

M. SCHADER

**Distance minimale entre partitions et préordonnances
dans un ensemble fini**

Mathématiques et sciences humaines, tome 67 (1979), p. 39-47

http://www.numdam.org/item?id=MSH_1979__67__39_0

© Centre d'analyse et de mathématiques sociales de l'EHESS, 1979, tous droits réservés.

L'accès aux archives de la revue « Mathématiques et sciences humaines » (<http://msh.revues.org/>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

DISTANCE MINIMALE ENTRE PARTITIONS ET PREORDONNANCES DANS UN ENSEMBLE FINI

M. SCHADER *

1. INTRODUCTION

Beaucoup de problèmes de classification suggèrent d'exprimer la similarité ou dissimilarité d'objets par une préordonnance $\lesssim$ sur l'ensemble (fini) X d'objets** au lieu d'utiliser un indice de similarité à valeurs réelles (cf [5]).

Dans ce cas, on est devant le problème de construire, à partir d'une préordonnance $\lesssim$ donnée sur X ,

(1) une hiérarchie de X

(c.à.d. une famille $P_0, P_1, \dots, P_q$ de partitions de X qui satisfasse à $P_0 = \{\{x\} \mid x \in X\}$, $P_q = \{X\}$, $P_i \neq P_{i+1}$ et P_i plus fine que P_{i+1})

ou

(2) une partition P de X

telle que $(P_i)_i$ (respectivement P) représente aussi bien que possible les dissimilarités d'objets.

* Institut für Statistik und Mathematische Wirtschaftstheorie, Universität Augsburg, D-8900 Augsburg.

** On appelle préordonnance sur X un préordre total $\lesssim$ sur $Y := \{\{x, y\} \mid x, y \in X \text{ et } x \neq y\}$. $\{x, y\} < \{u, v\}$ exprime que x et y sont plus similaires que u et v .

Concernant le problème (1), on peut démontrer (cf [6]) qu'il existe une application bijective de l'ensemble des hiérarchies de X dans l'ensemble des ultra-préordonnances sur X (ce sont les préordonnances $\lesssim'$ sur X qui satisfont à $\{x,z\} \lesssim' \{x,y\}$ ou $\{x,z\} \lesssim' \{y,z\}$ pour tout $x,y,z \in X$).

Ainsi, pour la classification hiérarchique, on peut se limiter à l'analyse de l'ensemble $P(X)$ des préordonnances sur X . A la place de la hiérarchie $(P_i)_i$ on calcule l'ultrapréordonnance $\lesssim'$ correspondant à $(P_i)_i$.

L'ensemble fini $P(X)$ est ordonné par la relation R définie par

$$\lesssim_1 R \lesssim_2 : \Leftrightarrow G_{\lesssim_1} \subset G_{\lesssim_2}^*.$$

De plus, $P(X)$ est un sup-demi-treillis gradué et semi-modulaire supérieurement. Par conséquent $d : P(X)^2 \rightarrow \mathbb{R}_+$

$$d(\lesssim_1, \lesssim_2) := 2g(\sup\{\lesssim_1, \lesssim_2\}) - g(\lesssim_1) - g(\lesssim_2)$$

est une distance sur $P(X)$ qui respecte la structure de demi-treillis de $P(X)$ si g est une graduation quelconque de $P(X)$ (cf [3]).

D'après [6]

$$\lesssim' = \min\{\lesssim^* \mid \lesssim^* \text{ ultra-préordonnance et } \lesssim R \lesssim^*\}$$

est la solution au problème de minimiser $d(\lesssim, \lesssim')$ pour une préordonnance $\lesssim$ donnée à condition que $\lesssim'$ soit une ultra-préordonnance sur X ; ainsi la hiérarchie $(P_i)_i$ correspondant à $\lesssim'$ peut être considérée comme solution du problème (1).

Il est intéressant d'essayer de résoudre le problème (2) de manière analogue - c'est-à-dire en minimisant une distance "naturelle" sur $P(X)$.

* $G_{\lesssim} := \{(a,b) \mid a,b \in Y \text{ et } a \lesssim b\}$ désigne le graphe de $\lesssim$.

2. PARTITION ET ULTRA-PREORDONNANCE A DEUX CLASSES

Une ultra-préordonnance $\lesssim^+$ sur X sera dite ultra-préordonnance à deux classes si $|Y/\sim^+| = 2$ c'est-à-dire s'il y a précisément deux classes d'équivalence suivant $\sim^+$.

Or, on peut définir une application f de l'ensemble des ultra-préordonnances à deux classes sur X dans l'ensemble des partitions de X qui sont différentes de $\{X\}$ et de $\{\{x\} \mid x \in X\}$:

Si $\lesssim^+$ est une ultra-préordonnance à deux classes, on se rappelle que $Y/\sim^+ = \{E_1, E_2\}$. Nous supposons que ces classes sont numérotées de sorte que $\{x, y\} \in E_1$, $\{u, v\} \in E_2$ implique $\{x, y\} <^+ \{u, v\}$. Désignons par S la relation suivante sur X

$$x S y : \Leftrightarrow x = y \text{ ou } \{x, y\} \in E_1,$$

alors S est réflexif et symétrique. De plus, S est une relation transitive, car $\{x, y\} \in E_1$, $\{y, z\} \in E_1$ et $(\{x, z\} \lesssim^+ \{x, y\} \text{ ou } \{x, z\} \lesssim^+ \{y, z\})$ entraînent $\{x, z\} \in E_1$.

Autrement dit, S est une relation d'équivalence sur X et nous posons $f(\lesssim^+) := X/S$.

Evidemment f est injectif. En outre, f est surjectif puisque pour toute partition P de X ($\{X\} \neq P \neq \{\{x\} \mid x \in X\}$), il existe une ultra-préordonnance à deux classes $\lesssim^+$ caractérisée par $f(\lesssim^+) = P$:

$$\{x, y\} \lesssim^+ \{u, v\} : \Leftrightarrow \exists A \in P : x, y \in A \text{ ou } \nexists A \in P : u, v \in A.$$

En somme, f est une application bijective. Ainsi on peut essayer de résoudre le problème (2) en cherchant une approximation de la préordonnance donnée $\lesssim$ par une ultra-préordonnance à deux classes $\lesssim^+$. La partition cherchée sera $P = f(\lesssim^+)^*$.

* Il est clair que les partitions $\{\{x\} \mid x \in X\}$ et $\{X\}$ sont exclues par cette procédure..

3. LE DEMI-TREILLIS $P(X)$

On sait que l'ensemble $P(X)$ des préordonnances sur X est un sup-demi-treillis gradué, semi-modulaire supérieurement.

Exigeant que les éléments minimaux de $P(X)$ aient le niveau 0, on obtient la graduation suivante:

$$g(\preceq) := n(n-1)/2 - |Y/\sim|$$

où n désigne le nombre d'objets.

Il résulte, que les ultra-préordonnances à deux classes $\preceq^+$ (l'ensemble de ces préordonnances sera dans la suite désigné par $P^+(X)$) ont toutes le même niveau $n(n-1)/2 - 2$, c'est-à-dire qu'elles sont des prédécesseurs de $\sup P(X)$.

Au moyen des nombres de Stirling de seconde espèce $S(i,j)$ on peut déterminer le cardinal de l'ensemble des préordonnances du niveau $n(n-1)/2 - 2$ et le cardinal de $P^+(X)$.

Pour $\preceq \in P(X)$ et $g(\preceq) = n(n-1)/2 - 2$ on a $|Y/\sim| = 2$. Alors chaque élément de Y (il y en a au total $n(n-1)/2$) appartient à une des deux classes de $Y/\sim$, c.à.d.

$$|\{\preceq \mid \preceq \in P(X) \text{ et } g(\preceq) = n(n-1)/2 - 2\}| = 2^{n(n-1)/2} - 2.$$

D'autre part, le cardinal de l'ensemble des ultra-préordonnances à deux classes est égal au nombre des partitions de X en $2, 3, \dots, n-1$ classes et par conséquent (cf [4])

$$|P^+(X)| = S(n,2) + S(n,3) + \dots + S(n,n-1).$$

Par exemple pour $n \leq 8$ on obtient

n	nombre de préord. du niveau $n(n-1)/2 - 2$	nombre d'ultra- préord. à 2 classes
3	6	3
4	62	13
5	1 022	50
6	32 766	201
7	2 097 150	875
8	268 435 452	4 138

Notre problème était de trouver une ultra-préordonnance à deux classes $\lesssim^+$ qui soit "proche" d'une préordonnance donnée $\lesssim$ (au sens de la structure R de demi-treillis de $P(X)$). La solution consiste à minimiser $d(\lesssim, \lesssim^+)$, d étant une distance sur $P(X)$ qui respecte cette structure.

Comme dans [6] nous allons utiliser

$$d(\lesssim_1, \lesssim_2) = 2g(\sup\{\lesssim_1, \lesssim_2\}) - g(\lesssim_1) - g(\lesssim_2) \quad *$$

4. L'APPROXIMATION

Soit $\lesssim \neq \sup P(X)$ une préordonnance qui n'est pas élément de $P^+(X)$. On a donc $g(\lesssim) = n(n-1)/2 - m$ et $2 \leq m \leq n(n-1)/2$. Pour $\lesssim^+ \in P^+(X)$ on peut démontrer que

$$d(\lesssim, \lesssim^+) = \begin{cases} m-2 & , \quad \lesssim R \lesssim^+ \\ m & , \quad \lesssim \not R \lesssim^+ . \end{cases}$$

En effet, si $\lesssim R \lesssim^+$, alors $\sup\{\lesssim, \lesssim^+\} = \lesssim^+$, d'où $d(\lesssim, \lesssim^+) = g(\lesssim^+) - g(\lesssim) = m - 2$.

* Le graphe de $\sup\{\lesssim_1, \lesssim_2\}$ est la fermeture transitive de $G_{\lesssim_1} \cup G_{\lesssim_2}$.

Réciproquement, $\lesssim \nmid \lesssim^+$ implique $\sup \{\lesssim, \lesssim^+\} = \sup P(X)$, ce qui entraîne $d(\lesssim, \lesssim^+) = 2(n(n-1)/2 - 1) - (n(n-1)/2 - m) - (n(n-1)/2 - 2) = m$.

Avec $L := \{\lesssim^+ \mid \lesssim^+ \in P^+(X) \text{ et } \lesssim R \lesssim^+\}$ il faut donc considérer les deux cas $L = \emptyset$ et $L \neq \emptyset$.

Si $L = \emptyset$, toutes les ultra-préordonnances à deux classes ont la même distance à la préordonnance donnée, et par conséquent il n'y a pas de classification significative de X .

Le cas $L \neq \emptyset$ indique qu'il existe certaines partitions de X qui approchent $\lesssim$ mieux que toutes les autres partitions. On peut trouver ces éléments de L en examinant, pour chaque $\lesssim^+ \in P^+(X)$, si la propriété $\lesssim R \lesssim^+$ est vérifiée.

Mais, comme $|P^+(X)| = \sum_{j=2}^n S(n, j)$, il est impossible d'exécuter cet examen face à un nombre, disons, de $n \geq 100$.

Toutefois, il est possible d'établir un algorithme qui calcule L en $n(n-1)(n-2)/6$ opérations au maximum. Dans ce but rappelons que chaque élément de L est une ultra-préordonnance à deux classes.

L'ensemble $M := \{\lesssim^* \mid \lesssim^* \text{ ultra-préordonnance et } \lesssim R \lesssim^*\}$ possède un minimum noté $\lesssim'$. Cette relation $\lesssim'$ peut être calculée en commençant par $\lesssim' := \lesssim$ et en transformant $\{x, y\} \lesssim' \{y, z\} \prec' \{x, z\}$ en $\{x, y\} \lesssim' \{y, z\} \sim' \{x, z\}$ pour tout $x, y, z \in X$: $x \neq y$, $x \neq z$, $y \neq z$ (cf [6]).

Or, $g(\lesssim') \leq n(n-1)/2 - 2$ est équivalent à $L \neq \emptyset$. Si, par exemple, $g(\lesssim') = n(n-1)/2 - 2$, on a $L = \{\lesssim'\}$. D'autre part, si $g(\lesssim') = n(n-1)/2 - 2 - 1$ ($1 \leq l \leq m - 2$), il y a $l+2$ classes d'équivalence suivant $\sim'$, à savoir $C_1, C_2, \dots, C_{l+2}$.

Sans perte de généralité supposons que ces classes sont numérotées de manière que $i < j$, $\{x, y\} \in C_i$ et $\{u, v\} \in C_j$ implique $\{x, y\} \prec' \{u, v\}$. Les éléments de L sont donc les ultra-préordon-

nances $\lesssim^+$ telles que

$$Y/\sim^+ = \{E_1, E_2\} = \{C_1 \cup \dots \cup C_j, C_{j+1} \cup \dots \cup C_{l+2}\} \quad (j=1, 2, \dots, l+1).$$

Par comparaison des résultats des problèmes (1) et (2) on constate que les partitions $f(\lesssim^+)$ utilisées pour (2) comme classifications de X sont exactement celles qui forment la hiérarchie optimale du problème (1).

5. ALGORITHME ET EXEMPLE

Pour simplifier la description de l'algorithme nous posons $X = \{1, 2, \dots, n\}$; soient de plus $C_1, C_2, \dots, C_q$ les classes d'équivalence suivant $\sim$, qui seront de nouveau numérotées comme ci-dessus. Nous calculons les deux classes E_1, E_2 suivant $\sim^+$:

(A)

$x := 1, y := 2, z := 3.$

Aller à (B).

(B)

Calculer $i, j, k \in \{1, \dots, q\}$ de manière que $\{x, y\} \in C_i, \{y, z\} \in C_j$ et $\{x, z\} \in C_k$.

$i_1 := \min\{\max\{i, j\}, \max\{i, k\}, \max\{j, k\}\}, i_2 := \max\{i, j, k\}.$

Si $i_1 = i_2$, aller à (C).

$$C_{i_1} := \bigcup_{v=0}^{i_2-i_1} C_{i_1+v}$$

Si $i_2 < q$ alors $C_{i_1+v} := C_{i_2+v}$ pour $v \in \{1, 2, \dots, q-i_2\}.$

$q := q - i_2 + i_1.$

Aller à (C).

(C)

Si $q = 1$ STOP: Il n'y a pas de classification significative de X .

Si $z < n$ alors $z := z+1$, aller à (B)

Si $y < n-1$ alors $y := y+1, z := y+1$, aller à (B).

Si $x < n-2$ alors $x := x+1$, $y := x+1$, $z := x+2$, aller à (B).
Aller à (D).

(D)

Pour $j \in \{1, \dots, q-1\}$:

$$E_1 := \bigcup_{v=1}^j C_v, \quad E_2 := \bigcup_{v=j+1}^q C_v.$$

STOP.

Si, par exemple $X = \{1, \dots, 6\}$ et si $\preceq$ est la relation suivante

$$\{1,6\} \prec \{1,4\} \sim \{4,6\} \prec \{2,3\} \prec \{1,3\} \sim \{5,6\} \prec \{1,2\} \sim \{2,4\} \sim \{3,4\} \sim \{2,5\} \sim \{3,5\} \sim \{4,5\} \sim \{2,6\} \sim \{3,6\} \sim \{1,5\}$$

(cf [5], p. 55)

alors l'algorithme commence par $q = 6$ et les classes

$$C_1 = \{\{1,6\}\}$$

$$C_2 = \{\{1,4\}, \{4,6\}\}$$

$$C_3 = \{\{2,3\}\}$$

$$C_4 = \{\{1,3\}, \{5,6\}\}$$

$$C_5 = \{\{1,2\}, \{2,4\}, \{3,4\}, \{2,5\}, \{3,5\}, \{4,5\}, \{2,6\}\}$$

$$C_6 = \{\{3,6\}, \{1,5\}\}$$

et s'arrête avec $q = 4$ et

$$C_1 = \{\{1,6\}\}$$

$$C_2 = \{\{1,4\}, \{4,6\}\}$$

$$C_3 = \{\{2,3\}\}$$

$$C_4 = \{\{1,2\}, \{1,3\}, \{1,5\}, \{2,4\}, \{2,5\}, \{2,6\}, \{3,4\}, \{3,5\}, \{3,6\}, \{4,5\}, \{4,6\}\}$$

ce qui donne les trois partitions possibles

$$\{\{1,6\}, \{2\}, \{3\}, \{4\}, \{5\}\}$$

$$\{\{1,4,6\}, \{2\}, \{3\}, \{5\}\}$$

$$\{\{1,4,6\}, \{2,3\}, \{5\}\}.$$

BIBLIOGRAPHIE

- [1] BARBUT, M., MONJARDET, B., Ordre et classification, Paris, Hachette, 1970.
- [2] BIRKHOFF, G., Lattice theory, Providence, American Mathematical Society, 1973.
- [3] COMYN, G., VAN DORPE, J.C., "Valuation et semi-modularité dans les demi-treillis", Mathématiques et Sciences humaines, 56, 1976, 63-75.
- [4] EISEN, M., Elementary combinatorial analysis, New York, Gordon and Breach, 1969.
- [5] LERMAN, I.C., Les bases de la classification automatique, Paris, Gauthier-Villars, 1970.
- [6] SCHADER, M., "Hierarchical Analysis: Classification with Ordinal object dissimilarities", à paraître dans Metrika, 1979.