Bordures statistiques pour la fouille incrémentale de données dans les Data Streams

Jean-Emile Symphor*, Pierre-Alain Laur*

*GRIMAAG-Dépt Scientifique Interfacultaire,
Université des Antilles et de la Guyane, Campus de Schoelcher,
B.P. 7209, 97275 Schoelcher Cedex, Martinique, France
{je.symphor,palaur}@martinique.univ-ag.fr.

Résumé. Récemment la communauté Extraction de Connaissances s'est intéressée à de nouveaux modèles où les données arrivent séquentiellement sous la forme d'un flot rapide et continu, $i.e.$ les data streams. L'une des particularités importantes de ces flots est que seule une quantité d'information partielle est disponible au cours du temps. Ainsi après différentes mises à jour successives, il devient indispensable de considérer l'incertitude inhérente à l'information retenue. Dans cet article, nous introduisons une nouvelle approche statistique en biaisant les valeurs supports pour les motifs fréquents. Cette dernière a l'avantage de maximiser l'un des deux paramètres (précision ou rappel) déterminés par l'utilisateur tout en limitant la dégradation sur le paramètre non choisi. Pour cela, nous définissons les notions de bordures statistiques. Celles-ci constituent les ensembles de motifs candidats qui s'avèrent très pertinents à utiliser dans le cas de la mise à jour incrémentale des streams. Les différentes expérimentations effectuées dans le cadre de recherche de motifs séquentiels ont montré l'intérêt de l'approche et le potentiel des techniques utilisées.

1 Introduction

Ces dix dernières années un grand nombre de travaux ont été proposés pour rechercher des motifs fréquents dans de grandes bases de données. En fonction des domaines d'applications les motifs extraits sont soit des itemsets (Srikant, 1995; Zaki, 2001; Pei et al., 2001; Ayres et al., 2002) soit des séquences (Agrawal et al., 1993; Han et al., 2000). Récemment les travaux issus de la communauté des chercheurs en base de données et en fouille de données considèrent le cas des data streams où l'acquisition des données s'effectue de façon régulière, continue ou incrémentalement et cela sur une durée longue voire éventuellement illimitée.

Compte tenu de la grande quantité d'information mise en jeu dans le cas des data streams, le problème de l'extraction de motifs fréquents est toujours d'actualité ((Li et al., 2004; Jin et al., 2003; Demaine et al., 2002; Manku et Motwani, 2002; Golab et Ozsu, 2003; Karp et al., 2003)). Dans ce contexte, un motif est dit *θ-fréquent* s'il est observé au moins une fraction θ, appelée support du motif, sur tout le stream. Le paramètre $theta$, tel que $0 < \theta < 1$, est fixé par l'utilisateur.