Classification supervisée et non supervisée des données de grande dimension

Charles BOUVEYRON[1] & Stéphane GIRARD[2]

[1] SAMOS-MATISSE, CES, UMR CNRS 8174
Université Paris 1 (Panthéon-Sorbonne)
90 rue de Tolbiac, 75634 Paris Cedex 13, France

[2] Mistis, INRIA Rhône-Alpes & LJK
655 avenue de l'Europe, 38330 Saint-Ismier Cedex, France

Résumé Cet article est consacré à la classification des données de grande dimension. Supposant que de telles données vivent dans des sous-espaces de dimensions intrinsèques inférieures à la dimension de l'espace original, nous proposons une re-paramétrisation du modèle de mélange gaussien. En forçant certains paramètres à être communs dans une même classe ou entre les classes, nous exhibons une famille de modèles adaptés aux données de grande dimension, allant du modèle le plus général au plus parcimonieux. Ces modèles gaussiens sont ensuite utilisés pour la classification supervisée ou non-supervisée. La nature de notre re-paramétrisation permet aux méthodes ainsi construites de ne pas être perturbées par le mauvais conditionnement ou la singularité des matrices de covariance empiriques des classes et d'être efficaces en terme de temps de calcul.

Mots-clefs : Classification supervisée et non supervisée, fléau de la dimension, modèle de mélange gaussien, modèle parcimonieux.

1 Introduction

La classification de données situées dans un espace de grande dimension est un problème délicat qui apparaît dans de nombreuses sciences telles que l'analyse d'images. Dans cet article, nous focalisons notre attention sur les modèles probabilistes [12]. Parmi ceux-ci, le modèle de mélange gaussien est le plus populaire [33] bien que son comportement dans la pratique soit décevant lorsque la taille de l'échantillon est faible en regard du nombre de paramètres à estimer. Ce phénomène bien connu est appelé « fléau de la dimension » ou *curse of dimensionality* depuis les travaux de Bellman [3]. On pourra consulter [36, 37] pour une étude théorique de l'effet de la dimension en classification supervisée (ou discrimination).

Pour éviter le sur-ajustement des modèles, il est nécessaire de trouver un compromis entre le nombre de paramètres à estimer et la généricité du modèle. Nous proposons ici un modèle de mélange gaussien parcimonieux permettant de représenter le sous-espace propre à chacune des classes. Les paramètres de ce modèle sont estimés par maximum de vraisemblance ou par l'algorithme EM [20] selon que l'on soit confronté à un problème de classification supervisée ou non-supervisée. Les méthodes de classification ainsi construites sont baptisées respectivement HDDA pour *High Dimensional Discriminant Analysis* et HDDC pour *High Dimensional Data Clustering*. Notons qu'il est possible de contraindre le modèle afin de limiter davantage le nombre de paramètres à estimer. Ainsi, il sera