ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

A Class of Kernels for Sets of Vectors

Frédéric Desobry[1], Manuel Davy[2] and William J. Fitzgerald[1]

1- Signal Processing group, Engineering Department
Cambridge University, Trumpington street, CB2 1PZ Cambridge, UK

2- Laboratoire d'Automatique, de Génie Informatique et Signal, CNRS
BP48, cité scientifique, 59651 Villeneuve d'Ascq cedex, France

Abstract. In some important applications such as speaker recognition or image texture classification, the data to be processed are sets of vectors. As opposed to standard settings where the data are individual vectors, it is difficult to design a reliable kernel between sets of vectors of possibly different cardinality. In this paper, we build kernels between sets of vectors from probability density functions level sets estimated for each set of vectors, where a pdf level set is (roughly) a part of the space where most of the data lie.

1 Introduction

Various practical situations involve the comparison of two sets of vectors. A first usual approach consists of comparing each vector in one set to each vector in the other set using a kernel, leading to various algorithms such as the two class classification support vector machines. In a second approach, each of the two sets of vectors is seen as a single data point, and a higher level kernel is designed so as to compare the two sets. In the first situation, the similarity of the two sets is summarized in the so-called *kernel matrix*, whereas in the second situation, the similarity is summarized by a single real value. Many applications require such high level kernel defined on sets of vectors. Important examples are speaker recognition (where one speaker is represented by a set of vectors in the 20-30 dimensional space of cepstral coefficients) or image texture classification.

Such applications share several important properties. Firstly, all vectors in the two sets lie in the same space (denoted $\mathcal{X}$). Secondly, there may not be the same number of vectors in the two sets. In the following, we denote by $\boldsymbol{x} = \{x_1, \dots x_m\}$ with size m and $\boldsymbol{x'} = \{x'_1, \dots x'_{m'}\}$ with size m' the two sets to be compared using a kernel denoted $K(\boldsymbol{x}, \boldsymbol{x'})$.

Several previous approaches have been proposed to build the kernel $K(\cdot, \cdot)$. The most popular one consists of 1) estimating the probability density functions (pdfs) according to which $\boldsymbol{x}$ and $\boldsymbol{x'}$ are distributed, and 2) applying to them a kernel defined on the space of such pdfs. More verbosely, one may estimate, say, Gaussian mixtures from the set of vectors, and then derive a kernel based on a divergence measure or a likelihood ratio between the estimated pdfs. These approaches suffer however from several drawbacks. In particular, estimating pdfs in a space $\mathcal{X}$ of large dimension is tough, especially when the number of data (m or m') is smaller than the dimension of $\mathcal{X}$. A usual (unsatisfactory) solution consists of replacing pdfs by histograms and building the kernel from the similarity measures based on histograms [1]. It may be emphasized that

numerical computation of similarity measures between pdfs can be difficult when the pdfs do not belong to the exponential family. Another argument may be that if underlying pdfs can be approximated well enough, a more natural way to use them would be to adopt a Bayesian framework.

In this paper, we propose to derive a kernel $K(\cdot,\cdot)$ between sets of vectors based on level sets[1] of the underlying pdfs. Briefly, this work builds on the following idea: two sets of vectors are essentially similar if they occupy the same part of the space $\mathcal{X}$. In other words, the similarity between the sets $\boldsymbol{x}$ and $\boldsymbol{x'}$ can be evaluated as the similarity between level sets estimated from the vectors in $\boldsymbol{x}$ and $\boldsymbol{x'}$. This approach, first proposed in [2] as a contrast function between sets of vectors, does not require the estimation of densities in possibly large dimension spaces.

This paper is organized as follows. Section 2 recalls some necessary elements about level sets and presents a level set estimation algorithm (one-class Support Vector Machine). Section 3 proposes some metrics between sets of vectors. These metrics are built using estimated level sets, and they are computed extrinsically in feature space. They will be used in Section 4 to derive kernels $K(\cdot,\cdot)$. In Section 5, we discuss various aspects of the proposed kernels $K(\cdot,\cdot)$. Section 6 presents some simulation results and conclusions.

2 Level sets of probability density functions

Let $p(\cdot)$ a pdf over $\mathcal{X}$. The ϵ-level set of $p(\cdot)$ is the subset S of $\mathcal{X}$ defined as $S = \{x \in \mathcal{X} \text{ s.t. } p(x) \geq \epsilon\}$. As $p(\cdot)$ is usually unknown, it is more convenient to define (equivalently) the level set S in terms of the distribution P which $p(\cdot)$ derives from: S is the minimum (Lebesgue) measure subset of $\mathcal{X}$ with P-measure $1-\nu$, where ϵ and ν are of course related. This definition makes easier the estimation of S from a set $\boldsymbol{x} = \{x_1, \ldots x_m\}$ as, asymptotically with m, one only needs to counts how many of the x_i's actually are in S. Of course, for purposes of decision involving two, or more, level sets, these have to be defined with fixed ν, not for fixed ϵ.

As P is unknown, one needs to estimate the set S; we denote S_m an estimate of S based on the learning set $\boldsymbol{x} = \{x_1, \ldots, x_m\}$, where the samples x_i's are distributed i.i.d. according to P. Using a pdf $p_m(\cdot)$ estimated from $\boldsymbol{x}$ so as to define S_m as $\{x \in \mathcal{X}, p_m(x) \geq \epsilon\}$ is irrelevant insofar as the density estimation step is precisely what we want to avoid when comparing the sets $\boldsymbol{x}$ and $\boldsymbol{x'}$. A classical approach is then to select S_m in a predefined class of sets (balls, ellipsoids, convexes), or, alternatively, to define its boundary as $\{x \in \mathcal{X}, f_m(x) = 0\}$ and select f_m in a class of functions (piecewise polynomials, poor classes in reproducing kernel Hilbert spaces).

A suitable level-set estimation technique must be (strongly) consistent, achieve fast rates of convergence (if possible, independent of the dimension of $\mathcal{X}$) and lead to a practicable and computationally cheap algorithm. As our final objective is to build $K(\boldsymbol{x}, \boldsymbol{x'})$ as a kernel between estimated level sets S_m and S'_m, we

[1] A level set of a pdf can be seen as the part of the space that contains a given fraction of the probability mass indicated by the pdf. This is more precisely defined in Section 2.

need to be able to define and compute a dissimilarity measure between S_m and S'_m. These sets being subsets of $\mathcal{X}$ in possibly large dimension whose similarity (based, say, on their intersection) can be extremely difficult to compute.

In the following, we show that the reproducing kernel Hilbert space (r.k.h.s.) framework enables an easy and aesthetic solution to these problems. We briefly recall that, in the context we address, a r.k.h.s. $\mathcal{H}$ is a vector space of real-valued functions defined over $\mathcal{X}$, in which the evaluation functional is a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the reproducing property $\langle f(\cdot), k(x, \cdot)\rangle_{\mathcal{H}} = f(x), \forall f \in \mathcal{H}, \forall x \in \mathcal{X}$. We assume that k is such that $k(x, x) = 1 \,\forall\, x \in \mathcal{X}$. In $\mathcal{X}$, we select the boundary of the estimated level set S_m (which is equivalent to considering S_m itself) such that it follows $\{x \in \mathcal{X}, f_m(x) = 0\}$, where f_m is a function in $\mathcal{H}$ obtained as the solution of the optimization problem:

$$\min_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} c(x_i, f(x_i)) + \lambda \|f\|^2_{\mathcal{H}} \tag{1}$$

where c is a cost function such as the hinge loss. Examples of such boundary estimators from a function in $\mathcal{H}$ are one-class Support Vector Machines (SVMs), see [2] and references therein. In this framework, the P-measure $(1 - \nu)$ is directly related to the regularization parameter λ of the optimization problem. Whatever the framework adopted, solving (1) implies that Wahba's representer theorem holds, that is:

$$f_m(x) = \sum_{i=1}^{m} \alpha_i k(x, x_i) - b \text{ with the } \alpha_i\text{'s and } b \text{ in } \mathbb{R} \tag{2}$$

At this step, we have characterized (independently) $\boldsymbol{x}$ and $\boldsymbol{x'}$ by estimated level sets S_m and S'_m of underlying densities, rather than by the densities themselves. Strong motivation about the soundness is provided in [2]; briefly, this is mainly because estimating a level set is easier than estimating the pdf. However, we now need to be able to compute a dissimilarity measure between these sets, which is addressed in next section. In the remainder of this article, we further assume that k is such that the kernel matrix between x_i's $(i = 1, \ldots, m)$ is with rank m whatever m is[2].

3 Metrics and other dissimilarity measures between level sets

In this section, we build dissimilarity measures (e.g., metrics), between the two subsets of $\mathcal{X}$ respectively defined as $S_m = \{x \in \mathcal{X}, \sum_{i=1}^{m} \alpha_i k(x, x_i) - b_m \geq 0\}$ and $S'_m = \{x \in \mathcal{X}, \sum_{i=1}^{m'} \alpha'_i k(x, x'_i) - b'_{m'} \geq 0\}$. The main difficulty is, however, that S_m and S'_m are, say, unions of connected subsets of $\mathcal{X}$, which makes extremely difficult the computation of their dissimilarity[3]. This difficulty can be overcome

[2]This assumption is made only to ensure a simpler presentation, and the results presented here hold for more general kernels.

[3]We do not consider the dissimilarity $d(S, S') = \frac{1}{m+m'} \#(x \notin S' \cup x' \notin S)$ because it is too poor in most situations.

by computing dissimilarity between p_m and S'_m in $\mathcal{H}$ rather than in $\mathcal{X}$. We denote $\mathbf{S}_m$ and $\mathbf{S'}_m$ the respective images of S_m and S'_m in $\mathcal{H}$.

In $\mathcal{H}$, all $k(x_i,\cdot)$'s and $k(x'_i,\cdot)$'s lie on a sphere with radius one as $\|k(x_i,\cdot)\|^2_{\mathcal{H}} = \langle k(x_i,\cdot), k(x_i,\cdot)\rangle_{\mathcal{H}} = k(x_i,x_i) = 1$. This remark together with the reproducing property enables easy computation of dissimilarity measures between $\mathbf{S}_m$ and $\mathbf{S'}_m$. In the following of this section, we focus on three dissimilarity measures, namely the symmetric difference measure $d_1(\cdot,\cdot)$, Hausdorff measure $d_\infty(\cdot,\cdot)$ and a contrast measure $d_c(\cdot,\cdot)$. Both $d_1(\cdot,\cdot)$ and $d_\infty(\cdot,\cdot)$ are metrics if defined over compact sets.

Most of the following calculations are geometrically derived from Fig. 1, which is a simplified two-dimension view of $\mathbf{S}_m$ and $\mathbf{S'}_m$ in $\mathcal{H}$. In particular,

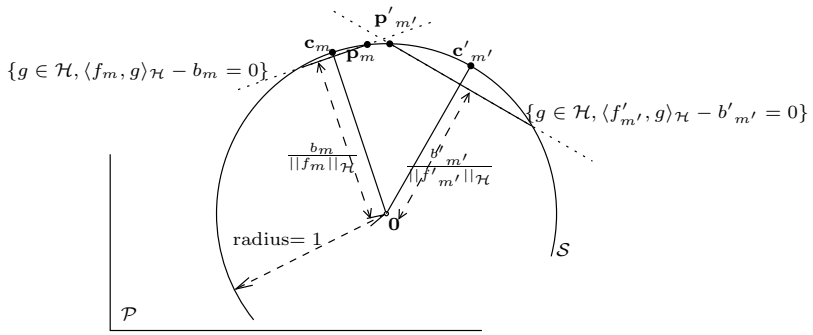

Fig. 1: Estimation of the level sets $\mathbf{S}_m$ and $\mathbf{S'}_m$ in $\mathcal{H}$ using ν-one class support vector machines.

most derivations arise from the fact that the arc distances $d_{\text{arc}}(\mathbf{c}_m, \mathbf{c'}_{m'})$ and $d_{\text{arc}}(\mathbf{c}_m, \mathbf{p}_m)$ can be computed in input space, see [2].

• *Symmetric difference measure* $d_1(\cdot,\cdot)$*:* This metric is defined as $d_1(\mathbf{S}_m, \mathbf{S'}_m) = \mu\left((\mathbf{S}_m \backslash \mathbf{S'}_m) \cup (\mathbf{S'}_m \backslash \mathbf{S}_m)\right)$, thus it can be evaluated by computing a volume in $\mathcal{H}$. Though $\mathcal{H}$ is infinite-dimensional, the $k(x_i,\cdot)$'s and $k(x'_i,\cdot)$'s actually lie in its subspace spanned by $\{k(x_1,\cdot),\ldots,k(x_m,\cdot),k(x'_1,\cdot),\ldots,k(x'_{m'},\cdot)\}$ where Lebesgue measure is properly defined. As a result, the volume is proportional to, e.g., when $\mathbf{c}_m$, $\mathbf{p'}_{m'}$, $\mathbf{p}_m$ and $\mathbf{c'}_{m'}$ are ordered : $\left(d_{\text{arc}}(\mathbf{c}_m, \mathbf{p'}_{m'}) + d_{\text{arc}}(\mathbf{c'}_{m'}, \mathbf{p}_m)\right)^{(m+m')}$. Embedding in a subspace of dimension $(m+m')$ is only a matter of convenience, to define the measure properly: whatever m and m', what we compare are the arc distances, therefore we may consider instead:

$$\left(d_{\text{arc}}(\mathbf{c}_m, \mathbf{p'}_{m'}) + d_{\text{arc}}(\mathbf{c'}_{m'}, \mathbf{p}_m)\right) \tag{3}$$

which makes it possible to compare sets of different cardinality without any further normalization. The other relative positions of the points $\mathbf{c}_m$, $\mathbf{p'}_{m'}$, $\mathbf{p}_m$ and $\mathbf{c'}_{m'}$ (which can easily be determined) lead to similar expressions.

• *Hausdorff measures $d_\infty(\cdot,\cdot)$:* It is defined as follows:

$$\begin{aligned} d_\infty(\mathbf{S}_m, \mathbf{S'}_m) &= \max\left(\max_{g\in\mathbf{S}_m}\min_{g'\in\mathbf{S'}_m}\|g-g'\|_{\mathcal{H}}, \max_{g'\in\mathbf{S'}_m}\min_{g\in\mathbf{S}_m}\|g-g'\|_{\mathcal{H}}\right) \\ &= \max\left(d_{\text{arc}}(\mathbf{c}_m,\mathbf{p}_m) + d_{\text{arc}}(\mathbf{c}_m,\mathbf{p'}_{m'}), \mathbf{c'}_{m'},\mathbf{p'}_{m'}) + d_{\text{arc}}(\mathbf{c'}_{m'},\mathbf{p}_m)\right) \end{aligned}$$

where the second equality is again derived for the same relative position of $\mathbf{c}_m$, $\mathbf{p'}_{m'}$, $\mathbf{p}_m$ and $\mathbf{c'}_{m'}$.

• *Contrast measure $d_c(\cdot,\cdot)$:* Finally, we also use a contrast measure between $\mathbf{S}$ and $\mathbf{S'}$, defined as:

$$d_c(\mathbf{S},\mathbf{S'}) = \frac{d_{\text{arc}}(\mathbf{c}_m,\mathbf{c'}_{m'})^2}{d_{\text{arc}}(\mathbf{c}_m,\mathbf{p}_m)^2 + d_{\text{arc}}(\mathbf{c'}_{m'},\mathbf{p'}_{m'})^2} \tag{4}$$

This contrast function was first introduced in [2] to achieve change detection.

4 From dissimilarity measures to kernels

Using exponentiated dissimilarity measures as, e.g., in [3, 4], we define a kernel between sets of samples as a kernel between level set of their underlying distribution:

$$k(S,S') = \exp\left(-\frac{d(S,S')^2}{2\sigma^2}\right) \tag{5}$$

with $d(\cdot,\cdot)$ any dissimilarity measure between sets such as, e.g., $d_1(\cdot,\cdot)$, $d_\infty(\cdot,\cdot)$ or $d_c(\cdot,\cdot)$. Eq. (5) clearly defines a symmetric positive similarity measure between level sets. The remainder of this section deals with positive definiteness of those kernels.

Symmetric difference measure. Rewriting $d_1(S,S')$ as: $d_1(G,G') = \mu\left((G\backslash G')\cup(G'\backslash G)\right) = \int_{x\in\mathcal{X}}(\mathbb{1}_G(x)-\mathbb{1}_{G'}(x))^2 d\mu(x) = \|\mathbb{1}_G(x)-\mathbb{1}_{G'}(x)\|^2_{L^2(\mathcal{X})}$ makes it possible to conclude (by [5]) of the positive definiteness of the kernel built on $d_1(\cdot,\cdot)$.

Hausdorff and contrast measures. We were not able to conclude w.r.t. to positive definiteness of the kernels based on $d_\infty(\cdot,\cdot)$ and $d_c(\cdot,\cdot)$, however developments in [6] makes it possible to use such kernels with corresponding theoretical framework. Practical results further confirm this.

5 Discussion

Pdf level sets based decision. The idea of plugging in pdf level sets instead of the pdf themselves is not new. It was proposed as a theoretical approach in [7] for outlier detection (with density support estimates using balls), and in [2] for a practical and efficient change detection approach. Theoretical argument for this choice is related to the corresponding rates of convergence (see [8]).

Other approaches. Many approaches propose a kernel between densities. However, computing a dissimilarity between densities can somewhat be very difficult. As a result, authors are force to use very simple density estimates (histograms in [1]), or make unnecessary crude assumptions such as gaussianity in $\mathcal{H}$ (see, e.g., [10]).

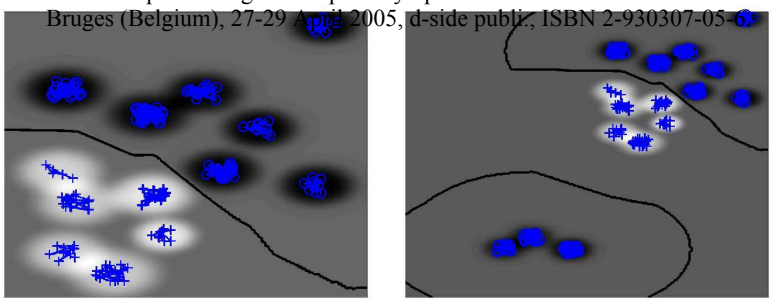

Fig. 2: Two-class SVM classification between sets of 2D points with different size, using the kernel built on d_c. Toy examples correspond to linearly (left) and non-linearly separable situations.

6 Toy examples and Conclusion

In figure 2, we illustrate the use of a pdf level-set based kernel between sets of points. The data are sets of 2D points with different size : a SV level-set estimator is first trained *independently* on each of them. These level sets are the regions in white for one class and in black for the other. They are then used as the inputs of a classic two-class SVM which yields the frontier plotted in black.

Further work include extensive simulations to compare the efficiency of the proposed kernels wrt density-based kernels such as those described in, e.g., [10, 1]. Relevant fields applications includes image processing tasks such as object recognition, and speech processing.

References

[1] M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In *AISTATS 2005*, page 8, January 2005.

[2] F. Desobry, M. Davy, and C. Doncarli. An on-line kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 2004. to appear.

[3] J. Lafferty and G. Lebanon. Diffusion kernels on statistical manifolds. Technical Report CMU-CS-04-101, School of Computer Science, 2004.

[4] B. Haasdonk and C. Bahlmann. Learning with distance substitution kernels. *Pattern Recognition - Proc. of the 26th DAGM Symposium, Tübingen, Germany, August/September 2004*, 2004.

[5] I.J. Schoenberg. Metric spaces and completely monotone functions. *The Annals of Mathematics*, 39(4):811–841, October 1938.

[6] C.S. Ong, X. Mary, S. Canu, and A.J. Smola. Learning with non-positive kernels. In *Proceedings of the 21st International Conference on Machine Learning*, pages 639–646, 2004.

[7] L. Devroye and G.L. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488, June 1980.

[8] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.

[9] R. Kondor and T. Jebara. A kernel between sets of vectors. In *Proceedings of the ICML*, 2003.