

Échantillonnage de motifs d’intervalles selon la fréquence

Djawad Bekkoucha*, Abdelkader Ouali*, Bruno Crémilleux*, Patrice Boizumault*

Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ
GREYC UMR6072, F-14000 Caen, France
*prénom.nom@unicaen.fr

Résumé. L’échantillonnage de motifs est une approche récente pour mener à bien un processus de découverte de motifs en limitant le nombre de motifs produits et en favorisant l’interactivité avec l’analyste. Le principe est de tirer aléatoirement un motif proportionnellement à un intérêt donné, par exemple la mesure de fréquence. Dans cet article, nous proposons FIPS, la première méthode d’échantillonnage de motifs d’intervalles à partir de données numériques. FIPS s’appuie sur le principe des approches en plusieurs étapes (Boley et al.). Avec des données numériques, la difficulté majeure est de déterminer le nombre exact de motifs d’intervalles couvrant un objet. Nous montrons formellement que FIPS tire des motifs d’intervalles proportionnellement à la fréquence. Les expérimentations, répondant à plusieurs questions de recherche, montrent la qualité des motifs tirés par FIPS pour la découverte d’information.

1 Introduction

Les spécialistes en science des données sont des acteurs majeurs de la transformation de données pour la mise en évidence d’informations, ou mieux, de connaissances. En pratique, les analystes veulent interagir (visualiser, sélectionner, explorer) non seulement avec les données mais aussi avec les motifs ou modèles supportés par les données. Pour mener à bien de tels processus, il est nécessaire de disposer de méthodes d’extraction de motifs quasi instantanées afin que l’analyste n’attende pas, sinon il décroche du système. L’échantillonnage de motifs est une réponse à ce défi (Dzyuba et al., 2017).

L’échantillonnage de motifs consiste à tirer aléatoirement un motif proportionnellement à une mesure d’intérêt comme par exemple la fréquence. Ainsi un motif \mathcal{V}_1 deux fois plus fréquent qu’un motif \mathcal{V}_2 aura deux fois plus de chance d’être tiré. Une approche naïve pour effectuer l’échantillonnage de motifs est de générer l’ensemble des motifs puis d’effectuer un tirage pondéré selon la mesure d’intérêt. Une telle méthode échoue en pratique compte-tenu de la taille gigantesque des espaces de recherche même pour des langages de motifs aussi simples que les motifs ensemblistes. Une réponse à ce problème a été apportée par Boley et al. qui ont défini une approche fondée sur deux tirages successifs. Cette stratégie, à condition d’effectuer une décomposition appropriée au problème étudié, assure d’effectuer un tirage exact selon la distribution résultante de la mesure d’intérêt. Les exemples présentés dans Boley et al. concernent des motifs ensemblistes, c’est-à-dire des données décrites par des valeurs booléennes.