SVM feature selection for multidimensional EEG Data

Nisrine Jrad, Ronald Phlypo, Marco Congedo

SVM FEATURE SELECTION FOR MULTIDIMENSIONAL EEG DATA

Nisrine Jrad, Ronald Phlypo, Marco Congedo

Vision and Brain Signal Processing (ViBS), GIPSA-Lab, CNRS UMR 5216, Grenoble Universities
961 rue de la Houille Blanche, 38402 GRENOBLE Cedex, France

ABSTRACT

In many machine learning applications, like Brain - Computer Interfaces (BCI), only high-dimensional noisy data are available rendering the discrimination task non-trivial. In this work, we focus on feature selection, more precisely on optimal electrode selection and weighting, as an efficient tool to improve the BCI classification procedure. The proposed framework closely integrates spatial feature selection and weighting within the classification task itself. Spatial weights are considered as hyper-parameters to be learned by a Support Vector Machine (SVM). The resulting spatially weighted SVM (sw-SVM) is then designed to maximize the margin between classes whilst minimizing the generalization error. Experimental studies on eight Error Related Potential (ErrP) data sets, illustrate the efficiency of the sw-SVM from a physiological and a machine learning point of view.

***Index Terms*—** Support Vector Machines, spatial filters, feature extraction, Brain Computer Interfaces.

1. INTRODUCTION

The problem of learning a robust classifier from high dimensional noisy data arises in many machine learning applications. In this study, we focus on Brain Computer Interfaces (BCI). The majority of BCI applications make use of Electroencephalography (EEG) signals to categorize two or more classes of cerebral activity and associate them with simple computer commands. The EEG is a relatively cheap recording system, measuring the potential field at the scalp, which is an instantaneous reflection of the electrocerebral activity. It generally consists of recordings from 8 to 128 electrodes, sampling the potential field at a sample rate of 128Hz to several kHz. Unfortunately, an inherent default of this recording setup is its high sensitivity to noise. Classification of EEG brain signals thus requires adequate processing techniques to tackle the problems of feature reduction and noise cancellation.

In general, the classification is dealt with in two parts [1, 2]. Firstly, data preprocessing admits the representation of the data in an adequate form. This part generally includes a feature selection and/or a feature extraction, known as filtering. Features might be of frequential, temporal and/or spatial nature. Secondly, this data representations are fed to a classifier, which is often borrowed from the machine learning community without inquiring much into possible improvements that could be done, thus resulting in classifiers that are not fully exploiting the characteristics of the data in hand.

Concerning spatial filtering techniques, signal-processing criteria like signal-to-noise ratio [1] and ratio of class variances [2], have frequently been employed. They rely on the physical propagation model assuming that brain activities are in a quasi-linear and instantaneous relation to the recorded signals.

As for feature selection and noise cancellation, some algorithms have been proposed or adapted to the BCI scheme. For instance, genetic algorithms [3] have been successfully adapted for BCI spatial feature selection. Moreover, recent researches treat the problem of how to rate the relevance of features in terms of classification performance, using a zero norm optimization [4] or a recursive feature elimination [5]. More recently, a sensor selection procedure, based on the SNR, has been proposed in [6].

A common characteristic of all these feature extraction and selection methods is that they do not directly optimize a discrimination function. In fact, they construct or select features according to criteria that do not consider the overall BCI classification performance. A relation might be found between the objective functions yielding optimal spatial filters and class separability. But this relation has, to the best of our knowledge, never been addressed explicitly before.

In this work, we focus on the optimal selection and weighting of spatial features so as to improve the separability of the classes. The problem is treated within the SVM classification task itself. By introducing the spatial feature weights as hyper-parameters in a Support Vector Machine (SVM), they can be optimized for the specific classification problem in hand. In our spatially weighted SVM (sw-SVM), no assumption is made regarding the EEG data structure, i.e., the approach is completely data-driven. sw-SVM offers a highly flexible approach, in that it can handle any kind of features, thus adapting to any kind of EEG-based BCI (P300, motor imagery, SSVEP, etc.). To show the evaluate the sw-

This work has been supported through the project OpenViBE2 of the ANR (National Research Agency), France.

SVM framework, experiments are conducted on 8 subjects in a controlled Error-related Potential (ErrP) scenario [7]. The proposed sw-SVM algorithm is compared to a classical SVM approach to ascertain its efficiency.

The remainder of this paper is organized as follows. The proposed sw-SVM framework is introduced in Section 2. The sw-SVM optimization problem and one variant of the possible solutions are presented. Section 3 accounts for BCI data sets description and discusses the so obtained experimental results. Finally, Section 4 holds our conclusions.

2. METHOD

2.1. Problem description

BCI applications with two classes of action provide a training set of labeled trials from which a decision function should be learned that correctly classifies unlabeled trials. Let us consider an EEG trial recorded over s electrodes in a short time period of T samples as a matrix $\tilde{\boldsymbol{X}}_p \in \mathbb{R}^{S\times T}$. A pattern $\boldsymbol{x}_p \in \mathbb{R}^{d\times 1}$ will be obtained by unfolding the matrix $\tilde{\boldsymbol{X}}_p$, as such identifying $\mathbb{R}^{S\times T}$ with $\mathbb{R}^{d\times 1}$, where $d = ST$. $\boldsymbol{x}_p$ thus contains the commplete spatio-temporal recorded information of a single trial. Hence, the entire available set of data can be denoted $\{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_p, y_p), ..., (\boldsymbol{x}_P, y_P)\}$ with $y_p \in \{-1, 1\}$ the class labels.

Our task consists in finding the spatial features and an appropriate weighting function that maximize the separation margin between classes. Thus, we aim at finding a matrix $\boldsymbol{D} \in \mathbb{R}^{d\times d}$ of electrode weights assigned to each of the trials $\boldsymbol{x}_p$ so that $\{\boldsymbol{D}\boldsymbol{x}_p\}_{p=1}^{P}$ maximize the margin of the SVM.

2.2. sw-SVM: spatially weighted-SVM

The central idea of classical SVM is to separate data by finding a vector $\boldsymbol{w} \in \mathbb{R}^{d\times 1}$ and an offset $b \in \mathbb{R}$ of a hyperplane which provides us with the largest possible margin between classes and minimizes the number of misclassified patterns. The proposed sw-SVM suggests to involve spatial feature weights in the primal SVM optimization problem and tunes these weights as hyper-parameters of SVM.

For the application of EEG electrode selection, time features belonging to a same EEG electrode, hereafter indexed by s, have to be dealt with in a congeneric way so that a spatial interpretation of the solution becomes possible. The resulting matrix $\boldsymbol{D}$ is thus diagonal with S different unknowns repeated in T diagonal blocks of size $S \times S$, each containing the spatial filter $\mathbf{d} \in \mathbb{R}^{S\times 1}$ on their diagonal.

According to the above assumptions, the matrix $\boldsymbol{D}$ can be learned by solving the sw-SVM optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{w},b,\xi,\boldsymbol{D}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{p=1}^{P}\xi_p \\ \text{subject to} \quad & y_p(\langle \boldsymbol{w}, \boldsymbol{D}\boldsymbol{x}_p\rangle + b) \geq 1 - \xi_p \quad \forall p \in \{1,\dots,P\} \\ \text{and} \quad & \xi_p \geq 0 \quad \forall p \in \{1,\dots,P\} \\ \text{and} \quad & \sum_{s=1}^{S} D_{s,s}^2 = 1 \quad \forall s \in \{1,\dots,S\}, \end{aligned} \tag{1}$$

where ξ_p are called slack variables that ensure the problem has a solution in case the data is not linearly-separable, and C is the regularization parameter that controls the trade-off between a low training error and a large margin.

The objective function is not convex with respect to all parameters jointly, but is in each of its parameters. Hence, we proceed by alternating the search for a solution of (1). For $\boldsymbol{D}$ fixed, the problem is reduced to a ℓ_1 soft margin SVM with the only difference being that $\boldsymbol{x}_p$ is replaced by $\boldsymbol{D}\boldsymbol{x}_p$ in the inequality constraint. The primal and dual objective functions of such a problem are convex, and their solution can be obtained by any of the available SVM algorithms. Here we opt for simpleSVM [8]. Let us denote by $J(\boldsymbol{D})$ the optimal value of this problem. By setting to zero the derivates of the Lagrangian with respect to the primal variables, the efficient optimization problem of the dual formulation yielding $J(\boldsymbol{D})$ can be formulated as follows :

$$J(\boldsymbol{D}) = \begin{cases} \max_{\boldsymbol{\alpha}} \mathbf{1}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{Y}^T\boldsymbol{X}^T\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{X}\boldsymbol{Y}\boldsymbol{\alpha} \\ \text{subject to} \quad \boldsymbol{y}^T\boldsymbol{\alpha} = 0 \\ \text{and} \quad 0 \leq \alpha_p \leq C \quad \forall p \in \{1,\dots,P\}, \end{cases}$$

where $\boldsymbol{\alpha}$ is the vector of Lagrangian multipliers, $\boldsymbol{Y} = \text{Diag}(y_1, ..., y_P)$ is the matrix containing the trial labels on its diagonal, $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_P\}$ and $\boldsymbol{y}^T = \{y_1, ..., y_P\}$.

The value $J(\boldsymbol{D})$ is thus obtained for a given $\boldsymbol{\alpha}$ by solving the following problem:

$$\min_{\boldsymbol{D}} J(\boldsymbol{D}) \quad \text{subject to} \sum_{s=1}^{S} D_{s,s}^2 = 1. \tag{2}$$

By setting $\tilde{\boldsymbol{D}} = \boldsymbol{D}^T\boldsymbol{D}$, problem (2) reduces to a minimization problem under ℓ_1 constraints over $\tilde{\boldsymbol{D}}$. This is clearly an instance of the Multiple Kernel Learning (MKL) problem proposed in [9] with a homogeneous degree 1 polynomial kernel. Authors of [9] prove, with positive assumptions on the kernel matrices, that the search for the optimal $\tilde{\boldsymbol{D}}$ is convex, yielding fast convergence toward the optimal conditional solution. Hence, the optimization problem can be solved efficiently using a gradient descent algorithm as in SimpleMKL [9].

We initialize $\boldsymbol{D}$ as $S^{-1}\mathbb{I}$, where $\mathbb{I}$ is the identity matrix. We proceed by alternating the search for a solution of (1) for

$\boldsymbol{D}$, given a fixed $\boldsymbol{\alpha}$ and for $\boldsymbol{\alpha}$ given previous estimation of $\boldsymbol{D}$. The alternating optimization scheme is stopped when the ℓ_2 norm of the changes on $\mathbf{D}$ becomes negligible.

3. EXPERIMENTAL RESULTS

3.1. ErrP data set

The proposed method was evaluated on a visual feedback ErrP [7] experiment. Eight BCI-naif healthy subjects performed the experiment. They had to retain the position of a sequence of digits displayed in square boxes on a computer screen in front of them. Then, the sequence disappeared, a target digit was shown and subjects were asked to click on the box where it previously appeared. If the answer was correct, the chosen box background color turned into green, otherwise it turned into red. The number of digits composing the sequences continuously adapted to subject performance with an algorithm tuned to allow around 20% errors for all subjects, regardless the working memory ability and limits.

The experiment involved 2 sessions that lasted together approximately half an hour. Each session consisted of 6 blocks of 6 trials, for a total of $6 \times 6 \times 2 = 72$ trials. Recordings of the EEG were made from 31 electrodes using the extended $10/20$ international system. Raw EEG potentials were first re-referenced to the common average. A window of 1000ms posterior to the stimulus has been considered for each trial since studies on feedback ErrP report two peaks around 250ms and $300 - 500$ms as main components of the evoked potential. Then, a $1 - 10$Hz 4^{th} order butterworth filter was applied as error related potentials are known to be a relatively slow cortical potential. Finally, EEG samples were averaged in 16 continuous equally spaced windows. No artifact rejection algorithm was applied and all trials were kept for analysis.

3.2. Results

A sw-SVM 5 Cross-Validation was performed with different values for the regularization parameter C. A 5 Cross-validation experiments with classical SVM (without spatial feature selection) was carried out as a comparative baseline.

Table 1 reports the 5 top weighted electrodes for the 8 subjects and the number of spatial features selected for 1 of the 5 partitions. Figure 1 shows the electrode weights averaged across the 5 partitions as topographic maps (associated with that value of C allowing the highest classification rate). Table 1 and Figure 1 show that, according to sw-SVM, 11 spatial features at most suffice to capture error-related potentials in this data set. This confirms results in [10], where it was found that a substantial reduction of sensors performs well in noisy data and when the training set is small. Electrode selection is also known to be strongly subject-dependent. However, in 6 out of 8, the subjects' central area holds the strongest weight. This is in accordance with current knowledge on ErrP.

Subject	5 top weighted electrodes					# of elec.
S1	CP3	C3	P7	CP4	TP7	10
S2	F7	TP8	T4	P3	C3	9
S3	P8	CPz	T3	FP2	FC3	10
S4	O1	Cz	TP8	FP2	FCz	11
S5	O1	FCz	P7	FC3	Oz	6
S6	P7	F3	FCz	FC4	P4	9
S7	Cz					1
S8	F8	FPz	Cz	Ft7	F7	7

Table 1. For the 8 subjects and a given partition: the 5 top weighted electrodes, and the number of selected electrodes according to sw-SVM.

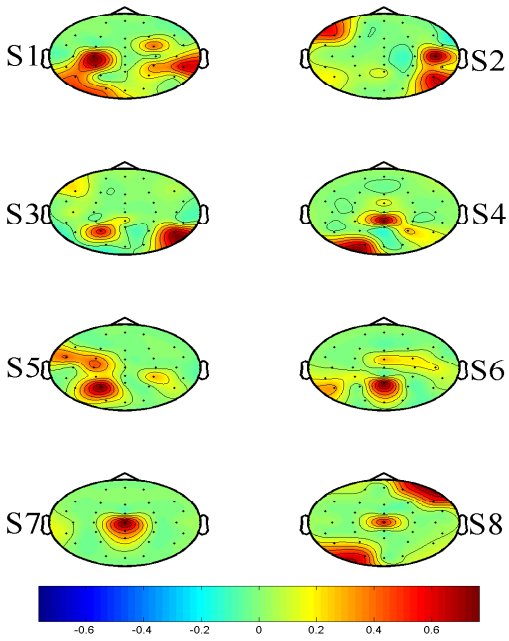

Fig. 1. Topographical maps of the weights averaged across the 5 partitions. Each map refers to a subject (S1 to S8).

Subject	SVM	sw-SVM	p-value
S1	72.86± 3.19	78.57 ±10.10	0.1778
S2	68.57± 14.81	77.14 ± 11.74	0.0327
S3	58.57± 10.59	74.29± 6.39	0.0042
S4	87.14± 12.78	90.00± 10.83	0.1778
S5	60.00±6.39	78.57± 13.36	0.0486
S6	85.71± 8.75	91.43± 7.82	0.0161
S7	68.57± 27.01	81.43± 8.14	0.2859
S8	64.29± 11.29	74.29± 9.58	0.3111
Mean	70.71±11.85	80.72±9.75	

Table 2. Performances of classical SVM and sw-SVM for the 8 subjects. t-test p-values of sw-SVM and baseline SVM are also reported.

Table 2 reports the single trial recognition rates (mean and standard deviations) for the 8 subjects. Classification accuracy is between 74% and 91%, averaging to about 81% for sw-SVM. These figures have been achieved with a relatively low number of features (from 1 to 11 electrodes). Noteworthy is also that available data include a small number of trials and even a smaller number of errors (less than 20% of the available data). Thus, as expected, the cross-validation variance is high. It will be interesting to consider more in depth the performance of sw-SVM on larger data-sets.

Table 2 reports also the 5 Cross-Validation performance provided by a classical SVM approach where all electrodes are used. A repeated-measure t-test has been performed to test the null hypothesis of no difference in the performance of the two methods. The p-value of such null hypothesis being true is also reported. While tolerating a 0.05 type II error level we can reject the null hypothesis for 4 out of the 8 subjects. Combining the 8 individual p-values with the combination function proposed by Edgington and Fisher (see [11] for details), yields a combined p-value of 0.000024 and 0.000157, respectively.[1] In conclusion, as compared to a standard SVM, on this data set sw-SVM both yields a significant dimensionality reduction and a considerable performance improvement.

4. CONCLUSION

In this paper, EEG spatial feature selection and weighting was considered from a machine learning point of view. Feature weights are introduced in the SVM theoretical framework and tuned as hyper-parameters of SVM. Hence, they maximize the margin between classes and minimize generalization error. The proposed method guarantees an automatic and reliable selection and appropriate weighting of spatial features. Though sw-SVM was designed as a spatial feature selector for BCI applications, there is no reason to restrict it for this particular application.

Experimental results on Error-related Potentials data sets illustrate the efficiency of the method. The algorithm performs well in terms of both spatial feature selection and classification accuracy. It is a promising tool for flexible and robust data classification that could perform well even with a small number of training observations. These results motivate further research that may aim to extend SVM toward a spatio-temporal filtering SVM. Hence, all relevant aspects of brain post-stimulus dynamics recorded in an EEG could be modeled in a supervised learning fashion.

Acknowledgment

The authors would like to acknowledge Sandra Rousseau from GIPSA-Lab Grenoble INP for providing the data.

[1] These combining functions express the probability to obtain the 8 observed individual p-values under the combined null hypotheses that all 8 individual null hypotheses are simultaneously and independently true.

5. REFERENCES

[1] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "xDAWN algorithm to enhance evoked potentials: Application to brain computer interface," *IEEE Trans. Biomedical Engineering*, vol. 56, no. 8, pp. 2035–2043, 2009.

[2] C. Gouy-Pailler, M. Congedo, C. Brunner, C. Jutten, and G. Pfurtscheller, "Nonstationary brain source separation for multiclass motor imagery," *IEEE Trans. on Biomedical Engineering*, vol. 57, no. 2, pp. 469–478, 2010.

[3] M. Schröder, M. Bogdan, W. Rosenstiel, T. Hinterberger, and N. Birbaumer, "Automated EEG feature selection for brain computer interfaces," in *Proc. of the 1st Int. IEEE EMBS Conf. on Neural Engineering*, 2003, pp. 626–629.

[4] J. Weston, A. Elisseeff, B. Schölkopf, and Pack Kaelbling, "Use of the zero-norm with linear models and kernel methods," *Journal of Machine Learning Research*, vol. 3, pp. 1439–1461, 2003.

[5] A. Rakotomamonjy and V. Guigue, "BCI Competition III : Dataset II - Ensemble of SVMs for BCI P300 speller," *IEEE Trans. Biomedical Engineering*, vol. 55, no. 3, pp. 1147–1154, 2008.

[6] H. Cecotti, B. Rivet, M. Congedo, C. Jutten, O. Bertrand, J. Mattout, and E. Maby, "Suboptimal sensor subset evaluation in a P300 brain-computer interface," in *Proc. of European signal Processing Conf. (EUSIPCO)*, 2010, pp. 924–928.

[7] W.H.R. Miltner, C.H. Braun, and M.G.H. Coles, "Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a generic neural system for error detection," *Journal of Cognitive Neuroscience*, vol. 9, pp. 788–798, 1997.

[8] S. Vishwanathan, A. J. Smola, and M. Murty., "SimpleSVM," in *Int. Conf. on Machine Learning*, 2003.

[9] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[10] C. Sannell, T. Dickhaus, S. Halder, E.M. Hammer, K.R. Müller, and B. Blankertz, "On optimal channel configurations for SMR-based brain-computer interfaces," *Brain Topogr.*, vol. 23, no. 2, pp. 186–193, 2010.

[11] M. Congedo, J.F. Lubar, and D. Joffe, "Low-resolution electromagnetic tomography neurofeedback," *IEEE Trans. on Neuronal Systems and Rehabilitation Engineering*, vol. 12, no. 4, pp. 387–397, 2004.