CONDITIONALLY POSITIVE DEFINITE KERNELS FOR SVM BASED IMAGE RECOGNITION

Sabri Boughorbel, Jean-Philippe Tarel, Nozha Boujemaa

Sabri.Boughorbel@inria.fr, Tarel@lcpc.fr, Nozha.Boujemaa@inria.fr

ABSTRACT

Kernel based methods such as Support Vector Machine (SVM) have provided successful tools for solving many recognition problems. One of the reason of this success is the use of kernels. Positive definiteness has to be checked for kernels to be suitable for most of these methods. For instance for SVM, the use of a positive definite kernel insures that the optimized problem is convex and thus the obtained solution is unique. Alternative class of kernels called conditionally positive definite have been studied for a long time from the theoretical point of view and have drawn attention from the community only in the last decade. We propose a new kernel, named log kernel, which seems particularly interesting for images. Moreover, we prove that this new kernel is a conditionally positive definite kernel as well as the power kernel. Finally, we show from experimentations that using conditionally positive definite kernels allows us to outperform classical positive definite kernels.

1. INTRODUCTION

Support Vector Machine (SVM) [1] is one of the latest and most successful algorithm in computer vision. It is providing good solutions to many image recognition problems. SVM has a solid theoretical framework [2] which helps to analyze and understand why it works so well. The basic idea behind SVM is to build a classifier that maximizes the margin between positive and negative examples. Large margin classifiers have proved to yield to good generalization capacity which means a good ability to discover the true underlying data distribution. Formally, SVM algorithm boils down to minimize quadratic problem:

$$W(\boldsymbol{\alpha}) = -\sum_{i=1}^{\ell} \alpha_i + \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \qquad (1)$$

with respect to Kuhn-Tucker coefficients $\boldsymbol{\alpha}$, under the equilibrium constraint:

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0 \qquad (2)$$

Thanks to Muscle NoE for funding.

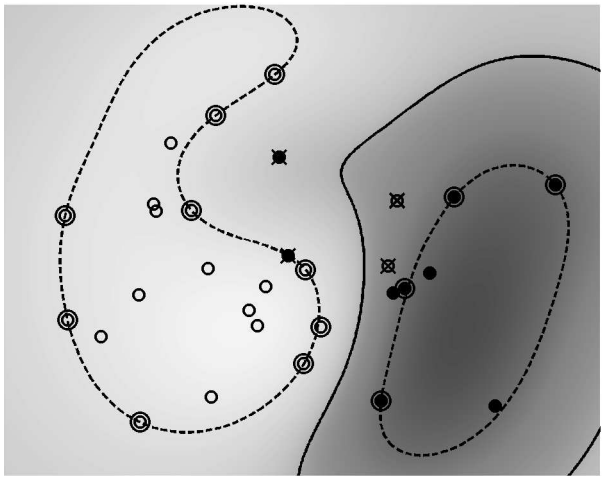

Fig. 1. Illustration of SVM recognition on a 2D toy problem in presence of outliers. Positive example are represented with white circles and negative example are represented with black filled circles. Solid line represents the decision boundary of the SVM classifier. Dotted lines depict the edge of the margin, support vectors are surrounded and misclassified examples are crossed.

where $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell) \in \mathcal{X} \times \{\pm 1\}$ is the training set. The SVM decision function $f(\boldsymbol{x}) \in \{\pm 1\}$ is expressed as:

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i \in SV} \alpha_i^0 y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b_0 \right) \qquad (3)$$

where $\boldsymbol{\alpha}_i^0$ are the optimal coefficients obtained by the minimization of (1). b_0 is the shift coefficient which can be computed with respect to the training set. SV is the set of indexes corresponding to non-zero α_i^0 since the Kuhn-Tucker condition have to be checked:

$$\alpha_i^0 [y_i (\sum_{j \in SV} \alpha_j^0 y_j K(\boldsymbol{x}_j, \boldsymbol{x}_i) + b_0) - 1] = 0.$$

Training examples corresponding to non-zero α_i are called support vectors.

As long as $K(\boldsymbol{x}, \boldsymbol{x}')$ is a positive definite kernel, $W(\boldsymbol{\alpha})$ is convex with respect to $\boldsymbol{\alpha}$. Therefore the minimization is

0-7803-9332-5/05/$20.00 ©2005 IEEE

achieved at a unique minimum. It has been proved that for any positive definite kernel, there exists a mapping function such as the kernel can be written as dot product, i.e. $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j)$. The function Φ usually maps feature space $\mathcal{X}$ into a high dimension space.

To illustrate this, we consider the well-known example of the polynomial kernel of degree 2 on $\mathbb{R}^2$:

$$\begin{aligned} K(\boldsymbol{x}, \boldsymbol{x}') &= (\boldsymbol{x} \cdot \boldsymbol{x}')^2 \\ &= (x_1 x_1' + x_2 x_2')^2 \\ &= \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x}') \end{aligned}$$

with $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$, $\Phi(\boldsymbol{x}) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)$. $K(\boldsymbol{x}, \boldsymbol{x}')$ is thus positive definite, since it can be written as a dot product on $\mathbb{R}^3$. Generally, the mapping Φ is used implicitly which means that the kernel computation does not require the explicit expression of Φ. Only definite positiveness of the kernel must be checked. This is called the kernel trick [3]. The SVM classifier is in fact only a linear classifier on the mapped space, see Fig. 2. Now, let us recall the

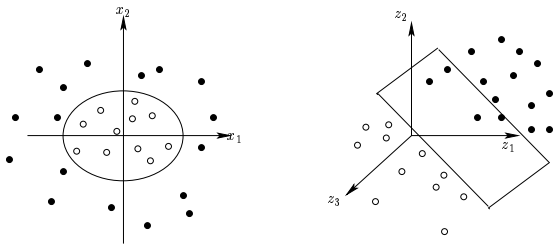

Fig. 2. 2D Data on the left is non linearly separable. Using the polynomial kernel of degree 2, data is mapped in $\mathbb{R}^3$ on the right. The SVM separates mapped positive and negative examples in $\mathbb{R}^3$ by hyperplane. This hyperplane corresponds to a curve in $\mathbb{R}^2$.

definition of a positive definite kernel:

Definition 1. Let $\mathcal{X}$ be a nonempty set. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *positive definite kernel* if and only if K is symmetric (i.e $K(\boldsymbol{x}, \boldsymbol{x}') = K(\boldsymbol{x}', \boldsymbol{x})$ for all $\boldsymbol{x},\ \boldsymbol{x}' \in \mathcal{X}$) and

$$\sum_{j,k=1}^{n} c_j c_k K(\boldsymbol{x}_j, \boldsymbol{x}_k) \geq 0,$$

for all $n \geq 1,\ c_1, \ldots, c_n \in \mathbb{R}$, and $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n \in \mathcal{X}$.

2. CONDITIONALLY POSITIVE DEFINITE KERNELS

We reviewed in the introduction how positive definite kernels are suitable for SVM. Looking at the equilibrium constraint (2), it is clear that the domain to which the vector $\boldsymbol{c}$ belongs can be restrained. This remark leads to define the family of conditionally positive definite kernels [4]. In the following, we review the main properties of this family and its connections with positive definite family.

Definition 2. Let $\mathcal{X}$ be a nonempty set. A kernel K is called *conditionally positive definite* if and only if it is symmetric and

$$\sum_{j,k=1}^{n} c_j c_k K(\boldsymbol{x}_j, \boldsymbol{x}_k) \geq 0$$

for $n \geq 1,\ c_1, \ldots, c_n \in \mathbb{R}$ with $\sum_{j=1}^{n} c_j = 0$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$.

We now give a basic example of conditionally positive definite kernel which will be of importance in the following, and we recall the proof of its definite positiveness from [4], page 69. **Example:**

$$\begin{aligned} K(\boldsymbol{x}, \boldsymbol{x}') &= -\|\boldsymbol{x} - \boldsymbol{x}'\|^2 \\ &= -\|\boldsymbol{x}\|^2 - \|\boldsymbol{x}'\|^2 + 2\langle \boldsymbol{x}, \boldsymbol{x}' \rangle \end{aligned}$$

Proof. *To check definition 2, we consider* $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subseteq \mathcal{X}$ *with* $\sum_{j=1}^{n} c_j = 0$*, and we have:*

$$\begin{aligned} \sum_{j,k=1}^{n} c_j c_k K(\boldsymbol{x}_j, \boldsymbol{x}_k) &= -\sum_{j,k=1}^{n} c_j c_k ||\boldsymbol{x}_j||^2 \\ &\quad - \sum_{j,k=1}^{n} c_j c_k ||\boldsymbol{x}_k||^2 + 2 \sum_{j,k=1}^{n} c_j c_k \langle \boldsymbol{x}_j, \boldsymbol{x}_k \rangle \\ &= -\underbrace{\sum_{k=1}^{n} c_k}_{=0} \sum_{j=1}^{n} c_j ||\boldsymbol{x}_j||^2 - \underbrace{\sum_{j=1}^{n} c_j}_{=0} \sum_{k=1}^{n} c_k ||\boldsymbol{x}_k||^2 \\ &\quad + 2 \sum_{j,k=1}^{n} c_j c_k \langle \boldsymbol{x}_j, \boldsymbol{x}_k \rangle = 2 \sum_{j,k=1}^{n} c_j c_k \langle \boldsymbol{x}_j, \boldsymbol{x}_k \rangle \geq 0 \end{aligned}$$

This proves that minus of the square of the Euclidean distance is conditionally positive definite. □

In [3], conditionally positive definite kernels are used with SVM algorithm. In the following, we explain why conditionally positive definite kernels are also suitable for SVM algorithm. We firstly recall, from [4] page 74, an important relation between positive definite and conditionally positive definite kernels:

Proposition 1. *Let K be a symmetric kernel on $\mathcal{X} \times \mathcal{X}$. Then for any $\boldsymbol{x}_0 \in \mathcal{X}$, we set*

$$\begin{aligned} \tilde{K}(\boldsymbol{x}, \boldsymbol{x}') &= \frac{1}{2}[K(\boldsymbol{x}, \boldsymbol{x}') - K(\boldsymbol{x}, \boldsymbol{x}_0) \\ &\quad - K(\boldsymbol{x}', \boldsymbol{x}_0) + K(\boldsymbol{x}_0, \boldsymbol{x}_0)] \end{aligned} \tag{4}$$

$\tilde{K}$ is positive definite if and only if K is conditionally positive definite.

This Proposition 1 presents a strong and interesting link between positive and conditionally positive definite families since it gives a necessary and sufficient condition. It can be used with advantages when designing new kernels for SVM.

From that, let us show how conditionally positive definite kernels are suitable for SVM. Assume that K and $\tilde{K}$ are two kernels satisfying the Proposition 1. We consider the SVM problem (1) using a conditionally positive definite kernel:

$$\widetilde{W}(\boldsymbol{\alpha}) = -\sum_{i=1}^{\ell} \alpha_i + \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j \tilde{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

under constraint, $\sum_{i=1}^{\ell} \alpha_i y_i = 0$. We now replace $\tilde{K}$ by its expression from (4). For any $\boldsymbol{x}_0 \in \mathcal{X}$, we thus have:

$$\begin{aligned} \tilde{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \frac{1}{2}[K(\boldsymbol{x}_i, \boldsymbol{x}_j) - K(\boldsymbol{x}_i, \boldsymbol{x}_0) \\ &\quad -K(\boldsymbol{x}_j, \boldsymbol{x}_0) + K(\boldsymbol{x}_0, \boldsymbol{x}_0)] \end{aligned}$$

Terms corresponding to $K(\boldsymbol{x}_i, \boldsymbol{x}_0)$, $K(\boldsymbol{x}_j, \boldsymbol{x}_0)$ and $K(\boldsymbol{x}_0, \boldsymbol{x}_0)$ vanish according to the constraint (2). Formally:

$$\begin{aligned} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_0) &= \underbrace{\sum_{j=1}^{\ell} \alpha_j y_j}_{=0} \sum_{i=1}^{\ell} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}_0) \\ &= 0. \end{aligned}$$

Similarly,

$$\sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_0, \boldsymbol{x}_j) = 0.$$

Therefore, only the term corresponding to $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ remains, and we obtain:

$$\widetilde{W}(\boldsymbol{\alpha}) = -\sum_{i=1}^{\ell} \alpha_i + \frac{1}{4} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) = 2W(\frac{1}{2}\boldsymbol{\alpha})$$

Hence $\widetilde{W}(\boldsymbol{\alpha})$ is rewritten with respect to the associated positive definite kernel K only. Thus with SVM, the use a conditionally positive definite kernel is equivalent to the use of the associated positive definite kernel. This also proves that conditionally positive definite kernels can be used for SVM algorithm.

3. POWER AND LOG KERNELS

In this section, we investigate properties of conditionally positive definite family. We focus on two particular kernels:

- Power distance kernel introduced first in [3]:

$$K_{\text{Power}}(\boldsymbol{x}, \boldsymbol{x}') = -\|\boldsymbol{x} - \boldsymbol{x}'\|^{\beta}. \tag{5}$$

  This kernel leads to scale invariant SVM classifier, as it can be shown by direct extension of [**?**].

- New kernel we named Log kernel:

$$K_{\text{Log}}(\boldsymbol{x}, \boldsymbol{x}') = -\log\left(1 + \|\boldsymbol{x} - \boldsymbol{x}'\|^{\beta}\right). \tag{6}$$

To prove that the above kernels are conditionally positive definite, we recall from [4] page 75 and 78:

Theorem 1. *Let $\mathcal{X}$ be a nonempty set and let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be symmetric. Then K is conditionally positive definite if and only if $\exp(uK)$ is positive definite for all $u > 0$.*

Proposition 2. *If $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is conditionally positive definite and satisfies $K(x, x) \leq 0$ for $\boldsymbol{x} \in \mathcal{X}$ then so it is of $-(-K)^{\beta}$ for $0 < \beta < 1$ and of $-\ln(1 - K)$.*

Proof. *Considering a conditionally positive definite kernel K such that $K(\boldsymbol{x}, \boldsymbol{y}) \leq 0$ for any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{X}$, then it follows from Theorem 1 that $\exp(uK)$ is positive definite, and thus conditionally positive definite. The constant kernel is obviously conditionally positive definite. Thus, $\exp(uK) - 1$ is conditionally positive definite as the sum of two conditionally positive definite kernels. For $0 < \beta < 1$ and $s < 0$, we can write:*

$$\begin{aligned} -(-s)^{\beta} &= \frac{\beta}{\Gamma(1-\beta)} \int_0^{+\infty} (e^{us} - 1) \frac{du}{u^{\beta+1}} \\ -\ln(1-s) &= \int_0^{+\infty} (e^{us} - 1) \frac{e^{-u}}{u} du \end{aligned}$$

Then , by replacing s with K, we can deduce that $-(-K)^{\beta}$ and $-\ln(1 - K)$ are conditionally positive definite kernels as a sum of conditionally positive definite ones. □

By applying the previous proposition to conditionally positive definite kernel $K(\boldsymbol{x}, \boldsymbol{x}') = -\|\boldsymbol{x} - \boldsymbol{x}'\|^2$ introduced in Sec. 2, we deduce that the Log (6) and Power distance (5) kernels are conditionally positive definite for $0 < \beta \leq 2$.

4. EXPERIMENTS

We compared performances of Power and Log kernels, for image recognition tasks. The tests have been carried on an image database containing 5 classes from Corel database, with an additional texture class of grasses, as shown in Fig. 3. Each class contains 100 images. Images are described using an RBG color histogram with a size of $4^3 = 64$ bins. A 3-fold cross validation is applied to estimate the errors rates. We considered the recognition problem of one class-vs-the others. Comparisons are performed with

Fig. 3. Each row presents one of the 6 classes used for experiments (castles, sun rises, grasses, mountains, birds, water falls).

respect to the following kernels: RBF kernel $K_{\text{RBF}}(x,y) = e^{-\|x-y\|^2/2\sigma^2}$, Laplace kernel [5] $K_{\text{Lap}}(x,y) = e^{-\|x-y\|/\sigma}$, Power and Log kernels. Tab. 1 summarizes the average performances of the different kernels. We tuned two parameters to obtain the best validation error: 1) the SVM regularization coefficient and the kernel hyper-parameter (β, σ, and d) (see Fig. 4). The Log and Power kernels lead to better performances than the other kernels. Tab. 2 presents the best class confusion obtained for the Log kernel. Sunrises, Grasses and Birds classes are well recognized. Some confusions appear between Castles, Mountains and Waterfalls classes due to the presence of similar colors.

Kernels	valid. error	test error
RBF	24.38±0.54	24.33±1.06
Laplace	23.5±0.82	23.66±0.89
Power	21.88±0.15	**21.44± 2.16**
Log	21.77±0.20	**21.12±1.70**

Table 1. Average validation and test errors for the different kernels.

	castle	sun	grass	mount	bird	water
castle	71	4	0	9	5	6
sun	5	84	0	0	0	0
grass	0	0	100	0	1	0
mount	12	6	0	72	3	19
bird	5	4	0	5	82	4
water	7	2	0	14	9	71

Table 2. Best class confusion matrix using the Log kernel.

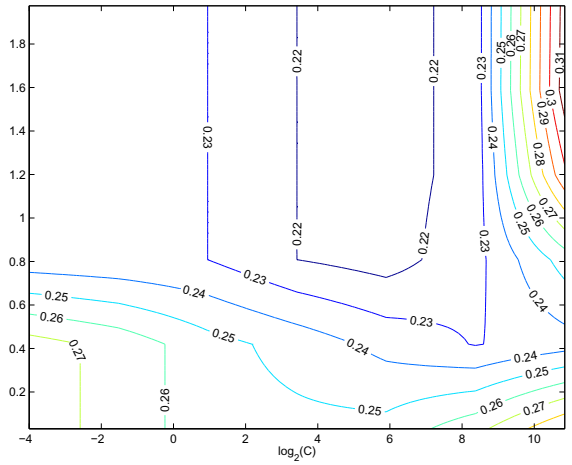

Fig. 4. Average validation error with respect to $\log_2(C)$ and β for Log kernel. C is the SVM regularization coefficient.

5. CONCLUSION

We have summarized, mainly from [4], several of the important properties of conditionally positive definite kernels. In particular, conditionally positive definite kernels have been proved to be suitable for SVM algorithm. Moreover, conditionally positive definite kernels have many interesting properties related with positive definite kernels. These properties provides very powerful tools to design both new conditionally positive definite kernels and new positive definite kernels. We proposed in particular a new kernel in the context of SVM, we named the Log kernel which seems to perform particularly well in our image recognition tests.

6. REFERENCES

[1] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[2] B. Scholkopf and A. Smola, *Learning with kernels*, MIT University Press Cambridge, 2002.

[3] B. Scholkopf, "The kernel trick for distances," in *NIPS*, 2000, pp. 301–307.

[4] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups: Theory of Positive Defi nite and Related Functions*, Springer-Verlag, 1984.

[5] O. Chapelle, P. Haffner, and V. Vapnik, "Svms for histogram-based image classifi cation," *IEEE Transactions on Neural Networks, 1999. special issue on Support Vectors*, 1999.