SUPPORT VECTOR EEG CLASSIFICATION IN THE FOURIER AND TIME-FREQUENCY CORRELATION DOMAINS

Gary N. Garcia, Touradj Ebrahimi and Jean-Marc Vesin
Swiss Federal Institute of Technology - EPFL, CH-1015, Lausanne, Switzerland
Gary.garcia@epfl.ch, Touradj.Ebrahimi@epfl.ch,Jean-Marc.Vesin@epfl.ch

Abstract In this paper we use support vector learning machines (SVM) for classifying EEG signals corresponding to imagined motor movements. The parameters of an SVM Kernel are optimized for minimizing a theoretical error bound. Fourier features and correlative time-frequency based features are extracted from EEG signals and compared with respect to their discriminatory power.

Keywords – Direct brain-computer communication, EEG classification, SVM, optimal SVM parameters choice, time-frequency correlation.

I. INTRODUCTION

A noninvasive electroencephalogram (EEG) based brain-computer communication device (henceforth called brain-computer interface BCI) can be subdivided into three subsystems, namely EEG acquisition, EEG signal processing and the output subsystem (Figure 1).

The EEG acquisition subsystem is responsible for gathering and digitizing the EEG signals measured at the scalp. EEG signals are composed of the single signals measured at different electrodes placed on the scalp according to the ten-twenty international system [1].

Digital EEG is fed into the signal processing subsystem where it is preprocessed and classified among a predetermined set of classes. The classification result (a label indicating the most probable class) is sent to the output subsystem which executes the action associated with the class label.

Each class corresponds to a mental activity (MA). Usually, a BCI is operated with a small number of MAs that correspond to imagined motor tasks [2].

Successful operation of a BCI depends on the judicious choice of features that are extracted from EEG signals, the classification strategy and the user himself who has to modulate his mental activity so as to make the BCI accomplish his intents [2]. The first two points involve knowledge of brain's electrophysiology and machine learning. The latter point deals with the feedback provided to the user during the training [3].

In the case of imagined motor tasks, the Fourier analysis of each EEG component appears to be adequate for the classification [2]. Good results using the joint correlative time-frequency representation (CTFR) of EEG were also reported in [4].

The classification strategy is adapted to the nature of the extracted features and must allow for continuous updating of its intrinsic parameters. Indeed, the EEG associated with a specific MA can present short term and long term variations as a result of different brain background activities [2].

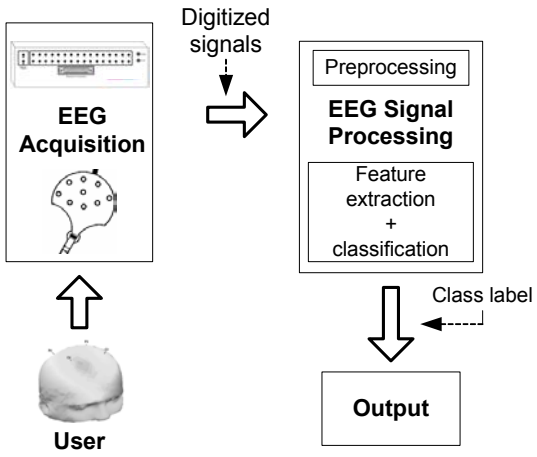

Figure 1. Parts of a BCI.

Recent advances in machine learning research have pointed out the advantages of support vector machines (SVM) over other classification techniques [5]. Solid theoretical foundations, good generalization capabilities and easy parameters updating are among the most appealing qualities of SVMs for BCI applications.

In this paper, we focus on the application of SVMs to EEG classification in the Fourier and in the CTFR domains.

II. FOURIER ANALYSIS AND CTFR OF EEG

We note an EEG signal as $X(t)=\left[x_1(t) \quad ... \quad x_N(t)\right]^{\mathbf{T}}$ where **T** is the transpose operator, $x_n(t)$ is the signal measured at the n^{th} electrode and N is the number of electrodes.

A simple Fourier analysis of $X(t)$ (such as reported in [2] and [3]) consists in computing the power at some frequency bands of each of the components of $X(t)$. If we note by N_B the number of frequency bands, the result of this analysis is a set of values that can be arranged into a vector of length NN_B. When no prior information about the optimal N_B and the length of the frequency bands is available, one can compute the power corresponding to the uniformly spaced 2 Hz bands ranging from 2 to 40 Hz.

The CTFR of $X(t)$ is defined as

$$A_X(\theta,\tau)=\int X\left(t+\tau/2\right)\cdot X^{\mathbf{H}}\left(t-\tau/2\right)e^{i\theta t}dt \qquad (1)$$

where θ and τ are the frequency and time lags respectively, "$\cdot$" denotes the ordinary matrix multiplication and $^{\mathbf{H}}$ is the

Hermitian operator. If we note by N_τ and N_θ the number of time and frequency lags for which $A_X(\theta,\tau)$ is computed, we obtain N^2 matrices of dimension N_θ x N_τ. As in the Fourier case, these values can be arranged into a single vector. When no prior information about the optimal values for N_θ and N_τ exists, $A_X(\theta,\tau)$ can be sampled at the same rate as $X(t)$.

The CTFR measures the degree of similarity between two time-frequency shifted versions of $X(t)$. Besides the spectral information, the CTFR provides information about the time-frequency interactions between the components of $X(t)$. Thus, with the CTFR the EEG components are not independently analyzed (as in the Fourier case) but their relationship is also taken into account.

An important drawback of the CTFR resides in its relative high sensitivity to noise. Consequently, the most important values of the CTFR, in terms of classification must be selected [4].

By virtue of the above considerations, the result of the Fourier analysis or the CTFR applied to an EEG signal $X(t)$ is a feature vector that we note $\boldsymbol{x}$.

III. Support Vector Machines for Classification

For the sake of explanatory convenience we only consider the two-class classification problem. Multi-class classification for a small number of classes (as in BCI applications) can be done with multiple pair-wise comparisons.

A more complete description of the SVM theory can be found in [6].

Given a set of labeled observations (training set) $\{(\boldsymbol{x}_l, y_l)\ ;\ 1 \le l \le L\}$ ($\boldsymbol{x}_l \in \mathbb{R}^D$ is the observed vector and $y_l \in \{-1,1\}$ its label); one has to estimate a decision function $f_{\alpha^*} : \mathbb{R}^D \to \{-1,1\}$ chosen from a set of admissible decision functions $\{f_\alpha\}$ (α is a vector of parameters) such that f_{α^*} will correctly classify unseen vectors $\boldsymbol{x}$.

The risk associated with a function f_α is

$$R(\alpha) = \int \frac{1}{2}\left|y - f_\alpha(\boldsymbol{x})\right| dP(\boldsymbol{x}, y) \tag{2}$$

where $P(\boldsymbol{x}, y)$ is the (usually unknown) probability distribution of pairs $(\boldsymbol{x}, y)$. The empirical risk is defined as the mean error rate on the training set.

$$\mathrm{R}_{emp}(\alpha) = \frac{1}{2L}\sum_{l=1}^{L}\left|1 - f_\alpha(\boldsymbol{x}_l)\right| \tag{3}$$

For some η such that $0 < \eta < 1$, with probability of at least $(1-\eta)$, the following bound holds [6]:

$$R(\alpha) \le \mathrm{R}_{emp}(\alpha) + \sqrt{\frac{h\left[\log(2L/h)+1\right] - \log(\eta/4)}{L}} \tag{4}$$

where h is the Vapnik-Chervonenkis (VC) dimension (a measure of the learning capacity of the set $\{f_\alpha\}$). The second term in the right side of (4) is called the VC confidence.

While the empirical error can become arbitrary small by choosing $\{f_\alpha\}$ with large h, the VC confidence term increases with h. The SVM approach for reaching a good compromise consists in keeping the value of the empirical risk fixed (e.g. equal to zero) and minimize the VC confidence.

In the linearly separable case, the decision functions can be written as $f_{(w,b)}(\boldsymbol{x}) = sign(\langle w, \boldsymbol{x}\rangle + b)$ ($\langle w, \boldsymbol{x}\rangle$ is the inner product of $w, \boldsymbol{x} \in \mathbb{R}^D$). One can show that the following bound on h holds [6]:

$$h \le \min\left(R^2\|w\|^2, D\right) + 1 \tag{5}$$

where R is the radius of the smallest sphere containing the training vectors.

The parameters of the decision function can be found by minimizing $\|w\|^2$ subject to $y_l f_{(w,b)}(\boldsymbol{x}_l) \ge 1$; $1 \le l \le L$. In this formulation, training errors are not allowed.

In order to allow for training errors one introduces the non-negative lack variables $\xi_{l=1,\ldots,L}$ such that the parameters w and b are now found by minimizing

$$\Phi\left(w, \boldsymbol{\xi} = [\xi_1,\ldots,\xi_L]^{\mathbf{T}}\right) = \frac{1}{2}\|w\|^2 + C\sum_{l=1}^{L}\xi_l \tag{6}$$

(C is a user defined constant controlling the learning capacity) under the constraints

$$\xi_{l=1,\ldots,L} \ge 0$$
$$y_l f_{(w,b)}(\boldsymbol{x}_l) \ge 1 - \xi_l \ ;\ 1 \le l \le L$$

Introducing positive Lagrange multipliers for the constraints and taking the derivatives with respect to $\|w\|$, the solution of (6) is

$$w = \sum_{l=1}^{L} y_l \alpha_l \boldsymbol{x}_l \tag{7}$$

where the α_l are found by solving the dual problem of (6): maximize

$$W\left(\boldsymbol{\alpha} = [\alpha_1,\ldots,\alpha_L]^{\mathbf{T}}\right) = \boldsymbol{\alpha}^{\mathbf{T}} \cdot \mathbf{1}_L - \frac{1}{2}\boldsymbol{\alpha}^{\mathbf{T}} \cdot \Lambda \cdot \boldsymbol{\alpha} \tag{8}$$

($\mathbf{1}_L$ is the L x 1 matrix with unitary elements and Λ is a L x L matrix whose elements are $\Lambda_{jk} = \langle \boldsymbol{x}_j, \boldsymbol{x}_k \rangle$) under the constraints

$$0 \le \alpha_{l=1,\ldots,L} \le C$$
$$\boldsymbol{\alpha}^{\mathbf{T}} \cdot Y = 0 \ ;\ Y = [y_1,\ldots,y_L]^{\mathbf{T}}$$

The Karush-Kuhn-Tucker (KKT) conditions imply that the solution of (8) is sparse (some of the α_l 's are equal to zero).

An $\boldsymbol{x}_l$ associated to a nonzero α_l is called a support vector (SV); in this case the corresponding ξ_l is equal to zero and $y_l f_{(w,b)}(\boldsymbol{x}_l) = 1$. The latter equality allows us to determine b.

The optimal decision function can be expressed as

$$f(\boldsymbol{x}) = sign\left(\sum_{l;\alpha_l>0} y_l \alpha_l \langle \boldsymbol{x}_l, \boldsymbol{x} \rangle + b \right) \quad (9)$$

As the optimal decision function depend on the SVs only, one says that the training set is entirely characterized by the SVs. Because of this property, it is easy to build a new set of support vectors by adding the old support vectors to a new training set.

It is important to note that the optimization problem Eq. (8) and the decision function Eq. (9) only involve inner products. This property allows us to apply the above solution to the nonlinearly separable case.

When the data is not linearly separable it is projected to a high dimensional space (Ң) so that it becomes linearly separable in Ң. The mapping $\psi : \mathbb{R}^D \to$ Ң does not need to be calculated explicitly. Instead, a Kernel function that computes the inner product in Ң is defined: $K(\boldsymbol{x}_j, \boldsymbol{x}_k) = \langle \psi(\boldsymbol{x}_j), \psi(\boldsymbol{x}_k) \rangle$. In this case, the solution is obtained by replacing the inner products in Eq. (8) and Eq. (9) by the Kernel functions.

A popular Kernel in SVM classification is the radial basis kernel

$$K_{RBF}(\boldsymbol{x}_j, \boldsymbol{x}_k) = \exp\left(-\theta \left\| \boldsymbol{x}_j - \boldsymbol{x}_k \right\|^2\right) \quad (10)$$

In this study, we consider a multiple parameter radial basis kernel [8]

$$K(\boldsymbol{x}_j, \boldsymbol{x}_k) = K_{jk} = \exp\left(-(\boldsymbol{x}_j - \boldsymbol{x}_k)^{\mathbf{T}} \cdot \Theta \cdot (\boldsymbol{x}_j - \boldsymbol{x}_k)\right) \quad (11)$$

where Θ is a D x D diagonal matrix with elements $\{\theta_d \geq 0 \ ; \ 1 \leq d \leq D\}$.

The elements of Θ are found by optimizing a bound on the generalization error [8]. The optimal value of θ_d measures the discriminative power of the d^{th} component of vectors $\boldsymbol{x}$. If θ_d is zero we can safely remove the corresponding component. This can be considered as a feature selection step.

The process of optimization is briefly presented below. For a more complete description the reader is referred to [8].

In [9] Vapnik proposed the following upper bound on the leave-one-out error.

$$P = \frac{1}{L} R^2 \|w\|^2 \quad (12)$$

The radius R of the smallest sphere containing the training vectors can be found by solving the following optimization problem:

$$R^2 = \underbrace{\max}_{\boldsymbol{\beta}} \left[\sum_{l=1}^{L} \beta_l K_{ll} - \sum_{j,k=1}^{L} \beta_j \beta_k K_{jk} \right] \quad (13)$$

under constraints

$$\sum_{l=1}^{L} \beta_l = 1 \ ; \ \beta_{l=1,\ldots,L} \geq 0$$

In order to find the optimal value for θ_d, one can compute the derivative of P with respect to θ_d and perform a gradient step algorithm. As the optimal solution of Eq. (8), that we note $\boldsymbol{\alpha}$ implicitly depends on θ_d the chain rule is applied.

$$\frac{\partial P}{\partial \theta_d} = \left.\frac{\partial P}{\partial \theta_d}\right|_{\boldsymbol{\alpha} \text{ fixed}} + \frac{\partial P}{\partial \boldsymbol{\alpha}} \frac{\partial \boldsymbol{\alpha}}{\partial \theta_d} \quad (14)$$

The following results can be obtained

$$\left.\frac{\partial P}{\partial \theta_d}\right|_{\boldsymbol{\alpha} \text{ fixed}} = -R^2 \boldsymbol{\alpha}^{\mathbf{T}} \cdot \frac{\partial}{\partial \theta_d} \tilde{K} \cdot \boldsymbol{\alpha} + \|w\|^2 \left(\sum_{l=i}^{L} \beta_l \frac{\partial K_{ll}}{\partial \theta_d} - \sum_{j,k=i}^{L} \beta_j \beta_k \frac{\partial K_{jk}}{\partial \theta_d} \right)$$

$$\frac{\partial P}{\partial \boldsymbol{\alpha}} = 2R^2 \left(\mathbf{1}_L - \tilde{K} \cdot \boldsymbol{\alpha} \right)$$

$$\frac{\partial \boldsymbol{\alpha}}{\partial \theta_d} = -\Omega^{-1} \cdot \frac{\partial}{\partial \theta_d} \Omega \cdot \begin{bmatrix} \boldsymbol{\alpha}_{\alpha_l \neq 0} & b \end{bmatrix}^{\mathbf{T}} \ ; \ \Omega = \begin{bmatrix} K_{\alpha_l \neq 0} & Y_{\alpha_l \neq 0} \\ Y_{\alpha_l \neq 0}^{\mathbf{T}} & 0 \end{bmatrix}$$

where $\tilde{K}$ is the matrix whose elements are $\tilde{K}_{jk} = y_j y_k K_{jk}$, $K_{\alpha_l \neq 0}$ is the matrix obtained after removing the elements corresponding to the nonsupport vectors from K and $Y_{\alpha_l \neq 0}$ is the matrix whose elements are the labels of the SVs.

The parameter θ_d is updated as follows:

$$\theta_d \leftarrow \theta_d - \varepsilon \frac{\partial P}{\partial \theta_d}$$

where ε is a user defined learning factor.

IV. Results and Discussion

Two male right handed subjects participated in the experiments. The signals from electrodes C3 and C4 [1] were measured at a rate of 128 Hz. The reference was placed in Cz. In addition, an electrode was placed at each eyebrow for detecting ocular artefacts.

The subjects were asked to perform two types of imagined mental activities, namely left and right index finger movement (MA1 and MA2 respectively). Half second segments of EEG (EEG trials) were classified.

Five hundred (artefacts-free) EEG trials per MA and per subject were selected for experimentation.

The training set was composed of 200 randomly selected EEG trials per MA, the generalization error rate was estimated on the remaining 600 EEG trials.

For the Fourier analysis, we computed the power of the two Herz frequency bands uniformly spaced in the 2 to 40 Hz frequency range. Thus, a vector of 38 components was obtained for each EEG trial.

As the modulus of the CTFR (for real signals) is symmetrical with respect to the origin we can compute the CTFR at 32 time lags and 32 frequency lags. Thus, a vector of 4096 components was obtained for each EEG trial.

The estimated error rate associated with the Fourier and CTFR analysis are represented in Figure 2. The theoretical error bound (Eq.(12)) is also represented for comparison.

Furthermore, the error rate obtained with a linear discriminant analysis (LDA) based classification is depicted in Figure 3.

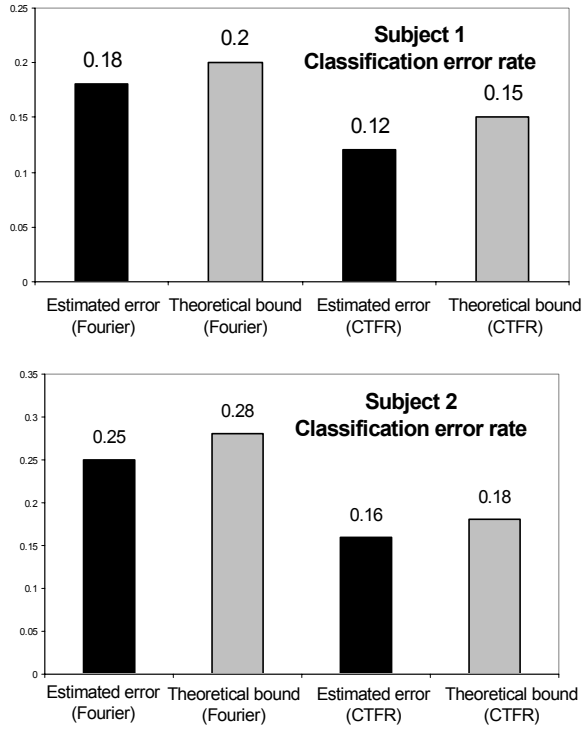

Figure 2. Classification error rate for Fourier and CTFR based features.

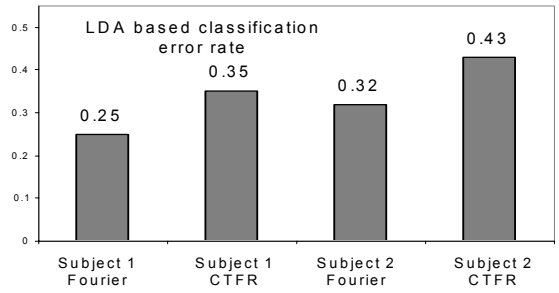

Figure 3. Error rate for the LDA based classifier.

From the results reported on Figure 3, one can say that the Fourier analysis provides better results when the classes are assumed to be linearly separable. This is an obvious result since each feature vector's component is considered independently in the LDA based classifier, and the Fourier components reflect global characteristics of the signal while the CTFR components are local measures that are dependent one another.

In Figure 2 the error obtained with the CTFR analysis is smaller than in the Fourier case. This means that the CTFRs corresponding to the two MAs considered here are better discriminated when nonlinear separation is assumed. Indeed, the classification error for the CTFR based features is considerably smaller when a SVM allowing for nonlinear decision boundaries is used.

V. Conclusions and Future Work

The flexibility requirements imposed on the classification strategy, in the framework of BCI applications are satisfactorily fulfilled by an SVM based classifier. The solid theoretical foundations of the SVM allow us to optimize several parameters of a Kernel function using analytical methods. The overfitting is cleverly avoided by controlling the trade-off between the training error minimization and the learning capacity of the decision functions. Finally, the decision function parameters can be easily updated because they depend on the SVs only.

Features based on the time-frequency interaction between the signals coming from different electrodes provide better results in terms of classification error rate. These features can efficiently separate the classes when nonlinear decision boundaries are constructed.

Kernel based methods can be used in the context of novelty detection for outliers detection [5]. We intend to utilize this approach for providing the feedback during the training sessions, in order to make the user confine his mental activity, corresponding to a given MA into a small region in the feature space.

References

[1] H.H. Jasper, "The Ten-Twenty electrode system of the international federation," *EEG and Clin. Neurophysiol.*, vol. 10, pp. 371-375, 1958.

[2] J.R. Wolpaw, et al, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, pp. 767-791, 2002.

[3] C. Neuper, A. Schloegl and G. Pfurtscheller, "Enhancement of Left-Right Sensorimotor EEG Differences During Feedback-Regulated Motor Imagery," *Clin. Neurophysiol.*, vol. 16, pp. 373-382, 1999.

[4] G. Garcia, T. Ebrahimi and J-M. Vesin, "Classification of EEG signals in the ambiguity domain for brain computer interface applications," *IEEE Int. Conf. Dig. Signal Proc.*, vol. 1, pp. 301-305, 2002.

[5] B. Schölkopf and A. Smola, *Learning with Kernels*, MIT Press, 2002.

[6] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 2000.

[7] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, 2, pp. 121-167, 1998.

[8] O. Chapelle, et al, "Choosing Multiple Parameters for Support Vector Machines," *Machine Learning*, 46, pp. 131-159, 2002.

[9] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.