PUBLICATIONS MATHÉMATIQUES ET INFORMATIQUES DE RENNES

BASAVANNEPA TALLUR

Classification et corrélation Application à l'analyse d'un questionnaire parents-élèves sur l'enseignement des mathématiques

Publications de l'Institut de recherche mathématiques de Rennes, 1993-1994, fascicule 3 « Fascicule de didactique des mathématiques », , exp. n° 5, p. 1-18

<http://www.numdam.org/item?id=PSMIR_1993-1994___3_A5_0>

Classification et corrélation
Application à l'analyse d'un questionnaire parents-élèves sur l'enseignement des mathématiques

Basavannepa TALLUR
I.R.I.S.A. - Campus de Beaulieu - 35042 RENNES Cédex

Résumé
La notion de similarité ou de distance est fondamentale dans l'analyse classificatoire des données. Cette notion de similarité ou de distance doit prendre en compte d'une facon fidèle la nature des données a analyser par une représentation mathématique appropriée. Les méthodes de classification hiérarchique basées sur la notion de distance et du critère d'"inertie expliquée" utilisent une représentation euclidienne des données avec une métrique adaptée.
La méthode A.V.L. ("Algorithe de la Vraisemblance des Liens") développée par I.C. Lerman (Lerman 1981) utilise plutôt une notion de similarité entre les éléments a classifier. La représentation ensembliste des données et une hypothèse de non association conduit a définir une mesure de similarité. Par cette démarche on retrouve la notion de corrélation comme une mesure de similarité entre variables numériques. Nous avons utilisé cette notion de corrélation pour construire un indice de similarité ainsi qu'un indice d'aggrégation pour la classification des lignes et colonnes d'un tableau des fréquences. Après avoir presenté la construction des indices, on exposera les résultats d'une enquête psycho-pédagogique menée auprès des élèves de 6ème d'une part, et des parents des élèves d'autre part afin d'analyser leur perception de l'enseignement des mathématiques. Ces résultats ont été publiés dans le rapport de l'IRISA numéro 177 (Juillet 1982) dont on reprend quelques passages dans cet article.
Le paragraphe 1 introduit la méthode AVL de I.C.Lerman et présente la démarche générale de construction de l'indice de similarité entre variables descriptives. Ensuite une généralisation de cet indice au cas des tableaux de contingence est présentée au paragraphe 2. Les paragraphes 3 et 4 décrivent respectivement l'algorithme directe basé sur la corrélation (A.B.C.) pour la classification des lignes et/ou colonnes d'un tableau de contingence et l'application aux données de l'enquête psycho-pédagogique. La thèse d'état de B.Tallur (1988) traite en détail les problèmes liés à l'analyse exploratoire de tableaux de contingence par la classification.

1 INTRODUCTION

1.1 Schéma de construction de l'indice de similarité

I.C. Lerman (Lerman, 1981, Chapitre 2) décrit le schéma général conduisant à l'expression de l'indice de proximité entre variables dont on peut distinguer les quatre étapes suivantes :

1°) Représentation des variables par des structures mathématiques adéquates.

2°) Choix d'un indice "brut" de proximité.

3°) Au couple de structures observées à comparer, on associe un couple de structures aléatoires et donc à l'indice "brut" on associe un indice brut aléatoire sous une hypothèse appropriée d'indépendance (ou d'absence de lien) pour construire l'indice centré et réduit.

4°) Indice de proximité probabiliste (ou la "vraisemblance" d'une similarité).

1. Représentation mathématique.

Une variable descriptive est représentée par une partie d'un ensemble ou par une pondération sur ce dernier. Selon que l'ensemble de représentation soit l'ensemble E des individus ou l'ensemble produit ExE, on distingue deux grandes catégories de variables.

On peut classer dans la première catégorie l'attribut descriptif a qui peut être représenté par la partie E des individus qui le possèdent, et la variable numérique X qui définit une pondération sur E en associant à chaque individu i le nombre X(i), valeur de la variable X pour l'individu i.

Toute variable qualitative discrète qui définit une relation binaire sur E est susceptible d'une représentation dans ExE. Ainsi on peut classer dans la deuxième catégorie :

- la variable "rang" qui définit un ordre total 0 sur ExE que l'on représente par son graphe.

- la variable qualitative ordinale qui définit un préordre total w sur E que nous représentons par la partie R(w) de ExE définie par $R(w)=\Sigma\{ E_i \times E_j / i<j\}$ (somme ensembliste) où E_i est la $i^{ème}$ classe du préordre formée des individus possédant la $i^{ème}$ modalité de la variable.

- la variable qualitative nominale qui définit une partition π sur E que nous représentons dans l'ensemble $F = P_2(E)$ des parties à deux éléments de E par :
$R(\pi)=\Sigma P_2(E_i)$ (somme ensembliste)
où $P_2(E_i)$ est l'ensemble des paires réunies dans la $i^{ème}$ classe.

- d'une manière générale, la variable pondération sur ExE peut être représentée par une matrice carrée

$$\{\mu_{xy} /(x, y)\in ExE\}$$

Soit (a,b) un couple de variables définissant le même type de structure sur E et (α, β) le couple de structures définies sur E par (a, b). A α (resp. β) nous associons l'ensemble A (resp. B) des structures sur E de même type et ayant les mêmes caractéristiques cardinales que α (resp. β).

Le couple (a, b) sera représenté par un couple de parties $(R(\alpha), R(\beta))$ de E ou de ExE selon qu'il s'agisse de variables de la 1ère ou de la 2ème catégorie.

2ème étape. Définition d'un indice "brut" de proximité.

On introduit l'indice brut de similarité entre a et b:

$$s=card(R(\alpha) \cap R(\beta)) \qquad .....(1\text{-}1)$$

3ème étape. Hypothèse d'indépendance.

On associe les deux variables aléatoires duales à l'indice brut de proximité s, sous l'hypothèse d'absence de lien (ou d'indépendance) :

$$S_\alpha = \text{card}(R(\alpha) \cap R(\beta')) \qquad \text{.....(1-2)}$$

$$S_\beta = \text{card}(R(\alpha') \cap R(\beta))$$

où α' (resp. β') est un élément aléatoire dans l'ensemble A (resp. B) muni d'une probabilité uniformément répartie. Les v.a. S_α et S_β ont la même distribution et on montre (Lerman 1981, chapitre 2) sous des conditions assez générales que cette distribution commune est asymptotiquement normal.

indice centré et réduit :

En désignant par E(S) et $\sigma(S)$ la moyenne et la variance de S_α (ou S_β) on définit l'indice centré et réduit:

$$Q(\alpha, \beta) = \{s - E(S)\} / \sigma(S) \qquad \text{.....(1-3)}$$

La réduction globale des similarités:

Pour n "assez grand" l'indice $Q(\alpha, \beta)$ approché par $\sqrt{\rho(\alpha, \beta)}$, où $\rho(\alpha, \beta)$ désigne le coéfficient de corrélation entre les variables associées aux structures α et β, prend des valeurs élevées, en valeur absolue, et l'indice $P(\alpha, \beta)$ défini à l'étape 4 ci-dessous ne permet pas une discrimination suffisante.

En réalité, Lerman (1984) montre - pour tenir compte du contexte - que l'h. a. l. doit directement associer à l'ensemble A des variables observées un ensemble A^* de v. a. indépendantes, et non pas seulement à deux d'entre elles sans référence aux autres, et rapporter l'association entre deux variables à l'ensemble des associations mutuelles. Cela peut se faire notamment au moyen d'une réduction globale des similarités:

$$Q(\alpha, \beta) \rightarrow \{Q(\alpha, \beta) - \text{Moy}(Q)\}/\sqrt{\text{Var}(Q)}$$

où Moy(Q) et Var(Q) désignent, respectivement, la moyenne et la variance des valeurs de l'indice $Q(\alpha, \beta)$, pour toutes les paires de structures observées.

Sous certaines conditions on montre (cf. Le Calvé, 1976; Lerman, 1984) la normalité asymptotique de l'indice $Q(\alpha, \beta)$ réduit globalement comme ci-dessus.

4. Indice de similarité probabiliste.

L'indice de similarité définitif utilisé dans l'A.V.L. s'écrit :

$$P(\alpha, \beta) = \Pr\{S<s/N\} \qquad \text{.....(1-4)}$$

où S est l'une des deux v.a. duales de même loi S_α et S_β et N désigne l'hypothèse d'absence de lien (ou d'indépendance) exprimée ci-dessus.

Le degré de ressemblance exprimé par l'indice $P(\alpha, \beta)$, entre les variables a et b est d'autant plus grand que la valeur de s est invraisemblablement grande, relativement à l'hypothèse d'indépendance N.

En utilisant le caractère asymptotiquement normal de la variable S, on a l'approximation suivante :

$$P(\alpha, \beta) \approx \phi(Q(\alpha, \beta)) \qquad \text{.....(1-5)}$$

où ϕ est la fonction de répartition de la loi normale N(0, 1)

étape 1 : représentation.

Les variables numériques, comme les attributs descriptifs appartiennent à la première catégorie de variables à savoir celles qui peuvent être représentées au niveau de l'ensemble E. La variable numérique est représentée par une mesure sur E alors qu'un attribut est représenté par une partie de E.

Une variable numérique X est définie par sa distribution $(X_1, X_2, ..., X_n)$ sur l'ensemble d'objets E, où X_i est la mesure de la variable X sur l'objet i.

étape 2 : indice brut.

La base de construction de l'indice de similarité entre deux variables numériques X et Y dont les distributions sur E sont respectivement :

$(X_1, X_2, ..., X_n)$ et $(Y_1, Y_2, ..., Y_n)$

sera

$$s_{XY} = \Sigma\{X_i Y_i / 1 \le i \le n\} \qquad .....(1-6)$$

dont on examinera la distribution sous l'hypothèse d'indépendance.

étape 3 : Sous l'hypothèse d'indépendance entre X et Y, les distributions de ces variables sont fixées mais la position relative de l'une par rapport à l'autre est inconnue.

Les deux variables aléatoires duales associées à s_{xy} sous l'hypothèse d'indépendance sont :

$$S_X = \Sigma\{X_i Y_{\sigma(i)} / 1 \le i \le n\} \quad \text{et} \quad S_Y = \Sigma\{X_{\sigma(i)} Y_i / 1 \le i \le n\} \qquad .....(1-7)$$

où $(\sigma(1), ..., \sigma(n))$ est un élément aléatoire pris dans l'ensemble de toutes les permutations de (1, 2, ..., n).

La distribution de S_X est la même que celle de S_Y.

On trouve l'espérance mathématique et la variance de la variable S_X :

$$E[S(X)] = n\mu_X \cdot \mu_Y \quad \text{et} \quad V[S(X)] = (n^2/(n-1))\sigma^2_X \sigma^2_Y \qquad .....(1-8)$$

où μ_X (resp μ_Y) et σ_X^2 (resp. σ_Y^2) sont la moyenne et la variance de la variable X (resp. Y).

L'indice centré et réduit de proximité entre X et Y s'écrit :

$$Q(X, Y) = (s_{XY} - n\mu_X\mu_Y)/\sqrt{[(n^2/(n-1))\sigma_X^2\sigma^2_Y]}$$
$$= \sqrt{(n-1)}\rho(X, Y) \qquad .....(1-9)$$

où $\rho(X, Y)$ est le coefficient de corrélation linéaire entre X et Y.

Le caractère asymptotiquement normal de la variable Q(X, Y) est démontré sous des conditions assez générales dans le théorème de A.Wald et J. Wolfowitz (1944).

étape 4 : Indice de proximité entre X et Y sera approché par:

$P(X, Y) \simeq \phi (Q(X, Y))$

REMARQUE. Nous ne nous étendrons pas ici, sur le problème de construction d'un indice de similarité entre objets ou individus qui a été abordé dans Lerman et Peter (1985) et Lerman (1987).

2 CAS D'UN TABLEAU DE CONTINGENCE

2.1. REPRESENTATION GEOMETRIQUE D'UN TABLEAU DE CONTINGENCE.

Considérons le tableau de contingence

$$K_{IJ} = \{k_{ij}/ (i, j) \in IxJ\} \qquad \text{----(2-1)}$$

où les k_{ij} sont des cardinaux des classes $C_i \cap C'_j$ du croisement de deux partitions $\{C_1, C_2, ..., C_n\}$ et $\{C'_1, C'_2, ..., C'_m\}$ respectivement en n = card(I) et m = card(J) classes.

Pour représenter géométriquement ce tableau des données, on fait jouer à l'un des deux ensembles I ou J, le rôle de l'ensemble des objets et à l'autre le rôle de l'ensemble des variables.

Supposons, pour fixer les idées, que I joue le rôle de l'ensemble des individus et J celui des variables. Nous utiliserons la même représentation de I à travers J, que celle qui est adoptée dans le cadre de l'analyse des correspondances.

Nous noterons, pour tout $(i, j) \in IxJ$,

$$k_{i.} = \Sigma \{k_{ij}/ j \in J\} \ , \ k_{.j} = \Sigma \{k_{ij}/ i \in I\}$$
$$k_{..} = \Sigma \{k_{ij}/ (i, j) \in IxJ\}$$
$$f_{ij} = k_{ij}/ k_{..}, \ f_{i.} = k_{i.}/k_{..}, \ f_{.j} = k_{.j}/k_{..} \qquad \text{.....(2-2)}$$
$$f^i_J = (f^i_j/ j \in J) \text{ où } f^i_j = f_{ij}/ f_{i.}$$

f^i_J : "proïil de i à travers J"

On associe à l'ensemble I d'objets le nuage N(I) dans R^m muni de la métrique diagonale $\{(1/f_{.j})/ j \in J\}$:

$$N(I) = \{(f^i_J, f_{i.}) / i \in I\} \qquad \text{.....(2-3)}$$

où $f_{i.}$ est le poids affecté au point f^i_J, pour tout $i \in I$.

J.L. Philoche a proposé une représentation plus symétrique (journées de statistique, Paris Juin 1979) de l'ensemble I par le nuage M(I) dans R^m :

$$M(I) = \{(l_{iJ}, f_{i.}) / i \in I\} \qquad \text{.....(2-4)}$$

où $l_{iJ} = (r_{ij} / j \in J)$, $r_{ij} = f^i_j / f_{i.} f_{.j}$;

l'objet i de I est représenté par le point l_{iJ} de R^m.

L'analyse du nuage M(I) dans R^m, muni de la métrique diagonale $(f_{.j} / j \in J)$ est équivalente à l'analyse du nuage N(I) muni de la métrique diagonale $\{(1/f_{.j}) / j \in J\}$.

Nous verrons que, quelle que soit la représentation géométrique retenue - la représentation classique ou 'symétrique' - nous aboutissons aux mêmes indices de similarités.

Considérons pour fixer les idées, le problème de la classification de l'ensemble J. On peut noter une *différence essentielle* entre notre méthode et la méthode ascendante de classification hiérarchique de J.P. Benzecri et M. Jambu pour ce type de données. Cette dernière est conçu à partir de la représentation de l'ensemble de points-objets J dans R^n et utilise les concepts de distance au carré et d'inertie expliquée au sens de la métrique du χ^2, tandis que notre méthode se conçoit au niveau de $(R^m)^*$ et utilise le concept de la corrélation, conformément à la représentation définie plus haut.

2.2. INDICE DE PROXIMITE ENTRE COLONNES OU LIGNES D'UN TABLEAU DE CONTINGENCE.

Considérons pour fixer les idées le problème de la classification de l'ensemble J de colonnes. J jouera donc le rôle de variables descriptives et nous allons considérer la représentation de I à travers J définie par le nuage N(I). On va associer à la colonne j, $1 \leq j \leq m$, une variable numérique X_j dont la mesure sur l'individu i, $1 \leq i \leq n$, vaut f^i_j. La distribution de la variable X_j $(1 \leq j \leq m)$ est donnée par la suite des valeurs prises :

$$\{f^1_j, f^2_j, \ldots, f^i_j, \ldots, f^m_j\}$$

avec les fréquences relatives :

$$\{f_{1.}, f_{2.}, \ldots\ldots, f_{n.}\}$$

L'indice de proximité entre les éléments j et h de J n'est autre que le coefficient de corrélation $\rho(j,h)$ entre les variables numériques X_j et X_h dont nous allons préciser l'expression.

En effet, la moyenne et la variance de la variable X_j sont respectivement :

$$\mu_j = \Sigma [f_{i.} f^i_j / i \in I] = f_{.j}$$

$$\sigma^2_j = \Sigma [f_{i.} (f^i_j - f_{.j})^2 / i \in I] = \Sigma [(f^2_{ij} / f_{i.}) / i \in I] - f^2_{.j} \qquad \ldots\ldots(2\text{-}5)$$

D'autre part, la covariance entre X_j et X_h est donnée par :

$$\sigma_{jh} = \Sigma \{f_{i.} (f^i_j - f_{.j})(f^i_h - f_{.h}) / i \in I\}$$

$$= \Sigma \{(f_{ij} f_{ih} / f_{i.}) / i \in I\} - f_{.j} f_{.h} \qquad \ldots\ldots(2\text{-}6)$$

On obtient finalement :

$$\rho(j, h) = \sigma_{jh} / (\sigma_j \sigma_h)$$

$$\rho(j, h) = \frac{\Sigma\{(f_{ij}f_{ih}/f_{i.}) / i \in I\} - f_{.j}f_{.h}}{[\{\Sigma[(f^2_{ij}/f_{i.}) / i \in I] - f^2_{.j}\}\{\Sigma[(f^2_{ih}/f_{i.}) / i \in I] - f^2_{.h}\}]^{1/2}} \quad \ldots\ldots(2\text{-}7)$$

Cet indice dépend de la représentation euclidienne de I à travers J. Avant d'adopter définitivement l'indice (2-7) ci-dessus, nous avons expérimenté d'autres indices tels que le coefficient de corrélation entre les colonnes et la covariance entre les variables associées qui n'ont pas donné des résultats satisfaisants. On vérifie aisément les deux propriétés suivantes de cet indice :

Propriété 1 : L'indice de proximité $\rho(j, h)$ défini ci-dessus est invariant par rapport à la métrique dont peut être muni R^m pour l'évaluation des distances entre éléments de I.

Propriété 2 : La représentation euclidienne de I à travers J au moyen du nuage M(I) conduit au même indice $\rho(j, h)$ de proximité entre les éléments j et h de J, que (2-7) ci-dessus.

Pour obtenir l'indice de proximité entre éléments de I on considère, de façon analogue la représentation euclidienne de J à travers I au moyen du nuage N(J) ou M(J).

A partir de (2-7) on peut directement écrire l'expression de l'indice de proximité entre les lignes i et i' de I en intervertissant les rôles de I et de J :

$$\rho(i, i') = \frac{\Sigma\{(f_{ij}f_{i'j}/f_{.j}) / j \in J\} - f_{i.}f_{i'.}}{[\{\Sigma[(f^2_{ij}/f_{.j}) / j \in J] - f^2_{i.}\}\{\Sigma[(f^2_{i'j}/f_{.j}) / j \in J] - f^2_{i'.}\}]^{1/2}} \quad ; 1 \le i, i' \le n.$$

$\ldots\ldots(2\text{-}8)$

Les valeurs de l'indice (2-7) (resp. (2-8)) entre colonnes (resp. lignes) joueront le même rôle que celles de l'indice "centré et réduit" Q(a, b) défini dans le paragraphe 1.1 du chapitre 1.

Au lieu de considérer la représentation de l'ensemble I à travers J par le nuage N(I) dans R^m, si on considère la représentation de J à travers I par le nuage N(J) dans R^n muni de la métrique du χ^2, on peut associer à la colonne j un "individu" défini par son profil f^j_I, $(1 \le j \le m)$. Lerman et Peter (1985) ont montré que l'indice de similarité entre les colonnes j et h obtenu en tant que le coéfficient de corrélation entre les variables X_j et X_h définies ci-dessus est identique à celui défini par le cosinus de l'angle entre les vecteurs f^j_I et f^h_I par rapport au centre de gravité du nuage $g_I = \{f_i / i \in I\}$.

Propriété 3 : L'indice de proximité (2-7) (resp. (2-8)) entre colonnes (resp. lignes) vérifie la propriété d'équivalence distributionnelle :

En remplaçant les colonnes (resp. lignes) de "profils" identiques par leur somme, les valeurs de l'indice de proximité entre lignes (resp. colonnes) ne sont pas modifiées.

3 ALGORITHME BASÉ SUR LA CORRÉLATION: A.B.C.

Nous avions suggéré (cf Lerman-Tallur, 1980) une façon d'associer un arbre de classification directement à l'indice de proximité que nous venons d'exposer dans les précédents paragraphes de ce chapitre. La structure des données qui nous intéresse ici est celle qu'on peut assimiler à un tableau de contingence ou une juxtaposition, selon une ou deux directions, des tableaux de contingence.

La méthode que nous proposons ici est une méthode naturelle de construction de l'arbre de classification hiérarchique. Elle est comparable à des méthodes basées sur les carrés des distances entre les centres de gravité des classes, mais au lieu de minimiser un critère basé sur la distance, on maximise un critère basé sur la corrélation.

Soit le tableau de contingence (2-1) :

$K_{IJ} = \{k_{ij} / (i, j) \in I \times J\}$
avec card (I) = n et card (J) = m

Supposons, pour fixer les idées, que l'ensemble à classifier soit celui des colonnes. Nous avons vu que, pour la définition de l'indice de proximité entre les éléments de J, nous avons associé à chaque élément $j \in J$ une variable numérique X_j en passant par une représentation géométrique du tableau (c'est-à-dire, en nous plaçant dans le nuage N(I) des profils des lignes). Le coefficient de corrélation $\rho(j, j')$ entre les variables X_j et $X_{j'}$ associées respectivement aux colonnes j et j' de J est, à un coefficient multiplicatif près, l'indice centré et réduit Q(j, j'). Alors que l'A.V.L. se réfère à une échelle de probabilité fournie par la fonction de répartition $\phi(Q(j, j'))$ de la loi N(0,1), la méthode proposée ici se réfère directement à l'indice de corrélation.

3.1 Algorithme basé sur la corrélation (A.B.C.)

Le principe de l'A.B.C. consiste à agréger à chaque pas les deux éléments ou deux classes d'éléments les plus proches au sens de l'indice d'agrégation défini de la façon suivante.

Définition. Soient C et D deux classes formées à un certain niveau (pas) de l'algorithme où C est formée par la réunion de p colonnes $C_1, C_2, ..., C_p$ et D est formée par la réunion de q colonnes $D_1, D_2, ..., D_q$. Notons X_{Ci} $(1 \le i \le p)$ (respectivement X_{Dj} $(1 \le j \le q)$) la variable numérique associée à la colonne C_i (resp. D_j), $X_C = \Sigma\{X_{Ci} / 1 \le i \le p\}$ et $X_D = \Sigma\{X_{Dj} / 1 \le j \le q\}$.
Alors, l'indice d'agrégation S_{CD} entre les classes C et D sera

$$S_{CD} = \rho(X_C, X_D)/\sqrt{(pq)} \qquad \text{---(3-1)}$$

où ρ désigne le coefficient de corrélation.

On peut résumer l'algorithme de la façon suivante :

Pas 1. Au départ chaque colonne constitue une classe. On réunit les deux colonnes les plus proches au sens de l'indice d'agrégation (ξ_{-1}) (c-à-d. celles qui réalise le maximum de cet indice) avec p = q = 1 :

$S_{jh} = \rho(X_j, X_h)$ pour tout (j, h)$\in$ JxJ, 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 formée des colonnes C_i ($1 \leq i \leq p$) sera la variable X_C somme des variables X_{Ci} ($1 \leq i \leq p$). Remarquons que cela revient à remplacer les colonnes réunies C_i ($1 \leq i \leq p$) par leur somme et à associer à cette colonne somme une variable numérique conformément au paragraphe 2.2.

Pas 3. On réunit les classes les plus proches au sens de l'indice d'agrégation (ξ_{-1}) (c-à-d. pour lesquelles cet indice est maximum) et on retourne au pas 2 jusqu'à ce que toutes les classes soient réunies.

Définition. La hiérarchie de partitions sera dite "sans inversion" si la valeur maximale de l'indice d'agrégation fondé sur les similarités à un pas s quelconque est inférieure à la valeur maximale de cet indice à tous les pas inférieurs à s. Autrement dit, l'algorithme est sans inversions si la suite des valeurs maximales de l'indice d'agrégation aux niveaux successifs est strictement décroissante, lorsque l'indice d'agrégation est basé sur les similarités.

4. APPLICATION AUX TABLEAUX BINAIRES

Le domaine très vaste d'application de la méthode de l'Analyse des correspondances, qui, on le sait, a été fondamentalement développé pour le traitement des tableaux de contingence montre que les tableaux binaires (présence-absence codées 1-0) peuvent être assimilés à ces derniers pour analyser les nuages des profils. Inspiré par ceci, nous avons essayé notre méthode de classification sur des données de type 0-1 provenant d'une enquête psycho-pédagogique en mathématiques. Les résultats de cette analyse, forts intéressants, sont confrontés à ceux de l'Analyse des correspondances d'une part et à ceux de la classification par l'A.V.L. de l'autre. S'agissant d'éprouver l'efficacité et l'opportunité d'un nouvel outil, il est bien nécessaire de comparer sa performance avec quelques outils standards bien connus. Nous exposerons les résultats de cette application de manière détaillée dans les paragraphes suivants.

4 APPLICATION

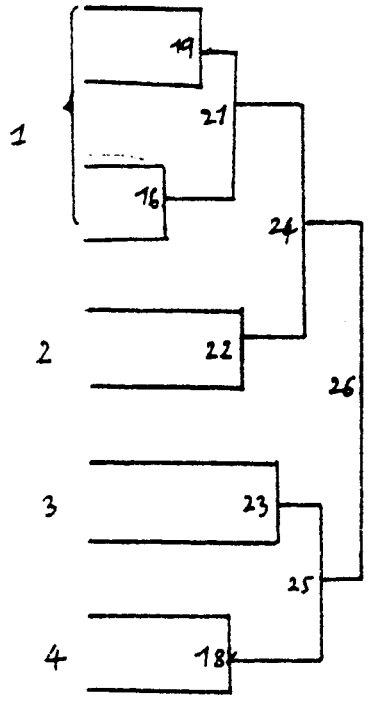
4.1 DIDACTIQUE DES MATHÉMATIQUES

On présentera ci-dessous une application de la mthode de classification exposée dans le paragraphe précedent á une enquête psychopédagogique. Le but de cette enqête était d'analyser de quelle manière l'enseignement des mathématiques en classe de $6^{ème}$est perçu par les enfants eux mêmes, les enseignants et les parents.
Une première enquête menée par Jacques Degouys (Université de Haute Bretagne), Régis Gras (IRMAR) et Marcel Postic (Université de Haute Bretagne) en 1978 avait permis l'établissement d'une échelle d'attitude.

L'étude actuelle porte sur un échantillon de 163 élèves. Le questionnaire se compose de 3 parties : un ensemble de questions destinées à l'enfant, un ensemble de questions aux parents et une troisième partie destirée à l'enseignant. Chaque question a deux réponses possibles : oui et non. Nous présenterons ci-dessous l'analyse par classification hiérarchique, selon l'algorithme basé sur la corrélation portant sur deux parties distinctes du questionnaire : réponses des enfants et réponses des parents.

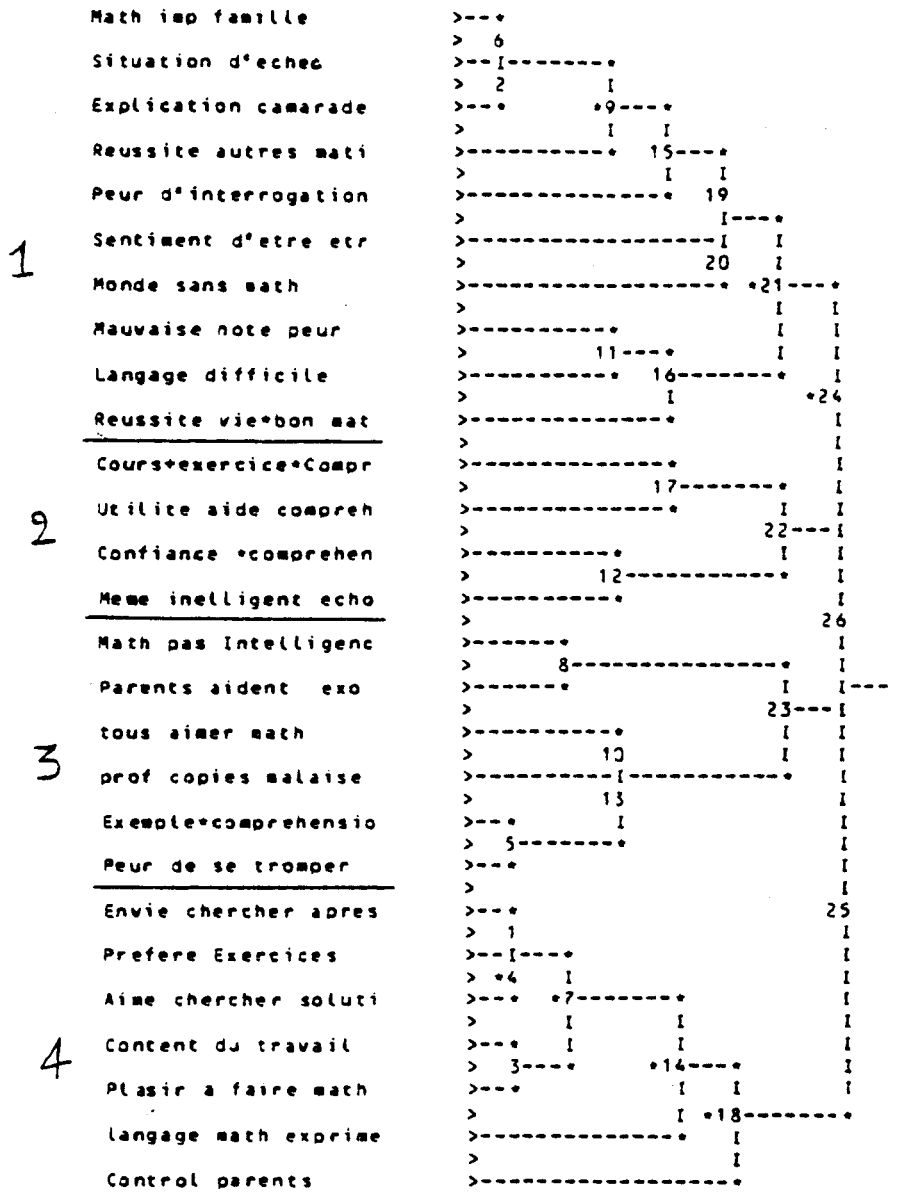
4.1.1 RÉPONSES DES ENFANTS :

Il a été retenu un ensemble de 27 questions auxquelles les réponses sont données par oui ou par non.(voir la liste des questions en annexe). La figure 1 représente l'arbre, condensé à ses noeuds significatifs, de la classification hiérarchique des questions (ou plutôt, il s'agit des attributs définis par des réponses positives aux questions).



Au niveau 22 la statistique globale atteint le maximum et on distingue quatre grandes classes. (voir schéma ci-contre).
Classe 1-Cette classe regroupe l'ensemble des opinions négatives à l'égard des mathématiques. Cet ensemble des opinions négatives reste marqué par celles de la famille du fait de la jeunesse des enfants. Notons que le trait dominant est l'extériorité des mathématiques par rapport à l'enfant. C'est en situation d'échec que l'enfant se prononce ici. Les mathématiques

Figure 1. Classification des réponses des enfants par la méthode A.B.C.



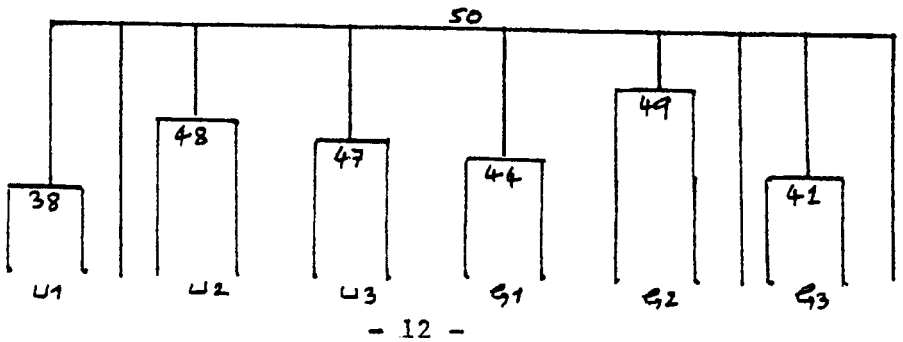
constituent un monde à part, extérieur à celui de l'enfant. C'est ainsi que d'autres que lui (famille,professeur, camarade) peuvent tenir entre ces mondes des rôles médiateurs. L'enfant ici a des difficultés. La réussite sociale par les mathématiques, l'ésotérisme de leur langage représentent les grands traits du mythe mathématique.

Classe 2 - Cette fois, dans l'ensemble des opinions plutôt tournées vers une représentation positive des mathématiques, la forme générale des opinions est "subordination". Si les maths ou leur enseignement ont telle propriété, alors tout est versé en faveur de la compréhension. Celle-ci, en effet, est successivement subordonnée à la qualité du cours et des exercices, à la confiance que l'on ressent, à l'utilité des mathématiques elles-mêmes.
Classe 3 - Cette fois l'élève consent à s'engager sur cette terre où les maths sont apparemment terrifiantes : "tout le monde pourrait aimer les maths" ; il ne s'exclut pas a priori. Pourtant, dans cet engagement, le sentiment de crainte, le coût de l'engagement accompagnent le moment de distribution des copies et celui de prise de parole.
Classe 4- Cette fois aucune restriction ne vient troubler les opinions très favorables aux mathématiques. L'expression de plaisir apparaît 3 fois dans ces opinïons ; de plus la forme personnelle ("je", "mes"...) employée 6 fois sur 7 montre à quel point, l'enfant s'implique dans la réussite, même si ses parents contrôlent ses résultats.

4.1.2 RÉPONSES DES PARENTS

L'ensemble des 58 réponses des parents a été classifié par l'algorithme présenté dans cet article.

Les six grandes classes se situent l'une par rapport à l'autre de la manière suivante :



INTERPRÉTATION (voir figure 2 : l'arbre de classification).

Contrairement à l'enquête faite auprès des enfants, ce n'est pas la réussite en maths qui établit une partition dans l'ensemble des opinions des parents. Elle n'est pas ressentie par eux, comme discriminatrice dans le champ de l'attitude. Par contre, c'est plutôt l'utilité des disciplines, donc les racines de l'avenir et de la réussite sociale qui est en jeu.

Classe U_1 Les mathématiques et les Sciences naturelles constituent les disciplines les plus utiles pour l'avenir. Les résultats étant jugés bons, la meilleure voie de réalisation de ce projet passe par la filière C. Les parents projettent peut-être leur foi dans l'avenir social, sur l'intérêt que l'enfant porterait aux mathématiques, les langues étant jugées comme une gêne dans le projet. Notons que cette classe se renforce à l'ultime niveau, du dernier élément du tryptique scientifique : la physique, dont on sait qu'elle constitue un pilier de la filière C.

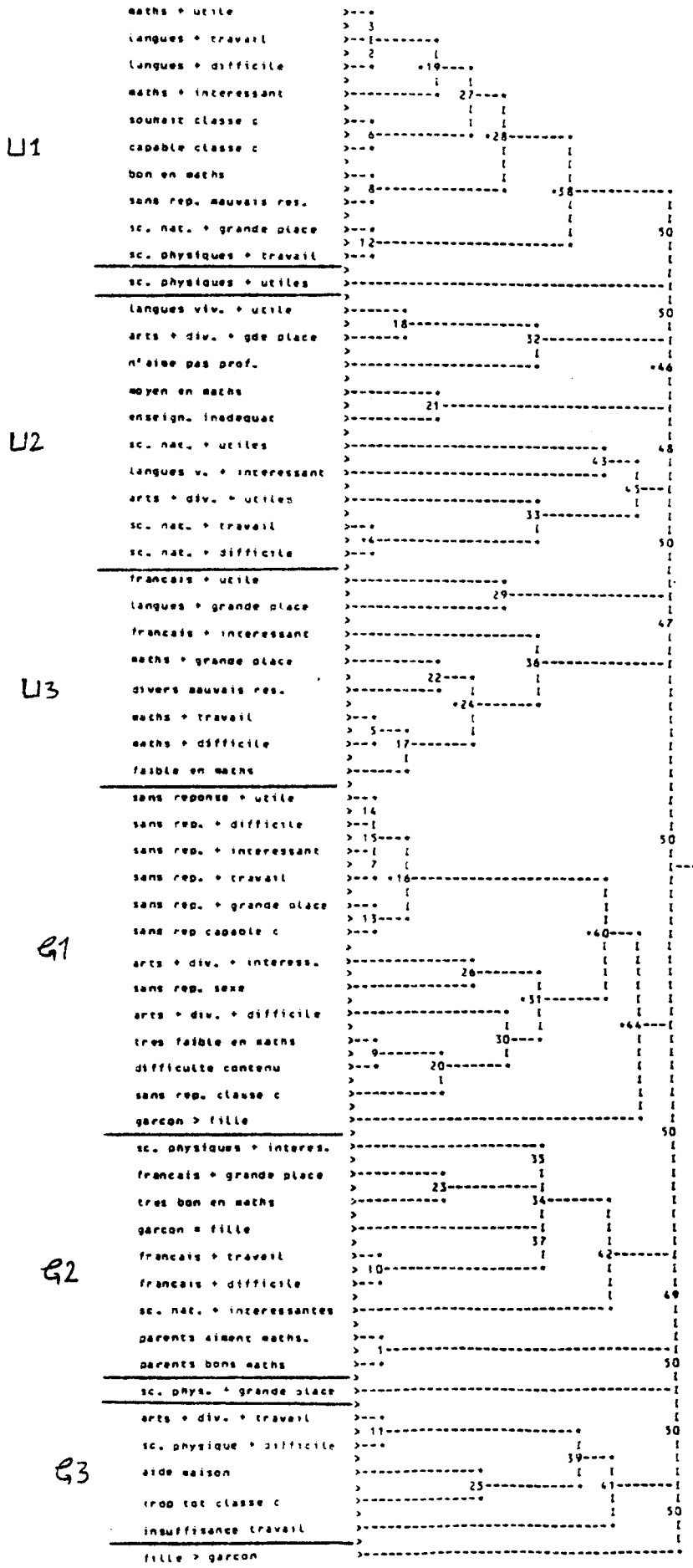
Classe U_2 Les résultats en mathématiques sont, cette fois, moyens. Mais refusant de responsabiliser l'enfant, et,peut-être,eux-mêmes pour des raisons héréditaires, le maître et son enseignant sont culpabilisés. La voie "langues vivantes" est, dans ce cas, déclarée la plus utile pour l'avenir.

Classe U_3 Cette fois la planche de salut social est le français ; mais cet intérêt semble l'alibi contre le pouvoir des mathématiques qui se voient accusées des maux pernicieux : excès de travail, trop grande place, difficulté etc...échec, bien évidemment.

Classe G_1 Cette classe, constituée surtout d'absence de réponses, est l'image d'une population détachée pour des raisons culturelles peut-être, de la préoccupation d'une liaison école-société. Comment sans culpabilisation excessive, pourrait-on subordonner la réussite sociale à la réussite scolaire lorsque, en particulier, l'enfant est très faible en mathématiques ?

Classe G_2 Ici par contre, la réussite est excellente, mais comme naturelle, prise comme héréditaire, les parents aimaient les mathématiques et réussissaient dans cette discipline. A quoi bon dans ce cas où la sérénité sociale a sa place, où l'on peut aimer les mathématiques et les sciences de façon gratuite, discriminer les filles des garçons : l'enseignement des maths a la même importance pour l'un que pour l'autre. Par contre le français semble occuper trop de place dans l'enseignement.

Figure 2. Classification des réponses des parents par la méthode A.B.C.



Classe G_3 Quelques opinions peu structurées se retrouvent ici. C'est une sous-classe des incertitudes où l'enfant recevant une aide à la maison ne manifeste pas ses aptitudes, avec une grande fiabilité.

RECONNAISSANCE

Tous mes remerciements à Régis GRAS de l'aide indispensable qu'il m'a apportée dans l'interprétation des résultats de l'enquête psychopédagogique. C'est lui qui m'a rendu les données accessibles, et les conclusions sont de lui.

5 Annexe

5.1 Questionnaire enfants

CODE ATTRIBUT	SA SIGNIFICATION
Math imp famille	Si je travaille en maths c'est à cause de l'importance qu'on lui donne dans ma famille
Math pas intélligence	Être doué en maths n'est pas une affaire d'intélligence
Cours+exercice=comprehension	Quand on comprend les cours et qu'on revoit les exercices on est sûr de comprendre les maths
Situation d'échec	En maths on est toujours en situation d'échec
Mauvaise note peur	Lorsque j'ai une mauvaise note en maths je n'ose pas le dire à la maison
Tous aimer math	Tout le monde pourrait aimer les maths
Monde sans math	On peut imaginer un monde sans mathématiques
Langage difficile	Ce qui est le plus difficile en mathématiques, c'est le langage
Prof copies malaise	On n'est jamais à l'aise quand le prof rend les copies de maths
Envie chercher après cours	Après les cours de maths, j'ai envie de continuer à chercher
Explication camarade	Je préfère demander une explication à un camarade plutôt qu'à un prof
Exemple+compréhension	On est sûr de comprendre quand le prof nous donne un exemple
Réussite vie + bon en maths	Pour réussir dans la vie il faut être bon en maths
Contrle parents	Mes parents contrôlent souvents mes résultats en mathématiques
Réussite autres matières	On n'a pas besoin de maths si on réussit dans d'autres matières
Aime chercher solution	J'aime chercher la solution des problèmes des maths
Peur d'intérrogation	Lorsqu je sais que je vais être intérrogé en maths je n'ai pas envie d'aller en cours
Confiance + compréhension	Pour comprendre les maths, il faut se sentir en confiance
Peur de se tromper	Quand on répond en math, on a peur de se tromper
Préfère exercices	Ce que je préfère en maths, c'est faire des exercices
Parents aident en exo	Mes parents m'aident parfois à faire mes exercices de maths
Content du travail	Je suis content au travail que je fais en mathématiques
Même intélligent échoue	Même s'ils sont intélligents, certains élèves peuvent échouer en mathématiques
Langage math exprimer claire	Avec le langage mathématique on peut s'exprimer clairement
Sentiment d'être étranger	En cours de mathématiques, on a le sentiment d'être étranger
Plaisir à faire math	J'éprouve du plaisir à faire des mathématiques
Utilité aide compréhension	On ne peut comprendre les mathématiques que si on voit à quoi elles servent

5.2 Questionnaire parents

Question	Réponse	Code Attribut
Quelle matière vous semble la plus utile pour l'avenir de votre enfant?	Mathématiques	Maths+ utile
	Langues vivantes	Langues viv + utile
	Français	Français + utile
	Sciences physiques	Sc. physiques + utiles
	Sciences naturelles	Sc. nat. + utiles
	Arts + divers	Arts + div. + utiles
	Sans reponse	Sans reponse + utile
Quelle matière semble la plus intéressante?	Mathématiques	Maths + intéressant
	Langues vivantes	Langues v. + intéressant
	Français	Français + intéressant
	Sciences physiques	Sc. physiques + intéressant
	Sciences naturelles	Sc. nat. + intéressant
	Arts + divers	Arts + div. + intéressant
	Sans reponse	Sans reponse + intéressant
Quel est l'enseignement qui vous paraît occuper la plus grande place en 6ème?	Mathématiques	Maths + grande place
	Langues vivantes	Langues viv + grande place
	Français	Français + grande place
	Sciences physiques	Sc. physiques + grande place
	Sciences naturelles	Sc. nat. + grande place
	Arts + divers	Arts + div. + grande place
	Sans reponse	Sans reponse + grande place
Quelle est la matière qui, cette année lui donne le plus de travail?	Mathématiques	Maths + travail
	Langues vivantes	Langues viv + travail
	Français	Français + travail
	Sciences physiques	Sc. physiques + travail
	Sciences naturelles	Sc. nat. + travail
	Arts + divers	Arts + div. + travail
	Sans reponse	Sans reponse + travail
Quelle est la matière qui semble présenter le plus de difficultés à votre enfant en 6ème?	Mathématiques	Maths + difficle
	Langues vivantes	Langues viv + difficle
	Français	Français +difficle
	Sciences physiques	Sc. physiques + difficle
	Sciences naturelles	Sc. nat. + difficle
	Arts + divers	Arts + div. + difficle
	Sans reponse	Sans reponse + difficle
Comment estimez-vous l'importance de l'enseignement des mathématiques?	Plus important pour les garçons que pour les filles	garçon > fille
	Aussi important pour les garçons que pour les filles	garçon = fille
	Moins important pour les garçons que pour les filles	fille > garçon
	sans opinion	sans rep sexe
Comment jugez-vous votre enfant en mathématiques?	très bon en mathématiques	très bon en maths
	bon en mathématiques	bon en maths
	Moyen en mathématiques	moyen en maths
	Faible en mathématiques	faible en maths
	Très faible en mathématiques	très faible en maths

D'autres attributs du questionnaire parents

ATTRIBUT	CODE ATTRIBUT
Si notre enfant réussit mal en mathématiques c'est parce que :	
Il ne travaille pas suffisamment	insuffisance travail
La méthode d'enseignement ne lui convient pas	enseign inadéquat
le contenu des cours trop difficile pour lui	difficulté
Il n'aime pas l'enseignant de maths cette année	n'aime pas prof
Raisons diverses, mauvais résultats	divers mauvais rés
Je ne sais pas	sans rep mauvais rés
L'enfant demande de l'aider dans son travail en maths à la maison	aide maison
J'aimais les maths quand j'étais élève	parents aiment maths
Je réussissais en maths	parents bon maths
Je souhaite que notre enfant suive une classe de type C	souhait classe C
Je ne sais pas s'il doiit suivre une classe C	sans rep classe C
Je pense qu'il est capable de suivre une classe C	capable classe C
Je ne sais pas s'il est capable de suivre une classe C	sans rep capable C
Il est trop tôt pour savoir s'il est capable de suivre une classe C	trop tôt classe C

6 BIBLIOGRAPHIE

Lerman I.C., Tallur B. (1980): Classification des éléments constitutifs d'une juxtaposition de tableaux de contingence *Revue de Statistique Appliquée, vol.XXVIII, no. 3*

Lerman I.C. (1981): Classification et analyse ordinale des données. *Dunod, Paris*

Lerman I.C., Peter P. (1985): Élaboration et logiciel d'un indice de similarité entre objets d'un type quelconque. Application au problème de concensus en classification. *Publication interne n0. 262, IRISA, Université de Rennes 1*

Lerman I.C. (1987): Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque.Application au problème de concensus en classification. *Revue de Statistique Appliquée, vol.XXXV, no. 2*

Tallur B. (1982): Un nouvel algorithme de classification hiérarchique des éléments constitutifs de tableau de contingence basé sur la corrélation. *Publication interne n0. 177, IRISA, Université de Rennes 1*

Tallur B. (1988): Contribution à l'analyse exploratoire de tableaux de contingence par la classification. *Thèse d'état, Université de Rennes 1*

Wald A., Wolfowitz J. (1944): Statistical tests based on permutations of the observations. *Ann. Math. Stat., vol. 15*