#400

Available online at www.sciencedirect.com
SCIENCE DIRECT®
NEUROCOMPUTING
ELSEVIER
Neurocomputing 61 (2004) 479–484
www.elsevier.com/locate/neucom

Letters

Classification of individual trials based on the best independent component of EEG-recorded sentences

Dik Kin Wong*, Marcos Perreau Guimaraes, E. Timothy Uy, Patrick Suppes

Center for Study of Language and Information, Stanford University, Stanford, CA, USA

Available online 20 August 2004

Abstract

Significant results for individual trials were achieved in recognizing brain waves using single-channel pereeptron-based classifiers (SCC) on electroencephalography (EEG) data generated by visual sentences as stimuli. By the use of a multiehannel classifier (MCC), classification rates were greatly improved for the subjects with bipolar settings, while the improvement for subjects with monopolar settings were small. We also present here an alternative multichannel scheme using independent component analysis (ICA). Sources were estimated by computing a linear eombination of EEG channels. Therefore, multichannel data were taken into account even when one independent component was used. Unlike multichannel perceptron classifiers which were only useful for bipolar recordings, the best ICA-component classifier (ICA SCC) was effective for both monopolar and bipolar settings.

Keywords: EEG; Language; Individual trials; ICA; Multichannel classifier; Monopolar; Bipolar

*Corresponding author.
E-mail address: dkwong@stanford.edu (D.K. Wong).

0925-2312/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2004.06.004

1. Introduction

In electroencephalography (EEG), potential differences are measured between pairs of electrodes. Placement of the electrodes on the scalp is described by the standard 10–20 system [7]. There are two connection methodologies: monopolar and bipolar. In our monopolar settings, the mastoid (behind the ear) was used as the common reference of the electrodes. For our bipolar settings, differences between neighboring electrode pairs were measured to better eliminate common-mode noise.

Analysis of individual-trial results for a 24-sentence and a 48-sentence experiment is reported in this letter. The experiments and averaged-trial results have been previously reported in Ref. [9]. In both experiments, English sentences on geography were presented visually word-by-word, with the onset of words set by the timing of the corresponding auditory words. Subjects had to evaluate the truth of each sentence by pressing one of two keys on the keyboard. For the 24-sentence experiment, with a few exceptions, the recordings were done using monopolar settings. For the 48-sentence experiment, bipolar recordings were made. In both experiments, there were 10 trials of each sentence for each subject, resulting in 240 trials for each subject of the 24-sentence experiment and 480 for the 48-sentence experiment. In the 24-sentence experiment, there were 7 subjects with monopolar settings, 3 with bipolar settings. There were 10 subjects for the 48-sentence experiment with bipolar settings.

2. Methods

ICA was originally proposed in the literature to solve the blind-source separation problem [3]. The key assumptions are that the sources are fixed in location, that the propagation between the sources and sensors are approximately instantaneous, and that the sensor-recorded signals are a linear mix of the sources. To introduce uniqueness to the unmixing, ICA also requires that the sources are statistically independent. Different ICA algorithms include FastICA [6], Infomax [1], and SOBI [2], each based on different principles of estimating this statistical independence. We used Infomax, implemented in EEGLAB [4], to decompose our recordings before applying a single-channel classifier (SCC). Infomax, which can be implemented as an unsupervised neural network processor, reduces statistical dependency by maximizing the mutual information that the output Y contains about the input X. The mapping between the electrode recordings and the independent components, called the unmixing matrix A, is computed by the algorithm. Independent components are obtained by computing $z = Ax$. More detailed discussions of ICA can be found in Ref. [1].

All trials were low-pass filtered to avoid aliasing before downsampling 16 times. This reduction was reasonable, as we were able to optimize classification in prior experiments [9] using bandpass filters with high cutoff frequencies of no more than 30 Hz. After downsampling and removal of low-frequency noise below 1 Hz, trials were split into three sets: training, validation and test. For both experiments, the 10

trials for each sentence were split 4/3/3. The downsampled time-series data were scaled to a range between −1 and +1 and used to train multioutput perceptrons, one for each channel. For linear perceptrons, training can be done simply by solving a set of linear equations. The number of outputs computed by each multioutput perceptron was equal to the number of classes, and a classification was made on the validation set by selecting the class with the maximum output. By selecting the best channel with the best classification rates, a best SCC is defined. Extending to a multichannel (MCC) was trivial. Instead of selecting the best channel, k best channels were chosen. Data from the corresponding channels were concatenated, resulting in longer input vectors to be fed into larger multioutput perceptrons. The optimal parameter k was determined by the validation set. In the case of the best ICA-component classifier (ICA SCC), trials in the training and validation sets were used to estimate the unmixing matrix A by ICA. The unmixing matrix was applied to all trials, projecting the trials onto the space of the approximately-independent components. We then applied the best single-channel classifer (SCC) on the scaled data with a range between −1 and +1.

For each permutation, the superset of the training and validation data after unmixing was used to compute the matrix Z, with each row being a single trial of the best independent component. For the ith perceptron, w_i was the weight vector and $\hat{y}_i$ the desired output vector for all trials in the matrix Z. Minimization was done for all the trials in Z in order to solve for the weights using Tikhonov Regularization [10], in which $\lambda^2 = 40$ for the subjects of both experiments. The regularization parameter λ was introduced to avoid overfitting; similar regularization methods are used in learning, e.g., neural networks [5,8]. We minimized the regularized objective function $G_\lambda(w_i) = \|Zw_i - \hat{y}\|^2 + \lambda^2\|w_i\|^2$.

For ICA, the unmixing matrix A was computed by the superset of the training and validation set, composed of 70% of the trials in either experiment. By inverting the unmixing matrix A, the mixing matrix A^{-1} was estimated. A scalp map, by projecting the corresponding column of A^{-1} onto the locations of the electrodes, is commonly used to show the spatial location of a component. For the best monopolar subject, S14, we identified two scalp maps, one for an obvious eyeblink and the other the best component in Fig. 1. We did not attempt to draw a conclusion about the actual location or the invariance across subjects, as both require the use of dipole fitting. However, the scalp map provided important evidence to weaken the hypothesis of

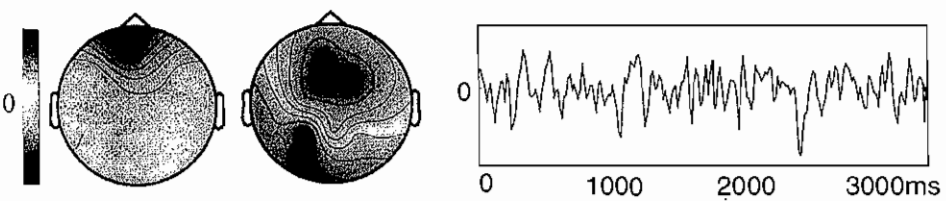

Fig. 1. Two scalp maps, the left one corresponding to an obvious eyeblink component and the right one to the best component. The best ICA component for a single trial of the first sentence is shown on the far right.

the best component being a by-product of an eyeblink. In Fig. 1, we also show the best ICA component of the first sentence.

3. Results

We show results using the three methods, SCC, MCC and ICA SCC, along with their significance in Table 1. Since the probability distribution of the joint classification results of the 10 permutations cannot be derived without further assumptions about the sampling dependency, we conservatively report the results as if they were based on a single permutation. The statistical significance of a result is commonly expressed by either the number of standard deviations from the mean under the null hypothesis or the statistical p-value given the null hypothesis. The probability, i.e., p-value, of the null hypothesis, which is that the observed probabilities are at chance level, is computed by $P(Y \geqslant k) = 1 - \sum_{j=0}^{k-1} \binom{n}{j} p^j q^{n-j}$,

Table 1
Significance for (a) monopolar and (b) bipolar settings. All the results shown for the monopolar setting are from the 24-sentence experiment

	Standard deviations			$p<10^{-n}$		
	SCC	MCC	ICA SCC	SCC	MCC	ICA SCC
(a) monopolar						
S10	5	8	6	−4	−8	−4
S11	9	9	15	−9	−9	−20
S12	12	12	15	−14	−14	−20
S13	5	7	10	−3	−6	−10
S14	9	9	17	−9	−9	−24
S15	2	4	7	−1	−3	−6
S19	9	10	15	−8	−10	−20
(b) bipolar						
S16	4	7	5	−3	−6	−3
S17	4	7	2	−3	−7	−2
S18	24	29	21	−40	−55	−32
S10	13	18	14	−14	−24	−17
S11	1	2	5	−1	−1	−4
S12	14	23	14	−16	−34	−17
S13	4	11	7	−3	−11	−6
S16	7	9	7	−6	−9	−6
S18	35	43	47	−65	−86	−97
S24	9	12	12	−9	−13	−13
S25	5	11	10	−4	−12	−11
S26	15	26	15	−18	−42	−18
S27	9	14	13	−9	−17	−16

The results for the bipolar setting are from the 48-sentence experiment, except the first three results (subjects S16, S17, S18) which are from the 24-sentence experiment.

where n is the number of test trials, k is the number of correct classifications and p is the chance probability (not the same p as the p-value). Here, the chance probability p is simply $1/N$ where N is the number of classes. For the 24- and 48-sentence experiments, $n = 72$ and 144, $N = 24$ and 48. In general, statisticians regard a p-value of 10^{-3} or smaller to be significant. Similarly, physicists usually consider at least 3 or 4 standard deviations to be a respectable result. Both measures are given in Table 1, but we discuss and summarize only the p-values.

Based on these standards, a majority of the results, as can be seen from Table 1, are highly significant. For monopolar settings (Table 1a) using SCC, 6 of the 7 subjects had a p-value better than 10^{-3}, i.e. $p < 10^{-3}$, and 4 of these 6 subjects had p-values $p < 10^{-8}$. Results are slightly improved by using MCC, and much further with ICA SCC. With ICA SCC, all subjects had p-values $p < 10^{-4}$, and 4 subjects had p-values $p < 10^{-20}$.

For bipolar settings (Table 1b) with SCC, 12 out of 13 experimental sessions (some with the same subject, i.e., subjects S16 and S18) had p-values better than 10^{-3}. The two best cases with SCC were in fact different sessions of the same subject S18, with $p < 10^{-40}$ and 10^{-65}, respectively. The levels of significance were improved by using either MCC or ICA SCC. With MCC, 12 out of the 13 experimental sessions achieved p-values better than 10^{-6}, with 5 cases obtaining p-values better than 10^{-20}. The best p-value achieved with MCC is $p < 10^{-86}$. With ICA SCC, improvements in terms of significance are generally not as good as those for MCC. On the other hand, for S18, ICA SCC, $p < 10^{-97}$, which is the best in the table.

In addition to the levels of significance, the achieved classification rates are shown in Fig. 2. The best classifications rates for both experiments were achieved by bipolar settings. In the 24-sentence experiment, MCC for S18 achieved the best recognition of 72% correct with a chance level of $\frac{1}{24}$. In the 48-sentence experiment, ICA SCC for the same subject S18 achieved the best recognition of 58% correct with a chance level

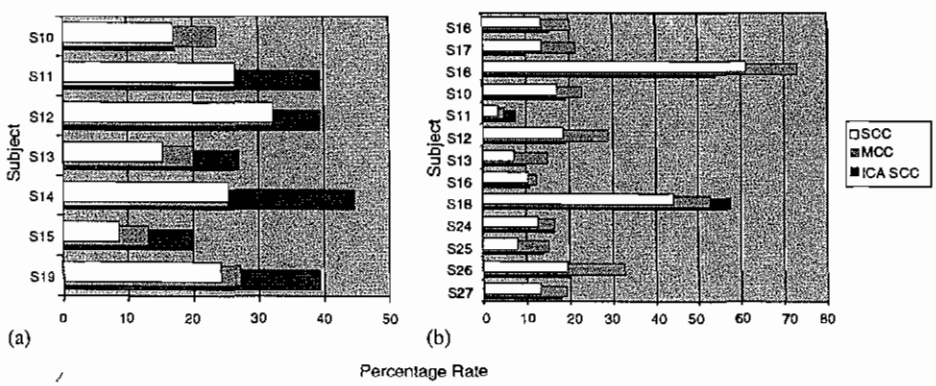

Fig. 2. Percent classification rates of individual trials for a single-channel classifier (SCC), a multichannel classifier (MCC) and a single-component classifier (ICA SCC) for (a) monopolar and (b) bipolar settings. All the results shown for the monopolar setting are from the 24-sentence experiment. The results for the bipolar setting are from the 48-sentence experiment, except the first three results (S16, S17, S18) which are from the 24-sentence experiment.

of $\frac{1}{48}$. Overall, these histograms show that ICA SCC outperforms MCC for monopolar settings, while MCC is a slightly better scheme when bipolar settings are used.

References

[1] A.J. Bell, T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (1995) 1129–1159.
[2] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, E. Moulines, A blind source separation technique using second-order statistics, IEEE Trans. Signal Process. 45 (1997) 434–444.
[3] P. Comon, Independent component analysis, a new conccpt?, Signal Process. 36 (1994) 287–314.
[4] A. Delorme, S. Makeig, EEGLAB: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis, J. Neurosci. Meth. 134 (2004) 9–21.
[5] F. Girosi, M. Jones, T. Poggio, Regularization theory and neural networks architectures, Neural Comput. 7 (1995) 219–269.
[6] A. Hyvarinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Comput. 9 (1997) 1483–1492.
[7] H.H. Jasper, The ten–twenty electrode placement of the international federation, Electroencephalogr. Clin. Neurophysiol. 10 (1958) 371–375.
[8] D.J.C. MacKay, Bayesian interpolation, Neural Comput. 4 (1992) 415–447.
[9] P. Suppes, B. Han, J. Epelboim, Z.-L. Lu, Invariance between subjects of brain wave representations of language, Proc. US. Nat. Acad. Sci. 96 (1999) 12953–12958 available on the websitc: www.pnas.org.
[10] A.N. Tikhonov, V.Y. Arsenin, Solutions of Ill-posed Problems, V H Winston and Sons, Washington D.C., 1977.