HAL
open science

On Recursive Edit Distance Kernels with Application to Time Series Classification

Pierre-François Marteau, Sylvie Gibet

To cite this version:

Pierre-François Marteau, Sylvie Gibet. On Recursive Edit Distance Kernels with Application to Time Series Classification. 2013. hal-00486916v8

HAL Id: hal-00486916
https://hal.science/hal-00486916v8

Preprint submitted on 1 Aug 2013 (v8), last revised 25 May 2014 (v12)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On Recursive Edit Distance Kernels with Application to Time Series Classification

Pierre-François Marteau, *Member, IEEE* and Sylvie Gibet, *Member, IEEE*

Abstract—This paper proposes some extensions to the work on kernels dedicated to string or time series global alignment based on the aggregation of scores obtained by local alignments. The extensions we propose allow to construct, from classical recursive definition of elastic distances, recursive edit distance (or time-warp) kernels that are positive definite if some sufficient conditions are satisfied. The sufficient conditions we end-up with are original and weaker than those proposed in earlier works, although a recursive regularizing term is required to get the proof of the positive definiteness as a direct consequence of the Haussler's convolution theorem. The classification experiment we conducted on three classical time warp distances (two of which being metrics), using Support Vector Machine classifier, leads to conclude that, when the pairwise distance matrix obtained from the training data is *far* from definiteness, the positive definite recursive elastic kernels outperform in general the distance substituting kernels for the classical elastic distances we have tested.

Index Terms—Edit distance, Dynamic Time Warping, Recursive kernel, Time series classification, Support Vector Machine.

1 Introduction

Elastic similarity measures such as Dynamic Time Warping (DTW) or Edit Distances have proved to be quite efficient compared to non-elastic similarity measures such as Euclidean measures or LP norms when addressing tasks that require the matching of time series data, in particular time series clustering and classification. A wide scope of applications in different domains, such as physics, chemistry, finance, bio-informatics, network monitoring, etc, have demonstrated the benefits of using elastic measures. A natural question to ask at this point is whether we can develop and use *elastic* kernel methods based from such elastic distances. This requires to address the existence and construction of Reproducing Edit Distance Hilbert Spaces for a given elastic measure, i.e. a functional vector space endowed with the dot product in which edit distances or time warping algorithms can be defined. Unfortunately it seems that common elastic measures that are derived from DTW or more generally dynamic programming recursive algorithms are not directly induced by an inner product of any sort, even when such measures are metrics. One can conjecture that it is not possible to embed time series in an Hilbert space having a time-warp capability using these classical elastic measures, but nevertheless, we can propose (close) regularized variants for which such kernel construction construction is possible.

This paper aims at (re-)exploring this issue and, following earlier works ([14], [12], [36], [10], [11]) proposes Recursive Edit Distance Kernels (REDK) constructions that try to preserve the properties of elastic measures from which they are derived, while offering the possibility of embedding time series in Time Warped Hilbert Spaces. The main contributions of the paper are as follows

1) we verify the indefiniteness of the main time-warp measures used in the literature,
2) we propose a new method to construct positive definite kernels from classical time-warp or edit measures. This method is quite general and less restrictive than previous ones, although it requires to introduce a recursive regularizing term that allows to prove the definiteness of our kernels as a direct consequence of the Haussler's convolution theorem,
3) we show that the regularizing approach we develop applies also when a symmetric *corridor* is used to limit the search space used to evaluate the elastic measure,
4) we experiment and compare the proposed kernels on some time series classification tasks using a large variety of time series data sets to estimate in practice the benefit we can expect from such kernels.

The paper is organized as follows: the second section synthesizes the related works; the third section introduces the notation and mathematical background that is used throughout the paper. The fourth section develops the construction of a general REDK and details some instantiations derived from classical elastic measures. The fifth section gathers classification experimentation on a wide range of time series data and compares REDK accuracies with classical elastic measures. The sixth section proposes a conclusion and further research perspectives. Appendix A states the indefiniteness of classical elastic measures, and appendix B presents the proof of our main results.

• *P.F. Marteau and Sylvie Gibet are with the Univ. Bretagne Sud, UMR 6074, IRISA, F-56000 Vannes, France.*
E-mail: {Pierre-Francois.Marteau, Sylvie.Gibet}(AT)univ-ubs.fr

2 Related works

During the last decades, the use of kernel-based methods in pattern analysis has provided numerous results and fruitful applications in various domains such as biology, statistics, networking, signal processing, etc. Some of these domains, such as bioinformatics, or more generally domains that rely on sequence or time series models, require the analysis and processing of variable length vectors, sequences or timestamped data. Various methods and algorithms have been developed to quantify the similarity of such objects. From the original dynamic programming [2] implementation of the symbolic edit distance [18] by Wagner and Fisher [37], the Smith and Waterman (SW) algorithm [32] has been designed to evaluate the similarity between two symbolic sequences by means of a local gap alignment. More efficient local heuristics have since been proposed to meet the massive symbolic data challenge, such as BLAST [1] or FASTA [22]. Similarly, dynamic time warping measures have been developed to evaluate similarity between digital time series or timestamped data [35], [26], and more recently [6], [20] propose elastic metrics dedicated to such digital data.

Our ability to construct kernels with elastic or time-warp properties from such *elastic distances* allowing to embed time series into vector spaces (Euclidean or not) has attracted attention (e.g. [12][15][13]). Indeed, significant benefits can be expected from applications of kernel-based machine learning algorithms to variable length data, or more generally data for which some elastic matching has a meaning. Among the kernel machine algorithms applicable to discrimination or regression tasks, Support Vector Machines (SVM) [33], [4], [28] are still reported to yield state-of-the art performances although their accuracy greatly depends on the exploited kernel.

The definition of *good* kernels from known elastic or time-warp distances applicable to data of variable lengths has been a major challenge since the 1990s. The notion of 'goodness' has rapidly been associated to the concept of definiteness. Basically SVM algorithms involve an optimization process whose solution is proved to be uniquely defined if and only if the kernel is positive definite: in that case the objective function to optimize is quadratic and the optimization problem convex. Nevertheless, if the definiteness of kernels is an issue, in practice, many situations exist where definite kernels are not applicable. This seems to be the case for the main elastic measures traditionally used to estimate the similarity of objects of variable lengths. A pragmatic approach consists of using indefinite kernels [12][13][40], although contradictory results have been reported about the impact of definiteness or indefiniteness of kernels on the empirical performances of SVMs. The sub-optimality of the non-convex optimization process is possibly one of the causes leading to these un-guaranteed performances [38], [12]. Regulation procedures have been proposed to locally approximate indefinite kernel functions by definite ones with some benefits. Among others, some approaches apply direct spectral transformations to indefinite kernels. These methods [39] [8] consist in i) flipping the negative eigenvalues or shifting the eigenvalues using the minimal shift value required to make the spectrum of eigenvalues positive, and ii) reconstructing the kernel with the original eigenvectors in order to produce a positive semidefinite kernel. Yet, in general, 'convexification' procedures are difficult to interpret geometrically and the expected effect of the original indefinite kernel may be lost. Some theoretical highlights have been provided through approaches that consist in embedding the data into a pseudo-Euclidean (pE) space and in formulating the classification problem with an indefinite kernel, such as minimizing the distance between convex hulls formed from the two categories of data embedded in the pE space [13]. The geometric interpretation results in a constructive method allowing for the understanding, and in some cases the prediction of the classification behavior of an indefinite kernel SVM in the corresponding pE space.

Some other works like [31], [17] address the construction of elastic kernels for time series analysis through a decomposition of time series as a sum of local low degree polynomials and, using a resampling process, the piecewise approximation of the time series are embedded into a proper so-called Reproducing Kernel Hilbert Space in which the SVM is learned.

Other approaches try to use directly the elastic distance into the kernel construction, without any approximation or resampling process.These works are based on the work of Haussler on convolution kernels [14] defined on a set of discrete structures such as strings, trees, or graphs. The iterative method that is developed is generative, as it allows for the building of complex kernels from the convolution of simple local kernels. Following the work of Haussler [14], Saigo et al [25] define, from the Smith and Waterman algorithm [32], a kernel to detect local alignment between strings by convolving simpler kernels. These authors show that the Smith and Waterman distance measure, dedicated to determining similar regions between two nucleotide or protein sequences, is not definite, but is nevertheless connected to the logarithm of a point-wise limit of a series of definite convolution kernels. Cuturi et al. [11] have adapted this approach to times series alignments covering the DTW elastic distance. In fact, these previous studies have very general implications, the first one being that classical elastic measures can also be understood as the limit of a series of definite convolution kernels.

In this paper we tackle a new approach to the positive definiteness regularization of elastic kernels that follows directly the lines of the Haussler's theorem on convolution kernels. The condition to obtain positive definiteness is weaker than the one proposed in previous works [11], although it requires the addition of an explicit recursive regularizing term. This term is easy to evaluate and has

the same complexity than the recursive equation that defines the elastic distance. Our regularization strategy is thus quite general, independent from the data (although a parameter can be optimized), and can be applied to a very large family of editing or dynamic time warping like distances. It applies also to some of their variants which exploit a symmetric *corridor* to speed-up the computation.

3 Notations and Mathematical Backgrounds

We give in this section commonly used definitions, with few details, for metric, kernel and definiteness, sequence set, and classical elastic measures.

3.1 Kernel and definiteness

A very large literature exists on kernels, among which [3], [28] and [29] present a large synthesis of major results. We give hereinafter some basic definitions and some results derived from these previous references.

Definition 3.1: A kernel on a non empty set U refers to a complex (or real) valued symmetric function $\varphi(x,y) : U \times U \to \mathbb{C}$ (or $\mathbb{R}$).

Definition 3.2: Let U be a non empty set. A function $\varphi : U \times U \to \mathbb{C}$ is called a positive (resp. negative) definite kernel if and only if it is Hermitian (i.e. $\varphi(x,y) = \overline{\varphi(y,x)}$ where the *overline* stands for the conjugate number) for all x and y in U and $\sum_{i,j=1}^{n} c_i \bar{c_j} \varphi(x_i,x_j) \geq 0$ (resp. $\sum_{i,j=1}^{n} c_i \bar{c_j} \varphi(x_i,x_j) \leq 0$), for all n in $\mathbb{N}$, $(x_1,x_2,...,x_n) \in U^n$ and $(c_1,c_2,...,c_n) \in \mathbb{C}^n$.

Definition 3.3: Let U be a non empty set. A function $\varphi : U \times U \to \mathbb{C}$ is called a conditionally positive (resp. conditionally negative) definite kernel if and only if it is Hermitian (i.e. $\varphi(x,y) = \overline{\varphi(y,x)}$ for all x and y in U) and $\sum_{i,j=1}^{n} c_i \bar{c_j} \varphi(x_i,x_j) \geq 0$ (resp. $\sum_{i,j=1}^{n} c_i \bar{c_j} \varphi(x_i,x_j) \leq 0$), for all $n \geq 2$ in $\mathbb{N}$, $(x_1,x_2,...,x_n) \in U^n$ and $(c_1,c_2,...,c_n) \in \mathbb{C}^n$ with $\sum_{i=1}^{n} c_i = 0$.

In the last two above definitions, it is easy to show that it is sufficient to consider mutually different elements in U, i.e. collections of distinct elements $x_1,x_2,...,x_n$. This is what we will consider for the remaining of the paper.

Definition 3.4: A positive (resp. negative) definite kernel defined on a finite set U is also called a positive (resp. negative) semidefinite matrix. Similarly, a positive (resp. negative) conditionally definite kernel defined on a finite set is also called a positive (resp. negative) conditionally semidefinite matrix.

For convenience sake, we will use p.d. and c.p.d. for positive definite and conditionally positive definite to characterize either a kernel or a matrix having these properties.

Constructing p.d. kernels from c.p.d. kernels is quite straightforward. For instance, if $-\varphi$ is a c.p.d. kernel on a set U and $x_0 \in U$ then, according to [3], $\psi(x,y) = \varphi(x,x_0) + \overline{\varphi(y,x_0)} - \varphi(x,y) - \varphi(x_0,x_0)$ is a p.d. kernel, so are $e^{(\psi(x,y))}$ and $e^{-\varphi(x,y)}$. The converse is also true. Furthermore, $e^{-t\varphi(x,y)}$ is p.d. for $t > 0$ if $-\varphi$ is c.p.d. We will precisely use this last result to construct p.d. kernels from classical elastic distances.

3.2 Sequence set

Definition 3.5: Let $\mathbb{U}$ be the set of finite sequences (symbolic sequences or time series): $\mathbb{U} = \{A_1^p | p \in \mathbb{N}\}$. A_1^p is a sequence with discrete index varying between 1 and p. We note Ω the empty sequence (with null length) and by convention $A_1^0 = \Omega$ so that Ω is a member of set $\mathbb{U}$. $|A|$ denotes the length of the sequence A. Let $\mathbb{U}_p = \{A \in \mathbb{U} \mid |A| \leq p\}$ be the set of sequences whose length is shorter or equal to p.

Definition 3.6: Let A be a finite sequence. Let $A(i)$ be the ith element (symbol or sample) of sequence A. We will consider that $A(i) \in S \times T$ where S embeds the multidimensional space variables (either symbolic or numeric) and $T \subset \mathbb{R}$ embeds the time stamp variable, so that we can write $A(i) = (a(i), t(i))$ where $a(i) \in S$ and $t(i) \in T$, with the condition that $t(i) > t(j)$ whenever $i > j$ (time stamps strictly increase in the sequence of samples). A_i^j with $i \leq j$ is the subsequence consisting of the i_{th} through the j_{th} element (inclusive) of A. So $A_i^j = A(i)A(i+1)...A(j)$. Λ denotes the null element. A_i^j with $i > j$ is the null time series, e.g. Ω.

3.3 General Edit/Elastic distance on a sequence set

Definition 3.7: An edit operation is a pair $(a,b) \neq (\Lambda,\Lambda)$ of sequence elements, written $a \to b$. Sequence B results from the application of the edit operation $a \to b$ into sequence A, written $A \Rightarrow B$ via $a \to b$, if $A = \sigma a \tau$ and $B = \sigma b \tau$ for some sub-sequences σ and τ. We call $a \to b$ a substitution operation if $a \neq \Lambda$ and $b \neq \Lambda$, a delete operation if $b = \Lambda$, an insert operation if $a = \Lambda$.

For any pair of sequences A_1^p, B_1^q, for which we consider the extensions A_0^p, B_0^q whose first element is the null symbol Λ, and for each elementary edit operation related to position $0 \leq i \leq p$ in sequence A and to position $0 \leq j \leq q$ in sequence B is associated a cost value $\Gamma_{A(i)\to B(j)}(A_1^p, B_1^q)$, or $\Gamma_{A(i)\to\Lambda,j}(A_1^p, B_1^q)$ or $\Gamma_{\Lambda,i\to B(j)}(A_1^p, B_1^q) \in \mathbb{R}$. To simplify this writing we will simply write $\Gamma(A(i) \to B(j))$, $\Gamma(A(i) \to \Lambda)$ or $\Gamma(\Lambda \to B(j))$ although this will not be fully appropriate in general.

Definition 3.8: A function $\delta : \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ is called an edit distance defined on $\mathbb{U}$ if, for any pair of sequences

A_1^p, B_1^q, the following recursive equation is satisfied

$$\delta(A_1^p, B_1^q) = \\ \text{Min} \begin{cases} \delta(A_1^{p-1}, B_1^q) + \Gamma(A(p) \to \Lambda) & del \\ \delta(A_1^{p-1}, B_1^{q-1}) + \Gamma(A(p) \to B(q)) & sub \\ \delta(A_1^p, B_1^{q-1}) + \Gamma(\Lambda \to B(q)) & ins \end{cases}$$

Note that not all edit/elastic distances are metric. In particular, the dynamic time warping distance does not satisfy the triangle inequality.

3.3.1 Levenshtein distance

The Levenshtein distance $\delta_{lev}(x, y)$ has been defined for string matching. For this edit distance, the *delete* and *insert* operations induce unitary costs, i.e. $\Gamma(A(p) \to \Lambda) = \Gamma(\Lambda \to B(q)) = 1$ while the *substitution* cost is null if $A(p) = B(q)$ or 1 otherwise.

3.3.2 Dynamic time warping

The DTW similarity measure δ_{dtw} [35][26] is defined according to the previous notations such as:

$$\begin{aligned} \delta_{dtw}(A_1^p, B_1^q) &= d_{LP}(a(p), b(q)) \\ &+ \text{Min} \begin{cases} \delta_{dtw}(A_1^{p-1}, B_1^q) & del \\ \delta_{dtw}(A_1^{p-1}, B_1^{q-1}) & sub \\ \delta_{dtw}(A_1^p, B_1^{q-1}) & ins \end{cases} \end{aligned}$$

where $d_{LP}(a(p), b(q)$ is the LP norm in $\mathbb{R}^k$, and so for DTW, $\Gamma(A(p) \to \Lambda) = \Gamma(A(p) \to B(q)) = \Gamma(\Lambda \to B(q)) = d_{LP}(a(p), b(q))$. Let us note that the time stamp values are not used, therefore the costs of each edit operation involve vectors a and b in S instead of vectors (a, t_a) and (b, t_b) in $S \times T$. One of the main restrictions of δ_{dtw} is that it does not comply with the triangle inequality as shown in [6].

3.3.3 Edit Distance with real penalty

$$\delta_{erp}(A_1^p, B_1^q) = \\ \text{Min} \begin{cases} \delta_{erp}(A_1^{p-1}, B_1^q) + \Gamma(A(p) \to \Lambda) & del \\ \delta_{erp}(A_1^{p-1}, B_1^{q-1}) + \Gamma(A(p) \to B(q)) & sub \\ \delta_{erp}(A_1^p, B_1^{q-1}) + \Gamma(\Lambda \to B(q)) & ins \end{cases}$$

with

$$\begin{aligned} &\Gamma(A(p) \to \Lambda) = d_{LP}(a(p), g) \\ &\Gamma(A(p) \to B(q) = d_{LP}(a(p), b(q) \\ &\Gamma(\Lambda \to B(q)) = d_{LP}(g, b(q)) \end{aligned}$$

where g is a constant in S and $d_{LP}(x, y)$ is the LP norm of vector $(x - y)$ in S.

Note that the time stamp coordinate is not taken into account, therefore δ_{erp} is a distance on S but not on $S \times T$. Thus, the cost of each edit operation involves vectors a and b in $\mathbb{R}^k$ instead of vectors (a, t_a) and (b, t_n) in $\mathbb{R}^{k+1}$.

According to the authors of ERP [6], the constant g should be set to 0 for some intuitive geometric interpretation and in order to preserve the mean value of the transformed time series when adding gap samples.

3.3.4 Time warp edit distance

Time Warp Edit Distance (TWED) [19], [20] is defined similarly to the edit distance defined for string [18][37]. The similarity between any two time series A and B of finite length, respectively p and q is defined as:

$$\delta_{twed}(A_1^p, B_1^q) = \\ \text{Min} \begin{cases} \delta_{twed}(A_1^{p-1}, B_1^q) + \Gamma(A(p) \to \Lambda) & del_A \\ \delta_{twed}(A_1^{p-1}, B_1^{q-1}) + \Gamma(A(p) \to B(q)) & subs \\ \delta_{twed}(A_1^p, B_1^{q-1}) + \Gamma(\Lambda \to B(q)) & del_B \end{cases}$$

with

$$\begin{aligned} &\Gamma(A(p) \to \Lambda) = d(A(p), A(p-1)) + \lambda \\ &\Gamma(A(p) \to B(q)) = d(A(p), B(q)) + d(A(p-1), B(q-1)) \\ &\Gamma(\Lambda \to B(q)) = d(B(q), B(q-1)) + \lambda \end{aligned}$$

where λ is a positive constant that represents a gap penalty and ν is a non negative constant which characterizes the *stiffness* of the δ_{twed} elastic measure. The time stamps are exploited to evaluate $d(A(p), B(q))$. In practice, $d(A(p), B(q)) = d_{LP}(a(p), b(q)) + \nu d_{LP}(t(p), t(q))$

3.4 Indefiniteness of elastic distance kernels

In appendix A, we give counter examples, one for each elastic distance we have previously defined, showing that these distances do not lead to definite kernels. This demonstrates that the metric properties of a distance defined on $\mathbb{U}$, in particular the triangle inequality, are not sufficient conditions to establish definiteness (conditionally or not) of the associated distance kernel. One could conjecture that elastic distances cannot be definite (conditionally or not), possibly because of the presence of the $\max$ or $\min$ operators in the recursive equation. In the following sections, we will see that replacing these $\min$ or $\max$ operators by a sum operator allows, under some conditions, for the construction of series of positive definite kernels whose limit is quite directly connected to the previously addressed elastic distance kernels.

4 Constructing positive definite kernels from elastic distance

The main idea leading to the construction of positive definite kernels from a given elastic distance defined on $\mathbb{U}$ is to replace the $\min$ or $\max$ operator into the recursive equation defining the elastic distance by a summation ($\sum$) operator. Instead of keeping only one of the best alignment paths, the new kernel will sum up the costs of all the existing sub-sequence alignment paths with some weighting factor that will favor *good* alignments while penalizing bad alignments. In addition, this weighting factor can be optimized. This principle has been applied successfully to the Smith and Waterman symbolic distance which is also known to be indefinite [25], and more recently to dynamic time warping kernels for time series alignment [11]. If, basically, we are following the same objective, the approach that we propose is quite different since it develops a regularization principle which is

simply and directly expressed into the recursive equation defining the elastic distance. First we replace the *min* (or *max*) operator by a summation operator and introduce a symmetric *corridor* function to optionally limit the summation. Then we add a regularizing recursive term such that the proof of the positive definiteness property can be understood as a direct consequence of the Haussler's convolution theorem, as shown in appendix B. Basically, everything comes back to the proper census of pairs of symmetric alignment paths.

4.1 Recursive accumulation of *f-cost* products

Definition 4.1: A function $\mathcal{C}(.,.) : \mathbb{U} \times \mathbb{U} \rightarrow \mathbb{R}$ is called a Recursive Accumulation of *f-cost* Products (RAfP) if, for any pair of sequences A_1^p, B_1^q, there exist a function $f : \mathbb{R} \rightarrow \mathbb{R}$ and three indicator functions $h_d : \mathbb{N}^2 \rightarrow \{0,1\}$, $h_s : \mathbb{N}^2 \rightarrow \{0,1\}$ and $h_i : \mathbb{N}^2 \rightarrow \{0,1\}$ such that the following recursive equation is satisfied:

$$\mathcal{C}(A_1^p, B_1^q) = \sum \begin{cases} h_d(p,q)\mathcal{C}(A_1^{p-1}, B_1^q) f(\Gamma(A(p) \rightarrow \Lambda)) & del \\ h_s(p,q)\mathcal{C}(A_1^{p-1}, B_1^{q-1}) f(\Gamma(A(p) \rightarrow B(q))) & sub \\ h_i(p,q)\mathcal{C}(A_1^{p}, B_1^{q-1}) f(\Gamma(\Lambda \rightarrow B(q))) & ins \end{cases} \quad (1)$$

This recursive definition relies on an initialization. To that end we set $\mathcal{C}(\Omega, \Omega) = \xi$, where ξ is a real constant, and Ω the null sequence. Note that if no path satisfying the indicator functions exists when aligning time series A and B, then $\mathcal{C}(A, B) = 0$.

Furthermore, according to the indicator functions h_d, h_s and h_i, this type of construction sums up the multiplication of the local quantities $f(\Gamma(A(i) \rightarrow B(j)))$ that we call *f-costs* for all or for some of the possible alignment paths that exist between the two time series, the concept of alignment path being precisely defined in appendix B (see definition B.3).

In this paper we are specifically concerned with the following two symmetric RAfP families. The first one, characterized by $\forall (p,q) \in \mathbb{N}^2$, $h_d(p,q) = h(p-1,q)$, $h_i(q,p) = h(q-1,p)$ and $h_s(p,q) = h(p-1,q-1)$, accumulates the *f-cost* products along all the alignment paths between the two time series (Eq.2) that exist inside the symmetric *corridor* defined by $h(p,q)$. For the case depicted in Figure 1, $h(p,q) = 1$, whenever $|p-q| \leq 3$, 0 otherwise. We denote this RAfP function $\mathcal{C}_h$.

$$\mathcal{C}_h(A_1^p, B_1^q) = \sum \begin{cases} h(p-1,q)\mathcal{C}_h(A_1^{p-1}, B_1^q) f(\Gamma(A(p) \rightarrow \Lambda)) \\ h(p-1,q-1)\mathcal{C}_h(A_1^{p-1}, B_1^{q-1}) f(\Gamma(A(p) \rightarrow B(q))) \\ h(p,q-1)\mathcal{C}_h(A_1^{p}, B_1^{q-1}) f(\Gamma(\Lambda \rightarrow B(q))) \end{cases} \quad (2)$$

The second RAfP function accumulates the *f-cost* products of all the paths that do not include any substitution outside the diagonal. Basically the substitutions are only allowed when $p = q$. This function is thus characterized by $\forall (p,q) \in \mathbb{N}^2$, $h_d(p,q) = h(p-1,q)$, $h_i(q,p) = h(p,q-1)$ where h is a symmetric function, and $h_s(p,q) = \delta_{p,q}$, where δ is the Kronecker's symbol (Eq.3). We denote this second RAfP function $\mathcal{C}_{h,\delta}$. As an example, the red dotted path in Figure 1 is taken cared of by $\mathcal{C}_{h,\delta}$.

$$\mathcal{C}_{h,\delta}(A_1^p, B_1^q) = \sum \begin{cases} h(p-1,q)\mathcal{C}_{h,\delta}(A_1^{p-1}, B_1^q) f(\Gamma(A(p) \rightarrow \Lambda)) \\ \delta_{p,q}\mathcal{C}_{h,\delta}(A_1^{p-1}, B_1^{q-1}) f(\Gamma(A(p) \rightarrow B(q))) \\ h(p,q-1)\mathcal{C}_{h,\delta}(A_1^{p}, B_1^{q-1}) f(\Gamma(\Lambda \rightarrow B(q))) \end{cases} \quad (3)$$

4.2 Recursive Edit Distance Kernel - Main result

Definition 4.2: We call Recursive Edit Distance Kernel (REDK) the function $\mathcal{K}(.,.) = \mathcal{C}_h(.,.) + \mathcal{C}_{h,\delta}(.,.) : \mathbb{U} \times \mathbb{U} \rightarrow \mathbb{R}$.

The following theorem, that relates to the ones presented in [11] and [30], establishes sufficient conditions on $f(\Gamma(a \rightarrow b))$ for a REDK function to be definite positive, and thus is a basis for the construction of definite REDK. The proof of this theorem, that is given in Appendix B, is a direct consequence of the Hausslers *R-convolution* theorem [14]. The sufficient condition that is proposed here is quite general and less restrictive than those proposed by [11] and [30]. The avenue that is taken here only requires that the local kernel $k(x,y)$ should be PD. In previous works, the condition for positive definiteness is that the local kernel $k(x,y)$ and the ratio $k(x,y)/(k(x,y)+1)$ should be jointly PD. Our new condition is obviously weaker than the one proposed in [9]. Furthermore, our REDK construction applies to a wider range of edit or time warp distances developed either for symbolic sequence or time series matching.

Theorem 4.3: Definiteness of REDK:

REDK is a positive definite kernel on $\mathbb{U} \times \mathbb{U}$ iif the local kernel $k(x,y) = f(\Gamma(x \rightarrow y))$ is a positive definite kernel on $((S \times T) \cup \{\Lambda\})^2$.

A sketch of proof for theorem 4.3 is given in the appendix B.

As the cost function Γ is, in general, conditionally negative definite, choosing $f(.)$ for the exponential ensures that $f(\Gamma(x \rightarrow y))$ is a positive definite kernel [27].

Other functions can be used, such as the Inverse Multi Quadric kernel $k(x,y) = \frac{1}{\sqrt{(\Gamma(x \rightarrow y))^2 + \theta^2}}$. As with the exponential (Gaussian or Laplace) kernel, the Inverse Multi Quadric kernel results in a positive definite matrix with full rank [21] and thus forms a infinite dimension feature space. Note here that Cuturi et al. [11] state that to ensure the definiteness of the DTW-REDK kernel, not only $f(\Gamma(x \rightarrow y))$, but also $f(\Gamma(x \rightarrow y))/(1 + f(\Gamma(x \rightarrow y)))$ need to be p.d. kernels which forms a stronger condition than the one we propose.

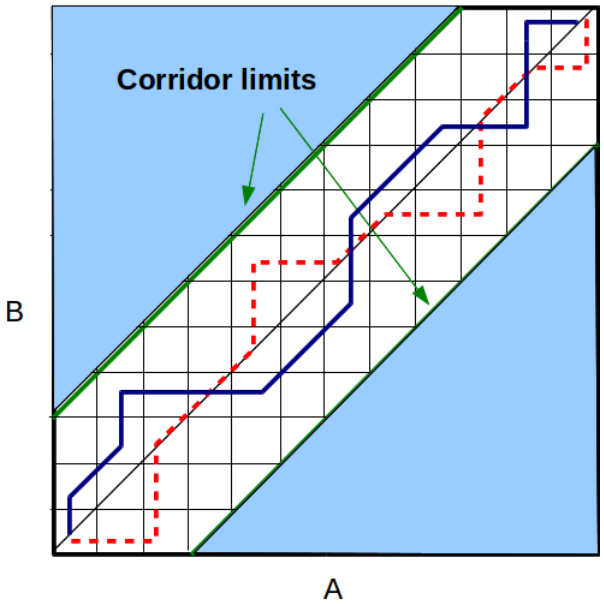

Fig. 1. Example of alignment paths existing when matching time series A and B. In red dotted line a path whose editing *f-cost* product is accumulated by $\mathcal{C}_h$ and $\mathcal{C}_{h,\delta}$, in blue plain line, a path whose editing cost product is accumulated by $\mathcal{C}_h$ but not $\mathcal{C}_{h,\delta}$. A symmetric corridor of any form can be used to reduced the number of admissible alignment paths that are taken care of by a RAfP

4.2.1 Computational cost of REDK

Although the number of paths that are summed up exponentially increases with the lengths $|A|$ and $|B|$ of the time series that are evaluated, the recursive computation of any RAfP function and therefore of $\mathcal{K}(A,B)$ lead to a quadratic computational cost $O(|A||B|)$, e.g. $O(N^2)$ if N is the average length of the time series that are considered. This quadratic complexity can be reduced to a linear complexity by limiting the number of alignment paths to consider in the recursion. This can be achieved when using a search corridor [26] as far as the kernel remains symmetric, which is the case when processing time series of equal lengths and restraining the search space using, for instance, a fixed size corridor symmetrically displayed around the diagonal as shown in Fig. 1. Note that any kind of corridor can be specified using the indicator functions h_d, h_s and h_i.

4.3 *Exponentiated* REDK

Definition 4.4: The following equations define the two RAfP function instances entering into the construction of the exponentiated REDK.

$$\mathcal{C}_{e,h}(A_1^p, B_1^q) = \sum \begin{cases} h(p-1,q)\mathcal{C}_{e,h}(A_1^{p-1}, B_1^q)\ e^{-\nu'\Gamma(A(p)\rightarrow\Lambda)} \\ h(p-1,q-1)\mathcal{C}_{e,h}(A_1^{p-1}, B_1^{q-1})\ e^{-\nu'\Gamma(A(p)\rightarrow B(q))} \\ h(p,q-1)\mathcal{C}_{e,h}(A_1^{p}, B_1^{q-1})\ e^{-\nu'\Gamma(\Lambda\rightarrow B(q))} \end{cases}$$

$$\mathcal{C}_{e,h,\delta}(A_1^p, B_1^q) = \sum \begin{cases} h(p-1,q)\mathcal{C}_{e,h,\delta}(A_1^{p-1}, B_1^q)\ e^{-\nu'\Gamma(A(p)\rightarrow\Lambda)} \\ \delta_{pq}\mathcal{C}_{e,h,\delta}(A_1^{p-1}, B_1^{q-1})\ e^{-\nu'\Gamma(A(p)\rightarrow B(q))} \\ h(p,q-1)\mathcal{C}_{e,h,\delta}(A_1^{p}, B_1^{q-1})\ e^{-\nu'\Gamma(\Lambda\rightarrow B(q))} \end{cases} \quad (4)$$

where ν' is a *stiffness* parameter that weighs the contribution of the local elementary costs. The larger ν' is, the more the kernel is selective around the optimal paths. At the limit, when $\nu' \rightarrow \infty$, only the optimal path costs are summed up by the kernel. Note that, as is generally seen, several optimal paths leading to the same global cost exist, therefore $\lim_{\nu'\rightarrow+\infty} -1/\nu' log(\mathcal{K}_e(A,B))$ does not coincide with the elastic distance δ that involves the same corresponding elementary costs, although we expect it to be close.

The alignment cost of two null time series being 0, we set $\mathcal{C}_{e,h}(\Omega,\Omega) = \mathcal{C}_{e,h,\delta}(\Omega,\Omega) = 1$ in this exponentiation context.

Proposition 4.5: Definiteness of the exponentiated REDK instance, $\mathcal{K}_e$:
$\mathcal{K}_e(.,.)$ is positive definite for the cost functions $\Gamma(A(p) \rightarrow \Lambda)$, $\Gamma(A(p) \rightarrow B(q))$ and $\Gamma(\Lambda \rightarrow B(q))$ involved in the computation of the δ_{lev}, δ_{dtw}, δ_{erp} and δ_{twed} distances.

The REDKs constructed from these distances are referred respectively to $REDK_{lev}$, $REDK_{erp}$, $REDK_{dtw}$, $REDK_{twed}$ in the rest of the paper.

The proof of proposition 4.5 is straightforward and is omitted.

4.3.1 Interpretation of the exponentiated REDK

In a REDK kernel, each alignment path is assigned with a product of *f-costs* that is the multiplication of the local cost functions attached to each edge of the path. For exponentiated REDK, the local cost function, e.g. $e^{-\nu'\Gamma(A(p)\rightarrow B(q))}$ can be interpreted, up to a normalizing constant, as a probability to align symbol $A(p)$ with symbol $B(q)$, and the value affected to each path can be interpreted as the probability of a specific alignment between two sequences. In this case the REDK, that sums up the probability of all possible alignment paths between two sequences, can be interpreted as a matching probability between two sequences. This probabilistic interpretation suggests an analogy between REDK and the *alpha-beta* algorithm designed to learn HMM models: while the Viterbi's algorithm that uses a *max* operator in a dynamic programming implementation (just like the *DTW* algorithm) evaluates only the probability of the best alignment path, the *alpha-beta* algorithm is based on the summation of the probabilities of all possible alignment paths. As reported in [25], the main drawbacks of these kind of kernels is the vanishing of the product of local cost functions (that

are lower than one) when comparing long sequences. When considering Gram-matrix (obtained from pairwise distances on finite sets) this leads to a matrix that suffers from the diagonal dominance problem, i.e. the fact that the kernel value decreases extremely fast when the similarity slightly decreases.

5 Classification experiments

The purpose of this experiment is to estimate the benefit one can expect from using PD elastic kernels instead of indefinite ones in SVM classification. It is clear that the the Sequential Minimal Optimization (SMO) involved in the SVM learning procedure [23] [7] is able to handle indefinite kernels, but, as the convergence toward a global extremum is not guaranteed (the optimization is problem is not convex), we can expect some better accuracy for definite kernels especially when the pairwise distance matrix derived from the train data set is far from definiteness. This is precisely what our experiment targets to demonstrate.

To that end, we empirically evaluate the effectiveness of some REDK kernels comparatively to Gaussian Radial Basis Function (RBF) Kernels or elastic distance substituting kernels [12] using some classification tasks on a set of time series coming from quite different application fields. The classification task we have considered consists of assigning one of the possible categories to an unknown time series for the 20 data sets available at the UCR repository [16]. As time is not explicitly given for these datasets, we used the index value of the samples as the time stamps for the whole experiment.

For each dataset, a training subset (TRAIN) is defined as well as an independent testing subset (TEST). We use the training sets to train two kinds of classifiers:

- the first one is a first near neighbor (1-NN) classifier: first we select a training data set containing time series for which the correct category is known. To assign a category to an unknown time series selected from a testing data set (different from the train set), we select its nearest neighbor (in the sense of a distance or similarity measure) within the training data set, then, assign the associated category to its nearest neighbor. For that experiment, a leave one out procedure is performed on the training dataset to optimized the meta parameters of the considered comparability measure.
- the second one is a SVM classifier [4], [34] configured with a Gaussian RBF kernel whose parameters are $C > 0$, a trade-off between regularization and constraint violation and σ that determines the width of the Gaussian function. To determine the C and σ hyper parameter values, we adopt a 10-folded cross-validation method on each training subset. According to this procedure, given a predefined training set TRAIN and a test set TEST, we adapt the meta parameters based on the training set TRAIN: we first divide TRAIN into 5 stratified subsets $TRAIN_1, TRAIN_2, \cdots, TRAIN_5$; then for each subset $TRAIN_i$ we use it as a new test set, and regard $(TRAIN - TRAIN_i)$ as a new training set; Based on the average error rate obtained on the five classification tasks, the optimal values of meta parameters are selected as the ones leading to the minimal average error rate.

We have used the LIBSVM library [5] to implement the SVM classifiers.

5.1 Experimenting with REDK

We tested the exponentiated REDK kernel based on the $\delta_{erp}, \delta_{dtw}, \delta_{twed}$ distance costs. We consider respectively the positive definite $REDK_{erp}, REDK_{dtw}, REDK_{twed}$ kernels. Our experiment compares classification errors on the test data for

- the first near neighbor classifiers based on the $\delta_{erp}, \delta_{dtw}, \delta_{twed}$ distance measures (1-NN δ_{erp}, 1-NN δ_{dtw} and 1-NN δ_{twed}),
- the SVM classifiers using Gaussian distance substituting kernels based on the same distances [40], e.g. SVM δ_{erp}, SVM δ_{dtw}, SVM δ_{twed},
- and their corresponding REDK kernel, SVM $REDK_{erp}$, SVM $REDK_{dtw}$, SVM $REDK_{twed}$.

For δ_{erp}, δ_{twed}, $REDK_{erp}$ and $REDKtwed$ we used the L1-norm, while the L2-norm has been implemented for δ_{dtw} and $REDK_{dtw}$, a classical choice for DTW [24].

5.1.1 Meta parameters

For δ_{erp} kernel, meta parameter g is optimized for each dataset on the train data by minimizing the classification error rate of a first near neighbor classifier using a Leave One Out (LOO) procedure. For this kernel, g is selected in $\{-3, -2.99, -2.98, \cdots, 2.98, 2.99, 3\}$. This optimized value is also used for comparison with the $REDK_{erp}$ kernel.

For δ_{twed} kernel, meta parameters λ and ν are optimized for each dataset on the train data by minimizing the classification error rate of a first near neighbor classifier. For our experiment, the *stiffness* value (ν) is selected from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and λ is selected from $\{0, .25, .5, .75, 1.0\}$. If different (ν, λ) values lead to the minimal error rate estimated for the training data then the pairs containing the highest ν value are selected first, then the pair with the highest λ value is finally selected. These optimized (λ, ν)values are also used for comparability purposes with the $REDK_{twed}$ kernel.

The kernels exploited by the SVM classifiers are the Gaussian Radial Basis Function (RBF) kernels $K(A, B) = e^{\delta(A,B)^2/(2\sigma^2)}$ where δ stands for $\delta_{erp}, \delta_{dtw}, \delta_{twed}$, $REDK_{erp}$, $REDK_{dtw}$, $REDK_{twed}$. To limit the search space for the SVM meta parameters, the pairwise distances are normalized using the *min*

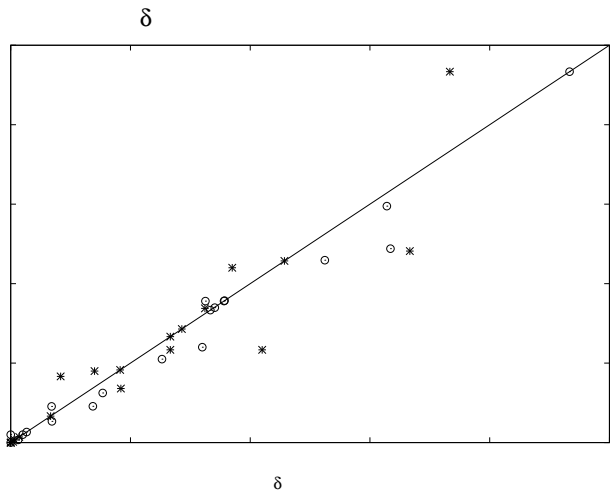

Fig. 2. Comparison of error rates (in %) between two SVM classifiers, the first one based on the δ_{erp} substituting kernel (SVM δ_{erp}), and the second one based on a REDK kernel induced by the δ_{erp} distance (SVM $REDK_{erp}$). The straight line has a slope of 1.0 and dots correspond, for the pair of classifiers, to the error rates on the train (star) or test (circle) data sets. A dot below (resp. above) the straight line indicates that SVM $REDK_{erp}$ has a lower (resp. higher) error rate than distance SVM δ_{erp}

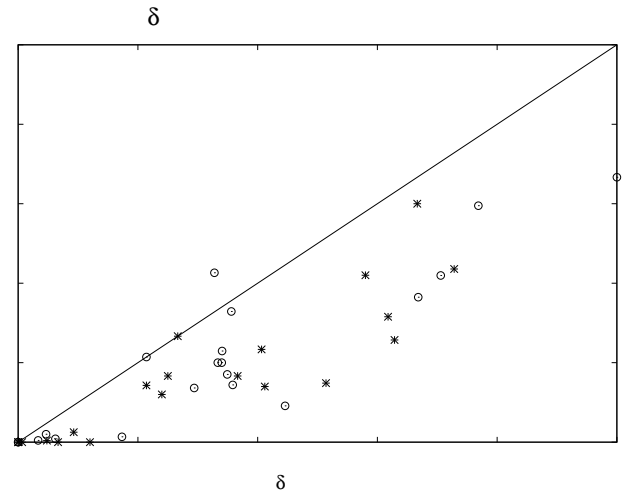

Fig. 3. Comparison of error rates (in %) between two SVM classifiers, the first one based on the δ_{dtw} substituting kernel (SVM δ_{dtw}), and the second one based on a REDK induced by the δ_{dtw} distance (SVM $REDK_{dtw}$)). The straight line has a slope of 1.0 and dots correspond, for the pair of classifiers, to the error rates on the train (star) or test (circle) data sets. A dot below (resp. above) the straight line indicates that SVM $REDK_{dtw}$ has a lower (resp. higher) error rate than distance δ_{dtw}

and *max* values calculated on the TRAIN data, basically $\delta = (\delta - \delta_{min})/(\delta_{max} - \delta_{min})$.

Meta parameter C is selected from $\{2^{-5}, 2^{-4}, ..., 1, 2, ..., 2^{10}\}$, and σ^2 from $\{2^{-5}, 2^{-4}, ..., 1, 2, ..., 2^{14}\}$. The best values are obtained using the 10-folds cross validation procedure.

For the $REDK_{erp}$, $REDK_{dtw}$ and $REDK_{twed}$ kernels, meta parameter $1/\nu'$ is selected from the discrete set $S = \{10^{-5}, 10^{-4}, ..., 1, 10, 100\}$.

The optimization procedure is as follows:

- for each value in S, we train a SVM $REDK_*$ classifier on the training dataset using the previously described 10-folded cross validation procedure to select the SVM meta parameters *cost* and σ and the average of the classification error is recorded.
- the best σ, C and ν' values are the ones that lead to the minimal average error.

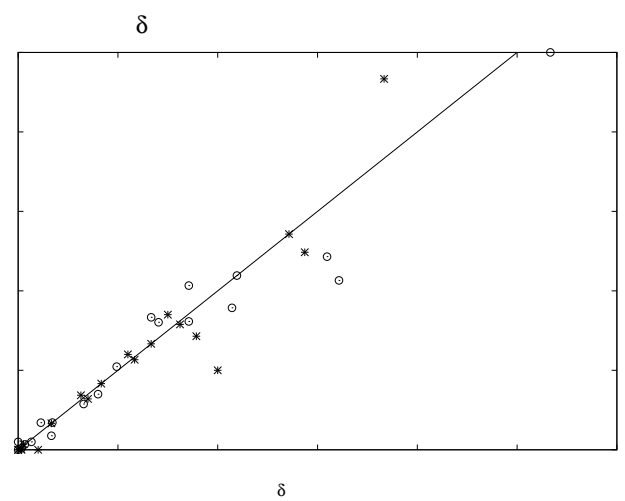

Fig. 4. Comparison of error rates (in %) between two SVM classifiers, the first one based on the δ_{twed} substituting kernel (SVM δ_{twed}), and the second one based on a REDK induced by the δ_{erp} distance (SVM $REDK_{twed}$). The straight line has a slope of 1.0 and dots correspond, for the pair of classifiers, to the error rates on the train (star) or test (circle) data sets. A dot below (resp. above) the straight line indicates that SVM $REDK_{twed}$ has a lower (resp. higher) error rate than distance SVM δ_{twed}

5.2 Discussion

5.2.1 REDK experiment analysis

Tables 1 and 2 show the classification error rates obtained for the tested methods, e.g. the first near neighbor classifier based on the δ_{erp}, δ_{dtw} and δ_{twed} distances (1-NN δ_{erp}, 1-NN δ_{dtw} and 1-NN δ_{twed}), the Gaussian RBF kernel SVM based on the same distances (SVM δ_{erp}, SVM δ_{dtw} and SVM δ_{twed}) and Euclidean distance, and the Gaussian RBF kernel SVM based on the REDK kernels (SVM $REDK_{erp}$, SVM $REDK_{dtw}$ and SVM $REDK_{twed}$).

TABLE 1
Comparative study using the UCR datasets: classification error rates (in %) obtained using the first near neighbor classification rule and a SVM classifier for the erp, $REDK_{erp}$, dtw and $REDK_{dtw}$ kernels. Two scores are given S1|S2: the first one, S1, is evaluated on the training data, while the second one, S2, is evaluated on the test data.

DATASET	1-NN δ_{erp}	SVM δ_{erp}	SVM $REDK_{erp}$	1-NN δ_{dtw}	SVM δ_{dtw}	SVM $REDK_{dtw}$
Synthetic control	0.67|3.7	**0.33|1**	**0.33|1**	0|2.33	**0**|2.33	**0|1**
Gun-Point	6.12|4	**0|1.33**	**0|1.33**	18.36|9.3	6|8.67	**0|.67**
CBF	0|0.33	**3.33**|3.44	**3.33|2.67**	0|0.33	3.33|1.67	**0|0.22**
Face (all)	10.73|20.18	**.71**|17.04	**.71|16.98**	6.8|19.23	4.64|17.46	**1.25|8.52**
OSU Leaf	30.15|40.08	**18.5**|31.41	22|**29.75**	29|44.63	29|38.43	**21|29.75**
Swedish Leaf	11.02|12	9.2|7.68	**6.8|6.24**	24.65|20.8	20.6|17.92	**7|7.2**
50 Words	19.38|28.13	**16.22|16.26**	16.89|17.80	33.18|31	30.89|33.41	**15.78|18.24**
Trace	10.01|17	**0|0**	**0**|1	0|0	**0|0**	**0|0**
Two Patterns	0|0	**0|0**	**0|0**	0|0	**0|0**	**0|0**
Wafer	.1|0.9	.2|0.62	**0.1|0.34**	1.4|2.01	2.4|3.11	**0.2|0.42**
face (four)	4.35|10.2	**4.17|3.41**	8.33|4.55	26.09|17.05	12.5|14.7	**8.33|6.82**
Ligthing2	11.86|14.75	13.33|26.23	**11.67|22.95**	13.56|13.1	18.33|**16.39**	**8.33**|21.31
Ligthing7	23.19|30.1	**22.86|17.81**	**22.86|17.81**	33.33|27.4	31.43|17.80	**12.86|16.44**
ECG	10.01|13	**7**|16	9|**12**	23.23|23	12|17	**6|10**
Adiac	35.99|37.85	33.33|31.71	**24.10|24.4**	40.62|39.64	36.41|35.29	**21.79|20.97**
Yoga	14.05|14.7	21|12.63	**11.67|10.5**	16.37|16.4	20.33|17.03	**11.67|11.47**
Fish	16.09|12	**9.14**|6.86	**9.14|4.57**	26.44|16.57	25.71|22.29	**7.43|4.57**
Coffee	25.93|25	**14.29|17.86**	**14.29|17.86**	14.81|17.86	10.71|**10.71**	**7.14|10.71**
OliveOil	17.24|16.67	**13.33|16.67**	**13.33|16.67**	13.79|13.33	**13.33**|16.67	**13.33|10**
Beef	68.97|50	**36.67|46.67**	46.67|**46.67**	46.67 |50	33.33|50	**30|33.33**
# Best Scores	-	**15**/10	**15/17**	-	4/4	**20/19**
# Uniquely Best Scores	-	3/4	**5/10**	-	0/1	**16/16**

In this experiment, we show that the SVM classifiers clearly outperform in general the 1-NN classifiers that are used as a baseline. But the interesting results reported in tables 1 and 2 and figures 2, 3 and 4 is that SVM $REDK_{erp}$ and SVM $REDK_{twed}$ perform slightly better than SVM δ_{erp} and SVM δ_{twed} respectively, and the SVM $REDK_{dtw}$ is clearly much efficient than the SVM δ_{dtw}. This could come from the fact that δ_{erp} and δ_{twed} are metrics but not δ_{dtw}, but another explanation could be related to the the SMO optimization involved into the SVM learning. SVM δ_{dtw} poorly behaves compared to the other tested classifiers, probably because the SVM optimization process does not perform well. Nevertheless, the $REDK_{dtw}$ kernel based on δ_{dtw} greatly improves the classification results. To further explore the potential impact of indefiniteness on classification rates, we give in table 3 two quantified hints of deviation to conditionally definiteness for the distance-matrices corresponding to the δ_{dtw}, δ_{erp} and δ_{twed} distances. Since to be conditionally definite (negative) a pairwise distance-matrix should have a single positive eigenvalue, the first hint is the number of positive eigenvalues $\#Pev$ (we give also as a reference the total number of eigenvalues, $\#Ev$). The second hint, $\Delta_p = 100 * \frac{\sum_{ev_i>0}(ev_i) - \text{ArgMax}_{ev_i>0}\{ev_i\}}{\sum_{ev_i>0} ev_i}$, where ev_i is an eigenvalue of the distance matrix, quantifies the weight of the extra positive eigenvalues relatively to the weight of the total number of positive eigenvalues. Therefore, a conditionally definite (negative) matrix should be such that simultaneously $\#Pev = 0$ and $\Delta_p = 0$.

By examining the distance-matrices corresponding to each training datasets and for each distances $\delta_{dtw}(A, B)$, $\delta_{erp}(A, B)$ and $\delta_{twed}(A, B)$, we can show that the $\delta_d tw$ kernel is much further away from a conditionally definite matrix than the δ_{erp} and δ_{twed} kernels. The distance that is closer to conditional definiteness is the δ_{twed} distance. This is clearly quantifiable by computing the number of positive eigenvalues, as well as their amplitudes, of the pairwise distance matrices. For datasets such as *FaceAll*, *OSULeaf*, *SwedishLeaf*, *50words*, *Adiac*, etc., all the three distances lead to indefinite pairwise distance matrices. The REDK regularization brings in general a significant improvement, specifically when the number and amplitudes of the extra positive eigenvalues are high. Furthermore, for datasets of small sizes (such as *CBF*, *Beef*, *Coffee*, *OliveOil*, etc.), δ_{erp} and δ_{twed} produce conditionally definite matrices where $\delta_d tw$ does not. One can see that the regularization brought by REDK is much more effective on $\delta_d tw$ on these data set. Nevertheless, some datasets are better classified by SVM that use directly the distance substituting kernel instead of the derived REDK regularized kernel.

To conclude, our experiment clearly shows that for problems involving Gram-matrices *far* from definiteness (more than 5-10% of eigenvalues are positive, see Table 4), the REDK regularized version outperforms the original elastic distance substituting kernel. This is particularly true for δ_{dtw}. For δ_{twed}, or even δ_{erp} distances, most of the training sets lead to pairwise distance-matrices that are either conditionally definite negative or close to CND matrices (with very few extra positive eigenvalues). Nevertheless, in most cases, the REDK-δ_{erp} and REDK-δ_{twed} outperformed δ_{erp} and δ_{twed} respectively, showing some robustness of the REDK regularization. The main drawback of REDK is the extra parameter ν'.

TABLE 2
Comparative study using the UCR datasets: classification error rates (in %) obtained using the first near neighbor classification rule and a SVM classifier for the δ_{twed} and $REDK_{twed}$ kernels. Two scores are given S1|S2: the first one, S1, is evaluated on the training data, while the second one, S2, is evaluated on the test data.

DATASET	1-NN δ_{twed}	SVM δ_{twed}	SVM $REDK_{twed}$
Synthetic control	1|2.33	**0**|1.33	**0**|**1**
Gun-Point	0|1.33	2|**0.67**	**0**|**0.67**
CBF	0|0.67	**3.33**|3.33	**3.33**|**1.77**
Face (all)	1.43|18.93	**0.54**|17.10	0.71|**16.15**
OSU Leaf	17.59|24.79	**15**|**17.36**	17|20.66
Swedish Leaf	8.82|10.24	7|6.56	**6.4**|**5.76**
50 Words	18.26|18.9	16.22|**14.07**	**15.78**|16.04
Trace	1|5	**0**|**0**	**0**|1
Two Patterns	0|0.12	**0**|**0**	**0**|**0**
Wafer	0.1|.86	0.3|0.34	**0.1**|**0.2**
face (four)	8.7|3.41	**8.33**|**2.27**	**8.33**|3.41
Ligthing2	13.56|21.31	20|32.15	**10**|**21.31**
Ligthing7	24.64|24.66	**27.14**|**21.92**	**27.14**|**21.92**
ECG	13.13|10	**11**|8	12|**7**
Adiac	36.25|37.6	28.72|30.95	**24.85**|**24.30**
Yoga	19.06|12.97	11.67|**9.87**	**11.33**|10.46
Fish	12.07|5.14	**6.29**|**3.43**	6.86|**3.43**
Coffee	18.52|21.43	17.86|21.43	**14.29**|**17.86**
OliveOil	11.11|16.67	**13.33**|**13.33**	**13.33**|16.67
Beef	58.62|53.3	**36.67**|53.33	46.67|**50**
# Best Scores	-	12|10	**15**|**14**
# Uniquely Best Scores	-	5|5	**8**|**10**

This extra parameter ν' makes the search for an optimal setting on the train data more difficult and requires more learning data to converge. The trade-off between learning and generalization is therefore more complex to optimize.

6 Conclusion

Following the work on convolution kernels [14] and local alignment kernels defined for string processing around the Smith and Waterman algorithm [32] [25], or defined for time series on the basis of the DTW similarity measure [11], we have addressed the definiteness of elastic distances through the construction of positive definite kernels built from the recursive definitions of elastic (editing) distances themselves and achieve some extension and generalization of these previous results. Contrary to other approaches that rely on the regularization of the Gram-matrix [39][8], this approach does not depend at all on the data, excpet for the optimization of a regularizing parameter (ν'). By adding an extra recursive term we achieve some simple sufficient conditions, that are weaker than those proposed in [25] or [11], allowing to build positive definite exponentiated REDK. These conditions are basically satisfied by any classical elastic distances defined by a recursive equation. In particular they can be applied to the edit distance, the well known Dynamic Time Warping measure and to some variants such as the Edit Distance With Real penalty and the Time Warp Edit Distance, the latter two being metrics as well as the symbolic edit distance. Furthermore, our results apply when a symmetric *corridor* is used to limit the search space required to evaluate the elastic distance, thus reducing consequently the computation time.

The experiments conducted on a variety of time series datasets show that the positive definite REDKs outperforms in general the indefinite elastic distances from which they are derived, when considering 1-NN and SVM classification tasks. This is mostly striking when the Gram-matrix evaluated on the train data sets is *far* from definiteness, which is the case when DTW is involved.

Appendix A
Indefiniteness of some elastic measures

A.1 The Levenshtein distance

The Levenshtein distance kernel $\varphi(x,y) = \delta_{lev}(x,y)$ is known to be indefinite. Below, we discuss the first known counter-example produced by [9]. Let us consider the subset of sequences $V = \{abc, bad, dab, adc, bcd\}$ that leads to the following distance matrix

$$M_{lev}^V = \begin{pmatrix} 0 & 3 & 2 & 1 & 2 \\ 3 & 0 & 2 & 2 & 1 \\ 2 & 2 & 0 & 3 & 3 \\ 1 & 2 & 3 & 0 & 3 \\ 2 & 1 & 3 & 3 & 0 \end{pmatrix} \tag{5}$$

and consider coefficient vectors C and D in $\mathbb{R}^5$ such that
$C = [1, 1, -2/3, -2/3, -2/3]$ with $\sum_{i=1}^5 c_i = 0$ and
$D = [1/3, 2/3, 1/3, -2/3, -2/3]$ with $\sum_{i=1}^5 d_i = 0$.

Clearly $CM_{lev}^V C^T = 2/3 > 0$ and $DM_{lev}^V D^T = -4/3 < 0$, showing that M_{lev}^V has no definiteness.

TABLE 3
Analysis of the deviation to conditionally definiteness for the distance-matrices associated to the δ_{dtw}, δ_{erp} and δ_{twed} distances. We report for each dataset the number of positive eigenvalues ($\#Pev$) relatively to the total number of eigenvalues ($\#Ev$)and the deviation to definiteness estimated as Δ_p that expresses in %. The expectation is a single positive eigenvalue, $\#Pev = 1$, corresponding to $\Delta_p = 0\%$.

DATASET	δ_{dtw}		δ_{erp}		δ_{twed}	
-	#Pev/#Ev	Δ_p	#Pev/#Ev	Δ_p	#Pev/#Ev	Δ_p
Synthetic control	110/300	15.66%	8/300	.16%	6/300	.22 %
Gun-Point	23/50	2.54%	1/50	0%	1/50	0%
CBF	5/30	3.36%	1/30	0%	1/30	0%
Face (all)	242/560	26.6%	83/560	2.42%	41/560	1.89%
OSU Leaf	96/200	31.79%	29/200	2.97%	16/200	.89%
Swedish Leaf	206/500	17.04%	24/500	.68%	23/500	.41%
50 Words	218/450	34.03%	119/450	9.54%	93/450	4.85%
Trace	43/100	5.42%	1/100	0%	1/100	0%
Two Patterns	453/1000	36.7%	259/1000	13.8%	226/1000	9.85%
Wafer	497/1000	14.84%	137/1000	1.29%	39/1000	.04%
face (four)	2/24	.74%	1/24	0%	1/24	0%
Ligthing2	20/60	13.44%	1/60	0%	1/60	0%
Ligthing7	24/70	14.25%	1/70	0%	1/70	0%
ECG	38/100	14.7%	1/100	0%	1/100	0%
Adiac	159/390	5.54%	26/390	.82%	39/390	.69%
Yoga	142/300	23.4%	29/300	3.17%	10/300	.41%
Fish	71/175	17.57%	1/175	0%	1/175	0%
Coffee	12/28	8.83%	1/28	0%	1/28	0%
OliveOil	4/30	.24%	1/30	0%	1/30	0%
Beef	15/30	6.17%	1/30	0%	1/30	0%

A.2 The Dynamic Time Warping distance

The DTW kernel $\varphi(x, y) = \delta_{dtw}(x, y)$ is also known not to be conditionally definite. The following example demonstrates this known result. Let us consider the subset of sequences $V = \{01, 012, 0122, 01222\}$.

Then the DTW distance matrix evaluated on V is

$$M^V_{dtw} = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 \end{pmatrix} \tag{6}$$

and consider coefficient vectors C and D in $\mathbb{R}^4$ such that
$C = [1/4, -3/8, -1/8, 1/4]$ with $\sum_{i=1}^4 c_i = 0$ and
$D = [-1/4, -1/4, 1/4, 1/4]$ with $\sum_{i=1}^4 d_i = 0$. Clearly $CM^V_{dtw}C^T = 2/32 > 0$ and $DM^V_{dtw}D^T = -1/2 < 0$, showing that M^V_{dtw} has no definiteness.

A.3 The Time Warp Edit Distance

Similarly, it is easy to find simple counter examples that show that TWED kernels are not definite.

Let us consider the subset of sequences $V = \{010, 012, 103, 301, 032, 123, 023, 003, 302, 321\}$.

For the TWED metric, with $\nu = 1.0$ and $\lambda = 0.0$ we get the following matrix:

$$M^V_{twed} = \begin{pmatrix} 0 & 2 & 7 & 9 & 6 & 7 & 5 & 5 & 10 & 9 \\ 2 & 0 & 5 & 9 & 4 & 5 & 3 & 3 & 8 & 9 \\ 7 & 5 & 0 & 6 & 7 & 4 & 6 & 2 & 5 & 10 \\ 9 & 9 & 6 & 0 & 13 & 10 & 12 & 8 & 1 & 4 \\ 6 & 4 & 7 & 13 & 0 & 5 & 3 & 5 & 12 & 9 \\ 7 & 5 & 4 & 10 & 5 & 0 & 2 & 6 & 9 & 6 \\ 5 & 3 & 6 & 12 & 3 & 2 & 0 & 4 & 11 & 8 \\ 5 & 3 & 2 & 8 & 5 & 6 & 4 & 0 & 7 & 10 \\ 10 & 8 & 5 & 1 & 12 & 9 & 11 & 7 & 0 & 5 \\ 9 & 9 & 10 & 4 & 9 & 6 & 8 & 10 & 5 & 0 \end{pmatrix} \tag{7}$$

The eigenvalue spectrum for this matrix is the following:
$\{$ 4.62, 0.04, -2.14, -0.98, -0.72, -0.37, -0.19, -0.17, -0.06, -0.03 $\}$. This spectrum contains 2 strictly positive eigenvalues, showing that M^V_{twed} has no definiteness.

A.4 The Edit Distance with Real Penalty

For the ERP metric, with $g = 0.0$ we get the following matrix:

$$M^V_{erp} = \begin{pmatrix} 0 & 2 & 3 & 3 & 4 & 5 & 4 & 2 & 4 & 5 \\ 2 & 0 & 3 & 5 & 2 & 3 & 2 & 2 & 4 & 5 \\ 3 & 3 & 0 & 4 & 3 & 2 & 3 & 1 & 3 & 4 \\ 3 & 5 & 4 & 0 & 7 & 6 & 7 & 5 & 1 & 2 \\ 4 & 2 & 3 & 7 & 0 & 3 & 2 & 2 & 6 & 5 \\ 5 & 3 & 2 & 6 & 3 & 0 & 1 & 3 & 5 & 4 \\ 4 & 2 & 3 & 7 & 2 & 1 & 0 & 2 & 6 & 5 \\ 2 & 2 & 1 & 5 & 2 & 3 & 2 & 0 & 4 & 5 \\ 4 & 4 & 3 & 1 & 6 & 5 & 6 & 4 & 0 & 1 \\ 5 & 5 & 4 & 2 & 5 & 4 & 5 & 5 & 1 & 0 \end{pmatrix} \tag{8}$$

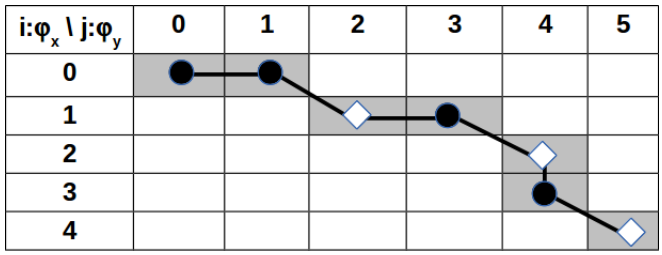

Fig. 5. Example of an alignment path corresponding to the alignment map $(0,0)(0,1)(1,2)(1,3)(2,4)(3,4)(4,5)$. The black squares correspond to substitution operations and black circles to either deletion or insertion operations.

The eigenvalue spectrum for this matrix is the following:
$\{$ 4.63, 0.02, $1.39e-17$, -2.21, -0.97, -0.56, -0.41, -0.26, -0.17, -0.08 $\}$. This spectrum contains 3 strictly positive eigenvalues (although the third positive eigenvalue which is very small could be the result of the imprecision of the used diagonalization algorithm), showing that M^V_{erp} has no definiteness.

Appendix B
Proof of our main result

B.1 Proof of theorem 4.3

Definition B.1: Let $\mathbb{U}_n$ be the subset of $\mathbb{U}$ containing all the sequences whose lengths are lower or equal to n.

Definition B.2: For all $(a,b) \in ((S \times T) \cup \{\Lambda\})^2$ let $\kappa(a,b)$ be the local kernel defined as follows:
$\kappa(a,b) = f(\Gamma(a \to b))$
whre Λ stands for the *null* symbol.

Definition B.3: Let π be an ordered alignment map between two finite non empty sequences of successive integers of the form $0,..,n$. Basically π is a finite sequence of pairs of integers $\pi(l) = (i_l, j_l)$ for $l \in \{0,...,|\pi|-1\}$, satisfying the following conditions

- $0 \leq i_l, \forall l \in 0,..,|\pi|-1$
- $i_l \leq i_{l-1}+1, \forall l \in 1,..,|\pi|-1$
- $j_l \leq j_{l-1}+1, \forall l \in 1,..,|\pi|-1$
- $i_{l-1} < i_l$ or $j_{l-1} < j_l, \forall l \in \{1,..,|\pi|-1\}$

$\pi_x(l) = i_l$ and $\pi_y(l) = j_l$ are the two coordinate access functions for the lth pair of mapped integers so that $\pi(l) = (\pi_x(l), \pi_y(l))$.

For all $n \geq 1$, let $\mathcal{M}_n$ be the set of alignment maps π such that the two sets of integers mapped by π are $\{1 \cdots n\} \times \{1 \cdots n\}$. By convention we set $\mathcal{M}_0 = \emptyset$.

As shown in figure 5, there exists a direct correspondence between an alignment map and an alignment path i.e. a finite sequence of editing operations.

Definition B.4: For all n we introduce two embeddings, basically two vectorized representations, $\varphi_{\pi_x} : \mathbb{U}_n \to ((S \times T) \cup \{\Lambda\})^{2n}$ and $\varphi_{\pi_y} : \mathbb{U}_n \to ((S \times T) \cup \{\Lambda\})^{2n}$ induced by any mapping $\pi \in \mathcal{M}_n$. These mappings are uniquely obtained using algorithm 1.

Algorithm 1 For any $n > 0$ and any sequence $A \in \mathbb{U}_n$, uniquely provides two embeddings $\varphi_{\pi_x}(A)$ and $\varphi_{\pi_y}(A)$ in $((S \times T) \cup \{\Lambda\})^{2n}$ given any alignment map $\pi \in \mathcal{M}_n$.

```
Require: n > 0,  π ∈ M_n,  A ∈ U_n
  l ← 1; i_l ← 1; j_l ← 1;
  while l ≤ 2n do
    if  l < |π| then
      if π_x(l) = π_x(l − 1) + 1 and π_y(l) = π_y(l − 1) + 1
      then
        if π_x(l) < |A| then
          φ_{π_x}(A)_l = A(π_x(l)); i_l ← i_l + 1;
        else
          φ_{π_x}(A)_l = Λ;
        end if
        if π_y(l) < |A| then
          φ_{π_y}(A)_l = A(π_y(l)); j_l ← j_l + 1;
        else
          φ_{π_y}(A)_l = Λ;
        end if
      else if π_x(l) = π_x(l − 1) and π_y(l) = π_y(l − 1) + 1
      then
        φ_{π_x}(A)_l = Λ;
        if π_y(l) < |A| then
          φ_{π_y}(A)_l = A(π_y(l)); j_l ← j_l + 1;
        else
          φ_{π_y}(A)_l = Λ;
        end if
      else
        φ_{π_y}(A)_l = Λ;
        if π_x(l) < |A| then
          φ_{π_x}(A)_l = A(π_x(l)); i_l ← i_l + 1;
        else
          φ_{π_x}(A)_l = Λ;
        end if
      end if
    end if
    if l ≥ |π| and i_l < |A| then
      φ_{π_x}(A)_l = A(i_l); i_l ← i_l + 1;
    else
      φ_{π_x}(A)_l = Λ;
    end if
    if l ≥ |π| and j_l < |A| then
      φ_{π_y}(A)_l = A(j_l); j_l ← j_l + 1;
    else
      φ_{π_x}(A)_l = Λ;
    end if
    l ← l + 1;
  end while
  return  φ_{π_x}(A), φ_{π_y}(A)
```

For example, for a sequence A of length 5 the two embeddings in $((S \times T) \cup \{\Lambda\})^{2\times 5}$ corresponding to the

alignment path depicted in Figure 5 are

$$\varphi_{\pi_x}(A) = \Lambda A(1)\Lambda A(2)A(3)A(4)A(5)\Lambda\Lambda\Lambda$$
$$\varphi_{\pi_y}(A) = A(1)A(2)A(3)A(4)\Lambda A(5)\Lambda\Lambda\Lambda\Lambda$$

where Λ is the *null* symbol.

Then, for any $A \in \mathbb{U}_n$ and $\pi \in \mathcal{M}_n$, we call $\mathcal{P}_\pi(A) = \{\varphi_{\pi_x}(A), \varphi_{\pi_y}(A)\}$ the set of embeddings (or parts) for sequence A induced by π. Note that all these embeddings are sequences whose lengths are $2n$. Even if cases for which $\varphi_{\pi_x}(A) = \varphi_{\pi_y}(A)$ exist (e.g. the main diagonal path), we always consider $\varphi_{\pi_x}(A)$ and $\varphi_{\pi_y}(A)$ as distinct complementary parts of A.

Proposition B.5: If the local kernel $k(x, y) = f(\Gamma(x \to y))$ is positive definite on $((S \times T) \cup \{\Lambda\})^2$ then $\forall n \geq 1$ and $\forall \pi \in \mathcal{M}_n$

$$K_\pi(A, B) = \sum_{\varphi(A)\in\mathcal{P}_\pi(A)} \sum_{\varphi(B)\in\mathcal{P}_\pi(B)} \prod_{p=1}^{2n} k(\varphi(A)_p, \varphi(B)_p) \tag{9}$$

is a p.d. kernel on $(\mathbb{U}_n)^2$.

Proof of lemma B.5 is a direct consequence of the Haussler's *R-convolution* kernel theorem [14]. Indeed, since $k(x, y)$ is a p.d. kernel on $((S \times T) \cup \{\Lambda\})^2$, and, considering the sets of parts $\mathcal{P}_\pi(A)$ and $\mathcal{P}_\pi(B)$ associated respectively to the sequences A and B, the conditions for the Haussler's *R-convolution* are satisfied.

Note that $K_\pi(A, B)$ simply rewrites as

$$\begin{aligned} K_\pi(A, B) = &\prod_{p=1}^{2n} k(\varphi_{\pi_x}(A)_p, \varphi_{\pi_y}(B)_p)) \\ &+ \prod_{p=1}^{2n} k(\varphi_{\pi_y}(A)_p, \varphi_{\pi_x}(B)_p)) \\ &+ \prod_{p=1}^{2n} k(\varphi_{\pi_x}(A)_p, \varphi_{\pi_x}(B)_p)) \\ &+ \prod_{p=1}^{2n} k(\varphi_{\pi_y}(A)_p, \varphi_{\pi_y}(B)_p)) \end{aligned} \tag{10}$$

Let $\mathcal{M}_{n,h} \subset \mathcal{M}_n$ be the subset of paths that are restricted by the symmetric corridor induced by any symmetric indicator function h entering into the construction of the RAfP functions $\mathcal{C}_h(.,.)$ and $\mathcal{C}_{h\delta}(.,.)$.

Proposition B.6: For any $n \in \mathbb{N}$, any $\pi \in \mathcal{M}_{n,h}$, and any $(A, B) \in (\mathbb{U}_n)^2$, we have

$$\mathcal{K}(A, B) = \frac{1}{2} \sum_{\pi\in\mathcal{M}_{n,h}} K_\pi(A, B) \tag{11}$$

We state first that $\forall \pi \in \mathcal{M}_{n,h}$, $\exists! \tilde{\pi} \in \mathcal{M}_{n,h}$ such that $\pi_x = \tilde{\pi}_y$ and $\pi_y = \tilde{\pi}_x$. $(\pi, \tilde{\pi})$ is a pair of symmetrical paths with respect to the main diagonal.

Thus, since $\mathcal{C}_h(.,.)$ evaluates the sum of the *cost-product* along all paths in $\mathcal{M}_{n,h}$, that is

$$\mathcal{C}_h(A, B) = \sum_{\pi\in\mathcal{M}_{n,h}} \prod_{p=1}^{2n} k(\varphi_{\pi_x}(A)_p, \varphi_{\pi_y}(B)_p) \tag{12}$$

we can rewrite rewrites

$$\begin{aligned} \mathcal{C}_h(A, B) = &\sum_{\pi\in\mathcal{M}_{n,h}} \left(\frac{1}{2} \prod_{p=1}^{2n} k(\varphi_{\pi_x}(A)_p, \varphi_{\pi_y}(B)_p) \right. \\ &\left. + \frac{1}{2} \prod_{p=1}^{2n} k(\varphi_{\tilde{\pi}_x}(A)_p, \varphi_{\tilde{\pi}_y}(B)_p) \right) \\ = &\frac{1}{2} \sum_{\pi\in\mathcal{M}_{n,h}} \left(\prod_{p=1}^{2n} k(\varphi_{\pi_x}(A)_p, \varphi_{\pi_y}(B)_p) \right. \\ &\left. + \prod_{p=1}^{2n} k(\varphi_{\pi_y}(A)_p, \varphi_{\pi_x}(B)_p) \right) \end{aligned}$$

Similarly, for $\mathcal{C}_{h\delta}(.,.)$ we get

$$\begin{aligned} \mathcal{C}_{h,\delta}(A, B) = \frac{1}{2} \sum_{\pi\in\mathcal{M}_{n,h}} &\left(\prod_{p=1}^{2n} k(\varphi_{\pi_x}(A)_p, \varphi_{\pi_x}(B)_p) \right. \\ &\left. \prod_{p=1}^{2n} k(\varphi_{\pi_y}(A)_p, \varphi_{\pi_y}(B)_p) \right) \end{aligned}$$

Hence, the expected result.

Proof of Proposition 4.3 (Definiteness of REDK)

Proposition B.6 states that $\mathcal{K}(.,.)$ is a finite sum of p.d. kernels defined on $(\mathbb{U}_n)^2$. According to the closure properties under the addition of p.d. kernels, we show that $\mathcal{K}(.,.)$ is a p.d. kernel defined on $(\mathbb{U}_n)^2$.

As this result is true for all n, one can see $\mathcal{K}(.,.)$ as a point-wise limit as n tends toward infinity of a p.d. kernel defined on $(\mathbb{U}_n)^2$. This establishes that $\mathcal{K}(.,.B)$ is a p.d. kernel on $\mathbb{U}^2$ $\square$.

ACKNOWLEDGMENTS

The authors thank the French Ministry of Research, the Brittany Region, the General Council of Morbihan and the European Regional Development Fund that had partially fund this research. The authors also thank Dr. E. Keogh's team at UC Riverside for making available the time series data sets that have been used in this study.

REFERENCES

[1] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. A basic local alignment serach tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[2] R. Bellman. *Dynamic Programming*. Princeton Univ Press, 1957. New Jersey.

[3] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*, volume 100 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, April 1984.

[4] Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In *COLT*, pages 144–152, 1992.

[5] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001.

[6] L. Chen and R. Ng. On the marriage of lp-norm and edit distance. In *Proceedings of the 30th International Conference on Very Large Data Bases*, pages 792–801, 2004.

[7] Pai-Hsuen Chen, Rong-En Fan, and Chih-Jen Lin. A study on smo-type decomposition methods for support vector machines. *Neural Networks, IEEE Transactions on*, 17(4):893–908, 2006.

[8] Yihua Chen, Eric K. Garcia, Maya R. Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based classification: Concepts and algorithms. *J. Mach. Learn. Res.*, 10:747–776, June 2009.

[9] Corinna Cortes, Patrick Haffner, and Mehryar Mohri. Positive Definite Rational Kernels. In *Proceedings of COLT'03*, volume 2777 of *Lecture Notes in Computer Science*, pages 41–56, Washington D.C., August 2003. Springer, Heidelberg, Germany.

[10] Corinna Cortes, Patrick Haffner, and Mehryar Mohri. Rational kernels: Theory and algorithms. *Journal of Machine Learning Research*, 5:1035–1062, 2004.

[11] M. Cuturi, J-P. Vert, O. Birkenes, and T. Matsui. A Kernel for Time Series Based on Global Alignments. In *Proceedings of ICASSP'07*, Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, pages II–413 – II–416, Honolulu, HI, April 2007. IEEE.

[12] B. Haasdonk and C. Bahlmann. Learning with distance substitution kernels. In *DAGM-Symposium*, pages 220–227, 2004.

[13] Bernard Haasdonk. Feature space interpretation of svms with indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:482–492, 2005.

[14] D. Haussler. Convolution kernels on discrete structures. Technical report, University of California, Santa Cruz, 2008. Technical Report.

[15] Akira Hayashi, Yuko Mizuhara, and Nobuo Suematsu. Embedding time series data for classification. In *MLDM*, pages 356–365, 2005.

[16] E. J. Keogh, X. Xi, L. Wei, and C.A. Ratanamahatana. The UCR time series classification-clustering datasets, 2006. http://wwwcs.ucr.edu/ eamonn/time_series_data/.

[17] Karthik Kumara, Rahul Agrawal, and Chiranjib Bhattacharyya. A large margin approach for writer independent online handwriting classification. *Pattern Recogn. Lett.*, 29:933–937, May 2008.

[18] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR, 163(4):845-848, 1965 (Russian)*, pages 707–710, 1966. English translation in Soviet Physics Doklady, 10(8).

[19] P. F. Marteau. Time warp edit distance. Technical report, VALORIA, Universite de Bretagne Sud, 2008. Technical Report valoriaUBS-2008-3v, http://hal.archives-ouvertes.fr/hal-00258669/fr/.

[20] P. F. Marteau. Time warp edit distance with stiffness adjustment for time series matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):306–318, 2009.

[21] C. A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive functions. *Constructive Approximation*, 2:11–22, 1986.

[22] W. Pearson. Rapid and sensitive sequence comparisons with fasp and fasta. *Methods Enzymol*, 183:63–98, 1990.

[23] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

[24] C. A. Ratanamahatana and E. J. Keogh. Making time-series classification more accurate using learned constraints. In *Proceedings of the Fourth SIAM International Conference on Data Mining (SDM'04)*, pages 11–22, 2004.

[25] H. Saigo, J.P. Vert, N. Ueda, and T. Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20:1682–1689, 2004.

[26] H. Sakoe and S. Chiba. A dynamic programming approach to continuous speech recognition. In *Proceedings of the 7th International Congress of Acoustic*, pages 65–68, 1971.

[27] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, nov 1938.

[28] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[29] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.

[30] Kilho Shin and Tetsuji Kuboyama. A generalization of haussler's convolution kernel: mapping kernel. In *ICML*, pages 944–951, 2008.

[31] K.R. Sivaramakrishnan and C. Bhattacharyya. Time series classification for online tamil handwritten character recognition a kernel based approach. In Nikhil R. Pal, Nikola Kasabov, Rajani K. Mudi, Srimanta Pal, and Swapan K. Parui, editors, *Neural Information Processing*, volume 3316 of *Lecture Notes in Computer Science*, pages 800–805. Springer Berlin / Heidelberg, 2004.

[32] T. Smith and Waterman M. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[33] Vladimir Vapnik. *Statistical Learning Theory*. Wiley-Interscience, ISBN 0-471-03003-1, 1989.

[34] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[35] V. M. Velichko and N. G. Zagoruyko. Automatic recognition of 200 words. *International Journal of Man-Machine Studies*, 2:223–234, 1970.

[36] J. P. Vert, H. Saigo, and T. Akutsu. Local alignment kernels for biological sequences. In B. Scholkopf, K. Tsuda, and J.P. Vert, editors, *Kernel Methods in Computational Biology*, pages 131–154. MIT Press, 2004.

[37] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21:168–173, 1973.

[38] Adam Woznica, Alexandros Kalousis, and Melanie Hilario. Distances and (indefinite) kernels for sets of objects. *Data Mining, IEEE International Conference on*, 0:1151–1156, 2006.

[39] Gang Wu, Edward Y. Chang, and Zhihua Zhang. Learning with non-metric proximity matrices. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 411–414, New York, NY, USA, 2005. ACM.

[40] Dongyu Zhang, Wangmeng Zuo, David Zhang, and Hongzhi Zhang. Time series classification using support vector machine with gaussian elastic metric kernel. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, ICPR '10, pages 29–32, Washington, DC, USA, 2010. IEEE Computer Society.