Classification ascendante hiérarchique à noyaux et pistes pour un meilleur passage à l'échelle

Julien Ah-Pine & Xinyu Wang

Laboratoire ERIC, 5, avenue Pierre Mendès France 69676 Bron Cedex, France
{julien.ah-pine,xinyu.wang}@univ-lyon2.fr

Résumé. Nous nous intéressons au problème de la classification ascendante hiérarchique d'un ensemble d'individus représentés dans un espace euclidien. Nous donnons une expression de la formule de Lance et Williams en fonction de produits scalaires plutôt qu'en termes de distances. Nous établissons les conditions dans lesquelles cette nouvelle expression est équivalence à la méthode initiale. L'intérêt de cette approche est double. Tout d'abord, nous pouvons étendre naturellement les techniques classiques de classification ascendante hiérarchique aux fonctions noyaux. Ensuite, le raisonnement sur des matrices de produits scalaires est davantage propice à la définition de méthodes de seuillage de mesures de proximités. Nous proposons alors de prétraiter la matrice de proximités de façon à la rendre éparse afin de permettre un meilleur passage à l'échelle de ces techniques de classification.

Mots-clés. Classification ascendante hiérarchique, Méthodes à noyaux, Matrices éparses, Passage à l'échelle.

Abstract. We are interested in clustering a set of points represented in an euclidean space by means of agglomerative hierarchical techniques. We establish an expression of the Lance and Williams formula using dot products instead of distances. We state the conditions for which this new formula is equivalent to the original one. The interest of such an approach is twofold. Firstly, we can naturally extend agglomerative hierarchical clustering techniques to kernel functions. Secondly, reasoning in terms of dot products allows us to better design thresholding strategies of proximity values. Thereby, we propose to sparsify the proximity matrix in the goal of making these clustering techniques more scalable.

Keywords. Agglomerative hierarchical clustering, Kernel methods, Sparse matrices, Scalability.

1 La CAH et la formule de Lance et Williams

Nous nous intéressons au problème de classification automatique d'un ensemble de n individus dont nous connaissons les mesures de proximité deux à deux. Parmi les différents

types de techniques, nous étudions les méthodes hiérarchiques dont le but est de construire un arbre binaire composé de n partitions emboîtées. Il existe deux façons de construire un tel arbre. La première consiste à partir de la partition triviale à une classe et de scinder récursivement les classes en deux jusqu'à obtenir la partition triviale à n classes. La seconde méthode, à l'inverse, part de la partition à n classes et regroupe successivement deux classes en une seule jusqu'à obtenir la partition à une classe. Cette deuxième approche est dite classification ascendante hiérarchique (CAH) et nous portons notre étude sur ce type de techniques (voir Murtagh et Contreras (2012) pour un état de l'art récent sur la classification hiérarchique).

La CAH prend en entrée une matrice carrée symétrique de taille n que l'on notera par $\mathbf{D}$ et qui comporte les mesures de dissimilarités entre chaque paire d'individus. Puis, elle construit l'arbre binaire en regroupant à chaque itération deux classes. A chaque pas, l'algorithme transforme la matrice $\mathbf{D}$ de sorte à calculer les mesures de dissimilarités entre la classe nouvellement formée et les autres classes restantes. La sortie de la procédure correspond à une suite de partitions emboîtées qui est représentée par une structure appelée dendrogramme. Notons par k une classe d'individus. Remarquons que k peut être réduit à un singleton et auquel cas il s'agira de l'individu $\mathbf{x}_k$. A chaque itération, l'algorithme de CAH effectue les deux étapes clefs suivantes:

1. Déterminer les deux classes les plus proches en résolvant:

$$\min_{k,l} \mathbf{D}_{kl} \tag{1}$$

2. Fusionner les deux classes k et l les plus proches et calculer les dissimilarités entre la nouvelle classe constituée (kl) et les autres classes restantes m:

$$\forall m : \mathbf{D}_{(kl)m} \tag{2}$$

Dans le cadre de la CAH, la formule de Lance et Williams (1967) occupe une place centrale. Il s'agit d'une équation de récurrence dépendant de 4 paramètres réels et permettant d'unifier de nombreuses méthodes qui sont les suivantes: "single link", "complete link", "average" (UPGMA), "Mcquitty" (WPGMA), "centroid" (UPGMC), "median" (WPGMC) et "Ward".

Cette formule permet de calculer de façon itérative les mesures de dissimilarités (2):

$$\mathbf{D}_{(kl)m} = \alpha_k \mathbf{D}_{km} + \alpha_l \mathbf{D}_{lm} + \beta \mathbf{D}_{kl} + \gamma |\mathbf{D}_{km} - \mathbf{D}_{lm}| \tag{3}$$

Par exemple, la méthode "centroid" correspond aux paramètres suivants: $\alpha_k = \frac{|k|}{|k|+|l|}, \alpha_l = \frac{|l|}{|k|+|l|}, \beta = 0, \gamma = 0$; où $|k|$ est le cardinale de la classe k.

2 L'extension de la CAH aux fonctions noyaux

Dans cette communication, nous nous restreignons aux matrices $\mathbf{D}$ dont les mesures de dissimilarités sont des distances euclidiennes. Autrement dit, nous supposons que pour deux individus $\mathbf{x}_i$ et $\mathbf{x}_j$, nous avons $\mathbf{D}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$ qui peut également s'écrire $\mathbf{D}_{ij} = \langle \mathbf{x}_i, \mathbf{x}_i \rangle + \langle \mathbf{x}_j, \mathbf{x}_j \rangle - 2\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ où $\langle ., . \rangle$ est le produit scalaire associé à la norme $\|.\|$. Par abus de langage, nous utiliserons similarités et produits scalaires de façon équivalente.

Nous proposons d'exprimer la formule de Lance et Williams en termes de similarités plutôt qu'en termes de distances. Notre objectif est d'apporter une vision quelque peu différente de cette méthode générique afin de mettre à profit certains avantages liés à l'utilisation de produits scalaires.

En effet, la première extension que nous suggérons consiste à associer à la CAH la riche gamme d'outils proposés dans le cadre des méthodes à noyaux. Notons par $\mathbb{E}$ l'espace euclidien initial dans lequel sont représentés les individus. L'idée principale des méthodes à noyaux est de projeter les individus dans un espace $\mathbb{F}$ qui est de plus grande dimension que $\mathbb{E}$. Ceci est fort utile lorsque la représentation dans $\mathbb{E}$ présente des limites pour des tâches de classification comme le cas de groupes d'individus non linéairement séparables.

La projection des individus dans $\mathbb{F}$ s'opère par une application $\phi : \mathbb{E} \to \mathbb{F}$. Cependant, il n'est pas utile de représenter explicitement les individus dans $\mathbb{F}$. Dans le cadre des méthodes à noyaux, c'est la matrice des produits scalaires dans $\mathbb{F}$ qui est fondamentale pour mettre en oeuvre les différentes techniques. Or, il est possible de déterminer cette matrice de similarités sans passer par la représentation ϕ. On utilise en fait une fonction noyau $K : \mathbb{E} \times \mathbb{E} \to \mathbb{R}$ qui pour toute paire d'individus $\mathbf{x}_i$ et $\mathbf{x}_j$ est telle que: $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. On parle alors de l'astuce du noyau ("kernel trick").

Nous supposerons dans la suite que nous disposons d'une matrice carrée symétrique de taille n notée $\mathbf{S}$ et qui contient les mesures des produits scalaires entre chaque couple d'individus obtenus par une fonction noyau K. Nous avons ainsi la définition suivante, $\forall i, j = 1, \ldots, n : \mathbf{S}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$.

Nous montrons alors que la procédure suivante basée sur des similarités et incluant 8 paramètres réels $a_k, a_l, a_{kl}, a_k, a_l, a_{kl}, c, d$, permet d'obtenir les mêmes résultats que la CAH basée sur la formule de Lance et Williams:

1. Déterminer les deux classes les plus proches en résolvant:

$$\max_{k,l} c\mathbf{S}_{kl} + d\left[\mathbf{S}_{kk} + \mathbf{S}_{ll}\right] \tag{4}$$

2. Fusionner les deux classes k et l les plus proches et calculer les similarités entre la nouvelle classe (kl) et les autres classes m ainsi qu'avec elle-même (auto-similarité) à l'aide, respectivement, des formules de récurrence suivantes:

$$\mathbf{S}_{(kl)m} = a_k\mathbf{S}_{km} + a_l\mathbf{S}_{lm} + a_{kl}|\mathbf{S}_{km} - \mathbf{S}_{lm}| \tag{5}$$

$$\mathbf{S}_{(kl)(kl)} = b_k\mathbf{S}_{kk} + b_l\mathbf{S}_{ll} + b_{kl}\mathbf{S}_{kl} \tag{6}$$

Nous présentons dans cette communication les correspondances des différentes techniques entre la formule de Lance et Williams et la méthode introduite ci-dessus. A titre illustratif dans le cadre de ce résumé long, nous donnons les paramètres pour la méthode "centroid". Celle-ci est obtenue quelque soit **S**, en utilisant les valeurs des paramètres suivantes: $a_k = \frac{|k|}{|k|+|l|}$, $a_l = \frac{|l|}{|k|+|l|}$, $a_{kl} = 0$ dans (5); $b_k = \frac{|k|^2}{(|k|+|l|)^2}$, $b_l = \frac{|l|^2}{(|k|+|l|)^2}$, $b_{kl} = \frac{2|k||l|}{(|k|+|l|)^2}$ dans (6) et $c = 2$, $d = -1$ dans (4).

3 Seuiller les similarités pour réduire la complexité

L'algorithme classique de CAH repose sur le stockage de la matrice carrée des distances **D** qui est nécessairement dense. Cette approche a une complexité en mémoire en $O(n^2)$ et une complexité temporelle en $O(n^3)$. Le passage à l'échelle de cet algorithme est donc très limité. Pour pallier à cet inconvénient, des algorithmes plus performants ont été définis dans les années 1980. Ainsi, un algorithme alternatif basé sur les "nearest neighbor chains" permet de trouver une solution identique pour de nombreuses méthodes de CAH avec une complexité temporelle réduite à $O(n^2)$ et une complexité en mémoire restant en revanche de l'ordre de $O(n^2)$ (voir par exemple Murtagh (1983) pour une synthèse de ces algorithmes). Malgré ces avancées, une double complexité quadratique limite encore fortement l'utilisation de la CAH pour la classification automatique de données massives.

La deuxième extension proposée dans cette communication concerne donc la question d'un meilleur passage à l'échelle de la CAH. Dans notre cas, nous proposons de réduire la taille de la matrice de proximités donnée en entrée de la CAH. Cela permettrait également de diminuer la complexité temporelle. Pour ce faire, la stratégie générale que nous adoptons est de prétraiter la matrice de proximités en remplaçant des valeurs par zéro afin de la rendre éparse. Dans cette perspective, nous soutenons qu'il est plus avantageux de raisonner en termes de similarités qu'en termes de dissimilarités. En effet, ce sont les valeurs de proximités qui sont les moins pertinentes pour la classification qui devraient être écrêtées. S'il s'agit de mesures de similarités, cela revient donc naturellement à rendre nulle des mesures très faibles. En revanche, si l'on écrête des dissimilarités, cela reviendrait à remplacer des valeurs très grandes par zéro. Or ceci est incohérent.

Ensuite, le raisonnement avec des produits scalaires permet de suggérer des pistes pour des méthodes simples de seuillage. Par exemple, il est souvent d'usage en analyse de données de centrer le nuage des individus en positionnant le barycentre au centre du repère. Cette translation n'a pas d'incidence sur les mesures de distances. Or ceci n'est pas le cas des similarités. Dans ce nouveau repère, les produits scalaires sont de signes différents et ceux-ci nous indiquent (par rapport au barycentre) si les vecteurs individus suivent une direction relativement similaire ou non. Une première approche de seuillage consisterait par exemple, à écrêter les produits scalaires de signes négatifs. De ce fait, la matrice devient davantage éparse et seuls les produits scalaires des vecteurs suivant les mêmes directions sont retenus.

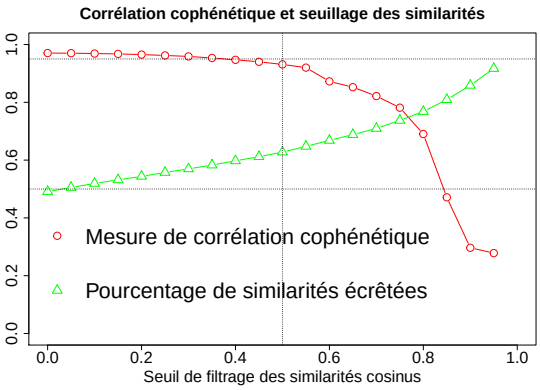

Figure 1: Résultats de la CAH avec seuillage des similarités sur le jeu de données Iris.

Nous proposons alors d'étudier, à partir de quelques exemples numériques, l'impact de ce type de seuillage sur les résultats de la CAH en mesurant de combien l'arbre binaire obtenu diffère de l'arbre binaire de référence qui est le résultat de la CAH obtenu sans seuillage.

Nous montrons dans la Figure 1 les résultats pour le jeu de données Iris avec une matrice de similarités basées sur la mesure cosinus des données centrées-réduites et une CAH basée sur la méthode "centroid". Nous faisons varier le seuil de filtrage de 0 à 0.95 avec un pas de 0.05. Nous utilisons le coefficient de corrélation cophénétique proposé par Sokal et Rohlf (1962) pour mesurer l'écart entre le dendrogramme de référence et celui obtenu avec une matrice de similarités éparses. Nous traçons également le pourcentage des produits scalaires de la matrice $\mathbf{S}$ qui sont écrêtées pour chaque valeur de seuil.

Sur cet exemple, si nous retenons uniquement les produits scalaires positifs (seuil nul) comme discuté précédemment, près de la moitié des mesures les plus faibles sont écrêtées alors que l'arbre binaire que l'on obtient a une corrélation cophénétique de plus de 0.95 avec l'arbre binaire de référence.

Bibliographie

[1] Lance, G. N., and W. T. Williams. (1967), A general theory of classificatory sorting strategies: 1. Hierarchical systems, *The Computer Journal*, 9, 373-380.
[2] Murtagh, F. (1983), A survey of recent advances in hierarchical clustering algorithms, *The Computer Journal*, 26(4), 354-359.
[3] Murtagh, F. and Contreras, P. (2012), Algorithms for clustering: an overview, *WIREs Data Mining Knowledge Discovery*, 2, 86-97.
[4] Sokal, R. R. and Rohlf, F. J. (1962), The comparison of dendrograms by objective methods. *Taxon*, 11, 33-40.