RAIRO. RECHERCHE OPÉRATIONNELLE

ISRAËL-CÉSAR LERMAN

Formules de réactualisation en cas d'agrégations multiples

RAIRO. Recherche opérationnelle, tome 23, n° 2 (1989), p. 151-163

<http://www.numdam.org/item?id=RO_1989__23_2_151_0>

Recherche opérationnelle/Operations Research
(vol. 23, n° 2, 1989, p. 151 à 163)

FORMULES DE RÉACTUALISATION EN CAS D'AGRÉGATIONS MULTIPLES (*)

par Israël-César LERMAN (1)

Résumé. — *On considère la situation — assez fréquente — où dans une procédure de construction ascendante d'un arbre de classification hiérarchique, plusieurs paires de classes réalisent « simultanément » à un même niveau la même plus grande valeur de la proximité entre classes. Alors que classiquement plusieurs réactualisations sont nécessaires — de la matrice des proximités entre classes — avant le passage au niveau suivant, on propose ici des formules — associées aux critères de proximité entre classes les plus classiques — permettant une seule réactualisation.*

Mots clés : Classification Ascendante Hiérarchique; Indices d'association entre classes.

Abstract. — *We consider the frequently observed situation where several class pairs correspond — simultaneously — to the highest value of the proximity index between classes, in the building of hierarchical classification, by an agglomerative procedure. The usual way consists of applying several times a reactualization formulae, before going on the next level of the classification tree. In this paper, we propose only one reactualization, by the means of a formulae taking into account the joining of more than two classes. This formulae is established for the most classical association criteria between classes.*

Keywords : Hierarchical Classification; Proximity Indices between classes.

I. INTRODUCTION

La donnée d'un ensemble fini E d'objets, muni d'un indice de distance (resp. proximité) entre parties disjointes de ce dernier, permet la construction ascendante hiérarchique d'un arbre de classification sur E.

Le principe d'un tel algorithme est très simple : on démarre de la partition discrète où chaque classe contient un seul objet et on réunit à chaque pas les deux classes les plus proches. La phase finale de l'algorithme est celle où

(*) Reçu novembre 1988.
(1) I.R.I.S.A., Campus Universitaire de Beaulieu, avenue du Général Leclerc, 35042 Rennes Cedex.

Recherche opérationnelle/Operations Research, 0399-0559/89/02 151 13/$ 3.30

tous les objets se trouvent réunis dans une même classe définissant l'ensemble plein E.

Cet algorithme devient un outil très puissant d'analyse des données dès lors qu'on a finement conçu la notion de proximité entre parties disjointes de E, en tenant compte des aspects formels et statistiques dans la représentation des objets de E.

Signalons tout de suite que la situation à laquelle on s'intéresse ici est celle où à un même pas de l'algorithme, plusieurs paires de classes réalisent « en même temps » la plus grande proximité. Ce cas se rencontre beaucoup plus fréquemment qu'on ne le croit, surtout si E est un « gros » ensemble, décrit par un « petit » nombre de variables qualitatives. Dans cette situation, il y a lieu bien sûr de fusionner à un même niveau de l'arbre toutes les paires de classes qui réalisent la plus grande proximité. Le problème traité ici concerne la gestion de cette agrégation multiple.

Le déroulement classique de l'algorithme repose sur la réactualisation pas à pas de la matrice des indices de distance (resp. proximité) entre classes déjà formées.

Dans le cas généralement traité, on considère la réactualisation de cette matrice après la fusion d'exactement deux classes et cela même si plusieurs classes doivent être réunies à un même niveau.

Dans ce cas, supposons qu'à une étape donnée, on ait déjà formé la partition $\{C_j/1 \leqq j \leqq k\}$ dont on dispose de la matrice des distances

$$\{\Delta(h, j)/1 \leqq h < j \leqq k\} \tag{1}$$

ainsi que de celle des masses ou cardinaux des classes

$$\{m_j/1 \leqq j \leqq k\}, \tag{2}$$

où on a noté une même classe par son indice.

La fusion des deux classes C_h et $C_{h'}$ nécessite — pour la réactualisation — la connaissance de la suite

$$\{\Delta(h \cup h', j)/j \neq h \text{ et } j \neq h'\}, \tag{3}$$

où on a noté $h \cup h'$ pour $C_h \cup C_{h'}$ et j pour C_j.

Une formule telle que

$$\Delta(h \cup h', j) = f[\Delta(h, j), \Delta(h', j), \Delta(h, h'), M_h, M_{h'}, M_j] \tag{4}$$

où f est une fonction numérique, est du plus grand intérêt puisqu'elle permet — à partir de (1) et (2) — de réactualiser directement (1) et (2).

Lance et Williams [Lance & Williams (1967)] ont proposé une formule dite de récurrence telle que (4) en lui donnant une expression linéaire par rapport aux trois premiers arguments. Cette formule est plus précisément la suivante :

$$\Delta(h \cup h', j) = \alpha \Delta(h, j) + \beta \Delta(h', j) + \gamma \Delta(h, h'),$$

où α, β et γ sont des fonctions de M_h, $M_{h'}$ et M_j.

Cette formule regroupe un ensemble de coefficients classiques pour l'association des classes dans une procédure ascendante hiérarchique. Des généralisations ont été apportées à une telle formule [Jambu (1978), Tricot & Donegani (1988)]. Mais quel que soit l'intérêt formel de telles formules de récurrence, nous ne comprenons pas complètement le souci d'enfermer la fonction f de (4) ci-dessus dans une forme analytique figée; et ceci pour deux raisons :

(i) on peut trouver des coefficients d'association entre classes dont la formule de réactualisation n'obéit pas nécessairement à une expression analytique posée *a priori*. La conception de tels coefficients doit davantage obéir à la nature mathématique de la représentation des classes qu'à une formule de récurrence de forme analytique préfixée. Ainsi le critère de la vraisemblance du lien maximal présenté au paragraphe IV ci-dessous n'obéit à aucune des formules déjà proposées.

(ii) Le praticien se limite en général à une ou deux procédures d'agrégation compte tenu de la nature de ses données et de la représentation mathématique qu'il adopte pour ces dernières.

Nous allons quant à nous considérer directement cinq indices d'association des classes parmi les plus classiques : « distance minimale », « diamètre », « moyenne des distances », « inertie expliquée » et « vraisemblance du lien maximal ».

Si on imagine qu'à une étape de l'algorithme, on ait à fusionner « en même temps », respectivement, $n_1, n_2, \ldots, n_l$ classes et si on ne dispose que de la formule (4), on doit procéder à $[(n_1-1)+(n_2-1)+\ldots+(n_l-1)]$ réactualisations sur une matrice de distances (resp. proximités) entre classes, où le nombre de classes diminue à chaque fois d'une unité. Par contre, si on dispose d'une formule telle que

$$\Delta(j_1 \cup \ldots \cup j_r, h_1 \cup \ldots \cup h_s) = f[\Delta(j_l, h_m), M_{j_l}, M h_m / 1 \leqq l \leqq r, 1 \leqq m \leqq s], \quad (6)$$

où $j_1, \ldots, j_r$ (resp. $h_1, \ldots, h_s$) sont r (resp. s) classes qu'il y a lieu de fusionner « en même temps », on n'aura plus à effectuer qu'une seule réactualisation globale.

G. de Rham [Rham (1980)] a proposé d'étendre la formule de Lance et Williams dans un cas particulier de celui que nous envisageons ici, puisqu'il s'agit simplement d'exprimer $\Delta(j_1 \cup j_2, h_1 \cup h_2)$.

II. CRITÈRES DE LA DISTANCE MINIMALE, DU DIAMÈTRE ET DE LA MOYENNE DES DISTANCES

II.1. Critère de la distance minimale

On suppose que l'ensemble E des objets à classifier se trouve muni d'une distance δ qu'on étend à une dissimilarité Δ_1 entre parties disjointes de E, au moyen de la formule :

$$\Delta_1(C, D) = \min\{\delta(x, y)/(x, y) \in C \times D\}, \tag{1}$$

où C et D sont deux parties disjointes de E.

Cherchons maintenant à obtenir le correspondant de la formule (6) ci-dessus. On a

$$\Delta_1(j_1 \cup \ldots \cup j_r, h_1 \cup \ldots \cup h_s) = \min\{\delta(x, y)/(x, y) \in (C_{j_1} \cup \ldots \cup C_{j_r}) \times (C_{h_1} \cup \ldots \cup C_{h_s})\} = \min\{\Delta_1(j_l, h_m)/1 \leqq l \leqq r,\ 1 \leqq m \leqq s\}; \tag{2}$$

en effet, le minimum général est égal au minimum des minimums.

II.2. Critère du diamètre

Dans ce cas la proximité entre deux classes C et D est mesurée par la petitesse de la plus grande distance permettant de relier un élément de C à un élément de D. Désignons par Δ_2 la distance correspondante entre parties disjointes de E :

$$\Delta_2(C, D) = \max\{\delta(x, y)/(x, y) \in C \times D\}, \tag{3}$$

où C et D sont deux parties disjointes de E.

Relativement à des notations déjà introduites, on a :

$$\Delta_2(j_1 \cup \ldots \cup j_r,\ h_1 \cup \ldots \cup h_s)$$
$$= \max\{\delta(x, y)/(x, y) \in (C_{j_1} \cup \ldots \cup C_{j_r}) \times (C_{h_1} \cup \ldots \cup C_{h_s})\}$$
$$= \max\{\Delta_2(j_l, h_m)/1 \leqq l \leqq r,\ 1 \leqq m \leqq s\}; \quad (4)$$

en effet, le maximum général est égal au maximum des maximums.

Remarque. — On peut montrer que si C et D sont deux classes présentes à un niveau donné de l'arbre hiérarchique des classifications (obtenu par agrégations successives des paires de classes les plus proches au sens de Δ_2), $\Delta_2(C, D)$ est bien le diamètre de $C \cup D$; c'est-à-dire, $\max\{\delta(x, y)/(x, y) \in (C \cup D) \times (C \cup D)\}$.

II.3. Critère de la moyenne des distances

La distance δ sur E peut être étendue en une distance Δ_3 sur l'ensemble des parties de E au moyen de la formule

$$\Delta_3(C, D) = \frac{1}{|C| \times |D|} \sum \{\delta(x, y)/(x, y) \in C \times D\}, \quad (5)$$

où C et D sont deux parties de E de cardinaux respectifs $|C|$ et $|D|$.

Nous avons à déterminer le correspondant de la formule (6) du paragraphe précédent. On a

$$\Delta_3(j_1 \cup \ldots \cup j_r,\ h_1 \cup \ldots \cup h_s)$$
$$= \frac{1}{(M_{j_1} + \ldots + M_{j_r})(M_{h_1} + \ldots + M_{h_s})}$$
$$\times \sum \{\delta(x, y)/(x, y) \in (C_{j_1} \cup \ldots \cup C_{j_r}) \times (C_{h_1} \cup \ldots \cup C_{h_s})\}$$
$$= \frac{1}{(M_{j_1} + \ldots + M_{j_r})(M_{h_1} + \ldots + M_{h_s})}$$
$$\times \sum_{1 \leqq l \leqq r}, \sum_{1 \leqq m \leqq s} \sum \{\delta(x, y)/(x, y) \in C_{j_l} \times C_{h_m}\}$$
$$= \sum_{1 \leqq l \leqq r}, \sum_{1 \leqq m \leqq s} \nu_{lm} \Delta_3(j_l, h_m), \quad (6)$$

où

$$\nu_{lm} = \frac{M_{j_l} M_{hm}}{(M_{j_1} + \ldots + M_{j_r})(M_{h_1} + \ldots + M_{h_s})}$$

$1 \leqq l \leqq r$, $1 \leqq m \leqq s$. On a

$$\sum_{1 \leqq l \leqq r}, \sum_{1 \leqq m \leqq s} \nu_{lm} = 1. \tag{7}$$

III. CRITÈRE DE L'INERTIE EXPLIQUÉE

Ce critère est dû à Ward [Ward (1963)]. L'ensemble E des objets est supposé représenté par un nuage de points dans un espace euclidien. L'inertie expliquée par une classification $P = \{ C_j / 1 \leqq j \leqq k \}$ de E, est donnée par la formule

$$\mathscr{I}(P) = \sum_{1 \leqq j \leqq k} M_j \delta^2 (G_j, G), \tag{1}$$

où m_j est la masse de la j-ième classe dont le centre de gravité est noté G_j. G est le centre de gravité du nuage total. On a

$$G = \frac{1}{\mu} \sum_{1 \leqq j \leqq k} M_j G_j, \tag{2}$$

où μ désigne la masse totale du nuage.

On démontre (dans presque tous les ouvrages portant sur la classification, voir par exemple [Lerman (1981)], que la perte de l'inertie expliquée résultant de la fusion de deux classes C_j et C_h est égale à

$$\Delta(j, h) = \frac{M_j M_h}{(M_j + M_h)} \delta^2 (G_j, G_h). \tag{3}$$

Dans ces conditions, à chaque pas de la construction hiérarchique des classifications, on réunira les paires de classes qui rendent minimal l'indice (3). La formule de réactualisation résultant de la fusion des deux classes d'indices

j et h est la suivante :

$$\Delta(j \cup h,\ l) = \frac{1}{M_j + M_h + M_l}[(M_j + M_l)\,\Delta(j,\ l) + (M_h + M_l)\,\Delta(h,\ l) - M_l\,\Delta(j,\ h)]. \quad (4)$$

Il s'agit d'étendre cette formule en cas d'agrégations multiples, où, avec des notations utilisées, le premier membre prend la forme $\Delta(j_1 \cup \ldots \cup j_r,\ h_1 \cup \ldots \cup h_s)$.

THÉORÈME : *On a la formule de réactualisation*

$$\Delta(j_1 \cup \ldots \cup j_r,\ h_1 \cup \ldots \cup h_s) = \frac{1}{S_{r,s}}\Big[\sum\{(M_{j_l} + M_{h_m})\,\Delta(j_l,\ h_m)/1 \leqq l \leqq r,\ 1 \leqq m \leqq s\} - \sum\left\{(M_{j_l} + M_{j_{l'}})\frac{V_s}{U_r}\Delta(j_l,\ j_{l'})/1 \leqq l < l' \leqq r\right\} - \sum\left\{(M_{h_m} + M_{h_{m'}}) \times \frac{U_r}{V_s}\Delta(h_m,\ h_{m'})/1 \leqq m < m' \leqq s\right\}\Big] \quad (5)$$

où $S_{r,s} = U_r + V_s$, avec

$$U_r = \sum_{1 \leqq l \leqq r} M_{j_l} \text{ et } V_s = \sum_{1 \leqq m \leqq s} M_{h_m}.$$

Dans le cas où $r = 2$ et $s = 1$, la formule (5) précédente devient

$$\Delta(j_1 \cup j_2,\ h_1) = \frac{1}{M_{j_1} + M_{j_2} + M_{h_1}} \times [(M_{j_1} + M_{h_1})\,\Delta(j_1,\ h_1) + (M_{j_2} + M_{h_1})\,\Delta(j_2,\ h_1) - (M_{j_1} + M_{j_2})\frac{M_{h_1}}{(M_{j_1} + M_{j_2})}\Delta(j_1,\ j_2)] \quad (6)$$

qui se réduit à la formule (4) ci-dessus.

Nous allons maintenant prouver – avec les notations ci-dessus – que si la formule est vraie jusqu'à $(r,\ s)$, elle est également vraie de $(r,\ s+1)$, ce qui correspond à l'établissement de la formule par récurrence. On a, en appliquant

la formule (4) ci-dessus :

$$(7)\quad \Delta(j_1 \cup \ldots \cup j_r,\ h_1 \cup \ldots \cup h_s \cup h_{s+1}) = \frac{1}{S_{r,\,s+1}}[S_{r,\,s}\Delta(j_1 \cup \ldots \cup j_r,\ h_1 \cup \ldots \cup h_s) + (U_r + M_{h_{s+1}})\Delta(j_1 \cup \ldots \cup j_r,\ h_{s+1}) - U_r\Delta(h_1 \cup \ldots \cup h_s,\ h_{s+1})] \quad (7)$$

Nous allons examiner, respectivement, les facteurs multiplicatifs de chacun des termes de la forme

(i) $\Delta(j_l, h_m)$ pour $1 \leqq l \leqq r,\ 1 \leqq m \leqq s$
(ii) $\Delta(j_l, h_{s+1})$ pour $1 \leqq l \leqq r$
(iii) $\Delta(j_l, j_{l'})$ pour $1 \leqq l < l' \leqq r$
(iv) $\Delta(h_{m'}, h_{m'})$ pour $1 \leqq m < m' \leqq s$
(v) $\Delta(h_{m'}, h_{s+1})$ pour $1 \leqq m \leqq s$

et montrer qu'ils sont de même type que ceux correspondants de la formule (5).

Le coefficient de $\Delta(j_l, h_m)$ — pour $1 \leqq l \leqq r,\ 1 \leqq m \leqq s$ — s'obtient à partir du développement que permet la formule (5) du premier terme sous le signe crochet de la formule (7). Il est égal à

$$\frac{1}{S_{r,\,s+1}} \times S_{rs} \times \frac{1}{S_{rs}} \times (M_{j_l} + M_{h_m}) = \frac{1}{S_{r,\,s+1}} \times (M_{j_l} + M_{h_m}) \quad (8)$$

qui correspond bien à la formule (5), en passant de s à $(s+1)$.

Considérons à présent le coefficient de $\Delta(j_l, h_{s+1})$, pour $1 \leqq l \leqq r$. Ce dernier s'obtient à partir du développement — conformément à la formule (5) — du deuxième terme sous le signe crochet de la formule (7). Il est égal à

$$\frac{1}{S_{r,\,s+1}} \times (U_r + M_{h_{s+1}}) \times \frac{1}{(U_r + M_{h_{s+1}})} \times (M_{j_l} + M_{h_{s+1}})\Delta(j_l, h_{s+1}) = \frac{1}{S_{r,\,s+1}}(M_{j_l} + M_{h_{s+1}})\Delta(j_l, h_{s+1}), \quad (9)$$

qui correspond bien à la formule (5), en passant de s à $(s+1)$.

Considérons à présent le coefficient de $\Delta(j_l, j_{l'})$, pour $1 \leqq l < l' \leqq r$. Il s'obtient à partir du développement que permet la formule (5), du deuxième

et troisième terme sous le signe crochet de la formule (7). On obtient :

$$\frac{1}{S_{r,\,s+1}} \times \left[-(M_{j_l} + M_{j_{l'}}) \frac{V_s}{U_r} - (M_{j_l} + M_{j_{l'}}) \frac{h_{s+1}}{U_r} \right]$$

$$= \frac{1}{S_{r,\,s+1}} \left\{ -(M_{j_l} + M_{j_{l'}}) \frac{V_{s+1}}{U_r} \right\}, \quad (10)$$

qui correspond bien à la formule (5), en passant de s à $(s+1)$.

Maintenant, considérons le coefficient de $\Delta(h_m, h_{m'})$, pour $1 \leqq m < m' \leqq s$. On l'obtient à partir du premier et du troisième terme sous le signe crochet de la formule (7). En utilisant l'expression (5) on a :

$$\frac{1}{S_{r,\,s+1}} \left\{ \left[-(M_{h_m} + M_{h_{m'}}) \times \frac{U_r}{V_s} \right] - U_r \times \frac{1}{V_{s+1}} \times \left[-(M_{h_m} + M_{h_{m'}}) \frac{h_{s+1}}{V_s} \right] \right\}$$

$$= \frac{1}{S_{r,\,s+1}} \left[-(M_{h_m} + M_{h_{m'}}) \times \frac{U_r}{V_s} \left(1 - \frac{h_{s+1}}{V_{s+1}} \right) \right]$$

$$= \frac{1}{S_{r,\,s+1}} \left[-(M_{h_m} + M_{h_{m'}}) \times \frac{U_r}{V_{s+1}} \right], \quad (11)$$

qui correspond bien à la formule (5), en passant de s à $(s+1)$.

Il reste maintenant le cas (v) où il y a lieu de déterminer le coefficient de $\Delta(h_m, h_{s+1})$ pour $1 \leqq m \leqq s$. Ce dernier s'obtient en développant le troisième terme sous le signe crochet de (7), conformément à la formule (5). On obtient :

$$\frac{1}{S_{r,\,s+1}} \times \left[-U_r \times \frac{1}{V_{s+1}} \times (M_{h_m} + M_{h_{s+1}}) \right], \quad (12)$$

qui correspond bien à la formule (5), en passant de s à $(s+1)$.

Par conséquent la formule (5) est complètement établie quels que soient r et s.

IV. CRITÈRE DE LA VRAISEMBLANCE DU LIEN MAXIMAL

Commençons par rappeler l'expression de ce critère d'association entre deux parties disjointes C et D de l'ensemble E des unités statistiques à classifier; il peut s'agir de l'ensemble des objets ou de l'ensemble des variables de description.

On suppose qu'on soit parvenu à la définition d'un indice de similarité $\mathscr{S}$ sur E. Dans le cas où la donnée est un indice de dissimilarité $\mathscr{D}$, on posera tout simplement $\mathscr{S} = -\mathscr{D}$. On procède alors à ce que nous appelons la réduction globale des similarités sur l'ensemble $F = P_2(E)$ des parties à deux éléments sur E. Soient $\text{moy}(\mathscr{S})$ et $\text{var}(\mathscr{S})$ la moyenne et la variance de $\mathscr{S}$ sur F; n désignant le cardinal de E, on a :

$$\text{moy}(\mathscr{S}) = \frac{2}{n(n-1)} \Sigma \{ \mathscr{S}(x, y) / \{ x, y \} \in F \}$$

$$\text{var}(\mathscr{S}) = \frac{2}{n(n-1)} \Sigma \{ (\mathscr{S}(x, y) - \text{moy}(\mathscr{S}))^2 / \{ x, y \} \in F \} \tag{1}$$

On pose alors

$$Q_s(x, y) = \frac{[\mathscr{S}(x, y) - \text{moy}(\mathscr{S})]}{\sqrt{\text{var}(\mathscr{S})}} \tag{2}$$

pour tout $\{ x, y \}$ de F.

On se place alors par rapport à une hypothèse d'absence de liaison où à l'ensemble E des objets on associe un ensemble E^* de n objets aléatoires indépendants, respectivement de « mêmes types statistiques » que les objets de E. Dans ce cadre, on admet l'approximation normale $\mathscr{N}(0, 1)$ pour la distribution de $Q_s(x^*, y^*)$, où x^* et y^* sont les deux objets aléatoires indépendants respectivement associés à x et à y. Cette approximation est justifiée au mieux dans le contexte de la conception des indices sous-jacents à la méthode de classicfication basée sur la vraisemblance des liens.

La table des indices définitifs se réfère à une échelle de probabilité. Elle se met sous la forme :

$$\{ P(x, y) / \{ x, y \} \in F \}, \tag{3}$$

où

$$P(x, y) = \Phi [Q_s(x, y)], \tag{4}$$

pour tout $\{x, y\}$ appartenant à F, où Φ est la fonction de répartition de la loi normale centrée et réduite.

La base de la constitution de l'indice de comparaison entre deux classes C et D est — dans l'algorithme de la vraisemblance du lien maximal — donné par

$$p(C, D)=\max\{P(x, y)/(x, y)\in C\times D\}. \tag{5}$$

L'indice définitif s'obtient à partir de la loi de probabilité de l'indice aléatoire $p(C^*, D^*)$, où C^* et D^* sont associés à C et à D conformément à l'hypothèse d'absence de liaison. Si t est un nombre compris entre 0 et 1, on a

$$\Pr\{p(C^*, D^*)<t\}=t^{|C|\times|D|} \tag{6}$$

De sorte que l'indice de comparaison entre les deux classes C et D s'écrit

$$P(C, D)=p(C, D)^{|C|\times|D|} \tag{7}$$

[Lerman (1970), (1981)].

F. Nicolaü [Nicolaü (1980)] a étudié d'autres fonctions de base que celle (5) que nous avions considéré par exemple la moyenne de $\{P(x, y)/(x, y)\in C\times D\}$.

Dans la construction ascendante hiérarchique de l'arbre des classifications, seul l'ordre des valeurs de l'indice (7) appliqué aux différentes paires de classes en présence à un niveau donné, intervient. Par conséquent, toute fonction strictement croissante de (7) donne un indice équivalent pour la construction de l'arbre. Des raisons de précision calcul nous conduisent à prendre la fonction strictement croissante $f(x)=-\mathrm{Log}[-\mathrm{Log}(x)]$ pour $0<x<1$, dont l'échelle des valeurs est $]-\infty, +\infty[$. On note

$$\pi(C, D)=f[P(C, D)]. \tag{8}$$

Les notations sont les mêmes qu'au paragraphe III. Relativement à une partition $P=\{C_j/1\leqq j\leqq k\}$ de l'ensemble E à classifier, on note M_j le cardinal ou la masse de la classe C_j. $j_1\cup j_2\cup\ldots\cup j_r$ (resp. $h_1\cup h_2\cup\ldots\cup h_s$) indiquera $C_{j_1}\cup C_{j_2}\cup\ldots\cup C_{j_r}$

(resp. $C_{h_1}\cup C_{h_2}\cup\ldots\cup C_{h_s}$).

On pose

$$U_r = \sum_{1 \leqq l \leqq r} M_{j_l} \quad \text{et} \quad V_s = \sum_{1 \leqq m \leqq s} M_{h_m}.$$

On a la formule suivante de récurrence

$$\pi(j_1 \cup \ldots \cup j_r,\ h_1 \cup \ldots \cup h_s) = -\operatorname{Log}(U_r) - \operatorname{Log}(V_s) + \max\{\pi(j_l,\ h_M) + \operatorname{Log}(M_{j_l}) + \operatorname{Log}(M_{h_m}) / 1 \leqq l \leqq r,\ 1 \leqq m \leqq s\} \quad (9)$$

Cette formule est trivialement vraie pour $r=1$ et $s=1$. La formule de réactualisation généralement utilisée concerne $r=2$ et $s=1$, où elle devient

$$\pi(j_1 \cup j_2,\ h_1) = -\operatorname{Log}(M_{j_1} + M_{j_2}) - \operatorname{Log}(M_{h_1}) + \max\{\pi(j_1,\ h_1) + \operatorname{Log}(M_{j_1}) + \operatorname{Log}(M_{h_1}),\ \pi(j_2,\ h_1) + \operatorname{Log}(M_{j_2}) + \operatorname{Log}(M_{h_1})\}. \quad (10)$$

Cette dernière formule résulte de la propriété : le maximum de deux maximums est égal au maximum général.

Nous allons maintenant supposer que la formule (9) est vraie jusqu'à (r, s) (r et s entiers positifs) et démontrer qu'elle peut s'étendre jusqu'à $(r, s+1)$, ce qui établira qu'elle est toujours vraie. En appliquant (10) on a :

$$\pi(j_1 \cup \ldots \cup j_r,\ h_1 \cup \ldots \cup h_s \cup h_{s+1}) = -\operatorname{Log} U_r - \operatorname{Log} V_{s+1} + \max\{\pi(j_1 \cup \ldots \cup j_r,\ h_1 \cup \ldots \cup h_s) + \operatorname{Log} U_r + \operatorname{Log} V_s,\ \pi(j_1 \cup \ldots \cup j_r,\ h_{s+1}) + \operatorname{Log} U_r + \operatorname{Log} M_{h_{s+1}}\}.$$

En vertu de la formule (9) — valable jusqu'à (r, s) — on obtient :

$$= -\operatorname{Log} U_r - \operatorname{Log} V_{s+1} + \max\{\max\{\pi(j_l,\ h_m) + \operatorname{Log}(M_{j_l}) + \operatorname{Log}(M_{h_m}) / 1 \leqq l \leqq r,\ 1 \leqq m \leqq s\},\ \max\{\pi(j_l,\ h_{s+1}) + \operatorname{Log}(M_j) + \operatorname{Log}(M_{h_{s+1}}) / 1 \leqq l \leqq r\}\}$$

$$= -\operatorname{Log} U_r - \operatorname{Log} V_{s+1} + \max\{\pi(j_l,\ h_m) + \operatorname{Log}(M_{j_l}) + \operatorname{Log}(M_{h_m}) / 1 \leqq l \leqq r,\ 1 \leqq m \leqq (s+1)\}. \quad (11)$$

C.Q.F.D.

V. CONCLUSION

Nous avons déjà mentionné dans l'introduction l'intérêt de nos formules pour la construction ascendante d'un arbre de classification hiérarchique portant sur un « gros » ensemble décrit par un « petit » nombre de variables

qualitatives. Dans ce cas en effet, plusieurs agrégations multiples peuvent se produire à un même niveau de la construction.

Le besoin de ces formules est encore plus nécessaire dans le cas de l'application de l'algorithme de classification hiérarchique dit des « voisins réciproques » où — à chaque pas — il y a lieu de fusionner les paires de classes telles que l'une des composantes est la plus proche voisine de l'autre.

On sait que l'agrégation — à un niveau donné — d'une paire de classes plutôt que d'une autre, peut modifier sensiblement l'allure de l'arbre en ce qui concerne la suite des agrégations successives à ce niveau. Or il se peut — qu'à un niveau donné — plusieurs paires de classes réalisent une proximité très voisine de la plus grande proximité observée. Plus précisément, si π_k est la plus grande proximité au niveau k et si ε_k est un nombre positif considéré suffisamment « petit » devant π_k. On peut envisager de fusionner directement l'ensemble des paires de classes dont la proximité se situe dans l'intervalle $](\pi_k-\varepsilon_k), \pi_k]$. Nos formules de réactualisation seront alors très utiles.

BIBLIOGRAPHIE

1. M. JAMBU, *Classification automatique pour l'analyse des données*, t. 1, Dunod, Paris, 1978.
2. G. N. LANCE et W. T. WILLIAMS, *A General Theory of Classification Sorting Strategies: 1 = hierarchical Systems, 2 = Clustering Systems*, Computer Journal, Vol. 9-10, 1967, pp. 373-380.
3. I. C. LERMAN, *Sur l'analyse des données préalable à une classification automatique. Proposition d'une nouvelle mesure de similarité*, Rev. Math. et Sc. Hum., 8e année, n° 32, Paris, 1970.
4. I. C. LERMAN, *Classification et analyse ordinale des données*, Dunod, Paris, 1981.
5. F. NICOLAÜ, *Criterios de anàlise classificatoria hierarquica baseados na funçao de distribuiçao*, 1980, Faculté des Sciences de Lisbonne, thèse de doctorat soutenue en février 1981.
6. C. DE RHAM, *La classification hiérarachique ascendante selon la méthode des voisins réciproques*, Les Cahiers de l'Analyse des Données, vol. V, 1980, n° 2, p. 135-144.
7. M. L. TRICOT et M. DONEGANI, *Présentation unifiée des indices de proximité entre classes en classification hiérarchique ascendante*, Chaire de Statistique, Département de Mathématiques, École Polytechnique Fédérale de Lausanne, 1988.