ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

The one-sided mean kernel: a positive definite kernel for time series

Nicolas Chrysanthos[1,2], Pierre Beauseroy[1], Hichem Snoussi[1], Edith Grall[1], Fabrice Ferrand[2] *

1- Institut Charles Delaunay (UMR CNRS 6279 STMR), LM2S
Université de Technologie de Troyes, 10010 Troyes, France

2- Sagem Défense Sécurité
18-20 Quai du Point du Jour
92100 Boulogne-Bilancourt, France

Abstract. We propose in this paper a new kernel for time series on structured data in the dynamic time warping family. We demonstrate using the theory of infinitely divisible kernels that this kernel is positive definite, that it is a radial basis kernel and that it reduces to a product kernel when comparing two sequences of the same length. Finally we compare this kernel with the global alignment kernel in a classification task using support vector machines.

1 Introduction

Kernel methods have proven extremely useful for dealing with a wide range of problems, in particular they have been used to extract non linear features using extensions of linear methods and to treat structured data as vectors in a Hilbert space. In this publication we deal with the case of data organized as sequences, whose elements lie in an arbitrary space $\mathcal{X}$, which can be a space of structured data. We only assume $\mathcal{X}$ is endowed with a positive definite (abbreviated p.d.) kernel k. We denote by $\mathcal{X}^*$ the space of finite sequences with elements in $\mathcal{X}$, such that $\mathcal{X}^* = \cup_{i=1}^{\infty} \mathcal{X}^i$. The goal is now to design a kernel k^* on $\mathcal{X}^*$ with suitable properties.

Let $\boldsymbol{x} = (x_1, \ldots, x_l)$ and $\boldsymbol{x}' = (x'_1, \ldots, x'_m)$ two elements of $\mathcal{X}^*$. In the general case, these two elements may not have the same length, and thus one cannot use traditional vector-based approaches such as a Gaussian kernel in an Euclidian space to compare these sequences. One solution is to define alignments between sequences. An alignment associates elements from one sequence to elements in another sequence such that the order of elements is preserved. Alignments can either introduce gaps or repetitions.

Attempts to deal with this problem resulted in the well-known dynamic time warping kernel [1], which seeks the best alignment between two sequences. Although this kernel has been extensively used by practitioners it has been demonstrated recently [2] that it is in fact not p.d. Since then some researchers have proposed alternatives, such as for example the global alignment (abbreviated g.a.) kernel [3]. In this work we shall consider a particular kind of alignments

*This work was funded by Sagem Défense Sécurité and the ANRT, the french National Association for Research and Technology.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

with repetitions, these in which *only the shortest sequence can have repeated elements*, hence the name "one-sided". In this case we will not use the classical formalism of alignments, but rather refer to what we call "dilatation operators". These will be precisely defined in Section 3.2, but we can already define them informally: a dilatation operator is a function that maps a finite sequence to a longer finite sequence by repeating one or more of its elements while keeping the order. We denote by $\xi'_{l\to m}$ the set of dilatation operators that map sequences of length l to sequences of length m. One can easily see that the cardinal of this set can be defined with binomial coefficients: $|\xi'_{l\to m}| = \binom{m-1}{l-1}$.

2 The one-sided mean alignment kernel

2.1 Practical case: real values with Gaussian kernel

We start by giving an example of the one-sided mean kernel in the case where elements of sequences are real values: $\mathcal{X} = \mathbb{R}$. This is useful to get a sense of how this kernel is represented in most practical cases, before we delve into the more abstract setting of infinitely divisible kernels. Let $\boldsymbol{x}$ and $\boldsymbol{x}'$ two elements of $\mathcal{X}^*$. We shall refer to the shortest and longest elements of $(\boldsymbol{x}, \boldsymbol{x}')$ as $\boldsymbol{x}_{\mathrm{l}}$ and $\boldsymbol{x}_{\mathrm{m}}$ respectively, with $l \leq m$ denoting the respective lengths of the sequences. Note that $\boldsymbol{x}_{\mathrm{l}}$ and $\boldsymbol{x}_{\mathrm{m}}$ are both elements of $\mathcal{X}^*$ whereas x_l and x_m refer respectively to the l^{th} and m^{th} components of $\boldsymbol{x}$. In the real case the one-sided kernel k^* is defined as:

$$k^*(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{1}{|\xi'_{l\to m}|}\sum_{\epsilon\in\xi'_{l\to m}}\frac{1}{m}\|\epsilon(\boldsymbol{x}_{\mathrm{l}}) - \boldsymbol{x}_{\mathrm{m}}\|^2\right) \tag{1}$$

First note that for two sequences of the same length m the kernel evaluation reduces to the classic Gaussian kernel: $k^*(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\frac{1}{m}\|\boldsymbol{x} - \boldsymbol{x}'\|^2)$. Note also that as the shorter sequence plays a special role, this equation is not symmetric w.r.t. $\boldsymbol{x}_{\mathrm{l}}$ and $\boldsymbol{x}_{\mathrm{m}}$; however it is indeed symmetric w.r.t. $\boldsymbol{x}$ and $\boldsymbol{x}'$. This example illustrates interesting properties of this kernel: as it is defined using means of distances, comparing sequences of diverse lengths will always yield values that are in the same range; and this has important consequences in terms of consistency when studying sub-sampling of continuous time series.

2.2 Abstract case: infinitely divisible kernels

Definition 1 *Let K be a p.d. kernel on $\mathcal{X}\times\mathcal{X}$. The kernel K is called* infinitely divisible *if for each positive integer n there exists a p.d. kernel K_n such that $K = K_n^n$.*

Next, for any kernel k on $\mathcal{X}\times\mathcal{X}$, and any integer $m \geq 1$ we denote by $k_m : \mathcal{X}^m \times \mathcal{X}^m \to \mathbb{R}$ the product kernel defined as $k_m(\boldsymbol{x}, \boldsymbol{x}') = k(x_1, x'_1) \cdot \ldots \cdot k(x_m, x'_m)$. We can now state a more general definition of the one-sided mean kernel.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Definition 2 *Let k be a kernel on $\mathcal{X} \times \mathcal{X}$. The one-sided mean alignment kernel k^* is a kernel on $\mathcal{X}^* \times \mathcal{X}^*$ defined as the* geometric mean *of all one-sided alignment scores:*

$$k^*(\boldsymbol{x}, \boldsymbol{x}') = \left[\prod_{\epsilon \in \xi'_{l \to m}} k_m(\epsilon(\boldsymbol{x}_\mathrm{l}), \boldsymbol{x}_\mathrm{m})^{\frac{1}{m}} \right]^{\frac{1}{|\xi'_{l \to m}|}} \tag{2}$$

In the literature k is sometimes called the *base kernel* with respect to k^*. As our main contribution we state the following theorem:

Theorem 1 *The one-sided mean kernel k^* verifies the following properties:*

1. *If k is p.d. and infinitely divisible, then k^* is p.d.,*
2. *When comparing two sequences $\boldsymbol{x}$ and $\boldsymbol{x}'$* of the same length *m, k^* reduces to the product kernel: $k^*(\boldsymbol{x}, \boldsymbol{x}') = k_m(\boldsymbol{x}, \boldsymbol{x}')^{\frac{1}{m}}$.*
3. *If k is a radial basis kernel [1], then k^* is a radial basis kernel.*

Of course, the Gaussian kernel $k(x, y) = \exp(-(x-y)^2)$ is itself infinitely divisible [4], and it suffices to express the product as the exponentiation of distances to be lead to Equation 1.

3 Demonstration of the main theorem

3.1 Principle

In order to prove that the one-sided mean kernel is p.d. we will prove that its evaluation over any finite dataset $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N)$ of any size N yields a p.d. matrix. We denote by n the length of longest sequence in the dataset. Informally, using the theory of infinitely divisible kernels we will "divide" the values of kernel evaluations $k^*(\boldsymbol{x}^i, \boldsymbol{x}^j)$ into sufficiently small parts that will be rearranged to expose the fact that the Gram matrix can be expressed as a Schür product of many p.d. matrices. Indeed, one can see that the kernel is already defined as a product of other kernels, however this product is indexed by a set $\xi'_{l \to m}$ which depends on the particular pair of samples $(\boldsymbol{x}_\mathrm{l}, \boldsymbol{x}_\mathrm{m})$ being considered. Thus our task shall be to rewrite this product such that it is indexed by $\xi_{1 \to n}$, a set independent of the pair of samples considered.

3.2 Formal definition of dilatation operators

We shall define the set of dilatation operators in a recursive manner. First for any positive integer l indicating the length of a sequence, we denote by ϵ_i^l the operator that dilates a sequence of length l by repeating once its i^{th} element:

$$\epsilon_i^l : \begin{array}{rl} \mathcal{X}^l & \to \mathcal{X}^{l+1} \\ a_1 a_2 \ldots a_l & \mapsto a_1 a_2 \ldots a_i a_i \ldots a_l \end{array}$$

[1] In the sense of Haussler [4] a (generalized) radial basis kernel is a kernel with values in $[0, 1]$ and equal to 1 only on the diagonal $\{(x, x),\, x \in \mathcal{X}\}$.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

For the sake of our demonstration we will have to extend slightly this definition by enlarging the support of ϵ_i^l to all of $\mathcal{X}^*$:

$$\epsilon_i^{*l}(\boldsymbol{x}) = \begin{cases} \epsilon_i^l(\boldsymbol{x}) & \text{if } |\,\boldsymbol{x}\,| = l \\ \boldsymbol{x} & \text{if } |\,\boldsymbol{x}\,| \neq l \end{cases} \tag{3}$$

For the sake of clarity we shall from now on omit the star exponent and denote by ϵ_i^l the extended operator. Next we denote by $\xi_{l\to l+1}$ the set of all dilatation operators that map a sequence of length l to a sequence of length $l+1$. Thus $\xi_{l\to l+1} = \{\epsilon_i^l,\, i \in [\![1, l]\!]\}$.

Let $l < m$ two integers. In order to define the set of dilatation operators that map a sequence of length l to a sequence of length m, we state that one such operator first dilates a sequence of length l to a sequence of length $l+1$, then to a sequence $l+2$, etc. until a sequence of length m is reached. Recursively it can be defined as:

$$\xi_{l\to m} = \{\epsilon' \circ \epsilon, \quad \epsilon \in \xi_{l\to m-1} \wedge \epsilon' \in \xi_{m-1\to m}\} \tag{4}$$

Finally for consistency, we have $\xi_{l\to l} = \{\mathrm{Id}_{\mathcal{X}^*}\}$.

Combinatorial considerations For the sake of the demonstration we consider for example $\epsilon_1^3 \circ \epsilon_2^2$ and $\epsilon_3^3 \circ \epsilon_1^2$ to be two *different* elements of $\xi_{2\to 4}$ although they are identical in the mathematical sense since they both represent the same input-output relation. Thus the cardinal of $\xi_{l\to m}$ is $|\xi_{l\to m}| = (m-1)(m-2)\dots l = \frac{(m-1)!}{(l-1)!}$. We denote by $\xi'_{l\to m}$ the set of dilatation operators *without repetition* such that $|\xi'_{l\to m}| = \binom{m-1}{l-1}$ and such that for each element in $\xi'_{l\to m}$ there are exactly $(m-l)!$ identical elements in $\xi_{l\to m}$.

3.3 Developments

We first start by replacing $\boldsymbol{x}_\mathrm{m}$ by $\epsilon(\boldsymbol{x}_\mathrm{m})$ which does not change the values by virtue of Equation 3; and then by replacing $\xi'_{l\to m}$ by $\xi_{l\to m}$ which does not change the value of the geometric mean as discussed in Section 3.2, so that we obtain $k^*(\boldsymbol{x}, \boldsymbol{x}') = \prod_{\epsilon\in\xi_{l\to m}} k_m(\epsilon(\boldsymbol{x}_\mathrm{l}), \epsilon(\boldsymbol{x}_\mathrm{m}))^{\frac{1}{m\cdot|\xi_{l\to m}|}}$.

Next, we change the index set of the product from $\xi_{l\to m}$ to $\xi_{1\to m}$. As both elements $\boldsymbol{x}_\mathrm{l}$ and $\boldsymbol{x}_\mathrm{m}$ have length strictly superior to $l-1$ this results according to Equation 3 to elements of $\xi_{l\to m}$ being repeated exactly $(l-1)!$ times, which we account for by changing the exponent and which leads to $k^*(\boldsymbol{x}, \boldsymbol{x}') = \prod_{\epsilon\in\xi_{1\to m}} k_m(\epsilon(\boldsymbol{x}_\mathrm{l}), \epsilon(\boldsymbol{x}_\mathrm{m}))^{\frac{1}{m\cdot(l-1)!\cdot|\xi_{l\to m}|}}$. Finally, as $|\xi_{l\to m}| = \frac{(m-1)!}{(l-1)!}$, we have:

$$k^*(\boldsymbol{x}, \boldsymbol{x}') = \prod_{\epsilon\in\xi_{1\to m}} k_m(\epsilon(\boldsymbol{x}_\mathrm{l}), \epsilon(\boldsymbol{x}_\mathrm{m}))^{\frac{1}{m!}} \tag{5}$$

The next step is to prove a certain identity by induction. Let $p \geq m > l$. Using Equation 4 we can decompose any element of $\xi_{1\to p+1}$ such that:

$$\prod_{\epsilon\in\xi_{1\to p+1}} k_{p+1}(\epsilon(\boldsymbol{x}_\mathrm{l}), \epsilon(\boldsymbol{x}_\mathrm{m})) = \prod_{\epsilon'\in\xi_{1\to p}} \prod_{\epsilon''\in\xi_{p\to p+1}} k_{p+1}(\epsilon''(\epsilon'(\boldsymbol{x}_\mathrm{l})), \epsilon''(\epsilon'(\boldsymbol{x}_\mathrm{m}))) \tag{6}$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Then by breaking down the definition of k_{p+1} and rearranging the terms in the product one can easily see that for any $\boldsymbol{x}_p$, $\boldsymbol{x}'_p$ in $\mathcal{X}^p$:

$$\prod_{\epsilon'' \in \xi_{p \to p+1}} k_{p+1}(\epsilon''(\boldsymbol{x}_p), \epsilon''(\boldsymbol{x}'_p)) = k_p(\boldsymbol{x}_p, \boldsymbol{x}'_p)^{p+1} \tag{7}$$

By applying Equation 7 to $\boldsymbol{x}_p = \epsilon'(\boldsymbol{x}_\mathrm{l})$ and $\boldsymbol{x}'_p = \epsilon'(\boldsymbol{x}_\mathrm{m})$, combining with Equation 6 and elevating to the power $\frac{1}{(p+1)!}$ we obtain:

$$\prod_{\epsilon \in \xi_{1 \to p+1}} k_{p+1}(\epsilon(\boldsymbol{x}_\mathrm{l}), \epsilon(\boldsymbol{x}_\mathrm{m}))^{\frac{1}{(p+1)!}} = \prod_{\epsilon \in \xi_{1 \to p}} k_p(\epsilon(\boldsymbol{x}_\mathrm{l}), \epsilon(\boldsymbol{x}_\mathrm{m}))^{\frac{1}{p!}}$$

By induction from $p = m$ until $p = n$, and by using Equation 5 we obtain that $k^*(\boldsymbol{x}, \boldsymbol{x}') = \prod_{\epsilon \in \xi_{1 \to n}} k_n(\epsilon(\boldsymbol{x}_\mathrm{l}), \epsilon(\boldsymbol{x}_\mathrm{m}))^{\frac{1}{n!}}$. As k_n is symmetric we are finally lead to another expression for the one-sided mean kernel:

$$k^*(\boldsymbol{x}, \boldsymbol{x}') = \prod_{\epsilon \in \xi_{1 \to n}} k_n(\epsilon(\boldsymbol{x}), \epsilon(\boldsymbol{x}'))^{\frac{1}{n!}} \tag{8}$$

Recall that n is the length of the longest sequence in the dataset, thus Equation 8 is valid for any pair of samples $(\boldsymbol{x}, \boldsymbol{x}')$ in the dataset.

3.4 Conclusion of the demonstration

For any $\epsilon \in \xi_{1 \to n}$, denote by $K_{\epsilon,N}$ the $N \times N$ Gram matrix obtained by evaluation of the kernel $k_n^{\frac{1}{n!}}$ over the samples $\epsilon(\boldsymbol{x}^1), \ldots, \epsilon(\boldsymbol{x}^N)$. As demonstrated in [4], a product of p.d. infinitely divisible kernels is p.d. and infinitely divisible; thus k_n is infinitely divisible, which proves that $k_n^{\frac{1}{n!}}$ is a p.d. kernel. Thus for any $\epsilon \in \xi_{1 \to n}$, $K_{\epsilon,N}$ is a p.d. matrix.

Now let us define $K_N = \bigotimes_{\epsilon \in \xi_{1 \to n}} K_{\epsilon,N}$ the Schür product (entrywise product) of the $(n-1)!$ aforementioned matrices. According to Equation 8, we have that $K_N = \left(k^*(\boldsymbol{x}^i, \boldsymbol{x}^j)\right)_{i,j}$. The Schür product of p.d. matrices is a p.d. matrix [4], thus K_N is p.d., which concludes the demonstration of the first property.

The second and third properties are easily deduced from the fact that between two sequences of the same length there exists only one one-sided alignment.

4 Experiments

In this section we carry experiments for comparing the one-sided mean kernel with the g.a. kernel on the Japanese vowels dataset [5]. The database contains utterances by nine male speakers of two Japanese vowels 'a' and 'e' successively. Each utterance is described as a time series of length varying from 7 to 29 observations, each observation consists in 12 LPC cepstrum coefficients. The task is to guess which of the nine speakers pronounces a new utterance of 'a' or 'e'. We use the data divided in 270 samples for training and 370 for testing. We use a one-against-one classification scheme as implemented in the Scikit-learn

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

One-sided mean kernel			
$d_{med} \backslash C$	0.1	1	10
0.5	0.946	0.978	0.986
1.0	0.959	0.978	0.981
2.0	0.962	0.981	0.978

Global alignment kernel			
$d_{med} \backslash C$	0.1	1	10
0.5	0.978	0.957	0.954
1.0	0.978	0.970	0.970
2.0	0.983	0.981	0.981

Fig. 1: Correct detection ratios

library. As the base kernel we use a Gaussian kernel that we normalize with a distance d in the set $\{0.5 \cdot d_{med}, d_{med}, 2.0 \cdot d_{med}\}$ where d_{med} is defined as $d_{med} = \text{median}_{i,j}\{\text{mean}_{t,t'}\{\|x_t^i - x_{t'}^j\|^2\}\}$ with i and j being indexes for samples and t and t' indexes for time. Concerning the g.a. kernel, as advised in [3] we take the logarithm of the values. We also experiment with different values for the regularization parameter C for the training of the SVM. As we can see in Figure 1 the one-sided performs favorably compared to the g.a. kernel.

5 Conclusion

The kernel we propose is in the same family as the g.a. kernel; it can just as well handle time series whose elements are structured data because it is defined from a base kernel k. With only mild requirements on k we have demonstrated that k^* is both p.d. and a radial basis kernel. Because it is defined using means instead of sums the one-sided mean kernel does not suffer from diagonal dominance issues, and thus can readily be used in practice; whereas it is common to take the logarithm of the values of the g.a. kernel which breaks its p.d. property. In addition the one-sided mean kernel seems to compare favorably in terms of performance on real-world datasets. Of course there are many applications where the fact that only the shorter sequence can have repeated states is an issue, for example one would not want to use the one-sided kernel for applications such as protein sequence analysis. However, when dealing with for example the sampling of continuous processes, one can obtain meaningful results.

References

[1] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978.

[2] J-P Vert. The optimal assignment kernel is not positive definite. *arXiv preprint arXiv:0801.4061*, 2008.

[3] M. Cuturi, J-P Vert, O. Birkenes, and T. Matsui. A kernel for time series based on global alignments. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 2, pages II–413. IEEE, 2007.

[4] D. Haussler. Convolution kernels on discrete structures. Technical report, Department of Computer Science, University of California at Santa Cruz, 1999.

[5] Mineichi Kudo, Jun Toyama, and Masaru Shimbo. Multidimensional curve classification using passing-through regions. *Pattern Recognition Letters*, 20(11):1103–1111, 1999.