DOI: 10.1017/S1431927605502939
Microsc Microanal 11(Suppl 2), 2005

Building Order in Large Image Data Sets: Classification Techniques at Work

José María Carazo.

National Center for Biotechnology (CSIC-CNB), Campus Universidad Autónoma, 28049 Madrid, Spain

Introduction.- The theme of this tutorial, "Building Order in Large Image Data Sets", come as a consequence of the drive towards automation and digital control and recording that characterize modern microscopy approaches. The widespread introduction of these techniques result in the technical capability to obtain a large number of images under defined sample and observation conditions. However, large collections of raw data are a valuable asset only if we are able to extract new knowledge out of them, and one way to extract this knowledge is by detecting patterns in the data such that the data were grouped according to these patterns. When we have large collections of data sets, normally organized in data bases, the term "data mining" is used to refer to this pattern finding approach.

Intuitively, users would like to discover "relevant/significant" patterns in the data. However, there is no single way to code what "relevant" or "significant" mean to different users, although it is clear that the user "opinion" is key. Somehow, the user has to be included in the data mining process. In fact, we will review a number of techniques and it will become clear that those techniques that produce a graphical and intuitive result, in many cases grouping, are preferred in practical cases. Excellent introductory books on this subject are cited in [1]. In the following I will sketch the basic steps of pattern detection in the data, mainly focused on the goal to detect reasonably homogeneous groups among the image samples. Of course, they will be treated in detailed along the tutorial.

Making a simplification: Dimensionality reduction.- The combination of a rich experimental reality with many degrees of freedom with a large collection of these images result in a combinatorial explosion of the complexity of the overall pattern recognition study of these data sets. This large complexity precludes their analysis by straight forward, brute force, approaches. Therefore, ways to reduce this complexity are needed, resulting in the so called dimensionality reduction approaches. The central idea is to change the input original images for a simplified version of them in which only the most pronounced trends of variability are kept. There exists many different ways to achieve such a reduction, from linear methods such as Principal Components Analysis (PCA) [3] to non-linear ones such as Self-Organizing neural networks (SOM's) [4] (Figure 1), to mention just two of the most widely used approaches. We will review some of them, including PCA and SOMs, highlighting the "simplification" process that the images go through (Figure 1). Naturally, the success of this approach will be extremely dependant on the ability to perform that simplification such as the key discriminative information among different samples is retained. As before, "relevant" trends of variability are sought, and methods that present these trends in a graphical manner will be preferred in most of the application.

Clustering and classification: Working in these reduced spaces the task will be to group images into "beams" such as the elements of the same beam are more similar among them than to elements of

other beams, and here we will review from classical hierarchical ascendant classification methods (HAC) to fuzzy partional ones such as Fuzzy C-means.

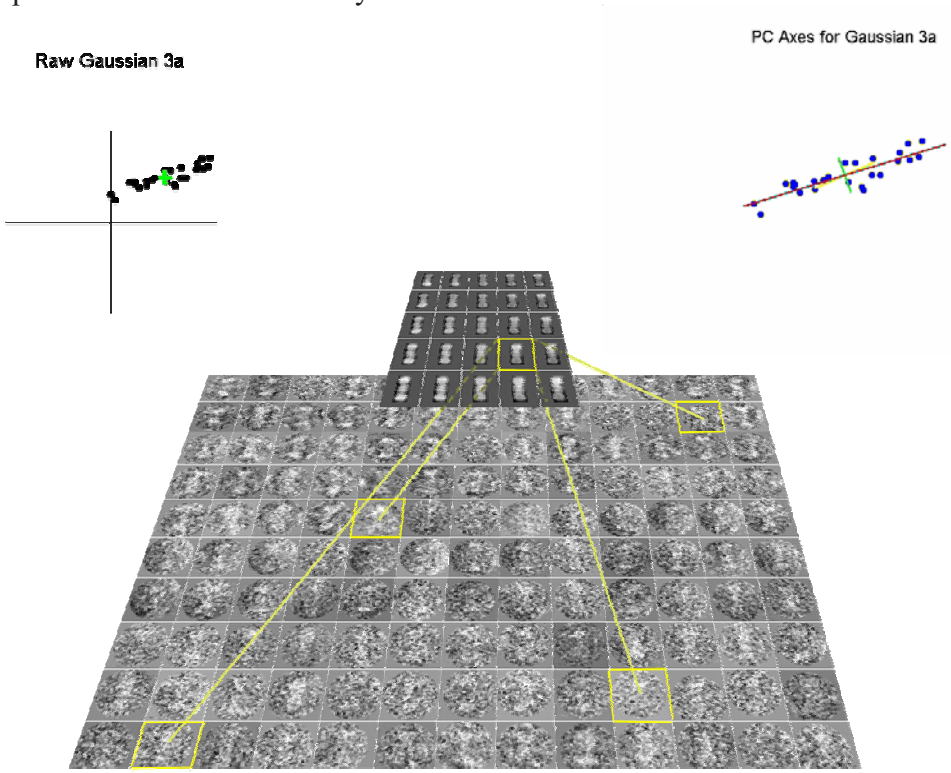

Figure 1: On the Top Left: a two dimensional plot of a set of samples characterized just by two degrees of freedom represented as their projections onto a system of coordinates of two arbitrary orthogonal axis. On the Top Right, the result of rotating these two orthogonal axis such as the first one explain most of the "information" that the samples contain. This is what PCA does. Bottom part: An application of Self Organizing Neural Network to the analysis of a large collection of noisy images. The output is a non-lineal projection of the noisy input images into a reduced space of a much smaller number of much cleaner images. The trends of variability in the input image population can be deduced observing the trends in the reduced dimensionality space.

References

[1] O. Duda, P. E. Hart and D. G. Stork, Pattern Classification (2nd ed.), John Wiley and Sons, 2001.

[2] M. Friedman and A. Kandel, Introduction to Pattern Recognition, statistical, structural, neural and fuzzy logic approaches, World Scientific, Singapore, 1999.

[3] H. Hotelling, Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, vol. 24, pp. 417-441, 498-520, 1933.

[4] T. Kohonen, *Self-Organizing maps*, Second ed: Springer-Verlag., 1997.

[5] Acknowledgements: I acknowledge finantial support from Spanish Comisión Interministerial de Ciencia y Tecnología (BFU2004-00217/BMC), Comunidad Autónoma de Madrid (07B/0032/2002; GR/SAL/0342/2004), National Institute of Health (HL67465-01 ; HL70472), European Union (-2003-508833; FP6-502828), and Fundación BBVA (UCAM2004030013).

https://doi.org/10.1017/S1431927605502939 Published online by Cambridge University Press