Clustering de séries temporelles par construction de dictionnaire

Étienne Goffinet*,** Mustapha Lebbah*
Hanane Azzag*, Loïc Giraldi**

*Laboratoire Informatique de Paris Nord
99 Avenue Jean Baptiste Clément
93430 Villetaneuse
www.lipn.univ-paris13.fr
**Renault SAS
1 Avenue du Golf, 78280 Guyancourt

Résumé. La classification non-supervisée est un domaine qui regroupe les méthodes d'analyses de données dont l'objectif est la recherche de groupes d'observations similaires dans un jeu de données. Lorsque les données considérées sont issues de l'observation d'un phénomène à différents instants, elles sont appelées des séries temporelles : par exemple l'évolution du cours du temps d'une action boursière, de données météorologiques... Dans certains cas, ces séries peuvent alterner différentes phases de fonctionnement distinctes, que l'on appelle des régimes : par exemple, l'observation de la vitesse d'une voiture qui peut montrer des phases d'accélération, une vitesse de croisière, des phases de freinage, etc... Nous présentons dans cet article une méthode dédiée à l'analyse de ce dernier type de séries temporelles et qui est basée sur la combinaison de trois étapes : la segmentation individuelle des séries temporelles, le recodage dans un dictionnaire de régimes communs et le clustering des séquences catégorielles ainsi produites. Notre contribution inclut également une stratégie innovante de sélection de modèle pour la segmentation. Nous présentons les différents avantages de cette méthode et les résultats obtenus sur des jeux de données publics.

1 Introduction

L'attention portée à l'analyse de séries temporelles a augmenté drastiquement ces dernières décennies en même temps que la capacité à les produire. Dans le cas industriel cela peut s'expliquer par la baisse des coûts des capteurs ajoutée au besoin de certifier les équipements avec toujours plus de rigueur. L'augmentation de la taille et de la complexité des données à analyser ont rendu nécessaire le développement d'outils pour fournir une aide précieuse à la décision des experts du métier.

La classification non-supervisée (ou clustering) de séries temporelles est un outil qui a pour but de partitionner un jeu de données en des groupes d'observations temporelles "similaires", ce qui constitue une première phase dans la compréhension de sa structure. Définir la similarité