REVUE DE STATISTIQUE APPLIQUÉE

BASAVANNEPPA TALLUR

Un nouvel indice d'agrégation en classification ascendante hiérarchique

Revue de statistique appliquée, tome 34, n° 2 (1986), p. 53-62

<http://www.numdam.org/item?id=RSA_1986__34_2_53_0>

UN NOUVEL INDICE D'AGRÉGATION EN CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

Basavanneppa TALLUR

I.R.I.S.A., Université de Rennes I,
Campus de Beaulieu, Avenue du Général Leclerc,
35042 Rennes Cédex

RÉSUMÉ

Un indice de proximité entre les lignes (ou colonnes) d'un tableau de contingence a été élaboré par I.C. LERMAN et B. TALLUR (R.S.A., Vol. XXVIII, n° 3, 1980) en vue d'une classification hiérarchique par l'Algorithme de la Vraisemblance des Liens (A.V.L.).

Nous présentons dans le présent article un nouvel indice d'agrégation directement basé sur l'indice de corrélation qui sera utilisé par un algorithme de classification ascendante hiérarchique, A.B.C. (Algorithme Basé sur la Corrélation), des éléments constitutifs d'un tableau de contingence. Cet algorithme (A.B.C.) sera ensuite comparé à l'A.V.L. à travers un exemple d'application aux données réelles.

SUMMARY

I.C. LERMAN and B. TALLUR (R.S.A., Vol. XXVIII, 3, 1980) have developed an index of proximity between rows (or columns) of contingency tables in view of hierarchical classification by Likelihood Link Algorithm (L.L.A.).

In the present paper, we define a new agregation criterion and describe an Algorithm of hierarchical classification Based on Correlation (A.B.C.). This algorithm is then compared to the L.L.A. by means of an application to the real data.

INTRODUCTION

Dans le cadre de l'Algorithme de la Vraisemblance des Liens (A.V.L.), nous avons développé un indice de proximité entre les lignes (resp. colonnes) d'un tableau de contingence qui tient compte de la représentation mathématique fidèle de la structure d'un tel tableau des données ([5]). Cet indice a ensuite été étendu aux différentes situations de juxtaposition de plusieurs tableaux de contingence. De nombreuses applications ont été réalisées dans les domaines aussi divers que la sociologie, l'épidémiologie, la médecine préventive, la géographie économique et l'économie rurale.

Le présent article propose un nouvel indice d'agrégation directement basé sur l'indice de proximité mentionné ci-dessus, permettant d'obtenir une classification ascendante hiérarchique.

Nous rappellerons brièvement dans le §1 l'indice de proximité entre les lignes (resp. colonnes) d'un tableau de contingence et d'une juxtaposition de

Mots-clés : **Classification Ascendante Hiérarchique (C.A.H.), Algorithme de la Vraisemblance des Liens (A.V.L.), Indice d'agrégation, Hiérarchie, Inversion.**

plusieurs tableaux de contingence. Nous présenterons l'indice d'agrégation entre les classes et l'algorithme basé sur la corrélation dans le §2. Le §3 sera consacré à une étude comparative de cet algorithme avec l'A.V.L. à travers un exemple d'application.

1. INDICE DE PROXIMITÉ (RAPPELS)

Considérons un tableau de contingence K_{IJ} :

$$K_{IJ} = \{k_{ij} / (i, j) \varepsilon IxJ\} \tag{1}$$

avec card (I) = n, card (J) = m

On utilisera les notations classiques suivantes :

$k_{i.} = \Sigma \{k_{ij}/j \varepsilon J\}$, $k_{.j} = \Sigma \{k_{ij}/i \varepsilon I\}$

$k_{..} = \Sigma \{k_{ij}/(i, j) \varepsilon IxJ\}$

$f_{ij} = k_{ij}/k_{..}$, $p_{i.} = k_{i.}/k_{..}$, $p_{.j} = k_{.j}/k_{..}$ (2)

$f^i_J = \{f^i_j/j \varepsilon J\}$ où $f_{ij}/p_{i.} = f^i_j$

f^i_J est le « profil » de la ligne i à travers l'ensemble J

La représentation mathématique de ce tableau de données est celle qui est fournie dans le cadre de l'Analyse des Correspondances. Considérons, pour fixer les idées, le cas où l'ensemble à classifier est l'ensemble J des colonnes. Nous associerons alors à l'ensemble I, le nuage N (I) de leurs profils dans R^m muni de la métrique diagonale des $(1/p_{.j}/j \varepsilon J)$ (la métrique du χ^2). Le nuage N (I) est défini par

$$N(I) = \{(f^i_J, p_{i.}) \; i \varepsilon I\} \tag{3}$$

où $p_{i.}$ est la masse affectée au point f^i_J pour tout i ε I.

L'ensemble I joue le rôle des individus et J, celui des variables.

L'idée originale consiste à assimiler une colonne j ε J à une variable numérique X_j en se situant dans le nuage N (I); la j-ème coordonnée du profil de l'individu i, f^i_j n'est autre que la i-ème valeur observée de la variable X_j. L'indice de proximité S_{jh} entre les colonnes j et h sera défini par le coefficient de corrélation entre les variables X_j et X_h qui leur sont associées respectivement, et en respectant les pondérations attachées à chacun des individus; l'expression explicite de ce coefficient est la suivante :

$$S_{jh} = \rho_{x_j, x_h} = \frac{\{\sum_{i \varepsilon I} (f_{ij} f_{ih}/p_{i.}) - p_{.j} p_{.h}\}}{[\{\sum_{i \varepsilon I} (f^2_{ij}/p_{i.}) - p^2_{.j}\}\{\sum_{i \varepsilon I} (f^2_{ih}/p_{i.}) - p^2_{.h}\}]^{1/2}} \tag{4}$$

Pour plus de détails, et pour son extension au cas de juxtaposition de plusieurs tableaux de contingence, nous renvoyons les lecteurs à la référence bibliographique [5].

2. L'INDICE D'AGRÉGATION ET L'ALGORITHME DE CLASSIFICATION BASÉ SUR LA CORRÉLATION

Supposons que l'ensemble à classifier soit celui des colonnes du tableau K_{IJ}. A chaque colonne $j \in J$ on associe une variable numérique X_j (voir §1 ci-dessus).

L'indice d'agrégation sera défini de la façon suivante.

Définition

Supposons qu'au pas (ou niveau) s on ait les classes C et D suivantes :

C est formée par la réunion de p colonnes $C_1, C_2, \ldots, C_p$
et D est formée par la réunion de q colonnes $D_1, D_2, \ldots, D_q$.

Alors, l'indice d'agrégation S_{CD} entre les classes C et D sera

$$S_{CD} = \left(\frac{1}{\sqrt{pq}}\right) \rho_{X_C, X_D} \tag{5}$$

où X_C (respectivement, X_D) est la variable somme des X_{C_i} $(1 \leqslant i \leqslant p)$ (respectivement, X_{D_j} $(1 \leqslant j \leqslant q)$), et X_{C_i} (respectivement X_{D_j}) est la variable associée à la colonne C_i (respectivement D_j).

Algorithme de Classification Ascendante Hiérarchique Basé sur la Corrélation (A.B.C.)

Pas 1. Au départ chaque colonne constitue une classe. On réunit les deux colonnes les plus proches au sens de l'indice d'agrégation (5) (c'est-à-dire, les deux colonnes qui maximisent l'indice S) avec $p = q = 1$:

$$S_{x_j x_h} = \rho_{x_j, x_h} \text{ pour tout } (j, h) \in J \times J,\ j \neq h.$$

Pas 2. On associe à la classe formée au pas précédent, la somme des variables associées aux colonnes constituant cette classe. Ainsi, la variable associée à la classe C réunissant les colonnes C_i $(1 \leqslant i \leqslant p)$ sera la variable X_C somme des variables X_{C_i} $(1 \leqslant i \leqslant p)$ associées aux colonnes C_i. Remarquons que cela revient à remplacer les colonnes réunies C_i $(1 \leqslant i \leqslant p)$ par leur somme et à associer à cette nouvelle colonne une variable numérique (comme au §1).

Pas 3. On réunit les classes les plus proches au sens de l'indice d'agrégation (5) et on retourne au Pas 2 jusqu'à ce que toutes les classes soient réunies.

Nous montrerons, en annexe, que l'algorithme décrit ci-dessus (et basé sur l'indice d'agrégation (5)) produit une hiérarchie « sans inversion », tant que l'indice S est positif ou nul, au sens suivant (remarquons que dans le cas où l'indice devient négatif, la fusion des classes n'a plus de sens) :

Définition

La hiérarchie de partitions sera dite « sans inversion » si la valeur maximale (resp. minimale) de l'indice d'agrégation fondé sur les similarités

(resp. dissimilarités) à un pas s quelconque est inférieure (resp. supérieure) à la valeur maximale (resp. minimale) de cet indice à tous les pas inférieurs à s.

Théorème

L'algorithme A.B.C. basé sur l'indice d'agrégation (5) produit une hiérarchie de partitions « sans inversion » au sens de la définition ci-dessus, tant que l'indice s est positif ou nul.

La démonstration de ce théorème est fournie en annexe.

Remarque

Malgré le fait que la hiérarchie de partitions produite par l'algorithme A.B.C. puisse présenter des inversions lorsque l'indice S devient négatif, cet algorithme est intéressant par sa démarche simple et naturelle pour les tableaux de contingence. Dans la pratique, la réunion des classes avec un indice de corrélation négatif n'étant pas intéressante, on pourra arrêter l'algorithme (ou tout au moins l'interprétation de l'arbre de classification) au niveau (pas) où le critère d'agrégation devient négatif.

Aides à l'interprétation

On peut utiliser les valeurs des statistiques « globale » et « locale », définies dans le cadre de l'A.V.L. (voir [3], chapitre 4) pour déterminer respectivement les meilleures partitions et les nœuds « significatifs ». L'arbre de classification sera condensé aux nœuds significatifs. Le degré de neutralité de chaque élément, mesuré par la variance des proximités de l'élément à tous les autres, sera calculé et les éléments seront rangés par ordre décroissant du degré de neutralité. Cette notion du degré de neutralité a également été définie dans le contexte de l'A.V.L. (voir [3], chapitre 3).

De plus, nous disposons de la valeur maximale de l'indice d'agrégation à chaque pas. Cette valeur est d'autant plus forte, si elle est positive, que la liaison entre les classes agrégées à un niveau donné est plus intense. Par contre, une valeur négative de cet indice avertit que les deux classes agrégées s'opposent entre elles.

3. APPLICATION AUX DONNÉES RÉELLES ET COMPARAISON AVEC L'A.V.L.

L'algorithme A.B.C. présenté ci-dessus a été appliqué à plusieurs tableaux des données et les résultats obtenus ont été toujours très satisfaisants. Nous présentons ci-dessous son application à l'étude sur l'hypertension artérielle et nous comparons les résultats avec ceux issus de l'A.V.L. Nous nous limiterons à la comparaison des suites de partitions obtenues par les deux algorithmes ainsi que les statistiques locales et globales relatives aux différents niveaux.

Etude sur l'hypertension artérielle

Cette étude importante dans le domaine de la médecine préventive a fait l'objet d'une communication au Colloque National des Centres d'Examens de Santé et dont les conclusions vont très probablement permettre de faire un pas décisif vers la prévention des maladies cardiovasculaires ([2]). Nous avons, dans cette étude, utilisé la méthode de classification par l'A.V.L. aussi bien sur des tableaux binaires que sur des tableaux de contingence. Nous avons repris ici l'analyse du tableau de contingence croisant les modalités de la variable « Tension Artérielle Systolique » (T.A.S.) avec l'ensemble des modalités des variables biologiques et sociales. La classification par l'A.V.L. de l'ensemble des modalités de la T.A.S. nous a notamment permis de formuler l'hypothèse d'existence du seuil de « sécurité » de la T.A.S. à 13 pour les femmes de 30 à 40 ans ([7]).

Nous reproduisons ci-dessous, l'arbre de classification issu de l'application de l'A.V.L. (Fig. 1) et celui produit par l'algorithme basé sur la corrélation (Fig. 2).

Les partitions obtenues aux différents niveaux sont différentes pour les deux méthodes, à l'exception des niveaux 1, 2 et 6.

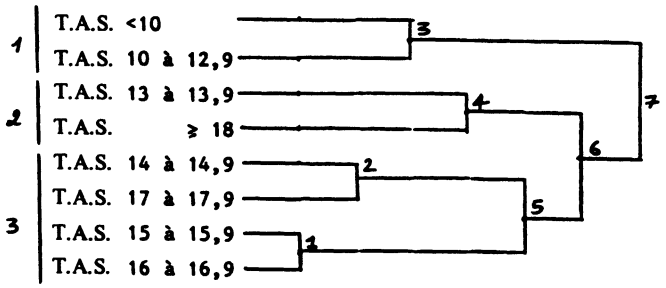

Figure 1. — Arbre détaillé de classification des modalités de la T.A.S. issu de l'A.V.L.

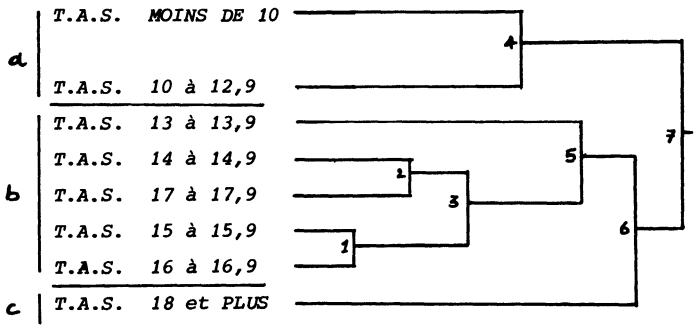

Figure 2. — Arbre détaillé de classification des modalités de la T.A.S. issu de l'algorithme A.B.C.

On peut faire les observations suivantes :

Les deux méthodes séparent les valeurs faibles (< 13) de la T.A.S. de celles fortes ($\geqslant 13$), au niveau 6 de l'arbre où la partition comporte deux classes.

Au niveau 5, les partitions en trois classes définies par les deux algorithmes sont peu différentes :

En effet, à ce niveau l'A.V.L. produit les trois classes suivantes des valeurs de la T.A.S. :

1) < 13,
2) 13 à 13,9 et $\geqslant 18$,
3) 14 à 17,9.

Puisque la modalité « T.A.S. $\geqslant 18$ » est la plus « neutre » de toutes et que son effectif est très faible, on peut considérer que la classe 2) concerne les valeurs de la T.A.S. allant de 13 à 13,9.

A ce même niveau 5, l'A.B.C. sépare les trois groupes de tension :

a) faible (< 13),
b) intermédiaire (13 à 17,9),
c) forte ($\geqslant 18$).

On remarque ici que le groupement des modalités est plus harmonieux que dans le cas précédent, et que la connexité des valeurs est mieux respectée. Mais si compte tenu du caractère « neutre » de la modalité « T.A.S. $\geqslant 18$ » et de son faible effectif, on ne retenait que les deux premières classes comme interprétables à ce niveau, et que l'on descendait au niveau 4 pour obtenir une partition en trois classes, on retomberait sur les mêmes classes que celles produites par l'A.V.L. (que l'on vient de décrire plus haut) (sans tenir compte de la modalité neutre « T.A.S. $\geqslant 18$ »).

TABLEAU
Comparaison de la Statistique Globale pour les deux méthodes

Niveau	Statistique Globale	
	A.V.L.	A.B.C.
1	1,671	1,671
2	2,141	2,141
3	2,265	2,911
4	2,364	3,104
5	3,265	4,210
6	4,317	4,317

Comparaison de la statistique globale

On peut observer qu'aux niveaux 3, 4 et 5 où les partitions obtenues par les deux algorithmes sont différentes, les valeurs de la statistique globale pour l'A.B.C. sont plus fortes que pour l'A.V.L. La statistique globale étant

une mesure d'adéquation entre une partition et la préordonnance (qui est la même pour les deux algorithmes) définie par l'indice de similarité sur l'ensemble de couples d'éléments à classifier, on peut considérer que les partitions sont plus proches de l'ordonnance dans le cas de l'A.B.C. Mais nous avons remarqué au cours de notre expérience pratique que ce qui est plus intéressant dans l'interprétation des résultats c'est l'évolution de la valeur de ce critère, et non pas sa valeur absolue.

4. CONCLUSION

Le nouvel indice d'agrégation (5) utilisé par l'algorithme A.B.C. dans la construction d'une classification ascendante hiérarchique a été obtenu d'une manière naturelle pour l'analyse des tableaux de contingence. L'A.B.C. étant particulièrement bien adapté à cette structure du tableau des données, il produit des classes plus harmonieuses que l'A.V.L. pour ce type de données. De nombreuses applications que nous avons effectuées montrent que cet indice pourrait donner des résultats intéressants dans le cas des tableaux des données numériques, des tableaux disjonctifs complets ainsi que des tableaux d'incidence (attributs descriptifs).

BIBLIOGRAPHIE

[1] J.P. BENZECRI et collaborateurs. — *Analyse des Données, Tome 2,* Analyse des Correspondances, Dunod, Paris (1973).

[2] M. CAILLET, L. MASSÉ, H. COURCOUX, E. COSTE, E. ABOU, B. TALLUR et B. DUPONT. — Importance du niveau de la Tension Artérielle Systolique dans la Sélection de Population-Cible en Médecine Préventive. *Communication du 5e Colloque National des Centres d'Examens de Santé,* Bordeaux, Juin 1981.

[3] I.C. LERMAN. — *Classification et Analyse Ordinale des Données,* Dunod (1981).

[4] I.C. LERMAN. — Programme de Classification Hiérarchique. *Rapport IRISA n° 148,* Université de Rennes I, Juin 1981.

[5] I.C. LERMAN et B. TALLUR. — Classification des Eléments Constitutifs d'une justaposition de tableaux de Contingence. *Rapport IRISA n° 127,* Université de Rennes I, 1980; R.S.A., Vol. XXVIII, N° 3, 1980.

[6] B. TALLUR. — « Etude de l'Agriculture Régionale Française », *Rapport IRISA n° 103,* Université de Rennes I, 1978.

[7] B. TALLUR. — Méthode d'Interprétation d'une Classification Hiérarchique d'Attributs-Modalités pour l'Explication d'une variable; Application à la Recherche du seuil critique de la Tension Systolique et des Indicateurs de Risques Cardiovasculaires. *Rapport IRISA n° 159,* Université de Rennes I, Janvier 1982; R.S.A., Vol. XXXI, N° 1, 1983.

ANNEXE

Démonstration du théorème

Supposons qu'au pas h on ait réuni les classes C et D en une classe $E = C \cup D$. Soient $\text{card}(C) = p$, $\text{card}(D) = q$. Supposons qu'au pas $h + 1$ les classes F et G soient réunies.

Nous allons montrer que la valeur maximale de l'indice S au pas $h + 1$ est inférieure à celle au pas h (c'est-à-dire que $S_{FG} < S_{CD}$).

Cas où $F \neq E$ et $G \neq E$

D'après l'algorithme, il est clair que

$$S_{FG} < S_{CD}$$

Cas où $F = E$ ou $G = E$

Supposons que $G = E$ avec $F \neq E$ et que $\text{card}(F) = r$.

Notons

$X_C = \sum_{i=1}^{p} X_i$ la variable associée à la classe C,

$$X_D = \sum_{j=1}^{q} Y_j \quad \text{et} \quad X_F = \sum_{k=1}^{r} Z_k$$

celles associées respectivement aux classes D et F.

Si $S^{(i)}$ désigne la valeur maximale de l'indice S au pas i, nous avons

$$S^{(h)} = S_{CD} = \left(\frac{1}{\sqrt{pq}}\right) \rho_{X_C, X_D}$$

$$= \left(\frac{1}{\sqrt{pq}}\right) \frac{\sum_{i=1}^{p} \sum_{j=1}^{q} \text{COV}(X_i, Y_j)}{\sqrt{\text{Var}\left(\sum_{i=1}^{p} X_i\right)} \sqrt{\text{Var}\left(\sum_{j=1}^{q} Y_j\right)}}$$

Posons

$\sigma_i^2 = \text{Var}(X_i)$, $\sigma_j^2 = \text{Var}(Y_j)$, $\sigma_{ij} = \text{COV}(X_i, Y_j)$

$\sigma_{ii'} = \text{COV}(X_i, X_{i'})$, $\sigma_{jj'} = \text{COV}(Y_j, Y_{j'})$

$$a = \sqrt{\sum_{i} \sigma_i^2 + \sum_{i \neq i'} \sigma_{ii'}} > 0 \text{ et}$$

$$b = \sqrt{\sum_{j} \sigma_j^2 + \sum_{j \neq j'} \sigma_{jj'}} > 0$$

On peut écrire :

$$S^{(h)} = \left(\frac{1}{\sqrt{pq}}\right) \frac{\sum_{i=1}^{p} \sum_{j=1}^{q} \sigma_{ij}}{ab} \tag{6}$$

Avec des notations analogues, nous avons :

$$S_{CF} = \left(\frac{1}{\sqrt{pr}}\right) \rho x_C, x_F$$

Posant

$\sigma_k^2 = \text{Var}(Z_k)$, $\sigma_{kk'} = \text{COV}(Z_k, Z_{k'})$, $\sigma_{ik} = \text{COV}(X_I, Z_k)$ et

$$c = \sqrt{\sum_{k} \sigma_k^2 + \sum_{k \neq k'} \sigma_{kk'}}$$

On a :

$$S_{CF} = \left(\frac{1}{\sqrt{pr}}\right) \frac{\sum_{i} \sum_{k} \sigma_{ij}}{ac} \tag{7}$$

De même, en posant $\sigma_{jk} = \text{COV}(Y_j, Z_k)$, on a :

$$S_{DF} = \left(\frac{1}{\sqrt{qr}}\right) \frac{\sum_{j} \sum_{k} \sigma_{jk}}{bc} \tag{8}$$

A partir de (6), on obtient :

$$\sum_{i} \sum_{j} \sigma_{ij} = \sqrt{pq}\, abS^{(h)} \tag{9}$$

Puisqu'au pas h, la valeur maximale de S est $S^{(h)}$, on a :

$$S_{CF} < S^{(h)} \Rightarrow \sum_{i=1}^{p} \sum_{k=1}^{r} \sigma_{ik} < \sqrt{pr}\, acS^{(h)} \tag{10}$$

et

$$S_{DF} < S^{(h)} \Rightarrow \sum_{j=1}^{q} \sum_{k=1}^{r} \sigma_{jk} < \sqrt{qr}\, bcS^{(h)}$$

Nous avons, d'autre part :

$$S^{(h+1)} = S_{(CUD)F} = \left(\frac{1}{\sqrt{(p+q)\,r}}\right) \rho(X_C + X_D), X_F$$

$$= \left(\frac{1}{\sqrt{(p+q)\,r}}\right) \frac{\sum_{i} \sum_{k} \sigma_{ik} + \sum_{j} \sum_{k} \sigma_{jk}}{\left(\sqrt{a^2 + b^2 + 2 \sum_{i} \sum_{j} \sigma_{ij}}\right) c}$$

$$< \left(\frac{1}{\sqrt{(p+q)\,r}}\right) \frac{(\sqrt{pr}\, ac + \sqrt{qr}\, bc)\, S^{(h)}}{(\sqrt{a^2 + b^2 + 2\sqrt{pq}\, abS^{(h)}})\, c}$$

(voir (9) et (10))

$$\Rightarrow S^{(h+1)} < \frac{(\sqrt{p}\, a + \sqrt{q}\, b)\, S^{(h)}}{\sqrt{p+q}\, \sqrt{(a^2 + b^2 + 2\sqrt{pq}\, abS^{(h)})}}$$

Le dénominateur du second membre peut s'écrire comme ci-dessous :

$$\sqrt{(\sqrt{p}\,a + \sqrt{q}\,b)^2 + (\sqrt{q}\,a - \sqrt{p}\,b)^2 + 2(p+q)\sqrt{pq}\,abS^{(h)}}.$$

D'où on a :

$$S^{(h+1)} < \frac{S^{(h)}}{\sqrt{\left[1 + \frac{(\sqrt{q}\,a - \sqrt{p}\,b)^2 + 2(p+q)\sqrt{pq}\,abS^{(h)}}{(\sqrt{p}\,a + \sqrt{q}\,b)^2}\right]}}$$

$< S^{(h)}$ si $S^{(h)}$ est positif ou nul.