K-Spectral Centroïd pour des données massives

Brieuc Conan-Guez, Alain Gély, Lydia Boudjeloud-Assala, Alexandre Blansché

Université de Lorraine - Site de Metz - île du Saulcy 57045 METZ CEDEX 1
brieuc.conan-guez@univ-lorraine.fr, alain.gely@univ-lorraine.fr,
lydia.boudjeloud-assala@univ-lorraine.fr, alexandre.blansche@univ-lorraine.fr
http://www.lita.univ-lorraine.fr/

Résumé. Nous nous intéressons à la classification non supervisée de séries chronologiques. Pour ce faire, nous utilisons l'algorithme K-Spectral Centroïd (K-SC), une variante des K-Means. K-Spectral Centroïd utilise une mesure de dissimilarité entre séries chronologiques, invariante par translation et par changement d'échelle. Cet algorithme est coûteux en temps de calcul : lors de la phase d'affectation, il nécessite de tester toutes les translations possibles pour identifier la meilleure ; lors de la phase de représentation, le calcul du nouveau barycentre nécessite l'extraction de la plus petite valeur propre d'une matrice. Nous proposons dans ce travail trois optimisations de K-SC. L'identification de la meilleure translation peut être réalisée efficacement en utilisant la transformée de Fourier discrète. Chaque matrice peut être calculée incrémentalement. Le calcul du nouveau barycentre peut s'effectuer à moindre coût grâce à la méthode de la puissance itérée. Ces trois optimisations fournissent exactement la même classification que K-SC.

1 Introduction

Dans ce travail, nous nous intéressons à la classification non supervisée de séries chronologiques issues de l'analyse de media-sociaux, pour faire émerger différents types de partages de l'information. Ce travail s'inscrit dans le cadre de l'ANR INFO-RSN dont l'objet principal est l'étude de la diffusion de l'information sur les réseaux socionumériques (Twitter).

Pendant 6 mois, les tweets citant l'URL d'un article de presse issu d'une liste prédéfinie de 32 médias ont été collectés. Plusieurs types de séries temporelles peuvent être définis. On peut par exemple associer une série temporelle à chacune des URL collectées, ou encore à chacun des hashtags apparaissant dans la collecte. La série temporelle est alors l'ensemble des informations d'horodatage des tweets citant un article, ou comportant un hashtag.

L'étude de séries chronologiques présente deux problématiques. L'une est sémantique et concerne le choix de la dissimilarité pour comparer les séries. L'autre, plus technique, concerne le temps de calcul pour la classification de ces séries.

D'un point de vue sémantique, la dissimilarité choisie doit permettre de mettre en évidence des comportements similaires, bien qu'à des échelles différentes. En effet, le volume de tweets engendré par un quotidien national est bien évidemment très différent de celui d'un journal de presse locale. D'autre part il faut permettre une certaine lattitude de façon à distinguer les