UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 08-70
Statistics and Econometrics Series 024
December 2008

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

LOCALLY LINEAR APPROXIMATION FOR KERNEL METHODS: THE RAILWAY KERNEL

Javier González and Alberto Muñoz

Abstract

In this paper we present a new kernel, the Railway Kernel, that works properly for general (nonlinear) classification problems, with the interesting property that acts locally as a linear kernel. In this way, we avoid potential problems due to the use of a general purpose kernel, like the RBF kernel, as the high dimension of the induced feature space. As a consequence, following our methodology the number of support vectors is much lower and, therefore, the generalization capability of the proposed kernel is higher than the obtained using RBF kernels. Experimental work is shown to support the theoretical issues.

Keywords: Support Vector Machines, Kernel Methods, Classification Problems

Universidad Carlos III de Madrid, c/ Madrid 126, 28903 Getafe (Madrid), e-mail adresses: (Alberto Muñoz) alberto.munoz@uc3m.es, (Javier González) javier.gonzalez@uc3m.es.

Acknowledgements: The research of Alberto Muñoz and Javier González was supported by Spanish Goverment grants 2006-03563-001, 2004-02934-001/002 and Madrid Goverment grant 2007-04084-001.

Locally Linear Approximation for Kernel Methods: The Railway Kernel

Alberto Munoz, Javier González

December 18, 2008

Abstract

In this paper we present a new kernel, the Railway Kernel, that works properly for general (nonlinear) classification problems, with the interesting property that acts locally as a linear kernel. In this way, we avoid potential problems due to the use of a general purpose kernel, like the RBF kernel, as the high dimension of the induced feature space. As a consequence, following our methodology the number of support vectors is much lower and, therefore, the generalizacion capability of the proposed kernel is higher than the obtained using RBF kernels. Experimental work is shown to support the theoretical issues.

1 Introduction

Support Vector Machines (SVM) have proven to be a successful method for the solution of a wide range of classification problems [3], [13]. The best available techniques use kernel combinations to produce another kernel matrix [13, 15], for training the SVM. Kernel combinations can be linear [10] or functional [13]. In the first case it is possible to obtain positive definite combiantion by restricting the combination to belong to the cone of the positive definite matrices. The resulting optimization problem can be solved by quadratic programming [10]. In the second case, for some kernel combinations the definite positiveness cannot be guaranteed and a transformation of the final kernel has to be done. Even though, these combinations work better in real examples that the simpler linear schemes [13].

Linear SVMs are optimal in the classical setting in which two normally distributed populations have to be separated. This assertion is supported by the fact that SVM classifier approaches the optimal Bayes rule and its generalization error converges to the optimal Bayes risk [11]. Our aim in this paper is to build a functional global kernel for general nonlinear classification problems that locally behaves as a linear (optimal) kernel and that do not require any posterior transformation to be positive definite. Within this approach we expect to avoid the problems due to the use of a general purpose kernel like the RBF kernel: in this latter case, the data are embedded in a high dimensional feature space and problems of overfitting and poor generalization may appear. Since the proposed kernel

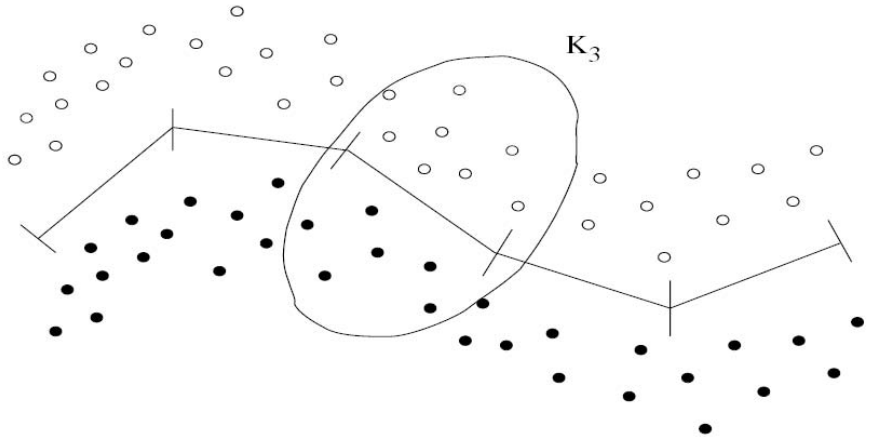


Figure 1: Illustration of the Railway Kernel performance.

behaves locally as a linear kernel, the good properties of the SVM classifier will be inherited by our method. In particular, the number of support vectors will be much lower and, therefore, the generalizacion capability will be higher than the obtained using RBF kernels.

To motivate our approximation, consider the situation presented in Figure 1. The decision function is clearly nonlinear. However, this function can be approximated locally by linear functions. For instance, a linear SVM (with kernel K_3) solves the classification problem in the oval area. We build a global kernel that will behave locally as the linear kernels whose decision functions are shown in the figure. We denote this kernel by '**Railway Kernel**'. The name for this kernel has been choosen because it is build like a railway where their wagons are the local decision functions.

This new Railway Kernel is presented as a particular case of a wider type of combinations where the kernels in the local areas are more complex. We focus here in the linear case because of its interesting properties we have already mentioned.

The paper is organized as follows. In Section 2 we briefly review the main concepts of Reproducing Kernel Hilbert Spaces and Support Vector Machines. The general framework for the proposed kernel is presented in Sections 3 and 4. The experimental setup and results on various artificial and real data sets are described in Section 5. Section 6 concludes.

2 Support Vector Machines in a nutshell

2.1 Reproducing Kernel Hilbert Spaces

There are several ways to introduce RKHS (see[1, 5, 21, 13]). In a nutshell, the essential ingredient for a Hilbert function space H to be a RKHS is the existence of a symmetric

positive definite function $K : X \times X \to \mathbb{R}$ named Mercer Kernel or reproducing kernel for H [1]. The elements of H, H_K in the sequel, can be expressed as finite linear combinations of the form $h = \sum_s \lambda_s K(x_s, \cdot)$ where $\lambda_s \in \mathbb{R}$ and $x_s \in X$.

Consider the linear integral operator T_K associated to the kernel K defined by $T_K(f) = \int_X K(\cdot, s) f(s) ds$. If we impose that $\int\int K^2(x, y) dx dy < \infty$, then T_K has a countable sequence of eigenvalues $\{\lambda_j\}$ and (orthonormal) eigenfunctions $\{\phi_j\}$ and K can be expressed as $K(x, y) = \sum_j \lambda_j \phi_j(x) \phi_j(y)$ (where the convergence is absolute and uniform).

2.2 Regularization and Support Vector Machines

The the starting point in this work is a two-class classification problem. Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a sample of n observations with $\mathbf{x}_i \in X$ (some input space) and $y_i \in Y \equiv \{1, -1\}$. Then, the classification problem can be solved by the Regularization Theory seeking the function f^* that solves the following functional optimization problem [5, 13] :

$$\arg \min_{f \in H_K} \frac{1}{n} \sum_{i=1}^{n} (y_i, f(\mathbf{x}_i))_+ + \gamma \|f\|_K^2 \,. \tag{1}$$

where $\gamma > 0$, $L(y_i, f(\mathbf{x}_i)) = (|c(\mathbf{x}_i) - y_i| - \varepsilon)_+$, $\varepsilon \geq 0$ and $\|f\|_K$ is the norm of the function f in H_K, a Reproducing Kernel Hilbert Space of Kernel K. Since $f \in H_K$ it holds that, for every $\mathbf{x} \in X$, $f(\mathbf{x}) = \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \mathbf{x})$, for appropriate $\mathbf{x}_i \in X$ and $\alpha_i \in \mathbb{R}$. Thus, calling $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ a given data set of points in X, and $K_{(S)} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j}$, then we will have $\|f\|_K^2 = \sum_{i=1}^n \sum_{i=1}^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) = \alpha^T K \alpha$. This approach is equivalent to the Support Vector Machines (SVM) originally proposed in [2].

Expression (1) measures the trade-off between the data error and the complexity of the solution (measured by $\|f\|_K^2$). For details, proofs and generalizations, refer to [7], [19], [6]. By the Representer Theorem ([7, 19]), the solution f^* to the functional optimization problem (1) exists, is unique and admits a representation of the form

$$f^*(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}), \quad \forall \mathbf{x} \in X \text{ where } \alpha_i \in \mathbb{R} \,. \tag{2}$$

In practice the solution to (1) is obtained by solving and quadratic problem and efficient methods specific for SVMs have been developed in the literature. In addition, due to the definition of the loss function, the solution to 1 generally depends on a small number of data called support vector. This implies that $\{\alpha_{1l}, ..., \alpha_{nl}\}$ generally contains a large number of zeros.

The relation kernel-RKHS is one to one. There exists a unique Hilbert space H_K of functions on X with reproducing kernel K [5]. This makes the election of the kernel K to be crucial to obtain the appropiate space to find the solution to (1).

3 Railway Kernel

In this section we will study our new type of locally linear kernel, the Railway Kernel. We define this kernel as a particular case of a more general class of local kerenls. We proceed as follows: First, the kernel is defined on 'simple' areas where the linear SVM works. Then the kernel is extended to the intersection of such 'pure' areas.

Next we introduce a special kernel that acts as an indicator function on the proccess. based on the use of an special type of indicator functions that help to determine which space area a point belong to, we show the kernel performance in an example with non intersection areas. Then, we analyze the case when intersection areas appear. Finally we propose a method to build the global kernel to solve the classification problem under consediration.

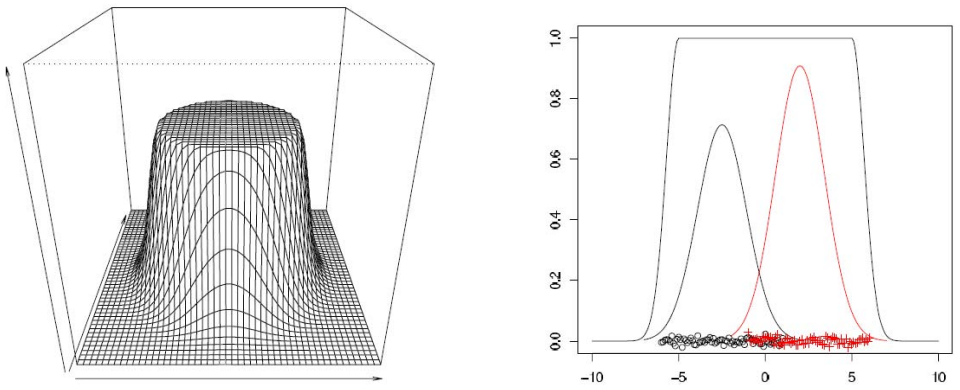
3.1 Indicator kernel functions

Given a data set, let assume that we are able to identify specific space areas where the problem can be solved using a linear SVM (see [13] for studing the convergence of SVMs with linear kernels to the optimal Bayes rule in a linear case). In this section we define a special indicator function to identify such areas. For the sake of simplicity only spherical areas are considered in this paper. The generalization to more elaborated shapes is straightforward. The indicator kernel function takes value 1 if the point under consideration is in the circular area defined by a given center and a radius, and decreases to zero quite fast as the distance to the center grows. Assumming smoothness in the boundary of the areas, we can define the following indicator kernel function $\lambda(x)$:

$$\lambda(x) = \begin{cases} 1 & \text{if} \|x-c\|^{1/2} \leq r \\ e^{-\gamma\left(\|x-c\|^2 - r^2\right)} & \text{if} \|x-c\|^{1/2} > r \end{cases}, \quad (3)$$

where $\|\cdot\|$ denotes the Euclidean distance, $x \in R^d$ is a sample point, $c \in R^d$ is the center of the sphere and $r > 0$ is the radius. Parameter $\gamma > 0$ is fixed in order to obtain a fast transition from 0 to 1 and, in this case, $\lambda(x)$ will aproximate an indicator function. It is immediate to check that $\lambda(x)$ is a kernel. The two dimensional case is shown in Figure 2a. Figure 2b represents a two-class classification problem in one dimension and the corresponding indicator function. If a SVM kernel built from this indicator function is used to solve the classification problem, points outside the indicator influence will not be considered.

These indicator kernel functions will help us to identify space areas where a type of kernel works. We concentrate on the linear case. As first approach, only circular areas are considered and we assume that there is no intersection among them. The use of a smooth contour is justified by theoretical properties of the final kernel to construct.



(a) (b)

Figure 2: Indicator kernel funcions. (a) 2D case. (b) 1D case for a two-class classification problem. Class density functions are shown.

3.2 Railway Kernel for a two areas problem

First consider the case of two areas without intersection. Kernel K_1 solves the classification problem in area A_1 and so does K_2 in area A_2. Let x and y be two sample data points. We define two functions: $H_1(x, y) = \lambda_1(x)\lambda_1(y)$ and $H_2(x, y) = \lambda_2(x)\lambda_2(y)$, where λ_1 and λ_2 are indicator kernel functions (with appropriate c and r parameters). The functions H_1 and H_2 take the value 1 when x and y belong to the same area, and 0 otherwise. In this particular case, we define the global Railway Kernel K_R as follows:

$$K_R(x, y) = H_1(x, y)K_1(x, y) + H_2(x, y)K_2(x, y)\,. \tag{4}$$

Notice that the new kernel is obtained as a functional combination of linear kernels.

The Railway Kernel will approximate piecewise a global non-linear function by local linear functions. that guaranties that $K(x, y)$ is semidefinite positive. As consequence, if the computed kernel has an interesting structure. Notice that $K_R(x, y)$ is a block-diagonal matrix. This fact can be used to improve the optimization method used to solve the SVM problem (see [20] for details about the SVM optimization problem).

By the Representer Theorem, the SVM solution takes the form: $f(x) = \sum_i \alpha_i K(x, x_i) + b$. In this case, due to the particular Railway Kernel structure the solution is given by:

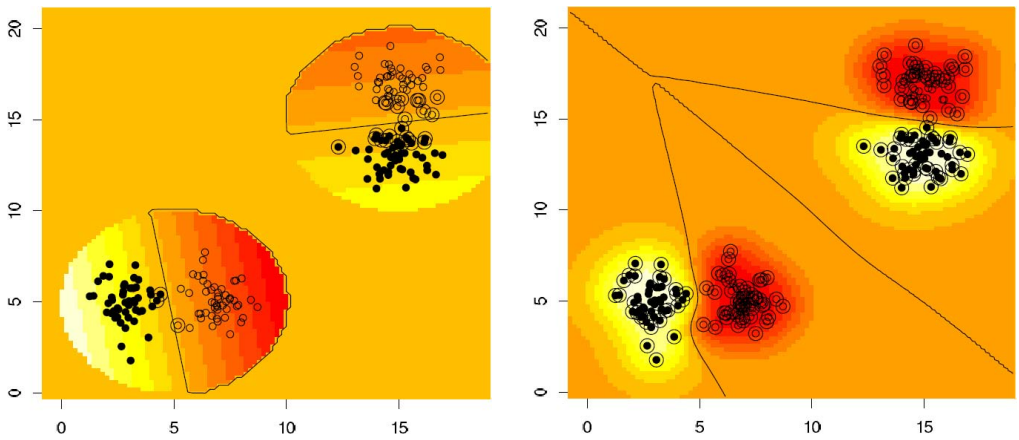
$$f(x) = \sum_{x_i \in A_1} \alpha_i K_1(x, x_i) + \sum_{x_j \in A_2} \alpha_j K_2(x, x_j) + b \tag{5}$$

Notice that K_R behaves like K_1 in the domain of indicator function H_1 and like K_2 in the domain of indicator function H_2.

We have not yet studied neither a multiarea problem, nor intersection between areas. These issues will be considered in Sections 3.4 and 3.5 .

3.3 A first example

Now we ilustrate the performance of this kernel in a simple example. In this example we generate four groups of observations (50 observations per group) corresponding to four bivariate normal distributions: $N(\mu_i, \Sigma_i)$ for group i, with $\mu_1 = (3, 5)$, $\mu_2 = (7, 5)$, $\mu_3 = (15, 17)$, $\mu_4 = (15, 13)$ respectively, and $\Sigma_1 = \Sigma_2 = diag(0.5, 1.5)$ and $\Sigma_3 = \Sigma_4 = diag(1.5, 0.5)$. Points in groups 1 and 3 belong to class $+1$ and points in groups 2 and 4 belong to class -1. Consider two areas defined by indicator kernel funcions with centers $c_1 = (5, 5)$, $c_2 = (15, 15)$ and radii $r_1 = r_2 = 5$ respectively. The point in this example is that the classes are linearly separable in each of these areas; however there is no a global proper linear kernel. In this case, the problem could be solved with a RBF kernel ($\sigma = 1$). Nevertheless when the Railway Kernel is used several advantages appear. The number of support vector is significatively lower than in the RBF case (13.5% vs. 73.5%). Figure 3a and 3b show the decision functions for the Railway and RBF kernels respectively. In

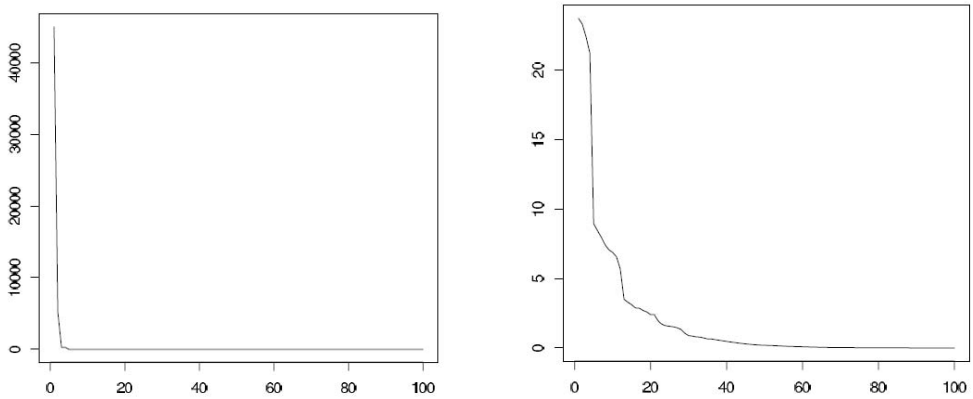


(a) Local linear kernels solution. (b) Global RBF solution.

Figure 3: Two solutions for a modified XOR problem (support vectors are highlighted).

addition, the number of positive eigenvalues of the kernel matrix is clearly lower using the Railway Kernel (2.0% vs. 25%). Therefore, the manifold induced by the Railway Kernel is of lower dimension than the obtained using the RBF kernel. Figures 4a and 4b show the eigenvalues for the Railway and RBF kernels respectively.

The resulting Railway Kernel matrix for this two areas problem is block diagonal. A graphical represetantion of the RK matrix for that example is shown in Figure 5. The two block correspond to two linear kernel applied to to the data of the two areas of the problem.



(a) Local linear kernel.

(b) RBF kernel matrix.

Figure 4: Eigenvalues of the kernel matrices for the modified XOR problem.

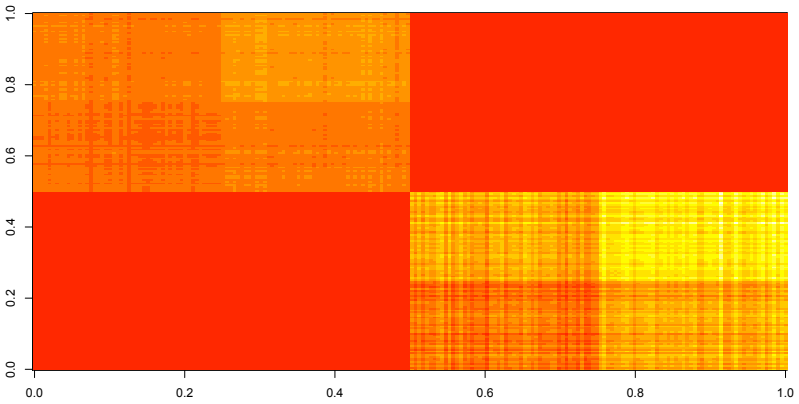


Figure 5: The Railway Kernel performance in an simple example with intersection.

3.4 Building the Railway Kernel in the intersections

In previous sections we have worked with very simple areas corresponding to different space areas. Next we deal with the problem of intersection between areas. Let A_1 and A_2 the areas under consideration. In this case, the Railway Kernel is built as follows:

$$K_R(x,y) = \begin{cases} K_1(x,y) & \text{if} \quad x,y \in A_1 \cap A_2^c\,, \\ K_2(x,y) & \text{if} \quad x,y \in A_1^c \cap A_2\,, \\ \frac{1}{2}(K_1(x,y)+K_2(x,y)) & \text{if} \quad x,y \in A_1 \cap A_2\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{6}$$

where A_i^c represents the complementary set of A_i.

Intersections between areas can be seen as areas where both kernels achieve the same performance, and should be equally weighted. Thus, the average of the kernels (which is a kernel [4]) is computed for points in the intersection. Figure 6a shows graphically the idea of intersection, and Figure 6b shows the Railway Kernel perfomance in a simple example.

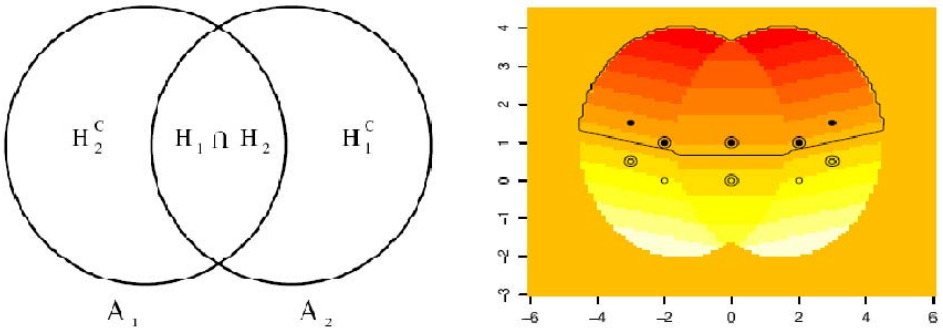


Figure 6: The Railway Kernel performance in an simple example with intersection.

The matrix computed in (6) is a semidefinite positive and block diagonal matrix. Thus, it comes from a Mercer kernel. It is possible to find an analytical expression for (6). Consider the example given in Figure 6a). Without loss of generality suppose that our sample is distributed in 3 zones: H_1^c, H_2^c and $H_1 \cdot H_2$, where $H_1^c(x,y)$ is the region of the space where the funcion H_1 vanishes and it is given by $H_1^c = (1-\lambda_1(x))(1-\lambda_1(y))$. Thus, it represents those points in A_2 and not in A_1. H_2^c represents those points in A_1 and not in A_2. The final kernel (K_R) will be the sum of three matrices. $K_R(x,y) = 0$ when x and y belong to different zones. In other case, $K_R(x,y)$ is exactly the kernel that works on the zone x and y belong to. The expression for the kernel is as follows:

$$\begin{aligned} K_R(x,y) \;=\; & H_1^c(x,y)K_2(x,y) + H_2^c(x,y)K_1(x,y) + \\ & (H_1(x,y)H_2(x,y))\frac{1}{2}(K_1(x,y)+K_2(x,y))\,. \end{aligned} \tag{7}$$

As before, the matrix corresponding to the kernel K_R aplied to any sample X is a block-diagonal matrix where each block is a kernel matrix. Then, K_R is a positive semidefinite matrix and thus a kernel. The generalisation of (7) to the case of more than 2 areas is straightforward.

Notice that, to compute the Railway Kernel it is enough to use the areas information and the local linear kernels on that areas. We have built a method to compute an aproximation to a nonlinear decision function with a sum of local linear hiperplanes.

3.5 Multiple areas Railway Kernel

In practical cases, the Railway Kernel is defined for multiple areas. Following eq. (7) a generalisation of K_R to a problem with a number p or ares is given by

$$\begin{aligned} K_R(x,y) &= \sum_{i=1}^{p} H_i(x,y)K_i(x,y) + \\ &+ \frac{1}{2}\sum_{i\neq j}^{p}(H_i(x,y)H_j(x,y)(K_i(x,y)+K_j(x,y)) \end{aligned} \tag{8}$$

where $H_i(x,y)$ for $i = 1, ..., p$ are the areas indicator functions and, $K_i(x,y) = x^T y$ (in our piece-wise kernel definition). When $K_i(x,y) = K(x,y)$, that is the same kernel is used in all the areas, the intersections can be considered as new independent zones. Therefore we can consider, without loss of generality, the Railway Kernel to be formed by $p + q$ independent zones, for q the number of intersections.

Proposition 1 *Let the Railway Kernel defined in eq. (10) for p non overlapped areas . Let $K_i(x,y) = \sum_{j=1}^{d} \mu_i \phi_j(x)\phi_j(y)$ for $i = 1, ..., p$ the kernel functions considered in the p areas. Then,*

$$K_R(x,y) = \sum_{i=1}^{p}\sum_{j=1}^{d} \mu_i \tilde{\phi}_j(x)\tilde{\phi}_j(y) \tag{9}$$

where $\tilde{\phi}_j(x) = \phi_j(x)\lambda_i(x)$, being the dimension of the space induced by K_R equal to $p \times d$.

Proof 1 *Let $K_i(x,y) = K(x,y) = \sum_{j=1}^{d} \mu_i \phi_j(x)\phi_j(y)$ for $i = 1, ..., p$ the local kernels of the p areas an their expansions via the Mecer's theorem. Since the areas are independent we can rewrite K_R as*

$$\begin{aligned} K_R(x,y) &= \sum_{i=1}^{p} H_i(x,y)K_i(x,y) + \\ &= \sum_{i=1}^{p} H_i(x,y)\sum_{j=1}^{d} \mu_i \phi_j(x)\phi_j(y) \qquad (10) \\ &= \sum_{i=1}^{p} \lambda_i(x)\lambda_i(y)\sum_{j=1}^{d} \mu_i \phi_j(x)\phi_j(y) \qquad (11) \end{aligned}$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{d} \mu_i \phi_j(x)\lambda_i(x)\phi_j(y)\lambda_i(y) \tag{12}$$

being $\tilde{\phi}_j(x) = \phi_j(x)\lambda_i(x)$ and the number of different eigenfunctions $p \times d$.

Remark 1 *This proposition is valid for any H(x,y) separable. That is for any H(x,y) such that*

$$H(x, y) = \lambda(x)\lambda(y) \tag{13}$$

for any x, y two sample points.

Example 1 *The railway kernel, for p areas, and $K(x, y) = x^T y$ is given by*

$$K_R(x, y) = \sum_{i=1}^{p} \lambda_i(x) x \lambda_i(y) y. \tag{14}$$

If the problem is originally in dimension 2 then $x = (x_1, x_2)$ and $y = (y_1, y_2)$. Then

$$K_R(x, y) = \sum_{i=1}^{p} \lambda_i(x)\lambda_i(y)(x_1 y_1 + x_2 y_2). \tag{15}$$

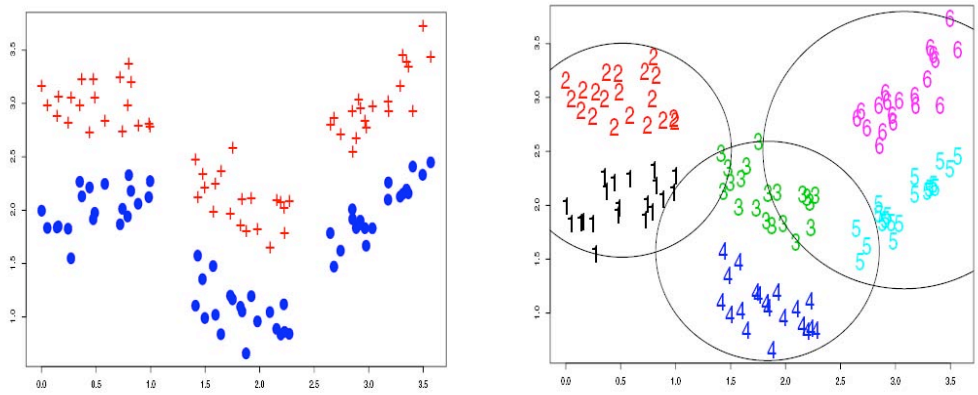
4 Areas Location

In Section 3 we have assumed that each point in the sample belongs to one or more previously defined areas. Now we present a local algorithm to detect such areas in a classification problem. The algorithm works in two main steps. First, single labelled areas are created. Second, the closest areas of different labels are joined in order to define the zones described in Section 2.1.

Following the ideas in [14], the procedure is based on the iteration of the K-means algorithm with increasing number of centroids until a condition is satisfied. This generates a partition of the space in M areas $A_1, \ldots, A_M$. The condition reflects the pureness of the regions found by K means. Interesting zones will be characterised by containing small (or zero in the case of pure areas) number of homogeneously distributed data with different label to the others. We apply a χ^2 -test to detect this situation. If the null hypothesis (small number of data of one class are randomly distributed among the data of the other) is not rejected for all the zones, the splitting procedure stops and the areas $A_1, \ldots, A_M$ automatically created.

Once the areas $A_1, \ldots, A_M$ have been built, the final areas are obtained by joining the nearest areas with different labels. This is done by comparing the minimum distances between the points of every single area to those of the areas with different label. This given rise to the zones defined by the functions $H(x, y)$. In order to obtain the indicator function kernels needed to build the Railway Kernels, centers and radii are needed. Centers are computed as the centroids, and radii in each area are computed as the maximum distance between the center and the farthest point in this area.

An example to illustrate of the performance of the algorithm is presented in Figure 7. Figure 7a presents a two class problem in two dimensions. In Figure 7b the result of applying the areas location algorithm is shown. The procedure is as follows. The K-means iteration stops when six centroids are used (since six perfect pure areas are found). After this, the final zones are obtained by joining the previous areas as is described above. Figure 7b shows the six final spherical areas detected.



(a) Initial problem.

(b) One Step of the algorithm.

Figure 7: An example of the Areas Location algorithm performance.

5 Experiments

To test the performance of the proposed method, a SVM (with the upper bound on the dual variables fixed to 1) has been trained on artificial and real data sets using the Railway Kernel matrix previously constructed.

We have compared the proposed methods with several techniques. In both experiments we have compared the Railway Kernel with two SVM classifiers built using RBF kernels in which the parameter has been tuned in two different ways. For the first classifier (SVM_1) the parameter σ is choosen as a function of the data dimension (see [17] and [15] for details). For the second (SVM_2), σ and the upper bound on the dual variables of the optimization problem are choosen following the ideas in [8]. We have also included the results for other 5 RBF kernels of parameters $\sigma = 0.5, 2.5, 5, 7.5$ and 10

In the second experiment we also used the same set of five kernels but we only included the values for the best and the worst test error case. We also included the average of them (AKM).

In order to compare the results with other techniques we estimated the test errors for a K-nn classifier, SVM of linear kernel, and the MARK-L procedure. Results are shown in Table 2.

5.1 Two areas with different scattering matrices

The first data set under consideration is presented in Figure 8 and corresponds to 400 points in $\mathbb{R}^2$. There are two areas of points (80% of the sample is in area A_1 and 20% is in area A_2). Each area A_i corresponds to a normal cloud. The first area center is $(0, 1)$ and the second group center is $(1, 1)$, while the diagonal covariance matrices are $\sigma_i^2 I$ where $\sigma_1 = 10^{-2}\sigma_2$, and $\sigma_2 = 1$. The first area center is $(0, 1)$ and the second group center is $(1, 1)$. The point on this example is that the areas do not coincide with the classes $\{-1, +1\}$ that are to be learned. Half of the points in each class belongs to area A_1, and the other half to area A_2. Within each area, the classes are linearly separable. Therefore, the only way to build a proper classifier for this data set is to take into account the area each point belongs to. We use 50% of the data for training and 50% for testing.

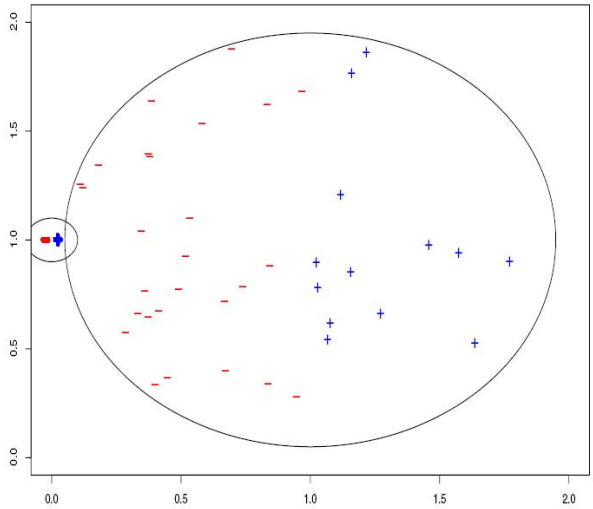


Figure 8: Two areas with different scattering matrices. The first area center is $(0, 1)$ and the second area center is $(1, 1)$. The areas do not coincide with the classes $\{-1, +1\}$.

To compare the performance of the Railway Kernel, consider a set of three RBF kernels with parameters σ =0.5, 5 and 10 respectively.

Table 1 shows the performance of our proposal for this data set. The results have been averaged over 10 runs. Given the geometry of the data, it is clear that is not possible to choose a unique best σ for the whole data set. As σ grows, the test error increases for the data contained in area A_1, and decreases within area A_2. The Railway Kernel clearly improves the best RBF kernel.

5.2 The breast cancer data set

In this section we have dealt with a database from the UCI Machine Learning Repository: the Breast Cancer data set [12]. The data set consists of 683 observations with 9 features each. We use 80% of the data for training and 20% for testing.

Table 1: Percentage of missclassified data and percentage of support vectors for the two different scattering data set: A_1 stands for the less scaterring group, A_2 stands for the most dispersive one.

	Train Error			**Test Error**			**Support Vectors**
Method	Total	A_1	A_2	Total	A_1	A_2	Total
$\mathbf{RBF}_{\sigma=0.5}$	2.4	3.0	0.0	13.4	4.1	51.0	39.2
$\mathbf{RBF}_{\sigma=2.5}$	3.0	3.8	0.0	12.6	6.5	41.5	62.1
$\mathbf{RBF}_{\sigma=5}$	4.6	5.8	0.0	13.6	8.6	35.0	82.6
$\mathbf{RBF}_{\sigma=7.5}$	14.9	18.4	0.5	18.7	22.6	20.5	94.6
$\mathbf{RBF}_{\sigma=10}$	29.1	36.2	0.5	36.0	44.1	10.0	94.4
Railway Kernel	3.7	3.6	15.6	4.2	0.1	20.6	14.1
$\mathbf{SVM}_1$	2.1	2.6	0.0	13.5	4.1	51.0	39.6
$\mathbf{SVM}_2$	2.1	2.6	0.0	11.0	3.3	41.5	37.6

Table 2 shows the performance of the Railway Kernel on this data set. Again, the results have been averaged over 10 runs. Our method clearly improve the RBF kernel with σ parameter choosen as a function of the data dimension. Our method does not take into account the penalization parameter of the SVM. However, our results are similar to the classification results obtained when both parameters, σ and the upper bound on the dual variables of the optimization problem, are choosen, but using significantly less support vectors.

6 Comments and Conclusions

In this paper we have presented a new kernel, the Railway Kernel. This global kernel takes advantage of the good generalization properties of the local linear kernels for classification tasks. We have shown that the potential problems due to the use of a general purpose kernel like the RBF kernel have been avoid. The generalization capability of the proposed kernel is higher than the obtained using RBF kernels. The method could be generalized by using alternative nonlinear local kernel. Further research will focus on the theoretical properties of the Railway Kernel and extensions.

References

[1] Aroszajn, N. *Theory of Reproducing Kernels.* Transactions of the American Mathematical Society, 68(3):337-404, 1950.

[2] Boser, B. E., Guyon, I. and Vapnik, V. *A training algorithm for optimal margin classifiers.* In Proc. Fifth ACM Workshop on Computational Learning Theory (COLT) 144.152. ACM Press, New York, 1992.

Table 2: Percentage of missclassified data, sensitivity (Sens.), specificity (Spec.) and percentage of support vectors for the cancer data set. Standard deviations in brackets.

Method	**Train**			**Test**			**Support Vectors**
	Error	Sens.	Spec.	Error	Sens.	Spec.	
Best RBF	2.3 (0.3)	0.979	0.976	3.1 (1.6)	0.976	0.966	13.6 (1.3)
Worst RBF	0.0 (0.0)	1.000	1.000	24.7 (2.3)	1.000	0.627	74.0 (2.4)
AKM	1.6 (0.3)	0.988	0.981	3.4 (1.5)	0.978	0.966	21.7 (1.2)
Railway Kernel	2.5 (0.3)	0.979	0.974	2.9 (0.4)	0.975	0.876	18.6 (3.6)
SVM	0.1 (0.1)	1.000	0.999	4.2 (1.4)	0.989	0.942	49.2 (1.0)
MARK-L	0.0 (0.0)	1.000	1.000	3.6 (1.2)	0.980	0.956	18.3 (0.0)
k**-NN**	2.7 (0.5)	0.961	0.980	3.4 (1.5)	0.949	0.974	— (—)
$\mathbf{SVM}_1$	0.1 (0.1)	1.000	0.999	4.2 (1.4)	0.989	0.942	49.2 (1.0)
$\mathbf{SVM}_2$	0.0 (0.0)	1.000	0.999	2.9 (1.6)	0.963	0.975	49.2 (1.0)

[3] C. Cortes and V. Vapnik. *Support Vector Networks.* Machine Learning, 20:273-297, 1995.

[4] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machine. Cambridge University Press, 2000.*

[5] Cucker, F. and Smale, S. *On the Mathematical Foundations of Learning.* Bulletin of the American Mathematical Society, 39(1):1-49, 2002.

[6] Cox, D. and O'Sullivan, F. *Asymptotic Analysis and Penalized Likelihood and Related Estimators.* Annals of Statistics, 18:1676-1695, 1990.

[7] Kimeldorf, G.S. and Wahba, G. *A Correspondence between Bayesian Estimation on Stochastic Processes and Smoothing by Splines.* Annals of Mathematical Statistics, 2:495-502, 1971.

[8] S.S. Keerthi and C. Lin. *Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel.* Neural Computation, 15:1667-1689, 2003.

[9] Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. *On combining classifiers.* IEEE Transactions on Pattern Analysis and Machine intelligence. 20(3), 226-239, 1998.

[10] Lanckriet, G.R.G., Cristianini, N., Bartlett, P., El Ghaoui, L., Jordan, M.I. *Learning the Kernel Matrix with Semidefinite Programming.* Journal of Machine Learning Research, 5,27-72, 2004.

[11] Y. Lin., G. Wahba, H. Zhang and Y. Lee. *Statistical Properties and Adaptive Tuning of Support Vector Machines.* Machine Learning, 48:115-136, 2002.

[12] O.L. Mangasarian and W.H. Wolberg. *Cancer diagnosis via linear programming.* SIAM News, 23 (5):1-18, 1990.

[13] Moguerza, J. M. and Muñoz, A. *Support Vector Machines with Applications.* Statistical Science, 21(3):322-357, 2006.

[14] A. Muñoz and J. M. Moguerza. *Combining Support Vector Machines and ARTMAP Architectures for Natural Classification* Springer-Verlag, LNCS, 2774, pp. 1621, 2003.

[15] A. Muñoz and J. M. Moguerza. *Estimation of High-Density Regions Using One Class Neighbor Machines* IEEE Transactions on Pattern Analysis and Machine Intelligence, 28 (3):476-480, 2006.

[16] Muñoz, A. and Martín de Diego, I. *From indefinite to Semi-Definite Matrices.* Springer-Verlag, LNCS 4109, 764-772, 2006.

[17] B. Schölkopf, J.C. Platt,J. Shawe-Taylor, A.J. Smola and R.C. Williamson. *Estimating the Support of a High Dimensional Distribution.* Neural Computation, 13(7):1443-1471 , 2001.

[18] B. Schölkopf, R. Herbrich, A. Smola and R. Williamson. *A Generalized Representer Theorem.* NeuroCOLT2 TR Series, NC2-TR2000-81, 2000.

[19] Schölkopf, B., Herbrich, R., Smola, A.J. and Williamson, R.C. *A Generalized Representer Theorem.* Lecture Notes in Artificial Intelligence, 2111:416-426, Springer, 2001.

[20] B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, 2002.

[21] Wahba, G. *Spline Models for Observational Data.* Series in Applied Mathematics, vol. 59, SIAM. Philadelphia, 1990.