Invited Session *Mining Large High Dimensional Data*

Conference ASMDA 05, Brest, France, May, 17-20, 2005

Organizer :

Stéphane Lallich, Université de Lyon, Laboratoire ERIC, Lyon 2,

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.

About the locality of kernels in high-dimensional spaces

Damien Francois, Vincent Wertz,
UCL – CESAME, Avenue G. Lemaitre, 4, B-1348 Louvain-la-Neuve, Belgium
{francois,wertz}@auto.ucl.ac.be
Michel Verleysen, UCL - Machine Learning Group
Place du Levant, 3, B-1348 Louvain-la-Neuve, Belgium
verleysen@dice.ucl.ac.be

Abstract. Gaussian kernels are widely used in many data analysis tools such as Radial-Basis Function networks, Support Vector Machines and many others. Gaussian kernels are most often deemed to provide a local measure of similarity between vectors. In this paper, we show that Gaussian kernels are adequate measures of similarity when the representation dimension of the space remains small, but that they fail to reach their goal in high-dimensional spaces. We suggest the use of p- Gaussian kernels that include a supplementary degree of freedom in order to adapt to the distribution of data in high-dimensional problems. The use of such more flexible kernel may greatly improve the numerical stability of algorithms, and also the discriminative power of distance- and neighbor-based data analysis methods.

Keywords: High dimensional spaces, Local Models, Gaussian Kernels.

Quality measure based on Kohonen maps for supervised learning of large high dimensional data

Elie Prudhomme and Stéphane Lallich
Université Lyon 2, Laboratoire ERIC
5, avenue Pierre Mendès-France, 69676 BRON Cedex, France
stephane.lallich@univ-lyon2.fr, elie.prudhomme@etu.univ-lyon2.fr

Abstract. In supervised learning, the prediction of the class is the ultimate goal. On a broader basis, a good learning methodology is expected to (1) enable a representation of the data in order to facilitate user's navigation within the data set and (2) contribute to the choice of examples and attributes, while ensuring a structured, understandable prediction. Various studies have shown how the so-called neighbourhood graph, from the predictors, gives ground to such a methodology (e.g.: the relative neighbourhood graph of Toussaint). However, the construction of such a graph (O(n3)) remains complex. Moreover, when the number of dimensions increases, distance becomes hard to compute and lose their selectivity. In the case of large high dimensional dataset, we propose to substitute a selforganized map built on the predictors to the neighbourhood graph. After a short reminder on the principles of the SOM for unsupervised learning, we analyse how it can found an optimized strategy of learning. Then we propose to use original statistics (narrowly correlated with the error in generalization) in order to assess the level of quality of this strategy. Diverse experiments highlight the feasibility of this approach, therefore reliable criterion are available for us to select relevant examples and attributes.

Keywords: supervised learning, Kohonen maps, statistical validation.