MATHÉMATIQUES ET SCIENCES HUMAINES

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>

Math. Sci. hum. (17^{e} année, n°67, *1979*, p.39-47)

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 $\precsim$ 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 $\precsim$ donnée sur X,

(1) une hiérarchie de X

(c.à.d. une famille $P_0, P_1, \ldots, P_q$ 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 $\precsim$ sur $Y := \{\{x,y\} \mid x,y \in X \text{ et } x \neq y\}$. $\{x,y\} \prec \{u,v\}$ 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 $\precsim'$ sur X qui satisfont à $\{x,z\} \precsim' \{x,y\}$ ou $\{x,z\} \precsim' \{y,z\}$ 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 $\precsim'$ correspondant à $(P_i)_i$.

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

$$\precsim_1 \; R \; \precsim_2 \; : \Longleftrightarrow \; G_{\precsim_1} \subset G_{\precsim_2} \; .^{*}$$

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

$$d(\precsim_1 , \precsim_2) := 2g(\sup\{\precsim_1 , \precsim_2\}) - g(\precsim_1) - g(\precsim_2)$$

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]

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

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

---

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

2. PARTITION ET ULTRA-PREORDONNANCE A DEUX CLASSES

Une ultra-préordonnance $\precsim^+$ 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 $\precsim^+$ 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 $\{x,y\} \prec^+ \{u,v\}$. Désignons par S la relation suivante sur X

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

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 ($\{x,z\} \precsim^+ \{x,y\}$ ou $\{x,z\} \precsim^+ \{y,z\}$) entraînent $\{x,z\} \in E_1$.

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

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 $\precsim^+$ caractérisée par $f(\precsim^+) = P$:

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

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

---

* 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:

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

où n désigne le nombre d'objets.

Il résulte, que les ultra-préordonnances à deux classes $\precsim^+$ (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 $\precsim \in P(X)$ et $g(\precsim) = n(n-1)/2 - 2$ 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.
$|\{ \precsim | \precsim \in P(X) \text{ et } g(\precsim) = n(n-1)/2 - 2 \}| = 2^{n(n-1)/2} - 2$.

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,...,n-1 classes et par conséquent (cf [4])

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

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

Comme dans [6] nous allons utiliser

$$d(\precsim_1 , \precsim_2) = 2g(\sup\{\precsim_1 , \precsim_2\}) - g(\precsim_1) - g(\precsim_2) \quad *$$

4. L'APPROXIMATION

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

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

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

---

* Le graphe de $\sup\{\precsim_1 , \precsim_2\}$ est la fermeture transitive de $G_{\precsim_1} \cup G_{\precsim_2}$.

Réciproquement, $\precsim \not R \precsim^+$ implique $\sup\{\precsim, \precsim^+\} = \sup P(X)$, ce qui entraîne $d(\precsim, \precsim^+) = 2(n(n-1)/2 - 1) - (n(n-1)/2 - m) - (n(n-1)/2 - 2) = m$.

Avec $L := \{\precsim^+ \mid \precsim^+ \in P^+(X)$ et $\precsim R \precsim^+\}$ 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 $\precsim$ mieux que toutes les autres partitions. On peut trouver ces éléments de L en examinant, pour chaque $\precsim^+ \in P^+(X)$, si la propriété $\precsim R \precsim^+$ 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 $M := \{\precsim^* \mid \precsim^*$ ultra-préordonnance et $\precsim R \precsim^*\}$ possède un minimum noté $\precsim'$. Cette relation $\precsim'$ peut être calculée en commençant par $\precsim' := \precsim$ et en transformant $\{x,y\} \precsim' \{y,z\} \prec' \{x,z\}$ en $\{x,y\} \precsim' \{y,z\} \sim' \{x,z\}$ pour tout $x,y,z \in X$: $x \neq y$, $x \neq z$, $y \neq z$ (cf [6]).

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

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
$Y/\sim^+ = \{E_1,E_2\} = \{C_1 \cup \ldots \cup C_j,\ C_{j+1} \cup \ldots \cup C_{l+2}\}$ $(j=1,2,\ldots,l+1)$.
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 chie optimale du problème (1).

5. ALGORITHME ET EXEMPLE

Pour simplifier la description de l'algorithme nous posons $X = \{1,2,\ldots,n\}$; soient de plus $C_1,C_2,\ldots,C_q$ 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 $i,j,k \in \{1,\ldots,q\}$ 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).

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

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

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

STOP.

Si, par exemple $X = \{1,\ldots,6\}$ et si $\precsim$ 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.