Statistique et analyse des données

Christophe Perruchet

Classification hiérarchique de structures mathématiques

Statistique et analyse des données, tome 7, n° 3 (1982), p. 55-67
<http://www.numdam.org/item?id=SAD_1982__7_3_55_0>

Statistiques et Analyse de données
1982 - Vol. 7 n° 3 pp. 55-67

CLASSIFICATION HIERARCHIQUE DE STRUCTURES MATHEMATIQUES

Christophe PERRUCHET

Departement de Mathématiques Appliquées pour les Télécommunications et l'Informatique

Centre National d'Etudes des Télécommunications

92131 ISSY LES MOULINEAUX

Resume : *Le domaine d'application des méthodes de classification s'est largement étendu ces quinze dernières années, mais toujours pour des données concrètes issues d'expériences, d'observations, ou de simulations. On présente ici des applications de la classification à des structures mathématiques. D'une part à l'ensemble des séries formelles définies sur un corps quelconque, d'autre part à l'ensemble des entiers muni de la distance p-adique.*

Abstract : *The application area of the methods of cluster analysis has largely developed during the last fifteen years, but always for concrete data resulting from experiments, observations, or simulations. This paper presents an application of hierarchical clustering to mathematical structures. On the one hand the set of formal series defined on a field, on the other hand the set of integers fit with the p-adic distance.*

Mots-clés : *Classification, Hiérarchie, Séries formelles, Distance p-adique, Entiers relatifs.*

0 - INTRODUCTION

Le domaine d'application des méthodes de classification s'est considérablement élargi ces quinze dernières années et touche aussi bien la zoologie que les télécommunications, la médecine que la géochimie, la psychologie sociale que l'informatique.

Pendant la même période s'est développée une théorie mathématique de la classi-

fication (Benzecri, Diday, Jambu, Lerman...), qui sert maintenant d'outil pour tous les travaux de recherche. Cette théorie se développe encore actuellement, par elle-même, et par ses emprunts à la statistique, la théorie des graphes, etc...

Cependant, les applications de la classification se sont toujours faites sur des données concrètes issues de l'expérience ou, plus simplement, de l'observation.

On propose ici une application de la classification hiérarchique à des structures mathématiques précises. D'une part à l'ensemble des séries formelles définies sur un corps quelconque, d'autre part à l'ensemble des entiers munis de la distance p-adique.

Outre un aspect purement technique, ce travail a aussi pour but de montrer comment, et en quoi, les méthodes de classification permettent d'appréhender la structure d'un champ de données strictement organisées.

I - HIERARCHIE INDICEE

On rappelle ici quelques notions suffisantes pour la compréhension de ce travail. Le lecteur intéressé trouvera un exposé détaillé dans [1], p. 119.

1.1. Hiérarchie de parties

- Soit E un ensemble, on appelle hiérarchie de parties de E un ensemble $\mathcal{H}$ de parties non vides de E vérifiant :

. l'axiome d'intersection :

$\forall \{A,B\} \subset \mathcal{H} : A \cap B \in \{A,B,\emptyset\}$

. l'axiome de réunion :

$\forall A \in \mathcal{H} : \cup \{B/B \in \mathcal{H}, B \neq A, B \subset A\} \in \{A,\emptyset\}$

i.e, tout élément de non minimal pour l'inclusion, est la réunion des éléments distincts de lui, inclus dans lui.

- la hiérarchie $\mathcal{H}$ est totale si on a :

$E \in \mathcal{H}$

et

$\forall x \in E : \{x\} \in \mathcal{H}$

- l'ensemble des successeurs de A est défini par :

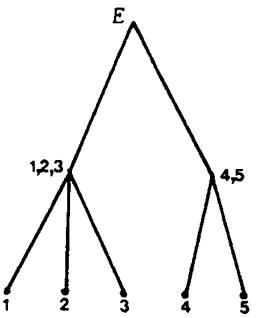
$Su(A) = \{ B \in \mathcal{H} / B \subset A, B \neq A \}$

- l'ensemble des successeurs immédiats de A est défini par :

$Sui(A) = \{B \in Su(A) / \forall C \in Su(A) : B \cap C \in \{\emptyset, C\}\}$

- Toute hiérarchie peut être représentée par un arbre, par exemple la hiérarchie suivante définie sur $E = \{1, 2, 3, 4, 5\}$; par :

$\mathcal{H}$ = {{1} ,{ 2} ,{ 3} ,{ 4} ,{ 5} ,{ 1, 2, 3} ,{ 4,5} , E}
est représentée par l'arbre :



Cette hiérarchie est totale, et on a :

Su (E) = {{1,2,3} , {4,5} ,{ 1} ,{ 2} ,{ 3} , {4} ,{ 5}}

Sui (E) ={{ 1,2,3} , { 4,5} }

Su ({4,5}) = Sui ({ 4,5}) ={{ 4} , { 5}}

- Rappelons enfin la proposition suivante que nous aurons à utiliser.

Un ensemble $\mathcal{H}$ de parties non vides d'un ensemble E est une hiérarchie totale sur E si et seulement si $\mathcal{H}$ vérifie :

- l'axiome d'intersection
- et $E \in \mathcal{H}$; $\forall x \in E$: $\{x\} \in \mathcal{H}$

1.2. Hiérarchie indicée

Une hiérarchie totale $\mathcal{H}$ sur un ensemble E munie d'une fonction v à valeurs réelles vérifiant :

. $\forall \{A,B\} \in \mathcal{H}$: $(A \subset B,\ A \neq B) \Longrightarrow v(A) < v(B)$

. $\forall x \in E$: $v(\{x\}) = 0$

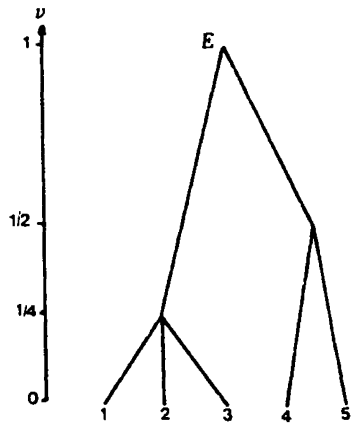
. $v(E) = 1$

est appelée hiérarchie indicée

- v est appelé indice de niveau ou indice de diamètre de la hiérarchie.
- Graphiquement l'indice de niveau se représente comme un axe associé à l'arbre.

Par exemple, si nous reprenons la hiérarchie illustrant le § 1.1., en définissant v par: v({1,2,3}) = 1/4 et v({ 4,5}) = 1/2

On obtient l'arbre indicé suivant :



2- HIERARCHIE DE SERIES

2.1. L'ultramétrique :

Considérons l'ensemble des séries formelles définies sur un corps K, on note K [[X]] cet ensemble.

Un élément de K [[X]] est S défini par :

$\forall x \in K : S(x) = \Sigma \{a_n x^n / n \in IN\}$

où $\{a_n\}$ est une suite d'éléments de K.

On se propose de déterminer la hiérarchie induite sur K [[X]] par une certaine distance ultramétrique.

Pour cela, on appelle "ordre de la série S", le rang de son premier terme non nul : $O(S) = \inf \{n / n \in IN, a_n \neq 0\}$

Proposition : soit d, défini par :

$\forall S \in K [[X]], \forall S' \in K [[X]] - \{S\} : d(S,S') = (\frac{1}{2})^{O(S-S')}$

$\forall S \in K [[X]]: d(S,S) = 0$

alors d est une distance ultramétrique sur K[[X]] .

Démonstration :

Pour tout triplet { S, S', S" } de K [[X]],

$d(S,S) = 0$ par definition

$d(S, S') = d(S', S)$, car $O(S - S') = O(S' - S)$

Enfin, si S et T sont respectivement de terme général a_n et b_n, S + T est de terme général $a_n + b_n$ dans K [[X]]

On a donc : $O(S + T) \geq \min (O(S), O(T))$

d'où en écrivant : $S - S' = S - S'' + S'' - S'$

$O(S - S') \geq \min (O(S - S''), O(S' - S''))$

ce qui est équivalent a :

$d(S, S') \leq \max (d(S,S''), d(S', S''))$

donc d est une distance ultramétrique sur K[[X]]

Nous avons ainsi muni K [[X]] d'une topologie ultramétrique, cette topologie n'est bien sûr pas unique en revanche il y a unicité de la hiérarchie indicée associée.

2.2. La hiérarchie indicée associée

2.2.1. Définition

Puisque d est ultramétrique, tout point d'une boule de K[[X]]peut être choisi comme centre de cette boule, le rayon d'une boule égale son diamètre et (K[[X]],d) est séparé.

Notons : B(S,r) la boule fermée de centre S et de rayon r de (K[[X]], d)

ray (B) = diam (B) = sup {d(S, S') / { S, S'} $\subset$ B }

le rayon ou le diamètre d'une boule B.

et $\mathfrak{B}$ l'ensemble des boules fermées de (K[[X]],d)

Proposition : $\mathfrak{B}$ est une hiérarchie indicée sur K[[X]] dont ray est l'indice de niveau.

Démonstration :

La démonstration de cette proposition, qui peut être faite dans un cas très général (équivalence entre hiérarchie indicée et distance ultramétrique, cf [1] p. 141, le passage au dénombrable n'offrant aucune difficulté), est ici explicitée dans le cas d'un ensemble de séries formelles muni d'une distance ultramétrique particulière.

Il faut montrer que :

1- $\forall$ {B, B'} $\subset \mathfrak{B}$: $B \cap B' = \emptyset$, ou $B \subset B'$, ou $B' \subset B$.

2- K [[X]] $\in \mathfrak{B}$, et $\forall S \in$K [[X]] : {S} $\in \mathfrak{B}$

3- ray est strictement croissant sur

1. Puisque d est ultramétrique, si $S \in B \cap B'$, S peut être choisi comme centre de B et de B'. Or, dans tout espace métrique, de deux boules concentriques, l'une est nécessairement incluse dans l'autre.

2. On a pour tout S : S = B (S, 0)

et K [[X]] est boule fermée de rayon 1, de centre quelconque puisque :

ray (K [[X]]) = sup {d (S, S') / { S,S'} ⊂ K [[X]]}

$$= \left(\frac{1}{2}\right)^{\inf 0(S-S')} = \left(\frac{1}{2}\right)^{0} = 1$$

3. On a :

$\forall$ {B,B'} ⊂ K [[X]] : (B ≠ B', B ⊂ B') $\Longrightarrow$ ray (B) < ray (B')

puisque, si S" ∈ B' - B, alors pour tout S ∈ B :

ray (B') ≥ d(S", S) > ray (B)

Remarquons enfin que les valeurs de l'indice de niveau de la hiérarchie $\mathfrak{B}$ sont :

$$1, \frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{2^n}, \ldots$$

2.2.2. Description

La hiérarchie ainsi définie peut être décrite de manière ascendante ou descendante.

De manière descendante, l'ensemble Sui K[[X]] des successeurs immédiats de K[[X]] est défini au niveau 1/2 comme l'ensemble des classes de séries ayant leur terme de rang zéro constant. Si les coefficients des séries sont à valeur dans un corps K, ces noeuds sont en nombre |K|

On a bien : $\forall$B ∈ Sui K [[X]] : diam (B) = sup { d(S,S') /{ S,S'} ⊂ B}

$$= 2^{-\inf \{0(S,S')/ \{ S,S'\} \subset B\}}$$

$$= 2^{-1} = 1/2$$

Et, $\forall$ {B,B'} ⊂ Sui K [[X]] : d(B,B') = inf { d(S,S')/S ∈ B, S' ∈ B'}

$$= 2^{-\sup \{0(S - S')/S \in B, S' \in B'\}}$$

$$= 2^{0} = 1$$

Car, pour tous S et S' appartenant respectivement à B et B', S et S' se distinguent dès leur terme de rang zéro, d'où :

0 (S - S') = 0 = sup { 0 (S - S') / S ∈ B, S' ∈ B' }

Chacun de ces noeuds se sépare ensuite en |K| noeuds de niveau 1/4. Chacun de ces noeuds est une classe de séries égales entre elles jusqu'à leur terme de rang un.

De manière générale, chaque noeud de niveau $1/2^n$ se sépare en |K| noeuds de niveau $1/2^{n+1}$, qui sont les classes de séries égales entre elles jusqu'à leur terme

de rang n.

Soit B tel que : diam (B) = $1/2^n$, alors :

$\forall \{B', B''\} \subset Sui(B)$:

diam (B') = sup $\{d(S, S') / \{S, S'\} \subset B'\}$

$= 2^{-\inf \{0(S - S') / \{S,S\} \subset B'\}}$

$= 1/2^{n+1}$

d(B', B") = inf $\{d(S', S") / S' \in B', S" \in B"\}$

$= 2^{-\sup \{0(S' - S") / S' \in B', S" \in B"\}}$

$= 1/2^n$

Car, si S' et S" appartiennent respectivement à B' et B", éléments de Sui(B), S' et S" sont dans B et sont égales jusqu'à leur terme de rang n-1.

De manière ascendante, il est clair que K[[X]] étant infini les premières agrégations ne peuvent être décrites. On pourrait dire que l'on y agrège les séries distinctes "égales à l'infini".

En général, le passage du niveau $1/2^n$ au niveau $1/2^{n-1}$ se fait en agrégeant les noeuds de niveau $1/2^n$ contenant les séries égales jusqu'à leur terme de rang n-2.

2.3. Exemple

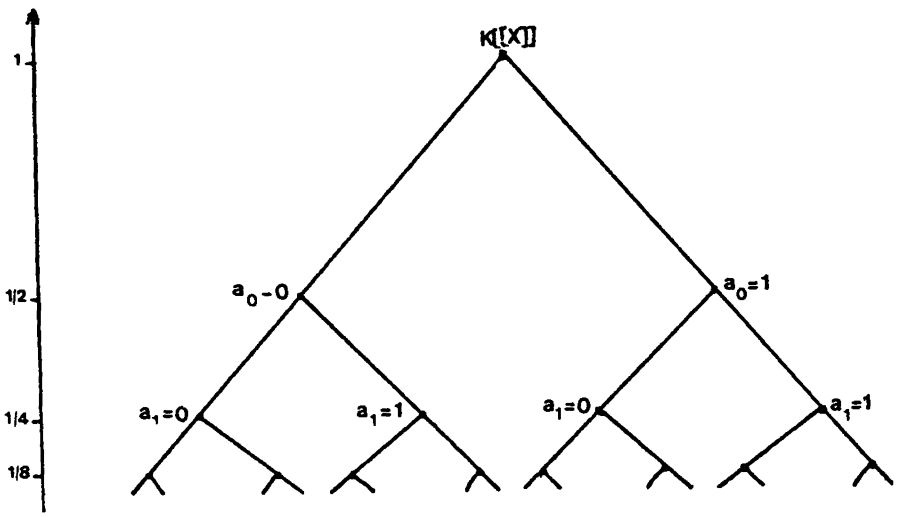
En choisissant le corps fini : K = { 0,1 } muni de

+	0	1
0	0	1
1	1	0

X	0	1
0	0	0
1	0	1

Si l'on note a_n le terme de rang n de la série S, l'arbre associé à la hiérarchie est un arbre binaire (tout noeud a deux successeurs immédiats) tel qu'à chaque niveau plusieurs noeuds, ici 2^n, sont définis.

Cet arbre peut être représenté de la manière suivante :



3 - HIERARCHIE D'ENTIERS

3.1. L'ultramétrique

Soit p un nombre premier ($p \in \mathbb{N}$), n un entier ($n \in \mathbb{Z}$), on note $0_p(n)$ l'exposant de p dans la décomposition en nombres premiers de n.

On a alors les propriétés suivantes :

$\forall \{n,m\} \subset \mathbb{Z}$

. $0_p(n) \in \mathbb{N}$ (1)

. $0_p(n) = 0_p(-n)$ (2)

. $0_p(nm) = 0_p(n) + 0_p(m)$ (3)

. enfin en écrivant : $n = p^{0_p(n)}\, n'$, où $0_p(n') = 0$

$m = p^{0_p(m)}\, m'$, où $0_p(m') = 0$

en supposant : $0_p(n) \leq 0_p(m)$,

alors :

$$n + m = p^{0_p(n)} \left(n' + m' p^{0_p(m) - 0_p(n)}\right)$$

d'où, pour $n + m \neq 0$: $0_p(n + m) \geq 0_p(n)$, d'après (3)

Soit : $0_p(n + m) \geq \inf (0_p(n), 0_p(m))$, si $n + m \neq 0$ (4)

On peut alors montrer la proposition suivante :

Proposition : d, application de $\mathbb{Z}^2$ dans $\mathbb{R}^+$ définie par :

$$\forall \{n,m\} \subset \mathbb{Z} \quad : \quad \begin{cases} d(n,n) = 0 \\ d(n,m) = \left(\frac{1}{2}\right)^{O_p(n-m)} , \; n \neq m \end{cases}$$

est une distance ultramétrique sur $\mathbb{Z}$, appelée distance p-adique.

Preuve :

On a $d(n,m) = d(m,n)$ d'après (2), et :

$\forall \{n,m,s\} \subset \mathbb{Z}$

$O_p(n-m) = O_p(n-s+s-m) \geq \inf (O_p(n-s), O_p(s-m))$,

si $n-s+s-m \neq 0$, i.e. $n \neq m$

d'où : $\max (-O_p(n-s), -O_p(s-m)) \geq -O_p(n-m)$

$\iff \max (2^{-O_p(n-s)}, 2^{-O_p(s-m)}) \geq d(n,m)$

car 2^x est croissante

$\iff \max (d(n,s), d(s,m)) \geq d(n,m)$

Cette dernière relation étant triviale pour $n = m$.

3.2. La hiérarchie associée

L'équivalence entre structure ultramétrique et hiérarchie indicée (pour une démonstration voir [1] p. 141, le passage au dénombrable n'offre aucune difficulté), nous permet d'écrire que $\mathfrak{B}$ l'ensemble des boules de $(\mathbb{Z},d)$ (ouvertes ou fermées puisque d est ultramétrique) est une hiérarchie sur $\mathbb{Z}$ dont ray est l'indice de niveau.

On rappelle que d étant ultramétrique, ray, rayon ou diamètre d'une boule B est défini par :

$$\text{ray } (B) = \sup \{ d(x,y) / \{x,y\} \subset B \}$$

et que tout point de B peut être choisi comme centre de B.

En remarquant que les valeurs de ray sont :

$$1, \frac{1}{2}, \ldots, \frac{1}{2^n}, \ldots$$

on constate que les classes de la hiérarchie $\mathfrak{B}$ de niveau $\frac{1}{2^n}$ sont les boules de rayon $\frac{1}{2^n}$.

$$\text{ray } (B) = \frac{1}{2^n} \iff \exists\, y \in \mathbb{Z} \; ; \; B = B\left(y, \frac{1}{2^n}\right)$$

$\Longleftrightarrow \exists y \in \mathbb{Z}$; $B = \{ x = y + p^{\alpha}q / \ \alpha \geq n,\ O_p(q) = 0 \}$

puisque : $d(x,y) \leq \frac{1}{2^n} \Longleftrightarrow O_p(x-y) \geq n$

$\Longleftrightarrow x-y = p^{\alpha}q$ où $\alpha \geq n,\ O_p(q) = 0$

p et q sont donc premiers entre eux puisque p est premier.

Montrons que l'on peut choisir comme centres des classes successeurs immédiats d'une classe B de niveau $\frac{1}{2^{n-1}}$ contenant un nombre x, les p nombres de B :

$x,\ x+p^{n-1},\ \ldots,\ x+kp^{n-1},\ \ldots,\ x+(p-1)\ p^{n-1}$

En notant [p] l'ensemble { 0,..., p },]p] l'ensemble {1,..., p} , et B(x) la classe de niveau $\frac{1}{2^n}$ contenant x, on a :

$\forall k \in [p-1]$, $\forall k' \in [p] - \{k\} : B(x+kp^{n-1}) \not\subset B(x+k'p^{n-1})$

Car sinon il existerait $q \in \mathbb{Z}$, et $\alpha \geq n$ tels que :

$x + kp^{n-1} = x + k'p^{n-1} + p^{\alpha}q$

$\Longleftrightarrow (k-k')\ p^{n-1} = p^{\alpha}q$

$\Longleftrightarrow k-k' = p^{\beta}q$ où $\beta \geq 1$

Or, on peut toujours supposer k-k' positif, comme $k-k' \leq p-1$, la dernière égalité est impossible.

Les p classes $\{ B(x+kp^{n-1}) \ / \ k \in [p-1] \}$ sont donc disjointes deux à deux puisqu'elles sont de centres distincts dans un espace ultramétrique.

Pour prouver que ce système forme Sui(B), il reste à montrer que :

$\forall y \in B$, $\exists k \in [p-1]$; $y \in B(x+kp^{n-1})$

or : $y \in B \Longrightarrow \exists \lambda \geq n-1,\ \exists q \in \mathbb{Z}$; $O_p(q) = 0,\ y = x+p^{\lambda}q$

recherchons donc $k \in [p-1]$, $\alpha \geq n$, $q' \in \mathbb{Z}$; $O_p(q') = 0$

tels que :

$y = x + kp^{n-1} + p^{\alpha}q'$

ceci est équivalent à :

$p^{\lambda}q = kp^{n-1} + p^{\alpha}q'$

$\Longleftrightarrow p^{n-1}\ (p^{\beta}q-k) = p^{\alpha}q'$ où $\beta = \lambda-n+1 \geq 1$

$\Longleftrightarrow p^{\beta}q = p^{\delta}q' + k$ où $\delta = \alpha-n+1 \geq 1$

La dernière équation représente la division euclidienne de $p^{\beta}q$ par p. Le quotient peut s'écrire $p^{\delta-1}q'$ avec $\delta-1$ positif et q' premier avec p puisque : $\beta \geq 1$. Le reste de la division est k : $k \leq p-1$.

On a donc montré que pour toute classe de niveau $\frac{1}{2^{n-1}}$, l'ensemble de ses sucesseurs immédiats est un ensemble de p classes de niveau $\frac{1}{2^n}$ dont on a trouvé un système de centres.

On en déduit que le nombre de classes de niveau $\frac{1}{2^n}$ est p^n.

Ces résultats peuvent s'interpréter de la manière suivante :

$x,y \in B$; $ray(B) = \frac{1}{2^n}$

$\iff x = y + p^{\alpha}q \qquad \alpha \geq n,\ 0_p(q) = 0$

$\iff x \equiv y\ (p^{\alpha})$

Les classes de la hiérarchie $\mathfrak{B}$ sont donc les classes de congruence modulo les puissances successives de p.

Ces classes s'emboitent en raison de la relation :

$n' > n \implies (x \equiv y(p^{n'}) \implies x \equiv y(p^n))$

La congruence modulo p^n étant une relation d'équivalence, elle induit une partition de $\mathbb{Z}$; l'ensemble de ces partitions pour $n \in \mathbb{N}$ constitue la hiérarchie $\mathfrak{B}$.

En particulier au niveau $\frac{1}{2^n}$, les restes possibles de la division euclidienne par p^n sont : $0, 1, \ldots, p^{n-1}$. Le nombre de classes de niveau $\frac{1}{2^n}$ est donc bien égal à p^n.

On en déduit que l'on peut choisir comme système de centres des classes de niveau $\frac{1}{2^n}$ toute suite de p^n nombres consécutifs et en particulier $]p^n]$, comme on va le voir dans l'exemple qui suit.

3.3. Exemple

Choisissons p = 2, on a $|Sui(\mathbb{Z})| = 2$, $\mathbb{Z} = B \cup B'$, où :

$B = \{n = 2k\}$ est l'ensemble des nombres pairs

$B' = \{n = 2k + 1\}$ est l'ensemble des nombres impairs.

On a bien :

$ray(B) = \sup \{d(n,m) / \{n,m\} \subset B\} = 2^{-\inf 0_2(n,m)}$

$0_2(n-m) = 0_2(2(k-k')) \geq 1$, et l'inf est atteint pour k-k' non multiple de 2,

e.g. $n = 2k = 6$, $m = 2k' = 16$ d'où :

$$\text{ray}(B) = \frac{1}{2}$$

De même : $\text{ray}(B') = \frac{1}{2}$, car $n-m = 2k + 1 - 2k' - 1 = 2(k-k')$, enfin :

$d(B,B') = \inf \{ d(n,n') / n \in B, n' \in B' \}$

$= 2^{-\sup 0_2(n-n')}$

or, pour tous n et n' : $n - n' = 2(k-k') - 1$ est impair, d'où :

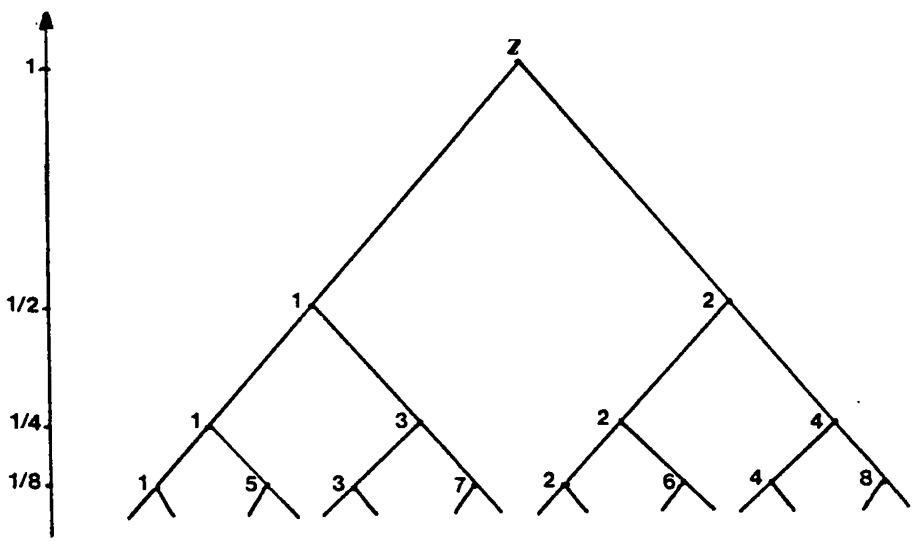
$d(B,B') = 2^{-0} = 1 = \text{ray}(\mathbb{Z})$

Au niveau $\frac{1}{4}$ les classes sont :

B(1), B(3) successeurs immédiats de B'

B(2), B(4) successeurs immmédiats de B

On obtient l'arbre suivant où chaque classe est repérée par un de ses centres :



BIBLIOGRAPHIE

[1] BENZECRI J.P. :
L'Analyse des Données. T1, La Taxinomie DUNOD, 1973

[2] CAILLEZ F. & PAGES J.P.
Introduction à l'Analyse des Données SMASH, 1976

[3] DIDAY E. et Coll. :
Optimisation en Classification Automatique (2 tomes) INRIA, 1979

[4] JAMBU M. :
Classification Automatique pour l'Analyse des Données
1 - Méthodes et Algorithmes

[5] avec LEBEAUX M.O. : 2- Logiciels DUNOD, 1978

[6] LERMAN I.C. :
Classification et analyse ordinale des données DUNOD, 1981.