A novel kernel-based method for local pattern extraction in random process signals

Majid M. Beigi [1] and Andreas Zell[1]

1- Computer Science Dept. - University of Tübingen
Sand 1, D-72076 - Tübingen, Germany

Abstract. We consider a class of random process signals which contain randomly position local similarities representing the texture of an object. Those repetitive parts may occur in speech, musical pieces and sonar signals. We suggest a warped time resolved spectrum kernel for extracting the subsequence similarity in time series in general, and as an example in biosonar signals. Having a set of those kernels for similarity extraction in different size of subsequences, we propose a new method to find an optimal linear combination and selection of those kernels. We formulate the optimal kernel selection via maximizing the Kernel Fisher Discriminant criterion (KFD) and use Mesh Adaptive Direct Search method (MADS) to solve the optimization problem. Our method is used for biosonar landmark classification with promising results.

1 Problem

Bats can distinguish objects by emitting a series of ultrasound signals (chirps) that generally sweep covering frequencies from 22 to 100 kHz. Inspired by the bat biosonar system, researchers have utilized ultrasonic sensing techniques for mobile robots (biomimetic robots) and tried to classify different textures and landmarks using received echo signals. We used a sonar head system consisting of three ultrasound transducers, one for emission chirp signals (Polaroid 7000), two for reception (Polaroid 6000) and tried to classify three trees as landmarks (Fig. 2.**a**). The emitted pulse was a linearly frequency modulated chirp sweeping from 20kHz to 120kHz in 1 ms. The reflected echo contains the information about the geometry of the tree. We passed the reflected echoes through a bank of 10 gammatone filters between 20 kHz and 120 kHz (Fig. 1.**a**). Then, they were delivered to half-wave rectifiers. After *frame blocking* (50% overlap for frames), we used a *Hamming window.* At last, we made a feature matrix of the average energy of each channel of gammatone filter bank in each frame. The task is to classify the echoes of each object using those features. As we see in Fig. 1.**b**, despite the seemingly randomness of those preprocessed signals, there are some local similarities (shown by p) in echoes from one tree. We should find the size of subsequences of the time series *independent of the positions of occurrences* that have maximum similarities in echoes of each object. The intuition behind our idea is that the texture and structure of the objects and, as an example, the size of leaves or branches, and so the energy reflected by them can be related to the size of the similar subsequence lying in the signals. Inspired by the solutions for a similar problem known as remote homology detection in protein families

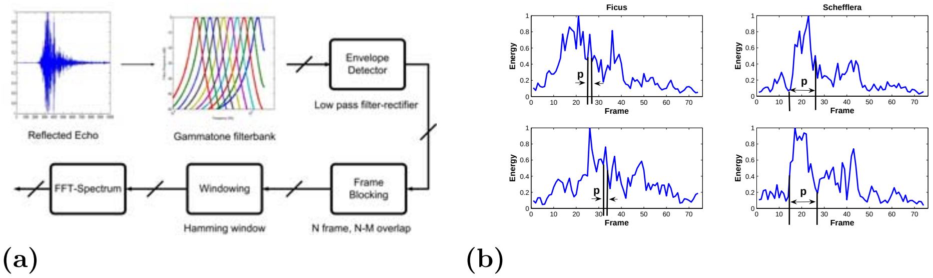

Fig. 1: **(a)** Block diagram of the preprocessing steps for reflected echoes. **(b)** The energy spectrum in each time frame for Ficus and Schefflera trees (output of gammatone filter centered around 50 kHz). Our suggested spectrum kernel tries to find the local similarities in window of size p in echoes of one object.

and the work of Lodhi et al. [1], we suggest a kernel called *warped time-resolved spectrum kernel* for our classification task.

2 Algorithms

2.1 Warped time resolved spectrum kernel

A time sequence $s = s_1 ... s_n$ is a sequence of data points at successive times with $s_i \in \Re^d$ where $1 \leq i \leq n$ and d is the dimension of data points. We denote $|s|$ the length of s, $s(i-p+1:i)$ the p-length subsequence of s from position $i-p+1$ to position i, $\mathbf{I}_p^{|s|}$ the set of indices defining all the p-long (*contiguous* or *non-contiguous*) subsequence of s: $\mathbf{I}_p^s = \{\mathbf{i} : \mathbf{i} \in \mathrm{N}^p, 1 \leq i_1 < ... < i_p \leq |s|\}$ and $u = s_{\mathbf{i}}$ as a subsequence of s in positions given by $\mathbf{i} = (i_1, ..., i_{|u|})$. The number of gaps in the subsequence is $g_{\mathbf{i}} = (i_{|u|} - i_1 + 1) - |\mathbf{i}|$. For example, if we consider $s = s_1 s_2 s_3 s_4 s_5$, $u = s_1 s_3 s_5$ is a subsequence of s in the positions $\mathbf{i} = (1, 3, 5)$ of length $|\mathbf{i}| = 3$ and $g_{\mathbf{i}} = 2$.

For $u \in \Sigma^{p \times d}$, the infinite set of all subsequences with size p and dimension d, the implicit embedding map ϕ brings s to a vector space F ($\phi : s \longrightarrow (\phi_u(s)) \in F$) and the u component ($u \in \Sigma^{p \times d}$) of our feature vector is: $\phi_u^p(s) = \sum_{\mathbf{i} \in \mathbf{I}_p^{|s|}} \varphi_u(s_{\mathbf{i}}) \gamma^{g_{\mathbf{i}}}$, where $\gamma \in (0,1)$ is a decay factor as a cost for warping (non-contiguousity) in the time series and φ is an implicit map that satisfies:

$$\kappa_p(s_{\mathbf{i}}, t_{\mathbf{j}}) =< \varphi_u(s_{\mathbf{i}}),\ \varphi_u(t_{\mathbf{j}}) > \quad \text{for} \quad \mathbf{i} \in \mathbf{I}_p^s,\ \mathbf{j} \in \mathbf{I}_p^t,\ u \in \Sigma^{p \times d}$$

in which κ_p is a kernel function that measures the local similarity between two p-length subsequences $s_{\mathbf{i}}$ and $t_{\mathbf{j}}$ of the time series in consideration. In words, $\phi_u^p(s)$ is a sum over all similarities between p-long subsequences of s and u. The dot product of those feature vectors represents the w*arped-time resolved p-spectrum kernel*:

$$\mathcal{K}_p(s,t) = \langle \phi_u^p(s), \phi_u^p(t) \rangle = \int_{\mathbf{R}^{d \times p}} \phi_u^p(s) \phi_u^p(t) du$$

$$= \sum_{\mathbf{i}\in\mathbf{I}_p^s} \sum_{\mathbf{j}\in\mathbf{I}_p^t} \gamma^{g_\mathbf{i}} \gamma^{g_\mathbf{j}} \int_{R^{d\times p}} \varphi_u(s_\mathbf{i})\varphi_u(t_\mathbf{j})du = \sum_{\mathbf{i}\in\mathbf{I}_p^s} \sum_{\mathbf{j}\in\mathbf{I}_p^t} \kappa_p(s_\mathbf{i}, t_\mathbf{j})\gamma^{g_\mathbf{i}+g_\mathbf{j}}$$

As we see from the above equation, the kernel adds all similarity scores between subsequences, considering their warping. Needless to say, the calculation of that kernel has a very high computational cost. We use dynamic programming to calculate it in an efficient manner and justifiable time. Considering the definitions of $\mathbf{I}_p^s$ and $\mathbf{I}_p^t$, we express the kernel using a suffix version of that:

$$\mathcal{K}_p(s,t) = \sum_{i=1}^{|s|} \sum_{j=1}^{|t|} \sum_{(\mathbf{i},\mathbf{j})\in\mathbf{I}_p^{s(1:i)}\times I_p^{t(1:j)}} \kappa_p(s_\mathbf{i}, t_\mathbf{j})\gamma^{g_\mathbf{i}+g_\mathbf{j}} = \sum_{i=1}^{|s|} \sum_{j=1}^{|t|} \mathcal{K}_p^S(s(1:i), t(1:j)$$

where:

$$\mathcal{K}_p^S(s(1:i), t(1:j)) = \sum_{(\mathbf{i},\mathbf{j})\in\mathbf{I}_p^{s(1:i)}\times\mathbf{I}_p^{t(1:j)}} \kappa_p(s_\mathbf{i}, t_\mathbf{j})\gamma^{g_\mathbf{i}+g_\mathbf{j}}$$

We consider $s' = s(1:|s'|)$, $t' = t(1:|t'|)$, $1 \le |s'| \le |s|$ and $1 \le |t'| \le |t|$ (prefixes of s and t). If we add a new data point x to the time series s', using the above equation we can calculate $\mathcal{K}_p(s'x, t')$:

$$\mathcal{K}_p(s'x, t') = \mathcal{K}_p(s', t') + \sum_{j=1}^{|t'|} \mathcal{K}_p^S(s'x, t'(1:j))$$

We accept a constraint on choosing the kernel function $\kappa_p(s_\mathbf{i}, t_\mathbf{j})$, we suppose: $\kappa_p(s_\mathbf{i}, t_\mathbf{j}) = \prod_{i=1}^{p} \kappa^*(s_{\mathbf{i}_i}, t_{\mathbf{j}_i})$, in which κ^* is an arbitrary function that measures the similarity between two data points of the time series. In this study, as a suitable and arbitrary selection we consider $\kappa^*(s_{\mathbf{i}_i}, t_{\mathbf{j}_i}) = \exp\frac{-(s_{\mathbf{i}_i}-t_{\mathbf{j}_i})^2}{2\sigma^2}$ to measure the similarity between two data points, then:

$$\kappa_p(s_\mathbf{i}, t_\mathbf{j}) = \prod_{i=1}^{p} \kappa^*(s_{\mathbf{i}_i}, t_{\mathbf{j}_i}) = \exp\left(-\frac{||s_\mathbf{i}-t_\mathbf{j}||^2}{2\sigma^2}\right)$$

That, $\kappa_p(s_\mathbf{i}, t_\mathbf{j})$ is a gaussian kernel of width σ and suitable for measuring the local similarity of subsequences in time series. Then, if we add another new data point y to the time series t', considering the assumption for κ_p and the above definition of $\mathcal{K}_p^S$, it can be shown:

$$\mathcal{K}_p^S(s'x, t'y) = \kappa^*(x,y) \sum_{i=1}^{|s'|} \sum_{j=1}^{|t'|} \gamma^{|s'|-i+|t'|-j} \mathcal{K}_{p-1}^S(s'(1:i), t'(1:j))$$

It means when new points are added, to measure new p-suffix kernel, we must calculate similarities of $p-1$ length subsequences in the suffixes considering the degrees of warping. To evaluate $\mathcal{K}_p^S$ recursively, we define:

$$\mathcal{K}_p^{Sw}(k,l) = \sum_{i=1}^{k} \sum_{j=1}^{l} \gamma^{k-i+l-j} \mathcal{K}_{p-1}^S(s'(1:i), t'(1:j))$$

Then: $$\mathcal{K}_p^S(s'x, t'y) = \kappa^*(x,y)\mathcal{K}_p^{Sw}(|s'|, |t'|)$$

to express the above kernel recursively, we use the relation:

$$\sum_{i=1}^{a}\sum_{j=1}^{b} f(i,j) = f(a,b) + \sum_{i=1}^{a-1}\sum_{j=1}^{b} f(i,j) + \sum_{i=1}^{a}\sum_{j=1}^{b-1} f(i,j) - \sum_{i=1}^{a-1}\sum_{j=1}^{b-1} f(i,j)$$

let $f(i,j) = \gamma^{k-i+l-j}\mathcal{K}_{p-1}^S(s'(1:i), t'(1:j))$, $a = k$ and $b = l$, we have:

Algorithm: *Recursive computation of the warped time resolved spectrum kernel.*

$$\mathcal{K}_p^{Sw}(k,l) = \mathcal{K}_{p-1}^{S}(s'(1:k),t'(1:l)) + \gamma\mathcal{K}_p^{Sw}(k,l-1) + \gamma\mathcal{K}_p^{Sw}(k-1,l)$$
$$-\gamma^2\mathcal{K}_p^{Sw}\ \ (k-1,l-1)$$

$$\mathcal{K}_p^{S}(s'x,t'y) = \kappa^*(x,y)(x,y)\mathcal{K}_p^{Sw}(|s'|,|t'|)$$
$$\mathcal{K}_p(s'x,t') = \mathcal{K}_p(s',t') + \textstyle\sum_{j=1}^{|t'|}\mathcal{K}_p^{S}(s'x,t'(1:j))$$

$$\begin{array}{rcl}
\mathcal{K}_0^{S}(s',t') & = & 1 \quad \textbf{for} \quad \textbf{all} \quad s',t', \\
\mathcal{K}_i^{S}(s',t') & = & 0, \quad \textbf{if} \ \min(|s'|,|t'|) < i, \\
\mathcal{K}_i(s',t') & = & 0, \quad \textbf{if} \ \min(|s'|,|t'|) < i,
\end{array}$$

The computation of the kernel follows a dynamic programming technique with the order of $O(p|s||t|)$. We have recursions over the prefixes of the time series and the lengths of the subsequences and we do the routine above until $x = s_{|s|}$ and $|t'| = |t|$.

To prevent that with larger sizes of subsequences the kernel achieves a higher similarity score we normalize the kernel, $\mathcal{K}_i^{norm}(s,t) = \frac{\mathcal{K}_i(s,t)}{\sqrt{\mathcal{K}_i(s,s)\mathcal{K}_i(t,t)}}$. This operation scales the similarities in the range [0,1]. In practice and specially in our classification task, it makes sense to consider the similarity of subsequences having different sizes and calculate a linear combination of different i-spectrum kernels with different weighting $\theta_i \geq 0$. The weighted kernel is:

$$K(s,t) = \sum_{i=1}^{p} \theta_i \mathcal{K}_i^{norm}(s,t) \tag{1}$$

Finding suitable values of the parameters θ_i is a case of more general problem known as optimal kernel selection. In the following subsection we suggest our new algorithm to solve this problem.

2.2 Fisher discriminant based optimal kernel selection

The Kernel Fisher Discriminant [2] is a non-linear extension of the Linear Fisher Discriminant Analysis. Given a set of n_+ positive training data $\chi_+ \subset \mathbb{R}^d$ and a set of n_- negative data $\chi_- \subset \mathbb{R}^d$, ($n = n_+ + n_-$, all data), and a map $\phi :\ \mathbb{R}^d \rightarrow\ \mathcal{F}$, the aim is to find a direction $w = \sum\limits_{i=1}^{n} \alpha_i \phi(x_i)$ in the feature space $\mathcal{F}$ given by weights $\alpha = [\alpha_1, ..., \alpha_n]$ which maximizes the separation of the mean scaled in the feature space and minimizes the variance in that direction. For that, the criterion $J(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T (N + \lambda I)\alpha}$ should be maximized [2]. The parameter λ is a regulation factor and M and N (defined in [2]) are gained in terms of the kernel matrix K, where $K_{i,j} = k(x_i, x_j) =< \phi(x_i), \phi(x_j) >$. If we consider:

$$D = \left[\begin{array}{cc} I_{n_+} - \frac{1}{n_+} 1_{n_+} 1_{n_+}^T & 0 \\ 0 & I_{n_-} - \frac{1}{n_-} 1_{n_-} 1_{n_-}^T \end{array} \right]_{n \times n}, \quad y = \left[\begin{array}{c} (1/n_+) 1_{n_+} \\ (-1/n_-) 1_{n_-} \end{array} \right]_{n \times 1}$$

where 1_n and I_n denote the vector of all ones and the identity operator in $\mathbb{R}^d$, respectively. It can be proven (not shown here):

$$\alpha_{\max} = (KDK + \lambda I)^{-1} Ky \quad \text{and} \quad J_{\max}(K) = \alpha_{\max}^{'} Ky$$
$$J_{\max}(K) = y' K (KD'K + \lambda I)^{-1} Ky \tag{2}$$

If we consider the variable K as a linear combination of a set of kernel matrices, in the next step, we try to find the matrix K, which maximizes the above equation. Considering equations 1 and 2, the problem of finding the optimal kernel in term of maximizing the Fisher discriminant ratio can be written as:

$$\begin{array}{ll} \min & f(\sum_{i=1}^{l} \theta_i \mathcal{K}_i^{norm}) = -J_{\max}(\sum_{i=1}^{l} \theta_i \mathcal{K}_i^{norm}) \\ \text{subject to} & \theta \succeq 0, \quad 1^T \theta = 1 \end{array}$$

It is easy to prove the convexity of the above objective function. We suppose $f(x,y) = x'y^{-1}x$, $h(K) = Ky$ and $g(K) = KD'K + \lambda I$, considering the convexity of f, h and g, we conclude the convexity of $f(h,g)$ and so the above objective function. Then, any local optimum answer for the objective function is a global one of that, too. To solve the problem we use a mesh adaptive direct search (MADS) method. It computes a series of points that get closer and closer to the optimal point. The algorithm searches a set of random selected points, called a mesh, around the current point-the point computed at the previous step of the algorithm. The mesh is formed by adding the current point to a scalar multiple of a set of vectors called a pattern and the point in the mesh that improves the objective function becomes the current point at the next step. The routine continues until a stopping criterion is fulfilled [4].

3 Experiment and results

We gathered the sonar data, 720 echoes for each tree shown in (Fig. 2.**a**). After the preprocessing steps for each echo (Fig. 1.**a**), we have a time series in which each point is a time frame and its value is an array of the average energy of each channel of gammatone filter. We selected randomly 100 echoes of each tree and then calculated $\mathcal{K}_i^{norm}(s[m], s[n])$ for $i \in [1, l]$, $m, n \in [1, 100]$ and $\sigma \in \{1, 10, 100, 1000\}$ where $s[m]$ and $s[n]$ are the m-th and n-th of preprocessed echoes and l is the length of the time series (in our experiment 90). Using the optimal kernel selection noted above, we found the optimal value for θ_i in equation 1 and σ and calculated the matrix K:

$$\mathbf{K}(i,j) = K(s[i], s[j]) = \textstyle\sum_{k=1}^{l} \theta_i^{opt} \mathcal{K}_l^{norm}(s[i], s[j]) \qquad i,\, j \in [1, 300]$$

where $s[i]$ is i-th echo, for Ficus echoes $i \in$[1,100], for Bamboo $i \in$ [101,200] and for Schefflera $i \in$[201,300]. In this study, we found that suitable values for σ are in the range [10,100] and for γ (as warping cost) in the range [0.1,.2].

A SVM learns a classification function $f(x)$ of the form:

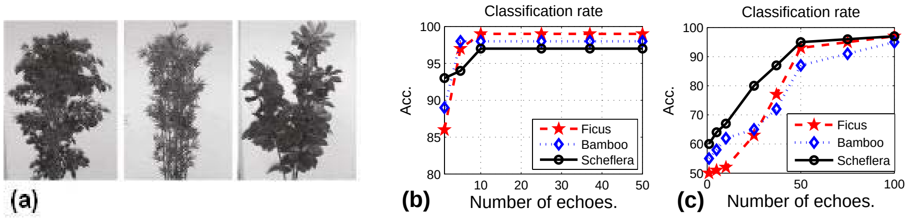

Fig. 2: **(a)** Ficus (left), bamboo (middle) and Schefflera (right) trees are our biosonar landmarks. **(b)** The overalaccuracy of classifiers using different numbers of echoes for testing with the Warped Time-resolved spectrum kernel. **(c)** The accuracy of classifier via Template matching using acoustic images of echoes (Wang et al. [5]).

$$f(x) = \sum_{i;x_i \in \chi_+} \lambda_i K(x, x_i) - \sum_{i;x_i \in \chi_-} \lambda_i K(x, x_i) \tag{3}$$

where non-negative λ_i weights are computed during training by maximizing a quadratic objective function and $K(.,.)$ is the kernel. Given this function, a new data x is predicted to belong to the positive dataset, if the value of $f(x)$ is positive, otherwise it belongs to the negative dataset. After training the classifier, we used the remaining data (1860 echoes) for testing. Fig. 2**a** shows the average accuracy of the classifier based on the number of echoes as observation. It shows a high accuracy even for a low number of echoes. Comparing with the previous works of our group (Wang et al. [3, 5]), it shows a notable improvement in accuracy. The best result for classification gained before was through template matching in 2D biosonar acoustic images (using a 2D Discrete Cosine Transform). The classification was made via extracting the maximum normalized cross correlation between the acoustic templates (Fig. 2**c**). As shown in Fig. 2**b**, we could get higher accuracy in both single and repeated observations (even with fewer echoes). With our suggested kernel we could extract patterns and similarities in the subsequences of time series, considering their warping and without dependency on the order of those subsequences.

References

[1] H. Lodhi and C. Saunders J. Shawe-Taylor and N. Cristianini and C. Watkins, Text Classification using String Kernels, *Jounal of Machine Learning Research*, 419-444, 2002.

[2] S. Mika, G. Rätsch, J.Weston, B. Schölkopf, and K.-R. Müller, Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, pages 41-48. IEEE, 1999a.

[3] M. Wang and A. Zell, Sequential sensing with Biosonar for natural landmark classification *IEEE International Workshop on safty, security and rescue robotics (SSRR 2005)*, pages 137-142,2005.

[4] C. Audet and J.E Dennis Jr., Mesh adaptive direct search algorithms for constrained optimization, *SIAM J. optim.*, Vol.17, No. 1, pp.188-217, 2006.

[5] M. Wang, *Natural Landmark Classification with a Biosonar based Mobile Robot*. PhD thesis,University of Tübingen, 2006.