REVUE DE STATISTIQUE APPLIQUÉE

SAID CHAH

Comparaisons par triplets en classification automatique

Revue de statistique appliquée, tome 34, nº 1 (1986), p. 61-79
<http://www.numdam.org/item?id=RSA_1986__34_1_61_0>

© Société française de statistique, 1986, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) 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/

COMPARAISONS PAR TRIPLETS EN CLASSIFICATION AUTOMATIQUE

Said CHAH

Centre Scientifique IBM-France
36 Av. R. Poincaré 75116 PARIS

1. INTRODUCTION

En agrégation des similarités [7] (méthodes de classification sur des données qualitatives) l'information utilisée en vue d'effectuer la classification se présente sous forme de m tableaux des comparaisons par paires : pour chaque paire {i, j} on retient le fait que les éléments i et j « se ressemblent ou non » relativement à chacune des m variables qualitatives V_k ($k = 1, m$) décrivant l'ensemble E.

En théorie d'agrégation des préordonnances (classification de donnés hétérogènes) [2], les éléments de l'ensemble à classifier sont considérés par quadruplés : pour chaque quadruplé (i, j, k, l) on retient celle des pairs {i, j} et {k, l} dont les éléments « se ressemblent » le plus relativement à chacune des variables de description. Partant de ce type d'informations les méthodes d'agrégation des préordonnances permettent de classifier un ensemble décrit par des variables hétérogènes.

Dans le cas d'un ensemble E homogène muni d'un indice de proximité S (une distance ou un indice de similarité) l'information précédente s'écrit sous la forme suivante : pour chaque quadruplé (i, j, k, l) on retient le fait que les éléments de la paire {i, j} sont plus ressemblants ou moins ressemblants, au sens de l'indice S, que ceux de la paire {k, l}.

Dans cette publication une nouvelle structure basée sur des comparaisons par triplets sera proposée. Elle peut être décrite de la manière suivante : **Pour chaque triplet (i, j, k) d'éléments de l'ensemble à classifier on connaît celui des éléments i et k qui « ressemble le plus » à l'élément j relativement à chacune des variables de description.** Généralement l'information dont on dispose est constituée de m structures de ce type. Cette nouvelle approche permettra d'introduire de nouveaux critères de classification opérant soit sur des données hétérogènes, si l'ensemble à classifier est décrit par des variables, soit sur des tableaux de comparaisons par triplets lorsque les donnés se présentent directement sous cette forme.

Si l'ensemble à classifier est muni d'un indice de proximité S (S étant une distance si les variables de description sont quantitatives ou un indice de similarité si les variables sont qualitatives) cette information s'écrit tout simplement sous la forme suivante : **pour chaque triplet (i, j, k) on retient celui**

des éléments i et k qui ressemble le plus, au sens de l'indice S, à l'élément j. On montrera que la connaissance de cette nouvelle structure, dite triordonnance induite sur E par l'indice de similarité S, suffit pour établir une classification sur un ensemble homogène.

2. TYPOLOGIE SUR LES TRIORDONNANCES

2.1. Définition

Soit E un ensemble à n éléments, on désigne par F l'ensemble des paires d'éléments de E et par G le sous ensemble de F × F suivant :

$$G = \{(\{i, j\}, \{j, k\}) \mid i, j, k \in E\}$$

Une triordonnance sur E est une relation de préordre partiel, définie sur F, **comparant uniquement des éléments de G.** L'ensemble des triordonnances définies sur E sera noté $\Omega(E)$.

2.2. Codage d'une triordonnance

Pour représenter une triordonnance P, on utilisera le codage $[T_1]$ suivant :

$$T(i, j, k) = \begin{cases} 1 & \text{si} \quad \{i, j\} >_P \{j, k\} \\ 0 & \text{si} \quad \{i, j\} =_P \{j, k\} \\ -1 & \text{si} \quad \{i, j\} <_P \{j, k\} \end{cases}$$

On notera $\Phi(E)$ l'ensemble des représentants, selon le codage précédent, des éléments de $\Omega(E)$.

2.3. Typologie sur les triordonnances

L'objet de ce paragraphe est d'introduire les notions de triordonnances induites par, un indice de similarité, une relation binaire quelconque définie sur E, ou une variable de description de nature quelconque.

2.3.1. Triordonnance induite par un indice de similarité

S étant un indice de similarité défini sur E × E, on appelle triordonnance induite sur E par l'indice de similarité S, notée P_s, le préordre partiel portant sur G suivant :

$$(\{i, j\}, \{j, k\}) \in G \qquad \{i, j\}\ P_s\ \{j, k\} \iff S(i, j) \geqslant S(j, k)$$

La variable T_s représentant la triordonnance P_s selon le codage $[T_1]$ s'écrit :

$$T_s(i, j, k) = \begin{cases} 1 & \text{si} \quad S(i, j) > S(j, k) \\ 0 & \text{si} \quad S(i, j) = S(j, k) \\ -1 & \text{si} \quad S(i, j) < S(j, k) \end{cases}$$

2.3.2. Triordonnance induite par une relation binaire

R étant une relation binaire définie sur E, on appelle triordonnance sur E induite par la relation R le préordre partiel portant sur G suivant :

$$\{i, j\}\ P_R\ \{j, k\} \iff iRj \text{ et } j\not{R}k \text{ (ex aequo si } iRj \text{ et } jRk \text{ ou } i\not{R}j \text{ et } j\not{R}k)$$

La variable T_R associée à la triordonnance P_R selon le codage $[T_1]$ se présente sous la forme suivante :

$$T_R(i, j, k) = \begin{cases} 1 & \text{si } iRj \text{ et } j\not{R}k \\ 0 & \text{si } (iRj \text{ et } jRk) \text{ ou } (i\not{R}j \text{ et } j\not{R}k) \\ -1 & \text{si } i\not{R}j \text{ et } jRk \end{cases}$$

iRj désigne le fait que i n'est pas en relation avec j.

Remarque :

$$T_R(i, j, k) = Y(i, j) - Y(j, k) = \mathrm{Sign}(Y(i, j) - Y(j, k))$$

$$Y(i, j) = \begin{cases} 1 & \text{si } iRj \\ 0 & \text{sinon.} \end{cases}$$

L'ensemble des triordonnances induites sur E par l'ensemble des relations d'équivalence définies sur E (ou partitions de E) est un sous ensemble de $\Omega(E)$ que l'on notera Ω_e. Il lui correspond un sous ensemble de $\Phi(E)$ noté Φ_e (ensemble des représentants des triordonnances induites sur E par l'ensemble des partitions de E).

2.3.3. Triordonnance induite par une variable qualitative

Il s'agit de la triordonnance associée à la relation d'équivalence induite sur E par la variable qualitative V.

2.3.4. Triordonnances induites par une variable ordinale

A partir d'une variable ordinale V définie sur E, on peut extraire deux types d'informations : une information relative à la ressemblance entre éléments de E **(si $|r(i) - r(j)| < |r(k) - r(j)|$ alors j est plus proche de i que de k)** où r désigne l'application rang induite par V sur E, la seconde information concerne le classement (préordre) induit par V sur E. De ce fait à une variable ordinale V on peut associer deux triordonnances différentes, la première définie sur E par :

$$\{i, j\}\ P_v\ \{j, k\} \iff |r(i) - r(j)| \leqslant |r(j) - r(k)|$$

contient une information concernant la ressemblance entre les éléments de E. La variable T_v correspondante est donnée par :

$$T_v(i, j, k) = \begin{cases} 1 & \text{si } |r(i) - r(j)| < |r(j) - r(k)| \\ 0 & \text{si } |r(i) - r(j)| = |r(j) - r(k)| \\ -1 & \text{si } |r(i) - r(j)| > |r(j) - r(k)| \end{cases}$$

La deuxième triordonnance est celle induite sur E par la relation de préordre définie sur E par la variable V. Lors d'un problème de recherche de partitions c'est la première triordonnance que l'on retiendra comme information relative à la ressemblance des éléments de l'ensemble à classifier extraite de la

variable ordinale V. Dans le cas d'un problème de recherche d'un ordre optimale (agrégation d'ordres) c'est la deuxième triordonnance que l'on retiendra.

2.3.5. Triordonnance induite par une variable quantitative

V étant une variable quantitative définie sur E, on appelle triordonnance induite par V sur E le préordre partiel portant sur G suivant :

$$\{i, j\}\ P_v\ \{j, k\} \Longleftrightarrow |V(i) - V(j)| \leqslant |V(j) - V(k)|$$

où V(i) désigne la valeur prise par la variable V pour l'élément i de E. La triordonnance P_v induite par la variable quantitative V est représentée (cas du codage T_1) par la variable T_v suivante :

$$T_v(i, j, k) = \begin{cases} 1 & \text{si} \quad |V(i) - V(j)| < |V(j) - V(k)| \\ 0 & \text{si} \quad |V(i) - V(j)| = |V(j) - V(k)| \\ -1 & \text{si} \quad |V(i) - V(j)| > |V(j) - V(k)| \end{cases}$$

Remarque :

n désigne le cardinal de E.

m désigne le nombre des variables de description.

$N = \sum_i \sum_j 1 = n^2$: si l'on considère tous les couples (i, j).

$N = \sum_{i \leqslant j} \sum 1 = \frac{n(n+1)}{2}$: si l'on considère toutes les paires {i, j}.

$N = \sum_{i < j} \sum 1 = \frac{n(n-1)}{2}$: si l'on ne considère que les paires $\{i, j\} \mid i \neq j$.

$M = \sum_i \sum_j \sum_k 1 = n^3$: si l'on considère tous les triplets (i, j, k).

2.3.6. Remarque importante

Dans le cas d'un problème où les couples (i, j) et (j, i) ne jouent pas le même rôle (voir exemple à la fin de l'article) la triordonnance doit être définie comme une relation de préordre partiel définie sur l'ensemble F **des couples** d'éléments de E, comparant uniquement les éléments de l'ensemble G suivant :

$$G = \{((i, j),(j, k)) \mid i, j, k \in E\}$$

Cela revient à donner aux notations {i, j} utilisées dans cet article le sens de couple et non de paire.

3. CRITÈRES D'ADÉQUATION ENTRE TRIORDONNANCES

Deux approches différentes seront proposées en vu d'introduire des mesures d'adéquation entre triordonnances. Une première approche, dite corrélationnelle, permet de construire deux mesures d'adéquation : l'une

induite par la covariance et l'autre par le coefficient de corrélation. La seconde approche est contingentielle, elle permet d'associer à chaque critère de contingence (2×2) une mesure d'adéquation entre Triordonnances.

3.1. Approche corrélationnelle du problème

Etant données deux triordonnances P et Q, on désigne par T_p et T_q les variables ternaires représentant P et Q (selon le codage T_1), soit :

$$T_p(i, j, k) = \begin{cases} 1 & \text{si} \quad \{i, j\} >_p \{j, k\} \\ 0 & \text{si} \quad \{i, j\} =_p \{j, k\} \\ -1 & \text{si} \quad \{i, j\} <_p \{j, k\} \end{cases}$$

$$T_q(i, j, k) = \begin{cases} 1 & \text{si} \quad \{i, j\} >_q \{j, k\} \\ 0 & \text{si} \quad \{i, j\} =_q \{j, k\} \\ -1 & \text{si} \quad \{i, j\} <_q \{j, k\} \end{cases}$$

3.1.1. Critère d'adéquation induit par la covariance

La covariance induit une mesure d'adéquation sur l'ensemble $[\Omega(E)]^2$ notée Ψ_{Cov} donnée par :

$$\Psi_{Cov}(P, Q) = \text{Cov}(T_p, T_q) = \frac{\sum\sum\sum T_p(i, j, k)\, T_q(i, j, k)}{M}$$

3.1.2. Critère d'adéquation induit par le coefficient de corrélation

Au coefficient de corrélation correspond une mesure d'adéquation entre triordonnances définie par :

$$\Psi_{Cor}(P, Q) = \text{Cor}(T_p, T_q) = \frac{\text{Cov}(T_p, T_q)}{\sqrt{\text{Var}(T_p)}\,\sqrt{\text{Var}(T_q)}}$$

$$= \frac{\sum\sum\sum T_p(i, j, k)\, T_q(i, j, k)}{\sqrt{\sum\sum\sum T_p^2(i, j, k)}\,\sqrt{\sum\sum\sum T_q^2(i, j, k)}}$$

3.2. Approche contingentielle du problème

Pour construire des mesures d'adéquation entre triordonnances une seconde approche de nature contingentielle peut être utilisée. Elle consiste, dans un premier temps, à considérer que les deux triordonnances P et Q ne comportent pas d'ex-aequo, à construire la table de contingence (2×2) croisant les modalités des variables binaires T_p et T_q (représentant les deux triordonnances P et Q), et à mesurer l'adéquation entre les triordonnances P et Q par l'assocation, au sens d'un critère de contingence (2×2), qui existe entre les variables T_p et T_q. Ces mêmes critères seront utilisés pour mesurer l'adéquation entre deux triordonnances quelconques (avec ex-æquo). Pour ce faire, il suffit de poser $T_p(i, j, k) = 0$ si $\{i, j\} =_p \{j, k\}$ et $T_q(i, j, k) = 0$ si $\{i, j\} =_q \{j, k\}$. **Ce qui revient à annuler l'effet des ex-æquo sur les critères considérés.**

La table de contingence, en question, se présente comme suit :

		T_q		
		1	-1	
T_p	1	n_{11}	n_{1-1}	$n_{1.}$
	-1	n_{-11}	n_{-1-1}	$n_{-1.}$
		$n_{.1}$	$n_{.-1}$	M

En fonction des notations $T_p(i, j, k)$, $T_q(i, j, k)$, les quantités n_{11}, n_{1-1}, $n_{1.}$ ...etc, sont données par :

$$4n_{11} = \sum_i \sum_j \sum_k [1 + T_p(i, j, k)\, T_q(i, j, k)] = 4n_{-1-1}$$

$$4n_{1-1} = \sum_i \sum_j \sum_k [1 - T_p(i, j, k)\, T_q(i, j, k)] = 4n_{-11}$$

$$n_{1.} = n_{.1} = n_{.-1} = n_{-1.} = \frac{M}{2}$$

Tout critère de contingence, défini sur le tableau précédent, donne lieu à une mesure d'adéquation entre la triordonnance P et la triordonnance Q. De sorte que, si l'on désigne par C un indice d'association quelconque sur une table de contingence (2 × 2) : BELSON, PEARSON, RAND, JORDAN... etc., il lui correspond une mesure d'adéquation entre triordonnances notée Ψ_C et définie par :

$$\Psi_C(P, Q) = C(T_p, T_q)$$

La formulation en fonction des notations $T_p(i, j, k)$ et $T_q(i, j, k)$ des critères d'adéquation entre triordonnances obtenus par le biais de l'approche contingentielle, ainsi que l'étude des équivalences entre ces critères ont été effectuées par l'auteur dans [3].

4. AGRÉGATION DES TRIORDONNANCES

Etant donné un ensemble de m triordonnances définies sur E, notées $P_r (r = 1, m)$, le problème d'agrégation des triordonnances consiste à calculer parmi toutes les triordonnances d'un type donné noté Ω_t celles qui s'ajustent le mieux, au sens d'une mesure d'adéquation entre triordonnances, à l'ensemble des triordonnances $P_r (r = 1, m)$ données.

Si l'on désigne par Ψ_C une mesure d'adéquation entre triordonnances et par Φ_t l'ensemble des représentants selon le codage $[T_l]$ des éléments de Ω_t, le problème général d'agrégation des triordonnances, au sens du critère Ψ_C, s'écrit de la manière suivante :

$$\operatorname*{Max}_{P \in \Omega_t} \sum_r \Psi_C(P, P_r) \; <=> \; \operatorname*{Max}_{T \in \Phi_t} \sum_r C(T, T_r)$$

où C désigne soit la covariance ou le coefficient de corrélation dans le cas de l'approche corrélationnelle soit un critère de contingence dans le cas de l'approche contigentielle. Les triordonnances $P_r (r = 1, m)$ sont les données

du problème et P parcourt l'ensemble Ω_t (donné). Il existe donc autant de problèmes d'agrégation de triordonnances que de critères d'adéquation entre triordonnances différents.

5. COMPARAISONS PAR TRIPLETS EN CLASSIFICATION DE DONNÉES HOMOGÈNES

5.1. Problème de classification

E étant un ensemble à n éléments décrits par m variables homogènes et muni d'un indice de similarité noté S, le problème que l'on se pose consiste à effectuer une classification automatique (reconnaissance automatique d'une structure de partition) sur E. Pour ce faire nous proposons la démarche suivante.

On calculera parmi toutes les partitions possibles de E celles dont les triordonnances associées s'ajustent le mieux, au sens d'un critère d'adéquation entre triordonnances, à la triordonnance P_s, induite par l'indice de proximité S. S pouvant être ou un indice de similarité si les variables de description sont qualitatives ou une distance si les variables sont quantitatives.

Le modèle proposé apparaît comme une application particulière du problème général d'agrégation des triordonnances : l'ensemble des triordonnances à ajuster est réduit à l'unique triordonnance P_s, celles que l'on recherche sont du type Ω_e. Au sens d'un critère du type Ψ_C, le problème d'optimisation associé à la recherche des partitions s'ajustant le mieux à la triordonnance P_s s'écrit de la manière suivante :

$$\underset{P \in \Omega_e}{\text{Max}}\ \Psi_C(P, P_s)$$

De ce fait il existe, a priori, autant de critères de classification, fondés sur les triordonnances et opérant sur des données homogènes, que de mesures d'adéquation différentes entre triordonnances. Commençons par étudier le problème au sens des mesures Ψ_{Cov} et Ψ_{Cor}.

5.2. Critère induit par la covariance

Le problème de la recherche des triordonnances du type Ω_e qui s'ajustent le mieux, au sens de la mesure Ψ_{Cov}, à la triordonnance P_s s'écrit :

$$\underset{P \in \Omega_e}{\text{Max}}\ \Psi_{Cov}(P, P_s) \ <\ =\ >\ \underset{T \in \Phi_e}{\text{Max}}\ \text{Cov}(T, T_S)$$

T_s est une donnée du problème et T parcourt l'ensemble Φ_e (des représentants des triordonnances induites par les partitions de E). Si l'on désigne par Y la variable binaire représentant une partition R de E, alors :

$$T(i, j, k) = Y(i, j) - Y(j, k)$$

$$M\,E(T) = \sum_i \sum_j \sum_k T(i, j, k) = \sum_i \sum_j \sum_k [Y(i, j) - Y(j, k)] = 0.$$

$$
\begin{aligned}
2M\,E(T_s) &= \sum_i\sum_j\sum_k T_s(i,j,k) + \sum_i\sum_j\sum_k T_s(k,j,i) \\
&= \sum_i\sum_j\sum_k [T_s(i,j,k) + T_s(k,j,i)] = 0. \\
M\,Cov(T,T_s) &= \sum_i\sum_j\sum_k T(i,j,k)\,T_s(i,j,k) = \sum_i\sum_j\sum_k Y(i,j)\,T_s(i,j,k) \\
&- \sum_i\sum_j\sum_k Y(j,k)\,T_s(i,j,k) = \sum_i\sum_j\sum_k Y(i,j)\,T_s(i,j,k) \\
&- \sum_i\sum_j\sum_k Y(i,j)\,T_s(k,i,j) \\
&= \sum_i\sum_j \left[\sum_k T_s(i,j,k) - \sum_k T_s(k,i,j)\right] Y(i,j).
\end{aligned}
$$

Comme T_s parcourt l'ensemble des triordonnances induites par les partitions de E, Y parcourt l'ensemble des partitions de E. Cela donne lieu à la proposition suivante.

Proposition

Le problème de la recherche des partitions s'ajustant le mieux, au sens du critère Ψ_{Cov} à la triordonnance P_s est équivalent au programme linéaire suivant :

$$
\begin{aligned}
&\text{Max} \sum_{i\neq j}\sum \left[\sum_k T_s(i,j,k) - \sum_k T_s(k,i,j)\right] Y(i,j) \\
&Y(i,j) = Y(j,i) \\
&Y(i,j) + Y(j,k) - Y(i,k) \leqslant 1 \qquad i \neq j \neq k \\
&Y(i,j) \in \{0,1\}
\end{aligned}
$$

La résolution du programme ci-dessus relève les mêmes algorithmes utilisés en agrégation des similarités [7].

5.2.1. Interprétation du critère obtenu

Corollaire

Le programme linéaire précédent s'écrit de la manière suivante :

$$
\begin{cases}
\underset{Y\ \text{Partition}}{\text{Max}} \sum\sum [B(i,j) - \overline{B}(i,j)]\,Y(i,j) \\
B(i,j) = 2b(i,j) + f(i,j) \\
\overline{B}(i,j) = 2\overline{b}(i,j) + \overline{f}(i,j)
\end{cases}
$$

$$
\begin{aligned}
b(i,j) &= \text{Card}\,\{k \mid k \in E \quad \text{et} \quad S(i,j) > \max\{S(i,k), S(j,k)\}\} \\
\overline{b}(i,j) &= \text{Card}\,\{k \mid k \in E \quad \text{et} \quad S(i,j) < \min\{S(i,k), S(j,k)\}\} \\
f(i,j) &= \text{Card}\,\{k \mid k \in E \quad \text{et} \quad S(i,j) = \max\{S(i,k), S(j,k)\}\} \\
\overline{f}(i,j) &= \text{Card}\,\{k \mid k \in E \quad \text{et} \quad S(i,j) = \min\{S(i,k), S(j,k)\}\}
\end{aligned}
$$

Si l'on désigne par :

B_i la boule (ouverte) de centre i et de diamètre la dissimilarité entre i et j ($d(i,j) = 1\text{-}S(i,j)$). B_j la boule (ouverte) de centre j et de diamètre la

dissimilarité entre i et j. Fr $(B_i \cup B_j)$ l'ensemble des éléments de E appartenant à la frontière de $B_i \cup B_j$. Fr $(B_i \cap B_j)$ l'ensemble des éléments de E appartenant à la frontière de $B_i \cap B_j$

alors :

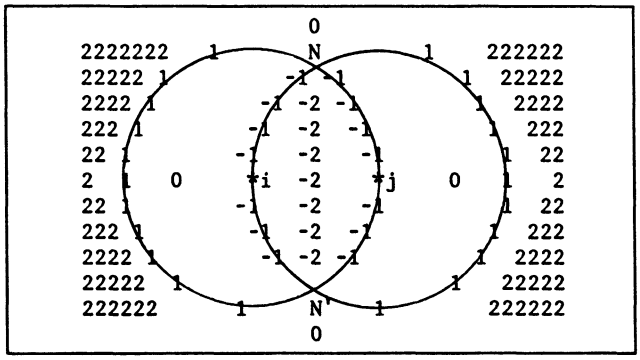
$$b(i,j) = n - [\text{card}\,\{B_i \cup B_j\} + \text{card}\,\{\text{Fr}(B_i \cup B_j)\}]$$

$$\overline{b}(i,j) = \text{Card}\,\{B_i \cap B_j\}$$

$$f(i,j) = \text{Card}\,\{\text{Fr}(B_i \cup B_j)\}$$

$$\overline{f}(i,j) = \text{Card}\,\{\text{Fr}(B_i \cap B_j)\}$$

soit :



Autrement dit :

La contribution d'un observateur situé à l'extérieur des boules B_i et B_j est + 2.
La contribution d'un observateur situé sur la frontière de $(B_i \cup B_j)$ est + 1.
La contribution d'un observateur situé à l'intérieur de $(B_i \cap B_j)$ est − 2.
La contribution d'un observateur situé sur la frontière de $(B_i \cap B_j)$ est − 1.
La contribution des autres éléments est nulle.

Notons enfin que les deux points N et N′ appartiennent à la fois à f (i, j) et $\overline{f}$(i, j) par conséquent leurs contributions sont nulles.

5.2.2. Règles des comparaisons par paires correspondantes

En l'absence de la contrainte linéaire « Y partition », la variable Y* qui réalise le maximum de la fonction économique :

$$\sum_{i \neq j}\sum [B(i,j) - \overline{B}(i,j)]\, Y$$

s'obtient par simple application de la règle des comparaisons par paires suivante :

$$B(i,j) > \overline{B}(i,j) \quad \text{alors} \quad Y^*(i,j) = 1.$$

A cause des effets du type effet Condorcet :

$$B(i,j) > \overline{B}(i,j) \quad \text{et} \quad B(j,k) > \overline{B}(j,k) \quad \text{mais} \quad B(k,i) < \overline{B}(k,i)$$

la solution Y*, ainsi obtenue, n'est pas une partition, d'où la nécessité d'introduire la contrainte « Y partition ».

S'il n'y a pas d'ex-aequo ($f(i, j) = \overline{f}(i, j) = 0$), alors la règle des comparaisons par paires précédente peut être énoncée de la manière suivante :

si

le nombre d'éléments de E situés à l'extérieur des boules B_i et B_j est supérieur au nombre d'éléments de E contenus dans l'intersection des boules B_i et B_j

alors

i et j sont réunis dans la partition optimale

5.2.3. Sur quelques aspects du critère

La règle des comparaisons par paire qui découle du critère (5.2) fait des éléments de l'ensemble à classifier tantôt des observateurs (juges) tantôt des points observés (éléments soumis à un jugement). En effet ces mêmes éléments sont consultés chaque fois qu'il s'agit de classifier deux éléments i et j (la question étant : i et j se ressemblent-ils « suffisamment » pour les réunir dans la partition optimale). En d'autres termes le jury est constitué des éléments de l'ensemble à classifier eux-mêmes. Quant à la logique du jugement émis par chacun des éléments consultés, elle s'explique clairement par le fait que **le jugement émis par un observateur dépend de sa position par rapport aux éléments observés i et j.** Un observateur k dont les distances qui le séparent des points observés i et j sont supérieures à celle qui existe entre i et j est favorable au classement de i avec j (puisqu'il se trouve éloigné aussi bien de i que de j). Tandis qu'un observateur dont les distances qui le séparent des deux points observés i et j sont inférieures à celle qui existe entre i et j est défavorable à ce classement (puisqu'il se trouve « proche » aussi bien de i que de j). Enfin, la faible contribution (1 ou -1 au lieu de 2 et -2) des éléments situés sur les frontières (cas limites) s'explique par leur « impartialité ». Notre observateur est d'autant plus sûr de son jugement que la distance qui le sépare des éléments observés i et j est grande (les critères (5.2 et 5.2.1.) ne tiennent pas compte de ce fait, dans (5.4) on construit un critère qui tient compte de l'éloignement de l'observateur). Si la somme des contributions des éléments favorables au classement de i avec j est supérieure à la somme des contributions des éléments non favorables à ce classement (aspect majoritaire du critère) alors i et j devraient (sauf pour quelques exceptions dont on minimisera le nombre, qui sont inévitables par suite des effets condorcet) être réunis dans la partition optimale.

Remarque :

$$\Psi_{Cov}(P, Q) = \frac{\sum \sum \sum \text{Sign}\,[Y(i, j) - Y(j, k)]\,\text{Sign}\,[S(i, j) - S(j, k)]}{M}$$

5.3. Critère induit par le coefficient de corrélation

Au sens de la mesure Ψ_{Cor} le problème du calcul des partitions s'ajustant le mieux à la triordonnance P_s s'écrit de la manière suivante :

$$\underset{P \in \Omega_e}{\text{Max}}\ \Psi_{Cor}(P, P_s) <=> \underset{T \in \Phi_e}{\text{Max}}\ \text{Cor}(T, T_s)$$

Proposition

La recherche des partitions s'ajustant le mieux à la triordonnance P_s, au sens de la mesure Ψ_{Cor}, donne lieu au problème d'optimisation suivant :

$$\underset{T \in \Phi_\varepsilon}{\text{Max}} \frac{\sum\sum\sum T(i,j,k)\, T_s(i,j,k)}{\sqrt{\sum\sum\sum T^2(i,j,k)}\ \sqrt{\sum\sum\sum T_s^2(i,j,k)}}$$

5.4. Critères tenant compte de l'éloignement d'un observateur

5.4.1. Poids d'un observateur éloigné

Revenons sur le critère :

$$\sum_i \sum_j [B(i,j) - \overline{B}(i,j)]\, Y(i,j)$$

Dans (5.2.3) nous avons vu que pour classifier deux éléments i et j (c'est-à-dire pour calculer $Y^*(i,j)$) tout se passe comme si l'on scrutait l'opinion d'un ensemble d'observateurs constitué par les éléments de E. Et que l'opinion émise par chaque observateur consulté dépend de sa position par rapport aux points observés i et j : l'opinion d'un observateur « éloigné » est positive, tandis que celle d'un observateur « proche » est négative. **L'observateur est d'autant plus sûr de son jugement que « la distance » qui le sépare des points observés i et j est grande.** Relativement au critère (5.3), deux observateurs k et k′ tels que k′ est plus « éloigné » que k de la paire {i, j}, ont la même contribution. Autrement dit ce critère ne tient pas compte du degré d'éloignement d'un observateur.

Le problème que l'on se pose est de construire un nouveau critère qui tienne compte du degré d'éloignement de chaque observateur. Pour ce faire il suffit de pondérer le critère :

$$\sum\sum\sum T(i,j,k)\, T_s(i,j,k)$$

par la quantité $|S(i,j) - S(j,k)|$. Le critère résultant s'écrit :

$$\sum\sum\sum |S(i,j) - S(j,k)|\, T(i,j,k)\, T_s(i,j,k)$$

montrons que le nouveau critère répond bien à nos espérances, pour cela considérons deux observateurs k et k′ tels que k′ est plus « éloigné » que k de la paire observée {i, j}, alors :

$S(i,j) > S(j,k) => T(i,j,k) = 1$ le coût étant égal à $S(i,j) - S(j,k)$

$S(i,j) > S(j,k') => T(i,j,k') = 1$ le coût étant égal à $S(i,j) - S(j,k')$

or :

$$S(j,k) > S(j,k') => S(i,j) - S(j,k') > S(i,j) - S(j,k).$$

Ce qui montre que le coût du triplet (i, j, k) pour la nouvelle fonction économique est moins important que celui du triplet (i, j, k′), c'est-à-dire que, relativement au nouveau critère, l'opinion de l'observateur k′ a plus de poids

que celle de l'observateur k. Alors que, relativement au critère initial les deux observateurs k et k' possèdent le même poids.

5.4.2. Critère correspondant

Proposition

Le problème d'optimisation relevant de la recherche des partitions de E qui maximisent le critère :

$$\sum \sum \sum \mid S(i,j) - S(j,k) \mid T_s(i,j,k)\, T(i,j,k)$$

est équivalent au programme linéaire suivant :

$$\text{Max } 2n \sum_{i \neq j}\sum \left\{ S(i,j) - \left[\frac{S(i,.) + S(.,j)}{2n} \right] \right\} Y(i,j)$$

$$Y(i,j) = Y(j,i)$$

$$Y(i,j) + Y(j,k) - Y(i,k) \leqslant 1 \quad i \neq j \neq k$$

$$Y(i,j) \in \{0, 1\}$$

$$S(i,.) = \sum_k S(i,k),\ S(.,j) = \sum_k S(k,j)$$

6. COMPARAISONS PAR TRIPLETS EN CLASSIFICATION DE DONNÉES HÉTÉROGÈNES

6.1. Enoncé du problème

L'objet de ce paragraphe est de montrer que le problème général d'agrégation des triordonnances peut servir de modèle pour **classifier un ensemble E décrit par m variables hétérogènes.** Les variables de description peuvent être qualitatives, quantitatives ou ordinales. Pour ce faire nous proposons la démarche suivante :

A chacune des variables de description $Y_r (r = 1, m)$, qu'elle soit qualitative, quantitative ou ordinale, on associe la triordonnance P_r induite par Y_r sur E. De même à toute partition Y de E on associe la triordonnance P induite par Y sur E. En vue de classifier E on cherchera parmi toutes les partitions possibles de E celles dont les triordonnances induites sur E s'ajustent le mieux, au sens d'une mesure d'adéquation entre triordonnances du type Ψ_c, à l'ensemble des triordonnances $P_r (r = 1, m)$.

Le modèle proposé apparaît comme une application particulière du problème général d'agrégation des triordonnances : l'ensemble des triordonnances à ajuster est constitué par les m triordonnances $P_r (r = 1, m)$, induites par les m variables hétérogènes. Celles que l'on recherche sont du type Ω_e. Au sens d'un critère du type Ψ_c, le problème d'optimisation qui en découle s'écrit :

$$\underset{P \in \Omega_e}{\text{Max}} \sum_r \Psi_C(P_r, P)$$

A chaque critère d'adéquation entre triordonnances correspond donc un critère de classification, **à priori différent,** opérant sur des données hétérogènes. Dans

un premier temps nous proposons d'étudier le problème au sens des mesures Ψ_{Cov} et Ψ_{Cor}.

6.2. Critère hétérogène induit par la covariance

Le problème de la recherche des triordonnances du type Ω_e qui s'ajustent le mieux, au sens de la mesure Ψ_{Cov}, à l'ensemble des triordonnances $P_r (r = 1, m)$ s'écrit :

$$\underset{P \in \Omega_e}{\text{Max}} \sum_r \Psi_{Cov}(P, Pr) < = > \underset{T \in \Phi_e}{\text{Max}} \sum_r \text{Cov}(T, Tr)$$

$T_r (r = 1, m)$ sont les données du problème et T parcourt l'ensemble Φ_e (des représentants des triordonnances induites par les partitions de E).

Proposition

Le problème de la recherche des partitions s'ajustant le mieux, au sens du critère Ψ_{Cov}, à l'ensemble des triordonnances $P_r (r = 1, m)$ est équivalent au programme linéaire suivant :

$$\text{Max} \sum_{i \neq j}\sum \left\{ \sum_r \left[\sum_k T_r(i, j, k) - \sum_k T_r(k, i, j) \right] \right\} Y(i, j)$$

$$Y(i, j) = Y(j, i)$$

$$Y(i, j) + Y(j, k) - Y(i, k) \leqslant 1 \quad i \neq j \neq k$$

$$Y(i, j) \in \{0, 1\}$$

6.2.1. Interprétation du critère obtenu

Corollaire

Le programme linéaire précédent s'écrit de la manière suivante :

$$\underset{Y \text{ partition}}{\text{Max}} \sum_{i \neq j}\sum \left[\sum_r B_r(i, j) - \sum_r \overline{B}_r(i, j) \right] Y(i, j)$$

$$B_r(i, j) = 2b_r(i, j) + f_r(i, j)$$

$$\overline{B}_r(i, j) = 2\overline{b}_r(i, j) + \overline{f}_r(i, j)$$

$$b_r(i, j) = \text{Card}\,\{k \mid k \in E \quad \text{et} \quad \{i, j\} >_r \{i, k\}, \{i, j\} >_r \{j, k\}\}$$

$$\overline{b}_r(i, j) = \text{Card}\,\{k \mid k \in E \quad \text{et} \quad \{i, j\} <_r \{i, k\}, \{i, j\} <_r \{j, k\}\}$$

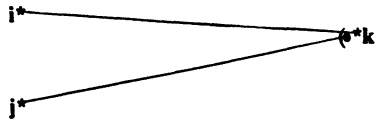
$$f_r(i, j) = \text{Card}\,\{k \mid k \in E \quad \text{et} \quad [\{i, j\} =_r \{i, k\} \quad \text{et} \quad \{i, j\} \geqslant_r \{j, k\}]$$

ou $[\{i, j\} \geqslant_r \{i, k\} \quad \text{et} \quad \{i, j\} =_r \{j, k\}]\}$

$$\overline{f}_r(i, j) = \text{Card}\,\{k \mid k \in E \quad \text{et} \quad [\{i, j\} =_r \{i, k\} \quad \text{et} \quad \{i, j\} \leqslant_r \{j, k\}]$$

ou $[\{i, j\} \leqslant_r \{i, k\} \quad \text{et} \quad \{i, j\} =_r \{j, k\}]\}$

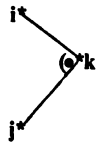
b_r correspond au nombre des configurations de ce type :



Rappelons que la contribution de chaque élément k se trouvant dans cette position est $+2$.

$f(i, j)$ correspond au cardinal de la « frontière » de l'ensemble constitué par les éléments de la configuration précédente. La contribution de chaque élément de cette frontière est $+1$.

$\overline{b}_r(i, j)$ correspond au nombre des configurations de ce type :



La contribution d'un élément se trouvant dans cette position est -2.

$\overline{f}(i, j)$ correspond au cardinal de la « frontière » de l'ensemble des éléments de la configuration précédente. La contribution des éléments de cette frontière étant -1.

Remarque

Les programmes linéaires (6.2) et (6.2.1) restent valables dans le cas ou le problème étudié contient des données manquantes : il suffit, pour chaque couple (i, j), de faire varier l'indice r sur les variables V_r pour lesquelles $V_r(i)$ et $V_r(j)$ sont données.

6.2.2. Calcul des contributions

Le coefficient $\sum_r [B_r(i, j) - \overline{B}_r(i, j)]$, coût de la fonction économique (6.2.1), représente la somme des contributions de chacune des variables de description ($B_r(i, j) - \overline{B}_r(i, j)$ est la contribution de la variable Y_r). Bien sûr on peut se contenter de l'expression :

$$\sum_k T_r(i, j, k) - \sum_k T_r(k, i, j)$$

pour calculer la contribution de la variable Y_r au coût de la fonction économique (6.2.1), cependant il serait intéressant d'interpréter concrètement la contribution de chaque type de variables. Pour ce faire supposons que E est décrit par trois variables : Y_1 quantitative, Y_2 ordinale et Y_3 qualitative.

Cas d'une variable quantitative

La première variable Y_1 étant numérique, on ordonne les points à classifier sur un axe selon leurs valeurs croissantes. La quantité

$|Y_1(i) - Y_1(j)|$ peut servir de dissimilarité entre i et j. On repère l'intervalle $[t, t']$ $(d(t, i) = d(i, j) = d(j, t'))$:

$$+^2 +^2 +^2 +^2 +^1 t] \text{———} i^* -_1 -_2 -_2 -_2 -_1$$
$$*j \text{———} [t' +^1 +^2 +^2 +^2 +^2 +^2 +^2 \text{———}$$

La contribution d'un élément situé entre i et j est négative (-2) : ces éléments ont une opinion négative quant au classement de i avec j. La contribution d'un élément situé sur la frontière de l'intervalle [i, j] est aussi négative mais moins faible $(-1$ au lieu de $-2)$ que celle des éléments situés à l'intérieur de l'intervalle. La contribution d'un élément situé à l'extérieur de l'intervalle $[t, t']$ est positive $(+2)$: ces éléments sont favorables au classement de i avec j. La contribution d'un élément situé sur la frontière du segment $[t, t']$ est aussi positive mais plus faible $(+1$ au lieu de $+2)$ que celle des éléments situés à l'extérieur de l'intervalle. Enfin les autres éléments ont une contribution nulle. De sorte que la contribution :

$$B_1(i, j) - \overline{B}_1(i, j)$$

de la variable Y_1 au coefficient $B(i, j) - B'(i, j)$, cas de l'exemple ci-dessus, s'écrit :

$$B_1(i, j) - \overline{B}_1(i, j) = [2 \times (4 + 6) + (1 + 1)] - [2 \times 3 + (1 + 1)] = 14.$$

Cas d'une variable ordinale

Le même procédé peut être utilisé pour expliciter le calcul de la contribution $B_2(i, j) - \overline{B}_2(i, j)$ de la variable ordinale V_2 au coefficient $B(i, j) - \overline{B}(i, j)$, il suffit pour cela de considérer comme dissimilarité entre i et j la quantité $|r(i) - r(j)|$ (r désigne l'application qui à un élément i de E fait correspondre son rang dans l'ordre induit par la variable ordinale Y_2 sur E) :

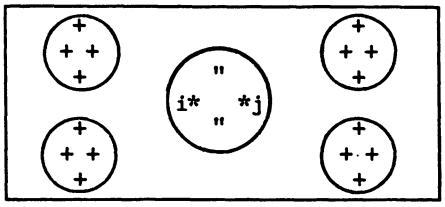
$$+^2 +^2 +^2 +^2 +^1 t] \text{———} i^* -_1 -_2 -_2 -_2 -_1$$
$$*j \text{———} [t' +^1 +^2 +^2 +^2 +^2 \text{———}$$

Dans le cas de l'exemple ci-dessus on a :

$$B_1(i, j) - \overline{B}_1(i, j) = [2 \times (4 + 4) + (1 + 1)] - [2 \times 4 + (1 + 1)] = 8$$

Cas d'une variable qualitative

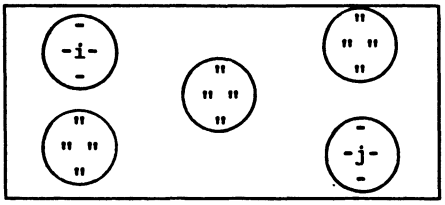
Pour la variable qualitative Y_3 les choses se passent un peu différemment. Supposons que la variable Y_3 possède cinq modalités et que i et j aient la même modalité. On considère la partition induite sur E par la variable Y_3, soit :



i et j étant dans la même classe, tous les éléments de E ayant une modalité différente de celle de i ont une contribution positive (+2) : ils sont favorables au classement de i avec j, les autres ont une contribution nulle. Dans le cas de l'exemple ci-dessus la contribution de la variable Y_3 est donnée par :

$$B_3(i, j) - \overline{B}_3(i, j) = 2 \times 16 .$$

Cas où les éléments i et j ne possèdent pas la même modalité :



Si i et j appartiennent à deux classes différentes alors l'opinion de tous les éléments ayant la même modalité que i ou j est négative (−1). Les autres éléments sont neutres, dans l'exemple ci-dessus on a :

$$B_3(i, j) - \overline{B}_3(i, j) = -8 .$$

Dans ce cas $b(i, j) = \overline{b}(i, j) = 0$, $f(i, j) = n - 8 = 12$ et $\overline{f}(i, j) = n = 20$. Au total le coefficient $\sum_r B_r(i, j) - \sum_r \overline{B}_r(i, j)$ est égal à la somme des contributions de toutes les variables de description, si l'on considère que les éléments i et j ont des modalités différentes relativement à la variable qualitative alors :

$$\sum_r B_r(i, j) - \sum_r \overline{B}_r(i, j) = 14 + 8 - 8 = 14$$

6.2.3. Règle des comparaisons par paires correspondante

En l'absence de la contrainte « Y partition », la variable Y* qui réalise le maximum de la fonction économique :

$$\sum_i \sum_j \left[\sum_r B_r(i, j) - \sum_r \overline{B}_r(i, j) \right] Y(i, j)$$

s'obtient par simple application de la règle des comparaisons par paires suivante :

$$\text{si} \quad \sum_r B_r(i, j) > \sum_r \overline{B}_r(i, j) \quad \text{alors} \quad Y^*(i, j) = 1 .$$

A cause des effets du type effet Condorcet : la solution Y*, ainsi obtenue, n'est pas une partition, d'où la nécessité d'introduire la contrainte linéaire « Y partition ».

6.2.4. Cas ou toutes les variables sont qualitatives

Proposition

Dans le cas ou toutes les variables de description sont qualitatives les programmes linéaires (6.2) et (6.2.1.) s'écrivent tout simplement sous la forme

suivante :

$$\text{Max } 2n \sum_{i \neq j}\sum \left\{ C(i,j) - \left[\frac{C(i,.) + C(.,j)}{2n} \right] \right\} Y(i,j)$$

$$Y(i,j) = Y(j,i)$$

$$Y(i,j) + Y(j,k) - Y(i,k) \leqslant 1 \qquad i \neq j \neq k$$

$$Y(i,j) \in \{0, 1\}$$

$$C(i,j) = \sum_r C_r(i,j),\ C(i,.) = \sum_r C_r(i,.),\ C(.,j)$$

$$= \sum_r C_r(.,j),\ C_r(i,.) = \sum_j C_r(i,j),\ C_r(.,j) = \sum_i C_r(i,j)$$

$$C_r(i,j) = \begin{cases} 1 & \text{si } V_r(i) = V_r(j) \\ 0 & \text{si } V_r(i) \neq V_r(j) \end{cases}$$

Le critère (6.2), introduit à l'origine pour classifier des données hétérogènes, équivaut au critère dérivé de Jordan [20] dans le cas où toutes les variables de description sont qualitatives. Si l'on désigne par c et y deux variables qualitatives, l'expression contingentielle du critère dérivé de Jordan telle qu'elle a été donnée par F. Marcotorchino [20] s'écrit :

$$dJ(c, y) = \sum_{u=1}^{p} \sum_{v=1}^{q} \left[\frac{n_{uv}}{n} \right] \left[n_{uv} - \frac{n_{u.}\, n_{.v}}{n} \right]$$

6.3. Critère induit par le coefficient de corrélation

Au sens de la mesure Ψ_{Cor} le problème du calcul des partitions s'ajustant le mieux à l'ensemble des triordonnances $P_r (r = 1, m)$ s'écrit de la manière suivante :

$$\underset{P \in \Omega_e}{\text{Max}} \sum_r \Psi_{Cor}(P, Pr) <=> \underset{T \in \Phi_e}{\text{Max}} \sum_r \text{Cor}(T, Tr)$$

$$<=> \underset{T \in \Phi_e}{\text{Max}} \sum_r \frac{\sum\sum\sum T(i,j,k)\, T_r(i,j,k)}{\sqrt{\sum\sum\sum T^2(i,j,k)}\ \sqrt{\sum\sum\sum T_r^2(i,j,k)}}$$

7. APPLICATION EN RECONNAISSANCE DE LA PAROLE

Les techniques des comparaisons par triplets ont été utilisées en vue de classifier un ensemble de 38 phonèmes connaissant la matrice des confusions entre phonèmes notée Mc. Il s'agit d'une matrice **non symétrique** croisant les phonèmes et où Mc(i, j) représente le nombre de fois où l'on reconnait le phonème i alors que c'est le phonème j qui a été prononcé. Dans ce problème particulier les couples (i, j) et (j, i) ne jouent pas le même rôle : $Mc(i,j) \neq Mc(j,i)$, c'est pourquoi il faut considérer l'ensemble des couples des phonèmes et non celui des paires. L'information extraite de la matrice des confusions en vue d'établir une classification des phonèmes se présente de la manière suivante : pour chaque triplet de phonèmes (i, j, k) on retient celui des **phonè-**

mes prononcés i et k qui a été le plus confondu avec **le phonème reconnu** j. Cette information apparait comme une triordonnance sur l'ensemble des phonèmes et peut être codée de la manière suivante :

$$T_{Mc}(i, j, k) = \begin{cases} 1 & \text{si} \quad Mc(j, i) > Mc(j, k) \\ 0 & \text{si} \quad Mc(j, i) = Mc(j, k) \\ -1 & \text{si} \quad Mc(j, i) < Mc(j, k) \end{cases}$$

Pour classifier l'ensemble des phonèmes il suffit de résoudre le programme linéaire suivant :

$$\text{Max} \sum_{i \neq j} \sum \left[\sum_k T_{Mc}(i, j, k) - \sum_k T_{Mc}(k, i, j) \right] Y(i, j)$$

$$Y(i, j) = Y(j, i)$$

$$Y(i, j) + Y(j, k) - Y(i, k) \leqslant 1 \qquad i \neq j \neq k$$

$$Y(i, j) \in \{0, 1\}$$

ou d'utiliser la même heuristique utilisée en Agrégation des similarités [7]. Il s'agit du programme linéaire (5.2), où l'on a remplacé $T_s(i, j, k)$ par $T_{Mc}(i, j, k)$ et $T_s(k, i, j)$ par $T_{Mc}(k, i, j)$.

Pour classifier l'ensemble des phonèmes à partir de la matrice des confusions une seconde approche est possible. Elle consiste à extraire de la matrice des confusions l'information suivante : pour chaque triplet de phonèmes (i, j, k) on retient celui des **phonèmes reconnus** i et k qui a été le plus confondu avec **le phonème prononcé** j. Cette information apparait comme une triordonnance sur l'ensemble des phonèmes qui s'écrit :

$$T_{Mc}(i, j, k) = \begin{cases} 1 & \text{si} \quad Mc(i, j) > Mc(k, j) \\ 0 & \text{si} \quad Mc(i, j) = Mc(k, j) \\ -1 & \text{si} \quad Mc(i, j) < Mc(k, j) \end{cases}$$

Partant de cette seconde triordonnance (différente de la première) on peut établir une classification des phonèmes en utilisant le programme linéaire ci-dessus.

8. RÉFÉRENCES

[1] S. CHAH. — *Calcul des Partitions Optimales d'un Critère d'Adéquation à une Préordonnance,* Publication de l'ISUP, Vol. XXIX, Fasicule 1, 1984.

[2] S. CHAH. — Agrégation des Préordonnances, *Etude IBM,* N° 63, Mai 1984.

[3] S. CHAH. — Comparaisons par Triplets en Classification Automatique, *Etude IBM,* N° 86, Mai 1985.

[4] S. CHAH. — Critères de Classification sur des Données Hétérogènes, *Revue de Statistique Appliquée,* volume XXXIII, N° 2, 1985.

[5] J.L. CHANDON, S. PINSON. — *Analyse Typologique : Théorie et Applications,* Masson, Paris, 1981.

[6] I.C. LERMAN. — *Classification et Analyse Ordinale des Données.* Dunod, Paris, 1981.

[7] F. MARCOTORCHINO, P. MICHAUD. — Agrégation de Similarités en Classification Automatique, *R.S.A.,* Vol. XXX, N° 2, 1982.

[8] F. MARCOTORCHINO. — Utilisation des Comparaisons par Paires en Statistique des Contingences, *Partie (I), Etude IBM F069,* 1984.

[9] F. MARCOTORCHINO. — Utilisation des Comparaisons par Paires en Statistique des Contingences, *Partie (II), Etude IBM F071,* 1984.

[10] F. MARCOTORCHINO. — Utilisation des Comparaisons par Paires en Statistique des Contingences, *Partie (III), Etude IBM F081,* 1985.

[11] P. MICHAUD. — Agrégation à la majorité : Hommage à Condorcet, *Etude IBM N° F051,* 1982.