FUSION OF EUCLEDIAN METRICS IN FEATURELESS DATA ANALYSIS: AN EQUIVALENT OF THE CLASSICAL PROBLEM OF FEATURE SELECTION

V.V. Mottl[1], O.S. Seredin[2], O.V. Krasotkina[2] , I.B. Muchnik[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] Rutgers University, USA, muchnik@dimacs.rutgers.edu

The problem of embedding the given set of objects into a linear space with inner product by choosing an appropriate kernel function or set of features is considered as the major challenge in both kernel-based and feature-based approach to estimating dependences in data sets of arbitrary kind. The main idea is to treat several kernels or, in particular, several numerical features on the same set of objects as Cartesian product of the respective number of linear spaces, each supplied with a specific kernel function as a specific inner product.

Introduction

The problem of finding empirical dependences $y(\omega):\Omega\to Y$ in a set of entities $\omega\in\Omega$ is one of the glowing problems of modern informatics. Let a given data set be the set of experimentally measured values of a characteristic $y_j=y(\omega_j)$ within an accessible subset of objects $\Omega^*=\{\omega_1,...,\omega_N\}$. It is required to continue this function onto the entire set Ω for it would be possible to estimate this characteristic $\hat{y}(\omega)$ for objects $\omega\in\Omega\setminus\Omega^*$ not represented in the original data set [1,2]. In particular, if $y(\omega)$ takes values from a finite set, the problem is usually called the pattern recognition problem, and in the case of a real-valued characteristic it is referred to as the problem of regression estimation.

Machine learning techniques based on the potential function method [1], in particular, the support vector method resulting from it [2], essentially exploit the assumption that the universe of entities can be represented as a linear space. As instruments of introducing linear operations in the set of entities $\omega\in\Omega$ serve either a vector of observable features $\mathbf{x}(\omega)\in R^n$ or a two-argument function $K(\omega',\omega'')$ called the potential function (kernel function in the English-language literature). The potential function method as particular case of Yu. I. Zhuravlev's method of estimate computing in precedent-based machine learning [3]. The specificity of a potential function consists in that its values can be immediately interpreted as inner product in a hypothetic linear space without the intervening notion of a feature vector. The continuation of the function rests on the Euclidean metric in the set of entities

$$\rho(\omega',\omega'')=\left[K(\omega',\omega')+K(\omega'',\omega'')-2K(\omega',\omega'')\right]^{1/2},\quad(1)$$

Such metric expresses a specific compactness hypothesis, namely, the assumption that, for any two close entities, the values of the goal characteristic will be also close in most cases.

There is usually much freedom in measuring dissimilarity of entities, and, thus, several heuristic potential functions can be heuristically suggested within the bounds of the same data analysis problem. However, the choice of features, each of which defines, actually, a simplest potential function, is also ever heuristic. The aim of this work is to study the ways of fusing the given set of potential functions and to organize, thereby, a concurrence of several compactness hypotheses in finding empirical regularities in the given set of entities. The main requirement here is to avoid discrete selection of potential functions or features with the purpose of achieving acceptable computational complexity of the fusion algorithm.

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).

The linear space produced by a potential function

A potential function $K(\omega',\omega'')$ on a set of entities of arbitrary kind $\omega \in \Omega$ can be defined as a real-valued function $\Omega \times \Omega \to \mathbb{R}$ possessing two principal properties – symmetry $K(\omega',\omega'') = K(\omega'',\omega')$ and nonnegative definiteness of the matrix $\left[K(\omega_i,\omega_j); i,j=1,...,m\right]$ for any finite collection of entities $\{\omega_1,...,\omega_m\} \subset \Omega$. The function (1) produced by a potential function is a metric [4], and, so, the set of entities Ω supplied with a potential function becomes a metric space.

The mechanism by which any potential function $K(\omega',\omega'')$ embeds the given set Ω into a real linear space with inner product $\Omega \subseteq \tilde{\Omega}$ is specified in [4]. The null element $\phi \in \Omega$ and linear operations $\omega' + \omega'': \tilde{\Omega} \times \tilde{\Omega} \to \tilde{\Omega}$ and $c\omega: R \times \tilde{\Omega} \to \tilde{\Omega}$ are defined in $\tilde{\Omega}$ in a special way, whereas the role of inner product is played by the potential function itself. The dimensionality of the linear space $\tilde{\Omega}$ is the maximum number of elements $\{\omega_1,...,\omega_m\} \subset \tilde{\Omega}$ for which the matrix $\left[K(\omega_i,\omega_j);\ i,j=1,...,m\right]$ can be positive definite.

The class of linear decision rules in the linear space produced by a potential function

The convenience of a potential function as a means of measuring dissimilarity of any two objects by the respective Euclidean metric (1) consists in that it involves the notion of a linear function $y(\omega): \Omega \to \mathbb{R}$ in the set of entities of any kind. In this Section, we consider a class of decision rules in the extended linear space $\tilde{\Omega}$ supplied with linear operations and inner product produced by a continuation of the given potential function. The class of linear function in $\tilde{\Omega}$ is defined by two parameters $\vartheta \in \tilde{\Omega}$ and $b \in \mathbb{R}$

$$y(\omega \mid \vartheta, b) = K(\vartheta,\omega) + b, \omega \in \Omega. \qquad (2)$$

We shall call parameter ϑ the direction element of the linear function. If the real value of the linear function is immediately treated as the goal characteristic of an entity, the choice of parameters $\vartheta \in \tilde{\Omega}$ and $b \in \mathbb{R}$ determines a regression dependence. If the sign of the linear function is understood as the goal characteristic, the parameters specify a classification of the set of entities into two classes: $y(\omega) = K(\vartheta,\omega) + b > 0 \to$ class 1, $y(\omega) \le 0 \to$ class 2.

It should me marked that such a way of specifying a linear function on the set of all entities Ω is constructive only if the role of the direction element is played by an accessible entity participating in the training set $\vartheta \in \Omega^* = \{\omega_1,...,\omega_N\} \subset \Omega \subset \tilde{\Omega}$ or, at least, an entity that exists in the real world $\vartheta \in \Omega \subset \tilde{\Omega}$. However, in this case, the class of functions $\{y(\omega \mid \vartheta, b): \vartheta \in \Omega, b \in \mathbb{R}\}$ will be, generally speaking, incorrect in the sense that the functions from this class will, most likely, not reproduce the values of the goal function $y(\omega_j \mid \vartheta, b) \ne y_j$ specified in the training set $\{(y_j,\omega_j); j=1,...,N\}$. Embedding the set of real-world entities into a linear space $\tilde{\Omega} \supset \Omega$ allows for speaking about "imaginary" linear combinations of entities $\vartheta = \sum_{i=1}^{m} a_i \omega_i$ in the sense of linear operations induced in the extended set $\tilde{\Omega}$ by the potential function $K(\omega',\omega'')$, i.e. about linear combinations of original incorrect functions $y(\omega \mid \omega_i, b)$. Such an embedding is a particular realization of Yu.I. Zhuravlev's general idea of constructing correct algebras over sets of incorrect functions (algorithms) [5], which, in this case, exploits special properties of potential functions.

As we shall see below, the only reasonable choice of ϑ is, as a rule, a linear combination of objects represented in the training set $\hat{\vartheta} = \sum_{j=1}^{N} a_j \omega_j$.

Cartesian product of linear spaces produced by several potential functions

Let $K_i(\omega',\omega'')$, $i=1,...,n$, be the kernel functions defined on the same set of objects $\omega \in \Omega$ by different experts. These kernel functions embed the set Ω into different linear spaces $\Omega \subset \tilde{\Omega}_i$, $i=1,...,n$, with different inner products and, respectively, different linear operations. It is convenient to treat the n linear spaces jointly as Cartesian product

$$\tilde{\tilde{\Omega}} = \Omega_1 \times ... \times \tilde{\Omega}_n = \left\{\bar{\omega} = <\omega_1,...,\omega_n> : \omega_i \in \tilde{\Omega}_i\right\} \qquad (3)$$

formed by ordered n-tuples of elements from $\tilde{\Omega}_1,...,\tilde{\Omega}_n$. The kernel function (i.e. inner prod-

uct) in this linear space can be defined as the sum of the kernel functions (inner products) of the corresponding components in any two n-tuples $\bar{\omega}'=<\omega'_1,...,\omega'_n>$ and $\bar{\omega}''=<\omega''_1,...,\omega''_n>$:

$$K(\bar{\omega}',\bar{\omega}'') = \sum_{i=1}^{n} K_i(\omega'_i,\omega''_i),\ \bar{\omega}',\bar{\omega}'' \in \tilde{\tilde{\Omega}}. \quad (4)$$

The dimensionality of the combined linear space $\tilde{\tilde{\Omega}}$ (3) will not exceed the sum of dimensionalities of the particular linear spaces.

A really existing object $\omega \in \Omega$ will be represented by its n-fold repetition $\bar{\omega} =< \omega,...,\omega > \in \tilde{\tilde{\Omega}}$. Then any real-valued linear function $\Omega \to \mathbb{R}$ is specified by the choice of parameters $\bar{\vartheta} \in \tilde{\tilde{\Omega}}$ and $b \in \mathbb{R}$

$$y(\omega) = K(\bar{\vartheta},\bar{\omega}) + b = \sum_{i=1}^{n} K_i(\vartheta_i,\omega) + b, \quad (5)$$

where $\bar{\vartheta}$ is a combination of imaginary elements of particular linear spaces $\bar{\vartheta}=<\vartheta_1,...,\vartheta_n>$, $\vartheta_i \in \tilde{\Omega}_i$, produced by particular kernel functions $K_i(\omega',\omega'')$ in $\tilde{\Omega}_i$.

Thus, to define a numerical dependence over a set of objects of any kind by combining several kernel functions $K_i(\omega',\omega'')$, we have, first of all, to choose, as parameters, one element in each of linear spaces $\vartheta_i \in \tilde{\Omega}_i$ into which the kernel functions embed the original set $\Omega \subseteq \tilde{\Omega}_i$. It should be marked that the less the norm of the ith parameter in its linear space $\| \vartheta_i \|^2 = K_i(\vartheta_i,\vartheta_i)$, the less the influence of the respective summand on the value of the function (5). If $K(\vartheta_i,\vartheta_i) \to 0$, i.e. $\vartheta_i \cong \phi_i \in \tilde{\Omega}_i$, the ith kernel function will practically not affect the function.

This means that the parametric family of numerical functions (5) implies also an instrument of emphasizing "adequate" kernel functions with respect to the available observations and suppressing "inadequate" ones. Which kernel functions should be considered as adequate is the key question for providing a good generalization performance of the decision rule when it is applied to objects not represented in the training set.

Fusion of potential functions

If the total dimensionality of the combined extended linear space $\tilde{\tilde{\Omega}}$ (3) is greater than the number of objects in the training set $\{(\omega_j, y_j);\ y_j \in \mathbb{R}, j=1,...,N\}$ or $\{(\omega_j, g_j);\ g_j \in \{-1, 1\},\ j=1,...,N\}$ there always exist linear functions (5) that exactly reproduce the trainer's data. Following the widely adopted principle [2], we shall prefer the function with the minimum norm of the direction element under the constraints of the training set:

$$\begin{cases} \| \bar{\vartheta} \|^2 \to \min,\ \bar{\vartheta} =< \vartheta_1,...,\vartheta_n > \in \tilde{\tilde{\Omega}}, \\ \sum_{i=1}^{n} K_i(\vartheta_i,\omega_j) + b = y_j \\ \quad \text{или } g_j \sum_{i=1}^{n} K_i(\vartheta_i,\omega_j) + b \geq \text{const}. \end{cases} \quad (6)$$

The simplest version of norm follows from (4)

$$\| \bar{\vartheta} \|^2 = \sum_{i=1}^{n} K_i(\vartheta_i,\vartheta_i), \quad (7)$$

but any linear combination of kernel functions with nonnegative coefficients also possesses all the properties of norm $\| \bar{\vartheta} \|^2 = \sum_{i=1}^{n} (1/r_i) K_i(\vartheta_i,\vartheta_i)$. In this case, the criterion (6) will try to avoid kernels with small $_i$. If $r_i = 0$, the respective kernel does not participate in forming the goal function.

The idea of adaptive training consists in jointly inferring the direction elements ϑ_i and the weights $_i$ from the training set by additionally penalizing large weights:

$$\begin{cases} \sum_{i=1}^{n} \left[(1/r_i) K_i(\vartheta_i,\vartheta_i) + \log r_i \right] \to \min(\vartheta_i, r_i), \\ \sum_{i=1}^{n} K_i(\vartheta_i,\omega_j) + b = y_j \\ \text{или } g_j \sum_{i=1}^{n} K_i(\vartheta_i,\omega_j) + b \geq \text{const},\ j=1,...,N. \end{cases} \quad (8)$$

It can be shown that the following iterative procedure solves both regression estimation and pattern recognition problem:

$$\vartheta_i^{(k)} = r_i^{(k-1)} \sum_{j=1}^{N} \lambda_j^{(k)} \omega_j \text{ или } \vartheta_i^{(k)} = r_i^{(k-1)} \sum_{j=1}^{N} \lambda_j^{(k)} g_j \omega_j, \quad (9)$$

$$r_i^{(k)} = (r_i^{(k-1)})^2 \sum_{j=1}^{N} \sum_{l=1}^{N} K_i(\omega_j,\omega_l) \lambda_j^{(k)} \lambda_l^{(k)}, \quad (10)$$

where the real numbers $\lambda_1^{(k)},...,\lambda_N^{(k)}$ are the Lagrange multipliers (nonnegative in the case of pattern recognition) found as solutions of the respective dual problem. Updating the constant $b^{(k)}$ doesn't offer any difficulty. As we see, the abstract variables $\vartheta_i^{(k)} \in \tilde{\Omega}_i$ (9) are linear combinations of the objects of the training set in the sense of linear operations induced by the kernel functions as inner products in the respective linear spaces. Substitution of (9) and (10) into (5) eliminates $\vartheta_i^{(k)}$ and gives the completely constructive estimate of the sought function, respec-

tively, for regression estimation and pattern recognition:

$$\hat{y}^{(k)}(\omega)=\sum_{j=1}^{N}\lambda_j^{(k)}\sum_{i=1}^{n}r_i^{(k)}K_i(\omega_j,\omega\ )+b^{(k)},$$
$$\hat{y}^{(k)}(\omega)=\sum_{j=1}^{N}\lambda_j^{(k)}g_j\sum_{i=1}^{n}r_i^{(k)}K_i(\omega_j,\omega)+b^{(k)}\begin{cases}>0,\\<0.\end{cases}\quad(11)$$

As a rule, the process converges in 10-15 steps and displays a pronounced tendency to suppressing the weights at "abundant" potential functions $r_i \to 0$ along with emphasizing $r_i \gg 0$ the kernel functions which are "adequate" to the trainer's data. This fact provides a computationally effective selection of potential functions without straightforward discrete choice of their subsets.

A particular case: Feature selection as kernel fusion

There is no insurmountable barrier between the featureless kernel-based way of forming parametric families of numerical functions on a set of objects of any kind and the usual parametric family of linear functions on the set of objects represented by vectors of their numerical features. The latter way is particular case of the former one. Indeed, a numerical feature $x(\omega):\Omega\to\mathbb{R}$ is equivalent to the simplest kernel function in the form of product $K(\omega',\omega'')=x(\omega')x(\omega'')$ that embeds the set of objects into a one-dimensional linear space $\Omega\subseteq\tilde{\Omega}$. Respectively, a vector of features $\mathbf{x}(\omega)=\left[x_1(\omega)\cdots x_n(\omega)\right]$ gives n kernel functions at once $K_i(\omega',\omega'')=x_i(\omega')x_i(\omega'')$ and n versions of such an embedding $\Omega\subseteq\tilde{\Omega}_i$. The choice of one object in each of these spaces $\vartheta_i\in\tilde{\Omega}_i$, $i=1,...,n$, namely, n real numbers $\left(x_1(\vartheta_1)\cdots x_n(\vartheta_n)\right)\in\mathbb{R}^n$, along with a numerical constant $b\in\mathbb{R}$ specifies a linear function on the set of objects: $y(\omega)=\sum_{i=1}^{n}K_i(\vartheta_i,\omega)+b=\sum_{i=1}^{n}a_ix_i(\omega)+b$ where $a_i=x_i(\vartheta_i)$. The less the ith coefficient, i.e. the norm of the ith imaginary object $\|\vartheta_i\|=x_i(\vartheta_i)$, the less is the contribution of this feature $x_i(\omega)$ to the value of the function.

Conclusions

Treating the universal set of "all feasible" objects as a linear space practically wipes out the difference between a set of kernels and a set of features and, so, between the featureless and feature-based approach to data analysis. The featureless multi-kernel approach replaces the problem of choosing the features by that of choosing the kernels. According to which of these two problems is easier, the feature-based or the featureless approach should be preferred.

References

1. M.A. Aizerman, E. M. Braverman, L. I. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control, 1964, Vol. 25, pp. 821-837.
2. V. Vapnik. Statistical Learning Theory. John-Wiley & Sons, Inc. 1998.
3. Yu.I. Zhuravlev, V.V. Nikiforov. Recognition algorithms based on computing estimates. Kibernetika, 1971, No. 3 (in Russian).
4. V.V. Mottl. Metric spaces admitting linear operations and inner product. Doklady Mathematics, Vol. 67, No. 1, 2003, pp. 140–143.
5. Yu.I. Zhuravlev. Correct algebras over sets of inaccurate (heuristic) algorithms. I,II,III. Kibernetika, 1977, No. 4,6; 1978, No 2 (in Russian).