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, 7] on spectroscopic data and astrophysics. The methods are implemented in a R package [8, 9] which is freely available on the CRAN archive.

Finally, the extension to the classification of non necessarily quantitative data is investigated in [10]. Applications are developped in grassland classification [11, 12], verbal autopsy [13], and hyperspectral remote sensing [14].

References

[1] C. Bouveyron, G. Celeux, and S. Girard. Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA. *Pattern Recognition Letters*, 32(14):1706–1713, 2011.

[2] C. Bouveyron, S. Girard, and C. Schmid. High dimensional data clustering. *Computational Statistics and Data Analysis*, 52:502–519, 2007.

[3] C. Bouveyron, S. Girard, and C. Schmid. High dimensional discriminant analysis. *Communication in Statistics - Theory and Methods*, 36(14):2607–2623, 2007.

[4] C. Bouveyron, S. Girard, and C. Schmid. Class-specific subspace discriminant analysis for high-dimensional data. In C. Saunder et al., editor, *Lecture Notes in Computer Science*, volume 3940, pages 139–150. Springer-Verlag, Berlin Heidelberg, 2006.

[5] C. Bouveyron and S. Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, 42(11):2649–2658, 2009.

[6] J. Jacques, C. Bouveyron, S. Girard, O. Devos, L. Duponchel, and C. Ruckebusch. Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data. *Journal of Chemometrics*, 24:719–727, 2010.

[7] S. Girard and J. Saracco. Supervised and unsupervised classification using mixture models. In D. Fraix-Burnet and S. Girard, editors, *Statistics for astrophysics - Clustering and classification*, volume 77, pages 69–90. EDP Sciences, 2016.

[8] L. Bergé, C. Bouveyron, and S. Girard. HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. *Journal of Statistical Software*, 46(6):1–29, 2012.

[9] C. Bouveyron and S. Girard. Classification supervisée et non supervisée des données de grande dimension. *La revue de Modulad*, 40:81–102, 2009.

[10] C. Bouveyron, M. Fauvel, and S. Girard. Kernel discriminant analysis and clustering with parsimonious Gaussian process models. *Statistics and Computing*, 25:1143–1162, 2015.

[11] M. Lopes, M. Fauvel, S. Girard, and D. Sheeren. Object-based classification of grasslands from high resolution satellite image time series using Gaussian mean map kernels. *Remote Sensing*, 9(7), 2017.

[12] M. Lopes, M. Fauvel, A. Ouin, and S. Girard. Spectro-temporal heterogeneity measures from dense high spatial resolution satellite image time series: Application to grassland species diversity estimation. *Remote Sensing*, 9(993), 2017.

[13] S. Sylla, S. Girard, A. Diongue, A. Diallo, and C. Sokhna. A classification method for binary predictors combining similarity measures and mixture models. *Dependence Modeling*, 3:240–255, 2015.

[14] M. Fauvel, C. Bouveyron, and S. Girard. Parsimonious Gaussian process models for the classification of hyperspectral remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 12(12):2423–2427, 2015.