Kent Academic Repository

Phan, Huy, Do, Quan, Do, The-Luan and Vu, Duc-Lung (2013) *Metric Learning for Automatic Sleep Stage Classification.* **In: 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2013). . pp. 5025-5028. IEEE E-ISBN 978-1-4577-0216-7.**

Metric Learning for Automatic Sleep Stage Classification

Huy Phan[1], Quan Do[2], The-Luan Do[3], and Duc-Lung Vu[4]

Abstract— **We introduce in this paper a metric learning approach for automatic sleep stage classification based on single-channel EEG data. We show that learning a global metric from training data instead of using the default Euclidean metric, the k-nearest neighbor classification rule outperforms state-of-the-art methods on Sleep-EDF dataset with various classification settings. The overall accuracy for Awake/Sleep and 4-class classification setting are 98.32 % and 94.49 % respectively. Furthermore, the superior accuracy is achieved by performing classification on a low-dimensional feature space derived from time and frequency domains and without the need for artifact removal as a preprocessing step.**

I. INTRODUCTION

The study of sleep is highly important in health care since sleep disorders affect the well-being and productivity of many individuals. The foundation of sleep classification was first laid in 1953 [7], since then has remained an important research topic. Sleep scoring based on polysomnography can be visually performed by a human expert to classify every 30-second epochs of EEG data into different sleep stages, following Rechtschaffen and Kales (R&K) rules [6] and based on the structure of the signal. However, it is a very time-consuming and labor-intensive task.

During the last decades, different approaches for automatic sleep stage classification using EEG signal have been proposed. Most of them are similar in the way that features characterizing each EEG data epoch will be first extracted, followed by a classification algorithm to assign class label to each data epoch. Feature extraction usually relies on time-domain analysis [14], spectral analysis [13], wavelet decomposition [8] [12], and even unsupervised feature learning [15]. Typical classification algorithms are neural networks [8], hidden makov models [9], k-means clustering [10], k-nearest neighbors (kNN) [11], support vector machines [12] to name a few. Comparative study on performance of these classifiers has also been conducted in [11]. Sleep stage classification based EEG data also involves in using single-channel EEG signal [8], multi-channel EEG signals [12] [13], and multimodal combination with other signals such as EOG, EMG and ECG signals [12].

In this paper, we re-visit the fundamental issue of machine learning: how to measure dissimilarity/similarity between samples. Without prior knowledge, Euclidean distance is implicitly employed in most proposed classifiers to measure the dissimilarities between examples represented as vector inputs. Although the Euclidean distance is convenient and intuitive, it ignores the fact that the semantic meaning of "similarity" is inherently task- and data-dependent [1]. Ideally, the distance metric should be adapted to the particular problem. Inspired by this, we show in this paper learning a global distance metric from labeled examples and using it in kNN classification significantly improve performance for sleep staging although only a few features extracted from single-channel EEG data. We also study the effects of dimensionality reduction, which is implicit during learning distance metric, on the classification accuracy.

II. EEG SIGNAL, SLEEP STAGES, AND SLEEP EEG DATASET

A. EEG Signal

Based on electrical recordings taken on the scalp of a subject, the Electroencephalogram (EEG) signal measurements are able to provide information about activities of the brain. It is the most important signal in sleep stage classification no matter manual scoring by human experts or automatic classification systems. Analyzing the information obtained from the EEG measurements can help carry out inference and studies about sleep.

B. Sleep Stages

According to R&K sleep scoring standard [6], sleep are divided into two major stages named rapid eye movement (REM) and non-rapid eye movement (NREM). Further, NREM is divided into four sub-stages: 1, 2, 3, and 4, making up to totally 6 sleep stages (including awake).

Stage 1, usually lasting between 1-5 minutes and contributing 4-5% of total sleep, is a transition stage between wakefulness and sleep. It consists of a low-voltage EEG tracing with well-defined alpha and theta activity, occasional vertex spikes, and slow rolling eye movements (SEMs) [8].

Stage 2 is considered as the "baseline" of sleep and comprises 45-55% of complete sleep duration. It is characterized by a relatively low-voltage, mixed-frequency EEG background buried in the occurrence of sleep spindles [16] and K-complexes [17]. Alternatively, high-voltage delta waves may appear up to 20% of Stage 2.

Stage 3 constitutes 4-6% of total sleep duration and usually only appears in the first one-third of the sleep episode. During at least 20% and at most 50% of this stage, EEG signals exhibit strongly discriminative characteristics with ≤ 2 Hz frequencies and ≥ 75 V amplitudes (delta waves).

Stage 4 contributes 12-15% to the total sleep duration. Characteristics of Stages 3 and 4 are quite similar and they are known as slow wave sleep (SWS). The difference is that, during Stage 4, delta waves cover $\geq 50\%$ of the record.

[1,2,3,4]H. Phan, Q. Do, T. L. Do, and D. L. Vu are with Department of Computer Engineering, University of Information Technology, Km 20, Ha Noi highway, Linh Trung ward, Thu Duc district, HCMC, Vietnam `huypq at uit.edu.vn, quando at uit.edu.vn, luandt at uit.edu.vn, lungvd at uit.edu.vn`

REM, in which dreaming occurs, is characterized by rapid eye movements under closed eyelids, motor atonia, and low voltage EEG patterns. REM constitutes 20-25% of a normal sleep night. During REM sleep, the brain activity is reversed from Stage 4 to a pattern similar to Stage 1 [8].

C. Sleep EEG Dataset

The experiments presented in this paper are based on the Sleep-EDF database [3] obtained from the PhysioBank online resource. We only used four recordings: sc4002e0, sc4012e0, sc4102e0, and sc4112e0 which were recorded in 1989 over the course of one full day from healthy male and female pioneers between 21 and 35 years old. These recordings include horizontal electrooculogram (EOG), Fpz-Cz and Pz-O channels of EEG, submental-electromyogram (EMG) envelope, oronasal airflow, and rectal body temperature. Since we aimed at illustrating efficiency of learning a metric for sleep staging with single EEG channel, we only use Fpz-Oz EEG sampled at 100 Hz for analysis. The hypnogram data accompanying with the dataset was used as ground truth to evaluate performance of the classification algorithm in the experiments. It was created by manual scoring according to R&K using the two EEG channels.

III. SLEEP STAGE CLASSIFICATION MODEL

A. Feature Extraction

We derive the following features for each 30-second single-channel EEG data epoch.

Statistical measures These include variance, skewness, and kurtosis. The variance is to characterize the spread of the data while the skewness represents the asymmetry around the sample mean. The kurtosis measures how the distribution is prone to outliers.

Spindle score Similar to the features used in [2], the overall spindle score presents the percentage of the signal classified as spindle activity. Spindle activity refers to segments of the EEG signal with two peaks and two troughs created by the difference between five consecutive points changes from positive to negative in straight. In order to search for these patterns at different frequencies, a lag parameter was used. This parameter indicates the width in sample points of each rise or fall. In this study, with data sampled at 100 Hz, the lag parameter was set to 5 to allow for detection of spindles in 8-12 Hz frequency range.

Permutation entropy (PE) The permutation entropy is to measure the "uncertainty' of the EEG signal. Similar to the spindle score, it searches for patterns such as peaks, troughs, and slopes in the signal but its values depends on the distribution of these patterns. An equal distribution of all patterns will produce a maximum value while a minimum value will be induced when only a single pattern is present. Here we consider two PE measures respective to two lag parameter values of 1 and 2. As in [4], the threshold was set at 1% of the interquartile range of the data.

Power in different frequency bands We computed the total power in five frequency bands, including delta (up to 4 Hz), theta (4-7.5 Hz), alpha (7.5-12 Hz), beta (12-26 Hz), and gamma (above 26 Hz) [18], adding other five features.

Total power The total power in all five frequency bands was also computed as used as a feature.

Properties of log power These features present characteristics of the log of the power spectral density (PSD) on delta and alpha bands [5]. First, we performed linear fitting in delta (0.5-4 Hz) and alpha (8-17 Hz) frequency bands to determine the slope and offset. After that, we could extract the maximum values of the PSD above the linear estimate in the delta and alpha frequency ranges. Generally, a prominent peak will produce a large value. Two frequency values correspond to these maximum values are also extracted.

Power fractions The low and high power fractions with respect to low frequency ranges (delta and theta ranges) and high frequency ranges (beta and gamma ranges) were also computed by summing the power in the individual ranges and dividing by the total power.

By this feature extraction, each 30-second EEG data epoch is represented by a 17-dimensional vector inputs. These inputs are then used in classification algorithm. To motivate further research, the Matlab source code will be made publicly available.

B. Metric Learning with Large Margin Nearest Neighbor

Given a training data set of instances labeled with their true class labels, the kNN algorithm assigns the class label to a new data instance. The class label is obtained from the majority vote of the k closest instances in training data. In order to measure the closeness, a distance metric needs to be pre-defined. Without prior knowledge, a Euclidean distance is implicitly employed. However, as discussed, the distance metric should be adapted to the particular problem and learned from the training data. Large Margin Nearest Neighbors (LMNN) algorithm [1] can learn this global metric in a supervised fashion to improve the classification accuracy of the kNN rule.

Let the training data consist of a data set $D = (x_1, y_1), \ldots, (x_n, y_n) \subset \mathcal{R}^m \times \mathcal{C}$, where the set of possible classes is $\mathcal{C} = 1, \ldots, c$. LMNN learns a metric of type:

$$D(x_i, x_j) = (x_i - x_j)^T \mathbf{M}(x_i - x_j) = ||\mathbf{L}(x_i - x_j)||^2 \quad (1)$$

The matrix $\mathbf{M}$ needs to be positive semi-definite for the metric $D(\cdot, \cdot)$ to be well defined. In the special case of the Euclidean metric, the matrix M is identity. For generalization, the metric $D(\cdot, \cdot)$ is often referred as Mahalanobis metric.

LMNN optimizes the matrix $\mathbf{M}$ with two objectives: minimize the distances between examples in the same class, and in the meantime keep examples from different classes far away. Fig. 1 show an example of such an optimization.

With the learned metric in Fig. 1, the input vector $\vec{x_i}$ is surrounded by training instances of the same class. If it was a test sample, it would be classified correctly under $k = 3$ nearest neighbor rule. It leads to the final optimization problem as in (2):
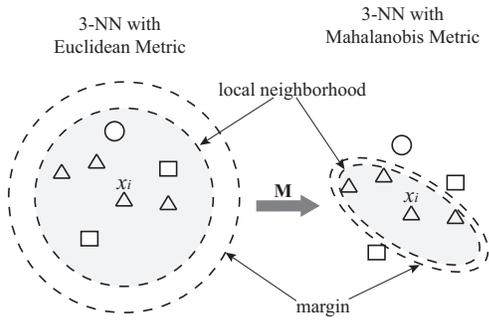
Fig. 1. Illustration of 3-NN algorithm with Euclidean and Mahanalobis metrics.

$$\text{minimize} \quad \sum_{ij} \eta_{ij}(x_i - x_j)^T M(x_i - x_j)$$

$$+ C \sum_{ijl} \eta_{ij}(1 - y_{il})\xi_{ijl} \quad (2)$$

$$\text{subject to} \quad \begin{cases} \text{(i) } (x_i - x_l)^T M(x_i - x_l) \\ \quad - (x_i - x_j)^T M(x_i - x_j) \geq 1 - \xi_{ijl} \\ \text{(ii) } \xi_{ijl} \geq 0 \\ \text{(iii) } M \succeq 0 \end{cases}$$

where the binary value y_{ij} indicates whether samples x_i and x_j are in the same class and the binary value η_{ij} indicating whether x_j is a selected nearby neighbor of x_i with the same class, and ξ_{ijl} are slack variables.

Intuitively, the first term in the objective function is to minimize the distances between all training samples and their selected neighbors. The second term is to maximize the margin between same-class distances (x_i to x_j) and different-class distances (x_i to x_l) of all training samples. The margin is relaxed by slack variables and is of exactly one unit fixed by the scale of the matrix $\mathbf{M}$. Any alternative choice $C > 0$ would result in rescaling of $\mathbf{M}$ by a factor of $1/C$. We use the implementation of this algorithm provided by the authors of [1], thanks for their efforts.

Furthermore, LMNN can be used as a supervised dimensionality reduction by optimizing on matrix $\mathbf{L}$ rather than matrix $\mathbf{M} = \mathbf{L}^T\mathbf{L}$ and constrain $\mathbf{L}$ to be rectangular of size $r \times m$, where r is the desired output dimensionality which is presumed to be smaller than the input dimensionality, m.

IV. EXPERIMENT

A. 4-class Classification

For Stage 1 of NREM sleep and REM sleep, EEG signals are similar and, thus, can be merged into one class. Hence, we attempt to classify four sleep stages consisting of Awake, Stage1 + REM, Stage 2 and Slow Wave Stage (SWS). In addition, this partition is consistent with the previous work in [8] using the same dataset, allow us properly comparing performance. It is also considerable to notice that we neither preprocess the data for artifact removal nor apply bootstrapping to lessen affects of imbalanced data as in [8].

The number of data epochs of each class in the dataset is tabulated in Table I. We randomly divide the data set into training set (70 %) and test set (30 %). The features are

TABLE I
NUMBER OF DATA EPOCHS OF 4 CLASSES IN THE DATASET

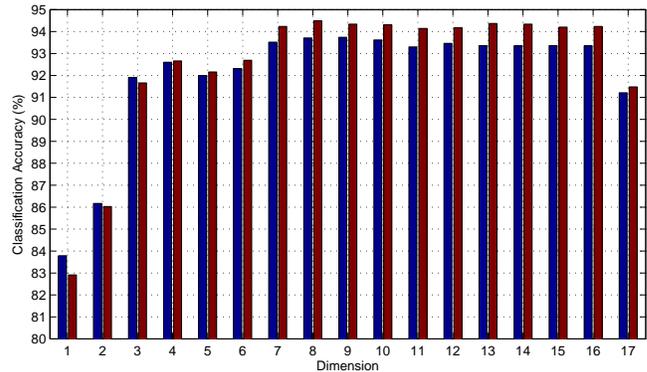| Wake	Stage 1 + REM	Stage 2	SWS
7722	1027	2036	529

Fig. 2. Training and testing accuracy (blue and red bars respectively) according to different dimensionality reduction settings.

extracted from 30-second segments of Pz-Oz channel. The transformation matrix is learned from the training data and the test data is used to evaluate the classification accuracy.

Since dimensionality of the feature space is quite small (17 dimensions), we investigate different dimensionality reduction by setting r to be running from 1 to 17 with a step of 1. The variation of overall training and testing classification accuracy according to different dimensionality is illustrated in the Fig. 2. The highest testing accuracy, 94.49 %, is obtained at $r = 8$ with the learned Mahalanobis metric, compared to 91.48 % by using the default Euclidean metric. For the sake of comparison, this result outperforms not only the work using Sleep-EDF dataset with similar setting [8] with the average accuracy of 93 %, but also other works using private multi-channel recorded data such as [12] with the average accuracy of 93 %. For further detail, Fig.3 exhibits 4-class classification confusion matrix over the test set with $r = 8$. Out of 4 classes, "Wake" and "S1 + REM" classes are most and least discriminative with the classification accuracy of 99.43 % and 75.83 % respectively. The superior overall testing accuracy is owning to that the "Wake" class' contributes the strongest weight due to its largest cardinality.

B. Awake/Sleep Classification

Considering Awake/sleep classification setting, similar experimental study is conducted. The whole dataset, out of which the number of data epochs of "Awake" and "Sleep" classes are 7722 and 3592 respectively, is divided into training set (70 %) and test set (30 %). The features are again extracted from 30-second segments of Pz-Oz channel. The study of dimensionality reduction shows that the average testing accuracy is up to 98.32 % with $r = 11$. This result outperforms the classification accuracy of 95 % reported in a recent study [12] which uses their own recorded multi-channel EEG data.

C. Discussion

It is worth noting that artifact removal as having been done in [8], which removed most artifact epochs, would

Fig. 3. 4-class classification confusion matrix respective to $r = 8$.

further benefit the learning process. Artifact epochs which are contaminated by eye movements, blinks, muscle, heart and line noise [19] are semi-automatically removed to refine the data. However, it would greatly degrade the automatic of the classification system since we need to visually inspect the data epochs before feeding to the system.

Other machine learning techniques may also benefit this proposed metric learning approach. As can be seen from Table I, the number of data epochs of the "Wake" class is much larger than that of the "SWS" class. Boostrapping [20] is a technique to generates more data from the original dataset to reduce the skewness of the dataset when cardinality of a class is much smaller comparing to other classes. As a result, it facilitates the learning process. Exploring other features which better represents the data is a general technique to boost the performance of a classification system. Higher feature space for sleep stage classification that has been investigated in literature work such as [14] [8] can be readily integrate into the metric learning framework. The features can be also learned from the data itself instead of feature engineering [15]. In addition, multi-channel EEG and other signals like EOG, EMG and ECG have been proved to be useful for not only visually sleep scoring [21] but also automatic sleep stage classification [18]. By learning a metric over multimodal data, we can increase accuracy and reliability of the classification system.

V. CONCLUSIONS

A metric learning approach has been proposed to address automatic sleep stage classification. By learning a Mahalanobis distance metric from labeled samples, the kNN classification rule empirically outperforms the classification accuracy reported in other literature works using similar settings. Furthermore, we do not perform any preprocessing steps such as artifact removal or boostrapping. This excellent result indicates the potential of metric learning in addressing biomedical signal processing problems.

REFERENCES

[1] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification". J. of Machine Learning Research, vol. 10, pp. 207-244, Jan. 2009.

[2] E. C. McKay, J. W. Sleigh, L. J. Voss, and J. P. Barnard, "Episodic waveforms in the electroencephalogram during general anaesthesia: A study of patterns of response to noxious stimuli". Anaesthesia and Intensive Care, vol. 38, no. 1, pp. 102-112, Jan 2010.

[3] B. Kemp, "The Sleep-EDF Database". http://www.physionet.org/physiobank/database/sleep-edf/. Accessed January 2012.

[4] E. Olofsen, J. W. Sleigh, and A. Dahan, "Permutation entropy of the electroencephalogram: A measure of anaesthetic drug effect". British J. of Anaesthesia, vol, 101, no. 6, pp. 810-821, 2008.

[5] K. Leslie, J. Sleigh, M. Paech, L. Voss, C. Lim, and C, Sleigh, "Dreaming and electroencephalographic changes during anesthesia maintained with propofol or desflurane". Anesthesiology, vol. 111, no. 3, pp. 547-555, 2009.

[6] A. Rechtschaffen and A. Kales, "A Manual of Standardized Terminology, Technique and Scoring System for Sleep Stages of Human Subjects". Public Health Service, U.S, Government Printing Office, Washington, DC, 1968.

[7] E. Aserinsky and N. Kleitman, "Regularly occurring periods of eye motility, and concomitant phenomena during sleep". Science, vol. 118, pp. 273-274, 1953.

[8] F. Ebrahimi, M. Mikaeili, E. Estrada, and H. Nazeran, "Automatic Sleep Stage Classification Based on EEG Signals by Using Neural Networks and Wavelet Packet Coefficients," Proc. IEEE EMBS'08, pp. 1151-1154, 2008.

[9] L. Doroshenkov, V. Konyshev, S. Selishchev, "Classification of human sleep stages based on EEG processing using hidden markov models". Biomedical Engineering, vol.41, no. 1, pp. 25-28, 2007.

[10] S. Gunes, K. Polat, and S. Yosunkaya, "Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting". Expert Systems with Applications, vol.37, no. 12, 2010.

[11] G. Becq, S. Charbonnier, F. Chapotot, A. Buguet, L. Bourdon, and P. Baconnier, "Comparison between five classifiers for automatic scoring of human sleep recordings," In: S.K. Halgamuge, L. Wang (Eds.), Studies in Computational Intelligence (SCI), vol.4: Classification and Clustering for Knowledge Discovery, Springer-Verlag, pp. 113-127, 2005.

[12] S. Khalighi, T. Sousa, D. Oliveira, G. Pires, and U. Nunes, "Efficient feature selection for sleep staging based on maximal overlap discrete wavelet transform and SVM," Proc. in EBMC'11, pp. 3306-3309, 2011.

[13] Y. Li, K.M. Wong, and H. de Bruin, "Electroencephalogram signals classification for sleep state decision a Riemannian geometry approach". IET Signal Processing, vol. 6, no. 4, pp. 288-299, 2012.

[14] C. Vural and M. Yildiz, "Determination of Sleep Stage Separation Ability of Features Extracted from EEG Signals Using Principle Component Analysis". J. of Medical Systems, vol. 34, no. 1, pp. 83-89, 2010.

[15] M. Langkvist, L. Karlsson, and A. Loutfi, "Sleep Stage Classification Using Unsupervised Feature Learning". Advances in Artificial Neural Systems, doi:10.1155/2012/107046, 2012.

[16] L. De Gennaro and M. Ferrara, "Sleep spindles: an overview". Sleep medicine reviews, vol. 7, no. 5, pp. 423-440, 2003.

[17] A. L. Loomis, E. N. Harvey, and G. A. Hobart, "Cerebral states during sleep as studies by human brain potentials". J. of Experimental Psychology, vol. 21, no. 2, pp. 127-144, 1937.

[18] B. A. Lopour, S. Tasoglu, H. E. Kirsch, J. W. Sleigh, and A. J. Szeri, "A continuous mapping of sleep states through association of EEG with a mesoscale cortical model". J. of Computational Neuroscience, vol. 30, no. 2, pp. 471-487, 2011.

[19] T-P. Jung, S. Makeig, C. Humphries, T. W. Lee, M. J. McKeown, V. Iragui, and T. J. Sejnowski, "Removing Electroencephalographic Artifacts by Blind Source Separation," Psychophysiology, vol. 37, pp. 163-178, 2000.

[20] Y. Hamamoto, "A bootstrap technique for nearest neighbor classifier design". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 1, pp. 73-79, 1997.

[21] J. Shepard, Jr. M.D, "Atlas of Sleep Medicine". Futura Publishing Company, 1991.