Logiciel d'Aide à l'Évaluation des Catégorisations

Julien Velcin, William Vacher, Jean-Gabriel Ganascia

LIP6 - 104, avenue du président Kennedy - 75016 Paris
{Julien.Velcin, Jean-Gabriel.Ganascia}@lip6.fr, William.Vacher@free.fr
http://www-poleia.lip6.fr/~velcin

Les méthodes de classification automatique sont employées dans des domaines variés et de nombreux algorithmes ont été proposés dans la littérature. Au milieu de cette "jungle", il semble parfois difficile à un simple utilisateur de choisir quel algorithme est le plus adapté à ses besoins. Depuis le milieu des années 90, une nouvelle thématique de recherches, appelée *clustering validity*, tente de répondre à ce genre d'interrogation en proposant des indices pour juger de la qualité des catégorisations obtenues. Mais le choix est parfois difficile entre ces indices et il peut s'avérer délicat de prendre la bonne décision. C'est pourquoi nous proposons un logiciel adapté à cette problématique d'évaluation.

1 Evaluer les catégorisations

La validation manuelle n'est pas forcément toujours faisable ou souhaitable. C'est pourquoi il convient de prendre en considération des méthodes automatiques quantitatives afin de donner une idée de la qualité des catégorisations. Nous nous basons sur la distinction entre critères "externes" et "internes" faite par Halkidi et al. (2002). Alors que les premiers reposent sur l'hypothèse d'une partition idéale des données (étiquettes données par l'utilisateur, par exemple), les seconds n'utilisent aucune information *a priori* pour juger de la qualité des catégorisations. C'est cette seconde approche que nous avons choisi d'adopter dans notre logiciel.

Contrairement à l'approche externe, aucun étiquetage préalable des données ne permet ici de comparer le résultat du clustering à un quelconque modèle idéal. De nombreux indices de validité ont été proposés et des travaux récents attestent de la vitalité de cette perspective de recherche. Ils se basent sur la recherche, thème classique en apprentissage non supervisé, d'un compromis entre les principes de similarité intra-classe et de dissimilarité inter-classes. Des indices caractéristiques de cette approche interne sont les indices de Dünn, Davies-Bouldin et Hubert modifié, qui ont été implémentés dans notre logiciel.

2 Logiciel et expérimentations

L'objectif du logiciel que nous proposons est d'aider l'utilisateur à comparer différentes partitions d'un même jeu de données sur la base de critères internes. Ces partitions peuvent être les résultats obtenus à l'aide d'un ou de plusieurs algorithmes de classification automatique, tels les k-means ou EM. Les données d'entrée sont, d'une part, la définition du langage de description et des exemples d'apprentissage décrits à l'aide de ce langage, et, d'autre part, les