Construction de descripteurs à partir du coclustering pour la classification supervisée de séries temporelles

Dominique Gay*, Marc Boullé*

*Orange Labs, Lannion, FRANCE
prenom.nom@orange.com

Résumé. Nous présentons un processus de construction de descripteurs pour la classification supervisée de séries temporelles. Ce processus est libre de tout paramétrage utilisateur et se décompose en trois étapes : *(i)* à partir des données originales, nous générons de multiples nouvelles représentations simples ; *(ii)* sur chacune de ces représentations, nous appliquons un algorithme de co-clustering ; *(iii)* à partir des résultats de co-clustering, nous construisons de nouveaux descripteurs pour les séries temporelles. Nous obtenons une nouvelle base de données objets-attributs dont les objets (identifiant les séries temporelles) sont décrits par des attributs issus des diverses représentations générées. Nous utilisons un classifieur Bayésien sur cette nouvelle base de données. Nous montrons expérimentalement que ce processus offre de très bonnes performances prédictives comparées à l'état de l'art.

1 Introduction

La classification de séries temporelles (TSC) est un sujet qui a été intensivement étudié durant les dernières années. Le but est de prédire la classe d'un objet (une série temporelle ou courbe) $\tau_i = \langle (t_1, x_1), (t_2, x_2), \ldots, (t_{m_i}, x_{m_i}) \rangle$ (où $x_k, (k = 1..m_i)$ est la valeur de la courbe au temps t_k), étant donné un ensemble de séries temporelles labellisées d'apprentissage. Les problèmes de TSC sont différents des problèmes de classification supervisée dans les bases transactionnelles puisqu'il y a une dépendance temporelle entre les attributs ; ainsi l'ordre des attributs importe. La TSC est applicable dans de nombreux domaines dont les données sont des séries temporelles : e.g., pour le diagnostic médical (par exemple la classification d'électrocardiogramme de patients) mais aussi dans d'autres domaines comme la maintenance de machines industrielles, la finance, la météo, ... Le grand nombre d'applications a succité de nombreuses approches ; toutefois la majorité de la communauté s'est attachée à suivre le processus suivant (Liao, 2005) : *(i)* choisir une nouvelle représentation des données, *(ii)* choisir une mesure de similarité (ou une distance) pour comparer deux séries temporelles et enfin *(iii)* utiliser l'algorithme (NN) du plus proche voisin (avec la mesure choisie sur la représentation choisie) comme classifieur. Ding et al. (2008) propose un état de l'art des différentes représentations et mesures ainsi qu'une étude expérimentale comparative basée sur le classifieur NN. Il en ressort que le classifieur NN couplé avec une distance Euclidienne ou Dynamic Time Warping (DTW) présente les meilleures performances prédictives pour les problèmes de TSC.