Sparse Subspace K-means (SSKM)

Abdul Wahab Diallo & Ndèye Niang & Mory OUATTARA
Présenté par Mory OUATTARA
Affiliation : Laboratoire de mathématiques et Informatique, Université NANGUI-ABROGOUA, Abidjan, Côte d'Ivoire

Résumé

La multiplication des sources d'information et le développement de nouvelles technologies ont engendré des bases données complexes, souvent caractérisées par un nombre de variables relativement élevé par rapport aux individus.
L'objectif de ce travail a été de développer des méthodes de classification adaptées à ces jeux de données de grande dimension.

Nous présenterons d'abord les méthodes classiques de classification non supervisée et leur limites. Ensuite nous aborderons les méthodes de subspace clustering adaptées au cas de la grande dimension, plus précisément lorsque les individus sont décrits par des sous-espaces de variables. Lorsque le nombre de variables devient très grand et en présence de variables de bruit, des méthodes dites "sparse" telles que le Sparse K-means permettent de sélectionner les variables pertinentes pour caractériser les partitions. Nous nous sommes inspirés de la méthode sparse k-means pour proposer une nouvelle approche de sparse subspace clustering appelée Sparse Subspace K-means (SSKM) qui est basée sur une modification de la fonction de coût de l'algorithme Sparse K-means. SSKM permet de déterminer les sous espaces caractéristiques au niveau des classes plutôt que de la partition globale. La méthode proposée est ensuite illustrée sur des données simulées et sur un jeu de données réelles. Dans sa comparaison avec les méthodes de la littérature, SSKM se montre aussi bonne ou meilleure tant au niveau des indices de qualité de partition que de détection de variables pertinentes.