Extraction de règles d'association séquentielle à l'aide de modèles semi-paramétriques à risques proportionnels

Nicolas S. Müller*, Matthias Studer*, Gilbert Ritschard*, Alexis Gabadinho*

*Institut d'études démographiques et des parcours de vie, Université de Genève
{nicolas.muller, matthias.studer, gilbert.ritschard, alexis.gabadinho}@unige.ch

Résumé. La recherche de liens entre objets fréquents a été popularisée par les méthodes d'extraction de règles d'association. Dans le cas de séquences d'événements, les méthodes de fouille permettent d'extraire des sous-séquences qui peuvent ensuite être exprimées sous la forme de règles d'association séquentielle entre événements. Cette utilisation de la fouille de séquences pour la recherche de liens entre des événements pose deux problèmes. Premièrement, le critère principal utilisé pour sélectionner les sous-séquences d'événements est la fréquence, or les occurrences de certains événements peuvent être fortement liées entre elles même lorsqu'elles sont peu fréquentes. Deuxièmement, les mesures actuelles utilisées pour caractériser les règles d'association ne tiennent pas compte du caractère temporel des données, comme l'importance du *timing* des événements ou le problème des données censurées. Dans cet article, nous proposons une méthode pour rechercher des liens significatifs entre des événements à l'aide de modèles de durée. Les règles d'association sont construites à partir des motifs séquentiels observés dans un ensemble de séquences. L'influence sur le risque que l'événement « conclusion » se produise après le ou les événements « prémisse » est estimée à l'aide d'un modèle semi-paramétrique à risques proportionnels. Outre la présentation de la méthode, l'article propose une comparaison avec d'autres mesures d'association [1].

1 Introduction

La recherche de motifs fréquents et de règles d'association entre des objets a fait l'objet de nombreux travaux en data mining. Son extension à la recherche de motifs séquentiels fréquents a été également un domaine en plein essor ces dernières années. En revanche, la caractérisation des règles d'association séquentielle restent un problème moins exploré. Il existe néanmoins des critères, comme la confiance ou le rappel, qui ont été reformulés dans le cadre de la fouille de règles d'association à l'intérieur d'une séquence unique (Mannila et al., 1997). Toujours dans le cadre d'une séquence unique, Blanchard et al. (2008) ont développé un indice d'intensité d'implication séquentielle inspiré de l'indice d'implication (Gras et al., 2004).

Nous proposons dans cet article une nouvelle méthode qui extrait des règles d'association entre événements à partir de séquences multiples. Nous utilisons pour cela des modèles

[1] Etude soutenue par le Fonds national suisse de la recherche (FNS) FN-100015-122230