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.

Méthodes	α_i	α_j	β	γ
single	1/2	1/2	0	-1/2
complete	1/2	1/2	0	1/2
average	$\frac{\lvert C_i\rvert}{\lvert C_i\rvert+\lvert C_j\rvert}$	$\frac{\lvert C_j\rvert}{\lvert C_i\rvert+\lvert C_j\rvert}$	0	0
Mcquitty	1/2	1/2	0	0
centroid	$\frac{\lvert C_i\rvert}{\lvert C_i\rvert+\lvert C_j\rvert}$	$\frac{\lvert C_j\rvert}{\lvert C_i\rvert+\lvert C_j\rvert}$	$-\frac{\lvert C_i\rvert\lvert C_j\rvert}{(\lvert C_i\rvert+\lvert C_j\rvert)^2}$	0
median	1/2	1/2	-1/4	0
Ward	$\frac{\lvert C_i\rvert+\lvert C_k\rvert}{\lvert C_i\rvert+\lvert C_j\rvert+\lvert C_k\rvert}$	$\frac{\lvert C_j\rvert+\lvert C_k\rvert}{\lvert C_i\rvert+\lvert C_j\rvert+\lvert C_k\rvert}$	$-\frac{\lvert C_k\rvert}{\lvert C_i\rvert+\lvert C_j\rvert+\lvert C_k\rvert}$	0

TAB. 1 – *Formule de Lance et Williams : méthodes et paramètres.*

La procédure “bottom-up” décrite ci-dessus avec la formule LW forme l’approche classique de la CAH. Cette dernière est simple et flexible mais elle est coûteuse en mémoire, en temps de traitement et ne passe pas à l’échelle dans le cas des grandes masses de données. En effet, la matrice de dissimilarité est nécessairement dense et coûte en mémoire $O(N^2)$ tandis que la procédure “bottom-up” a une complexité en $O(N^3)$.

Dans cet article, nous proposons une reformulation de l’approche classique que nous venons de rappeler mais qui permet d’outrepasser les limites évoquées ci-dessus. Notre idée principale est de définir une expression équivalente de la formule LW en termes de similarités plutôt qu’en termes de dissimilarités. Dans cette perspective, nous supposons que les objets sont représentés dans un espace euclidien et que la dissimilarité est mesurée par le carré de la distance euclidienne entre les vecteurs normés. Dans ce cas, le produit scalaire associé est le cosinus de l’angle formé par les vecteurs et peut-être ainsi vu comme une mesure de similarité.

Notre approche présente un double avantage. D’une part, elle permet d’étendre naturellement la CAH à des fonctions noyau permettant ainsi de traiter plus efficacement les données qui sont non linéairement séparables dans l’espace de description initial. D’autre part, elle permet de définir des stratégies d’écrêtage de la matrice de similarité S qui est rendu creuse. Cette dernière est alors plus légère en mémoire et nous pouvons également améliorer les temps de traitement comme nous l’expliquerons par la suite. Afin d’illustrer les propriétés de notre approche, nous appliquons celle-ci sur des tâches de classification automatique de données textuelles. Nos résultats montrent que notre méthode permet de réduire la complexité de la CAH et d’obtenir de meilleurs résultats dans de nombreux cas.

La suite de l’article est organisée de la façon suivante. En section 2, nous introduisons les différents ingrédients de notre approche. Puis, en section 3, nous présentons les résultats des expériences que nous avons menées sur trois jeux de données classiques. Nous concluons en section 4 par une discussion et une esquisse des travaux futurs.

2 Notre approche

Nous supposons que les objets sont représentés dans un espace euclidien $\mathcal{I}$ de dimension p et de produit scalaire noté $\langle .,. \rangle$. Pour deux vecteurs $x, y \in \mathcal{D}$, leur similarité est définie par :

$$S(x,y) = \langle \frac{x}{\|x\|}, \frac{y}{\|y\|} \rangle \tag{3}$$

Remarquons que $\forall x \in \mathcal{D}$, $S(x,x) = 1$. Ainsi, les vecteurs ont tous la même norme ce qui constitue une condition importante dans notre cas. La matrice S des produits scalaires est associée à la matrice D des carrés des distances euclidiennes comme suit :

$$D(x,y) = \|\frac{x}{\|x\|} - \frac{y}{\|y\|}\|^2 = \quad S(x,x) + S(y,y) - 2S(x,y) = 2(1 - S(x,y)) \tag{4}$$

Sous cette condition, nous présentons une procédure "bottom-up" fondée sur deux formules de récurrence qui produisent un arbre binaire équivalent à l'approche classique. Partant d'un arbre à N feuilles, nous regroupons itérativement les couples de groupes d'objets (C_i, C_j) vérifiant :

$$(C_i, C_j) = \underset{(C_k, C_l)}{\arg\max}\, S(C_k, C_l) - \frac{1}{2}(S(C_k, C_k) + S(C_l, C_l)) \tag{5}$$

Lorsque le nouveau nœud $C_{(ij)} = C_i \cup C_j$ est ajouté à l'arbre, les similarités entre $C_{(ij)}$ et les groupes existants ainsi qu'avec lui même sont obtenues par les formules suivantes :

$$\begin{aligned} S(C_{(ij)}, C_k) &= \alpha_i S(C_i, C_k) + \alpha_j S(C_j, C_k) + \beta S(C_i, C_j) \\ &\quad -\gamma |S(C_i, C_k) - S(C_j, C_k)| \qquad (6) \\ S(C_{(ij)}, C_{(ij)}) &= \delta_i S(C_i, C_i) + \delta_j S(C_j, C_j) \qquad (7) \end{aligned}$$

Dans notre cas, les sept méthodes de CAH sont définies avec les paramètres listés dans les Tableaux 1 et 2. Les paramètres $\alpha_i, \alpha_j, \beta, \gamma$ dans (6) sont les mêmes que ceux énoncés dans le Tableau 1 alors que les nouveaux paramètres δ_i, δ_j dans (7) sont introduits dans le Tableau 2. Notons que, exceptées les méthodes "median" et "centroid", δ_i, δ_j peuvent être choisis arbitrairement à condition que $\delta_i + \delta_j = 1$.

Dans cette reformulation de la formule LW, il est nécessaire d'avoir deux formules de récurrence distinctes, l'une pour $S(C_{(ij)}, C_k)$ et l'autre pour $S(C_{(ij)}, C_{(ij)})$, afin d'obtenir l'équivalence entre la recherche du minimum dans (1) et celle du maximum dans (5). La preuve de cette équivalence s'obtient en fait en montrant que $S(C_k, C_l) - \frac{1}{2}(S(C_k, C_k) + S(C_l, C_l))$ dans (5) est égal à $-\frac{1}{2}D(C_k, C_l)$ dans (1).

Méthodes	δ_i	δ_j
single	1/2	1/2
complete	1/2	1/2
average	1/2	1/2
Mcquitty	1/2	1/2
centroid	$\frac{\|C_i\|^2}{(\|C_i\|+\|C_j\|)^2}$	$\frac{\|C_j\|^2}{(\|C_i\|+\|C_j\|)^2}$
median	1/4	1/4
Ward	1/2	1/2

TAB. 2 – *Formule basée sur les similarités cosinus : méthodes et paramètres δ_i et δ_j.*

Comme nous raisonnons désormais avec des produits scalaires, nous pouvons étendre naturellement notre approche à des fonctions noyau. On note K une matrice de produits scalaires (ou matrice de Gram) de taille N et pour deux objets $x, y \in \mathcal{D}, K(x,y) = \langle \phi(x), \phi(y) \rangle$ où

$\phi : \mathcal{I} \rightarrow \mathcal{F}$ et $\mathcal{F}$ est un espace de Hilbert de dimension $q > p$ (q pouvant être infini). La matrice S contenant des similarités cosinus dans l'espace $\mathcal{F}$ peut alors être facilement obtenue en utilisant l'astuce du noyau : $\forall x, y \in \mathcal{D}, S(x,y) = K(x,y)/\sqrt{K(x,x)K(y,y)}$.

Ensuite, de façon générale, S peut contenir des valeurs négatives. Dans ce cas, soit $m < 0$ la plus petite de ces valeurs. Il est toujours possible de transformer S de façon à n'avoir que des valeurs positives : $\forall x, y \in \mathcal{D}, S(x,y) \leftarrow (S(x,y) + |m|)/(1 + |m|)$. Comme cette application est monotone croissante, la matrice S transformée reste une matrice de Gram.

Supposons désormais que les valeurs de S sont comprises entre 0 et 1. Nous écrêtons S selon un paramètre de seuillage $\tau \in [0,1]$. Ainsi, toute valeur inférieure ou égale à τ est remplacée par 0. La matrice S devient creuse [2] et la complexité en mémoire est réduite *de facto* à $O(M)$, M étant le nombre de paires d'objets dont la similarité est strictement positive.

Pour améliorer la complexité en temps de traitement, nous proposons de restreindre la recherche du couple de groupes d'objets à fusionner aux seules paires dont la similarité est strictement positive. Nous introduisons pour cela l'ensemble $\mathbb{S} = \{(C_k, C_l) : S(C_k, C_l) > 0\}$. L'équation (5) est alors remplacée par :

$$(C_i, C_j) = \underset{(C_k,C_l)\in\mathbb{S}}{\arg\max}\, S(C_k, C_l) - \frac{1}{2}(S(C_k, C_k) + S(C_l, C_l)) \tag{8}$$

Dans ce cas, la complexité de notre procédure "bottom-up" est réduite à $O(NM)$.

3 Expériences

Les objectifs de nos expériences sont de démontrer que, sous la condition énoncée précédemment : (i) notre méthode basée sur (5), (6) et (7) est équivalente à la procédure classique fondée sur (1) et (2) ; (ii) l'écrêtage de S et notre approche utilisant (8), (6) et (7) permet de réduire considérablement les coûts en mémoire et en temps de traitement et de produire des résultats de bonne qualité en comparaison de la méthode classique.

Pour cela, nous expérimentons sur des tâches de classification automatique de données textuelles. Les documents sont représentés dans un espace vectoriel "sacs de mots". Les termes sont les différentes dimension de l'espace et les coordonnées des vecteurs sont les nombres d'occurrence des termes dans les documents. Notons que les termes étant apparus dans moins de 0.2% et dans plus de 95% des documents de la collection ont été retirés. Les différentes cas à l'étude sont :

— Reuters, 10 groupes, 2446 documents, 2547 termes.
— Smart, 3 groupes, 3893 documents, 3025 termes.
— 20ng, 15 groupes, 4483 documents, 4455 termes.

Étant donné une matrice documents-termes, S [3] et D sont déterminés par (3) et (4).

Le résultat de l'approche classique basée sur D est calculé (résultat de référence) ainsi que les résultats de notre approche basée sur S avec différents niveaux de seuillage τ. Ce paramètre est choisi de façon à ce que 0%, 10%, 25%, 50%, 75%, 90% de S soit écrêtées.

2. Notons que si nous devions appliquer ce même principe à partir de la matrice de dissimilarité D, ce sont les valeurs les plus grandes et au-dessus d'un seuil qu'il aurait fallu écrêter. Or ceci n'est pas raisonnable.

3. Les documents étant représentés par des vecteurs de composantes positives, les cosinus sont donc de valeurs positives.

Pour mesurer la proximité entre l'arbre obtenu par la méthode classique et ceux obtenus avec notre approche, nous utilisons la valeur absolue de la corrélation cophénétique (CC).

Afin d'évaluer la qualité d'un résultat, nous coupons l'arbre obtenu pour obtenir la partition avec le nombre correct de groupes et nous comparons celle-ci avec la vérité terrain. L'index de Rand corrigé (ARI) est une mesure classique dans ce cas.

Nous avons utilisé deux fonctions noyau : linéaire et gaussien [4].

En raison de la restriction en nombre de pages [5], nous ne pouvons pas montrer l'ensemble des résultats de nos expériences. Cependant, voici des observations importantes que nous pouvons en faire :

— lorsque 0% de S est écrêté nous retrouvons le même résultat que l'approche classique,
— l'usage de la mémoire et les temps de traitement diminuent lorsque S est de plus en plus creuse,
— la méthode "single link" présente un comportement particulier : même en écrêtant S de 90% de ses valeurs, nous obtenons le même résultat que l'approche classique,
— les mesures ARI sont instables selon les méthodes et jeux de données mais nous obtenons souvent des améliorations,
— "average" et "ward" sont les méthodes les plus performantes en général.

Concernant les deux derniers points, nous donnons plus précisément dans la Tableau 3 les meilleures performances observées. Il est important de noter que celles-ci sont souvent obtenues par notre méthode avec une matrice S largement écrêtée.

	Méthode	Noyau	τ	Mem%	Temps%	CC	ARI
Reuters	Average	Gaussien	0	0	0	1	**0.543**
	Average	Gaussien	0.99	-75	-62	0.81	0.539
Smart	Average	Linéaire	0	0	0	1	0.939
	Average	Linéaire	0.078	-90	-85	0.96	**0.944**
20ng	Ward	Gaussien	0	0	0	1	0.100
	Ward	Gaussien	0.99	-50	-47	0.26	**0.154**

TAB. 3 – *Meilleures valeurs ARI pour chaque collection, quand S est dense et quand S est écrêtée avec dans ce cas la diminution relative en mémoire et en temps de traitement.*

4 Discussion et travaux futurs

Notre méthode repose sur une expression de la formule LW en termes de similarités cosinus et sur l'écrêtage de la matrice de similarité correspondante. En théorie, elle améliore le passage à l'échelle de la CAH et permet également l'extension de celle-ci à des fonctions noyau. En pratique, nous avons pu constater ces améliorations dans le cas de la classification automatique de documents. De surcroît, nous pouvons également observer que notre approche permet d'aboutir à plusieurs reprises à de meilleurs résultats en terme de qualité. Nous pouvons expliquer ce phénomène par deux points : (i) l'écrêtage de S permet de réduire le bruit ; (ii) notre méthode peut-être vue telle une sorte de "fermeture transitive" qui repose sur le principe

4. Le paramètre γ du noyau gaussien est fixé à $1/p$.
5. Cet article est une version française et écourtée de Ah-Pine et Wang (2016).

"les amis de mes amis sont mes amis" ce qui permet de mieux tenir compte de la géométrie intrinsèque aux données.

Toutefois, il est possible qu'un mauvais choix du paramètre de seuillage conduise à une dégradation de la qualité. Un point critique, et donc un des travaux futurs, concerne le choix du paramètre τ mais également l'application de diverses autres stratégies d'écrêtage comme la restriction aux k plus proches voisins. Parmi les travaux en cours, nous poursuivons l'effort d'amélioration du passage à l'échelle de ce type de classification automatique en implémentant notre approche dans une architecture distribuée en utilisant l'outil Apache Spark.

Références

Ah-Pine, J. et X. Wang (2016). Similarity based hierarchical clustering with an application to text collections. In *Advances in Intelligent Data Analysis XV - 15th International Symposium, IDA 2016, Stockholm, Sweden, October 13-15, 2016, Proceedings*, pp. 320–331.

Cristianini, N. et J. Shawe-Taylor (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Lance, G. N. et W. T. Williams (1967). A general theory of classificatory sorting strategies ii. clustering systems. *The computer journal 10*(3), 271–277.

Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv :1109.2378*.

Murtagh, F. et P. Contreras (2012). Algorithms for hierarchical clustering : an overview. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery 2*(1), 86–97.

Xu, R., D. Wunsch, et al. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on 16*(3), 645–678.

Summary

Lance-Williams formula is a framework that unifies seven schemes of agglomerative hierarchical clustering. In this paper, we establish a new expression of this formula using cosine similarities instead of distances. We state conditions under which the new formula is equivalent to the original one. The interest of our approach is twofold. Firstly, we can naturally extend agglomerative hierarchical clustering techniques to kernel functions. Secondly, reasoning in terms of similarities allows us to design thresholding strategies on proximity values. Thereby, we propose to sparsify the similarity matrix in the goal of making these clustering techniques more efficient. We apply our approach to text clustering tasks. Our results show that sparsifying the inner product matrix considerably decreases memory usage and shortens running time while assuring the clustering quality.