IDIAP RESEARCH REPORT

A Kernel Classifier for Distributions

Alexei Pozdnoukhov Samy Bengio

IDIAP–RR 05-32

June 2005

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-32

A Kernel Classifier for Distributions

Alexei Pozdnoukhov Samy Bengio

June 2005

Abstract. This paper presents a new algorithm for classifying distributions. The algorithm combines the principle of margin maximization and a kernel trick, applied to distributions. Thus, it combines the discriminative power of support vector machines and the well-developed framework of generative models. It can be applied to a number of real-life tasks which include data represented as distributions. The algorithm can also be applied for introducing some prior knowledge on invariances into a discriminative model. We illustrate this approach in details for the case of Gaussian distributions, using a toy problem. We also present experiments devoted to the real-life problem of invariant image classification.

1 Introduction

Large margin classifiers such as Support Vector Machines (SVMs) have shown to yield state-of-the-art performance in various classification tasks (Burges, 1998). Moreover, classification tasks are known to be in general better solved by discriminant approaches (which aim at providing a model that minimizes some error on the training set of positive and negative examples), than by generative approaches (which aim at providing class-specific models that maximize the likelihood of the corresponding class training data).

On the other hand there are still several well-known classification tasks for which the current state-of-the-art is based on generative models. This is in general due to one of many possible reasons. For instance, in speech recognition, for which the best solutions are based on Hidden Markov Models and Gaussian Mixture Models trained by EM (Rabiner, 1989), one faces datasets of potentially millions of frame examples, which cannot be handled by SVMs given the (at least) quadratic training time and space complexity.

Another example is the task of speaker verification, which consists of determining whether the voice of a given person corresponds to the claimed ID. One problem with this task is that one can in general only collect very small (and thus non representative) amount of data of a given client, but very large datasets of potential impostors, which makes the problem highly imbalanced. State-of-the-art systems are thus based on the likelihood ratio of generative models of the two classes, where the client model is often adapted from the impostor model, given the lack of client specific data.

Based on these facts, there have been several attempts at integrating generative and discriminant approaches into one framework. One such attempt is based on the Fisher Kernel (Jaakkola & Haussler, 1999), where an SVM with a specific kernel is trained on examples which are the derivative of the log likelihood of the generative models of each class with respect to the parameters of the models.

Furthermore, in many classification tasks, feature extraction procedures sometimes result in huge sets of features, which can be hardly processed in the raw representation. One solution often used is to model them with distributions. In the field of image processing, this is the case for invariant features, extracted at some points of interest of an image. This approach is extensively used in object categorization problems. One of the direct solutions for this problem is to build a kernel classifier (SVM) by defining the kernel over distributions, using either KL-divergence or similar methods.

Another field of applications concerns invariant learning. While constructing invariant learning algorithms, it is reasonable to deal with the whole set of patterns which can be obtained from every given training sample by applying some transformation (such as translation or rotation). These sets are usually considered as some manifolds in the input space (Graepel & Herbrich, 2003; Fung, Mangasarian & Shavlik, 2002). A "soft" representation of these manifolds as distributions could be used instead.

Facing these problems, we feel it is worth developing a discriminative classifier which would directly deal with generative models, i.e. distributions. The approach presented in this paper exploits some nice performance of margin-based methods by constructing a maximum margin solution for a set of distributions labeled into two classes. Experiments are concentrated on the task of invariant image classification.

The rest of the paper is organized as follows. In Section 2, the notion of margin maximization for distributions is defined. Some general facts, which provide a foundation for an approximate approach, described in Section 3, are also presented. Section 4 justifies the proposed approximations. Next, we consider the particular important case of Gaussian distributions in Section 5. Section 6 is devoted to experiments, followed by discussion and conclusions in Section 7.

2 Margin Maximization for Distributions

The margin maximization principle is based on results from the Statistical Learning Theory (Vapnik, 1998), 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. The proof of the margin maximization principle for the considered problem is out of the scope of this paper.

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 $\{f = \boldsymbol{w}\boldsymbol{x} + b\}$, 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:

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

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

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 Optimization Problem

The method of Lagrange multipliers can be applied to solve the problem stated in Section 2.1. Introducing the Lagrange multipliers $\{\alpha_i\}$ and $\{\beta_i\}$, one obtains the optimization problem of finding the saddle point of Lagrangian. Differentiation in b and $\boldsymbol{w}$ gives:

$$\sum_{i=1}^{L}\alpha_i \frac{\partial}{\partial b}\int\limits_{y_i(\boldsymbol{w}\boldsymbol{x}+b)\geq 1} p(\boldsymbol{x}|\boldsymbol{x}_i,\boldsymbol{r}_i)dx = 0, \tag{4}$$

$$\sum_{i=1}^{L}\alpha_i \frac{\partial}{\partial \boldsymbol{w}}\int\limits_{y_i(\boldsymbol{w}\boldsymbol{x}+b)\geq 1} p(\boldsymbol{x}|\boldsymbol{x}_i,\boldsymbol{r}_i)dx = \boldsymbol{w}. \tag{5}$$

Analogously to the standard SVM, the multipliers β_i vanish, resulting in a box-type constraints for the weights α_i. Next, introducing the following notation:

$$I(\boldsymbol{w},\boldsymbol{x}_i,t) = \int\limits_{y_i(\boldsymbol{w}\boldsymbol{x}+b)=1-t} p(\boldsymbol{x}|\boldsymbol{x}_i,\boldsymbol{r}_i)dx, \tag{6}$$

one yields the following optimization problem:

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{L}\int_{\xi_i}^{\infty} I(\boldsymbol{w},\boldsymbol{x}_i,t)dt, \tag{7}$$

$$s.t. \sum_{i=1}^{L} \alpha_i y_i I(\boldsymbol{w}, \boldsymbol{x}_i, \xi_i) = 0, \tag{8}$$

$$\sum_{i=1}^{L} \alpha_i y_i \boldsymbol{x}_i^{\star} I(\boldsymbol{w}, \boldsymbol{x}_i, \xi_i) = \boldsymbol{w}, \tag{9}$$

$$0 \leq \alpha_i \leq C. \tag{10}$$

Generally, this problem can not be reduced to the dual variable formulation since there is no closed form solution for $\boldsymbol{w}$. However, for a number of applications and particular types of distributions we will approach this problem by an approximate solution. Let us note, however, that for the case of traditional data samples, $p(\boldsymbol{x}|\boldsymbol{x}_i,\boldsymbol{r}_i)=\delta(\boldsymbol{x}\text{-}\boldsymbol{x}_i)$, the problem (7)-(10) reduces to the standard soft margin SVM.

2.3 Iterative Solution

The general approach for solving the optimization problem (1)-(3), or its equivalent (7)-(10), is to apply an iterative procedure to obtain an approximate solution. This type of optimization approach has been described in (Bezdec & Hathaway, 2003), and applied in (Bi & Zhang, 2004). Generally, the nature of SVM-related methods is that they try to find the Support Vectors, i.e. the samples which lie closest to the discrimination surface. 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 (Graepel & Herbrich, 2003; Fung, Mangasarian & Shavlik, 2002; Bhattacharyya, Pannagadatta & Smola, 2004) for examples of such solution 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.

Here we show that a similar approach holds for the case of distributions. Let us consider the following result.

Lemma. Consider the optimization problem (1)-(3). There exists a set of samples $\{\boldsymbol{x}_i^*,\ i{=}1,\ldots L\}$ such that the optimal separating hyper-plane $\boldsymbol{w}^*$ for the set $\{\boldsymbol{x}_{\mathrm{i}}^*\}$ coincides with the solution of the problem (1)-(3).

Proof. If the dimensionality of the feature space is less than the number of (non degenarative) samples, then the proof is trivial. Otherwise, for high (infinite) dimensional feature spaces, let $\boldsymbol{w}$ be the solution of (1)-(3). Since $p(\boldsymbol{x}|\boldsymbol{x}_i,\boldsymbol{r}_i)$ is a p.d.f., then for any $\boldsymbol{w}$, and any fixed $\boldsymbol{x}_i$ and $p(\boldsymbol{x}|\boldsymbol{x}_i,\boldsymbol{r}_i)$, function $I(\boldsymbol{w},\boldsymbol{x}_i,t)$ is continuous and monotonically increases with t. Then, according to the Weicherstrass theorem, there exists $\boldsymbol{x}_i^*$ such that $I(\boldsymbol{w},\boldsymbol{x}_i,-y_i\boldsymbol{w}\boldsymbol{x}_i^*)=\eta$. Thus, the optimization problem (1) can be reformulated in terms of minimizing the regularized risk functional with the cost function $c(.,.,.)$ that only depends on the choice of $\boldsymbol{x}_i^*$ for any fixed $\boldsymbol{x}_i$, and $p(\boldsymbol{x}|\boldsymbol{x}_i,\boldsymbol{r}_i)$: $c(\boldsymbol{x}_i, y_i, \boldsymbol{w}) = C\sum\limits_{i=1}^{L} \max(0, \eta - I(\boldsymbol{w}, \boldsymbol{x}_i, -y_i\boldsymbol{w}\boldsymbol{x}_i^*))$, and hence the semi-parametric generalized representer theorem (Scholkopf, Herbrich & Smola, 2001) can be applied. Thus, there exists a representation $\boldsymbol{w} = \sum\limits_{i=1}^{L} \alpha_i \boldsymbol{x}_i^* + b$ which coincides with the solution of the problem (1)-(3).

The considered result, however, does not provide a way for finding neither $\boldsymbol{x}_i^*$ nor α_{i} in the desired representation.

A general iterative scheme includes iterations through a series of (currently) optimal samples for a given approximation to the hyper-plane. Then the margin is maximised again for the modified samples, etc. The major disadvantage of this type of approaches is the convergence of the described procedure. Even if it is possible to prove that these iterations converge, the rate of convergence is unknown and may appear to be unreasonably low. One of the successful attempts is (Tsochantaridis et al., 2004), where the rate of convergence in the similar problem was estimated. However, the problem considered in (Tsochantaridis et al., 2004) is a quite specific case, where iterations are carried out in the context of structured *outputs*.

We thus propose a simple 2-step method to obtain an approximate solution to (1)-(3).

3 Hyper-plane Projection Method

From now on, let us deal with 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) for the training set of means $\{\boldsymbol{x}_i, y_i\}$ in the feature space induced by $K(.,.)$. 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:

- solve a standard SVM optimization problem for the means $\boldsymbol{x}_i$. The solution is $(\boldsymbol{w}_0, b_0)$;
- calculate the projections of $p(\boldsymbol{x}|\boldsymbol{x}_i, \boldsymbol{r}_i)$ on $\boldsymbol{w}_0$. This results in a 1-D optimization problem (see Section 3.1);
- solve the 1-D problem according to the given value of η;
- compute the inverse projection. This results in a modified training set $\boldsymbol{x}_i^*$ (see Section 3.2);
- solve a standard SVM optimization problem for the samples $\boldsymbol{x}_i^*$.

The detailed explanation of the projection steps is presented below. Please also refer to Figure 1 for an illustration.

3.1 Direct Projection

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

$$\mu_j = E[\boldsymbol{w}_0 \Phi(\boldsymbol{x}_j) + b_0] = \sum_{i=1}^{L} y_i \alpha_i \int_{\mathrm{X}} K(\boldsymbol{x}_j, \boldsymbol{x}_i) p(\boldsymbol{x}|\boldsymbol{x}_i, \boldsymbol{r}_i) dx + b_0, \tag{11}$$

$$\sigma_j^2 = E[(\boldsymbol{w}_0 \Phi(\boldsymbol{x}_j) - \mu_j)^2] = \sum_{i,k=1}^{L} y_i y_k \alpha_i \alpha_k \int_{\mathrm{X}^2} K(\boldsymbol{x}_i, \boldsymbol{x}_k) p(\boldsymbol{x}|\boldsymbol{x}_i, \boldsymbol{r}_i) p(\boldsymbol{x}'|\boldsymbol{x}_k, \boldsymbol{r}_k) dx dx' - \mu_j^2. \tag{12}$$

These 1-D p.d.f.s 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

$$\int_{y_j f(\boldsymbol{x}) \geq \chi_j} \pi(\chi|\mu_j, \sigma_j) d\chi \geq \eta. \tag{13}$$

It can be solved easily and results in some threshold constant c_η^j such that $\chi_j = f(\boldsymbol{x}_j^*) = c_\eta^j$. The difficulty arises for the original samples that have been classified incorrectly. Currently, we propose to neglect these samples. The reasoning is simple: one would not like to update the classification based on doubtful sample distributions, which means have not been classified correctly.

3.2 Inverse Projection

Next, given the set of χ_j one has to find an inverse projection of χ_j into the feature space. Obviously, this transformation is not unique and some criteria are required to define it. At this step, it is hard to control the margin; while the constraint (2) can still be satisfied. One would like to find $\boldsymbol{x}_{\mathrm{j}}^*$ such that the inequality in (2) holds (or 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 to obtain the inverse projection $\boldsymbol{x}_j^*$ of the χ_j:

$$\begin{aligned} &\boldsymbol{x}_j^* = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{x}_j, \boldsymbol{r}_j), \\ &s.t.\ f(\boldsymbol{x}) = c_\eta^j \end{aligned} \tag{14}$$

Due to a lack of space, we omit the formal reasoning behind this criterion. A useful intuition 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 (14) is a constrained optimization problem, which has to be solved. It results in the desired inverse projections $\boldsymbol{x}_j^*$ which form the new training set. The standard SVM solution for the obtained training set approximates the solution of the initial problem (1).

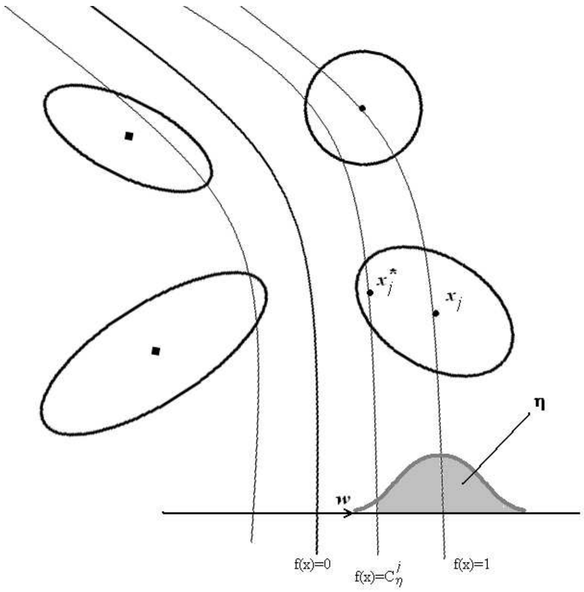

Figure 1: The illustration of the hyper-plane projection method. Refer to the text for the notations.

4 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 results in the following algorithm. The exact linear optimization problem reduces to:

$$\min \tfrac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{L} \alpha_i (1 - \operatorname{erf} \tfrac{y_i(\boldsymbol{w}\boldsymbol{x}_i + b) - 1}{\sqrt{2\boldsymbol{w}^T \Sigma_i^{-1} \boldsymbol{w}}}), \tag{15}$$
$$\sum_{i=1}^{L} y_i \alpha_i \left(\boldsymbol{w}^T \Sigma_i^{-1} \boldsymbol{w}\right)^{-\frac{1}{2}} \exp(-\tfrac{(y_i(\boldsymbol{w}\boldsymbol{x}_i + b) - 1)^2}{2\boldsymbol{w}^T \Sigma_i^{-1} \boldsymbol{w}}) = 0,$$
$$\boldsymbol{w} = \sum_{i=1}^{L} y_i \alpha_i \boldsymbol{x}_i^\star \left(\boldsymbol{w}^T \Sigma_i^{-1} \boldsymbol{w}\right)^{-\frac{1}{2}} \exp(-\tfrac{(y_i(\boldsymbol{w}\boldsymbol{x}_i + b) - 1)^2}{2\boldsymbol{w}^T \Sigma_i^{-1} \boldsymbol{w}}).$$

Note that the term in the exponent is a Mahalanobis distance from $\boldsymbol{x}_{\mathrm{i}}$ to the line $y_i(\boldsymbol{w}\boldsymbol{x}_i + b) = 1$. This formulation can be used to control the precision of the approximate solutions. We now present the non-linear version of the algorithm using the Gaussian isotropic RBF kernel with parameter δ. The method of hyper-plane projections can be applied using the following results for the direct step:

$$\begin{aligned} \mu_j &= \sum_{i=1}^{L} y_i \alpha_i D(\boldsymbol{x}_j, \boldsymbol{x}_i) + b, \\ \sigma_j^2 &= \sum_{i,k=1}^{L} y_i y_k \alpha_i \alpha_k M(\boldsymbol{x}_j, \boldsymbol{x}_i, \boldsymbol{x}_k) - \mu_j^2, \end{aligned} \tag{16}$$

where

$$\begin{aligned} D(\boldsymbol{x}_k, \boldsymbol{x}_i) &= \left|\Sigma_i^{-1} + \delta\right|^{-\frac{1}{2}} \\ &\quad \exp(-\frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{x}_i)^T \delta \Sigma_i^{-1} (\Sigma_i^{-1} + \delta)^{-1} (\boldsymbol{x}_k - \boldsymbol{x}_i)), \\ M(\boldsymbol{x}_k, \boldsymbol{x}_i, \boldsymbol{x}_j) &= \left|\Sigma_i^{-1}\Sigma_j^{-1} + \delta(\Sigma_i^{-1} + \Sigma_j^{-1})\right|^{-\frac{1}{2}} \\ &\quad \exp(-\frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{x}_i)^T \Omega (\boldsymbol{x}_k - \boldsymbol{x}_i)), \text{ where} \\ \Omega &= \delta \Sigma_i^{-1}((\Sigma_i^{-1}\Sigma_k^{-1})^{-1}(\Sigma_i^{-1} + \Sigma_k^{-1}) + \delta)^{-1}. \end{aligned} \tag{17}$$

The inverse projection can then be carried out by solving the following optimization problem:

$$\begin{aligned} \boldsymbol{x}_j^* &= \arg\min_{\boldsymbol{x}} (\boldsymbol{x} - \boldsymbol{x}_j)^T \Sigma_j^{-1} (\boldsymbol{x} - \boldsymbol{x}_j), \\ s.t. &\sum_{i=1}^{L} y_i \alpha_i \exp(-\delta(\boldsymbol{x} - \boldsymbol{x}_i)^2) + b = c_\eta^j. \end{aligned} \tag{18}$$

This problem has the following approximate analytical solution:

$$\begin{aligned} \boldsymbol{x}_j^* &= (I + 2\gamma\delta c_\eta^j \Sigma_i)^{-1} \\ &\quad (\boldsymbol{x}_j + 2\gamma\delta \sum_{i=1}^{L} y_i \alpha_i \exp(-\delta(\boldsymbol{x}_j - \boldsymbol{x}_i)^2) \Sigma_i \boldsymbol{x}_i), \end{aligned} \tag{19}$$

for some positive constant γ, which has to be chosen to satisfy the constraint in (18).

Despite the cumbersome expressions above, the real computations are significantly simplified, since for high-dimensional input data diagonal covariance matrices are often used.

4.1 Links to Related Methods

A related general problem was considered by Leen (1995) with a different aspect of modifying the risk functional according to some prior input distribution. A very promising (although tricky) approach is known as Vicinal Risk Minimization (Vapnik, 2000).

The most related methods for the particular case of Gaussian distributions discrimination were proposed recently by Bhattacharyya, Pannagadatta and Smola (2004) and Bi and Zhang (2004) and deal with uncertain data. The training samples are considered to be given with some uncertainty, presented in a form of Gaussian distributions.

The method of Bhattacharyya, Pannagadatta and Smola (2004), which is aimed at classifing datasets with missing (uncertain) samples, considers margin maximization under the following constraints:

$$Pr[y_i(\boldsymbol{w}\boldsymbol{x}_i + b) \geq 1 - \xi_i] \geq \eta, \;\; \xi_i \geq 0. \tag{20}$$

As one can see, this formulation is similar to the constraint (2). The relevance can be shown by appling the Chebyshev inequality in order to obtain the corresponding deterministic constraint from (20).

Bi and Zhang (2004) instead deal with the constraint

$$Pr[y_i(\boldsymbol{w}\boldsymbol{x}_i + b) \geq 1 - \xi_i] \leq \eta, \;\; \xi_i \geq 0. \tag{21}$$

This type of approach leads to more robust models, since, in the end, the less certain samples obtain the least weights. However, this intuition is hardly applicable for the problems we are aiming to. It was developed for a specific medical applications considered by the authors.

Despite these differences, both models lead to a Second Cone Programming Optimization (SOCP) problem. This optimization problem can be solved numerically by Interior Point methods (Nesterov & Nemirovskii, 1993), which are, however, quite costly in terms of computational time. The proposed approximate approach involves standard SVM QP optimization only.

Moreover, due to the computational costs of the SCOP, both papers mention the need of an approximate solution. The approximation is also based on modifying the means. However, the update rule of Bi and Zhang (2004) is very straightforward, as it suggests updating the samples along the $\boldsymbol{w}$ without taking into account any information on the covariance of the parent distribution.

The related update rule can be easily derived from the approximate formulation proposed by Bhattacharyya, Pannagadatta and Smola (2004). This, however, was not done by the authors. Nevertheless, let us note that for the linear case of (18) the exact solution of the inverse projection for the Gaussians is given by

$$\boldsymbol{x}_j^* = \boldsymbol{x}_j + \frac{c_\eta^j - (\boldsymbol{w}\boldsymbol{x}_j + b)}{\boldsymbol{w}^T \Sigma_j \boldsymbol{w}} \Sigma_j \boldsymbol{w}. \tag{22}$$

This linear case almost perfectly coincides with the result which could be derived from the formulation considered by Bhattacharyya, Pannagadatta and Smola (2004).

The significant difference with our method lies in the way the algorithms were kernelized. Nonlinearity through the kernel trick is introduced under a number of assumptions in all the methods. It is not possible to “kernelize” the initial algorithms directly. This is also the case for the general original problem (7)-(10), presented above (and for the particular case of Gaussian distributions as well). However, the developed approximate procedure deals with precise kernels directly by using the feature space of the original SVM for making projections, hence this drawback is partly avoided.

Finally, as it was mentioned above, a number of papers devoted to invariant learning are based on discriminating different objects in the input space. Methods aimed at direct margin maximization were recently proposed by Graepel and Herbrich (2003), Fung, Mangasarian and Shavlik (2003). Furthermore, a general method for defining an SVM kernel function for pairs of distributions was presented by Kondor, Jebara and Howard (2004).

5 Experiments

As mentioned in the introduction, there is a broad field of applications of the proposed approach. This includes problems from speech processing, biometric client identification, object categorization, etc.

However, the presented and some related approaches can also be used for handling invariances in pattern recognition problems. A similar setting can also be applied for the task when data sample are given with some uncertainty of a known nature. Here, we present experiments on invariant face image classification.

Let us start with a simple synthesized 2-D data example, which nicely illustrates the presented approach.

5.1 Toy Data Experiments

The task to be solved is a 2-class classification of 2-D Gaussians (see Figure 2). The dataset contains several samples of each class; their covariances are illustrated by ellipses around the means. The modified training set of $\{\boldsymbol{x}_{-}^{*}\}$ is shown with black dots, which form a curve for different value of γ. The samples $\boldsymbol{x}_j^*$ coincides with the means $\boldsymbol{x}_j$ for $\gamma = 0$, and tend to the decision boundary according to the covariance of their parent distribution, as γ increases. Final samples (shown by filled boxes) correspond to the $\eta = 0.9$. The modified decision boundary $f^*(\boldsymbol{x}) = 0$ is shown with a thick line.

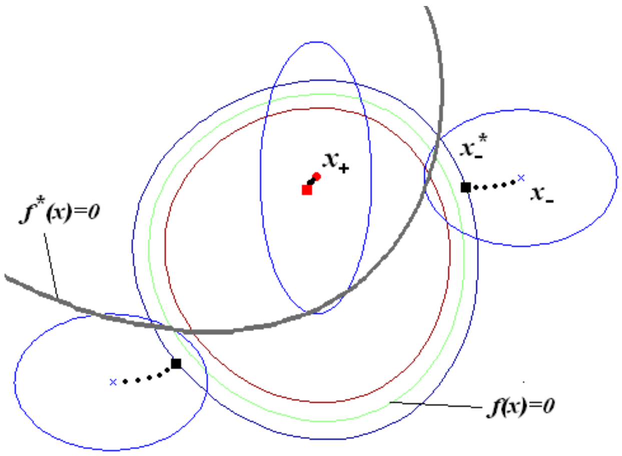

Figure 2: Toy Data Discrimination

5.2 Invariant Image Classification

The task of invariant image classification is still a challenging problem in computer vision. While a number of approaches for dealing with simple images like Optical Characters exist, methods for complex real-life images are still under development.

A natural way to model invariances is to consider how the desired transformation change the input samples. Generally, this dependence is very complex and highly non linear, hence difficult to model. Linear approximations are thus used instead.

5.2.1 Tangent Vectors

Suppose we are given grey-scaled images on the plane (ζ, ψ). The intensity of the image is defined by some function $U(\xi, \psi)$. It provides a high-dimensional input vector $\boldsymbol{x}$ for a given discrete set of coordinates (ξ, ψ). Then, it is possible to introduce the invariant transformations by the corresponding *tangent vectors* (Simard et al., 1998). Consider the transformation t_α defined by the set of parameters α in some region of $D \in R^2$:

$$t_\alpha : D \in R^2 \mapsto t_\alpha(D) \in R^2. \tag{23}$$

This transformation is assumed to be differentiable with respect to α and $(\zeta, \psi) \in D$, and reduces to the identity transformation for some value of α^0. The linear approximation to the invariant manifold is thus

$$S(U, \alpha) = U + \sum_{j=1}^{J} (\alpha_j - \alpha_j^0) L_{\alpha_j}(U), \tag{24}$$

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

$$L_{\alpha_j}(U) = \left.\frac{\partial S(U,\alpha)}{\partial \alpha_j}\right|_{\alpha=\alpha_0}. \tag{25}$$

Tangent vectors ℓ_i^j can be obtained by discretising the result of applying the operators L_{α_j} to the continuous image U which correspond to a discrete sample $\boldsymbol{x}_i$. Hence, ℓ_i^j denotes the tangent vector, which corresponds to the j^{th} invariant transformation of the sample $\boldsymbol{x}_i$.

The corresponding tangent vector based Gaussian distribution is as follows:

$$\begin{array}{l} P(\boldsymbol{x}|\boldsymbol{x}_i, \{\ell_i^1, \dots \ell_i^J\}) = \frac{\exp(-(\boldsymbol{x}-\boldsymbol{x}_i)^T L_{x_i}^{-1}(\boldsymbol{x}-\boldsymbol{x}_i))}{(2\pi)^{N/2}|L_x|^{1/2}}, \\ \text{where } L_x^{-1} = \sum_{j=1}^{J} \frac{\ell_i^j \ell_i^{j\,T}}{2\gamma_j^2 \ell_i^{j\,2}}. \end{array} \tag{26}$$

We consider these distributions as a training set for our algorithm. Parameter γ_j controls the effective width of the distribution for the given direction of ℓ_i^j. It can be fixed to some value according to some prior knowledge, since the resulting transformed images can be visualized. We used the same value of γ for all the invariances. Generally, the resulting Gaussians have full-ranked covariances, that dramatically slows down the overall computations.

5.2.2 Face Data Classification

We conducted experiments using images of the faces detected from movie scenes using a face detector, described in (Schneiderman & Kanade, 2000). There is a total of 2899 images in the database. The data is available at [http://www.robots.ox.ac.uk/∼vgg/data]. The original image dimension is 81 by 81 pixels, and a grey-scale level is 8 bit. We present an approach to the problem of binary classification of the main actor against all the other captured images. Example images and their corresponding labels are presented in Figure 3.

We used the following experimental setting: the training set consisted of 300 samples, taken randomly from the database. The rest 2599 images were used as a testing set. Two basic invariant transformations were considered: scalings and rotations. Finite difference vectors, obtained as a difference of an original and a transformed images were used instead of the original tangent vectors. The reason is that real tangent vectors fail for such complex and non-smooth images as faces when the transformation is more than infinitely small. Different tangent vectors were used for the rotations to the left and right, as well as for the zoom in and zoom out scalings.

Figure 3: Examples of training (left) and testing (right) face data images with the corresponding class labels below.

The parameters of the algorithms were chosen according to the minimum of cross-validation error over the training set, resulting in the following values: $\delta = 2 \cdot 10^{-5}$, $C = 100$, $\gamma = 1000$. Table 1 presents training and testing errors obtained with SVM with Gaussian RBF kernel (SVM), SVM trained with virtual samples (VSV SVM), and the developed method (SVM Gauss). We consider the

Table 1: Classification accuracies for the compared margin-based algorithms.

Algorithm	Tr.Err.,%	Test.Err.,%	Time, s
SVM	0.5	11.2	0.75
VSV SVM	0.4	9.9	9.6
SVM Gauss	0.5	9.7	2.8

Virtual SV method as a state-of-the-art approach to invariant learning with SVM-based methods. Given unlimited computational resources, this is currently the method of practical choice.

While both methods are statistically significantly better than baseline SVM (with 95% confidence), no significant improvement was observed in comparison with Virtual Support Vectors in terms of the testing error. However, the proposed approach is faster in terms of operational time. This advantage is even more significant if the covariance matrices are diagonal.

6 Discussion and Conclusions

While classical SVMs discriminate between example points of two classes, we proposed in this paper a novel SVM formulation to discriminate between example distributions of two classes, while still keeping advantages of SVMs such as margin maximization and kernel trick. This extension can be used for many different settings, including principled incorporation of invariances described by distributions, which was illustrated in this paper. Other possible uses of this model include the possibility to maximize the margin for problems that were traditionally solved by generative models and log likelihood ratios such as speech processing.

Since the direct solution was not tractable, the paper presented an approach for an approximate solution of the optimization problem. This approach consists of two simple projection steps, resulting in a modified training set. Thus, possible problems with the convergence of an iterative scheme are avoided. Next, the algorithm was turned into a nonlinear version through the usual "kernel trick". The feature space of the original SVM (trained on the means of the distributions) is exploited for the latter. The algorithm demands a standard SVM QP solver only. The case of Gaussian distributions was considered in details, and some links to related research were provided.

The algorithm was applied to the real problem of invariant face image classification. The knowledge on invariances was incorporated into the algorithm by considering a special type of distributions, based on tangent vectors. The comparison to the state-of-the-art virtual support vector was provided. Currently, the method is found to be competitive with the state-of-the-art, and the proposed solution was preferable in terms of computational time.

Concerning invariant learning problems, the basic advantage of the proposed method is that it maximizes the margin between invariance-modelling distributions directly. Note that approaches based on measuring the pair-wise overlap between distributions are subject to the curse of dimensionality if the linear approximation to the invariant transformations is used.

For a practitioner, the algorithm provides a nice feedback. As $\boldsymbol{x}_j^*$ are known, these samples can be visualized. An interesting question is whether these samples coincide with human's intuition to be the most discriminative. To our knowledge, the answer is positive most of the times.

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

Bezdec, J. and Hathaway, R. (2003). Convergence of alternating optimization. *Neural, Parallel Sci.Comput.,*, *11*, 351–368.

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.

Burges, C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, Vol. 2, Number 2, p. 121-167, Kluwer Academic Publishers.

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.

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.

Leen, T.K., (1995). From data distributions to regularization in invariant learning. *Neural Computation*, vol. 7, no. 5, pp. 974-981.

Jaakkola, T. and Haussler, D., (1999). Exploiting generative models in discriminative classifiers. In: M.S.Kearns, S.A.Solla, D.A.Cohn (eds.) *Advances in Neural Information Processing Systems*, vol. 11, pp. 487-493, MIT Press.

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

Nesterov, Y., Nemirovskii, A., (1993) Interior Point Algorithms in Convex Programming. *Studies in Applied Mathematics, 13*, SIAM, Philadelphia.

Rabiner, L.R., (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE*, vol. 77, 2, February 1989, pp. 257-286.

Scholkopf, B., Herbrich, R., and Smola, A., (2001) A Generalized Representer Theorem. In: Helmbold, D. and Williamson, B., (eds.) *Proc. of COLT/EuroCOLT 2001*, *LNAI2111*, pp. 416-426, Springer-Verlag, Berlin.

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

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

Tsochantaridis, I., Hofmann, T., Joachims, T., Yasemin, A., (2004). Support Vector Machine Learning for Interdependent and Structured Output Spaces. *21th Int. Conf. on Machine Learning,* Banff, Canada.

Vapnik, V., (1998). Statistical Learning Theory. J.Wiley, NY, 1998.

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