Learning to Compare using Operator-Valued Large-Margin Classifiers

Andreas Maurer[*]
Adalbertstraße 55
D-80799 München
andreasmaurer@compuserve.com

Abstract

The proposed method uses homonymous and heteronymous example-pairs to train a linear preprocessor on a kernel-induced Hilbert space. The algorithm seeks to optimize the expected performance of elementary classifiers to be generated from single future training examples. The method is justified by PAC-style generalization guarantees and the resulting algorithm has been tested on problems of geometrically invariant pattern recognition and face verification.

1 Introduction

A striking feature of human vision is the ability to reliably recognize complex objects after having seen only a single representative specimen, often under a different spatial perspective. Attempts to explain this astonishing generalization performance rely on the concept of transfer. From the earliest childhood on the semantic equivalence of visual phenomena involving different scales and perspectives of identical objects is observed, leading to an accumulation of experience which can be applied to the recognition of novel categories [10].

The potential utility of machine learning algorithms possessing a similar power of meta-generalization is obvious. One possible approach [5] uses the accumulated experience to learn a representation Φ, mapping the input data to the euclidean space $\mathbb{R}^d$. Once a training example for a novel category or concept has been presented, the class membership of any other specimen can be determined by thresholding the euclidean distance $\|\Phi(x) - \Phi(x')\|$ between the representations of example and specimen.

The method presented here is of this kind and belongs to the category of kernel-techniques [6]: A fixed positive definite kernel defines a map ψ to embed the input data in a Hilbert space H. The training data is then used to learn a linear transformation $T : H \to \mathbb{R}^d$ and the combined map $\Phi = T \circ \psi$ is employed to represent future data.

Similar to [5] training is based on homonymous and heteronymous pairs of examples. In section 2 we present a probabilistic model for the generation of such pairs and derive a compact expression for the risk incurred by using a particular linear transformations T.

[*] www.andreas-maurer.eu

The Hilbert-Schmidt inner product casts this risk functional into a form dependent only on the operator T^*T. This form is equivalent to the more familiar risk of vector-valued classifiers (such as SVM's). This equivalence leads to generalization guarantees and gives rise to a regularized training algorithm for the transformation T.

The algorithm has been tried with good results on the recognition of digits under arbitrary rotations, scalings and combined rotations and scalings, the recognition of objects under spatial rotations and face recognition.

Some of this work can be found in [8]. The algorithm and the experimental results are new.

2 Risk functionals for representations

$\mathcal{X}$ will denote the input space containing the data to be processed. We will not require a fixed alphabet of labels $\mathcal{Y}$ and a distribution μ on $\mathcal{X} \times \mathcal{Y}$, as is customary in learning theory, instead we assume that there is a *pair oracle*, that is a probability distribution ρ on $\mathcal{X}^2 \times \{-1, 1\}$ with the following interpretation:

- $\rho(x, x', r)$ is the probability to encounter the pair $x, x' \in \mathcal{X}$, being *homonymous* if $r = 1$ (having the same name, label, class or category) or being *heteronymous* if $r = -1$ (of different name, label, class or category).

The labeled pairs in $\mathcal{X}^2 \times \{-1, 1\}$ are also called *equivalence constraints* by some authors (e.g. [3]). Here it suffices to postulate the existence of ρ as an axiom, but is is possible to derive ρ from multi-class learning tasks or multi-task environments to justify the use of the risk functional defined below (see [8]).

Now fix, once and for all, a feature-map $\psi : \mathcal{X} \to H$ where H is a Hilbert space, such that $\|\psi(x) - \psi(x')\| \leq 1, \forall x, x' \in \mathcal{X}$. This boundedness assumption is a technical requirement for our theoretical results and not implicitely exploited by our algorithm.

Let T be an arbitrary linear transformation $T : H \to \mathbb{R}^d$ for some integer d. We want to decide if two inputs x and x' are homonymous or heteronymous and base this decision only on the distance between the represented points $T\psi(x)$ and $T\psi(x')$. Evidently this involves thresholding this distance at a fixed positive constant, and because this constant can be absorbed in the transformation we can assume it to be one. The probability of error of this decision rule, as (x, x', r) are drawn from the pair oracle ρ, is then given by

$$R(T, \rho) = \Pr_{(x,x',r)\sim\rho} \{r(1 - \|T\psi(x) - T\psi(x')\|) \leq 0\}. \tag{1}$$

Any bound on this risk functional can be converted to an error bound for elementary threshold classifiers, as soon as we are furnished with (single) examples of the classes involved.

Since the precise nature of the distribution ρ is hidden from the learner, the transformation T has to be chosen on the basis of a finite sample $S = ((x_1, x_1', r_1), ..., (x_m, x_m', r_m)) \in (\mathcal{X}^2 \times \{-1, 1\})^m$ of labeled pairs generated in m independent, identical trials of ρ.

The problem of selecting T to minimize (1) is equivalent to learning a classifier for the binary classification problem implied by the pair oracle ρ with input space $\mathcal{X}^2$ and a hypothesis space parametrized by linear transformations. It is also equivalent to learning the pseudometric $(x, x') \mapsto \|T\psi(x) - T\psi(x')\|$ and the kernel $(x, x') \mapsto \langle T^*T\psi(x), \psi(x')\rangle$. The loss of generality in our approach just corresponds to the selection of a constrained hypothesis space for the purpose of better generalization.

3 Operator-valued large-margin classifiers

With H_2 we denote the real vector space of symmetric operators on H satisfying $\sum_{i=1}^{\infty} \|Te_i\|^2 \leq \infty$ for every orthonormal basis $(e_i)_{i=1}^{\infty}$ of H. For $S, T \in H_2$ and an orthonormal basis (e_i) the series $\sum_i \langle Se_i, Te_i \rangle$ is absolutely summable and independent of the chosen basis. The number $\langle S, T \rangle_2 = \sum \langle Se_i, Te_i \rangle$ defines an inner product on H_2, called the Hilbert-Schmidt inner product, making H_2 into a Hilbert space whose elements are called Hilbert-Schmidt operators. We denote the corresponding norm with $\|.\|_2$ (see Reed and Simon [9] for background on functional analysis).

For every $v \in H$ we define an operator Q_v by $Q_v w = \langle w, v \rangle v$. Then $Q_v \in H_2$ and $\|Q_v\|_2 = \|v\|^2$. If $T : H \to \mathbb{R}^d$ then $T^*T \in H_2$ and $\langle T^*T, Q_v \rangle_2 = \|Tv\|^2$, a fundamentally important formula, which allows to rewrite the risk functional (1) as

$$R(T, \rho) = \Pr_{(x, x', r) \sim \rho} \left\{ r \left(1 - \left\langle T^*T, Q_{\psi(x) - \psi(x')} \right\rangle_2 \right) \leq 0 \right\}.$$

On H_2 the positive operator T^*T can therefore be regarded as a linear classifier, and the risk above is just the expected error of this classifier on the labeled data-point $\left(Q_{\psi(x)-\psi(x')}, r\right) \in H_2 \times \{-1, 1\}$ as (x, x', r) are drawn from the distribution ρ. Inspired by the theory of linear large-margin classifiers we fix a Lipschitz function $f \geq 1_{(-\infty, 0]}$ with Lipschitz constant c_f and define the empirical risk functional $\hat{L}_f$ for a positive semidefinite operator $P \in H_2$ and a sample $S = ((x_1, x_1', r_1), ..., (x_m, x_m', r_m)) \in \left(\mathcal{X}^2 \times \{-1, 1\}\right)^m$

$$\hat{L}_f(P, S) = \frac{1}{m} \sum_{i=1}^{m} f\left(r_i \left(1 - \left\langle P, Q_{\psi(x_i) - \psi(x_i')} \right\rangle_2 \right) \right). \tag{2}$$

Theorem 1: Let $\mathcal{T}$ be some class of linear transformations $T : H \to \mathbb{R}^d$. Then for every $\delta > 0$ we have with probability greater than $1 - \delta$ in a sample $S \sim \rho^m$, that for every $T \in \mathcal{T}$

$$R(T, \rho) \leq \hat{L}_f(T^*T, S) + \frac{1}{\sqrt{m}} \left(4\, c_f \sup_{A \in \mathcal{T}} \|A^*A\|_2 + \sqrt{\frac{\ln(1/\delta)}{2}} \right).$$

The proof (see [8], Theorem 3) just transcribes the results in [2] to the setting of Hilbert-Schmidt operators. Using standard techniques (e.g. Lemma 15.5 in [1]) the theorem can be converted into a uniform upper bound for *all* linear transformations. Up to a logarithmic term minimization of this upper bound is equivalent to minimization of a regularized objective function of the form

$$\Lambda(T) = \Omega(T^*T) = \hat{L}_f(T^*T, S) + \frac{\lambda}{\sqrt{m}} \|T^*T\|_2,$$

where $\lambda > 0$ is a regularization parameter. This is the method proposed by the author.

4 A training algorithm

Any minimizer of Ω must be a linear combination of the $Q_{\psi(x_i) - \psi(x_i')}$ and therefore leave invariant the linear span M of the vectors $\psi(x_i) - \psi(x_i')$. A minimizer T of Λ must thus vanish on $M^\perp$ and its range can be at most m-dimensional. We can therefore replace H by M (now that we have the dimension-independent bound above) and set $d = m$. The most general form of the optimal T is

$$Tz = \sum_{i=1}^{d} \langle z, v_i \rangle e_i,$$

Table 1: Learning algorithm

```
Given sample $S$, regularization parameter $\lambda$, margin $\gamma$, learning rate $\theta$
initialize $\lambda' = \lambda / \sqrt{|S|}$
initialize $T = (v_1, ..., v_d)$ (where the $v_i$ are row-vectors)
repeat
    Compute $\|T^*T\|_2 = \left(\sum_{ij} \langle v_i, v_j \rangle^2\right)^{1/2}$
    For $i = 1, ..., d$ compute $w_i = 2 \|T^*T\|_2^{-1} \sum_j \langle v_i, v_j \rangle v_i$
    Fetch $(x, x', r)$ from sample $S$
    For $i = 1, ..., d$ compute $a_i \leftarrow \langle v_i, x - x' \rangle$
    Compute $b \leftarrow \sum_{i=1}^{d} a_i^2$
    If $r (1 - b) < \gamma$
        then for $i := 1, ..., d$ do $v_i \leftarrow v_i - \theta \left(\frac{r}{\gamma} a_i (x - x') + \lambda' w_i\right)$
        else for $i := 1, ..., d$ do $v_i \leftarrow v_i - \theta \lambda' w_i$
until convergence
```

where e_i is the canonical basis of $\mathbb{R}^d$ and the vectors v_i (which completely specify the transformation T), are linear combinations of the $\psi(x_i) - \psi(x_i')$, the coefficients of which have to be determined by the learning algorithm.

From now on we take f to be the hinge-loss with margin γ, that is $f(t) = 1 - t/\gamma$ if $t < \gamma$ and $f(t) = 0$ otherwise. Then f is convex and has Lipschitz constant $1/\gamma$. It follows that Ω is a convex functional on the positive semidefinite operators on M.To minimize Λ we use gradient descent. Despite the fact that Λ is not convex, the gradient technique is unlikely to become trapped in a local minimum:

Proposition 2: Suppose Ω is a continuous functional on the set of positive semidefinite $m \times m$ matrices. For any $m \times m$ matrix T define $\Lambda(T) = \Omega(T^*T)$. If Λ attains a stable minimum at T, then Ω attains a stable minimum at T^*T.

Proof: To arrive at a contradiction assume that Λ attains a stable minimum at T, but Ω does not attain a stable minimum at T^*T. Then there is a sequence A_n of positive semidefinite matrices such that $A_n \to T^*T$ and $\Omega(A_n) < \Omega(T^*T)$. So $A_n^{1/2} \to (T^*T)^{1/2} = |T|$. By polar decomposition write $T = U|T|$, with U unitary and define $T_n = UA_n^{1/2}$. Then $T_n \to U|T| = T$, but $\Lambda(T_n) = \Omega(T_n^*T_n) = \Omega(A_n) < \Lambda(T^*T)$, so Λ cannot attain a stable minimum at T $\square$

So if gradient descent arrives at a stable minimum T of Λ, then it must have found a global minimum by the convexity of Ω. Computing the gradient of the objective function with respect to the variables v_k then yields the algorithm given in Table 1.

Note that with any standard SVM algorithm Ω could be minimized over *all* of H_2. Minimization of $\Lambda(T) = \Omega(T^*T)$ however requires that the minimum be attained at a positive semidefinite operator $A = T^*T$ in H_2, a constraint which a standard SVM algorithm does not respect. The problem can nevertheless be cast in the form of convex optimization, because the positive semidefinite operators in H_2 form a convex set H_2^+ and Ω is convex. If we manage to deal with the difficult positivity constraint,we can then take as a solution $T = A^{1/2}$ for any minimizer A of Ω in H_2^+. By the above proposition this will not lead to global minimizers with different metric properties as our algorithm.

The regularizer penalizes the dimensionality of the sought transformation. It can therefore be expected that the optimal transformation is dimensionally sparse, in the sense that only few of its singular values are significantly different from zero. In practice this was indeed found to be the case: Typically only 8-36 singular values of T were larger than 2% of the spectral norm (maximal singular value).

5 Experiments with images

All the experiments reported below were carried out with margin $\gamma = 1$ and the regularization parameter $\lambda = 0.005$. The gradient descent was carried out for 10^6 steps with a constant learning rate $\theta = 0.01$.

Pixel vectors were normalized, otherwise there was no preprocessing. The feature map ψ was effected by the Gaussian RBF-kernel $\kappa(x, y) = (1/2) \exp\left(-4 |x - y|^2\right)$, where $|x - y|$ is the euclidean metric on pixel vectors.

Each experiment used a set of labeled training data from which homonymous and heteronymous pairs were generated at random with equal frequency and fed into the algorithm to produce the operator T. The sparsity mentioned in Table 2 is the number of singular values of T larger than 2% of the spectral norm.

The resulting representation $T \circ \psi$ was then applied to the pixel-vectors in the test set, which were equidimensional to those in the training set. Test and training set had no overlapping categories.

On the test set we measured two properties of the representation: The area under the ROC-curve (ROC area $T \circ \psi$) for the distance as a detector of class-equality. This can be regarded as an estimator for the probability that a homonymous pair is represented at a closer distance than an independently chosen heteronymous pair. The other quantity is the error (error $T \circ \psi$) of nearest neighbor classifiers when each category of the test set is represented by a *single* example, averaged over 10 runs with randomly chosen examples. Both quantities were also measured for the unrepresented but normalized input pixel vectors (ROC area input, error input).

Three experiments were made with planar geometrically invariant character recognition: Rotation invariance, scale invariance and combined rotation and scale invariance. For each invariance randomly transformed images of printed alpha characters were used for the training set and randomly transformed images of printed digits were used for the test set, the digit 9 being eliminated whenever rotation invariance was involved. Scaling ranged over a linear factor of two. All images hat 28x28 pixels.

The COIL-100 dataset (converted to grayscale) was used to test object recognition with invariance under certain spatial rotations. Every other image of Objects 1 to 80 was used for training, 81 to 100 for testing, as in [7].

The ATT dataset was used for an experiment with face-recognition. As in [5] images of the subjects numbered from 1 to 35 were used for training, those from 36 to 40 for testing.

The results of these experiments are summarized in table 2. The method significantly increased the ROC area and reduced the recognition error in all cases by comparison to the unprocessed input vectors. The greatest improvement was in the case of planar rotation and scale invariances. The results on the ATT dataset are very good and comparable to those reported in [5], but the ATT dataset is simply too easy and the more interesting AR-Purdue dataset was not available. The results with the COIL database are better than those reported in [7].

It must be emphasized that our method makes no assumption on the specific properties of image data, such as high correlations for neighboring pixels: In contrast to the approaches described in [5] and [7], our method would yield the same results if the images were sub-

Table 2: Experimental results

type of	rotation	scale	rot.+scale	spatial rot.	face
experiment	invariance	invariance	invariance	invariance	recognition
training set	alpha	alpha	alpha	COIL $\leq$80	ATT 1-35
nr of categories	20	52	20	80	35
nr of examples	2000	1560	4000	2880	350
sparsity of T	9	36	8	46	36
test set	digits\9	digits	digits\9	COIL $\geq$81	ATT 36-40
nr of categories	9	10	9	20	5
nr of examples	900	300	1800	720	50
ROC area input	0.597	0.69	0.54	0.845	0.934
ROC area $T \circ \psi$	0.9994	0.993	0.982	0.989	0.997
error input	0.716	0.508	0.822	0.375	0.113
error $T \circ \psi$	0.008	0.036	0.093	0.093	0

jected to any fixed but unknown permutation of pixel indices. The method can thus be used in any other domain where kernel techniques apply.

References

[1] M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, UK, 1999.

[2] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 2002.

[3] A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall. Learning a Mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research* 6: 937-965, 2005.

[4] J.Baxter, A model of inductive bias learning, *Journal of Artificial Intelligence Research* 12: 149-198, 2000

[5] S. Chopra, R. Hadsell and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification, *CVPR,* 2005

[6] Nello Cristianini and John Shawe-Taylor, Support Vector Machines, *Cambridge University Press*, 2000.

[7] F. Fleuret and G. Blanchard. Pattern recognition from one example by chopping, *NIPS*, 2005.

[8] A. Maurer, Generalization bounds for subspace selection and hyperbolic PCA. *Subspace, Latent Structure and Feature Selection. LNCS* 3940: 185-197, Springer, 2006

[9] Michael Reed and Barry Simon. *Functional Analysis*, part I of *Methods of Mathematical Physics, Academic Press*, 1980.

[10] A. Robins, Transfer in Cognition, in *Learning to Learn*, S. Thrun, L. Pratt Eds. Springer 1998.

[11] E. P. Xing, A. Y. Ng, M. I. Jordan, S. Russel. Distance metric learning, with application to clustering with side information. In S. Becker, S. Thrun, K. Obermayer, eds, *Advances in Neural Information Processing Systems* 14, Cambridge, MA, 2002. MIT Press.