Sur les différentes expressions formelles d'une hiérarchie binaire symétrique ou implicative

I.C. Lerman

IRISA, Campus Universitaire de Beaulieu, 35042 Rennes Cédex
lerman@irisa.fr

Mots clés : classification ascendante hiérarchique, arbre implicatif, hiérarchie binaire.

La Classification Ascendante Hiérarchique (CAH) est un outil puissant d'analyse des données. Elle a connu un développement fulgurant ces trente dernières années. Face à un tableau des données décrivant un ensemble $\mathcal{O}$ d'objets élémentaires (resp., $\mathcal{C}$ de catégories) par un ensemble $\mathcal{V}$ de variables (attributs ou descripteurs), la CAH permet l'organisation en classes et sous classes de proximité aussi bien de l'ensemble $\mathcal{O}$ des objets (resp., $\mathcal{C}$ de catégories) que de l'ensemble $\mathcal{V}$ des variables descriptives. L'analyse des données qui en résulte suppose la reconnaissance de classes ou sous-classes "significatives" issues de l'une ou de l'autre des des deux classifications hiérarchiques duales. Elle suppose également la mise en correspondance de ces deux dernières.

Entre parties disjointes de l'ensemble E à traiter, on établit un indice de dissimilarité δ dépendant de la description fournie par le tableau des données. δ a un caractère parfaitement symétrique par rapport aux deux arguments à comparer.

La notion intuitive de hiérarchie implicative est apparue au début des années 90 ([2]). Sa construction est en tout point analogue à celle d'une CAH ; mais à une différence fondamentale près que nous allons préciser. Si $\mathcal{A}$ désigne l'ensemble à organiser ($\mathcal{A}$ est généralement un ensemble d'attributs booléens), $\mathcal{A}$ est muni d'un indice de similarité orientée (un indice d'implication) σ qui est ensuite étendu à la comparaison de deux segments ordonnés et disjoints de $\mathcal{A}$.

L'organisation synthétique des résultats que la CAH procure est un arbre indicé de classifications qui s'emboîtent sur E. Chaque classification occupe un niveau de l'arbre. L'interprétation de l'expert repose pour l'essentiel sur une indexation de la suite croissante des niveaux de l'arbre au moyen d'une section commençante de la suite des entiers. Le principe de la construction de l'arbre est binaire : une classe apparaissant à un niveau donné, résulte de la fusion de deux sous classes filles qui sont déjà apparues à des niveaux inférieurs. Ces deux dernières sont repérées comme étant les plus proches au sens de l'indice δ. Il arrive souvent qu'on propose une indexation numérique des niveaux de l'arbre à partir d'un indice de stratification qui correspond à la valeur de δ pour les deux sous classes filles dont la jonction se produit au niveau concerné. Nous nous limitons quant à nous à une indexation ordinale parce que, pour une part, comme nous venons de le mentionner, c'est cette indexation qui joue le rôle le plus important dans l'interprétation de l'expert. D'autre part, cette forme ordinale d'indexation permet une plus grande clarté formelle dans le passage entre une hiérarchie binaire symétrique et un arbre implicatif.

Il existe différentes expressions formelles de la structure dégagée par une CAH classique : hiérarchie de parties (indicée ou non), arbre de classification (indicé ou non),

distance ultramétrique.

La formalisation proposée dans ([3]) et ([1]) de la structure dégagée par une hiérarchie implicative s'apparente à celle d'une hiérarchie binaire non indicée de parties. Dans ([4]) nous avons repris la formalisation avec un éclairage nouveau d'une façon argumentée et nourrie en faisant le parallèle le plus étroit possible avec la notion de classification hiérarchique binaire. Ainsi, on se rend clairement compte de ce qui change en passant du cas symétrique au cas orienté. C'est précisément l'objet de notre présentation.

À cette fin et de façon systématique le cas classique est repris en rappelant ou en précisant des définitions telles que celles de hiérarchie binaire (non indicée ou indicée), d'arbre de classification binaire (non indicé ou indicé) ou de distance ultramétrique. On supposera - comme évoqué ci-dessus - que l'indexation ou la valeur de la distance ultramétrique décrit une section commençante de l'intervalle $\mathbb{N}$ des entiers. Nous introduisons une formalisation - nouvelle à notre connaissance - en termes de hiérarchie de fourches binaires. En effet, une telle expression formelle est particulièrement adaptée au transfert entre la notion de hiérarchie binaire symétrique et la notion de hiérarchie implicative ; la notion de fourche binaire symétrique donnant une notion de fourche binaire orientée.

Toujours dans le cas symétrique, nous précisons les règles de passage entre les différentes expressions formelles. Ainsi, on a une équivalence de représentation entre une hiérarchie binaire (non indicée) de parties et une hiérarchie de fourches binaires symétriques. On a également une équivalence de représentation entre une hiérarchie binaire indicée, un arbre binaire de classification indicé et la distance ultramétrique associée. Cependant, la question se pose de faire correspondre de façon compatible, à une hiérarchie binaire non indicée de parties d'un ensemble E, que nous notons ici $\mathcal{H}_b(E)$, un arbre de classification binaire indicé ordinalement sur E. Il y a plusieurs solutions à ce problème. Elles sont énumérables. Nous proposons précisément un algorithme pour réaliser - de proche en proche - l'une quelconque des solutions possibles. Cet algorithme réalise un "accrochage" de ce que nous appelons les "chaînes complètes" de la hiérarchie binaire. Une telle chaîne complète est une suite orientée par inclusion de parties de $\mathcal{H}_b(E)$, telle qu'il n'existe pas de parties de $\mathcal{H}_b(E)$, strictement comprise entre deux parties consécutives et telle que le premier (resp., le dernier) élément est un singleton (resp., la partie pleine E).

Dans ces conditions, la transposition de la formalisation dans le cas implicatif ou orienté se déduit de façon claire et naturelle. Comme exprimé ci-dessus, ce qui se transporte le plus directement est la notion de hiérarchie de fourches binaires. Ces dernières sont symétriques dans le cas classique. Elles deviennent orientées dans le cas implicatif. De cette dernière (hiérarchie de fourches orientées) dérive un ordre total sur l'ensemble organisé que nous désignerons comme ci-dessus par $\mathcal{A}$. Le correspondant de la hiérarchie binaire de parties de E est une hiérarchie binaire d'intervalles de $\mathcal{A}$, lequel est totalement ordonné. La notion d'arbre binaire indicé (resp., non indicé), de classifications devient une notion d'arbre binaire indicé (resp., non indicé) orienté de classifications ; chaque jonction entre deux classes devient orientée de gauche à droite. La caractérisation d'une distance ultramétrique est maintenant relative à un triangle orienté de la forme (x, y, z) où le plus petit des vecteurs est soit $\vec{xy}$, soit $\vec{yz}$; les deux autres étant égaux. On a finalement le tableau de correspondance :

Cas symétrique	**Cas orienté**
Fourche binaire symétrique	Fourche binaire orientée
Hiérarchie de fourches binaires symétriques	Hiérarchie de fourches binaires orientées
Hiérarchie binaire indicée (resp.,non indicée) de parties	Hiérarchie binaire indicée (resp.,non indicée) d'intervalles
Chaînes complète de parties	Chaînes complète d'intervalles
Distance ultramétrique	Distance ultramétrique orientée

Nous donnons ci-dessous les graphiques de deux arbres binaires où le premier est symétrique et le second est orienté. Nous avons choisi les mêmes associations pour chacun des deux arbres. Cependant, elles sont orientées pour l'arbre implicatif.

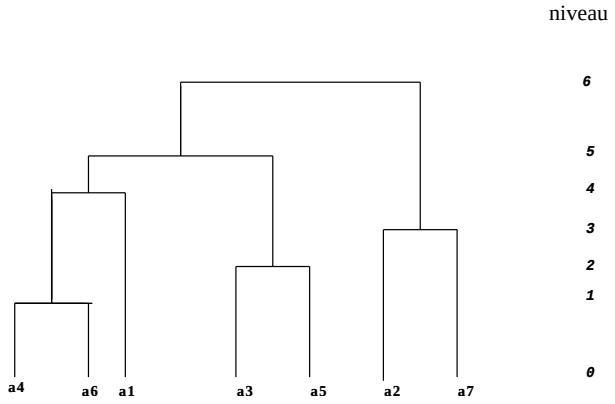

Figure 1: Arbre symétrique de classification binaire

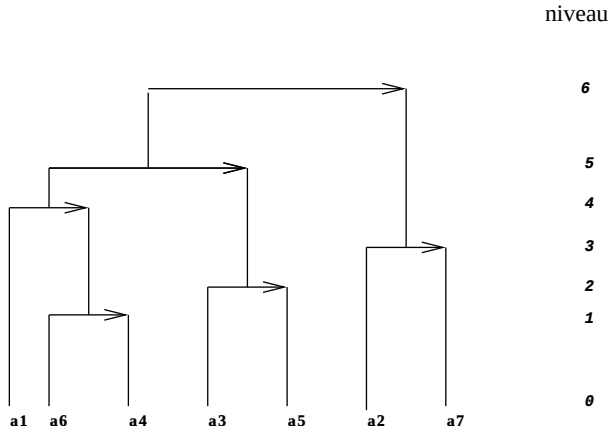

Figure 2: Arbre de classification binaire implicative

Remerciements

Nous remercions Pascale Kuntz (Professeur à l'École Polytechnique de l'Université de Nantes) pour l'échange que nous avons pu avoir autour de [4] et qui nous ont permis de préciser certains aspects.

Références

[1] R. Gras, P. Kuntz, "Discovering r-rules with a directed hierarchy", *Soft Computing* (5): March 2006, 453-460.

[2] R. Gras, A. Larher, "L'implication statistique, une nouvelle méthode d'analyse des données", *Mathématiques et Sciences Humaines* (120), 1993, 5-31

[3] P. Kuntz, "Classification hiérarchique orientée en ASI", In R. Gras, F. Spagnolo and J. David editors, *Troisièmes Rencontres Internationales - A.S.I. Analyse Statistique Implicative*, Università degli Studi di Palermo, 2005, 53-62

[4] I.C. Lerman, "Analyse logique, combinatoire et statistique de la construction d'une hiérarchie implicative ; niveaux et noeuds significatifs", *Publication Interne Irisa* n^0 1827, novembre 2006, 45 pages