Classifying n-back EEG data using entropy and mutual information features

Liang Wu[1], Predrag Neskovic[1], Etienne Reyes[1],
Elena Festa[2], and William Heindel[2] *

Department of Physics and Institute for Brain and Neural Systems[1],
Department of Psychology[2], Brown University, Providence, RI 02906

Abstract. In this work we show that entropy (H) and mutual information (MI) can be used as methods for extracting spatially localized features for classification purposes. In order to increase accuracy of entropy estimation, we use a Bayesian approach with a Dirichlet prior to derive estimation equations. We calculate the H and MI features for each electrode (H) and pair of electrodes (MI) in three frequency bands and use them to train the Naive Bayes classifier. We test the H and MI features on one/five trial long segments of n-back memory EEG signals and show that they outperform power spectrum and linear correlation features respectively.

1 Introduction

Electroencephalography (EEG) is a non-invasive technique that was first used by Hans Berger in 1929 to record electrical activity of the human brain. Since then, the EEG has been successfully used in numerous applications such as medical diagnosis and EEG-based brain-computer interfaces [1]. In these applications, extraction of informative and discriminative features plays a very important role.

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 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]).

Of particular interest for analysis of EEG signals are the techniques that can capture interactions among different brain areas, such as the correlation analysis and the MI. The main advantage of the MI over standard correlation methods (such as the coherence analysis) is that it captures not only linear but also nonlinear dependencies without requiring the specification of any kind of dependence. Perhaps the main drawback of using entropy and MI is that it is often difficult to accurately estimate them from the data. In general, the first step in calculating the entropy of a variable X is to discretize it by dividing it into bins. If the number of samples (N) is large and the number of bins (K) satisfies the condition $K << N$, one can safely use a frequency-based approach.

*This work was supported by the Research Seed Fund Award from Brown University.

However, if that condition is not satisfied, one has to devise a different procedure for entropy estimation. In this work we describe one such method, namely a Bayesian approach with a Dirichlet prior for estimating entropy and therefore MI.

The central aim of this paper is to demonstrate that entropy and mutual information can be used as methods for extracting spatially localized features for classification purposes. More specifically, we use entropy to characterize the outputs of each electrode in isolation and MI to capture dependences between different electrodes. We then use these features to train a Naive Bayes (NB) classifier to associate a given segment of the n-back memory EEG data with both the subject and the task. We compare the entropy and MI features to traditional features such as the power spectrum and linear correlation in three different frequency bands and show that the entropy and MI features outperform the PS and LC features respectively in all bands.

2 Entropy and Mutual Information

Entropy. The entropy is a non-negative quantity and measures the uncertainty of a random variable. If we denote with symbol X a discrete random variable that takes values from a set $\{x_i\}$ and with $P\{X = x_i\} = p(x_i)$ a probability that the variable X takes a specific value x_i the entropy of X is defined as

$$H(X) = -\sum_{x_i} p(x_i) \log p(x_i). \tag{1}$$

For example, if the variable X represents a specific electrode and $\{x_i\}$ is the collection of measurements from the electrode then if the entropy of the variable X is zero that means that each possible measurement occurs with a probability of either 1 or 0. In other words, a zero value of entropy indicates that there is no uncertainty related to the outputs. Similarly, higher values of entropy correspond to higher uncertainty of the outputs (the measurements) of the electrode.
Mutual Information. The mutual information is the reduction in the uncertainty of X due to the knowledge of Y

$$I(X;Y) = H(Y) - H(Y|X) = \sum_{x_i, y_j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}. \tag{2}$$

It is clear that the MI is a symmetric function $I(X;Y) = I(Y;X)$ and equal to zero if the variables are independent. The MI of a random variable with itself is just the entropy of the random variable, i.e. $I(X;X) = H(X) - H(X|X) = H(X)$, and for that reason entropy is sometimes called *self-information.*

3 Entropy Estimation

We now derive the expression that we use to calculate the entropy. We divide the electrode outputs into K bins and denote with vector $\mathbf{n} = n_i$ the number of counts per bin so that $N = \sum_i^K n_i$. With variable $\mathbf{q} = \{q_i\}$ we denote the vector consisting of true (unknown) probabilities of the states, $1 \leq i \leq K$. To estimate the expected entropy we use a Bayesian approach

$$\hat{H} = \int d\mathbf{q} H(\mathbf{q}) \frac{P(\mathbf{n}|\mathbf{q})P(\mathbf{q})}{P(\mathbf{n})}, \tag{3}$$

where the normalization term is $P(\mathbf{n}) = \int d\mathbf{q}' P(\mathbf{n}|\mathbf{q}')P(\mathbf{q}')$. We assume that the observations are repeatedly and independently sampled from the distribution $\mathbf{p}$ which means that $\mathbf{n}$ is multinomially distributed. Therefore, $P(\mathbf{n}|\mathbf{q}) = N! \prod_{i=1}^K (q_i^{n_i}/n_i!)$. We choose for the prior the Dirichlet distribution $Dir(\mathbf{q}) \propto \prod_i q_i^{\beta-1}$. In order to enforce the non-negativity of $\mathbf{q}$ we include the Heaviside function ($\theta(x) = 1$ for $x \geq 0$ and 0 otherwise) and in order to enforce the condition that $\sum_i q_i = 1$ we include the delta function $\Delta(\mathbf{q}) \equiv \delta(\sum_i q_i - 1)$. Our prior then becomes $P(\mathbf{q}) \propto \Delta(\mathbf{q}) \prod_i \theta(q_i) q_i^{\beta-1}$. It is interesting to note that for both the multinomial and for the Dirichlet distribution with $\beta = 0$ the maximum likelihood estimates of the probabilities are frequencies, $q_i = n_i/N$.

The parameter β reflects the prior knowledge of the number of data points in each bin, the effective number of observations. In our experiments, we set this parameter to zero, $\beta = 0$. However, rather then choosing a specific value for the parameter β, it has been shown [7] that better estimates can be obtained by averaging over this parameter. Since the objective of this work is to demonstrate the usefulness of entropy-based features for the classification purposes and not the accuracy of the estimation, we did not implement this more elaborate approach.

Calculating the integral in Eq. (3) is quite difficult and one might be tempted to estimate the simpler quantity, the densities $\hat{\mathbf{q}} = \{\hat{q}_i\}$, and then just plug them in the expression for the entropy. Unfortunately, this quick solution is incorrect. This is due to the fact that the entropy is a nonlinear function of the probabilities and therefore entropy of an average is generally not the same as the average of entropy ($H(\hat{\mathbf{q}}) \neq \hat{H}(\mathbf{n})$). Fortunately, calculating moments is rather easy [8]. The expected value of the α moment of q is

$$E(q_j^\alpha) \equiv \frac{\int d\mathbf{q} q_j^\alpha P(\mathbf{q}) \prod_i q_i^{n_i}}{\int d\mathbf{q} P(\mathbf{q}) \prod_i q_i^{n_i}} = \frac{\Gamma(n_j + \beta + \alpha)\Gamma(n + m\beta)}{\Gamma(n_j + \beta)\Gamma(n + m\beta + \alpha)}, \tag{4}$$

where we use as the prior the Dirichlet distribution. Then, noting that $q_j \ln q_j$ can be written as $\partial(q_j^\alpha)/\partial\alpha|_{\alpha=1}$, the expected value of the entropy becomes

$$E(H) = -\sum_{j=1}^K E(q_j \ln q_j) = \frac{1}{N + K\beta} \sum_{j=1}^K (n_j + \beta)(\psi(N + K\beta + 1) - \psi(n_j + \beta + 1)),$$

where $\psi(z) = \frac{d \ln \Gamma(z)}{dz}$ is a poly-gamma function. As one can see, the expression for the expected entropy is quite simple and is not computationally expensive. It depends both on the data (counts n_j) and our prior knowledge. Generalizing this expression to MI is straightforward and amounts to replacing the count per bin n_j with counts per two bins, $n_{i,j}$.

3.1 Data Acquisition

Six subjects (ages 20-24, 5 females and 1 male), performed an n-back memory task while the EEG was recorded. The n-back task requires subjects to decide whether a currently present stimulus matches one presented n trials previously. Cognitive memory load was manipulated across different blocks of trials by varying the number of previous trials back in a sequence of trials (from 0-back to 3-back).

Vertical and horizontal eye movements were recorded by electro-oculography via bipolar electrodes placed on the external canthi of the eyes (horizontal EOG) and on the inferior and superior areas of the ocular orbit (vertical EOG). Scalp electrical activity (EEG) was recorded from 58 tin electrodes mounted in an electrode cap (Quick-cap, Neuromedical Supplies, Inc.). Electrode positions included the standard 10-20 International System locations and additional intermediate positions. Recordings were performed with a linked mastoid reference. 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 there are around 1,125 sampling points per trial.

4 Results

In this section we present the effectiveness of different features for the classification of EEG signals. The objective is 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 different bands: A (1-20Hz), B (1-50Hz), and C (1-80Hz). Within each band we then extracted the following features: Power Spectrum (PS), Entropy (H), Linear Correlation (LC), and Mutual Information (MI).

Power Spectrum and Entropy are calculated for each electrode separately whereas the LC and MI are calculated for each pair of electrodes. If we represent the outputs of two electrodes with X and Y, then the LC between the two electrodes is calculated as $LC(i,j) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$.

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 NB classifier outputs the probability that a given sequence of EEG data, represented with features $(f_1, ..., f_F)$, belongs to specific class c_k,

$$p(c_k|(f_1, ..., f_F)) = \frac{\prod_i^F p(f_i|c_k)p(c_k)}{p(f_1, ..., f_F)}. \tag{5}$$

We model the likelihood terms using univariate Gaussian distributions $N(\mu_i, \sigma_i)$ and calculate the mean μ_i and the variance σ_i using maximum likelihood estimates from the data. 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.

The main assumption here is that the value of each feature is independent of the value of any other feature given the class. Despite this unrealistic (naive) assumption, the NB has been successfully applied in many practical situations and it has been shown that it is comparable to much more sophisticated classifiers [10]. Furthermore, the NB classifier can easily deal with high-dimensional feature vectors and can be trained using a relatively small number of examples.

To evaluate a classifier we use a leave-one-out method. In this work, we present results using one/five trial long segments.

Features:	PS	H	LC	MI	MI+H
Band A:	64.5%	71.9%	54.9%	56.1%	58.0%
Band B:	85.7%	87.0%	67.4%	75.3%	76.3%
Band C:	89.2%	89.0%	71.1%	77.6%	78.2%

Table 1: Single trial classification rates with NB classifier.

In the first experiment, we compare the entropy features to power spectrum features, Table 1 (first two columns). Both PS and H features are extracted from single trial EEG segments and from three different frequency bands. The number of PS as well as H features is 62. One can see that entropy features outperform power spectrum features in lower frequency bands while the performance equalizes as one includes information from higher frequencies.

In the second experiment, we contrast LC features against MI features, Table 1 (columns three and four). The number of LC and MI features is 1,891 since we calculate them only between *different* electrodes and LC(X,X) and $MI(X;X) = H(X)$ are excluded. As one can see, the MI features outperform linear correlation features in all frequency bands. Note that while in the first experiment the features represent each electrode in isolation the features in the second experiment capture dependences between different electrodes. Although the entropy features outperform all other features, we cannot draw a conclusion that H features are necessarily more discriminative compared to MI or LC features since the number of the MI and LC features is orders of magnitude larger compared to H features. Indeed, adding H features to MI features (fifth column) improves results only marginally which suggests that lower performance of LC features is likely due to higher dimensionality of the feature space.

In addition, MI features require significantly more sampling points for accurate estimation compared to H features. In order to evaluate the importance of the size of the training/testing segment on the performance, we repeat the previous two experiments but now using five trial long segments, Table 2. As expected, the classification rates are higher for all the features and bands, and the trend remains the same: the entropy outperforms the power spectra for

Feature type:	PS	H	LC	MI	MI+H
Band A:	75.1%	85.9%	72.5%	80.8%	81.5%
Band B:	91.5%	92.4%	82.1%	86.1%	86.8%
Band C:	94.4%	93.3%	84.2%	88.2%	88.7%

Table 2: Classification rates with NB classifier using five trials long segments.

lower frequencies while the MI features outperforms LC features in all frequency bands.

5 Conclusions

In this work we demonstrated that entropy and mutual information can be used to extract discriminative features that can be useful for classification purposes. In order to increase accuracy of entropy estimation, we used a Bayesian approach with a Dirichlet prior and derived the estimation equations. We calculated H and MI features for each electrode (H) and pair of electrodes (MI) and used them to train the NB classifier. We tested the H and MI features on one/five trial long EEG segments and showed that for n-back memory tasks they outperform power spectrum and linear correlation features respectively.

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] I. Nemenman, W. Bialek, and R. R. Steveninck. Entropy and information in neural spike trains: Progress on the sampling problem. *Phys. Rev. E*, 69:56111–56116, 2004.

[8] D. H. Wolpert and D. R. Wolf. Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E*, 52:6841–6854, 1995.

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

[10] D. J. Hand and K. Yu. Idiot's bayes - not so stupid after all? *International Statistical Review*, 69:385–399, 2001.