Feature Selection and Classification in Brain Computer Interfaces by a Genetic Algorithm

Luca Citi[1], Riccardo Poli[2], Caterina Cinel[3], Francisco Sepulveda[2]

[1]Department of Electronic Engineering, University of Florence, Italy
[2] Department of Computer Science, University of Essex, UK
[3]Department of Psychology, University of Essex, UK

Abstract. In this paper we explore the use of evolutionary algorithms in a wrapper-based selection of features and the classification of P300 signals in Brain Computer Interfaces. In particular we focus on a paradigm that uses the P300 potential associated to particular visual stimuli for hands free text entering. In our experiments the GA has found new ways to process and combine EEG signals to improve P300 detection accuracy.

1. Introduction

Brain Computer Interfaces (BCIs) can be divided into dependent and independent types [1]. In the former, activity in the various motor pathways is needed for generating the EEG signals that will carry information pertaining to a given task (see, e.g., [2]), whereas in the latter, relevant EEG will arise regardless of the activity pattern in motor pathways. Within the independent BCI realm, P300 potentials have provided a relatively robust means to detect user's intentions concerning the choice of objects within a visual field. To this end, Donchin and others [3, 4] have developed a protocol whereby a subject is shown a matrix of characters or symbols, where rows and columns flash periodically in random order (see, for example, Fig. 1). Large P300 potentials are then observed only in response to the matrix element the subject has chosen, regardless of where the gaze is directed. Matrix size effects on the P300 amplitude potential have been recently investigated as well [5].

In the present study, we aimed at simultaneously selecting P300-based features and discovering classification technique for maximized recognition performance. The setup was as described in [4].

2. Methods

In our work we used the 2nd Wadsworth BCI Dataset from the BCI2003 competition [6]. This contains three sessions recorded using the paradigm described in [4]. We used the 19 standard channels of the 10-20 system.

Our objective was to maximally emphasize the P300 signal w.r.t. to background noise and other potentials for the purpose of brain-activity based dictation of characters. In order to achieve this we applied two pre-processing stages. The first stage consisted in extracting a one second epoch starting from the stimulus, applying a 30th order lowpass FIR filter (F_{pass} = 34 Hz, F_{stop} = 47 Hz, $W_{\text{pass}} = W_{\text{stop}} = 1$) and skipping every other sample. We then applied the rbio3.3 Continuous Wavelet Transform (CWT) to every channel using 30 different scales from within the range [2,40]. CWT was chosen because the base functions have similarities with the typical shapes of the P300 complex. We then kept the 40 samples between 270ms and 590ms obtaining a 19×30×40 matrix of features **V**.

Naturally, **V** represents an enormous number of features, which could trouble even the best classification techniques. So, a feature selection stage is required. We used a *wrapper approach* to feature selection and classification [7] where a subset of the features is selected, a classifier is realized, its performance evaluated and the process is iterated until both the features and the classifier are sufficiently good (this is different from a *filter approach* where the subset of features is optimized separately from the classifier). In our approach we used a Genetic Algorithm (GA) [8] to perform this joint optimization of features and classifier. In order to allow the exploitation of both linear and non-linear relationships between the features, we used a *polynomial classifier* where a subset of the features are combined in a polynomial of the form

$$P(\mathbf{V}) = a_0 + \sum_{h=1}^{N} a_h \prod_{k=1}^{M} \mathbf{V}(c_{h,k}, s_{h,k}, t_{h,k})^{e_{h,k}}$$

where: a_h are coefficients; $c_{h,k}$, $s_{h,k}$, $t_{h,k}$ are the channel, scale and time indexes of a feature in the matrix **V**; and $e_{h,k}$ are integers in {-3,-2,-1,0,1,2,+3}. The output of the polynomial was squashed in interval [-1,1]. If the result was greater than a threshold σ the trial was classified as target. By allowing a GA to optimize both the real-valued coefficients a_h and the $N \times M$ integer matrices $c_{h,k}$, $s_{h,k}$, $t_{h,k}$ and $e_{h,k}$ we effectively performed the feature selection and the classifier optimization stages jointly.

We used blend crossover (where the value of the offspring parameters is the result of interpolating the parents' parameters) to perform the search. Parents were chosen by tournament selection. Mutation was implemented as crossover between an individual from the population and a randomly generated one. The objective function was the mean (over all the trials in the training set) of the square of the difference between the squashed output of the polynomial and the correct output. The population size was 20,000.

To test the generalization of the system we used 5-fold cross validation using 4 of the 5 runs of session 10 of the dataset as training set (selecting all target trials and choosing randomly the same number of non-targets) and the other run as validation set (using all trials).

3. Results

In most runs the GA evolved (near-) linear classifiers.[1] There can be two reasons for this: a) linear classifiers perform better or b) linear terms are easier to discover.[2] Since all our efforts to evolve non-linear components failed, we believe the first explanation is more likely.

When we set N=2 we obtained equations like

$$P(\mathbf{V}) = -0.335 - 0.159 \cdot \mathbf{V}(16,15,11) + 0.100 \cdot \mathbf{V}(10,15,11)$$

This classifies a trial as target if the weighted difference between channels T6 and C4 of the correlation with the mother wavelet stretched 17 times and shifted by approximately 380ms is greater than σ–0.335 (with σ=0, TP=0.77 and FP=0.24 on validation set). As CWT is linear, the equation can be seen as calculating the correlation between the weighted difference between the two channels and the mother wavelet stretched and shifted. This suggests that the difference between T6 and C4 is important for the purpose of P300 detection.

Fig.2 shows the signals recorded in T6 (bottom) and C4 (top) in the presence (solid line) and in the absence of P300 (to reduce the noise, plots are averages over multiple trials[3]). Fig. 3 shows the weighted difference between these signals (top) as well as the appropriately stretched and scaled wavelet mentioned above (bottom). The non-target plots for C4 and T6 are very similar (and in-phase). On the contrary the target plots are quite different. So, subtraction tends to cancel the non-target signal and to enhance the target one: exactly what we need for a reliable detection of the P300. When a P300 is present, the signal resulting from the subtraction has a shape similar to the wavelet in Fig. 3, so convolution with it further strengthens our classifier.

Table 1 shows the results obtained with a 5-fold cross-validation for polynomials with N=3 and N=4 linear terms. The value of σ can be used to trade true positives (TP) for false positives (FP). We tested two criteria to set σ optimally: *a)* the maximum rate of correct outputs (MaxCorr); *b)* the maximum mutual channel information (MaxInfo)

$$I(S,R) = H(S) - H(S/R) = \sum_{s_i \in S} \sum_{r_j \in R} P\{s_i, r_j\} \lg_2 \frac{P\{s_i, r_j\}}{P\{s_i\} P\{r_j\}}$$

where S is a stimulus on the screen and output R the response provided by the detector. In both cases we set Pr{S=target}=1/6 because 1/6 is the target frequency in the Donchin speller paradigm [4].

From the table we can see that the rate of correct classification for our classifiers is up to 87.62%, which compares well with the results reported by others on similar datasets. It is interesting to note that the MaxCorr criterion favors specificity excessively, as clearly shown by the fact that $I(S,R)$ is significantly reduced w.r.t. the

[1] Linear terms are obtained when in a term of the polynomial a factor has exponent 1 and all others have exponent 0. Since this is a complex configuration, to help evolution we later added a pure a linear part to the general polynomial.

[2] A second order term, for example, can lead to a very big product that needs to be paired with a small coefficient.

[3] We averaged all the trials in the dataset where either exactly one (solid lines) or exactly zero (dashed lines) P300 potentials where present within the time interval shown.

maximum achievable (e.g. 0.146 vs. 0.163). We can also see that the use of four features improves $I(S,R)$.

4. Discussion

The results reported in Table 1 are quite encouraging [10,11,12], but these are still a far cry from what we need to achieve fast and reliable brain-computer interfaces. These would have a huge number of potential applications, particularly in the area of communication aids for people with severe motor disabilities.

Why can we not do better? Certainly one reason is the enormous amount of noise and variability present in EEG signals. For example, these are extremely small in amplitude, their acquisition requires a good contact between skin and electrodes, which is very hard to achieve on all electrodes, and a considerable amplification. Also, muscular noise (e.g. the blinking of an eye, swallowing, etc.) can completely cover brain activity. However, we believe these are not the only reasons why the accuracy and reliability of BCI detection systems cannot be improved beyond a certain limit: some perceptual phenomena including attentional blink, repetition blindness and other effects caused by attentional limits can interfere with character identification in Farwell and Donchin's P300-based speller paradigm [3]. In particular in our experiments we have found evidence for the existence of "near targets".

To study the possible influence of perceptual errors we first split the signals into (partially overlapping) trials lasting 1s and starting from a stimulus (the flashing of a row or a column of letters on the screen). We then grouped the trials in the dataset into 12 classes on the basis of which of the 6 rows and 6 columns flashed. The trials representing each row (column) where further subdivided on the basis of the row (column) of the target chosen by the subject, thereby producing 12×6=72 classes. In order to reduce noise we averaged the trials within each class.[4] We concentrated our analysis on both the signals recorded in the Cz channel and the weighted difference $\Delta=0.1\times C4-0.159\times T6$ between channels T6 and C4, which, in our experiments, we had found to have high significance for the purpose of P300 detection.

What should one expect to see when plotting the averaged signals for each class? In theory, out of all the trials where column *c* flashed, only those where column *c* actually contained the target should present a P300 and likewise for the rows. Indeed this is what we observed, as illustrated in Fig. 4, where the *red* plot represents the Δ signal averaged over all the trials where the row containing the target flashed, which confirms the presence of P300s (similar results were obtained for Cz). However, the *green* plot, which represents the Δ signal averaged over all the trials where the row that flashed was *adjacent* to the one containing the target, differs significantly from the remaining plots (representing situations where the target was further away from

[4] Different rows and columns had different numbers of stimuli, namely: 180, 165, 75, 165, 0 and 0 (top to bottom) for the rows, and 150, 150, 105, 30, 75 and 75 (left to right) for the columns.

the flashing column), effectively presenting a large P300-like wave peaked at 300ms.[5]

What generated these P300-like waves in the presence of *near-target* stimuli? A plausible explanation is to attribute these to attentional orienting mechanisms [9], where the subject's visual system, being unable to focus attention only on precisely the target letter, generated P300 (surprise, attentional orienting) signals. We suspect these perceptual errors may be a reason for the limited single-trial performance shown by automated P300 detectors. This, of course, drammatically increases the number of repetitions needed for reliable recognition.

However, as Fig. 4 suggests, EEG signals may contain information regarding the *degree of targetness* of stimuli. So, spurious P300s will not necessarily work against BCI if, in the future, we will be able to exploit this information.

5. Conclusions

In this paper we have explored the use of evolutionary algorithms to aid the selection of features and the classification of P300 signals in BCI. This approach has confirmed the usefulness of linear detectors, while at the same time revealing the importance of selecting certain EEG channels and using their differences to cancel non-P300 components. The evolved classifiers have shown state-of-the-art performance.

Bibliography

[1] Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain-computer interfaces for communication and control. Clin. Neurophys. 2002; 113:767-791.

[2] Sutter EE. The brain response interface: communication through visually induced electrical brain responses. J Microcomput. Appl. 1992;15:31–45.

[3] Farwell LA, Donchin E. Talking off the top of your head: toward a mental prothesis utilizing event-related brain potentials. Electroenceph. Clin. Neurophysiol. 1988; 70:510–523.

[4] Donchin E, Spencer KM, Wijesinghe R. The mental prosthesis: Assessing the speed of a P300-based brain–computer interface. IEEE Trans. Rehab. Eng. 2000; 8:174–179.

[5] Allison BZ, Pineda JA. ERPs evoked by different matrix sizes: Implications for a brain computer interface (BCI) system. IEEE Trans. Neur. Sys. Rehab. Eng. 2003; 11(2):110-113.

[6] Documentation 2nd Wadsworth BCI Dataset http://ida.first.fraunhofer.de/projects/bci/competition/albany_desc/albany_desc_ii.pdf

[7] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence 97(1-2) (1997), 273–324.

[8] Goldberg DE. Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reading, 1989.

[9] Posner, MI, Snyder, CRR., Davidson, BJ. Attention and the detection of signals. *Journal of*

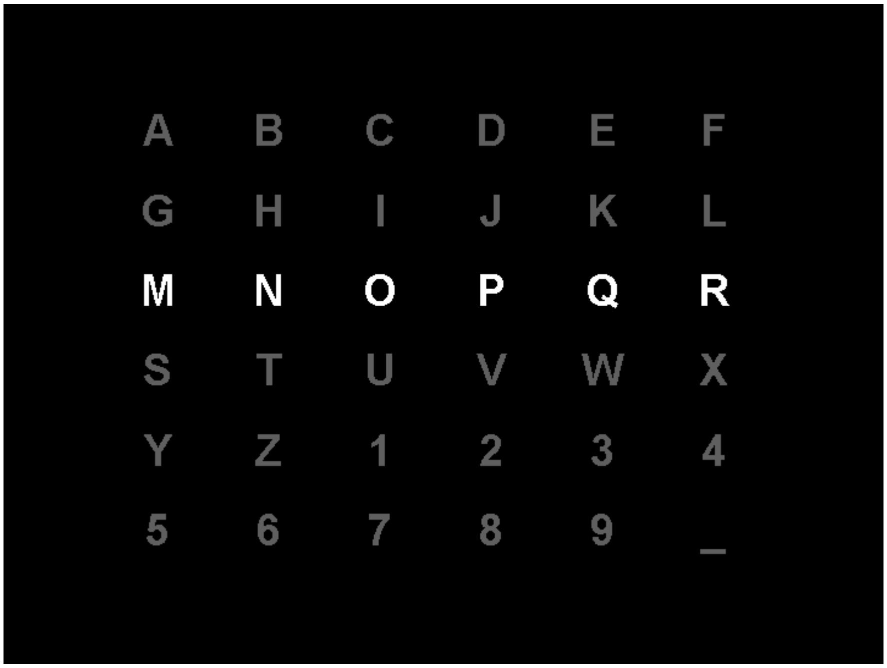
[5] Stimuli at distance 2 (*blue* plot) seem to contain less information on the whereabouts of the target.

Experimental Psychology: General: 1980: 109:160-174.

[10] Bayliss, JD. A flexible brain-computer interface. PhD thesis, Department of Computer Science, University of Rochester, 2001.

[11] Meinicke, P., Kaper, M., Hoppe, F., Heumann, M., and Ritter. H. Improving Transfer Rates in Brain Computer Interfacing: a Case Study. Neural Information Processing Systems (NIPS) 2002.

[12] Beverina, F., Palmas, G., Silvoni, S., Piccione, F., and Giove, S. User adaptive BCIs: SSVEP and P300 based interfaces. PsychNology Journal: 1(4):331-354, 2003.



A	B	C	D	E	F
G	H	I	J	K	L
M	**N**	**O**	**P**	**Q**	**R**
S	T	U	V	W	X
Y	Z	1	2	3	4
5	6	7	8	9	_

Figure 1. Example of display used in Donchin's P300-based speller paradigm.

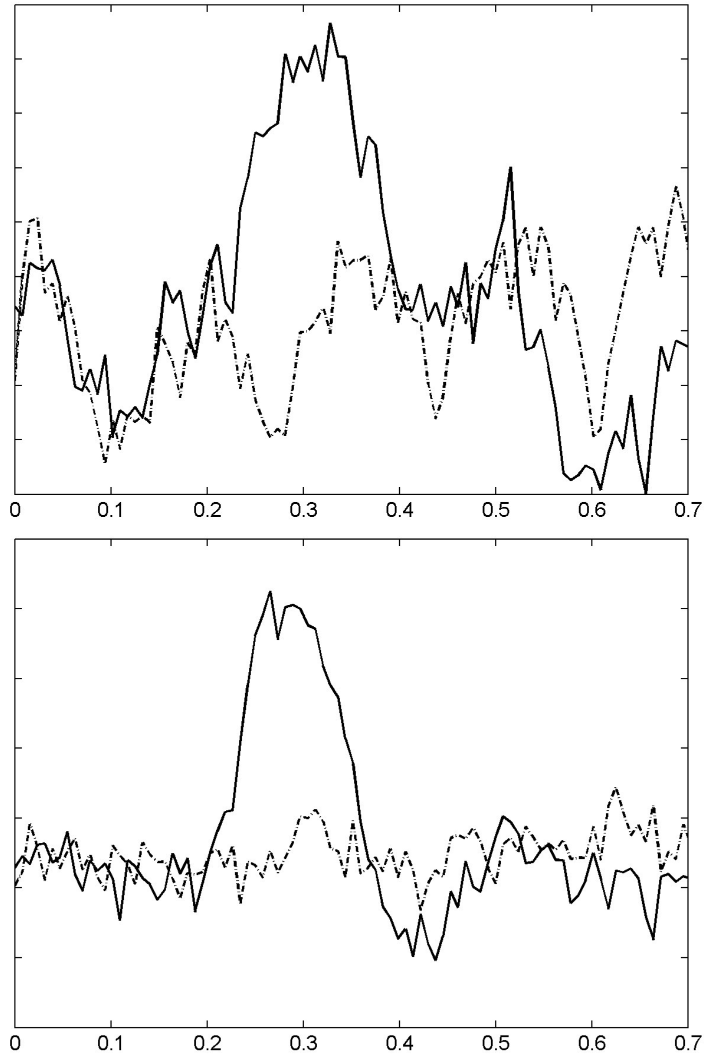


Figure 2. Average signals recorded in T6 (bottom) and C4 (top) in the presence (solid lines) and in the absence (dashed lines) of P300.

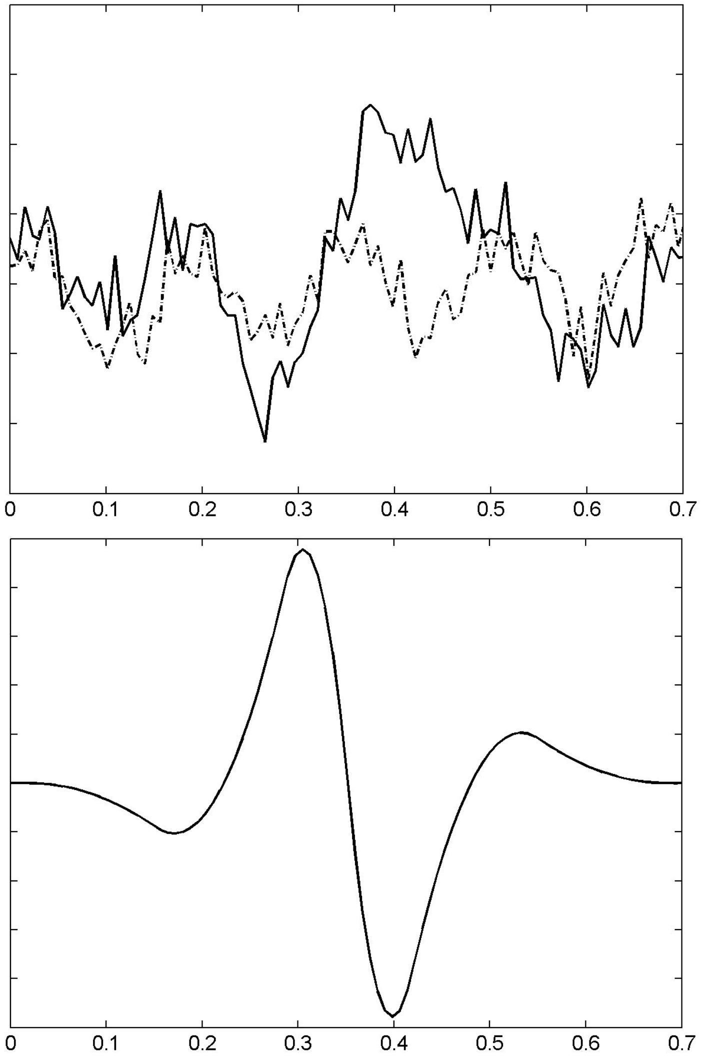


Figure 3. Average weighted difference between the C4 and T6 signals (top) and the mother wavelet stretched 17 times and shifted by approximately 380ms (bottom) which was selected by the GA for optimal P300 detection.

Table 1. Results obtained by the GA with a 5-fold cross-validation for polynomials with *N*=3 and *N*=4 linear terms

		***N*=3**		***N*=4**	
		MaxCorr	***MaxInfo***	***MaxCorr***	***MaxInfo***
TP	mean	51.63%	65.61%	51.82%	70.58%
	std	6.89%	6.35%	5.58%	6.94%
FP	mean	6.00%	11.94%	5.22%	13.15%
	std	1.93%	4.64%	1.64%	3.99%
Correct	mean	86.94%	84.32%	87.62%	84.14%
	std	1.33%	2.99%	1.17%	2.45%
***I*(*S*,*R*)**	mean	0.137	0.148	0.146	0.163
	std	0.026	0.016	0.023	0.018

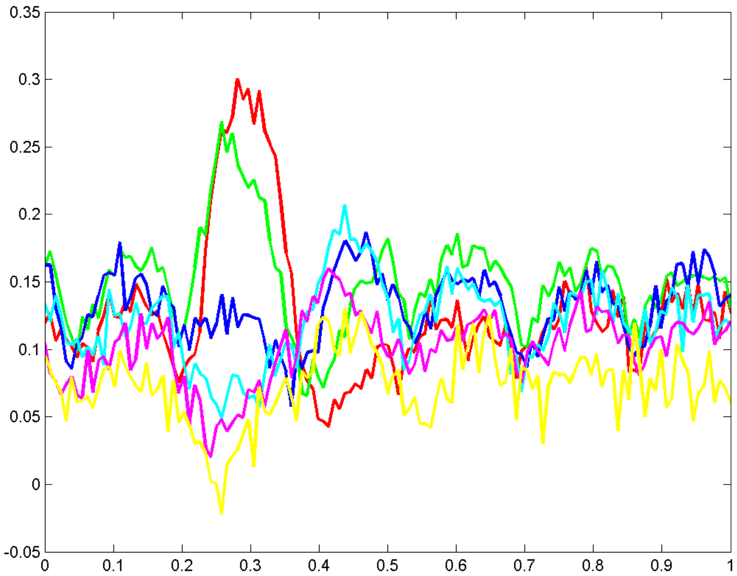


Figure 4. Average P300 signals recorded for target (red plot), near-target (green plot) and non-target (remaining plots) signals.