Echantillonnage de motifs séquentiels sous contrainte sur la norme

Lamine Diop[***], Cheikh Talibouya Diop[**], Arnaud Giacometti[*], Dominique Li[*], Arnaud Soulet[*]

[*]Université de Tours, France
{arnaud.giacometti, dominique.li, arnaud.soulet}@univ-tours.fr
[**]Université Gaston Berger de Saint-Louis, Sénégal
{diop.lamine3,cheikh-talibouya.diop}@ugb.edu.sn

Résumé. L'échantillonnage de motifs est une méthode non-exhaustive pour découvrir des motifs pertinents qui assure une bonne interactivité tout en offrant des garanties statistiques fortes grâce à sa nature aléatoire. Curieusement, une telle approche explorée pour les motifs ensemblistes et les sous-graphes ne l'a pas encore été pour les données séquentielles. Dans cet article, nous proposons la première méthode d'échantillonnage de motifs séquentiels. Outre le passage aux séquences, l'originalité de notre approche est d'introduire une contrainte sur la norme pour maîtriser la longueur des motifs tirés et éviter l'écueil de la « longue traîne ». Nous démontrons que notre méthode fondée sur une procédure aléatoire en deux étapes effectue un tirage exact. Malgré le recours à un échantillonnage avec rejet, les expérimentations montrent qu'elle reste performante.

1 Introduction

Les motifs séquentiels ont été introduits par Agrawal et Srikant (1995) il y a plus de 20 ans et leur utilité a été prouvée dans différents domaines de recherche et d'applications comme la fouille d'usage du Web, la fouille de textes, la bioinformatique, la détection de fraudes, etc. Depuis la première publication, de nombreuses méthodes ont optimisé l'extraction des motifs séquentiels (Zaki, 2001; Pei et al., 2001) et ont introduit des variantes (Lo et al., 2008; Gomariz et al., 2013). Malgré toutes ces avancées, l'extraction des motifs séquentiels reste une tâche coûteuse qui génère souvent trop de motifs. Cette limite aussi atteinte par l'extraction des motifs ensemblistes a été contournée par l'échantillonnage de motifs. Une telle approche tire un nombre limité de motifs où la probabilité de tirer un motif est proportionnelle à sa fréquence. Cette approche a l'avantage de contrôler la taille de la sortie et d'apporter une collection de motifs qui reflète l'intégralité de l'espace de recherche. A notre connaissance, une telle approche n'a encore pas été envisagée pour les motifs séquentiels.

Adapter la procédure d'échantillonnage de motifs en deux étapes (Boley et al., 2011) aux données séquentielles n'est pas trivial. D'une part, une limite importante de l'échantillonnage de motifs est d'avoir tendance à retourner des motifs rares correspondant à la longue traîne. En effet, la longue traîne signifie que la très grande majorité des motifs ont une fréquence très faible et elle occulte les motifs les plus fréquents. Ce problème est exacerbé dans le cas