REVUE DE STATISTIQUE APPLIQUÉE

SAID CHAH

Critères de classification sur des données hétérogènes

Revue de statistique appliquée, tome 33, nº 2 (1985), p. 19-36
<http://www.numdam.org/item?id=RSA_1985__33_2_19_0>

CRITERES DE CLASSIFICATION SUR DES DONNEES HETEROGENES

Said CHAH

Centre Scientifique IBM-France
36, Av. R. Poincaré 75116 Paris

S'il existe actuellement un nombre important d'algorithmes pour classifier un ensemble d'objets décrits par des variables homogènes, il n'en est pas de même lorsque les variables de description sont hétérogènes. La plupart des artifices existants font appel à des techniques de codage ou de quantification en vue de rendre homogènes les variables en question et traiter le problème par des algorithmes conçus à l'origine pour des données homogènes. Le but de cet article est de proposer une nouvelle approche, fondée sur la théorie d'agrégation des préordonnances introduite par l'auteur dans [1], permettant d'établir un ensemble de critères de classification opérant sur des données hétérogènes. On montrera que, utilisée sur des données homogènes, l'approche en question redonne un critère de classification connu.

Rappelons qu'une préordonnance définie sur un ensemble E à n éléments est une relation de préordre total définie sur l'ensemble, noté F, des paires d'éléments de E. On peut définir diverses techniques susceptibles d'induire une préordonnance sur un ensemble E. L'approche que nous proposons en vue d'aborder le problème de classification de données hétérogènes est la suivante :

A chaque variable Vr $(r = 1, m)$, qu'elle soit quantitative, qualitative nominale ou qualitative ordinale, on associera, selon une technique que l'on précisera, une préordonnance notée Pr, dite préordonnance induite par la variable Vr sur l'ensemble à classifier E. Et l'on cherchera parmi tous les éléments de l'ensemble des préordonnances induites par les partitions de E, noté π_e ceux qui s'ajustent le mieux, au sens d'une mesure d'association entre préordonnances que l'on précisera, à l'ensemble des préordonnances Pr $(r = 1, m)$. Il s'agit d'une application particulière de la théorie d'Agrégation des Préordonnances. Ainsi au sens d'un critère d'association entre préordonnances, noté Ψ_C, le problème de classification de l'ensemble E décrit par les m variables hétérogènes Vr $(r = 1, m)$ s'écrit de la manière suivante :

$$\text{(CDH)} \qquad \boxed{\underset{\pi_e}{\text{Max}} \sum_{r} \Psi_C(P, Pr)}$$

Les préordonnances Pr $(r = 1, m)$ sont des données du problème, P parcourt l'ensemble π_e. On montrera que certains problèmes du type (CDH) peuvent être mis sous forme d'un programme linéaire que l'on peut résoudre soit directement $(n \leqslant 80)$, soit par la même heuristique utilisée en Agrégation des Similarités par F. Marcotorchino et P. Michaud $(n \leqslant 350)$.

Mots-clés : **Variables Hétérogènes, Agrégation des préordonnances, programmation linéaire.**

1. NOTION DE PREORDONNANCES INDUITES PAR DES VARIABLES

1.1. Définition

E étant un ensemble à n éléments et F l'ensemble des paires d'éléments de E, on appelle préordonnance définie sur E une relation de préordre total, notée $\leqslant$, définie sur F. De la même façon on appelle ordonnance sur E une relation d'ordre total, définie sur F. L'ensemble des préordonnances définies sur E sera noté $\pi(E)$ (ou π si aucune confusion n'est à craindre).

1.2. Codages d'une préordonnance

Selon l'usage que l'on veut en faire, une préordonnance P peut être codée d'une façon binaire ou ternaire. On peut définir deux codages ternaires, équivalents, notés T1 et T2, et deux codages binaires notés B1 et B2.

1.2.1. Codages ternaires (T1 et T2) :

$$\forall i, j, k, l \in E \quad Tp(i, j, k, l) = \begin{cases} 1 & \text{si } \{i, j\} >_p \{k, l\} \\ 0 & \text{si } \{i, j\} =_p \{k, l\} \\ -1 & \text{si } \{i, j\} <_p \{k, l\} \end{cases} \quad [T1]$$

Un deuxième codage, équivalent au premier, est donné par :

$$\forall i, j, k, l \in E \quad Op(i, j, k, l) = \begin{cases} 1 & \text{si } \{i, j\} >_p \{k, l\} \\ 1/2 & \text{si } \{i, j\} =_p \{k, l\} \\ 0 & \text{si } \{i, j\} <_p \{k, l\} \end{cases} \quad [T2]$$

Le symbole $>_p$ désigne le fait que la paire $\{i, j\}$ est classée après la paire $\{k, l\}$ dans l'ordre induit par P sur F, $=_p$ désigne le fait que les deux paires sont ex aequo, $<_p$ désigne le fait que la paire $\{i, j\}$ est classée avant la paire $\{k, l\}$.

1.2.2. Remarque

$$Tp(i, j, k, l) = Op(i, j, k, l) - Op(k, l, i, j)$$
$$Op(i, j, k, l) = 1 - Op(k, l, i, j)$$

Dans le cas d'une préordonnance induite par un indice de similarité S ($S(i, j)$ et $S(k, l)$ connus), on retrouve le codage (T1), sous le nom de "Signature des données", dans un article de L. Guttman [6] sur les problèmes d'analyse des proximités.

1.2.3. Codages binaires (B1 et B2)

$$\forall i, j, k, l \in E \quad Tp(i, j, k, l) = \begin{cases} 1 & \text{si } \{i, j\} \geqslant_p \{k, l\} \\ 0 & \text{sinon} \end{cases}$$

$$\forall i, j, k, l \in E \quad Tp(i, j, k, l) = \begin{cases} 1 & \text{si } \{i, j\} >_p \{k, l\} \\ 0 & \text{sinon} \end{cases}$$

1.2.4. Cas d'une ordonnance

Si la donnée est une ordonnance alors tous les codages précédents sont équivalents.

On notera $\Gamma(E)$ l'ensemble des représentants, selon un codage donné, des éléments de $\Pi(E)$.

1.3. Typologie sur les préordonnances

L'objet de ce paragraphe est d'introduire la notion de préordonnance induite par une variable de description V, de nature quelconque. Il s'agit d'une technique susceptible d'induire des préordonnances sur un ensemble E, dont nous aurons besoin dans cet article. D'autres techniques ont été présentées par l'auteur dans [1].

1.3.1. Préordonnance induite par une relation binaire

R étant une relation binaire définie sur E. On appelle préordonnance sur E induite par la relation R, le préordre total sur F, noté Pr et défini par :

$$\{i,j\} \,\mathrm{Pr}\, \{k,l\} \Leftrightarrow iRj \quad \text{et} \quad k\not{R}l$$

(exaequo si iRj et kRl ou $i\not{R}j$ et $k\not{R}l$). La variable Tr associée à la préordonnance Pv se présente sous la forme suivante :

$$\forall i,j,k,l \in E \quad \mathrm{Tr}(i,j,k,l) = \begin{cases} 1 \text{ si} & iRj \text{ et } k\not{R}l \\ 0 \text{ si} & \begin{cases} iRj \text{ et } KRl \\ \text{ou} \\ i\not{R}j \text{ et } k\not{R}l \end{cases} \\ -1 \text{ si} & i\not{R}j \text{ et } KRl \end{cases}$$

1.3.1.1. Remarque

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

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

L'ensemble des préordonnances induites sur E par l'ensemble des relations d'équivalence définies sur E (ou partitions de E) est un sous ensemble de $\Pi(E)$ que l'on notera Π_e. Il lui correspond un sous ensemble de $\Gamma(E)$ noté Γ_e (ensemble de représentants des préordonnances induites sur E par l'ensemble des partitions de E).

1.3.2. Préordonnance induite par une variable quantitative

V étant une variable quantitative définie sur E, on appelle préordonnance induite par V sur E le préordre total défini sur F par :

$$\{i,j\} \,\mathrm{Pv}\, \{k,l\} \Leftrightarrow |V(i) - V(j)| \leqslant |V(k) - V(l)|$$

où V (i) désigne la valeur prise par la variable V pour l'élément i de E. La préordonnance Pv induite par la variable quantitative V est représentée (cas du codage Tl) par la variable Tv suivante :

$$Tv(i,j,k,l) = \begin{cases} 1 \text{ si } |V(i) - V(j)| < |V(k) - V(l)| \\ 0 \text{ si } \qquad\qquad = \\ -1 \text{ si } \qquad\qquad > \end{cases}$$

1.3.2.1. Remarque

$$Tv(i,j,k,l) = -Tv(k,l,i,j)$$

1.3.3. Préordonnances induites par une variable ordinale

Selon l'usage que l'on veut en faire, on peut définir deux préordonnances, différentes, induites par une variable ordinale V. La première, la plus naturelle, est celle associée à la relation de préordre induite par V sur E (1.3.4). La deuxième (celle dont il s'agira dans cet article) s'obtient de la manière suivante : on désigne par r l'application qui à un élément i de E fait correspondre son rang dans l'ordre induit par V sur E ; la relation Pv définie sur F par :

$$\{i,j\} \, Pv \, \{k,l\} \Leftrightarrow |r(i) - r(j)| \leqslant |r(k) - r(l)|$$

est une préordonnance sur E. La variable Tv correspondante est donnée par :

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

Le choix entre les deux approches précédentes se fera en fonction du problème traité.

1.3.4. Préordonnance induite par une variable qualitative

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

2. CRITERES D'ASSOCIATION ENTRE PREORDONNANCES

Le but de ce paragraphe est d'introduire des mesures d'association entre préordonnances, pour ce faire nous proposons deux approches. La première fait appel à la covariance et au coefficient de corrélation pour définir deux mesures d'association entre préordonnances. La seconde approche consiste à associer dans un premier temps, à chaque critère de contingence (2X2) une mesure d'association entre ordronnances et, dans un deuxième temps, à chaque critère de contingence (2X2), dans le cas d'un codage binaire de la préordonnance, et (3X3), dans le cas d'un codage ternaire, une mesure d'association entre préordonnances.

2.1. Approche corrélationnelle du problème d'association entre préordonnances

Etant donné deux préordonnances P et Q, on désigne par Tp et Tq les variables ternaires représentant les préordonnances P et Q (cas du codage T1), lesquelles sont définies par :

$$\forall\, i,j,k,l \in E \quad Tp(i,j,kl) = \begin{cases} 1 \text{ si } \{i,j\} >_p \{k,l\} \\ 0 \text{ si } \quad =_p \\ -1 \text{ si } \quad <_p \end{cases}$$

$$\forall\, i,j,k,l \in E \quad Tq(i,j,k,l) = \begin{cases} 1 \text{ si } \{i,j\} >_q \{k,l\} \\ 0 \text{ si } \quad =_q \\ -1 \text{ si } \quad <_q \end{cases}$$

La covariance et le coefficient de corrélation induisent sur l'ensemble $\Pi(E)$ deux mesures d'association, notées Ψ_{cov} et Ψ_{cor}, données par :

$$\Psi_{cov}(P,Q) = \text{Cov}(Tp,Tq)$$

$$\Psi_{cor}(P,Q) = \text{Cor}(Tp,Tq)$$

Si l'on désigne par M le cardinal de FXF (cf. 2.1.2), comme $E(Tp) = 0$, on a :

$$M \cdot \Psi_{cov}(P,Q) = \Sigma\Sigma\Sigma\Sigma Tp(i,j,k,l)\, Tq(i,j,k,l)$$

$$\Psi_{cor}(P,Q) = \text{Cor}(Tp,Tq) = \text{Cov}(Tp,Tq)/\sqrt{\text{Var}(Tp)}\sqrt{\text{Var}(Tq)}$$

avec : $\text{Var}(Tp) = \Sigma\Sigma\Sigma\Sigma\,(Tp(i,j,k,l))^2/M$.

Si le codage utilisé est ternaire du type T2 alors :

$$M \cdot \Psi_{cov}(P,Q) = \Sigma\Sigma\Sigma\Sigma\,[Op(i,j,k,l) - 1/2]\,[Oq(i,j,k,l) - 1/2]$$

En effet : $E(Op) = \Sigma\Sigma\Sigma\Sigma\, Op(i,j,k,l)/M = 1/2$.

$$\Psi_{cor}(P,Q) = \Psi_{cov}(P,Q)/\sqrt{\Sigma\,(Op(i,j,k,l) - 1/2)^2}\sqrt{\Sigma\,(Oq(i,j,k,l) - 1/2)^2}$$

Pour une simple raison de commodité et de simplicité des calculs et des formules on préfèrera, dorénavant, le codage T1 au codage T2. Les formulations définitives des mesures Ψ_{cov} et Ψ_{cov} sont, par conséquent, celles issues du codage T1.

2.1.1. Proposition

Si P et Q sont des ordonnances alors :

$$\boxed{\Psi_{Cor}(P,Q) = \Sigma\Sigma\Sigma\Sigma Tp(i,j,k,l)\; Tq(i,j,k,l)/M}$$

En effet : si P et Q sont des ordonnances alors

$$T^2 p(i,j,k,) = T^2 q(i,j,k,l) = 1.$$

M désigne le cardinal de FXF, sa valeur dépend de la convention adoptée, voir la remarque suivante (2.1.2).

2.1.2. Remarque

$$N = \sum_{i} \sum_{j} 1 = n^2 : \text{ si l'on considère tous les couples } (i,j).$$

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

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

$$M = \sum \sum_{\{i,j\}} \sum \sum_{\{k,l\}} 1 = N^2 : \text{ si l'on considère tous les couples } (\{i,j\}, \{k,l\})$$

$$M = \sum \sum_{\{i,j\} \neq \{k,l\}} \sum \sum 1 = N(N-1) : \text{ si l'on considère les couples } (\{i,j\}, \{k,l\})/$$

$\{i,j\} \neq \{k,l\}$.

Généralement la convention est : $N = n(n-1)/2$ et $M = N(N-1)$, mais toute autre convention est possible.

2.1.3. Remarque

Si P et Q sont des ordonnances alors :

$$\Sigma\Sigma\Sigma\Sigma \, | Op(i,j,k,l) - Oq(i,j,k,l) | = 2M [1 - \Psi_{cov}(P,Q)]$$

2.2. Aperçu sur l'approche contingentielle

P et Q étant deux ordonnances, définies sur E, en vue de définir des mesures d'association entre P et Q, on établit la table de contingence (2X2), croisant les modalités des variables binaires Tp et Tq (représentant les ordonnances P et Q, selon le codage Tl). La table de contingence, en question, se présente comme suit :

		Tq		
		1	−1	
Tp	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 $Tp(i,j,k,l)$ et $Tq(i,j,k,l)$, les quantités n_{11}, n_{1-1}, n_{-11} ..... etc., sont données par :

$$4n_{11} = \sum_i \sum_j \sum_k \sum_l [1 + Tp(i,j,k,l)\, Tq(i,j,k,l)]$$

$$= 4n_{-1-1}.$$

$$4n_{1-1} = \sum_i \sum_j \sum_k \sum_l [1 - Tp(i,j,k,l)\ Tq(i,j,k,l)]$$

$$= 4n_{-11}$$

$$n_{.1} = n_{.-1} = n_{1.} = n_{-1.} = M/2.$$

Démonstration :

$$\begin{aligned}
4n_{11} &= \Sigma\Sigma\Sigma\Sigma(1 + Tp(i,j,k,l))(1 + Tq(i,j,k,l)) \\
&= \Sigma\Sigma\Sigma\Sigma[1 + Tp(i,j,k,l) + Tq(i,j,k,l) \\
&+ Tp(i,j,k,l)\, Tq(i,j,k,l)] \\
&= \Sigma\Sigma\Sigma\Sigma[1 + Tp(i,j,k,l)\, Tq(i,j,k,l)].
\end{aligned}$$

$$\begin{aligned}
4n_{-1-1} &= \Sigma\Sigma\Sigma\Sigma(1 - Tp(i,j,k,l))(1 - Tq(i,j,k,l)) \\
&= \Sigma\Sigma\Sigma\Sigma[1 + Tp(i,j,k,l)\, Tq(i,j,k,l)]
\end{aligned}$$

$$\begin{aligned}
4n_{-11} &= \Sigma\Sigma\Sigma\Sigma(1 - Tp(i,j,k,l))(1 + Tq(i,j,k,l)) \\
&= \Sigma\Sigma\Sigma\Sigma[1 - Tp(i,j,k,l)\, Tq(i,j,k,l)]
\end{aligned}$$

$$\begin{aligned}
4n_{1-1} &= \Sigma\Sigma\Sigma\Sigma(1 + Tp(i,j,k,l))(1 - Tq(i,j,k,l)) \\
&= \Sigma\Sigma\Sigma\Sigma[1 - Tp(i,j,k,l)\, Tq(i,j,k,l)]
\end{aligned}$$

$$\begin{aligned}
4n_{.1} &= \Sigma\Sigma\Sigma\Sigma(1 + Tp(i,j,k,l)\, Tq(i,j,k,l)) \\
&+ \Sigma\Sigma\Sigma\Sigma(1 + Tp(i,j,k,l)\, Tq(i,j,k,l)) \\
&+ \Sigma\Sigma\Sigma\Sigma(1 - Tp(i,j,k,l)\, Tq(i,j,k,l)) \\
&= 2M.
\end{aligned}$$

Tout critère de contingence, défini sur le tableau précédent, donne lieu à une mesure d'association entre l'ordonnance P et l'ordonnance Q. De sorte que, si l'on désigne par C un indice d'association quelconque sur une table de contingence (2X2) : Belson, Pearson, Rand, Jordan... etc., il lui correspond une mesure d'association entre ordonnances notée Ψ_C et définie par :

$$\boxed{\Psi_C(P, Q) = C(Tp, Tq)}$$

L'objet des propositions qui suivent est d'établir, en utilisant les notations $Tp(i, j, k, l)$ et $Tq(i, j, k, l)$, les expressions des critères d'association entre ordonnances obtenus par le biais de cette approche. En vue de réduire le volume de cet article nous donnerons sans démonstration quelques résultats concernant cette approche. Pour plus de détails voir [1].

2.2.1. Mesure induite par le critère d'ajustement simple Ψ_a

$$\Psi_a(P, Q) = A(Tp, Tq) = [n_{11} + n_{-1-1}]/M$$

si l'on remplace n_{11} et n_{-1-1} par leurs valeurs on obtient :

$$4M\Psi_a(P, Q) = 2\sum_i \sum_j \sum_k \sum_l [1 + Tp(i, j, k, l)\, Tq(i, j, k, l)]$$

les indices i, j, k et l sous le signe $\Sigma\Sigma\Sigma\Sigma$ indiquent que celui-ci porte sur les couples ({i, j} {k, l}), dont le nombre, noté M, dépend de la convention adoptée, voir (2.1.2). Ces indices seront omis dans les notations si aucune confusion n'est à craindre.

D'où :

$$\Psi_a(P, Q) = A(Tp, Tq) = [\Sigma\Sigma\Sigma\Sigma\, Tp(i, j, k, l)\, Tq(i, j, k, l)]/2M + 1/2$$
$$= \Psi_{cov}(P, Q)/2 + 1/2$$

2.2.2. Mesure d'association induite par le critère de Yule Ψ_y :

Sur une table de contingence (2X2) le critère de Yule [5] est donné par :

$$\Psi_y(P, Q) = \frac{[n_{11}\, n_{-1-1} - n_{1-1}\, n_{-11}]}{[n_{11}\, n_{-1-1} + n_{1-1}\, n_{-11}]}$$

Le critère de Yule induit une mesure d'association entre ordonnances notée Ψ_y, définie par :

$$\boxed{\Psi_y(P, Q) = \frac{2M\, \Sigma\Sigma\Sigma\Sigma\, Tp(i, j, k, l)\, Tq(i, j, k, l)}{M^2 + [\Sigma\Sigma\Sigma\Sigma\, Tp(i, j, k, l)\, Tq(i, j, k, l)]^2}}$$

2.2.3. Cas d'un second critère de Yule Ψ'_y

Sur une table de contingence (2X2), Yule a proposé un second critère [5], il s'écrit de la manière suivante :

$$\frac{[\sqrt{n_{11}\, n_{-1-1}} - \sqrt{n_{1-1}\, n_{-11}}]}{[\sqrt{n_{11}\, n_{-1-1}} + \sqrt{n_{1-1}\, n_{-11}}]}$$

Le second critère de Yule induit une mesure d'association entre ordonnances notée Ψ'_y définie par :

$$\boxed{\Psi'_y(P, Q) = \Sigma\Sigma\Sigma\Sigma\, Tp(i, j, k, l)\, Tq(i, j, k, l)/M = \Psi_{cov}(P, Q)}$$

2.2.4. Mesure d'association induite par d'autres critères

En utilisant les formulations, en fonction de Tp(i, j, k, l) et Tq(i, j, k, l), des quantités n_{11}, n_{1-1}, n_{-11}... etc., on peut établir l'ensemble des résultats suivants :

Cas du critère de Kendall :

$$M\,\Psi_k(P, Q) = M \cdot K(Tp, Tq) = [n_{11}\, n_{-1-1} - n_{1-1}\, n_{-11}]$$

$$\Psi_k(P, Q) = \Sigma\Sigma\Sigma\Sigma\, Tp(i, j, k, l)\, Tq(i, j, k, l)/4$$

Cas du critère de BELSON :

$$\Psi_b(P, Q) = B(Tp, Tq) = 4[K(Tp, Tq)]^2 = 4[\Psi_k(P, Q)]^2$$

$$\Psi_b(P, Q) = [\Sigma\Sigma\Sigma\Sigma\, Tp(i, j, k, l)\, Tq(i, j, k, l)]^2/4$$

Cas du critère de PEARSON :

$$\Psi_{\chi^2}(P, Q) = \chi^2(Tp, Tq) = [\Sigma\Sigma\Sigma\Sigma\, Tp(i, j, k, l)\, Tq(i, j, k, l)]^2/M$$

D'autres critères quadratiques ont été étudiés par l'auteur dans [1].

2.3. Approche contingentielle des problèmes d'association entre préordonnances

L'approche contingentielle (table de contingence (2X2)) s'adapte parfaitement bien aux problèmes d'association entre ordonnances car le codage d'une ordonnance est binaire et unique. Cela n'exclut pas la possibilité de l'utilisation de la dite approche en vue de mesurer l'association entre deux préordonnances, pour ce faire il suffit d'adopter l'une des trois conventions suivantes :

2.3.1. Première convention

Elle consiste à utiliser l'un des codages binaires B1 ou B2 pour représenter les préordonnances. Cette approche permet de continuer à travailler sur une table de contingence (2X2) mais, elle pose un problème de choix entre deux codages possibles.

2.3.2. Deuxième convention

Le codage le mieux adapté à une préordonnance est certes le codage ternaire : car il est unique et tient compte des ex aequo. Si l'on veut contourner le problème du choix du codage posé par la première possibilité on peut adopter une seconde qui consiste à établir la table de contingence (3X3), croisant les modalités des variables ternaires Tp et Tp (représentant les préordonnances P et Q). La table de contingence en question se présente de la manière suivante :

		Tq			
		-1	0	1	
Tp	-1	n_{-1-1}	n_{-10}	n_{-11}	$n_{-1.}$
	0	n_{0-1}	n_{00}	n_{01}	$n_{0.}$
	1	n_{1-1}	n_{10}	n_{11}	$n_{1.}$
		$n_{.-1}$	$n_{.0}$	$n_{.1}$	

La même démarche que celle utilisée précédemment permet d'associer à tout critère de contingence une mesure d'association entre préordonnances. Si l'on désigne par C un critère de contingence quelconque alors :

$$\Psi_C(P, Q) = C(Tp, Tq)$$

Le critère C est calculé sur la table de contingence (3X3) ci-dessus. Si l'on convient d'utiliser cette approche il est préférable de garder les notations contingentielles ; en effet les expressions des critères en notations : $Tp(i, j, k, l)$ et $Tq(i, j, k, l)$ sont compliquées et ne se prêtent à aucune simplification.

2.3.3. Troisième convention

Une dernière possibilité consiste à utiliser les mêmes mesures d'association obtenues dans le cas de deux ordonnances pour mesurer l'association entre deux préordonnances. Pour cela il suffit de poser en cas d'ex aequo $Tp(i, j, k, l) = 0$ (codage (T1)), ce qui annule l'effet des ex aequo sur les mesures considérées. Cette démarche est la plus pratique pour les calculs. C'est cette convention que l'on adoptera, implicitement, si aucune indication n'est donnée sur la nature de la convention adoptée.

2.4. Relation entre l'approche contingentielle et l'approche corrélationnelle

Si P et Q sont des ordonnances alors le codage des ordonnances P et Q est unique et l'on a :

$$\begin{aligned} 4\Psi_k(P, Q) &= M \cdot \Psi_{cov}(P, Q) = M\Psi_{cor}(P, Q) = M \cdot \Psi'_y(P, Q) \\ &= M[2\Psi_a(P, Q) - 1] \end{aligned}$$

3. CRITERES DE CLASSIFICATION SUR DES DONNEES HETEROGENES

3.1. Problème d'agrégation des préordonnances

Etant donné un ensemble de m préordonnances définies sur E, notées $Pr(r = 1, m)$, le problème d'agrégation des préordonnances consiste à calculer les préordonnances d'un type donné, noté Π_t, qui s'ajustent le mieux, au sens d'une mesure d'association entre préordonnances, à l'ensemble des préordonnances Pr $(r = 1, m)$ données. Si l'on désigne par Ψ_C une mesure d'association entre préordonnances le problème général d'agrégation des préordonnances s'écrit de la manière suivante :

$$\boxed{\underset{\Pi_t}{\text{Max}} \sum_r \Psi_C(P, Pr)}$$

Dans ce programme les Pr $(r = 1, m)$ sont les données du problème et P parcourt l'ensemble des préordonnances du type Π_t (donné). Ainsi, au sens du critère Ψ_{cor}, le problème d'agrégation des préordonnances s'écrit :

$$\operatorname*{Max}_{\Pi_t} \sum_r \Psi_{cor}(P, Pr) \Leftrightarrow \operatorname*{Max}_{\Gamma_t} \sum_r \operatorname{Cor}(T, Tr)$$

Ou $\{Tr/r = 1, m\}$ désigne l'ensemble des variables représentant les préordonnances Pr et T parcourt l'ensemble, noté Γ_t, des variables représentant les préordonnances du type Π_t.

3.2. Problème de classification de données hétérogènes

3.2.1. Enoncé du problème

E étant un ensemble à n éléments décrits par un ensemble de m variables hétérogènes, notées Vr (r = 1, m). Il s'agit d'effectuer une classification automatique (recherche de partitions) sur E.

L'approche que nous proposons en vue d'aborder et résoudre le problème en question est la suivante :

A chaque variable Vr (r = 1, m), qu'elle soit quantitative, qualitative, ou ordinale, on associe la préordonnance notée Pr induite par Vr sur E. Et l'on cherchera parmi tous les éléments de Π_e (ensemble des préordonnances induites par les partitions de E) ceux qui s'ajustent le mieux, au sens d'une mesure d'association entre préordonnances du type Ψ_C, à l'ensemble des préordonnances Pr (r = 1, m). On reconnaît un cas particulier du problème général d'aggrégation des préordonnances rappelé ci-dessus, introduit par l'auteur dans [1].

Rappelons qu'un élément P de Π_e est représenté, selon le codage (T1), par la variable ternaire suivante :

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

où R désigne la relation d'équivalence sur E induisant la préordonnance P et où chacune des préordonnances Pr (r = 1, m), induites par les variables Vr (r = 1, m), sont représentées, selon le codage (T1) par la variable ternaire notée Tr, donnée par :

$$Tr(i,j,k,l) = \begin{cases} 1 \text{ si } \{i,j\} >_{Vr} \{k,l\} \\ 0 \text{ si } \quad =_{Vr} \\ -1 \text{ si } \quad <_{Vr} \end{cases}$$

Compte tenu de ces notations, le problème de la recherche des partitions s'ajustant le mieux, au sens d'une mesure d'association entre préordonnances du type Ψ_C, à l'ensemble des préordonnances Pr (r = 1, m) induites par les m variables hétérogènes Vr, s'écrit de la manière suivante :

$$\boxed{\operatorname*{Max}_{\Pi_e} \sum_r \Psi_C(P, Pr) \Leftrightarrow \operatorname*{Max}_{\Gamma_e} \sum_r C(T, Tr)} \qquad \text{(CDH)}$$

3.2.2. Problème (CDH) au sens de la mesure Ψ_{cov}

Au sens de la mesure Ψ_{cov} le problème (CDH) s'écrit de la manière suivante :

$$\underset{\Pi_e}{\mathrm{Max}} \sum_r \Psi_{cov} P, Pr) \Leftrightarrow \underset{\Gamma_e}{\mathrm{Max}} \sum_r \mathrm{Cov}(T, Tr)$$

$Tr\,(r = 1, m)$ sont des données du problème, T parcourt l'ensemble Γ_e. (des représentants des préordonnances induites par les partitions de E). Or si l'on désigne par Y la variable binaire représentant une partition P de E (1.3.4.1), alors :

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

d'où :

$$\begin{aligned} M.\,\mathrm{Cov}(T, Tr) &= 2\Sigma\Sigma\, Y(i,j)\, \Sigma\Sigma\, Tr(i,j,k,l) \\ &= 2\Sigma\Sigma\, Y(i,j)\, [\Sigma\Sigma\, Or(i,j,k,l) - \Sigma\Sigma\, Or(k,l,i,j)] \end{aligned}$$

et

$$\begin{aligned} M \sum_r \mathrm{Cov}(T, Tr) &= \sum_r \{2\Sigma\Sigma\, Y(i,j)\, \Sigma\Sigma\, Tr(i,j,k,l)\} \\ &= 2 \sum_{i\neq j} Y(i,j)\, [\sum_r \sum_{k\neq l} Or(i,j,k,l) - \sum_r \sum_{k\neq l} Or(k,l,i,j)] \end{aligned}$$

Comme T parcourt l'ensemble des préordonnances induites par les partitions de E, Y parcourt l'ensemble des partitions de E. Cela donne lieu à la proposition fondamentale suivante :

3.2.2.1. Proposition

Le problème de la recherche des partitions s'ajustant le mieux, au sens du critère Ψ_{cov}, à l'ensemble des préordonnances $Pr\,(r = 1, m)$ induites par les variables Vr est équivalent au programme linéaire suivant :

$$\boxed{\begin{array}{ll} \mathrm{Max} \sum\limits_{i\neq j} Y(i,j) \{\sum\limits_r [\sum\limits_{k\neq l} Or(i,j,k,l)] - \sum\limits_r [\sum\limits_{k\neq l} Or(k,l,i,j)]\} & \\ Y(i,j) = Y(j,i) & i \neq j \\ Y(i,j) + Y(j,k) - Y(i,k) \leqslant 1 & i \neq j, j \neq k, k \neq 1 \\ Y(i,j) \in \{0, 1\} & \end{array}}$$

Les contraintes décrivent le fait que la variable binaire Y (inconnue du problème) représente une partition de E, voir [8].

La résolution de ce programme peut se faire soit par programmation linéaire ($n \leqslant 80$), soit par la même heuristique ($n \leqslant 350$), développée par F.Marcotorchino et P. Michaud, pour résoudre les problèmes d'Agrégation des Similarités [8].

3.2.2.2. Corollaire

D'après la proposition précédente, en l'absence d'un effet du type suivant :

$$\sum_r [\sum\sum_{k\neq l} Or(i,i',k,l)] > \sum_r [\sum\sum_{k\neq l} Or(k,l,i,i')]$$

et

$$\sum [\sum\sum Or(i',i'',k,l)] > \sum [\sum\sum Or(k,l,i',i'')]$$

mais

$$\sum [\sum\sum Or(i,i'',k,l)] < \sum [\sum\sum Or(k,l,i,i'')]$$

le problème de classification de l'ensemble E décrit par les variables hétérogènes Vr relève de la règle d'agrégation suivante :

$\sum_r [\sum\sum_{k\neq l} Or(i,j,k,l)] > \sum_r [\sum\sum_{k\neq l} Or(k,l,i,j)]$
alors
les éléments i et j de E sont réunis dans la partition optimale

3.2.2.3. Remarque

$$\begin{aligned} CH(i,j) &= \sum_r [\sum\sum_{k\neq l} Or(i,j,k,l) - \sum\sum Or(k,l,i,j)] \\ &= \sum_r [\sum\sum_{k\neq l} (2\, Or(i,j,k,l) - 1)] \\ &= 2\,(\sum [\sum\sum Or(i,j,k,l)] - mN/2) \end{aligned}$$

Ce qui fait que la règle précédente peut s'écrire sous la forme suivante :

$$\sum_r [\sum\sum_{k\neq l} Or(i,j,k,l)] > mN/2 \Leftrightarrow iRj \text{ dans la partition optimale}$$

Si, pour un problème particulier, la structure des données est telle que l'effet précédent n'a jamais lieu, les partitions optimales s'obtiennent par simple application de l'une des règles précédentes. Il convient donc d'effectuer une vérification avant d'utiliser la programmation linéaire ou l'heuristique.

3.2.2.4. Corollaire

Le programme linéaire correspondant à la recherche des partitions s'ajustant le mieux à l'ensemble des préordonnances Pr s'écrit de la manière suivante :

$$\text{Max} \sum\sum_{i\neq j} CH(i,j)\, Y(i,j)$$

CH (i, j) se calcule de la manière suivante :

On établit les m préordres totaux induits sur F (ensemble des paires d'éléments de E) par les variables Vr (r = 1, m), et, relativement à chaque préordre Pr, on compte le nombre des paires {k, l} classées avant la paire {i, j} et l'on note S(i, j) la somme de ces nombres, puis l'on compte le nombre des paires {k, l} classées après la paire {i, j} et l'on note I (i, j) la somme de ces nombres, enfin

CH (i,j) n'est autre que la différence entre S (i,j) et I (i,j). Autrement dit on a :

$$\mathrm{CH}(i,j) = S(i,j) - I(i,j)$$

$$S(i,j) \quad = \sum_{r} [\sum_{k \neq 1}\sum \mathrm{Or}(i,j,k,l)]$$

$$I(i,j) \quad = \sum_{r} [\sum_{k \neq 1}\sum \mathrm{Or}(k,l,i,j)]$$

Un exemple de calcul des coefficients CH (i, j) est donné dans (3.2.2.7).

Si cette nouvelle approche trouve tout son intérêt dans les problèmes de classifiction sur des données hétérogènes, rien ne nous empêche d'utiliser cette même approche pour classifier un ensemble décrit par des variables homogènes. Au contraire on doit être curieux de savoir à quoi correspond ce critère dans le cas particulier où les variables de description sont homogènes.

Supposons que les m variables Vr sont toutes qualitatives, et notons Cr la variable binaire suivante :

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

Mr (i) désigne l'application qui à un élément i de E fait correspondre sa modalité relativement à la variable qualitative Vr. Or, comme :

$$\mathrm{Tr}(i,j,k,l) = \mathrm{Cr}(i,j) - \mathrm{Cr}(k,l)$$

on a :

$$\sum_{i \neq j}\sum Y(i,j) \sum_{r} [\sum_{k \neq l}\sum \mathrm{Tr}(i,j,k,l)]$$

$$= \sum\sum Y(i,j) \sum [\sum\sum (\mathrm{Cr}(i,j) - \mathrm{Cr}(k,l))]$$

$$= \sum_{i \neq j}\sum Y(i,j) \sum_{k \neq \ell}\sum [\sum_{r} \mathrm{Cr}(i,j) - \sum_{r} \mathrm{Cr}(k,1)]$$

$$= \sum\sum Y(i,j) [\sum\sum (C(i,j) - C(k,l))]$$

$$(C(i,j) = \sum_{r} \mathrm{Cr}(i,j) \text{ et } C(k,l) = \sum_{r} \mathrm{Cr}(k,l))$$

$$= \sum\sum Y(i,j) [N\,C(i,j) - \sum\sum C(k,l)]$$

$$= N \sum\sum Y(i,j) [C(i,j) - \bar{C}]$$

On reconnaît le critère de l'écart à la moyenne introduit par l'auteur dans [2] comme critère d'adéquation à une préordonnance pondérée.

3.2.2.5. Proposition

Soit à classifier un ensemble E décrit par m variables qualitatives notées Vr. Si, pour résoudre ce problème, l'on adopte la démarche qui consiste à calculer les partitions s'ajustant le mieux, au sens de la mesure Ψ_{cov}, à l'ensemble des m préordonnances Pr, induites sur E par les m variables qualitatives Vr, alors le problème d'optimisation associé est linéaire, il correspond au critère de l'écart à la moyenne et s'écrit de la manière suivante :

$$\boxed{\begin{array}{ll} \mathrm{Max} \sum_{i \neq j}\sum Y(i,j) [C(i,j) - \bar{C}] & \\ Y(i,j) = Y(j,i) & i \neq j \\ Y(i,j) + Y(j,k) - Y(i,k) \leqslant 1 & i \neq j,\ j \neq k,\ k \neq 1 \\ Y(i,j) \in \{0,1\} & \end{array}}$$

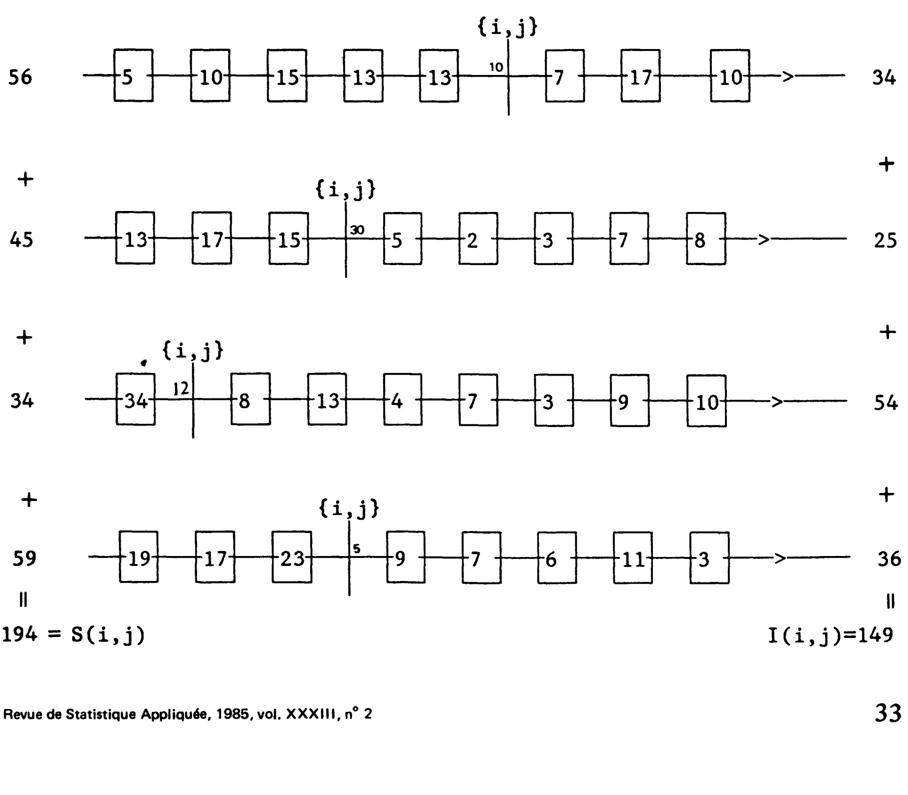
Pour mieux comprendre le contenu de cette proposition il convient d'effectuer une comparaison entre ce problème et celui des partitions centrales de S. Regnier. Pour classifier un ensemble E décrit par m paramètres qualitatifs, S. Regnier [11] propose de calculer les partitions qui s'ajustent le mieux, au sens de la différence symétrique, à l'ensemble des m partitions induites sur E par les variables qualitatives Vr ($r = 1, m$). On sait que [10] la résolution complète de ce problème donne lieu au critère de la majorité de Condorcet. Si, au lieu de chercher les partitions s'ajustant le mieux à l'ensemble des m partitions induites par les m variables qualitatives Vr, ce qui donne lieu au critère de la majorité de Condorcet, on cherche les partitions (ou les préordonnances associées) s'ajustant le mieux, au sens de la mesure Ψ_{cov}, à l'ensemble des m préordonnances Pr ($r = 1, m$) induites par les m variables qualitatives, alors on retrouve le critère des écarts à la moyenne. Notons que ce résultat reste vrai dans le cas d'un problème d'Agrégation de Relations binaires quelconques, seul le sens de la quantité $C(i, j)$ change : $C(i, j)$ est le nombre des relations pour lesquelles $i R j$.

3.2.2.6. Remarque

Antérieurement, sur un autre type de préoccupation (Relation de synonymie), F. Marcotorchino montre que le problème de la recherche de la partition s'ajustant le mieux à une relation quelconque, au sens du critère de W.F. de la Vega linéarisé [3], équivaut au critère des écarts à la moyenne.

3.2.2.7. Exemple de calcul des coefficients CH (i, j)

Dans cet exemple : $m = 4$, $n = 10$, $n = n^2 = 100$ (on peut se limiter à $N = n(n - 1)/2 = 45$) le contenu des boites indique le nombre des ex aequo.



On a par exemple :

$$\Sigma\Sigma O3(i,j,k,l) = 34 + 12/2 = 40 \qquad \Sigma\Sigma O3(k,l,i,j) = 54 + 12/2 = 60$$

$$\Sigma\Sigma T3(i,j,k,l) = 34 - 54 = -20 = \Sigma\Sigma O3(i,j,k,l) - \Sigma\Sigma)3(k,l,i,j)$$

$$CH(i,j) = [194 + (10 + 30 + 12 + 5)/2] - [149 + (10 + 30 + 12 + 5)/2] = \\ = 194 - 149 = 45$$

Si toutes les variables sont qualitatives il convient d'utiliser directement la formule : $CH(i,j) = [C(i,j) - \overline{C}]$. Où $C(i,j)$ désigne le nombre des variables pour lesquelles les éléments i et j ont la même modalité. $\overline{C}$ désigne la moyenne des $C(i,j)$.

3.3. Problème (CDH) au sens de la mesure Ψ_{cor}

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

$$\underset{\Pi_e}{\mathrm{Max}} \sum_r \Psi_{cor}(P, Pr) \Longleftrightarrow \underset{\Gamma_e}{\mathrm{Max}} \sum_r \mathrm{Cor}(T, Tr)$$

3.4. Problème (CDH) au sens d'autres mesures

Les problèmes (CDH) au sens des mesures d'association entre préordonnances : Ψ_A, Ψ_K, Ψ'_Y (cas de la convention 2.3.3) sont équivalents au programme linéaire 3.2.2.1, soit :

$$\mathrm{Max} \sum \Psi_A(T, Tr) \Longleftrightarrow \mathrm{Max} \sum \Psi_K(T, Tr) \Longleftrightarrow \mathrm{Max} \sum \Psi'_Y(T, Tr) \\ \Longleftrightarrow \mathrm{Max} \sum \Psi_{cov}(T, Tr)$$

4. EXEMPLE D'APPLICATION

Il s'agit de classifier un ensemble de 40 micro ordinateurs, chacun étant décrit par 14 variables : 10 qualitatives et 4 numériques. Les données de ce problème proviennent du numéro spécial hors-série de la revue L'Ordinateur Individuel (84/85).

Les variables qualitatives

Si aucune indication n'est donnée alors :

La modalité "0" indique l'absence du caractère.
La modalité "1" indique que le caractère existe en option.
La modalité "2" indique la présence du caractère.

1 Ecran en couleur : {0, 1, 2}.
2 Système d'exploitation disquette CP/M : {0, 1, 2}.
3 Système d'exploitation disquette MS-DOS : {0, 1, 2}.
4 Système d'exploitation disquette "Autre" : {0, 1, 2}.

5 Processeur : {1 (8 Bits), 2 (16 Bits), 3 (32 Bits)}.
6 Interface parallèle : {0, 1, 2}.
7 Interface série : {0, 1, 2}.
8 Interface IEEE 488 : {0, 1, 2}.
9 Disque dur : {0 (s'il n'existe pas), 1 (5 Mo), 2 (10 Mo)}.
10 Nombre des lecteurs de disquettes : {1, 2}.

Les variables numériques

11 Le prix (en Francs).
12 Mémoire vive : configuration (Ko).
13 Mémoire vive : maximum (Ko).
14 Mémoire de masse : unité de disquette (Ko).

Tableau des données

PAP	0	0	2	0	2	1	1	1	0	1	20000	192	512	720
QX 10	1	2	0	0	1	2	1	1	0	2	23500	192	250	320
MACINTOSH	0	0	0	2	3	0	2	0	0	1	26000	128	512	400
TI PC	2	2	2	0	2	2	1	0	0	1	26300	128	768	320
PAP (2)	0	0	2	0	2	1	1	1	0	2	27200	192	512	720
APRICOT	0	0	2	0	2	2	2	1	0	2	28400	256	768	315
Z 150	0	0	2	0	2	2	2	0	0	2	28500	320	640	360
GOUPIL 3	0	2	0	0	1	2	2	0	0	2	29700	64	1024	360
APPLE 3	0	0	0	2	1	1	2	1	1	1	35000	256	256	140
TANDY 2000	1	0	2	0	2	2	2	0	0	2	30200	128	768	720
IBM PC	1	1	2	0	2	2	1	0	0	2	36100	128	640	320
TI PC(2)	2	2	2	0	2	2	1	0	0	2	39000	256	768	320
APPLE 2E	1	1	0	2	1	1	1	1	1	1	39400	128	832	140
TELE PC	1	0	2	0	2	2	2	1	2	1	59200	256	640	360
PAP(3)	0	0	2	0	2	1	1	1	2	1	47400	192	512	720
IBM PC(2)	1	1	2	0	2	2	1	0	2	1	51000	128	640	320
Z150(2)	0	2	2	0	2	2	2	0	2	2	51500	320	640	360
TANDY 2000 (2)	1	0	2	0	2	2	2	0	2	2	52200	128	768	720
VICTOR S1	1	2	2	0	2	2	2	1	2	2	66000	256	896	1228
T 200	0	2	0	0	1	2	2	0	0	2	22500	64	64	256
AS 100	2	1	2	0	2	2	2	2	0	2	32000	128	512	640
MZ 35	1	0	2	2	1	2	2	0	0	2	34000	136	372	400
BASIS 108	1	0	0	2	1	2	2	1	0	2	28500	384	384	160
LISA 2	0	0	0	2	3	2	1	1	0	1	35500	512	512	400
EUROPE PC	1	0	2	0	2	2	2	0	2	1	47400	128	1024	327
PSI 80	0	2	0	0	1	2	2	0	0	2	47800	80	256	308
CORONA PC 2	1	1	2	0	2	2	2	1	2	2	45000	256	512	320
OPLITE	1	0	2	0	2	2	2	0	0	2	33500	256	640	360
HORIZON	0	2	0	0	1	2	2	0	0	2	35000	64	576	360
FOXY	1	1	2	0	2	2	2	1	2	1	51000	256	1024	360
SKS 2500	0	2	0	0	1	1	2	1	0	2	32000	64	256	800
ZEPHYR	0	2	0	0	1	2	2	0	0	2	41400	64	64	640
MBC 4050	0	2	0	0	2	2	2	0	0	2	35600	256	1024	640
SANCO 8000	0	2	0	0	1	2	2	0	0	2	26100	70	192	400
IPC MODELE 15	0	2	0	0	1	0	2	0	0	2	43000	64	512	782
DESKTOP 10	0	1	2	0	2	0	2	1	0	2	44800	128	768	360
LISA 2-S	0	0	0	2	3	2	1	1	1	1	47400	512	1024	400
NEC PC 8000	0	0	0	2	1	2	1	1	0	2	31800	32	64	320
M 20	0	0	0	2	2	2	2	1	0	2	21600	128	512	286
TRS 80 MODELE 120	0	0	2	1	2	2	0	0	1		32000	80	768	422

La partition optimale :

La classe 1	PAP PAP(2) PAP(3)
La classe 2	QX 10 TI PC APRICOT Z 150 GOUPIL 3 IBM PC TI PC(2) T 200 AS 100 MZ 35 OPLITE HORIZON SKS 2500 ZEPHYR MBC 4050 M 20 PSI 80 DESKTOP 10 IPC MODELE 15 TANDY 2000 TRS MODELE 80 SANCO 8000 NEC PC 8000
La classe 3	MACINTOSH LISA-2 LISA 2-S
La classe 4	APPLE 3 APPLE 2E BASIS 108
La classe 5	TELE PC Z 150 (2) VICTORS S1 EUROPE PC FOXY IBM PC (2) CORONA PC (2) TANDY 2000 (2)

REFERENCES

[1] S. CHAH. – Agrégation des Préordonnances, *Etude IBN n° F063,* Mai 1984.

[2] S. CHAH. – Optimisation linéaire en classification automatique. *Thèse de Doctorat de troisième cycle,* Université Pierre et Marie Curie, Paris VI, 1983.

[3] S. CHAH. – Calcul des partitions optimales d'un critère d'adéquation à une préordonnance, *Publication de l'ISUP,* Vol XXIX-Fascicule 1, 1984.

[4] J.L. CHANDON, F.F. BOCTOR. – Approximation d'une préordonnance par une partition, *publication de l'I.A.E. d'Aix-Marseille,* n° 279, (2^e version) 1983.

[5] L.A. GOODMAN, W.H. KRUSKAL. – *Mesures of Association for Cross Classification,* Springer, Verlag, New-York, Heidelberg Berlin, 1979.

[6] L. GUTTMAN. – A general nonmetric technique for finding the smallest coordinate space for a configuration of points", *Psychometrika,* Vol. 33, n° 4, 468-507, 1981.

[7] I.C. LERMAN. – *Classification et analyse ordinale des données.* Dunod, Paris, 1981.

[8] F. MARCOTORCHINO. – Agrégation des Similarités, *Thèse de Doctorat d'Etat.* Université Paris VI, Paris 1981.

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

[10] P. MICHAUD. – Agrégation à la majorité : Hommage à Condorcet, *Etude IBM,* n° FO51, 1982.

[11] S. REGNIER. – Sur Quelques aspects Mathématiques des Problèmes de Classification Automatique, *I.C.C. Bulletin,* Vol 4, 175-191, Rome, 1965.

[12] W.F. DE LA VEGA. – Techniques de Classification automatique utilisant un indice de ressemblance. *Revue Française de Sociologie,* 1976.

[13] W.F. DE LA VEGA, S. REGNIER. – Préhension et interprétation de plusieurs classifications d'un même ensemble de données, *Compte-rendu de contrat D.G.R.S.T.,* Paris, A.D.I.S.H., 1976.