CLASSIFICATION OF ELECTROENCEPHALOGRAPHY (EEG) SIGNALS FOR DIFFERENT MENTAL ACTIVITIES USING KULLBACK LEIBLER (KL) DIVERGENCE

Anjum Gupta

gupta@spawar.navy.mil
SSC San Diego
53150 Systems CT (Room 309)
San Diego CA 92152

Shibin Parameswaran

sparames@ucsd.edu
University of California, San Diego
Dept. of Electrical Engineering
9500 Gilman Dr La Jolla CA 92092 USA

Cheng-Han Lee

chl079@ucsd.edu
University of California, San Diego
Dept. of Computer Sciences and Eng.
9500 Gilman Dr La Jolla CA 92092 USA

ABSTRACT

Automatic classification of electroencephalography (EEG) signals, for different type of mental activities, is an active area of research and has many applications such as brain computer interface (BCI) and medical diagnoses. We introduce a simple yet effective way to use Kullback-Leibler (KL) divergence in the classification of raw EEG signals. We show that k-nearest neighbor (k-NN) algorithm with KL divergence as the distance measure, when used using our feature vectors, gives competitive classification accuracy and consistently outperforms the more commonly used Euclidean k-NN. We also develop and demonstrate the use of a KL-based kernel to classify EEG data using support vector machines (SVMs). Our KL-distance based kernel compares favorably to other well established kernels such as linear and radial basis function (RBF) kernel. The EEG data, used in our experiments for classification, was recorded while the subject performed 5 different mental activities such as math problem solving, letter composing, 3-D block rotation, counting and resting (baseline). We present classification results for this data set that are obtained by using raw EEG data with no explicit artifact removal in the pre-processing steps.

Index Terms— Electroencephalography, Pattern classification, Kullback-Liebler (KL) divergence, Support Vector Machines, Brain Computer Interface

1. INTRODUCTION

Automatic classification of different mental tasks using the brain waves as recorded in the EEG signals is an important and challenging problem. EEG signals are widely used for diagnosing neurological disorder such as epilepsy, sleep apnea etc. as well as in the field of Brain Computer Interface. EEG signals offer an inexpensive and relatively easy way of reading brain activity compared to methods such as fMRI, Magnetoencephalography (MEG) and Electrocorticographic (ECoG). These reasons make EEG a popular and an important resource for identifying and classifying the brain activity.

Processing and classification of EEG waves has been an active topic of research in the area of signal processing as well as machine learning. EEG signals are obtained by placing 2 to 64 sensors on the scalp of a subject and record the underlying brain activity. The classification task normally involves classifying a segment of the signal, normally ranging from 1 to 5 seconds in length. Obtaining a clean feature vector can be challenging as the EEG signals often record various artifacts that are introduced due to small muscle movements and eye blinks etc. Various methods for artifact removal have been developed that use many different filters and methods such as independent component analysis (ICA) [1] etc. Removing artifacts from the EEG signals is often an important pre-processing step before getting a feature vector [2]. For the classification of EEG signals, various machine learning based algorithms such as k-nearest neighbor, neural networks [3, 4], support vector machines [5] and Bayesian networks [6, **?**] have been used in the past.

In this paper, we show that using the KL-divergence as the underlying distance measure for the classification of EEG signals. We also propose rather simple but robust feature vector that can be efficiently computed from the signal using simple Fourier transform. We show the effectiveness of using KL-divergence for classification by using it with two algorithms, namely, k-NN with KL-divergence as the distance measure

and SVMs with KL kernel.

The KL divergence has been a well accepted distance measure between two probability distributions. We first show a method of obtaining a transformed feature set from the EEG waves that behaves like a probability distribution. Both of our classification methods using KL divergence consistently outperform the more commonly used Euclidean distance measure. We also demonstrate the classification performance obtained using KL kernel, which is comparable to well accepted kernels like radial basis function (RBF), linear and polynomial.

In the next section, we go over our data set and the method for obtaining and reasoning behind the new power-spectrum based feature set are explained. In section 3, the classification algorithms used are briefly described. In section 4, the results obtained from our test are presented and section 5 contains our conclusions and give some ideas for future research.

2. DATASET AND FEATURE EXTRACTION

The data used to conduct our experiments were originally collected by Keirn and Aunon [7]. The EEG recordings came from healthy human subjects performing 5 different mental tasks. These 5 mental tasks were:

1. Baseline: Do nothing specific and relax as much as possible
2. Multiplication task: Solve a given non-trivial multiplication problem without vocalizing or making any movements
3. Letter composing task: Compose a letter to a friend or a relative without vocalizing
4. Rotation task: Visualize a 3-D block figure and rotate it around an axis
5. Counting task: Imagine a blackboard and visualize, without vocalizing, numbers being written on it in a sequential order

The data was obtained from 6 EEG sensors placed on the scalp and 1 EOG sensor placed near the eyes. The 6 sensors were placed in positions C3, C4, P3, P4, O1 and O2 as defined by the 10-20 sensor placement system [8]. The data was recorded at 250 Hz and for 10 seconds during each mental task. Each task was repeated five times per session and the subjects attended two such sessions resulting in 10 trials per task per subject.

During feature extraction stage, each task is split into 2-second long windows with 90% overlap between the adjacent windows, giving 40 windows of 2 seconds. The feature vectors are derived from the power spectrum of these 2-second windows. Specifically, each feature vector consists of normalized power spectral density (PSD) in the frequency range from 8 Hz to 32 Hz of each of the 6 channels concatenated together. The range of 8 Hz to 32 Hz is chosen as most of the brain activity for different mental task is found in this frequency range [9].The PSD histograms are obtained with a frequency resolution of 2 Hz and hence each channel has 12 frequency components. It is important to note that the normalization is done separately for each individual channel. This represents the distribution of the power among the different frequency bands per channel and this behaves like a discrete probability distribution which represents the percent power contained in each frequency band. These 12 dimensional vectors from the 6 channels are concatenated together to form one 96 dimensional feature vector representing one 2 second window.

Each subject demonstrates a very different EEG wave composition for different mental tasks. So for the classification task all the data is used from subject number 6. We split the data set of 10 trials into training and testing sets in two different ways. First way was to randomly choose 10% of the data as the testing data and other 90% as the training data and repeat this process 20 times to generate a cumulative accuracy. Second method was to use 9 out of 10 trials as the training data and 1 out of 10 trials as the testing set. Each of the 10 trials was used as a test trial and the accuracy was obtained by averaging the accuracy obtained from each of the 10 trials. Note that the first way of getting training and testing set gives us a much higher accuracy since the testing and training points can be chosen from the same trial which can show a high degree of similarity thus making the classification task easier. However, both the methods for obtaining training and test sets have been used in the past research and both are important in demonstrating a particular capability and robustness of the classification algorithm.

3. CLASSIFICATION METHODS

We use two widely recognized classification algorithms for classifying the mental tasks, namely, k-NN and SVM. The underlying distance of choice for both methods is the KL-divergence. To be able to use the KL-divergence with SVMs, we implemented a KL-distance based kernel. Our results show that using KL divergence is a promising, yet unexplored, field for EEG classification. With no pre-processing step for artifact removal, our results using the KL-divergence show better classification accuracy than the more commonly used Euclidean distance.

3.1. KL Divergence Based Metric

KL Divergence is used to measure the distance between two probability distributions [10]. It has strong foundations in information theory. KL divergence between two distributions p

and q is defined as

$$KL(p,q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) \quad (1)$$

KL divergence is not symmetric. To make it a symmetric distance measure, we use averaging as shown below,

$$KL_{dist}(p,q) = \frac{KL(p,q) + KL(q,p)}{2} \quad (2)$$

In our case, the probability distributions are the power distribution in each of the channels. Hence, to find the KL distance between two feature vectors, we take the sum of KL distances between the sub-feature vectors representing the PSD's per channel. For instance let a point $\mathbf{x} = [\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_m}]'$ be the feature vector where each component vector $\mathbf{x_i}$ represents the normalized power spectral density of i^{th} channel. Then, the KL distance between two points $\mathbf{x}$ and $\mathbf{y}$ is given by

$$KL_{dist}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{m} \frac{KL(\mathbf{x_i},\mathbf{y_i}) + KL(\mathbf{y_i},\mathbf{x_i})}{2} \quad (3)$$

3.2. K-Nearest Neighbor Classifier

For the k-nearest neighbor classifier, we classify each test point to the class that majority of its closest k neighbors belong to. The number of k is varied from 3 to 9. In the case that there is a tie, the farthest neighbor is thrown out of consideration until the tie is broken and there is a clear majority. As noted earlier, the distance between two points is calculated using equation 3.

3.3. SVM with KL Distance Based Kernel

We followed the method recommended in [11] in kernelizing our KL distance. The KL kernel for EEG is obtained by exponentiating the KL distance given in equation 3.

$$K_{kldist}(\mathbf{x},\mathbf{y}) = exp^{(-a \cdot KL_{dist}(\mathbf{x},\mathbf{y})+b)} \quad (4)$$

where scaling factor $a > 0$ is similar to the variance of gaussian kernel and b is a shifting factor. The transformed data using the KL kernel is then classified using one-against-one multiclass SVM classifier. The parameters a and b were chosen minimized training set errors using a simple grid search. The results were compared against other well known kernels such as linear, polynomial and RBF kernels to demonstrate the effectiveness our KL distance based kernel for EEG.

3.4. Smoothing For Classification

We also experimented with "smoothing," where each 2 second was labeled using the average label of 3 second long data consisting of five 2 second windows with 1.8 second overlap. The smoothing shows a small improvement in the overall accuracy. The results presented in this paper, however, are obtained without any smoothing.

	Base	Math	Letter	Rotate	Count
Base	1				
Math	93.25	1			
Letter	68.00	87.75	1		
Rotate	97.63	89.00	93.88	1	
Count	82.75	68.50	76.38	85.75	1

Table 1. Pairwise Results - Euclidean Distance

	Base	Math	Letter	Rotate	Count
Base	1				
Math	95.00	1			
Letter	72.00	87.00	1		
Rotate	98.63	90.38	93.63	1	
Count	82.63	76.13	75.50	86.25	1

Table 2. Pairwise Results - KL Distance

4. RESULTS

Our experiments are conducted by separating time segments of training and testing from each other. Sometimes the results are obtained by allowing time segments from the same trial span both training and testing set. As expected, in the case where the time segments are distributed randomly among the test and training set, we obtain a much higher accuracy, generally exceeding 95% for 5 class classification. Table 3 shows the results from the randomly created testing and training sets. This distinction is important since many classification methods, that may show good classification in the randomized testing and training sets, sometimes fail to produce good results when tested across trials, where segments from any particular trial is not shared among the test and training set.

4.1. K-nearest neighbor results

We conduct the classification experiments by using the two distinct classifiers and using different sets of activities. Our feature vector and distance measure was first tested in pairwise classification. For pairwise classification, only two classes were used at a time for the classification. Table 1 shows the classification results for each pair of classes. The results listed in the table 1 are from 5-NN classifier using the Euclidean distance and table 2 shows the results from the KL-distance 5-NN classifier.

	KL-Divergence	Euclidean
3-nn	99.42	98.17
5-nn	98.75	96.58
7-nn	97.42	95.5
9-nn	96.08	92.83

Table 3. Accuracy With Randomly Created Test and Training Sets

5-Classes		3-Classes	
KL Div	Euclidean	KL Div	Euclidean
68.68	67.75	86.17	85.33

Table 4. Accuracy Results using KL and Euclidean Distance

	KL	RBF	Linear
Accuracy	0.78	0.75	0.73

Table 5. Accuracy Results For Different SVM Kernels

A subset of 3 classes was created based on the pairwise classification analysis. The table shows that pair of class 1 and 3 and the pair of class 3 and 5 do not show a good separation. The 3 classes included in that set were (1) Math (2) Letter and (3) Rotate. We also run the k-nearest neighbor with all 5 classes. Table 4 shows the results for the 5-class and a 3-class classification accuracy using both KL and Euclidean distance.

4.2. SVM results

The SVM results were obtained using three different kernels, namely, radial basis function (RBF), Linear, and our KL-kernel as described in the previous section. To avoid any bias towards any one kernel, the parameters a and b were optimized individually for each of these kernels separately. The parameters were optimized using a simple grid search. Each of the 10 trial was used as the test set and the results were averaged over all the 10 runs. Table 5 shows the accuracy results for each of the kernels. These accuracy results are for all 5 classes.

5. CONCLUSION AND FUTURE WORK

We show that KL-divergence is a viable and an effective distance measure to classify EEG signals. We show that this distance measure outperforms the Euclidean distance in K-nearest neighbor and the support vector machines using a KL distance based kernel outperforms RBF and linear kernels. More research can be done to develop more extensive feature set. Some more pre-processing steps such as artifact removal can also be implemented and may further improve the classification. The accuracy results from KL-divergence based classifiers generalize very well for classifying new trials from the same subject.

6. REFERENCES

[1] Tzyy-Ping Jung, Colin Humphries, Te-Won Lee, Scott Makeig, Martin J. McKeown, Vicente Iragui, and Terrence J. Sejnowski, "Extended ICA removes artifacts from electroencephalographic recordings," in *Advances in Neural Information Processing Systems*, Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, Eds. 1998, vol. 10, The MIT Press.

[2] P.S. Hammon and V.R. de Sa, "Preprocessing and meta-classification for brain-computer interfaces," *Biomedical Engineering, IEEE Transactions on*, vol. 54, no. 3, pp. 518–525, March 2007.

[3] Charles W. Anderson and Zlatko Sijercic, "Classification of eeg signals from four subjects during five mental tasks," in *In Proceedings of the conference on engineering applications in neural networks (EANN'96*, 1996, pp. 407–414.

[4] R. Palaniappan, "Brain computer interface design using band powers extracted during mental tasks," *Neural Engineering, 2005. Conference Proceedings. 2nd International IEEE EMBS Conference on*, pp. 321–324, March 2005.

[5] T N Lal, M Schroder, T Hinterberger, J Weston, M Bogdan, N Birbaumer, and B Scholkopf, "Support vector channel selection in bci," *IEEE Transactions on Biomedical Engineering*, , no. 51, pp. 1003–1010, 2004.

[6] Pradeep Shenoy and Rajesh P. N. Rao, "Dynamic bayesian networks for brain-computer interfaces," in *Advances in Neural Information Processing Systems 17*, Lawrence K. Saul, Yair Weiss, and Léon Bottou, Eds., pp. 1265–1272. MIT Press, Cambridge, MA, 2005.

[7] Z.A. Keirn and J.I. Aunon, "A new mode of communication between man and his surroundings," in *IEEE Transactions on Biomedical Engineering*, 1990, vol. 37(12), pp. 1209–1214.

[8] H. Jasper, "The ten twenty electrode system of the international federation," *Electroencephalographic Clinical Neurophysiology*, vol. 10, pp. 371–375, 1958.

[9] Bruce J. Fisch and Rainer Spehlmann, *Fisch and Spehlmanns EEG primer: Basic principles of digital and analog EEG*, Elsevier Health Sciences, 1999.

[10] S Kullback and R A Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, , no. 22, pp. 79–86, 1951.

[11] A. B. Chan, N. Vasconcelos, and P.J. Moreno, "A family of probabilistic kernels based on information divergence," Tech. Rep. SVCL-TR-2004-01, June 2004.