ESANN'1996 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-25-26 April 1996, D-Facto public., ISBN 2-9600049-6-5, pp. 229-236

INTERPRETING DATA THROUGH NEURAL AND STATISTICAL TOOLS

Anne Guérin-Dugué, Carlos Aviles-Cruz, Patricia M. Palagi
INPG-TIRF
46 Avenue Félix Viallet
F - 38031 Grenoble Cedex
email : guerin@tirf.inpg.fr

Abstract

We present here a set of neural and statistical techniques allowing simple interpretations of the structure of multidimensional databases. These techniques concern the estimation of the intrinsic dimension and the representation through a transformation into a low dimensional space (by a linear or a non-linear transformation). Two examples are given in the image analysis domain.

1. Introduction

Data analysis proposes a powerful set of techniques for representation, discrimination and classification. With neural techniques, we now have non-linear operators which complete these more classical linear methods in statistical data analysis. With all these tools and combinations of tools, the choice becomes difficult for selecting the appropriate technique for a given problem. This is one of the aims of data analysis, to provide indices and representations allowing a suitable choice of the processing techniques, by the joint interpretation of the results.

In this paper, we consider the domain of pattern recognition from multidimensional data. We show how several techniques (fractal geometry [10, 12], mean distances to the k-nearest neighbours [4, 11], Principal Components Analysis [4], Self-Organising Features Map [8,9], Curvilinear Component Analysis [3]) can provide important information on the intrinsic structure of a database. Here, we focus on the determination of the intrinsic dimension and on the linear or non-linear relations between the different features in the database.

This paper is divided into three parts. In the following section, we will briefly describe these different techniques and their interpretation. Then, an illustration will be given with two databases from an image segmentation application. Finally, a conclusion will present the perspectives of this work where the collected results must provide information both on the structure of the multidimensional database and also on the data generation process.

2. Description of the different techniques

We will describe different methods used to characterise high dimensional databases, from very simple parameters (such as inertia) to more complex ones (from non-linear transformations). The description will be illustrated by two examples with two databases (database 1 and 2). These databases have been generated from an application in image analysis where four different objects (classes) in the scene have been characterised with 16 features for database 1 [5] and with 18 features for database 2 [6]. For both databases, there are 100 samples per class.

2.1 Inertia parameters

Inertia parameters represent very classical and simple parameters in data analysis of multimodal distributions. These inertia (global, between-class and within-class inertia) give information on dispersion of the samples and on separability of the classes. Therefore, these inertia associated with the distances between the centres of gravity of each class (Fisher's coefficients) can give a very rough idea of the overlapping between classes [4]. In the the case of simple hyperspheric clusters, these parameters are well-adapted, but are not adapted to more complex structures (for example, fig. 1 for a 2D-database).

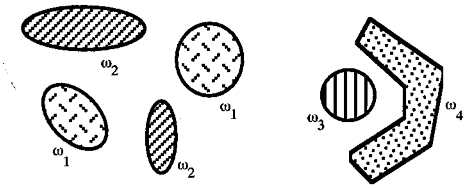


Fig. 1. Examples of clusters with complex shapes.

2.2 Intrinsic dimension

The intrinsic dimension represents the minimum number of independent variables underlying the process of data generation. When these intrinsic variables are linearly transformed to obtain the real database, the Principal Components Analysis (PCA) provides a powerful technique for dimension reduction. The evolution of the restored inertia versus the number of output dimensions produces a curve that gives an idea of the number of linearly correlated dimensions: a breakpoint, more or less sharp, is present in the curve (see fig. 2a and section 2.3.1).

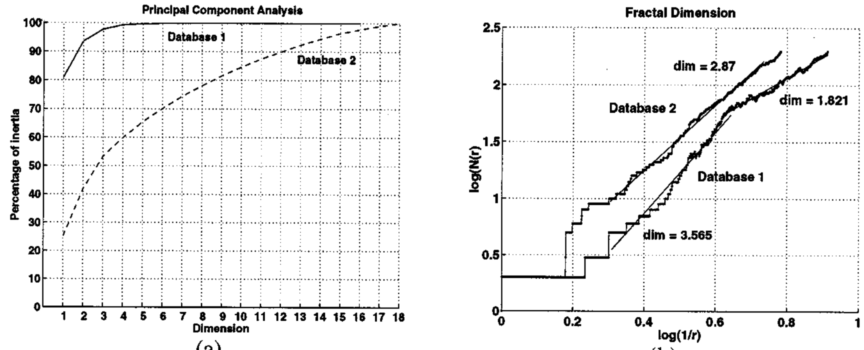
In [4], Fukunaga has proposed an estimation of the intrinsic dimension as a local dimension according to an approach which is similar to a local linearisation of a non-linear data structure. This dimension (d_i) for a distribution X is computed from the evolution of the mean distance between a sample and its k^{th} nearest neighbour, and the mean distance between a sample and its $(k+1)^{st}$ nearest neighbour (eq. 1). Other similar techniques have been proposed [7] to iteratively reach the intrinsic dimension, instead of a rough computation by averaging over several values of k. Here, we have implemented the Fukunaga's original method and this first estimation will be compared with other approaches: the fractal dimension, and the relations between distances with Curvilinear Components Analysis (see section 2.3.2). This analysis on the two databases gives similar results for an estimated value (d_i) of around 3 or 4 for both.

$$\frac{E\left\{d_{(k+1)NN}(X)\right\}}{E\left\{d_{kNN}(X)\right\}} \cong 1 + \frac{1}{k.d_i} \qquad (1)$$

This estimation is valid if the distribution X is locally uniform. More general methods [10] have been proposed based on the concept of the fractal dimension [12]. We consider here the similarity dimension. A hyper volume A is stated fractal if A is the union of *N(r)* non-overlapping copies of itself. Each copy is similar to A scaled down by a ratio *r*. The fractal dimension of A (*d_f*) is then

$$d_f = \lim_{r \to 0} \frac{\log(N(r))}{\log(1/r)}. \qquad (2)$$

In practice, d_f is the slope of a regression line from the graph *log(N(r))* versus *log(1/r)*. This is the "box counting method". This technique gives valid results if the number of available samples is sufficient. If not, the shape of the graph shows that the linear regression is not suitable. The range of the variable *r* , in which the slope will be estimated, gives the scale of this measure. If the scale is too coarse, the estimation is not accurate. With a scale too fine, the measure is very sensitive to noise. This is another argument towards the confrontation of different estimations. For the two databases, figure 2.b shows the evolution *log(N(r))* versus l*og(1/r)*. For database 1, at a median scale, the estimation of the fractal dimension is rather 4 (3.56). We do not consider the measure "1.82" estimated at a resolution too fine. For database 2, the estimation is rather 3 (2.87).



(a) (b)

Fig. 2. (a) Evolution of the percentage of restored inertia by PCA, (b) Estimation of the fractal dimension.

2.3 Representation into a low dimensional space

If the input dimension is more than 3, direct representation is visually inefficient. There exists techniques allowing a dynamic representation by rotation around 3 axes randomly chosen. Nevertheless here, we consider more classical "static" techniques through a reduction of dimension into a lower dimensional output space. Firstly, the information of the intrinsic dimension can help to determine of this output dimension. Secondly, the application of these techniques can produce new indices to determine the best output dimension, as will be shown afterwards.

2.3.1 Two well-known representations

By means of Principal Components Analysis (PCA), an input multidimensional space can be linearly transformed by projection into an output subspace from the most significant uncorrelated axes. On the evolution of the percentage of explained variance after PCA, we can see a different behaviour between the two databases (fig. 2.a). The input features seem to be more linearly correlated in database 1 than in database 2. Moreover, 3 or 4 features seem to be sufficient to restore database 1.

In the domain of artificial neural networks, Kohonen's Self Organising Features Map (SOFM) can be viewed as an non-linear extension of PCA [1]. SOFM's provide a data mapping between the high dimensional input space and a low dimensional output space [8]. In this mapping, both the number of samples and the dimension are reduced. When the intrinsic dimension of the data structure is higher than the dimension of the output map, the obtained organisation will fail to completely represent the initial structure. However, simple representations can be realised where the "neurons" are in a two or three dimensional space. Recently, based on 2D-SOFM, Kraaijveld et al. proposed [9] a new representation from the organisation on a 2D-map. A grey level is associated with each neuron (like a pixel in an image) which is proportional to the maximum distance in the SOFM between the current neuron and its four closest neighbours (E, W, N and S). Then, for a dense cluster, the associated spatial region will be rather dark. The boundaries between two different clusters will be enhanced by a whiter area (see examples in fig. 3 left). If the data are not initially and naturally structured into clusters, the produced image will be seen as very noisy. But, due to the problem of the possible mismatching between the dimensions, we can notice that if a cluster is visible in this image then it is really well separated in the initial space, but the inverse implication is not always true.

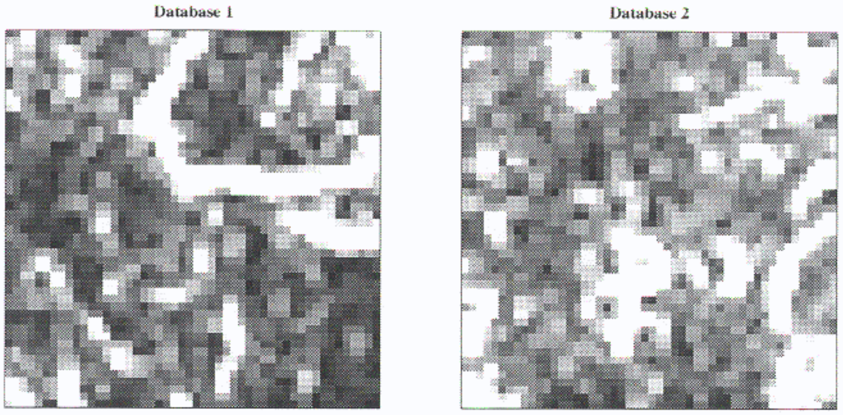


Fig. 3. Projection image from 2D-SOFM (50 x 50).

For both databases, the dimension of the output map (2D) is less than the intrinsic dimension. So, this mismatching does not allow a clear representation of the clustering. This is above all true for database 2: nothing can be deduced concerning the clustering (fig. 3 right). For database 1, one cluster for class one (top-right) is clearly visible (fig. 3 left).

2.3.2 A more suitable representation

The *major drawback* of SOFM is that the projection is realised from a predetermined shape of the map and from a predetermined structure of neighbourhood. These *a priori* configurations act upon the quality of the provided projection because these initial choices must fit the input data structure (some shape in some submanifold) which is a priori unknown. To overcome this drawback, Demartines [3] has proposed a new model called "Curvilinear Component Analysis". The principle is opposite to SOFM : instead of quantising the input distribution by a map whose shape and neighbourhood are predefined, the input distribution is first quantised and then non-linearly transformed into a low dimensional space, where the neurons themselves "find" themselves their neighbourhood according to the input topology. The only parameters to be a priori fixed are the number of neurons and the output dimension. Their locations are free (not constrained by a structure of neighbourhood, as in SOFM) and are adjusted according to the minimisation of an energy function (eq. 3). For this purpose, the distance Y_{ij} between two samples *i* and *j* evaluated in the output space must match the associated distance X_{ij} in the input space. By means of a weighting function $F(Y_{ij})$, short ranged distances are favoured to longer ranged distances. So, it is possible to re-shape a data structure by unfolding it into an output space of lower dimension (see figure 4 for a simple example).

$$E = \frac{1}{2}\sum_{i}\sum_{j\neq i}(X_{ij} - Y_{ij})^2 \cdot F(Y_{ij}) \qquad (3)$$

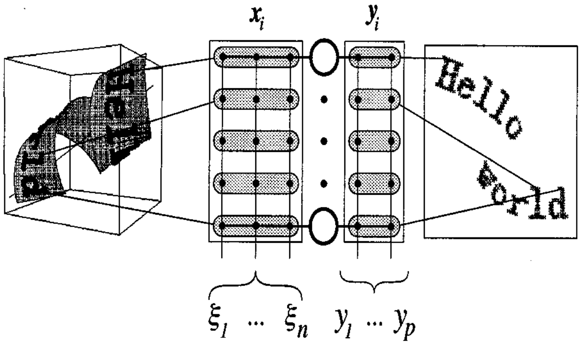


Fig. 4. CCA network, ξ_i are the synaptic weights from the vector quantisation, and y_i, the synaptic weights from the non-linear transformation. An illustration is given with a 3D non-linearly structured data-base projected into a 2D space matching the folded structure (from [3]).

The dimension of the output space can be determined from the evaluation of the intrinsic dimension. Moreover, in the case of the CCA, another feature can help this determination: the joint distribution of the distances X_{ij} and Y_{ij}, called the X-Y distribution (fig. 5.d). CCA preserves short range topology, so the X-Y distribution will be "thin" (X≅Y) for short distances X_{ij} and Y_{ij},. For a long range topology, X_{ij} and Y_{ij} will not be so well correlated. The evolution of the shape of this distribution versus the output dimension is another means to roughly estimate the instrinsic dimension. As the output dimension increases, the area of the linear regression will concern longer distances.

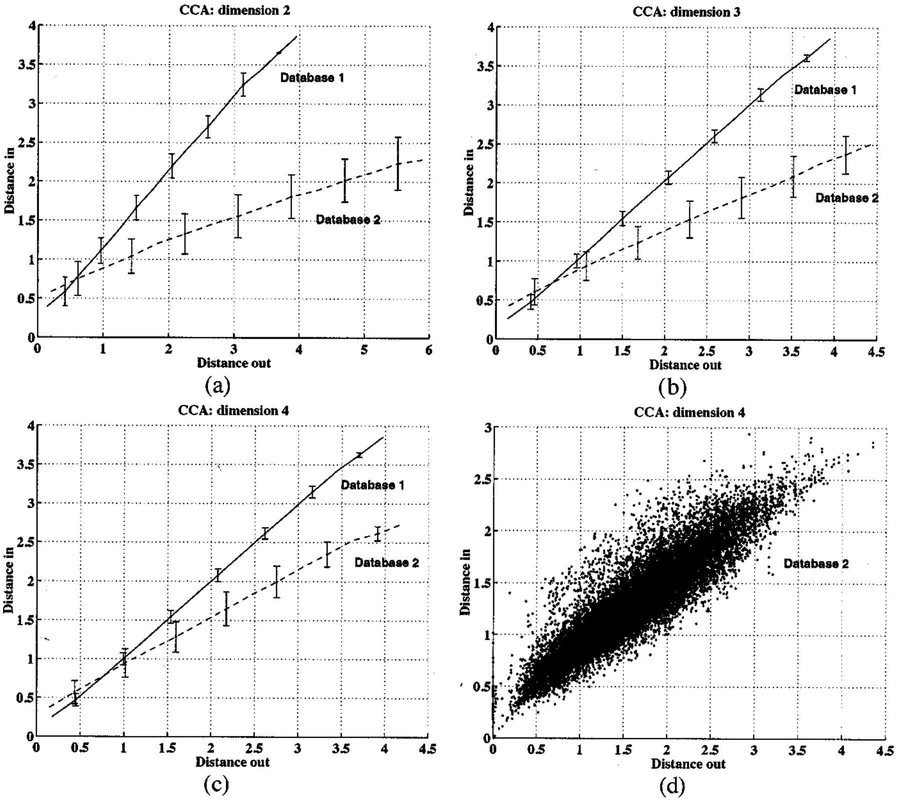


Fig. 5. (a)-(b)-(c)Evolution of the X-Y distribution when the output dimension increases from 2 to 4, (d) X-Y distribution obtained from distances between the 400 samples of database 2. The output dimension is 4.

In figures 5a-b-c, this distribution has been plotted for 3 different output dimensions (2, 3 and 4). To be clear, the representation of the distribution has been simplified : the "Distance out" (Y) axis has been split into 20 bins. In each bin, the means of X_{ij} and Y_{ij} are plotted and are associated to the standard deviation of X_{ij}. To simplify the graphics, only 7 bars of standard deviation have been drawn.

For database 1, we observe a mean X-Y distribution close to the bisecting line. This is true from the output dimension equal to 3. Moreover, the standard deviations decrease when the output dimension increases : the joint distribution X-Y becomes "thinner". These results confirm the interpretation of PCA. The 16 features of database 1 are linearly generated from very few independent features (around 3). For database 2, the situation is more complex. The X-Y distribution shows a growing evolution with a saturation effect of for the longer distances. This means that CCA has non-linearly re-shaped the input database 2 and thus a perfect preservation of topology is not realised for longer distances. This effect is even more visible when the output dimension is small compared to the intrinsic dimension. In figures 5a-b-c, we observe that the distribution X-Y for database 2 changes towards the bisecting line, and also the standard deviation decreases when the output dimension increases. The quality of the output representation doesn't significantly change with an output dimension greater than 4. For database 2, the 18 features are non-linearly generated from around 4 independent features.

3. Recapitulation of the results

With these two databases, two very different data structures have been observed. Both databases can be reduced with 3 or 4 intrinsic independent parameters. Database 1 is linearly structured from these parameters and database 2 is non-linearly structured.

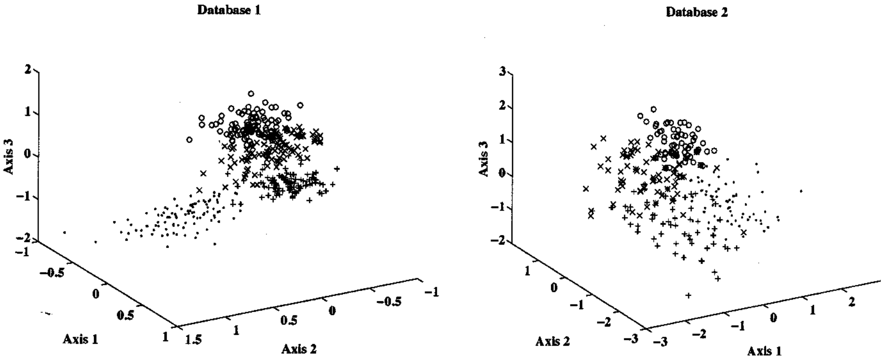


Fig. 6. Non linear transformation by CCA into an output 3D-subspace

In database 1, class 1 ('.') and class 2 ('+') are well clustered (see SOFM analysis in figure 3 left). For database 2, the clusters are less dense and seem to be closer. Here, while the databases contain few samples (400), CCA was implemented without the vector quantisation stage. It is interesting to have information about the intrinsic dimension. Then, there still remains the problem of performing the suitable data transformation into the intrinsic space. CCA is a candidate for this transformation.

4. Conclusions

This set of techniques is not exhaustive but represents coherent set for a first step in data analysis. From these results, different levels of interpretation can be performed at the level of the data structures and at the level of the data generation process. Here, we have only discussed the first level, and in particular the estimation of the intrinsic dimension and the linearity or the non-linearity of the database structure. We have shown how different measures (fractal geometry, distances from the K-nearest neighbours) can allow us to estimate this dimension and how PCA and CCA can give features for this calculation. On the one hand, the use of the evolution of the restored inertia after a PCA is very classical ; on the other hand, the use of information from the X-Y distribution with CCA is more original and useful. The confrontation of linear and non-linear methods in data analysis is very informative for the interpretation of the databases structure.

With thes reduction dimension techniques, the input databases are transformed into a subspace with 2 or 3 dimensions, then the direct representation can provide information on the clustering. But, the interpretations must be done knowing a possible mismatching between the resulting dimension and the intrinsic dimension. An alternative representation based on the distances between input samples, can be built by hierarchical classification [11]. These techniques are interesting because the representation is realised in the original input multidimensional space (there is no distortion due to a space transformation). This group of methods produce dendrograms where at each level, samples are agglomerated to form clusters according to the rules of the smaller distances between the clusters.

The perspective of this work is to continue reserach into the interpretation of the data structure to better formalise the link between the data structure and the choice and configuration of the processing techniques particularly in the case of unsupervised and supervised learning.

References

1. F. Blayo, P. Demartines: "Data analysis: How to compare Kohonen neural networks to other techniques", IWANN'91, vol 540, Springer Verlag, 1991.
2. P. Brodatz: "Textures: A Photographic Album for Artists and Designers", Dover Publications Inc., june 1966.
3. P. Demartines, J. Hérault: "CCA: Curvilinear Component Analysis", XV Colloque Gretsi, Juan-Les-Pins, pp. 921-924, septembre1995.
4. K. Fukunaga: "Introduction to Statistical Pattern Recognition", Academic Press Inc., San Diego, 2nd Edition, 1990.
5. A. Guérin-Dugué, C. Aviles-Cruz: "High order statistics from natural textured images", Workshop on System Identification and High Order Statistics, Sophia Antipolis, France, sept. 1993.
6. A. Guérin-Dugué, P. M. Palagi: "Texture Segmentation using Pyramidal Gabor Functions and Self-Organisation Feature Maps", Neural Processing Letters, vol 1, n° 1, pp. 25-29, 1994.
7. K. W. Pettis, T. A. Bailley, A. K. Jain, R. C. Dubes: "An Intrinsic Dimensionality Estimator from Near-Neighbor Information", IEEE Trans on PAMI, vol 1, n° 1, pp. 25-37, january 1979.
8. T. Kohonen: "The self-organizing maps", Proc. of the IEEE, vol 78, n°9, pp. 1481-1497, 1990.
9. M. A. Kraaijveld, J. Mao, A. K. Jain: "A Nonlinear Projection Method Based on Kohonen's Topology Preserving Maps", IEEE Trans on NN, vol 6, n° 3, pp. 548-559, may 1995.
10. B. B. Mandelbrot: "Fractal geometry of nature", Freeman, San Francisco, 1982.
11. K. M. Mardia, J. T. Kent, J. M. Bibby:□"Multivariate Analysis", Academic Press, San Diego, 1989.
12. C. Trichot: "Courbes et dimension fractale", Springer Verlag, 1993.