Classification in high dimension

Stéphane Girard
INRIA Rhône-Alpes & LJK (team MISTIS).
655, avenue de l’Europe, Montbonnot. 38334 Saint-Ismier Cedex, France
Stephane.Girard@inria.fr

Abstract: This report summarizes my contributions in high dimensional classification and/or clustering.

Clustering in high-dimensional spaces is a recurrent problem in many fields of science, for example in image analysis. Indeed, the data used in image analysis are often high-dimensional and this penalizes clustering methods. In this paper, we focus on model-based clustering method. Popular clustering methods are based on the Gaussian mixture model and show a disappointing behavior when the size of the dataset is too small compared to the number of parameters to estimate. This well-known phenomenon is called curse of dimensionality.

To avoid overfitting, it is necessary to find a balance between the number of parameters to estimate and the generality of the model. I proposed a Gaussian mixture model which takes into account the specific subspace in which each cluster is located and therefore limits the number of parameters to estimate. The Expectation-Maximization (EM) algorithm is used for parameter estimation and the intrinsic dimension of each group is determined automatically either with the scree-test of Cattell or by maximum likelihood [1]. This allows to derive a robust clustering method in high-dimensional spaces, called High Dimensional Data Clustering (HDDC) [2]. The method has also been adapted to supervised classification (HDDA – High Dimensional Data Analysis) [3, 4] and to the label noise situation [5]. In order to further limit the number of parameters, it is possible to make additional assumptions on the model. We can for example assume that classes are spherical in their subspaces or fix some parameters to be common between classes. Finally, HDDA and HDDC are evaluated and compared to standard clustering or classification methods on artificial and real datasets. These approaches are shown to outperform existing clustering methods [6]. The methods are implemented in a R package [7, 8] which is freely available on the CRAN archive. Finally, the extension to the classification of non necessarily quantitative data is investigated in [9].
References


