Iterative Majorization Approach to the Distance-based Discriminant Analysis

Serhiy Kosinov[1], Stéphane Marchand-Maillet[1], and Thierry Pun[1]

Computer Vision and Multimedia Lab, University of Geneva,
24 Rue du General-Dufour, CH-1211, Geneva 4, Switzerland,

Abstract. This paper proposes a method of finding a discriminative linear transformation that enhances the data's degree of conformance to the compactness hypothesis and its inverse. The problem formulation relies on inter-observation distances only, which is shown to improve non-parametric and non-linear classifier perfromance on benchmark and real-world data sets. The proposed approach is suitable for both binary and multiple-category classification problems, and can be applied as a dimensionality reduction technique. In the latter case, the number of necessary discriminative dimensions can be determined exactly. The sought transformation is found as a solution to an optimization problem using iterative majorization.

1 Introduction

Efficient algorithms, developed originally in the field of multidimensional scaling (MDS), quickly gained popularity and paved their way into discriminant analysis. Koontz and Fukunaga (1972), as well as Cox and Ferry (1993) proposed to include class membership information in the MDS procedure and recover a discriminative transformation by fitting *a posteriori* a linear or quadratic model to the obtained reduced-dimensionality configuration. The wide-spread use of guaranteed-convergence optimization techniques in MDS sparked the development of more advanced discriminant analysis methods, such as one put forward by Webb (1995), that integrated the two stages of scaling and model fitting, and determined the sought transformation as a part of the MDS optimization. These methods, however, focused mostly on deriving the transformation without adapting it to the specific properties of the classifier that is subsequently applied to the observations in the the transformed space. In addition to that, these techniques do not explicitly answer the question of how many dimensions are needed to distinguish among a given set of classes.

In order to address these issues, we propose a method that relies on an efficient optimization technique developed in the field of MDS and focuses on finding a discriminative transformation based on the compactness hypothesis (see Arkadev and Braverman (1966)). The proposed method differs from the above work in that it specifically aims at improving the accuracy of the non-parametric type of classifiers, such as nearest neighbor (NN), Fix and Hodges (1951), and can determine exactly the number of necessary discriminative dimensions, since feature selection is embedded in the process of deriving the sought transformation.

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 a criterion based on the compactness hypothesis. Then, in Section 3, we demonstrate how the method of iterative majorization (IM) can be used to find a solution that optimizes the chosen criterion. Section 4 describes the extensions of the proposed approach for dimensionality reduction and multiple class discriminant analysis, whereas the details of our experiments are provided in Section 5.

2 Problem formulation

Suppose that we seek to distinguish between two classes represented by matrices X and Y having N_X and N_Y rows of m-dimensional observations, respectively. For this purpose, we are looking for a transformation matrix $T \in \mathbb{R}^{m\times k}$, $k \ll m$, such that $\{T : X \mapsto X', Y \mapsto Y'\}$, that eventuates in compactness within members of one class, and separation within members of different classes.

While the above preamble may fit just about any class-separating transformation method profile (e.g., Duda and Hart (1973)), we must emphasize several important assertions that distinguish the presented method and naturally lead to the problem formulation that follows. First of all, we must reiterate that our primary goal is to improve the NN performance on the task of discriminant analysis. Therefore, 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, Zhou and Huang (2001). More formally, these requirements can be accommodated by an optimization criterion expressed in terms of distances among the observations from the two datasets as follows:

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

where the numerator and denominator of (1) represent the geometric means of the within- and between-class distances defined as $\sqrt{(x_i - x_j)TT^T(x_i - x_j)^T}$ and $\sqrt{(x_i - y_j)TT^T(x_i - y_j)^T}$, respectively, and $\Psi(\cdot)$ denotes a Huber robust estimation function, Huber (1964), parametrized by a positive constant c. The choice of Huber function in (1) is motivated by the fact that at c the function switches from quadratic to linear penalty allowing to mitigate the consequences of an implicit unimodality assumption that the formulation of the numerator of (1) may lead to. In the logarithmic form, criterion (1) is written as:

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

Our preliminary studies, Kosinov (2003), 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, adverse dependence on the initial value, 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 (Borg and Groenen (1997), de Leeuw (1977), Heiser (1995)), which we discuss in the following section.

3 Iterative majorization

It can be verified that majorization remains valid under additive decomposition. 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 the logarithm and Huber function are majorizable by linear and quadratic functions, respectively, Heiser (1995). This fact makes it possible to derive a majorizing function of $S_W(T)$ as follows:

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

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, that in this case is equal to 1 if $\Psi(d_{ij}^W(\bar{T})) < c$ or $c/\Psi(d_{ij}^W(\bar{T}))$ otherwise, and K_1 is a constant term with respect to T. Switching to matrix notation and defining a square symmetric design matrix B dependent on $\bar{T}$ (see Kosinov (2003) for derivation details) let us rewrite the majorizing expression of $S_W(T)$ in its final form:

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

An attempt to majorize $-S_B(T)$ directly runs into problems due to the difficulties of finding a proper quadratic majorizing function of the negative logarithm. As a practical solution, we replace the neg-logarithm with its piece-wise linear approximation (see Figure 1, left panel), which, in turn, can

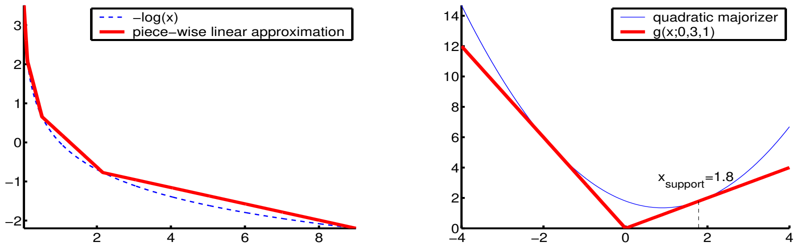


Fig. 1. Majorization of piecewise-linear approximation of $-log(x)$

be represented as a sum of the functions defined as:

$$g(x; x_0, l, r) = \begin{cases} r(x - x_0) & \text{if } x \geq x_0, \\ -l(x - x_0) & \text{if } x < x_0; \end{cases} \quad (5)$$

where $l + r > 0$, to ensure convexity. It is easy to see that the family of functions defined in (5) is one of the many possible generalizations of the absolute value function $|x|$, the former being equivalent to the latter whenever $x_0 = 0$ and $l = r = 1$. Similarly to $|x|$, $g(x; x_0, l, r)$ can be majorized by a quadratic $ax^2 + bx + c$ with coefficients $a > 0$, b and c determined from the majorization requirements (see an example in Figure 1, right panel). Finally, $-S_B(T)$ expressed in terms of the above quadratics can be majorized by the following function, written in matrix notation as:

$$\mu_{-S_B}(T, \bar{T}) = \mathbf{tr}(T^T Z^T G Z T) - \mathbf{tr}(T^T Z^T C Z \bar{T}) + K_2, \quad (6)$$

where Z is the matrix obtained by joining X and Y together, row-wise, and G, C are design matrices dependent on $\bar{T}$, see Kosinov (2003) for derivation details and a description of an alternative faster method based on Taylor series expansion.

Finally, combining results (4) and (6), we obtain a majorizing function of the $\log J(T)$ optimization criterion:

$$\begin{aligned} \mu_{\log J}(T, \bar{T}) &= \alpha \mu_{S_W} + \beta \mu_{-S_B} \\ &= \frac{\alpha}{2} \mathbf{tr}\left(T^T X^T B X T\right) + \beta \mathbf{tr}(T^T Z^T G Z T) \\ &\quad - \beta \mathbf{tr}(T^T Z^T C Z \bar{T}) + K_3, \end{aligned} \quad (7)$$

that is used to find an optimal transformation T minimizing $\log J(T)$ criterion via the iterative procedure described in Heiser (1995), and, thus, constitutes the core of the proposed distance-based discriminant analysis (DDA) method.

While at every iteration it is possible to minimize (7) by solving a system of linear equations, it is often recommended, Krogh and Hertz (1992), 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, for which efficient solution methods exist, Rojas et al. (2000), Hager (2001).

4 Dimensionality reduction and multiple-class setting

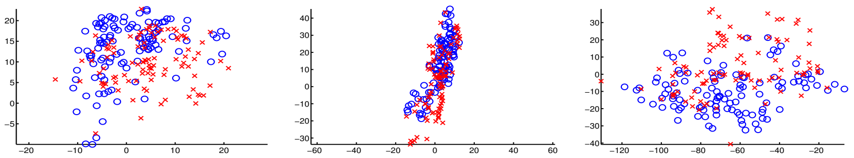
For any $T \in \mathbb{R}^{m \times k}$, $k < m$, the proposed method has an additional advantage of being a dimensionality reduction technique. Moreover, the value of k, i.e., the exact number of dimensions the data can be reduced to without loss of discriminatory power with respect to (2), is precisely determined by the number of non-zero singular values of T. Indeed, the distances between the transformed observations may be viewed as distances between the original observations in a different metric TT^T, that can be expressed as $TT^T = USV^TVSU^T = U_k S_k^2 U_k^T$ using the singular value decomposition of T. The obtained expression reveals that the effect of the full-dimensional transformation T is captured by the first k left-singular vectors of T scaled by the corresponding non-zero singular values, whose number gives an answer to the question of how many dimensions are needed in the transformed space.

While the above discussion is concentrated mostly on the two-class configuration, it is straightforward to generalize the presented formulation to a multiple-class discriminant analysis setting, for the number of classes $K \geq 2$:

$$\log J_K(T) = \sum_{i=1}^{K-1} \left(\alpha^{(i)} S_W(T)^{(i)} - \beta^{(i)} S_B(T)^{(i)} \right). \tag{8}$$

5 Experimental results

Our empirical analysis was based on data sets from the UCI Machine Learning Repository, Blake and Merz (1998). First of all, we verified that the solutions of the optimization problem formulated in Section 2 found by the proposed method were of better quality compared to those of generic techniques, confirming the results reported by Van Deun and Groenen (2003), and Webb (1995). Indeed, numerous random initializations of the gradient search led to inferior as well as unstable results reflected in higher values of $\log J$ (see Figure 2), while the IM-based method proved nearly insensitive to the choice of the initial supporting point and regularly reached far lower criterion values maintaining convergence property at all times, as illustrated in Figure 3. We also validated the proposed dimensionality reduction technique by analysing how the classification performance varied with respect to



(a) $\log J = -0.17$ (b) $\log J = -0.22$ (c) $\log J = -0.19$

Fig. 2. Two-dimensional discriminative projections of the Sonar data set: inferior solutions found by the gradient descent method

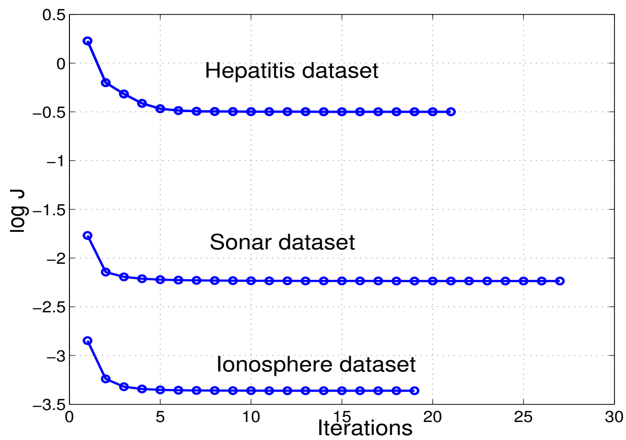


Fig. 3. Convergence of the IM procedure in the DDA method

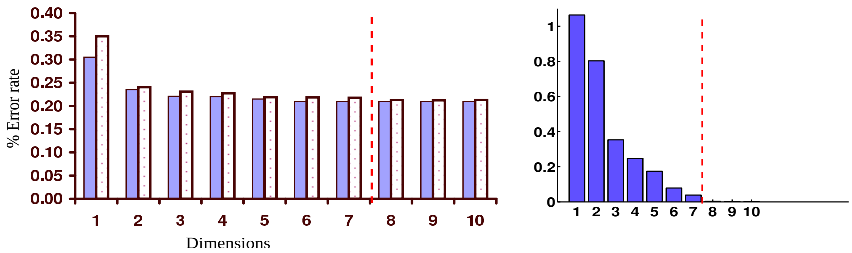


Fig. 4. Dimensionality reduction experiments: classification performance results (left) and singular values of $T \in \mathbb{R}^{m \times m}$ (right). The dashed lines mark the boundary that determines the dimensionality of the transformed space.

k, the dimensionality of the transformed space, and how it was related to the number of non-zero singular values of the full-dimensional transformation, an example of which for the Sonar data set is depicted in Figure 4. The right pane plots 10 largest out of 60 singular values of the full-dimensional transformation, in descending order, while the left diagram[1] shows the results of 10-fold cross-validation experiments with respect to the transformed space dimensionality. It is easy to see that the singular values beyond the 7^{th} are virtually zero. And as the diagram on the left confirms, adding dimensions beyond 7 no longer improves the classification performance (confirmed by Chow test at 99% confidence).

The experiments with the rest of the UCI data sets compared 10-fold cross-validation classification performance of the nearest neighbor classifier in the original feature space (denoted as NN) and that achieved in the transformed space derived by the proposed distance-based discriminant analysis method (denoted henceforth as DDA+NN). Therefore, the goal of this analysis was to assess the effect of applying a DDA transformation on the accuracy of the NN classifier. The error rates of NN and DDA+NN data classification experiments are presented in Table 1, showing a consistent improvement in

Table 1. Classification results for UCI data sets

Data set	Classes	% Error of NN	% Error of DDA+NN
Hepatitis	2	29.57	0.00
Ionosphere	2	13.56	7.14
Diabetes	2	30.39	27.11
Heart	2	40.74	21.11
Monk's P1	2	14.58	0.69
Balance	3	21.45	3.06
Iris	3	4.00	3.33
DNA	3	23.86	6.07
Vehicle	4	35.58	24.70

performance. A separate set of experiments (see Kosinov (2003) for details) using the ETH80 database also revealed the importance of the length constraint, introduced in Section 3 to avoid overfitting. Both unconstrained and length-constrained solutions found by the DDA procedure lead to zero error rate on the training data, but turned out to perform quite differently on the test data sets, on which the length-constrained version of the proposed method demonstrated up to 20% better classification accuracy. Additionally, the results of our more recent experiments reveal that the DDA combined with an SVM classifier, Cristianini and Shawe-Taylor (2000), produces a smaller number of support vectors in the solutions found via the transformed space, which leads to better classification accuracy.

[1] Dot-filled bars denote performance achieved by fixing k *a priori*, while shaded bars show results obtained from a k-truncated SVD of the full-dimensional transformation.

References

ARKADEV, A., BRAVERMAN, E. (1966): Computers and Patter Recognition. Thompson, Washington, D.C.

BLAKE, C., MERZ, C. (1998), UCI Repository of machine learning databases.

BORG, I., GROENEN, P. J. F. (1997): Modern Multidimensional Scaling. New York, Springer.

COX, T., FERRY, G. (1993): Discriminant analysis using nonmetric multidimensional scaling. *Pattern Recognition*, *26*(1), 145–153.

CRISTIANINI, N., SHAWE-TAYLOR, J. (2000): An introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press.

DE LEEUW, J. (1977): Applications of convex analysis to multidimensional scaling. *Recent Developments in Statistics*, 133–145.

DUDA, R. O., HART, P. E. (1973): Pattern Classification and Scene Analysis. John Wiley.

FISHER, R. A. (1936): The Use of Multiple Measures in Taxonomic Problems. *Ann. Eugenics*, *7*, 179–188.

FIX, E., HODGES, J. (1951): Discriminatory analysis: Nonparametric discrimination: Consistency properties. Tech. Rep. 4, USAF School of Aviation Medicine.

HAGER, W. (2001): Minimizing quadratic over a sphere. *SIAM Journal on Optimization*, *12*(1), 188–208.

HEISER, W. (1995): Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. *Recent advances in descriptive multivariate analysis*, 157–189.

HUBER, P. (1964): Robust estimation of a location parameter. *Annals of Mathematical Statistics*, *35*, 73–101.

KOONTZ, W., FUKUNAGA, K. (1972): A nonlinear feature extraction algorithm using distance information. *IEEE Trans. Comput.*, *21*(1), 56–63.

KOSINOV, S. (2003): Visual object recognition using distance-based discriminant analysis. Tech. Rep. 03.07, Computer Vision and Multimedia Laboratory, Computing Centre, University of Geneva, Rue Général Dufour, 24, CH-1211, Geneva, Switzerland.

KROGH, A., HERTZ, J. A. (1992): A Simple Weight Decay Can Improve Generalization. In: J. E. Moody, S. J. Hanson, R. P. Lippmann (eds.), *Advances in Neural Information Processing Systems*, vol. 4, 950–957. Morgan Kaufmann Publishers, Inc.

ROJAS, M., SANTOS, S., SORENSEN, D. (2000): A New Matrix-Free Algorithm for the Large-Scale Trust-Region Subproblem. *SIAM Journal on Optimization*, *11*(3), 611–646.

VAN DEUN, K., GROENEN, P. J. F. (2003): Majorization Algorithms for Inspecting Circles, Ellipses, Squares, Rectangles, and Rhombi. Tech. rep., Econometric Institute Report EI 2003-35.

WEBB, A. (1995): Multidimensional scaling by iterative majorization using radial basis functions. *Pattern Recognition*, *28*(5), 753–759.

ZHOU, X., HUANG, T. (2001): Small Sample Learning during Multimedia Retrieval using BiasMap. In: *IEEE Computer Vision and Pattern Recognition (CVPR'01)*, Hawaii.