Apport des modèles locaux pour les K-moyennes prédictives

Vincent Lemaire, Oumaima Alaoui Ismaili

2 Avenue Pierre Marzin, 22300 Lannion
(vincent,oumaima)@orange.com

Résumé. Dans le cadre du clustering prédictif, pour attribuer la classe aux groupes formés à la fin de la phase d'apprentissage, le vote majoritaire est la méthode communément utilisée. Cependant, cette approche comporte certaines limitations qui influent directement sur la qualité des résultats obtenus en termes de prédiction. Pour surmonter ce problème, nous proposons d'incorporer des modèles prédictifs localement dans les clusters formés afin d'améliorer la qualité prédictive du modèle global. Les résultats expérimentaux montrent que cette incorporation permet d'obtenir des résultats (en termes de prédiction) significativement meilleurs par rapport à ceux obtenus en utilisant le vote majoritaire ainsi que des résultats très compétitifs avec ceux obtenus par des algorithmes performants d'apprentissage supervisé "similaires". Ceci est effectué sans dégrader le pouvoir descriptif (explicatif) du modèle global.

1 Introduction

L'algorithme des K-moyennes prédictives (Eick et al., 2004; Al-Harbi et Rayward-Smith, 2006; Alaoui Ismaili, 2016) est une version modifiée de l'algorithme des K-moyennes standard. Il vise à décrire et à prédire d'une manière simultanée.

L'idée est de générer dans la phase d'apprentissage un nombre minimal de clusters compacts dont les instances doivent appartenir à la même classe. Ces clusters vont servir par la suite à décrire les données et à prédire la classe des nouvelles instances (voir la figure 1).

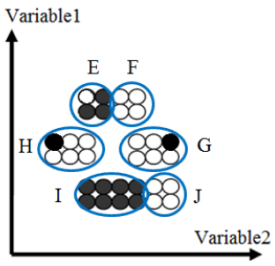

FIG. 1 – *Objectif du clustering prédictif*

La méthode communément utilisée dans la littérature permettant d'attribuer la classe aux clusters formés par l'algorithme des K-moyennes prédictives est le vote majoritaire. Bien que cette approche parvienne à obtenir de bons résultats, celle-ci a également certaines limites. On citera par example :

— pour le taux de bonne classification (ACC) : si un cluster contient Q% d'instances de la classe C1 et 100-Q% d'instances de la classe C2, alors l'utilisation du vote majoritaire va produire un taux de mauvaise classification très important (Q). La présence d'un modèle local à ce cluster devrait permettre de mieux discriminer les exemples selon leur classe d'appartenance. Ceci est très visible pour les clusters E, H, G de la figure 1.