MATHÉMATIQUES ET SCIENCES HUMAINES

I. C. LERMAN

H-classificabilité

Mathématiques et sciences humaines, tome 27 (1969), p. 21-28

<http://www.numdam.org/item?id=MSH_1969__27__21_0>

M.S.H. (7e année, n° 27, 1969, p. 21-28)

H-CLASSIFICABILITÉ

par

I. C. LERMAN[1]

RÉSUMÉ

Le but principal de la classification automatique est de découvrir, sur une population finie E, d'objets décrits au moyen d'un ensemble fini d'attributs, une hiérarchie de classifications emboitées respectant de manière satisfaisante les ressemblances entre objets ; c'est-à-dire telle que deux objets se trouvent réunis à un niveau d'autant plus élevé de la hiérarchie que leur ressemblance est grande. E est en général un échantillon d'un ensemble E' plus vaste ; on peut dans ces conditions se poser le problème de juger si au vu des données, il n'est pas artificiel *d'imposer sur E une structure de chaîne de partitions censée respecter les ressemblances entre objets. D'où la proposition de statistiques susceptibles de mesurer l'aptitude de la population étudiée à être organisée en une hiérarchie de classifications ; et, la définition d'un test éprouvant l'hypothèse de la non-existence d'une* h-*classification* « *naturelle* ». *La donnée de base que nous considérons est une préordonnance sur E ; c'est-à-dire un préordre total sur l'ensemble F des paires d'objets distincts de E.*

Notre propos est surtout de mettre l'accent sur un problème que semblent négliger ou ignorer les taxinomistes et de proposer dans le cadre de notre point de vue, une tentative de solution. Lorsque le taxinomiste se trouve devant différentes classifications, d'une même population d'objets, assez voisines entre elles et obtenues par différentes méthodes ; il ne peut s'empêcher de penser : « voilà une population « bien classifiable ». »

Sous le nom de *h*-classificabilité, nous désignerons l'aptitude d'une population E d'objets à être organisée en une hiérarchie de classifications respectant de manière satisfaisante les ressemblances entre objets (à être *h*-classifiée dirons-nous plus brièvement).

Une telle définition reste par trop vague, l'objet de ce qui suit sert à la préciser et à lui donner un aspect calcul.

1. — PRÉLIMINAIRES.

Relativement à une visée classificatoire, on suppose établi pour la description de l'ensemble fini E d'objets, ($|\,E\,| = n$), un ensemble fini A d'attributs; $A = \{ a_1, a_2, \ldots, a_j, \ldots, a_p \}$,($|\,A\,| = p$). Chaque objet x de E sera représenté par le sous-ensemble X, ($X \subset A$), des attributs qu'il possède. La représentation est ainsi définie par une application de E dans l'ensemble P (A) des parties de A. On peut d'une manière équivalente associer à chaque objet x le vecteur logique a (x) = ($x_1, x_2, \ldots, x_j, \ldots, x_p$) où x_j vaut 1 si l'objet x possède l'attribut a_j et 0 sinon, a (x) est un point du cube $\{0,1\}^p$. De la sorte E nous est donné comme un échantillon dans P (A) ou dans $\{ 0,1 \}^p$.

1. Centre de Mathématiques Appliquées et de Calcul de la Maison des Sciences de l'Homme.

Introduisons ici l'ensemble F des paires d'objets distincts de E (i.e. des parties à deux éléments de E) :

$$F = \{ \{ x, y \} \mid x \in E, y \in E, x \neq y \}.$$

1. — *La donnée de base.*

La donnée de base considérée est une préordonnance sur E ; c'est-à-dire un préordre total sur F, celui pour lequel $\{ x, y \} \leqslant \{ z, t \}$ si et seulement si les deux objets x et y se ressemblent davantage que z et t, quel que soit le couple $(\{ x, y \}, \{z, t \})$ du produit cartésien $F \times F$. Ce préordre peut-être fourni directement par le spécialiste qui comparerait les différentes paires d'objets. Il peut être associé au choix d'une distance ou d'un coefficient de dissimilarité de la façon suivante :

$$\forall (\{ x, y \}, \{ z, t \}) \in F \times F, \{ x, y \} \leqslant \{ z, t \} \Leftrightarrow d(x, y) \leqslant d(z, t).$$

Plus généralement la préordonnance sur E peut-être associée à un indicateur de ressemblance qui se présente sous la forme d'un couple ou d'un triplet d'indices de similarité.

Si nous insistons sur une telle donnée, c'est qu'elle est relativement stable ; nous voulons dire que lorsqu'on remplace un indice par un autre, la préordonnance associée à une similarité fluctue peu et cela d'autant moins que la variance du nombre d'attributs possédés par un même objet, dans E, est petite [cf. [3]].

Par ailleurs, nous avons pu montrer que cette donnée est la plus générale permettant d'obtenir des résultats féconds du point de vue de la classification automatique.

2. — *L'objet recherché.*

L'idéal de ce que semblent rechercher les taxinomistes est une hiérarchie de classifications emboîtées respectant de manière satisfaisante les ressemblances entre objets, c'est-à-dire telle que deux objets se trouvent réunis à un niveau d'autant plus élevé de la hiérarchie que leur ressemblance est grande. Mathématiquement cet objet est une chaîne C de partitions de moins en moins fines dans le treillis des partitions sur E, $\mathcal{P}(E)$.

$$C = (P_0, P_1, \ldots, P_k, P_{k+1}, \ldots).$$

P_0 est la partition la plus fine, celle pour laquelle chaque classe contient exactement un élément de E, les classes de P_{k+1} sont obtenues par réunion de classes de P_k. La longueur d'une chaîne (nombre de partitions distinctes) est inférieure ou égale à n, $(n = \mid E \mid)$; elle est égale à n (chaîne maximale) si et seulement si la partition P_{k+1} se déduit de la partition P_k par réunion d'exactement deux classes de P_k.

Munissons l'ensemble $E \times E$ de la fonction $d : E \times E \to \mathbf{N}$, définie comme suit : $d(x, y)$ est le plus petit entier k tel que x et y soient dans une même classe de P_k. $d(x, y)$ définit une distance sur E qui satisfait l'inégalité ultramétrique.

$$d(x, z) \leqslant \max [d(x, y), d(y, z)]$$

pour tout x, y et z de E. Inversement, si d est une distance ultramétrique donnée sur E, on peut lui associer canoniquement une chaîne de partitions sur E ; l'un quelconque des niveaux de la chaîne étant défini par la relation d'équivalence :

$x \equiv y$ si et seulement si $d(x, y) \leqslant \rho$ où ρ est un nombre réel positif. En faisant parcourir à ρ le demi-axe réel positif $(\mathbf{R} +)$, on obtiendra successivement les différentes partitions de la chaîne.

Nous avons ainsi une image du treillis des partitions $\mathcal{P}(E)$ dans l'ensemble des espaces ultramétriques (E, d). Une telle représentation ne nous satisfait pas encore, pour deux raisons. D'une part,

l'ensemble représenté et celui représentant ne sont pas équipotents (sur E, l'ensemble des chaînes de partitions est fini alors que l'ensemble des distances ultramétriques est infini continu) ; d'autre part, on aimerait que l'objet recherché soit de même nature que la donnée de base, ici une préordonnance sur E ; d'où la

2.1. — *Définition.*

Une préordonnance ω sur E est dite ultramétrique si la condition suivante est satisfaite :

$$(\forall\ x, y, \text{ et } z \in E),\ r(x, y) \leqslant j \text{ et } r(y, z) \leqslant j \Rightarrow r(x, z) \leqslant j$$

où j est un entier donné et $r(x, y)$ désigne le rang de la paire $\{ x, y \}$ pour ω.

Les termes de la définition ci-dessus se justifient par la

2.2. — *Proposition.*

La condition nécessaire et suffisante pour qu'une distance d, définie sur E, soit ultramétrique est que la préordonnance associée le soit.

Il suffit de se rappeler la définition d'une préordonnance associée à une distance, sur E, pour établir cette proposition.

Les différentes sections strictement commençantes d'une préordonnance ultramétrique ωu sont saturées ; nous voulons dire que si R est une telle section

$$\{ x, y \} \in R \text{ et } \{ y, z \} \in R \Rightarrow \{ x, z \} \in R.$$

Or la donnée d'une chaîne de partitions est équivalente à la donnée d'une suite croissante (au sens de l'inclusion) de parties de F saturées.

Si $R_1 \subset R_2 \subset \ldots \subset R_k \subset R_{k+1} = F$ définit une telle suite, R_i est l'ensemble des paires réunies par la i^{eme} partition de la chaîne, P_i ; P_i est plus fine que P_{i+1}. Une préordonnance ultramétrique est en fait la classe d'équivalence formée par l'ensemble des distances ultramétriques qui définissent une même chaîne de partitions.

Dans ce cadre, le problème de la recherche d'une hiérarchie de classification respectant au mieux les ressemblances entre objets se pose comme suit : « Déterminer la préordonnance ultramétrique ωu la plus « proche » de ω (préordonnance de base). La notion de proximité à laquelle, il semble, qu'on puisse accorder le plus de confiance est définie par la donnée d'une distance sur l'ensemble Ω des préordres totaux sur F.

$$(\forall\ \omega \text{ et } \omega \text{ de } \Omega),\ d(\omega, \omega) = \left| \text{gr.}(\omega)\ \Delta\ \text{gr.}(\omega) \right|$$

où le second membre désigne le cardinal de la différence symétrique des graphes de ω et ω dans $F \times F$.

D'un point de vue statistique, il s'agit là d'un problème d'estimation non paramétrique ; ce n'est point cette question qui nous préoccupe ici.

2. — H-CLASSIFICABILITÉ [1].

1. — *Introduction.*

Rappelons que l'ensemble A des attributs est ici fixé une fois pour toute, que E est éventuellement un échantillon d'une population E′ plus vaste et que la donnée de base considérée est une préordonnance ω sur E.

1. On trouvera dans [3] un traitement plus détaillé de cette question.

Le premier problème que devrait se poser le taxinomiste est celui de juger si, au vu des données (ici de ω), il n'est pas « artificiel » d'imposer sur E une structure de chaîne de partitions censée respecter les ressemblances entre objets. Expliquons nous sur ce point à l'aide d'une illustration :

« Si E est un ensemble de points pris « au hasard », selon une loi uniforme dans le plan euclidien, le problème de chercher une h-classification sur E, muni de la distance euclidienne ou de l'ordonnance associée, n'a aucun intérêt. La seule hiérarchie de classifications admissible dans ce cas est la plus grossière : celle à deux niveaux, dont le premier définit la partition la plus fine sur E et le second, la moins fine. » Le problème posé est celui de trouver un test assez puissant permettant d'éprouver l'hypothèse N de l'« indépendance des objets » relativement à leur description, soit de la non existence d'une h-classification « naturelle ». Nous préciserons plus loin cette hypothèse et le test. Mais faisons remarquer que la nature statistique du problème posé est très classique. En effet, par exemple, pour un ensemble de points observés dans le plan euclidien, il existe bien un test statistique qui éprouve l'hypothèse de l'indépendance des points au profit de celle de la régression linéaire. Un tel test serait à effectuer avant l'ajustement de l'ensemble des points par une droite.

Notons enfin que si (E, ω) est suffisamment h-classifiable, on peut s'attendre à ce qu'un algorithme donné sous optimal pour un critère, fournisse la préordonnance ultramétrique optimale, vis-à-vis de l'un quelconque des critères. Cette considération intuitive est renforcée par le fait qu'il arrive souvent que différentes méthodes fournissent des hiérarchies de classifications très voisines.

Nous commencerons par proposer une statistique susceptible de mesurer la classification de (E, ω). L'étude de la distribution de la statistique dans l'hypothèse N nous permettra de proposer un test qui éprouve cette hypothèse.

2. — *Écartement entre la structure de ω et celle d'une préordonnance ultramétrique.*

Une condition nécessaire et suffisante pour qu'une distance d, définie sur E, soit ultramétrique est que tout triangle soit isocèle, la base étant le plus petit des côtés. Cette proposition devient, relativement à la préordonnance ω : ω est ultramétrique, si et seulement si, pour tout triplet $\{ x, y, z \}$, (partie à trois éléments de E), pour lequel $\{ x, y \} \leqslant \{ y, z \} \leqslant \{ x, z \}$, on a encore $\{ x, z \} \leqslant \{ y, z \}$, c'est-à-dire $\{ x, z \}$ et $\{ y, z \}$ dans une même classe du préordre ω.

Soit dans ces conditions J l'ensemble des parties à trois éléments de E ; considérons l'application τ de J dans l'ensemble des intervalles ouverts de F, pour ω, qui, à chaque triplet $\{ x, y, z \}$ associe l'intervalle ouvert $]M(x, y, z), S(x, y, z)[$ où $M(x, y, z)$ désigne la paire médiane et $S(x, y, z)$ la paire de rang le plus élevé parmi les trois paires $\{ x, y \}$, $\{ y, z \}$ et $\{ z, x \}$. Une telle application peut être matérialisée par un tableau T dont les colonnes portent les éléments de F dans l'ordre où ils se présentent pour ω (ainsi chaque colonne porte le nom d'une paire et les différentes colonnes correspondant à une même classe du préordre ω sont regroupées et séparées de celles correspondantes à la classe suivante. Chaque ligne du tableau T représente un élément de J. Le tableau est à $n(n-1)/2$ colonnes et $n(n-1)(n-2)/6$ lignes, ($|E| = n$). Chaque intervalle du type $]M(x, y, z), S(x, y, z)[$ est porté par la ligne du tableau $\{x, y, z\}$.

Exemple.

Soit $E = \{a, b, c, d, e\}$ et ω la préordonnance sur E :

$$\{ a, d \} = \{ a, c \} < \{ a, e \} < \{ c, e \} < \{ b, d \} = \{ c, d \} < \{ b, c \} < \{ d, e \} < \{ a, b \} < \{ b, e \}$$

Pour une même ligne correspondante par exemple à $\{ x, y, z \}$, les trois points indiquent les trois paires $\{ x, y \}$, $\{ y, z \}$ et $\{ x, z \}$; la suite des croix l'intervalle $]M(x, y, z), S(x, y, z)[$.

Une mesure de l'« écartement » entre la structure de ω et celle d'une préordonnance ultramétrique va être définie à partir d'une mesure sur F, dépendant de ω.

TABLEAU T

J	F									
	(a, d)	(a, c)	(a, e)	(c, e)	(b, d)	(c, d)	(b, c)	(d, e)	(a, b)	(b, e)
(a, b, c)		.					.	X	.	
(a, b, d)	.				.		X	X	.	
(a, b, e)			.						.	.
(a, c, d)	.	.	X	X		.				
(a, c, e)		.	.	.						
(a, d, e)	.		.	X	X	X	X	.		
(b, c, d)					.	.	.			
(b, c, e)				.			.	X	X	.
(b, d, e)					.			.	X	.
(c, d, e)				.		.	X	.		

2.1. — *Statistiques mesurant l'écartement.*

Nous venons de dire qu'un triplet { x, y, z } pour lequel on a { x, y } ⩽ { y, z } ⩽ { x, z }, pour ω, est tel que l'intervalle] { y, z }, { x, z } [est vide si ω est ultramétrique. Vis-à-vis d'un tel triplet, la préordonnance ω est d'autant « moins ultramétrique » que le cardinal de] { y, z {, { x, z { [défini sur ω, est grand. Pour tenir compte de l'ensemble J de tous les triplets on peut adopter comme mesure de l' « écartement » entre la structure de ω et celle d'une préordonnance ultramétrique :

$$H(\omega) = \sum_{J} |] M(x, y, z), S(x, y, z) [|$$

$$(| J | = n(n-1)(n-2)/6 \quad \text{si} \quad | E | = n).$$

Référons-nous à la représentation de l'application τ par le tableau T. H (ω) est le nombre de croix dans le tableau T. H (ω) a été définie en lisant T ligne par ligne. Si nous regardons le tableau verticalement (colonne par colonne), H (ω) apparaît comme une mesure sur F, affectant à chaque élément p de F le cardinal m_p du sous-ensemble de J dont chaque élément { x, y, z } est tel que M (x, y, z) < p < S (x, y, z), pour ω. En d'autres termes m_p est le nombre de fois où la paire p intervient pour séparer strictement une paire M (x, y, z) d'une paire S (x, y, z).

D'où l'idée de caractériser la classificabilité de E par *la distribution des entiers* m_p, exactement par la suite décroissante des entiers m_p, $p \in F$; une telle suite sera notée D (ω).

Dans l'exemple ci-dessus : H (ω) = 13 et D (ω) = (3, 3, 2, 2, 1, 1, 1, 0, 0, ...).

Si ω est ultramétrique, H (ω) est nul et le vecteur D (ω) a toutes ses composantes nulles. H (ω) permet de préordonner totalement l'ensemble, $\mathscr{E}_n$, des couples (E, ω), où E est un ensemble de cardinal n et ω une préordonnance sur E. D (ω) fournit sur l'ensemble $\mathscr{E}_n$ un préordre seulement partiel, néanmoins plus sûr que le préordre total précédent ; défini de la façon suivante : (E, ω) ≤ (E', ω') (i.e. (E, ω) plus classifiable que (E', ω') dans $\mathscr{E}_n$) si et seulement si D (ω) ≤ D (ω') (i.e. di ≤ d'i pour tout i, où di (resp. d'i) est le ième élément de la suite D (ω) (resp. D' (ω')).)

Il est manifeste que le préordre total défini à partir de H (ω) est compatible avec ce préordre partiel :

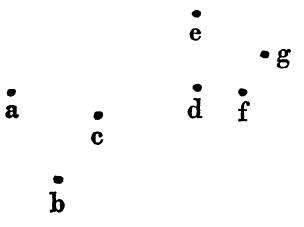
$$D(\omega) \leqslant D(\omega') \Rightarrow H(\omega) \leqslant H(\omega').$$

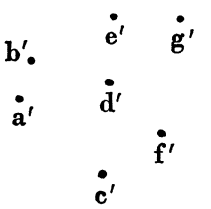
2.2.2. — *Illustration géométrique.*

Nous figurons ci-dessous deux ensembles E et E' de 7 points chacun, dans le plan euclidien.

$$E = \{ a, b, c, d, e, f, g \}, \qquad E' = \{ a', b', c', d', e', f', g' \}.$$

L'ensemble E (resp. E') est muni de la préordonnance ω (resp. ω') associée à la distance euclidienne définie sur E (resp. sur E'). Pour un même ensemble, le nombre de paires de points distincts est 21 et le nombre de triplets 35.





E E'

La préordonnance ω :

(d, f) = (f, g) < (b, c) < (d, e) = (d, g) < (e, g) < (e, f) < (a, c) < (c, d) < (a, b) < (b, d) < (c, e) = (c, f) < (b, f) = (c, g) < (a, d) < (b, e) < (a, e) < (b, g) < (a, f) < (a, g).

La préordonnance ω' :

(a, b) < (e, g) < (d, e) < (c, f) = (d, f) < (b, d) < (a, d) < (b, e) = (c, d) < (d, g) < (a, c) < (f, g) < (e, f) = (a, e) < (b, c) < (a, f) < (b, f) = (b, g) < (c, e) < (a, g) = (c, g).

Il semble, visuellement, que (E, ω) soit « mieux classifiable » que (E', ω'). En effet, on pourrait distinguer pour E, la classification ({ a, b, c }, { d, e, f, g }).

On a trouvé pour (E, ω) :

$$H(\omega) = 77 ; D(\omega) = (10, 10, 9, 7, 7, 7, 6, 6, 4, 4, 4, 1, 1, 1, 0, 0, \ldots)$$

et pour (E', ω') :

$$H(\omega') = 128 ; D(\omega') = (14, 13, 13, 12, 11, 11, 9, 9, 9, 7, 5, 5, 5, 3, 1, 1, 0, 0, \ldots).$$

On a bien :

$$D(\omega) < D(\omega').$$

3. — *Test de l'hypothèse N de la non existence d'une h-classification « naturelle ».*

Nous allons commencer par préciser une telle hypothèse. Rappelons que E nous est donné comme un échantillon dans P (A), ensemble des parties de l'ensemble des attributs. Sur P (A) les objets d'une population « bien classifiable » apparaissent comme des « îlots sporadiques ». Il semble par conséquent naturel d'exiger que, pour l'hypothèse N, E corresponde à un échantillon de n éléments aléatoires indépendants dans P (A) muni d'une mesure adéquate. Notre première réaction a été de vouloir adopter sur P (A) une mesure uniforme ; cependant si, par exemple, le nombre d'attributs possédés par un même objet de E était invariable, soit a, il est plus cohérent de concentrer toute la masse sur le niveau du simplexe P (A) correspondant à a. Par conséquent, soit la partition de E selon le nombre d'attributs possédés par un même objet : $E = E_1 + E_2 + \ldots + E_j + \ldots + E_k$, où tout objet de E_j possède exactement a_j attributs. Si n_j est le cardinal de E_j, $(n_1 + n_2 + \ldots + n_k = n)$, affectons au niveau du simplexe P (A), formé des parties de A ayant a_j éléments, la masse $\frac{n_j}{n}$ que nous répartissons d'une manière uniforme sur les différents éléments de ce niveau. L'hypothèse N ainsi exprimée semble assez bien refléter la non classificabilité. En effet, reportons nous au tableau d'incidence des données Objets X Attributs (E x A), à n lignes et p colonnes et considérons l'ensemble $\mathcal{C}$ des tableaux n x p formés de 0 et de 1 pour lesquels le nombre de 1 de la ligne i est celui de la ligne correspondante du tableau E x A ; l'hypothèse N consiste à considérer notre tableau des données comme issu de $\mathcal{C}$ muni du modèle uniforme.

On peut effectuer le test sur la base de la statistique H (ω) ou de la distribution D (ω). Considérons pour cela un échantillon E′ de n éléments aléatoires indépendants (tirage avec remise), dans P (A) muni de la mesure de probabilité ci-dessus définie, et établissons sur E′ la préordonnance ω′ de la même façon que l'a été ω sur E (avec le même indicateur de ressemblance). Dans la mesure où on connaît la distribution de H (ω) sur l'ensemble des échantillons de type E′, on pourra effectuer le test au vu de la valeur de H (ω). Sinon, E′ simulera l'hypothèse N et le test est défini à partir de E′.

Il semble plus satisfaisant d'effectuer le test en comparant les deux distributions D (ω) et D (ω′) où D (ω′) est associé à l'échantillon E′ muni de ω′, défini ci-dessus. Dans la mesure où D (ω) est à gauche de D (ω′), (D (ω) < D (ω′) ; on conclura à la classificabilité de la population étudiée.

Si n est grand, c'est sur des simulations de D (ω) et D (ω′) qu'on basera le test. Il reste que les calculs sont assez importants et dépendent d'une simulation par échantillonnage ; d'où l'intérêt de connaître dans l'hypothèse N la distribution théorique des valeurs m_p, $p \in F$.

Désignons par M la variable $M_p / |J|$. Nous avons déterminé la distribution théorique de M dans l'hypothèse N, dans un cas particulier important [cf. 3].

Ce cas particulier est celui où toute la masse de la mesure de probabilité, définie ci-dessus, est concentrée sur un même niveau du simplexe P (A) ; il correspond à la situation où le nombre d'attributs possédés par un même objet est invariable dans la population étudiée. Ce cas est très important car la préordonnance associée est un invariant quel que soit l'indicateur de ressemblance (fonction de s, u et v) choisi pour la définir. Le spécialiste peut se ramener à ce cas dans la mesure où il peut se contenter d'une famille de hiérarchies de classifications, chacune établie pour la sous-population d'objets ayant un nombre d'attributs fixé ; cela est d'autant plus aisé que le nombre d'attributs possédés par un même objet aura une faible variance dans la population étudiée.

BIBLIOGRAPHIE

[1] J. P. Benzecri, *Reconnaissance des formes et classification automatique*, Cours I.S.U.P. (1968-1969), Paris.

[2] B. JAULIN, « Mesure de la ressemblance en archéologie ». *Colloque International, C.N.R.S. Archéologie et Calculateurs.* Marseille, 7-12 avril 1969.

[3] I. C. LERMAN, *Les bases de la classification automatique*, Gauthier-Villars, Collection Programmation, Paris, 1970.

[4] S. RÉGNIER, « Non fécondité du modèle statistique général dans la classification automatique ». *Colloque International, C.N.R.S. Archéologie et Calculateurs.* Marseille, 7-12 avril 1969.

[5] M. ROUX, *Algorithme pour construire une hiérarchie particulière.* Thèse 3e cycle, déc. 1968, I.S.U.P., Paris.

[6] W. F. de la VEGA, « Techniques de classification automatique utilisant un indice de ressemblance », *Revue Française de Sociologie*, déc. 1967.