Optimal EEG feature selection from average distance between events and non-events

John LaRocco, Carrie R. H. Innes, Philip J. Bones, *Senior Member, IEEE*,
Stephen Weddell, *Member, IEEE*, Richard D. Jones, *Senior Member, IEEE*

Abstract— **Biosignal classification systems often have to deal with extraneous features, highly imbalanced datasets, and a low SNR. A robust feature selection/reduction method is a crucial step in this process. Sets of artificial data were generated to test a prototype EEG-based microsleep detection system, consisting of a combination of EEG and 2-s bursts of 15-Hz sinusoids of varied signal-to-noise ratios (SNRs) ranging from 16 to 0.03. The balance between events and non-events was varied between evenly balanced and highly imbalanced (e.g., events occurring only 2% of the time). Features were spectral estimates of various EEG bands (e.g., alpha band power) or ratios between them. A total of 34 features for each of the 16 channels yielded a total of 544 features. Five minutes of EEG from a total of eight subjects were used in the generation of the artificial data. Several feature reduction and classifier structures were investigated. Taking only a single feature corresponding to the maximum of average distance between events and non-events (ADEN) on unbalanced data yielded a phi correlation of 0.94 on the mock data with an SNR of 0.3, compared with a phi coefficient of 0.00 for principal component analysis (PCA). ADEN consistently outperformed alternative system configurations, independent of the classifier utilized. While ADEN's high performance may be due to the nature of the artificial dataset, this simulation has demonstrated strong potential compared to other feature selection/reduction methods.**

I. Introduction

Microsleeps are complete breaks in responsiveness for 0.5-15 s. They can lead to multiple fatalities in certain occupational fields (e.g., transportation and military) due to their need for extended continuous vigilance. Therefore, an automatic microsleep detection system may assist in the reduction of poor performance and occupational fatalities. An EEG-based microsleep detector offers advantages over a video-based detector, such as speed and temporal resolution. This paper represents an investigation on the performance of feature selection and classification algorithms upon a simulated dataset.

This research represents a progression of prior work on EEG-based lapse detection [1, 2]. A set of software modules was developed to expand system combinations and permutations beyond earlier work. Before EEG data from prior studies were investigated [2], it was considered essential to validate and optimize the software modules.

The occurrence of microsleeps is usually rare relative to non-events, forming highly imbalanced data. An artificial "gold standard" dataset was implemented, with artificial events superimposed on real EEG data. Different parameters of the artificial dataset were varied, such as the ratio of events to non-events and the signal to noise ratio. The purpose of this testing was to confirm that the system works correctly on a dataset of precisely known events and to determine how signal power and class balances affect performance.

The complete microsleep detection system involves preprocessing, feature extraction, feature selection/reduction, and classification steps, as shown in Fig. 1. The preprocessing step includes signal acquisition, filtering, and artifact removal. The feature extraction step takes the processed EEG data and returns a set of features based upon an algorithmic process, such as spectral power estimates. More than one set of features can result from one set of data, forming a matrix of different types of feature sets. The number of features is reduced/selected in various ways, such as PCA, so as to minimize and optimize the number of features given to the classifier without losing key information in the feature set. A set of fewer features reduces the computational complexity and improves the system response times. The final step is pattern recognition. Based upon prior training, each set of features is assigned a category determined by the classification algorithm. In these respects, the system is similar to a brain-computer interface (BCI).

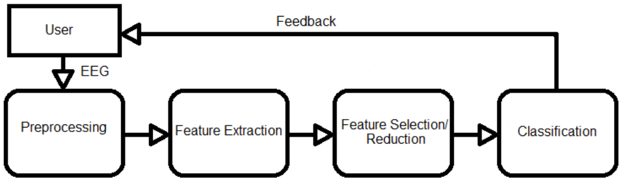

Fig. 1. Microsleep detection system.

The first step was the implementation of a system to train,

This work was supported by the Canterbury Medical Research Foundation and University of Canterbury (UC).

John LaRocco is with the New Zealand Brain Research Institute (NZBRI), and UC Department of Electrical and Computer Engineering (ECE), Christchurch, New Zealand (e-mail: john.larocco@nzbri.org).

Carrie Innes is with the NZBRI, Christchurch Hospital Department of Medical Physics and Bioengineering (MPBE), UC ECE Department, Christchurch, New Zealand (e-mail: carrie.innes@nzbri.org).

Philip Bones is with the UC ECE Department, Christchurch, New Zealand (e-mail: phil.bones@canterbury.ac.nz).

Stephen Weddell is with the UC ECE Department, Christchurch, New Zealand (e-mail: steve.weddell@canterbury.ac.nz).

Richard Jones is with the NZBRI, Christchurch Hospital MPBE Department, and UC Departments of ECE and Psychology, Christchurch, New Zealand (e-mail: richard.jones@nzbri.org).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

test, and validate the performance of a module. Leave-one-out analysis was used to measure the accuracy of each system configuration, by reserving the data from one subject for testing and training on the others. Each configuration comprised a different arrangement of feature extraction, feature selection, and classification techniques and was trained on different subsets of data. For this, an artificial dataset was needed, as the preprocessing and feature extraction modules would be effectively bypassed.

This paper documents an evaluation of feature selection/reduction methods, including a novel approach, based upon classifier performance on the artificial gold-standard dataset.

II. Methods

A. Artificial Dataset

The artificial data was generated to loosely approximate the microsleep detection task. The advantage of using an artificial dataset was the ability to exactly control the parameters of the event to be detected. The event for the artificial dataset was a 15 Hz sine wave, lasting for 2 s. Five minutes of 16-channel EEG data were taken from 8 subjects, and further subdivided into 2-s segments. A total of 6 segments had the sine wave added to all channels, resulting in 2% of the time being events and 98% non-events. A total of 34 EEG band-derived spectral features were then taken from each segment for each channel, resulting in 544 features for 300 segments for each subject.

The sine-wave amplitude was adjusted relative to the EEG to meet a signal-to-noise amplitude ratio (SNR) of 16, 3, 1, 0.3, and 0.03. The first four of these are shown in Fig. 2. The prevalence of non-events could bias the classifier, so balanced datasets were required for comparison. Five other artificial datasets were generated, identical to the previously described ones, except with equal numbers of events and non-events. Class balance was achieved by repeating events and randomly deleting a random subset of non-events until the ratio of events to non-events was balanced. If trends were present in both balanced and unbalanced data, evidence could potentially be stronger.

B. Evaluation Criteria

Software toolset validation utilized performance metrics from prior research [1, 2]. The performances of each classifier were averaged together after leave-one-out cross-validation. The performance metrics were: mean accuracy, sensitivity, specificity, selectivity, and phi correlation. The phi correlation coefficient was the primary single measure of classifier effectiveness, due to being independent of class distributions and being the best integrated measure of the other performance metrics. A phi of 1.00 indicates perfect performance in successfully identifying all events and non-events.

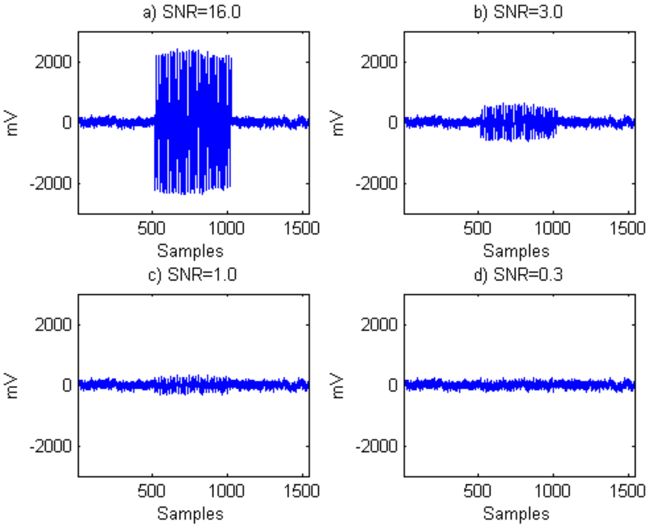

Fig. 2. A combination of EEG and Events: a) “Very Easy” (SNR=16.0), (b) “Easy” (SNR=3.0), (c) “Medium” (SNR=1.0), and (d) “Hard” (SNR=0.3).

C. Feature Selection

Feature selection and reduction methods can be separated into supervised and unsupervised methods. Supervised methods, such as common spatial patterns (CSP) [3], use *a priori* knowledge of classes in the training data to either select a subset of features or generate meta-features. An unsupervised method, such as PCA, does not use knowledge of class labels to operate. In the benchmark established by prior research [1, 2], PCA was used as the feature reduction method. It is considered that a supervised method of feature reduction with data normalization performs better than unsupervised and supervised methods lacking data normalization [4].

In addition to PCA, a simple supervised selection/reduction solution was explored by comparing the normalized differences between averaged features of each class. A subset of features corresponding to the largest average differences between events and non-events (ADEN) was retained, and all other features in the training and testing data were discarded. ADEN required the user to define U features to retain. The training data $\mathbf{X}$ (of F features and M observations) were normalized via z-score transform, and then features corresponding to events and nonevents were separated into $\mathbf{X}_e$ and $\mathbf{X}_n$. Each was averaged to form mean feature vectors (F long), $\bar{\mathbf{x}}_e$ and $\bar{\mathbf{x}}_n$. The difference formed a single vector, $\Delta\mathbf{x}_f$.

$$\Delta\mathbf{x}_f = \text{abs}(\bar{\mathbf{x}}_{e,f} - \bar{\mathbf{x}}_{n,f}). \quad (1)$$

Training data $\mathbf{X}$ were reduced to a matrix of U features and M observations, with all remaining features based on the indices f of the U terms in $\Delta\mathbf{x}_f$. The testing data would likewise be reduced to u features, selected from the f indices corresponding to features in the training data.

While ADEN might select collinear features, it is considered that this can potentially achieve a greater measure of robustness. Originally, only the feature corresponding to the maximal averaged-distance (ADEN$_1$) in the training data was retained. However, multiple ADEN features were then investigated, specifically features with the 10 highest averaged distances ($ADEN_{10}$).

D. Configurations

The feature reduction modules tested were PCA, $ADEN_1$, and $ADEN_{10}$. A single classifier was used but the pattern recognition algorithms explored were: linear discriminant analysis (LDA), radial basis functions (RBF), support vector machines (SVM) with a Gaussian kernel (SVMG), and SVM with a polynomial kernel (SVMP). Configurations covering each of the feature reduction and pattern recognition modules were incorporated.

III. Results

Classification performance by averaged 8-fold cross-validation of the unbalanced data (SNR=0.3), via LDA and 3 feature selection/reduction modules are presented in Fig. 3. For both balanced and unbalanced datasets, ADEN performed substantially higher than PCA. In particular, configurations incorporating ADEN were the only ones able to classify the hard dataset (SNR=0.3): $ADEN_1$ yielded a phi correlation of 0.94 compared with 0.00 for PCA. On LDA, $ADEN_{10}$ features achieved a marginally higher phi of 0.96 on the hard unbalanced data over 0.94 for $ADEN_1$. However, no configuration was able to correctly classify the balanced or unbalanced "very hard" datasets (SNR=0.03).

Results from the balanced and unbalanced "hard" datasets (SNR=0.3) for the 4 pattern recognition algorithms are presented in Table 1.

Performance metrics for $ADEN_1$ were high across all pattern recognition modules for the unbalanced "hard" dataset, with only marginal differences between LDA, RBF, and both SVM kernels. However, performance metrics varied greatly across datasets. As expected, classification performance reduced as the signal became weaker relative to the background noise.

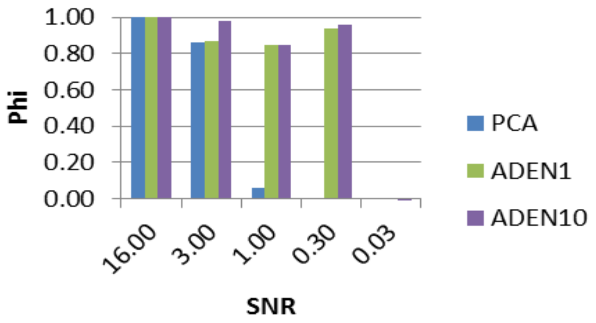

Fig. 3. Classification performance for different feature selection /reduction methods with LDA on unbalanced data (SNR=0.3).

Table 1. Classifier Performance for Different Pattern Recognition Algorithms on the Hard Data (SNR=0.3).

a) $ADEN_1$ (Balanced data)

	LDA	***RBF***	***SVMG***	***SVMP***
Sensitivity	0.90	0.94	0.88	0.88
Specificity	1.00	0.99	1.00	0.99
Selectivity	0.99	0.99	0.87	0.87
Phi	0.91	0.94	0.86	0.86

b) $ADEN_1$ (Unbalanced data)

	LDA	***RBF***	***SVMG***	***SVMP***
Sensitivity	0.92	0.90	0.94	0.90
Specificity	1.00	1.00	1.00	1.00
Selectivity	0.97	0.93	0.97	0.97
Phi	0.94	0.90	0.95	0.93

IV. Discussion

In summary, average distance feature selection modules provided higher performance scores, independent of class balance or pattern recognition module. In contrast with $ADEN_1$, PCA substantially dropped in performance when faced with the "hard" dataset (SNR=0.3). However, neither $ADEN_1$ nor $ADEN_{10}$ features were able to classify the "very hard" dataset (SNR=0.03), irrespective of whether it was balanced or unbalanced.

$ADEN_{10}$ resulted in the highest phi coefficient, 0.96, out of tests on the unbalanced datasets. On the balanced data (SNR=0.3), phi was similarly high at 0.90. The similar performance range for ADEN suggested independence from class distribution.

As the simulated events decreased in amplitude, they became harder to discern from the background EEG. However, perplexingly, a small *increase* in performance was seen in the "hard" dataset (SNR=0.3) relative to the "medium" dataset (SNR=1.0) for both $ADEN_1$ and $ADEN_{10}$. We were unable to determine the reason for this. Neither $ADEN_1$ nor $ADEN_{10}$ could correctly classify the "very hard" artificial data (SNR=0.03) in either the balanced or unbalanced case. The amplitude of the event appears to have dropped to a point where it cannot be distinguished from the background EEG.

All four pattern recognition modules – LDA, RBF, SVMG, SVMP – performed similarly, indicating that classification performance is strongly dependent upon feature reduction rather than type of classifier or class balance.

Performance for PCA dropped completely on the "hard" data (SNR=0.3), whereas ADEN did not. This strongly suggests that unsupervised generation of meta-features, as opposed to the supervised selection of a subset of existing features, may effectively lose or 'hide' useful information. The maximum ADEN – i.e., $ADEN_1$ – allows the detection of the single feature and electrode location corresponding to the events with the most consistent absolute distance between two classes. When applied to larger and more complex datasets, a greater number of ADEN features could be selected. ADEN consistently outperformed alternative system configurations, independent of the classifier structure utilized.

While ADEN's high performance may be in part due to the nature of the artificial dataset, ADEN clearly has advantages over the prior benchmark of PCA. It is likely that multiple ADEN features would contain redundant information but this may be advantageous. It may combine signals corresponding to the same event across multiple channels, increasing robustness and probability of successful detection. The eigenvalues found by PCA are combinations of multiple features, many of which are likely to be noise. Average distance feature selection methods may prove more suitable for the microsleep detection task than unsupervised feature reduction methods.

V. Conclusion

This study has demonstrated that despite success on the "hard" (0.3) dataset with ADEN, other methods could be applied to the system. Bandpass filtering the signal and rejecting ocular artifacts may improve performance but the microsleep identification task is likely to be substantially more difficult than the 0.3 dataset. As such, future research will be directed at examining the feature extraction process and potentially implementing new feature reduction techniques, such as combining ADEN with genetic algorithms [5, 6]. Given its potential to select optimal spectral and spatial features, ADEN is likely to be of value in other biosignal classification application systems, such as BCIs.

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, 2007.

[2] M. T. Peiris, P. R. Davidson, P. J. Bones, and R. D. Jones, "Detection of lapses in responsiveness from the EEG," *J. Neural Eng.,* vol. 8 (016003), pp. 1-15, 2011.

[3] K. Yin, J. Wu, and J.-C. Zhang, "A framework of common spatial patterns based on support vector decomposition machine," in *Proc. Int. Conf. on Machine Learning and Cybernetics,* vol. 6, pp. 3434-3438, 2008.

[4] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized common spatial patterns with generic learning for EEG signal classification," in *Proc. Ann. Int. Conf. of the IEEE Eng. Med. Biol. Soc.,* vol. 31, pp. 6599-6602, 2009.

[5] S. Parini, L. Maggi, and G. Andreoni, "An automated method for relevant frequency bands identification based on genetic algorithms and dedicated to the motor imagery BCI protocol," in *Proc. Ann. Int. Conf. of the IEEE Eng. Med. Biol. Soc.,* vol. 29, pp. 2512-2515, 2007.

[6] L. Wang, G. Xu, J. Wang, S. Yang, and W. Yan, "Motor imagery BCI research based on Hilbert-Huang Transform and Genetic Algorithm," in *Proc. Int. Conf. Bioinf. Biomed. Eng.,* vol. 5, pp. 1-4, 2011.