AN EVOLUTIONARY APPROACH TO FEATURE SELECTION AND CLASSIFICATION IN P300-BASED BCI

Luca Citi[1], Riccardo Poli[2], Francisco Sepulveda[2]

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

SUMMARY
We explore the use of evolutionary algorithms in the selection of features and the classification of P300 signals in BCI. As a result we have found new ways to process and combine EEG signals to improve detection.

INTRODUCTION
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 that flash periodically (in groups). 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].

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 evoked 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_{pass} = W_{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×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).

RESULTS AND DISCUSSION
In most runs the GA evolved (near-)linear classifiers.[1]

[1] Linear terms are obtained when in a term of the polynomial a factor has exponent 1 and all others have exponent 0. Since

There can be two reasons for this: a) linear classifiers perform better or b) linear terms are easier to discover.[2] Since all our effort 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.1 shows the signals recorded in T6 (lower left) and C4 (upper left) in the presence (solid line) and in the absence of P300 (to reduce the noise, plots are averages over multiple trials), their weighted difference (upper right) and an appropriately stretched and scaled wavelet. 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. 1, so convolution with it further strengthen 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 maximum achievable (e.g. 0.146 vs. 0.163). We can also see that 4 features improve $I(S,R)$.

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.

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.

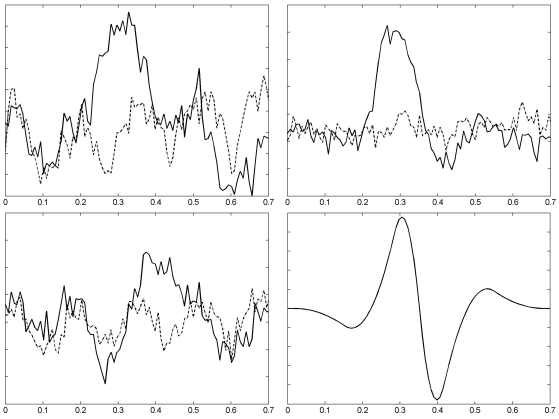

Figure 1.

REFERENCES

[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.

Table 1.

		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