Revue française d'automatique informatique recherche opérationnelle. Informatique théorique

M. Gondran

Valeurs propres et vecteurs propres en classification hiérarchique

Revue française d'automatique informatique recherche opérationnelle. Informatique théorique, tome 10, n° R1 (1976), p. 39-46

<http://www.numdam.org/item?id=ITA_1976__10_1_39_0>

R.A.I.R.O. Informatique Théorique
(vol. 10, n° 3, mars 1976, p. 39 à 46)

VALEURS PROPRES ET VECTEURS PROPRES EN CLASSIFICATION HIÉRARCHIQUE (*)

par M. GONDRAN ([1])

Communiqué par J. BERSTEL

Résumé. — *On montre dans cette note qu'un niveau d'une classification hiérarchique basée sur une ultramétrique correspond à une valeur propre de la matrice des distances. Alors les vecteurs propres générateurs du semi-module correspondant à cette valeur propre sont en bijection avec les différents types de ce niveau.*

Cette interprétation des niveaux et des types d'une classification hiérarchique permet ainsi de lier l'analyse factorielle et la classification dans une théorie unifiée.

1. INTRODUCTION

Les deux approches classiques de l'analyse des données, analyse factorielle et classification paraissent à première vue étrangères.

L'analyse factorielle semble s'intéresser plus au quantitatif et la classification au qualitatif.

Nous essayons de montrer dans cette note que les approches sont très voisines. En effet, on montrera que la recherche d'une classification hiérarchique basée sur une ultramétrique se ramène à la recherche des valeurs propres et des vecteurs propres de la matrice des dissimilarités dans une structure algébrique appropriée (semi-module sur un semi-anneau, *cf.* Gondran [1], Gondran et Minoux [2]).

Ainsi la différence entre les deux approches ne se trouve en fait que dans la structure algébrique choisie pour le traitement des données.

Au paragraphe 2, nous introduirons la structure algébrique sous-jacente à une ultramétrique. C'est le semi-anneau $(S, \oplus, \bigstar)$, où $S = \mathbf{R}^+ \, U \, \{ +\infty \}$ et où les opérations $\oplus$ et $\bigstar$ correspondent aux opérations « min » et « max ». Nous montrerons alors la liaison de l'ultramétrique sous dominante avec cette structure algébrique, puis nous rappellerons les liaisons de cette ultramétrique avec l'arbre de poids minimal et les classifications hiérarchiques indicées.

Au paragraphe 3, nous établirons le théorème fondamental liant les valeurs propres et les vecteurs propres de la matrice des dissimilarités avec les niveaux et les différents types d'une classification hiérarchique.

(*) Reçu mai 1975.

([1]) Électricité de France, Direction des Études et Recherches, Service informatique et Mathématiques appliquées, Clamart.

Ainsi un niveau d'une classification hiérarchique basée sur l'ultramétrique sous dominante correspond à une valeur propre de la matrice des dissimilarités. Alors les vecteurs propres générateurs du semi-module correspondant à cette valeur propre sont en bijection avec les différentes classes de ce niveau.

Ce théorème a en fait une interprétation plus vaste car à toute hiérarchie indicée correspond une distance ultramétrique (*cf.* Benzecri [13]). On peut alors conclure que les niveaux et les classes de toute classification hiérarchique sont en bijection avec les valeurs propres et les vecteurs propres de la matrice des distances de l'ultramétrique correspondante.

Enfin au paragraphe 4, nous donnons un exemple illustrant les résultats précédents.

2. PROPRIÉTÉS DES CLASSIFICATIONS LIÉES A UNE ULTRAMÉTRIQUE

2.1. La structure algébrique sous-jacente

Soit un ensemble de n objets. On définit entre deux objets quelconques i, j, un indice de dissimilarité $a_{ij} \in \mathbf{R}^+$, indice qui peut ne pas satisfaire les axiomes de la distance.

La matrice $A = (a_{ij})$ peut être considérée comme la matrice d'incidence généralisée d'un graphe symétrique G dont les arêtes seront valuées par les a_{ij}.

La distance ultramétrique sous dominante $\delta(i, j)$ peut alors être définie de la façon suivante (*cf.* Roux [10]) :

$$\delta(i, j) = \min_{\Pi \in C_{ij}} [\max(a_{i_1 i_2}, a_{i_2 i_3}, \ldots, a_{i_r, i_{r+1}})] \tag{2.1}$$

où C_{ij} représente l'ensemble des chemins de i à j; Π représente le chemin $(i_1, i_2, i_3, \ldots, i_r, i_{r+1})$ avec $i_1 = i$, $i_{r+1} = j$.

Considérons le semi-anneau $(S, \oplus, \bigstar)$, où $S = \mathbf{R}^+ \, U \, \{+\infty\}$ et ou les opérations « addition » et « multiplication » correspondent aux opérations « min » et « max » ($\oplus$ = min, $\bigstar$ = max).

$\oplus$ admet l'élément neutre $\varepsilon = +\infty$ appelé élément nul et $\bigstar$ admet comme élément neutre $e = 0$, appelé unité.

De plus ε est absorbant pour $\bigstar$.

On définira l'addition et la multiplication pour les matrices carrées d'ordre n à éléments dans S à partir des lois $\oplus$ et $\bigstar$. L'ensemble des matrices $M_n(S)$ ainsi défini admet alors la même structure que S avec comme élément nul la matrice

$$\Sigma = \begin{bmatrix} \varepsilon & \ldots & \varepsilon \\ \vdots & & \vdots \\ \varepsilon & \ldots & \varepsilon \end{bmatrix}$$

et comme unité la matrice

$$E = \begin{bmatrix} e & & \varepsilon \\ & \ddots & \\ \varepsilon & & e \end{bmatrix}$$

On notera encore $\oplus$ et $\bigstar$ les opérations induites sur $M_n(S)$.

On peut alors montrer (*cf.* Carré [3], Gondran [4], [5]) que la matrice de la distance ultramétrique sous dominante est

$$A^* = A \oplus A^2 \oplus \ldots \oplus A^{n-1} = A \oplus A^2 \oplus \ldots \oplus A^n = \ldots = A^{n-1} \quad (2.2)$$

et vérifie les équations

$$A^* = AA^* \oplus A = A^* A \oplus A, \quad (2.3)$$

on a même ici, puisque $A^* = A^* \oplus E$,

$$A^* = AA^* = A^* A = A^* A^*. \quad (2.4)$$

A^* peut alors être calculé à partir de (2.3) par la méthode de Gauss (*cf.* Carré [3], Gondran [4], [5]).

2.2. L'arbre de poids minimum

En considérant a_{ij} comme la capacité inférieure de l'arête (i, j), $\delta(i, j)$ représente *la capacité du chemin de capacité minimale entre i et j.*

Considérons alors l'arbre de poids minimum du graphe ayant pour longueur les a_{ij}. On a alors la propriété suivante (*cf.* Kalaba [6], Hu [7]) : *un chemin de capacité minimale entre les sommets i et j est obtenu en considérant le chemin liant i à j dans l'arbre de poids minimum.*

La recherche d'un arbre de poids minimum permet donc d'obtenir la distance ultramétrique sous dominante entre tous les objets.

Nous allons tirer quelques conséquences très importantes de la propriété précédente.

(*a*) Un arbre ayant $n-1$ arêtes, la matrice A^* a au plus n coefficients distincts (les $n-1$ longueurs des arêtes de l'arbre de poids minimal et l'élément unité e).

Si la matrice A a tous ses coefficients distincts (à part les $a_{ii} = e = 0$ et ceux égaux à $\varepsilon = +\infty$), A^* a alors exactement n coefficients distincts.

(*b*) D'après l'algorithme de Kruskal [8], l'arbre de poids minimum et donc l'ultramétrique sous dominante, ne dépendent que de l' « ordonnance » entre les indices de dissimilarité a_{ij} (*cf.* les algorithmes de Johnson [11] et de Lerman [12]).

(*c*) Ordonnons les p ($p \leqq n$) éléments de A^* :

$$\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_k \geqq \ldots \geqq \lambda_p = e = 0.$$

Alors l'arbre de classification de la classification hiérarchique liée à l'ultramétrique sous dominante aura p niveaux et les indices de ces niveaux seront les λ_k.

L'arbre de classification sera obtenu en connectant à chaque étape les sommets adjacents de l'arbre de poids minimum pris dans l'ordre des λ_k croissants.

La classification de niveau λ_k correspond aux composantes connexes de l'arbre de poids minimum auquel on a enlevé les arêtes plus grandes que λ_k (*cf.* [14]).

Nous donnons dans [9] un algorithme simple et efficace basé sur les remarques précédentes.

3. VALEURS PROPRES ET VECTEURS PROPRES EN TYPOLOGIE

Un vecteur $V \in S^n$ est un *vecteur propre* de la matrice A dans le semi-anneau S si $V \neq \varepsilon$ et s'il existe $\mu \in S$, appelé *valeur propre*, tel que

$$A V = \mu V, \tag{3.1}$$

où μV correspond au vecteur de composante $\mu \bigstar v_i$.

LEMME 1 : *Pour tout* $\mu \in S$, *le vecteur* $\begin{pmatrix}\mu \\ \vdots \\ \mu\end{pmatrix}$ *est vecteur propre de* A.

Démonstration :

$$a_{ij} \bigstar \mu \geqq \mu$$

donc

$$\sum_{j=1}^{n} a_{ij} \mu \geqq \mu,$$

et comme

$$a_{ii} = 0, \qquad \sum_{j=1}^{n} a_{ij} \mu = \mu,$$

donc

$$A \begin{pmatrix}\mu \\ \vdots \\ \mu\end{pmatrix} = \mu \begin{pmatrix}\mu \\ \vdots \\ \mu\end{pmatrix}. \quad \square$$

Toutes les valeurs de S sont donc valeurs propres.

Soit $\mathscr{V}_\mu$ l'ensemble des vecteurs propres associés à la valeur propre μ.

LEMME 2 : *Quel que soit $V \in \mathscr{V}_\mu$, on a*

$$V = \mu V = A V = A^* V. \tag{3.2}$$

Démonstration : (3.1) entraîne pour tout i :

$$\sum_j a_{ij} v_j = \mu v_i,$$

c'est-à-dire : $v_i = a_{ii} v_i \geqq \mu v_i$, donc $v_i = \mu v_i$.

$V = AV$ entraîne alors : $V = A^2 V = A^3 V = \ldots = A^* V$. □

Posons

$$V_\mu^i = \mu (A^*)^i. \tag{3.3}$$

μ étant donné, on peut associer à tout objet i, le vecteur V_μ^i.

LEMME 3 : *Deux objets i et j sont dans la même classe de niveau μ de l'arbre de classification si et seulement si leurs vecteurs associés V_μ^i et V_μ^j sont égaux.*

Démonstration : immédiate en considérant la liaison entre l'arbre de poids minimal et la matrice A^*.

LEMME 4 : *$\mathscr{V}_\mu$ est identique au semi-module engendré par les V_μ^i.*

Démonstration : (i) Montrons d'abord que tout élément du semi-module est un vecteur de $\mathscr{V}_\mu$.

Tout élément W du semi-module engendré par les V_μ^i est de la forme

$$W = \sum_{i=1}^{n} x_i V_\mu^i.$$

Comme d'après (2.4), $(A^*)^i = A (A^*)^i$ et que $\mu^2 = \mu$, on a

$$A V_\mu^i = \mu A (A^*)^i = \mu (A^*)^i = \mu\mu (A^*)^i = \mu V_\mu^i, \tag{3.4}$$

on en déduit

$$A W = \mu W.$$

(ii) Montrons maintenant que tout vecteur de $\mathscr{V}_\mu$ est engendré par les V_μ^i.

Immédiat, car d'après (3.2), on a $V = A^* \mu V$, c'est-à-dire :

$$V = \sum_{i=1}^{n} v_i V_\mu^i. \quad \square \tag{3.5}$$

LEMME 5 : *Un vecteur V_μ^i ne peut être engendré par aucune combinaison d'autres vecteurs.*

Démonstration : Supposons que $V_\mu^i = \sum \gamma_k W^k$ avec $W^k \in \mathscr{V}_\mu$.

(i) Puisque la composante i de V_μ^i est μ, il existe k' tel que $(\gamma^{k'} W^{k'})_i = \mu$; ce qui entraîne $\gamma^{k'} \leqq \mu$.

Comme $W^{k'} \in \mathscr{V}_\mu$, on a alors en utilisant le lemme 2, $\gamma^{k'} W^{k'} = W^{k'}$; donc $V_\mu^i \leqq W^{k'}$.

(ii) D'après (3.5), on a

$$W^{k'} = \sum_i (W^{k'})_i V_\mu^i$$

et comme $(W^{k'})_i = \mu$, on déduit

$$W^{k'} \leqq \mu V_\mu^i = V_\mu^i.$$

Des deux inégalités précédentes, on déduit $W^{k'} = V_\mu^i$. □

THÉORÈME : *Un niveau d'une classification hiérarchique basée sur une ultramétrique correspond à une valeur propre de la matrice des distances. Il existe alors une base unique du semi-module des vecteurs propres correspondants; chacun de ces vecteurs propres définit un type de la classification hiérarchique de ce niveau.*

Remarque : Les valeurs propres et les vecteurs propres de A^* sont identiques à celles de A. Cela permet de donner une interprétation plus vaste du théorème précédent.

En effet comme à toute hiérarchie indicée correspond une ultramétrique (*cf.* Benzecri [13]), on peut conclure que les niveaux et les classes de *toute* classification hiérarchique sont en bijection avec les valeurs propres et les vecteurs propres de la matrice des distances de l'ultramétrique correspondante.

4. EXEMPLE

Considérons la matrice des dissimilarités suivantes :

$$A = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \\ i \end{array} \begin{array}{c} \begin{array}{ccccccccc} a & b & c & d & e & f & g & h & i \end{array} \\ \begin{bmatrix} 0 & & & & & & & & \\ 7 & 0 & & & & & & & \\ 5 & 2 & 0 & & & & & & \\ 8 & 10 & 7 & 0 & & & & & \\ 10 & 9 & 11 & 8 & 0 & & & & \\ 8 & 9 & 10 & 4 & 9 & 0 & & & \\ 10 & 10 & 9 & 11 & 5 & 10 & 0 & & \\ 12 & 11 & 9 & 11 & 1 & 9 & 6 & 0 & \\ 10 & 9 & 9 & 10 & 6 & 7 & 3 & 6 & 0 \end{bmatrix} \end{array}$$

L'arbre de poids minimum est alors :

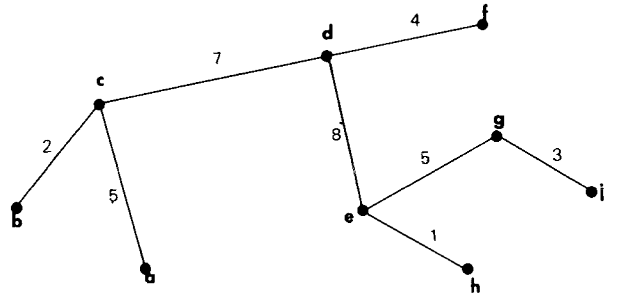

Figure 1

Les classes du niveau 5 sont alors les composantes convexes suivantes :

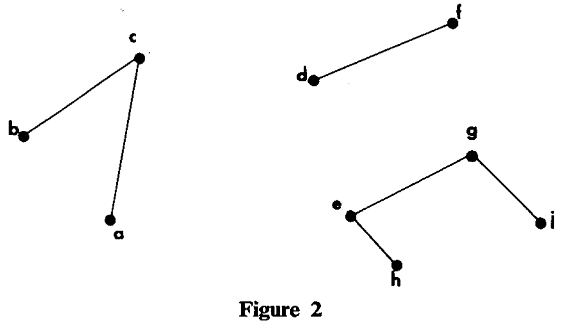

Figure 2

et le système des vecteurs propres associés est

$$\begin{bmatrix}5\\5\\5\\7\\8\\7\\8\\8\\8\end{bmatrix} \text{ pour } abc, \quad \begin{bmatrix}7\\7\\7\\5\\8\\5\\8\\8\\8\end{bmatrix} \text{ pour } df, \quad \begin{bmatrix}8\\8\\8\\8\\5\\8\\5\\5\\5\end{bmatrix} \text{ pour } eghi.$$

BIBLIOGRAPHIE

1. M. GONDRAN. *Théorème de Perron-Frobenius dans les semi-anneaux*. Note E.D.F. (à paraître).
2. M. GONDRAN et M. MINOUX. *Valeurs propres et vecteurs propres en théorie des graphes*. Note E.D.F. HI 1941/02, 22 septembre 1975.

3. B. A. CARRE. *An Algebra for Network Routing Problems.* J. Inst. Maths. Applics, vol. 7, 1971, p. 273-294.
4. M. GONDRAN. *Algèbre linéaire et cheminement dans un graphe.* R.A.I.R.O., 9e année, janvier 1975, V-1, p. 81-103.
5. M. GONDRAN. *Algèbre des chemins et algorithmes.* Note E.D.F. HI 1753/02, 10 août 1974; parue dans Combinatorial Programming: Methods and Applications, D. Reidel Publishing Co., Dordrecht, Hollande, 1975. B. Roy, éditeur, en anglais sous le titre *Path Algebra and Algorithms.*
6. R. KALABA. *On some Communication Network Problems.* Ch. 21 in *Combinatorial Analysis.* Proc. Sympos. Appl. Math., vol. 10, American Mathematical Society, Rhode, Island, 1960.
7. T. C. HU. *The Maximum Capacity Route Problem.* Ops. Res., vol. 9, 1961, p. 898-900.
8. J. B. KRUSKAL Jr. *On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem.* Proc. Amer. Math. Soc., n° 7, 1956, p. 40-50.
9. P. COLLOMB et M. GONDRAN. *Un algorithme efficace pour l'arbre de classification.* Note E.D.F. HI 1857/02 du 26 juin 1975.
10. M. ROUX. *Un algorithme pour construire une hiérarchie particulière.* Thèse de 3e cycle (L.S.M. I.S.U.P.), 1968.
11. S. C. JOHNSON. *Hierarchical Clustering Schemes.* Psychometrica, vol. 32, 1967, p. 241-243.
12. I. C. LERMAN. *Les bases de la classification automatique.* Gauthier-Villars, Paris, 1970.
13. J. P. BENZECRI. *L'Analyse des données,* tome 1 = Taxinomie, Dunod, Paris, 1974.
14. C3E. *Analyse des données multidimensionnelles,* tome III, 1972.