MATHEMATICALLY CORRECT METHODS OF SIMILARITY MEASUREMENT ON SETS OF SIGNALS AND SYMBOL SEQUENCES OF DIFFERENT LENGTH.

V.V. Mottl[1], O.S. Seredin[2], V.V. Sulimova[3]

[1] Scientific Council "Cybernetics", Vavilov St. 40, 117968, Moscow, vmottl@yahoo.com
[2] Tula State University, Lenin Ave. 92, 300600, Tula, seredin@uic.tula.ru
[3] Tula State University, Lenin Ave. 92, 300600 Tula, sulimova@tula.net

Embedding a set of objects of arbitrary kind into a linear space by choosing an appropriate two-argument function possessing properties of inner product (kernel function) is a convenient approach to solving most glowing problems of modern informatics such as that of finding empirical regularities in sets of signals and symbolic sequences of different length. However, constructing kernel functions is no easy problem. In this work, we propose a sufficiently universal probabilistic principle of kernel function construction on sets of signals and symbol sequences of different length, which is based on interpretation of every object as effect of random transformation of another object from the same set.

Introduction

The characteristic feature of the problem of finding empirical regularities [5] in sets of signals and symbol sequences is that the initial presentation of objects does not allow for forming a priori feature space of their sufficiently informative characteristics, which would satisfy the requirements of the compactness hypothesis. Examples of such objects are face shots obtained under different shooting angles and with various face expressions, handwritten symbols and signatures presented by pen motion paths, amino acid sequences forming protein polymeric molecules. The featureless approach to data analysis [1] is most adequate for such problems. It is based on finding some digital similarity measure of object pairs and practically does not need evident indication of their feature vectors.

The idea of featureless machine learning was proposed in works of M.A. Iserman, E.M. Braverman and L.I. Rosonoer under the name of the potential function method [2], and later it was used in the support vector machine method created by V.N. Vapnik [4]. These methods are based on the notion of so-called kernel function, any real-valued function of two arguments $K(\alpha',\alpha'')$ on a set of object pairs $\alpha',\alpha''\in A$ whose values can be interpreted as inner product of elements of some hypothetical linear space representing initial entities of the real world.

However, constructing a kernel function on a set of objects of arbitrary kind, which would reflect the claimed conception of the objects' similarity and dissimilarity, is quite an involved problem.

Random transformation of the set of objects

In this paper, we propose a sufficiently universal probabilistic principle of kernel function construction. The essence of the approach consist in treating every object $\alpha\in A$ as effect of a random transformation of another object of the same set $\vartheta\to\alpha$, $\vartheta\in A$. The only specificity required of the set of objects A is the possibility to treat it as a probabilistic space and, hence, to define probability distributions by their densities with respect to an appropriate

This work is supported by the Russian Foundation of Basic Research (Grant 02-01-00107), Grant of the President of the Russian Federation for young scientists No. MK-3173.2994.09 (O.S. Seredin) and the Federal Scientific Program "Research and Development in Priority Directions of Science and Technology" in 2002-2006 (Oriented Basic Research).

σ-finite measure. In particular, if the set A is finite or countable, the role of the probability density with respect to the counting measure is played by the probabilities of single events.

We will suppose, that the probability distribution density $\zeta(\vartheta)$, defined in the object set A, formally expresses the assumed popularity frequency of each object. A natural way to estimate the similarity of two object $\alpha',\alpha'' \in A$ is to evaluate the likelihood of the hypothesis that they origin from the same unknown randomly chosen object $\vartheta \in A$ as the common prototype:

$$K(\alpha',\alpha'') = \int_{\vartheta\in A} \zeta(\vartheta)\psi(\alpha' \mid \vartheta)\psi(\alpha'' \mid \vartheta)d\vartheta. \quad (1)$$

The family of conditional distribution densities $\psi(\alpha \mid \vartheta)$, that determines the random transformation $\vartheta \to \alpha$, associates each object α with a function $x_\alpha(\vartheta) = \psi(\alpha \mid \vartheta)$ on the entire set A, which can be naturally considered as representation of the object in the linear space of all real functions in A. Then, the two-argument likelihood function (1) is inner product of two corresponding functions with weight $\zeta(\vartheta)$ and, so, is a kernel function on the set of objects.

A natural way to choose the probability distribution $\zeta(\vartheta)$ follows from an additional assumption on the properties of random transformation $\psi(\alpha \mid \vartheta)$. We shall say that this transformation is ergodic, if there exists the final probability density $\zeta(\alpha)$ that meets the integral equation

$$\zeta(\alpha) = \int_{\vartheta\in A} \psi(\alpha \mid \vartheta)\zeta(\vartheta)d\vartheta, \quad (2)$$

and reversible, if the condition

$$\zeta(\vartheta)\psi(\alpha \mid \vartheta) = \zeta(\alpha)\psi(\vartheta \mid \alpha) \quad (3)$$

holds for all $\vartheta,\alpha \in A$. Such terminology is suggested by the fact that condition (2) is equivalent to ergodicity of the Markov random process $(\alpha_s,\ s = 1,\ 2,\ 3,\ ...)$ defined by transition distributions $\psi(\alpha_s \mid \alpha_{s-1})$, and $\zeta(\alpha)$ is its final distribution on A. The condition (3) is equivalent to the reversibility of this Markov process, i.e. coincidence of its probabilistic properties in the forward and reverse direction.

The reversibility of the random transformation $\psi(\alpha \mid \vartheta)$ allows to express the kernel function through the double-step transformation $\alpha' \to \vartheta \to \alpha''$ defined by the family of conditional distribution densities $\psi^{[2]}(\alpha'' \mid \alpha')$:

$K(\alpha',\alpha'')=\zeta(\alpha')\psi^{[2]}(\alpha''|\alpha')=\zeta(\alpha'')\psi^{[2]}(\alpha'|\alpha'')$.

Such definition of the kernel function is equivalent to (1), but it is preferable in some cases.

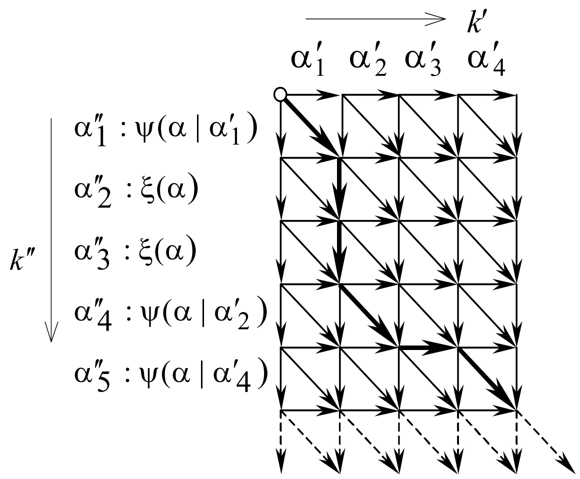
Kernel function on set of sequences of arbitrary finite length

Kernel function on set of sequences of arbitrary finite length $\omega = (\alpha_k,\ k = 1,...,N) \in \Omega$ constituted from elements of the given set $\alpha \in A$, in which probability kernel function has been already defined, can be constructed on basis of the random transformation of sequences $\varphi(\omega'' \mid \omega')$. You should notice, any transformation of some initial sequence $\omega' = (\alpha'_k, k = 1,...,N')$ to other sequence of other length $\omega'' = (\alpha''_k,\ k = 1,...,N'')$ inevitably have to be connected not only with elements change $\alpha'_{k'} \to \alpha''_{k''}$, but with removal of some elements from ω' and addition new elements into ω''. Rules, which regulate removal and addition of elements, inevitably have to be different for symbol sequences and signals by virtue of its nature considerable difference. The difference consists not so much in character of the set A, from which elements $\alpha_k \in A$ are scooped, as in a meaning, which is put in element combination of these elements in the unified regulated structure $\omega = (\alpha_k,\ k = 1,...,N)$.

This paper considers only case of symbol sequences. Taking into account a signal specificity leads to not great changes of the scheme considered here.

A model of the random transformation of the sequence $\omega' = (\alpha'_k,\ k = 1,...,N')$ in other sequence $\omega'' = (\alpha''_k, k = 1,...,N'')$ can be constructed as the system of two probabilistic mechanisms. One of them is the random process $w = (h_t,\ t = 1,2,3,...)$, formed by independent choice of one of three variants of its continuation $h_t \in H = \{h, h', h''\}$ on each step. It defines a transformation structure, i.e. regulates random choice of positions, on which elements are removed and added. The second mechanism defines random law of element choice on each position in sequence changed. In the model considered the first mechanism is general one for symbol sequences and signals, thus the difference consists only in the second mechanism

The structure of transformation of one sequence in other one can be represented as a random path on the graph (figure). Here diagonal advancement corresponds with transformation of the element of the first sequence in the respective element of the second sequence in accordance with the given random transformation on the set of objects A. Horizontal advancement corresponds with gap of the next element of the first sequence, and vertical advancement corresponds with forming of new element of the second sequence, not depending on elements of the first one, in accordance with the final distribution on the set of objects $\xi(\alpha)$.



The structure of sequence transformation as the random path on the graph and probabilistic rules of element's choice of new sequence

In the result of the random process w realization initial sequence of length N' is transformed to new sequence of other random length N''. Probabilistic properties of the random process w completely define the conditional probability distribution on set of new sequence $p(N''|N')$ lengths.

We proved, that the sequence transformation $\varphi(\omega''|\omega')$ is ergodic and reversible one, if the random process of sequences lengths transformation $p(N''|N')$ is ergodic and reversible process.

Specifically, this characteristic exists, if the random process w formed by independent choice of one of three variant of its continuation h, h' or h'' on each step, moreover the probability of symbol gap in initial sequence, i.e. choice h', is less than the probability of new symbol insertion in transformed sequence, i.e. choice h''. Then the distribution density

$$\zeta(\omega'') = \int_{\vartheta\in\Omega} \varphi(\omega''|\omega')\zeta(\omega')d\omega'$$

is the final distribution on the set of sequences of different length, and the function of two-arguments $K(\omega',\omega'') = \zeta(\omega')\varphi(\omega''|\omega')$ is the kernel function.

The considered random principle of the kernel function constructing was realized on practice by the example of kernel function constructing on set of signatures, presented by pen motion paths, which form signals of different length.

You should notice, the application of the considered approach is inevitably connected with appearance of some computing problems, and its decision is not at all trivial. Specifically, by virtue of the probabilistic character of the kernel function, which is considered here, its values for different pairs of signatures are greatly small magnitudes, moreover they differ one from another on whole orders. Such values are not only inconvenient for their interpretation, but they also do practically impossible the construction of decision rules of recognition on their basis. For overcoming this obstacle we consider the initial kernel function as function referring to class of so-called radial kernel functions, and then we replace it by the linear kernel function, which generate the given radial one.

The linear kernel function is the simplest kind of kernel functions, its role can be played by the two-argument real function $\mu(\omega',\omega'')$, named the commonness of two elements ω' and ω'' with respect to some center ϕ [3]

$$\mu(\omega',\omega'') = \frac{1}{2}\left[r^2(\omega',\phi) + r^2(\phi,\omega'') - r^2(\omega',\omega'')\right]$$

where $r(\omega',\omega'')$ is the function of dissimilarity of two objects, possessing the metrics characteristics.

We name the kernel function of kind $K_{lin}(\omega',\omega'') = \mu(\omega',\omega'')$ the linear one, as its using in the identification problem results in the family of linear decision rules. Along with the linear kernel function, other kind of kernel functions are also traditionally considered, such as the parametric families of polynomial

$$K_{pol}(\omega',\omega'') = \left[\mu(\omega',\omega'') + 1\right]^{\alpha}$$

and radial kernel functions

$$K_{rad}(\omega',\omega'') = \exp\left[-\alpha r^2(\omega',\omega'')\right].$$

We assume that the constructed kernel function is the radial one. Then we find the unknown linear kernel function, which was formed from it by the respective transformation

$$K_{lin}(\omega',\omega'') = -\frac{1}{2\alpha}\left[\ln K_{rad}(\omega',\phi) + \right. \\ \left. + \ln K_{rad}(\phi,\omega'') - \ln K_{rad}(\omega',\omega'')\right]. \quad (4)$$

It should be noticed that the expression (4) contains a combination of logarithms of the initial kernel function values, and, as the consequence, values of the new kernel function become comparable by magnitude and convenient for their using when solving the problem of finding empirical regularities in sets of signals and symbol sequences of different length.

References

1. Duin R.P.W, De Ridder D., Tax D.M.J. Experiments with a featureless approach to pattern recognition. Pattern Recognition Letters, 1997, Vol. 18, No. 11-13, pp. 1159-1166.
2. Iserman M.A., Braverman I.M., Rosonauer L.I. The method of potential functions in the machine learning theory. M.: Nauka, 1970, 384 c. (In Russian).
3. Mottl V.V. Metric spaces admitting linear operations and inner product. //Doklady Mathematics, Vol. 67, № 1, 2003, pp. 140-143.
4. Vapnik V. Statistical Learning Theory. John-Wiley & Sons, Inc., 1998.
5. Vapnik V. The dependences reconstruction on empirical data. M.: Nauka, 1979. (In Russian)
6. Watkins C. Dynamic alignment kernels. In: A.J. Smola, P.L. Bartlett, B. Schlkopf, and D. Schuurmans, Ed. Advances in Large Margin Classifiers. MIT Press, 2000, pp. 39-50.