# KMOD - A Two-parameter SVM Kernel for Pattern Recognition

N.E. Ayat [1,2] M. Cheriet [1] C.Y. Suen [2]

[1] LIVIA, École de Technologie Supérieure, 1100, rue Notre Dame Ouest, Montreal, H3C 1K3, Canada

[2] CENPARMI, Concordia University, 1455 de Maisonneuve Blvd West, Montreal, H3G 1M8, Canada

Emails: ayat@livia.etsmtl.ca, cheriet@gpa.etsmtl.ca, suen@cenparmi.concordia.ca

## Abstract

*It has been shown that Support Vector Machine theory optimizes a smoothness functional hypothesis through kernel application. We present KMOD, a two-parameter SVM kernel with distinctive properties of good discrimination between patterns while preserving the data neighborhood information. In classi£cation problems, the experiments we carried out on the Breast Cancer benchmark produced better performance than RBF kernel and some state of the art classi£ers. As well, it also generated favorable results when subjected to a 10-class problem of recognizing handwritten digits in the NIST database.*

## 1 Introduction

The SVM (Support Vector Machine) is a powerful classi£er that provides a linear separation in an augmented space, different from the original one, by means of some de£ned kernels. These kernels map the data vectors into a high-dimensional space, of possibly in£nite dimension, where a linear separation is more likely. This process amounts to £nding a non-linear frontier in the original input space. The £nal decision function is as: $f(x) = \sum_i \alpha_i y_i k(x_i, x)$; where $x_i$,$y_i$ and $k$ represent respectively a support vector, its corresponding label and a given kernel (Table 1). The parameters $\alpha_i$ are the solution of the following quadratic optimization problem to be maximized: $L_D = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$ and must satisfy $0 \leq \alpha_i \leq C$ where $C$ is a penalization parameter [7].

KMOD (kernel with moderate decreasing) has been introduced in [1]. The present work is an extension of it where we study the spectral behavior of KMOD and proceed experiments on two databases. The motivation is two fold. First, we do believe that controlling the spatial behavior of kernels is of major interest as long as the complexity and distribution of the data in hand vary widely from one classi£cation problem to another. In particular, when the data is sparse, it is important to capture the closeness information of all its points through the kernel application. KMOD is a two-parameter SVM kernel that allows such a behavior. As well, we explain intuitively its behavior with respect to the duality between spatial similarity in the original space and correlation in the augmented space. This additional precision let the SVM deal better with sparse and Cauchy distributed data. Another part of the work emphasizes on the spectral properties of KMOD and their connection with the entropy numbers. In section 2, we present our kernel and give an intuitive analysis of its characteristics. In section 3, we analyze the spectral behavior of KMOD and its connection with the entropy numbers theory. As well, we make some remarks about the generalization of the classi£er with respect to the kernel behavior in the frequency domain. In section 4 and 5, we give an experimental study of the classi£er performance on two real-life benchmarks: a two-class problem through UCI Breast Cancer database and a ten-class problem through NIST digit image database. The results on both data show the relative superiority of KMOD w.r.t alike kernels. In section 6, we summarize the work.

| Kernel | Formula |
|---|---|
| linear | $k(x,y) = x.y$ |
| sigmo"d | $k(x,y) = \tanh(ax.y + b)$ |
| polynomial | $k(x,y) = (1 + x.y)^d$ |
| RBF | $k(x,y) = \exp(-a\|x-y\|^2)$ |
| exponential RBF | $k(x,y) = \exp(-a\|x-y\|)$ |

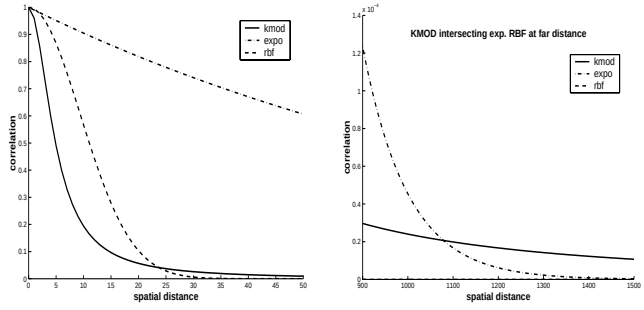
**Table 1.** *Common kernels*

## 2 KMOD: theoretical analysis

In general, the function that embeds the original space into the augmented feature space is unknown. The existence of such a function for a given kernel is assured by Mercer's theorem. This kernel must express a dot product in the feature space.

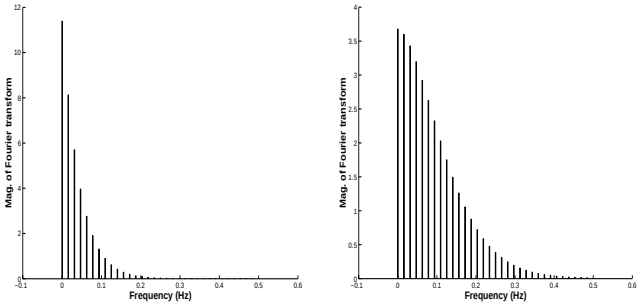
All kernels found in the literature are either dot product based functions ($k(x,y) = k(x.y)$) or distance based functions ($k(x,y) = k(||x-y||)$). By adopting the latter

formulation, knowing an estimation of the Euclidean distance between two points in the original space, we £nd how much they are correlated in the augmented space. The following question however arises: Is the kernel spatial behavior of any importance? In most common distance based kernels (e.g. RBF), points very close to each other are strongly correlated whereas points far apart have almost uncorrelated images in the augmented space. Our £rst concern is to force the image of the original points to be linearly separable in the augmented space. In order to possess such a behavior, a kernel must turn points which are very close to each other in the original space into weakly correlated points (as weak as possible) while still maintaining the closeness information from being zero. To meet this challenge, we need the following couple of features: a quick decrease in the neighborhood of zero and a moderate decrease near in£nity. The RBF kernel satis£es correctly the £rst requirement but not the second, whereas exponential RBF (Table 1) does not satify both requirements (Figure 1). Alternatively, we propose KMOD whose analytic expression is $kmod(x, y) = a[\exp(\frac{\gamma}{\|x-y\|^2+\sigma^2}) - 1]$; where $a$ is a normalization constant equal to $\frac{1}{\exp(\frac{\gamma}{\sigma^2})-1}$; $\gamma$ and $\sigma$ are two positive reals that control the decreasing behavior of the kernel. In particular, $\sigma$ is a scale space parameter that de£nes a gate surface around zero whereas $\gamma$ controls the decreasing speed around zero. The $-1$ bias ensures KMOD converges toward 0 at in£nity. It is worth to mention that the second property of KMOD allows capturing more closeness information from the data points. RBF kernel, however, is penalizing quickly intermediate neighborhood. Figure 1⟨b⟩ shows that in far neighborhood RBF function reaches a zero value. KMOD, however, is decreasing moderately and intersecting Exponential RBF at an arbitrary point. It is worthwhile to point out that the behavior of KMOD remains the same, even though its pro£le changes with its parameters.




⟨a⟩ ⟨b⟩

**Figure 1.** *⟨a⟩:Correlation in feature space vs. spatial distance in input space;⟨b⟩:KMOD preserving the far points closeness information*




⟨a⟩ ⟨b⟩

**Figure 2.** *Spectrum of KMOD ⟨a⟩: for $\gamma = 30$ and $\sigma = 5$; ⟨b⟩: for $\gamma = 190$ and $\sigma = 5$*

## 3 What about $\gamma$?

Recently, many authors have £gured out the similarity between SVM model and regularization theory, i.e. the maximization of the margin through the minimization of $\|w\|^2$ is a smoothness functional hypothesis where we refer to smoothness as a measure of the oscillatory behavior of a function [6]; i.e. the regularization properties of any kernel goes along with its spectral properties. In other hand, it has been shown that the properties of kernel spectrum can be used to make statements about the generalization error of the associated SVM. The connection is that a less pessimist bound on the generalization error, called the entropy numbers, is bounded by the decay rate of the Gram-Schmidt matrix eigenvalues [8]. As well, it is established that the decay rate of the Gram-Schmidt matrix eigenvalues is proportional to the decay rate of the kernel spectrum values. Thus, the faster this decay is, the smaller the expected error is. This space-spectral duality enables the analysis of SVM kernel properties in the frequency domain. In order to get denumerable spectral values, we simulated the discrete spectrum of KMOD by considering its periodic extension inside the spatial domain. Depending on the value of $\gamma$, we report in Figure 2 two different cases of the Fourier transform pro£les. Figure 2⟨a⟩ exhibits the spectral behavior of KMOD for small values of $\gamma$. The asymptotic curve in that case decreases exponentially i.e. the decay rate of the Gram-Schmidt matrix eigenvalues $\lambda_i$ is $O(e^{-\alpha i})$ for some $\alpha \geq 0$. The $n^{\text{th}}$ entropy number $\varepsilon_n$ is then $\ln \varepsilon_n^{-1} = O(\ln^{\frac{1}{2}} n)$ [8]. The plot in Figure 2⟨b⟩ shows that for medium values of $\gamma$, the asymptotic behavior decays as rapidly as $O(e^{-\alpha i^2})$. Hence, the bound on the entropy numbers is tighter and the model with such parameter values is expected to have a smaller generalization error. For large values of $\gamma$, KMOD spectrum tends to be a gate function with a width proportional to $\frac{1}{\sigma^2}$. Therefore, the Fourier Transform decay rate is slower and we expect a greater generalization error. The

optimum value for $\gamma$ may then belong to an intermediate span of values. However, the entropy numbers theory does not allow us to make a priori statements about the choice exact of KMOD parameters values. This is a model selection problem that need either a cross-validation procedure or optimization of an upper bound on the error.

## 4 Breast cancer benchmark

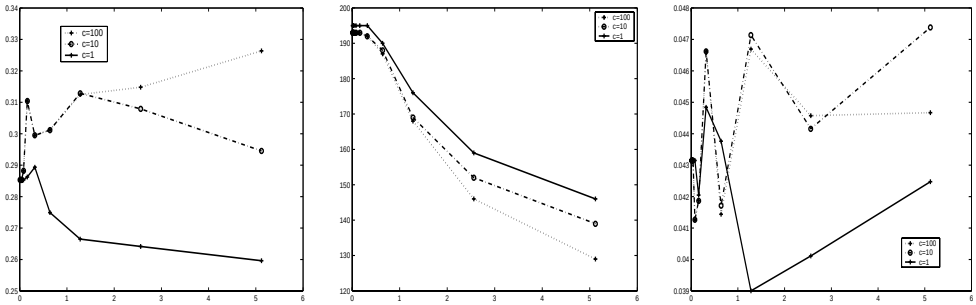
In order to assess our kernel, we perform simulations on the UCI Breast Cancer Database which is a hard classsi£cation problem. This data is a binary classi£cation problem partitioned into 100 different realizations of training and testing sets of respectively 200 and 77 examples each, upon which we train the classi£er and then compute its test error. To estimate the precision of the classi£er we perform successive training and testing throughout the 100 realizations and then compute the average and deviation of the generalization error. This protocol of assessment was already used in [5]. Hence we can compare our results with those of [5]. During the experiment, we used 4 different values for $C$ (cf. §1 ) : 0.1, 1, 10, 100; 50 values for $\sigma$ starting at 0.01 and 10 values for $\gamma$ starting at 0.001. We report in £gure 3 ⟨a⟩ the variation of the error w.r.t. $\sigma$ for C=1, 10 and 100. The error for C=0.1 is constant and equal to 0.287 so we did not report its curve. Except for C=100, the error decreases w.r.t. $\sigma$ until it reaches a minimum value (which is not in the scope of the £gure 3 ⟨a⟩). As well, C=1 gives the least generalization error with a value of 25.4±4.4 %. A value that is better than all published results on the Breast Cancer database [5] and which we recall below (see Table 2). Nevertheless, it is worth mentioning that a better value for C would be found using some gradient descent procedures that minimize an empirical estimate of the error. This is not the focus of the present paper. Moreover, notice in £gure 3 ⟨b⟩ that C=1, gives roughly, 10 and 20 more support vectors than C=10 and C=100. Clearly, this contrasts with a pessimist bound on the generalization error given by $\frac{\sharp sv}{l}$ in [7]; where $\sharp sv$ is the support vectors number and $l$ is the total number of training examples. Contrary to what one would expect, the error is smaller. In £gure 3 ⟨c⟩, we report the variation of the spread of the error w.r.t. $\sigma$. As well, we proceed a Student signi£cance test at 0.05 con£dence between KMOD SVM and other classi£ers related to the errors in the table. The plus sign beside the values in the table indicates whether or not the error is signi£cantly different from KMOD's one.

## 5 Recognition of handwritten digits

Support Vector Machine is a binary classi£er which is useful for a two-class data only. However, k-class pattern recognition problems (where one has $k \geq 3$ classes) such as the digit recognition task could be solved using a voting scheme method based on combining many binary decision functions. One possible approach is to consider a collection of $k$ binary classi£cation problems. $k$ classi£ers can then be constructed, one for each class. The $i^{th}$ classi£er constructs a hyper-plane between the class $i$ and the $k-1$ other classes. A majority vote across the classi£ers is then applied to classify a new example. Alternatively, $\frac{k(k-1)}{2}$ hyperplanes can be constructed, separating the classes from each other and similarly an appropriate voting scheme could be used. Clearly, a digit recognition system using this strategy needs building 45 different models, one for each pair of classes. This scheme was already used to solve multiclass recognition problems with linear decision functions as in the Ho-Kashyap classi£er. It is commonly referred as "Pairwise strategy" in contrast with the well-known "One Against Others strategy"[3, 4]. We tried KMOD as well as RBF kernel and a polynomial kernel on NIST database using both of the learning strategies. We used a subset of 20,000 images from the hsf_123 part for training and 10,000 images from the hsf_7 part for testing. From each image we extract 272 features that well characterize both local and morphological shapes present in the digit image [1]. These values are fed to every SVM model. During the "Pairwise strategy" of learning, 45 training processes are run to build the whole pairs' models. During classi£cation, we used an appropriate combination scheme that consists of £nding the class $k$ for which all the pairs' models $(k, j)$ with $0 \leq j \leq 9$ have a positive output. In the "One Against Others strategy", 10 different models are built, one for each class. We use a simple combination scheme given by $C_j = Arg\max_i(O_i)$; where $C_j$ is the resulting class label and $O_i$ is the $i^{th}$ svm output. The tested example will belong to the class for which the corresponding model output is maximal. No reject option was considered in this strategy. We report in Table 3 the best results using both of the learning strategies. We picked up empirically the corresponding kernel parameters using a £nite number of values to reduce the test error. Notice the increase of recognition rate near 1% for the "Pairwise strategy". KMOD does slighly better than other kernels in general. Moreover, in order to evaluate the signi£cance of the results we did a z-normal test between KMOD and other kernels' recognition rates. Contrary to the Student test, this one does not take into account the variability throughout different testing sets. We assume that 10,000 testing examples are suf£cient to pass over this condition. The plus sign beside the values in the table indicates whether or not the error is signi£cantly different from KMOD's one. Recall that, our kernel is signi£cantly better than RBF and polynomial kernel of degree 3 at a con£dence level of 0.05.

## 6 Conclusion

Whilst it is always possible to assume that the data fed into a SVM have bounded support, its sparseness inside the original space can vary widely, depending on its distribution, the feature extraction method and the dif£culty of the




$\langle a \rangle$ $\langle b \rangle$ $\langle c \rangle$

**Figure 3.** *$\langle a \rangle$: Mean of error vs. $\sigma$; $\langle b \rangle$: Mean of Support vectors number vs. $\sigma$; $\langle c \rangle$: Spread of error vs. $\sigma$*

| classi£er | KMOD svm | RBF svm | RBF network | Adaboost | Adaboost Reg |
|---|---|---|---|---|---|
| error | 25.4±4.4 | 26.0±4.7 | + 27.6±4.7 | + 30.4±4.7 | 26.5±4.5 |

**Table 2.** *Performance of KMOD (column 1) vs. other classi£ers on Breast Cancer database ([5])*

| strategy / kernel | one-against-others | pairwise |
|---|---|---|
| KMOD | 97.77 | 98.56 |
| polynomial (d=4) | 97.42 | 98.40 |
| polynomial (d=3) | + 96.77 | + 97.88 |
| RBF | + 96.91 | + 98.03 |

**Table 3.** *Recognition rates (in percentage) on NIST database*

problem on hand. We do believe that kernels preserving the all data on closeness information while still penalizing the far neighborhood are more reliable. KMOD is a two-parameter kernel that allows both of the characteristics through a varying speed of decay around zero and a moderate decrease toward in£nity. A spectral study of the behavior of KMOD has shown that for intermediate values of $\gamma$ we expect smaller generalization error. Experiments done on the UCI Breast Cancer database yield the best results yet established. Furthermore, we handle a large-scale classi£cation problem by exploring a digit recognition task which shows better performance for KMOD among most common kernels. Moreover, a procedure for automatic optimization of KMOD parameters has been devised, to ensure separation frontiers to £t the data more effectively, [2].

**Acknowledgments:** This research was supported by the NSERC of Canada and the FCAR program of the Ministry of Education of Quebec.

## References

[1] N.E. Ayat, M. Cheriet, and C.Y. Suen. Kmod- a new support vector machine kernel for pattern recognition. Application to digit image recognition. In *ICDAR*, pages 1215–1219, Seattle,USA, Sept. 2001.

[2] N.E. Ayat, M. Cheriet, and C.Y. Suen. Empirical error based optimization of SVM kernels. Application to digit image recognition. In the $8^{th}$ IWFHR, Niagara-on-the-lake, 2002.

[3] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classi£ers. In *Fifth Annual Workshop on Computational Learning Theory*, Pittsburg, 1992.

[4] U. Kre$\beta$el. Pairwise classi£cation and support vector machines. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 15, pages 255–268. 1999.

[5] G. R$\ddot{a}$tsch, T. Onoda, and K.-R. M$\ddot{u}$ller. Soft margins for adaboost. *Machine Learning*, 43(3):287–320, 2001.

[6] A. Smola. *Learning with Kernels*. PhD thesis, GMD First, Berlin, Germany, 1998.

[7] V. Vapnik. *The Nature of Statistical Learning Theory*. NY, USA, 1995.

[8] R. C. Williamson, A. J. Smola, and B. Scholkopf. Entropy numbers, operators and support vector kernels. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 9, pages 127–144. 1999.