RC24302 (W0707-060) July 6, 2007
Computer Science

IBM Research Report

Subject-Adaptive Real-Time Sleep Stage Classification Based on Conditional Random Field

Gang Luo, Wanli Min
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Subject-Adaptive Real-Time Sleep Stage Classification Based on Conditional Random Field

Gang Luo, PhD, Wanli Min, PhD
IBM T.J. Watson Research Center, Hawthorne, NY
{luog, wanlimin}@us.ibm.com

Abstract

Sleep staging is the pattern recognition task of classifying sleep recordings into sleep stages. This task is one of the most important steps in sleep analysis. It is crucial for the diagnosis and treatment of various sleep disorders, and also relates closely to brain-machine interfaces. We report an automatic, online sleep stager using electroencephalogram (EEG) signal based on a recently-developed statistical pattern recognition method, conditional random field, and novel potential functions that have explicit physical meanings. Using sleep recordings from human subjects and birds, we show that the average classification accuracy of our sleep stager almost approaches the theoretical limit and is about 8% higher than that of existing systems. Moreover, for a new subject s_{new} with limited training data D_{new}, we perform subject adaptation to improve classification accuracy. Our idea is to use the knowledge learned from old subjects to obtain from D_{new} a regulated estimate of CRF's parameters. Using sleep recordings from human subjects, we show that even without any D_{new}, our sleep stager can achieve an average classification accuracy of 70% on s_{new}. This accuracy increases with the size of D_{new} and eventually becomes close to the theoretical limit.

1. Introduction

Sleep is indispensable to everybody. As have been reported in Ancoli-Israel and Roth[1] that is consistent with other national studies, about one-third of Americans had some kind of sleep problem. Hence, the study of sleep pattern, much of which is through sleep recordings, has consistently been an important research topic.

A typical sleep recording has one or more channels of electroencephalogram (EEG) waves coming from electrodes. Sleep staging is the pattern recognition task of classifying sleep recordings into sleep stages (e.g., wake, sleep) continuously. This task is crucial for the diagnosis and treatment of various sleep disorders[19]. In addition, it relates closely to both intensive care unit monitoring of brain activity[20] and brain-machine interfaces[2]. In the latter case, successful classification can facilitate disabled people to control computers. Sleep staging is also of special interest to the study of avian bird song system[3] and the evolutionary theory of mammalian sleep[4].

Many statistical pattern recognition methods, such as autoregression[5], Kullback-Leibler divergence-based nearest-neighbor classification[6], and hidden Markov model (HMM)[7], have been used to build an automatic, online sleep stager. Despite all these efforts, existing sleep stagers can only achieve an average classification accuracy below 80%[8, 19], which is insufficient for physicians to quickly and correctly diagnose sleep disorders by establishing a clear classification of the problem. (In brain-computer interfaces, incorrect EEG wave classification can cause computers to receive wrong instructions.) In this work, we present an automatic, online sleep stager based on a recently-developed statistical pattern recognition method, conditional random field (CRF), and novel potential functions that have explicit physical meanings. According to our testing results on single-channel sleep recordings from human beings and birds, our sleep stager can achieve an average classification accuracy that almost approaches the theoretical limit[9] and is about 8% higher than that of existing systems.

One challenge for sleep staging is that in practice, we often have enough training data D_{old} from several old subjects s_{old} but very limited training data D_{new} from a new subject s_{new}, as it often takes several days or several weeks to manually label sufficient D_{new} for s_{new}[19]. In this case, it is undesirable to train the parameter vector Θ of the CRF by only using D_{new}. Instead, we perform *subject adaptation* to improve the classification accuracy on s_{new}[10]. Our high-level idea is to use the knowledge on Θ that is learned from D_{old} to obtain a regulated estimate of Θ from D_{new}. In this way, the classification accuracy on s_{new} increases with the size of D_{new} and eventually becomes close to the theoretical limit[9]. Especially, even without any D_{new}, the average accuracy on s_{new} can be 70% according to our test results on sleep recordings from human beings.

CRF was originally proposed by the natural language processing community in 2001[11]. In contrary to HMM, CRF directly models the probabilities of possible label sequences given an observation sequence, without making unnecessary independence

assumptions on the observation elements. Consequently, CRF overcomes HMM's shortcoming of being unable to represent multiple interacting features or long-range dependencies among the observation elements. To the best of our knowledge, neither the application of CRF nor subject adaptation has been studied before in EEG wave classification. Also, so far no method has been reported that can achieve satisfactory accuracy on both human beings and birds for EEG wave classification.

The rest of the paper is organized as follows. Section 2 provides a brief review of CRF. Section 3 presents our automatic, online sleep stager based on CRF for a single subject. Section 4 describes the subject adaptation technique. Section 5 discusses feature extraction. We evaluate the performance of our techniques in Section 6 and conclude in Section 7.

2. Review of CRF

We first review the concept of CRF. Let X be the observation sequence, and Y be the corresponding label (state) sequence. The CRF definition in Lafferty *et al.*[11] is as follows:

Definition. *Let* $G = (V, E)$ *be a graph such that* $Y = (y_v)_{v \in V}$, *so that Y is indexed by the vertices of G. Then* (X, Y) *is a conditional random field in case, when conditioned on X, the random variables* y_v *obey the Markov property with respect to the graph:* $P(y_v \mid X, y_w, w \neq v) = P(y_v \mid X, y_w, w \sim v)$, *where w~v means that w and v are neighbors in G.*

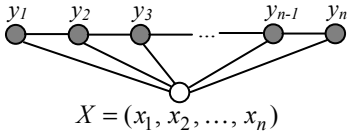

Figure 1. Graphical structure of a linear-chain CRF.

A special case of CRF is the linear-chain CRF (LCRF)[11] as shown in Figure 1, where the graph G is a linear chain so that each y_i has exactly two neighbors: y_{i-1} and y_{i+1}. As has been shown in Lafferty *et al.*[11], in this case, the distribution of the label sequence Y given the observation sequence X has the following form:

$$p(Y \mid X) \propto \exp\{\sum_{i=1}^{n}[\sum_{j=1}^{k_1}\lambda_j f_j(y_{i-1}, y_i, X, i) + \sum_{j=1}^{k_2}\mu_j g_j(y_i, X, i)]\}.$$

Here, f_j and g_j are called potential functions. λ_j and μ_j are parameters. The selection of appropriate potential functions is both application-dependent and critical to the success of the CRF method.

3. CRF-based sleep stager for a single subject

In our sleep stager, we use linear-chain CRFs. In this case, $X = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n)$ is the observation sequence, where each element $\vec{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}]^T$ is an m-dimensional vector that represents the observed EEG wave signal (possibly after some transformation) at time point i ($1 \le i \le n$). $Y = (y_1, y_2, \ldots, y_n)$ is the label sequence. Each y_i ($1 \le i \le n$) belongs to the sleep stage space S (e.g., {wake, REM, NREM}) and represents the sleep stage at time point i that needs to be labeled.

Our sleep stager uses the following two kinds of potential functions, the first one is for f_j and the second one is for g_j:

(1) $1_{y_{i-1}=s}1_{y_i=t}$ ($s, t \in S$),

(2) $1_{y_i=t}x_{i,h}$ ($t \in S, 1 \le h \le m$).

Here, the indicator function $1_{y_i=t} = \begin{cases} 1 & (if\ y_i = t) \\ 0 & (if\ y_i \neq t) \end{cases}$.

For each i ($1 \le i \le n$), the number of potential functions is $k = |S|^2 + |S|m$. Our intuition is that local features are often the most important ones. Hence, at any time point i ($1 \le i \le n$), we focus on the local observation elements and only consider the first-order term $\vec{x}_i$. Also, these potential functions are easy to compute, which is important for online classification. In fact, these potential functions can be justified from the statistical mechanics perspective: (1) The term $\exp\{\lambda_{s,t}1_{y_{i-1}=s}1_{y_i=t}\}$ can be viewed as the spontaneous transition probability from state s to state t. (2) As discussed below, our X is the power spectral density, a quantity associated with energy. Hence, the term $\exp\{\mu_{t,h}1_{y_i=t}x_{i,h}\}$ can be viewed as an analogy to the Boltzmann factor $P(E) \propto \exp(-\beta E)$, which is related to the probability for a canonical ensemble to be in a state with energy E[12].

Given the $k = k_1 + k_2$ potential functions, parameter estimation (i.e., learning λ_j's and μ_j's from a labeled training data set) and inference making (i.e., given X, computing the most likely Y) in the CRF are performed using the forward-backward dynamic programming and Viterbi algorithms, as described in Lafferty *et al.*[11] and Sha and Pereira[18].

4. Subject adaptation

Next, we discuss subject adaptation. This technique combines the (usually sufficient) training data sequence (X_{old}, Y_{old}) from several old subjects s_{old} with the (possibly insufficient) training data sequence

(X_{new}, Y_{new}) from a new subject s_{new} to improve the classification accuracy on s_{new}. Let Θ be the column parameter vector of the CRF that contains λ_j's and μ_j's. $L_{old}(\Theta) = \ln p(Y_{old} \mid X_{old}, \Theta)$ and $L_{new}(\Theta) = \ln p(Y_{new} \mid X_{new}, \Theta)$ are the log-likelihood functions for s_{old} and s_{new}, respectively. Let $\hat{\Theta}$ denote the maximum-likelihood estimator (MLE) of Θ on s_{old}. A theorem about MLE[13] asserts that $\hat{\Theta}$ asymptotically follows a normal distribution, whose mean vector and covariance matrix are Θ and $\Sigma = -(\nabla^2 L_{old})^{-1}$, respectively. Here, $\nabla^2 L_{old}$ is the Hessian matrix of $L_{old}(\Theta)$. This can be viewed as a prior of Θ when we fit the same model to s_{new}. The corresponding probability density function is

$$p(\Theta) \propto \exp\{-(\Theta - \hat{\Theta})^T \cdot \Sigma^{-1} \cdot (\Theta - \hat{\Theta}) / 2\}$$
$$= \exp\{(\Theta - \hat{\Theta})^T \cdot \nabla^2 L_{old} \cdot (\Theta - \hat{\Theta}) / 2\}.$$

From Bayes' theorem, the posterior distribution of Θ is

$$p(\Theta \mid X_{new}, Y_{new}) \propto p(Y_{new} \mid X_{new}, \Theta) p(\Theta)$$
$$\propto \exp\{L_{new}(\Theta) + (\Theta - \hat{\Theta})^T \cdot \nabla^2 L_{old} \cdot (\Theta - \hat{\Theta}) / 2\}.$$

The gradient of $L_{old}(\Theta)$, ∇L_{old}, can be efficiently computed using a backward-forward dynamic programming method[11]. $\nabla^2 L_{old}$ can be computed numerically by taking difference quotients of ∇L_{old}. Then we can obtain the point estimate Θ for s_{new} by maximizing $L_{new}(\Theta) + (\Theta - \hat{\Theta})^T \cdot \nabla^2 L_{old} \cdot (\Theta - \hat{\Theta}) / 2$ (e.g., using the BFGS method).

5. Data collection and transformation

We applied our sleep stager to four 24-hour human sleep recordings in the sleep-EDF database[14] whose subject ids are sc4002e0, sc4012e0, sc4102e0, and sc4112e0. Each recording was from a different, healthy Caucasian male or female (21-35 years old) without any medication. The raw data has sampling rate 100Hz and a sleep stage is assigned for each 30-second epoch by a human scorer. The sleep stage space S = {wake, REM, NREM1, NREM2, NREM3, NREM4}.

Due to its large size and often existing artifacts, each EEG recording is first transformed to capture the embedded, useful information. This process is called feature extraction. The most popular signal processing techniques for feature extraction include wavelet transform, fast Fourier transform[15], zero-crossing, parametric waveform recognition[16], etc. We adopted an approach based on power spectral properties of the EEG signal. The Thompson multi-taper method[17] is applied to 3-second moving window to obtain the localized power spectral density (PSD) with between-window-shift of 2.7 seconds. Consequently, we have 1,333 data points for each hour's sleep recording. Figure 2 shows the average log PSD for each stage. For each frequency f and each time point i, the logarithm of the PSD is normalized across time to obtain the Z score $Z_{f,i}$, where normalization is performed by first subtracting the mean and then dividing by the standard deviation.

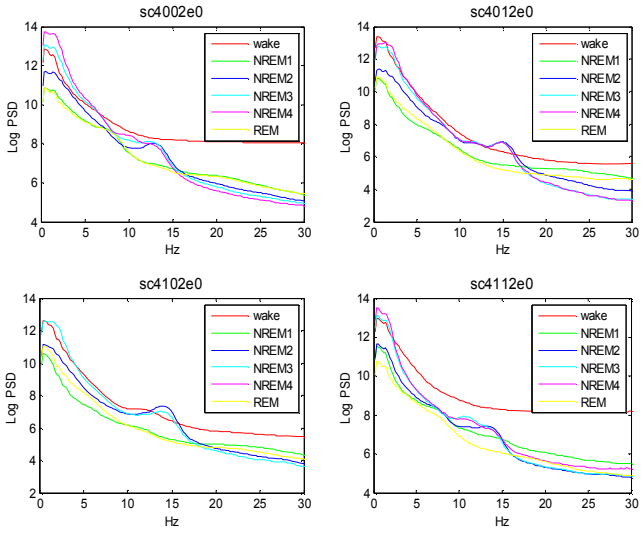

Figure 2. Stage-specific average logarithmic power spectral density of four human subjects.

We choose m=6 disjoint frequency bands: 0.2Hz-4Hz, 4.2Hz-8Hz, 8.2Hz-12Hz, 12.2Hz-16Hz, 16.2Hz-23Hz, and 23.2Hz-29Hz, which jointly contain 99% of the power of EEG waves. The justifications for selecting these frequency bands are as follows. First, as Figure 2 shows, the PSD curves of various stages are well separated within these bands. Second, it is well known that human sleep is characterized into different stages based on the frequency content of the delta-wave (0Hz-4Hz), theta-wave (4Hz–8Hz), alpha-wave (8Hz–13Hz), beta1-wave (13Hz–22Hz), and beta2-wave (22Hz–35Hz), which are similar to our frequency bands. Hence, the features contained within these bands should provide enough discrimination power for stage classification.

For the jth ($1 \le j \le 6$) band, at time point i, let $\tilde{x}_{i,j}$ denote the maximum Z score within this band. That is,

$$\tilde{x}_{i,j} = \max\{Z_{f,i}, \textit{ for all frequencies } f \textit{ in the } j\textit{th band}\}.$$

Since occasionally the recording has very large noise caused by movement, we truncate $\tilde{x}_{i,j}$ by

$$x_{i,j} = sign(\tilde{x}_{i,j}) \min\{|\tilde{x}_{i,j}|, A\}, \text{ where } A = 5.$$

Vector $\vec{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}]^T$ is the transformed observation element at time point *i*. The classification of the sleep recording is based on the x_i's across time.

We also applied our sleep stager to bird sleep data provided by Daniel Margoliash at the University of Chicago. The single-channel EEG raw data was collected at the forebrains of two zebra finches, each with one or two whole nights' sleep (8-hour period). The sampling rate is 1,000Hz and a sleep stage is assigned to each 3-second epoch by a human scorer. The sleep stage space S = {wake, REM, NREM}. We choose $m=4$ disjoint frequency bands: 1Hz-5Hz, 5.5Hz-10Hz, 10.5Hz-20Hz, and 20.5Hz-30Hz. In the Thompson multi-taper method[17], the between-window-shift is 0.9s.

6. Results

Our experiments were performed on a computer with one 2.2GHz Intel Core™ Duo T2600 processor and 2GB of memory. Feature extraction code is written in Matlab R2006a and classification code is written in R. For each human subject, the training data contains four segments, each of 30 minutes. Two tests were performed on two disjoint test data segments, each of 60 minutes. For each bird subject, the training data contains three segments, each of 10 minutes. Three tests were performed on three disjoint test data segments, each of 10 minutes. For both human subjects and bird subjects, each sleep stage has sufficient occurrences in every test data segment. For comparison, we also applied the widely used benchmark classifier of Gaussian Observation Hidden Markov Models (GOHMM)[7] to the same features as we used for the CRF classifier.

The feature extraction time for each 30-minute data segment is 80 seconds. The training time of the CRF classifier varies from 38 seconds to 230 seconds and labeling on test data takes less than one second. Thus, the CRF classifier can be used online. Table 1 and Table 2 report the accuracy obtained by the HMM classifier and the CRF classifier on human data and bird data, respectively. We obtained higher classification accuracy on the bird data than on the human data, which is expected in view of fewer sleep stages of birds. The same experiment is repeated using the feature of minimum Z-score in each frequency band and the results are similar. In most cases the CRF classifier achieves better accuracy than the HMM classifier with average improvement of about 8%. The average accuracy of the CRF classifier (83.7% for human data and 88.2% for bird data) already approaches the limit of automated sleep staging method, as there is only 80%-90% interscorer agreement in manual staging[9]. The HMM classifier, however, has an advantage of shorter training time, normally 30 seconds, which is expected given its strong model assumption of Gaussian observation.

Table 1. Classification accuracy on human data.

classification accuracy	Test 1		Test 2		average accuracy of two tests	
subject id	CRF	HMM	CRF	HMM	CRF	HMM
sc4002e0	81.6%	69.8%	77.8%	66.9%	79.7%	68.4%
sc4012e0	87.0%	72.4%	71.5%	72.7%	79.3%	72.6%
sc4102e0	89.7%	83.5%	86.7%	82.7%	88.2%	83.1%
sc4112e0	82.2%	69.1%	93.0%	88.9%	87.6%	79.0%
average accuracy of four subjects					83.7%	75.8%

Table 2. Classification accuracy on bird data.

classifi-cation accuracy	Test 1		Test 2		Test 3		average accuracy of three tests	
	CRF	HMM	CRF	HMM	CRF	HMM	CRF	HMM
subject 1	88.0%	82.0%	88.3%	89.5%	94.6%	72.7%	90.3%	81.4%
subject 2	86.7%	71.7%	89.2%	82.5%	82.6%	69.0%	86.2%	74.4%
average accuracy of two subjects							88.2%	77.9%

We evaluated our subject adaptation technique using the human data. In each test, we treated one human subject as the new subject s_{new} and varied the length L of s_{new}'s training data sequence D_{new} from 0 to 120 minutes. The other three human subjects are treated as old subjects and their entire training data sequences are used to obtain Θ's prior distribution. The classification accuracy achieved by subject adaptation on the test data sequence of s_{new} is called the *adaptation accuracy*. When s_{new}'s entire training data sequence is used to train the CRF without subject adaptation, the accuracy obtained by the CRF

classifier on the test data sequence of s_{new} is called the *final accuracy*. Two tests were performed for each human subject. In six of these eight tests, even when $L=0$ (i.e., no training data from s_{new}), we can obtain an adaptation accuracy between 70% and 90%, which is close to the final accuracy and improves slightly when L becomes larger. Figure 3 shows the classification accuracy for the other two tests (test 2 of sc4012e0 and test 1 of sc4112e0). There, the adaptation accuracy is below 50% when $L=0$. When L becomes larger, the adaptation accuracy improves and eventually reaches the final accuracy.

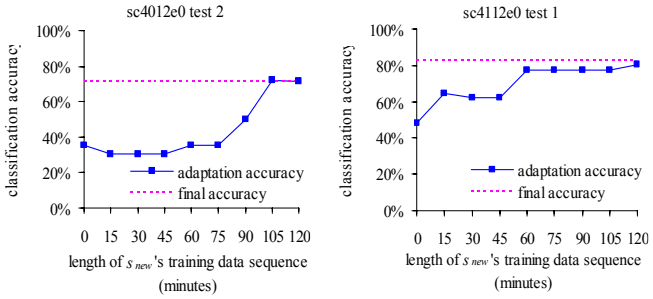

Figure 3. Classification accuracy achieved by subject adaptation.

7. Conclusion

One advantage of CRF is that the user can define potential functions that appropriately fit the specific application. This paper proposes using CRF and novel potential functions that have explicit physical meanings to perform the pattern recognition task of sleep staging. On both human subjects and birds, the average classification accuracy of our sleep stager almost approaches the theoretical limit and is about 8% higher than that of existing systems. Moreover, for a new subject s_{new} with limited training data D_{new}, we propose performing subject adaptation to improve classification accuracy. Even without any D_{new}, the average accuracy on s_{new} can be 70%. This accuracy increases with the size of D_{new} and eventually becomes close to the theoretical limit.

Acknowledgements

We thank Xing Wei and Zhenghua Fu for helpful discussions, and Daniel Margoliash for providing the bird sleep data.

References

1. Ancoli-Israel S, Roth T. Characteristics of insomnia in the United States: results of the 1991 national sleep foundation survey. Sleep 22 Suppl. 2, S347-S353, 1999.
2. Is this the bionic man? Nature 442, 164-171, 2006.
3. Rauske PL, Shea SD, and Margoliash D. State and neuronal class-dependent reconfiguration in the avian song system. J. NeuroPhysiology 89, 1688-1701, 2003.
4. Crick F, Mitchison G. The function of dream sleep. Nature 304, 111-114, 1983.
5. Sergejew AA, Tsoi AC. Markovian analysis of EEG signal dynamics in obsessive-compulsive disorder. In Advances in Processing and Pattern Analysis of Biological Signals, I. Gath, G.F. Inbar, Eds. (Plenum, New York, 1995), pp. 33-44.
6. Gersch W, Martinelli F, Yonemoto J, Low MD, and Ewan JA Mc. Automatic classification of electroencephalograms: Kullback-Leibler nearest neighbor rules. Science 205(4402), 193-195, 1979.
7. Penny WD, Roberts SJ. Gaussian observation hidden Markov models for EEG analysis. Technical report TR-98-12, Imperial College, London, 1998.
8. Becq G, Charbonnier S, Chapotot F, Buguet A, Bourdon L, and Baconnier P. Comparison between five classifiers for automatic scoring of human sleep recordings. FSKD 2002: 616-620.
9. Schaltenbrand N, Lengelle R, and Toussaint M *et al.* Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients. Sleep 19(1), 26-35, 1996.
10. Rabiner LR, Juang BH. Fundamentals of speech recognition. Prentice Hall, Englewood Cliffs, NJ, 1993.
11. Lafferty JD, McCallum A, and Pereira FC. Conditional random fields: probabilistic models for segmenting and labeling sequence data. ICML 2001: 282-289.
12. Huang K. Statistical Mechanics, 2nd Edition. John Wiley & Sons, New York, 1987.
13. Bickel PJ, Ritov Y, and Rydén T. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. Ann. Statist. 26(4), 1614-1635, 1998.
14. Kemp B. The Sleep-EDF Database. http://www.physionet.org/physiobank/database/sleep-edf, 2006.
15. Shaker MM. EEG waves classifier using wavelet transform and Fourier transform. IJBS 1(2), 85-90, 2006.
16. Hanaoka M, Kobayashi M, and Yamazaki H. Automatic sleep stage scoring based on waveform recognition method and decision-tree learning. Systems and Computers in Japan 33(11), 1-13, 2002.
17. Thomson DJ. Spectrum estimation and harmonic analysis. Proc. IEEE 70(9), 1055-1096, 1982.
18. Sha F, Pereira FC. Shallow parsing with conditional random fields. HLT-NAACL 2003: 134-141.
19. Penzel T, Kesper K, Gross V, Becker HF, and Vogelmeier C. Problems in automatic sleep scoring applied to sleep apnea. EBMS 2003: 358-361.
20. Scheuer ML. Continuous EEG monitoring in the intensive care unit. Epilepsia 43 Suppl. 3, 114-127, 2002.