Extraction d'itemsets distinctifs dans les flux de données

Chongsheng Zhang[1] et Florent Masseglia

INRIA Sophia Antipolis-Méditerranée
2004 Route des lucioles - BP 93
06902 Sophia Antipolis Cedex
{Prenom.Nom@sophia.inria.fr},

Résumé. L'extraction d'itemsets distinctifs est un sujet de recherche récent qui connait plusieurs algorithmes pour les données statiques (Knobbe et Ho, 2006; Heikinheimo et al., 2007). Ces solutions ne sont toutefois pas conçues pour le cas des flux de données, pour lesquels les temps de réponse doivent être aussi faibles que possible. Nous considérons le problème de l'extraction d'itemsets distinctifs dans les flux, qui peut avoir de nombreuses applications dans la sélection de variables, la classification ou encore la recherche d'information. Nous proposons l'heuristique *IDkF* (Itemsets Distinctifs dans les Flux) et des résultats d'expérimentations en comparaison d'une technique de la littérature.

1 Introduction

Les flux de données sont des sources produisant de grandes quantités de données à une grande vitesse. Ces caractéristiques nous obligent non seulement à travailler dans un espace mémoire limité par rapport à ces données mais aussi à envisager des méthodes de traitement qui optimisent un compromis indispensable entre la vitesse d'exécution et la précision des résultats. Un des défis les plus importants se situe certainement dans l'aspect évolutif des données mais aussi de leur distribution et des concepts qu'elles véhiculent. Ce dernier point incite à l'élaboration de techniques de traitement efficaces et adaptatives.

La fouille de flux permet d'extraire des connaissances dans les flux en garantissant des temps de réponse appropriés et une marge d'erreur la plus faible possible. Parmi ces connaissances, citons par exemple les itemsets fréquents (Teng et al., 2003; Giannella et al., 2003) (qui seront présentés plus en détail en section 3). Un des problèmes de l'extraction d'itemsets (fréquents ou distinctifs) vient du grand nombre de motifs extraits. Dans Knobbe et Ho (2006), les auteurs proposent l'extraction de motifs appelés Miki (Maximally Informative k-Itemsets). Ces itemsets sont différents des itemsets fréquents dans la mesure où ils minimisent la redondance tout en exprimant une information maximale sur les données. Cette information est mesurée en terme d'entropie.

Exemple 1 *Considérons une application de recherche de documents à partir des informations de la figure 1. Dans cette table, la valeur «1» signifie que la variable est présente dans le*

[1]Ces travaux sont issus du projet MIDAS ANR07-MDCO-008, financé par l ANR.