Salient EEG Channel Selection in Brain Computer Interfaces by Mutual Information Maximization

Tian Lan[1], Deniz Erdogmus[1], Andre Adami[1], Misha Pavel[1], Santosh Mathan[2]
[1]BME Department, Oregon Health & Science University, Beaverton, Oregon, USA
[2]Human Centered Systems, Honeywell Laboratories, Minneapolis, Minnesota, USA

***Abstract* – Modern brain computer interface (BCI) applications use information obtained from the user's electroencephalogram (EEG) to estimate the mental states. Selecting an optimal subset of the EEG channels instead of using all of them is especially important for ambulatory EEG where the user is mobile due to reduced data communication and computational load requirements. In addition, elimination of irrelevant sensors improves the robustness of the classification system by reducing dimensionality. In this paper, we propose a filter approach for EEG channel selection using mutual information (MI) maximization. This method ranks the EEG channels, such that the MI between the selected sensors and class labels is maximized. This selection criterion is known to reduce classification error. We employ a computationally efficient approach for MI estimation and EEG channel ranking. This approach is illustrated on EEG data recorded from three subjects performing two mental tasks. Experiment results show that the proposed approach works well and the position of the selected channels using the proposed method is consistent with the expected cortical areas for the mental tasks.**

***Keywords* – EEG channel selection, independent component analysis, mutual information, brain computer interface**

I. Introduction

Brain computer interface (BCI) is a system that can estimate user intent directly from brain activity in real time. The goal of BCIs is to enhance communication between human users and computers, as well as to enable disabled people utilize such systems. This technology requires the ability to accurately and rapidly recognize the mental state from brain signals. Although BCIs based on many measurement modalities are possible, research focus is mainly on the use of noninvasive electroencephalogram (EEG), which has the added advantage of convenience and relatively low cost. However, EEG signals recorded in the laboratory environment are prominently different from the EEG signals recorded in the real world in the following aspects: in the laboratory (1) the experimental setup is controlled facilitating better performances, (2) various precautions to improve signal quality can be implemented, (3) large-scale data collection, analysis, and signal processing hardware and software can be utilized. For practical applications of this technology, these artificial environmental and design restrictions must be relaxed, since light-weight low-power implementations impose size and complexity reduction requirements. In real world applications, EEG signals could be very noisy and contaminated by various motion artifacts. All these factors make it extremely difficult to estimate the cognitive or mental state from recorded EEG signals, hence limit the application of BCIs.

Many methods have been successfully used to reduce the noise in EEG signals, such as adaptive noise/artifact cancellation. However, these methods have limit capabilities for classification-relevant preprocessing. We could partly solve the problem of distractive features and artifacts by eliminating the irrelevant and redundant information in the features, and in turn increase the robustness of the classification system. Feature extraction methods can greatly affect signal-to-noise ratio. Good methods enhance the signal and reduce central nervous system (CNS) and non-CNS noise [1]. There are two aspects for feature extraction: selecting the most salient EEG channels, while eliminating the irrelevant and redundant EEG channels and extracting time/frequency characteristics useful for classification from the selected EEG channels. An important reason for selecting EEG channels is that in some BCI applications, the users are required to execute some tasks while moving, which means a wireless or reduced-size EEG collection, recording, and preprocessing equipment must be employed. The capacity of the wireless channel and the real-time data processing requirement make it impossible to transmit and process all of the EEG channels. Therefore, selecting the most salient EEG channels becomes a critical problem in BCI applications. In this paper, we will focus on the first aspect and introduce a salient EEG channel selection method using MI maximization.

EEG channel selection can be treated as a feature selection problem. However, unlike usual feature selection, it is necessary to treat all features coming from a channel together [2], because each EEG channel may contain more than one feature (e.g., different frequency bands of activity). Many methods have been proposed to solve this problem, such as genetic algorithms and support vector machines [2-4]. The feature selection methods found in the literature can be characterized as wrapper or filter. In the wrapper approach, feature selection is coupled with a specific classifier, resulting in an inflexible design, as well as increased off-line computational burden. On the contrary, the filter approach, which selects features by optimizing some criterion is independent of the classifier, hence is more flexible. In this paper, we take the filter approach and use the MI between features and class labels as the criterion to rank EEG channels in descending order according to their contribution to class discriminability. This criterion is motivated by Fano's and Hellman & Raviv's bounds, which demonstrate that the classification error is bounded from above by this MI [5,6]. The proposed method has the following advantages: (1) it is independent of the following classifier topology; (2) it is computationally effective and efficient since fast ICA and entropy estimation routines are employed; (3) sequential channel evaluation strategy eliminates the combinatorial explosion problem (at the cost of possible sub-optimality in degenerate situations).

The proposed approach is applied to EEG-based BCIs in the context of cognitive augmentation. The experimental results show that this method can greatly improve the classification performance compared to using EEG channels selected based on physiological experience in the BCI literature. Furthermore, the position of the selected channels is highly consistent with the expected cortical areas in the mental tasks.

II. EEG Data Collection

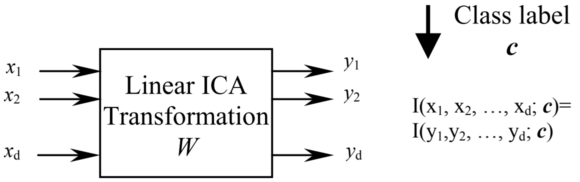
Experimental Setup and Mental Tasks

EEG data is collected while three subjects S_1-S_3 execute two mental tasks, Larson and n-back [7-9]. In the Larson task, the subjects are required to maintain a mental count according to the presented configuration of images on the monitor. The combination of mental activities during this task includes Attention, Encoding, Rehearsal, Retrieval, and Match. The complexity of the task is divided into two classes, low and high workloads, which depend on the inter-stimuli interval. In the n-back task, subjects are required to match the letter in either spatial location or verbal identity in the previous trials. The easy task only requires matching any of the previous trials, involving the combination of mental activities include Attention, and Match, which is defined as low workload. The difficult task requires matching the third previous trials, and involves a complex combination of metal activities that includes Attention, Encoding, Rehearsal, Retrieval, and Match, which is defined as high workload.

EEG data is collected using a BioSemi Active Two system (http://www.biosemi.com) using a 30 channel EEG cap and eye electrodes. Vertical and horizontal eye movements and blinks were recorded with electrodes below and lateral to the left eye. EEG is sampled and recorded at 256Hz from 30 channels.

Data processing

EEG signals are pre-processed to remove eye blinks using an adaptive linear filter based on the Widrow-Hoff training rule (LMS) [10]. 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. The power spectral density (PSD) of the EEG signals, estimated using the Welch method [11] 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), 30-44Hz (gamma). These bands, sampled every 0.25 seconds, are used as the basic features for cognitive classification. The particular selection of the frequency bands is based on well-established interpretations of EEG signals in prior cognitive and clinical contexts [12].



Class label
c
x_1
x_2
x_d
Linear ICA Transformation W
y_1
y_2
y_d
$I(x_1, x_2, \ldots, x_d; c) = I(y_1, y_2, \ldots, y_d; c)$

Fig. 1 Linear ICA transformation and mutual information estimation.

III. Method

Mutual Information Estimation

Our goal is to find an optimal subset of EEG channels that minimizes the classification error, which is *equivalent* to maximizing the MI between input features and class labels:

$$\max_{\mathbf{x}} I(x_{i1},\ldots,x_{id};c) \tag{1}$$

Where $\mathbf{x}$ is the feature vector, c is the class label, and d is the number of EEG channels retained. However, since features are generally mutually dependent, it is very difficult to estimate MI directly from the joint entropy of higher dimensional random variables. Furthermore, the pdf estimation of n dimensional random vector typically requires a large number of data samples (exponential in the dimensionality). This makes it impractical to estimate the joint pdf of higher dimensional random vectors.

An intuitive and efficient method to solve this problem is to use an indirect way: linear ICA transformation plus one-dimensional entropy estimation. The block diagram of this method is shown in Figure 1. This method exploits the two facts: 1) linear ICA transformation does not change the mutual information; 2) if the components of the random vector are independent, the joint and joint-conditional entropies become the sum of marginal and marginal-conditional entropies, [1] as shown in (2) and (3)

$$I(x_1,x_2,...,x_d;\mathbf{c}) = I(y_1,y_2,...,y_d;\mathbf{c}) \tag{2}$$

$$\begin{aligned} I(y_1,y_2,\ldots,y_d;\mathbf{c}) \approx & \sum_j \sum_d H(y_{dj}) \\ & - \sum_c P_c \sum_j \sum_d H(y_{dj} \mid \mathbf{c}) \end{aligned} \tag{3}$$

where $\mathbf{x}$ is the original EEG channels, $\mathbf{y}$ is the transformed feature vector, d is the number of EEG channels being considered, j is the number of features for each EEG channel, and c is the class label.

Many effective and efficient algorithms based on a variety of assumptions, including maximization of non-Gaussianity and minimization of mutual information, exist

[1] Here we assume that the same ICA transformation achieves independence in the overall conditional distributions simultaneously.

to solve the ICA problem [13-15]. Those utilizing fourth order cumulants could be compactly formulated in the form of a generalized eigen-decomposition problem that gives the ICA solution in an analytical form [16].

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

$$\mathbf{R_x W} = \mathbf{Q_x W \Lambda} \qquad (4)$$

where $\mathbf{R_x}$ is the covariance matrix and $\mathbf{Q_x}$ is the cumulant matrix estimated using sample averages: $\mathbf{Q_x}$=E[$\mathbf{x^T x x x^T}$]-$\mathbf{R_x}$tr($\mathbf{R_x}$)-E[$\mathbf{x x^T}$]E[$\mathbf{x x^T}$]-$\mathbf{R_x R_x}$. Given the estimates for these matrices, the ICA solution can be easily determined using efficient generalized eigendecomposition 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 [13]. 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:

$$\hat{H}(Y) = \frac{1}{N-m} \sum_{i=1}^{N-m} \log \frac{(N+1)(y_{(i+m)} - y_{(i)})}{m} \qquad (5)$$

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

Salient EEG channel Ranking

Using the proposed MI estimation method described above, the salient EEG channel ranking can be achieved in a recurrent manner:

A. Estimating the MI between all features in one EEG channel and class labels. Repeat this process for all channels, find the channel with maximum MI, and mark it as opt-sub1 (optimal subset of 1 channel).

B. Pick one in the remaining EEG channels, combine it with opt-sub1 to form sub2 (subset of 2 channels). Estimate MI between all features in sub2 and class labels. Repeat this process for all remaining channels, find the channel with maximum MI, and mark it as opt-sub2.

C. Repeat Step B by increasing one channel at a time, until all EEG channels are ranked in the sense of MI maximization.

This procedure results in an ordering of EEG channels such that the first d channels have maximal MI with class labels (approximations include linear ICA induces independence and no degenerate feature pairs such as the XOR problem exist). The choice of d to be used in the application is dependent on the requirement for classification performance and computational cost. In our experiments, we typically observed that only up to 10 of the 30 channels contribute novel discriminative information, while the other 20 do not increase MI.

Using this search strategy, the computational complexity is (n+1)n/2 (n is the total number of EEG channels) instead of the 2^n of exhaustive evaluation. Another advantage of this method is that, since the EEG ranking is independent of the classifier, it is computationally efficient (it does not require repeated classifier training) and it does not require re-ranking when we use another classifier, bringing design flexibility to the table. The pseudocode of the proposed EEG channel selection method is shown in Table. 1.

Table 1. Pseudocode for the process of channel selection. (n: number of EEG channels, d: number of ranked channels, x: feature space, c: class labels)

```
d=1;
x=[ ];
while (d<n)
    for i=1 to n-d+1
        x=[x, $x_i$];
        calculate $I$(x;c);
    end
    x=max$_i$(I(x;c));
    d=d+1;
end
```

IV. Experiments And Results

As mentioned in part II, we have collected EEG data in 6 combinations, corresponding to one of the two mental tasks for three subjects. Each case consists of about 3000 data samples in a 150 dimensional feature space (30 EEG channels × 5 frequency bands) with two classes: low and high workloads. We applied the proposed EEG channel selection approach on these data files. It is well known that the optimal EEG channels vary for different mental tasks and different subjects. We first applied the approach on individual subject-task combinations, and obtained specialized EEG channel rankings, called Local n (n is the number of the EEG channels). As an evaluation for the ability to select optimal channels across tasks and subjects, we also mixed all data files together and applied this approach to get a new ranking, called Global n. An instance of Local 10 and Global 10 EEG channels are shown in Table 2. The 7 channels from physiological experience Phy 7 are also listed as a reference [17].

To validate the proposed method, we employed a committee of 3 classifiers: GMM, KNN, and KDE, together with majority vote and decision fusion on the selected EEG channels [18]. We divided each case into five parts, using four parts as the training set to train the classifiers, and the remaining one part as the testing set. Using the jackknife approach, we rotated the training data and testing data, and combined the results together as our final classification results. The confusion matrix is estimated and the correct classification rate can be calculated by weighted sum of the main diagonal entries of the confusion matrix. The classification rates for different data files under different subset of EEG channels are shown in Table 3. The arithmetic average of 6 correct classification rates within one selection of EEG channels is also listed as an overall channel selection evaluation.

Table 2 shows that the optimal channels are different across tasks and subjects, nevertheless, there exist common channels among these subsets. The classification results in Table 3 clearly show that the proposed method for channel selection is much superior to literature-based selection. The average classification rate also shows that the Local subsets are superior to Global subsets as expected, since they are optimal for that particular case. There is a trade-off between the classification performance and the computational cost in selecting the number of the EEG channels. Furthermore, more EEG channels might also introduce irrelevant information (after 10 channels in this case), which will compromise the robustness of the classification system.

Table 2. Optimal EEG channels illustration. Phy 7: 7 EEG channels from physiological experience; Local 10: 10 best EEG channels evaluated from individual data file; Global 10: 10 best EEG channels evaluated from all data files

Phy 7			Cz, P3, P4, Pz, O2, PO4, F7
Local 10	S_1	Larson	CP5, Fp2, FC5, Fp1, C4, P4, F7, AF3, P7, FC6
		n-back	AF3, FC5, Fp1, Fp2, F8, F7, FC6, O1, CP6, P4
	S_2	Larson	Fp2, O1, AF4, F7, C3, PO3, FC6, CP2, C4, Pz
		n-back	C4, O1, F8, Fz, F3, FC5, FC1, C3, Cz, CP1
	S_3	Larson	Fp2, F8, F7, FC5, FC6, AF3, C3, F4, P4, AF4
		n-back	CP5, F8, C4, FC6, Fp2, FC5, P3, AF4, C3, P7
Global 10			Fp2, FC5, O1, F3, FC6, F8, F7, AF3, O2, CP6

Table 3 Classification rate for three subjects: S_1, S_2 and S_3, in two mental tasks: Larson and n-back, for different subsets of EEG channels. Average is arithmetic average of the 6 classification rate for a particular EEG channel subset.

		Phy 7	7 Local	10 Local	7 Global	10 Global
S_1	Larson	0.74	0.90	0.86	0.84	0.78
	n-back	0.78	0.89	0.87	0.85	0.83
S_2	Larson	0.64	0.80	0.80	0.75	0.77
	n-back	0.73	0.89	0.89	0.88	0.86
S_3	Larson	0.48	0.71	0.76	0.76	0.73
	n-back	0.55	0.75	0.80	0.80	0.78
Average		0.65	0.82	0.83	0.81	0.79

V. Conclusion

In this paper, we presented an MI maximization method for EEG channel selection in BCI application. By exploiting the facts that MI does not change due to the linear ICA transformation, and that the joint entropy of the independent components can be estimated by summing the marginal entropies of each component, we can estimate the MI between the EEG channels and the class labels in an efficient way. A sequential incremental ranking strategy is also applied to reduce the number of total MI evaluations. Since the proposed feature selection method is a filter approach, it is independent of the classifiers, thus it does not require re-ranking the channels when using other classifiers.

Experiments results shows that this method yields superior classification results compared with the results for channels selected from physiology experience. Although this method does not use any prior knowledge about the brain activities, the selected EEG sites exhibit high consistency with expected cortical areas for these mental tasks: most channels are selected from the frontal sites, which are generally associated with working memory tasks.

Acknowledgement

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

References

[1] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, T.M. Vaughan, Brain–computer interfaces for communication and control. Clinical Neurophysiology 113, pp. 767-791, 2002.

[2] T.N. Lal, M. Schroder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, B. Scholkopf, Support Vector Channel Selection in BCI. Max Planck Institute for Biological Cybernetics, Tech. Rep. N. 120, Dec. 2003.

[3] Michael Schroder, Martin Bogdan, Wolfgang Rosenstiel, Thilo Hinterberger, and Niels Birbaumer, Automated EEG Feature Selection for Brain Computer Interfaces, Proceedings of 1st International IEEE EMBS Conference on Neural Engineering, Capri Island, Italy, Mar. 20-22, 2003.

[4] Luca Citi, Riccardo Poli, Caterina Cinel, Francisco Sepulveda, Feature Selection and Classification in Brain Computer Interfaces by Genetic Algorithm, Proceedings of Genetic and Evolutionary Computation Conference – GECCO 2004, Seattle.

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

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

[7] E. Halgren, C. Boujon, J. Clarke, C. Wang, and P. Chauvel, Rapid Distributed Fronto-parieto-occipital Processing Stages During Working Memory in Humans. Cerebral Cortex, Vol. 12, No. 7, Jul. 2002.

[8] Alan Gevins, Michael E. Smith, Linda McEvoy, and Daphne 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, No. 4, Jun. 1997

[9] Alan Gevins, Michael E. Smith, Neurophysiological Measures of Working Memory and Individual Differences in Cognitive Ability and Cognitive Style. Cerebral Cortex, Vol. 10, No. 9, Sep. 2000.

[10] B. Widrow and M. E. Hoff, "Adaptive switching circuits," in IRE WESCON Convention Record, 1960, pp. 96-104.

[11] 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.

[12] 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.

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

[14] 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.

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

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

[17] 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.

[18] Deniz Erdogmus, Andre Adami, Michael Pavel, Tian Lan, Santosh Mathan, Stephen Whitlow, Michael Dorneich, Cognitive State Estimation Based on EEG for Augmented Cognition. Proceedings of the 2nd International IEEE EMBS Conference on Neural Engineering, Arlington, Virginia, Mar. 16-19, 2005.