

Chapitre 5 : Analyse Implicative Séquentielle

Julien Blanchard, Fabrice Guillet, Régis Gras

Equipe Connaissances & Décision (COD)

LINA – FRE CNRS 2729 – Ecole Polytechnique de l'Université de Nantes

julien.blanchard@polytech.univ-nantes.fr

Résumé. La découverte de motifs fréquents dans des séquences (généralement des séquences temporelles d'évènements) est l'une des tâches majeures de la fouille de données. Dans cet article, nous nous intéressons à l'évaluation de la qualité des règles séquentielles. Nous proposons une mesure inédite nommée *SI* qui évalue la significativité des règles au regard d'un modèle probabiliste. Les simulations numériques montrent que *SI* a des caractéristiques uniques en comparaison aux autres mesures de qualité de règles séquentielles.

1 Introduction

La découverte de motifs fréquents dans des séquences symboliques (généralement des séquences temporelles d'évènements) est l'une des tâches majeures de la fouille de données. Les travaux de recherche dans ce domaine se divisent en deux catégories :

- la découverte d'*épisodes* fréquents dans une longue séquence d'évènements (approche initiée par Mannila, Toivonen, et Verkamo Mannila et al. (1995) Mannila et Toivonen (1996)),
- la découverte de *motifs séquentiels* fréquents dans un ensemble de séquences d'évènements (approche initiée par Agrawal et Srikant Agrawal et Srikant (1995) Srikant et Agrawal (1996)).

Episodes et *motifs séquentiels* sont des structures séquentielles, c'est-à-dire définies avec un ordre (partiel ou total). Une telle structure peut être par exemple :

petit-déjeuner then déjeuner then dîner

La structure est qualifiée par sa fréquence (ou support) et généralement par des contraintes sur les positions des évènements, comme une fenêtre maximale de temps "moins de 12 heures séparent *petit-déjeuner* et *dîner*" Srikant et Agrawal (1996) Mannila et al. (1997) Das et al. (1998) Höppner (2002) Sun et al. (2003).

La différence entre *épisodes* et *motifs séquentiels* réside principalement dans la mesure de leur support : la fréquence des *épisodes* est intra-séquence Mannila et al. (1997) Das et al. (1998) Weiss (2002) Höppner (2002) Sun et al. (2003) Yang et al. (2003), alors que la fréquence des *motifs séquentiels* est inter-séquences Agrawal et Srikant (1995) Srikant et Agrawal (1996) Spiliopoulou (1999) Zaki (2001) Han et al. (2005) (voir Joshi et al. (1999) pour une synthèse sur les différentes manières d'évaluer la fréquence). Ainsi, les algorithmes d'extraction d'*épisode* fréquents recherchent des structures qui se répètent souvent à l'intérieur d'une