Hiérarchies pour la classification supervisée

Christophe Osswald et Arnaud Martin

Laboratoire E^3I^2, ENSIETA
2 rue François Verny 29806 Brest Cedex 9
Christophe.Osswald@ensieta.fr et Arnaud.Martin@ensieta.fr

RÉSUMÉ. *Les méthodes pour construire une hiérarchie sont nombreuses. Ici, nous explorons les fonctions de Lance et Williams pour déterminer des paramètres de construction d'une classification automatique adaptés à des données constituées d'imagettes sonar. Nous considérons que la qualité d'une hiérarchie pour nos données est liée au fait qu'elle contient des classes homogènes relativement à l'étiquetage existant. Lorsque nous utilisons ces paramètres pour construire une hiérarchie mêlant données étiquetées et non étiquetées, la même mesure de qualité de règle permet de proposer des étiquettes pour les données apparaissant dans les classes les plus homogènes.*

MOTS-CLÉS : *Hiérarchies, fonctions d'agrégation, mesure de qualité, caractérisation des fonds sous-marins.*

Introduction

La caractérisation des fonds sous-marins permet de constituer des cartes à l'usage des sédimentologues, de la navigation sous-marine autonome ou de la détection de pollution. La carte est composée de grandes images sonar (voir Martin *et al.* [MSL04]), découpée en 4003 imagettes de 64x384 pixels. La distinction entre deux types de sédiments (sable, rocher, cailloutis, ride, vase, ombre) est souvent malaisée, même pour un expert. Les types de sédiment sont présents de façon variée : le sable représente près de 55% des imagettes, les rochers 21%, les cailloutis moins de 1%. De plus, 40% des imagettes contiennent plus d'un type de sédiment : l'étiquette indique alors uniquement le sédiment occupant la plus grande surface sur l'imagette.

Ici, nous mettons en oeuvre une méthode de classification supervisée qui autorise un étiquetage multiple (plusieurs propositions sur une seule imagette) ou aucune étiquette. Ce type d'étiquetage devient particulièrement pertinent dans le cadre d'une fusion multi-capteur ultérieure.

Nous rappelons tout d'abord les mécanismes d'un algorithme de classification ascendante hiérarchique, ainsi que le formalisme de Lance et Williams pour les fonctions d'agrégation utilisées dans ce cadre. Nous donnons des résultats concernant les espaces de définition de ces fonctions, et présentons une famille de ces fonctions que nous utilisons pour couvrir les méthodes usuelles. A l'aide de la mesure de Jaccard, nous optimisons notre méthode de construction de hiérarchie dans une optique de classification supervisée.

1. Classification Ascendante Hiérarchique

Les algorithmes de classification ascendante hiérarchique (CAH) constituent des méthodes usuelles pour construire un système de classes à partir d'un ensemble d'objets sur lequel on peut évaluer une dissimilarité $d(x, y)$. L'algorithme construit une distance entre classes : $d(\{x\}, \{y\}) = d(x, y)$. A chaque étape il fusionne les deux classes les plus proches en une seule classe, et réévalue les dissimilarités entre cette nouvelle classe et les classes préexistantes. Les classes $C = A \cup B$ ainsi créées, munies de l'indice $d(A, B)$ forment une hiérarchie indicée ; la dissimilarité associée est une ultramétrique.

1.1. *Fonctions de Lance et Williams*

Lance et Williams [LW67] synthétisent de nombreuses méthodes d'évaluation de distance inter-classes par la formule :

$$d^p(C,D) = \alpha_A d^p(A,D) + \alpha_B d^p(B,D) + \beta d^p(A,B) + \gamma |d^p(A,D) - d^p(B,D)|$$

Chen [Che96] restreint les formes des termes α, β et γ à des fonctions ne dépendant que de paramètre issus des cardinaux des classes, afin de garantir que les valeurs obtenues amènent bien à une ultramétrique. La plupart des algorithmes usuels de CAH peuvent être définis par ces trois fonctions α, β et γ ainsi qu'un réel p.

$$r_A = \frac{|A|}{|A \cup B|} \quad r_B = \frac{|B|}{|A \cup B|} \quad r_D = \frac{|D|}{|A \cup B|} \quad p \neq 0$$
$$d^p(C,D) = \alpha(r_A, r_D) d^p(A,D) + \alpha(r_B, r_D) d^p(B,D) + \beta(r_A, r_B, r_D) d^p(A,B) + \gamma(r_D) |d^p(A,D) - d^p(B,D)|$$

Algorithme	$\alpha(u,w)$	$\beta(u,v,w)$	$\gamma(w)$	p
lien simple	$-1/2$	0	$1/2$	1
lien complet	$1/2$	0	$1/2$	1
méthode de Ward	$(u+w)/(1+w)$	$-w/(1+w)$	0	2

1.2. *Fonctions admissibles et internes*

Pour s'assurer qu'il n'y ait pas d'inversion lors de la CAH, *i.e.* qu'on n'ait pas $A \subsetneq B$ avec $f(A) > f(B)$, il faut que l'algorithme soit monotone (Dragut [Dra01]). On parle alors d'une fonction admissible :

i. $\alpha(u,w) + \alpha(1-u,w) + \beta(u,1-u,w) \geqslant 1$
ii. $\alpha(u,w) \geqslant 0$
iii. $\gamma(w) \geqslant \max\{-\alpha(u,w), -\alpha(1-u,w)\}$

Un algorithme de Lance et Williams conserve l'espace si :

$$\min\{d(A,D), d(B,D)\} \leqslant d(A \cup B, D) \leqslant \max\{d(A,D), d(B,D)\}$$

Les ultramétriques sont alors des points fixes. C'est le cas pour le lien simple et le lien complet, mais pas pour la méthode de Ward.

De nombreux algorithmes de CAH ne peuvent s'écrire comme des algorithmes de Lance et Williams. Notamment, toute fonction d'agrégation interne amène à un algorithme de CAH conservant l'espace. Osswald [Oss03] étudie une famille de telles fonctions qui utilisent les médianes.

2. Hiérarchies

Le lien simple est connu pour favoriser les hiérarchies déséquilibrées : dès que l'algorithme engendre une classe d'une taille conséquente, la plupart des agrégations ultérieures vont se faire avec cette classe. Cela est dû au fait que si x est un élément de A et y un élément de B, on a $d(x,y) \geqslant d(A,B)$ et $d(A,B)$ est souvent petit lorsque A et B sont de cardinal important. A l'inverse, pour le lien complet comme pour la méthode de Ward, on a $d(x,y) \leqslant d(A,B)$ et l'effet obtenu est inverse : les hiérarchies équilibrées sont favorisées par l'algorithme.

Ici, la mesure de distance entre nos imagettes est la distance euclidienne entre les vecteurs constitués par les paramètres extraits d'un filtrage de Gabor [MSL04]. Les hiérarchies obtenues sur un ensemble de 24 imagettes pour lesquelles tous les types de sédiments sont représentés à quatre reprises, avec éventuellement des frontières sont représentées en figure 1.

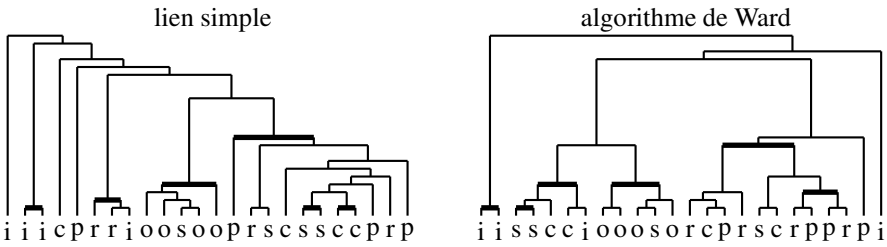

FIG.1. *Hiérarchies construites sur 24 imagettes avec frontières*

2.1. *Mesure de qualité d'une hiérarchie*

Nous considérons qu'une hiérarchie est d'autant plus fidèle à l'analyse de l'expert qu'elle contient des classes représentant au mieux les types de sédiments. L'intérêt d'une hiérarchie est donc défini par sa forme, *i.e.* l'ensemble des classes qu'elle contient. La valeur des indices des classes n'a donc pas d'intérêt pour cet objectif.

Pour chaque type de sédiment i, on considère la classe A de la hiérarchie qui « ressemble » le plus à l'ensemble M_i des imagettes de type i. Pour ce faire, nous utilisons la mesure de Jaccard $J(A \iff M_i)$, où $P(A)$ représente la proportion des éléments de A par rapport à l'ensemble tout entier, pour obtenir une mesure globale $q(H)$ de la qualité d'une hiérarchie. Elle est assez résistante au nombre d'objets considérés [TKS02] :

$$J(A \iff M_i) = \frac{P(A \cap M_i)}{P(A) + P(M_i) - P(A \cap M_i)}$$

$$q(H) = \prod_i \max_{A \in H} J(A \iff M_i)$$

2.2. *Paramètres pour les fonctions de Lance et Williams*

Nous plongeons les méthodes d'agrégation dans un espace constitué par quatre segments de l'espace des fonctions admissibles. Le tableau suivant définit ces quatre segments, le réel $x \in [0, 1]$ est le paramètre qui permet de parcourir chaque segment.

Algorithme	$\alpha(u, w)$	$\beta(u, v, w)$	$\gamma(w)$	p
simple à complet	$-1/2 + x$	0	$1/2$	1
complet à intermédiaire	$\frac{1-x}{2} + x\frac{u+w/2}{1+w}$	0	$\frac{x-1}{2}$	1
intermédiaire à Ward	$\frac{u+(1+x)w/2}{1+w}$	$\frac{-xw}{1+w}$	0	2
Ward à Ward-γ_1	$(u+w)/(1+w)$	$-w/(1+w)$	x	2

Le premier segment est composé de fonctions internes, qui conservent donc l'espace. Les deux suivants lient le lien complet à la méthode de Ward, et dilatent l'espace. Le dernier segment modifie l'algorithme de Ward en augmentant la valeur de $\gamma(w)$ et donc la propension de l'algorithme à construire une hiérarchie équilibrée. Il se conclut par la méthode Ward-γ_1, qui est une telle extension de la méthode de Ward

Une combinaison linéaire de la fonction d'agrégation du lien complet et de celle de la méthode de Ward n'est pas toujours admissible. Le lien intermédiaire permet de rester au sein des fonctions admissibles qui dilatent l'espace en liant ces deux algorithmes usuels. Le lien moyen est une méthode qui conserve l'espace, et s'il est possible de paramétrer le passage du lien simple au lien complet en passant par le lien moyen, les fonctions de Lance et Williams obtenues perdent largement en lisibilité par rapport à celles que nous utilisons. Dans la mesure où les résultats obtenus mettent en exergue l'intérêt des méthodes qui ne conservent pas l'espace, nous n'avons pas retenu le lien moyen parmi notre famille de fonctions d'agrégation.

La mesure de qualité ne tient compte que de la forme de la hiérarchie produite pour un paramètre x. Ainsi, pour un segment et un ensemble d'imagettes donnés, l'algorithme de Lance et Williams considéré produit une hiérarchie $H(x)$, et la fonction qui à x associe $q(H(x))$ est constante par morceaux. La figure 2 montre les moyennes de $q(H(x))$ pour 100 jeux de 47 imagettes.

Dans le cadre de cette étude, nous ne conservons que les imagettes ne contenant qu'un seul type de sédiment : cela permet de se limiter aux imagettes pour lesquelles les informations fournies par l'expert sont les plus fiables, et pour lesquelles le jeu de paramètres extraits est efficace.

Les différents types de sédiment sont présents en proportions très différentes. Afin d'analyser le comportement des méthodes de CAH vis-à-vis des classes de tailles variées, nous utilisons des échantillons dans lesquels le nombre d'imagette de type i est proportionnel à n_i^λ où n_i est le nombre d'imagettes de type i total, et où λ prend les valeurs 0.3, 0.5 et 0.7. La présence de classes naturelles de tailles déséquilibrées n'est donc en rien une gêne pour cette méthode, alors qu'elle amène souvent à des difficultés dans les applications de classification supervisée.

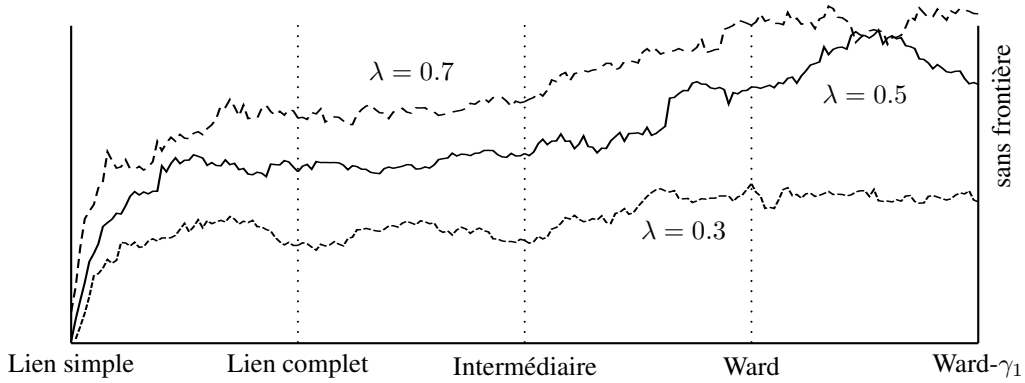

FIG.2. *Qualité des hiérarchies obtenues sur 47 imagettes*

3. Conclusion

Cette démarche nous permet d'associer classification supervisée et classification non-supervisée, en optimisant les paramètres de classification non-supervisée sur un sous-ensemble de données connu. Par la suite, la mesure de Jaccard nous permet d'identifier pour chaque type de sédiment une meilleure classe dans la hiérarchie. L'étiquetage ainsi réalisé n'est pas univoque, et les imagettes peuvent aisément recevoir zéro ou deux étiquettes. Nous obtenons une structure qui étend minimalement les partitionnements, et qui est assez simple à prendre en compte dans le cadre de la fusion de classifieurs.

Références

[Che96] Z. Chen. Space-conserving agglomerative algorithms. *Journal of Classification*, 13 :157–168, 1996.

[Dra01] A. Dragut. Characterization of a set of algorithms verifying the internal similarity. *Mathematical Reports*, 53(3-4) :225–232, 2001.

[LW67] G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies. *the Computer Journal*, 9(4) :373–380, 1967. and vol. 10, n°3, pp 271-277.

[MSL04] A. Martin, G. Sévellec, and I. Leblond. Characteristics vs decision fusion for sea-bottom characterization. In *colloque caractérisation in-situ des fonds sous-marins*, Brest, France, 2004.

[Oss03] C. Osswald. Robustesse aux variations de méthode pour la classification hiérarchique. In *XXXVèmes Journées de Statistiques*, Lyon, France, 2003.

[TKS02] P.-T. Tan, V. Kumar, and J. Srivastana. Selecting the right interestingness measure for association patterns. In *SIGKDD'02*, Edmonton, Canada, 2002.