Proceedings of the IV Brazilian Conference on Neural Networks - IV Congresso Brasileiro de Redes Neurais
pp. 084-089, July 20-22, 1999 - ITA, São José dos Campos - SP - Brazil

A Competitive Wavelet Layer for Pattern Clustering

Roberto Kawakami Harrop Galvão, Takashi Yoneyama
CTA - ITA - Div. Eletrônica - 12228-900 - São José dos Campos - SP
E-mails: kawakami@ele.ita.cta.br, takashi@ele.ita.cta.br

Abstract

A competitive "wavelet layer" is proposed for pattern clustering. It exploits the representation capabilities of adaptive wavelets to generate template approximations for each cluster of data. A brief review of adaptive wavelet representations, as well as some insight into local minima problems, is provided. The method is illustrated by a simple clustering problem, in which step responses of dynamic systems are discriminated with basis on the presence of parasitic oscillations. The results suggest that the wavelet layer exhibits superior performance than the conventional competitive neural layers when patterns exhibit a low signal-to-noise ratio.

1. Introduction

Recently, important bridges have been established between the field of artificial neural networks (ANN's) and wavelets [1],[2]. Wavelet Theory comprises a set of techniques aimed at developing efficient representations of signals through the use of elementary functions that are localized both in frequency and in time [3]. A remarkable feature of wavelet-based signal processing is that it mimics several natural phenomena found in biological sensory systems [4], [5]. This has been a strong motivation for its use with Artificial Intelligence tools.

By spreading 1-D signals across a 2-D time-scale map, wavelets allow the identification of features that can be used to (1) represent signals in a more compact way (data compression), (2) separate signals from noise (signal restoration, or denoising), and (3) recognize/classify signals (pattern analysis). In particular, clustering may be more efficiently performed after data undergo a wavelet pre-processing [6].

This paper proposes an algorithm for wavelet-based clustering which employs a competitive training scheme. Unlike the majority of works in this field, the use of wavelets is not restricted to a pre-processing stage. Instead, the representation capabilities of adaptive wavelets [1] are exploited to synthesize a typical element, or "template" for each cluster of signals.

A brief introduction to adaptive wavelet representations is provided, in order to help the presentation of the key concepts. Initialization and training schemes are described, and some insight into the sources of local minima problems is given.

For illustration purposes, the proposed clustering technique is applied to a simple problem, in which step responses of first-order dynamic systems are discriminated with basis on the existence/absence of parasitic oscillations. The results suggest that the wavelet layer presents better performance than conventional competitive neural networks when the patterns to be clustered (1) are of an oscillatory nature and (2) exhibit a low signal-to-noise ratio.

2. Notation

$$(\cdot) \equiv \text{continuous variable } [\cdot] \equiv \text{discrete time index}$$
$$L^2(\mathbb{R}) \equiv \{f : \mathbb{R} \rightarrow \mathbb{R} \text{ s.t. } \int_{-\infty}^{+\infty} (f(t))^2\, dt < \infty\}$$
$$\langle f, g\rangle \equiv \int_{-\infty}^{+\infty} f(t)\overline{g(t)}dt$$
$$\hat{f}(\omega) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt$$

3. Mathematical Preliminaries

A fundamental result of wavelet theory states that, if a function $\psi \in L^2(\mathbb{R})$ satisfies the condition

$$\int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty \qquad (1)$$

then any finite-energy signal f can be recovered from its inner products with rescalled and time-shifted versions of ψ, that is:

$$f = C_{\psi}^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a^{-2} \langle f, \psi_{a,b}\rangle\, \psi_{a,b} da db$$

where:

$$\psi_{a,b}(\cdot) \triangleq |a|^{-2}\, \psi\left(\frac{\cdot - b}{a}\right);\ a \in \mathbb{R}^*, b \in \mathbb{R} \qquad (2)$$

$$C_{\psi} \triangleq 2\pi \int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega \qquad (3)$$

Parameters a and b are denominated *scale* and *translation*. Functions $\psi_{a,b}$ are called wavelets, derived from the mother wavelet ψ. Remark that reducing the scale has the effect of compressing ψ and spreading $\hat{\psi}$.

Eq.(1) essentially implies $\hat{\psi}(0) = 0$, i.e., ψ should have zero mean. Moreover, it can be shown [3], that, if ψ decays reasonably fast in the time and frequency domains, than there exists countable sets of scales A and translations B such that:

$$f = \sum_{a \in A, b \in B} w(a,b)\psi_{a,b} \tag{4}$$

Coefficients $w(a,b)$ can be obtained as the inner products of f with "decomposition" functions $\tilde{\psi}_{a,b}$. The reconstruction of f from these coefficients is numerically stable, so the wavelet series can be truncated at some point, when the approximation error is found to be acceptable. One then says that f is approximated by a finite combination of wavelets.

If f is given in a discrete-time form, the coefficients of its wavelet expansion can be obtained in a fast manner by using a special bank of digital filters. Such algorithm computes the coefficients in dyadic scales (powers of two) and employs successive downsampling operations to eliminate redundancy and thus increase the speed of calculus [7].

4. Adaptive Wavelet Representations

The filter bank method for finding the coefficients of a wavelet expansion has two main drawbacks:

1) Downsampling makes the coefficients variant to time shifts [7].

2) Signal features in intermediary scales (that is, scales that are not powers of two) may not be adequately represented in the coefficients (fig.1) [8].

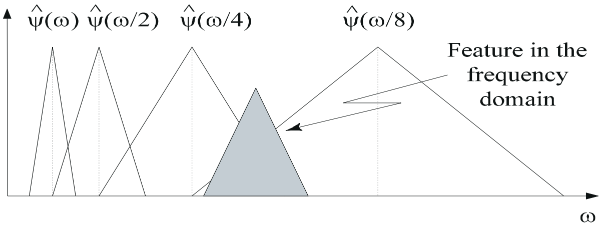


Figure 1: Some features may not be adequately represented by wavelets in dyadic scales.

An alternative procedure consists of approximating the signal by a finite combination of *adaptive wavelets* [1]. In this approach the sets of scales A and translations B are not chosen *a priori* (fig.2). Instead, they are obtained by using numerical optimization algorithms which seek to minimize the approximation error.

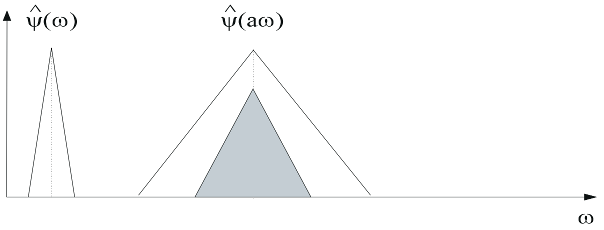


Figure 2: Adaptive wavelets are an alternative to represent features that are not located in dyadic scales.

Suppose that a discrete-time signal $f[\cdot]$ is to be approximated, over the interval $\mathcal{D}_f = [0, T] \subset \mathbb{Z}$, by a linear combination of K wavelets derived from a mother wavelet ψ. This representation problem can be stated as follows:

Find sets of scales $A = \{a_k; k = 1, ..., K\}$, translations $B = \{b_k; k = 1, ..., K\}$ and coefficients $W = \{w_k; k = 1, ..., K\}$ which minimize the cost function:

$$E(A,B,W) \triangleq \frac{1}{2}\sum_{t=0}^{T}\left(e_a^{A,B,W}[t]\right)^2 \tag{5}$$

where

$$e_a^{A,B,W}[t] \triangleq f[t] - \sum_{k=1}^{K} w_k \psi\left(\frac{t-b_k}{a_k}\right) \tag{6}$$

Remark that t denotes a discrete time index (a unity sampling time is assumed, without loss of generality). Also, since the number of wavelets is finite, the normalizing factor $|a_k|^{-1/2}$ was included in the coefficient w_k. For simplicity in deriving gradient equations, henceforth let $t_k \triangleq (t - b_k)/a_k$.

If one chooses the real Morlet mother wavelet (a modulated gaussian):

$$\psi(t) \triangleq \cos(\Omega t)\exp(-0.5t^2) \tag{7}$$

then the partial derivatives of the cost function with respect to the parameters involved are:

$$\frac{\partial E}{\partial a_k} = -\sum_{t=0}^{T} t_k M_k[t] \tag{8}$$

$$\frac{\partial E}{\partial b_k} = -\sum_{t=0}^{T} M_k[t] \tag{9}$$

$$\frac{\partial E}{\partial w_k} = -\sum_{t=0}^{T} e_a[t]\psi(t_k) \tag{10}$$

where

$$M_k[t] \triangleq e_a[t]\frac{w_k}{a_k}\left[\Omega\sin(\Omega t_k)e^{-0.5t_k^2} + t_k\psi(t_k)\right] \tag{11}$$

The optimization technique adopted in the present work is the conventional Gradient-Descent method, i.e.:

$$X^{i+1} = X^i - \lambda_X \nabla_X^i E \tag{12}$$

where $\nabla_X^i E$ is the gradient of the cost with respect to parameter vector X ($X = A, B$, or W). Different step sizes λ_X may be required for A, B, and W.

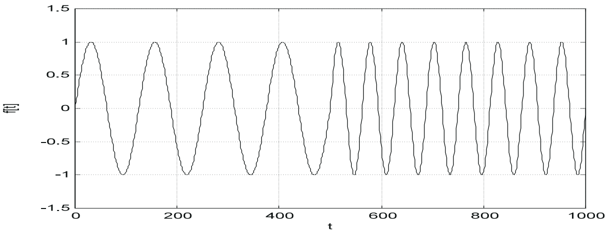


Figure 3: Example signal.

4.1. Parameter Initialization

Good parameter initialization is very important to accelerate the convergence of numerical optimization algorithms and reduce the risk of capture by local minima that are not global optimum points. The following example gives an idea of the types of local minima which arise in adaptive wavelet representations.

Fig.3 depicts a signal consisting of two sinusoids, of frequencies 0.05 and $0.10\ rad/s$. Here, $T = 1000$.

Suppose one is trying to represent this signal by a sum of two real Morlet wavelets with $\Omega = 2.5$. Suppose also that the following parameters are adopted for the optimization algorithm:

$$\begin{array}{ccc} 200 \text{ iterations} & \lambda_A = \lambda_B = 1 & \lambda_W = 10^{-3} \\ W^0 = [1,1] & B^0 = [250, 750] & A^0 = [a^0, a^0] \end{array}$$

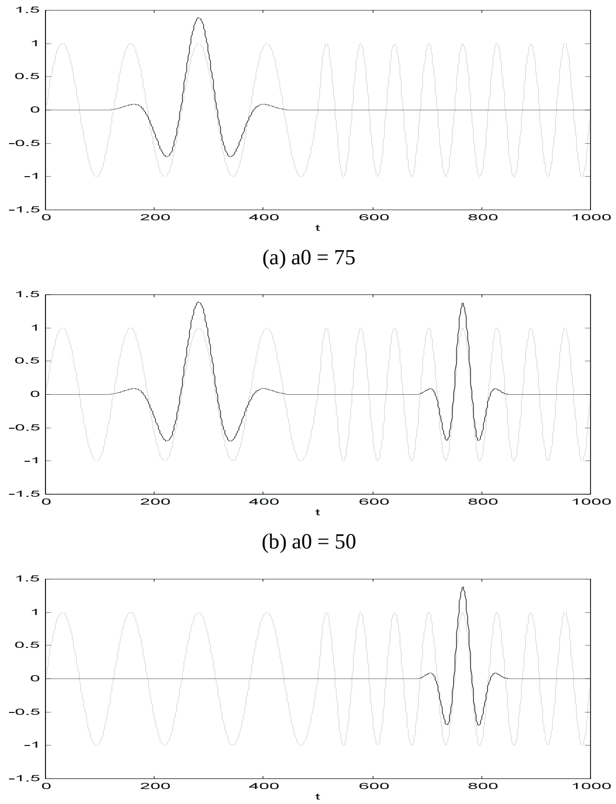
Results corresponding to three different initial scales are seen in fig.4. Note that, depending on the initial condition, a wavelet may not manage to “resonate” with any part of the signal. In this case, the solution found is to “annihilate” this wavelet.

Table 1 shows the resulting parameters at the end of the optimization, as well as the costs attained in each case. The values marked in boldface are those responsible for the “annihilation” of a wavelet, which can be due to a decrease of either the weight w_k or the scale a_k. Note that when this happens, parameters not related to the annihilation remain almost unchanged with respect to their initial values.

Table 1: Results of using different initial scales

a^0	75	50	12.5
a$_1$	53	53	**0.05**
b$_1$	281	281	253
w$_1$	1.39	1.39	0.78
a$_2$	72	27	27
b$_2$	750	766	766
w$_2$	**0.0003**	1.37	1.38
E	204	182	227

The concept of “wavelet anihilation” is similar to the turning off of neurons in conventional neural networks: if a node is not contributing to cost reduction, then the network may try to de-activate it. In ANN this is done either by decreasing the synaptic weights or by moving the decision surface of the neuron outside the region where patterns are found [9].



(a) a0 = 75

(b) a0 = 50

(c) a0 = 12.5

Figure 4: Effect of different initial scales

In order to cover a wide range of possible values for the scale parameter, the following multiresolutional initialization scheme is adopted in the present work:

1) *Specify L, the number of multiresolution levels to be initially employed and $K_{1,}$ the number of biased wavelets to be used at the lowest scale of the multiresolution. K_1 must be a multiple of 2^{L-1}.*

2) *Initial positions are set in a dyadic lattice, according to the following equations:*

$$\begin{aligned} b^0_{k_1} &= \frac{2k_1 - 1}{2K_1}T;\ k_1 = 1, ..., K_1 \\ b^0_{K_1+k_2} &= \frac{2k_2 - 1}{2K_2}T;\ k_2 = 1, ..., K_2 \\ b^0_{K_1+K_2+\cdots+k_L} &= \frac{2k_L - 1}{2K_L}T;\ k_L = 1, ..., K_L \end{aligned}$$

where $K_2 = K_1/2, ..., K_L = K_{L-1}/2$.

3) *Initial scales are set such that the union of the effective supports at each level covers D_f. This can be achieved, with $1 : 2$ overlapping, by making $a^0 = 2T(K_l R)^{-1}$ at the l^{th} resolution level (R is the effective support width of the mother wavelet employed). An example of an initial grid is depicted in fig.5.*

4) *Coefficients w_k are initialized with small random values, taken from a normal distribution with zero mean.*

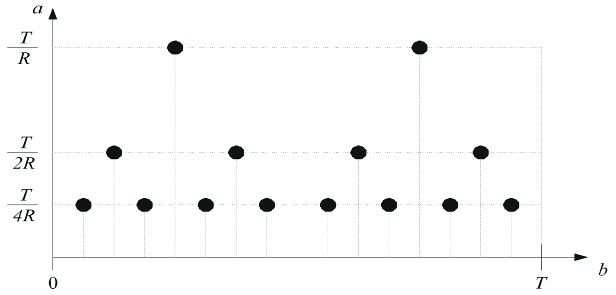


Figure 5: Initial Wavelet Grid.

5. The Competitive Wavelet Layer

Fig.6 depicts the architecture of the proposed competitive wavelet layer (CWL). It consists of G groups of K wavelets, each group with its own set of scales A_g, translations B_g, and weights W_g. Henceforth, to help clarify the relation between such architecture and conventional neural layers, the linear combination of the wavelets in a group will be called a wavelet neuron, or "wavelon" [2]. The g^{th} wavelon in the layer will be denoted by y_g.

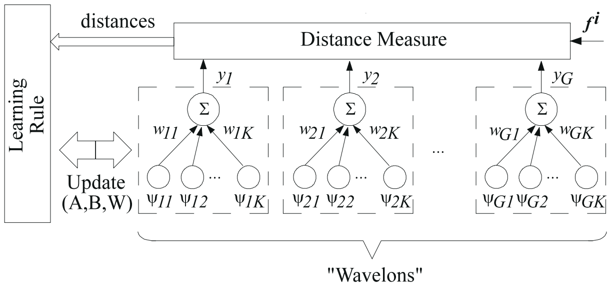


Figure 6: Architecture of an unsupervised wavelet layer with G "wavelons".

When a training pattern is presented to the CWL, its distance (norm of the difference) to each wavelon is computed. The wavelon with the smallest distance is declared the "winner" (best match to the pattern) and has its parameters (A_g, B_g, W_g) updated in order to further decrease the distance to the pattern.

This winner-take-all strategy, however, often does not yield an adequate separation of clusters, since the same wavelon may keep winning for patterns that should be assigned to different classes. An alternative is to adjust both the winner's and losers' parameters in an inverse proportion to their approximation errors (*leaky competitive learning* [10]). Another possibility is to employ the concept of neighbourhood, that is, to arrange the layer topologically and update not only the parameters of the winner wavelon, but also of its neighbours [10].

Thus, letting J_g be the distance between wavelon y_g and the i^{th} training pattern, the update law for the parameter vector X_g ($X_g = A_g, B_g$, or W_g) is:

$$X_g^{i+1} = X_g^i - \eta \lambda_X \nabla_X J_g \tag{13}$$

where $\eta = 1$ for the winner wavelon. For a loser wavelon, η may be zero ("winner-take-all"), a function of its position with respect to the winner (neighbourhood approach), or a function of the distance J_g (leaky competitive learning).

If one uses real Morlet wavelets and adopts as distance measure the cost in eq.(5), then the gradient equations and the initialization scheme for each wavelon are the same as those previously derived.

A final remark should be done with respect to wavelet-based clustering. Wavelets are a good choice for basis functions when signals to be clustered exhibit an oscillatory behaviour. However, if oscillations are small when compared, for instance, to trends present in the signal, it may be advisable to begin the clustering with another technique, such as conventional competitive networks. After re-centering data on the templates thus obtained, wavelets may then be used to perform a new clustering on the residual oscillations [5].

6. An Illustrative Example

Consider a problem in which one is to group the step inputs of first-order dynamic systems with basis on their time constant and on the presence or absence of parasitic oscillations. Such a situation could arise, for instance, during the quality control of a set of electric devices. As an example, suppose that half of the systems have a time constant of $1.0s$ and the other half, a time constant of $0.5s$. Within each group, half of the systems has a parasitic dynamic with the following transfer function:

$$G_{parasitic}(s) = \frac{100}{s^2 + 4s + 100} \tag{14}$$

For some reason, this parasitic subsystem is excited when the output of the system reaches 0.5.

With this setting, there are 4 classes of time responses to be considered:

a) Slow, no parasitic oscillation

b) Fast, no parasitic oscillation

c) Slow, parasitic oscillation present

d) Fast, parasitic oscillation present

Fig.7 depicts these four classes, which were sampled with a period of $0.01s$ (to fit the signals in the framework employed in the present section, the discrete time index is used in the horizontal axis). In this case, $T = 500$.

White gaussian noise with zero mean and a standard deviation of 0.1 was added to the step responses, in order to generate a set of patterns to be clustered. 50 patterns were taken from each of the 4 classes, thus generating a set of 200 patterns. Examples of the noisy patterns are shown in fig.8.

Remark that the patterns to be clustered in this example have oscillations only as a minor component. Thus, a preliminary clustering using a conventional competitive layer with 2 neurons is first performed. Letting W be the (501×1) vector of synaptic weights of the neuron, the

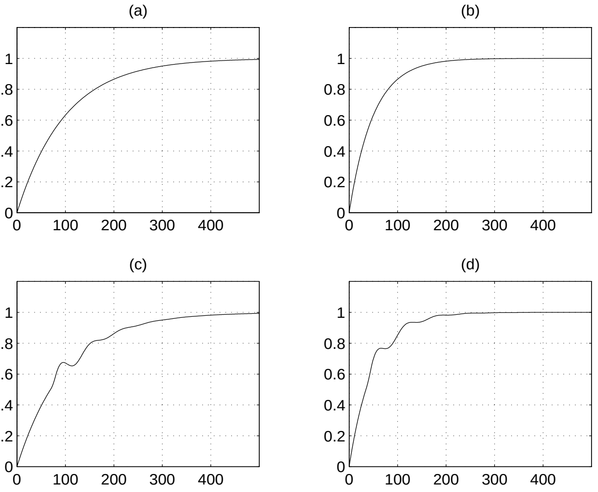


Figure 7: Step Responses: (a),(b) without parasitic oscillations; (c),(d) with parasitic oscillations.

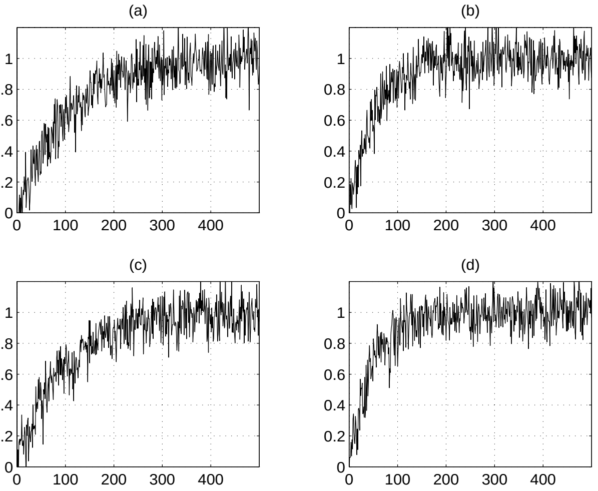


Figure 8: Noisy Step Responses: (a),(b) without parasitic oscillations; (c),(d) with parasitic oscillations.

update rule after the presentation of the i^{th} pattern is:

$$W^{i+1} = W^i + \eta\lambda \left[f^i - W^i\right] \qquad (15)$$

where $\eta = 1$ (winner neuron) or $\eta = (0.97)^i$ (loser neuron). Initial weights are taken from a normal distribution with a standard deviation of 0.01.

Using $\lambda = 0.15$, a perfect separation between slow and fast step responses was obtained. The templates can be seen in fig.9.

The next step consists of re-centering each cluster on the respective template. Fig.10 depicts typical signals that remain after template subtraction.

As it can be seen, these "residual signals" display a more pronounced oscillatory characteristic. Now, each of the two clusters obtained ("slow" and "fast" step responses) will be refined with basis on the presence or absence of parasitic oscillations. Table 2 brings the errors committed when cluster refinement was done with the conventional competitive layer. The best results obtained, by varying the learning step λ, are marked in boldface ($\lambda = 0.02$).

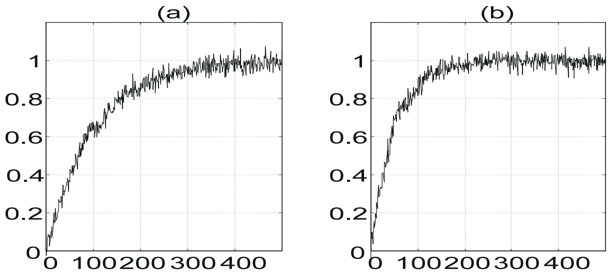


Figure 9: Initial Templates: (a) Slow, (b) Fast.

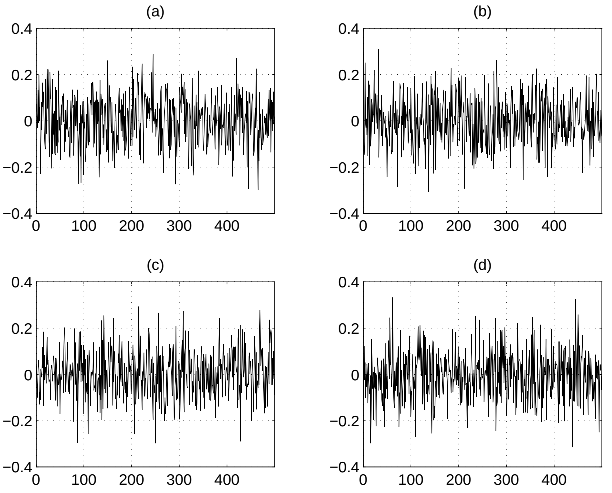


Figure 10: Examples of signals remaining after the subtraction of the respective class template.

Table 2: Neural Layer: Percentual Errors

$\lambda\ (\times 10^{-3})$	1.0	2.0	5.0	10	15	**20**	50
Slow	43	42	49	46	41	**31**	48
Fast	48	50	50	46	48	**45**	46

Table 3 brings the errors committed when cluster refinement was done with a wavelet layer with 2 wavelons. Each wavelon had seven Morlet wavelets distributed among 3 resolution levels. Initial weights were taken from a normal distribution with a standard deviation of 0.01 and training parameters[1] were set to $\lambda_W = 10^{-2}, \lambda_A = \lambda_B = 1000$. Parameter η was chosen as above, that is $\eta = 1$ (winner wavelon) or $\eta = (0.97)^i$ (loser wavelon). The effect of using different frequencies Ω for the mother wavelet was also studied. Again, the best results (on the average) are in boldface ($\Omega = 5$).

[1] Large values for λ_A e λ_B are needed because, in the gradient equations for A and B, the scale appears on the denominator. Due to the initialization scheme employed, initial scales are larger than one.

Table 3: Wavelet Layer: Percentual Errors

Ω	2.5	**5**	7
Slow	42	**6**	18
Fast	9	**30**	45

As can be seen, if Ω is conveniently chosen, the residual signals are better processed by the wavelet layer than by a conventional competitive layer. The reason is made clear in fig.11, which depicts the 4 wavelons obtained (two to refine each cluster) plotted over the parasitic oscillations. Remark how the wavelons are in phase or contra-phase with the oscillation: this “resonance” is the cornerstone for the good performance of the wavelet layer. Note that the objective of using a multiresolutional structure inside the wavelons is precisely to increase the probability of a wavelet entering in resonance with a feature that carries discriminatory information.

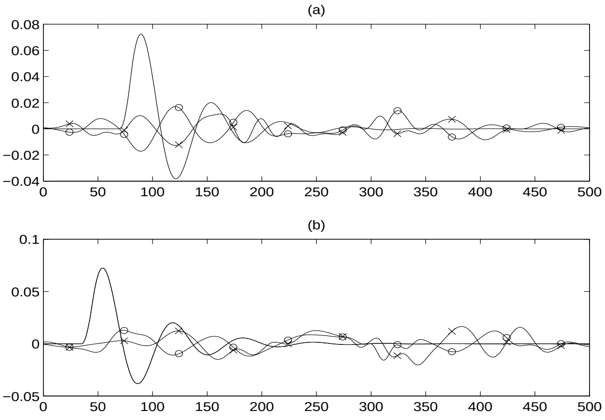


Figure 11: Wavelons for the presence (×) or abscence (O) of parasitic oscillations. (a) Slow Step Response, (b) Fast Step Response.

Fig.12 depicts a template obtained with the conventional competitive layer. It is interesting to point out that, while a neuron in the competitive layer has 501 weights (one for each sample in the analyzed signals), the wavelons employed had solely $3 \times 7 = 21$ parameters each. Thus, although the neural layer had the capability of arriving at the wavelet-synthesized templates, the large number of degrees of freedom made it very sensitive to the noise.

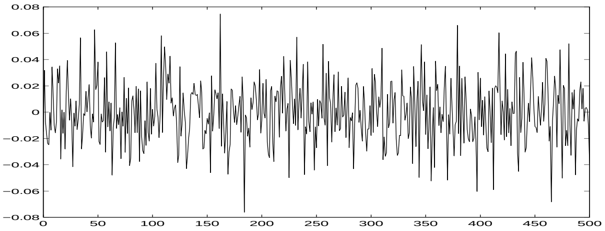


Figure 12: A template obtained when refining the clusters with the conventional competitive layer.

7. Concluding Remarks

Any template synthesized by wavelets can also be obtained with a conventional neural layer. However, due to the very nature of the basis functions employed, the CWL is expected to yield better results when the signals to be clustered display oscillatory characteristics. The performance of the CWL depends on a good choice for the central frequency of the mother wavelet (Ω), but this restriction could be alleviated by using an adaptive Ω [11].

Possible applications for the CWL might include, for instance, clustering of biomedical signals, such as electrocardiographic patterns. Some research in this direction was carried out in [5], which also exemplifies how to interpret the knowledge embedded in a wavelet layer.

Future works could exploit the use of wavelets in other ANN paradigms [10], such as learning vector quantization and adaptive resonance theory (ART). ART appears as an interesting possibility, since wavelet clustering is intrinsically based on resonance.

References

[1] H. H. Szu, B. Telfer, and S. Kadambe. Neural network adaptive wavelets for signal representation and classification. *Optical Engineering*, 31(9):1907–1916, Sep. 1992.

[2] Q. Zhang and A. Benveniste. Wavelet networks. *IEEE Trans. Neural Networks*, 3(6):889–898, Nov. 1992.

[3] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, Philadelphia, 1992.

[4] H. H. Szu. Intelligent neural networks using smart wavelets preprocessing. In *Proc. 6th Int. Conf. On Tools with Artificial Intelligence*, pages 348–349, Los Alamitos, 1994.

[5] R. K. H. Galvão. *Wavelet-Based Techniques for Adaptive Feature Extraction and Pattern Recognition*. PhD thesis, ITA, São José dos Campos, 1999.

[6] A. V. Chagas, M. C. Bossan, and J. Nadal. Agrupamento de batimentos cardíacos do eletrocardiograma utilizando uma camada de kohonen. In *Anais III CBRN - Congresso Brasileiro de Redes Neurais, Florianópolis, 21-24 Jul. 1997*, pages 185–189, 1997.

[7] G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Wellesley, 1996.

[8] H. Weng and K. M. Lau. Wavelets, period doubling, and time-frequency localization with application to organization of convection over the tropical western pacific. *J. Atmospheric Sciences*, 51(17):2523–2541, Sep. 1994.

[9] C. L. Nascimento Jr. *Artificial Neural Networks in Control and Optimization*. PhD thesis, UMIST, Manchester, UK, 1994.

[10] J. M. Zurada. *Introduction to Artificial Neural Systems*. West Publishing Co., St. Paul, 1992.

[11] H. Dickhaus and H. Heinrich. Classifying biosignals with wavelet networks - a method for noninvasive diagnosis. *IEEE Engineering in Medicine and Biology Magazine*, 15(5):103–111, Sep./Oct. 1996.