

Pourquoi les modèles de mélange pour la classification ?

Christophe Biernacki

CNRS & Université de Lille 1, Villeneuve d'Ascq, France
biernack@math.univ-lille1.fr

Résumé Les modèles de mélange apportent une réponse rigoureuse, flexible et interprétable pour les multiples besoins de la classification : classification supervisée ou non, nature des données, choix du nombre de groupes, *etc.* Les domaines d'applications sont de plus en plus nombreux, aidés en cela par le développement de solutions logicielles adaptées.

Mots clés : classification supervisée, classification non supervisée, algorithme EM, choix de modèles.

Abstract Mixture models provide a mathematical-based, flexible and meaningful approach for the wide variety of classification requirements: Unsupervised or supervised classification, data features, number of classes selection, *etc.* Fields in which mixture models have been successfully applied are numerous and specific softwares are now available.

Keywords: Classification, clustering, EM algorithm, model selection.

1 Introduction

La classification a pris aujourd'hui une place importante en analyse des données *exploratoire* et *décisionnelle*, tant au niveau des domaines d'applications que des développements méthodologiques. L'objectif exploratoire est typiquement représenté par la classification dite *non supervisée*, ou classification automatique, qui vise à découvrir une partition hypothétique dans un ensemble d'objets. Le but décisionnel quant à lui se réfère plutôt à la classification *supervisée*, ou analyse discriminante, qui cherche généralement à affecter tout nouvel objet à des groupes préalablement définis. Pour cela, il est donc nécessaire de disposer de méthodes s'adaptant à la diversité des données à analyser comme l'abondance, ou au contraire la rareté, des objets (*individus*) disponibles ainsi que le nombre et/ou le type de descripteurs (*variables*) pour chacun d'entre eux.

Grâce à leur flexibilité, les mélanges finis de distributions de probabilité répondent à ces exigences (voir [36] et les nombreuses références associées). Ils sont devenus aujourd'hui un outil populaire et utilisé avec succès dans un nombre croissant de disciplines comme l'astronomie, la biologie, la génétique, l'économie, les sciences de l'ingénieur, le marketing, la reconnaissance d'images... En outre, le fait que des logiciels comme MIXMOD¹ [5] mettent à disposition ces méthodes sous forme de code performant et portable

¹Site web du logiciel MIXMOD : <http://www-math.univ-fcomte.fr/mixmod/index.php>