Classification ascendante hiérarchique à noyaux et une application aux données textuelles

Julien Ah-Pine*, Xinyu Wang*

*Université de Lyon, Université de Lyon 2, ERIC EA 3083
5, avenue Pierre Mendès France 69676 Bron Cedex, France
Julien.Ah-Pine,Xinyu.Wang@univ-lyon2.fr

Résumé. La formule de Lance et Williams permet d'unifier plusieurs méthodes de classification ascendante hiérarchique (CAH). Dans cet article, nous supposons que les données sont représentées dans un espace euclidien et nous établissons une nouvelle expression de cette formule en utilisant les similarités cosinus au lieu des distances euclidiennes au carré. Notre approche présente les avantages suivants. D'une part, elle permet d'étendre naturellement les méthodes classiques de CAH aux fonctions noyau. D'autre part, elle permet d'appliquer des méthodes d'écrêtage permettant de rendre la matrice de similarités creuse afin d'améliorer la complexité de la CAH. L'application de notre approche sur des tâches de classification automatique de données textuelles montre d'une part, que le passage à l'échelle est amélioré en mémoire et en temps de traitement; d'autre part, que la qualité des résultats est préservée voire améliorée.

1 Introduction

Soit un ensemble d'objets $\mathcal{D}$ constitué de N éléments. Soit D la matrice des dissimilarités entre chaque paire d'objets. La procédure classique de la classification ascendante hiérarchique (CAH) initialise un arbre par N feuilles [1] puis, à chaque itération, elle regroupe le couple de groupes d'objets (C_i, C_j) dont la distance est la plus petite :

$$(C_i, C_j) = \underset{(C_k, C_l)}{\arg\min}\, D(C_k, C_l) \quad (1)$$

Un nouveau groupe $C_{(ij)} = C_i \cup C_j$ est créé et un nœud est ajouté à l'arbre binaire. Il faut ensuite calculer les dissimilarités entre $C_{(ij)}$ et les groupes existants. Il existe plusieurs techniques de CAH selon la façon dont on définit la dissimilarité entre deux groupes. Cependant, Lance et Williams (Lance et Williams, 1967) ont montré que la majorité d'entre elles pouvaient être généralisées par la formule (formule LW) suivante :

$$D(C_{(ij)}, C_k) = \alpha_i D(C_i, C_k) + \alpha_j D(C_j, C_k) + \beta D(C_i, C_j) + \gamma |D(C_i, C_k) - D(C_j, C_k)| \quad (2)$$

Nous présentons dans le Tableau 1, la définition des sept méthodes que nous étudions dans le cadre de la formule LW.

1. Les N objets sont vus comme des singletons.