Prédiction d'événements distants basée sur des règles séquentielles

Lina Fahed*, Philippe Lenca*, Yannis Haralambous*, Riwal Lefort**, Marie-Laure Tallec**

*IMT-Atlantique, Lab-STICC, F-29238 Brest, France
{lina.fahed, philippe.lenca, yannis.haralambous}@imt-atlantique.fr
**Crédit Mutuel ARKEA, Pôle Innovation et Opération, service Datalabs
{riwal.lefort, marie-laure.tallec}@arkea.com

Résumé. Dans cet article, nous nous concentrons sur la prédiction d'évènements distants à travers la fouille de règles séquentielles en proposant l'algorithme *D-SR* (Distant Sequential Rules). L'originalité de *D-SR* réside dans le fait qu'il fouille les règles avec un conséquent temporellement distant de l'antécédent, en appliquant une contrainte de gap minimal entre les deux éléments. Nous proposons d'intégrer *D-SR* dans les algorithmes traditionnels de fouille de règles en tant qu'étape de post-traitement ou pendant le processus de fouille. Les expérimentations montrent l'efficacité de *D-SR* en termes de scalabilité et de performance en prédiction.

1 Introduction

La fouille de données est un domaine qui a pour but la recherche de motifs, de tendances ou de relations cachées dans les données. Depuis son introduction en 1993 par Agrawal (Agrawal et al., 1993), la fouille de motifs séquentiels dans une base de séquences reste toujours aujourd'hui un domaine actif (Boudane et al., 2017). Un motif séquentiel, noté $P = <p_1, \ldots, p_k>$, est une liste ordonnée d'événements. Une occurrence du motif dans une séquence est la série d'instants d'apparition des événements qui le composent dans la limité d'une fenêtre de taille w de la séquence. Le support d'un motif, $supp(P)$, est le nombre de ses occurrences. Un motif est considéré comme fréquent si $\mathrm{supp}(P) \geq \mathrm{minsupp}$ où $\mathrm{minsupp}$ est un seuil prédéfini.

Mis à part la fouille de motifs séquentiels, il est également possible de fouiller des règles d'association séquentielles, appelées « règles séquentielles » et notées $R : P \rightarrow Q$ où $P = <p_1, \ldots, p_k>$ et $Q = <q_1, \ldots, q_e>$. Une règle séquentielle indique que si un ou plusieurs évènements arrivent dans un ordre donné (l'antécédent de la règle), alors un ou plusieurs évènements sont susceptibles d'arriver (le conséquent) avec une certaine probabilité, appelée la confiance de la règle : $\mathrm{conf}(P \rightarrow Q) = \frac{\mathrm{supp}(P \cdot Q)}{\mathrm{supp}(P)}$. Une règle est confiante si $\mathrm{conf}(R) \geq \mathrm{minconf}$, où $\mathrm{minconf}$ est un seuil prédéfini. Les règles séquentielles sont souvent utilisées pour prédire des évènements futurs (le conséquent des règles) (Mannila et al., 1997).

La tâche de fouille de règles séquentielles est souvent décomposée en deux sous-tâches (Agrawal et al., 1993) : la fouille de motifs séquentiels fréquents et la génération de règles confiantes à partir de motifs (Agrawal et al., 1993). Dans la fouille de motifs séquentiels, et