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.

Summary

In this paper, we deal with the problem of curves clustering. We propose a nonparametric method which partitions the curves into clusters and discretizes the dimensions of the curve points into intervals. The cross-product of these partitions forms a data-grid which is obtained using a Bayesian model selection approach while making no assumption regarding the curves. Finally, a post-processing technique, aiming at reducing the number of clusters in order to improve the interpretability of the clustering, is proposed. It consists in optimally merging the clusters step by step, which corresponds to an agglomerative hierarchical classification whose dissimilarity measure is the variation of the criterion. Interestingly this measure is none other than the sum of the Kullback-Leibler divergences between clusters distributions before and after the merges. The practical interest of the approach for functional data exploratory analysis is presented on an artificial and a real world dataset.