ESTIMATING COGNITIVE STATE USING EEG SIGNALS

Tian Lan, Andre Adami, Deniz Erdogmus, Misha Pavel

Biomedical Engineering Department, Oregon Health & Science University
20000 NW Walker Rd. Beaverton, OR 97006, USA

ABSTRACT

Using EEG signals to estimate cognitive state has drawn increasing attention in recently years, especially in the context of brain-computer interface (BCI) design. However, this goal is extremely difficult because, in addition to the complex relationships between the cognitive state and EEG signals that yields the non-stationarity of the features extracted from EEG signals, there are artefacts introduced by eye blinks and head and body motion. In this paper, we present a classification system, which can estimate the subject's cognitive state from the measured EEG signals. In the proposed system, a mutual information based method is employed to reduce the dimensionality of the features as well as to increase the robustness of the system. A committee of three classifiers was implemented and the majority voting results of the committee are taken to be the final decisions. The results of a preliminary test with data from freely moving subjects performing various tasks as opposed to the strictly controlled experimental set-ups of BCI provide strong support for this approach.

1. INTRODUCTION

The Electroencephalogram (EEG) is a recording of the electrical potentials on the scalp, revealing the electrical activity of the brain tissue. Based on the evidence that EEG appears to reflect aspects of cognitive processes and may differentiate among mental activities and cognitive loads, much research effort has focused on exploiting the information content of this signal for understanding the functioning of the brain as well as for clinical diagnostics. In particular, Brain Computer Interface (BCI) for communication and control is a typical application, in which a person interacts with the computer directly without utilizing physical manners. In this paper, we focus on an alternative application: augmented cognition. Here, the goal is to enhance the task-related performance of a human user through computer assistance based on the assessments of the user's cognitive level.

EEG has many advantages in measuring brain activity including the convenience and low cost. However, it is very difficult to estimate cognitive or mental state from EEG signals for a number of reasons. Specifically, EEG signals (1) contain noise as a result of the movement of the electrode on the scalp; (2) are contaminated with eye blinks or other muscular activities; (3) are not stationary. These difficulties are amplified in the augmented cognition application, because the subjects are freely moving around rather than bring in constrained environment as in a strictly controlled typical BCI experimental setup.

In this paper, we will present a classification system, based on state-of-the-art signal processing and machine learning approaches, to solve the cognitive state estimation problem in augmented cognition applications. The proposed method is also applicable to BCI problems.

2. METHOD

The cognitive state estimation system we propose contains four parts: preprocessing, feature extraction/selection, classification, and postprocessing. Preprocessing is used to filter out noise and remove the artefacts. Feature extraction and selection generates features from the clean EEG signals and selects useful features. For classification, a committee of 3 classifiers are employed and the majority voting of the committee is adopted as the final decision. Postprocessing exploits the prior knowledge to improve the classification results. The schematic diagram of the proposed system is shown in Fig. 1.

2.1 Preprocessing

The experimental setting involves a user outfitted with wearable monitoring, communication, and mobile computing equipment walking outside. The monitoring equipment is a BioSemi ActiveTwo EEG system with 32 electrodes (http://www.biosemi.com/). Vertical and horizontal eye movements and blinks were recorded with electrodes below and lateral to the left eye. Although there are several other sensors, our effort focuses on extracting information and estimating cognitive state from EEG signals.

EEG is sampled and recorded at 256Hz from 7 channels (CZ, P3, P4, PZ, O2, P04, F7) while the subject is moving around and performing various prescribed tasks (such as visual search and communicating on the radio). These sites are selected based on a saliency analysis of EEG collected from various subjects performing cognitive test battery tasks [1]. EEG signals are preprocessed to remove eye blinks using adaptive filters [2]. Information from the VEOGLB ocular reference channel was used as the noise reference source for the adaptive ocular filter. DC drifts were removed using high-pass filters (0.5Hz cut-off). A band-pass filter (between 2Hz and 50Hz) was also employed, as this interval is generally associated with cognitive activity.

2.2 Feature extraction and selection

Feature extraction is a process focused on discovering a pattern that can differentiate the various classes, while feature

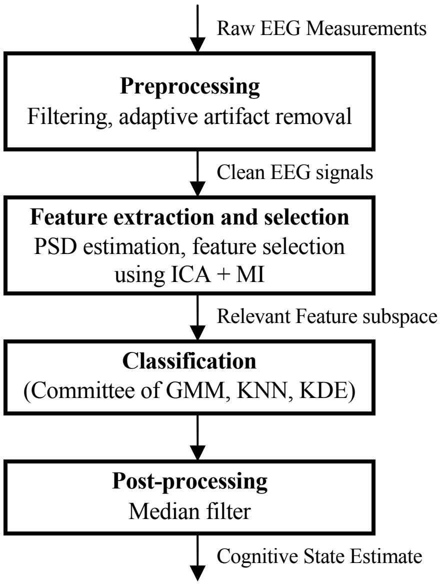
Raw EEG Measurements
Preprocessing
Filtering, adaptive artifact removal
Clean EEG signals
Feature extraction and selection
PSD estimation, feature selection using ICA + MI
Relevant Feature subspace
Classification
(Committee of GMM, KNN, KDE)
Post-processing
Median filter
Cognitive State Estimate

Fig. 1. Schematic diagram of the cognitive state estimation

selection is to find reduced dimensionality optimal feature vectors to keep the useful information and eliminate irrelevant information, in order to reduce the computational load and increase the robustness of the classification system.

Feature Extraction: Usually, two approaches are used to extract the features from EEG signals. The first approach is based on the characteristic P300 signal that appears in the EEG approximately 300ms following the occurrence of an event, referred to as event-related potentials (ERP). Since this approach relies on a device that interacts between a subject and the stimuli corresponding to the event, it is impractical in our augmented cognition setting. Hence, we employ the second approach: spatial and spectrum analysis. The power spectral density of signals from the seven different sites is used as features in this application.

The PSD of the clean EEG signals, estimated using the Welch method [3] with 50%-overlapping 1-second windows, is integrated over 5 frequency bands: 4-8Hz (theta), 8-12Hz (alpha), 12-16Hz (low beta), 16-30Hz (high beta), and 30-44Hz (gamma). These bands, sampled every 0.1 seconds, are used as the basic features for the classification. The particular selection of the frequency bands is based on well-established interpretations of EEG signals in existing cognitive and clinical research literature [4].

Feature Selection: These PSD features constitute a high dimensional vector (5×7=35 in our application) that contains information pertinent to the classification of cognitive states, as well as irrelevant components and noise. Direct classification using such input features is undesirable, since the unwanted components have an adverse effect on the overall classification performance and the generalization ability of the system. Consequently, an intelligent and practical technique for extracting the relevant information from these features is necessary.

Feature selection and dimensionality reduction has been shown to be an effective way to improve robustness and has been an active field of research in pattern recognition. This can be achieved by feature transformation method. The transformation generates either a new feature space, or a subset of the original, which can be treated as a special case of the former situation. This transformation can be linear or nonlinear. Linear transformations have been widely used due to their simplicity. While nonlinear transformations attract increasingly more attention, usually, linear projections are preferred provided that they yield satisfactory results.

There are many existing linear transformation methods for dimensionality reduction. Principle component analysis (PCA) is a widely used dimensionality reduction technique [5,6]. However, the projections it finds are not necessarily related to the class labels, therefore, it is not particularly useful in pattern recognition. Linear discriminant analysis (LDA) attempts to eliminate this shortcoming of PCA by finding linear projections that maximize class separability under the Gaussian distribution assumption [7]. The LDA projections are optimized based on the means and the covariance matrices of classes, which are not descriptive of an arbitrary probability density function (pdf). Independent component analysis (ICA) has also been used as a tool to find linear transformations that maximize the statistical independence of random variables [8,9]. However, like PCA, the projection that ICA finds also has no necessary relationship with class labels, and it is not able to enhance class separability [10].

Optimal feature selection coupled with a specific classifier topology, namely the *wrapper* approach, results in a combinatorial computational requirement; thus, is unsuitable for adaptive learning of feature projections. On the contrary, the filter approach, which selects features by optimizing some criterion is independent of classifier, hence is more flexible. Since we will employ a committee of classifiers, the filter approach is more suitable for this application.

In the filter approach, it is important to optimize a criterion that is relevant to Bayes risk, which is typically measured by the probability of error. A suitable criterion is mutual information (MI) between the projected features and the class labels (defined in (1)), which is motivated by lower and upper bounds in information theory that relate this quantity to probability of error [11,12]. Several MI based methods has been developed for feature selection [13-17]. However, since features are generally mutually dependent, feature selection in this manner is typically suboptimal in the sense of maximum joint mutual information principle.

$$I_S(\mathbf{y};c) = H_S(\mathbf{y}) - \sum_c p_c H_S(\mathbf{y} \mid c) \qquad (1)$$

If the components of the random vector $\mathbf{y}$ in (1) are independent, the joint and joint-conditional entropy becomes the sum of marginal and marginal-conditional entropies. Thus, the joint mutual information of a feature vector with the class labels is equal to the sum of marginal mutual information of each individual feature with the class labels

$$I_S(\mathbf{y};c) = \sum_{i=1}^{n} I_S(y_i;c) \qquad (2)$$

where $I_S(y_i;c) = H_S(y_i) - \sum_c p_c H_S(y_i \mid c)$, and y_i is the i^{th} component of $\mathbf{y}$. In feature selection, we exploit this fact by combining independent component analysis (ICA) transformation with a sample-spacing based entropy estimator [18].

Many effective and efficient algorithms based on a variety of assumptions including maximization of non-Gaussianity, minimization of mutual information, nonstationarity of the sources, etc., exist to solve the ICA problem [18-20]. All these could be compactly formulated in the form of a generalized eigen-decomposition problem that gives the ICA solution in an analytical form [21]. Therefore, this formulation will be employed in this paper.

According to this formulation, one possible assumption set that leads to an ICA solution utilizes the higher order statistics (specifically fourth-order cumulants). Under this set of assumptions, the separation matrix **W** is the solution to the following generalized eigen-decomposition problem:

$$\mathbf{R_x W} = \mathbf{Q_x W \Lambda} \tag{3}$$

where $\mathbf{R_x}$ is the covariance matrix and $\mathbf{Q_x}$ is the cumulant matrix estimated using sample averages: $\mathbf{Q_x}$=E[$\mathbf{x}^\mathbf{T}\mathbf{x}\mathbf{x}\mathbf{x}^\mathbf{T}$]-$\mathbf{R_x}$tr($\mathbf{R_x}$)-E[$\mathbf{x}\mathbf{x}^\mathbf{T}$]E[$\mathbf{x}\mathbf{x}^\mathbf{T}$]-$\mathbf{R_x}\mathbf{R_x}$. Given the estimates for these matrices, the ICA solution can be easily determined using efficient generalized eigen-decomposition algorithms (or using the *eig* command in Matlab).

There exist many entropy estimators in the literature for single-dimensional variables. Here, we use an estimator based on sample-spacing, which stems from order statistics. This estimator is selected because of its consistency, rapid asymptotic convergence, and simplicity. Consider a random variable *Y*. Given a set of iid samples of *Y* $\{y_1,\ldots,y_N\}$, first these samples are sorted in increasing order such that $y_{(1)}\leq\ldots\leq y_{(N)}$. The *m*-spacing entropy estimator is given by:

$$\hat{H}(Y) = \frac{1}{N-m}\sum_{i=1}^{N-m}\log\frac{(N+1)(y_{(i+m)}-y_{(i)})}{m} \tag{4}$$

The selection of the parameter *m* is determined by a bias-variance trade-off and typically, $m=\sqrt{N}$.

2.3 Classification

Since we do not know the distribution of the projected features of EEG signals, the comparison and combination of both parametric and nonparametric classifiers is desirable in classification process. The committee consists of three classifiers denoted by GMM, KNN, and KDE. The selected optimal feature vectors are used as inputs to the committee of classifiers. The GMM classifier implements a parametric Bayes classifier [22] assuming that each class distribution can be described by a GMM with 4 Gaussian components that is fit to the data from each class using the Expectation-Maximization algorithm [23]. The KNN classifier decides based on the votes from $3C+1$ neighbours, where *C* is the number of classes and each vote is weighted inversely proportional to the class prior p_c of the contributing neighbours. It is well known that the KNN classifier asymptotically approaches the optimal Bayes classification error rate [22]. The KDE classifier implements a nonparametric Bayes classifier assuming that the distribution of each class is given by a kernel density estimate [24] (using Gaussian kernels whose bandwidth parameters are selected according to Silverman's rule-of-thumb [25]).

In our experiments, we find that KDE classifier has better performance than the other two classifiers in most cases. Therefore, the committee decision strategy is that majority vote is preferred in general and in the case of no majority agreement, the KDE decision is adopted as final decision. In real-time application, a committee decision is offered at a rate of 10Hz.

2.4 Postprocessing

Postprocessing exploits prior context knowledge to improve the classification performance. Assuming that the mental state does not fluctuate within a given 2-second interval, it is possible to improve classification performance by a median filter that smoothens the decisions offered by the committee. The application of a median filter also introduces the inherent assumption that the integer class labels are assigned to cognitive tasks (the classes) in correlation with their actual corresponding cognitive loads. This postprocessing step increases performance significantly.

3. RESULTS

In order to illustrate the performance of the proposed cognitive state estimation system, we present results from an experiment where the subject is required to execute four predetermined tasks: *slow walking*, *navigating and counting, communicating with radio*, and *studying mission map*. During tasks, EEG signals are recorded when the subject is performing one of the four tasks listed above. Each task is assigned a class label from number 1 to 4 respectively. After preprocessing and PSD estimation mentioned in the second section, approximately 6000 samples are obtained, each with 35 dimensional inputs and a desired class label. A randomly selected one third of these samples are used as training set for feature selection and classification, and the remaining two thirds samples are used as testing set. The feature selection is achieved by the proposed ICA transformation and MI-sorting algorithm. The correct classification rates for different dimensionality of optimally selected features are evaluated using the classifier committee over 10 Monte Carlo runs. The results are shown in Fig. 2, from which we can see that an accuracy of 80% is achieved with 12 dimensions, while the remaining 23 dimensions do not significantly contribute to the classification accuracy.

The classification results based on 10, 12, and 35 dimensional optimally selected features are compared in Table 1 via the confusion matrix of the classification results; the *ij*th entry shows P(decide class j | true class i). The classification results illustrated here shows that this feature selection method is able to capture the low-dimensional relevant components in the original feature space. In some experiments, the correct classification rate reaches the best performance when using 5 to 10 optimal features. After that, the classification rate decreases when the number of features increases. This indicates that the high dimensionality feature also introduces irrelevant and confusing information, which may impair the classification accuracy; hence mutual information based feature projections improve classifier robustness. We also compare the proposed feature selected method with Mermaid-SIG algorithm [26]. The classification results show that the classification performances are similar. However, the

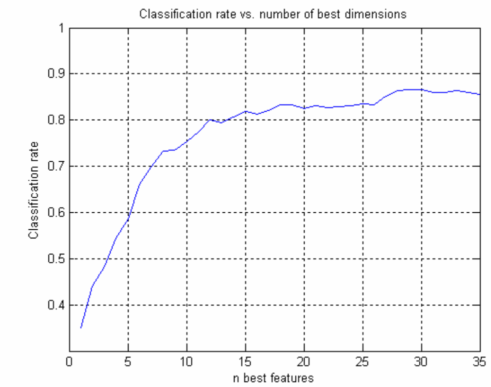

Figure 2. Correct classification rate vs. dimensionality of optimally selected features.

Number of dimension	Confusion Matrix
10-dim input	$\begin{bmatrix} 0.38 & 0.33 & 0.25 & 0.04 \\ 0.03 & 0.82 & 0.15 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0.01 & 0.24 & 0.75 \end{bmatrix}$
12-dim input	$\begin{bmatrix} 0.6 & 0.22 & 0.17 & 0.01 \\ 0.01 & 0.91 & 0.08 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0.18 & 0.82 \end{bmatrix}$
35-dim input	$\begin{bmatrix} 0.6 & 0.29 & 0.1 & 0.01 \\ 0.02 & 0.83 & 0.15 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0.02 & 0.98 \end{bmatrix}$

Table 1. Confusion matrix for classifiers on 4 cognitive states using 10, 12, and 35 dimensional input feature vectors.

ICA transformation in combination with MI sorting algorithm is much faster and computationally efficient, which is critical in real-time applications. In another experiment where the stationary and mobile cases are considered separately with 3 tasks in each case, the correct classification rate reaches more than 95% based on 3 dimensional optimal features.

4. DISCUSSION

In this paper, we present a classification system, which robustly estimates the cognitive state using power spectrum density of EEG signals. Experimental results suggest that our system can achieve good performance in classifying 4 different actions of an ambulatory subject. The system can work online after training.

Although the preliminary results are satisfactory, the nonstationarity of EEG signals presents a future challenge in finding robust features that will allow cross-session and cross-subject generalization. Our future research will focus on finding stationary representations of the features corresponding to the different cognitive states. In addition, we plan to combine the EEG-based analysis with additional sources of information including signals representing the movements and pose of the user as wells as those sensing aspects of the environment such ambient light, auditory noise and temperature.

. **Acknowledgments:** This work was supported by DARPA under contract DAAD-16-03-C-0054. The data was collected at the Human-Centered Systems Lab., Honeywell.

REFERENCES

[1] C.A. Russell, S.G. Gustafson, “Selecting Salient Features of Psychophysiological Measures” , Air Force Research Laboratory Technical Report (AFRL-HE-WP-TR-2001-0136), 2001.

[2] P. He, G. Wilson, C. Russell, “Removal of Ocular Artifacts from Electroencephalogram by Adaptive Filtering”, Medical & Biologial Engineering & Computing, vol. 42, pp. 407-412, 2004.

[3] P. Welch, “The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short Modified Periodograms”, IEEE Transactions on Audio and Electroacoustics, vol. 15, no. 2, pp. 70-73, 1967.

[4] A. Gevins, M.E. Smith, L.McEvoy, D. Yu, “High Resolution EEG Mapping of Cortical Activation Related to Working Memory: Effects of Task Difficulty, Type of Processing, and Practice”, Cerebral Cortex, vol. 7, pp. 374-385, 1997.

[5] E. Oja, *Subspace Methods of Pattern Recognition*, Wiley, 1983.

[6] P.A. Devijver, J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice Hall, London, 1982.

[7] K. Fukunaga. *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, New York, 1990.

[8] R. Everson, S. Roberts, “Independent Component Analysis: A Flexible Nonlinearity and Decorrelating Manifold Approach”, *Neural Computation*, vol. 11, no. 8, pp. 1957-1983, 2003.

[9] A. Hyvärinen, E. Oja, P. Hoyer, J. Hurri, “Image Feature Extraction by Sparse Coding and Independent Component Analysis”, Proceedings of ICPR'98, pp. 1268-1273, 1998.

[10] K. Torkkola, “Feature Extraction by Non-Parametric Mutual Information Maximization”, Journal of Machine Learning Research, vol. 3, pp. 1415-1438, 2003.

[11] R. M. Fano, *Transmission of Information: A Statistical Theory of Communications*. Wiley, New York, 1961.

[12] M.E. Hellman, J. Raviv, “Probability of Error Equivocation and the Chernoff Bound”. IEEE Transactions on Information Theory, vol. 16, pp. 368-372, 1970.

[13] R. Battiti, “Using Mutual Information for Selecting Features in Supervised Neural Networks learning”, IEEE Transactions on Neural Networks, vol. 5, no. 4, pp. 537-550, 1994.

[14] A. Ai-ani, M. Deriche, “An Optimal Feature Selection Technique Using the Concept of Mutual Information”, Proc. of ISSPA, pp. 477-480, 2001.

[15] N. Kwak, C-H. Choi, “Input Feature Selection for Classification Problems”, IEEE Trans. Neural Networks, vol, 13, no. 1, pp. 143-159, 2002.

[16] H.H. Yang, J. Moody, “Feature Selection Based on Joint Mutual Information”, in Advances in Intelligent Data Analysis and Computational Intelligent Methods and Applications, Rochester, New York, 1999.

[17] H.H. Yang, J. Moody, “Data Visualization and Feature Selection: New Algorithms for Nongaussian Data”, Proc. NIPS’00, pp. 687-693, 2000.

[18] E.G. Learned-Miller, J.W. Fisher III, “ICA Using Spacings Estimates of Entropy”, J. Machine Learning Research, vol. 4, pp.1271-1295, 2003.

[19] K.E. Hild II, D. Erdogmus, J.C. Principe, “Blind Source Separation Using Renyi's Mutual Information”, IEEE Signal Processing Letters, vol. 8, no. 6, pp. 174-176, 2001.

[20] A. Hyvärinen, E. Oja, “A Fast Fixed Point Algorithm for Independent Component Analysis”, Neural Computation, vol. 9, no. 7, pp. 1483-1492, 1997.

[21] L. Parra, P. Sajda, “Blind Source Separation via Generalized Eigenvalue Decomposition”, Journal of Machine Learning Research, vol. 4, pp. 1261-1269, 2003.

[22] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, 2nd ed., Wiley, 2000.

[23] A.P. Dempster, N.M. Laird, D.B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” Journal of the Royal Statistical Society, vol. 39, pp. 1-38, 1977.

[24] E. Parzen, “On Estimation of a Probability Density Function and Mode”, in *Time Series Analysis Papers*, Holden-Day, 1967.

[25] B.W. Silverman, “Density Estimation for Statistics and Data Analysis,” Chapman and Hall, London, 1986.

[26] K.E. Hild II, *Blind Source Separation of Convolutive Mixtures Using Renyi’s Divergence*, PhD Dissertation, University of Florida, 2003.