34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Adaptive Automatic Sleep Stage Classification under Covariate Shift

Sirvan Khalighi, *Student Member, IEEE,* Teresa Sousa, Urbano Nunes, *Senior Member, IEEE*

Abstract —**Current automatic sleep stage classification (ASSC) methods that rely on polysomnographic (PSG) signals suffer from inter-subject differences that make them unreliable in facing with new and different subjects. A novel adaptive sleep scoring method based on unsupervised domain adaptation, aiming to be robust to inter-subject variability, is proposed. We assume that the sleep quality variants follow a covariate shift model, where only the sleep features distribution change in the training and test phases. The maximum overlap discrete wavelet transform (MODWT) is applied to extract relevant features from EEG, EOG and EMG signals. A set of significant features are selected by minimum-redundancy maximum-relevance (mRMR) which is a powerful feature selection method. Finally, an instance-weighting method, namely the importance weighted kernel logistic regression (IWKLR) is applied for the purpose of obtaining adaptation in classification. The classification results using leave one out cross-validation (LOOCV), show that the proposed method performs at the state-of-the art in the field of ASSC.**

I. Introduction

The development of automatic sleep stage classification (ASSC) and monitoring based on Rechtschaffen and Kales standard (R&K) [1] and the American Academy of Sleep Medicine (AASM) has consistently been an important research topic. ASSC is highly desirable, to save time and improve the agreement levels of sleep scoring versus the traditional scoring by experts. Most of the ASSC methods are based on electroencephalographic (EEG) records, sometimes in combination with electrooculographic (EOG) and electromyographic (EMG) records. They categorize sleep-wake cycle in awake, non rapid eye movement (NREM) and rapid eye movement (REM) sleep stages. NREM sleep is further divided into three stages: N1, N2 and N3 [2]. Current ASSC methods reported in scientific publications are based on classical supervised and unsupervised learning approaches like linear discriminate analysis (LDA), hidden markov model (HMM), fuzzy clustering or kernel methods such as artificial neural networks (ANN), and support vector machine (SVM) [3-10]. These ASSC methods that rely on PSG signals are inadequate to handle inter-subject variability. In these methods, it is implicitly assumed that existent kernel is completely correct, and without any kernel error. Since sleep features may vary due to health problems like apnea, recording environment changes and subjects' physical conditions, the training and test probability distributions are not necessarily the same in practice. Actually, probability distributions of training and test subjects are related to each other in some sense, and it can be learned something about test probability distribution via the training set. One of the assumptions is to consider a connection between train and test domains based on instance weighting for covariate shift [11]. Covariate shift methods, reweight training samples in the training domain to minimize the error of predictions and to match with a new test domain. These methods firstly, estimate density ratio from the test and training domains, then the estimated density ratio is used to resample the training samples, or to train with weighted examples [12]. An adaptive ASSC approach has been developed based on unsupervised domain adaptation. The main goal is to cope with variations between a new subject and training set, aiming to improve sleep stage classification in two applications: sleep/awake detection and multiclass sleep stage classification. In both cases the classification is based on six EEG, two EOG channels and one EMG channel by using temporal, parametric and time-frequency features. A set of significant transformed and normalized features are selected by a minimum-redundancy maximum-relevance (mRMR) algorithm [13]. To cope with the non-stationarity, a weighted version of kernel logistic regression (KLR) , known as importance weight kernel logistic regression (IWKLR), is used as classifier [14].

This work has been supported by the QREN funded project SLEEPTIGHT, with FEDER reference CENTRO-01-0202-FEDER-011530, and by the Portuguese Technology Science Foundation (FCT).

The authors are with the Institute for Systems and Robotics (ISR-UC), University of Coimbra, 3030-290 Coimbra, Portugal.
e-mails: {skhalighi,tsousa,urbano}@isr.uc.pt.

II. Materials and Methodology

The proposed system is organized in various interoperating parts as described in the following.

A. Data Collection

Data from all-night PSG records, each with a duration around 8 hours (acquired by a SomnoStar Pro; Viasys SensorMedics), were provided by the Laboratory of Sleep from Hospital Centre of Coimbra. All EEG, EOG and EMG recordings were performed with a sampling rate of 200 Hz. The dataset comprises data from eight subjects, six males and two females with ages between 22 and 76 years old (mean = 50 years; STD = 19.05 years) acquired since year of 2009 until 2011. In 87% of the subjects, sleep apnea events were reported. The international 10-20 standard electrode placement system was used for EEG recording. Six EEG, two EOG and one EMG channels were used in our evaluation: F3-A2, C3-A2, O1-A2, F4-A1, C4-A1, and O2-

A1, right EOG (R-EOG)-A1, left EOG (L-EOG)-A2 and chin EMG signal X1 for all the subjects.

B. Structure of the Proposed Method

After applying common preprocessing, such as a notch filter at 50 Hz, band-pass Butterworth filter with lower cutoff of 0.5 Hz and higher cutoff of 45 Hz, and segmenting the signals in 30 second epochs, some features are extracted using several methods in the time-frequency, temporal and frequency domain as will be described below. PSG signals are traditionally analyzed in the frequency domain, since each sleep stage is characterized by a specific pattern of frequency contents. Moreover, PSG signals are non-stationary; therefore time-frequency transformations like wavelets are very useful. Due to superiority of the MODWT [10], [17], [18] versus discrete wavelet transform (DWT), a MODWT of depth 6 with Daubechies order four (db4) is applied to every 30 second epochs with a sampling rate of 200 Hz. The frequency ranges are broken down within δ range (<4 Hz), θ range (4–8 Hz), α range (8–13 Hz) and β range (13–30 Hz). To represent the time-frequency distribution of the EEG, EOG and EMG signals, features such as energy [10], percent of energy [10], mean and standard deviation are extracted from each sub-band. Furthermore, relative spectral power [7], harmonic parameters [4], percentile 25, 50, 75 [7], and skewness [7], are extracted as the temporal and frequency based features. To reduce the influence of extreme values, the transformation $\boldsymbol{X} = \log(\boldsymbol{Y})$ [19], is carried out where Y denotes the original matrix of features, and $\boldsymbol{X} = \{x_{ij}; i = 1,2,\dots,N \text{ and } j = 1,2,\dots,M\}$ (where N and M denote the number of subjects and the number of features, respectively), and to avoid features in greater numeric ranges dominating those in smaller numeric ranges, the transformed features, are normalized to the interval 0-1 [20]. Furthermore, a reduction in the dimension of the raw input variable is done by using the mRMR algorithm [13]. As illustrated in Figure 1, to handle the adaptive classification, covariate shift adaptation method is followed. Covariate shift is defined as a situation where the same observation $x \in X$, (X denotes a set of observations), with the same conditional distribution Y, (where Y denotes class labels) are in training and test domains. However, the marginal distributions of x may be different in source and the target domains. Formally, it assumes that $P_{tr}(Y|X = x) = P_{te}(Y|X = x)$ for all $x \in X$, but $P_{tr}(X) \neq P_{te}(X)$ [14]. At first glance, it may appear that covariate shift is not a problem in ASSC systems, because, we are only interested in $P(Y|X)$; and it assumes $P_{tr}(Y|X) = P_{te}(Y|X)$; but there is one question, why would the classifier learned from the source domain does not perform well on the target domain, even if $P_{tr}(X) \neq P_{te}(X)$. Shimodaira in [11] showed that this covariate shift becomes a problem when poorly specified models are used. In classification under covariate shift, the ratio $P_{te}(x,y)/P_{tr}(x,y)$ in the main equation of optimal model selection can be rewritten as follows [15]:

$$\frac{P_{te}(x,y)}{P_{tr}(x,y)} = \frac{P_{te}(y)}{P_{tr}(y)} \frac{P_{te}(x|y)}{P_{tr}(x|y)} = \frac{P_{te}(y)}{P_{tr}(y)} . \quad (1)$$

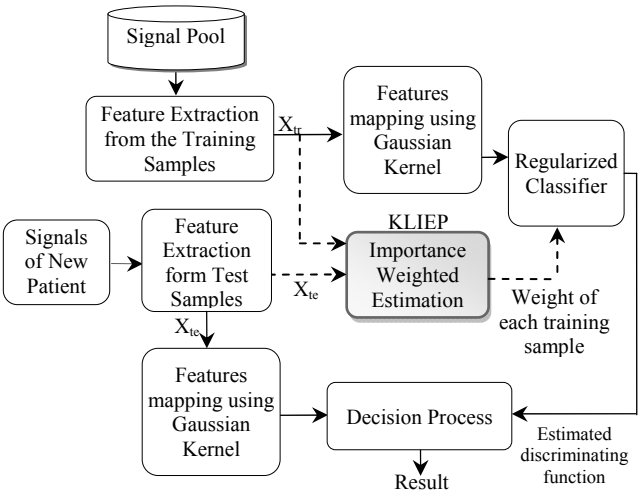

Figure 1. Adaptive Structure

Therefore, weighting each training instance by (1) could be useful. Actually, the influence of covariate shift can be asymptotically canceled by weighting the log-likelihood terms by accurately estimating the density ratio which is called the *importance* [16]:

$$W(X) = \frac{P_{te}(X)}{P_{tr}(X)} \quad (2)$$

where $P_{te}(X)$ and $P_{tr}(X)$ are test and training input densities, respectively. The importance weight $W(X)$ is unknown in practice, and needs to be estimated from data. A naïve approach consists on estimating the training and test densities, separately from training and test input samples, and then estimating the importance weight by taking the ratio of the estimated densities. However, direct density estimation is known to be a hard problem, particularly in high-dimensional cases [16]. Therefore, some other reweighting approaches can be used such as minimizing classification error of $P_{tr}(X)$ versus $P_{te}(X)$, minimizing the Maximum Mean Discrepancy (MMD) between $P_{tr}(X)$ and $P_{te}(X)$, and minimizing Kullback-Leibler (KL) divergence between $P_{tr}(X)$ and $P_{te}(X)$ [16]. The Kullback-Leibler importance estimation procedure (KLIEP) [16] was used once it integrates a built-in model selection procedure. This method allows us to directly learn the importance weight function, without going through the density estimation. Due to highly effects of the Gaussian width of KLIEP over the performance of importance weight estimation, the best value is calculated by cross validation [14].

C. Importance Weight Kernel Logistic Regression

Kernel logistic regression (KLR) is a kernelized variant of logistic regression. In KLR, the input vector is mapped to a high-dimensional space (feature space) and the logistic regression problem is solved in the feature space; the similarity in feature space can be implicitly computed via the kernel trick. The kernel trick allows converting a linear algorithm into a non-linear, keeping the computational simplicity. As mentioned, using importance sampling, the expectation over training samples is weighted according to their importance in the test distribution. Thus, by applying

TABLE I. SELECTED FEATURES USING MRMR. ALL FEATURES ARE EXTRACTED FROM SIX EEG, TWO EOG AND ONE EMG CHANNELS

Features	*MODWT*	*Harmonics*	*Relative Power*	*Percentile*	*Skewness*
Existent	**16**	24	45	120	180
Selected	6	16	**40**	**50**	**64**

TABLE II. EFFECT OF EMG AND EOG SIGNALS ON THE SLEEP SCORING

Sleep Stages	Sensitivity		
	With EMG &EOG	*Without EMG &withEOG*	*Without EOG &With EMG*
W	91.63	90.22	92.26
S1	47.73	49.31	42.91
S2	74.93	76.17	71.77
S3	85.11	83.89	75.65
REM	82.40	75.38	75.91

TABLE III. STATISTICAL ANALYSIS RESULT OF BINARY CLASSIFICATION

Sleep Stages	Sensitivity		
	SVM	*KLR*	*IWKLR*
Awake	**89.27**	75.45	73.75
Sleep	95.34	96.16	**96.55**

TABLE IV. STATISTICAL ANALYSIS RESULT OF MULTICLASS CLASSIFICATION

Sleep Stages	Sensitivity			Specificity		
	SVM	*KLR*	*IWKLR*	*SVM*	*KLR*	*IWKLR*
W	86.92	72.07	74.75	94.61	95.79	96.07
S1	44.41	44.19	41.06	95.50	93.99	94.82
S2	74.13	74.04	73.85	92.16	88.68	88.98
S3	88.33	87.73	89.31	96.16	95.25	94.19
REM	77.55	75.61	72.82	94.51	92.51	93.20

importance sampling to KLR, a weighted version of KLR, namely IWKLR is obtained [14].

Moreover, the IWKLR model contains two tuning parameters: the *kernel width* and the *regularization parameter*. Usually these tuning parameters are optimized based on cross validation. However, ordinary cross validation is no longer unbiased due to covariate shift; therefore, is not reliable as a model selection method. To cope with this problem, importance weighted cross validation (IWCV) [16] is used for unbiased model selection.

III. EXPERIMENTAL RESULTS AND DISCUSSION

The performance of the proposed algorithm was assessed using the eight subjects' dataset, mentioned in section II.*A*. In our experiments, a fourth order Daubechies with MODWT decomposition was adopted. KLR, IWKLR algorithms [14], and Libsvm toolbox [21], were used in the classification phase. For KLR and IWKLR, Gaussian kernel width 0.34, and regularization parameter 0.01 were used, and for SVM, the sigmoid kernel degree and C parameters were set to 0.13 and 1.25 respectively, as they produced the best observed results. The classification accuracy was determined using leave-one subject-out cross-validation (LOOCV).

The extracted feature sets and corresponding selected features, using mRMR method, are presented in Table I. A total of 385 (45 per each channel) features were extracted for each subject. The transformed and normalized feature matrix is fed into the feature selector. The total number of selected features by mRMR method was 159 for sleep/wake classification, and 176 for multiclass classification which has provided the best average accuracy when applying a grid search. As illustrated in Table I, the most relevant features are extracted from MODWT decomposition (64 selected features), Harmonic parameters (50), and relative power (40) and the least effective ones is Skewness (6 selected features). Relative power showed the best ratio of selected by extracted features. Some improvements were verified by using features extracted from EOG and EMG signals. As shown in Table II, by including the EOG signals almost 10% and 8% improvements in the classification of N3 and REM-stages were verified, respectively; which is due to remarkably differences of ocular movements in these two stages. The use of the EMG signal improved the classification accuracy of REM and awake stages around 8% and 4%, respectively. In these two stages, the EEG activity is almost similar, but the EMG activity is completely different (high and low muscle tone, respectively).

Sensitivity and specificity results of the proposed ASSC algorithm are shown in Tables III and IV. The results were obtained stage by stage using three different classification methods. For the sleep/awake detection case, as shown in Table III, SVM gave a better result in detecting the awake stage (89.29%) in comparison with KLR (75.45%) and IWKLR (73.75%).

Table IV gives a detailed comparison of sensitivity and specificity for the three analyzed classifiers: IWKRL, KLR, and SVM. The sensitivity value of awake stage is the most significant difference between these methods; where the sensitivity of using SVM (86.9%) is almost 12% and 15% better than the adaptive method using IWKLR (74.7%) and KLR (72.0%), respectively. It means that SVM-based method has a better ability to detect the awake stage. Furthermore, for the other sleep stages, the sensitivity and specificity values for the different methods were almost similar.

As shown in Figure 2, the average accuracies of IWKLR in multiclass and sleep/awake applications are better than KLR, which confirm that, adaptation based on importance weighting can improve the accuracies of ASSC. On the other hand, in more than 50% of the subjects, the average accuracies of the adaptive method and SVM-based method are almost similar, which is due to the model adaptation of IWKLR. In multiclass classification, the adaptive method based on IWKLR shows similar performance in N1, N2, N3 and REM stages, and in sleep/awake discrimination a better performance was observed in detecting the sleep stage. This experiment indicates that the inter-subject variation of PSG signals has higher expression in sleep stage, and detection of the sleep stages namely N1, N2, N3 and REM, had the highest effect in reducing the subject-independent performance. Although average accuracies of IWKLR and SVM are almost similar, the SVM-based method showed a

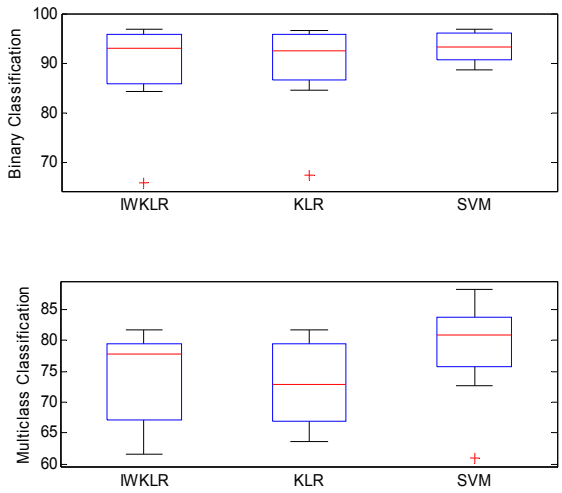

Figure 2. Accuracy results of binary and multiclass classification methods.

larger impact on the accuracy than the adaptive approach. Therefore, in practice, this method, for the different subjects, indicated a less balanced response than SVM. It could be due to negative transfer and weakness of KLR method.

The Cohen's Kappa (*k*) values varied from 0.65 for multiclass classification in IWKLR approach to 0.85 for binary classification in SVM-based method, which represents substantial and almost perfect concordance.

IV. Conclusion And Future Work

An adaptive sleep scoring method, based on unsupervised domain-adaptation, was proposed. To determine the validity of ASSC under covariate shift adaptation, IWKLR which is an instance of unsupervised domain adaptation methods was compared with KLR and SVM. For this purpose, several feature extraction methods have been applied. Features with higher positive impact in classification accuracy were the MODWT decomposition, harmonic parameters and relative power. Transformation and normalization in the feature domain played an important role in the remarkably improvement of classification accuracy. The integration of EOG and EMG channels indicated an improvement in classification, mainly for REM stage. The KLR-based adaptive approaches showed promising results in the multiclass sleep stage classification, namely with sensitivities average around 70%, with maximum of 89.31% for N3 stage and with minimum sensitivity of 44.19% for classification of N1 stage. However, it needs to be optimized in order to provide better results, especially as regards the classification of the awake state. In sleep/awake application, for detecting awake stage, the result of adaptive method (73.75%) was worse than the SVM-based method (89.27%). However, adaptive method indicated a better result (96.55%) in comparison to SVM (95.34%), in classifying the sleep stage. As a future work, the proposed adaptive method has to be validated in a larger dataset. Moreover, due to the observed quality of SVM, we are planning to investigate on SVM-based adaptive approaches aiming to improve ASSC results.

References

[1] A. Rechtschaffen, A. Kales, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects", *Bethesda, MD, U.S. National Institute of Neurological Diseases and Blindness, Neurol. Inform. Netw.*, 1968.

[2] C. Iber, Ancoli-Israel S, Chesson A., Quan SF, "The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications", *1th: Westchester, Illinois: American Academy of Sleep Medicine*, 2007.

[3] L. Fraiwan, K, Lweesy, N. Khasawneh, M. Fraiwan, H. Wenz, H. Dickhaus, "Classification of sleep stages using multi-wavelet time frequency entropy and LDA", *Methods of information in Medicine*, Vol.49, No.3,2010.

[4] F. Chapotot, G. Becq, "Automated sleep-wake staging combining robust feature extraction, artificial neural network classification, and flexible decision rules", *International Journal of Adaptive Control and signal Processing*, Vol. 24, Issue 5, May 2010.

[5] R. Agarwal, J. Gotman, "Computer Assisted sleep staging", *IEEE Transactions on Biomedical Engineering,* Vol. 48, No.12, 2001.

[6] E.F. Behbahani, A. Bastani ,"EEG Feature extraction using Hadamard transform for classification of sleep stages", *European Journal of Scientific Research*, Vol.50 No.2, pp.218-224, 2011.

[7] L. Zoubek, S. Charbonnier, S. Lesecq, A. Buguet, F. Chapolot, "Feature selection for sleep/wake stages classification using data driven methods", *Biomedical Signal Processing and Control*, Vol. 2, issue 3, Page 171-179, 2007.

[8] S. Gunes, K. Polat, S. Yosunkaya, "Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting", *Expert Systems with Applications*, Vol.37,Issue 12, 2010.

[9] H. Jo, J. Park, C. Lee, S. Ann, S. Yoo, "Genetic fuzzy classifier for sleep stage identification", *Computers in Biology and Medicine,* Vol. 40(7), pp. 629–634, July 2010.

[10] S. Khalighi, T. Sousa, D. Oliveria, G.Pires, U.Nunes, "Efficient feature selection for sleep staging based on maximal overlap discrete Wavelet Transform and SVM", *in IEEE EMBS'11,* 2011.

[11] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function", *Journal of Statistical Planning and Inference*, 90(2):PP 227–244, October 2000.

[12] M. Sugiyama, M. Krauledat, K.R. Muller, "Covariate shift adaptation by importance weighted cross validation", *Journal of Machine Learning Research*, vol.8, pp.985-1005, May 2007.

[13] H. Peng, F. Long, C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-Redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.27 (8), pp. 1226–1238, 2005.

[14] M. Yamada, M. Sugiyama, T. Matsui, "Semi-supervised speaker identification under covariate shift", *Signal Processing,* vol.90, no.8, pp.2353-2361, 2010.

[15] J. Jiang, "A literature survey on domain adaptation of statistical classifiers," online, Mar. 2008. Available: http://sifaka.cs.uiuc.edu/jiang4/domainadaptation/survey/dasurvey.pdf

[16] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. Bunau, "Direct Importance Estimation for Covariate Shift Adaptation", *Annals of the Institute of Statistical Mathematics*, Vol.60, no.4, pp.699-746, 2008.

[17] Percival, D. B., and Walden, A. T., *Wavelet Methods for Time Series Analysis*, Cambridge University Press, 2000.

[18] Saif Ahmad, "Temporal Pattern Identification and Summarization Method for Complex Time Serial Data," Ph.D. dissertation, University.of Surrey.

[19] G. Becq, S. Charbonnier, F. Chapotot, A. Buguet, L. Bourdon, P. Baconnier, "*Comparison between five classifiers for automatic scoring of human sleep recordings*", Studies in Computational Intelligence (SCI), Vol.4, Springer–Verlag, pp. 113–127, 2005.

[20] S. Aksoy, R. M. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval", *Pattern Recognition Letters*, 22, pp. 563–582, 2001.

[21] C. Chang and C.J. Lin, LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.