doi: 10.1111/j.1365-2869.2008.00706.x

Algorithms for sleep–wake identification using actigraphy: a comparative study and new results

JOËLLE TILMANNE[1], JÉRÔME URBAIN[1], MAYURESH V. KOTHARE[2], ALAIN VANDE WOUWER[3] and SANJEEV V. KOTHARE[4]

[1]Service de Théorie des Circuits et Traitement du Signal, Faculté Polytechnique de Mons, Mons, Belgium, [2]Department of Chemical Engineering, Lehigh University, Bethlehem, PA, USA, [3]Service d'Automatique, Faculté Polytechnique de Mons, Mons, Belgium and [4]Sleep Center for Children, Children's Hospital, and Department of Neurology, Harvard Medical School, Boston, MA, USA

Accepted in revised form 22 September 2008; received 12 March 2008

SUMMARY The aim of this study was to investigate two new scoring algorithms employing artificial neural networks and decision trees for distinguishing sleep and wake states in infants using actigraphy and to validate and compare the performance of the proposed algorithms with known actigraphy scoring algorithms. The study employed previously recorded longitudinal physiological infant data set from the Collaborative Home Infant Monitoring Evaluation (CHIME) study conducted between 1994 and 1998 [http://dccwww.bumc.bu.edu/ChimeNisp/Main_Chime.asp; *Sleep* **26** (1997) 553] at five clinical sites around the USA. The original CHIME data set contains recordings of 1079 infants <1 year old. In our study, we used the overnight polysomnography scored data and ankle actimeter (Alice 3) raw data for 354 infants from this data set. The participants were heterogeneous and grouped into four categories: healthy term, preterm, siblings of SIDS and infants with apparent life-threatening events (apnea of infancy). The selection of the most discriminant actigraphy features was carried out using Fisher's discriminant analysis. Approximately 80% of all the epochs were used to train the artificial neural network and decision tree models. The models were then validated on the remaining 20% of the epochs. The use of artificial neural networks and decision trees was able to capture potentially nonlinear classification characteristics, when compared to the previously reported linear combination methods and hence showed improved performance. The quality of sleep–wake scoring was further improved by including more wake epochs in the training phase and by employing rescoring rules to remove artifacts. The large size of the database (approximately 337 000 epochs for 354 patients) provided a solid basis for determining the efficacy of actigraphy in sleep scoring. The study also suggested that artificial neural networks and decision trees could be much more routinely utilized in the context of clinical sleep search.

KEYWORDS actigraphy, artificial neural networks, decision trees, sleep diagnosis, sleep–wake scoring

Correspondence: Dr Sanjeev V. Kothare, Sleep Center for Children, Children's Hospital, and Department of Neurology, Harvard Medical School, Fegan 9, 300 Longwood Avenue, Boston, MA 02115, USA. Tel.: (617) 355 6663; fax: (617) 730 0463; e-mail: sanjeev.kothare@childrens.harvard.edu

At the time of this study Dr Sanjeev V. Kothare was with the Section of Sleep Medicine, Division of Neurology, Department of Pediatrics, St Christopher's Hospital for Children, Drexel University College of Medicine, Philadelphia, PA, USA.

INTRODUCTION

Polysomnography (PSG), the gold standard for evaluating sleep disorders, suffers from several drawbacks: it is cost and labor-intensive, usually performed in a sleep laboratory and, due to the many sensors and wires placed on the patient, can disrupt the very sleep architecture it is designed to measure (so-called 'first night effect'). As indicated by several recent reviews and the newly updated practice parameters of the American

Academy of Sleep Medicine (AASM) (Ancoli-Israel *et al*., 2003; Littner *et al*., 2003; Morgenthaler *et al*., 2007; Thorpy *et al*., 1995), actigraphy has been widely recognized as a low cost alternative to conventional PSG for screening of sleep disorders, with special emphasis on sleep–wake cycles and more specifically, insomnia.

Actigraphy is based on the fundamental premise that the presence of movements indicates wakefulness and the absence of movements indicates sleep. Actigraphs or actimeters are miniature computerized wristwatch-like devices, most commonly worn on the wrist or ankle, which allow for up to several weeks of continuous recording of limb activity from which sleep and wake can be scored. Although actigraphy provides much less information than a full PSG and only measures sleep indirectly, there are a number of scenarios where it is particularly suitable due to its low cost and non-invasive characteristics, as documented in Littner *et al*. (2003), Ancoli-Israel (2000) and references therein. In particular, for infants below the age of 1 year when electroencephalogram (EEG) patterns are not well-established, actigraphy can provide a viable tool for screening of sleep–wake cycles as a measure for assessing insomnia/hypersomnia (Kevin *et al*., 2007).

Since the publication of the first automatic scoring algorithm for actigraphy by Webster *et al*. (1982), various computer algorithms have been developed to automatically score sleep and wake from the recorded raw actigraphy movement data (see Ancoli-Israel *et al*., 2003; Littner *et al*., 2003; Morgenthaler *et al*., 2007; Thorpy *et al*., 1995 for an extensive bibliography). Notable among the reported actigraphy algorithms are the ones developed by Sadeh *et al*. (1989, 1994, 1995) covering not only adult normal and abnormal subjects, but also infants, the algorithm of Cole *et al*. (1992), the Actigraph Data Analysis Software (ADAS) by Jean-Louis *et al*. (1996), the algorithm reported by Sazonov *et al*. (2004) and various commercially available algorithms accompanying the different actimeters available in the market.

A variety of new actimeter devices have been introduced into the market over the years, and each device must be appropriately calibrated and validated against PSG to give meaningful sleep–wake scores. However, unlike PSG that uses the Rechtschaffen and Kales (1968) rules to get standardized scoring, no such consensus currently exists for scoring actigraphic signals. Part of the difficulty appears to lie in the metric used to evaluate the quality of scoring. Many studies use the agreement rate (defined as 'accuracy' in the next section) to compare the quality of an actigraphy algorithm against the reference PSG. However, as discussed in de Souza *et al*. (2003), accuracy alone may provide misleading conclusions. For example, in studies with healthy subjects who have normal night sleep, even if the whole night of actigraphy recording is scored as sleep without detecting any wake epoch, one can get accuracy as high as 92% as pointed out in Sadeh *et al*. (1989). Thus evaluation parameters beyond agreement rate, such as sensitivity and specificity (de Souza *et al*., 2003), must be considered to fully understand the performance of a scoring algorithm.

To a large extent, the aforementioned problems arise from the difficulty in gathering large amount of informative data covering a wide spectrum of subjects to provide statistically significant and convincing validation of actigraphic scoring algorithms. Most studies typically use recordings from 10 to 50 patients. The limited availability of data is an even greater problem in the infant age group.

The objectives of our work are: (i) to develop two new pattern recognition algorithms for scoring of actigraphy data with the ability to model a possibly nonlinear classification border between sleep and wake states; (ii) to employ comprehensive objective criteria to evaluate the quality of our proposed algorithms; (iii) to validate our proposed algorithms on a clinical data set larger than has been reported in the literature for actigraphy studies; (iv) to compare the performance of our proposed scoring algorithms with previously reported actigraphy algorithms, under a range of possible training and validation scenarios to illustrate the inherent trade-offs in the scoring algorithms; and (v) to present the new techniques in an accessible and tutorial fashion for the clinical sleep research community to readily appreciate the broader utility of the techniques.

This paper is organized as follows. The following section introduces the Collaborative Home Infant Monitoring Evaluation (CHIME) database, the several quality indicators that will be used to assess the performance of the actigraphy scoring algorithms, and the algorithms of Sadeh *et al*. (1994) and Sazonov *et al*. (2004). Next, the proposed pattern recognition methods, the multilayer perceptron (MLP) and the decision tree are introduced in a tutorial fashion. Results section is dedicated to the analysis of the scoring results and the comparative evaluation of the various algorithms. Discussion section presents the interpretation of our results relative to published results, and finally in the Conclusions section, we draw conclusions from our study along with a discussion of possible directions for future research.

METHODS

The CHIME database

The data we used were recorded by a multicenter collaborative group of sleep laboratories for the CHIME study (CHIME-website, 2004; Crowell *et al*., 1997) between 1994 and 1998. The aim of the CHIME study was to answer questions about the possible role of home infant monitors in preventing sudden infant death syndrome (SIDS). The recordings consisted of full overnight PSG (21 channels) for over 1000 infants below 1 year of age, with gestational ages in the range of 23–42 weeks. The recording also contained the binary signal of a Healthdyne actimeter placed on the ankle, with ones representing movement and zeros standing for no movement, at a sampling frequency of 2 Hz.

We had 438 files at our disposal that contained the reference PSG scoring necessary for our study. From these files, 84 had to be discarded because of one or more of the following

reasons: (i) portions of data missing; (ii) incorrect time vector that was incompatible with the PSG recording; (iii) flat (zero) or non-binary signal from the actimeter; (iv) visually grossly incorrect actimeter readings based on a comparison with reference scoring.

The actimeter signal of the 354 remaining files was divided into 30-s epochs (with no overlap). We defined the activity value of one epoch as the sum of its actigraphy samples. As the actigraphy signal is binary and sampled at 2 Hz, the activity of a 30-s epoch is an integer between 0 and 60. Reference PSG scoring of each 30-s epoch was previously recorded by the CHIME staff in the data files in the form of one of the four sleep–wake states (quiet sleep, active sleep, indeterminate sleep[1] and wake) (see Clancy *et al.*, 2003). As the aim of our study was to distinguish sleep from wake, and not to recognize the sleep stages, we translated these scores to 'wake' (for the wake state) or 'sleep' (for the three other states). The reasons for this choice are the following: (i) our goal is to design algorithms generalizable to children and adults who present different sleep stages than infants; (ii) the distinction between sleep and wake without further classification of sleep stages is sufficiently useful for a number of clinical applications such as monitoring insomnia, circadian disorders, effects of treatments, etc.; (iii) the 2-Hz binary signal provided by the actimeters used in this study is limited and unlikely to enable a reliable classification amongst the sleep stages.

The resulting database of 354 infants consisted of 336 958 epochs, 70% of which were reference scored as sleep and 30% as wake. Of the 354 infants, 80 were healthy term, 125 were preterm, 69 were siblings of SIDS and 61 were infants with apparent life-threatening events (apnea of infancy). At the time of the recording, 161 were younger than 2 months, 92 between 2 and 3 months, 80 between 3 and 6 months, and two over 6 months. The remaining 19 infants did not have their health status and age data in their files, but were all <12 months old.

Metrics for evaluation of actigraphy algorithms

Several measures can be used to evaluate the performance of an algorithm. For classification problems, the confusion matrix (Kohavi and Provost, 1998) is a typical evaluation tool and evaluation measures can be directly computed from this matrix. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class (Kohavi and Provost, 1998). The confusion matrix used in our case, taking the sleep class as a positive result, is represented in Table 1.

Various ratios computed from this confusion matrix are very useful in evaluating and understanding the performance of an algorithm.

[1]Indeterminate sleep was scored when the infant was undoubtedly sleeping but the sleep stage, quiet or active, was uncertain.

Table 1 Confusion matrix

	Predicted class	
Actual class	*Sleep*	*Wake*
Sleep	TP	FN
Wake	FP	TN

TP, true positive; FN, false negative; FP, false positive; TN, true negative.

Accuracy (Acc). This is the proportion of objects that are correctly classified by the algorithm. Accuracy is also often called 'agreement rate' in the literature.

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \quad (1)$$

Sensitivity (Sen). This is the proportion of actual positive objects that are correctly predicted as positive.

$$\mathrm{Sen} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (2)$$

Specificity (Spe). This is the proportion of actual negative objects that are correctly predicted as negative.

$$\mathrm{Spe} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \quad (3)$$

In our case, sensitivity will indicate the efficiency of an algorithm to detect the sleep epochs, while specificity will represent its ability to detect wake epochs.

Positive predictive value (PPV). This is the probability that an object classified as positive (sleep in our case) by the algorithm is actually positive.

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (4)$$

Negative predictive value (NPV). This is the probability that an object classified as negative is actually negative.

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \quad (5)$$

As discussed in Introduction, the agreement rate (or 'accuracy') alone may not fully capture the quality of an actigraphy algorithm, and in fact could result in unrealistically good but misleading outcomes (de Souza *et al.*, 2003). Additional parameters, such as the ones defined above, provide a more comprehensive evaluation of the performance and trade-offs inherent in actigraphy algorithms. We will primarily use accuracy, specificity and sensitivity in our evaluations but we will also report PPV and NPV as additional information.

Linear combination methods

In this section, we will describe how one can apply two previously reported actigraphy algorithms (Sadeh *et al.*, 1994; Sazonov *et al.*, 2004) to the CHIME data set. These two

algorithms use linear combinations of features computed from the actigraphic signal to distinguish sleep from wake. As the original algorithms in Sadeh *et al*. (1994) and Sazonov *et al*. (2004) were optimized for different actimeters than the one used in our chosen data set, we will recalculate the parameters in these algorithms using the CHIME data set before testing them for their predictive capabilities. The results of applying these algorithms to our chosen data set are presented in the Results section.

Parameter optimization

To find the best set of parameters in a linear combination algorithm during the supervised[2] training phase, an appropriate cost function is minimized or maximized. One classical cost function that can be minimized is the sum of the squares of the errors (SSE) between the output of the algorithm and the known or true output. If there are N epochs in total, and if t_i and p_i are respectively the target (true or PSG scored, 0 for wake, 1 for sleep) and the predicted values of the ith epoch, the parameters of the classification function will be chosen so as to minimize the sum of the squares of the prediction errors (SSE):

$$\mathrm{SSE} = \sum_{i=1}^{N} (t_i - p_i)^2 \tag{6}$$

However, the high proportion of sleep epochs in our training database is likely to lead to algorithm parameters that overpredict sleep, due to an excessive weighting on sleep epochs. This typically leads to a high accuracy value but generally low specificity compared with sensitivity. In order to reach a better trade-off between sensitivity and specificity and to improve the detection of wake epochs, we have chosen to maximize an alternate cost function, the sum of sensitivity and specificity (SESP), in this training phase.

Sadeh's algorithm

Sadeh *et al*. (1994) developed an algorithm suited for a particular wrist actigraph (Ambulatory Monitoring, Ardsley, NY, USA). They created a discriminant function by first identifying the five-most efficient actigraphy-derived variables and then carrying out an analysis to propose the following discriminant function:

$$\mathrm{SI} = 7.601 - 0.065\mu - 0.056\sigma - 0.0703\mathrm{LogAct} - 1.08\mathrm{nat}, \tag{7}$$

where SI is the sleep indicator of the current epoch (if SI ≥ 0, the current epoch is classified as sleep); μ is the mean activity on a 11-min window centered on the current epoch; σ is the standard deviation of activity for the last 6 min; LogAct is the natural logarithm of the activity of the current epoch increased by 1 and nat is the number of epochs that satisfy the criterion $50 \leq$ epoch activity < 100 in an 11-min window centered on the current epoch.

The algorithm was evaluated on 16 healthy children and adolescents and gave accuracy, sensitivity and specificity of 91.16%, 94.95% and 74.5% respectively.

Sadeh's algorithm was designed for 1-min epochs, because the actimeter was set in the 'zero-crossing' mode and the activity count was calculated and scored with a period of 1 min. As the PSG standard is 30-s epochs, Sadeh's original algorithm can be compared with reference PSG scores by applying the following rule: if at least one of the two adjacent 30-s epochs was scored as wake by PSG, the corresponding 1-min epoch is reference scored as wake, otherwise as sleep.

Although not included in the original algorithm of Sadeh, the output SI from equation (7) can be transformed into the standard range 0–1 to represent the probability of sleep (PS) by passing it through a sigmoid function (Fig. 1):

$$\mathrm{PS}\ (\mathrm{SI}) = \frac{1}{1 + \mathrm{e}^{-\mathrm{SI}}} \tag{8}$$

Thus, if PS < 0.5, the epoch is classified as wake and as sleep otherwise.

While Sadeh's algorithm was derived for a different actimeter, the discriminant variables that it identified are of interest for actigraphy, independent of the actimeter used. To verify this, we apply Sadeh's unaltered algorithm using equation (7) (appropriately scaled for the magnitude of the actimeter reading) and also apply Sadeh's algorithm optimized for the values of the parameters in equation (7) specifically to match the CHIME data set. The results for both of these scenarios are presented in the Results section.

Sazonov's algorithm

Using the data from the CHIME study, but from another actimeter placed on the diaper of each infant, Sazonov *et al*.

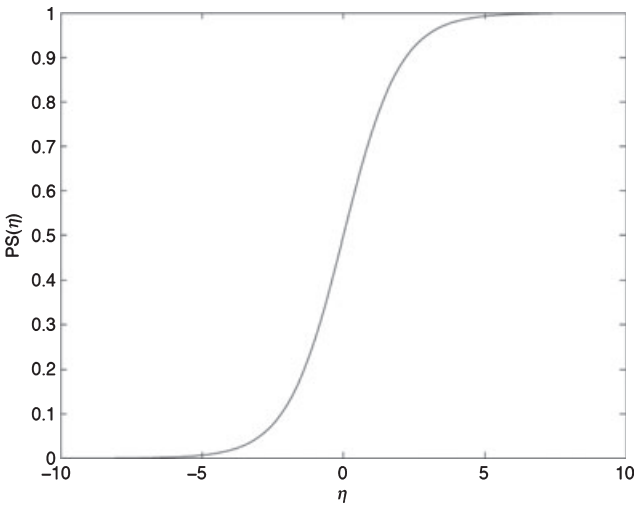

Figure 1. Sigmoid function.

[2]*Supervised* training is a training phase during which objects are presented to the algorithm together with their desired output (in our case, 1 for sleep and 0 for wake).

(2004) developed an algorithm for sleep–wake scoring using data from 26 CHIME subjects. The signal used is named the 'position' signal in the CHIME study and is a continuous measurement of acceleration, sampled at 50 Hz. This signal is very different from the actimeter signal we are using for our study, which is a binary signal from the ankle of the infant, sampled at 2 Hz.

The proposed algorithm of Sazonov, using the 'position' signal, was based on using the activity values of the current epoch and eight preceding epochs in a logistic regression. A linear combination of the chosen features was passed through a sigmoid function (equation 10, Fig. 1) to obtain a PS as follows:

$$\begin{aligned}\eta = {} & 1.99604 - 0.1945\text{maxACC}_0 - 0.09746\text{maxACC}_{-1} \\ & - 0.09975\text{maxACC}_{-2} - 0.10194\text{maxACC}_{-3} \\ & - 0.08917\text{maxACC}_{-4} - 0.08108\text{maxACC}_{-5} \\ & - 0.07494\text{maxACC}_{-6} - 0.073\text{maxACC}_{-7} \\ & - 0.10207\text{maxACC}_{-8}\end{aligned} \tag{9}$$

$$\text{PS}(\eta) = \frac{1}{1 + \text{e}^{-\eta}}, \tag{10}$$

where maxACC_{-i} is the maximum of the position signal in the epoch located i epochs before the current epoch. PS is the predicted probability of sleep.

Sazonov *et al.* (2004) reported the following performance with their algorithm: 75.4% accuracy, 93.2% sensitivity, 41% specificity.

As with Sadeh's algorithm, we adapted Sazonov's algorithm, originally developed for the 'position' signal, to the 'actimeter' signal in our study. As the 'actimeter' signal is binary, the maximum of the signal for each epoch is not meaningful. Instead, we used the activity signal, which is the sum of the samples of the actimeter signal for each epoch to evaluate Sazonov's algorithm.

Once again, as with Sadeh's algorithm, and in the spirit of validating if the chosen variables in Sazonov's algorithm are indeed generally useful for actigraphy, we applied Sazonov's unaltered algorithm using equation (9) (appropriately replacing maxACC with the activity) and, in addition, we also optimized the values of the parameters in equation (9) specifically to match the CHIME actimeter signal for our chosen data set of 354 subjects. The results for both of these scenarios are presented in the Results section.

Pattern recognition methods

We propose the use of two pattern recognition algorithms, namely, artificial neural networks (ANNs) and decision trees, to perform actigraphy-based sleep–wake classification. There are a number of reasons why the use of these techniques from the pattern recognition literature is appropriate in our context: (i) they are input/output modeling techniques that build the model exclusively from data (data-driven); (ii) they are particularly well suited for scenarios where the system cannot be completely characterized by first principle models, while empirical evidence suggests that the input/output system behavior is nonlinear; (iii) they are capable of generating possibly nonlinear classification borders between two or more classes; (iv) classification functions of the kind that maps actigraphy-derived variables to PS scores can, in theory, be approximated with arbitrary precision using neural network models of appropriate complexity ('universal approximation theorem for neural networks'; Lloyd, 2003). Similarly, decision trees with sufficiently many nodes and leaves can provide comparable accuracy (Breslow and Aha, 1997).

Artificial neurons

The idea of artificial neurons and neural networks originated from the notion that emulating brain functions and neuronal activity would allow the achievement of a high-level of pattern recognition comparable to human recognition. An artificial neuron is a model of a biological neuron and serves as a building block in this recognition task. As shown in Fig. 2, the inputs x_i (i = 1,..., d) ('dendrites') are multiplied by synaptic weights w_i, summed and compared to a threshold θ, yielding the intermediate variable u:

$$u = \sum_{i=1}^{d} w_i x_i - \theta \tag{11}$$

The output y of the neuron is a nonlinear function $\phi(u)$ of this intermediate variable. Typically, the activation function ϕ is chosen to be a sigmoid function as in Fig. 1. This model captures the essence of a biological neuron by ensuring that the neuron is excited (output = 1) if the weighted sum of the 'dendritic' inputs surpasses the threshold, while the neuron remains quiet (output = 0) otherwise.

A single neuron can be 'trained' to distinguish between two classes (1 and 0) by appropriately choosing the synaptic weights w_i, using data sets of inputs and outputs. This system is then called a perceptron.

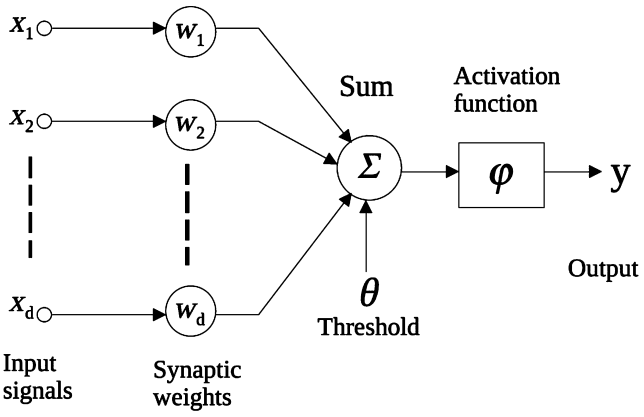

Figure 2. Structure of an artificial neuron.

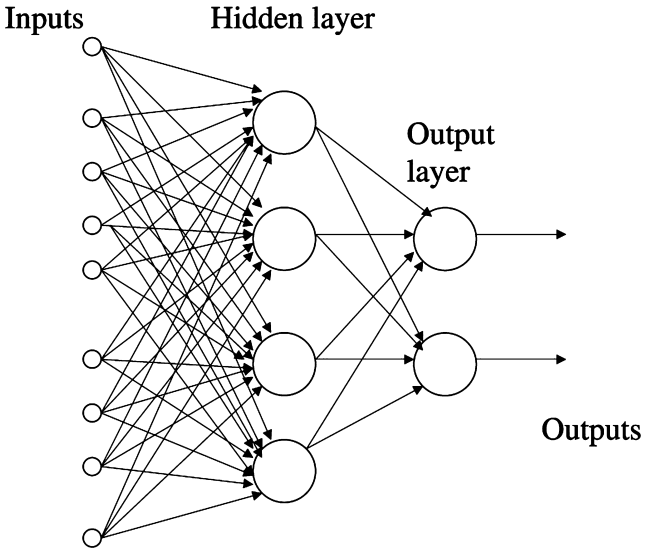

Figure 3. Multilayer perceptron with one hidden layer: each circle is one artificial neuron.

Multilayer perceptrons

A single perceptron as described above is quite similar to any linear combination method. However, when several layers of artificial neurons are arranged in series, parallel and cascade configurations, one obtains a 'multilayer perceptron', shown in Fig. 3 in its simplest configuration (only two layers), which has the capability of modeling highly nonlinear classification borders between two or more classes. This is also called an ANN.

Each neuron in the MLP in Fig. 3 takes as inputs the outputs from the neurons of the previous layer and sends its own output to all the neurons of the following layer. The last layer of neurons is called the output layer; all preceding layers are called hidden layers as their outputs are invisible. The nonlinearity of the activation function ϕ in one or more of the neurons in the hidden layers makes the overall map from input to output nonlinear, thereby giving the MLP the capability to model complex nonlinear functions.

For our sleep–wake classification problem, we need one output neuron from the MLP to get a probability of sleep, using appropriate actigraphy-based variables as inputs to the MLP. Thus, the task of building an MLP classifier for sleep–wake scoring using actigraphy amounts to choosing the most appropriate actigraphy features as inputs to the MLP, followed by the choice of the number of hidden layers and number of neurons in each layer of the MLP and, finally, tuning of the weights for each of the neurons by 'training' the MLP using data.

An important property of an MLP is that one hidden layer is sufficient to approximate, with arbitrary precision, any continuous function mapping inputs to outputs, through appropriate 'learning' or training using data (Lloyd, 2003). However, this property does not indicate how many neurons are required in the hidden layer, which can be large in practice. Sometimes, neural networks with more than one hidden layer are more effective, but their learning process is more difficult. We employed MLPs with one hidden layer and one neuron in the output layer, as we only have two classes to discriminate. Trial and error was used to determine the optimal number of neurons.

For our study, we evaluated 25 features derived from actigraphy as possible inputs to our MLP model. Of these, the five-most discriminant features were determined using Fisher's discriminant analysis and used as inputs for our MLP model with one hidden layer. The weights in our MLP model were determined by minimizing the SSE between the MLP output and reference PSG sleep–wake scoring, using the NETLAB Toolbox (Nabney, 2002), freely available on the web for use under MATLAB® as well as using the Neural Network Toolbox of MATLAB® (http://www.mathworks.com/products/neuralnet). The results are presented in the Results section.

Decision trees

A decision tree is a pattern recognition or classification algorithm, similar in spirit to an MLP, which is capable of mapping the input space into predefined classes using data-based learning. However, unlike MLPs, decision trees are expressed in terms of rules (IF ... THEN ... ELSE ...) that are easier to understand and interpret than ANNs. Decision trees have been studied extensively in various disciplines that include statistics, machine learning, data mining and pattern recognition (Breslow and Aha, 1997).

A typical decision tree is shown in Fig. 4 and consists of one or more terminal nodes called *leaf nodes* (shown with rectangles) and intermediate nodes called *decision nodes* (shown with ovals). The leaf nodes indicate the final classification reached by the classifier while the decision nodes specify a test based on which the subsequent branches are followed. The example in Fig. 4 illustrates the following situation: a person needs to choose between three modes of transport to reach a final destination, depending on the distance to the destination, weather conditions and easy accessibility by public

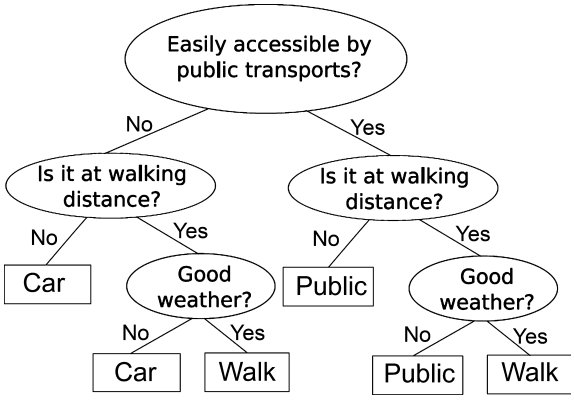

Figure 4. Example of decision tree: which transportation means should I take?

transport. The tree shown is developed from a training data set gathering the features (distance, access by public transport, weather) and the decision taken (walk, take the car or use public transport). The decision process starts at the top root node with the three features, and as we descend the tree, the decision nodes recursively partition the instance space into successively purer sets, until we reach the final classification indicating which transportation means is the best choice.

The training phase of a decision tree from data is called the induction phase and involves the development of a sufficient number of decision nodes and the choice of the 'splitting criterion' at each node so as to optimally classify new data points with minimal error. Extensive discussion of the various criteria used to determine the splitting functions, 'pruning' of the resulting tree to manage complexity and techniques to achieve reasonable trade-off between accuracy, classifier complexity and noise sensitivity goes beyond the scope of our paper. Several of these details can be found in recent surveys (Breslow and Aha, 1997; Rokach and Maimon, 2005) and also in several standard texts on machine learning (Mitchell, 2008).

For our sleep–wake classification problem, the terminal nodes of our decision tree classifier are one of two classes, viz. sleep or wake. The task of building a decision tree classifier using actigraphy involves choosing the most appropriate actigraphy features as inputs to the decision tree, followed by the choice of the number and levels of decision nodes, choice of the splitting function at each node and, finally, 'pruning' of the resulting decision tree to manage the classifier complexity.

For the decision tree induction, we evaluated 25 features derived from actigraphy as possible inputs of which nine features were ultimately used as inputs for induction of the decision tree. The actual construction of the decision tree was made using the Statistics Toolbox in MATLAB®. The results are presented in the following section.

RESULTS

Here, we summarize and compare the results obtained from the various algorithms described in the previous section, when applied to the database of 354 infants from the CHIME study as described in section 'The CHIME database'. We recall that we have 336 958 epochs of 30 s of which 236 869 ($\approx$ 70%) are reference scored as sleep and 100 089 ($\approx$ 30%) are reference scored as wake. We begin with the classification results obtained using linear combination methods.

Linear combination methods

Sadeh's algorithm

Using Sadeh's original unaltered algorithm (Sadeh *et al*., 1994) applied to our data set with 1-min epochs, we obtained an overall global accuracy of 77.6%, sensitivity of 89%, specificity of 52%, PPV of 81% and NPV of 68%. Note that, in this case, the 30-s epoch reference scoring in our data set was translated to 1-min epoch reference as discussed in 'Sadeh's algorithm' in the Methods section. Also, the actigraphy measurements had to be renormalized because the maximum activity in a given epoch in the original work of Sadeh *et al*. (1994) is not the same as in our case.

By computing the same features in Sadeh's original equation but this time for 30-s epochs and renormalizing the equation to account for the different maximum activity in Sadeh's data and in the CHIME data, we obtained a global accuracy of 75.3%, sensitivity of 81.3%, specificity of 61.2%, PPV of 83.2% and NPV of 57.9%.

These results by themselves are quite impressive given that the original equations of Sadeh *et al*. (1994) were developed for a different actimeter. The results suggest that the variables μ, σ, LogAct and nat are indeed the most appropriate discriminant variables to be used in any classification algorithm, regardless of the specific actimeter used, and this is also confirmed later in our own models developed independently. However, to provide a more realistic evaluation of the classification capabilities of Sadeh's algorithm, we decided to recompute the parameters in Sadeh's original discriminant function in equation (7) using our chosen CHIME data. We retained the same five features in the discriminant function but slightly modified one of them – 'nat' to represent the number of epochs in a 10.5-min window centered on the current epoch, which have an appropriately rescaled activity ≥48. This change was motivated by an analysis based on Fisher's generalized criterion for determining the most discriminant form of this feature.

We decided to use 30-s epochs, in keeping with the gold standard in sleep studies. Using 80% randomly selected epochs from the set of 336 958 epochs available, we minimized the SSE (see equation 6) and thereby obtained the following modified discriminant function of Sadeh:

$$\mathrm{SI} = 1.574 - 0.0056\mu - 0.006\sigma - 0.088\mathrm{LogAct} - 0.0854\mathrm{nat} \tag{12}$$

Using this equation, we classified the remaining 20% of the epochs not used in the parameter fitting for validation and achieved the following performance: 78.9% accuracy, 93.8% sensitivity, 43.7% specificity, 79.8% PPV and 73.8% NPV.

Alternatively, by maximizing SESP (see 'Parameter optimization', section), the modified discriminant function of Sadeh became:

$$\mathrm{SI} = 44.4398 - 0.302\mu - 0.2835\sigma - 2.8459\mathrm{LogAct} - 5.2709\mathrm{nat} \tag{13}$$

Validating this equation on the remaining 20% of the epochs not used in the parameter fitting gave the following performance: 76.3% accuracy, 83.4% sensitivity, 59.2% specificity, 82.9% PPV and 60.2% NPV. The results are summarized in Table 2.

Sazonov's algorithm

We evaluated Sazonov's algorithm (see 'Sazonov's algorithm' in the Methods section) using the same scenarios as in the

Table 2 Performance (%) of Sadeh's algorithm optimized for various objectives, using 30-s epochs

Optimization	*Acc*	*Sen*	*Spe*	*PPV*	*NPV*
None	75.3	81.3	61.2	83.2	57.9
Min SSE	78.9	93.8	43.7	79.8	73.8
Max Acc	78.9	94.3	42.5	79.5	75.8
Max SESP	76.2	83.4	59.2	82.9	60.2

Acc, accuracy; Sen, sensitivity; Spe, specificity; PPV, positive predictive value; NPV, negative predictive value; SSE, sum of the squares of the errors; SESP, sum of sensitivity and specificity.

previous section, i.e. using the original unaltered equation of Sazonov (equation 9) and using equation (9) with parameters that were optimized for various objectives. Additionally, we extended the features included in Sazonov's equation by incorporating not only the current and previous eight epochs as in the original equation (9), but also the eight following epochs. Indeed, the original algorithm of Sazonov used the basic principle that the patient is more likely to be asleep or awake if the preceding epochs show the same sleep–wake state. We extended this principle to incorporate subsequent epochs as well and re-estimated the parameters. As in the previous section, we used 80% randomly selected epochs to optimize the chosen objective and then used the remaining 20% of the epochs for validation. The results for these various scenarios are summarized in Table 3.

As expected, tuning the algorithm improves the results, the extended algorithm is slightly better than the original and using the SESP cost functions gives better trade-offs between sensitivity and specificity than SSE.

Pattern recognition methods

Multilayer perceptrons

In order to develop an MLP classifier, we first created a list of 25 features derived from actigraphy data. These features included the ones used by Sadeh *et al*. (1994) and Sazonov *et al*. (2004) as well as a dozen new features that we added. The discriminant power of these 25 features was evaluated using Fisher's generalized discriminant analysis criterion (Hair *et al*., 2006). Table 4 lists these features along with their discriminant power.

Table 3 Performance (%) of Sazonov's algorithm

Algorithm	*Tuning*	*Acc*	*Sen*	*Spe*	*PPV*	*NPV*
Original	None	77	87.1	53.1	81.5	63.5
Original	SSE	78.3	94.9	39.1	78.6	76.3
Original	SESP	75.7	83.4	57.5	82.3	59.5
Extended	SSE	78.7	94.7	40.7	79.1	76.5
Extended	SESP	75.2	81.5	60.3	82.9	57.9

Acc, accuracy; Sen, sensitivity; PPV, positive predictive value; NPV, negative predictive value; SSE, sum of the squares of the errors; SESP, sum of sensitivity and specificity.

The learning or training phase of the MLP is performed through supervised training. The proposed actigraphy features are computed for each epoch from the training data set and serve as the inputs to the MLP. The reference PSG scores (sleep = 1, wake = 0) serve as the corresponding target or 'true' value of the output of the MLP. The training phase then involves computing the weights and thresholds in the MLP so as to minimize a chosen cost function. In our case, we chose to minimize the SSE as shown in equation (6).

There are several numerically efficient methods available to solve the problem of minimizing the SSE objective function in order to obtain the optimal parameters in the MLP. We used two software packages, viz. NETLAB (Nabney, 2002) and the Neural Network Toolbox of MATLAB® (http://www.mathworks.com/products/neuralnet). Both these packages work under the MATLAB® environment for technical computing and provide extensive capabilities for building, optimizing and validating neural networks using a variety of numerical solving options.

A few points are worth mentioning in the context of training of MLPs:

1 The patterns in the training database and the proportion of classes (sleep versus wake) should be representative of what the trained MLP will be faced with for its eventual classification task.
2 The database for training the MLP must be large enough – several thousands of patterns of each class are often required to reach good classification performance. This is perhaps the reason why few attempts have been reported on the use of MLPs for sleep–wake as large actigraphy databases are difficult to obtain.
3 The training phase should be judiciously terminated when no further improvement in the cost function is observed. This avoids 'over-learning' that has the tendency to increase noise sensitivity. Typically, after each training iteration, the MLP is validated using a portion of the database not used in the training, until the cost function on the validation data no longer shows any substantial improvement.
4 There is currently no algorithm that can determine the optimal number of neurons and number of inner layers in the MLP for a given desired accuracy of the classifier. Typically, we fix the size of the neural network (number of layers and neurons in each layer) and then optimally find the weights and thresholds.
5 There is no guarantee that the learning process will lead to the global minimum of the cost function as the optimization problem is nonlinear due to the nonlinear nature of the activation functions.
6 The learning phase of the MLP may be computationally demanding and can take a long time, but once trained, the actual use of the resulting MLP for classification is relatively fast as it involves simple additions, multiplications and functional evaluations to get the output value.

Table 5 presents the performance achieved with MLP, trained with 80% of the available epochs with realistic proportions of sleep and wake epochs and validated on the

Table 4 Proposed actigraphy features with their discriminant power *D*

N	*Feature*	*D*
1	Activity of current epoch	0.1381
2	Sum of activities in a 10.5-min centered window	0.2212
3	Activity of current minus previous epoch	0.00001
4	Activity of current minus next epoch	0.0006
5	Mean activity of the file	0.005
6	Activity of current epoch divided by the number of periods of successive one-value signal in this epoch	0.0698
7	Same as feature 6 in a 5.5-min centered window	0.0998
8	Standard deviation of activity in a 10.5-min centered window	0.2289
9	Number of epochs in centered window with an activity ≥9 and ≤16	0.0688
10–14	Activity of epoch located respectively 5, 4, 3, 2, 1 epochs before the current one	0.0988, 0.1047, 0.113, 0.126, 0.136
15–19	Activity of epoch located respectively 1, 2, 3, 4, 5 epochs after the current one	0.1215, 0.101, 0.083, 0.0714, 0.064
20,21	Max, min epoch activity in a 10.5-min centered window	0.2333, 0.0158
22	Number of epochs in a 10.5-min centered window with activity value greater than five times the mean activity	0.2167
23	Longer one-period in epoch	0.1049
24	Number of one values in actigraphic signal in a 5.5-min centered window that are not between 2 zeros	0.0987
25	Natural logarithm of the activity of current epoch, incremented by 1	0.1762

remaining 20% of the epochs. The table shows performance with different numbers of neurons in the single hidden layer, with the proposed set of 25 input features from Table 4. The computations were carried out using both the software packages (NETLAB and the Neural Network Toolbox) in MATLAB® to crosscheck results. Overall, NETLAB provided better models that reached lower values of the SSE cost function.

The table also includes performance measures using only five of the most discriminant features (features 2, 8, 20, 22, 25 from Table 4). Three of these five features were also used by Sadeh *et al*. (1994), confirming the choices made by Sadeh. We then utilized Fisher's generalized criterion to determine the optimal length of the window for features 2, 8, 20 and 22. The resulting five-most discriminant attributes are the following:

Feature 2 optimized: sum of all the activities of a 37-epoch centered window, $D = 0.2426$.

Feature 8 optimized: activity standard deviation on a 25-epoch centered window, $D = 0.2531$.

Feature 20 optimized: maximum epoch activity on a 19-epoch centered window, $D = 0.2427$.

Feature 22 optimized: number of epochs in a 47-epoch centered window that have an activity superior to 2.025 times the mean activity of the file, $D = 0.3804$.

Feature 25: logarithm of the current epoch activity increased by one, $D = 0.1762$.

We see that the results with the five best features are better than those with 25 input features. Table 5 also shows that the performance with only five neurons in the hidden layer is close to that achieved with 10 neurons and our experiments showed no advantage in increasing this number. As expected, there is a big difference between the sensitivity and specificity.

Table 6 shows the results obtained when varying the proportion of wake in the learning database (PWL), for an MLP with five neurons in the hidden layer and using the five input features listed above. As expected, the specificity dramatically increases with the proportion of wake in the learning database, while sensitivity falls and global accuracy slightly decreases. Varying PWL enables us to reach very good trade-offs between sensitivity and specificity, without really worsening the accuracy. We note that including a greater proportion of wake epochs during training contradicts the common guideline that the MLP should be trained with data that are representative of the anticipated data it will have to

Table 5 Performance (in %) of multilayer perceptron trained with realistic proportion of sleep and wake epochs

Input features	*Hidden neurons*	*Acc*	*Sen*	*Spe*	*PPV*	*NPV*
25	5	79.1	94.4	43	79.7	76.5
25	10	77.5	85.6	58.4	83	63.1
5	5	80.3	92.6	51.1	81.9	74.4
5	10	80.5	92.7	51.6	82	74.7

Acc, accuracy; Sen, sensitivity; Spe, specificity; PPV, positive predictive value; NPV, negative predictive value.

Table 6 Performance of multilayer perceptron for increasing proportions of wake epochs in learning data from top to bottom

PWL	*Acc*	*Sen*	*Spe*	*PPV*	*NPV*
30	80.3	92.6	51.1	81.9	74.4
39	79.8	88.6	58.8	83.7	68.5
44	79.1	85.7	63.3	84.8	65
49	78.1	82.4	67.8	85.9	61.8
51	77.5	81.1	68.9	86.1	60.5
56	75	75.8	73.2	87	56

All values are in %.

PWL, proportion of wake in learning database; Acc, accuracy; Sen, sensitivity; Spe, specificity; PPV, positive predictive value; NPV, negative predictive value.

classify. Thus, accuracy is expected to drop, but wake classification improves.

Decision trees

The same 25 features listed in Table 4 were used as inputs to the training or induction phase of the decision tree classifier. The classical approach for inducing a decision tree is to choose splitting functions at the intermediate nodes so as to minimize either the total entropy or total impurity of the terminal nodes. In the limit, if all terminal nodes are completely pure, the tree entropy or tree impurity is zero, an ideal case. More realistically, induction develops sufficiently many nodes and depth in the decision tree so that the terminal node classes are as pure as possible. For our induction task, we chose to minimize Gini's impurity (Breslow and Aha, 1997; Mitchell, 2008).

We used the Statistics Toolbox in MATLAB® for building our decision tree classifier from data by using Gini's impurity as the splitting function. The Statistics Toolbox also contains various functions to build the tree, prune it and validate the resulting classifier. A few points are worth listing in the context of induction of decision trees:

1 There does not exist a computationally efficient method for determining the optimal number of nodes and decision tree depth needed for a prespecified accuracy (Rokach and Maimon, 2005), but heuristics can be applied through trials. Reducing the number of nodes from a very detailed decision tree by removing less relevant nodes is called 'pruning' and is accomplished using such heuristics.
2 Decision trees have a slight advantage over MLPs in that they require less data for training or induction. On the other hand, decision trees are more sensitive to noise and the patterns presented in the training data. Thus, an error in the decision function at a higher node will propagate to the lower nodes uncorrected leading to poor classification performance.
3 'Bagging' (also called 'bagstrap aggregating') algorithms for decision trees are employed to reduce the effect of noise sensitivity (Mitchell, 2008). In this approach, several training sets are created randomly from the original training set of epochs, allowing an epoch to belong to several training sets. A decision tree is built with each training set and the predicted class of an unknown object is the majority of the classes predicted from the different trees. This is known to reduce the variance of the classification results, improve stability and reduce sensitivity to noise and specific training patterns.
4 Decision trees can directly handle categorical or qualitative data ('sleep' or 'wake', 'sunny' or 'rainy') unlike MLPs that require numerical data.

Decision trees were trained with 80% of the available epochs with realistic proportions of sleep and wake epochs and validated on the remaining 20% of the epochs. The inputs to the decision tree classifier were the 25 actigraphy features from Table 4. Table 7 presents the results obtained with the full tree and with a tree pruned to a more suitable number of nodes.

Table 7 Performance (in %) of decision trees with 25 input features

Levels	*Nodes*	*Acc*	*Sen*	*Spe*	*PPV*	*NPV*
288	27923	79.2	88.9	56.3	82.8	68.2
29	185	80.3	94.6	46.7	80.7	78.4

Acc, accuracy; Sen, sensitivity; Spe, specificity; PPV, positive predictive value; NPV, negative predictive value.

Table 8 Performance (in %) of decision trees with the five and nine input features

Levels	*Nodes*	*Acc*	*Sen*	*Spe*	*PPV*	*NPV*
Five-most discriminant features (same as in MLP)						
244	21397	81.7	92.1	57.1	83.5	75.4
24	131	80.1	94	47.4	80.9	76.9
Proposed nine features						
232	12603	85.1	91.8	69	87.6	78
48	355	82.1	92.2	58.1	83.9	75.8

MLP, multilayer perceptron; Acc, accuracy; Sen, sensitivity; Spe, specificity; PPV, positive predictive value; NPV, negative predictive value.

The characteristic 'level' of each tree is given by MATLAB, and represents the number of branches that can still be pruned.

It was noted from the obtained decision trees that only features 2, 5, 7, 10, 14 and 22 appeared in the upper nodes of the decision tree. To these features, we added the five-most discriminant features similar to the case with MLPs. The resulting nine features (2, 5, 7, 8, 10, 14, 20, 22, 25) were optimized for their window sizes using Fisher's discriminant analysis and gave the optimized windows reported in the previous section on Multilayer perceptrons.

Table 8 presents the performance of decision trees trained with realistic proportions of sleep and wake epochs and using the five-most discriminant features (used with MLP) and the nine features discussed above.

We see that the best results are achieved with the nine features selected for decision trees. However, sleep is again much better detected than wake. To overcome this problem, we increased the proportion of wake epochs in the learning phase (PWL). Table 9 summarizes the results obtained when the decision tree is trained with increasing proportion of wake epochs in the training data. The improvement in specificity is apparent at the expense of sensitivity, thereby indicating the inherent trade-off in the classifier performance.

'Bagging' to reduce sensitivity to noise and specific training features did not improve the results appreciably. We believe this was because the epochs used for training were already sufficiently random.

Performance summary

Table 10 gathers the highest accuracy and SESP achieved by each algorithm. It is obvious that pattern recognition methods, and in particular decision trees, improve the classification

Table 9 Performance of decision trees for increasing proportions of wake epochs in learning data from top to bottom

PWL	*Acc*	*Sen*	*Spe*	*PPV*	*NPV*
30	82.1	92.2	58.1	83.9	75.8
39	81.4	87.4	67.2	86.4	69.1
44	80.8	85.2	70.3	87.2	66.5
49	78.9	80.7	74.6	88.3	61.9
56	76.9	77	76.7	88.7	58.3

Nine input features are used. All values are in %.
Acc, accuracy; Sen, sensitivity; Spe, specificity; PPV, positive predictive value; NPV, negative predictive value.

Table 10 Summary table: best performance (in %) achieved by each algorithm when considering accuracy or the sum of sensitivity and specificity

Algorithm	*Highest accuracy*	*Highest Sen + Spe*
Sadeh	78.9	142.6
Sazonov	78.7	141.8
MLP (five hidden neurons)	80.3	150.2
'Small' decision tree	82.1	155.5

MLP, multilayer perceptron; Sen, sensitivity; Spe, specificity.

performance significantly. The advantage of using these methods is already substantial when considering accuracy alone, but the most impressive improvement lies in the better detection of wake we can achieve without deteriorating the detection of sleep, as indicated by the higher SESP, surpassing the linear methods by more than 10%.

Rescoring rules

Noting the trend of actigraphy to overestimate sleep, Webster *et al*. (1982) proposed the use of postprocessing rules to rescore previously classified epochs. These 'rescoring rules' have been shown in several studies to improve specificity by several percentage points by rescoring sleep epochs as wake epochs. The rules are listed below:

After at least 4 min (respectively 10 and 15 min) scored wake, the first minute (respectively 3 and 4 min) scored sleep is rescored wake.
Six minutes or less (10 min or less) scored sleep are rescored wake if they are surrounded by at least 10 min (respectively 20 min) before and after scored wake.

Table 11 summarizes the results obtained by applying these rescoring rules to the decision tree classifier and the MLP classifier. The decision tree and MLP were trained with two-third of the infant files randomly chosen by cross-validating with the remaining one-sixth of the infant files randomly chosen and then rescoring was evaluated on the last one-sixth of the infant files. We see that specificity and accuracy improves with rescoring for decision trees but the improvement is smaller for MLP. The performance of the decision tree

Table 11 Performance (%) of a decision tree taking nine input features with 39 levels and 217 nodes, and multilayer perceptron with five input features with five hidden neurons, trained with real proportions of sleep and wake, before and after rescoring

Evaluation	*Acc*	*Sen*	*Spe*	*PPV*	*NPV*
Decision tree with nine input features					
Before rescoring	80.7	91.3	54.3	83.3	80.7
After rescoring	81	90.1	58.3	84.3	70.3
MLP with five input features					
Before rescoring	80.5	91.9	52.2	82.7	72
After rescoring	80.6	90.6	55.9	83.6	70.4

MLP, multilayer perceptron; Acc, accuracy; Sen, sensitivity; Spe, specificity; PPV, positive predictive value; NPV, negative predictive value.

before rescoring is slightly worse than the figures presented previously because, in this case, we excluded entire files from the training rather than randomly picking epochs across all files. The MLP did not suffer from this training process, which reinforces the idea that decision trees are more sensitive to the training set than MLPs. The results presented in this section are close to the previously reported results, which indicate a good generalizability of the performance of the developed algorithms.

Influence of age and health group

It has been shown previously by Sadeh *et al*. (1995) that developing different algorithms to score sleep–wake in infants of different ages did not improve the performance compared to developing only one algorithm for all the infants. We adopted this same approach and developed only one classification model applicable to the entire infant group. Similarly, we used randomly selected epochs for training our classification models, without any consideration of the age or health status of the chosen infant file.

Tables 12 and 13 summarize the performance of the decision tree with 48 levels, for the various age and health groups. The column 'NI' stands for number of infants available for the given age or health group. The results presented in this table do not strictly satisfy the cross-validation principle since the epochs used for training the decision tree are included in the

Table 12 Performance (%) of best decision tree, with 48 levels, for the different age groups

Age (months)	*NI*	*Acc*	*Sen*	*Spe*	*PPV*	*NPV*
< 2	161	82	91.9	60	83.6	77
2–3	92	81.7	91.8	57.7	83.8	74.7
> 3	82	84.2	93.7	57.3	86.1	76.4

Acc, accuracy; Sen, sensitivity; Spe, specificity; PPV, positive predictive value; NPV, negative predictive value; NI, number of infants available for the given age or health group.

Table 13 Performance (%) of best decision tree, with 48 levels, for the different health groups

Group	*NI*	*Acc*	*Sen*	*Spe*	*PPV*	*NPV*
Healthy	80	86.9	94.1	65.5	89	78.7
SIDS	69	82.1	93.3	56.9	83	79.1
Apnea	61	83.6	93.3	56	85.7	74.7
Premature	125	79.3	90	57.7	81	74.3

Acc, accuracy; Sen, sensitivity; Spe, specificity; PPV, positive predictive value; NPV, negative predictive value; SIDS, sudden infant death syndrome; NI, number of infants available for the given age or health group.

results. But they give a good indication of how the decision tree performs qualitatively for these groups.

The overall trend appears to be a slightly better classification performance for infants older than 3 months and a bigger influence of health status: the best performance is achieved for healthy infants with considerably worse classification for premature infants.

DISCUSSION

The linear classification algorithms proposed by Sadeh *et al.* (1994) and Sazonov *et al.* (2004) performed well with our chosen data even without optimization of the parameters, thus confirming the choices of the key discriminant variables for the classification task, regardless of the actimeter used. Tuning these parameters enabled us to improve the results, and to reach better trade-offs between sensitivity and specificity depending on the cost function chosen for optimization. The results show a slight advantage for Sadeh's algorithm, which uses more discriminant features than Sazonov's chosen features. However, the results presented in the previous section show that, as expected, pattern recognition methods have the potential to improve the results substantially due to their ability to generate nonlinear classification borders, using very similar discriminant variables for the classification task.

All methods have a tendency to overpredict sleep. We investigated three ways of reaching better trade-offs between sensitivity and specificity. The first involved changing the cost function used for training the algorithm so that more importance is given to the detection of wake. This method was successfully used with linear algorithms. The second approach involved increasing the proportion of wake epochs in the learning database. This method yielded excellent results with pattern recognition techniques. And thirdly, we evaluated rescoring rules and confirmed that specificity improves by a few percentage points.

Each algorithm evaluated in this study needed a relatively time-consuming and computationally demanding learning phase (up to several hours). However, once trained, the resulting classifier can then rapidly classify new objects: the developed neural networks can compute the classes of 418 000 epochs (3500 h of recording) in 1 s, while the developed decision trees predict the classes of 120 000 epochs (1000 h of recording) in 1 s.

The performance we achieved was better than that reported by Sazonov *et al.* (2004), who worked with the same database as we did. It is worth noting that the study of Sazonov *et al.* is the only one performed on a population of healthy and non-healthy infants. Several other studies have reported results for healthy infants only, e.g. see So *et al.* (2005) and Sadeh *et al.* (1995).

By comparing sensitivity and specificity, the performance of our classifier is in the same range as that obtained by So *et al.* (2005). However, their database contained very few wake epochs ($< 10\%$). This is why with similar sensitivity and specificity, the global accuracy we reached is lower than theirs. In fact, So *et al.* obtained a global accuracy lower than the percentage of sleep epochs, which means that they would have reached a better accuracy by merely scoring all the epochs as sleep. This illustrates the fallacy in using global accuracy alone to compare classifier performance.

However, our results are slightly worse than those of Sadeh *et al.* (1995) (88.9% of accuracy and 82.8% of specificity for infants younger than 3 months, which is the major age group of the CHIME database). Again, it is important to notice that Sadeh *et al.* worked with healthy children and that their database contained 79.1% of sleep epochs. Also, Sadeh *et al.* removed the epochs scored as uncertain and the transition epochs between sleep and wake, which represented 5.2% of their database and are the epochs where prediction errors are the most likely to occur. Additionally, the actimeter used in their study appeared to give much more information than the one used for the CHIME study. Several other studies involving adults (Cole *et al.*, 1992; Jean-Louis *et al.*, 2001) also yielded a greater accuracy but poor trade-offs between sensitivity and specificity.

We reiterate that we designed and evaluated our algorithms with a very large database. This was necessary to train pattern recognition algorithms (especially MLPs) and is important to ensure that our results are stable. To the best of our knowledge, such a large database has not been previously employed in sleep–wake classifier design.

Concerning the features, our experiments confirmed that Sadeh's features perform well. Indeed, three of the five-most discriminant features we obtained were inspired by the features suggested by Sadeh. Nevertheless, the most discriminant feature involved the mean (average) activity of the entire overnight recording of a subject. No previous study mentions the use of this feature based on a global trend of the recording. This feature is very important to correctly detect sleep when there are frequent small movements, and correctly predict wake when there are few movements. This is illustrated in Figs 5 and 6.

Figure 5 presents a recording (number 41090) with very little activity and it appears that each time there is a non-zero activity, the epoch must be scored as wake. The decision trees perform rather well for this file. Figure 6 shows a recording (number 20145) with frequent activity, and in fact, some sleep

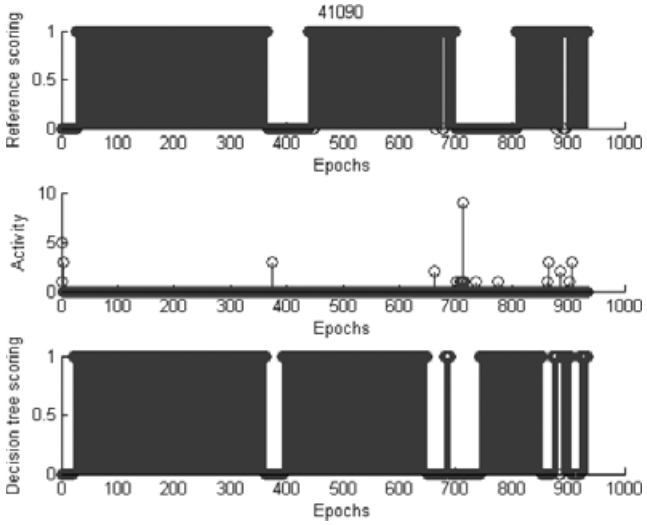

Figure 5. Reference scoring, activity signal and decision tree scoring for a file with few movements.

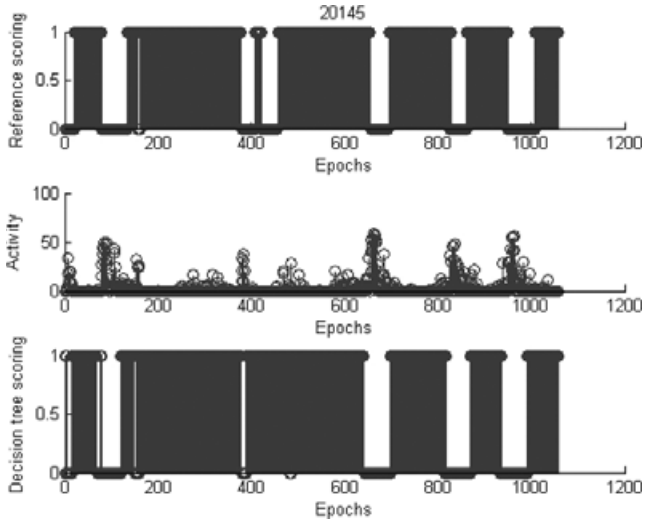

Figure 6. Reference scoring, activity signal and decision tree scoring for file with many movements.

epochs have higher activity than wake epochs in Fig. 5. It can be seen, however, that the decision tree was able to adapt to this very different kind of file and generate a very good classification.

CONCLUSIONS

We have proposed two new pattern recognition classifiers and applied them to a large training and validation database of infants to demonstrate their improved sleep–wake scoring performance. The algorithms are robust and suggest that neural network and decision tree classifiers can find broader applicability in the context of clinical sleep research, much in the spirit of traditional statistical methods, which are now been widely employed in a clinical setting.

A variety of advanced actimeters has been developed since the CHIME study, and provides much more detailed measurements. We believe that validating and testing our algorithms with state-of-the art actimeters would further confirm the efficacy of our models. New features suited for more precise actigraphy signals (not just binary signal but, e.g. the amplitude of the movement) could be developed. The actimeter used in the CHIME study was placed on the ankle. Current practice is to use the non-dominant hand for actigraphy studies. Developing and validating our proposed classifiers for this scenario would be needed to generalize our results to wrist actigraphy. Validating our proposed classifiers on adult populations in which sleep–wake cycles are well developed would be a natural testing scenario. And finally, fusion of actigraphy with another signal (e.g. EEG) could be investigated to improve the predictive capabilities of actigraphy.

ACKNOWLEDGEMENTS

Partial financial support from the P. C. Rossin Professorship and the Frank Hook Professorship at Lehigh University for the visiting student support of J. Tilmanne and J. Urbain is gratefully acknowledged. We also acknowledge financial support from the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The funding agencies do not assume responsibility for the scientific content of published articles. And finally, we acknowledge the CHIME team for providing us with their database and also in helping us interpret the data.

CONFLICT OF INTEREST

None of the authors have any conflicts of interest to declare.

REFERENCES

Ancoli-Israel, S. Actigraphy. In: M. H. Kryger, T. Roth and W. C. Dement (Eds) *Principles and Practice of Sleep Medicine*. W. B. Saunders, Philadelphia, 2000: 1295–1301.

Ancoli-Israel, S., Cole, R., Alessi, C., Chambers, M., Moorcroft, W. and Pollak, C. P. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*, 2003, 26: 342–392.

Breslow, L. A. and Aha, D. W. Simplifying decision trees: a survey. *Knowl. Eng. Rev.*, 1997, 12: 1–40.

CHIME-website. 2004. Available at: http://dccwww.bumc.bu.edu/ChimeNisp/Main_Chime.asp (last accessed 18 October 2008).

Clancy, R. R., Bergquist, R. R. and Dlugos, D. J. Neonatal encephalography. In: J. S. Ebersole and T. A. Pedley (Eds) *Current Practice of Clinical Electroencephalography*, 3rd edn. LippinCott Williams & Wilkins, Philadelphia, 2003: 160–234.

Cole, R. J., Kripke, D. F., Gruen, W., Mullaney, D. J. and Gillin, J. C. Technical note: automatic sleep/wake identification from wrist actigraphy. *Sleep*, 1992, 15: 461–469.

Crowell, D., Brooks, L. J., Colton, T., Corwin, M. J., Hoppenbrouwers, T. T., Hunt, C. E., Kapuniai, L. E., Lister, G., Neuman, M. R., Peucker, M., Ward, S. L. D., Weese-Mayer, D. E., Willinger, M.

and the CHIME Steering Committee. Infant polysomnography: reliability. *Sleep*, 1997, 20: 553–560.

Hair, J. F., Tatham, R. L., Anderson, R. E. and Black, W. *Multivariate Data Analysis*, 6th edn. Prentice Hall, Upper Saddle River, NJ, 2006.

Jean-Louis, G., von Gizycki, H., Zizi, F., Fookson, J., Spielman, A., Nunes, J., Fullilove, R. and Taub, H. Determination of sleep and wakefulness with the actigraph data analysis software (ADAS). *Sleep*, 1996, 19: 739–743.

Jean-Louis, G., Kripke, D. F., Mason, W. J., Elliott, J. A. and Youngstedt, S. D. Sleep estimation from wrist movement quantified by different actigraphic modalities. *J. Neurosci. Meth.*, 2001, 105: 185–191.

Kevin, S., Adamson, T. M. and Horne, R. S. C. The use of actigraphy for assessment of the development of sleep/wake patterns in infants during the first 12 months of life. *J. Sleep Res.*, 2007, 16: 181–187.

Kohavi, R. and Provost, F. Glossary of terms. *Mach. Learn.*, 1998, 30: 271–274.

Littner, M., Kushida, C. A., Anderson, W. M., Bailey, D., Berry, R. B., Davila, D. G., Hirshkowitz, M., Kapen, S., Kramer, M., Loube, D., Wise, M. and Johnson, S. F. Practice parameters for the role of actigraphy in the study of sleep and circadian rhythms: an update for 2002. *Sleep*, 2003, 26: 337–341.

Lloyd, J. W. *Logic for Learning: Learning Comprehensive Theories from Structured Data*. Springer, Berlin, 2003: 207–241.

Mitchell, T. M. *Machine Learning*. McGraw-Hill series in computer science, McGraw Hill, WCB, Boston, 2008.

Morgenthaler, T., Alessi, C., Friedman, L., Owens, J., Kapur, V., Boehlecke, B., Brown, T., Chesson, A., Coleman, J., Lee-Chiong, T., Pancer, J. and Swick, T. J. Practice parameters for the role of actigraphy in the study of sleep and sleep disorders: an update for 2007. *Sleep*, 2007, 30: 519–528.

Nabney, I. T. *NETLAB: Algorithms for Pattern Recognition. Advances in Pattern Recognition*. Springer, New York, 2002.

Rechtschaffen, A. and Kales, A. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. U.S. Department of Health, Education & Welfare, National Institute of Health Publication No. 204, Washington DC, 1968.

Rokach, L. and Maimon, O. Top-down induction of decision trees classifiers – a survey. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.*, 2005, 35: 476–487.

Sadeh, A., Alster, J., Urbach, D. and Lavie, P. Actigraphically based automatic bedtime sleep-wake scoring: validity and clinical applications. *J. Ambul. Monit.*, 1989, 2: 209–216.

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

Sadeh, A., Acebo, C., Seifer, R., Aytur, S. and Carskadon, M. A. Activity-based assessment of sleep-wake patterns during the first year of life. *Infant Behav. Dev.*, 1995, 18: 329–337.

Sazonov, E. S., Sazonova, N. S., Schuckers, S. A. C., Neuman, M. and CHIME study group. Activity-based sleepwake identification in infants. *Physiol. Meas.*, 2004, 25: 1291–1304.

So, K., Buckley, P., Adamson, T. M. and Horne, R. S. C. Actigraphy correctly predicts sleep behavior in infants who are younger than six months, when compared with polysomnography. *Pediatr. Res.*, 2005, 58: 761–765.

de Souza, L., Benedito-Silva, A. A., Pires, M. L. N., Poyares, D., Tufil, S. and Calil, H. M. Further validation of actigraphy for sleep studies. *Sleep*, 2003, 26: 81–85.

Thorpy, M., Chesson, A., Derderian, S., Kader, G., Millman, R., Potolicchio, S., Rosen, G. and Strollo, P. J. Practice parameters for the use of actigraphy in the clinical assessment of sleep disorders. *Sleep*, 1995, 18: 285–287.

Webster, J. B., Kripke, D. F., Messin, S., Mullaney, D. J. and Wyborney, G. An activity-based sleep monitor system for ambulatory use. *Sleep*, 1982, 5: 389–399.