

Echantillonnage de motifs avec une contrainte de fréquence

Arnaud Soulet

Université de Tours, LIFAT, Blois
firstname.lastname@univ-tours.fr

Résumé. L'échantillonnage de motifs est une technique récente de découverte de motifs favorisant l'interactivité avec l'utilisateur. Son principe est de tirer aléatoirement un motif proportionnellement à son intérêt. Malheureusement, les tirages peuvent se focaliser sur une partie de l'espace de recherche avec des motifs peu fréquents mais extrêmement nombreux. Il serait bien possible d'échantillonner des motifs et d'éliminer ceux non-fréquents, mais le taux de rejet s'avère souvent trop élevé. Dans cet article, nous proposons la première méthode efficace d'échantillonnage de motifs avec une contrainte de fréquence minimale. Elle s'appuie sur la suppression des items non-fréquents (opération de réduction) et sur la projection de la base de données sur chacun des items (opération de projection). En combinant ces deux opérations, nous proposons une méthode générique qui revient à éliminer tous les motifs contenant un couple non-fréquent d'items. Nos expérimentations montrent que notre méthode réduit considérablement le taux de rejet.

1 Introduction

Ces dernières années, une large partie des méthodes de découverte de motifs s'est orientée vers des approches favorisant l'interaction avec l'utilisateur afin d'intégrer ses retours dans le processus de découverte (Leeuwen, 2014). Pour cela, il est nécessaire de disposer de techniques d'extraction non-exhaustives afin d'être assez rapide pour extraire des motifs à la volée tout en s'appuyant sur des mesures d'intérêt complexes pouvant être mises à jour à chaque itération. Typiquement, ces méthodes d'extraction s'appuient sur des algorithmes de recherche en faisceau (Leeuwen, 2014), de recherche arborescente Monte-Carlo (Bosc et al., 2018) ou encore, d'échantillonnage de motifs (Boley et al., 2011). En particulier, les méthodes fondées sur l'échantillonnage de motifs ont reçu beaucoup d'attention avec la proposition de systèmes interactifs (Giacometti et Soulet, 2017; Dzyuba et al., 2017) ou d'algorithmes anytime pour extraire des données aberrantes (Giacometti et Soulet, 2016).

Plus précisément, l'échantillonnage de motifs consiste à tirer un motif X avec une probabilité proportionnelle à son intérêt $m(X)$. Par exemple avec la fréquence, un motif X deux fois plus fréquent qu'un motif Y aura deux fois plus de chance d'être tiré. Malheureusement, l'échantillonnage de motifs a souvent tendance à se focaliser sur des parties de l'espace de recherche avec une densité forte de motifs peu intéressants (i.e., beaucoup de motifs avec des valeurs faibles pour m). Pour minimiser ce phénomène, appelé « malédiction de la longue traîne », Diop et al. (2018) ont proposé d'ajouter une contrainte de longueur maximale sur les