Improving Kernel Classifiers for Object Categorization Problems

Alexei Pozdnoukhov Alexei.Pozdnoukhov@idiap.ch
Samy Bengio Samy.Bengio@idiap.ch
IDIAP Research Institute, Rue du Simplon, 4, Martigny, CH-1920, Switzerland

Abstract

This paper presents an approach for improving the performance of kernel classifiers applied to object categorization problems. The approach is based on the use of distributions centered around each training points, which are exploited for inter-class invariant image representation with local invariant features. Furthermore, we propose an extensive use of unlabeled images for improving the SVM-based classifier.

1. Introduction

Facing the growing amount of image databases and their sizes, the problem of constructing an effective data structuring scheme arises. One solution is to base it on an object categorization approach. The aim is then to classify the given set of images into some predefined categories, based on their image content. This classification task needs to be inter-class discriminative as much as possible, while providing good intra-class invariance to object appearance, lighting conditions, and background noise.

Generally, in image processing, feature extraction procedures result in huge sets of features, which can be hardly processed in their raw representation. Hence, they are often handled instead using distributions. In the field of image processing, this is the case when using invariant descriptors, e.g. SIFT (Lowe, 2004) and JETs (Schmid & Mohr, 1997), computed at automatically detected interest points of the image. These invariant descriptors have recently been used with good success in object categorization tasks (Opelt et. al.; Csurka et. al., 2004).

Appearing in *Proceedings of the ICML 2005 workshop on Learning with Partially Classified Training Data*, Bonn, Germany, 2005.

One solution to the use these distributions in kernel based methods for this problem is to build a kernel classifier (such as an SVM) by defining a kernel on distributions, using for instance the KL-divergence or other related distances such as the Bhattacharya affinity (Kondor, Jebara, & Howard, 2004). However, such classification tasks would benefit from a more direct application of margin-inspired learning criteria.

We thus present here an approach which directly maximizes the margin between distributions. Given Gaussian distributions, a large margin classifier can be built using an alternative formulation of SVMs, already mentioned by (Vapnik, 2000), but which leads to a Second Order Cone Programming (SOCP) optimization problem (Bhattacharyya, Pannagadatta & Smola, 2004; Bi & Zhang, 2004). Unfortunately, this is a time-consuming optimization task as compared to quadratic programming (QP) techniques, considering the efficient SVM-specific algorithms for solving the latter. We thus propose instead an approximate solution to this optimization problem, which not only avoids solving a complicated SOCP optimization problem, but also presents a nice feed-back for a practitioner.

Using a separate unlabeled large dataset of images, the algorithm suggests the most valuable locations which, being considered as new training samples, provide margin maximization between distributions. The unlabeled samples which are closest to the latter can then be used in an active learning fashion in order to enhance the performance of the classifier.

There are two major innovative aspects presented in this paper. Firstly, a kernel method for classifying distributions is proposed in Sections 2 and 3. Secondly, we explore the use of kernel methods for the challenging problem of improving object categorization models by introducing unlabeled data points into the training set (Section 4). An experimental setting for the problem of object categorization is then presented in Section 5.

2. Kernel Classifier for Distributions

The margin maximization principle is based on results from the Statistical Learning Theory (Vapnik, 2000), and provides a way to minimize the complexity of the model by bounding the VC-dimension of the modeling function.

Intuitively, the same approach can be used for other learning tasks. We thus now present a definition of margin for distributions, and provide a way for constructing learning algorithms based on this margin. The proof of the margin maximization principle for the considered problem is out of the scope of this paper. The general solution to the problem would include introducing functional data input spaces and corresponding generalization bounds.

Suppose one is given a training set of L probability distribution functions $p(\boldsymbol{x}|\boldsymbol{x}_i,\boldsymbol{r}_i)$, centered at $\boldsymbol{x}_i$ and specified by some parameters $\boldsymbol{r}_i$. We also associate some label y_i for each distribution. These are {+1, -1} for binary classification problem.

2.1. Linear Decision Functions

Consider the set of linear decision functions $\{f = \boldsymbol{wx} + b\}$, where $\boldsymbol{w}$ is a weight vector, and b is a constant threshold. The actual decision is usually taken according to $sign(f)$.

Consider the optimization problem:

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{L}\xi_i \tag{1}$$

subject to the following constraints:

$$\int\limits_{y_i(\boldsymbol{wx}+b)\geq 1} p(\boldsymbol{x}|\boldsymbol{x}_i,\boldsymbol{r}_i)dx \geq \eta - \xi_i, i = 1, ...L, \tag{2}$$

$$\xi_i \geq 0, i = 1, ...L. \tag{3}$$

The first constraint corresponds to the fact that η-quantile of the distribution lies outside the margin, not taking into account the slack variable ξ_i. These slack variables are equivalent to the analogue trick done in soft margin formulation of the Support Vector Machine.

2.2. Iterative Solution

The general approach for solving the optimization problem (1)-(3) is to apply an iterative procedure in order to obtain an approximate solution. This type of optimization approach has already been applied, for instance in (Bi & Zhang, 2004). Note however that the nature of SVM-related methods is that they try to find Support Vectors, i.e. samples which lie closest to the discriminative surface. Thus, when discriminating some subsets $S(\boldsymbol{x}_i)$, constraints of the type $\max\limits_{\boldsymbol{x}\in S(\boldsymbol{x}_i)} [y_i(\boldsymbol{wx} + b)] \geq 1 - \xi_i$ can be used. See for instance (Graepel & Herbrich, 2003; Fung, Mangasarian & Shavlik, 2002; Bhattacharyya, Pannagadatta & Smola, 2004) where such a solution was applied for different types of $S(\boldsymbol{x}_i)$. Solving problems with this type of constraints is roughly equivalent to the task of finding the "optimal" or "effective" sample from the subset.

A similar approach holds for the case of distributions. There exists a representation of the hyper-plane in terms of some samples $\boldsymbol{x}_i^*$ which coincide with the solution of the problem (1)-(3). We propose here a simple 2-step method to obtain an approximate solution to (1)-(3).

2.3. Hyper-plane Projection Method

¿From now on, let us consider the kernelized version of the proposed algorithm. Let $K(.,.)$ be a reproducing positive definite kernel. Let some $(\boldsymbol{w}_0, b_0)$ define the optimal separating hyper-plane in the feature space induced by $K(.,.)$ for the training set of means and targets $\{\boldsymbol{x}_i, y_i\}$. Actually, this is given by the set of Lagrange multipliers $\{\alpha_i\}$, obtained by solving the standard SVM optimization. The proposed scheme is as follows:

1. Solve a standard SVM optimization problem for the means $\boldsymbol{x}_i$. The obtained solution is $(\boldsymbol{w}_0, b_0)$.
2. Calculate the projections of $p(\boldsymbol{x}|\boldsymbol{x}_i, \boldsymbol{r}_i)$ on $\boldsymbol{w}_0$. This results in a simple 1-D optimization problem.
3. Solve the 1-D problem according to the given value of η.
4. Compute the inverse projection. This results in a modified training set $\boldsymbol{x}_i^*$.
5. Solve a standard SVM optimization problem using the original and the modified samples $\boldsymbol{x}_i^*$.

Detailed explanation of the projection steps is presented below.

2.4. Direct Projection

Consider the following averages in the feature space, which provide the means and variances of some 1-D distribution $\pi(\chi|\mu_j, \sigma_j)$.

$$\mu_j = E[\boldsymbol{w}_0 \Phi(\boldsymbol{x}_j) + b_0] = \sum_{i=1}^{L} y_i \alpha_i \int_{\mathrm{X}} K(\boldsymbol{x}_j, \boldsymbol{x}_i) p(\boldsymbol{x}|\boldsymbol{x}_i, \boldsymbol{r}_i) dx + b_0, \tag{4}$$

$$\sigma_j^2 = E[(\boldsymbol{w}_0 \Phi(\boldsymbol{x}_j) - \mu_j)^2] = \sum_{i,k=1}^{L} y_i y_k \alpha_i \alpha_k \int_{\mathrm{X}^2} K(\boldsymbol{x}_i, \boldsymbol{x}_k) p(\boldsymbol{x}|\boldsymbol{x}_i, \boldsymbol{r}_i) p(\boldsymbol{x}'|\boldsymbol{x}_k, \boldsymbol{r}_k) dx dx' - \mu_j^2. \tag{5}$$

These 1-D pdfs correspond to $p(\boldsymbol{x}|\boldsymbol{x}_j,\boldsymbol{r}_j)$ being projected to the 1-D subspace defined by $\boldsymbol{w}_0$. Given these projections, the constraints (2) can be (currently) satisfied by taking the χ_j in 1-D space such that

$$\int_{y_j f(\boldsymbol{x}) \geq \chi_j} \pi(\chi|\mu_j, \sigma_j) d\chi \geq \eta. \tag{6}$$

It can be solved easily and results in some threshold constant c_η^j such that $\chi_j = f(\boldsymbol{x}_j^*) = c_\eta^j$.

2.5. Inverse Projection

Given the set of χ_j, we now need to find an inverse projection of χ_j back into the feature space. Obviously, this transformation is not unique and some criterion is required in order to define it more precisely. At this step, it is hard to control the margin, but the constraint (2) can still be satisfied. One would thus like to find $\boldsymbol{x}_{\mathrm{j}}^*$ such that the inequality in (2) holds (or is violated as slightly as possible) over variations in $\boldsymbol{w}$ and b. For the majority of distributions which are used in real-life problems, the following criterion can be used in order to obtain the inverse projection $\boldsymbol{x}_j^*$ of the χ_j:

$$\boldsymbol{x}_j^* = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{x}_j, \boldsymbol{r}_j), \quad s.t.\ f(\boldsymbol{x}) = c_\eta^j \tag{7}$$

A useful intuition behind this criterion is as follows: if $\boldsymbol{x}_j^*$ is fixed at the maximum of $p(\boldsymbol{x}|\boldsymbol{x}_j,\boldsymbol{r}_j)$ at the surface $f(\boldsymbol{x}) = c_\eta^j$, then the integral in the left part of (2) is less likely to change. Problem (7) is a constrained optimization problem, which needs to be solved. It results in the desired inverse projections $\boldsymbol{x}_j^*$ which form the new training set. A standard SVM solution for the obtained training set then approximates the solution of the initial problem (1).

3. Discrimination of Gaussian Distributions

For the particular case of Gaussian distributions, $p(\boldsymbol{x}|\boldsymbol{x}_i,\boldsymbol{r}_i) = \mathcal{N}(\boldsymbol{x}_i,\Sigma_i)$, the presented scheme can be applied easily. The direct projection step requires integrating Gaussians which is feasible in closed form and is not presented here. Afterward, the inverse projection can be carried out by solving the following optimization problem:

$$\boldsymbol{x}_j^* = \arg\min_{\boldsymbol{x}} (\boldsymbol{x} - \boldsymbol{x}_j)^T \Sigma_j^{-1} (\boldsymbol{x} - \boldsymbol{x}_j), \quad s.t.\ \sum_{i=1}^{L} y_i \alpha_i \exp(-\delta(\boldsymbol{x} - \boldsymbol{x}_i)^2) + b = c_\eta^j. \tag{8}$$

This problem has the following approximate analytical solution:

$$\boldsymbol{x}_j^* = (I + 2\gamma\delta c_\eta^j \Sigma_i)^{-1} (\boldsymbol{x}_j + 2\gamma\delta \sum_{i=1}^{L} y_i \alpha_i \exp(-\delta(\boldsymbol{x}_j - \boldsymbol{x}_i)^2) \Sigma_i \boldsymbol{x}_i), \tag{9}$$

for some positive constant γ, which has to be chosen in order to satisfy the constraint in (8). The computations are significantly simplified, since for high-dimensional input data diagonal covariance matrices are often used.

4. Partly Labelled Data

We will use the unlabelled training data in the Active Learning style. Concerning SVMs, Active Leaning approaches exploit margin properties to optimize the version space by including respective queries from the pool of data (Tong, 2001). One of the surprising properties observed in SVM Active Learning, is that an SVM, trained on the algorithmically chosen patterns outperforms SVM, trained on the whole dataset. It empirically justifies the approaches aimed at optimizing the training set by selecting the appropriate patterns for training.

To make use of unlabelled data, we use an equation for modifying the means (9). Instead of substituting the original training sample $\boldsymbol{x}_j$ with the suggested "virtual" sample $\boldsymbol{x}_j^*$, we will take the real-life unlabelled sample, closest to $\boldsymbol{x}_j^*$. This may appear to be the time-consuming step of the algorithm. However, since data are presented as distributions, simple comparison of the corresponding probabilities can be considered.

Note that opposite to the common Active Learning scheme, when one sample is added to the training set each time, we add one sample for every obtained Support Vector. These samples have to be labelled then. To avoid user participation, the cooresponding labels of the "parent" samples $\boldsymbol{x}_j$ can be assigned. The labelling can be reconsidered for the samples which both obtain upper bound C weights and still classified incorrectly by the re-trained SVM. Otherwise, these samples can be neglected and the samples $\boldsymbol{x}_j^*$, originally suggested by (9), can be taken.

5. Object Categorization

Local invariant image descriptors provide reasonable performance for object detection and categorization tasks (Lowe, 2004; Schmid & Mohr, 1997). It is evident to try to benefit from both invariant features and a powerful classifier such as an SVM. A number of approaches has been proposed for the latter (Wallraven, Caputo & Graf, 2003; Csurka et. al., 2004). The common idea is to define some kernel for the sets of features.

Preliminary experimental results of (Eichhorn & Chapelle, 2004) suggest the following choice of the approach, which combines SIFT features (Lowe, 2004), modeled with Gaussian distributions, and a Bhattacharya kernel, which can be obtained in the closed form for this type of distributions (Kondor, Jebara, & Howard, 2004). However, with the growing number of dimension, the influence of the variance vanishes, and the Bhattacharya kernel for two Gaussians converges to the kernel, which mainly depends on the distance between the respective means. Therefore, it is still partly subject to the curse of dimensionality problem. Therefore, the proposed direct margin maximization approach is a promising alternative. These considerations will be verified experimentally.

The expected results of the ongoing experiments represent the decrease of the testing error while adding the samples from the unlabelled set according to the proposed method.

6. Conclusions

We presented an approach to improve the kernel-based solutions for object categorization problems. The method takes advantage of the direct margin maximization between distributions. These distributions are used to model the SIFT features, extracted from the images. Currently, SIFT features are known to be one the most successful features used for object categorization. Unlabelled data are used in the Active Learning manner. The samples are added into the training set according to the suggested update rule.

The advantage of the proposed scheme is a possibility to obtain a nice feed-back while constructing a classifier for object categories. The scheme also gives promising possibilities for feature selection. Currently, SIFT features in kernel classifiers are used ad-hoc for solving object categorization problems. The presented approach provides better understanding of SIFT features usage.

The expected experimental results are aimed at illustrating the efficiency of the approach. The described experimental setup will be completed throughly and the obtained results will be presented at the workshop.

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 supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss OFES. It was also partially funded by the Swiss NCCR project (IM)2.

References

Bhattacharyya, C., Pannagadatta, K. S., Smola, A. (2004) A Second Order Cone Programming Formulation for Classifying Missing Data. *Proc. of Neural Inf. Proc. Systems.*, MIT press, Cambridge.

Bi, J. and Zhang, T. (2004). Support Vector Classification with Input Data Uncertainty. *Proc. of Neural Inf. Proc. Systems.*, MIT press, Cambridge.

Csurka, G., Bray, C., Dance, C., Fan, L. (2004) Visual categorization with bags of keypoints The 8th European Conference on Computer Vision - ECCV, Prague, May 11-14, 2004.

Graepel, T., and Herbrich., R., (2003). Invariant Pattern Recognition by semidefinite programming machines. *Advances in Neural Information Processing Systems*, vol. 16, Cambridge, MA, MIT Press.

Eichhorn, J., Chapelle, O.: Object categorization with SVM: kernels for local features. MPI Technical Report, July, 2004.

Fung, G., Mangasarian, O.L., and Shavlik, J., (2002). Knowledge-based support vector machines classifiers. *Advances in Neural Information Processing Systems*, vol. 15, Cambridge, MA, MIT Press.

Kondor, R., Jebara, T., Howard, A., (2004). Probability Product Kernels *Journal of Vachine Learning Research*, 5(2004), pp. 819-844.

Lowe, D., (2004). Distinctive image features from scale-invariant keypoints *International Journal of Computer Vision*, 60, 2 (2004), pp. 91-110.

Opelt, A., Fussenegger, M., Pinz, A., Auer, P. (2004) Weak Hypotheses and Boosting for Generic Object Detection and Recognition The 8th European Conference on Computer Vision - ECCV, pp. 71-84, Prague, 2004.

Schmid, C., Mohr., R. (1997) Local greyvalue invariants for image retrieval. *Trans. on Pattern Analysis and Machine Intelligence*, 19(5):530-534, 1997.

Tong, S. (2001) Active Learning: Theory and Applications. A dissertation submitted to the department of computer science. Stanford University.

Vapnik, V., (2000). The Nature of Statistical Learning Theory. Second edition, Springer-Verlag, NY.

Wallraven, C., Caputo, B., Graf, A.B.A., (2003) Recognition with local features: the kernel recipe. *ICCV 2003 Proceedings*, vol. 2, pp. 257-264, IEEE press, 2003.