Clustering hiérarchique non paramétrique de données fonctionnelles

Marc Boullé *, Romain Guigourès *,**, Fabrice Rossi **

*Orange Labs
2 avenue Pierre Marzin
22300 Lannion
{prenom.nom}@orange.com

**SAMM, Université Paris 1
90 rue Tolbiac
75013 Paris
{prenom.nom}@univ-paris1.fr

Résumé. Dans cet article, il est question de clustering de courbes. Nous proposons une méthode non paramétrique qui segmente les courbes en clusters et discrétise en intervalles les variables continues décrivant les points de la courbe. Le produit cartésien de ces partitions forme une grille de données qui est inférée en utilisant une approche Bayésienne de sélection de modèle ne faisant aucune hypothèse concernant les courbes. Enfin, une technique de post-traitement, visant à réduire le nombre de clusters dans le but d'améliorer l'interprétabilité des clusters, est proposée. Elle consiste à fusionner successivement et de façon optimale les clusters, ce qui revient à réaliser une classification hiérarchique ascendante dont la mesure de dissimilarité correspond à la variation du critère. De manière intéressante, cette mesure est en fait une somme pondérée de divergences de Kullback-Leibler entre les distributions des clusters avant et après fusions. L'intérêt de l'approche dans le cadre de l'analyse exploratoire de données fonctionnelles est illustré par un jeu de données artificiel et réel.

1 Introduction

En analyse de données fonctionnelles (Ramsay et Silverman (2005)), les observations sont des fonctions (ou des courbes). Les données fonctionnelles sont présentes dans de nombreux domaines comme par exemple l'enregistrement des précipitations d'une station météorologique ou encore la surveillance de matériel, où chaque courbe est une série temporelle liée à une quantité physique enregistrée à fréquence spécifiée.
Les Méthodes d'analyse exploratoire pour les grandes bases de données fonctionnelles sont nécessaires dans de nombreuses applications pratiques comme par exemple, la surveillance de la consommation électrique (Hébrail et al. (2010)). Elles réduisent la complexité des données en combinant des techniques de clustering avec des méthodes d'approximation de fonction, modélisant par exemple un ensemble de données fonctionnelles par des courbes prototypiques, comme par exemple un ensemble de segments linéaires ou de splines. Dans ce type d'approches, à la fois le nombre de prototypes et le nombre de segments sont des paramètres utilisateur. D'un côté, cela limite pour l'utilisateur le risque d'obtenir des clusters trop complexes mais cela peut également induire un sous-apprentissage du modèle par rapport aux données.
Des approches Bayésiennes non paramétriques basées sur des processus de Dirichlet ont