ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Translation invariant classification of non-stationary signals

Vincent Guigue, Alain Rakotomamonjy and Stéphane Canu *

Laboratoire Perception, Systèmes, Information - CNRS - FRE 2645
Avenue de l'Université, 76801 St Étienne du Rouvray

Abstract. Non-stationary signal classification is a difficult and complex problem. On top of that, we add the following hypothesis: each signal includes a discriminant waveform, the time location of which is random and unknown. This is a problem that may arise in Brain Computer Interface (BCI). The aim of this article is to provide a new description to classify this kind of data. This representation must characterize the waveform without reference to the absolute time location of the pattern in the signal. We will show that it is possible to create a signal description using graphs on a time-scale representation. The definition of an inner product between graphs is then required to implement classification algorithm. Our experimental results showed that this approach is very promising.

1 Introduction

The classification of non-stationary signals is a common and difficult problem in signal processing. Classical statistical descriptors like mean, or Fourier coefficients are not efficient descriptors for such data where time-dependent information are needed. Usual approaches consist in using Time-Frequency (TFRs) or Time-Scale Representations (TSRs). Both solutions lead to a high dimensional classification problem. Furthermore, if the discriminative part of the signal is a transient signal, the time location of which is unknown and variable, then a translation invariant classifier is required. This situation occurs in a classical Brain Computer Interface (BCI) problem: the P300 speller paradigm [1].

Various time-frequency strategies have been implemented within this context. The modulation frequency defined by Sukittanon et al. [2] can be seen as a model of each TFR frequency. Hory et al. [3] propose to model the TFR using a mixture of χ^2 distributions, to focus on discriminant patterns. Michel et al. [4] use a graphical structure to characterize the pattern skeleton of the TFR. Another approach consists in working in the time-scale plane: Mallat [5] introduced a TSR based on wavelet maxima. Saito and Coifman [6] propose to optimize the representation to improve the classification results. Crouse et al. [7] obtain good results for signal classification based on the Hidden Markov Model (HMM) for each TSR scale.

*This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

We propose to use TSR in order to cope with the non-stationary signals. Then we will reduce the representation dimension, by means of TSR Gaussian modeling. Finally, the building of a graph between Gaussian models introduces comparative time information, which enable us to deal with the translation invariance problem.

Davy et al. [8], show the interest of using Support Vector Machines (SVMs) for non-stationary signal classification. Since we are using a graph representation of the TSR, we will adapt the graph kernel of Kashima et al. [9] as a SVM kernel. We will compare the results with the k nearest neighbors (knn) universal classifier. SVMs use the above mentioned inner products, and knn use the induced distances.

Section 2 presents the data used in this paper and the details of the graphical description. Section 3 deals with the definition of the inner product in graph space. We will focus on alternative methods in section 4. Finally, we will compare the results for classifying the signals of the different algorithms (section 5) and give some conclusions (section 6).

2 Data and data description

2.1 Building the simulated data

In many signal classification applications, data are composed of a pattern and noise. The pattern shape is characteristic, whereas its time location is unknown and random. For instance, when we try to identify a response to a stimulus in an electroencephalogram (EEG) [1], the time location of this response is unknown.

Hence, we worked on artificial data that present such characteristics. The signals $S(t)$ (exponential decreasing chirps) are generated according to:

$$S(t) = m_{u,v}(t-\tau)\Gamma(t-\tau) + b(t) \text{ with: } m_{u,v}(t) = e^{-\alpha t}\cos((u+vt)t+\phi) \quad (1)$$

where $\Gamma(t)$ is the step function and $b(t)$ a Gaussian white noise (standard deviation σ_b).

The objective of this work is to discriminate two classes of these signals which vary in u and v. In class 1 we have $u = 1 \cdot 10^{-3}, v = 2 \cdot 10^{-3}$ and in class -1 we have $u = 5 \cdot 10^{-4}, v = 6 \cdot 10^{-4}$. τ is drawn according to uniform distribution. We compare two datasets with $\sigma_b = 0.02$ and $\sigma_b = 0.2$, which lead respectively to signal to noise ratio (SNR) of 14.08dB and -25.75dB.

2.2 Time-scale representation

Time-scale representation is a decomposition of a signal $S(t)$ over elementary functions that are well concentrated in time and frequency [5]. Given a wavelet ψ located on scale a and time b, let $P_{a,b}$ be the projection of $S(t)$ over the analytic function $\psi_{a,b}$. The set $\{P_{a,b}\}$ of coordinate $\{a, b\}$ define the time-scale plane. Finally, the time-scale plane is divided into k parts, with: $k = \text{card}(a)\text{card}(b)$. For more clarity, we re-index the notations: the set of coefficients becomes

$\{P_i\}_{i=1,\ldots,k}$ of time-scale coordinates (b_i, a_i). Squared coefficients P_i^2 correspond to the local energy.

Translation covariance is required to face the hypothesis of a pattern, the time location of which is random. A translation covariant representation r must verify: $r(S(t-\tau)) = r_\tau(S(t))$ where r_τ is the translated representation. Orthogonal wavelet transforms do not verify this property. For this reason we use continuous wavelet transform (CWT) to solve this problem (Fig. 2).

2.3 Graphs representation

Although time is important to describe the pattern, it is a penalty factor due to the random time position of the patterns. The solution consists in filtering discriminant information by using comparative time between selected regions of the plane. This novel graph representation r relies on a set of nodes $H = \{h_i\}_{i=1,\ldots,k}$ (with $h_i = \{P_i, a_i\}$) and a matrix $E = \{e_{h_i h_j}\}$ which defines the arcs between the nodes ($e_{h_i h_j} = \Delta_{t_{ij}} = b_j - b_i$):

$$r : \begin{array}{ccc} \mathcal{F} & \rightarrow & \mathcal{G} \\ \{P_i, a_i, b_i\} & \rightarrow & \{h_i\}, E \end{array} \qquad (2)$$

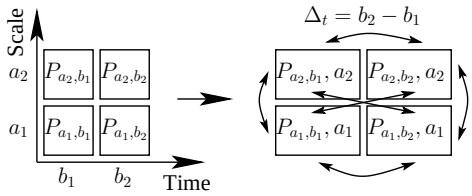

Fig. 1: From time-scale to graph representation. Scale a_i and coefficient P_i are gathered in the node h_i in graph space $\mathcal{G}$, the time description switch from absolute references b_i to comparative time measures $\Delta_{t_{ij}} = b_j - b_i$.

This graph is fully linked, each coefficient of the wavelet transform is a node linked to all the other coefficients. Nodes comprise scale location and weight informations, arcs are labeled with comparative time information. This graph is very large: it counts k nodes and $k(k-1)$ arcs.

Representation r has a very high dimensionality and the computation complexity of the inner product between graphs is closely related to the number of nodes in the graphs. Hence, it is necessary to reduce the representation size in order to compute efficiently the inner product in the graph space.

We limit the field of application to the case where the discriminant information lies in high energy regions of the time-scale representation, without taking into account possible interferences. Hence, we propose to reduce the dimensionality of the graph by modeling the CWT as a Gaussian mixture [10]. The modeling will enable us to reduce the dimension while keeping the TSR shape, focusing on high energy regions of the time-scale plane [3]. The nodes are composed of the covariance matrix of the region, the sum of the coefficients in the region, and the scale location of the Gaussian. The arc labels are Δ_t, the comparative time position between the Gaussian models (Fig. 2).

S. Mallat [5] showed that the local maxima in the continuous wavelet transform allows us to rebuild a denoised signal. This high energy modeling will reduce the size of the representation as well as the original signal noise. For more details about this graph representation, the reader can refer to [11].

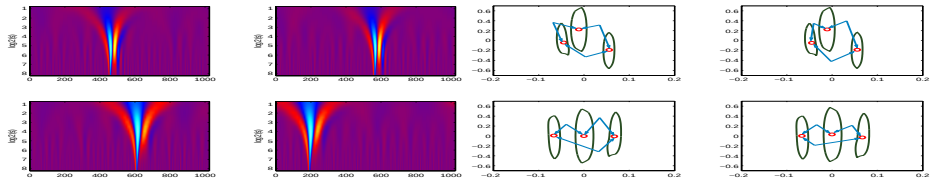

Fig. 2: Continuous wavelet transforms and graph representations. The two signals on top belong to class 1, the two bottom ones belong to class -1. Low noise level (SNR=14.08dB).

3 Distance and inner product between graphs

We use the inner product between graphical representations based on Kashima et al. article [9]. The idea is to compare two label sequences generated by two synchronized random walks on the two graphs. The operation is repeated until there is convergence of the result. The detailed computation of $K(G^1, G^2)$ is given in the paper of Kashima et al. . We use the norm induced by the inner product to define a distance between graphs.

4 Alternatives

In order to validate our approach, we implemented three other descriptions for the data.

Classical representations To justify a complex and costly approach, we implemented the simplest available description: the raw data. Another common description is based on statistical descriptors.

Translation arrangement It is clear that a simple inner product $\langle r(S_1), r(S_2)\rangle$ is not able to face the hypothesis of an unknown time located pattern. Hence, we build the set of S_2 translated representation $r_\tau(S_2)$ and we aim to find the highest inner product between the representation of S_1 and the set $r_\tau(S_2)$. Ideally, we would use the following translation invariant inner product:

$$\langle S_1, S_2\rangle = \max_{\tau\in\Omega} \left(k(r(S_1), r_\tau(S_2))\right) \tag{3}$$

where Ω is the set of translations. We apply this idea for both raw signals and wavelet transform. Given the fact that r is translation covariant, we only need to compute $r(S_2)$ once. However, this method is very slow when Ω becomes large.

Bag of vectors This method uses the nodes of the graph representation. Each

node is a vector containing the parameters of a Gaussian model. We build a representation which gathers those vectors. The arcs Δ_t are not taken into account. The inner product tries to match pairs of vectors from S_1 and S_2 representations. Wallraven et al. [12] have designed such a kernel:

$$K_{match} = \frac{1}{2}\left(K + K^T\right) \text{ with: } K(S_1, S_2) = \frac{1}{n_1}\sum_{i=1}^{n_1} \max_j \left(k(V_i^1, V_j^2)\right) \quad (4)$$

where n_k is the number of nodes, V_i^k is the i^{th} Gaussian parameter vector in time-scale representation of S_k.

Because of the max function, kernels in equation (3) and (4) are not positive definite. We approximate $\max_{y_j \in \mathcal{Y}} \left(k(x_i, y_j)\right)$ by $\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|} \exp\left(-\frac{||x_i - y_j||^2}{2\sigma^2}\right)$. Using this formulation and considering the bag of vectors as different instances of the same signal, this kernel can be seen as a Multiple Instance Kernel [13].

5 Results

|Learn.|/|Test.|	**1/1000**	**400/1000**	
Classifiers	**1-nn and SVM**	**1-nn**	**SVM**
Coef.	49.59% (±4.74)	30.76% (±1.53)	31.14% (±1.43)
Raw sig.	49.99% (±0.84)	48.10% (±2.01)	47.08% (±0.97)
Stat. descr.	47.16% ±(8.22)	35.54% (±2.12)	19.27% (±0.98)
T.A. (coef.)	46.12% (±3.39)	24.17% (±1.54)	23.57% (±1.38)
T.A. (sig.)	49.73% (±1.15)	43.17% (±1.99)	42.27% (±1.63)
Bag of vect.	29.40% (±15.24)	11.9% (±1.13)	8.29% (±0.80)
Graph	**13.98% (±13.26)**	6.66% (±0.64)	**5.25% (±0.35)**

Table 1: Misclassification rate on the test set, average over 30 runs. High noise level (SNR=-25.75dB). |Learn.|/|Test.|: number of data in learning and test set for each class, T.A.: translation arrangements .

The misclassification rates are given in table table 1 for highly noised signals (SNR=-25.75dB), using Support Vector Machine (SVM) [14] and k-nearest neighbors (knn). Results are averaged over 30 runs, on a 1000 signal test set. The size of learning set is variable (between 1 and 400 signals). The classification problem on high SNR signals (SNR = 14.08dB) is trivial even with the one signal learning set. Graph and bag of vectors kernel achieved 100% correct classification and demonstrate their abilities to describe the discriminant pattern whatever its position in the signal.

Table 1 shows that graph kernel combined with SVM is the best method for this problem, with the lowest misclassification rate and the lowest variance of the results. Graph kernel outperformed bag of vectors kernel by more than 3% whereas the only difference between the two methods is the time information.

6 Conclusions

Non vectorial descriptors open new perspectives in various fields. In the case of non-stationary patterns, randomly located in the signal, the graphical representation enable us to keep a time description without absolute reference. The results point out the interest of such a description and the efficiency of the graph kernel. This representation can be interesting for other applications in the field of non stationary signal classification. As a matter of fact, it is compact and focuses on discriminant part of the TSR.

One perspective of this work is to explore different ways to build the graph: we need to define a criterion (like Fisher's one) to determine which parts of the plane (or which Gaussians) are discriminants. This will enable us to treat the case where discriminant information resides in low energy part of the plane.

References

[1] L.A. Farwell and E. Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. In *Electroencephalography and Clinical Neurophysiology*, volume 70, pages 510–523, 1988.

[2] S. Sukittanon, L. Atlas, J. Pitton, and J. McLaughlin. Non-stationary signal classification using joint frequency analysis. In *ICASSP*, volume 6, pages 453–456, 2003.

[3] C. Hory. *Mixtures of Chi2 distributions for the interpretation of a time frequency representation*. PhD thesis, INPG, 2002.

[4] O. Michel, P. Flandrin, and A. Hero. Automatic extraction of time-frequency skeletons with minimal spanning trees. In *ICASSP*, volume 1, pages 89–92, 2000.

[5] S. Mallat. *A Wavelet Tour Of Signal Processing*. Academic Press, 1997.

[6] N. Saito and R.R. Coifman. Local discriminant bases. In *Wavelet Applications in Signal and Image Processing, Proc. SPIE 2303*, pages 2–14, 1994.

[7] M. Crouse, R. Nowak, and R. Baraniuk. Wavelet-based statistical signal processing using hidden markov models. *IEEE Transactions on Signal Processing*, 46(4):886–902, 1998.

[8] M. Davy, A. Gretton, A. Doucet, and P.J.W. Rayner. Optimised support vector machines for nonstationary signal classification. *IEEE Signal Processing Letters*, 9(12):442–445, December 2002.

[9] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *20th International Conference on Machine Learning*, pages 321–328. AAAI Press, 2003.

[10] H. Greenspan, J. Goldberger, and L. Ridel. In *Computer Vision and Image Understanding*, volume 84, pages 384–406, 2001.

[11] V. Guigue, A. Rakotomamonjy, and S. Canu. Translation invariant classification of non-stationary signals. Technical report, Lab. PSI, INSA de Rouen, `http://asi.insa-rouen.fr/ vguigue`, 2004.

[12] C. Wallraven, B. Caputo, and A.B.A. Graf. Recognition with local features : the kernel recipe. In *ICCV 2003 Proceedings*, volume 2, pages 257–264. IEEE Press, 2003.

[13] T. Gärtner, P.A. Flach, A. Kowalczyk, and A.J. Smola. Multi-instance kernels. In *ICML*, pages 179–186, 2002.

[14] V. N. Vapnik. *The Statisitcal Learning Theory*. Springer, 1998.