SNOW, un algorithme exploratoire pour le subspace clustering

Sylvain Dormieu *, Nicolas Labroche *

* UPMC Univ Paris 06, UMR 7606, LIP6
4 place Jussieu, 75005 Paris, France
sylvain.dormieu@gmail.com,
nicolas.labroche@lip6.fr

Résumé. Cet article propose un nouvel algorithme pour le problème de subspace clustering dénommé SNOW. Contrairement aux approches descendantes classiques, il ne repose pas sur l'hypothèse de localité et permet l'affectation d'une donnée à plusieurs clusters dans des sous-espaces différents. Les expérimentations préliminaires montrent que notre approche obtient de meilleurs résultats que l'algorithme COPAC sur une base de référence et a été appliquée sur une base de données réelles.

1 Introduction

Les méthodes de classification non supervisée - ou clustering - classiques synthétisent l'information en construisant des groupes de données. Ces données sont le plus souvent définies par un ensemble d'attributs et les clusters résultants sont donc déterminés également dans l'espace des attributs. Plusieurs méthodes comme la pondération ou la sélection d'attributs, ou des métriques adaptées permettent de modifier, limiter ou de supprimer l'influence de certains attributs, mais l'ensemble des clusters est généralement défini dans un même espace. Cependant, certains groupes peuvent n'être pertinents que dans un sous-ensemble des attributs. Ce sous-ensemble d'attributs caractéristiques est appelé le *sous-espace* du cluster. Une donnée peut donc appartenir à plusieurs clusters définis dans des sous-espaces différents. Comme indiqué par (Kriegel et al., 2009), l'objectif des méthodes de subspace clustering est de découvrir tous les clusters dans tous les sous-espaces.

Par exemple, pour des données représentant des objets de différentes formes et de différentes couleurs, il est possible de déterminer plusieurs sous-espaces évidents basés sur la couleur, ou bien sur la forme ou enfin sur les deux attributs à la fois. Dans cet exemple, un carré rouge devrait pouvoir appartenir à la fois au cluster *rouge* si le sous-espace est limité à l'attribut de couleur et au cluster *carré* s'il est limité au sous-espace de forme.

Ce papier est organisé comme suit : la section 2 rappelle les principaux travaux conduits dans le domaine du subspace clustering. La section 3 décrit l'algorithme SNOW. La section 4 présente des résultats comparatifs avec l'algorithme COPAC sur des données artificielles et illustre les résultats de Snow sur des données réelles issues du UCI Machine Learning Repository. La section 5 conclut l'article et présente les perspectives.