Advanced Data Mining Methodology Based on Latent Variable Models

Antonino Staiano [1,2], Roberto Tagliaferri [1,2], Lara De Vinco[1] and Giuseppe Longo [3]

[1] Dipartimento di Matematica ed Informatica, Università di Salerno, Via Ponte don Melillo, 84084 Fisciano (Sa), Italy
astaiano@unisa.it
[2] INFM Unità di Salerno, Salerno, Italia
[3] Dipartimento di Scienze Fisiche, Università Federico II di Napoli, Italy

Keywords. Latent Variable Models, Probabilistic Principal Surfaces, Neural Networks, Data Mining, Genomics

Introduction

Aim of this paper is to show a powerful tool for data mining activities based on a nonlinear latent variable model, i.e. Probabilistic Principal Surfaces (PPS) [1], [2]. PPS builds a probability density function of a given data set of patterns, lying in a D-dimensional space, which can be expressed in terms of a limited number of latent variables lying in a Q-dimensional space. Usually, Q is 2 or 3 dimensional and thus the density function is used to visualize the data in the latent space. PPS have been fruitful exploited for classification as well as visualization and clustering of complex real high-D data [3] and represents a promising data mining tool for researchers in genetics and bioinformatics.

Spherical PPS

PPS generates a non linear manifold passing through the data points defined in terms of a number of latent variables and of a nonlinear mapping from latent space to data space. Depending upon dimensionality of the latent space (usually at most 3-dimensional) one has 1D, 2D or 3D manifolds. Among the 3D manifolds, PPS permits to build a spherical manifold where the latent variables are uniformly arranged on a unit sphere. This particular form of the manifold provides a very effective tool both for data classification and visualization, reducing the problems deriving from the curse of dimensionality when data dimension increases (data tend to be sparse and located at the periphery) with respect to other neural models, i.e. Self-Organizing Maps (SOM), or Generative Topographic Mapping [2], [3].

1. Spherical PPS for Classification

Spherical PPS can be used as classifier by determining a template spherical manifold for each class of the problem and assigning a new data point to the class of its nearest template manifold, or alternatively assigning a test data choosing the class with the maximum posterior class probability for a given new input. In [3] two combining schemes to build ensembles of spherical PPS aimed to combine the density estimation of each component to computed improved probability density estimations are proposed. The ensemble classifier gained meaningful improved performance with respect to the single PPS classifier.

2. Spherical PPS for Visualization

After the projection of the data on the sphere, it is advisable for a data analyzer, such as in genetics and biomedicine, to localize the most interesting data points, for example the ones lying far away from more dense areas, or the ones lying in the overlapping regions between clusters, and to gain some information about them, by

linking the data points on the sphere with their position in the database which contains all the information about the typology of the data. Therefore, we allow the user to:
-Easily interact with the data into the latent space, hence with the data on the sphere in several ways.
-Visualize the data probability density in the latent space so giving a first understanding about the clusters in the data. A first insight on the number of agglomerates localized on the spherical latent manifold is provided by the mean of the responsibilities for each latent variable.
-Fix a number of clusters and visualize it: once the user or a data analyzer has an overall idea of the number of clusters on the sphere, he can then exploit this information through the use of classical clustering techniques (such as hard or fuzzy k-means) to find out the prototypes of the clusters and the data therein contained. This task is accomplished by running the clustering algorithm on the projected data.
The application of these visualization possibilities (see Fig. 1 as an example) to gene expression data (in a way similar to the works described in [4],[5] which used SOM as neural model) is still in progress.

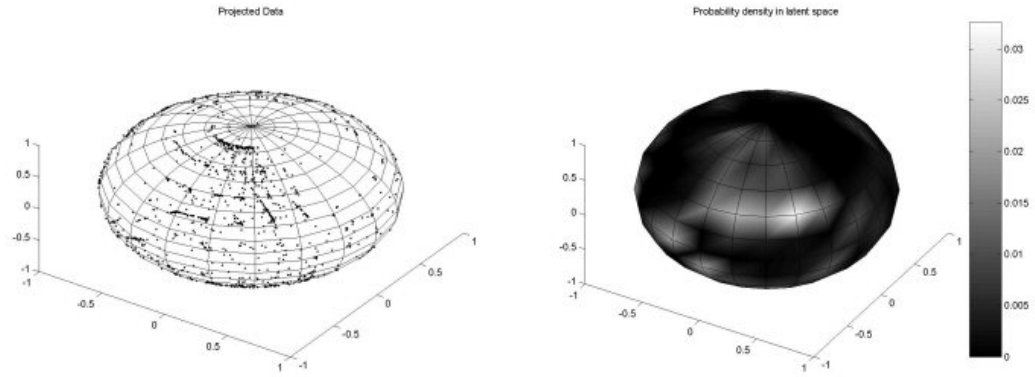

Fig. 1 Data point projections (left) and probability densities in the latent space (right)

References

[1] K. Chang, J. Ghosh, A unified Model for Probabilistic Principal Surfaces *, IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, NO. 1, 2001

[2] K. Chang, J. Ghosh, Nonlinear Dimensionality Reduction Using Probabilistic Principal Surfaces, PhD Thesis, Department of Electrical and Computer Engineering, The University of Texas at Austin, USA, 2000

[3] A. Staiano, Unsupervised Neural Networks for the Extraction of Scientific Information from Astronomical Data, PhD Thesis, Dipartimento di Matematica ed Informatica, Università di Salerno, Italy, 2004

[4] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, T. R. Golub, Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, *Proc. Natl. Acad. Sci., USA,* Vol. 96, 2907-2912, March, 1999

[5] P. Törönen, M. Kolehmainen, G. Wong, E. Castrén, Analysis of gene expression data using self-organizing maps*, FEBS Letters*, 451, 142-146, 1999