# Mean-shift : Clustering scalable et distribué

Gaël Beck, Hanane Azzag, Mustapha Lebbah, Tarn Duong, Christophe Cérin

Laboratoire d'Informatique de Paris Nord (LIPN)
Université Paris Nord – Paris 13, F-93430 Villetaneuse, France
Email: {beck, prénom.nom}@lipn.univ-paris13.fr

**Résumé.** Nous présentons dans ce papier un nouvel algorithme Mean-Shift utilisant les K-plus proches voisins pour la montée du gradient (NNMS : Nearest Neighbours Mean Shift). Le coût computationnel intensif de ce dernier a longtemps limité son utilisation sur des jeux de données complexes où un partitionnement en clusters non ellipsoïdaux serait bénéfique. Or, une implémentation scalable de l'algorithme ne compense pas l'augmentation du temps d'exécution en fonction de la taille du jeu de données en raison de sa complexité quadratique. Afin de pallier, ce problème nous avons introduit le "Locality Sensitive Hashing" (LSH) qui est une approximation de la recherche des K-plus proches voisins ainsi qu'une règle empirique pour le choix du K. La combinaison de ces améliorations au sein du NNMS offre l'opportunité d'un traitement pertinent aux problématiques du clustering appliquée aux données massives.

## 1 Introduction

L'objectif de la recherche non supervisée est d'affecter un label à des points non labélisés où le nombre et l'emplacement des clusters sont inconnus. Nous nous sommes concentrés sur un algorithme de clustering modal où le nombre de clusters est défini en terme de modes locaux de la fonction de densité de probabilité qui génère les données. Le plus connu des algorithmes de clustering modal est le $k$-means. Comme ce dernier est basé sur la distribution de mélange normale, il est contraint à trouver des clusters ellipsoidaux ce qui peut être inapproprié pour des jeux de données complexes. Le Mean-shift est une généralisation du $k$-means en raison de sa capacité à calculer des clusters de topologie aléatoire définis comme les bassins d'attractions des modes locaux générés par la montée de gradient de (Fukunaga et Hostetler, 1975). Afin de calculer les chemins de la montée de gradient, les $k$ plus proches voisins sont appropriés car ils s'adaptent à la topologie locale des données. La version actuelle des $k$ plus proches voisins Mean-shift contient des goulots d'étranglement posés par une grille de recherche multiple pour le choix d'un nombre de voisins optimal et par le calcul exact des k plus proches voisins. Nous proposons ici un nouvel algorithme qui résout ces gouffres computationnels : (a) une échelle normale efficace du choix du nombre des plus proches voisins qui évite la recherche en grille, (b) le locality sensitive hashing (LSH) qui est une version approximée des k plus proches voisins et (c) une implémentation MapReduce distribuée.