LARGE MARGIN WAVELET-BASED DICTIONARY FOR SIGNAL CLASSIFICATION

Florian Yger, Alain Rakotomamonjy

LITIS EA 4108,
Université de Rouen,
76800 Saint Etienne du Rouvray, France

ABSTRACT

This paper addresses the problem of automatic wavelet feature extraction for signal classication. We propose to jointly learn wavelet-based features (including scale and translation of the wavelet as well as its shape) and a decision function by casting the problem as a Multi-Kernel Learning problem. A novel active constraints algorithm is then proposed. Our method has been tested on a toy dataset and compared to classical methods with competitive results.

Index Terms— SVM, parametrized waveform, Multi-Kernel Learning

1. INTRODUCTION

In any pattern recognition problem, the choice of the features used for characterizing an object to be classified is of primary importance. Indeed, those features largely influence the performance of the pattern recognition system. Oftenly, features are extracted from original data and they are crafted so that they capture some informative characteristics about classes to be discriminated. For instance, for signal classification in Brain-Computer Interfaces, some works use wavelet or time-frequency transform of EEG signal as features [1].

For signal classification, wavelet decompositions are now classical features that are frequently used [2]. However, while being part of usual tools for feature crafting from signal, wavelet decompositions have to be carefully adapted to the problem at hand. Such an adaptation usually involves the choice of the wavelet decomposition tree as proposed by Saito et al. [3]. Furthermore, using classical wavelets such as Daubechies wavelet may lead to poor performance since those wavelets have been built in order to have properties that are not always necessary for achieving good classification performance. Hence, many works have recently looked for adapting the wavelet to their data to be classified. For instance, Neumann et al. [4] tune their wavelet by maximizing the distance between means of the classes in the wavelet feature space. Instead, Lucas et al. [5] consider the wavelet parameter as a parameter of their kernel-based classifier (a Support Vector Machine) and propose to set it by means of a cross-validation error criterion.

In this work, we also consider the problem of wavelet adaptation for wavelet-based signal classification. However, while most of the above-described approaches consider the wavelet adaptation problem independently to the the classification problem, in this work, we propose to jointly learn a combination of wavelet coefficients (including the shape of the mother wavelet) to be used for classification and the classifier in itself. For this purpose, we cast the problem as a multiple kernel learning where each kernel is a related to a wavelet coefficient resulting from a parametrized wavelet decomposition.

For this purpose, we first show how to build kernels from any wavelet decomposition. For selecting the optimal large-margin wavelet, we introduce the multiple wavelet kernel learning problem. Then, we propose an active constraint algorithm grounded on the KKT conditions. Our method is tested on a toy dataset and compared to other methods available in the literature.

2. LARGE MARGIN WAVELET KERNEL

2.1. From wavelet to kernel

Our objective in this work is to integrate the process of extracting wavelet features from a signal into the classifier learning process. Hence, our work can be interpreted as a method for selecting the best wavelet and the best elements of the resulting dictionary for a classification task. As described in the sequel, our selection criterion is a large margin criterion and thus, we need to introduce a kernel based on wavelet decomposition.

Let x and x' be to discrete signals belonging to $\mathbb{R}^d$ and $\phi_{\theta,s,t}$ be the wavelet resulting from the dilation at scale s and translation t of the orthogonal mother wavelet ϕ_θ. The role played by θ will be made clear in the sequel. Kernels from wavelet can be simply obtained by considering the (non)-linear mapping of the signals onto the wavelet decomposition. For instance, the following kernels can be built :

$$K_{\theta,s,t}(x, x') = \langle \phi_{\theta,s,t}, x \rangle \langle \phi_{\theta,s,t}, x' \rangle \qquad (1)$$

where $K_{\theta,s,t}$ is just a linear kernel on the wavelet coefficient obtained at scale s and translation t. If one only wants to take

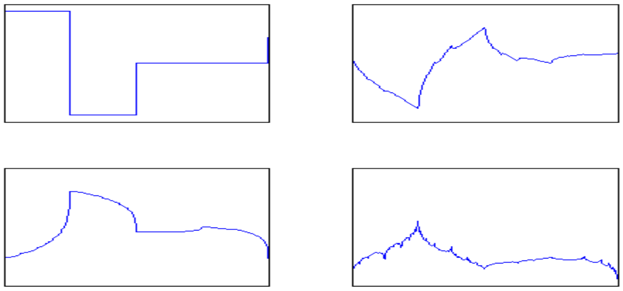

Fig. 1. Example of wavelet obtained from different θ

into account some frequency bands in the signal decomposition, then the following kernel can be used.

$$K_{\theta,s}(x,x') = \sum_t |\langle \phi_{\theta,s,t}, x\rangle| \cdot \sum_t |\langle \phi_{\theta,s,t}, x'\rangle| \qquad (2)$$

In the sequel, we will focus on kernels of the form given in equation (1) in a sake of clarity but our approach is general enough to be applied to any type of kernel.

Now that the kernels we use have been defined, we focus on how such kernels can be parametrized with respects to a mother wavelet.

As stated in [?], multi-resolution analysis based on the Fast Wavelet Transform algorithm computes the wavelet transform of a signal given a specified Quadrature Mirror Filter bank (QMF), with such filters being related to a single mother wavelet. Hence, there is a sort of mapping between wavelet and QM Filters. In their work, Sherlock et al. [7] have proposed an angular parametrization of QMFs and they have shown that any orthonormal wavelet decomposition can be generated using a proper set of angles. They have also introduced an algorithm for computing a QM Filter given some angular parameters. They also demonstrated that a $2M$ filter coefficients $\{h_i\}$ can be expressed in terms of M angular filters. Furthermore, they proved that in order for the QM Filter to generate an orthonormal wavelet basis, the constraint $\sum_i \theta_i = \frac{\pi}{4}$ has to be satisfied, which reduces the choice to $M-1$ free parameters. Figure 1 shows examples of wavelets generated by the Sherlock-Monro algorithm for $M = 4$. The two upper wavelets were generated with angular parameters $\frac{\pi}{2}$ and $\frac{\pi}{3}$ and the lower ones were selected during our experiments.

Now, for any vector θ of size $M-1$, with $\theta_i \in [0; 2\pi]$, we can define the kernel :

$$K_{\theta,s,t}(x,x') = \langle \phi_{\theta,s,t}, x\rangle \langle \phi_{\theta,s,t}, x'\rangle = c_{\theta,s,t}c'_{\theta,s,t} \qquad (3)$$

2.2. MKL for large margin wavelet selection

As defined in equation (1), the kernel $K_{\theta,s,t}$ is related only to a single wavelet. Hence, in order to take into account the full decomposition, we have to consider the set of kernel $\{K_{\theta,s,t}\}_{s,t}$ where s and t respectively spans all possible wavelet dilations and translations.

In this work, we build upon the Multiple Kernel Learning framework [8, 9] for combining all the kernels resulting for a full wavelet decomposition and for selecting the "best" wavelet mother by tuning appropriately the angular parameter θ. In the MKL framework, supposing that we have a training set $\{x_i, y_i\}_{i=1}^{\ell}$, our objective is to learn a decision function of the form

$$f(x) = \sum_{m \in \mathcal{M}} f_m(x; \theta_m, s_m, t_m)$$

where $\mathcal{M}$ is a set of index, $\{\theta_m, s_m, t_m\}$ a triplet which specifies a wavelet ϕ_{θ_m,s_m,t_m} and $f_m(x) = \sum_i \alpha_i y_i K_{\theta_m,s_m,t_m}(x, x_i)$.

Then, our problem boils down to the choice of the "large-margin" combination of wavelet kernel . In other word, we seek the best angular, scale and translation parameters for wavelet-based kernels that can be combined to maximize a large margin criterion. According to MKL, the problem to solve is then :

$$\begin{aligned} \min_{\{f_m\},b,\xi,d} \quad & \frac{1}{2}\sum_m \frac{1}{d_m}\|f_m\|^2_{\mathcal{H}_m} + C\sum_i \xi_i \\ \text{s.t.} \quad & y_i \sum_m f_m(x_i) + y_i b \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \\ & \sum_m d_m = 1 \ , \quad d_m \geq 0 \quad \forall m \ , \end{aligned} \qquad (4)$$

where each d_m controls the squared norm of f_m in the objective function. Within this framework, the decision function has the form :

$$f(x) = \sum_i \alpha_i y_i \left(\sum_{m \in \mathcal{M}} d_m K_{\theta_m,s_m,t_m}(x, x_i) \right) + b$$

Note that as explained in [9], the role of the penalty on d_m is to yield a sparse combination of kernels by setting to 0 many of these weights d_m.

The mother wavelet parametrization θ plays a central role in this problem and in how it can be solved. Indeed, in the Sherlock-Monro algorithm, θ can be considered as a continuous parameter.

Hence, in this sense, the number of kernels we have to deal with becomes infinite, and the problem becomes an Infinite Kernel Learning problem as introduced by Gehler et al. [?] instead of a multiple kernel learning.

Here, we consider a finite set of $\{\theta\}$ sampled from the space $[0, 2\pi]^{M-1}$. By doing so, we slighty relax the problem since we deal with a finite, although exponential, number of kernels. Owing to this approximation, we are able to apply the below-presented novel MKL algorithm that can handle a large number of kernels.

3. ACTIVE CONSTRAINTS MKL

3.1. Solving the main problem

The problem given in equation (4) has a smooth and convex objective function and it has linear constraints. As stated [9], it can be reformulated as the following optimization problem :

$$\begin{aligned} \min_{d} J(d) = & \begin{cases} \min\limits_{\{f_m\},b,\xi} \dfrac{1}{2}\sum\limits_m \dfrac{1}{d_m}\|f_m\|^2_{\mathcal{H}_m} + C\sum\limits_i \xi_i \\ \text{s.t. } y_i \sum\limits_m f_m(x_i) + y_i b \geq 1 - \xi_i \quad \forall i \\ \xi_i \geq 0 \quad \forall i \end{cases} \\ \text{s.t. } & \sum_m d_m = 1 \,, \quad d_m \geq 0 \quad \forall m \ , \end{aligned} \tag{5}$$

which is a non-linear problem with respects to d with simplex like constraints. We stress that $J(d)$ is a differentiable convex function but its evaluation needs the solution of a dual SVM problem with kernel $K = \sum_m d_m K_m$.

In what follows, we propose a novel approach for solving MKL problem that it is expected to improve existing MKL algorithms efficiency when dealing with many kernels but only few of them will be get positive weights d_m. Indeed, the KKT conditions of this optimization problem (5) imply that at optimality, we have

$$\begin{aligned} \frac{\partial J}{\partial d_m} &= -\lambda \quad \text{if } d_m > 0 \\ \frac{\partial J}{\partial d_m} &\geq -\lambda \quad \text{if } d_m = 0 \end{aligned} \tag{6}$$

with λ being the Lagrangian multiplier associated to the equality constraint and $\frac{\partial J}{\partial d_m} = -\frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j K_m(x_i, x_j)$. Owing to the sparsity-inducing penalization term on d_m, we note from the optimality conditions that, all kernels with nonzeros d_m, at optimality, should have their gradients at the same value. Hence, from these equations, we can evaluate whether a couple $\{d, \alpha\}$ is optimal or not.

Note that our constrained optimization problem is simpler if we know in advance which kernels are going to be active and which are not. Indeed, we can simply run a MKL algorithm using these active kernels in order to learn both the decision function and the kernel weights. Active constraint approaches for constrained optimization consists in starting from a guess on the kernel active sets and then iteratively update this set until optimality. Basically, the idea consists in optimizing MKL among a small working set of kernels (the active ones) and then in verifying if the resulting solution $\{d, \alpha\}$ satisfies all the other constraints given in equation (6).

Our approach is detailed in Algorithm 1. Denote by $\mathcal{K}_A = \{m | \, d_m > 0\}$ and by $\mathcal{K}_0 = \{m | \, d_m = 0\}$ with $\mathcal{K}_A \cap \mathcal{K}_0 = 0$. The active constraints algorithm consists then at each iteration in : i) train a MKL problem using only the working $\mathcal{K}_A$ kernel set, ii) check the optimality of the full problem, iii) then, if not optimal yet update the working set $\mathcal{K}_A$ by including a kernel which violates the KKT constraint. One can remove from $\mathcal{K}_A$ some coordinate so that $d_i = 0$. Steps i) to iii) are performed until equations (6) are satisfied up to a tolerance ε.

Algorithm 1 Active contraints MKL

```
Set $d_m = 0, \forall m$
Initialize randomly some $\mathcal{K}_A$ so that $\sum_{m \in \mathcal{K}_A} d_m = 1$
while not optimal do
    Solve MKL with according to $\mathcal{K}_A$ kernels
    Check KKT conditions of the full problem
    if not optimal then
        Among violating constraints choose, $d_u$
        $\mathcal{K}_A \leftarrow \mathcal{K}_A - \{i \in \mathcal{K}_A : d_i = 0\}$
        $\mathcal{K}_A \leftarrow \mathcal{K}_A \cup u$
    end if
    if $|\mathcal{K}_A|$ = maximal number of kernels then
        Break
    end if
end while
```

3.2. Optimality and Wavelet kernel selection

The main point of the algorithm is the update stage of the active kernel set. Indeed, if the solution is not optimal yet, it means that some gradients $\frac{\partial J}{\partial d_m}$ do not satisfy equations (6) up to a tolerance ε. Note that since the active set kernels have been optimized on step i), non-optimality can only occur for kernels that do not belong to active set, that is to say :

$$\exists m \in \mathcal{K}_0 \text{ so that } \quad \frac{\partial J}{\partial d_m} < -\lambda - \varepsilon$$

In our algorithm, for checking optimality and for selecting the kernel to add to the active set, we look for

$$u = \underset{m \in \mathcal{K}_0}{\operatorname{argmin}} -\frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j K_{\theta_m, s_m, t_m}(x_i, x_j) \tag{7}$$

and if $\frac{\partial J}{\partial d_u} < -\lambda - \varepsilon$, u is added to the active set $\mathcal{K}_A$, on the contrary, optimality conditions are satisfied. Resolution of problem 7 poses an important issue. Indeed, the problem is clearly non-convex and has discrete parameters.

In our experiments, since θ_m is a small-dimension vector, solving Equation 7 can still be performed by exhaustive search. Hence, for each θ_m, we do the wavelet transform of all training signals, build the corresponding Gram matrix for each couple $\{s_m, t_m\}$ and then compute $\frac{\partial J}{\partial d_m}$.

Note that according to this scheme, at each update stage, wavelet kernels that are added to the active set have different shapes (as parametrized by θ_m) and focus on different time-frequency locations (as parametrized by s_m and t_m).

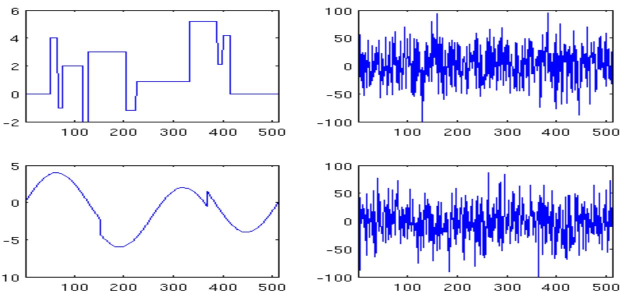

Fig. 2. Basis signals (left) of the toy dataset and their version corrupted with a Gaussian noise (right)

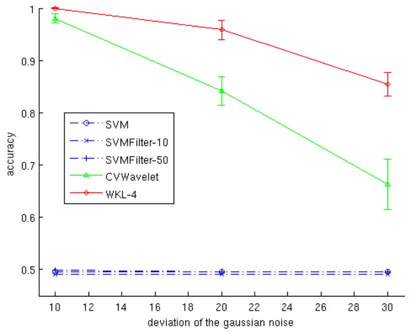

Fig. 3. Mean classification rate and standard deviation against standart deviation of the noise

4. EXPERIMENTAL RESULTS

The results presented here were obtained on a toy dataset of 1000 signals. In the figure 2, we show example of the basis signals to be classified (Blocks and HeaviSine of the Wavelab toolbox) and their corrupted version.

The generation of the dataset being quite simple, a naive approach would be to preprocess the signals with average filters. Hence we implemented several sizes of filters and classify the output with gaussian kernel, those methods are called *SVMFilter-N* (N being the size of the average filter). For the purposes of comparison, we have implemented the method described in [5] that is named *CVWavelet* in this section. This method uses a Gaussian kernel on features based on marginals of a DWT and applies a cross-validation process in order to choose the parameter θ of the QMF and the parameters of the kernel. In our experiments, we limited our method, called *WKL-4*, to add at most 15 kernels in the active constraint algorithm. Moreover, the kernels were generated with QMF filters of size 4.

Every method was trained on 100 signals selected in the dataset and tested on the remaining others. Classification rate have been computed as an average of 10 runs. The experiments were repeated three times with increasing noise variance. As we expected, the classification rate of each method drops as the variance of the noise rise. As illustrated on figure 3, *CVWavelet* and our method show more robustness to noisy data. Moreover, as is shown in the figure 3, our method achieves consistently better results than *CVWavelet*.

5. CONCLUSION

In this paper, we propose a novel approach to the problem of wavelet feature extraction for signal classification. Moreover, we have described an optimization algorithm able to handle such a problem.

Our active constraint algorithm performed well but its wavelet kernel selection step may be improved. In the next studies, we will try to overcome this issue.

Moreover, our work focuses on parametrized wavelet transforms but our method is general enough to be applied to other parametrized transforms. In future work, we plan to investigate other parametrized transforms in order to make our method more invariant to signal translation.

6. REFERENCES

[1] A. Cabrera and K. Dremstrup, “Auditory and spatial navigation imagery in braincomputer interface using optimized wavelets,” *Journal of Neuroscience Methods*, vol. 174, no. 1, pp. 135–146, 2008.

[2] Q. Xu, H. Zhou, Y. Wang, and J. Huang, “Fuzzy support vector machine for classification of eeg signals using wavelet-based features,” *Medical Engineering and Physics*, vol. 31, no. 7, pp. 858–865, 2009.

[3] N. Saito and R. Coifman, “Local discriminant bases and their applications,” *Journal of Math. Vision and Imaging*, vol. 5, no. 4, pp. 337–358, 1995.

[4] J. Neumann, C. Schnorr, and G. Steidl, “Efficient wavelet adaptation for hybrid wavelet-large margin classifiers,” *Pattern Recognition*, vol. 38, no. 11, pp. 1815–1830, 2005.

[5] M-F. Lucas, A. Gaufriau, S. Pascual, C. Doncarli, and D. Farina, “Multi-channel surface emg classification using support vector machines and signal-based wavelet optimization,” *Biomedical Signal Processing and Control*, vol. 3, no. 2, pp. 169 – 174, 2008.

[6] S. Mallat, *A wavelet tour of signal processing*, Academic Press, 1998.

[7] B. G. Sherlock and D. M. Monro, “On the space of orthonormal wavelets,” *IEEE Transactions on Signal Processing*, , no. 46, pp. 1716–1720, 1998.

[8] G. Lanckriet, N. Cristianini, L. El Ghaoui, P. Bartlett, and M. Jordan, “Learning the kernel matrix with semi-definite programming,” *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

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

[10] D. Vautrin, X. Artusi, M-F Lucas, and D. Farina, “A novel criterion of wavelet packet basis selection for signal classification with application to bci,” *IEEE Trans. Biomedical Engineering*, pp. 1–5, to appear.