8

Comparison Between Five Classifiers for Automatic Scoring of Human Sleep Recordings

Guillaume Becq[1], Sylvie Charbonnier[2], Florian Chapotot[1], Alain Buguet[4], and Lionel Bourdon[1] and Pierre Baconnier[3]

[1] Centre de Recherches du Service de Santé des Armées, 24 Avenue des Maquis du Grésivaudan, BP 87, 38702 La Tronche cedex, France
guillaume.becq@crssa.net
[2] Laboratoire d'Automatique de Grenoble, Ecole Nationale Supérieure d'Ingénieurs Electriciens de Grenoble, rue de la Piscine, BP 46, 38402 Saint Martin d'Hères, France
Sylvie.Charbonnier@lag.ensieg.inpg.fr
[3] Laboratoire du Traitement de l'Image de la Modélisation et de la Cognition, faculté de médecine, 38700 La Tronche, France
[4] Institut de Médecine Tropicale du Service de Santé des Armées, 13998 Marseille Armées, France

Abstract. The aim of this work is to compare the performances of 5 classifiers (linear and quadratic classifiers, k nearest neighbors, Parzen kernels and neural network) to score a set of 8 biological features extracted from EEG and EMG, in six classes corresponding to different sleep stages as to automatically elaborate an hypnogram and help the physician diagnosticate sleep disorders. The data base is composed of 17265 epochs of 20 s recorded from 4 patients. Each epoch has been classified by an expert into one of the six sleep stages. In order to evaluate the classifiers, learning and testing sets of fixed size are randomly drawn and are used to train and test the classifiers. After several trials, an estimation of the misclassification percentage and its variability is obtained (optimistically and pessimistically). Data transformations toward normal distribution are explored as an approach to deal with extreme values. It is shown that these transformations improve significantly the results of the classifiers based on data proximity.

Key words: Bayesian Classifiers, Error Estimation, Neural Networks, Normalization, Polysomnography, Representation, Sleep Staging

8.1 Introduction

In biology, taxonomy has been the source of numerous studies and still remains one of the predominant fields of research (genome studies). The development of multidimensional exploratory analyses, computing power and numerical solutions

Guillaume Becq et al.: *Comparison Between Five Classifiers for Automatic Scoring of Human Sleep Recordings*, Studies in Computational Intelligence (SCI) **4**, 113–127 (2005)
www.springerlink.com

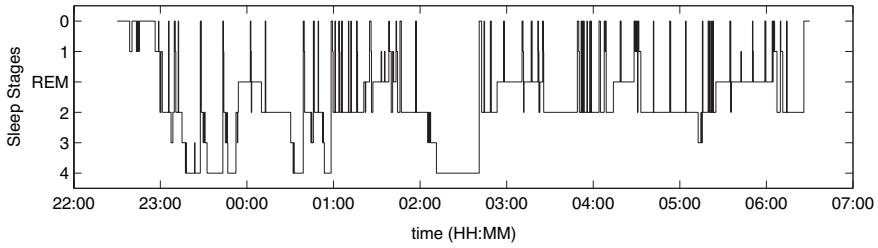


Fig. 8.1. A human hypnogram. Sleep-wake stage scoring has been realized by an expert into 6 different sleep stages from 22 h 30 min to 06 h 30 min over epochs of 20 s: 0-Wake and Movement Time, 1-stage 1 (transition from waking to sleeping), 2-stage 2, 3-stage 3, 4-stage 4 (stage 2, 3 and 4 are part of the orthodox sleep with more and more slow waves observed on the recording), REM–Rapid Eye Movement (or paradoxical) sleep (with rapid brain activity with or without rapid eye movements and muscle atonia)

can explain the growth of such studies. However, in the case of time series, one notices that relative few works have been developed to deal with clustering or classification techniques. One interesting source of such studies is the study of sleep, where several classification techniques have been tested to determine structures on real temporal data [24, 27].

The starting point of sleep studies has been the observation of the electrical activity of the brain measured by electrodes fixed on the scalp, during all night recordings. First observations showed that several patterns were similar from one individual to another, their distributions fluctuating throughout the night. Originally (about 1940) [20, 21], analog signals were plotted on pages of paper. At that time, sleep recordings consisted of huge blocks of paper. With the first discoveries and the evidence of different phases of electroencephalic activity during the night [2, 6, 7, 15], several techniques for electrodes placement were applied and various practices for classifying these activities sprang up. In order to extract the different patterns of such recordings, one expert was assigned to read signals page by page, and give a score corresponding to specific processes of the sleep activity of the brain. The result of this reading has been called an hypnogram and consists in a succession of stages through the night (see Fig. 8.1).

The advantages of working with hypnograms are: an extraction of information from raw data generated by polysomnographic (PSG: multi-channel sleep) recordings, an easier interpretation of the architecture of the night and a better vision of the organization of long term biological processes. It is then easier to discriminate strange charts from normal ones. Therefore, the hypnogram becomes a powerful tool for the diagnosis of sleep pathologies. Besides, the hypnogram, as a summary of the night, considerably reduces the storage of data and allows different laboratories to exchange results and share their knowledge. For that reason, a consensus for a standardization of the rules used to score PSG recordings was held in 1968, bringing together the different leaders in electroencephalography. It led to the creation of the manual by Rechtschaffen and Kales (R&K) [25] currently applied in the different sleep laboratories where pathologies, sleep disorders and untypical hypnograms are studied.

Since 1970, and the growth of computerized methods, interests have been initiated in order to score automatically polysomnographic recordings [11, 29, 30], allowing the expert to avoid spending too much time on this time-consuming work. But studies are still in progress and improvements have to be made. For a complete review of the history of sleep, the reader is referred to [14] where the author, speaking about automatic sleep staging, notes: "The task turned out to be much more difficult because of ambiguities, artifacts and variations in human scoring".

This study has been developed in order to understand the different difficulties encountered with real biological data, while comparing expert practices and machine learning algorithms. For that purpose, a comparative study of five classifiers for automatic analysis of human sleep recordings is presented where temporal data coming from different individuals are mixed together. The interest of transformations toward normal distribution is emphasized since they lead to homogeneous representations for the different selected features. In the first paragraph, the database, the different classifiers, the method chosen to evaluate the performances of the classifiers and the transformations toward normal distribution are presented. In the second paragraph, the results obtained are discussed.

8.2 Materials and Methods

8.2.1 Presentation of the Database

The study has been realized over $N = 11$ polysomnographic recordings available in our database (from 4 healthy subjects). Features were extracted from one EEG (electro-encephalogram, differential lead C3–A2) and one bipolar EMG (electromyogram, position chin), sampled at 200 Hz. The choice of these features has been made in accord with experts in an effort to test a minimal set of electrodes considered necessary for the scoring of sleep.

Eight features thought to represent important physiological processes calculated over epochs of 20 s have been considered and are reported in Table 8.1, where σ denotes the standard deviation and P_{rel} the relative power in a given frequency band. The different bands are: δ (0.5–4.5 Hz), θ (4.5–8.5 Hz), α (8.5–11.5 Hz), σ (11.5–15.5 Hz), β (15.5–22.0 Hz), γ (22.0–45.0 Hz) and corresponds to the ones generally employed in sleep and waking EEG spectral studies [3].

During these epochs of fixed temporal intervals ($\Delta t = 20$ s), EEG can be considered approximately stationary [22]. This assumption is fundamental for the estimation of the different retained features, both in time domain and in the spectral domain. In each epoch, a score has been attributed by an expert. This score is assigned from a set constituted of $K = 6$ classes representing the 6 different stages encountered during human sleep defined in regards with the conventional criteria of R&K [25]: 0-Wake and Movement Time, 1-stage 1, 2-stage 2, 3-stage 3, 4-stage 4, 5-Rapid Eye Movement sleep (or Paradoxical sleep). The different aspects of EEG and EMG signals are represented Fig. 8.2, in order to appreciate the variations of the different signals throughout human sleep.

Once all the signals have been segmented into epochs, preprocessed and their features extracted, we can represent any observation $\mathbf{x}$ by a state representation in an R^d space ($d = 8$ for our study) where $(.)^t$ denotes the transpose of the vector:

Table 8.1. Description of the features used in the study and their statistical values for a) raw data, b) with z-score normalisation and c) with transformations toward normal distribution

		a)				b)		c)		
	Feature	μ	σ	min	max	min	max	Transform.	min	max
F_1	$\sigma(EEG)$	16.87	13.67	4.69	227.31	−0.89	15.40	$\log(1+x)$	−1.97	5.38
F_2	$P_{\text{rel}}(EEG,\delta)$	0.69	0.16	0.01	0.99	−4.27	1.88	$\arcsin(\sqrt{x})$	−4.94	2.58
F_3	$P_{\text{rel}}(EEG,\theta)$	0.14	0.07	0.00	0.68	−1.92	7.46	$\arcsin(\sqrt{x})$	−2.93	5.60
F_4	$P_{\text{rel}}(EEG,\alpha)$	0.05	0.04	0.00	0.46	−1.34	10.65	$\log(\frac{x}{1-x})$	−4.91	2.95
F_5	$P_{\text{rel}}(EEG,\sigma)$	0.05	0.04	0.00	0.50	−1.20	11.54	$\log(\frac{x}{1-x})$	−4.53	2.87
F_6	$P_{\text{rel}}(EEG,\beta)$	0.04	0.04	0.00	0.94	−0.93	23.31	$\log(\frac{x}{1-x})$	−3.62	2.95
F_7	$P_{\text{rel}}(EEG,\gamma)$	0.06	0.10	0.00	1.35	−0.59	13.44	$\log(\frac{x}{1-x})$	−2.95	2.45
F_8	$\sigma(EMG)$	21.42	39.54	0.00	394.97	−0.54	9.45	$\log(1+x)$	−2.10	3.25

$$\mathbf{x} = (F_1, F_2, \ldots, F_d)^t \tag{8.1}$$

When regrouping both the temporal instants and the different individuals, we obtain an array of observations or statistical units [19] representing the database over which the classification study is done:

$$M = \begin{pmatrix} F_1(t_0(1)) & F_2(t_0(1)) & \cdots & F_d(t_0(1)) & C(t_0(1)) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ F_1(t_f(1)) & F_2(t_f(1)) & \cdots & F_d(t_f(1)) & C(t_f(1)) \\ \vdots & & \vdots & & \vdots \\ F_1(t_k(i)) & F_2(t_k(i)) & \cdots & F_d(t_k(i)) & C(t_k(i)) \\ \vdots & & \vdots & & \vdots \\ F_1(t_1(N)) & F_2(t_1(N)) & \cdots & F_d(t_1(N)) & C(t_1(N)) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ F_1(t_f(N)) & F_2(t_f(N)) & \cdots & F_d(t_f(N)) & C(t_f(N)) \end{pmatrix} \tag{8.2}$$

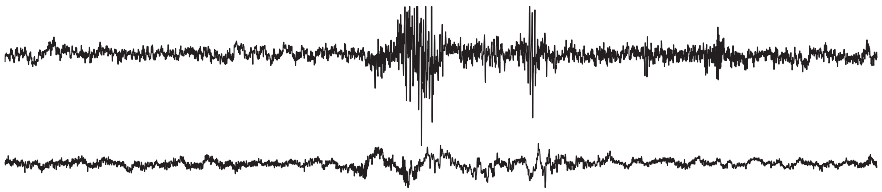
where $t_k(i) = t_0(i) + k\Delta t$ for individual i.

The database was then constituted of 17265 observations over 8 parameters in which we introduced into the last column the expert's classification. A visual display of such a database is given in Fig. 8.3, with features transformed into normal distribution (detailed in 8.2.4) for a better homogeneity of representation.

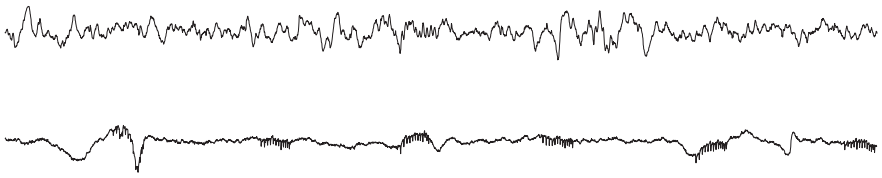
8.2.2 Learning and Testing Sets

In machine learning, the supervised learning approach tries to learn rules, statistics, mathematical models, with a computer, from a desired result. A database containing both the different features used to solve the problem and the corresponding desired results are used. The aim is to find the model that minimizes a criteria which is a function of the difference between the results calculated by the machine and the desired results.

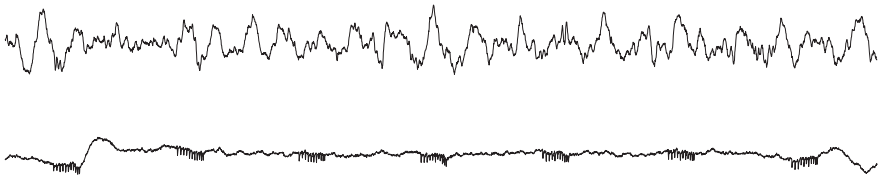
For this reason, it is common to separate the database into 2 sets: the first is used to induce the machine in a so called learning (or training) phase; the second



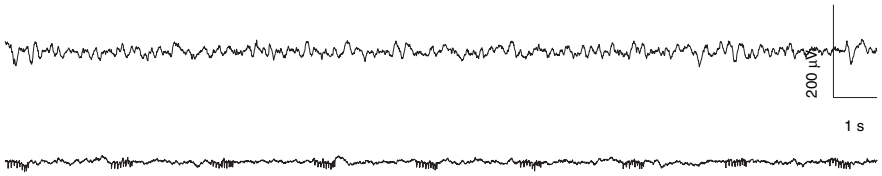
a) Stage 0: Waking and Movement Times, noisy signals.



b) Stage 2: Apparition of short bursting events and slow waves



c) Stage 3–4: Large slow waves.



d) Stage 5: Rhythmic θ activity and flat EMG.

Fig. 8.2. Electrophysiological behaviour during principal sleep stages. Each figure represents an epoch of 20 s. The same scale has been used for all figures

is used during a phase of validation (or test) for evaluating the performance with data that has not been used during the learning process. For a review of the different techniques for evaluating and preparing the data into learning set (LSet) and testing set (TSet) the reader is referred to [9, 10, 17].

Leave one out (N-1 vectors for learning and 1 vector for testing) or classical cross-validation (N/2 vectors for learning, and N/2 vectors for testing), can not be applied when working with a large database, such as ours, without huge computation times. We decided to randomly select a fixed number of data for the learning set and for the testing set, as is done in bootstrap techniques. The learning set will serve to train the classifiers, but also to calculate an optimist estimation (resubstitution techniques, empirical error) of the convergence of them. The testing set is used to obtain a pessimist estimation (cross validation techniques, real error) of them.

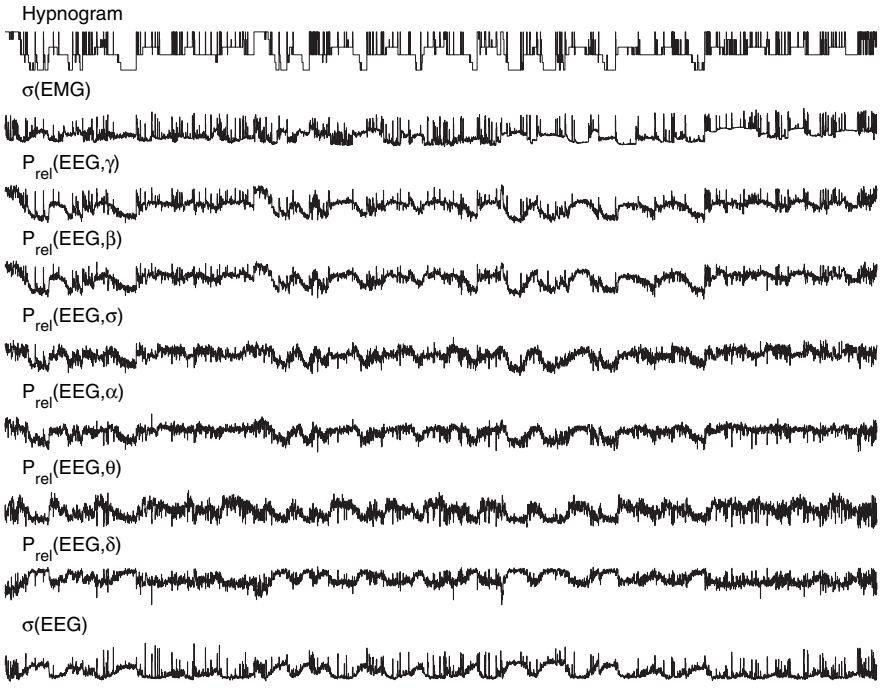


Fig. 8.3. Representation of 5000 elements of the database after transformation. Time and individuals are grouped together. When data is homogeneous, the influence of each feature is directly observable over the classification of the expert represented at the top of the figure

The difference with bootstrap techniques is that we do not reset the drawings after each drawing. This is done in order to obtain independence between estimation from the learning set and estimation from the testing set.

The choice of the number of data for the learning and testing sets can be obtained by looking over the stability of the performances of the classifiers. For a given size of the learning and testing sets, we trained a kNN classifier and a Parzen estimator. An estimation of the performance was realized over 30 subsets for both the optimistic and pessimistic error that are represented given with their standard deviation in Fig. 8.4. Classification errors reach $\approx 30\%$ and do not improve when the number of data in the sets increase over 500 samples when using both the kNN classifier or the Parzen estimator.

8.2.3 The Different Classifiers

Five common classifiers have been evaluated that can be regrouped into two distinct categories: the first category corresponds to the set of classifiers using probabilistic computations based upon the Bayes' rule to assign a class to a feature vector. The second category corresponds to classifiers delimiting regions into the representation

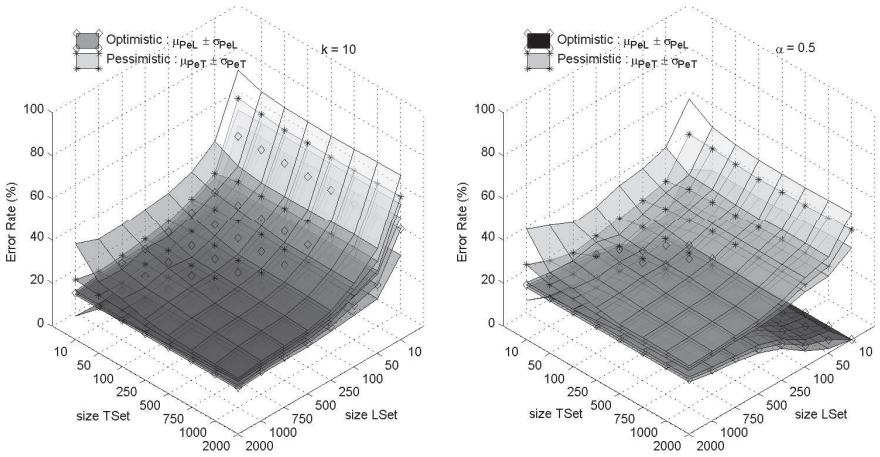


Fig. 8.4. Misclassification percentage in function of the size of the samples drawn from the database using the kNN classifier with $k = 10$ or the Parzen estimator with $\alpha = 0.5$. After 500 samples in the learning set and 500 samples in the testing set, performances are not improved

space by direct computation of frontiers. Explanation of the behavior of the different classifiers and the learning hypothesis can be found in [1, 5, 8]. Here we provide a short description of them.

Our study learning problem is to induce a classifier able to assign to a vector $\mathbf{x}$ of the representation space, a class $C \in \{\omega_i\}_{i=1}^{K}$ with respect to the knowledge database constituted of the data present in the learning set. We use the following notation: P for the probability, p the probability density, $E[.]$ the expectation operator, $|.|$ the determinant.

Bayes Rule-based Classifiers

The attribution of a vector $\mathbf{x}$ to a class is made using the Bayes' rule (8.3). The posterior conditional probability $P(\omega_i|\mathbf{x})$ is calculated for each of the K classes and the vector is given the class ω_i for which $P(\omega_i|\mathbf{x})$ is maximal (maximum a posteriori).

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})} \tag{8.3}$$

$$p(\mathbf{x}) = \sum_{i=1}^{K} p(\mathbf{x}|\omega_i)P(\omega_i) \tag{8.4}$$

The learning problem consists in estimating the conditional density function $p(\mathbf{x}|\omega_i)$ from the different samples of the learning set. The different classifiers depend on the hypotheses made on this density function.

Parametric Models

The probability density function is assumed to be a multidimensional Gaussian model.

$$p(\mathbf{x}|\omega_i) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{8.5}$$

Its parameters (mean $\boldsymbol{\mu}$ and covariance matrices Σ) are estimated with samples drawn from the learning set:

$$\boldsymbol{\mu} = E[\mathbf{x}] = (\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \ldots, \boldsymbol{\mu_d})^t \tag{8.6}$$

$$\Sigma = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] \tag{8.7}$$

Linear classifier: the covariance matrix $\Sigma = \Sigma_i$ is assumed to be the same for all classes. The resulting boundaries delimiting the classes are linear functions.

Quadratic classifier: the covariance matrix Σ_i is assumed to be different for each class and is estimated with representatives of each class in the learning set. The resulting boundaries delimiting the classes are quadratic functions.

Non-Parametric Models

The density function is described with

$$p(\mathbf{x}|\omega_i) = \frac{k_{n_i}}{n_i V_{n_i}} \tag{8.8}$$

with n_i number of representatives of class ω_i in the volume V_{n_i}.

k Nearest Neighbor (kNN) classifier: the probability density function is estimated by the volume occupied by a fixed number of neighbors (search of V_n with fixed k_n). It is simple to show that the decision obtained with the Bayes' rule maximization is equivalent to a voting kNN procedure. This procedure is a majority vote over the classes of the k nearest neighbors (present in the learning set) of the feature vector to classify.

Parzen estimator with Gaussian kernels: The probability density function is estimated by the sum of density kernels given a fixed volume V_n. To each sample $\mathbf{x_{i,j}}$ representative of class ω_i in the learning set, a density kernel $K(.)$ is associated. The sum over j of these n_i kernels gives the density of that class in that region and the probability density function is then

$$p(\mathbf{x}|\omega_i) = \frac{1}{n_i V_{n_i}} \sum_{j=1}^{n_i} K\left(\frac{\mathbf{x}-\mathbf{x_{i,j}}}{h_{n_i}}\right) \tag{8.9}$$

with

$$V_{n_i} = {h_{n_i}}^d = {n_i}^{-\alpha} \tag{8.10}$$

and the Gaussian kernel

$$K(u) = \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{1}{2}(u^2)\right) \tag{8.11}$$

These two methods require the tuning of a parameter: k, the number of neighbors for the k nearest neighbor estimator and α, for the Parzen estimator. The number of neighbors has been chosen to equal 10 and the parameter α has been set to 0.5, after evaluating the performance of the classifiers when incrementing the values of these parameters, as shown in Fig. 8.5.

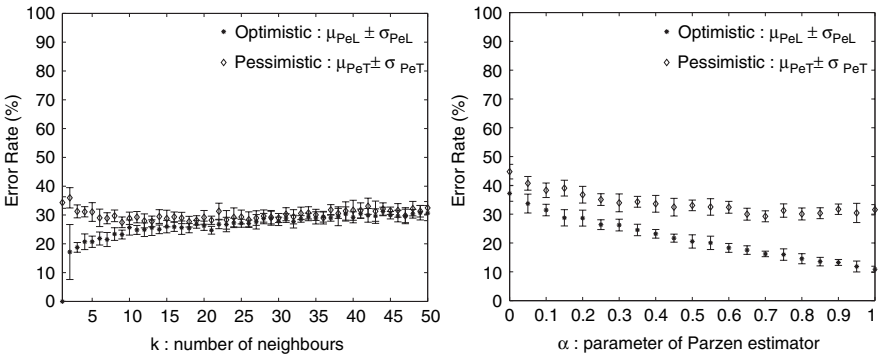


Fig. 8.5. Selection of the parameters for the k Nearest Neighbor classifier and for the Parzen estimator with Gaussian kernels. Size of the learning set and size of the test set have been set to $n_{LSet} = 500$ and $n_{TSet} = 500$. We retained $k = 10$ and $\alpha = 0.5$ for the comparison of the classifier because there is no improvement of the pessimistic error with greater values

Frontiers Based Classifiers

The frontiers of the classes in the multidimensional space are directly calculated from the data present in the learning set:

A multi layer perceptron (MLP): with 3 layers fully connected composed with 8 neurons in the input layer (hyperbolic tangent transfer function $y = 2/(1 + \exp(-2x)) - 1$), 6 neurons in the hidden layer (linear transfer function $y = x$) and 6 in the output layer (logarithmic sigmoid transfer function $y = 1/(1 + \exp(-x))$), trained by the feedforward backpropagation gradient algorithm. Weights were initiated randomly at the beginning of the learning phase. This structure is often used in discrimination [31] with an input layer connected to the representation space of $\mathbf{x}$ with $d = 8$ and the output layer connected to the desired class with $K = 6$. The choice of the number of neurons in the hidden layer has not been optimized in this study.

8.2.4 Transformations Towards Normal Distribution

Means, standard deviations, maximal and minimal values for the different retained features are given in Table 8.1. Inhomogeneity in raw data can be observed, as well as a wide spread of the data, which is typical with biological data. For example, even after doing a z-score defined by the transformation $\mathbf{z} = (\mathbf{x} - \mu)/(\sigma)$ (where μ is the mean of $\mathbf{x}$ and σ is its standard deviation), the maximal value of the sixth parameter is twenty three times the standard deviation.

In order to reduce the influence of extreme values, we applied transformations towards normal distribution on the whole set of the data, and for each parameter. These transformations are either $\log(x)$, $\log(1 + x)$, $\sqrt{x}$, $\sqrt[3]{x}$, $\log((x)/(1 - x))$, $1/(\sqrt{x})$, $\arcsin(\sqrt{x})$ depending on their effect over the different features. This was introduced by Theo Gasser [12] for normalization of EEG spectral parameters. These transformations perform more effectively, better than doing the simple z-score: the

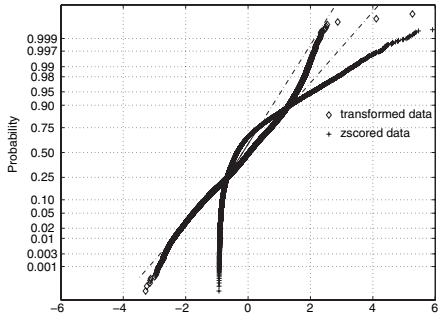


Fig. 8.6a. Effect of the transformations for $P_{\text{rel}}(EEG, \beta)$ plotted in a log-normal axis system, known as Henry plot (normal probability plot). The $\log(\frac{x}{1-x})$ transformation gives a better approximation to a normal distribution represented by a line than a simple z-score

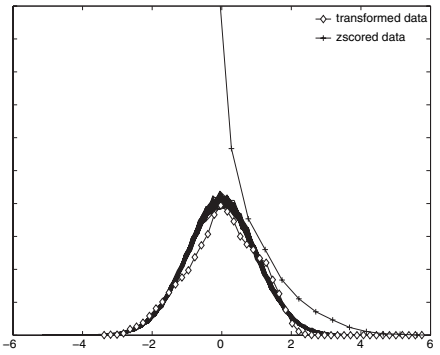


Fig. 8.6b. Effect of the transformation for $P_{\text{rel}}(EEG, \beta)$ over the distribution of the data as compared to z-scored data and data obtained from Gaussian distributions simulated with the same number of realisations

inter-individual variability is reduced with the advantage to reduce tails in distributions. The effects of these transformations can be seen in Table 8.1. The maximal value of the eight parameters after transformation is no more than 6 times the standard deviation. An example of such transformations over one feature can be seen in the Henry plot shown in Fig. 8.6a or from the density plot in Fig. 8.6b.

8.3 Results

A training set and a validation set, each made up of 500 vectors randomly chosen, was built. Each classifier was trained on the first set and applied on the validation set. The performance of the classifier is given by the classification error expressed in percentage on the training set (which is optimistic) and on the validation set (which is pessimistic). This procedure was achieved ten times, which provides two times ten values for each classifier and enables the estimation of mean and variance. The results from one classifier to the other is said to be different if means are statistically different.

8.3.1 Results with Raw Data

Results from raw data are presented in Fig. 8.7a. These correspond to the mean value of the classification error on the training set and on the validation set obtained over 10 trials. For MLP, at each trial, 10 classifiers were trained with a different initialization of the weights, and the network with the minimal classification error was selected, in order to ensure the convergence of the network.

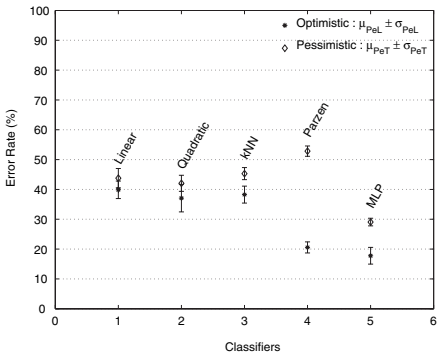


Fig. 8.7a. Means and standard deviations for the misclassification percentage obtained by 10 classifiers of each type on raw data: a) linear discrimination, b) quadratic discrimination, c) kNN with $k = 10$, d) Parzen estimator, with Gaussian kernel and e) Best MLP retained on each trial

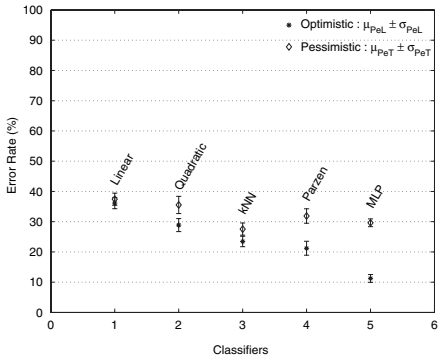


Fig. 8.7b. same as Fig. 8.7a with Transformed Data. All classifiers have improved their results (in terms of pessimistic error), except the neural network which obtains the same results

In Fig. 8.7a, the small standard deviation of the results tends to prove that the technique we use for the evaluation is appropriate. The best result is obtained with the neural network with $29 \pm 1\%$ of misclassification error on the validation set. It is significantly different from others ($p < 0.01$, using a Wilcoxon sign rank test for paired samples). Their results vary from $53 \pm 2\%$ for the Parzen estimator to $42 \pm 3\%$ for the quadratic classifier.

The large difference between optimistic and pessimistic estimation of the percentage for Parzen estimator (classifier d) ($p < 0.01$ Wilcoxon sign rank test for paired samples) shows that the error on the training set can definitely not be used to evaluate the performance of a classifier. Indeed, using this classifier, the vector from the training set participates too much in the decision for its classification. The high percentages obtained for optimistic estimation for the classifiers a) b) and c) shows that those classifiers do not perform well on the data. This can be explained by the large tails in the distribution and by the fact that the density probability functions of the classes cannot be fitted correctly by a multidimensional Gaussian model.

8.3.2 Results with Transformed Data

Results from data with transformations are presented in Fig. 8.7b. The transformations applied to the variables are given in Table 8.1. Classifiers were trained with new coordinates obtained after these transformations.

All classifiers increased their performances ($p < 0.01$, Wilcoxon rank sum test for independent samples) except the neural network which obtains the same results. The performance of the k nearest neighbor classifier and the Parzen classifier have been significantly improved. The pessimistic error decreases from $45 \pm 2\%$ to

$28 \pm 2\%$ for the k nearest neighbor classifier and from $53 \pm 2\%$ to $32 \pm 2\%$ with the Parzen estimator. The results obtained by the k nearest neighbor classifier are then equivalent to the results obtained with the neural network. The results of the linear and quadratic classifiers are slightly better with the transformed data. The misclassification error decreased from $44 \pm 3\%$ to $37 \pm 2\%$ with the linear classifier and from $42 \pm 3\%$ to $36 \pm 3\%$ with the quadratic classifier.

These results can be explained by the fact that the neural network is not sensitive to the distributions of the data; transformations have no effect on its ability to separate space into subspaces [26]. But, the effect of the transformation leads to an improvement of the speed of convergence for the optimization of the backpropagation during the learning process explained by a better homogeneity in the distribution of the weights and features in the input layer.

On the contrary, both the Parzen estimator and the k nearest neighbor estimator use the concept of data proximity to classify a new vector. They are then penalized by extreme values. When the extreme values are moved closer by transformation, their performances equal those of the neural network. The linear and the quadratic classifier make an assumption on the shape of the classes which is still not completely verified even after data transformation.

8.4 Discussion

Disagreements between human scorers are known to vary from 10% to 20% [28]. The results obtained by these classifiers do not enter this interval, but they are not very far from them. Besides, this study enables us to compare different techniques of classification.

The advantage of the neural network is that it does not require any data transformation. The results obtained are the same with the raw data or with the transformed data. The neural network can deal with a non-Gaussian probability density function and with extreme values. However, the selection of the best neural network and the optimization of the structure of the layers are not easy tasks and can be time consuming. Though the results obtained by the nearest neighbor classifier applied to homogeneous data are the best, this method requires storing a large amount of learning vectors in memory. This can make its application difficult in practice. The main advantage of classical statistical techniques (linear and quadratic discrimination) is that the algorithms are fast.

Why do results not enter the inherent interval of disagreement between scorers? One answer is that the hypothesis of the independence of the temporal epochs is not completely true because when experts score a recording, they intrinsically know the preceding page and score the new one in consequence. A way to take into account this temporal causality is to add new columns in the database corresponding to the preceding data of the parameters (switching from a state representation to a phase one). Another way is to introduce an inference table at the end of the classification process allowing certain transitions or rejecting others.

Another answer is that the parameters retained are not as discriminative as the ones chosen visually by an expert. This is a problem which is generally encountered in automatic classification as a means to replace a human classifier. Moreover, the classifiers were trained to classify data recorded on different subjects. They had to

deal not only with temporal dependence of the data as discussed before, but also with inter-individual variability. In our study, we have constructed and evaluated classifiers with no adaptation to one particular individual.

Experts are not so strict and often adapt their mind to fit the problem, but classifiers are built to fit optimized mathematical models from a learning database. The solutions proposed by these models can sometimes show the limits of the visual technique of human scoring and can be a way to refine expert knowledge.

For example, one visual interesting dilemna in the R&K manual is when a transition occured during an epoch from one stage to another: the rule is to assign the class to the predominant stage, i.e. the stage that lasts more than fifty percent of the epoch. When there are a lot of transitions, this results in many problems for the scorer and a lower productivity. This also raises doubt over the stationary hypotheses for the computing of temporal or spectral parameters. For more accurate precision, one can use recent segmentation techniques for temporal time series, where signals are segmented into non overlapping windows of variable lengths with respect to different criteria [4, 16, 18, 23]. But then, the estimation of the performance of the classifiers is not so easy.

Nowadays, R&K scoring proves its usefulness every day, but its limits are more and more admitted [13].

8.5 Conclusion

We have evaluated and compared the performance of five classifiers to automatically score polysomnographic data from various individuals into the six R&K sleep-wake stages. Though the results obtained (the misclassification percentage is about 30%) are not as good as the results obtained with the human scorers (the misclassification percentage is less than 20%), the results are interesting considering the amount of work human scoring requires. Automatic scoring may lighten the doctor's burden.

We showed that extreme values, frequently present in biological data, were a problem for all the evaluated classifiers, except for the neural network. To apply a transformation toward normal distribution appeared to be an interesting way to improve the performance of the classifiers. Both the neural network and the k nearest neighbor algorithm using transformed data gave good results. However, considering the information required to implement the two methods, we would recommend the use of the neural network.

Acknowledgments

The authors wish to thank Bénédicte Becq and Geoffrey Bramham for their unvaluable help with subtleties of the English language. They also want to thank the staff of the CRSSA/FH/PV division, who were present during this study, for their kind support.

References

1. Artioli E, Avanzolini G, Barbini P, Cevenini G, Gnudi G (1991) Classification of postoperative cardiac patients: comparative evaluation of four algorithms. Int J Biomed Comput 29:257–270
2. Aserinsky E, Kleitman N (1957) Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. Science 118:273–274
3. Chapotot F, Pigeau R, Canini F, Bourdon L, Buguet A (2003) Distinctive effects of modafinil and d-amphetamine on the homeostatic and circadian modulation of the human waking EEG. Psychopharmacology 166:127–138
4. Charbonnier S, Becq G, Biot L (2004) On-line segmentation algorithm for continuously monitored data in intensive care units. IEEE T Biomed Eng 51(3):484–492
5. Cornuéjols A, Miclet L (2002) Apprentissage artificiel, Concepts et algorithmes. Eyrolles Paris
6. Dement WC, Kleitman N (1957) Cyclic variations in EEG during sleep and their relation to eye movements, body motility and dreaming. Electroencephalogr Clin Neurophysiol 9:673–690
7. Dement WC (1958) The occurence of low voltage fast electroencephalogram patterns during behavioral sleep in the cat. Electroencephalogr Clin Neurophysiol 10:291–296
8. Dubuisson B (2001) Diagnostic, intelligence artificielle et reconnaissance de formes. Hermès science Europe Paris
9. Efron B (1983) Estimating the error rate of a prediction rule: improvement on crossvalidation. J Am Stat Ass 78(382):316–330
10. Efron B, Tibshirani R (1995) Crossvalidation and the bootstrap: Estimating the error rate of a prediction rule. Technical report (477) Statistics department Stanford University
11. Frost JD (1970) An automatic sleep analyser. Electroencephalogr Clin Neurophysiol 29:88–92
12. Gasser T, Bächer P, Möchs J (1982) Transformations towards the normal distribution of broad band spectral parameters of the eeg. Electroencephalogr Clin Neurophysiol 53:119–124
13. Himanen SL, Hasan J (2000) Limitations of Rechtschaffen and Kales. Sleep Med Rev 4(2):149–167
14. Hirshkowitz M (2000) Standing on the shoulders of giants: the *Standardized Sleep Manual* after 30 years. Sleep Med Rev 4(2):169–179
15. Jouvet M, Michel F, Courjon J (1959) Sur un stade d'activité électrique cérébrale rapide au cours du sommeil physiologique. C R Soc Biol 153:1024–1028
16. Keogh E, Chu S, Hart D, Pazzani M (2001) An online algorithm for segmenting time series. In: IEEE International Conference on Data Mining, pp. 289–296
17. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence, pp. 1137–1145.
18. Kohlmorgen J, Müller K, Rittweger J, Pawelzik K (2000) Identification of nonstationary dynamics in physiological recordings. Biol Cyber 83(1):73–84
19. Lebart L, Morineau A, Piron M (2000) Statistique exploratoire multidimensionnelle. Dunod Paris
20. Loomis AL, Harvey EN, Hobart G (1937) Cerebral stages during sleep, as studied by human brain potentials. J Exp Psychol 21:127–144

21. Loomis AL, Harvey EN, Hobart G (1938) Distribution of disturbance patterns in the human electroencephalogram, with special reference to sleep. J Neurophysiol 1:413–418
22. Mocks J, Gasser T (1984) How to select epochs of the eeg at rest for quantitative analysis. Electroencephalogr Clin Neurophysiol 58:89–92
23. Morik K (2000) The Representation race – preprocessing for handling time phenomena. 11th European Conference on Machine Learning 1810:4–19 Springer Berlin
24. Penzel T, Conradt R (2000) Computer based sleep recording and analysis. Sleep Med Rev 4(2):131–148
25. Rechtschaffen A, Kales A (1968) A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. US Government Printing Office Washington
26. Robert C, Guilpin C, Limoge A (1997) Comparison between conventional and neural network classifiers for rat sleep-wake stage discrimination. Neuropsychobiol 35:221–225
27. Robert C, Guilpin C, Limoge A (1999) Automated sleep staging systems in rats. J Neurosci Meth 88:111–122
28. Schaltenbrand N, Lengelle R, Toussaint M, Luthringer R, Carelli R, Jacqmin A, Lainey E, Muzet A, Macher JP (1996) Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients. Sleep 19(1):27–35
29. Smith JR, Negin M, Nevis AH (1969) Automatic analysis of sleep electroencephalograms by hybrid computation. IEEE T Syst Sci Cybern 5:278–284
30. Smith JR, Karacan I (1971) EEG sleep stage scoring by an automatic hybrid system. Electroencephalogr Clin Neurophysiol 31:231–237
31. Thiria S, Lechevallier Y, Gascuel O, Canu S (1997) Statistique et méthodes neuronales. Dunod, Paris