

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.

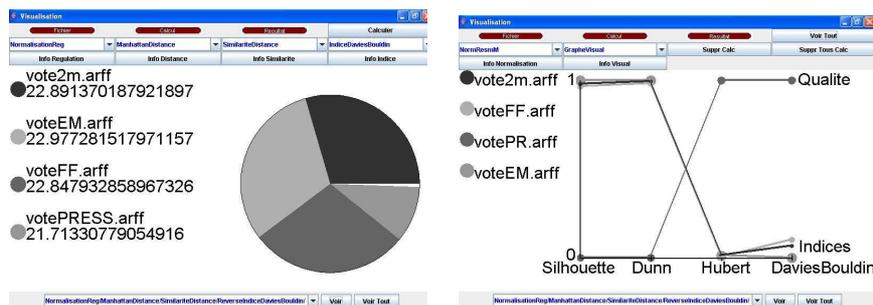
Contrairement à l'approche externe, aucun étiquetage préalable des données ne permet ici de comparer le résultat du clustering à un quelconque modèle idéal. De nombreux indices de validité ont été proposés et des travaux récents attestent de la vitalité de cette perspective de recherche. Ils se basent 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

Logiciel d'Aide à l'Évaluation des Catégorisations

partitions qui feront l'objet de la comparaison. L'évaluation repose sur trois composantes : l'indice utilisé, la mesure de distance (ou de similarité) choisie, ainsi que la normalisation effectuée sur les attributs numériques. Le logiciel permet de lancer plusieurs évaluations en même temps et propose, en sortie, une visualisation des résultats obtenus. La visualisation est différente suivant que l'on traite un ou plusieurs critères. De plus, le caractère évolutif de notre logiciel donne l'opportunité d'ajouter très facilement de nouveaux indices ou de nouvelles distances.



La figure ci-dessus présente les résultats obtenus avec quatre algorithmes (k-means, EM, Farthest-first et PRESS) sur la célèbre base "vote" du répertoire UCI. Elle permet de constater la supériorité de l'un des algorithmes dans le cas mono-critère (indice de Davies-Bouldin), ici celui qui a obtenu la plus petite des aires. Le cas multi-critères, par contre, semble indiquer deux types de résultats distincts. L'utilisation de notre logiciel peut ainsi suggérer à l'utilisateur d'étudier plus attentivement les raisons de cette différence.

3 Conclusion et perspectives

Nous présentons un logiciel pour aider l'utilisateur à comparer les résultats obtenus par des algorithmes de classification. La caractéristique principale de ce travail est son caractère évolutif : ajout de nouveaux indices, de nouvelles distances, etc. Dans les perspectives à court terme, nous souhaitons étendre le logiciel aux indices externes, tels la F-mesure ou les fonctions entropiques. A plus long terme, cet outil devrait nous permettre de comparer, non plus les partitions ou les algorithmes, mais directement les critères de pertinence. Ces derniers pourraient alors être regroupées et mis en relation avec la nature des données traitées (données clairsemées, bruitées, à grande dimension, etc.). Ceci devrait mener à une contribution concernant l'évaluation des techniques d'apprentissage non supervisé, évaluation qui présente encore de réelles difficultés au jour d'aujourd'hui.

Summary

This paper details a software that can assist the user for clustering comparison. It gives a clear visualization of different criteria (Dunn, Silhouette, etc.) calculated on one or more partitions of the data. The main feature is its modularity in three components: a quality criterion, a comparison measure and a normalization on numerical attributes. Furthermore, it allows the user to add its own items into those components.