MATHÉMATIQUES ET SCIENCES HUMAINES

I. C. LERMAN

Sur l'analyse des données préalable à une classification automatique (proposition d'une nouvelle mesure de similarité)

Mathématiques et sciences humaines, tome 32 (1970), p. 5-15

<http://www.numdam.org/item?id=MSH_1970__32__5_0>

M.S.H. (8e année, no 32, 1970, p. 5-15)

SUR L'ANALYSE DES DONNÉES PRÉALABLE A UNE CLASSIFICATION AUTOMATIQUE

(Proposition d'une nouvelle mesure de similarité)

par

I. C. LERMAN *

Une idée qui permet de préciser la notion de classe polythétique introduite par Beckner (1959) (cf. [1]), nous autorise ici à intervenir sur deux questions importantes et liées à la Taxinomie: le choix des attributs de description et le choix de la mesure de similarité.

I. REPRÉSENTATION DES DONNÉES

Relativement à une visée classificatoire, on suppose établi pour la description d'une population finie E d'objets ($| E | = n$), un ensemble fini A d'attributs:

$$A = \{ a_1, a_2, \ldots, a_i, \ldots, a_p \}, (| A | = p).$$

On retient pour un objet donné x sa description; c'est-à-dire le sous-ensemble X ($X \subset A$) des attributs qu'il possède. La représentation est ainsi définie au moyen d'une application de E dans l'ensemble $\mathscr{P}(A)$ des parties de A. De manière équivalente, on peut associer à chaque objet x le vecteur logique:

$$a(x) = (x_1, x_2, \ldots, x_i, \ldots, x_p)$$

où x_i est égal à 1 si l'objet x possède l'attribut a_i et 0 sinon; $a(x)$ est un point du cube $\{ 0, 1 \}^p$. De la sorte, E nous est transmis comme un échantillon dans $\mathscr{P}(A)$ ou dans $\{ 0, 1 \}^p$. Cette information est généralement consignée dans un « tableau de données » qui est une matrice d'incidence:

$$(\in_{ij}),\ i = 1, 2, \ldots, p \text{ et } j = 1, 2, \ldots, n;$$

où $\in_{ij} = 1$ si l'attribut a_i est présent chez l'objet codé j et 0 sinon. Ainsi, chaque attribut est représenté par une ligne de la matrice et chaque objet par une colonne.

* Centre de Mathématiques Appliquées et de Calcul, Maison des Sciences de l'Homme.

Définition

Par rapport à un même attribut a_i, deux objets x et y sont dits avoir une association positive (resp. négative) si a_i est présent (resp. absent) simultanément chez les deux objets; c'est-à-dire:

$$x_i = 1, y_i = 1 \text{ (resp. } x_i = 0, y_i = 0).$$

II. LES HYPOTHÈSES INITIALES

L'hypothèse fondamentale du spécialiste est que la population qu'il étudie a une aptitude suffisante à être organisée en une hiérarchie de classifications emboîtées de moins en moins fines qui respecte 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 que leur similarité est grande.

La ressemblance entre deux objets donnés sera perçue à partir des attributs de description que nous supposons établis de telle façon que seule une association positive contribue à la mesure de leur similarité. Cette circonstance correspond d'ailleurs à la situation la plus fréquente. Si par exemple, les différents caractères de la population étaient bivalents, les deux modalités d'un même caractère étant telles que:

a) également significatives de la ressemblance; ou bien,

b) l'une des deux modalités est significative de la ressemblance alors que l'autre ne l'est pas, on définira l'ensemble A des attributs en retenant pour chacun des caractères la ou les deux modalités significatives. De cette manière, la prise en compte des associations négatives peut être négligée.

III. LA NOTION DE CLASSE

Selon Beckner, une classe polythétique G d'une classification « naturelle » se réfère à un sous-ensemble B d'attributs tel que:

α) chaque élément de la classe possède une proportion importante (mais non fixée) d'attributs de B;

β) chaque attribut de B est présent chez une proportion importante (mais non fixée) d'éléments de G;

γ) il n'y a pas nécessairement un attribut de B qui soit possédé par tous les éléments de G.

Restreignant notre attention aux paires d'objets de G ou aux paires d'attributs de B, nous pouvons substituer à cette définition, la suivante:

α_1) deux objets donnés de la classe G possèdent simultanément une proportion importante d'attributs de B;

β_1) deux attributs donnés de B sont simultanément présents chez une proportion importante d'objets de G.

$(E_1, E_2, \ldots, E_k)$ est la partition de E en k classes qui définit la classification la plus significative (cf. [4] ch. 2 § VI. A.4.4. et cf. [5]), celle « naturelle » que vise Beckner; puisqu'à chacune des classes E_i

est associé un sous-ensemble A_i, des attributs auquel elle se réfère, à la famille des classes

$$\{ E_i \mid i = 1, 2, \ldots, k \}$$

correspond bijectivement une famille $\{ A_i \mid i = 1, 2, \ldots, k \}$ des parties de A.

Un même attribut étant plutôt spécifique d'une seule classe, on peut imposer à cette dernière famille d'être une partition. De plus, la fréquence des uns dans la restriction de la matrice à $\bigcup_{1 \leqslant i \leqslant k} (A_i \times E_i)$ (resp. au complémentaire de $\bigcup_{1 \leqslant i \leqslant k} (A_i \times E_i)$ est significativement grande (resp. petite).

Introduisons ici relativement à deux objets x et y le nombre $s = S(x, y)$ des attributs qu'ils possèdent en commun:

$$s = \sum_{1 \leqslant i \leqslant p} x_i \, y_i = \mid X \cap Y \mid$$

où X (resp. Y) est le sous-ensemble d'attributs possédés par x (resp. y).

De même, relativement à deux attributs a_l et a_m, définissons le nombre $\sigma = \Sigma (a_l, a_m)$ des objets possédant simultanément les deux attributs, $\sigma = \mid G \cap H \mid$ où G (resp. H) est l'ensemble des objets où a_l (resp. a_m) est présent.

Si l'on omet de préciser la classe et l'ensemble des attributs auquel cette classe se réfère, les conditions α_1 et β_1 deviennent:

α_2) deux objets d'une même classe (resp. de deux classes distinctes) ont une valeur de s « relativement » grande (resp. petite);

β_2) deux attributs d'une même classe d'attributs (resp. de deux classes distinctes) ont une valeur de σ « relativement » grande (resp. petite).

Il nous reste à donner un sens plus précis à l'adverbe « relativement » qu'on retrouve dans chacune des assertions ci-dessus. Ayant observé une valeur s associée à deux objets x et y ($s = S(x, y)$), comment juger si par exemple, une telle valeur est assez grande ?

X et Y sont les deux parties de l'ensemble A que définissent les deux objets x et y pour lesquelles on a: $\mid X \mid = l_x$ et $\mid Y \mid = l_y$; considérons comme nous l'avions fait dans l'étude de la Classificabilité (cf. [4] ch. IV § III), l'hypothèse N où X (resp. Y) serait pris dans l'ensemble des parties de A à l_x (resp. l_y) éléments, chacun de ces deux ensembles étant muni d'une probabilité uniformément répartie. La manière la plus objective pour répondre à la question posée est d'étudier dans l'hypothèse N la *vraisemblance* d'une valeur aussi grande que s; c'est-à-dire:

$$Pr\,[\mid X \cap Y \mid \geqslant s].$$

La valeur de s sera considérée d'autant plus grande que cette probabilité est plus petite; ou, ce qui revient au même, les deux objets x et y seront jugés d'autant plus proches que:

$$Pr\,[\mid X \cap Y \mid < s] = P(x, y)$$

est plus grande.

Dualement, on se placera dans l'ensemble des parties de E pour juger de la *vraisemblance* d'une valeur observée $\Sigma (a_l, a_m)$ aussi grande que σ et on désignera par $\Pi (a_l, a_m)$ la probabilité

$$Pr\,[\mid G \cap H \mid < \sigma]$$

calculée dans une hypothèse M duale de N.

En remplaçant l'adverbe « relativement » par $P(x, y)$ dans α_2) et par $\Pi(a_l, a_m)$ dans β_2) on aura achevé de donner un sens plus précis à ces énoncés.

Du point de vue calcul, n et p sont généralement assez grand pour admettre de manière sûre une distribution binomiale pour $|X \cap Y|$ (resp. $|G \cap H|$) dans l'hypothèse N (resp. M) de paramètres p et $u = |X| \times |Y|/p^2$ (resp. n et $v = |G| \times |H|/n^2$). En posant $\lambda = p \,.\, u$ et $\mu = n \,.\, v$, dans la mesure où λ (resp. μ) est trop petit vis-à-vis de p (resp. n) on adoptera pour la loi de la statistique $|X \cap Y|$ (resp. $|G \cap H|$) une approximation de type Poisson, sinon, une approximation par la loi normale.

Ces remarques nous conduisent à aborder le problème du choix des attributs.

IV. CHOIX DES ATTRIBUTS

Définition

Soit $(E_1, E_2, \ldots, E_i, \ldots, E_k)$ une partition de E. Si a est un attribut donné, désignons par f_i la proportion des objets de la classe E_i qui possèdent l'attribut a. $(f_1, f_2, \ldots, f_k)$ définira la distribution de la fréquence relative de présence de a sur les différentes classes.

L'attribut a discrimine d'autant mieux la classification que la dispersion de l'ensemble des valeurs f_i est plus grande par rapport à la dispersion de la fréquence de présence de a dans E.

Nous appellerons *signification* de l'attribut a par rapport à la classification $(E_1, E_2, \ldots, E_k)$ la quantité:

$$\boxed{\frac{1}{(k-1)} \sum_{i=1}^{k} (f_i - \bar{f})^2 \Big/ \bar{f}\,\frac{(1-\bar{f})}{n}} \qquad (1)$$

où:

$$\bar{f} = \frac{1}{k} \sum_{i=1}^{k} f_i;$$

rapport de l'estimation de la variance inter-classe sur la variance globale.

Il en résulte une pondération sur l'ensemble A des attributs où à chaque attribut a_i $(i = 1, 2, \ldots, p)$ est attaché un coefficient tel que (1).

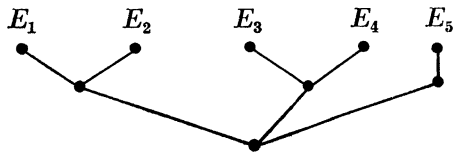
Si le choix des objets constituant la population E peut s'imposer de manière plus ou moins évidente au spécialiste, il en est tout autrement du choix des attributs de description. A ce sujet, deux questions se posent a priori au taxinomiste.

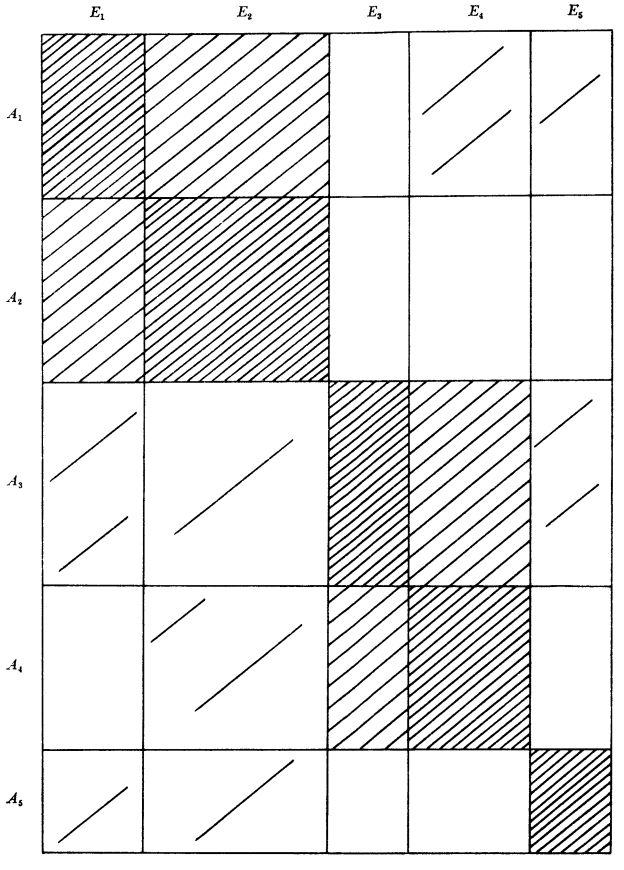
a) Quels sont les attributs pertinents vis-à-vis du problème étudié ?

b) Quelle importance convient-il d'accorder à chacun d'eux pour définir au mieux la classification ?

Selon notre point de vue, *a*) est un vrai problème alors que *b*) en est un faux. Pour nous justifier, supposons découverte, la classification recherchée et illustrons cette solution par le tableau d'incidence

ci-joint où la densité des hachures représente la densité des uns. La hiérarchie de classifications correspondante étant:





La première raison qui nous fait penser que *b*) est une fausse question, est que l'importance d'un attribut ne peut être définie dans l'absolu; elle est établie relativement à une classification comme nous l'avons exprimé ci-dessus [cf. (1)]. Par conséquent, il nous faudra connaître à l'avance la classification visée pour définir la bonne pondération. D'ailleurs la connaissance de cette pondération diminue très sensiblement l'intérêt d'une classification automatique.

Si un attribut est important, c'est-à-dire s'il est assez caractéristique d'une classe G relative à une classification « naturelle », sa présence chez un objet impliquera, en général, la présence de la plupart des attributs de la classe B des attributs, à laquelle G se réfère; ainsi que l'absence chez cet objet de la plupart des attributs du complémentaire de B dans A. Donc on peut s'attendre à ce que l'importance d'un attribut donné apparaisse dans une classification basée sur l'étude des ressemblances entre objets. Bien plus, la prise en compte d'une bonne pondération des attributs dans l'établissement d'une mesure de similarité accroît artificiellement l'importance propre de certains attributs. Enfin, dans cette hiérarchie de classifications que recherche la taxinomiste, un même attribut peut être plus ou moins discriminant selon les différents niveaux.

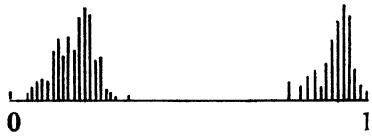
En établissant ses attributs de description, le chercheur en sciences humaines se rend tout à fait compte du caractère crucial du problème *a*). On ignore souvent si un attribut donné interviendra dans la formation des classes. Il en résulte un alourdissement du tableau des données par un accroissement du nombre d'attributs et, ce qui est plus grave, la présence d'attributs neutres qui perturbent la nature classifiable de la population. Les considérations du paragraphe précédent vont nous permettre de procéder au nettoyage de la matrice d'incidence des données.

V. NETTOYAGE DE LA MATRICE D'INCIDENCE DES DONNÉES

Supposons le problème résolu et adoptons pour illustrer cette solution le tableau d'incidence ci-dessus. Si $\{ b, c \}$ est une paire donnée d'attributs, on constate soit une nette proximité lorsque b et c sont relatifs à une même classe; soit une franche opposition lorsque les deux attributs sont relatifs à deux classes éloignées; donc une valeur de $\Pi\ (b, c)$ soit trop grande soit trop petite [cf. § III ci-dessus pour la définition de $\Pi\ (b, c)$]. Dans ces conditions, considérons pour un attribut donné a l'ensemble des paires d'attributs dont l'une des composantes est a:

$$Aa = \big\{ \{ a, c \} \,\big|\, c \in A - \{ a \} \big\}; \quad | Aa | = p - 1$$

et examinons l'allure de la distribution $\Pi\ (a)$ des valeurs de $\Pi\ (a, c)$ pour c parcourant $A - \{ a \}$. Pour cela, on portera sur un axe horizontal du plan l'intervalle [0, 1] des valeurs possibles de $\Pi\ (a, c)$ et sur un axe vertical le nombre d'éléments c pour lesquels on aura observé une valeur donnée de $\Pi\ (a, c)$. Une telle distribution sera portée vers les extrémités de l'intervalle [0, 1] comme essaie de la suggérer le diagramme suivant:



Si par exemple, il s'introduit dans notre tableau d'incidence, un attribut α neutre pour la classification, on devra généralement s'attendre à ce que la distribution $\Pi\ (\alpha)$ associée soit plus uniformément répartie entre 0 et 1 que celle $\Pi\ (a)$ attachée à un attribut pertinent pour la classification.

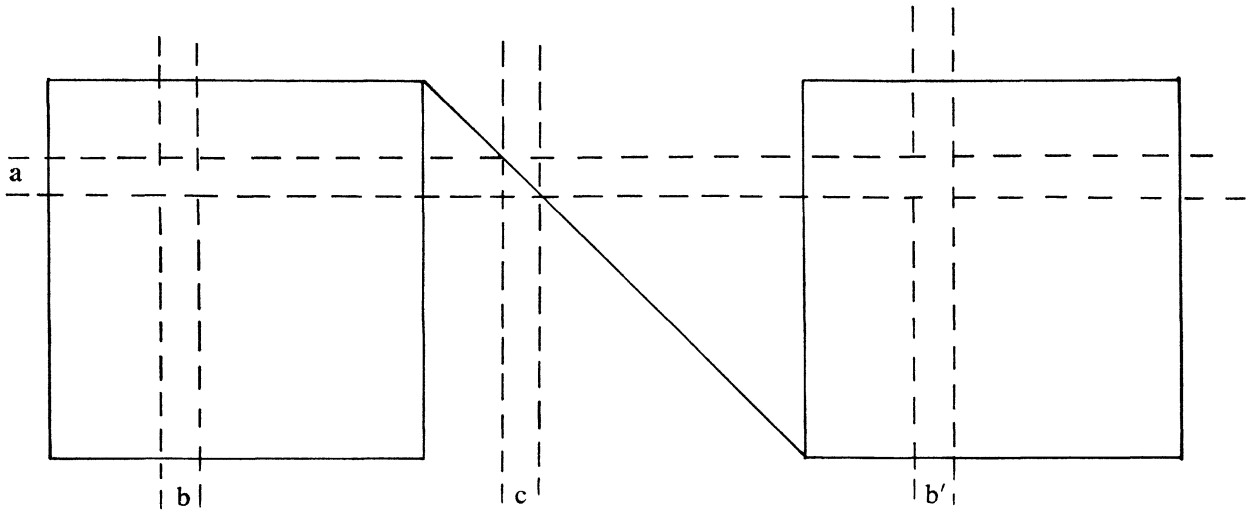
Par conséquent, dans un tableau correspondant à un cas réel, en étudiant pour chacun des attributs a la distribution $\Pi\ (a)$, il nous sera possible de détecter ceux des attributs pour lesquels $\Pi\ (a)$ n'est pas suffisamment *dispersée* (et ce, au moyen d'un coefficient de dispersion) et de les éjecter de notre étude.

On pourra dualement, par une technique analogue, nettoyer les colonnes du tableau d'incidence, on éliminera ainsi les objets les moins « typés ». On se trouvera finalement devant des données bien classifiables.

Exemple géométrique

Considérons la figure formée par deux surfaces carrées se déduisant l'une de l'autre par translation horizontale et reliées par une très mince bande oblique comme l'indique le dessin suivant. Rapportons

le plan de la figure à deux axes parallèles aux côtés de l'un des carrés et définissons une grille par un pavage du plan en petits carrés; a est une tranche horizontale de la grille, b, c et b' sont des tranches verticales. L'ensemble des attributs sera défini à partir de l'ensemble des tranches horizontales ou verticales. Si h (resp. k) est le nombre de tranches horizontales (resp. verticales), la description d'un point de la surface étudiée se fera au moyen d'un vecteur logique à $(k + h)$ composantes indexé sur l'ensemble des tranches où le 1 exprime pour le point son appartenance à une tranche.



On montrera que la technique précédente permet d'éliminer tous les attributs relatifs à des tranches telles que c.

On se rend compte que cette purification des données permet d'éviter l'effet de chaînage qu'on redoute dans l'application de l'algorithme « lexicographique » (cf. [4] ch. III § III).

VI. MESURE DE SIMILARITÉ ET PRÉORDONNANCE ASSOCIÉE

La mesure de similarité qui s'impose après l'analyse effectué au paragraphe III peut être définie par une application de l'ensemble F des paires d'objets distincts de E dans l'intervalle [0, 1] qui à chaque paire $\{ x, y \}$ affecte le nombre $P(x, y)$ qui, rappelons-le, est la probabilité:

$$Pr [| X \cap Y | < s \}$$

calculée dans l'hypothèse N (cf. § III), ou s est la valeur observée de l'indice $S(x, y)$.

Pour calculer effectivement $P(x, y)$ on aura à se référer soit à une table de la fonction de répartition d'une loi de Poisson, soit à celle d'une loi normale; on sait que c'est chose facile que d'introduire ces deux tables dans la mémoire d'un ordinateur. Nous allons tenter de faire sentir le progrès que représente cet indice par rapport à ceux déjà connus. Reprenons pour cela notre point de vue qui trouve son origine dans les travaux de R. N. Shepard et de J. P. Bénzecri (cf. [2] et [6]) où on ne retient comme information relative à la ressemblance des objets, qu'une préordonnance. Rappelons que cette donnée est un préordre total sur l'ensemble F des paires d'objets distincts de E pour lequel une paire p précède une paire q si les deux objets composant p se ressemblent davantage que ceux composant q.

Cherchant à synthétiser l'ensemble des indices de similarité proposés et à étudier l'influence du choix d'un indice sur la préordonnance associée; nous avons introduit (cf. [4] chap. I) relativement à deux objets x et y, en même temps que le paramètre s, les paramètres suivants:

t: cardinal du sous-ensemble des attributs non possédés par aucun des deux objets;

u (resp. v) cardinal du sous-ensemble des attributs possédés par l'objet x (resp. y) et non possédés par y (resp. x),

et nous avons défini une mesure de similarité comme une fonction réelle positive S définie sur l'ensemble $E \times E$ qui se présente sous la forme $(x, y) \to S(x, y) = \mathscr{S}(s, u, v)$ où la fonction $\mathscr{S}$ (s, u, v), définie sur le sous-ensemble de $\mathbf{N}^3$, $\{ s, u, v) \mid s + u + v \leqslant p \}$, est croissante par rapport à s, symétrique en u et v et décroissante par rapport à u; la croissance par rapport à s ou la décroissance par rapport à u étant stricte.

Nous avons montré que si le nombre d'attributs possédés par un même objet était invariable dans E, tous les indices de similarité étaient équivalents. Dans le cas où deux indices S et S' n'étaient pas équivalents, nous avons exprimé l'écart entre les deux préordonnances respectivement associées, $w(S)$ et $w(S')$, par le nombre d'inversions que présente $w(S')$ par rapport à $w(S)$. Parmi les indices de similarité qui se présentaient sous la forme $\mathscr{S}(s, u + v)$, $S(x, y) = s$ et $S'(x, y) = s + t$ étaient les deux pour lesquels les préordonnances respectivement associées étaient les plus écartées.

Si la variance du nombre d'attributs possédés par un même objet était petite, le nombre d'inversions que présente $w(s + t)$ par rapport à $w(s)$ était également petit.

Ces résultats étaient décisifs dans la pratique, lorsqu'on avait à traiter des questionnaires ou certains codes descriptifs d'objets pour lesquels le nombre d'attributs possédés par un même objet, dans la population étudiée, était sinon invariable, du moins de faible variance. Cependant le problème restait entier lorsque cette variance n'était pas négligeable. Un calcul théorique nous a permis de nous en rendre compte ainsi qu'un exemple concret qui portait sur « les caractéristiques des personnages-enfants à travers les contes d'enfants». Dans le cadre de cet exemple, P. Achard [cf. 3], cherchant à neutraliser dans la statistique s les effets de taille (nombre d'attributs possédés) qui rendaient trop ressemblants les objets de grosse taille, nous avait proposé de centrer et de réduire s en se référant à l'hypothèse N ; c'est-à-dire avec les notations du paragraphe III, d'adopter comme indice :

$$(s - pu) / \sqrt{pu(1 - u)}. \tag{1}$$

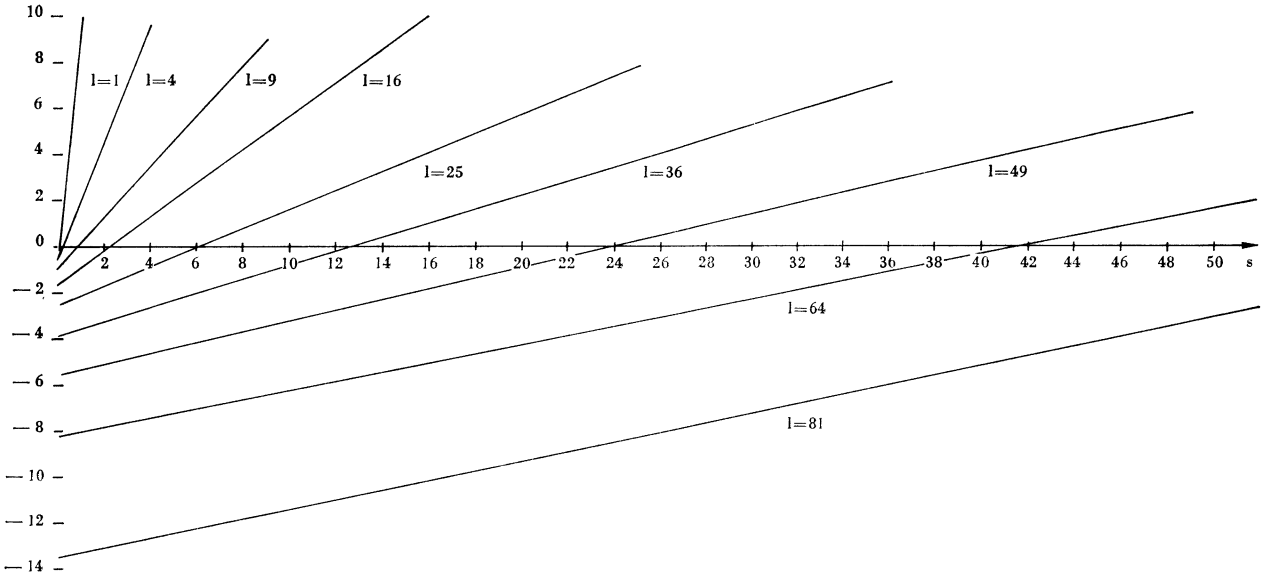
Toutefois simulant l'hypothèse N, nous avons remarqué une tendance de cet indice de similarité à rendre trop proches les objets de petite taille; remarque que nous allons confirmer par un calcul et un graphique.

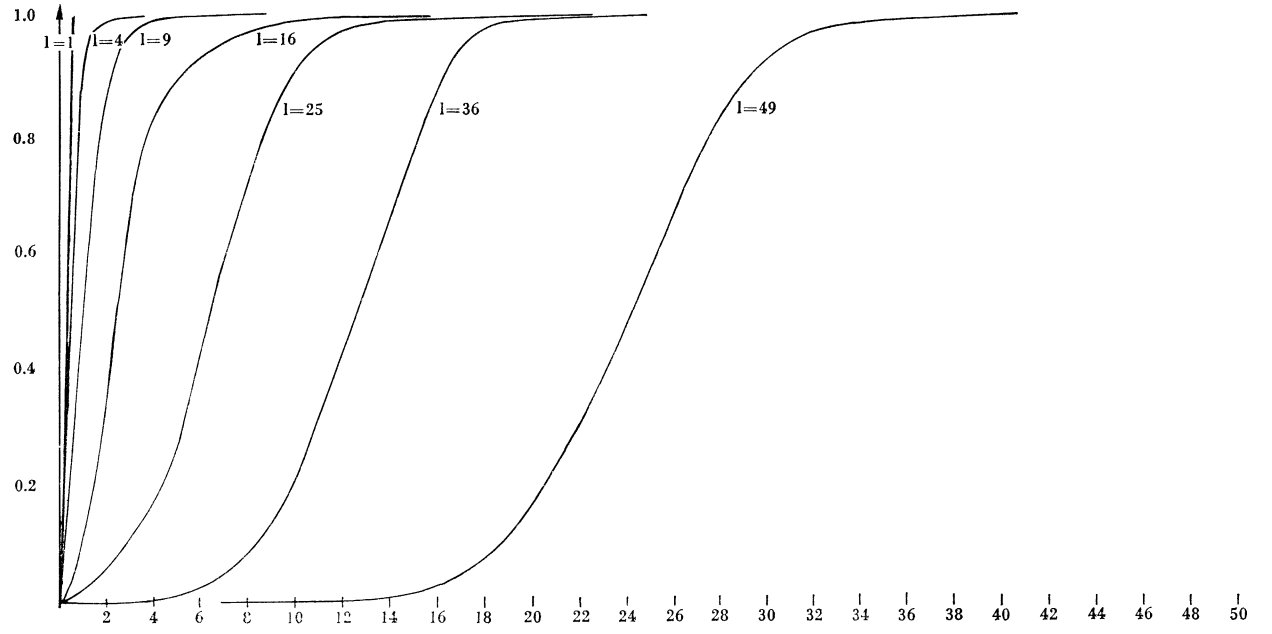
Supposons $n = 1\,000$ et soient deux paires d'objets $\{ x, y \}$ et $\{ x', y' \}$ telles que les deux composantes de la première (resp. seconde) paire aient une taille commune égale à 50 (resp. 500).

Les valeurs de l'indice s correspondant à une valeur égale à 1 pour la statistique (1) sont respectivement 4 pour la paire $\{ x, y \}$ et 264 pour la paire $\{ x', y' \}$. Or il semble intuitivement que deux objets de taille 500 qui ont 264 attributs communs se ressemblent davantage que deux objets de taille 50 qui n'ont que 4 attributs communs. Calculons la vraisemblance dans l'hypothèse N de chacun de ces deux résultats.

Pour la paire $\{ x, y \}$, $\lambda = p\,u = 2{,}5$ est trop petit devant $p = 1\,000$, utilisant la table de la loi de Poisson, on a: $P(x, y) = 0{,}76$; tandis que pour la paire $\{ x', y' \}$, $\lambda' = pu' = 250$ et la table de la loi normale fournit $P(x', y') = 0{,}84$. Du point de vue de notre mesure de similarité les objets x et y sont moins proches que x' et y'.

Le graphique 1 représente une famille de segments de droite. Une même droite définit la variation de l'indice (1) lorsque s varie, pour une paire d'objets de même taille l, ($p = 100$); alors qu'une courbe donnée du graphique 2 définit dans les mêmes conditions, la variation de la mesure $P(x, y)$. On notera que si le nombre d'attributs possédés par un même objet était constant dans E, la préordonnance associée à $P(x, y)$ est la même que celle associée à s qui est d'ailleurs, la même que celle associée à tout indice de la forme $\mathscr{S}(s, u, v)$. De plus, si la valeur du paramètre λ ne devenait pas trop petite par rapport à p dans E; la préordonnance associée à $P(x, y)$ est la même que celle associée à l'indice (1) ci-dessus.





VII. MESURE DE SIMILARITÉ ENTRE PARTIES DISJOINTES DE E

Cette nouvelle notion de mesure de similarité nous permet de définir, de façon naturelle, une mesure de similarité entre parties disjointes de E qui tient en particulier compte des cardinaux de ces parties.

Si G et H sont deux sous-ensembles disjoints de E de cardinaux respectifs g et h. Considérons l'ensemble des valeurs de $P(x, y)$ lorsque x parcourt G et y, H:

$$\{ P(x, y) \mid x \in G, y \in H \} \tag{1}$$

et désignons la plus grande de ces valeurs par:

$$P(G, H) = \max \{ P(x, y) \mid x \in G, y \in H \}. \tag{2}$$

On ne peut juger de la proximité des parties en cause en se référant uniquement à la valeur de $P(G, H)$, car une même valeur assez grande de $P(G, H)$ peut être naturelle si G et H sont de cardinal élevé et assez exceptionnelle si G et H sont de faible cardinal.

Par conséquent, nous allons comme précédemment, nous référer à l'hypothèse N. $P(G, H)$ peut se mettre sous la forme :

$$\max \{ \max \{ P(x, y) \mid y \in H \} \mid x \in G \} = \max \{ P(x, H) \mid x \in G \},$$

en notant :

$$P(x, H) = \max \{ P(x, y) \mid y \in H \};$$

or l'ensemble des valeurs $\{ P(x, y) \mid y \in H \}$ constitue dans l'hypothèse N, un échantillon de h points indépendants d'une variable aléatoire uniformément répartie entre 0 et 1 ; d'où :

$$Pr \{ P(x, H) < t \} = t^h \tag{3}$$

où :

$$0 < t < 1.$$

D'autre part, l'ensemble des valeurs $\{ P(x, H) \mid x \in G \}$ constitue dans l'hypothèse N, un échantillon de g points indépendants d'une variable aléatoire dont la fonction de répartition vient d'être précisée en (3) ; par conséquent :

$$Pr \{ P(G, H) < t \} = (t^h)^g = t^{hg}.$$

La probabilité d'observer pour $P(G, H)$ une valeur inférieure à t étant t^{gh} ; si $\bar{\omega}$ est la valeur observée de $P(G, H)$, on retiendra comme mesure de similarité entre G et H, $\bar{\omega}^{gh}$. Il sera intéressant d'appliquer avec une telle mesure de similarité entre classes, l'algorithme classique définissant une hiérarchie de classifications ascendante, où à chaque pas on réunit les deux classes les plus proches.

BIBLIOGRAPHIE

[1] BECKNER M., *The biological way of thought*, Columbia University Press, New York, 1959.

[2] BENZÉCRI J. P., *Analyse factorielle des proximités I et II*, Publications de l'Institut de Statistique de l'Université de Paris, XIII et XIV, 1964 et 1965.

[3] BENZÉCRI J. P., *Classification automatique et reconnaissances des formes*, cours I.S.U.P., 1968-1969.

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

[5] LERMAN I. C., " Typologie des personnages enfants à travers la littérature enfantine ", rapport interne, *C.M.A.C.*, Maison des Sciences de l'Homme.

[6] SHEPARD R. N., "The analysis of proximités : scaling with an unknown distance function", I et II, *Psychometrika*, 1962.

[7] SOKAL R. R. et SNEATH P. H. A., *Principles of numerical taxonomy*, Freeman, San Francisco and London, 1963.

[8] VÉGA W. F., de la, " Techniques de classification automatique utilisant un indice de ressemblance ", *Revue de Sociologie*, déc. 1967.

[9] ACHARD P., *Biais statistique des indices de similarité*. Note interne. C.M.A.C., Maison des Sciences de l'Homme, Paris, 1970.