Kernel Machines for Continuous and Discrete Variables

Francisco J. Ruiz[1], Cecilio Angulo[1], and Núria Agell[2]

[1] Grup de Recerca en Enginyeria del Coneixement, Universitat Politècnica de Catalunya. 08800 Vilanova i la Geltrú, España
fjruiz@mat.upc.es, cecilio.angulo@upc.es

[2] Grup de Recerca en Enginyeria del Coneixement, ESADE Universitat Ramon Llull. Av. Pedralbes 60-62. 08034 Barcelona, España
nuria.agell@esade.edu

Abstract. Kernel Machines, such as Support Vector Machines, have been frequently used, with considerable success, in situations in which the input variables were real values. Lately, these methods have also been extended to deal with discrete data such as string characters, microarray gene expressions, biosequences, etc. In this contribution we describe a new kernel allowing kernel machines to be applied in problems in which continuous and discrete variables, described mainly by its order of magnitude, but also by an interval, take part simultaneously. In addition, the structure of the features space induced by this kernel is also defined considering the nature of both continuous and discrete variables.

1 Introduction

Machine learning algorithms based on kernel functions (Kernel Machines) have been frequently used, with considerable success, in situations in which the input variables were real values. Polynomial and exponential kernels, including *Gaussian kernels*, are the most used in this kind of data. Implicitly, when a kernel function is employed, input variables from the original space are projected onto a different space, usually with a higher, even infinite, dimension, whose metric replaces the metric in the original input space. This projection is usually carried out in order to transform a non-linear problem into a linear one. This working space, called features space, is usually omitted during the training process, because it is not explicitly required to know either its structure, or its dimension. A direct classification or regression processing into a high dimension space is computationally expensive, so a projection mapping allowing us to work in this space in an implicit manner is a very appealing behaviour on using kernels.

Moreover, the projection of the original data in the input space into a different space, enables these learning machines to be used in situations where input variables belong to a space without an Euclidian metric, because it is only necessary to enclose a certain metric in the features space. This circumstance has motivated the use of these algorithms in the cases where variables are not real numbers, but string chains, images, protein chains or genes. On these occasions, the procedure implemented is to define a designed features space and a projection mapping, through which the appropriated kernel is developed.

Following this procedure, in a similar form to [1] for the *string kernels*, a kernel is proposed in this article to be used when only information about the order of magnitude of the variables implied in the problem is available. Next, it becomes obvious that this kernel converges on the continuous exponential kernel when the granularity of the discrete kernel tends to infinity, it being possible to apply this function both to real value and intervals. This relation will allow the defined kernel to be used in learning problems when continuous real data, orders of magnitude or intervals simultaneously appear.

In section 2, the kernel concept is introduced, in addition to, two different features spaces associated with continuous variables, a Reproducing Kernel Hilbert Space (RKHS) and another one associated with the Hilbert space $L_2(\mathbb{R})$ with the usual inner product. This last features space allows a generalization of the interval space which it is also introduced in this section. In section 3, the Qualitative Space of the Absolute Orders of Magnitude is briefly introduced and a suitable kernel is obtained to deal with data in this space. Section 4 is devoted to demonstrating the relationship between this function and those defined in section 2. Finally, some conclusions are extracted and a list of further research topics to be developed is enumerated.

2 Kernel Functions

Some learning machines, such as Support Vector Machines (SVM), can deal with a dual formulation of the learning problem [2]. In this dual approach, the decision function sought can be expressed by the inner product of the input data and the training patterns. In the linear case, the discriminant function can be written as,

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i \cdot y_i \cdot \langle \mathbf{x_i}, \mathbf{x} \rangle + b \tag{1}$$

where $\{\mathbf{x_i}, y_i\}$ are, respectively, the input variables and the known class of the training patterns.

If a new representation of the patterns' attributes, $\mathcal{X}$, in a new space is employed, $\mathcal{F} = \phi(\mathcal{X})$, the so-called *features space*, the originally non-separable problem can be converted into a separable one. The discriminant function will be linear in $\mathcal{F}$, even if it is not in the original space,

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i \cdot y_i \cdot \langle \phi(\mathbf{x_i}), \phi(\mathbf{x}) \rangle + b \tag{2}$$

The dual formulation allows the double step procedure of maps ϕ and the inner product to be replaced by the function $K(\mathbf{x_i}, \mathbf{x_j}) = \langle \phi(\mathbf{x_i}), \phi(\mathbf{x_j}) \rangle$ which is called the *kernel function*. Hence, the discriminant function will be,

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i \cdot y_i \cdot K(\mathbf{x_i}, \mathbf{x}) + b \tag{3}$$

When the attributes' space $\mathcal{X}$ is a finite space, $\mathcal{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$, function K is determined by the square matrix, $\mathcal{K} = (K(\mathbf{x_i}, \mathbf{x_j}))_{i,j=1}^{n}$. A necessary condition to assure that this function is a suitable kernel, i.e. it represents an inner product in a certain unknown features space, is that matrix $\mathcal{K}$ must be symmetric and semi-definite positive, i.e. their eigenvalues are not negative. This is because if a certain negative eigenvalue exists, then the squared norm of some vector in the features space would be negative. In general, a symmetric function K defined in $\mathcal{X} \times \mathcal{X} \in \mathrm{I\!R}^n \times \mathrm{I\!R}^n$ is a kernel if the Mercer condition is met,

$$\int_{\mathcal{X} \times \mathcal{X}} K(\mathbf{x}, \mathbf{z}) f(\mathbf{x}) f(\mathbf{z}) d\mathbf{x} d\mathbf{z} \geq 0 \quad \forall f \in L_2(\mathcal{X}) \tag{4}$$

where $L_2(\mathcal{X})$ is the Hilbert space of the squared integrable functions in $\mathcal{X}$. It can be demonstrated that this condition is equivalent to showing that for any finite subset of $\mathcal{X}$, the associated matrix $\mathcal{K}$ is symmetric semi-definite positive.

2.1 The RKHS Features Space

A fixed kernel does not univocally determine the representation of either the map ϕ, or the features space $\mathcal{F}$. One of the possible features space that can be associated with a fixed kernel is the so-called *Reproducing Kernel Hilbert Space* (RKHS). This space is a subset of the whole set of the real functions defined on $\mathcal{X}$, which will be called $\mathbf{R}^{\mathcal{X}}$ [3]. It is built from the map ϕ,

$$\begin{aligned} \phi : \mathcal{X} &\rightarrow \mathrm{I\!R}^{\mathcal{X}} \\ \mathbf{x} &\rightarrow K(\cdot, \mathbf{x}) \end{aligned} \tag{5}$$

i.e. for any value $\mathbf{x} \in \mathcal{X}$, there exists a map from $\mathcal{X}$ to $\mathrm{I\!R}$ defined on the kernel K.

The RKHS is the complete set of all the maps of $\mathrm{I\!R}^{\mathcal{X}}$ in the form $f(\cdot) = \sum_{i=1}^{m} \alpha_i \phi(\mathbf{x_i}) = \sum_{i=1}^{m} \alpha_i K(\cdot, \mathbf{x_i})$, with $m \in N$ and $\mathbf{x_1}, ...\mathbf{x_m} \in \mathcal{X}$, that is, the functions space in $\mathrm{I\!R}^{\mathcal{X}}$ being a linear combination of the maps $\phi(\mathbf{x_i})$, with $\mathbf{x_i} \in \mathcal{X}$. Each element in this space is determined by a finite set of real numbers $(\alpha_1, \alpha_2, ..., \alpha_m)$, not being necessarily unique because $\{\phi(\mathbf{x_1}), \phi(\mathbf{x_2}), ..., \}$ are generally not linearly independents. In this space, the inner product is defined in the following manner. Let $f, g \in RKHS$ be two functions in the form,

$$f = \sum_{i=1}^{m_1} \alpha_i \cdot k(\cdot, \mathbf{x_i}) \quad g = \sum_{i=1}^{m_2} \beta_i \cdot k(\cdot, \mathbf{x_i}) \tag{6}$$

then,

$$\langle f,g\rangle = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\alpha_i\cdot\beta_j\cdot K(\mathbf{x_i},\mathbf{x_j}) \tag{7}$$
$$= \sum_{j=1}^{m_2}\beta_j\cdot f(\mathbf{x_j}) = \sum_{i=1}^{m_1}\alpha_i\cdot g(\mathbf{x_i})$$

Equations 6 and 7 show that, although scalar numbers α_i and β_i are not univocally defined, the inner product is well defined. It is easy to demonstrate that this definition satisfies,

$$\langle\phi(\mathbf{x_i}),\phi(\mathbf{x_j})\rangle = K(\mathbf{x_i},\mathbf{x_j}) \tag{8}$$

2.2 A features space associated with the Hilbert space $L_2(\mathbb{R})$ with the usual inner product

The RKHS is only a possible features space. However, it is the most usual associated with any kind of kernel. A different methodology to built a features space associated with real values will be described below. In this case, inversely to the RKHS construction method, the features space $\mathcal{F}$ will be built not having the kernel as starting point. Nevertheless, it will be demonstrated that, by using this methodology, it is possible to reproduce some well-known kernels, such as the Gaussian one. The proposed features space will subsequently be related to a kernel defined on the Absolute Orders of Magnitude space.

Definition 1. *Let $\phi:\mathbb{R}\to L_2(\mathbb{R})$ be a map such that $\phi(x_0)=f_{x_0,\sigma}(x)$, with $f_{x_0,\sigma}(x)=F_\sigma(|x-x_0|)=F_\sigma(z)$, F being a decreasing function with respect to $z=|x-x_0|$ in $\mathbb{R}^+$, with $F_\sigma(x_0)=1$. Let σ be a parameter or a set of parameters. Then, function $f_{x_0,\sigma}(x)$ will be called* influence function*.*

So, the proposed features space is in $L_2(\mathbb{R})$ and it is composed of a set of functions $f_{x_0,\sigma}(x)$ reflecting the influence of x_0 on x. This function will be symmetric with respect to $x=x_0$ and its value decrease along the influence of x_0 on x decrease.

The kernel is now defined by using the usual inner product in $L_2(\mathbb{R})$,

$$K(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle=\int_{-\infty}^{\infty} f_{x_i,\sigma}(x)f_{x_j,\sigma}(x)dx \tag{9}$$

Let us illustrate the use of the influence function by definingb four possible exemples, *hard*, *triangular*,*exponential*, and *Gaussian* influence functions. Graphics in Figure 1 illustrate the shape of these functions. The shaded areas represent $K(x_i,x_j)$ for two particular values of x_i and x_j. Graphics on the right show $K(x_i,x_j)$ in front of $|x_i-x_j|$ for each of the four cases.

Example. *For the* hard *influence function*

$$f_{x_0,\sigma}(x)=\begin{cases}1\ if\quad |x-x_0|\le\sigma\\ \\ 0\ otherwise\end{cases} \tag{10}$$

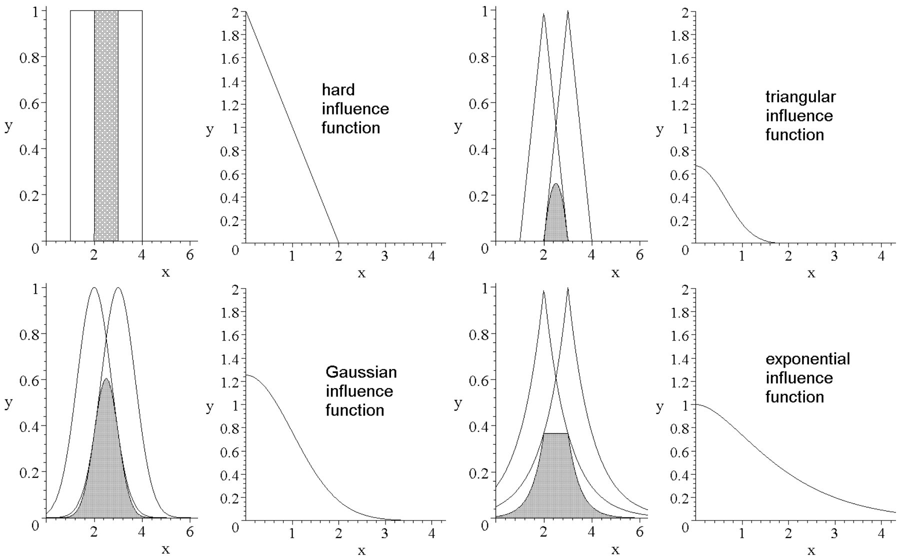

Fig. 1. Four different influence functions. On the left, *hard* and *triangular*, and on the right, *exponential* and *Gaussian* are represented. The interpretation of the associated kernel for two fixed values $\mathbf{x_i}$ and $\mathbf{x_j}$ is also illustrated (the area of shaded zone). On the right of each one, $K(x_0, x_0 + x)$ is represented according to x.

the associated kernel is

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \begin{cases} 2\sigma - |x_i - x_j| \; if \quad |x_i - x_j| \le 2\sigma \\ 0 \qquad\qquad\qquad otherwise \end{cases} \tag{11}$$

For the triangular *influence function*

$$f_{x_0,\sigma}(x) = \begin{cases} \frac{\sigma - |x - x_0|}{\sigma} \; if \quad |x - x_0| \le \sigma \\ 0 \qquad\qquad otherwise \end{cases} \tag{12}$$

the associated kernel will be

$$K(x_i, x_j) = \begin{cases} \frac{-6|x_i - x_j|^2\sigma + 3|x_i - x_j|^3 + 4\sigma^3}{6\sigma^2} \; if \quad |x_i - x_j| \le \sigma \\ \frac{(2\sigma - |x_i - x_j|)^3}{6\sigma^2} \qquad\qquad if \;\; \sigma < |x_i - x_j| \le 2\sigma \\ 0 \qquad\qquad\qquad\qquad otherwise \end{cases} \tag{13}$$

For the Gaussian *influence function*

$$f_{x_0,\sigma}(x) = e^{-\frac{(x - x_0)^2}{\sigma^2}} \tag{14}$$

the kernel will be

$$K(x_i, x_j) = \sigma\sqrt{\frac{\pi}{2}}e^{-\frac{(x_i-x_j)^2}{2\sigma^2}} \tag{15}$$

And, finally, for the exponential *influence function*

$$f_{x_0,\sigma}(x) = e^{-\frac{|x-x_0|}{\sigma}} \tag{16}$$

the kernel will be

$$K(x_i, x_j) = (|x_i - x_j| + \sigma) \cdot e^{-\frac{|x_i-x_j|}{\sigma}} \tag{17}$$

In the third case, the well-known Gaussian kernel is obtained in a different form from the standard use of the RKHS. In any case, the features space generated for the Gaussian kernel is also composed of Gaussian functions. However, in the case introduced, the width is smaller. This third example is, on the other hand, a new demonstration that the Gaussian function is effectively a kernel.

2.3 A Features Space for Intervals

Let $I(\mathbb{R})$ be the set of all the intervals on $\mathbb{R}$,

$$I(\mathbb{R}) = \{[a, b] \mid a \in \mathbb{R}, b \in \mathbb{R}, a \leq b\} \tag{18}$$

Definition 2. *Let $\phi : I(\mathbb{R}) \to L_2(\mathbb{R})$ be a map defined in such a way that $\phi([a, b]) = f_{[a,b],\sigma}(x)$ where $f_{[a,b],\sigma}(x) = f_{a,\sigma}(x)$ if $x < a$, $f_{[a,b],\sigma}(x) = 1$ if $a \leq x \leq b$, and $f_{[a,b],\sigma}(x) = f_{b,\sigma}(x)$ if $x > b$, for some influence function $f_{x_0,\sigma}$.*

For instance, if the exponential influence function is used, then

$$f_{[a,b],\sigma}(x) = \begin{cases} e^{-\frac{|x-a|}{\sigma}} & if \;\; x < a \\ 1 & if \;\; a \leq x \leq b \\ e^{-\frac{|x-b|}{\sigma}} & if \;\; x > b \end{cases} \tag{19}$$

By using the exponential function it is possible to express $f_{[a,b],\sigma}(x)$ as

$$f_{[a,b],\sigma}(x) = \frac{f_{a,2\sigma}(x) f_{b,2\sigma}(x)}{f_{b,2\sigma}(a)} \tag{20}$$

This new expression make the calculation of the matrix $\mathcal{K}$ easier , because,

$$\begin{aligned} K([a, b], [c, d]) &= \int_{-\infty}^{\infty} \frac{f_{a,2\sigma}(x) f_{b,2\sigma}(x) f_{c,2\sigma}(x) f_{d,2\sigma}(x)}{f_{b,2\sigma}(a) f_{d,2\sigma}(c)} dx \\ &= A \cdot \int_{-\infty}^{\infty} e^{-\frac{|x-a|+|x-b|+|x-c|+|x-d|}{2\sigma}} dx \end{aligned} \tag{21}$$

where

$$A = \frac{1}{f_{b,2\sigma}(a) f_{d,2\sigma}(c)} \tag{22}$$

In Figure 2 the shape of the image of an interval according to the map ϕ defined above can be observed. The shaded area corresponds to the value of the kernel for two fixed intervals. In the same figure, the value of $K([0,1],[x,1+x])$ with respect to x is also represented. It can be appreciated how K diminishes when distance between intervals increases.

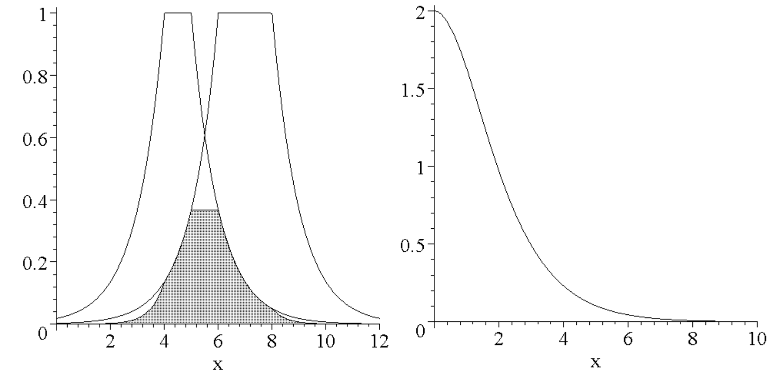

Fig. 2. Left. Representation in the features space of two intervals I_i andShaded area represents $K(I_i,I_j)$. Right. Representation of $K([0,1],[x,x+1])$ w.r.t. x.

3 Absolute Orders of Magnitude Model

One of the objectives of Qualitative Reasoning is to deal with problems in such a way that the relevance principle is conserved [4] , i.e., each variable involved in the problem is valued with the required level of precision. For this reason, the absolute orders of magnitude models [5] [6] work with a finite set of symbols or qualitative labels obtained via a partition of the real line, where any element of the partition is a basic label. These models provide a mathematical structure, which unifies sign algebra, and interval algebra trough a continuum of qualitative structures built from the rougher to the finest partition of the real line. In particular, the Absolute Orders of Magnitude Model of granularity n, $OM(n)$, is defined by a symmetric partition of the real line in $2n+1$ classes, from the real numbers $\{-a_{n-1}..., -a_1, 0, a_1..., a_{n-1}\}$. Each class is called a basic description or basic element, and is represented by a label of the set S_1,

$$S_1 = \{N_n, N_{n-1}, ...N_1, 0, P_1, ..., P_{n-1}, P_n\} \tag{23}$$

where

$$\begin{array}{c} N_n =]-\infty, -a_{n-1}], N_i =]-a_i, -a_{i-1}[, N_1 = [a_1, 0[, \\ 0 = \{0\}, \\ P_1 =]0, a_1], P_i =]a_{i-1}, a_i], P_n =]a_{n-1}, +\infty[\end{array} \tag{24}$$

Usually a linguistic label is associated with each of these classes, for instance small positive, medium positive, etc. Finally, this set is extended with all the possible convex subsets of the real line defined from the basic elements. So, the Quantity Space, S, is obtained by considering all the labels of the form,

$$I = [X, Y] \ ; \ \forall X, Y \in S_1 \ \ with \ \ X < Y, \tag{25}$$

where $X < Y$ represents $x < y \ \ \forall x \in X, \forall y \in Y$.

An order relationship, $\leq_P$, is defined in S, *to be more precise than*, given that $X, Y \in S$, X is more precise than Y, $X \leq_P Y$ if $X \subseteq Y$. From this relationship, the concept of *base of a qualitative label* can be defined $\forall X \in S - \{0\}$, as the set $B_X = \{B \in S_1 - \{0\} \mid B \leq_P X\}$. That is to say, B_X is composed of all the non-null basic labels contained X. On the other hand, the qualitative equality or *q-equality* is defined on pairs $X, Y \in S$, $X \approx Y$ if $X \cap Y \neq \emptyset$. This qualitative equality reflects the possibility that labels X and Y represent the same value.

The kernel construction will be induced from the concept of remoteness [7],

Definition 3. *Given a fixed $U \in S$, the remoteness with respect to U is defined as the map $a_U : S \to \mathbb{N}$ which for all $X \in S$, $a_U(X) = Card(B_{X_U}) - Card(B_X)$.*

Remoteness represents the number of non-null basic labels that had to be added to label X to obtain an element qualitatively equal to basic label U. The remoteness concept allows the map $\phi : S \to \mathcal{F} \subseteq [0,1]^{2n}$ to be defined in the following form,

Definition 4. *Given $X \in S - \{0\}$, and $\lambda \in [0,1]$,*

$$\phi(X) = (\lambda^{a_{N_n}(X)}, ..., \lambda^{a_{P_n}(X)})$$

.

This application reflects the global positioning of a qualitative label with respect to all the basic ones. The decay factor λ will produce that an increase in the remoteness between labels causing a diminution of the corresponding component.

The defined mapping ϕ permits us to build a kernel in the qualitative space of orders of magnitude in the following way, $K(X, Y) = \langle \phi(X), \phi(Y) \rangle$ where $\langle \cdot, \cdot \rangle$ is the usual inner product in $\mathbb{R}^{2n}$. This construction is directly extensible to a k-dimensional qualitative space $[OM(n)]^k$ to be used when patterns are given by k descriptions. In such a case, map ϕ will be a function $\phi : S^k \to \mathcal{F} \subseteq [0,1]^{2nk}$. For $X = (X_1, \ldots, X_k) \in S^k$, $\phi(X) = (\phi(X_1), ..., \phi(X_k))$.

4 Continuous limit of the Kernel in $OM(n)$

The features space $\mathcal{F}$ associated with the qualitative labels space S, defined in the past section, can be considered like a functions space, where functions are in the form $f_{\lambda,X_0} : S_1 - \{0\} \rightarrow [0,1]$. Hence,

$$\phi(X_0) = f_{\lambda,X_0}(X) = \lambda^{a_X(X_0)} \tag{26}$$

with $X \in S_1 - \{0\}$ and $X_0 \in S - \{0\}$.

Function $f_{\lambda,X_0}(X)$ can be considered as an influence function, like these introduced in subsection 2.3. It can be represented by associating each qualitative label X with a rectangle height $\lambda^{a_{X_0}(X)}$. In this manner, to apply the kernel to a pair of qualitative labels X_1 and X_2 is the same that adding the areas of the rectangles obtained by all the possible products of the highs of the rectangles associated with the basic labels in $\phi(X_1)$ and $\phi(X_2)$.

In Figure 3 the representation of the influence function for the non-basic labels $[N_2, N_1]$ and $[N_1, P_1]$ is illustrated, together with their inner product $\langle\phi[N_2, N_1], \phi[N_1, P_1]\rangle$ in the qualitative space $OM(3)$.

If it is considered that $a_{i+1} - a_i = \Delta$, a constant $\forall i$, and it is denoted x_i the midpoint of the basic label X_i, it can be written as,

$$\phi(X_0) = f_{\lambda,X_0}(X) = \lambda^{\frac{|x-x_0|}{\Delta}} \quad with \;\; X \in S_1 - \{0\} \tag{27}$$

This influence function can be related to those of the exponential function in Example 1, by using the equivalence,

$$\sigma = -\frac{\Delta}{ln\lambda} \tag{28}$$

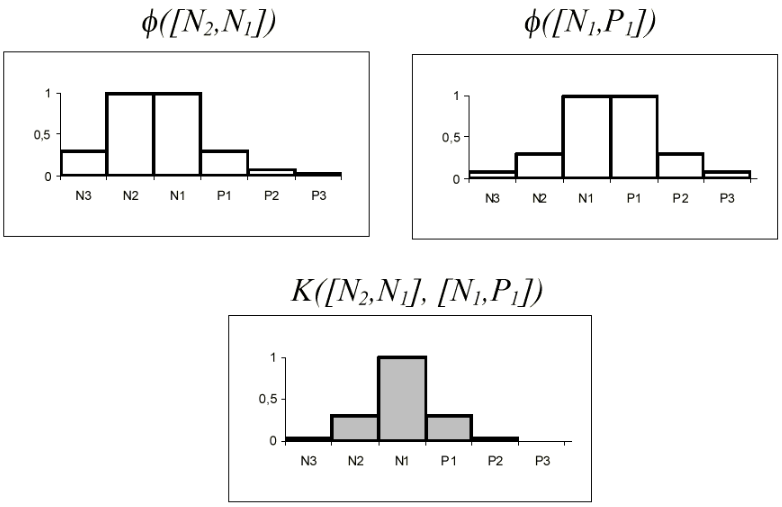

Fig. 3. Top. Representation in the features space of a pair of qualitative labels $[N_2, N_1]$ and $[N_1, P_1]$. Bottom. Shaded area represents $K([N_2, N_1], [N_1, P_1])$.

5 Conclusions and Further Work

A new methodology to obtain kernel functions has been proposed. It allows us to deal with either continuous variables, orders of magnitude, or those defined on real intervals. These kernels are obtained via a mapping from the original data set to the Hilbert space $L_2(\mathbb{R})$ with the usual inner product. In this manner, by using a Gaussian function, it is possible to reproduce the standard Gaussian kernel. A special case, the exponential function, has been pointed out because it has been demonstrated that it corresponds to the continuous limit of a kernel defined in the qualitative space of the absolutes orders of magnitude. The relationship obtained opens up a new challenge to develop learning methods based on kernels to be applied when data are simultaneously expressed in non-standard form, such as intervals, orders of magnitude, and real values.

Future work must be done about the definition of new concepts to measure the degree of remoteness between qualitative labels. Also, it is necessary to design procedures to chose different parameters of the decay parameter λ, depending on the length of the intervals defining basic labels.

References

1. Lodhi, H., Shawe-Taylor, J., Cristianini, N., Watkins, C.J.C.H.: Text classification using string kernels. NIPS (2000) 563–569
2. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and other kernel-base learning methods. Cambridge University Press (2000)
3. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press (2001)
4. Forbus, K.D.: Commonsense physics. Annals Revue of Computer Science (1988) 197–232
5. Agell, N.: Estructures matemàtiques per al model qualitatiu d'ordres de magnitud absoluts. PhD thesis, Universitat Politècnica de Catalunya (1998)
6. Travé-Massuyès, L., Dague, P., Guerrin, F.: Le Raisonnement Qualitatif pour les Sciences de l'Ingenieur. Hermès (1997)
7. Sánchez, M., Prats, F., Agell, N., Rovira, X.: Funciones kernel en espacios cualitativos de órdenes de magnitud: Aplicación a la evaluación del riesgo de crédito. Actas de la X Conferencia de la Asociación Española para la Inteligencia Artificial CAEPIA'03 **I** (2003) 69–78