Evaluation of distance-based discriminant analysis and its kernelized extension in visual object recognition

Serhiy Kosinov, Stéphane Marchand-Maillet

Department of Computer Science
University of Geneva, Switzerland
kosinov@cui.unige.ch

Abstract

This paper formulates the visual object recognition problem in the discriminant analysis framework and presents a kernelized version of the transformational approach of distance-based discriminant analysis. The sought transformation is found as a solution to an optimization problem formulated in terms of inter-observation distances only, using the technique of iterative majorization. The proposed approach is non-parametric, and can determine the dimensionality of the target space automatically since the process of feature extraction is fully embedded in the optimization procedure. Performance tests and experiments in the application of visual object and content-based image categorization demonstrate very competitive results in comparison to several popular existing techniques.

1. Introduction

Object recognition, as a fundamental computer vision problem, has long been a major focus of ongoing research, which lead to the development of a variety of methods and techniques proposed to date, e.g., [1, 2, 3, 4]. With the exception of several notable contributions, e.g., [5, 6], many approaches essentially treat the classifier as a black box completely isolated from the feature extraction process, which ultimately leads to suboptimal results. There exist numerous dimensionality-reducing data transformation methods originating from families as diverse as discriminant analysis techniques (LDA, DF-LDA, GDA), their advanced extensions (SHOSLIF, Fisherfaces), non-linear mappings (MDS, SOM) and neural networks (NeuroScale), yet, the answers to important questions, such as "How many dimensions are enough to discriminate among given classes?", still remain vague.

In order to address these issues, we investigate the problem of object categorization in the discriminant analysis framework and propose a method, alongside with its kernel-based extension, for finding a discriminative transformation of the original visual feature data. Based on the compactness hypothesis [7], the sought transformation specifically aims at improving the accuracy of the nearest neighbor (NN) classifier [8] and implicitly integrates the feature extraction process in the problem formulation. Additional constraints are imposed to prevent overfitting and thus improve generalization abilities of the proposed method.

The remainder of this paper is structured as follows. In section 2 we formulate the task of deriving a discriminant transformation as a problem of minimizing an asymmetric criterion based on the compactness hypothesis. In section 3 we review the iterative majorization method and demonstrate how it can be used to minimize the chosen criterion, while section 4 briefly describes a kernelized extension of the presented method. We detail our experimental results for both benchmark and real-world image data sets in section 5.

2. Problem formulation

Suppose that we seek to distinguish between two classes represented by data sets X and Y having N_X and N_Y m-dimensional observations, respectively. For this purpose, we are looking for such transformation matrix $T \in \mathbb{R}^{m \times k}$ such that $\{T : X \mapsto X', Y \mapsto Y'\}$, that the compactness hypothesis holds for either of the two classes in question, while its opposite is true for both.

Now, we must reiterate that our primary goal is to improve the NN performance on the task of discriminant analysis. This implies, first of all, that the sought problem formulation must relate only to the factors that directly influence the decisions made by the NN classifier, namely - the distances among observations. Secondly, in order to benefit as much as possible from the non-parametric nature of the NN, the sought formulation must not rely on the traditional class separability and scatter measures that use class means, weighted centroids or their variants which, in general, connote quite strong distributional assumptions. Finally, an asymmetric product form should be more preferable, justified as consistent with the properties of the data encountered in the target application area of multimedia retrieval and categorization [9]. More formally, these requirements can be accommodated by an optimization criterion expressed in terms of distances

among the observations from the two data sets as follows:

$$J(T) = \frac{\left(\prod_{i<j}^{N_X} \Psi\left(d_{ij}^W(T)\right)\right)^{\frac{2}{N_X(N_X-1)}}}{\left(\prod_{i=1}^{N_X}\prod_{j=1}^{N_Y} d_{ij}^B(T)\right)^{\frac{1}{N_X N_Y}}}, \qquad (1)$$

where the numerator and denominator of (1) represent the geometric means of the within- and between-class distances, respectively, and $\Psi(\cdot)$ denotes a Huber robust estimation function [10]. The choice of Huber function in (1) is motivated by its ability to switch from quadratic to linear penalty allowing to mitigate the consequences of an implicit unimodality assumption of the formulation. In the logarithmic form, criterion (1) is written as:

$$\begin{aligned}\log J(T) &= \frac{2}{N_X(N_X-1)}\sum_{i<j}^{N_X}\log\Psi\left(d_{ij}^W(T)\right) \\ &\quad -\frac{1}{N_X N_Y}\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}\log d_{ij}^B(T) \qquad (2)\\ &= \alpha S_W(T) - \beta S_B(T).\end{aligned}$$

Our previous studies [11, 12] have shown that neither straightforward gradient descent nor some of the state-of-the-art optimization routines are suitable for solving the above optimization problem mostly due to susceptibility to local minima, computational complexity and difficulties related to the discontinuities of the derivative of (2). However, by deriving some approximations of $S_W(T)$ and $S_B(T)$ one can make the task of minimizing $\log J(T)$ criterion amenable to a simple iterative procedure based on the majorization method, which we discuss in the following section.

3. Iterative majorization

3.1. General overview of the method

As stated in [13, 14, 15], the central idea of the majorization method is to replace the task of optimizing a complicated objective function $f(x)$ by an iterative sequence of simpler minimization problems in terms of the members of the family of auxiliary functions $\mu(x,\bar{x})$, where x and $\bar{x}$ vary in the same domain Ω. In order for $\mu(x,\bar{x})$ to qualify as a *majorizing function* of $f(x)$, the auxiliary function $\mu(x,\bar{x})$ is required to: (a) have a unique minimum, (b) always be greater than or equal to the original objective function, and (c) touch the surface of the original function at the *supporting point* $\bar{x}$.

Once an appropriate function $\mu(x,\bar{x})$ has been found, the iterative majorization algorithm proceeds as follows. After assigning an initial supporting point $\bar{x}$, the successor point x_s is found by minimizing $\mu(x,\bar{x})$. The obtained x_s subsequently becomes the next supporting point, and the process repeats until there is no improvement in the value of the objective function, i.e., convergence is reached.

3.2. Optimizing $\log J(T)$ by iterative majorization

It can be verified that majorization remains valid under additive decomposition [15]. Therefore, a possible strategy for majorizing (2) is to deal with $S_W(T)$ and $-S_B(T)$ separately and subsequently recombine their respective majorizing expressions.

We begin by noting that both logarithm and Huber distance are majorizable by linear and quadratic functions, respectively [15, 12]. This fact makes it possible to derive a majorizing function of $S_W(T)$ as follows:

$$\begin{aligned}S_W(T) &= \sum_{i<j}^{N_X}\log\Psi\left(d_{ij}^W(T)\right)\\ &\leq \sum_{i<j}^{N_X}\frac{\bar{w}_{ij}\cdot\left(d_{ij}^W(T)\right)^2}{2\Psi\left(d_{ij}^W(\bar{T})\right)} + K_1\\ &= \mu_{S_W}(T,\bar{T}), \qquad (3)\end{aligned}$$

where $T,\bar{T}\in\mathbb{R}^{m\times m}$, $\bar{T}$ is a supporting point for T, $\bar{w}_{ij}$ is a weight of the Huber function majorizer, as defined in [15], and K_1 is a constant that collects all of the terms that are irrelevant from the point of view of minimization with respect to T (see [12] for detailed derivations). In the matrix form, the above formulation can be expressed as:

$$\mu_{S_W}(T,\bar{T}) = \frac{1}{2}\mathbf{tr}\left(T^T X^T R X T\right) + K_1, \qquad (4)$$

where R is a square symmetric design matrix, as specified in [12].

As for $-S_B(T)$, we start out by expressing its every term using a second order Taylor series expansion[1] of the logarithm function around a supporting point $\bar{T}$. In the resulting formulation, the sum of Euclidean distances can be majorized by a rule based on the Cauchy-Schwarz inequality [13, 15] (again, see [12] for derivation details). In the matrix form, the resulting majorizing function of $-S_B(T)$ can be expressed as:

$$\begin{aligned}\mu_{-S_B}(T,\bar{T}) &= \frac{1}{2}\mathbf{tr}(T^T Z^T G Z T)\\ &\quad -2\mathbf{tr}(T^T Z^T G Z\bar{T}) + K_2, \qquad (5)\end{aligned}$$

[1] Note that the use of a Taylor series expansion instead of the negative logarithm makes it impossible to guarantee that majorization requirement (b) from section 3.1 is always satisfied, even though the minimum of the quadratic approximation lies in the direction of the steepest descent at all times. To counter rare yet theoretically possible consequences of such simplification we modify the quadratic Taylor series expansion whenever a problem is encountered so that the approximation becomes more conservative (i.e., the minimum is closer to the supporting point) and repeat current iteration.

where Z is the matrix obtained by joining X and Y together, row-wise, and G is a square symmetric design matrix of size $N = N_X + N_Y$, as defined in [12].

Finally, combining results (4) and (5), we obtain a majorizing function of the $\log J(T)$ optimization criterion:

$$\begin{aligned}\mu_{\log J}(T, \bar{T}) &= \alpha\mu_{S_W} + \beta\mu_{-S_B} \\ &= \frac{\alpha}{2}\mathbf{tr}\left(T^T X^T R X T\right) \\ &\quad + \frac{\beta}{2}\mathbf{tr}(T^T Z^T G Z T) \\ &\quad - 2\beta\mathbf{tr}(T^T Z^T G Z \bar{T}) + K_3, \quad (6)\end{aligned}$$

that can be used to find an optimal transformation T minimizing $\log J(T)$ criterion via the iterative procedure outlined in section 3.1. Similarly to the last terms in (4) and (5), K_3 is a constant that collects all of the other terms that are irrelevant from the point of view of minimization with respect to T.

While it is possible to minimize (6) by setting its derivative to zero and solving the resulting system of linear equations, it is often recommended [16] that a length-constrained solution be found, especially in the case of classifiers capable of achieving zero training error, to prevent overfitting. By incorporating the constraint into the Lagrangian, we obtain a standard trust-region subproblem formulation, for which efficient solution methods exist [17, 18]. Once solved for, an iterate of T can easily help determine the sufficient dimensionality of the target space k via the number non-zero singular values of T.

4. Kernel-based extension

According to the definition [19], kernel $K(x_i, x_j)$ implicitly projects x_i and x_j into some, possibly infinite-dimensional, feature space $\mathscr{F}$ and returns their dot product in that feature space $K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j)$, such that $\phi(x)$ is a mapping function that is never explicitly computed. Substituting in (1) the distances in the original feature space by those in $\mathscr{F}$ expressed solely via dot products $d_{ij}^{(\mathscr{F})} \equiv \sqrt{||\phi(x_i) - \phi(x_j)||^2} \equiv \sqrt{K_{ii} - 2K_{ij} + K_{jj}}$, we obtain a trivial kernel-based extension of the proposed method. Moreover, relying on the generalized Gaussian kernel $K_{ij} = e^{-\frac{d_{ij}}{2\sigma^2}}$, we derive a fomulation similar to (2), where $\log(x)$ is replaced by a concave function with similar properties from the point of view of majorization procedure. Therefore, the same algorithm is used, provided the necessary adjustments to the design matrices R and G.

5. Experimental results

For our object categorization experiments we chose a recently developed database ETH80 composed of entities corresponding to the basic level of human knowledge organization [20]. The database contains high-resolution color images of 80 objects from 8 different classes, for a total of 3280 images. The visual information for each image was represented by 286-dimensional feature vector containing 166 global color histogram and 120 Gabor filter texture descriptors extracted by the *Viper* system [21]. The training set comprised images taken one per class object viewed from a fixed position, while the rest was allocated to the test set. The results comparing NN classification performance in the original feature space (NN) and those derived by the proposed method (DDA+NN) or its kernel extension (KDDA+NN) are summarized in Table 1, showing an improvement over the baseline classifier.

Table 1: Results for the ETH80 image database

	% Error rate		
Object class	NN	KDDA+NN	DDA+NN
(1) Apple	4.47	1.00	0.75
(2) Car	14.47	5.00	5.78
(3) Cow	12.12	9.44	10.97
(4) Cup	3.09	0.94	2.22
(5) Dog	14.00	11.00	12.72
(6) Horse	14.47	9.44	13.16
(7) Pear	6.13	3.56	3.84
(8) Tomato	2.50	1.69	1.88

In addition to the tests mentioned above, we also explored empirically the influence of the DDA transformation on the performance of other popular classification methods, including NN as a baseline, on the real-world image categorization. For these experiments, three potentially overlapping image sets were selected from the Washington University annotated image collection [22], based on the presence of keywords "trees", "cars" and "ocean" in their annotation, testing every classifier by 10-fold cross-validation. The remarkable results of these experiments demonstrate that applying the DDA transformation not only consistenly improves NN classifier accuracy (as expected), but also provides a boost in performance to some more advanced non-linear classification methods, such as SVM [19], as shown in Table 2.

Table 2: Image categorization results

Classifier	% Error on image data set		
	Trees	Ocean	Cars
Fisher's LDA	43.89	45.56	17.72
NN	38.33	19.44	2.46
SVM (linear)	31.11	21.11	1.58
SVM (gaussian)	23.89	16.67	1.58
DDA+SVM (gaussian)	20.00	11.11	1.40
KDDA+NN	18.89	18.33	1.25
DDA+NN	18.86	18.33	1.23

6. Acknowledgements

This work is funded by the Swiss National Foundation (NCCR-IM2 and research grant 21-66648-01) and EU-IST project Webkit (FP5).

7. References

[1] B. Funt and G. Finlayson, "Color constant color indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 522–529, May 1995.

[2] T. Gevers and A. Smeulders, "Color-based object recognition," *Pattern Recognition*, vol. 32, no. 3, pp. 453–464, March 1999.

[3] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. of the International Conference on Computer Vision ICCV, Corfu*, 1999, pp. 1150–1157.

[4] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, 1998.

[5] K. Fukunaga and J. Mantock, "Nonparametric discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 5, pp. 671–678, 1983.

[6] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 607–616, 1996.

[7] A. Arkadev and E. Braverman, *Computers and Patter Recognition*. Washington, D.C.: Thompson, 1966.

[8] E. Fix and J. Hodges, "Discriminatory analysis: Nonparametric discrimination: Consistency properties," USAF School of Aviation Medicine, Tech. Rep. 4, February 1951.

[9] X. Zhou and T. Huang, "Small sample learning during multimedia retrieval using BiasMap," in *IEEE Computer Vision and Pattern Recognition (CVPR'01)*, Hawaii, 2001.

[10] P. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, vol. 35, pp. 73–101, 1964.

[11] S. Kosinov, S. Marchand-Maillet, and T. Pun, "Iterative majorization approach to the distance-based discriminant analysis," March 9–11 2004, presented by S. Kosinov at "Conference of the GfKl 2004", Dortmund, Germany.

[12] S. Kosinov, "Visual object recognition using distance-based discriminant analysis," Computer Vision and Multimedia Laboratory, Computing Centre, University of Geneva, Rue Général Dufour, 24, CH-1211, Geneva, Switzerland, Tech. Rep. 03.07, 2003.

[13] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling*. New York, Springer, 1997.

[14] K. van Deun and P. J. F. Groenen, "Majorization algorithms for inspecting circles, ellipses, squares, rectangles, and rhombi," Econometric Institute Report EI 2003-35, Tech. Rep., 2003.

[15] W. Heiser, "Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis," *Recent advances in descriptive multivariate analysis*, pp. 157–189, 1995.

[16] P. Bartlett, "For valid generalization, the size of the weights is more important than the size of the network," in *Advances in Neural Information Processing Systems 9*, 1997, pp. 134–140.

[17] M. Rojas, S. Santos, and D. Sorensen, "A new matrix-free algorithm for the large-scale trust-region subproblem," *SIAM Journal on Optimization*, vol. 11, no. 3, pp. 611–646, 2000.

[18] W. Hager, "Minimizing quadratic over a sphere," *SIAM Journal on Optimization*, vol. 12, no. 1, pp. 188–208, 2001.

[19] N. Cristianini and J. Shawe-Taylor, *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[20] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *International Conference on Computer Vision and Pattern Recognition (CVPR'03)*, Madison, Wisconsin, June 2003, pp. 409–415.

[21] D. M. Squire, W. Müller, H. Müller, and J. Raki, "Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback," in *The 11th Scandinavian Conference on Image Analysis*, Kangerlussuaq, Greenland, june 1999, pp. 143–149.

[22] Y. Li and L. G. Shapiro, "Object recognition for content-based image retrieval," in *Lecture Notes in Computer Science*. Springer-Verlag, to appear, 2004.