Approche de Clustering de Flux basée sur les Graphes de Voisinage

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

[*]Université de Lorraine, Laboratoire d'Informatique Théorique et Appliquée.
{ibrahim.louhi, lydia.boudjeloud-assala}@univ-lorraine.fr
[**]Luxembourg Institute of Science and Technology.
{ibrahim.louhi, thomas.tamisier}@list.lu

Dans plusieurs domaines les données sont générées d'une façon continue et souvent à une fréquence très rapide. Ce type de données est connu sous le nom de flux de données. Les flux de données sont caractérisés principalement par l'aspect temporel et par leur grande taille, ce qui rend le processus de clustering des éléments du flux une tâche laborieuse.

Traiter les éléments d'une manière séparée, au fur et à mesure de leur apparition, conduit souvent à des erreurs dans leur affectation aux nouveaux clusters. La principale idée de notre approche consiste à traiter un groupe de nouveaux éléments arrivant presque simultanément au lieu de traiter chaque élément séparément. Cela permet de prendre en compte les caractéristiques d'un groupe de données arrivant dans la même période temporelle. Nous supposons que deux éléments générés successivement sont probablement causés par les mêmes facteurs, ce qui implique qu'il y a de fortes chances qu'ils se ressemblent. Le but de notre approche est de construire incrémentalement un graphe de voisinage permettant de traiter et de visualiser le flux de données.

En premier lieu, nous attendons l'arrivée du premier groupe d'éléments (les groupes ont une taille fixe définie par l'utilisateur). Nous appliquons un clustering basé sur le voisinage sur les éléments du premier groupe : nous calculons la distance entre chaque couple d'éléments et nous considérons que deux éléments sont voisins si leur distance est inférieure à un seuil (qui est fixé également par l'utilisateur). Nous considérons que chaque ensemble de voisins constitue un cluster. Nous déterminons ensuite le centroid de chaque cluster (l'élément le plus proche du reste des éléments du cluster). Les clusters obtenus sont représentés dans un graphe de voisinage : pour chaque cluster, chaque élément est représenté par un noeud, les arêtes représentent la distance entre chaque élément et le centroid de son cluster.

Les éléments du groupe suivant sont traités, indépendamment dans un premier temps, avec le même processus que les éléments du premier groupe. De la même manière nous obtenons de nouveaux clusters et nous identifions également leurs centroids. Les nouveaux clusters sont utilisés pour mettre à jour le graphe de voisinage : nous calculons la distance entre chaque centroid des nouveaux clusters et les centroids des anciens clusters, si la distance entre deux centroids de clusters est inférieure au seuil, les deux clusters sont reliés. Cela se traduit par la création d'une arête entre les noeuds représentant les deux centroids. Dans le cas ou un nouveau cluster n'est similaire à aucun des anciens clusters, il est rajouté au graphe sans qu'il ne soit relié avec un autre cluster (ce qui représente l'apparition d'un nouveau cluster dans le