Visualisation and exploration of high-dimensional data using a “force directed placement”method: application to the analysis of genomic signatures

Sylvain Lespinats, Alain Giron, and Bernard Fertil

INSERM Unité 678, CHU Pitié-Salpêtrière
91 bd de l’hôpital, 75634 PARIS (France)

Abstract. Visualization of high-dimensional data is generally achieved by projection in a low dimensional space (usually 2 to 3 dimensions). Visualization is designed to facilitate the understanding of data sets by preserving some “essential”information. We have designed a non-linear multi-dimensional-scaling (MDS) tool relying on the force directed placement (FDP) algorithm to help dynamically discover features of interest in data sets. A user-driven relaxation of constraints built on the preservation of pairwise distances between data allows getting subjective representations of data that meet some specific angle. In a context of classification, we examine the impact of metric, sample size, and neighborhood preservation on the mapping of genomic signatures.

Keywords: Multi-Dimensional Scaling, Force Directed Placement, Classification, Proximity visualisation, Metric.

1 Introduction

High dimensional data raise unusual problems of analysis, given that some properties of the spaces they live in cannot be extrapolated from our current experience [Verleysen, 2001]. The notion of neighborhood in particular must be revised to take into account the number of dimensions. In particular (notably in the case of Euclidean spaces), we often face the problems of empty space and concentration of measure: when the number of dimensions is high, the neighborhood of each object is scarcely filled whereas most of the other objects are found in a thin outer shell. Distances between high dimensional objects are usually very concentrated around their average.

Exploration and analysis of high dimensional data are often made by means of dimension reduction techniques. Since human experience mostly deals with 3D space (and most data display devices are two-dimensional), finding a meaningful mapping of data in such low dimensional spaces is the issue. Principal component analysis (PCA), multidimensional scaling (MDS) [Cox and Cox, 1994], Kohonen maps (SOM) [Kohonen, 1997] are classic approaches in this context. In general, a loss function is defined to characterize the error in representing the dissimilarity between objects. It allows building the rules of projection from the original space of the data on to a smaller dimensional space. It is important to realize that any reduction of dimension leads to a subjective data representation. Depending on the purpose, different mappings may be achieved for the same set of data. For classification tasks, for example, the preservation of the neighborhood appears one of the aspects important to master. In this work, we examine some interesting mappings obtained by means of a nonlinear MDS-based projection. In particular, the consequences of dimension reduction on the classification of genomic signatures (256-dimension data originally) are analyzed.

2 Reduction of data dimension: principles ruling the present study

The approach that is presented here belongs to the MDS group of methods. It is thus advisable to define metrics for the original data space and for the target space (called output space thereafter), a lost function and a mapping algorithm. Usually, the characteristics of the data to be analyzed are to be considered to choose these various elements.

2.1 Data, metric and lost function

Data under investigation in this work concern the genomic signature. The whole set of short oligonucleotide frequencies observed in a DNA sequence is species-specific and is thus considered as a genomic signature (Deschavanne et al., Karlin et al.). The genomic signature characterizes the DNA molecule by 256 frequency variables, defined in the range [0-1]. Counts (and frequencies) of oligonucleotides can be displayed as parametric images allowing fast visual examination and comparison (http://genstyle.imed.jussieu.fr). It has been observed that the genomic signature results from a species-specific "writing style"[Deschavanne *et al.*, 1999]. Indeed, on one hand, the genomic signatures of species differ from one another; on the other hand, the majority of DNA segments isolated from the genome of a given species have comparable signatures. As a consequence, each species is given a genomic signature that can be derived from most of its available DNA fragments. The DNA style is obtained from the examination of relatively small chains of the genetic material. In practice, a sequence as short as 2000 nucleotides usually provides a good estimate.

The Euclidean metric allows showing statistically significant differences between species' genomic signatures [Deschavanne *et al.*, 1999]. This metric will thus be chosen to illustrate the method, for typical examples at first (projection from a 3D space towards a 2D space), then for the problem of classification of genomic signatures. In some instances, we may consider preserving only the rank order of distances between objets, not the exact values.

Such a procedure is found useful when the projection provides "unsatisfactory"results. The projection should then try matching the rank order of distances between objects in the two-dimensional output space to the rank order in the original space.

The lost function is defined as a weighted sum of errors over dissimilarities (distances or ranks) between all pairs of objects in the original space and the output space. Eventually, subsets of data may be considered to test the robustness of projection. A part of data is used to define the mapping whereas the remaining part serves checking representiveness of output space. In order to preferentially favor close proximity, a weighting scheme reducing the impact of errors related to large dissimilarities may be gradually applied during the phase of optimization. This approach takes benefit from the work by P. Demartine and J. Herault [Demartine and Herault, 1997] and T. Kohonen [Kohonen, 1997].

2.2 Loss function minimization algorithm

In general, the optimal position of data in the output space cannot be obtained analytically. It is necessary to implement a function minimization algorithm with widely recognized robustness and convergence aptitudes. Classically, in the context of MDS, one alternatively uses the generalized Newton-Raphson algorithm, TABU Search [Glover and Laguna, 1995], genetic algorithms [Goldberg, 1989] or simulated annealing [Dowsland, 1995].

Regarding our model (called FDP-MDS thereafter), we propose to set up a dynamic algorithm grounded on the "Force Directed Placement"paradigm (FDP) [Fruchterman and Reingold, 1999]. Firstly described at the beginning of the Eighties, the FPD method is yet popular in only a limited number of fields. In particular, it is extensively used for the design of printed circuits. It is on the other hand little known in the field of data analysis. The force directed placement metaphor may be clarified in the following way: the data to place in the output space are bounded by forces (materialized by springs for example) the magnitude of which are related to the satisfaction of dissimilarities. In the case of springs, length at rest corresponds to the dissimilarity between the connected objects in data space. Any departure from the resting value consequently results in a recall force contributing in the movement of object and accounting for the energy of the system. Starting from an initial state with the objects placed the most judiciously possible in output space, the system is allowed to relax towards a minimum state of energy for which the constraints of dissimilarities between objects are satisfied as much as possible. FPD algorithm is very interesting in the case of MDS, considering its speed of convergence and its possibilities to escape from local minima.

For problems dealing with few thousands of objects, it is possible to directly run the FDP algorithm with the whole set of data. For larger data collections, it is often interesting to select a subset of objects to coarsely define the topology of the output space, in a first step. Remaining data are subsequently positioned with respect to preceding ones, by preferentially satisfying local constraints. In our hands, the incremental approach shows up very effective, especially when initial objects are selected after clustering.

2.3 Non-linear projection achieved by FDP-MDS: examples

Two boxes: Data to be projected have three dimensions. Objects are organized to represent 2 cubic boxes with an open side not pointing in the same direction. Projection onto a 2D space with FDP-MDS correctly develops the 2 boxes and carries out a twist on a large scale (fig. 1). Relations of vicinity are satisfactorily preserved.

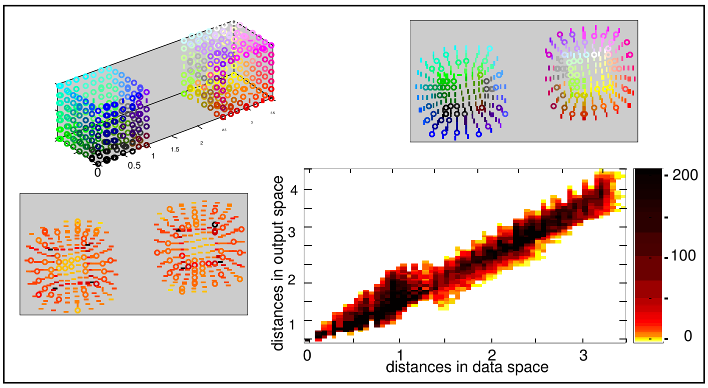

Fig. 1. Mapping of 2 3D open boxes in a 2D space. Upper left: original data (3D space), upper right: mapping (2D space), lower left, satisfaction of constraints on objects (satisfaction increases from black to white, LUT of fire), lower right, pairwise distances preservation (color codes for density of distances). NB: Colored figures are available from our WEB site <http://e6.imed.jussieu.fr/afficherpub.php/ASMDA05.pdf>

Earth globe: Data to be projected are the big cities around the word (3D). Projection accounts for local density of cities. The north hemisphere is properly developed (Fig. 2). Cities-free areas are distorted although continuity is preserved in most places (The grid is not used during the mapping construction).

2.4 Mapping high dimensional data: the genomic signature issues

The data concerned with this study belong to two families; the signatures of 5000 species constitute a subset of the diversity of ADN molecules on earth. The signature of a species, *B. subtilis*, is studied in detail. One thousand

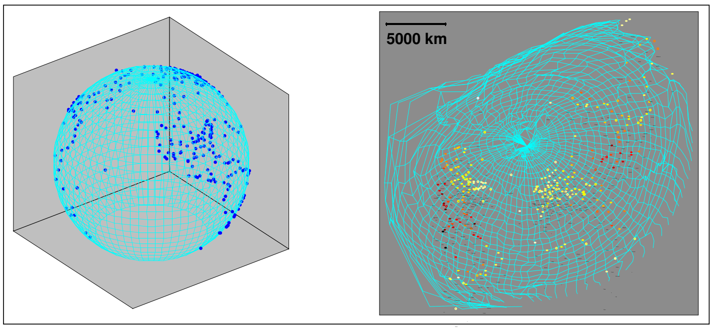

Fig. 2. Mapping of the earth globe (defined by the big cities) in a 2D space. Color indicates satisfaction of pairwise distances for the corresponding city (Color scheme is similar to Fig. 1).

eight hundred and twenty four signatures corresponding to the analysis of *B. subtilis* genome through a sliding window of preset size (5000 nucleotides) are calculated. The signature of each of these windows (called local signatures thereafter) generally displays the characteristics of *B. subtilis*. All signatures are defined by 256 frequency variables.

The first issue to be addressed in this work concerns the effect of sampling on the mapping of high dimensional data. Five hundred local signatures of *B. subtilis* are randomly selected to build a proximity preserving 2D output space (Fig 3, left panel). The 1324 remaining signatures are subsequently placed, using the FDP algorithm. It appears clearly that the mapping is not suitable to handle the diversity of local signatures of *B. subtilis*. Most of the signatures that were not considered for the mapping are concentrated around the center of the space, whereas a randomly placement would be expected. Obviously, pairwise distances between 500 local signatures are not enough to properly describe the proximity characteristics of these highly dimensional objects. New objects cannot fit in the output space. It must be pointed out that this peculiar behavior is not observed for the 5000 genomic signatures although their dimension is the same (result not shown). It is suggested that the intrinsic dimension of signatures is the key to explain this surprising result. Local signatures may stretch over most of the avaible dimensions (sampling effect) whereas variations among genomic signatures only concern specific directions characterizing the restricted set of possible pathways for species differentiation.

Surprisingly, switching from the Euclidean metric to the rank pseudo-metric solves the problem (Fig. 3, right panel)! It may be considered that the mapping obtained using the rank pseudo-metric is robust to sampling size, but additional experiments and theoretical developments are required to firmly conclude on this point.

The second issue deals with classification. Local signatures are expected similar to the genomic signature of the species they come from. It should be

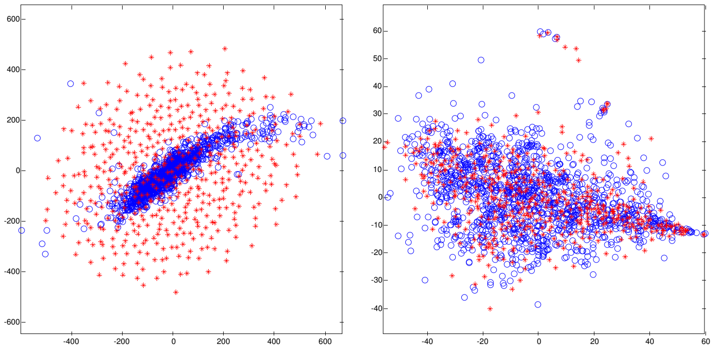

Fig. 3. Mapping of *B. subtilis* local signatures. Red crosses (500) are for signatures used to construct the mapping, blue circles (1324) are for additional local signatures placed afterwards.

subsequently possible to search for the species of origin of any local signature of *B. subtilis*, using a nearest neighbor classifier exploring the 5000 genomic signature set. Within the framework of this paper, 2 situations are considered: i) the mapping is learned using the species' signatures, ii) the mapping is learned with all available signatures, including *B. subtilis*' local signatures.

In data space (256 dimensions), only 64% of local signatures are correctly assigned to *B. subtilis*. In fact there are about one hundred of species in the hyper-sphere holding 95% of local *B. subtilis*' signatures, some of them being even very close to *B. subtilis*. It should be noted that an important subset of local signatures is misclassified for known biological reasons. When the space of projection is learned from the species' signatures, the rate of good classification falls to 0,7%(fig. 4, left panel). It is 24% when the space of projection is learned from the whole set of signatures (species and local, fig. 4, right panel). The zone devoted to local signatures in the output space is extended to satisfy constraints of distances between local signatures when they are included in the training sample. Even so, quality of classification remains poor.

3 Discussion et conclusion

The nonlinear approach of mapping described in this article was designed to preferentially preserve proximity. For small dimension problems, it appears that its effectiveness is quite good. It is unfortunately not the case for high dimension data where the learning sample size seems to be a critical parameter and the efficiency of local signature nearest neighbor classification is strongly reduced in the output space. The method of classification used in this work is particularly sensitive to "errors"of placement since only one "mis-placed"species may cause multiple classification errors. However, this situation is likely to occur many times in such dramatic reductions of dimension (256 towards 2-3). Considering that the growth of neighborhood with increasing radius (around every object) in a high dimensional space cannot be effectively matched in a low dimensional space, only data with a small intrinsic dimension may be properly mapped in a small dimensional Euclidean space.

An interesting alternative is proposed by H. Ritter and J. Walter [Ritter, 1999] [Walter and Ritter, 2002]: they use a 2-dimensional hyperbolic plane as output to simulate the singular growth of neighborhood of high dimensional space. The approach seems very promising. The learning sampling size is also an important parameter to master. Obviously, the conjunction of the empty space phenomenon with the singular growth of neighborhood in high dimensional space make the sampling phase (when required) particularly tricky. All together, it seems useful to recall that the analysis of the data resulting from consequent compression ratios must be carried out with infinite precautions.

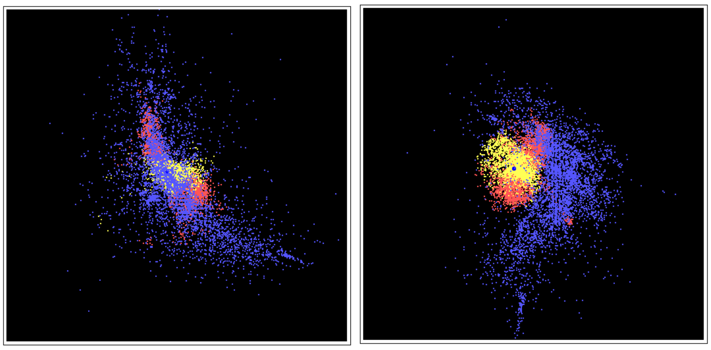

Fig. 4. mapping of genomic signatures in a small dimensional space: Species' signatures are in blue (dark), well-classified local *B. subtilis*' signatures (in the data space) are in yellow (light), mis-classified signatures are in red (see text). Left panel: mapping obtained with species' genomic signatures, right panel: mapping obtained with the full set of available signatures (species and local).

4 Acknowledgments

INSERM (Institutional funding), Action INTER EPST Bio-informatique 2001 Grant, contract N° 120910

References

[Cox and Cox, 1994]T. Cox and M. Cox. *Multidimensional Scaling.* Chapman & Hall, London, 1994.

[Demartine and Herault, 1997]P. Demartine and J. Herault. Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans. Neural Networks*, 8:148–154, 1997.

[Deschavanne *et al.*, 1999]P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.*, 16(10):1391–9, 1999.

[Dowsland, 1995]K.A. Dowsland. Simulated annealing. In C.R. Reeves, editor, *Modern techniques for combinatorial problems*, chapter 2. McGraw-Hill Book Company, Berkshire, 1995.

[Fruchterman and Reingold, 1999]T. Fruchterman and E. Reingold. Graph drawing by force-directed placement. *Software-Practice and Experience*, 21:1129–64, 1999.

[Glover and Laguna, 1995]F. Glover and M. Laguna. Tabu search. In C.R. Reeves, editor, *Modern euristic techniques for combinatorial problems*, chapter 3. McGraw-Hill Book Company, Berkshire, 1995.

[Goldberg, 1989]D.E. Goldberg. *Genetic algorithms in search, optimization and machine learning.* Addison-Wesley, Reading, Massachusetts, 1989.

[Kohonen, 1997]T. Kohonen. *self organizing maps.* Springer-Verlag, 1997.

[Ritter, 1999]H. Ritter. Self-organizing maps on non-euclidean spaces. In S. Oja and E. Kaski, editors, *Kohonen maps*, pages 97–110. Elsevier, Amsterdam, 1999.

[Verleysen, 2001]M. Verleysen. Learning high-dimensional data. In *NATO advance research workshop on limitation and future trends in neural computing*, Siena, Italy, 2001.

[Walter and Ritter, 2002]J.A. Walter and H. Ritter. On interactive visualization of high-dimensional data using the hyperbolic plane. In *SIKDD '02*, Edmonton, Alberta, Canada, 2002.