Large Margin Filtering for Signal Sequence Labeling

Rémi Flamary, Benjamin Labbé, Alain Rakotomamonjy

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

Signal Sequence Labeling

Problem: Obtaining a label for each sample of a signal while taking into account the sequentiality of the samples.
Current approaches:

- Hidden Markov Models [1] ,Conditional Random Fields [2].
- Segment the signal (change detection [3]) and decide the label of the regions afterward.

Our approach

- Take into account the temporal neighborhood of the sample in the decision (time-delay embedding).
- Jointly learn a temporal filter with the classifier: adapt to noise and delay.

Definitions

- $X \in \mathbb{R}^{N\times d}$ Feature matrix, d channels and N samples.
- $X_{i,j}$ value of channel j for the i^{th} sample.
- $\mathbf{y}_i$ label of the i^{th} sample.
- Filtering X w.r.t. F

$$\widetilde{X}_{i,j} = \sum_{m=1}^{f} F_{m,j}\, X_{i+1-m+n_0,j} \qquad (1)$$

$F \in \mathbb{R}^{f\times d}$ Filter matrix, d filters of length f

n_0 filter delay

- $\|.\|_{\mathcal{F}}$ is the Frobenius norm of a matrix.

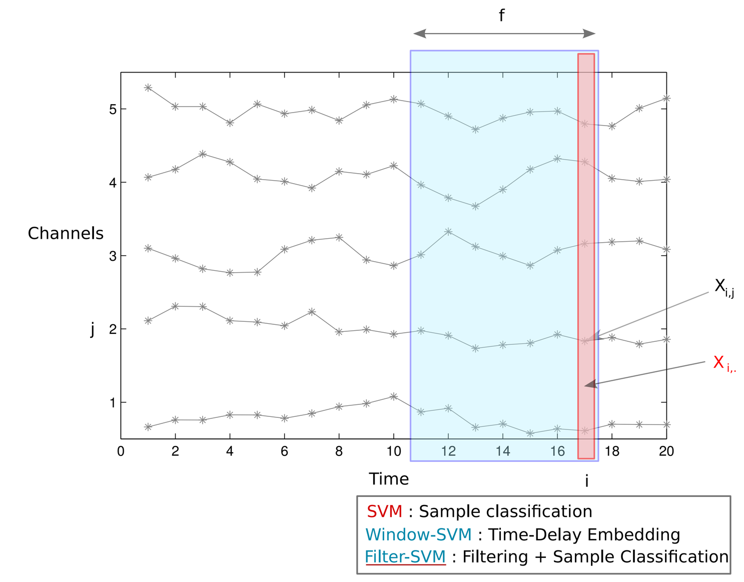

Figure 1: Sample classification *vs* Time window classification (at time i)

Window-SVM

- We learn a classifier (W, w_0) for a window of samples (Time-Delay Embedding).
- Decision function for the i^{th} sample of X:

$$g_W(i, X) = \sum_{m=1}^{f}\sum_{j=1}^{d} W_{m,j} X_{i+1-m+n_0,j} + w_0 \qquad (2)$$

where $W \in \mathbb{R}^{f\times d}$ and $w_0 \in \mathbb{R}$ are the classification parameters and f is the size of the time-window.

- Optimal function $g_W(.)$ obtained by minimizing:

$$J_{WSVM}(W) = \frac{1}{2}\|W\|^2_{\mathcal{F}} + \frac{C}{2}\sum_{i=1}^{N} H(\mathbf{y}, X, g_W, i)^2 \qquad (3)$$

w.r.t. (W, w_0) with $H(\mathbf{y}, X, g, i) = \max(0, 1 - \mathbf{y}_i g(i, X))$.

Filter-SVM

- We jointly learn a sample classifier $(\mathbf{w}, w_0)$ and a filtering F of the channels.
- Decision function for the i^{th} sample of X:

$$g_F(i, X) = \sum_{m=1}^{f}\sum_{j=1}^{d} \mathbf{w}_j F_{m,j} X_{i+1-m+n_0,j} + w_0 \qquad (4)$$

where $\mathbf{w}$ and w_0 are the parameters of the linear SVM classifier corresponding to a weighting of the channels.

- Optimal function $g_F(.)$ obtained by minimizing:

$$J_{FSVM} = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n} H(\mathbf{y}, X, g_F, i)^2 + \frac{\lambda}{2}\|F\|^2_{\mathcal{F}} \qquad (5)$$

w.r.t. $(F, \mathbf{w}, w_0)$ where λ is a regularization term.

Filter-SVM Solver

- Cost non-convex but convex w.r.t. $\mathbf{w}$ and w_0 when F is fixed.
- We define $J(F)$ that is differentiable [4]:

$$J(F) = \min_{\mathbf{w}, w_0} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n} H(\mathbf{y}, X, g_F, i)^2$$

- We minimize:

$$J(F) + \frac{\lambda}{2}\|F\|^2_{\mathcal{F}}$$

w.r.t. F using a gradient descent along F and a line search to find the optimal step.

Numerical Experiments on toy dataset

- *nbrel* discriminative signals with a switching mean $(-1, 1)$ among $nbtot = d$.
- σ Gaussian noise and time-lags applied to the channels.
- $f = 21$ and $n_0 = 11$ corresponding to a good average filtering.
- Test error is the average of 10 runs.

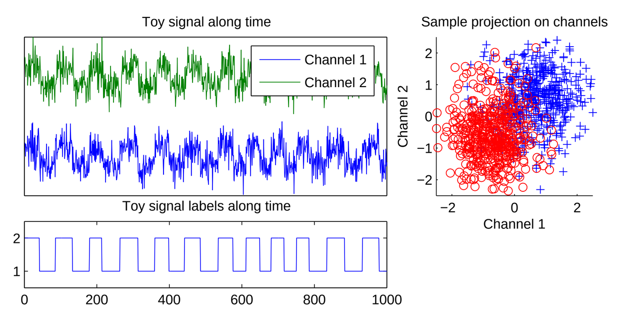

Figure 2: Toy example for $nbtot = nbrel = 2$ and $\sigma = 1$.

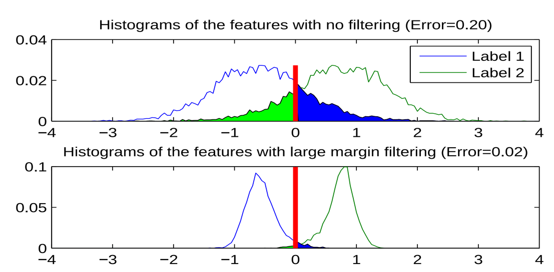

Figure 3: Histograms of the samples for the 2 possible labels of a 1D signal (with / without filtering).

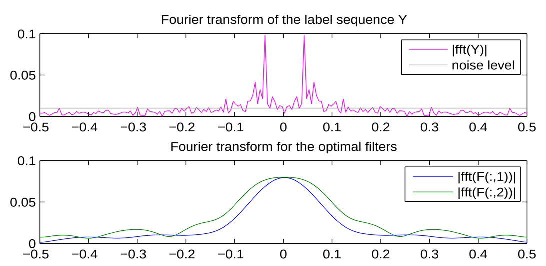

Figure 4: Fourier Transform of the discriminative information and of the impulse response of learned filters

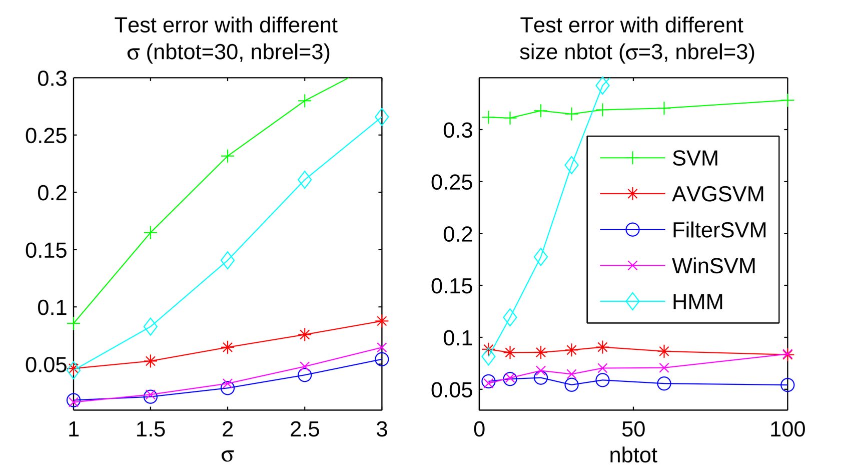

Figure 5: Test error for different σ values and for different number of channels *nbtot*.

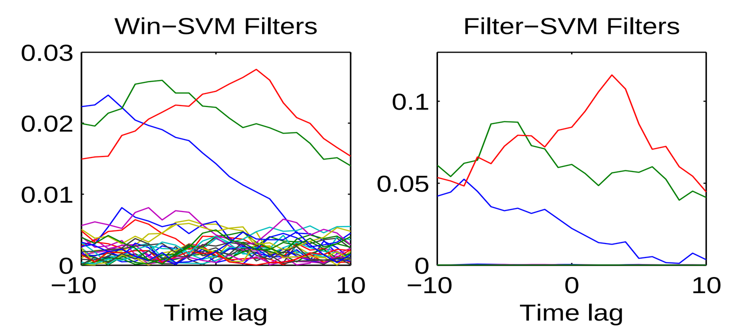

Figure 6: Coefficients of W (left) and coefficients F weighted by w (right).

Numerical Experiments on BCI dataset

- 3 classes, 3 subjects/tasks, 96 PSD channels.
- 9000 training samples, 3000 test samples.
- Test Error:

Method	Parameters	Sub 1	Sub 2	Sub3	Avg
BCI Comp.		0.2040	0.2969	0.4398	0.3135
SVM		0.2877	0.4283	0.5209	0.4123
Filter-SVM	$f = 8, \quad n_0 = 0$	0.2337	0.3589	0.4937	0.3621
	$f = 20, \quad n_0 = 0$	0.2021	0.2693	0.4381	0.3032
	$f = 50, \quad \mathbf{n_0 = 0}$	**0.1321**	**0.2382**	**0.4395**	**0.2699**
	$f = 100, \mathbf{n_0 = 50}$	**0.0537**	**0.1659**	**0.3859**	**0.2018**
Avg-SVM	$f = 100, n_0 = 50$	0.1544	0.2235	0.3870	0.2550

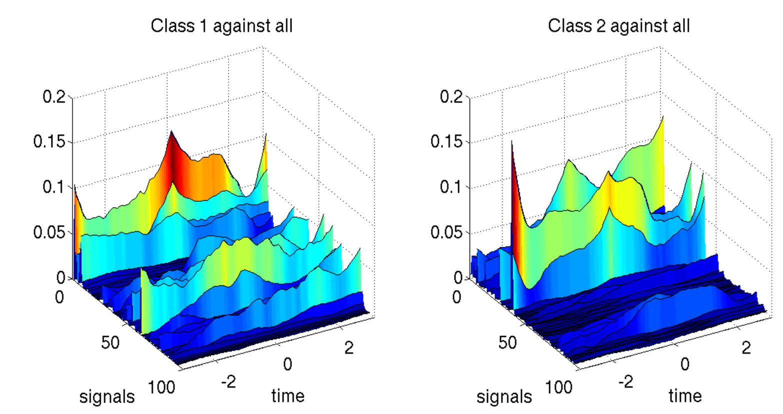

Figure 7: Discriminative Channel/Delay maps on BCI (F filter for subject 1): label 1 against all (left) and label 2 against all (right).

Conclusion

- On-line sequence labeling classifier.
- Better variable selection than classical SVM.
- Visualization of space/time discriminative maps.

Future works

- Non-linear SVM (kernel).
- Multi-task approach for F.

Bibliography

- O. Cappé, E. Moulines, and T. Rydèn. *Inference in Hidden Markov Models.* Springer, 2005.
- J. Lafferty, A.McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289, 2001.
- F. Desobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53:2961–2974, August 2005.
- J.F. Bonnans and A. Shapiro. Optimization problems with pertubation : A guided tour. *SIAM Review*, 40(2):202–227, 1998.