Cadre général et algorithmes de constructions pour des représentations symboliques adaptatives de séries temporelles

Bernard Hugueney[1]

Université PARIS-DAUPHINE
LAMSADE
Place du Maréchal de Lattre de Tassigny
75775 PARIS CEDEX 16
bernard.hugueney@lamsade.dauphine.fr,
http ://www.lamsade.dauphine.fr/~hugueney

Résumé Les séries temporelles constituent un domaine très important de la fouille de données. En effet, les très gros volumes de données numériques généralement entreposés ne se prêtent pas à une analyse directe. Dans un but à la fois de réduction de la dimensionnalité et d'extraction d'information, la fouille de données de séries temporelles donne généralement lieu à un changement de représentation des séries temporelles. Dans un objectif d'intelligibilité de l'information extraite lors du changement de représentation, on peut avoir recours à des représentations symboliques de séries temporelles. Nous proposons un cadre général de représentation de séries temporelles, ainsi que deux représentations particulières (Clustering-Based Symbolic Representations : CBSR et Segmentation-Based Symbolic Representation with Linear models of 0th order : SBSR-L0) s'inscrivant dans ce cadre général.

Keywords : fouille de données, séries temporelles, changements de représentations, représentations symboliques, recherche de motifs récurrents.

1 Introduction

Les séries temporelles constituent un domaine très actif de la fouille de données. En effet, les bases de données de séries temporelles sont caractérisées non seulement par leur très grand volume, mais aussi par le fait que les informations recherchées (tendances, corrélations,...) ne sont pas directement accessibles à partir des données brutes. Pour cette raison, des changements de représentation doivent être effectués. Nous nous intéressons plus particulièrement à des représentations symboliques, plutôt que numériques, car elles sont intelligibles par les utilisateurs. Nous présentons tout d'abord un cadre général permettant de formuler une très large gamme de représentations symboliques, notamment SAX (Symbolic Aggregate approXimation) qui est une représentation symbolique de séries temporelles classiquement utilisée. Nous proposons deux nouvelles représentations symboliques qui peuvent être considérées comme des extensions adaptatives de SAX : CBSR (Clustering-Based Symbolic Representations) et SBSR-L0 (Segmentation-Based Symbolic Representation with Linear models of 0th order). Pour chacune de ces représentations symboliques, nous proposons un algorithme de construction. En conclusion, nous suggérons