Ensemble learning based on overlapping clusters of subjects to predict microsleep states from EEG

Abdul Baseer Buriro, *Student Member*, *IEEE*, Reza Shoorangiz, *Member*, *IEEE*, Stephen J. Weddell, *Senior Member*, *IEEE*, Richard D. Jones, *Fellow*, *IEEE*

Abstract— **Microsleeps are brief and involuntary instances of complete loss of sleep-related consciousness. We present a novel approach of creating overlapping clusters of subjects and training of an ensemble classifier to enhance the prediction of microsleep states from EEG. Overlapping clusters are created using Kullback-Leibler divergence between responsive state features of each pair of training subjects. Highly correlated features within each overlapping cluster are discarded. The remaining features are selected via Fisher score based ranking followed by an average of 5-fold cross-validation areas under the curves of receiver operating characteristics (AUC$_{ROC}$) of a linear discriminant analysis (LDA) classifier. The decisions of LDA classifiers on overlapping clusters are fused using weighted average. We evaluated this new approach on 16-channel EEG data from 8 subjects who had performed a 1-D visuomotor task for two 1-h sessions. Joint entropy features were extracted from a 5-s window of EEG with steps of 0.25 s. Test performances were evaluated using leave-one-subject-out cross-validation. Our ensemble of overlapping clusters of subjects achieved a mean prediction performance, phi, of 0.42 compared with 0.39 for a single LDA classifier and 0.37 for generalized stacking.**

I. Introduction

Microsleeps are complete and unintentional sleep-related losses of consciousness of up to 15 s. Behavioural cues of microsleeps are eye closure, droopy eyes, and loss of visuomotor responsiveness [1, 2]. Although sleep restrictions increase the propensity of falling asleep, studies have shown that non-sleep-deprived and healthy people can also have frequent microsleeps [2, 3]. Drowsiness and fatigue substantially contribute to road crashes [4, 5]. The duration of microsleeps has also been reported to be highly correlated with the probability of accidents [6]. Microsleeps impair human performances to the extent that they can be fatal in extended-attention monotonous activities such as driving. Microsleep related accidents can potentially be averted if they can be non-invasively and accurately predicted.

Irrespective of feature reduction, feature selection, and classification techniques, the spectral components of EEG have frequently been used as features to analyze [7], detect [1, 8, 9], and predict [10, 11] microsleep states. However, we have found that pairwise joint entropy features were able to improve the prediction performance metrics (phi by 15% and AUC$_{ROC}$ by 3%) [12].

It has been well demonstrated that an ensemble learner/classifier, i.e., a combination of multiple classifiers, performs better than a single classifier [8, 9, 13, 14]. The creation of an ensemble classifier is based on the fusion of decisions from multiple classifiers and is constrained by uncorrelated classifier errors (diversity) [15]. Bootstrap aggregating (bagging), boosting, and generalized stacking are widely used ensemble techniques in machine learning literature. The diversity is achieved by training different (i.e., heterogeneous) classifiers on the same data set or using different partitions/subsets of the data to train the same classifier models. Data partitions are generally achieved through clustering techniques. The final decision is typically a majority or weighted vote of the individual decisions.

When the data of each training subject is large (e.g., > 10,000 feature points), a simple way of clustering is to treat the data of each subject as a cluster. Both Peiris et al. [8] and Ayyagari et al. [9] stacked the training subjects with multi-response linear regression (MLR) on pruned data to detect microsleep states at a resolution of 1.0 s. They achieved phis (stacked vs single classifier) of 0.39 vs 0.31 with 7 linear discriminant analysis (LDA) classifiers and 0.51 vs 0.38 with 7 leaky echo state network (ESN) classifiers respectively. Rahman et al. [15] partitioned the training data into overlapping clusters by repeatedly running the *k*-means algorithm, where each repetition was called a level.

The overall performance of stacked generalization based on data of individual training subjects may be limited when different subjects respond to a common task in a similar way. While its practicability may be a serious challenge when the number of training subjects grows. In a similar way, *k*-means clustering practically becomes a time consuming and redundant processing step if large data from different training subjects are concatenated. Besides, both numbers of clusters and levels are unknown in *k*-means based overlapping clusters.

In this paper, we propose a novel approach of ensemble learning that depends on overlapping clusters of training subjects and weighted voting. The idea is to create divergence-based overlapping clusters of individual training

* This work was supported by the New Zealand Brain Research Institute.

Abdul Baseer Buriro is with Department of Electrical and Computer Engineering at University of Canterbury, the Christchurch Neurotechnology Research Programme, and New Zealand Brain Research Institute, Christchurch, New Zealand, (e-mail: abdul.buriro@nzbri.org).

Reza Shoorangiz is with the Department of Electrical and Computer Engineering, University of Canterbury, the Christchurch Neurotechnology Research Programme, and the New Zealand Brain Research Institute, Christchurch, New Zealand (email: reza.shoorangiz@nzbri.org).

Stephen Weddell is with the Department of Electrical and Computer Engineering at University of Canterbury, and the Christchurch Neurotechnology Research Programme, Christchurch, New Zealand, (email: steve.weddell@canterbury.ac.nz).

Richard Jones is with the Department of Electrical and Computer Engineering at University of Canterbury, the Christchurch Neurotechnology Research Programme, and New Zealand Brain Research Institute, Christchurch, New Zealand, (email: richard.jones@nzbri.org).

978-1-5386-3646-6/18/$31.00 ©2018 IEEE

subjects. Divergence is a measure of dissimilarity between two distributions. If subjects X and Y, and subjects Y and Z respond to a common task in a similar way. Two clusters can intuitively be created that share the data of subject Y but, yet, still result in uncorrelated posterior probabilities. Two clusters of individual subjects, one being a subset or superset of the other cluster, are also likely to result in uncorrelated posterior probabilities due to different data distributions. This approach can be considered as an extension of weighted voting by clustering data in terms of individual training subjects.

This paper presents the concept of ensemble learning based on overlapping clusters of training subjects, and compares prediction performances (sensitivity, precision, phi, $\mathrm{AUC_{PR}}$, and $\mathrm{AUC_{ROC}}$) with a single LDA classifier and with generalized stacking.

II. Methods

A. Data

Fifteen non-sleep-deprived healthy subjects, aged 18-36 yr, used a steering wheel to track a 1-D pseudorandom target cursor as accurately as possible for two 1-h sessions, one week apart. During the task, EEG at 256 Hz, facial video at 25fps, and tracking error at 64 Hz were recorded. 16 EEG channels, namely Fp1, Fp2, F3, F4, F7, F8, C3, C4, O1, O2, P3, P4, T3, T4, T5, and T6 were placed per the international 10–20 system [2]. Data for the current study were from the 8 subjects who had at least one definite microsleep over the 2 sessions.

B. Gold Standard

Facial video and tracking error were used to form a gold standard, which comprised three classes: microsleep, responsive, and uncertain [10]. Incoherent tracking or unresponsiveness, along with a video rating of deep drowsiness was marked as a microsleep. Coherent tracking, irrespective of video rating was marked as a responsive. Data that could not be explained by either microsleep or responsive was marked as uncertain and discarded at the feature selection stage.

C. Feature Extraction and Preprocessing

Bandpass filtering of 1–45 Hz and artefact subspace reconstruction (ASR) were used to remove artefacts from the EEG [10]. The EEG signals were then decomposed into delta (0.5–4 Hz), theta (5–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–45 Hz) sub-bands, following the common average reference. Following the decimation to 128 Hz, EEG signals from each subband were segmented to 5 s epochs and with a step size of 0.25 s. Joint entropy features were extracted from each pair of EEG channels for each epoch and each subband. Joint entropy via k-nearest neighbour (kNN) with $k = 3$ was estimated as [16]

$$H(X) = \frac{d}{N}\sum_{i=1}^{N} \ln \epsilon_i - \psi(k) + \ln v + \ln N, \tag{1}$$

where N is the epoch length (i.e., 640 samples), d is the dimension of EEG time series X, ϵ is the distance between a sample point and its k^{th} neighbourhood, ψ is the digamma function, and v is the volume of a d-dimensional unit ball defined as

$$v = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}, \tag{2}$$

where Γ is the gamma function.

D. Overlapping Clustering

Training subjects were grouped into overlapping clusters according to the mean value of divergences between their responsive state features over alpha sub-band. Symmetric Kullback-Leibler divergence defined as

$$D_{KL} = D_{KL}(P||Q) + D_{KL}(Q||P) \tag{3}$$

was calculated via kNN ($k = 3$) using ITE toolbox [17]. The Kullback-Leibler divergence is defined as

$$D_{KL}(P||Q) = \sum_{i=1}^{N} P_i * \ln \frac{P_i}{Q_i}, \tag{4}$$

where P and Q are the probability distributions and N is number of responsive state features of a training subject.

The divergence matrix was then normalized with respect to the maximum value. Subjects of each row of the normalized divergence matrix were clustered if their normalized divergence was below a threshold of 0.15. Redundant clusters, i.e., clusters containing the same subjects, were discarded.

E. Feature Selection

M channels of EEG decomposed into B sub-bands gives $0.5BM(M\text{-}1)$ pairwise joint entropy features, i.e., 600 features per epoch. Irrelevant and redundant features result in poor classification performances. Features were therefore independently selected from the whole concatenated training data set for the single classifier, from each training subject data for generalized stacking, and from each cluster for ensemble of overlapping clusters of training subjects.

Linearly-correlated features ($|r| > 0.9$) from the training data set were discarded and the remaining features were sorted according to their Fisher scores. A higher Fisher score indicates a higher discrimination between classes. Features were then incrementally selected using 5-fold cross-validation $\mathrm{AUC_{ROC}}$ of an LDA classifier. The process was initiated by computing and then saving the mean $\mathrm{AUC_{ROC}}$ of the top-most feature. The process was iterated by combining the successive feature with the top-most feature and selecting it if the combined mean $\mathrm{AUC_{ROC}}$ was improved, otherwise, it was discarded.

F. Classification

In all approaches, LDA classifiers were trained on different combinations of data from 7 training subjects and tested on the 8th subject via the leave-one-subject-out cross-validation. Priors were incorporated to address the class imbalance in the training data sets. Prediction of microsleep was 0.25 s ahead of the gold standard, as shown in Fig. 1.

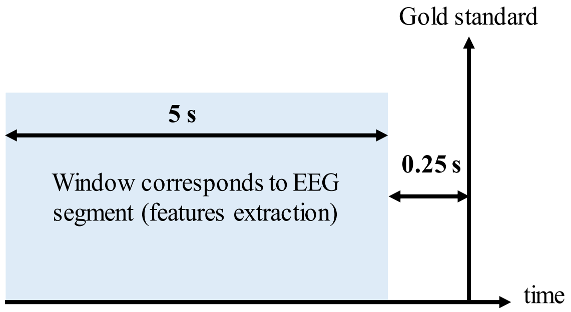

Fig. 1. Prediction of gold standard from corresponding epoch.

In the single classifier approach, training data of 7 subjects were concatenated. In generalized stacking one base (level-0) classifier per subject whereas, in ensemble of overlapping clusters of subjects, one classifier per cluster was used.

G. Fusion

The decision of individual classifiers on overlapping clusters were fused using weighted voting according to

$$\hat{Y} = \frac{1}{d}\sum_{i=1}^{d} a_i P(C|X, k_i), \qquad (5)$$

where d is the number of overlapping clusters k, a is the weight assigned to the posterior probabilities P of individual classifiers. C and X are the test class and test features respectively. Each weight was the reciprocal of the mean distance between a test point and centroids of the overlapping clusters of subjects. The weights are normalized with respect to their maximum values.

The decisions of individual base classifiers on training subjects were fused using meta (level-1) classifier. MLR was used as a meta classifier as it has proven to be an optimal level-1 classifier when fed with posterior probabilities (meta features) of the base (level-0) classifiers [13].

To address class imbalance at level-1 synthetic minority oversampling (SMOTE) was applied to the meta features. Coefficients (weights) of MLR and thresholds (to convert the continuous output of MLR into binary) were obtained via leave-one-training subject-out cross-validation. A threshold (from a range) at each cross-validation step was selected that gave the maximum correlation coefficient between MLR output and the cross-validation labels.

III. Results

Leave-one-subject-out cross-validation resulted in 8 training and test sessions. On average, our approach of overlapping clusters of subjects resulted in 3.9 (3-4) clusters per training session. In contrast, generalized stacking used 7 clusters in all the training sessions.

Metrics of microsleep state prediction performance (0.25 s ahead with a temporal resolution of 0.25 s) are presented in Table I. Both generalized stacking and ensemble of overlapping clusters of subjects gave better mean precision, $\mathrm{AUC_{ROC}}$ and $\mathrm{AUC_{PR}}$ than the single classifier approach. However, our proposed approach resulted in the best compromise between sensitivity and precision as indicated by a superior mean phi of 0.42.

TABLE I

Mean, Minimum, and Maximum Microsleep Prediction Performance (0.25 s) with Different Classification Approaches.

	Single classifier	Generalized stacking	Overlapping clusters
Sensitivity	0.71 (0.35-1.00)	0.59 (0.06-1.00)	0.68 (0.24-1.00)
Specificity	0.90 (0.58-0.99)	0.91 (0.64-1.00)	0.94 (0.73-1.00)
Precision	0.33 (0.00-0.95)	0.45 (0.02-0.97)	0.41 (0.02-0.98)
Phi	0.39 (0.04-0.80)	0.37 (0.05-0.87)	0.42 (0.06-0.88)
$\mathrm{AUC_{PR}}$	0.45 (0.02-0.96)	0.50 (0.02-0.97)	0.49 (0.02-0.98)
$\mathrm{AUC_{ROC}}$	0.94 (0.89-0.98)	0.96 (0.92-0.99)	0.95 (0.91-0.99)

Fig. 2 shows that phi accuracy is linearly related to the balance ratio between the classes. Subject-wise phi shows that our proposed ensemble of overlapping clusters of training subjects is effective in large class imbalance.

IV. Discussion

An ensemble of base classifiers trained on overlapping clusters of training subjects is proposed. To the best of our knowledge, joint entropy features combined with ensemble of overlapping clusters of subjects has resulted in the highest prediction phi of microsleep states.

Overlapping clusters are created using mean value of symmetric Kullback-Leibler divergences between joint entropy features of the training subjects. Features corresponding to the majority (responsive) class are chosen to get stable probability distributions involved in the

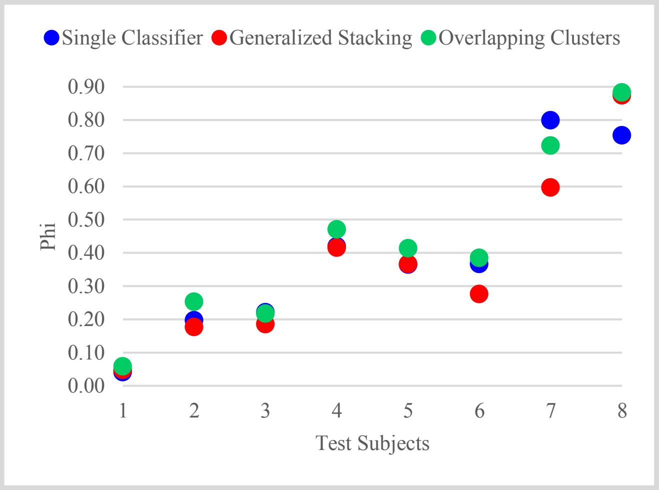

Fig. 2. Phi accuracy of different approaches for the independent test subjects, ordered with respect to their class imbalance ratios (number of microsleep states vs number of responsive states) of 1:813.40–1:2.26.

divergence. Features corresponding to minority class require less processing power but may result in fluctuating estimates of probability distributions for extremely small and varying number of sample points among different training subjects.

The slightly poorer phi performance of generalized stacking can be an indication of overfitting or lack of diversity among the base classifiers relative to a single classifier.

Practically, the ensemble of overlapping clusters of subjects can be advantageous over generalized stacking as it requires one classifier per cluster, whereas generalized stacking requires one base classifier per subject and one meta classifier per training session. A new training subject, irrespective of imbalance ratio, can easily be accommodated, based on its divergence with the other training subjects.

Compared to our previous work [12], with a single LDA classifier, slight improvement in mean sensitivity, phi, and AUC_{ROC} could be due to an increased number of folds in cross-validation to select features in the training sessions.

Irrespective of the classification approach, mean sensitivity, specificity, and test AUC_{ROC} (≥ 0.94) indicates that EEG-based microsleep state predictor incorporated with joint entropy features can confidently be used in critical applications in which prediction of true microsleep state is much more important than false microsleep state.

V. Conclusion

In this paper, we have presented a novel approach to create and train an ensemble classifier based on overlapping clusters. In this approach, base classifiers are trained on overlapping clusters obtained by grouping the data of different training subjects who performed similarly during all or most the task. The posterior probabilities of base classifiers are fused using weighted voting. The proposed approach resulted in better average phi metric compared to the single classifier and generalized stacking.

Despite these achievements, there is still some way to go to realize the accuracy desired for implementation in real-life applications.

References

[1] P. R. Davidson, R. D. Jones, and M. T. R. Peiris, "EEG-based lapse detection with high temporal resolution," *IEEE Trans. Biomed. Eng.,* vol. 54, pp. 832-839, May 2007.

[2] M. T. R. Peiris, R. D. Jones, P. R. Davidson, G. J. Carroll, and P. J. Bones, "Frequent lapses of responsiveness during an extended visuomotor tracking task in non-sleep-deprived subjects," *J. Sleep Res.,* vol. 15, pp. 291-300, Sep 2006.

[3] G. R. Poudel, C. R. H. Innes, P. J. Bones, R. Watts, and R. D. Jones, "Losing the struggle to stay swake: Divergent thalamic and cortical activity during microsleeps," *Hum. Brain Mapp.,* vol. 35, pp. 257-269, Jan. 2014.

[4] W. Vanlaar, H. Simpson, D. Mayhew, and R. Robertson, "Fatigued and drowsy driving: A survey of attitudes, opinions and behaviors," *J. Saf. Res.,* vol. 39, pp. 303-309, Jan. 2008.

[5] L. M. Swanson, C. Drake, and J. T. Arnedt, "Employment and drowsy driving: A survey of American workers," *Behav. Sleep Med.,* vol. 10, pp. 250-257, Oct. 2012.

[6] B. Sirois, U. Trutschel, D. Edwards, D. Sommer, and M. Golz, "Predicting accident probability from frequency of microsleep events," *World Congr. on Med. Phys. and Biomed. Eng.,* vol. 25, pp. 2284-2286, 2010.

[7] M. T. R. Peiris, R. D. Jones, P. R. Davidson, and P. J. Bones, "Detecting behavioral microsleeps from EEG power spectra," in *Conf. Proc. IEEE Eng. Med. Biol. Soc*, New York City, USA, Aug. 2006, pp. 5723-5726.

[8] M. T. R. Peiris, P. R. Davidson, P. J. Bones, and R. D. Jones, "Detection of lapses in responsiveness from the EEG," *J. Neural Eng.,* vol. 8, p. 016003(15pp), Feb 2011.

[9] S. S. D. P. Ayyagari, R. D. Jones, and S. J. Weddell, "Optimized echo state networks with leaky integrator neurons for EEG-based microsleep detection," in *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, Aug. 2015, pp. 3775-3778.

[10] R. Shoorangiz, S. J. Weddell, and R. D. Jones, "Prediction of microsleeps from EEG: Preliminary results," in *Conf. Proc. IEEE Eng. Med. Biol. Soc*, Aug. 2016, pp. 4650-4653.

[11] R. Shoorangiz, S. J. Weddell, and R. D. Jones, "Bayesian multi-subject analysis to predict microsleeps from EEG power spectral features," in *Conf. Proc. IEEE Eng. Med. Biol. Soc*, Jul. 2017, pp. 4183-4186.

[12] A. B. Buriro, S. J. Weddell, and R. D. Jones, "Prediction of microsleeps using pairwise joint entropy and mutual information between EEG channels," in *Conf. Proc. IEEE Eng. Med. Biol. Soc*, Jul. 2017, pp. 4495-4498.

[13] K. M. Ting and I. H. Witten, "Issues in stacked generalization," *J. Artif. Intell. Res.,* vol. 10, pp. 271-289, 1999.

[14] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?," *Machine Learning,* vol. 54, pp. 255-273, Mar. 2004.

[15] A. Rahman and B. Verma, "A novel ensemble classifier approach using weak classifier learning on overlapping clusters," in *Conf. Proc. Int. Jt. Conf. Neural Netw.*, Jul. 2010, pp. 1-7.

[16] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, "Nearest neighbor estimates of entropy," *Amer. J. Math. Management Sci.,* vol. 23, pp. 301-321, Feb. 2003.

[17] S. Zoltan, "Information theoretical estimators toolbox," *J. Mach. Learn. Res.,* vol. 15, pp. 283-287, Jan. 2014.