ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

Exploring the impact of alternative feature representations on BCI classification

A. Bahramisharif, M. van Gerven, and T. Heskes*

Radboud University Nijmegen – Institute for Computing and Information Sciences
Heyendaalseweg 135, 6525 AJ, Nijmegen – The Netherlands

Abstract. Classification performance in BCIs depends heavily on the features that are used as input to the employed classifier. If the BCI signal is extended in time, we may either use a representation of the signal at multiple time segments with a high risk of overfitting or averaged over time with a high risk of underfitting as input to the classifier. In this paper we present an empirical study which allows us to determine the right balance between these two representations. Using two BCI data sets, we show that our method can significantly improve classification performance.

1 Introduction

Brain computer interfacing (BCI) is a general term that is used when we have a direct connection between the brain and a computer [1]. In other words, the subject is given a task and the computer predicts task conditions based on measurements of brain activity recorded by various instruments. In these setups, we have to deal with a huge data set consisting of a multitude of noisy signals. A typical representation of such signals in (EEG-based) BCIs is the power spectrum, computed for multiple time segments and over multiple sensors.

The question we wish to address in this paper is how to make optimal use of the information contained within the power spectra as the signal evolves over time. From signal processing, it is known that if the measured signal is stationary then averaging over multiple time segments results in higher signal to noise ratio. This averaging method can also be applied in the frequency domain, in which case it is known as Bartlett's method [2]. This will reduce the variance of the periodogram at the price of reducing the time resolution. Averaging is the dominant approach in BCI data analysis (see e.g., [3]). Note that if the assumption of stationarity is violated, averaging may not be optimal and we may wish to use other approaches. Especially in BCI, the stationarity assumption is troublesome since brain activity may be highly non-stationary within a given trial. In that case, it could be wise to use the power spectra at individual time segments as input to a classifier. However, this leads to a much larger number of parameters in the classification with a high risk of overfitting.

The question then becomes which of both representations performs better in the context of BCI. To this end, we vary smoothly between using the segments as independent features on one hand and by averaging over the segments

*The authors gratefully acknowledge the support of the BrainGain Smart Mix Programme of the Netherlands Ministry of Economic Affairs and the Netherlands Ministry of Education, Culture and Science, the STW grant NET.7050 and the Vici grant 639.023.604.

on the other hand. This can be achieved by means of *regularization of differences*, which was used as a regularizer for logistic regression. The regularized logistic regression model was used for the classification of BCI data obtained for two distinct BCI paradigms. The first dataset consists of EEG data collected for subjects engaging in an imagined movement paradigm, which currently is one of the most often used BCI paradigms [4]. The second dataset consists of MEG data collected for subjects engaging in a relatively new covert attention paradigm [5]. We show that one can gradually move in the direction of concatenation, without loosing performance or, in some cases, obtaining even better classification performance.

The structure of the paper is as follows. In Section 2 we motivate and define the regularizer and describe how it is applied to logistic regression. In Section 3 the employed datasets are described in more detail. Experimental results for the BCI datasets are shown in Section 4. Conclusions follow in Section 5.

2 Methods

A signal is called stationary when the joint probability distribution of the signal does not vary over time. In other words, having a stationary signal and dividing it into segments, the mean and variance of each segment are the same as those of the whole signal. For stationary signals, it has been shown that averaging over segments results in higher signal to noise ratio [2]. Hence, our strategy would be to concatenate time segments to obtain one big feature vector when the signal is non-stationary and to average over the segments when the signal is stationary. We refer to both approaches as *concatenation* and *averaging* respectively.

In real life, most of the time we do not have exact knowledge about the signal. For instance, using a particular BCI paradigm, we do not know exactly what goes on in the brain of the subject. Is he or she using the same strategy over time? Is the brain responding exactly the same over time? Typically, it is unknown whether averaging, concatenation, or some weighted combination of both would result in the highest signal to noise ratio. To overcome this dilemma, we introduce a regularizer which allows us to vary smoothly between concatenation and averaging. This regularizer, called *regularization of differences*, is used in this paper as a penalization term for logistic regression. Logistic regression has been used before in BCI [7, 8] and solves a classification problem by expressing the probability of class membership as:

$$p(c \mid \boldsymbol{x}, \boldsymbol{\theta}) \propto \exp(\boldsymbol{\theta}_c^T \boldsymbol{x})$$

where $\boldsymbol{x} = (x_1, ..., x_M)$ is the set of features (including a bias term) and $\boldsymbol{\theta}_c$ is the parameter vector associated with class $c \in \{1, \dots, C\}$. A logistic regression model is trained by minimizing the following *loss function*:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \left(\log(\sum_{c=1}^{C} \exp(\boldsymbol{\theta}_c^T \boldsymbol{x}_n)) - \boldsymbol{\theta}_{c_n}^T \boldsymbol{x}_n \right)$$

which is written as a function of $\boldsymbol{\theta}$ because data $D = \{(\boldsymbol{x}_n, c_n)\}_{n=1}^{N}$ is assumed to be constant during training.

Often, the minimization is constrained by adding a regularizer $R(\boldsymbol{\theta})$; a well known example is Tikhonov regularization [9]. We define regularization of differences as a measure of differences between weights of features of different segments of the signal. Assume that we divide the signal into S segments, such that $\boldsymbol{\theta}_c^s$ corresponds to parameter vector associated with class c and time segment s. Then, the regularizer can be written in the following form:

$$R(\boldsymbol{\theta}) = \sum_{c=1}^{C} \sum_{s=1}^{S} \sum_{t=1}^{S} ||\boldsymbol{\theta}_c^s - \boldsymbol{\theta}_c^t||^2$$

where $|| \cdot ||$ is the Euclidean norm. Note that we could also use other distances instead of Euclidean, such as in Lasso [10] or Elastic Net [11] regularizers. However as the focus of the regularizer is to force the weights of different segments to be the same, there would be no major difference in the result when using different norms. The regularizer can easily be used with any classifier whose loss function is convex, guaranteeing a unique solution. In this paper, we use the regularizer in conjunction with logistic regression. The goal of training the classifier would be to minimize the *objective function*:

$$F(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \lambda R(\boldsymbol{\theta})$$

with regularization parameter λ. Large values of λ force parameters belonging to different segments to be the same, resembling averaging. Conversely, low values of λ results in treating each segment independently, resembling concatenation. The method is made available through the FieldTrip classification module[1].

3 BCI datasets

In this section, we describe the two paradigms which were used in the analysis.

3.1 Imagined movement

We conducted an experiment with ten subjects based on an imagined movement paradigm described in [6]. In this paper we restrict ourselves to the best five subjects according to previous results. We used multichannel EEG data recorded from 28 Ag/AgCl electrodes placed on the scalp according to the international 10-20 system using a left mastoid reference. The sampling frequency was 128 Hz. There were three tasks: right-hand imagined movement, left-hand imagined movement, and no imagined movement. Each subject performed around 60 trials for each task, each of which took 6 seconds (2 seconds before and 4 seconds after cue onset). For some of the subjects the data is quite unbalanced which will be taken into account in the analysis. Based on prior research [6, 4], we know that task-related activity takes place in the alpha band over the motor

[1] http://www.ru.nl/fcdonders/fieldtrip/modules

cortex. Therefore we filtered the data and used the power spectrum computed for 1 Hz frequency bands between 8 Hz and 14 Hz for EEG channels *C3*, *C4* and *Cz*. In order to prevent the influence of evoked response, we restricted ourselves to the 1–4 second interval after cue onset. The last thing to consider is how to choose the number of time segments and how to compute the power spectrum for segment. We used fixed-length time windows and chose to divide the three second interval into 23 segments, each of which has 50% overlap with its neighbors. To compensate for the effect of outliers we normalized the data to have zero mean and a standard deviation of one.

3.2 Covert attention

In the covert attention experiment, fifteen subjects had to covertly attend for 2500 ms to different locations in the visual field while maintaining gaze at a fixation cross. Again, we analyzed the results for the best five subjects only and focus on left versus right covert attention. Data was recorded using an MEG system which provides whole-head coverage using 275 gradiometers. For each subject, we collected about 128 trials per condition. Data was detrended and downsampled from 1200 Hz to 300 Hz. Based on prior knowledge [5], we used only 41 occipital sensors and focused again on the 8–14 Hz alpha band. We discarded the first 500 ms of the attention period to prevent the effect of evoked response and divided the remaining two second attention period into 15 segments each of which has 50% overlap with its neighbors. Again, data was normalized to have zero mean and a standard deviation of one.

4 Experimental results

In order to get insight into how classification performance depends on the feature representation, we computed the accuracy (proportion of correctly classified cases) using regularized logistic regression for different values of the regularization parameter λ, based on a five-fold cross-validation scheme [12]. For each subject we varied λ from 10^{-14} to 10^{9}, as the optimal value of λ is subject specific. The change in accuracy as a function of λ is shown in Fig. 1 for the imagined movement and covert attention datasets.

Figure 1 demonstrates that, for the datasets used, averaging seems to be better than concatenation. Furthermore, standard deviation of the accuracy for the covert attention paradigm is higher than that of the imagined movement paradigm, indicating a larger between-subject variability. Accuracy is lower for the imagined movement paradigm since this is a ternary instead of a binary classification problem (left, right and no movement). Results also show that regularization of differences is able to interpolate between the two extremes of concatenation and averaging and can thus in principle be used to select an optimal setting of the parameters. However, in order to pick the optimal λ we need both an *inner* and *outer* cross-validation. Specifically, we kept one fold out of five for testing and train based on the remaining four folds. Then we performed five-fold cross-validation using just the training data and calculated

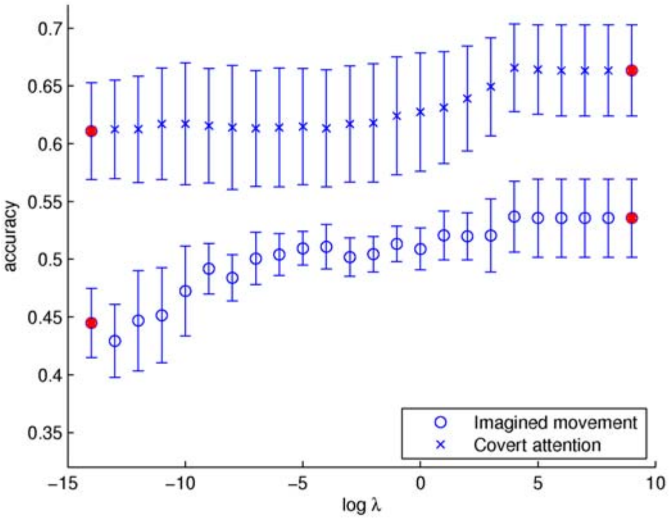

Fig. 1: The average change in accuracy as a function of λ for the imagined movement and covert attention datasets. Error bars show standard deviation. Results obtained using averaging and concatenation are denoted by filled circles.

the log-likelihood for each λ from 10^{-14} to 10^{9}. We used the smallest value of λ which gave the largest log-likelihood and retrained the classifier using all training data in order to test the classifier on the test data.

	Imagined movement				Covert attention		
Subject	Conc.	Avg.	Reg.	Subject	Conc.	Avg.	Reg.
1	0.38	0.46	0.48	6	**0.72**	**0.77**	**0.77**
2	0.40	**0.46**	**0.46**	7	**0.64**	**0.72**	**0.71**
3	**0.52**	**0.57**	**0.58**	8	0.53	**0.63**	**0.63**
4	**0.49**	**0.59**	**0.58**	9	0.52	**0.61**	**0.61**
5	0.44	**0.60**	**0.61**	10	**0.64**	0.59	**0.62**
Avg.	0.45	0.54	0.54	Avg.	0.61	0.66	0.67

Table 1: Classification performance using different strategies for the imagined movement and covert attention datasets. The bold ones are significantly better than assigning all data to the majority class (one-sided binomial test, $p = 0.05$).

The accuracy obtained using concatenation (lowest value of λ), averaging (highest value of λ), and regularization of differences for each subject using inner and outer cross-validation is given in Table 1. Significance levels were computed by comparing performance with a classifier that assigns each trial to the majority class using a one-sided binomial test [13]. Regularization of differences outperforms both concatenation and averaging in terms of significance (lower p-values on average) and average accuracy.

5 Conclusions

We used regularization of differences as a regularizer in the classification of BCI data based on logistic regression. In general, averaging tends to outperform concatenation, presumably due to the fact that signal to noise ratio increases by averaging over consecutive segments. However, this is afforded by the employed BCI paradigms since the signal of interest is relatively stable over a prolonged period (i.e., repetitive imagined movement and sustained covert attention). Nevertheless, even for the used datasets averaging is not always the best strategy for every subject. For example, for the covert attention dataset, concatenation gave a better performance for Subject 10. Using regularization of differences we are able to determine the right balance between concatenation and averaging, improving classification performance on average.

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(6):767-791, Elsevier, 2002.

[2] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*, Upper Saddle River, Prentice-Hall, 1996.

[3] T. N. Lal, M. Schröder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer and B. Schölkopf, Support vector channel selection in BCI, *IEEE Transactions on Biomedical Engineering*, 51(6):1003-1010, 2004.

[4] G. Pfurtscheller, C. Brunner, A. Schlögl and F. H. Lopes da Silva, Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks, *NeuroImage*, 31:153-159, Elsevier, 2006.

[5] S. P. Kelley, E. Lalor, R. B. Reilly and J. J. Foxe., Independent brain computer interface control using visual spatial attention-dependent modulations of parieto-occipital alpha, *Proceedings of the 2nd International IEEE EMBS Conference on Neural Engineering*, pages 667-670, 2005.

[6] G. Pfurtscheller and C. Neuper, Motor imagery activates primary sensorimotor area in humans, *Neuroscience letters*, 239(2-3):65-68, Elsevier, 1997.

[7] W. D. Penny, S. J. Roberts, E. A. Curran and M. J. Stokes, EEG-based communication: a pattern recognition approach, *IEEE Transactions on Rehabilitation Engineering*, 8(2):214-215, 2000.

[8] R. Tomioka, K. Aihara and K.-R. Müller, Logistic regression for single trial EEG classification, *Advances in Neural Information Processing Systems*, 19:1377-1384, MIT Press, 2007.

[9] A. N. Tikhonov, On the stability of inverse problems, *Doklady Akademii Nauk SSSR*, 39(5):195-198, 1943.

[10] R. Tibshirani, Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society Series B*, 58:267-288, Blackwell, 1996.

[11] H. Zou and T. Hastie, Regularization and variable selection via the Elastic Net, *Journal of the Royal Statistical Society Series B*, 67(2):301-320, Blackwell, 2005.

[12] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection,*Proceedings of the 14th International Joint Conference on Artificial Intelligence* (IJCAI), pages 1137-1145, 1995.

[13] S. L. Salzberg, On comparing classifiers: Pitfalls to avoid and a recommended approach, *Data Mining and Knowledge Discovery*, 1:317-327, Springer, 1997.