Découverte de sous-groupes à partir de données séquentielles par échantillonnage et optimisation locale

Romain Mathonat*,**, Jean-François Boulicaut*
Mehdi Kaytoue*,***

*Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
**Atos, 34 Rue de la Soie, 69100 Villeurbanne
***Infologic, 99 avenue de Lyon, 26500 Bourg-Lès-Valence, France
prenom.nom@insa-lyon.fr

Résumé. La découverte de règles caractéristiques d'une classe reste un problème difficile, particulièrement dans le cadre des données séquentielles (séquences d'ensembles). La découverte de sous-groupes est une bonne formalisation de cette tâche et de nombreux algorithmes dédiés ont été proposés ces 20 dernières années. Une exploration dite exhaustive est souvent inapplicable au vu de la taille de l'espace de recherche, et les méthodes heuristiques de référence, principalement les recherches en faisceau, posent des problèmes de paramétrage. Nous proposons une méthode d'échantillonnage depuis l'espace des motifs pour la découverte de sous-groupes dans des données séquentielles étiquetées. Celle-ci permet, entre autres, de trouver des optima locaux, ne nécessite pas de paramétrage, est indépendante de la mesure de qualité utilisée, et est simple à mettre en oeuvre. La validation empirique sur divers jeux de données nous permet de valider les qualités de cette approche.

1 Introduction

Les données séquentielles sont présentes dans de nombreux contextes applicatifs (analyses de textes ou de vidéos, exploitation de traces d'interactions, supervision de processus industriels, exploration de données en biologie moléculaire, etc). Le cas d'utilisation qui motive nos propres travaux est celui de la supervision industrielle, où les suites d'états du système étudié constituent une séquence. Nous voulons fouiller ces collections de séquences étiquetées par des experts métiers pour découvrir des règles sur les co-occurences de pannes. Ceci permettrait, d'une part, de mieux comprendre le système étudié, d'autre part, de pouvoir construire un moteur de règles explicables, offrant alors des perspectives de prédiction et d'anticipations de certains dysfonctionnements. Une formalisation simple de ce contexte est de considérer qu'il s'agit de découvrir des motifs séquentiels co-occurrents à une variable cible (étiquette ou classe). La découverte de règles caractérisant une classe ou une étiquette a été très étudiée (Novak et al. (2009)), notamment dans le cadre de la découverte de sous-groupes (Wrobel (1997)). La découverte de sous-groupes dans des données étiquetées consiste à trouver des motifs (e.g., des motifs séquentiels), également appelés descriptions, qui définissent en intention des objets