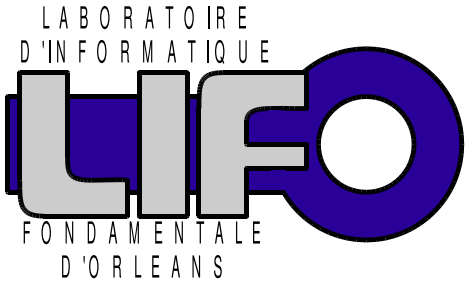

4 rue Léonard de Vinci
BP 6759
F-45067 Orléans Cedex 2
FRANCE
http://www.univ-orleans.fr/lifo

Rapport de Recherche

Relevant dimensions for classification and visualization

Lionel Martin, Matthieu Exbrayat
LIFO, Université d'Orléans

Rapport n^o **RR-2007-10**

Abstract

In this paper we introduce a dimensionality reduction method based on object class, which can be of interest both to define an appropriate distance measure or to visualize objects in a multidimensional space. The method we present is derived from data analysis techniques, such as Principal Components Analysis (PCA). In this paper we propose to consider only pairs of objects that belong to different classes, i.e. to maximise inter-classes distance. Moreover, we introduce a weight parameter that limits the influence of distant objects and favours the influence of close ones, in order to focus on local spatial organization, and thus raise the quality of nearest neighbour classification. Various tests results are presented, which show that a small set of characteristic dimensions can be sufficient to achieve a very satisfying good classification rate.

1 Introduction

Dimensionality reduction plays an important role in many machine learning problems. In this paper we propose an appropriate object representation method for supervised learning. Let us consider a set of observations $\{(x_i, y_i), i = 1..n\}$ where x_i stands for the representation (i.e. a set of features) of a given object and y_i stands for the class of this object. Given such a context, various tasks can be aimed at, the most frequent of which consisting in proposing a class y for a new observation x (classification); visualizing object (in a low dimension space) can also be of interest, in order to observe the general organization, to identify outliers or classes that are close together, etc.; last, we can look for coherent sub groups within classes (clustering).

For any of these goals, having a relevant description of objects is of first interest. Most of time, a more or less important part of the original features describing objects ($\{x_i\}$) has no or few links with the class of this latter, the main issue consisting thus in identifying significant data. A lot of work has been conducted in such ways as feature selection and similarity measures. Such methods present a potentially serious drawback : each feature is evaluated independently from others, using a score function (Guyon & Elisseeff, 2003). By the way, features that together bring information might be eliminated. On the other side, most of similarity measures do not take into account the class attribute (except such works as (Sebag & Schoenauer, 1994; Martin & Moal, 2001), where source or compiled features are weighed, this weight being computed according to a given heuristic).

In this paper we propose a method call "r-discriminant analysis", to compute a limited set of features, that are relevant with regard to class. These features are computed in the way that each of them consists of a combination of source features which maximises the overall (euclidean) distance amongst objects that belong to different classes (i.e. inter-class distance or between-class distance). As a consequence we first focus on numerical attributes, but non numerical ones can also be handled using a quantification mechanism [1].

We thus compute a projection subspace of a given dimension where the sum of inter-classes distance is maximized. A new distance measure is induced by this subspace. It is obvious that such a method is derived from Principal Components Analysis (PCA), where a projection subspace is computed that maximizes the total variance (i.e. the sum of the distance between each pair of objects). Our approach is distinct from linear discriminant analysis, which both maximizes inter-class variance and minimizes within-class variance.

When p source features are provided, *r-discriminant analysis* produces p orthogonal dimensions, sorted according to their (decreasing) influence w.r.t. the sum of inter-class distances, the influence of the last dimensions using to be negligible (but not null). We will thus introduce a simple criterion to determine the number of dimension of interest, and more sophisticated criteria will be mentioned.

The main drawback of a raw *r-discriminant analysis* resides in the fact that distant objects obviously influence inter-class variance more than close ones, which limits the discriminating ability of such an analysis. As a consequence we propose a weighting technique, that rises the influence of close objects, thus preserving local organization of objects from different classes. We call this weighted approach *wr-discriminant analysis*.

Validating *(w)r-discriminant analysis* is quite uneasy, as this approach does not constitute a classification method, but rather a data representation (or analysis) tool. We propose two validation methods:

[1] by replacing a k value feature by k binary features

- We first use a nearest-neighbour classification, where distance is computed, on one hand using the whole set of source features, and on the other hand using the more relevant dimensions of the subspace computed by *(w)r-discriminant analysis*. We will focus on the influence of the subspace dimension towards the good classification rate, showing that a limited set of dimensions can significantly raise this rate. We use various (standard) test sets of the UCI repository (Newman et al., 1998).
- We then present some screen snapshots of 3-D projections, using ACP and *(w)r-discriminant analysis*, highlighting the fact that, for some data sets, our approach can also offer a better spatial representation of objects.

In section 2 we will both formaly define r-discriminant and weighted r-discriminant analysis and demonstrate how they can be solved by searching for eigenvalues. Related work will be presented in section 3. In section 4 we will present various test results both in terms of supervised learning (nearest neighbour classification) and of visualization, as described above. Finally, we will sum up and propose future work guidelines in section 5.

2 R-discriminant analysis

Let us consider a set of observations $\{(x_i, y_i), i = 1..n\}$ where x_i stands for a given object (here: a point in $\mathbb{R}^p$) and y_i stands for its class. Let us denote $x_i = (x_{i1}, \ldots, x_{ip})$.

2.1 Principal Component Analysis (PCA) - a brief reminder on principle and technique

The main goal of PCA consists in the search of a 1-dimensional sub-space where the variance of the projected points is maximal. Practically speaking, man looks for a unit vector u such that, if d_u is the euclidean distance in the sub-space defined by u, the sum $\sum_i d_u(x_i, \overline{g})^2$ is maximized ($\overline{g}$ being the centroid of the points).

Let $h_u(x)$ denotes the projection of x in the sub-space defined by u (and containing the origin). From the matrix point of view, the sum can be expressed as (X' being the transposed matrix of X, and using the same notation for a point and the corresponding vector) :

$$\begin{aligned} \sum_i d(h_u(x_i), h_u(\overline{g}))^2 &= \sum_i ||h_u(x_i) - h_u(\overline{g})||^2 \\ &= ||h_u(x_i - \overline{g})||^2 \\ &= u'M'Mu \end{aligned} \tag{1}$$

and $M = (m_{ij})$ where $m_{ij} = x_{ij} - \overline{g}_i$ and $\overline{g}_i$ stands for the mean value of attributes of rank i, $\overline{g}_i = \frac{1}{n}\sum_j x_{ji}$ ($\overline{g}$ can be ignored, as long as the data use to be normalized, i.e. the mean is subtracted).

This problem is usually solved by looking for the highest eigen value of $M'M$, u being thus the eigen vector associated to this eigenvalue. The second most interesting dimension can then be found by looking for the next dimension, orthogonal to u, that have the maximal variance (which corresponds to the second highest eigenvalue of $M'M$ and its associated eigen vector), and so on.

2.2 R-discriminant analysis (RDA)

Considering PCA, we can underline the fact that, in the projection sub-space:

$$\sum_{i,j} d_u(x_i, x_j)^2 = 2n \sum_i d_u(x_i, \overline{g})^2 \tag{2}$$

which can be expressed as the sum of an inter-class (Σ_r) and a within-class (Σ_a) sums:

$$\begin{aligned} 2n \sum_i d_u(x_i, \overline{g})^2 &= \sum_{i,j \mid y_i \neq y_j} d_u(x_i, x_j)^2 \\ &+ \sum_{i,j \mid y_i = y_j} d_u(x_i, x_j)^2 \end{aligned} \quad (3)$$

Thus, maximizing the total variance is equivalent to maximizing the sum of the distance between each pair of objects.

With r-discriminant analysis, we propose to only consider and maximize the sum of inter-class distances, so that, in the subspace computed, the priority is given to a good relative representation of objects belonging to different classes.

We thus look for u, maximising a quadratic objective function under quadratic constraint:

$$\begin{cases} Max_{(u)} \sum_{i,j \ y_i \neq y_j} d_u(x_i, x_j)^2 \\ ||u||^2 = 1 \end{cases}$$

which will correspond to the dimension which preserves inter-class distances at most. We will then look for the next vector v, orthogonal to u ($||v||^2 = 1$) that maximizes the same expression, and so on.

2.3 Solving

We can clearly express the preceding sum in a matrix form, using equation (1). But the size of $M'M$ (to be diagonalized), which corresponds to the number of pairs of objects of distinct classes, is then of the order of n^2. Nevertheless, it can be expressed in a much simpler form, using equation (2), where the (double) sum of inter-objects distances can be replaced by the (simple) sum of the distance of each object to the centroid.

Let α be a class and $\overline{g_\alpha}_j$ the mean value of attribute j amongst objects of class α: $\overline{g_\alpha}_j = \frac{1}{n_\alpha} \sum_{i \mid y_i = \alpha} x_{ij}$ and $\overline{g_\alpha}$ is the centroid of points of class α. Let n_α be the number of objects of class α.

According to equation (2), once projection is done, the sum of distances amongst all objects of class α is:

$$\sum_{i,j \mid y_i = y_j = \alpha} d_u(x_i, x_j)^2 = 2n_\alpha \sum_{i \mid y_i = \alpha} d_u(x_i, \overline{g_\alpha})^2 \quad (4)$$

which, according to (1), is :

$$2n_\alpha \sum_{i \mid y_i = \alpha} d(h_u(x_i), h_u(\overline{g_\alpha}))^2 = 2n_\alpha u' M'_\alpha M_\alpha u$$

where $M_\alpha = (m^\alpha_{ij}) \in \mathbb{R}^{p \times p}$ with

$$m^\alpha_{ij} = \begin{cases} x_{ij} - \overline{g_\alpha}_j & is \quad y_i = \alpha \\ 0 & otherwise \end{cases}$$

Let $B_\alpha = (b^\alpha_{ij}) = n_\alpha M'_\alpha M_\alpha$: such that

$$b^\alpha_{ij} = n_\alpha \sum_{r \mid y_r = \alpha} (x_{ri} - \overline{g_\alpha}_i)(x_{rj} - \overline{g_\alpha}_j)$$

The total sum of within-class distances (after projection) is the sum, amongst all classes, if the expression given by (4):

$$\begin{aligned} \Sigma_a &= \sum_{\alpha=1}^{k} \sum_{i,j \mid y_i = y_j = \alpha} d(h(x_i), h(x_j))^2 \\ &= \sum_{\alpha=1}^{k} 2n_\alpha u' M'_\alpha M_\alpha u \end{aligned} \quad (5)$$

This expression can be expressed as $2u'\mathcal{M}_a u$, where $\mathcal{M}_a = (m_{ij}^a)$ is the sum of matrices $n_\alpha M'_\alpha M_\alpha$, with $\alpha = 1 \dots k$ (i.e. the sum matrices B_α as defined above):

$$m_{ij}^a = \sum_{\alpha=1}^{k} n_\alpha \sum_{r \mid y_r=\alpha} (x_{ri} - \overline{g_{\alpha i}})(x_{rj} - \overline{g_{\alpha j}}) \tag{6}$$

Last, the total sum of within-class distance (within the projection sub-space associated with vector u) can be expressed as: $u'\mathcal{M}_a u$. Since the variance of the complete data is given by $2nu'M'Mu$ (according to (1) and (2)), the sum of within-class distances, after projection, can be expressed the same way as its value is the difference of these two latter:

$$\begin{aligned} \Sigma_r &= 2nu'M'Mu - 2u'\mathcal{M}_a u \\ &= u'(2nM'M - 2\mathcal{M}_a)u \\ &= u'(2nM'M - 2\sum_\alpha n_\alpha M'_\alpha M_\alpha)u \end{aligned} \tag{7}$$

Maximizing this expression is obtained by searching extrema of quadratic form under quadratic constraint. We can notice that, the matrix to be diagonalized is symmetric (sum of symmetric matrices) of order p.

2.4 R-discriminant analysis: dual expression

There exists a so-called "dual" expression of PCA that consists in diagonalizing MM' instead of $M'M$, both of them having the same eigenvalues, and the eigen vectors of one being obtained from the ones of the other. This dual expression enables the diagonalization of the matrix of order $min(n, p)$.

Considering r-discriminant analysis, the matrix to be diagonalized is $nM'M - \mathcal{M}_a$. In fact, we have to maximize $u'(nM'M - \mathcal{M}_a)u$. Since matrices M'_{α_i} and M_{α_j} are orthogonal for each $\alpha_i \neq \alpha_j$, the expression of $\mathcal{M}_a$ can be rewritten: $M'_\Lambda M_\Lambda$ with $M_\Lambda = (m_{ij}^\Lambda) = \sum_{\alpha=1..k} \sqrt{n_\alpha} M_\alpha$.

We can also notice that $u'(nM'M - M'_\Lambda M_\Lambda)u$ can be expressed as $u'A'Bu$ where A and B are two matrices of dimension $2n \times p$, $A = (a_{ij})$, $B = (b_{ij})$ defined by:

$$a_{ij} = \begin{cases} \sqrt{n} m_{ij} & if \quad i \leq n \\ m_{i-n,j}^\Lambda & if \quad i > n \end{cases}$$

$$b_{ij} = \begin{cases} \sqrt{n} m_{ij} & if \quad i \leq n \\ -m_{i-n,j}^\Lambda & if \quad i > n \end{cases}$$

Thus we have to maximize $u'A'Bu$ under constraint $||u||^2 = 1$, which is equivalent to searching the eigenvalues of $A'B$. We can easily notice that $A'B$ and BA' have the same eigenvalues, BA' being of order $2n$.

Last, we can diagonalize BA'. Obtaining the eigen sub-spaces of $A'B$ using the eigen vectors BA' is then similar to the dual method of PCA (L. Lebart, 2000).

2.5 Weighted r-discriminant analysis (WRDA)

R-discriminant analysis aims at maximizing the sum of the square of distances amongst objects of different classes. We can see that a possible drawback of this approach lies in the fact that it favours the representation of the most distant objects. In effect, due to the criteria to maximize, the higher a distance, the higher its impact on the objective function. As a consequence, nearby objects could be very near or even merged in the representation subspace. In a context of classification (using a "nearest neighbour" method), this constitute a real drawback, as nearby objects of different classes should be well projected. Thus, rather than optimizing:

$$\Sigma_r = \sum_{i,j \mid y_i \neq y_j} d_u(x_i, x_j)^2$$

we propose to optimize

$$\Sigma_{r,w} = \sum_{i,j \mid y_i \neq y_j} w_{ij} d_u(x_i, x_j)^2$$

the weight w_{ij} growing when $d(x_i, x_j)$ decreases ($d(x_i, x_j)$ stands for the euclidean distance within the source space). We can notice that, if $d(x_i, x_j) = 0$, then $d_u(x_i, x_j)$ as no effect on the objective function (in this case, w_{ij} can be set to 1).

Practically speaking, we use $w_{ij} = 1/d^\sigma(x_i, x_j)$ where σ is a parameter that tunes the influence of nearby objects (the higher σ, the higher the influence of nearby objects, $\sigma \geq 0$).

We can not solve this problem using the former techniques. However the sum to maximize can be expressed with matrices, similarly to (1):

$$\begin{aligned} \sum_{i,j \mid y_i \neq y_j} & w_{ij}(h_u(x_i) - h_u(x_j))'(h_u(x_i) - h_u(x_j)) \\ & = u'N'Nu \end{aligned} \tag{8}$$

with N a matrix with p columns, the number of rows being equals to the number of pairs of objects of different class (of the order of n^2). Thus, each pair of objects (x_i, x_j) with $y_i \neq y_j$ corresponds to a row of matrix N:

$$\begin{array}{l} (\sqrt{w_{ij}}(x_{i1} - x_{j1}),\ \sqrt{w_{ij}}(x_{i2} - x_{j2})\ \ , \\ \qquad\qquad \ldots,\ \sqrt{w_{ij}}(x_{ip} - x_{jp})) \end{array}$$

We thus compute the eigen values of the matrix NN' (of order p, which may be relatively small); besides, building NN' implies going across N, having a size ($o(n^2 \times p)$.

We can notice that this expression offers another formulation of the preceding problem (non-weighted RDA, i.e. with weights $w_{ij} = 1$).

3 Related works

Feature selection has been applied to classification problems (Dash & Liu, 1997). Moreover, various studies have been conducted that introduce some discriminant aspects within analysis methods (L. Lebart, 2000; J.P. Nakache, 2003), called Fisher Linear Discriminant Analysis (LDA) (Fukunaga, 1990). In recent years, many approaches have been proposed to deal with high-dimensional problems (Ye & Xiong, 2006) and different covariance matrices for different classes (Kumar & Andreou, 1996).

These approaches are based on the decomposition of the sum to be maximized within PCA: $u'Tu$ where $T = M'M$ is the variance-covariance matrix of observations. Using the Huygens decomposition formula, the matrix can be expressed as the sum of the inter-class (E) and within-class (D) variance-covariance matrices: $u'Tu = u'Du + u'Eu$.
In this case, within-class variance must be maximized while inter-class variance must be minimized, in order to find a low dimensional representation space, where classes are more consistent and far from each other. Practically speaking, $u'Eu/u'Tu$ is maximized by searching for the higher eigenvalues of $T^{-1}E$.

We can notice that our approach is different from LDA, as shown on the following example. Consider 4 objects and 2 classes (+ and -) and the following (2D) dataset:

+ - - +

In this dataset, the means of the classes + and - are both exactly in the center of the figure and then the inter-class variance is null (but the sum of inter-class distances is not null). Then, if we search for a 1-dimensional subspace, the LDA gives the vertical subspace (minimizing within-class variance) while our method gives the horizontal subspace. This example illustrates the advantage of RDA when some classes are split into distant subclasses.

In the case of classification, (Mohri & Tanaka, 1994) uses this approach, projecting objects on the most discriminant axis (and thus using a one-dimensional representation space).
Nevertheless, this decomposition is obviously different from the one we use.

The representations proposed by a linear discriminant analysis will be of poor interest when within-class variance can hardly be reduced, and/or when inter-class variance is low. R-discriminant analysis fits

better to the case where classes are not scattered across space, and/or are split in sub-classes. Figure 1 presents the case of two classes (+ and -) and their centroids G+ and G-, where inter-variance is low while within-class variance is high (considering class +).
Projecting the objects from two to one dimensional space, linear discriminant analysis will use axis D_1 (which will mix classes), while r-discriminant analysis will use D_2 (which will produce a more class-consistent projection, and thus a better classification).

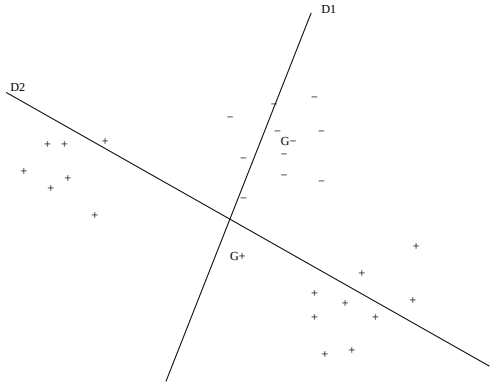

Figure 1: Example of dimensionality reduction

¿From another point of view, linear discriminant analysis consists of a PCA on classes centroids, weighted according to the size of classes (J.P. Nakache, 2003). By the way, the number of discriminant dimensions is bounded by the number of classes, which might be a too strong reduction.

4 Experimental results

4.1 Nearest neighbour

The tests we present aim at evaluating the gain introduced by weighted r-discriminant analysis (WRDA) within a classification process. These tests are based on standard data sets (numerical data sets with no missing value) available at UCI (Newman et al., 1998).
For each data set we realised a cross validation based on ten randomly generated subsets (the distribution of objects amongst classes being the same in each subset). We established a correct classification rate (using a 1-nearest neighbour method), based on the euclidean distance for subspaces from one to n dimensional sub-spaces (n being the number of features in the original data set), using first the most discriminant dimensions (WRDA) and the most significant (PCA).
Curves are based on the average value of twenty runs of the cross validation procedure. The weighting parameter σ varies from 0 to 10.

4.1.1 Glass and Zoo

The curves we obtain with these two data sets are quite characteristic of the classification by nearest neighbour with WRDA: when the number of dimensions growths, the correct classification rate first growths and then lowers until it reaches the rate corresponding to the classification within the original space. This highlights the interest of dimensionality reduction: a subspace can lead to better classification rate than the original set of features.

Figure 2 presents the results for *glass*. For a small number of dimensions (from 1 to 4) PCA is satisfying, compared to WRDA. But when the number of dimension growths (between 5 and 8), we can notice that a WRDA with an high σ raises the correct classification rate (77 %, compared to 66 % with PCA). We can notice that we obtain a clear peak with WRDA, for a relatively small number of dimension, that is higher that the highest correct classification rate observed with PCA.
Concerning *zoo* (fig. 3), PCA remains interesting if we only consider the most significant dimension, while WRDA brings a *noticeable gain* with 3 and even more with 4 dimensions (98 % of correct classification

with 4 dimensions and σ=4, vs. 92 % with PCA). As with glass, we observe a peak of the correct classification rate, followed by a slight decreasing when the number of dimensions increases, reaching 95 % when all available dimensions are used.

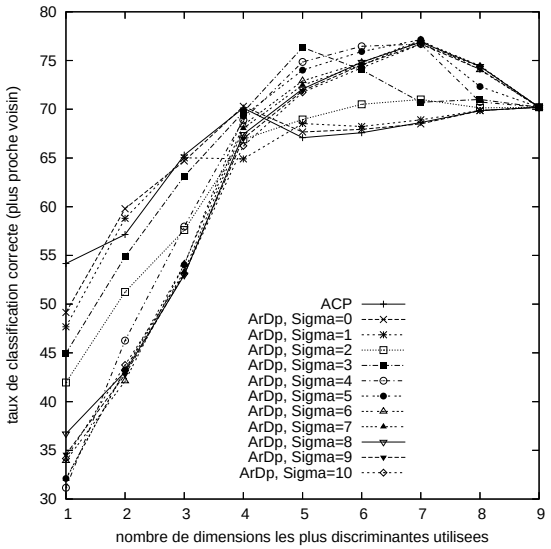

Figure 2: Correct classification for glass

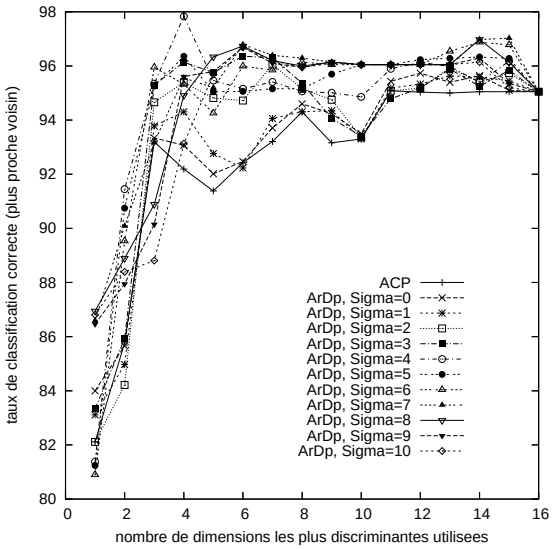

Figure 3: Correct classification for zoo

4.1.2 Tic-tac-toe and Pima

Figures 4 and 5 concern respectively *tic-tac-toe* and *pima indians diabete* data sets. They offer two characteristic shapes. With *tic-tac-toe*, a clear peak is followed by a strong decreasing, and finally a stabilization around the rate observed within the original space.
On the contrary, with *pima indians diabete*, there is no noticeable peak (except a very light one with PCA). Rates are growing regularly until they reach (using all the discriminant dimensions available) the rate observed within the original space. We may think that this is due to the fact that each original feature plays a role w.r.t. the object's class, and thus any dimensionality reduction leads to a loss of information and by the way of efficiency. We only observed this fact with 2 data sets (*pima-diabete* and *liver disorder*).

4.1.3 Synthetic overview of tests

Table 1 offers a synthetic comparison of correct classification rates with PCA and WRDA over a larger set of data sets, compared to classification within the original space. For each data set we have computed the correct classification rate for PCA and WRDA, σ running from 0 to 10, with a step of 0.5 ($\sigma = 0$ corresponds to RDA). Columns of this table indicate:

- for each data set : name, cardinality and dimension of the original space;
- for PCA: best correct classification rate obtained and dimension of the corresponding sub space;
- for WRDA: best correct classification rate obtained, dimension of the corresponding sub space and σ;
- correct classification rate with the original data (and euclidean distances).

When, for a given data set, the best rate was achieved with various values of σ and/or of the number of dimensions used, we chose to present the lowest values of these parameters. Some data sets (signaled by a *) do contain symbolic data. In this case quantification of symbolic data has been realized by replacing it by a set of binary attributes, each of these corresponding to a given value of the symbolic attribute.

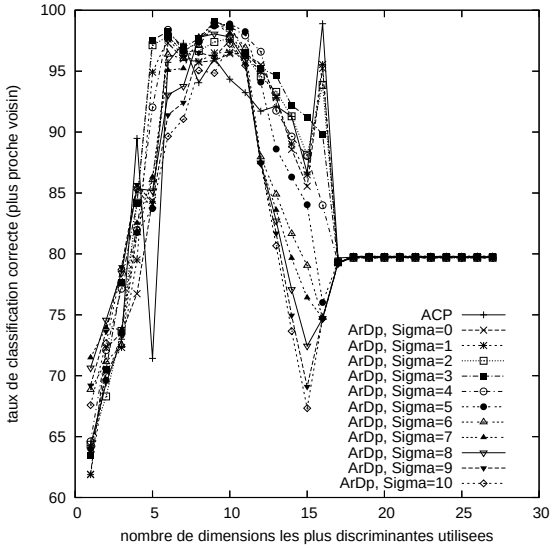

Figure 4: Correct classif. for tic-tac-toe

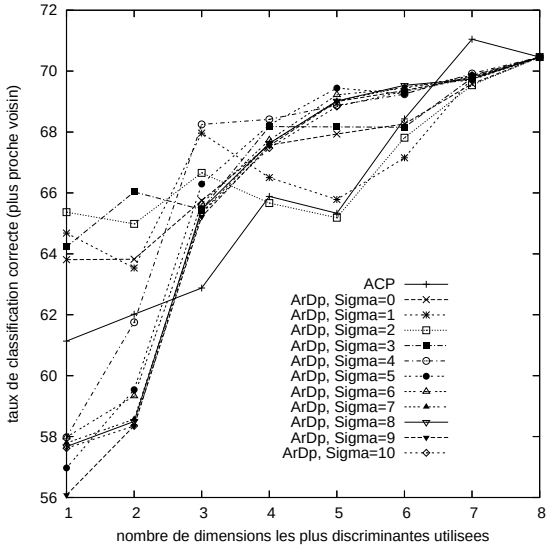

Figure 5: Correct classif. for pima diabete

In this case, the dimension of the original space we give corresponds to the dimension once quantization has been done.

We can notice that WRDA tends to offer higher performance (w.r.t. correct classification rate using one-nearest neighbour) than the original space. Its performance is also similar or higher than the one of PCA (better rate or smaller number of dimensions). The interest of weighting is also underlined by the fact that, in most cases, the best rate is reached with $\sigma > 0$. The counter performance observed with *balance scale* does *a priori* come from the very uniform distribution of objects amongst the original space, which limits discrimination possibilities.

These tests show that in many case WRDA can lead to a limited set of highly discriminant dimensions, and thus consistent w.r.t. classification. We consider that an automated parametering of WRDA, i.e. automated search of the number of dimensions and of the value of σ to be used, would be of great interest as a future work (see section 5).

Table 1: Best correct classification rate observed with PCA and wr-discriminant analysis

Set			PCA		wr-d analysis			source space
name	# of obj.	# of dim.	Max (%)	# of dim.	Max (%)	nb. dim.	σ	N.N. (%)
balance scale	625	4	95	2	88,8	2	3,5	79
cars (extrait) (*)	518	19	94,9	5	*96,3*	11	5	93,1
ecoli	336	7	81	7	81	*5*	2,5	80,9
glass	214	9	70,2	9	*77,1*	*7*	6	70,2
ionosphere	351	34	91,2	12	91,1	*8*	2	86,6
iris	150	4	94,5	4	*96,3*	*1*	4	94,5
liver disorder	345	6	62,5	6	62,5	6	0	62,5
pima-diabete	768	8	71	7	70,5	8	0	70,5
soybean (small)	47	35	100	2	100	2	0	100
teaching(*)	151	56	66,9	17	*67,4*	*16*	0,5	63,7
tic-tac-toe(*)	958	26	98,9	16	*99,1*	*9*	3	79,8
zoo	101	16	95,1	11	*97,8*	*4*	5	95,1
wine	178	13	95,5	5	*97,6*	*2*	4	95,2

4.2 Visualization

While WRDA appear to be efficient in the context of nearest-neighbour classification, its interest for visualization is to be show, as far as it introduces a bias in distance amongst objects. In order to show that WRDA might be anyway of interest considering visualization, we present a sample visualization of (small) soybean data set with PCA (fig. 6) and WRDA with $\sigma = 2$ (fig. 7). This data set consists of only 47 objects defined over 4 dimensions, and was thus easy to visualize in a small screen-shot. To distinguish objects among classes we use a different shape for each class.

PCA produces three groups. Two classes are clearly separated, the last two being more mixed. With WRDA the four classes are clearly separated. We can notice that WRDA clearly separate classes for $1 \leq \sigma \leq 3$, but complementary visualizations show that with higher values of σ, discrimination keep on growing, thus offering a slight side-effect: the axis that moves away the objects of the nearby classes becomes more and more significant, this stretching tending to visually move the two other classes closer.

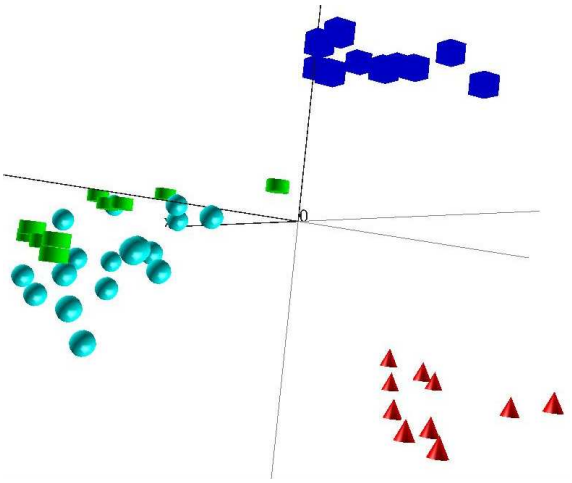

Figure 6: PCA for soybean (small)

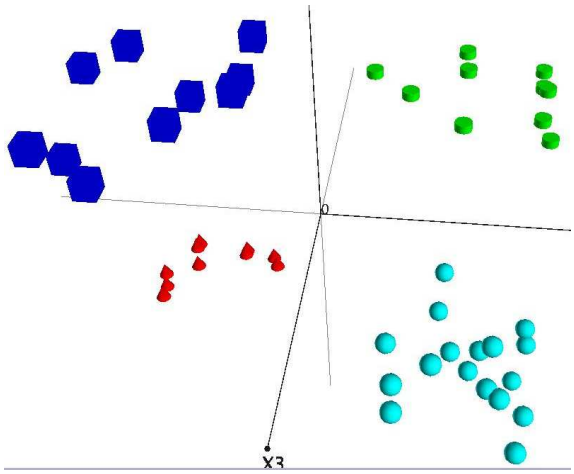

Figure 7: WRDA for soybean (small), $\sigma = 2$

5 Conclusion and future works

In this paper we have proposed a dimensionality reduction technique called weighted r-discriminant analysis, based on the maximization of inter-class distances. We have highlighted the relevance of this approach and of the weighting method we introduce to favour local organization of close objects that belong to different classes.
The tests we conducted have shown that WRDA does in many case improve results compared to a nearest neighbour search within the original representation space, in terms of the maximum correct classification rate and/or the number of dimensions used to reach this best rate. WRDA can also raise the quality of a spatial (2 or 3-D) visualization of objects compared to PCA.

In a classification framework, using WRDA supposes that two parameters are defined : the weighting coefficient and the number of dimensions aimed at. In this paper we studied, through several tests, the performance of WRDA when these parameters vary. Future work will consist in proposing strategies to estimate satisfying values for both of them, depending on the data set considered. This estimation could be based on the study of the eigenvalues associated with the most discriminant dimensions, using only the most significant ones (in many cases, the very first eigenvalues handle most of the variance).

¿From another point of view, as WRDA tends to render local organization of objects, one can wonder if a global maximization, even weighted, is still of interest. We could thus explore and propose local sets of dimensions depending on the space region observed. This would suggest three directions of study : first splitting space into areas (automatically, and/or manually), then searching the (locally) most discriminant dimensions, and last proposing and developing appropriated visualization tools.

References

Dash, M., & Liu, H. (1997). Feature selection for classification.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. New York: Academic Press.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, *3*, 1157–1182.

J.P. Nakache, J. C. (2003). *Statistique explicative applique*. Paris: Technip.

Kumar, N., & Andreou, A. (1996). On generalizations of linear discriminant analysis.

L. Lebart, A. Morineau, M. P. (2000). *Statistique exploratoire multidimensionnelle*. Paris: Dunod.

Martin, L., & Moal, F. (2001). A language-based similarity measure. *Machine Learning: ECML 2001, 12th European Conference on Machine Learning* (pp. 336–347). Springer.

Mohri, T., & Tanaka, H. (1994). An optimal weighting criterion of case indexing for both numeric and symbolic attributes.

Newman, D., Hettich, S., Blake, C., & Merz, C. (1998). UCI repository of machine learning databases.

Sebag, M., & Schoenauer, M. (1994). *Topics in case-based reasonning*, vol. 837 of *LNAI*, chapter A Rule-based Similarity Measure, 119–130. Springer-Verlag.

Ye, J., & Xiong, T. (2006). Null space versus orthogonal linear discriminant analysis. *ICML* (pp. 1073–1080).