MULTITASK LEARNING APPLIED TO SPATIAL FILTERING IN MOTOR IMAGERY BCI
A Preliminary Offline Study

Dieter Devlaminck, Bart Wyns
Electrical Energy, Systems and Automation, Ghent University, Technologiepark 913, 9052 Zwijnaarde - Gent, Belgium

Georges Otte
P.C. Dr. Guislain, Fr. Ferrerlaan 88A, 9000 Gent, Belgium

Patrick Santens
Department of Neurology, Ghent University Hospital, De Pintelaan 185, 9000 Gent, Belgium

Keywords: Multi-subject learning, Common Spatial Patterns (CSP), Brain-Computer Interfaces (BCI).

Abstract: Motor imagery based brain-computer interfaces (BCI) commonly use the common spatial pattern filter (CSP) as preprocessing step before feature extraction and classifiction. The CSP method is a supervised algorithm and therefore needs subject specific training data for calibration, which is very time consuming to collect. Instead of letting all that data and effort go to waste, the data of other subjects could be used to further improve results for new subjects. This problem setting is often encountered in multitask learning, from which we will borrow some ideas and apply it to the preprocessing phase.
This paper outlines the details of the multitask CSP algorithm and shows some results on data from the third BCI competition. In some of the subjects a clear improvement can be seen by using information of other subjects, while in some subjects the algorithm determines that a specific model is the best. We also compare the use of a global filter, which is constructed only with data of other subjects, with the case where we ommit any form of spatial filtering. Here, the global filter seems to boost performance in four of the five subjects.

1 INTRODUCTION

The development of BCI systems is an active research domain that has the goal to help people, suffering from severe disabilities, to restore the communication with their environment through an alternative interface. Such BCI systems can be divided in several categories based on the signal features they use. Some of these features like the P300 (Farwell and Donchin, 1988) and steady-state visual evoked potentials (SSVEP) (Kelly et al., 2005) are elicited naturally by external stimuli while others like the sensorimotor rhythms (SMR) can be independently generated by the subject. In case of SMR this can be achieved by performing the task of imagining different movements, such as left and right hand movement, or foot and tongue movement. The cortical areas involved in motor function (and also motor imagery) show a strong 8-12 Hz (or even 18-26 Hz) activity when the person is not performing any motor (imagery) task. However, when the person is engaged in a motor task the neural networks in the corresponding cortical areas are activated. This blocks the idle synchronized firing of the neurons and thus causes a measurable attenuation in those frequency bands. This decrease in power is also called event-related desynchronization (ERD) (Pfurtscheller and Lopes da Silva, 1999), the opposite is termed event-related synchronization (ERS). The location (electrode) of this feature depends on the type of motor task. For example, if a person moves his left arm, the brain region contralateral to the movement (around electrode C4) will display this ERD feature, while the neurons in the ipsilateral cortical motor area continue to fire synchronously.

Because of the low spatial resolution of electroenchephalography (EEG), a commonly used method to improve this resolution is the common

Devlaminck D., Wyns B., Otte G. and Santens P..
MULTITASK LEARNING APPLIED TO SPATIAL FILTERING IN MOTOR IMAGERY BCI - A Preliminary Offline Study.
DOI: 10.5220/0003335403780382
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing* (BIOSIGNALS-2011), pages 378-382
ISBN: 978-989-8425-35-5

spatial pattern (CSP) algorithm introduced by Koles (Koles, 1991) to detect abnormal EEG activity. Later, it was used for discrimination of imagined hand movement tasks (Müller-Gerking et al., 1999; Ramoser et al., 2000). Since then a lot of groups improved the basic CSP algorithm by extending it with temporal filtering (Dornhege et al., 2006), making it more robust for nonstationarities (Blankertz et al., 2008) or reducing calibration time by transferring knowledge learned during previous sessions (Krauledat et al.,). After almost a decade this method still proves it superiority judging from the results of the fourth BCI competition[1]. Still, this BCI setup is less accurate than the P300-based BCI and initially needs a longer training time. Some people are even unable to achieve proper control.

One way to further improve a subject specific CSP filter is to use the data recorded from other subjects, additionally to the subject's own data. To this end we will use some ideas of multitask learning, an active topic in machine learning (Evgeniou et al., 2005; Kato et al., 2008). In (Alamgir et al., 2010), the authors employed this concept to realize a classifier that was able to learn from multiple subjects, leading to an algorithm that performed well on new subjects even without training. The classifier could then be adapted when new data came available, reaching even higher classification accuracies with very few training samples. However, they did not apply any form of spatial filtering, using only features obtained from the EEG signal after filtering it in distinct pass-bands. We apply a similar idea in the preprocessing phase to construct spatial filters that make a trade-off between a global and subject specific filters.

In Section 2 we give the details of the multitask CSP algorithm. The results are then compared with the basic CSP algorithm in Section 3 on data of the third BCI competition. We highlight the strenghts and the weaknesses of the method in Section 4.

2 MULTITASK CSP

The goal of the basic CSP method is to learn a set of spatial filters for one subject that maximizes the signal variance for trials of one class while at the same time minimizes the signal variance for trials of the other classes. For the two class case, this can be formulated as follows

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T \Sigma^{(1)} \mathbf{w}}{\mathbf{w}^T \Sigma^{(2)} \mathbf{w}},$$

[1]On http://www.bbci.de/competition/iv/ you can find the data sets and results of the 4th BCI competition.

Algorithm 1: Rprop+.

$\eta_+ = 1.2$, $\eta_- = 0.5$, $\eta_{max} = 50$, $\eta_{min} = 10^{-30}$
initialize $\mathbf{w}$ as explaind in Section 2
repeat
  $t \leftarrow t+1$
  Compute gradient $\nabla R(\mathbf{w}) = \frac{\delta^{(t)} R(\mathbf{w})}{\delta \mathbf{w}}$
  for all $i = 1$ to $d(S+1)$ **do**
    if $\frac{\delta^{(t)} R(\mathbf{w})}{\delta w_i} \cdot \frac{\delta^{(t-1)} R(\mathbf{w})}{\delta w_i} > 0$ **then**
      $\eta_i \leftarrow \min(\eta_i \eta_+, \eta_{max})$
      $w_i \leftarrow w_i + \eta_i sign(\frac{\delta^{(t)} R(\mathbf{w})}{\delta w_i})$
    else if $\frac{\delta^{(t)} R(\mathbf{w})}{\delta w_i} \cdot \frac{\delta^{(t-1)} R(\mathbf{w})}{\delta w_i} < 0$ **then**
      $w_i \leftarrow w_i - \eta_i sign(\frac{\delta^{(t-1)} R(\mathbf{w})}{\delta w_i})$
      $\eta_i \leftarrow \max(\eta_i \eta_-, \eta_{min})$
      $\frac{\delta^{(t)} R(\mathbf{w})}{\delta w_i} \leftarrow 0$
    else
      $w_i \leftarrow w_i + \eta_i sign(\frac{\delta^{(t)} R(\mathbf{w})}{\delta w_i})$
    end if
  end for
until convergence

where $\Sigma^{(1)}$ and $\Sigma^{(2)}$ correspond to the covariance matrices of the trials corresponding to the first and respectively the second class.

We now want to use data of other subjects to improve the filters for specific subjects. To accomplish this, we first need a spatial filter $\mathbf{w}_s$ for each subject, which we decompose into the sum of a global and subject specific filter,

$$\mathbf{w}_s = \mathbf{w}_0 + \mathbf{v}_s ,$$

where $\mathbf{w}_0 \in \mathbb{R}^d$ represents the global spatial filter which is learned across all data (including those of other subjects) and $\mathbf{v}_s \in \mathbb{R}^d$ represents the subject specific filter. The number of channels is represented by d. A single optimization framework is proposed in which we learn both types of filters. This can be formulated as

$$\max_{\mathbf{w}_0, \mathbf{v}_s} \sum_{s=1}^{S} \frac{\mathbf{w}_s^T \Sigma_s^{(1)} \mathbf{w}_s}{\mathbf{w}_s^T \Sigma_s^{(2)} \mathbf{w}_s + \frac{1}{\lambda} ||\mathbf{w}_0||^2 + \lambda ||\mathbf{v}_s||^2}.$$

The parameter λ makes a trade-off between global or specific filters. For a high value of $\lambda >> 1$ the vector $\mathbf{v}_s$ is forced to zero and a global filter is constructed. When λ is very low (close to zero) the vector $\mathbf{w}_0$ is forced to zero and subject specific filters are computed. The number of subjects is denoted by S. This can be rewritten to a simpler form as,

$$\max_{\mathbf{w}} R(\mathbf{w}) = \max_{\mathbf{w}} \sum_{s=1}^{S} r_s(\mathbf{w}) = \max_{\mathbf{w}} \sum_{s=1}^{S} \frac{\mathbf{w}^T \bar{\Sigma}_s^{(1)} \mathbf{w}}{\mathbf{w}^T \bar{\Sigma}_s^{(2)}(\lambda) \mathbf{w}},$$

Table 1: Accuracy obtained by cross-validation for different parameter values λ.

subject \ λ	10^{-6}	10^{-4}	10^{-2}	1	10^{2}	10^{4}	10^{6}
aa	0.607	0.597	0.624	0.616	0.633	0.614	0.605
al	0.979	0.979	0.885	0.676	0.821	0.979	0.979
av	0.645	0.635	0.513	0.538	0.620	0.645	0.657
aw	0.681	0.710	0.643	0.618	0.647	0.666	0.683
ay	0.939	0.939	0.857	0.685	0.756	0.823	0.802

with

$$\mathbf{w}^T = \begin{pmatrix} \mathbf{w}_0^T & \mathbf{v}_1^T & \dots & \mathbf{v}_S^T \end{pmatrix},$$

$$\bar{\Sigma}_s^{(2)}(\lambda) = \bar{\Sigma}_s^{(2)} + \frac{1}{\lambda} D_0 + \lambda D_s,$$

$$\bar{\Sigma}_s^{(i)} = E_s \Sigma_s^{(i)} E_s^T$$

and

$$E_s = \begin{pmatrix} I_{d\times d} \\ \mathbf{0}_{(s-1)d\times d} \\ I_{d\times d} \\ \mathbf{0}_{(S-s)d\times d} \end{pmatrix},$$

$$D_0 = \begin{pmatrix} I_{d\times d} \\ \mathbf{0}_{Sd\times d} \end{pmatrix} \begin{pmatrix} I_{d\times d} & \mathbf{0}_{d\times Sd} \end{pmatrix},$$

$$D_s = \begin{pmatrix} \mathbf{0}_{sd\times d} \\ I_{d\times d} \\ \mathbf{0}_{(S-s)d\times d} \end{pmatrix} \begin{pmatrix} \mathbf{0}_{d\times sd} & I_{d\times d} & \mathbf{0}_{d\times(S-s)d} \end{pmatrix}.$$

We find the maximum through gradient search. To avoid finding the optimal step length in each iteration and speeding up convergence we employ the RProp+ algorithm, proposed in (Riedmiller and Braun, 1993) for supervised learning in feedforward artificial neural networks. The gradient can be computed as

$$\nabla R(\mathbf{w}) = \sum_{s=1}^{S} \frac{\bar{\Sigma}_s^{(1)}\mathbf{w} - r_s(\mathbf{w})(\bar{\Sigma}_s^{(2)} + \frac{1}{\lambda} D_0 + \lambda D_s)\mathbf{w}}{\mathbf{w}^T(\bar{\Sigma}_s^{(2)} + \frac{1}{\lambda} D_0 + \lambda D_s)\mathbf{w}}.$$

The RProp+ method is summarized in Algorithm 1 and uses the weight-backtracking approach. An intuitive way to intialize the component vector $\mathbf{w}_0$ in $\mathbf{w}$ is to take the average of the covariance matrices of all subjects and compute the best filter with the basic CSP algorithm. Initializing the other component vectors $\mathbf{v}_s$ in $\mathbf{w}$ is even easier, just run the basic CSP algorithm on the covariance matrices of each subject separately and select the best filter as starting point.

3 EXPERIMENTS

We use data of the third BCI competition[2], more precisely data set IVa. The set contains data recorded from 118 electrodes where the subjects performed two tasks: right hand motor imagery and foot imagery. Five subjects are included in the set and each subject recorded 280 trials. From each of these subjects, we use 100 trials for training and 180 for testing. To limit the number of parameters that needs to be computed by the RProp+ algorithm, the number of channels is reduced to 22. The ones selected are Fp1, Fpz, Fp2, F7, F3, Fz, F4, F8, T7, C3, Cz, C4, T8, P7, P3, Pz, P4, P8, POz, O1, Oz and O2. All remaining signals are band-pass filtered between 8 and 30 Hz.

The trade-off parameter λ is determined through cross-validation, which is the reason we still need a sufficient amount of data to accurately select the paramter value. For each subject only two spatial filters are computed: one for each class. The reason for the limit of one filter per class is the bad convergence of the algorithm after one iteration of projection deflation (a technique also use in principal component analysis to compute subsequent principal components). Table 1 shows the cross-validation accuracy for each subject and different parameters $\lambda \in \{10^{-6}, 10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}, 10^{6}\}$. Clearly, for some subjects a global filter is preferred (subject *av*), while for others a more intermediate filter is chosen (subject *aa*) or even a subject specific filter (subjects *aw* and *ay*). For subject *al* it does not matter which model parameter to choose as both global and specific filters perform equally well.

Figure 2 shows the spatial filters for two subjects *av* and *ay*, computed both with the basic CSP variant and with the multitask variant. As subject *ay* prefers a subject specific model, one can see that the multitask CSP variant (msCSP) converges to the same filter as the basic CSP variant (bCSP) for very low values of λ. However, for subject *av* the difference between the two filter variants can not be unnoticed. The global filters in the second and fourth column show a more physiological plausible solution, which is also supported by a higher accuracy on the test set as one can see in Table 3. In general, the multitask variant seems to improve the overall accuracy for each subject, except for subject *aa*, in which case a small decrease in performance is observed. The improvement in subjects such as *av* and *aw*, that initially do not perform well, can be due to the influence of subjects who do

[2]On http://www.bbci.de/competition/iii/ you can find the data sets and results of the 3e BCI competition.

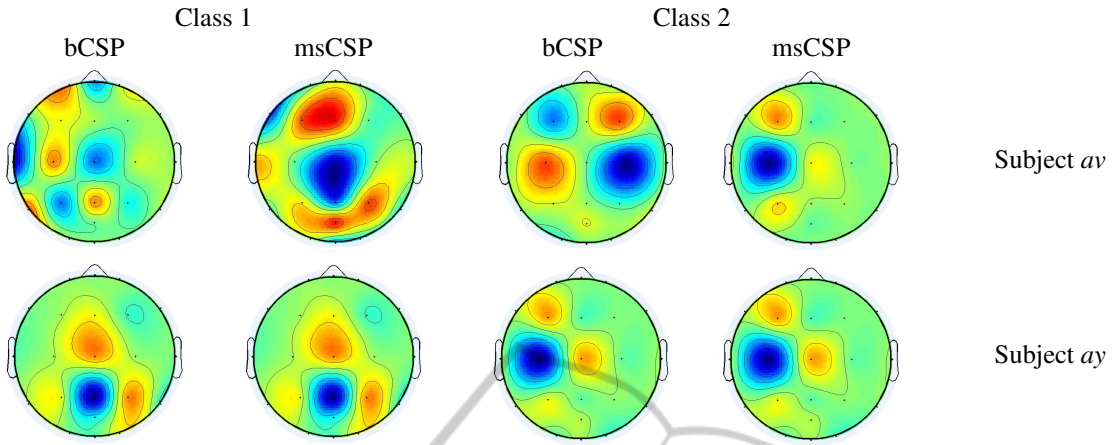

Table 2: The first row displays spatial filters for subject *av* and the second row for subject *ay*. The first and second column represent spatial filters for class one, the first one being a subject specific filter computed by the basic CSP algorithm (bCSP), while the second column displays the filter computed with the multitask CSP variant (msCSP). For subject *av* $\lambda = 10^6$ (corresponding to a global model) and for subject *ay* λ=10^{-6} (corresponding to a specific model). The third and fourth column then show the spatial filters for the second class.

perform well, such as *al* and *ay*. This idea is confirmed by looking at the msCSP filter of subject *av* for class two in Figure 2, which strongly resembles the subject specific filter of *ay*.

We also compare the application of a single global filter (for each class) with no spatial filtering. Here, the global filter is computed based on data of all subjects, except the data of the subject being tested ($\lambda = 10^{15}$). In this case, the training data is only used to build the classifier. When no spatial filtering is applied, we select the four channels C3, Cz, C4 and POz to compute the variance.

Although the global filter completely fails for subject *aw*, we see a clear improvement in all other subjects. This suggests it maybe possible to construct a global classifier in conjunction with this global filter to make predictions for new subjects without training. Predictions can then be further improved while new data comes available. Furthermore, it can potentially overcome the intial frustration of failure during earlier trials.

Table 3: Accuracy obtained on the test sets for each subject, comparing the basic CSP method with its multitask variant. Furthermore, two other methods are compared: the application of a single global model versus no spatial filtering.

method \ subject	*aa*	*al*	*av*	*aw*	*ay*
basic CSP	68.33	95.56	56.67	63.89	90.00
multitask CSP	64.44	95.56	67.78	73.89	90.00
no CSP	61.11	85.56	54.44	71.10	86.11
global CSP	66.67	93.33	66.11	53.89	90.56

4 CONCLUSIONS

We presented a multitask variant of the CSP algorithm that uses data recorded from multiple subjects to improve the results of a specific subject. The algorithm has two shortcomings. Firstly, the number of spatial filters that can be extracted is limited to one, but could potentially be overcome by using joint approximate diagonalization. Secondly, because the trade-off parameter is determined through cross-validation, the algorithm still needs sufficent training data to select the parameter reliably. However, with enough data to determine the trade-off parameter, we can see a clear improvement in all subjects except for one, where there's only a limited decrease in performance.

On top of that, we also employed the method for learning a single global filter based on data of all subjects except one, testing it on the remaining subject. In this case we can also see a clear improvement compared to the case where no spatial filtering is applied. This suggests that the method could be used to improve results when no training data is available. This is off course under the assumption that the classifier is also built from other subjects.

REFERENCES

Alamgir, M., Grosse-Wentrup, M., and Altun, Y. (2010). Multitask learning for brain-computer interfaces. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 13, pages 17–24, Sardinia, Italy.

Blankertz, B., Kawanabe, M., Tomioka, R., Hohlefeld, F., Nikulin, V., and Müller, K. (2008). Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing. In *Advances in Neural Information Processing Systems*, volume 20, pages 113–120, Vancouver, Canada.

Dornhege, G., Blankertz, B., Krauledat, M., Losch, F., Curio, G., and Müller, K. (2006). Optimizing spatio-temporal filters for improving Brain-Computer Interfacing. In *Advances in Neural Information Processing Systems*, volume 18, pages 315–322, Vancouver, Canada.

Evgeniou, T., Micchelli, C., and Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637.

Farwell, L. and Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523.

Kato, T., Kashima, H., Sugiyama, M., and Asai, K. (2008). Multi-task learning via conic programming. In *Advances in Neural Information Processing Systems 20*, volume 20, pages 737–744, Vancouver, Canada.

Kelly, S., Lalor, E., Reilly, R., and Foxe, J. (2005). Visual spatial attention tracking using high-density SSVEP data for independent brain-computer communication. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(2):172–178.

Koles, Z. (1991). The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. *Electroencephalography and Clinical Neurophysiology*, 79(6):440–447.

Krauledat, M., Schroder, M., Blankertz, B., and Müller, K. Reducing Calibration Time For Brain-Computer Interfaces: A Clustering Approach. In *Advances in Neural Information Processing Systems*, Vancouver, Canada.

Müller-Gerking, J., Pfurtscheller, G., and Flyvbjerg, H. (1999). Designing optimal spatial filters for single-trial eeg classification in a movement task. *Clinical Neurophysiology*, 110(5):787–798.

Pfurtscheller, G. and Lopes da Silva, F. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110:1842–1857.

Ramoser, H., Muller-Gerking, J., and Pfurtscheller, G. (2000). Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446.

Riedmiller, M. and Braun, H. (1993). A direct adaptive method for faster backpropagation learning: the RProp algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591, San Fransisco.