Traitement de Flux par un Graphe de Voisinage Incrémental

Ibrahim Louhi*,** Lydia Boudjeloud-Assala*
Thomas Tamisier**

*Université de Lorraine,
Laboratoire d'Informatique Théorique et Appliquée, LITA-EA 3097,
Metz, F-57045, France
{ibrahim.louhi, lydia.boudjeloud-assala}@univ-lorraine.fr
**Luxembourg Institute of Science and Technology,
41, rue du Brill, L-4422 Belvaux, Luxembourg
{ibrahim.louhi, thomas.tamisier}@list.lu

Résumé. Cet article s'intéresse au traitement et de la visualisation des flux de données en temps réel. Pour traiter les flux, nous proposons une nouvelle approche utilisant un clustering basé sur le voisinage. Au lieu de traiter les nouveaux éléments un par un, nous choisissons de traiter chaque groupe de nouveaux éléments simultanément. Un clustering est appliqué sur les éléments de chaque nouveau groupe en utilisant des graphes de voisinage. Les clusters obtenus sont ensuite utilisés dans la construction incrémentale d'un graphe représentant le flux de données. Le graphe du flux est visualisé en temps réel à l'aide de visualisations spécifiques reflétant le processus du traitement. En vue de valider l'approche, nous l'appliquons sur plusieurs jeux de données et la comparons avec divers algorithmes de clustering de flux.

1 Introduction

Ces dernières années ont connu de très grandes avancées technologiques, ce qui a contribué à l'augmentation du volume des données disponibles. Afin d'exploiter ces données brutes il est nécessaire d'en d'extraire des connaissances facilement compréhensibles. La fouille de données est l'étape la plus importante dans le processus d'extraction de connaissances. Plusieurs méthodes de fouille de données ont été proposées depuis que les chercheurs se sont intéressés à ce domaine. Le clustering est une tâche de fouille de données dont le principe est de diviser les données en sous-ensembles appelés *clusters*. Cette méthode considère que les éléments qui se ressemblent (selon certains critères) doivent être regroupés dans le même cluster (Berkhin, 2006). L'ensemble des clusters est une description de l'ensemble des données. Le clustering est généralement utilisé quand aucune information sur les classes n'est disponible et donc qu'aucune technique d'apprentissage ne peut être utilisée.

Dans plusieurs domaines les données sont produites d'une façon continue et souvent à une très grande vitesse. A titre d'exemple, Le suivi de la consommation énergétique, des opérations financières la géolocalisation des smartphones ou les capteurs météorologiques produisent une