KERNEL FUNCTIONS FOR SIGNALS AND SYMBOLIC SEQUENCES OF DIFFERENT LENGTH

V.V. Sulimova[1], V.V. Mottl[2], I.B. Muchnik[3]

[1] Tula State University
Tula, 300600, Lenin Ave., 92, sulimova@tula.net

[2] Computing Center of the Russian Academy of Sciences
Moscow, 119333, Vavilov St., 40, vmottl@yandex.ru

[3] Rutgers University
P.O. Box 8018, Piscataway, NJ 08855, USA, muchnik@dimacs.rutgers.edu

In data analysis, when quantitatively comparing two signals or symbolic sequences of different length, it is the traditional practice to evaluate the best similarity attainable via their alignment. Such a pair-wise similarity measure does not possess in principle the extremely desired property of inner product and, so, is useless for constructing kernel-based techniques which allow for harnessing the most developed linear methods of data analysis. In this work, we propose a mathematical framework for pair-wise similarity comparison between sequences of different length on the basis of computing a linear combination of conditional similarity values over all the alignments instead of the search for the best one. The conditions are proved under which the resulting similarity measure possesses all the properties if inner product.

Introduction

Sequences of different length $\omega=(\alpha_k, k=1,...,N_\omega)$ are a typical kind of objects in data analysis. The primitives $\alpha_k \in A$ making sequences may be real numbers, vectors or symbols from an alphabet. In the former two cases the sequences are scalar or vector signals, and in the latter case it is generally agreed to name them symbolic sequences.

Person identification via on-line signature processing, spoken command and continuous text recognition, prediction of protein properties, functions and structure from amino acid sequences, all these are well-known examples of data analysis problems concerned with the necessity to process sequences of different length.

A convenient instrument of solving such problems are the methods based on the notion of kernel function [1,2] – a real function of two arguments whose values for any finite collection of objects form a positive semidefinite matrix. A kernel function $K(\omega',\omega'')$ defined on the set arbitrary objects $\omega\in\Omega$ embeds the initial set into a linear space $\tilde{\Omega}\supseteq\Omega$ in which it acts as inner product. When dealing with objects of arbitrary kind, such an embedding allows for harnessing practically any classical data-analysis methods developed for linear spaces and avoid the intermediate stage of choosing a vector of numerical features $\mathbf{x}(\omega)\in R^n$ which would define the inner product in the traditional way $K(\omega',\omega'')=\left(\mathbf{x}(\omega')\right)^T\mathbf{x}(\omega'')$. The search for numerical features is especially problematic for sequences of different length.

Various ways of introducing a kernel function on the set of sequences of different length may be proposed. In this paper, we first outline a

This work is supported by INTAS grants No. 04-77-7347, 06-1000014-6563, and grants of the Russian Foundation for Basic Research No. 05-01-00679, 06-01-00412.

sufficiently general mathematical structure of kernel function, adequate to many practical problems, and then consider a particular case of that structure which leads to a simple algorithm of computing the kernel for any two given sequences.

Kernel on the set of primitives

Let us assume that the set of primitives $\alpha \in A$ has the structure of a linear space with an arbitrary inner product (kernel function) $\mu(\alpha',\alpha'')$, then the value $\left(\mu(\alpha,\alpha)\right)^{1/2}$ is a norm in it.

The interpretation of the set of primitives as an inner product linear space appears quite natural for signals which primordially are sequences of real number or vectors. It can be shown that this interpretation remains valid also for amino acid sequences of proteins, which are symbolic sequences over the alphabet of twenty amino acids existing in the nature, if the similarity between any two amino acids is measured as the probability of their common origin from the same unknown amino acid, as it is generally adopted in molecular biology [3].

The set of alignments of two sequences

If all sequences had the same length $\Omega=\left\{\omega=(\alpha_k,\ k=1,...,N=const)\right\}$ then, for instance, the product $K(\omega',\omega'')=\prod_{k=1}^{N}\mu(\alpha'_k,\alpha''_k)$ would possess all the properties of a kernel on the set Ω, because it is known that the product of any number of kernels is a kernel, too [4]. We deal, however, with a set of sequences of different length $\Omega=\left\{\omega=(\alpha_k,\ k=1,...,N_\omega)\right\}$, and such a method is applicable only after an alignment of the sequences being compared to a common length.

An alignment w of two sequences $\omega'=(\alpha'_k,k=1,...,N')$ and $\omega''=(\alpha''_k,k=1,...,N'')$, $\alpha'_k,\alpha''_k\in A$, is understood as bringing them to the same length via insertion of some "empty" aligning elements at some positions in each of the sequences with the respective renumbering of the elements: $\bar\omega'_w=(\bar\alpha'_{w,j},\ j=1,...,|w|)$ and $\bar\omega''_w=(\bar\alpha''_{w,j},\ j=1,...,|w|)$, where $|w|\ge\max\{N',N''\}$ is the common length of the aligned sequences. As the aligning element, we arbitrarily choose in the linear space of primitives an element $\alpha^0\in A$ of the unity norm $\mu(\alpha^0,\alpha^0)=1$.

The set of all alignments of a pair of sequences $<\omega',\omega''>$ of lengths N' and N'' will be denoted as $\mathcal{W}_{N'N''}$. Any alignment $w\in\mathcal{W}_{N'N''}$ may be represented as a path in the graph with horizontal, diagonal and vertical edges oriented from left to right and from top to bottom as shown in Figure 1. In this graph, the horizontal direction will be associated with the first sequence $\omega'=(\alpha'_k,k=1,...,N')$, and the vertical direction – with the second sequence $\omega''=(\alpha''_k,k=1,...,N'')$.

We shall consider any alignment $w\in\mathcal{W}_{N'N''}$ as a sequence of values from the three-element set: $w=(h_k,k=1,...,|w|)$, $h_k\in\{h,h',h''\}$. The value $h_k=h'$ means a horizontal step associated with insertion of an "empty" aligning element in the first sequence, the value $h_k=h$ is interpreted as a diagonal step which corresponds to the absence of insertions, and $h_k=h''$ signifies a vertical step denotative of insertion of an "empty" element in the second sequence. The symmetric analog of any alignment w resulting from replacement each step $h_k=h'$ by $h_k=h''$ and vise versa will be denoted by symbol w^T, so that $\mathcal{W}_{N''N'}=\{w^T: w\in\mathcal{W}_{N'N''}\}$.

A system of weights on the set of alignments and the structure of the kernel function

Let us associate any alignment w of two sequences ω' and ω'' with the value

$$K(\omega',\omega''|w)=K(\omega'',\omega'|w^T)=\prod_{j=1}^{|w|}\mu(\bar\alpha'_{w,j},\bar\alpha''_{w,j}),\quad(1)$$

meant as the measure of the alignment-dependent conditional similarity of two sequences.

Further, let us choose a system of non-negative weights of pair-wise alignments $p(w)\ge 1$

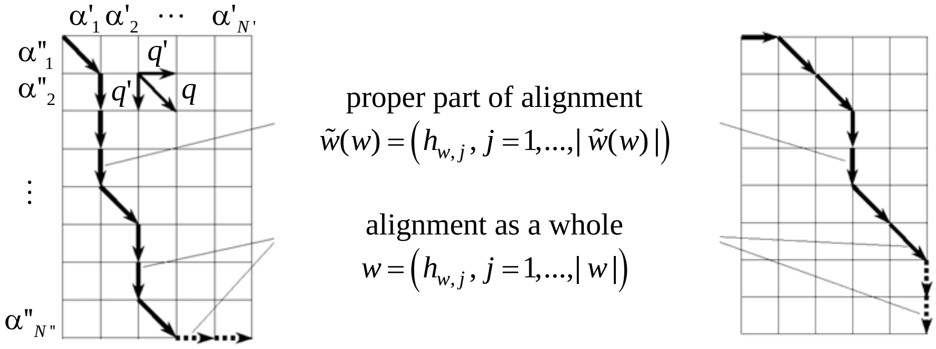
α'₁ α'₂ ⋯ α'_{N'}
α"₁
α"₂
⋮
α"_{N"}
q'
q'
q
proper part of alignment
$\tilde{w}(w)=\left(h_{w,j},\, j=1,\dots,|\,\tilde{w}(w)\,|\right)$
alignment as a whole
$w=\left(h_{w,j},\, j=1,\dots,|\,w\,|\right)$

Figure 1. Two different alignments of a pair of sequences.

which is common for all the pairs of lengths of sequences N' and $N"$, and expresses some a priori preferences on the set of different alignments of the same pair.

The traditional way of measuring the similarity between two sequences is based on the search for the alignment which maximizes their conditional similarity with respect to the weight $K(\omega',\omega")=\max_{w\in\mathcal{W}_{N'N"}} K(\omega',\omega"|w)$ [5], but this similarity measure will not possess the properties of a kernel function. In this work, instead of the maximization operation, we use the linear combination of the conditional similarities between two sequences over all the alignments:

$$K(\omega',\omega")=\sum_{w\in\mathcal{W}_{N'N"}} p(w)K(\omega',\omega"|w)\,. \qquad (2)$$

Let w be an alignment of sequences ω' and $\omega"$ having the lengths N' and $N"$. The beginning part of the alignment w up to the first tangency to the right or bottom boundary of the region $\mathcal{W}_{N'N"}$ (Figure 1) will be called its proper part and denoted by symbol $\tilde{w}(w)$.

We shall call the sequences ω' and $\omega"$ of lengths N' and $N"$ prolonged by alignment elements $\alpha^0\in A$ at the right up to some common length N the augmented sequences $\bar{\omega}'=(\alpha'_{k'},k'=1,\dots,N',\alpha'_{k'}=\alpha^0,k'>N')$ and $\bar{\omega}"=(\alpha"_{k"}, k"=1,\dots,N",\alpha"_{k"}=\alpha^0,k">N")$. All the alignments of the augmented sequences form the set $\mathcal{W}_{NN}$. Two alignments $w\in\mathcal{W}_{N'N"}$ and $\bar{w}\in\mathcal{W}_{NN}$ will be called equivalent and denoted as $w\sim\bar{w}$ if the proper part of alignment w is the beginning part of alignment $\bar{w}$.

A system of weights $p(w)$ will be said to be self-consistent if, first, the weights of symmetric alignments equal to each other $p(w)=p(w^T)$ and, second, any N', $N"$ and N such that $N\geq N'$ and $N\geq N"$ satisfy the condition $p(w)=\sum_{\bar{w}\in\mathcal{W}_{NN},\,w\sim\bar{w}} p(\bar{w})$, i.e. the weight of any alignment of the original sequences w is equal to the sum of the weights of equivalent alignments of the augmented sequences.

Theorem 1. For the linear combination $K(\omega',\omega")$ (2) of conditional similarity measures $K(\omega',\omega"\,|\,w)$ (**Ошибка! Источник ссылки не найден.**) to be a kernel function on the set of sequences over the linear space of primitives $\Omega=\{\omega=(\alpha_k,k=1,\dots,N_\omega)$, it is sufficient that the aligning element satisfies the condition $\mu(\alpha^0,\alpha^0)=1$ and the system of weights $p(w)$ is self-consistent.

At the same time, for some two-argument function $K(\omega',\omega")$ (2), which formally possesses the properties of kernel, would be also useful from the practical point of view, it is important to properly choose the original kernel $\mu(\alpha',\alpha")$ on the set of primitives $\alpha\in A$, the aligning element $\alpha^0\in A$, and the system of alignment weights $p(w)$.

Radial kernel on the set of primitives and multiplicative alignment weights

Let in the linear space of primitives with the null element $\phi\in A$ a Euclidean metric be defined, for instance, by some initial kernel

$\rho(\alpha',\alpha'') = \left[\kappa(\alpha',\alpha') + \kappa(\alpha'',\alpha'') - 2\kappa(\alpha',\alpha'')\right]$. It is known [1] that, in this case, the two-argument function

$$\mu(\alpha',\alpha'') = \exp\left[-\beta\rho^2(\alpha',\alpha'')\right] \qquad (3)$$

with any value of the parameter $\beta > 0$ is a kernel which embeds the linear space with inner product $\kappa(\alpha',\alpha'')$ into another linear space with inner product $\mu(\alpha',\alpha'')$.

The kernel function (3), which, by its structure, quantitatively expresses a pair-wise similarity between primitives with respect to the original metric $\rho(\alpha',\alpha'')$, is usually called the radial kernel.

The choice of the null element of the original linear space as the aligning element $\alpha^0 = \phi \in A$ meets the condition $\mu(\alpha^0,\alpha^0) = 1$ in Theorem 1.

We associate the non-negative numbers $q(h) = q$ and $q(h') = q(h'') = q'$ with each of three values of the variable h, h' and h''. The value $q > 1/3$ means the preference of the absence of insertions and deletions at each elementary step of comparing the sequences (Figure 1).

Let $w = (h_{w,j}, j = 1,...,|w|)$ be an arbitrary alignment, and $\tilde{w}(w)$ be its proper part. The weight of the alignment will be defined as the product

$$p(w) = \prod_{j=1}^{|\tilde{w}(w)|} q(h_{w,j}) . \qquad (4)$$

Theorem 2. The system of weights (4) is self-consistent.

So, the radial kernel on the set of primitives and the multiplicative system of weights meet all the requirements of Theorem 1 and define a kernel on the set of sequences of different length (2), which explicitly expresses the degree of their pair-wise similarity. The algorithm of computing the value of this kernel has the computational complexity proportional to the product of the lengths of the sequences being compared.

Refernences

1. Aizerman M.A., Braverman E.M., Rozonoer L. I. *Method of Potential Functions in the Theory of. Machine Learning* (in Russian). Nauka, Moscow, 1971.
2. Vapnik V. *Statistical Learning Theory*. New York: John-Wiley & Sons, Inc., 1998, 732 p.
3. Dayhoff M.O., Schwartz R.M., Orcutt B.C. A model for evolutionary change in proteins. *Atlas for Protein Sequence and Structure (M.O. Dayhoff, ed.)*, 1978, Vol. 5, pp. 345-352.
4. Haussler D. Convolution kernels on discrete structures. *Technical Report UCSC-CLR-99-10*, University of California at Santa Cruz, 1999.
5. Dubin R., Eddy S., Krogh A., Mitchison G. *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, 1998, 356 p.