Genetic Algorithms for Exploratory Data Analysis*

Alberto Perez-Jimenez and Juan-Carlos Perez-Cortes

Departamento de Informatica de Sistemas y Computadores
Universidad Politecnica de Valencia
Camino de Vera, s/n 46071 Valencia, Spain
{aperez,jcperez}@disca.upv.es

Abstract. Data projection is a commonly used technique applied to analyse high dimensional data. In the present work, we propose a new data projection method that uses genetic algorithms to find linear projections, providing meaningful representations of the original data. The proposed technique is compared with well known methods as Principal Components Analysis (PCA) and neural networks for non-linear discriminant analysis (NDA). A comparative study of these methods with several data sets is presented.

1 Introduction

Data projection is a commonly used technique applied to exploratory data analysis [3]. By projecting high dimensional data into a 2- or 3-dimensional space, a better understanding of the structure of the data can be acquired. Characteristics such as clustering tendency, intrinsic dimensionality, similarity among families or classes, etc. can be studied on a planar or tridimensional projection, which also can help to build a classifier or another statistical tool [12][8].

Data projection methods can be divided into linear and non-linear, depending on the nature of the mapping function [7]. They can also be classified as supervised or unsupervised, depending on whether the class information is taken into account or not. The best known linear methods are Principal Component Analysis, or PCA (unsupervised), Linear Discriminant Analysis or LDA (supervised) [3], and projections pursuit [2]. Schematically, PCA preserves as much variance of the data as possible, LDA tries to group patterns of the same class, separating them from the other classes, and, finally, projection pursuit tries to search projections in which points do not distribute normally. On the other hand, well known non-linear methods are: Sammon's Mapping (unsupervised) [10] , non-linear discriminant analysis, or NDA (supervised) [8] and Kohonen's self-organising map (unsupervised) [6]. Sammon's mapping tries to keep the distances among the observations using hill-climbing or neural networks methods [8][10], NDA obtains new features from the coefficients of the hidden layers of a multi-layer perceptron (MLP) and Kohonen Maps project data trying to preserve the topology.

* Work partially supported by the Spanish CICYT under grant TIC2000-1703-CO3-01

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 743–751, 2002.

In the present paper, a new linear supervised data projection method referred to as GLP (genetic linear projection) is proposed. The goal of this method is to find a set of linear projections maximising a certain criterion function. In this work, the accuracy of a Nearest Neighbour classifier has been used as the criterion to maximise. The optimisation is performed by means of a genetic algorithm (GA) [5] [4].

In Section 2 we describe the GLP algorithm, in Section 3 a comparison between a linear method (PCA), a non-linear method (NDA) and the proposed GLP algorithm over several data sets is presented. Finally, some conclusions and further works are presented in section 4.

2 Genetic Linear Projection (GLP)

A linear projection (LP) is defined as follow,

$$LP(x) = c_1 x_1 + c_2 x_1 + \ldots c_d x_d,$$

where x is a d-dimensional vector with components x_i, and c_i are the projections coefficients representing the projection axis.

The GLP searches for m (being m the projected space dimensionality) LP's at the same time, optimising the accuracy rate of a Nearest Neighbour classifier. The goal of using this criterion is to preserve the class structure of the data in the projected space. Since the projections obtained are always linear, the representation does not produce an excessive distortion of the original space and therefore the observed data is directly related to the original data.

This criterion does not impose the orthogonality of the projections, as opposed to methods such as PCA or LDA, neither forces the recomputation of the data distribution after choosing each new axis, as in Projection Pursuit.

The number of parameters to estimate by GLP is $m \times d$, since a linear projection is defined by d coefficients, being d the dimensionality of the original data, and m the dimension of the projected space. If we want to project high-dimensional data, the number of parameters to estimate will be large. For that reason, we propose a Genetic Algorithm to carry out the optimisation.

Genetic Algorithms have proved to be specially useful on large search spaces [4]. We have used a GA with the following properties:

- An individual is composed of m chromosomes representing the m LP's to search. Each chromosome contains d genes, holding each a binary string of b bits that encodes a coefficient of the LP in fixed point format.
- For the fitness function, the computed accuracy of a Nearest Neighbour classifier trained with the projected data obtained from the linear projection coded in the individual is used.
- As a genetic selection scheme, a rank-based strategy [9] has been used. In this strategy, the probability of being selected is computed from the rank position of the individuals. This method gave in our case a faster convergence than a fitness-proportionate method.

– Finally, the following setting are used for the rest of parameters: crossover probability is 0.6, mutation probability is 0.001, population size is 100, and the maximum number of generations is 300.

Finally, because to estimate the accuracy of a Nearest Neighbour classifier is a time consuming task. A micro-grain parallel GA [11] has been implemented to reduce computational time. In these algorithms several computers are used to compute individual fitness functions, obtaining a linear speedup.

3 Comparative Study

3.1 Methodology

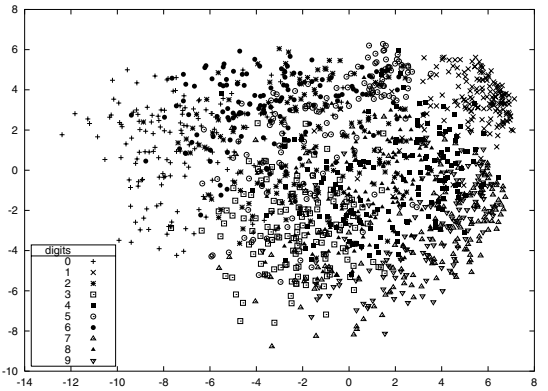
In this section our GLP method will be compared with the well known PCA (linear, unsupervised) and NDA (non-linear, supervised) methods. The three methods will be applied to four data sets in order to obtain 2-dimensional projections. The data sets used are described below.

– *Digits.* This is a high-dimensional data set containing 3000 patterns, representing 128×128 images of hand-written digits. Each pattern is obtained by resizing images to 14×14 and using gray values as features. The dimension of the data is 196.
– *IRIS.* This data set, obtained from the UCI repository [1], consists of 150 4-dimensional pattern from 3 classes. It contains four measurements on 50 flowers from each of three species of the Iris flower.
– *Cookies.* This synthetic corpus consists of two 10-dimensional normal distributions with

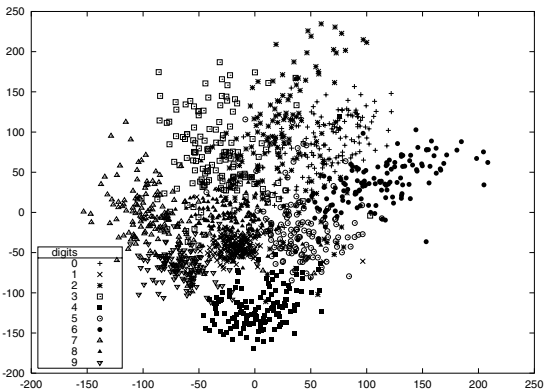
$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.0001 & 0 \cdots 0 \\ 0 & 1 \cdots 0 \\ \vdots & \vdots \ddots \vdots \\ 0 & 0 \cdots 1 \end{bmatrix},$$

$\mu_1 = (+0.1, 0, 0, \ldots)$ and $\mu_2 = (-0.1, 0, 0, \ldots)$, having each class 1000 patterns. These distributions represent two hyperspehers flattened (like cookies) in the dimension they are separated. This data set represents a well known case in which PCA does not work well because the maximal scattered axes are not the most significant.
– *Page Blocks.* This corpus, also obtained from the UCI repository, consists of 5473 10-dimensional patterns representing block documents. Each pattern is represented by 10 features representing geometrical and image properties of the segmented blocks. Blocks are classified into 5 classes.

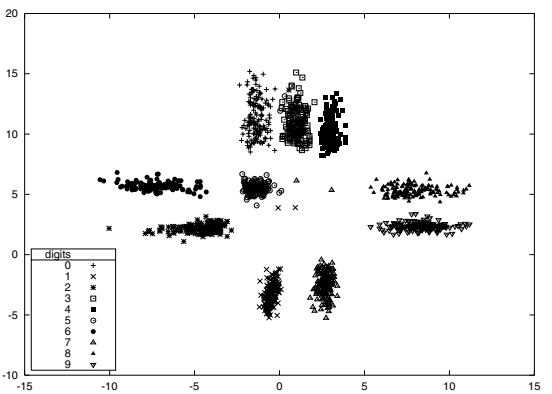
The performance of these methods will be first compared by means of visual judgement over the 2-dimensional projections obtained from the data sets. And then by means of the error rate of a Nearest Neighbour classifier (E_{NN}) computed for each data set in the original and projected spaces. This quantitative criterion shows how well the class structure is preserved by the projections.



a)

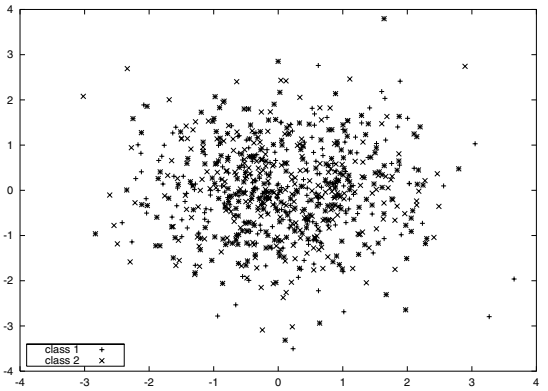


b)

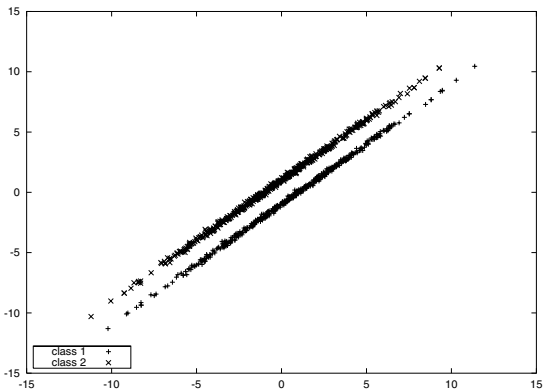


c)

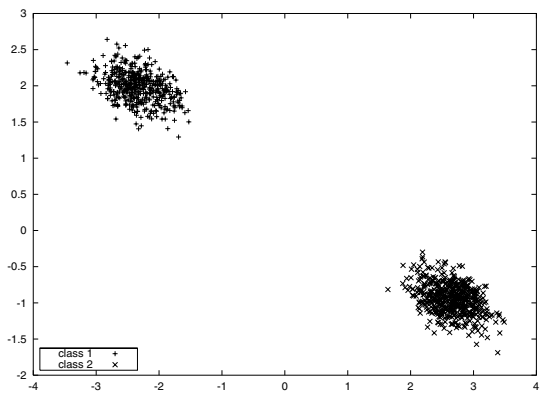
Fig. 1. *Digits* data set 2D projections using: a) PCA, b) GLP and c) NDA



a)

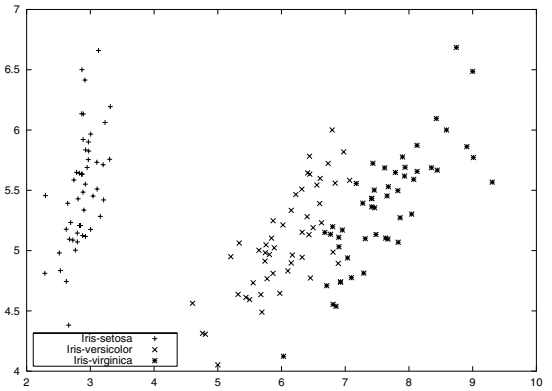


b)

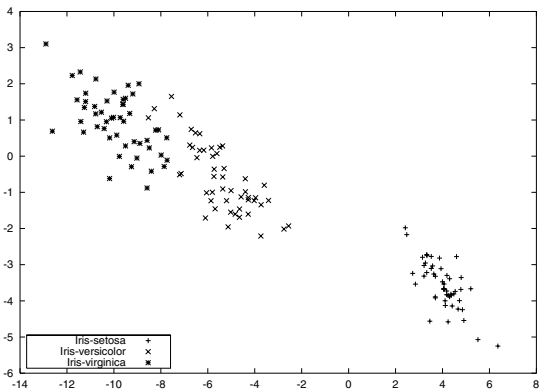


c)

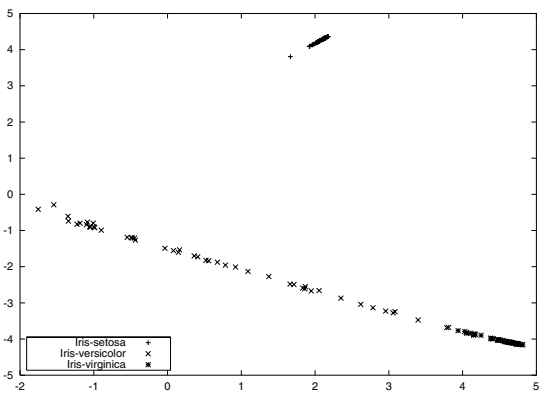
Fig. 2. *Cookies* data set 2D projections using: a) PCA, b) GLP and c) NDA



a)

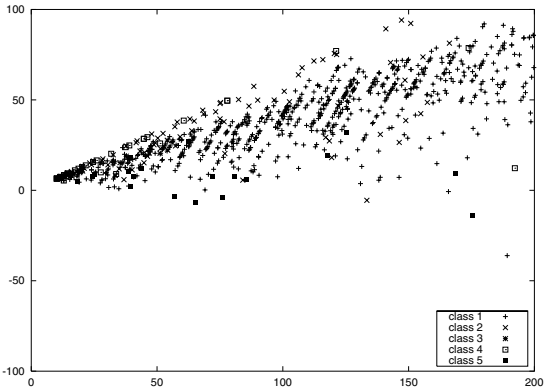


b)

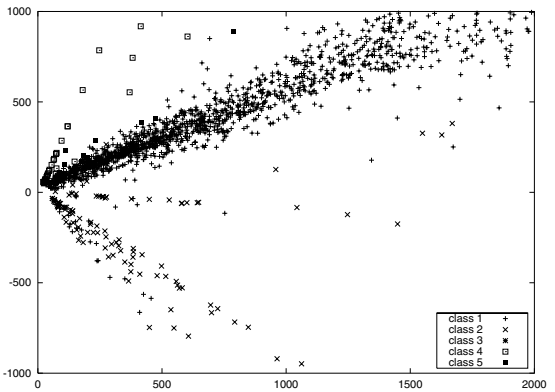


c)

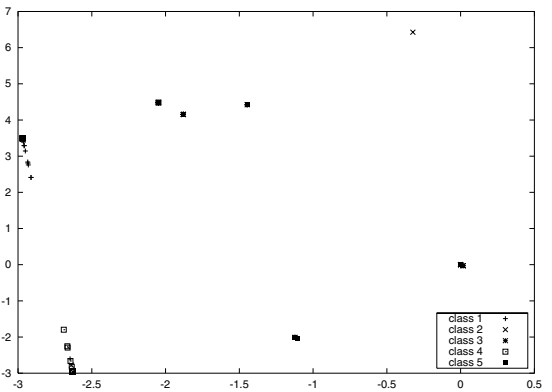
Fig. 3. *Iris* data set 2D projections using: a) PCA, b) GLP and c) NDA



a)



b)



c)

Fig. 4. *Page Blocks* data set 2D projections using: a) PCA, b) GLP and c) NDA

Table 1. Average error rates (%) of the Nearest Neighbour classifier (E_{NN}) computed over the four data sets

	Digits	Iris	Cookies	Page Blocks
ORIGINAL	3.3	4.0	0.4	9.2
PCA	56.3	4.0	42.7	11.4
GLP	24.0 ± 4.4	0.6 ± 0.6	0.3 ± 0.4	3.9 ± 0.8
NDA	0.2 ± 0.4	3.7 ± 1.3	0.0 ± 0.0	8.5 ± 1.5

3.2 Results

These data sets have been projected into a 2-dimensional space. In the case of GLP and NDA methods, 10 runs have been averaged for each data set with different initialisations values. The number of generations necessary to obtain GLP convergence for the *Digits*, *Cookies*, *Iris* and *Page Blocs* data sets was 300, 50, 25 and 50 respectively.

As can be seen from Figure 1a and 2a, PCA projections are not particularly meaningful for the *Digits* and *Cookies* data sets. In them, the directions of maximal data scatter are not interesting. Nevertheless, the projections obtained for the *Iris* and *Page blocks* data sets (Figures 3a and 4a) give an interesting view of the data structure. On the other hand, while GLP projection obtains a view of the *Iris* data set (Figure 3b) similar to the PCA projection, a more interesting view of the rest of data sets is obtained because the class information is considered. In Figure 2b, the cluster structure of the *Cookies* data set appears now clearly. In the same way, a much more meaningful view of the cluster structure from the *Digits* data set (Figure 1b) can be seen. Finally, the NDA projection shows the power of a supervised non-linear method extracting the cluster structure of the data sets. In the case of the *digits* data set, an remarkable view of its strong cluster structure can be seen (Figure 1c).

On the other hand, the study of E_{NN} values (Table 1) leads to similar conclusions. PCA obtains poor results for the *Digits* data set, this is not surprising considering that the original space is 196-dimensional. Results for the *Cookies* data set are particularly bad because the projection found by PCA, completely mixes the classes. GLP outperforms clearly PCA specially for this data set because the optimal projection is found. The NDA method shows that non-linear transformations are necessary to extract the class structure of the data when the intrinsic dimensionality is higher than the projected space dimensionality, this can be shown by the results obtained for data set *Digits*. For the remaining data sets, similar E_{NN} values as in the GLP method have been obtained. In some cases, the GLP method outperforms NDA, although the GLP algorithm is oriented to optimise this criterion, and therefore small differences of E_{NN} values are not important.

4 Conclusions

From the results obtained, it can be concluded that NDA projections outperform our GLP method for high dimensional data. In these cases, the NDA projection is able to extract the class structure even in a 2-dimensional projection. Nevertheless, we consider that NDA shows two important drawbacks. In the first place, because non-linear transformations are used, an important distortion of the original space is obtained, specially when projecting into a 2-dimensional space, trying to preserve the class structure. In these situations, a synthetic view of the configuration of real clusters is obtained. Moreover, the process of training an NDA neural network is not straightforward in many cases. The GLP method uses linear transformations, producing less distorted and more meaningful views of the original space (distortion can appear because the new axes are not necessarily orthogonal). Additionally, this method does not present the convergence problems of NDA networks. The PCA method is linear and does not present convergence problems, but it is an unsupervised method and therefore, the projections computed do not always show a good view of the class structure if the discriminant axes are not the ones with the higher variance.

References

1. C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998. http://www.ics.uci.edu/∼mlearn/ MLRepository.html, University of California, Irvine. 745
2. J. H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82(397), 1987. 743
3. K. Fukunaga. *Statistical Pattern Recognition*. Academic Press, second edition edition, 1990. 743
4. D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989. 744
5. J. H. Holland. *Adaptation in Natural and Artificial Systems*. Ann Tabor: The University of Michigan Press, 1975. 744
6. T. Kohonen. The self organizing map. *Proceedings IEEE*, pages 1464–1480, 1990. 743
7. B. Lerner, H. Guterman, M. Aladjem, I. Dinstein, and Y. Romen. On pattern classification with sammon's nonlinear mapping (an experimental study). *Pattern Recognition*, 31(4):371–381, 1998. 743
8. J. Mao and A. K. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks*, 6(2), 1995. 743
9. M. Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA, 1996. 744
10. J. W. Sammon. A non-linear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5), 1969. 743
11. L. Shyh-Chang, W. F. Punch III, and E. D. Goodman. Coarse-grain parallel genetic algorithms: Categorization and new approach. *Parallel and Distributed Processing*, 1994. 745
12. W. Siedlecki, K. Siedlecka, and J. Sklansky. An overview of mapping techniques for exploratory pattern analysis. *Pattern Recognition*, 21(5):411–429, 1988. 743