

Clustering semi-supervisé de séries temporelles multivariées par apprentissage profond

Dino Ienco*, Roberto Interdonato**

*INRAE, UMR TETIS, Univ. Montpellier
dino.ienco@inrae.fr

**CIRAD, UMR TETIS, Univ. Montpellier
roberto.interdonato@cirad.fr

Résumé. De grands volumes de données sont aujourd’hui produits par différents capteurs qui mesurent, généralement, plusieurs variables au cours du temps. Ces informations peuvent être organisées sous forme de séries temporelles multivariées. Recueillir suffisamment d’échantillons étiquetés, pour mettre en place une analyse d’apprentissage automatique supervisée pour ce type de données, est quasiment impossible. Dans ce contexte, les méthodes de clustering semi-supervisé représentent un outil particulièrement adapté pour exploiter au mieux la quantité réduite de connaissances à disposition. Dans le but d’améliorer l’analyse de séries temporelles multivariées dans ce contexte, nous proposons une approche de clustering semi-supervisé (sous contrainte) de séries temporelles basée sur des méthodes d’apprentissage profond utilisant des contraintes de type must- et cannot-link comme forme de supervision faible. L’évaluation expérimentale de cette méthode sur différents jeux de données a mis en évidence la plus value de notre proposition.

1 Introduction

De nos jours, l’acquisition continue de données massives est devenue un élément essentiel des processus dans plusieurs domaines d’application, tels que l’agriculture, la biochimie et la santé humaine. Ces données sont acquises par l’utilisation de capteurs de natures différentes (p.ex., des capteurs à distance, biochimiques ou portables), capables de produire des flux de données intégrant le suivi de plusieurs caractéristiques dans le temps. Lorsqu’il est nécessaire d’analyser ces données de manière automatique, la façon habituelle de les modéliser est de les organiser en séries temporelles multivariées. Ces structures de données suscitent un intérêt croissant, et plusieurs méthodes ont été proposées ces dernières années pour des tâches telles que la classification (Zhang et al., 2020) et la prévision de valeurs (Xie et al., 2021) d’une série temporelle multivariée. Cependant, le clustering de séries temporelles multivariées reste un problème ouvert. Dans les cas réels, étiqueter les données est une tâche coûteuse en termes de temps et de ressources, quel que soit le domaine d’application. C’est pourquoi le développement d’approches avancées de clustering non supervisé et semi-supervisé est devenu une priorité (Lampert et al., 2018), d’autant plus que ces méthodes permettent de caractériser un