ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

Feature extraction for EEG classification: representing electrode outputs as a Markov stochastic process

Liang Wu and Predrag Neskovic *

Department of Physics and Institute for Brain and Neural Systems
Brown University, Providence, RI 02906

Abstract. In this work we introduce a new model for representing EEG signals and extracting discriminative features. We treat the outputs of each electrode as a stochastic process and assume that the sequence of variables forming a process is stationary and Markov. To capture temporal dependences within an electrode, we use conditional entropy and to capture dependences between different electrodes we use conditional mutual information features of increasing complexities. We show that even when using a small number of sampling points for their estimation (e.g. a single trial) these features carry discriminative information. We test the usefulness of these features by classifying the EEG data from n-back memory tasks.

1 Introduction

Extracting informative and discriminative features from EEG signals is often of crucial importance for representing and classifying patterns of brain activations. Among the common techniques for analyzing EEG data and extracting features are power spectrum analysis [1], auto-regression (AR) analysis [2], and independent component analysis (ICA) [3]. Information theoretic methods, such as entropy (H) and mutual information (MI) have also been used to assess EEG signals and to discriminate between Alzheimer's and normal patients [4, 5]. Similarly, entropy has been used to characterize cognitive states and it has been shown that the entropy during the resting state is higher compared to the entropy during various cognitive tasks (e.g. the mental arithmetic task [6]).

In our previous work [7] we demonstrated that entropy and MI can be used to extract localized features and that MI can capture both linear and non-linear dependences between pairs of electrodes. We showed that those features outperform both power spectrum and linear correlation features when applied to classifying n-back EEG data. The main shortcoming of H and MI features is that they provide a very coarse representation of electrode outputs. For example, if we want to capture the "spatial" dependences among electrodes (i.e. between electrodes from different spatial locations) we can do it by representing electrode outputs as independent and identically distributed (i.i.d.) random variables, as

*This work was supported by the Research Seed Fund Award from Brown University. The authors thank Elena Festa and William Heindel for providing the data.

in [7], and then calculate the MI. However, if we want to capture temporal dependences, either among different electrodes or within the same electrode, we have to use a finer representation of electrode outputs.

The main objective of this paper is to propose one such solution and model outputs of each electrode as a stochastic process. Using a sequence of random variables clearly provides a much more accurate representation compared to the one that uses i.i.d. random variables to capture the outputs. However, this comes with a high computational price. For example, using our data, if we want to model a sequence that is only one trial long (around 1,125 sampling points), we would need over one thousand random variables. Obviously, such a sequence would not be of much use for practical purposes since estimation of entropy and mutual information would require prohibitive amounts of data. For that reason, in this work we assume that the stochastic process (representing the output of an electrode) is stationary and Markov. To capture temporal dependences within an electrode we use conditional entropy feature and to capture dependences between pairs of electrodes we construct mutual information-based features. We test the usefulness of these features by classifying the EEG data from n-back tasks and demonstrate their advantage over simple entropy and MI features that do not incorporate temporal information.

2 Spatio-temporal features

In this section we introduce features that we use to capture dependences within and between electrodes whose outputs are represented with stochastic processes. We define a stochastic process $\{X^t\}_{t=1}^T$ as an indexed sequence of random variables [8]. The process, for a given electrode, is characterized by the joint probability mass functions $P\{(X^1, X^2, ..., X^T) = (x^1, x^2, ..., x^T)\} = p(x^1, x^2, ..., x^T)$.

Suppose that we represent the outputs of one electrode with the sequence $(X^1, X^2, ..., X^T)$ and the outputs of another electrode with $(Y^1, Y^2, ..., Y^T)$. We can then calculate the dependence between these two electrodes by calculating the MI for the two sequences. Using the chain rule for MI [8] it is straightforward to calculate the MI between series $(X^1, X^2, ..., X^T)$ and $(Y^1, Y^2, ..., Y^T)$

$$
\begin{aligned}
& I(X^1, X^2, ..., X^T; Y^1, Y^2, ..., Y^T) = \sum_{j=1}^{T} I(X^j; Y^1, Y^2, ..., Y^T | X^{j-1}, .., X^1) = \\
& \sum_{i,j=1}^{T} I(X^j; Y^i | X^{j-1}, ..., X^1, Y^{i-1}, ..., Y^1) = \sum_{i=1}^{T} I(X^i; Y^i | X^{i-1}, Y^{i-1}, ..., X^1, Y^1) \\
& + \sum_{i=1}^{T-1} I(X^i; Y^{i+1} | X^{i-1}, Y^i, ..., X^1, Y^1) + \sum_{i=2}^{T} I(X^i; Y^{i-1} | X^{i-1}, Y^{i-2}, ..., X^1, Y^1) \\
& + ... + I(X^1; Y^T | Y^{T-1}, ..., Y^1) + I(X^T; Y^1 | X^{T-1}, ..., X^1) = \\
& \equiv I(\tau = 0) + I(\tau = -1) + I(\tau = 1) + ... + I(\tau = -(T-1)) + I(\tau = T-1).
\end{aligned}
$$

where τ is the time delay. In this work, we consider only the time delay $\tau = 0$.

In order to simplify the previous expressions, we will assume that the sequence of random variables is Markov

$$P\{X^{t+1} = x^{t+1}|X^t = x^t, ..., X^1 = x^1\} = P\{X^{t+1} = x^{t+1}|X^t = x^t\},$$

time invariant

$$P\{X^{t+1} = b|X^t = a\} = P\{X^+ = b|X = a\},$$

(for all t, and all a, b), and stationary [8]. As a consequence of these assumptions, each term in the summation of $I(\tau = 0)$ is equal and the average MI for a pair of variables is

$$\begin{aligned}\bar{I}(\tau = 0) = \frac{1}{T}\sum_{i=2}^{T} I(X^i; Y^i|X^{i-1}, Y^{i-1}) &\equiv I(X^+; Y^+|X, Y) \qquad (1)\\ &= H(X^+, X, Y) + H(Y^+, X, Y) - H(X, Y) - H(X^+, Y^+, X, Y).\end{aligned}$$

If we assume that X^+ and Y^+ are independent of Y given X then the previous expression reduces to

$$\begin{aligned}I(X^+; Y^+|X) &= H(X^+|X) - H(X^+|X, Y^+) \qquad (2)\\ &= H(X^+, X) + H(X, Y^+) - H(X^+, Y^+, X) - H(X).\end{aligned}$$

where $I(X^+; Y^+|X)$, in general, is not symmetric ($I(X^+; Y^+|X) \neq I(X^+; Y^+|Y)$) for any specific pair of electrodes. However both $I(X^+; Y^+|X)$ and $I(X^+; Y^+|Y)$ produce the same feature set once we go through all the electrodes.

Finally, assuming that the outputs are i.i.d. random variables, the mutual information of two times series reduces to

$$I(X; Y) = H(X) + H(Y) - H(X, Y). \qquad (3)$$

The entropy of a single random variable X, and joint entropy of (X, Y) is

$$H(X) = -\sum_{x_i} p(x_i)\log p(x_i), \qquad H(X, Y) = -\sum_{x_i, x_j} p(x_i, x_j)\log p(x_i, x_j).$$

With the lower index we label different values of a random variable and with the upper index we label random variables. Generalization to more than two variables is straightforward.

Note that for $X = Y$ Eq. (1) and Eq. (2) reduce to the conditional entropy (CH) of X^+ given X: $I(X^+; X^+|X, X) = I(X^+; X^+|X) = H(X^+|X)$, where

$$H(X^+|X) = -\sum_{i,j} p(x_i^+, x_j)\log\frac{p(x_i^+, x_j)}{p(x_j)}.$$

To estimate entropy, $E(H)$, we use a Bayesian approach with a Dirichlet prior, as described in [7].

2.1 Data Acquisition

Six subjects (ages 20-24, 5 females and 1 male), performed an n-back memory task while the EEG was recorded, as described in [7]. The n-back task requires subjects to decide whether a currently present stimulus matches one presented n trials previously. We used 4 different tasks reflecting 4 different memory loads (n = 0, 1, 2 and 3).

Electrical activity (EEG) was recorded with 62 electrodes using 10-20 International System. The EEG was amplified by battery-operated amplifiers (EMS, Inc.) with a gain of 46K through a bandpass of 0.01-100Hz. Electrode impedances were kept below $5k\Omega$ when possible. EEG was continuously acquired at a sampling rate of 512Hz and stored on a disk for offline analysis.

One session of EEG data recorded from one subject during one task includes 102 trials. The first 6 and the last 6 trials were ignored and therefore we use 90 trials per task. The length of each trial is about 2.2 seconds which means that for each electrode there are around 1,125 sampling points per trial.

3 Results

In this section we present the effectiveness of different features for the classification of EEG signals. The objective was to associate a segment of the EEG signal with both the subject and the memory task. Since we use six subjects and four tasks there are all together 24 classes, c_i, $i = 1, 24$.

The raw data is processed using the surface Laplacian [9] and filtered into three bands: A (1-20Hz), B (1-50Hz), and C (1-80Hz). Within each band we then extracted the following features: Conditional Entropy (CH), and Mutual Information features, MI(m), where m is the number of random variables used to calculate them. MI(3)= $I(X^+; Y^+|X)$ features combine information from 3 variables as defined in Eq. (2), and MI(4)= $I(X^+; Y^+|X, Y)$ features capture information from four variables Eq. (1). Note that MI features are always computed between different electrodes so there are 3,782 MI(3) features, and 1,891 MI(2) and MI(4) features. In addition, and for comparison purposes, we also extracted the power spectrum (PS), and entropy (H) features and the number of these features is 62.

Classification. Since the goal of this work is to contrast the effectiveness of different feature extraction methods, we use a classifier that is easy to implement and fast to train - a Naive Bayes (NB) classifier. The classifier uses Bayes rule to calculate the probability that a given EEG segment, represented with features $(f_1, ..., f_F)$, belongs to specific class c_k,

$$p(c_k|(f_1, ..., f_F)) = \frac{\prod_i^F p(f_i|c_k)p(c_k)}{p(f_1, ..., f_F)}. \tag{4}$$

As a consequence of the assumption that the value of each feature is independent of the value of any other feature given the class, the NB classifier can easily deal with high-dimensional feature vectors and can be trained with relatively small number of training examples.

We model the likelihoods using univariate Gaussian distributions $N(\mu_i, \sigma_i)$ and calculate the mean μ_i and the variance σ_i using maximum likelihood estimates. The parameters associated with each class are estimated using data from one task and one subject. The prior term, $p(c_k)$, is the same for all classes.

To evaluate a classifier we use a leave-one-out method. In this work, we present results using segments that are one (five) trials long.

Features:	MI(2)	MI(3)	MI(4)	MI(3)+CH	MI(4)+CH
Band A:	56.1%	70.9%	79.0%	70.7%	79.2%
Band B:	75.3%	86.5%	90.1%	87.1%	90.3%
Band C:	77.6%	89.3%	92.6%	89.3%	92.8%

Table 1: Classification rates from single trials.

In the first set of experiments, we compare the mutual information features of increasing complexity using single trial segments, Table 1. The MI(2) features capture only "spatial" information (dependences between electrodes from different spatial locations) whereas the MI(3) features, in addition, incorporate temporal information from one electrode. The MI(4) features are the most complex and include temporal information. It is important to note that the number of samples needed for accurate estimation of MI(4) features is significantly larger compared to other features, e.g. over two orders of magnitude compared to MI(2) features. It is remarkable that despite this fact, classification consistently improves with the increased complexity of the features. Adding temporal information from the electrodes in isolation (CH), as shown in the last two columns, does not significantly change the performance. This is to be expected since temporal information captured by CH features is implicitly contained in MI(3) and MI(4) features. Therefore, although MI(4) features capture more information compared to CH, they also capture more noise.

In order to evaluate the importance of the size of the EEG segment on performance, we repeated the previous experiments but now using five trials long segments, Table 2. The classification rates are significantly higher for all the features and bands, since a higher number of samples improves MI estimation.

In Table 3 we contrast PS, H, and CH features using single trial EEG segments (first three columns) and 5 trials long segments (last three columns). As one can see, the CH features outperform all other features, even the MI(4) features that capture much more information. However, one cannot draw the conclusion that CH features are more discriminative than MI(4) features because

Features:	MI(2)	MI(3)	MI(4)	MI(3)+CH	MI(4)+CH
Band A:	80.8%	85.4%	88.6%	86.4%	88.5%
Band B:	86.1%	94.4%	95.5%	93.9%	95.5%
Band C:	88.2%	95.8%	96.5%	95.3%	96.5%

Table 2: Classification rates using five trials long segments.

the dimensionality of the feature space is much larger for MI(4) features than for CH features (1,891 compared to 62) and the limited number of samples, for a given EEG segment, impacts MI(4) features more adversely than CH features.

Features:	PS	H	CH	PS	H	CH
Band A:	64.5%	71.9%	87.3%	75.1%	85.9%	95.5%
Band B:	85.7%	87.0%	94.9%	91.5%	92.4%	97.9%
Band C:	89.2%	89.0%	95.7%	94.4%	93.3%	98.1%

Table 3: Classification rates using 1 (columns 1-3) and 5 trials (columns 3-6).

4 Conclusions

In this work, we represented the output of each electrode as a stationary and Markov stochastic process, and used conditional entropy and conditional MI to capture dependences within and between electrodes. We showed that the performance of the classifier always increases with the complexity of the features and with the size of the training/testing segment. Furthermore, we showed that the most complex features (MI(4)) provide significant discriminative information even when estimated from a single trial.

References

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

[2] M. Akin and M. K. Kiymik. Application of periodogram and AR spectral analysis to EEG signals. *Journal of Medical Systems*, 24(4):247–256, 2000.

[3] S. Chiappa and D. Barber. Generative independent component analysis for EEG classification. In *European Symposium on Artificial Neural Networks*, pages 297–302, 2005.

[4] D. Abasolo, R. Hornero, P. Espino, D. Alvarez, and J. Poza. Entropy analysis of the EEG background activity in Alzheimers disease patients. *Physiol. Meas.*, 27:241–253, 2006.

[5] J. Jeong, J. C. Gore, and B. S. Peterson. Mutual information analysis of the EEG in patients with alzheimer's disease. *Clin. Neurophysiol.*, 112:827–835, 2001.

[6] T Inouye, K. Shinosaki, H. Sakamoto, S. Toi, A. Iyama, Y. Katsuda, and M. Hirano. Abnormality of background eeg determined by the entropy of power spectra in epileptic patients. *Electroencephalogr. Clin. Neurophysiol.*, 82:203–207, 1992.

[7] L. Wu, P. Neskovic, Etienne Reyes, Elena Festa, and William Heindel. Classifying n-back EEG data using entropy and mutual information features. In *European Symposium on Artificial Neural Networks*, 2007.

[8] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

[9] B. Hjorth. An on-line transformation of eeg scalp potentials into orthogonal source derivations. *Electroencephalog. Clin. Neurophysiol.*, 39(5):526–530, 1975.