Clustering topologique pour le flux de données

Mohammed Ghesmoune*, Mustapha Lebbah*, Hanene Azzag*

*Université Paris 13, Sorbonne Paris Cité
LIPN-UMR 7030 - CNRS
99, av. J-B Clément – F-93430 Villetaneuse, France
prénom.nom@lipn.univ-paris13.fr

Résumé. Actuellement, le clustering de flux de données devient le moyen le plus efficace pour partitionner un très grand ensemble de données. Dans cet article, nous présentons une nouvelle approche topologique, appelée G-Stream, pour le clustering de flux de données évolutives. La méthode proposée est une extension de l'algorithme GNG (Growing Neural Gas) pour gérer le flux de données. G-Stream permet de découvrir de manière incrémentale des clusters de formes arbitraires en ne faisant qu'une seule passe sur les données. Les performances de l'algorithme proposé sont évaluées à la fois sur des données synthétiques et réelles.

1 Introduction

Un flux de données est une séquence, potentiellement infinie, non-stationnaire (la distribution de probabilité des données peut changer au fil du temps) de données arrivant en continu. Dans le cas d'un flux, l'accès aléatoire aux données n'est pas possible et le stockage de toutes les données arrivant est infaisable. Le clustering de flux de données nécessite un processus capable de partitionner des observations de façon continue avec des restrictions au niveau de la mémoire et du temps. Dans la littérature, de nombreux algorithmes de clustering de flux de données ont été adaptés à partir des algorithmes de clustering traditionnel, par exemple, la méthode DbScan (Cao et al. (2006); Isaksson et al. (2012)) basée sur la densité, la méthode de partitionnement *k*-means (Ackermann et al. (2012)), ou encore la méthode basée sur le passage de message AP (Affinity Propagation) (Zhang et al. (2008)). Dans cet article, nous proposons le modèle G-Stream, qui permet de découvrir des clusters de formes arbitraires dans un flux de données en constante évolution. Les caractéristiques et les principaux avantages de G-Stream sont décrits ci-dessous : (a) La structure topologique qui est représentée par un graphe dans lequel chaque nœud représente un cluster. Les nœuds (clusters) voisins sont reliés par des arêtes. La taille du graphe est évolutive. (b) L'utilisation d'une fonction d'oubli afin de réduire l'impact des anciennes données dont la pertinence diminue au fil du temps. Les liens entre les nœuds sont également pondérés. (c) Contrairement à de nombreux algorithmes qui utilisent un nombre important de données pour initialiser leur modèle, G-Stream utilise seulement deux nœuds au départ. (d) Toutes les fonctions de G-Stream sont effectuées en-ligne. (e) L'utilisation de la notion de réservoir pour maintenir, de façon temporaire, les données très éloignées