MATHÉMATIQUES ET SCIENCES HUMAINES

M. BESSON

À propos des distances entre ensembles de parties

Mathématiques et sciences humaines, tome 42 (1973), p. 17-35

<http://www.numdam.org/item?id=MSH_1973__42__17_0>

Math. Sci. hum. (11e année, no 42, 1973, p. 17-35)

A PROPOS DES DISTANCES ENTRE ENSEMBLES DE PARTIES

par
M. BESSON [1]

RÉSUMÉ

Soit E un ensemble fini. Nous nous proposons ici d'établir des indices de distance entre les divers ensembles de parties de E. Nous généraliserons pour cela, à $\mathscr{P}\mathscr{P}(E)$, la notion de fonction caractéristique bien connue dans $\mathscr{P}(E)$.

SUMMARY

Let E be a finite set. We demonstrate a method to establish distances between different sets of sub-sets of E. We will apply in a generalised form the well-known characteristic function of $\mathscr{P}(E)$ to $\mathscr{P}\mathscr{P}(E)$.

1. Nous nous proposons d'étudier ici les distances entre ensembles de parties d'un ensemble E fini, que nous convenons d'appeler, pour simplifier l'écriture, les familles de E.

1.1. Soit E un ensemble fini
Soit R une famille de E

Un certain nombre de conditions sur R ordonnées par l'implication forment un sup demi-treillis explicité par le schéma 1.

1.2. Fonctions caractéristiques

Nous convenons d'adopter la définition suivante :
Une fonction caractéristique pour un certain type de famille de E est une application de E dans N^+ [2] qui caractérise chaque famille de ce type.

Établir une fonction caractéristique pour un certain type de famille de E, c'est donc trouver une injection de l'ensemble des familles de ce type dans l'ensemble des applications de E dans N^+.

1.3. Vecteurs caractéristiques

Soient $x_1, x_2, \ldots, x_n$ les éléments de E. Soit f une fonction caractéristique pour une certaine famille R sur E. Nous appellerons « Vecteur caractéristique » associé à (f, R) le vecteur :

$$\vec{V} = \{f(x_1), f(x_2) \ldots f(x_n)\}$$

1. Ecocentre de MacGregor Comarain. Directeur Nguyen Tien Phuc, 219, rue de Versailles, 92-Ville d'Avray.
2. Ensemble des entiers positifs ou nuls.

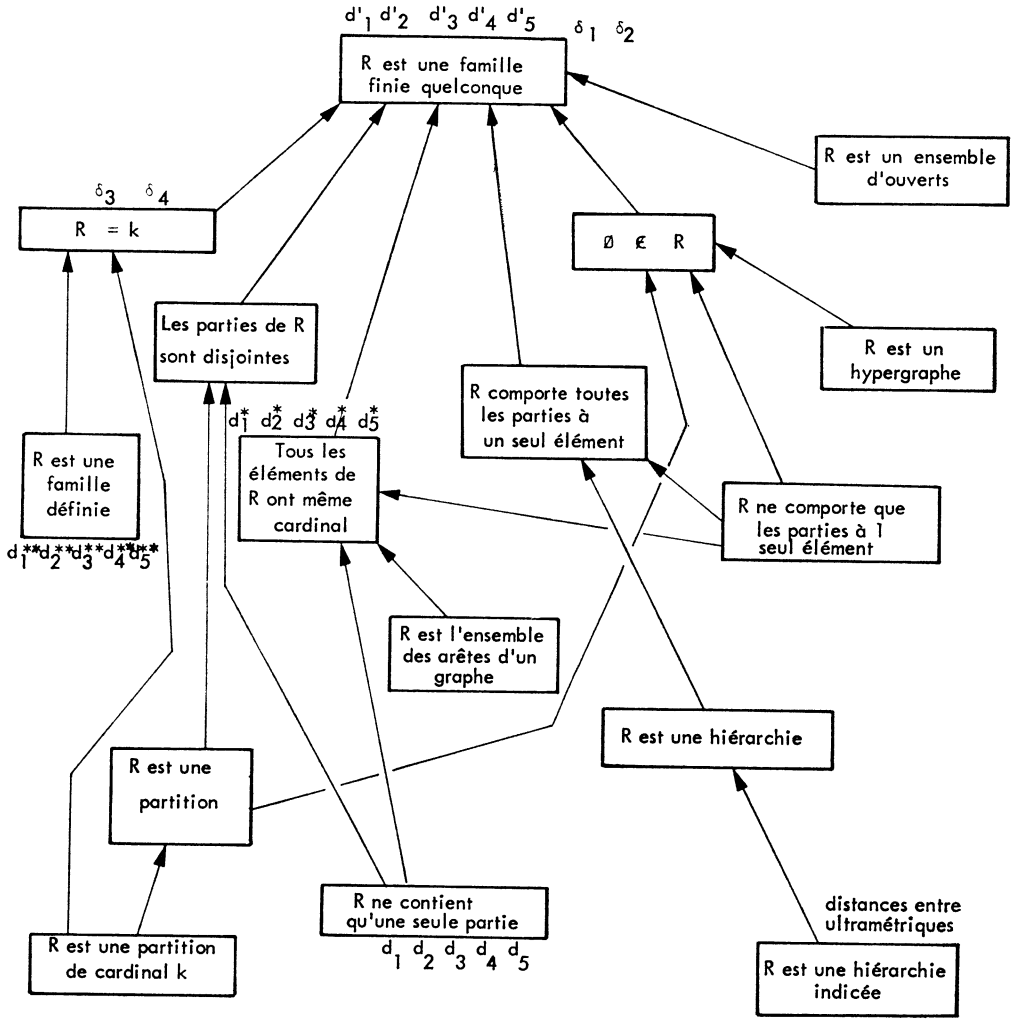


Schéma 1

Théorème

Soit T un certain type de familles sur E, et f_T une fonction caractéristique adaptée à ce type. Toute distance entre les vecteurs caractéristiques construits avec f_T, induit une distance entre les familles de type T.

Remarque : Ce théorème n'est qu'un cas particulier du théorème suivant :

«Soient E et E′ deux ensembles quelconques. Soit Ψ une injection de E dans E′. Soit d une distance sur E′

Définissons δ, application de E × E dans R^+, par l'égalité suivante :

$$(\forall\, x \in E)\ (\forall\, y \in E)\ (\delta\,(x, y) = d\,(\Psi\,(x), \Psi\,(y))$$

δ est une distance sur E.»

II. LES PARTIES

Nous nous proposons d'étudier maintenant les familles les plus simples : les parties (familles de cardinal 1).

2.1. *Fonctions caractéristiques*

La fonction caractéristique d'une partie X de E est l'application f_X de E dans [0, 1] telle que :

$(\forall\ x \in X)\ (f_X(x) = 1)$

$(\forall\ x \notin X)\ (f_X(x) = 0)$

Le vecteur caractéristique correspondant à une telle partie X est le vecteur :

$\vec{V}_f(X) = \{f_X(x_1), f_X(x_2) \ldots f_X(x_n)\}$

(vecteur qui ne comporte évidemment que des 0 ou des 1.)

2.2. *Distances induites par cette fonction caractéristique*

Soient X et Y deux parties quelconques de E.

Distance euclidienne :

Écrivons la distance euclidienne entre les vecteurs $V_f(X)$ et $V_f(Y)$:

$$d(V_f(X), V_f(Y)) = \sqrt{\sum_{i=1}^{n} (f_X(X_i) - f_Y(Y_i))^2}$$

$$= ||\vec{V}_f(X) - \vec{V}_f(Y)||$$

$$= \sqrt{|X \Delta Y|}\ [1]$$

Distance induite correspondante sur l'ensemble des parties de E

$$d_1(X, Y) = \sqrt{|X \Delta Y|}$$

Distance : somme des différences des coordonnées en valeur absolue

$$d_2(\vec{V}_f(X), \vec{V}_f(Y)) = \sum_{i=1}^{n} |f_X(x_i) - f_Y(x_i)|$$

$$= |X \Delta Y|$$

Distance : sup. des différences des coordonnées en valeur absolue

$$d(\vec{V}_f(X), \vec{V}_f(Y) = \underset{i=1,n}{Sup} (|f_X(x_i) - f_Y(x_i)|)$$

La distance induite sur l'ensemble des parties de E sera donc :

$$d_3(X, Y) = \begin{cases} 1 & \text{si } X \neq Y \\ 0 & \text{si } X = Y \end{cases}$$

2.3. *Autres distances*

Soient $x_1, x_2 \ldots x_n$ les éléments de E. Considérons trois parties A, B et C de E :

$A = \{x_1, x_2, x_3\}$

1. Δ représente la différence symétrique.

$B = \{x_4\}$

$C = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$

Appliquons-leur les distances d_1, d_2 et d_3 :

$|A \Delta B| = 4$

$|A \Delta C| = 4$

Donc : $d_1(A, B) = d_1(A, C)$

$d_2(A, B) = d_2(A, C)$

$d_3(A, B) = d_3(A, C) = 1$

Et cependant, n'avons-nous pas instinctivement l'impression que la ressemblance entre A et B est bien plus faible que celle entre A et C, qui ont trois éléments en commun, alors que A et B sont disjoints.

Pour pallier à cet inconvénient, et que des parties disjointes correspondent à des distances particulièrement fortes, nous avons cherché à établir d'autres distances (ou indices de distance).

Distance variant entre 0 et 1 :

Pour tout couple X et Y de parties de E, non-vides à la fois, posons :

$$d_4(X, Y) = \frac{|X \Delta Y|}{|X \cup Y|}^{1} \quad \text{et} \quad d_4(\varnothing, \varnothing) = 0$$

d_4 vaut 0 si et seulement si $X = Y$

d_4 vaut 1 si et seulement si $X \cap Y = \varnothing$

sans que X et Y soient simultanément vide.

Ainsi nous aurions dans l'exemple précédent

$$d_4(A, B) = \frac{4}{4} = 1$$

$$d_4(A, B) = \frac{4}{7}$$

Nous pouvons exprimer cette distance d avec les fonctions caractéristiques :

$$d_4(X, Y) = \frac{\sum_{i=1}^{n} (f_X(x_i) + f_Y(x_i) - 2 f_X(x_i)\, f_Y(x_i))}{\sum_{i=1}^{n} (f_X(x_i) + f_Y(x_i) - f_X(x_i)\, f_Y(x_i))}$$

ou avec les vecteurs caractéristiques :

$$d_4(X, Y) = \frac{(\vec{V}_f(X) + \vec{V}_f(Y))\, \vec{1}.^{2} - 2\, \vec{V}_f(X) . \vec{V}_f(Y)}{(\vec{V}_f(X) + \vec{V}_f(Y)) . \vec{1} - \vec{V}_f(X) . \vec{V}_f(Y)}$$

1. Cette distance a été utilisée également par M. Brissaud voir [6].
La mesure de similarité proposée par Jaccard en 1908 peut ainsi s'écrire : $\int (X, Y) = 1 - d_4(X, Y)$.

2. Vecteur unité.

d_4 vérifie bien en effet les axiomes d'une distance.

i) $d_4(X, Y) = 0 \Rightarrow X = Y$ (résultat trivial)

ii) $d_4(X, Y) = d_4(Y, X)$ (résultat trivial)

iii) d_4 vérifie l'inégalité triangulaire [1]

Indice de distance variant entre 0 et l'infini :

Pour tout couple X et Y de parties de E non simultanément vides, posons :

$$d_5(X, Y) = \frac{|X \Delta Y|^2}{|X \cap Y|} = \frac{\sum_{i=1}^{n} (f_X(x_i) + f_Y(x_i) - 2f_X(x_i) \cdot f_Y(x_i))}{\sum_{i=1}^{n} f_X(x_i) \cdot f_Y(x_i)}$$

$$= \frac{(\vec{V}_f(X) + \vec{V}_f(Y)) \cdot \vec{1} - 2\vec{V}_f(X) \cdot \vec{V}_f(Y)}{\vec{V}_f(X) \cdot \vec{V}_f(Y)}$$

et $d_5(X, X) = 0$.

Ainsi, nous aurions dans l'exemple précédent :

$d_5(A, B) \to \infty$

$d_5(A, C) = 4/3$

d_5 est bien un indice de distance sur l'ensemble des parties de E.

Mais d_5 n'est pas une distance : l'inégalité triangulaire n'est pas vérifiée.

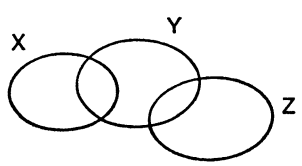
En effet, soient X, Y et Z trois parties d'un ensemble E fini quelconque telles que :

$X \cap Y \neq \varnothing$

$X \cap Z = \varnothing$

$Y \cap Z \neq \varnothing$

Nous avons bien sûr : $d_5(X, Z) > d_5(X, Y) + d_5(Y, Z)$



2.4. Corrélation entre deux vecteurs caractéristiques et indice de distance correspondant

Soit E un ensemble quelconque, à *n* éléments. Soient X et Y deux parties de E. Les séries statistiques formées par les diverses coordonnées de leurs vecteurs caractéristiques $V_f(X)$ et $V_f(Y)$ correspondent au coefficient de corrélation suivant :

---

1. Cette démonstration a été faite par G. Morlat, voir [7].
2. d_5 se trouve être l'inverse de la mesure de similarité proposée par Kuleyzinski en 1927, voir [4].

$$r\left(\vec{V}_f(X),\ \vec{V}_f(Y)\right) = \frac{\sum_{i=1}^{n} \left(f_X(x_i) - m_X\right)\left(f_Y(x_i) - m_Y\right)}{\sqrt{\sum_{i=1}^{n} \left(f_X(x_i) - m_X\right)^2}\ \sqrt{\sum_{i=1}^{n} \left(f_Y(x_i) - m_Y\right)^2}}$$

avec $m_X = \dfrac{\Sigma f_X(x_i)}{n} = \dfrac{|X|}{n}$

et $m_Y = \dfrac{\Sigma f_Y(x_i)}{n} = \dfrac{|Y|}{n}$

$$r\left(\vec{V}_f(X),\ \vec{V}_f(Y)\right) = \frac{\sum_{i=1}^{n} f_X(x_i)\left(f_Y(x_i) - \dfrac{|X| \cdot |Y|}{n}\right.}{\sqrt{\Sigma x_i^2 - \dfrac{|X|^2}{n}} \cdot \sqrt{\Sigma\, y_i^2 - \dfrac{|Y|^2}{n}}}$$

$$= \frac{n\,|X \cap Y| - |X| \cdot |Y|}{n \cdot \sqrt{|Y| - \dfrac{|Y|^2}{n}} \cdot \sqrt{|X| - \dfrac{|X|^2}{n}}}$$

$$\boxed{r\left(\vec{V}_f(X)\,,\ \vec{V}_f(Y)\right) = \frac{n|X \cup Y| - |X| \cdot |Y|}{\sqrt{|X| \cdot |Y|} \cdot \sqrt{|X| \cdot |Y|}}}$$

Remarque : Si l'écart-type d'une série statistique est nul, le coefficient de corrélation est considéré comme nul également. De même, ici, par une convention analogue, si une des deux parties X et Y est $\varnothing$ ou E, r est nul aussi.

Indice de distance correspondant :

Posons : $d_0(X, Y) = 1 - r\left(\vec{V}_f(X), \vec{V}_f(Y)\right)$

$$\boxed{d_0(X, Y) = \begin{cases} 1 + \dfrac{|X| \cdot |Y| - n\,|X \cup Y|}{\sqrt{|X| \cdot |Y|} \cdot \sqrt{|X| \cdot |Y|}} \\ 1 \text{ si } X = \varnothing \text{ ou E} \\ \text{ou} \\ \text{si } Y = \varnothing \text{ ou E} \end{cases}}$$

Il s'agit bien d'un indice de distance. En effet :

$d_0(X, Y) = d_0(Y, X)$

$d_0(X, X) = 0$

$d_0(X, Y) = 0 \Rightarrow \vec{V}_f(X)$ et $\vec{V}_f(Y)$ homothétiques sont donc égaux.

Cet indice de distance d_0 varie entre 0 et 2.

— Il vaut 2 si et seulement si :

$r(\vec{V}_f(X), \vec{V}_f(Y)) = -1$

Donc : $Y = \overline{X}$

— Il vaut 1 si et seulement si :

$$r(\vec{V}_f(X), \vec{V}_f(Y)) = \frac{|X| \cdot |Y|}{n}$$

(On pourrait dire alors que les parties X et Y sont «indépendantes», par analogie avec des séries statistiques quelconques.)

Mais est-ce une distance ?

C'est-à-dire : l'inégalité triangulaire est-elle vérifiée ?

La question reste ouverte.

2.5. Stabilité par passage au complémentaire

Une propriété assez intéressante des indices de distance entre parties d'un ensemble E, est leur stabilité par passage au complémentaire.

C'est-à-dire :

$(\forall X \subset E)\ (\forall Y \subset E)\ (d(\overline{X}, \overline{Y})) = d(X, Y)$

Puisque «$X \Delta Y = \overline{X} \Delta \overline{Y}$», les distances d_1, d_2 et d_3 seront stables. Par contre, les distances d_4, d_5 et d_0 ne le seront pas, en général.

— d_4 et d_5 ne vérifieront la propriété de stabilité que si :

$$|X \cap Y| = |\overline{X} \cap \overline{Y}|$$

— Tandis que d_0 ne la vérifiera que si :

$n\,|X \cap Y| - |X| \cdot |Y| = n\,|\overline{X} \cap \overline{Y}| - |\overline{X}| \cdot \overline{Y}|$

donc : $n\,|X \cap Y| = |X| \cdot |Y|$

(parties indépendantes).

III. LES PARTITIONS

Pour définir des distances dans des ensembles de partitions, nous essaierons d'abord de trouver des fonctions caractéristiques, correspondant à des vecteurs caractéristiques. Toute distance entre les vecteurs induira une distance entre les partitions.

Puis, nous essaierons par une démarche plus directe, de déterminer des indices de distance plus intéressants de divers points de vue.

3.1. Fonctions caractéristiques

Nous cherchons ici des fonctions de E dans N^+ qui caractérisent chaque partition.

Soit $E = \{x_1, x_2, ..., x_n\}$

Soit T une partition de E.

Pour tout élément y de E, posons :

$f_R(y) = Inf \{i / (\exists\ X \in R)\ ((y, x_i) \in X)\}$

C'est-à-dire : $f_R(y)$ est le plus petit des indices i tels que x_i soit réuni avec y dans la partition R.

Soit R′ une partition de E, différente de R. Donc :

$(\exists\ y \in E)\ (\exists\ z \in E)$ ((y, z) réunis par v et non par R′ ou (y, z) réunis par R′ et non par R).

Or, si (y, z) est un couple d'éléments de E réunis par R et non pas par R′ on a :

$$f_R(y) = f_R(z)$$

et $$f_{R'}(y) \neq f_{R'}(z)\ ).$$

Donc les fonctions f_R et $f_{R'}$ sont différentes.

Nous venons donc de définir une injection de l'ensemble des partitions dans l'ensemble des applications de E dans N^+, c'est-à-dire des fonctions caractéristiques pour les partitions de E.

Vecteur caractéristique correspondant :

$V_f(R) = (f_R(x_1), f_R(x_2) \ldots f_R(x_n))$.

Les distances euclidiennes, *somme*, et *sup* des différences des coordonnées en valeur absolue, impliquent évidemment des distances dans l'ensemble des partitions. Mais que représentent-elles dans la réalité ?

Il semble que les cas où elles sont bien adaptées sont assez rares.

3.2. Indices de distance

Acceptant de ne plus vérifier l'inégalité triangulaire, nous cherchons à établir maintenant de simples indices de distances, (plus significatifs) de la différence réelle entre deux partitions.

Soient Q et R deux partitions quelconques.

Posons :

$$\Psi'(Q, R) = \frac{1}{|R|}\left[\sum_{Z \in R}\ Inf_{X \in Q}\ \frac{|X\ \Delta\ Z|}{|X\ \cup\ Z|}\right] \text{ Si Q et R non vides à la fois.}$$

C'est-à-dire : pour chaque élément Z de R, nous cherchons l'élément X de Q tel que le rapport des cardinaux de $X \cap Z$ et de $X \cup Z$ soit le plus grand possible, Ψ' est la moyenne des quotients ainsi obtenus.

Puis, nous définissons des applications δ_1 et δ_2 par les relations :

$$\delta_1(R, Q) = \frac{1}{2}[\Psi(R, Q) + \Psi(Q, R)]$$

$$= \frac{1}{2|R|}\left[\sum_{X \in R} \underset{Z \in Q}{Inf}\, d_4(X, Z)\right] + \frac{1}{2|Q|}\left[\sum_{Z \in Q} \underset{X \in R}{Inf}\, d_4(X,Z)\right]$$

$$\delta_2(R, Q) = \frac{1}{2}\left[\frac{1}{1-\Psi(R, Q)} + \frac{1}{1-\Psi(Q, R)}\right] - 1$$

Démontrons que δ_1 et δ_2 sont des indices de distance sur l'ensemble des partitions :

i) $\Psi(R, Q) \in [O, 1]$

donc : $\delta_1(R, Q) \in [O, 1]$

donc : $\delta_1(R, Q) \geqslant O$ et $\delta_2(R, Q) \geqslant O$

ii) δ_1 et δ_2 sont symétriques par leur définition même

iii) δ_1 et δ_2 sont nuls et seulement si $R = Q$.

En effet :

$$(R = Q) \Rightarrow (\Psi(R, Q) = \Psi(Q, R) = 0)$$

donc : $(R = Q) \Rightarrow (\delta_1(R, Q) = 0$ et $\delta_2(R, Q) = 0)$

D'autre part :

$$(\delta_1(R, Q)) = 0 \Rightarrow (\Psi(R, Q) = \Psi(Q, R) = 0)$$

or : $(\Psi(R, Q) = 0) \Rightarrow (Q \supset R)$

et $(\Psi(Q, R) = 0) \Rightarrow (R \supset Q)$

donc : $(\delta_1(R, Q) = 0) \Rightarrow (Q = R)$

(même démonstration pour δ_2)—

δ_1 et δ_2 satisfont donc tous les axiomes des indices de distance.

Remarque : δ_1 vaut 1 si et seulement si :

$$\Psi(Q, R) = \Psi(R, Q) = 0$$

c'est-à-dire :

$$(\forall\, X \in R) \quad (\forall\, Z \in Q) \quad (X \cap Z = \varnothing)$$

or : $\underset{Z \in R}{(\cup Z)} \cap \underset{Z \in Q}{(\cup Z)} = \varnothing$ (et δ_2 est alors infini).

Ce qui n'arrive jamais puisque : $\underset{Z \in R}{\cup} Z = \underset{Z \in Q}{\cup} Z = E$

Généralisation

Il est immédiat de voir que ces indices de distances sont valables non seulement dans l'ensemble des partitions, mais aussi dans l'ensemble des familles quelconques. Leur application dans le cas particulier de l'ensemble des parties donne à nouveau les indices d_4 et d_5.

3.3. Exemple

Soit $E = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$

Soit $R = [\{x_1, x_4, x_5, x_6\}, \{x_2, x_3\}, \{x_7\}] \quad |R| = 3$

Soit $Q = [\{x_2, x_4, x_5, x_6, x_7\} \quad \{x_1, x_3\}] \quad |Q| = 2$

Distance induite

Vecteur caractéristique :

$$\vec{V}_f(R) = [1, 2, 2, 1, 1, 1, 7]$$

$$\vec{V}_f(Q) = [1, 2, 1, 2, 2, 2, 2]$$

Toute distance entre ces vecteurs à 7 dimensions induit une distance dans l'ensemble des partitions dont l'intérêt est en général faible.

Indice de distance

$$\Psi(R, Q) = \frac{1}{3}\left[\frac{3}{6} + \frac{2}{3} + \frac{4}{5}\right]$$

$$= \frac{1}{3}\left[\frac{15 + 20 + 24}{30}\right] = \frac{59}{90}$$

$$\Psi(Q, R) = \frac{1}{2}\left[\frac{3}{6} + \frac{2}{3}\right]$$

$$= \frac{1}{2}\left[\frac{3 + 4}{6}\right] = \frac{7}{12}$$

$$\delta_1(R, Q) = \frac{1}{2}\left[\frac{59}{90} + \frac{7}{12}\right] = 0{,}62$$

$$\delta_2(R, Q) = \frac{1}{2}\left[\frac{90}{31} + \frac{12}{5}\right] - 1 = 1{,}2 + 1{,}5 = 2{,}7.$$

3.4. Autres exemples

Soit $E = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$

Soit $R = [E]$

Soit $Q = [\{x_1\}, \{x_2\}, \{x_3\}, \ldots, \{x_7\}]$

R et Q sont encore deux partitions de E.

$$\Psi(Q, R) = \sum_{Z \in R} \operatorname{Inf}_{X \in Q}\left(\frac{|X \,\Delta\, Z|}{|X \cup Z|}\right)$$

$$= \frac{6}{7}$$

$$\Psi(R, Q) = \frac{1}{7}\left[\sum_{X \in Q} \operatorname{Inf}_{Z \in R}\left(\frac{|X \,\Delta\, Z|}{|X \cup Z|}\right)\right]$$

$$\Psi\,(Q, R) = \frac{1}{7}\left[7 \cdot \frac{6}{7}\right] = \frac{6}{7}$$

Donc : $$\delta_1\,(R, Q) = \frac{1}{2}\left[\frac{6}{7} + \frac{6}{7}\right] = \frac{6}{7} = \frac{|E| - 1}{|E|}$$

$$\delta_2\,(R, Q) = \frac{1}{2}\,[7 + 7] - 1 = 6 = |E| - 1$$

(voir annexe p. 34).

Considérons maintenant un autre exemple :

Soit $E = \{x_1, x_2, x_3, x_4\}$

Soit $R = [\{x_1, x_2\}, \{x_3, x_4\}]$

Soit $Q = [\{x_1, x_3\}, \{x_2, x_4\}]$

$$\Psi(R, Q) = \frac{1}{2}\left[\sum_{Z \in R}\ \mathit{Inf}_{X \in Q}\left(\frac{|X\ \Delta\ Z|}{|X\ \cup\ Z|}\right)\right]$$

$$= \frac{1}{2} \times 2\left[\frac{2}{3}\right] = \frac{2}{3} < \frac{|E| - 1}{|E|}$$

$$\Psi\,(Q, R) = \frac{2}{3}$$

Donc : $$\delta_1\,(R, Q) = \frac{2}{3}$$

$$\delta_2\,(R, Q) = \frac{1}{2}\,[3 + 3] - 1 = 2 < |E| - 1$$

IV. LES FAMILLES DE CARDINAL K

4.0. Un ensemble de k parties de E forme une famille de cardinal k. Comme pour les partitions nous déterminerons d'abord des distances à l'aide de fonction et vecteurs caractéristiques. Puis, déçus par leur manque de signification réelle, nous chercherons à établir des indices de distance plus représentatifs.

4.1. Soit R une famille de cardinal k. Ordonnons ces k parties suivant l'ordre lexicographique des éléments de E contenus :

$$R = \{A_1, A_2, \ldots, A_k\}$$

Soit α_i la fonction caractéristique habituelle associée à la partie A_i.

A tout élément x de E, par ces fonctions caractéristiques α_i, correspondent les chiffres (0 ou 1) :

$$\alpha_1(x), \alpha_2(x), \ldots, \alpha_k(x)$$

Considérons alors le nombre :

$$f(x) = \alpha_1(x)\ \alpha_2(x)\ \ldots \qquad \alpha_i(x)\ \ldots\ \alpha_k(x)$$

(ce nombre ne comporte que des zéros et des 1. On peut le considérer comme étant en système binaire et le traduire en système décimal. Mais cela n'a rien d'obligatoire.)

f(x) définit une fonction caractéristique, à laquelle correspond un vecteur caractéristique.

Mais il n'y a pas de raison de privilégier ainsi A_i par rapport à A_j si $i > j$.

C'est pourquoi les distances induites par celles entre vecteurs caractéristiques ne seront guère intéressantes.

Cependant, l'une de ces distances induites pallie à cet inconvénient. C'est la distance : nombre de différences entre les mêmes coordonnées de deux vecteurs.

4.2. Exemple

Soit $E = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$

Soit $R = [\{x_1, x_2, x_3, x_4, x_5, x_7\}, \{x_6, x_7\}]$

Soit $Q = [\{x_2, x_4, x_5, x_6, x_7\}, \{x_1, x_4, x_3\}]$

$|R| = |Q| = 2$

Distance induite :

Vecteur caractéristique :

$$V_f(R) = \begin{vmatrix} 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 01 \\ 11 \end{vmatrix} = \begin{vmatrix} 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 1 \\ 3 \end{vmatrix}$$

système binaire

$$V_f(Q) = \begin{vmatrix} 10 \\ 01 \\ 10 \\ 11 \\ 01 \\ 01 \\ 01 \end{vmatrix} = \begin{vmatrix} 2 \\ 1 \\ 2 \\ 3 \\ 1 \\ 1 \\ 1 \end{vmatrix}$$

système binaire

Toute distance entre ces vecteurs à 7 dimensions induit une distance dans l'ensemble des familles de cardinal k.

Calculons la distance : nombre de différences entre les mêmes coordonnées :

$$d_6(R, Q) = 0 + 1 + 0 + 1 + 1 + 0 + 1$$
$$= 4$$

Mais cette distance présente encore un inconvénient : celui de tenir compte d'une correspondance implicite entre les parties de R et les parties de Q.

C'est pourquoi nous la conseillerons particulièrement dans le cas des familles «définies» (paragraphe 5-3).

4.3. Indices de distance

Soient Q et R deux familles quelconques à k éléments de E. Les indices δ_1 et δ_2 (étudiés en 3.2) tenaient compte de l'erreur minimale que l'on faisait en associant à chaque élément de R un élément de Q, et à chaque élément de Q, un élément de R. Nous allons définir maintenant, avec une optique légèrement différente, des indices δ_3 et δ_4 qui se rapporteront à l'erreur minimale commise en associant bijectivement les éléments de R aux éléments de Q (ceci est possible puisque Q et R ont même cardinal : k).

Soit B (Q, R) l'ensemble des bijections de Q dans R.

Posons :

$$\Psi'(Q, R) = \underset{g \in B(Q, R)}{Inf} \left[\sum_{X \in Q} \frac{|X \Delta g(X)|}{|X \cup g(X)|} \right] \times \frac{1}{k}$$

Il en ressort que :

(i) $\Psi'(Q, R) = \Psi'(R, Q)$

(ii) Ψ' varie entre 0 et 1

Posons alors :

$$\delta_3(Q, R) = \Psi'(Q, R)$$

δ_3 est un indice de distance pour l'ensemble des familles de cardinal k, variant entre 0 et 1.

En effet :

(i) $\delta_3 \geqslant 0$

(ii) $\delta_3 = 0 \Rightarrow \Psi'(Q, R) = 1 \Rightarrow Q \subset R$

Mais comme $|Q| = |R| = k$

$$\delta_3 = 0 \Rightarrow Q = R$$

(iii) δ_3 est évidemment symétrique, puisque Ψ' l'est.

Remarque : δ_3 vaut 1 si et seulement si

$$\left(\bigcup_{X \in Q} X\right) \cap \left(\bigcup_{Y \in R} Y\right) = \varnothing$$

Si au lieu de désirer un indice de distance variant entre 0 et 1, nous en voulions un variant entre 0 et l'infini, il faudrait poser :

$$\delta_4(R, Q) = \frac{1}{1 - \Psi'(R, Q)} - 1$$

(il est trivial de vérifier qu'il s'agit bien d'un indice de distance).

Remarque : δ_4 est infini si et seulement si

$$\left(\bigcup_{Z \in Q} Z\right) \cap \left(\bigcup_{X \in R} X\right) = \varnothing$$

Cas particulier : si $k = 1$, nous nous trouvons à nouveau dans l'ensemble des parties de E, où δ_4 et δ_5 se confondent avec d_4 et d_5.

4.4. Reprenons l'exemple précédent (paragraphe 4.2.)

Indice de distance :

$$\Psi'(Q, R) = \frac{1}{2} \, Inf \left[\left(\frac{2}{7} + \frac{5}{5} \right), \left(\frac{3}{6} + \frac{3}{5} \right) \right]$$

$$= \frac{1}{2} \, Inf \left[\left(\frac{9}{7} \right), \left(\frac{15 + 18}{30} \right) \right] = \frac{1}{2} \, Inf \left[\frac{9}{7}, \frac{33}{30} \right]$$

$$= \frac{1}{2}\,\frac{33}{30} = \frac{33}{60} = 0,\ 5$$

$$\delta_3(Q, R) = \Psi'(Q, R) = 0{,}55$$

$$\delta_4(Q, R) = \frac{60}{27} - 1 = \frac{33}{27} = 1{,}2.$$

V. AUTRES FAMILLES

5.1. Hiérarchies indicées

Nous rappelons qu'une hiérarchie indicée est le couple formé par une hiérarchie $\mathscr{H}$[1] et une application φ de cette hiérarchie dans R^+ telle que :

(*i*) $(\forall\, X \in \mathscr{H})\ \ (|X| = 1 \Rightarrow \varphi(X) = 0)$

(*ii*) $(\forall\, X \in \mathscr{H})\ \ (\forall\, Y \in \mathscr{H})\ \ (Y \supset X \Rightarrow \varphi(X) \leqslant \varphi(Y))$

Nous savons d'autre part qu'une hiérarchie indicée est équivalente à une ultramétrique [2].

On peut donc la représenter par un vecteur à n^2 dimensions (les diverses valeurs de l'ultramétrique pour les n^2 couples d'éléments de E). Et toute distance entre ces vecteurs définira une distance dans une hiérarchie indicée.

5.2. Familles dont tous les éléments ont pour cardinal p

(ou bien : ensemble de p-uplets de E)

Soit R une telle famille :

$R = \{X_1, X_2 \ldots X_k\}$

avec :

$(\forall\, i \leqslant k)\ (|X_i| = p)$

Plaçons-nous dans E^p (ensemble de p-uplets d'éléments de E). Appelons Y_i l'ensemble des permutations des éléments de X_i (Y_i est une partie de E^p).

Posons alors :

$$Y = \bigcup_{i=1}^{i=k} Y_i\,; \quad Y_i = \bigcup_{i=1}^{i=k} \quad \text{(permutations des éléments de } X_i\text{)}.$$

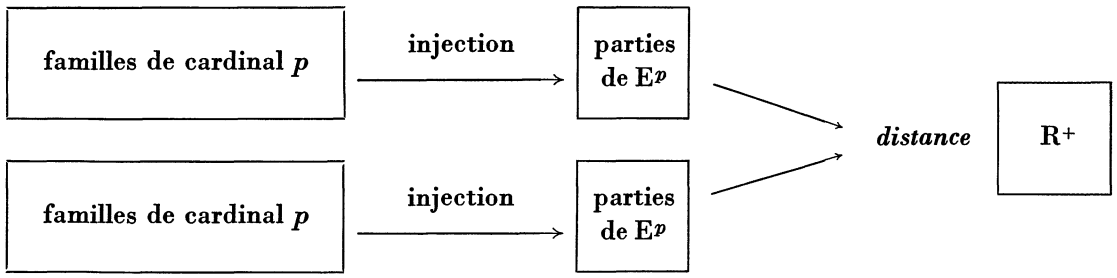
R correspond de façon injective à Y, qui est une partie de E^p.

Et nous connaissons d'autre part diverses distances entre les parties (d_1, d_2, d_3, d_4 et l'indice d_5). Chacune de ces distances induit par conséquent une distance dans l'ensemble des familles de cardinal k.

---

1. Voir [8].
2. Voir [7] annexe B.

Schéma explicatif



Étudions plus précisément ces distances

Soient R et R′ deux familles dont les éléments ont pour cardinal p.

Soient Y et Y′ les parties de E^p correspondantes de la façon que l'on vient de voir.

$$d_1^*(R, R') = d_1(Y, Y') = \sqrt{|Y \Delta Y'|}$$

$$= \sqrt{k! \,.\, |R \Delta R'|}$$

$$d_2^*(R, R') = d_2(Y, Y') = |Y \Delta Y'|$$

$$= k! \,.\, |R \Delta R'|$$

$$d_3^*(R, R') = d_3(Y, Y') = \begin{vmatrix} 1 \text{ si } R \neq R' \\ 0 \text{ si } R = R' \end{vmatrix}$$

$$d_4^*(R, R') = d_4(Y, Y') = \frac{k! \,|R \Delta R'|}{k! \,|R \cup R'|}$$

$$= \frac{|R \Delta R'|}{|R \cup R'|}$$

$$d_5^*(R, R') = d_5(Y, Y') = \frac{k! \,|R \Delta R'|}{k! \,|R \cap R'|}$$

$$= \frac{|R \Delta R'|}{|R \cap R'|}$$

Remarque : si $k = 1$, nous trouvons les mêmes distances entre R et R′ qu'entre les parties de E formées par la réunion des éléments de R, et celle de ceux de R′.

5.3. Familles définies

Nous appelons «famille définie d'ordre k», le couple formé par une certaine famille R de cardinal k et une bijection φ de R dans (1, 2 ..... k).

Posons alors, pour tout couple (R, φ) et (R′, φ') de familles d'ordre k

$$d_1^{**}(R, R') = \sum_{i=1}^{k} d_1(\varphi^{-1}(i), \varphi'^{-1}(i))$$

$$d_2^{**}(R, R') = \sum_{i=1}^{k} d_2(\varphi^{-1}(i), \varphi'^{-1}(i))$$

$$d_3^{**}(R, R') = \sum_{i=1}^{k} d_3(\varphi^{-1}(i), \varphi'^{-1}(i))$$

$$d^{**}_4(R, R') = \sum_{i=1}^{k} d_4(\varphi^{-1}(i), \varphi'^{-1}(i))$$

$$d_5^{**}(R, R') = \sum_{i=1}^{k} d_5(\varphi^{-1}(i), \varphi'^{-1}(i))$$

d_1, d_2, d_3 et d_4 étant des distances d_1^{**}, d_2^{**}, d_3^{**} et d_4^{**} le sont aussi (les axiomes de la distance se vérifient, il est trivial de le démontrer). Enfin, d_5 étant un indice de distance, d_5^{**} l'est aussi.

Remarque : une autre distance spécialement bien adaptée à ce cas précis est la distance d_6 (voir 4.2.).

5.4. Familles finies quelconques

Distances : chaque famille sur E étant une partie de l'ensemble des parties de E, nous pouvons appliquer entre les familles les distances d_1, d_2, d_3, d_4 ou d_5 définies entre les parties.

Nous aurons ainsi, pour tout couple (Q, R) de polyparties :

$$d'_1(Q, R) = \sqrt{|Q \Delta R|}$$

$$d'_2(Q, R) = |Q \Delta R|$$

$$d'_3(Q, R) = \begin{cases} 1 \text{ si } Q = R \\ 0 \text{ si } Q \neq R \end{cases}$$

$$d'_4(Q, R) = \frac{|Q \Delta R|}{|Q \cup R|}$$

$$d'_5(Q, R) = \frac{|Q \Delta R|}{|Q \cap R|}$$

Mais il y a 2^n parties dans un ensemble de n éléments. Et, par conséquent, des familles de cardinal faible, auront très peu de chances d'avoir des éléments communs. ($Q \cap R = \varnothing$).

Et nous aurons très souvent :

$$d'_1(Q, R) = \sqrt{|Q| + |R|}$$

$$d'_2(Q, R) = |Q| + |R|$$

$$d'_3(Q, R) = 1$$

$$d'_4(Q, R) = 1$$

$$d'_5(Q, R) = \infty$$

ce qui n'est guère intéressant.

Indices de distance : abandonnons encore l'inégalité triangulaire et appliquons les indices de distance δ_1 et δ_2 définis dans le cas particulier des partitions par les formules :

$$\Psi(Q, R) = \frac{1}{|R|}\left[\sum_{z \in R} \operatorname{Inf}_{x \in Q} \frac{|x \Delta z|}{|x \cup z|}\right]$$

$$\Psi(Q, \varnothing) = 1$$

$$\delta_1 (R, Q) = \frac{1}{2} [\Psi (R, Q) + \Psi (Q, R)]$$

$$\delta_2 (R, Q) = \frac{1}{2} \left[\frac{1}{1 - \Psi (R, Q)} + \frac{1}{1 - \Psi (Q, R)} \right] - 1$$

Ils s'appliquent sans aucune modification sur l'ensemble des familles finies quelconques.

Exemple :

Soit $E = \{x_1' \; x_2' \; x_3' \; x_4' \; x_5' \; x_6'\}$

et quatre familles sur E :

$Q = [\{x_1' \; x_2' \; x_3' \; x_4\}, \{x_5, x_6\}]$

$R = [\{x_1, x_2, x_3\}, \{x_3, x_4, x_5, x_6\}]$

$Q' = [\varnothing]$

$R' = [E]$

$Q \cap R = \varnothing$

$d'_1 (Q, R) = \sqrt{2 + 2} = 2$

$d'_2 (Q, R) = 2 + 2 = 4$

$d'_3 (Q, R) = 1$

$d'_4 (Q, R) = 1$

$d'_5 (Q, R) \to \infty$

$$\Psi (Q, R) = \frac{1}{2} \left[\frac{1}{4} + \frac{2}{4} \right] = \frac{3}{8}$$

$$\Psi (R, Q) = \frac{1}{2} \left[\frac{1}{4} + \frac{2}{4} \right] = \frac{3}{8}$$

$$\delta_1 (Q, R) = \frac{3}{8}$$

$$\delta_2 (Q, R) = \frac{1}{5} \left[\frac{8}{5} + \frac{8}{5} \right] - 1$$

$$\delta_2 (Q, R) = \frac{3}{5}$$

D'autre part : $Q' \cap R' = \varnothing$

Donc :

$d'_1 (Q', R') = \sqrt{2}$

$d'_2 (Q', R') = 2$

$d'_3 (Q', R') = 1$

$d'_4 (Q', R') = 1$

$d'_5 (Q', R') \to \infty$

D'autre part :

$\Psi (Q', R') = 1$

$\Psi (R', Q') = 1$

Donc :

$\delta_1 (Q', R') = 1$

$\delta_2 (Q', R') \to \infty$

les 5 premières distances donnant donc des résultats plus grands pour Q et R que pour Q′ et R′ (qui paraissent cependant plus «éloignés»). Tandis que δ_1 et δ_2 donnent des résultats plus conformes à ce que nous «sentons» être la réalité de l'éloignement.

VI. REMARQUE GÉNÉRALE

Reprenons le schéma du début de cette note et considérons les différents types de famille. Supposons que pour une famille donnée, le fait d'être du type T′ implique celui d'être du type T.

Alors l'ensemble des familles de type T′ est inclus dans l'ensemble des familles de type T.

Et toute distance (ou indice de distance) pour les familles de type T est aussi une distance (ou indice de distance) pour les familles de type T′.

CONCLUSION

Nous reprenons ici, le schéma initial complété avec l'indication des distances (ou indices de distance) qui s'y appliquent (voir page 18).

Le lecteur pourra ainsi aisément voir quels sont les distances et indices de distance qui s'appliquent pour chaque type de familles (en tenant compte de la remarque précédente).

Tel était le but que nous nous proposions ici.

ANNEXE

Remarque générale sur les indices δ_1 *et* δ_2

Soit E un ensemble quelconque à *n* éléments. Soient R et Q deux familles quelconques de E.

Rappel

Posons : $$\Psi (Q, R) = \frac{1}{|R|} \sum_{Z \in R} \; \mathit{Inf}_{X \in Q} \frac{|X \Delta Z|}{|X \cup Z|}$$

On a :

$$\delta_1 (R, Q) = \frac{1}{2} [\Psi (Q, R) + \Psi (R, Q)]$$

$$\delta_2 (R, Q) = \frac{1}{2} \left[\frac{1}{1 - \Psi (R, Q)} + \frac{1}{1 - \Psi (Q, R)} \right] - 1$$

δ_1 est un indice de distance qui varie entre 0 et 1 ($\delta_1 (\{\varnothing\}, \{E\}) = 1$).

Si Q et R sont des partitions, il ne pourra pas prendre la valeur 1.

Il nous semble que le maximum de δ_1 sera atteint

si $$R = \{E\}$$

et $$Q = [\{x_1\}, \{x_2\} \ldots\ldots \{x_n\}].$$

On aura alors : $\Psi (Q, R) = \dfrac{n-1}{n}$; $\Psi (R, Q) = \dfrac{n-1}{n}$

et $$\delta_1 (R, Q) = \frac{n-1}{n}$$

De même l'indice de distance δ_2 qui varie entre 0 et l'infini pour des familles quelconques, reste fini entre des partitions. Il nous semble que le maximum de δ_2 pour les partitions sera atteint dans le même cas particulier que le maximum de δ_1. Et nous aurons alors :

$$\delta_2 (Q, R) = n - 1.$$

Nous proposons donc ces deux conjectures.

Suivant ces indices, on pourrait dire que deux partitions ne peuvent être aussi «éloignées» que deux familles quelconques.

BIBLIOGRAPHIE

[1] Barbut, M., Monjardet, B., *Ordre et classification, algèbre et combinatoire*, Paris, Presses Universitaires de France, 1970, 2 tomes.

[2] Berge, C., *Graphes et hypergraphes*, Paris, Dunod, 1970.

[3] Choquet, M., *Topologie*, Paris, Masson, 1964.

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

[5] Roy, B., *Algèbre moderne et théorie des graphes*, Paris, Dunod, 1969.

[6] Brissaud, M., *Étude des populations et structures relationnelles*, Polycopié pour les sociologues, Faculté de Lyon, avril 1971.

[7] Besson, M., *Quelques méthodes d'analyses multicritères*, Thèse de 3e cycle, mars 1972, Université Paris.

[8] Roux, M., *Algorithme pour construire une hiérarchie particulière*, Thèse de 3e cycle, décembre 1968, Université Paris.