Consolidation Kernel

Aya El Dakdouki [a] and Yann Guermeur [b] and Nicolas Wicker [c,*]

[a]Jean Kuntzmann Laboratory (LJK), UFR Mathematics, Grenoble Alpes University, 38400 Saint-Martin-d'Hères, France [b]LORIA-CNRS, Campus Scientifique, BP239, 54506 Vandœuvre-lès-Nancy Cedex, France [c]Paul Painlevé Laboratory, Dept. of Mathematics, University of Lille, 59650 Lille, France

ARTICLE HISTORY
Compiled February 18, 2024

ABSTRACT
This paper introduces a new kernel for pattern classification. The *consolidation kernel* is designed to deal with a topological difficulty: a data set where some of the clouds of points associated with the different categories are parted in multiple clusters, possibly distant one from the other. It brings together such clusters. It is incorporated in a multi-class support vector machine. A comparative experimental study hightlights its appealing properties.

KEYWORDS
Pattern recognition and Translation invariance and Kernel engineering and K-means clustering and Multi-class support vector machines.

1. Introduction

Over the last three decades, both the theory and the practice of pattern classification have made rapid strides. Their joint progress is nicely illustrated by the solutions developed to take into account basic features of the data distribution. For instance, efficient methods are already available to deal with data endowed with a structure [Breiman, 2001], the manifold hypothesis [Pope et al., 2021], or exploit isotropy [Tropp, 2015].

In that context, kernel machines [Schölkopf and Smola, 2002, Hofmann et al., 2008] appear as models of choice. They benefit from the extensions that have made them capable of computing directly polytomies, i.e., performing pattern classification with a finite set of categories. An example is provided by the regularized kernel discriminant analysis (RKDA) [Ye et al., 2008]. However, the most popular family of multi-class kernel machines is the one of multi-class support vector machines (M-SVMs) [see Guermeur, 2012, Doğan et al., 2016, for a survey]. Those machines can be adapted to the specificities of the data through the choice of the kernel. This article introduces a new kernel designed to deal with situations where the description space $\mathcal{X}$ is included in the Euclidean space $\mathbb{R}^p$ and disconnected part of it, possibly distant, contain points with the same label. The *consolidation kernel* could be used as a kind of translation invariant kernel which at the same time does not allow the translation to make closer points of different categories. This behaviour is obtained by inferring information on the structure of the data (class conditional distributions) through clusterings performed category by category on a set of labelled points.

The rest of the paper is organized as follows. In the next section, we review existing works dealing with transformation invariant kernels. Section 3 introduces our new kernel. Its evaluation in the framework of a comparative study is the subject of Section 4. At last, we draw conclusions in Section 5.

2. State of the Art on the Transformation Invariant Kernels

Kernels have been designed for a variety of data: graphs [Kondor and Lafferty, 2002, Gärtner et al., 2003, Smola and Kondor, 2003], strings [Lodhi et al., 2002, Joachims, 1998, Salton et al., 1975] and of course images [Decoste and Schölkopf, 2002]. For a good review, we can refer to [Schölkopf and Smola, 2002]. When it comes to

*Corresponding author (nicolas.wicker@univ-lille.fr)

transformation invariance, the simplest idea is based on the generation of virtual examples [Poggio and Vetter, 1992, Niyogi et al., 1998]. In this approach, new examples are created using the transformation of interest (translation or rotation for example) to enlarge the training set. A variant of it, which applies to the computation of dichotomies only, is the *virtual support vector* method [Schölkopf et al., 1996]. There, the virtual examples are only generated from the support vectors (that utterly define the boundaries between the categories). The drawback is the enlarged memory and time complexities due to additionnal points. Very close kernels formalizing the idea of virtual support vectors are the *jittering kernels* [Decoste and Schölkopf, 2002, Schölkopf and Smola, 2002], where the transformation invariance is in the kernel itself. For instance, $\kappa^*(\mathbf{x}, \mathbf{x}')$ may be computed from a kernel κ using $T^* = \operatorname{argmin}_{T \in \mathcal{T}} \{\kappa(\mathbf{x}, \mathbf{x}) + \kappa(T\mathbf{x}', T\mathbf{x}') - 2\kappa(\mathbf{x}, T\mathbf{x}')\}$, where $\mathcal{T}$ is a transformation group and $\kappa^*(\mathbf{x}, \mathbf{x}')$ is equal to $\kappa(\mathbf{x}, T^*\mathbf{x}')$. Simard et al. [1998] introduced the *tangent distance* to incorporate a priori knowledge, including transformation invariances, into the distance measure. This distance was then incorporated in kernels by Haasdonk and Keysers [2002]. All these kernels can be generalized by computing an average kernel over any group of transformations. This gives rise to the *Haar-integration kernel* [Schulz-Mirbach, 1994, Haasdonk et al., 2005] defined for a standard kernel κ_0 and a transformation group $\mathcal{T}$, which contains the admissible transformations [see Schulz-Mirbach, 1994, for the formal definition]. The idea is to compute the average of the kernel output $\kappa_0(T\mathbf{x}, T'\mathbf{x}')$ over all pairwise combinations of the transformed examples $(T\mathbf{x}, T'\mathbf{x}'), \forall (T, T') \in \mathcal{T}^2$. The Haar-integration kernel κ of κ_0 with respect to $\mathcal{T}$ is thus

$$\kappa(\mathbf{x}, \mathbf{x}') = \int_{\mathcal{T}^2} \kappa_0(T\mathbf{x}, T'\mathbf{x}') \, \mathrm{d}T \mathrm{d}T',$$

under the condition of existence of the integral.

These kernels have sound foundations but lack flexibility for cases where there are no straightforward transformations to exploit. Besides, they do not depend on the nature of the task (supervised learning in our case). In the following section, we introduce a similar kernel, more flexible in that it depends on the learning task and captures explicitly the properties of translation invariance exhibited by the different categories.

3. Consolidation Kernel

The present work is inspired by our practice of (supervised) classification: the set of point associated with a category can be separated into several clusters for many complex reasons. The idea is here to design a kernel bringing together these clusters while keeping away clusters from different categories. It is implemented in the following way. Let $\mathcal{Y}$ denote the set of categories and let $s_m = \{(\mathbf{x}_i, y_i) : 1 \leqslant i \leqslant m\} \subset \mathcal{X} \times \mathcal{Y}$ be a set of labelled points. First, the subsets of s_m associated with the different categories are fragmented into a number of relevant clusters (by means of a clustering method). Second, a set of directions $\{\mathbf{c}_{i,2} - \mathbf{c}_{i,1} : i \in [\![1; M]\!]\}$ is obtained by application of two rules:

(1) $\mathbf{c}_{i,1}$ and $\mathbf{c}_{i,2}$ are prototypes of clusters associated with the same category;
(2) the vector $\mathbf{c}_{i,2} - \mathbf{c}_{i,1}$ does not connect two clusters associated with different categories.

Along each direction $\mathbf{c}_{i,2} - \mathbf{c}_{i,1}$, we want the kernel value to oscillate somehow according to the periodic function h_{d_i} with $d_i = \|\mathbf{c}_{i,2} - \mathbf{c}_{i,1}\|_2$ defined on $\mathbb{R}$ as follows:

$$\forall k \in \mathbb{Z}, \ \forall t \in [0, d_i), \quad h_{d_i}(kd_i + t) = \frac{4}{d_i^2} t^2 - \frac{4}{d_i} t + 1,$$

and depicted in Figure 1.

This function could be replaced by any similar d_i-periodic function with maximal value at 0, decreasing on $[0, d_i/2]$ and increasing on $[d_i/2, d_i]$. The purpose of this behaviour is to take into account the lengths of the admissible translations. With functions h_{d_i} at hand, the consolidation kernel can be defined in the following way.

Definition 1. *Let s_m be a set of labelled examples and $\{\mathbf{c}_{i,2} - \mathbf{c}_{i,1} : i \in [\![1; M]\!]\}$ the corresponding set of directions produced by the clustering method. The* consolidation kernel *κ, parameterized by $\lambda \in (0, 1)$, $\boldsymbol{\sigma} = (\sigma_i)_{0 \leqslant i \leqslant M} \in (\mathbb{R}_+^*)^{M+1}$ and $\boldsymbol{\tau} = (\tau_i)_{1 \leqslant i \leqslant M} \in \mathbb{R}_+^M$, is defined by:*

$$\forall (\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2, \ \kappa(\mathbf{x}, \mathbf{x}') = (1 - \lambda) \exp\left\{-\frac{1}{2} \frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{\sigma_0^2}\right\} + \lambda \sum_{i=1}^{M} \tau_i h_{d_i}\left(\langle \boldsymbol{\mu}_i, \mathbf{x} - \mathbf{x}' \rangle_2\right) \exp\left\{-\frac{1}{2} \frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{\sigma_i^2}\right\}$$

where $(\boldsymbol{\mu}_i)_{1 \leqslant i \leqslant M} = \left(\frac{1}{d_i}(\mathbf{c}_{i,2} - \mathbf{c}_{i,1})\right)_{1 \leqslant i \leqslant M}$.

The general idea of this definition, in line with those of the state-of-the-art transformation invariant kernels, is that when computing the similarity between two points, not only should their distance be taken into account

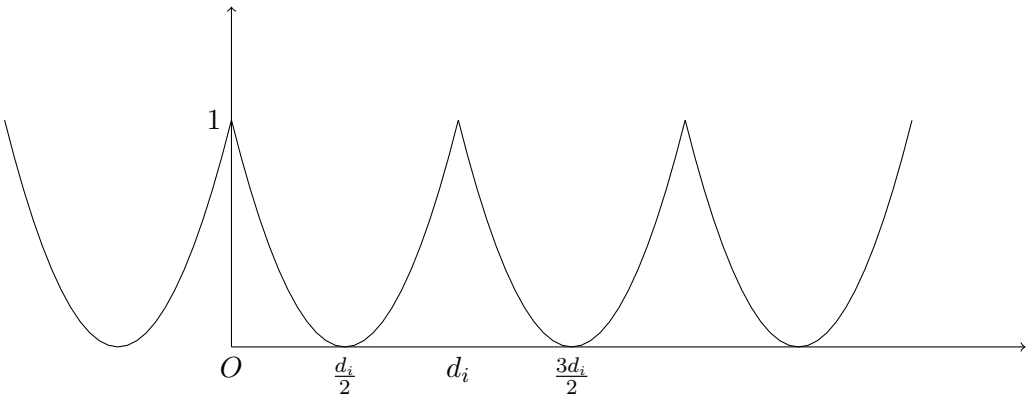


Figure 1.: Graph of function h_{d_i}.

but also other terms characterizing the data regularities. To prove that κ is a valid kernel, it suffices to replace the functions h_{d_i} with their Fourier expansions, giving:

$$\begin{aligned}\forall \left(\mathbf{x}, \mathbf{x}'\right) \in \mathcal{X}^2,\ \kappa\left(\mathbf{x}, \mathbf{x}'\right) &= (1-\lambda) \exp\left\{-\frac{1}{2}\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{\sigma_0^2}\right\} + \lambda \sum_{i=1}^{M} \tau_i \sum_{j=0}^{\infty} a_j \cos\left(2j\pi \left\langle \boldsymbol{\mu}_i, \mathbf{x}-\mathbf{x}' \right\rangle_2\right) \exp\left\{-\frac{1}{2}\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{\sigma_i^2}\right\} \\ &= (1-\lambda) \exp\left\{-\frac{1}{2}\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{\sigma_0^2}\right\} + \lambda \sum_{j=0}^{\infty} \sum_{i=1}^{M} \tau_i a_j \cos\left(2j\pi \left\langle \boldsymbol{\mu}_i, \mathbf{x}-\mathbf{x}' \right\rangle_2\right) \exp\left\{-\frac{1}{2}\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{\sigma_i^2}\right\},\end{aligned}$$

where $a_0 = \frac{1}{3}$ and $(a_j)_{1\leqslant j\leqslant\infty} = \left(\frac{4}{j^2\pi^2}\right)_{1\leqslant j\leqslant\infty}$. This alternative expression of the function makes simpler the comparison with the well-known spectral mixture (SM) kernel [Wilson and Adams, 2013] given by:

$$\forall \left(\mathbf{x}, \mathbf{x}'\right) \in \mathcal{X}^2,\ \kappa_{\mathrm{SM}}\left(\mathbf{x}, \mathbf{x}'\right) = \sum_{q=1}^{Q} a_q \frac{|\Sigma_q|^{1/2}}{(2\pi)^{p/2}} \cos\left(2\pi \left\langle \boldsymbol{\mu}_q, \mathbf{x}-\mathbf{x}' \right\rangle_2\right) \exp\left\{-\frac{1}{2}\left\|\Sigma_q^{1/2}\left(\mathbf{x}-\mathbf{x}'\right)\right\|_2^2\right\},$$

where the parameters $\boldsymbol{\theta} = \left\{a_q, \Sigma_q, \boldsymbol{\mu}_q\right\}$ are mixture weights, bandwidths and frequencies. The kernel κ appears as the convex combination of a Gaussian kernel and an infinite weighted sum of SM kernels. Since all the weights are positive, then according to Proposition 13.1 in Schölkopf and Smola [2002], κ is also a kernel. The SM kernel can discover quasi-periodic stationary structures. Our kernel is an extension of the Gaussian kernel that focuses on one kind of structure: translation invariance.

Next proposition illustrates the effect of the additional terms by showing that we can get $\kappa\left(\mathbf{x}'', \mathbf{x}\right) > \kappa\left(\mathbf{x}'', \mathbf{x}'\right)$ even though $\|\mathbf{x}''-\mathbf{x}\|_2 > \|\mathbf{x}''-\mathbf{x}'\|_2$.

Proposition 2 (Property of kernel κ)**.** *Suppose that κ is parameterized as follows. There exists $\sigma \in \mathbb{R}_+^*$ such that $\boldsymbol{\sigma} = \sigma \mathbf{1}_{M+1}$ and $\boldsymbol{\tau} = \frac{1}{M}\mathbf{1}_M$. Let $\mathbf{x}$, $\mathbf{x}'$ and $\mathbf{x}''$ be three points in $\mathcal{X}$ such that for every i in $[\![1; M]\!]$, $\langle \boldsymbol{\mu}_i, \mathbf{x}''-\mathbf{x}\rangle_2 \in d_i\mathbb{Z}$ and $\langle \boldsymbol{\mu}_i, \mathbf{x}''-\mathbf{x}'\rangle_2 \in \frac{d_i}{2} + d_i\mathbb{Z}$. Then $\|\mathbf{x}''-\mathbf{x}'\|_2^2 \in \left(\max\left\{0, \|\mathbf{x}''-\mathbf{x}\|_2^2 + 2\sigma^2 \ln(1-\lambda)\right\}, \|\mathbf{x}''-\mathbf{x}\|_2^2\right)$ implies that $\kappa\left(\mathbf{x}'', \mathbf{x}\right) > \kappa\left(\mathbf{x}'', \mathbf{x}'\right)$ although $\|\mathbf{x}''-\mathbf{x}\|_2 > \|\mathbf{x}''-\mathbf{x}'\|_2$.*

Proof.

$$\begin{aligned}\kappa\left(\mathbf{x}'', \mathbf{x}\right) - \kappa\left(\mathbf{x}'', \mathbf{x}'\right) &= \exp\left\{-\frac{1}{2}\frac{\|\mathbf{x}''-\mathbf{x}\|_2^2}{\sigma^2}\right\} - (1-\lambda)\exp\left\{-\frac{1}{2}\frac{\|\mathbf{x}''-\mathbf{x}'\|_2^2}{\sigma^2}\right\} \\ &> 0.\end{aligned}$$

□

Obviously, the conclusion of Proposition 2 can be achieved under other (weaker) conditions. Those selected only exhibit the advantage of being simple and easy to verify. An illustration is provided by the *chessboard problem* studied in the following section.

4. Experiments

The new kernel is assessed in the framework of a comparative study, where the reference is provided by the Gaussian kernel. Both kernels are incorporated in an M-SVM: the one of Weston and Watkins [1998], hereafter referred to as the WW-M-SVM. Our implementation of this machine can be found at the following address: `https://members.loria.fr/YGuermeur/WW-M-SVM.tar`.

4.1. Experimental Setup

In all the experiments below, the clustering method implemented to derive the directions $\boldsymbol{\mu}$ used by the consolidation kernel is the K-means algorithm. Only the value K of the number of clusters changes. Furthermore, model selection is minimal, so as to ease reproducibility. It is limited to the soft margin parameter C of the machine and the bandwidths of the radial basis functions (RBFs). The weights τ only take one value, $\frac{1}{M}$, and the coefficient λ of the convex combination is fixed to 0.1.

4.2. Standard Benchmark Data Sets

This experiment aims at comparing the selected combinations of machine and kernel with the state of the art. It is directly inspired by the one performed by Doğan et al. [2016] to compare nine M-SVMs equipped with a Gaussian kernel. Here, the nine M-SVMs are replaced with three machines. These machines are the WW-M-SVM equipped with the Gaussian kernel and the consolidation kernel, hereafter referred to as *our machines*, and the model identified as best (over the nine) by Doğan and his co-authors: a simplified implementation of the WW-M-SVM whose decision boundaries are linear (instead of affine), in the reproducing kernel Hilbert space (RKHS) spanned by the kernel. Over the twelve data sets from the UCI machine learning repository [Blake et al., 1998] initially used, only ten are kept: those without missing data. Their description is provided in Table 1. The experimental setup is also a five-fold cross validation, with the training set being split so as to produce a validation set for model selection. For each data set, the $M + 1$ RBFs of the consolidation kernel share one single value for their bandwidth. At last, the parameter K of the clustering method is set equal to 5.

Data set	**#Examples**	**#Attributes**	**#Classes**
Abalone	4177	8	3
Car Evaluation	1728	6	4
Glass Identification	214	9	6
Iris	150	4	3
Opt. Rec. of Handwritten Digits	5620	64	10
Page Blocks	5473	10	5
Landsat Satellite	6435	36	6
Image Segmentation	2310	19	7
Red Wine	1599	11	7
White Wine	4898	11	7

Table 1.: UCI data sets used in the experiments.

The results obtained for the three machines are given in Table 2. Here, *literature* designates the simplified variant of the WW-M-SVM used by Doğan et al. [2016]. Its test performances are those provided by the authors (using their own model selection procedure). The last column provides the values of the hyperparameters used for the machine equipped with the consolidation kernel.

It is easy to observe that the true WW-M-SVM outperforms the simplified variant (in fact all the nine machines used in Doğan et al. [2016]), on at least two data sets: Abalone and White Wine. On the contrary, for this M-SVM, the choice of the kernel makes little difference. The aim of the experiments of the next section is to assess this difference when the problem is known to be favourable to the new kernel.

4.3. Synthetic Data Sets

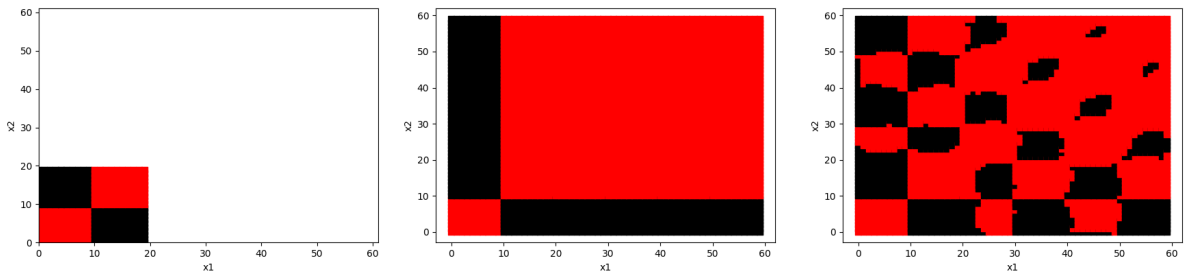
The first problem is a *chessboard problem*. This dichotomy computation consists in assigning to the points of a chessboard the color of the square to which they belong. For such a problem, both the clusters and the translation

Data set	Literature	Gaussian kernel	Kernel κ	C; $2\sigma^2$
Abalone	27.51	56.28	56.16	1.0; $0.6 \cdot p$
Car Evaluation	98.62	98.84	98.84	1.0; $0.4 \cdot p$
Glass Identification	68.78	68.40	68.36	0.5; $0.3 \cdot p$
Iris	96.35	96.00	96.00	1.0; $0.4 \cdot p$
Opt. Rec. of Handwritten Digits	98.77	98.72	98.72	1.0; $0.6 \cdot p$
Page blocks	96.83	96.47	96.47	1.0; $0.4 \cdot p$
Landsat Satellite	92.19	92.35	92.45	1.0; $0.08 \cdot p$
Image Segmentation	96.39	96.23	96.23	1.0; $0.08 \cdot p$
Red Wine	63.87	64.23	64.67	1.0; $0.4 \cdot p$
White Wine	64.86	66.42	66.62	0.8; $0.08 \cdot p$

Table 2.: Respective performances of the three classifiers.

invariances are obvious. We took benefit of that to parametrize the kernel in an optimized way. We consider two variants, both involving a 6×6 board, but differing in the nature of the training set. In the first case, this set is sampled in the four squares at the bottom left corner of the board. In the second case, the sampling involves fourteen squares randomly chosen among the thirty-six ones. In both cases, each square possesses a 10×10 grid of 100 points. The parameters choice is as follows: the directions of translation correspond to the two main diagonals and $(C, \sigma_0^2, \sigma_1^2, \sigma_2^2) = (1, 5, 250, 250)$. The two last bandwidths must be large enough to take into account a long-range dependence.

The classifications obtained are depicted in the last two panels of Figures 2 and 3. The superiority of the consolidation kernel over the Gaussian kernel is obvious as it is closer to reproduce the complete 6×6 chessboard. Interestingly, using the consolidation kernel the results are better when having only 4 squares to learn from compared to 14 (the generalization performances are 78.444% and 67.833% respectively). This is surprising but explainable as in the latter case, the periodic terms compete with the vanilla Gaussian term. This is particularly noticeable for points belonging to $[30, 50] \times [30, 50]$ where in the training set only one category is represented making it hard to affect points to the other category.

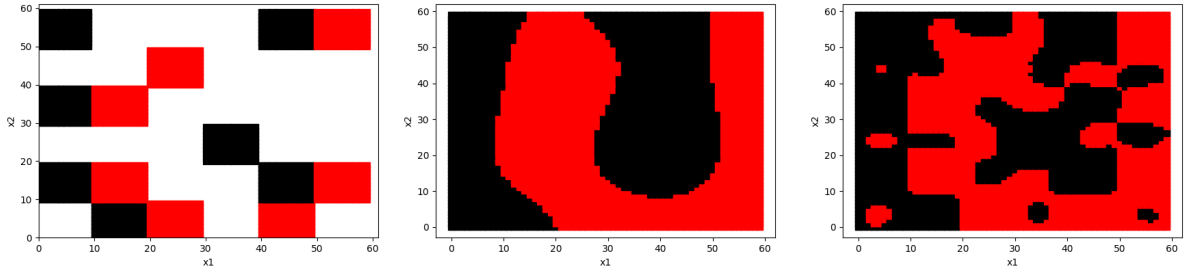


(a) Four-square training set (b) Gaussian kernel (c) Consolidation kernel

Figure 2.: Classifications with the Gaussian and the consolidation kernel for a four-square training set.

The second synthetic problem is the Madelon one, from the NIPS 2003 feature selection challenge [Guyon, 2003]. This is another two-category classification problem whose basic structure is described as follows. The data points are grouped in 32 clusters placed on the vertices of a five dimensional hypercube and randomly labelled +1 or -1. Once more, the data used are those available on the UCI repository website[2] and the number of clusters of the K-means algorithm is set equal to 5 (although 16 could have been more appropriate). A five-fold cross validation is performed on the union of the training and validation sets provided, corresponding to 2600 examples. For both kernels, model selection produces the same values for the two hyperparameters: $C = 1.0$ and $2\sigma^2 = 8.0 \cdot p$. The recognition rate obtained with the consolidation kernel is 67.85%, versus 58.00% with the Gaussian kernel. According to the two-sample proportion test, the superiority of the consolidation kernel over the Gaussian kernel is statistically significant with confidence exceeding 0.95.

[2] `https://archive.ics.uci.edu/dataset/171/madelon`



(a) Fourteen-square training set (b) Gaussian kernel (c) Consolidation kernel

Figure 3.: Classifications with the Gaussian and the consolidation kernel for a fourteen-square training set.

5. Conclusion

A new kernel has been introduced, which is designed to fit data sets where the clouds of points associated with the different categories exhibit the following behaviour. They are structured in clusters, possibly distant from each other and separated by clusters of other categories. The consolidation kernel can be seen as a convex combination of a Gaussian kernel and an infinite weighted sum of spectral mixture kernels. The main originality rests in the estimation of the parameters of the SM kernels, which is dedicated to the task of interest. It is non parametric, and based on a clustering of the clouds of points associated with the different categories. Experimental results show a performance indistinguishable from that of the Gaussian kernel on standard benchmarks which are not known to exhibit the behaviour considered. On the contrary, the gain in significant on famous artificial problems exhibiting this behaviour.

Our ongoing research deals with the empirical inference of the values of the hyperparameters K (clustering), λ, $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$. The final goal is to obtain an M-SVM capable of highlighting unknown structures in real-world data sets.

References

C. Blake, E. Keogh, and C.J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

D. Decoste and B. Schölkopf. Training invariant support vector machines. *Machine learning*, 46(1–3): 161–190, 2002.

U. Doğan, T. Glasmachers, and C. Igel. A unified view on multi-class support vector classification. *Journal of Machine Learning Research*, 17(45):1–32, 2016.

Y. Guermeur. A generic model of multi-class support vector machine. *International Journal of Intelligent Information and Database Systems*, 6(6):555–577, 2012.

I. Guyon. Design of experiments of this NIPS 2003 variable selection benchmark. Technical report, 2003. URL `http://clopinet.com/isabelle/Projects/NIPS2003/Slides/NIPS2003-Datasets.pdf`.

T. Gärtner, P. Flach, and S. Wrobel. *On graph kernels: Hardness results and efficient alternatives*. Springer, Berlin, Heidelberg, 2003.

B. Haasdonk and D. Keysers. Tangent distance kernels for support vector machines. In *Proceedings 16th International Conference on Pattern Recognition*, volume 2, 2002.

B. Haasdonk, A. Vossen, and H. Burkhardt. Invariance in kernel methods by Haar-integration kernels. In *Scandinavian Conference on Image Analysis*. Springer, Berlin, Heidelberg, 2005.

T. Hofmann, B. Scholkopf, and A.J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*. Springer, Berlin, Heidelberg, 1998.

R.I. Kondor and J.D. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *ICML'02*, pages 315–322, 2002.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of machine learning research*, pages 419–444, 2002.

P. Niyogi, F. Girosi, and T. Poggio. Incorporating prior information in machine learning by creating virtual examples. In *Proceedings of the IEEE 86.11*, pages 2196–2209, 1998.

T. Poggio and T. Vetter. Recognition and structure from one 2d model view: Observations on prototypes, object classes and symmetries, 1992.

P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning, 2021. URL `https://arxiv.org/abs/2104.08894`.

G. Salton, A. Anita Wong, and C.-S. Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, MA, 2002.

B. Schölkopf, C. Burges, and V. Vapnik. Incorporating invariances in support vector learning machines. In *International Conference on Artificial Neural Networks*. Springer, Berlin, Heidelberg, 1996.

H. Schulz-Mirbach. Constructing invariant features by averaging techniques. In *Proceedings of the 12th IAPR International. Conference on Pattern Recognition*, volume 2, 1994.

P. Simard, Y. Le Cun, J. Denker, and B. Victorri. *Transformation invariance in pattern recognition, tangent distance and tangent propagation*, volume 1524. Springer, 1998.

A.J. Smola and R. Kondor. *Kernels and regularization on graphs*. Springer, Berlin, Heidelberg, 2003.

J. Tropp. *An Introduction to Matrix Concentration Inequalities*. now publishers In, 2015.

J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science, 1998.

A.G. Wilson and R.P. Adams. Gaussian process kernels for pattern discovery and extrapolation. In *ICML*, 2013.

J. Ye, S. Ji, and J. Chen. Multi-class discriminant kernel learning via convex programming. *Journal of Machine Learning Research*, 9:719–758, 2008.