Tangent Vector Kernels for Invariant Image Classification with SVMs

Alexei Pozdnoukhov, Samy Bengio
IDIAP, Rue de Simplon, 4
CP592, CH-1920, Martigny, Switzerland
{pozd, bengio}@idiap.ch

Abstract

This paper presents an application of the general sample-to-object approach to the problem of invariant image classification. The approach results in defining new SVM kernels based on tangent vectors that take into account prior information on known invariances. Real data of face images are used for experiments. The presented approach integrates virtual sample and tangent distance methods. We observe a significant increase in performance with respect to standard approaches. The experiments also illustrate (as expected) that prior knowledge becomes more important as the amount of training data decreases.

1. Introduction

Prior knowledge is often used in machine learning algorithms to constrain models toward reasonable solutions. One such class of prior knowledge relates to invariances. These are transformations of the inputs that leave the outputs unchanged. The baseline methods that deal with invariances in the field of kernel methods are Virtual Support Vector (VSV) [1] and a method called "kernel jittering" [2]. We describe these and a number of other approaches below.

In this paper, we present yet another approach to the problem. Similarly to that done in the Tangent Distance approach [3] we use tangent vectors which correspond to local transformations we want to be invariant to. The main contribution is to provide such invariance through the use of special tangent vector kernels (TVK). The method does not lead to enlarged training sets and simply exploits standard SVM optimization algorithms.

The rest of the paper is organized as follows. We present some novel and state-of-the-art approaches to invariant learning in Section 2; we introduce a general sample-to-object concept (Section 2.1), explain the notion of *tangent vectors* and provide examples of using them for kernel construction in Sections 2.2-2.4. Section 3 presents some experiments on real images where we illustrate the performance of the proposed method in comparison to the current approaches. We also illustrate the importance of prior knowledge for small datasets in Section 3.2. Finally, Section 4 completes the paper with discussion and conclusions.

2. Approaches to Invariant Learning

One of the most widely used and probably most practically efficient approaches to invariant learning is to use specific task-dependent features in combination with a standard learning algorithm. At the same time, the general approaches to constructing a task-independent learning system that is invariant to some desired transformations is of particular interest. Some of the work that have been done in this direction in the field of kernel methods was mentioned in the introduction. Here we describe some other recent developments.

One of the well-known approaches to invariant learning is the Tangent Distance method [3]. It proposes to replace the Euclidean distance between data samples with a distance between the corresponding linear tangent manifolds defined by tangent vectors of the desired invariance transformation. This method was successfully applied to Optical Character Recognition (OCR) tasks. Its direct application for defining a kernel for SVMs was studied in [4]. Tangent Distance method, the baseline Virtual Support Vector method, as well as "kernel jittering" that combines virtual sample generation and kernel modification could also be considered as special cases of a general approach we describe below.

2.1. From Sample to Object

Suppose we have some understanding of our data that can be formalized as a transformation of the inputs that leaves the outputs unchanged. For example, in a 2D image classification task we are often given the evident knowledge that small rotations and scalings of the raw images do not affect the desired output class. Suppose the representation of the data (the set of features) allows us to describe the desired transformation as a mapping that leaves the outputs unchanged. The mapping applied to every sample produces a set of corresponding objects, which becomes a focus of our consideration. In other words, we assume that given some understanding of the data we are able to generalize each sample into the equivalence class - the object in the input space. This approach can be used to take prior knowledge into account in *kernel methods* by defining a kernel function between objects. A related attempt to derive a learning algorithm directly for objects was recently presented in [6].

Apart from the mentioned methods, some work has been done in [7] for defining a kernel between sets of vectors, but it was aimed at input representation and not to include invariances into the training algorithm. A novel approach that can be used for including invariances was recently presented in [8]. Vapnik's Vicinal Risk Minimization principle and derived SVM-based algorithm [5] is closely linked to the presented research and can also be considered as an implementation of this sample-to-object approach.

2.2. Tangent Vectors

Partly following the notation of [3], consider the transformation t_α defined by the set of parameters α in some region of $D \in R^2$:

$$t_\alpha : D \in R^2 \mapsto t_\alpha(D) \in R^2. \tag{1}$$

This transformation is assumed to be differentiable with respect to α and $(x, y) \in D$, and reduces to the identity transformation for some value of α^0. Then the object generated by this transformation and associated with an image U is defined by

$$S(U, \alpha) = U \circ t_\alpha^{-1},\ \alpha \in \Lambda, \tag{2}$$

where Λ is some admissible set of parameters α. Its corresponding linear approximation is

$$S_1(U, \alpha) = U + \sum_{j=1}^{J} (\alpha_j - \alpha_j^0) L_{\alpha_j}(U), \tag{3}$$

where $L_{\alpha_j}(U)$ are local transformations of U defined by:

$$L_{\alpha_j}(U) = \left.\frac{\partial S(U, \alpha)}{\partial \alpha_j}\right|_{\alpha=\alpha_0}. \tag{4}$$

Note that L_{α_j} are operators that generate the whole space of local transformations (a Lie algebra of local transformations).

Examples of transformations widely used in image processing such as rotations and scaling are shown below:

- Rotation:

$$t_\alpha = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix},\ L_\alpha^{rot} = y\frac{\partial}{\partial x} - x\frac{\partial}{\partial y}. \tag{5}$$

- Scaling:

$$t_\alpha = \begin{pmatrix} 1+\alpha & 0 \\ 0 & 1+\alpha \end{pmatrix},\ L_\alpha^{sc} = x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}. \tag{6}$$

2.3. Tangent Vector Kernels

Suppose the transformation we want to be invariant to defines a differentiable manifold in the input space. Hence the tangent vectors can be defined as described above, and the whole set of tangent vectors can be used to model all the local linear transformations of the given image. Let us define the following function H which gives the measure of proximity of a given vector x to the linear span of some vector x' generated with a tangent vector ℓ_j:

$$H(x|x', \ell_j) = e^{-\frac{(x-x')^2 \ell_j^2 - ((x-x')\cdot \ell_j)^2}{2\gamma_w^2 \ell_j^2}}, \tag{7}$$

where γ_w is the parameter related to the width of the proximity region.

The one-sided Tangent Vector Kernel (TVK) K_s which describes a similarity between the given sample x and an object based on sample x' generated by a set of corresponding tangent vectors $\{\ell_1, ...\ell_J\}$ can be defined as follows:

$$K_s(x, x') = e^{-\frac{(x-x')^2}{2\sigma^2}} \cdot \prod_{j=1}^{J} (\eta + H(x|x', \ell_j)) \tag{8}$$

where σ is a kernel bandwidth and real number $\eta \in [0, 1]$ defines the shape of the kernel.

Two-sided kernel K_d can be obtained by taking the average of two one-sided kernels:

$$K_d(x, x') = \frac{1}{2}(K_s(x, x') + K_s(x', x)). \tag{9}$$

The proposed kernel combines the advantages of both VSV and Tangent Distance approaches. In this approach we not only analytically include the Virtual SV into the model (without putting them into the data), but also take into account all the linear combinations of invariant transformations of interest. Moreover, using all the tangent vectors which correspond to linear transformations, one can take into account all the possible local linear transformations of an image.

Note that the proposed kernel (8) is not the only possible one to make use of the tangent vectors. Other kernels can be constructed in a similar way to the one presented by combining the terms (7) in a different manner.

2.4. Distribution-based Tangent Vector Kernels

Another method of kernel construction that directly implements the sample-to-object approach is to consider an object given by (8) if the latter is considered as a density function. The kernel can be obtained by measuring the overlap of two distributions that correspond to the object based on samples x and x' as it was proposed in [7] for sets of vectors. To do this we introduce the following distance measure between two distributions:

$$K_B(x', x'') = \int K_s(x, x')^\rho K_s(x, x'')^\rho dx. \tag{10}$$

Assuming $\eta = 0$ in (8) and applying normalization, we reduce (8) to a standard Gaussian:

$$\begin{aligned} K_s(x, x') &= \frac{e^{-(x-x')^T L_{x'}^{-1}(x-x')}}{(2\pi)^{N/2}|L_{x'}|^{1/2}}, \text{ where :} \\ L_{x'}^{-1} &= \left(\tfrac{1}{2\sigma^2} + \tfrac{J}{2\gamma_w^2}\right) I - \sum_{j=1}^{J} \frac{\ell_j \ell_j^T}{2\gamma_w^2 \ell_j^2}, \end{aligned} \tag{11}$$

where I is an identity matrix, and $|...|$ denotes the determinant. Hence the results of [7] can be directly applied to obtain a closed form of $K_B(x', x'')$:

$$\begin{aligned} K_B(x', x'') &= (2\pi)^{\frac{(1-2\rho)N}{2}} \left|\hat{L}\right|^{\frac{1}{2}} |L_{x'}|^{-\frac{\rho}{2}} |L_{x''}|^{-\frac{\rho}{2}} \\ &\exp(-\tfrac{\rho}{2} x'^T L_{x'}^{-1} x' - \tfrac{\rho}{2} x''^T L_{x''}^{-1} x'' + \tfrac{1}{2}\hat{x}^T \hat{L} \hat{x}) \end{aligned} \tag{12}$$

where $\hat{L} = (\rho L_{x'}^{-1} + \rho L_{x''}^{-1})^{-1}$ and $\hat{x} = \rho L_{x'}^{-1} x' + \rho L_{x''}^{-1} x''$.

A closed form equation for the distribution-based tangent vector kernel can also be derived for $\eta \neq 0$ and $\rho = 1$, which is more interesting but yields an even more cumbersome expression. However, an implementation of (12) still demands costly computations for high-dimensional input spaces. The experiments presented below will only use kernels based on (7)-(9).

3. Experiments

In order to test the proposed approach, we conducted experiments using images of the faces detected on every fifth frame of a movie using a face detector presented in [9]. Image dimension is 81 by 81 and gray scale level is 8 bits. There is a total of 2899 images in the database. The data is available at [http://www.robots.ox.ac.uk/∼vgg/data]. We present an approach to the problem of binary classification of the main actor against all the other images captured. Example faces and their corresponding label are presented in Figure 1. This task can be seen either as a person identification or an information retrieval task. We are not aimed at constructing a specific biometric identification or information retrieval system, though the proposed method could establish a foundation for them.

3.1. Person Identification

We compare standard SVM with RBF kernel, Virtual Support Vector method, Kernel Jittering, and the proposed approach. Two types of invariant transformations were studied: rotations (5) and scalings (6).

Figure 1. Some Images of the Database

The original Tangent Distance method was found to be just partly applicable to the task. The reason for that is its limitations in computing the tangent vectors. The input image has to be smooth enough to compute gradients that would approximate local transformations of the original image. The method worked well for binary images of digits, which were blurred with Gaussian filter for computing the gradients. We applied the method for our data using different Gaussian smoothing and found that the obtained approximation from these tangent vectors was not sufficient to described real transformations. Instead we generated virtual samples by applying a finite desired transformation and used them for computing the finite differences that were used to approximate the tangent vectors. Example transformed images obtained by rotations with original gradient-based tangent vectors and finite differences are shown in Figure 2. The first line in Figure 2 presents images obtained by applying direct calculation of tangent vectors according to (5). We can thus see that despite the accurate tuning of Gaussian filtering and other “tricks”, only very local rotations are reasonable.

Figure 2. Two Types of Virtual Images.

The second line in Figure 2 presents the original sample image x in the center; virtual samples obtained from x by applying rotations of 10 degrees are shown on the left and right of the figure. Let us denote them as $x + \ell_{left}^{exp}$ and $x + \ell_{right}^{exp}$. The intermediate images in between are $x + 0.5\ell_{left}^{exp}$ and $x + 0.5\ell_{right}^{exp}$. Since this approach implied that left and right rotations correspond to different tangent vectors, we used the following modified Tangent Vector Kernel:

$$K_s^{fd}(x, x') = e^{-\frac{(x-x')^2}{2\sigma^2}} + \sum_{j=1}^{J} H(x|x', \ell_j) \cdot e^{-\frac{(x-x'-\ell_j)^2}{2\gamma_r^2}}, \tag{13}$$

where we introduced one extra parameter γ_r corresponding to the length of proximity region and replaced product with a sum. The experiments with modified kernels based on products (as presented in (8)) led to similar results.

With a proper choice of parameters in (13) ($\gamma_w \sim \infty$, $\gamma_r = \sigma$), the resulted model is closely linked to VSV. The noticeable difference is that in the VSV approach every virtual sample is included in the decision function with its own weight, while in our case all the virtual samples form an object hence share the same weight

We then divided the dataset into 300 training and 2599 testing samples. The parameters of the algorithms were chosen according to the minimum of cross-validation error over the training set, resulting in $\sigma = 600$, $C = 100$. Parameters γ_w and γ_r in (13) can be chosen by the following heuristics: $\gamma_w \sim \sigma$, and $\gamma_r^2 \sim Var(\ell_{ij})$, i.e. the variance of tangent vectors. We used $\gamma_w = 500$ and $\gamma_r = 1000$.

Table 1 presents testing errors obtained with SVM with Gaussian RBF kernel (SVM), SVM trained with virtual samples (VSV SVM), SVM with jittered kernel (KJ SVM) and SVM with Tangent Vector Kernel (TVK SVM). The improvement of the testing error in comparison to the baseline SVM is statistically significant with a 95% confidence interval.

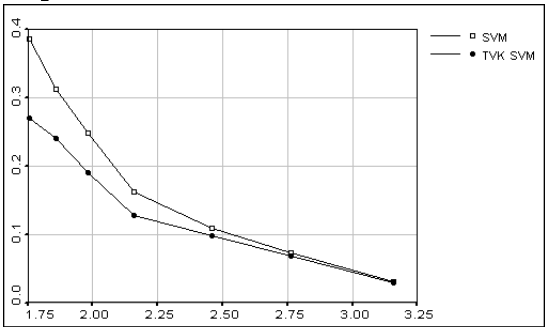
Table 1. Testing Error

	Algorithm	Testing Error, %
	SVM	11.2
	VSV SVM	9.8
	KJ SVM	10.0
	TVK SVM	9.7

3.2. The Importance of Prior Knowledge for Small Datasets

Another interesting experiment is to show the relative importance of prior knowledge with respect to the amount of available training data. We thus split the data using every N-th sample of the entire data for training, while the rest of the data were used for testing. Figure 3 shows the testing errors obtained for these different partitions. The X-axis in Figure 3 corresponds to the logarithm of the training set size and the Y-axis corresponds to the testing error. As expected, when the number of training examples is very small, prior knowledge is of prime importance, while its importance eventually decreases with increased amount of training examples.

Figure 3. SVM with RBF and TVK kernels.



4. Conclusion

In this paper we presented an application of the general sample-to-object approach to the problem of invariant image classification with kernel methods such as SVM. Experimental results on real data of face images yielded significant improvement with respect to the baseline SVM. The relative importance of prior knowledge with respect to the size of the training set was also illustrated. Future work will include an application of the described kernels for one-class SVM algorithms.

5. Acknowledgments

This research has been partially carried out in the framework of the European project LAVA, funded by the Swiss OFES project number 01.0412. It was also partially funded by the Swiss NCCR project (IM)2.

References

[1] B. Scholkopf, C. Burges, and V. Vapnik. Incorporating invariances in support vector learning machines. In C. von der Malsburg, W. von Seelen, J. C. Vorbruggen, and B. Sendhoff, (eds.), ICANN‘96, pp. 47-52, Berlin, 1996.

[2] D. DeCoste, M.C. Burl. Distortion-invariant recognition via jittered queries. In the proc. of *CVPR-2000*, June, 2000.

[3] P. Simard, Y. LeCun, J. Denker, B. Victorri. Transformation invariance in pattern recognition, tangent distance and tangent propagation. In G. Orr and K. Muller, (eds.), *Neural Networks: Tricks of the trade.* Springer, 1998.

[4] B. Haasdonk, D. Keysers. Tangent Distance Kernels for Support Vector Machines. In the proc. of ICPR’02, Vol.2, pp. 864-868.

[5] V. Vapnik. The Nature of Statistical Learning Theory. Second edition, Springer-Verlag, NY, 2000.

[6] T. Graepel, R. Herbrich. Invariant Pattern Recognition by Semidefinite Programming Machines. In the proc. of NIPS’03, in press.

[7] R. Kondor, T. Jebara. A kernel between sets of vectors. In the proc. of ICML-2003, Washington DC.

[8] L. Wolf, A. Shashua. Learning over Sets using Kernel Principal Angles. JMLR 4(Oct):913-931, 2003.

[9] H. Schneiderman, T. Kanade. A Statistical Method for 3D Object Detection Applied to Faces and Cars. In the proc. of CVPR-2000, pp.746-751.