify sleep stages, which is common knowledge among neurophysiologists. It seems that the Fourier transform is sufficient to extract relevant spectral information from EEG. Processing the signal with the Wavelet transform does not improve the classification accuracy. This can be explained by the fact that the same information is extracted by these two signal processing methods, with maybe a higher sensitivity of the Wavelet transform to EEG artifacts.

The method showed that EOG and EMG signals are especially important to discriminate PS phase from NREM stage I, which is in agreement with the R&K rules. EEG spectral information is not relevant to discriminate NREM stage I epochs from PS epochs. When only this information is used, many NREM stage I epochs are classified as PS epochs. This is illustrated in Fig. 5 that shows the scatter of NREM stage I and PS stage epochs of the database. The epochs, represented by the relative energies in the five frequency bands, were transformed using Principal Component Analysis, and each point (epoch) was projected on the first two components plane. PS and NREM stage I epochs cannot be separated, as it can be seen in Fig. 5.

From the neurophysiologists' point of view, EEG spectral components are the same in both stages most probably because both of them correspond to activated brain states, though their answer is not very clear about this point. It seems that falling asleep results from an active physiological process, which might be disturbed under particular conditions like sleep onset insomnia. Although not considered in the R&K sleep scoring manual, NREM stage I has been called "Skipped REM" by several authors who observed high frequency EEG activity during this transitional stage.

From SFS, the first two features that increase the classification accuracy, especially NREM stage I and PS accuracies, are the EMG entropy and the EOG entropy. Entropy

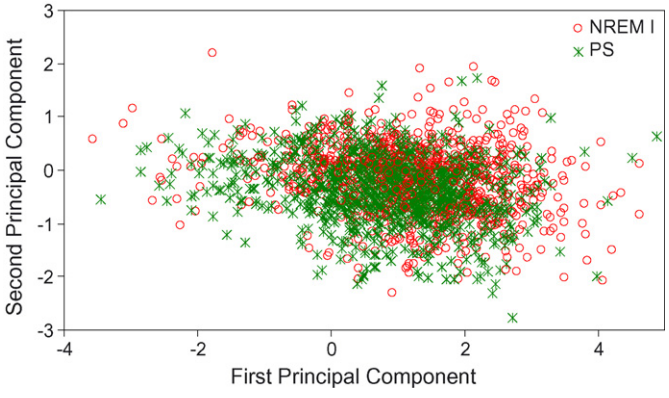

Fig. 5. Principal components analysis of EEG relative power features characterizing stages NREM I and PS.

is a measure of the signal variability: the more variant the signal, the higher the entropy. The selection of the entropy as the most discriminating feature is in agreement with the R&K rules that proposed to discriminate NREM stage I from PS using EMG as an index of muscular tone. During an episode of PS, the patient's skeletal muscles become atonic which is characterized by a flat EMG signal and a low EMG entropy. On the contrary, during NREM stage I, the patient's EMG activity is still elevated and reflected by a high entropy. The opposite information is extracted from EOG entropy. PS phase is characterized by rapid eye movements, which correspond to a high entropy. These rapid eye movements are not observed during the NREM stage I where the entropy is lower.

Fig. 6 shows a plot of NREM stage I and PS stages epochs of the database in two dimensions: EMG entropy, EOG entropy after transformation and normalisation. Though the two plots are superimposed, it is easy to see that a certain number of NREM stage I epochs can be distinguished from PS epochs, which explain the increase of 25% in NREM stage I accuracy without decreasing PS accuracy obtained with the SFS method, when these two features are added.

The kurtosis number of EOG is the third feature to be selected. Kurtosis is a measure of whether the distribution is peaked or flat relative to the normal distribution. The kurtosis of a signal measures the presence of irregular values, such as transitory sharp variations in the signal. Sharp variations related to the presence of rapid eyes movements (REMs) occur in the EOG during PS stage, and explain why EOG kurtosis is higher for some epochs of PS.

The 75th percentile provides some information about the amplitude of the signal. $\mathrm{Prctile75_{EEG}}$ provides an indication on the amplitude level of electrical brain activity and can be useful to discern relatively high amplitude activity during wakefulness and SWS stages.

Finally, let us note that the skewness number of any of the three PSG signals was irrelevant to discern the sleep stages. The reason for this can come from the shape of the physiological signals. The skewness number characterizes the degree of asymmetry of a distribution around its mean value. The PSG signals are more often than not symmetric, occasional signal asymmetries are not specific of any sleep/wake stage.

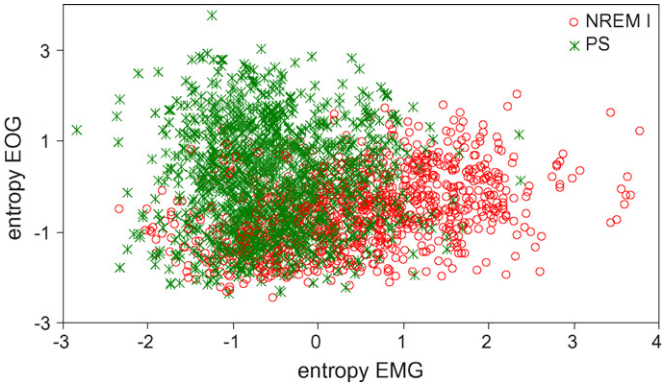

Fig. 6. stages NREM I and PS in a two-dimensional plane: EMG entropy vs. EOG entropy.

The use of two or three PSG signals requires the use of additional sensors, which can increase the patient discomfort during the recording. From this study, it is possible to conclude that EOG and EMG signals must be recorded if the discrimination of NREM stage I from PS is important for the undergoing study or for patient diagnosis purpose. If not, it may not be worth recording three signals, considering the small gain in performance accuracy.

The real challenge in automatic sleep analysis is now to be able to discriminate accurately NREM stage I from PS. Physicians are able to do so using the three PSG signals, EEG, EOG and EMG. Thus, a machine should be able to do so, if the appropriate features are selected, which constitute the future research. The three other stages – wakefulness, NREM stage II and SWS – are already correctly classified. Their classification accuracy is at least 85%. The errors occur on adjacent phases. These errors are due to periods of transitions from one sleep stage to another when it is difficult, even for a human expert, to make a decision.

6. Conclusion

In this study, data mining methods were applied on a large PSG recording database in order to select the most relevant features for sleep/wake stage classification. Methods were processed so as to be insensitive to the classifiers implemented and to the training set used. The results show that an appropriate selection of features improves the classification of sleep/wake stages. Relative power in different EEG frequency bands enables the correct classification of about 71% of the analyzed epoch. Adding the entropy of EMG, the entropy of EOG, the kurtosis of EOG, the 75th percentile of EEG and the standard deviation of EOG improves the classification accuracy by about 9%. These results are in agreement with the R&K rules, which are a standard for human sleep classification.

Future work should be oriented towards an improvement in discrimination of NREM stage I and paradoxical sleep. Indeed, the classification accuracy of these two stages is lower compared to the other sleep stages.

Acknowledgements

Special thanks are expressed to PhiTools (Strasbourg, France) for lending the PRANA software and the sleep recording database.

References

[1] A. Rechtschaffen, A. Kales, A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects, US Government Printing Office, Washington, 1969.
[2] M. Reite, D. Buysse, C. Reynolds, W. Mendelson, The use of polysomnography in the evaluation of insomnia, Sleep 18 (1995) 58–70.
[3] C. Robert, C. Guilpin, A. Limoge, Review of neural network applications in sleep research, J. Neurosci. Methods 79 (1998) 187–193.
[4] E. Oropesa, H.L. Cycon, M. Jobert, Sleep stage classification using wavelet transform and neural network, Int. Comput. Sci. Inst. (1999).

[5] N. Schaltenbrand, R. Lengelle, J.P. Macher, Neural network model: application to automatic analysis of human sleep, Comput. Biomed. Res. 26 (1993) 157–171.
[6] M. Schwaibold, J. Schochlin, A. Bolz, Automated sleep stage detection with a classical and a neural learning algorithm—methodological aspects, Biomed. Tech. (Berlin) 47 (2002) 318–320.
[7] J.D. Frost, An automatic sleep analyser, Electroencephalogr. Clin. Neurophysiol. 29 (1970) 88–92.
[8] J. Fell, J. Röschke, K. Mann, C. Schaffner, Discrimination of sleep stages: a comparison between spectral and nonlinear EEG measures, Electroencephalogr. Clin. Neurophysiol. 98 (1996) 401–410.
[9] M. Grözinger, J. Fell, J. Röschke, Neural net classification of REM sleep based on spectral measures as compared to nonlinear measures, Biol. Cybern. 85 (5) (2001) 335–341.
[10] R. Rosipal, Kernel-based Regression and Objective Nonlinear Measures to Assess Brain Functionnig, PhD thesis, University of Paisley, Scotland, 2001.
[11] G. Dumermuth, D. Lehmann, EEG power and coherence during non-REM and REM phases in humans in all-night sleep analyses, Eur. Neurol. 20 (6) (1981) 429–434.
[12] G. Dumermuth, B. Lange, D. Lehmann, C.A. Meier, R. Dinkelmann, L. Molinari, Spectral analysis of all-night sleep EEG in healthy adults, Eur. Neurol. 22 (1983) 322–339.
[13] T. Kobayashi, et al., Non-linear analysis of the sleep EEG, Psychiatry Clin. Neurosci. 53 (1999) 159–162.
[14] J. Röschke, J. Fell, P. Beckmann, The calculation of the first positive Lyapunov exponent in sleep EEG data, Electroencephalogr. Clin. Neurophysiol. 86 (1993) 348–352.
[15] T. Hastie, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2003.
[16] H.H. Jasper, Appendix to report to Committee on Clinical Examination in EEG: the ten–twenty electrode system of the International Federation, Electroencephalogr. Clin. Neurophysiol. 10 (1958) 371–375.
[17] G. Becq, S. Charbonnier, F. Chapotot, A. Buguet, L. Bourdon, P. Baconnier, Comparison between five classifiers for automatic scoring of human sleep recordings, in: S.K. Halgamuge, L. Wang (Eds.), Studies in Computational Intelligence (SCI), vol. 4: Classification and Clustering for Knowledge Discovery, Springer–Verlag, 2005, pp. 113–127.
[18] J. Mocks, T. Gasser, How to select epochs of the EEG at rest for quantitative analysis, Electroencephalogr. Clin. Neurophysiol. (1984) 89–92.
[19] F. Chapotot, R. Pigeau, F. Canini, L. Bourdon, A. Buguet, Distinctive effects of modafinil and d-amphetamine on the homeostatic and circadian modulation of the human waking EEG, Psychopharmacology 166 (2003) 127–138.
[20] J.W. Clark, The origin of biopotentials, in: J.G. Webster (Ed.), Medical Instrumentation, Application and Design, 3rd ed., John Wiley and Sons Inc., New York, 1998, , pp. 121–182, ISBN: 0-471-15368-0.
[21] R. Moddemeijer, On estimation of entropy and mutual information of continuous distributions, Signal Process. 16 (3) (1989) 233–246.
[22] T. Gasser, P. Bächer, J. Möchs, Transformations towards the normal distribution of broad band spectral parameters of the EEG, Electroencephalogr. Clin. Neurophysiol. (1982) 119–124.
[23] B. Dubuisson, Diagnostic, intelligence artificielle et reconnaissance de formes, Hermès Science Europe, Paris, 2001.
[24] A. Cornuéjols, L. Miclet, Y. Kodratoff, T. Mitchell, Apprentissage artificiel: Concepts et algorithmes, Eyrolles, 2002.