

R. GRAS

H. RATSIMBA-RAJOHN

**Analyse non symétrique de données par
l'implication statistique**

*Revue française d'automatique, d'informatique et de recherche
opérationnelle. Recherche opérationnelle*, tome 30, n° 3 (1996),
p. 217-232.

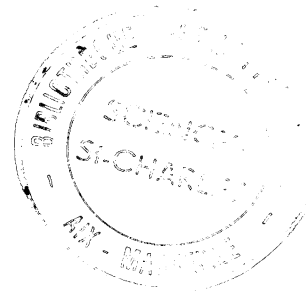
http://www.numdam.org/item?id=RO_1996__30_3_217_0

© AFCET, 1996, tous droits réservés.

L'accès aux archives de la revue « Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/legal.php>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>



ANALYSE NON SYMÉTRIQUE DE DONNÉES PAR L'IMPLICATION STATISTIQUE (*)

par R. GRAS ⁽¹⁾ et H. RATSIMBA-RAJOHN ⁽²⁾

Communiqué par Catherine ROUCAIROL

Résumé. – De nombreuses méthodes d'analyse de données permettent la classification de ces données selon un critère de ressemblance traduit par des indices différents. Cependant, dans de multiples situations réelles, se pose le problème de structuration d'un ensemble de variables observées sur un ensemble d'objets ou de sujets, structuration par inclusion ou par implication, du type : « si a alors b ». Notre approche, inspirée des travaux de I. C. Lerman, conduit à une classification orientée de ces variables, traduite par un graphe, puis à une classification orientée en classes de variables, traduite par une hiérarchie. On en examine les nœuds significatifs, ainsi que les contributions des objets-sujets et de certaines des catégories d'objets à l'apparition de classes particulières.

Mots clés : Implication, classification, hiérarchie, niveau significatif, contribution.

Abstract. – Many methods of data analysis built data organisations according to a criterion of resemblance measured by different indices. Mostly, these indices are symmetrical. However, in multiple real situations, we need to structure a set of variables or a set of variable classes according to inclusion or inference relation such that: "If a then b". Our approach, inspired by the works of I. C. Lerman, give us a direction to get an oriented classification of these variables, figured on a graph. Thus we obtain an oriented classification on different classes of variables represented by a hierarchy. In a next step, we examine the significant nodes of the latter as well as the contribution of subjects-objects considered individually or through a categorisation.

Keywords: Implication, classification, hierarchy, significant node, contribution.

Cet axe de recherche prend son origine, dans les années 1975-1979, à la faveur d'une problématique d'ordonnancement d'items proposés à 400 jeunes élèves, items supposés classés par complexité croissante selon la taxonomie

(*) Reçu en mars 1994.

⁽¹⁾ Institut de Recherche Mathématique de Rennes, Campus de Beaulieu, 35042 Rennes Cedex et IRESTE, route de Gachet, La Chantrerie, CP 3003, 44087 Nantes Cedex 03.

⁽²⁾ Laboratoire de didactique des Sciences et Techniques, 40, rue Lamartine, 33400 Talence.

de R. Gras qui affine et spécifie celle de Bloom (cf. [4]). Il s'agissait alors de valider cet ordre *a priori* par l'examen des performances de cette population d'élèves. Manifestement, il fallait alors disposer d'un outil non symétrique de traitement de données. La rencontre avec les approches d'I. C. Lerman (cf. [8] et [9]) du traitement de la ressemblance nous a permis de construire cet outil pour les variables binaires (réussite-échec) qu'il fallait classer, en rendant compte de l'in vraisemblance de la non-inférence. Dans le paragraphe 1, nous restituons la démarche adoptée par R. Gras dans sa thèse (cf. [4]). Dans les paragraphes suivants, nous présentons les développements ultérieurs des chercheurs de son équipe, qui font de l'analyse implicite une méthode ouverte aux spécialistes en Analyse de Données et aux chercheurs en Intelligence Artificielle.

1. IMPLICATION ENTRE VARIABLES BINAIRES ET EXTENSION À D'AUTRES TYPES

1.1. Modélisation dans le cas binaire

Considérons le croisement d'un ensemble E d'objets ou de sujets et d'un ensemble V de variables binaires. Soient a et b deux variables quelconques. On veut donner un sens à des expressions du type : « si a alors b », dans le cas où l'implication n'est pas stricte, *i.e.* dans le cas où l'ensemble des objets où a est vraie n'est pas contenu dans l'ensemble B où b est vraie. Pour ce faire, on associe à A et B de façon indépendante, comme le fait I. C. Lerman pour la similarité selon l'algorithme de la vraisemblance du lien (A.V.L. cf. [8]), deux parties aléatoires X et Y qui respectent de façon déterministe ou probabiliste les cardinaux respectifs de A et B . On notera \bar{B} , \bar{Y} (resp. \bar{b} ou \bar{a}) les complémentaires de B , Y dans E (resp. les négations de b ou a).

Principe d'admissibilité de l'implication

$(a \Rightarrow b)$ est admissible au niveau de confiance $1 - \alpha$ si et seulement si

$$\Pr [\text{Card} (X \cap \bar{Y}) \leq \text{card} (A \cap \bar{B})] \leq \alpha$$

Notons n , n_a , $n_{\bar{b}}$, $n_{a \wedge \bar{b}}$, les cardinaux respectifs de E , A , $\bar{B} \cap \bar{Y}$. Dans [9], il est établi que, pour un certain modèle de tirage, la variable aléatoire $\text{Card} (X \cap \bar{Y})$, de réalisation $n_{a \wedge \bar{b}}$, suit une loi de Poisson de

paramètre $n_a n_{\bar{b}}/n$. Si nous centrons et réduisons cette variable, pour des valeurs convenables des paramètres nous obtenons la variable :

$$Q(a, \bar{b}) = (\text{Card}(X \cap \bar{Y}) - n_a n_{\bar{b}}/n) / \sqrt{\frac{n_a n_{\bar{b}}}{n}}$$

qui, dans l'hypothèse d'indépendance, suit approximativement une loi normale centrée réduite. Soit $q(a, \bar{b})$ la valeur empirique de $Q(a, \bar{b})$ au cours du croisement $Ex V$ et $\varphi(a, \bar{b}) = \frac{1}{\sqrt{2\pi}} \int_{q(a, \bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$.

DÉFINITION 1: Soit a et b deux variables binaires. Le nombre $\varphi(a, \bar{b})$ est appelé *intensité d'implication* de a sur b . Ce nombre est égal à $1 - \Pr[Q(a, \bar{b}) \leq q(a, \bar{b})]$.

Le principe d'admissibilité de l'implication s'exprime alors de la façon suivante :

Principe d'admissibilité gaussienne de l'implication

L'implication $a \Rightarrow b$ est *admissible au niveau de confiance* $1 - \alpha$ si et seulement si

$$\varphi(a, \bar{b}) = 1 - \Pr[Q(a, \bar{b}) \leq q(a, \bar{b})] \geq 1 - \alpha$$

On montre que si $n_a \leq n_b$ alors $\varphi(a, \bar{b}) \geq \varphi(b, \bar{a})$ (cf. [5] et [7]).

Remarque 1 : Dans la pratique, on retient le plus souvent le niveau de confiance de .95.

Remarque 2 : Ce principe se distingue de ceux retenus par d'autres auteurs dont :

* J. Loevinger [10] : l'indice d'implication est $H(a, b) = 1 - \frac{n n_{a \wedge \bar{b}}}{n_a n_{\bar{b}}}$.

Mais cet indice est invariant dans toute dilatation des ensembles de référence, propriété qui s'oppose aux principes statistiques;

* J. Pearl [11], S. Acid [1], A. Gammernan et Z. Luo : le critère d'implication est mesuré par l'écart entre la distribution conjointe de a et b (et non pas a et \bar{b}) et leur distribution produit;

* J. G. Ganascia [3] évalue l'« incertitude » sur $a \Rightarrow b$ par l'indice $2\Pr[b|a] - 1$, indice qui ne sépare pas deux implications, l'une triviale, l'autre très informative.

Remarque 3 (cf. [5]) : La relation suivante entre le χ^2 d'indépendance et l'indice $q(a, \bar{b})$:

$$\frac{\chi^2}{q^2(a, \bar{b})} = \frac{n^2}{n_{\bar{a}} n_b}$$

montre la différence entre les deux concepts et donc la limitation de la seule considération de la case (a, \bar{b}) du tableau de croisement de a et b . Bien souvent, les psychologues ne retiennent que cette case pour étudier l'implication.

De la même façon, la relation : $\frac{\rho(a, \bar{b})}{q(a, \bar{b})} = -\sqrt{\frac{n}{n_{\bar{a}} n_b}}$ montre que, si corrélation et implication vont plutôt généralement dans le même sens, les deux concepts sont cependant bien distincts.

Remarque 4 : En appliquant le principe d'admissibilité gaussienne, on prouve par un prolongement par continuité que, si $B = E$, alors $\varphi(a, \bar{b}) = .5$. Ce résultat, s'il ne s'accorde pas avec une logique de l'inclusion, respecte par contre la sémantique originelle de l'implication.

1.2. Graphe d'implication

L'originalité et un des fruits de notre approche résident dans la capacité de l'intensité d'implication de permettre une structuration de l'ensemble des variables de V en un préordre partiel. En effet, pour un niveau de confiance $1 - \alpha$ donné, on peut définir une relation symétrique et transitive sur V telle qu'il lui corresponde un graphe transitif, pondéré par φ . Ce graphe est alors susceptible de représenter V de façon à mettre en évidence une structure globale et des sous-structures en chemins totalement ordonnés. L'utilisateur (cf. [6]) tire intérêt de ce type d'image qui facilite et enrichit l'interprétation.

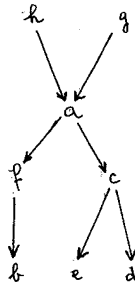


Figure 1.

1.3. Extension

Dans la thèse d'A. Larher [17], on étend la notion d'implication entre variables binaires à deux autres types :

* *des variables modales* associées à une logique multivalente; ici, les valeurs de vérité ne sont plus seulement 0 ou 1, mais toute valeur de l'intervalle $[0, 1]$;

* *des variables quantitatives ou numériques* correspondant sémantiquement, par exemple, à une fréquence ou un nombre d'occurrences d'un caractère chez un sujet.

Un même indice d'implication leur est associé et sa restriction aux variables binaires coïncide avec l'indice défini pour celles-ci.

2. IMPLICATION ENTRE CLASSES DE VARIABLES

Dans une perspective comparable à celle des classifications hiérarchiques, nous élargissons le concept d'implication statistique entre variables à celui d'implication entre classes de variables. Cependant afin de réduire l'effet de chaînage qui conduit à des associations forcées, nous ne considérerons que des classes admettant une bonne consistance, une bonne « cohésion implicative ». La modélisation de cette notion intuitive au sein d'une classe est choisie de telle façon qu'elle rende compte d'une orientation générale de ses éléments, orientation déjà signifiée par l'intensité d'implication. De plus, comme la cohésion s'oppose au désordre, elle est fondée sur le concept d'entropie (au sens de Shannon).

2.1. Cohésion implicative d'une classe à deux éléments

Supposons, par exemple, $n_a \leq n_b$ et considérons l'événement aléatoire $[Q(a, \bar{b}) \geq q(a, \bar{b})]$. Notons $p = \varphi(a, \bar{b})$. L'entropie de l'expérience où cet événement peut ou non se réaliser est :

$$\mathfrak{S} = -p \log_2 p - (1 - p) p \log_2 (1 - p)$$

DÉFINITION 2 : On appelle cohésion de la classe (a, b) et on note $c(a, b)$ le nombre :

$$* c(a, b) = \sqrt{1 - \mathfrak{S}^2} \text{ si } p \geq .50$$

$$* c(a, b) = 0 \text{ si } p \leq .50.$$

PROPOSITION : *La cohésion de la classe (a, b) est fonction croissante de l'intensité d'implication entre a et b .*

En prolongeant par continuité la cohésion c en tant que fonction de p , on obtient : si $p = 1$ alors $c(a, b) = 1$.

2.2. Cohésion implicative d'une classe quelconque

r variables a_1, \dots, a_r étant données, supposons que leurs effectifs soient classés par valeurs croissantes : $n_{a1} \leq \dots \leq n_{ar}$.

DÉFINITION 3 : On appelle cohésion de la classe $\mathbf{A} = (a_1, \dots, a_r)$ la moyenne géométrique des cohésions des classes à deux éléments qui peuvent être définies à partir de cette classe, soit : $C(\mathbf{A}) = [\prod c(a_i, a_j)]^{2/r(r-1)}$ pour $i \in \{1, \dots, r-1\}$, $j \in \{2, \dots, r\}$, $j > i$.

2.3. Implication entre classes

Nous souhaitons que l'implication entre classes intègre, d'une part, la qualité de cohésion de chacune, d'autre part, une liaison extrême entre les éléments des deux classes (nous choisissons la liaison maximale) et, en conséquence, les cardinaux des classes. A ces conditions, l'information restituée par l'implication est de bonne qualité.

Soit $\mathbf{A} = \{a_1, \dots, a_r\}$ et $\mathbf{B} = \{b_1, \dots, b_s\}$ supposées totalement ordonnées selon les cardinaux des référentiels de chaque variable, comme dans 2-2.

DÉFINITION 4 : On appelle implication de la classe \mathbf{A} sur la classe \mathbf{B} le nombre :

$$\Psi(\mathbf{A}, \mathbf{B}) = \{\sup \varphi(a_i, \bar{b}_j)\}^{rs} \times [C(\mathbf{A}) \times C(\mathbf{B})]^{1/2}$$

$$i \in \{1, \dots, r\}, \quad j \in \{1, \dots, s\}$$

Ainsi, l'implication entre \mathbf{A} et \mathbf{B} croît avec leurs cohésions, avec leur liaison maximale et décroît avec leurs cardinaux. Comme dans le cas de deux variables, on retiendra l'implication dans le sens où elle est maximale ($\mathbf{A} \Rightarrow \mathbf{B}$ ou $\mathbf{B} \Rightarrow \mathbf{A}$).

2.4. Hiérarchie implicative de classes

Un algorithme, basé sur le critère de la cohésion de classe, permet d'élaborer un arbre hiérarchique ascendant. Ainsi, au premier niveau, se réunissent, en une classe orientée par l'implication maximale, deux variables formant une classe dont la cohésion est maximale parmi toutes les cohésions des autres classes à deux termes. Au niveau suivant, on trouvera une classe à deux termes ou bien une classe à trois termes dont deux d'entre eux ont été réunis au niveau précédent, etc. En cas d'exæquos, la plus faible cardinalité définit le critère de priorité. L'élargissement d'une classe, contrairement aux hiérarchies de similarité, s'arrête dès que toute extension de cette classe conduit à une cohésion nulle.

Exemple :

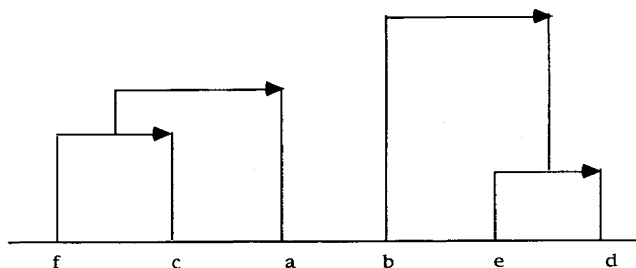


Figure 2.

3. NIVEAUX SIGNIFICATIFS

Étant donné la multiplicité des niveaux de formation des classes, il est indispensable de dégager ceux qui sont les plus pertinents par rapport à l'intention classificatrice du chercheur et eu égard aux critères choisis. Nous procédons (*cf.* [12]) alors de façon comparable à celle adoptée primitivement par I. C. Lerman et relativement à la hiérarchie de similarité mais en reconditionnant son approche.

3.1. Préordre cohésitif

Considérons l'ensemble V des variables $\{a_1, a_2, \dots, a_m\}$ et l'ensemble des couples (a, b) de $V \times V$ tels que $a \neq b$. Il existe $m(m-1)$ tels couples auxquels on a associé leurs cohésions $c(a, b)$ respectives.

DÉFINITION 5 : On appelle *préordre initial et global cohésif* sur $V \times V$ (ou préordonnance), le préordre Ω induit par l'application cohésion c sur $V \times V$.

Soit $G(\Omega)$ son graphe dans $V \times V$. D'après les paragraphes 1 et 2 qui précèdent, il s'ensuit que :

* d'une part, la classe de préordre correspondant à $c = 0$ contient tous les couples tels que $\varphi(a, \bar{b}) \leq 0,5$,

* d'autre part, si $n_a \leq n_b$ alors $c(b, a) \leq c(a, b)$.

Remarquons, par contre, que si $c(a, b) \leq c(c, d)$ on n'a pas nécessairement $c(b, a) \leq c(d, c)$ ou $c(b, a) \geq c(d, c)$.

3.2. Détermination des niveaux significatifs

Plaçons-nous à un niveau quelconque k de la hiérarchie. A ce niveau, se forme une classe de m_i variables ($2 \leq m_i \leq m$) dont la cohésion est moins bonne que celle des classes antérieurement formées, conformément à l'algorithme retenu, et meilleure que celles des classes à venir.

Soit Π_k la partition sur V définie à ce niveau constituée des classes qui y sont déjà formées et, éventuellement, des singletons non encore associés. Π_k est plus fine que Π_{k+1} .

Soit S_{Π_k} l'ensemble des couples séparés à ce niveau et R_{Π_k} l'ensemble des couples qui y sont réunis pour la première fois, étant entendu que l'on dira que le couple (a, b) est réuni si a et b appartiennent à la même classe du type $(\dots(\dots a, \dots))\dots b)\dots$.

L'ensemble $G(\Omega) \cap [S_{\Pi_k} \times R_{\Pi_k}]$ est constitué des couples de couples qui au niveau k respectent le préordre initial. Par exemple, si l'on a $c(e, f) < c(a, b)$ (donc $((e, f), (a, b)) \in G(\Omega)$) et si au niveau k , e et f sont séparés alors que a et b se réunissent dans la classe qui se forme, le couple $((e, f), (a, b))$ appartient à $G(\Omega) \cap [S_{\Pi_k} \times R_{\Pi_k}]$.

Comme il a été fait pour le cardinal de $A \cap \bar{B}$, associons (cf. I. C. Lerman [8]) au cardinal de $G(\Omega) \cap [S_{\Pi_k} \times R_{\Pi_k}]$ l'indice aléatoire $\text{card}[G(\Omega^*) \cap [S_{\Pi_k} \times R_{\Pi_k}]]$ où Ω^* est une préordonnance aléatoire dans l'ensemble, muni d'une probabilité uniforme, de toutes les préordonnances de même type cardinal que Ω . Cet indice a pour espérance $1/2 \text{card}[S_{\Pi_k} \times R_{\Pi_k}]$ et pour variance $\text{card}[S_{\Pi_k} \times R_{\Pi_k}] \text{card}[G(\Omega)]$.

Soit $s(\Omega, k)$ l'indice centré réduit obtenu :

$$\frac{(\text{card}[G(\Omega^*) \cap [S_{\Pi_k} \times R_{\Pi_k}]] - 1/2 \text{card}[S_{\Pi_k} \times R_{\Pi_k}])}{(\text{card}[S_{\Pi_k} \times R_{\Pi_k}] \text{card}[G(\Omega)])^{1/2}}.$$

DÉFINITION 6 : On appelle *nœud significatif* tout nœud correspondant à un maximum local de $s(\Omega, k)$ au cours de la constitution de la hiérarchie implicative. Nous dirons dans ce cas que la partition Π_k est en *résonance partielle* avec Ω .

Si, de plus, $G(\Omega) \cap [S_{\Pi_k} \times R_{\Pi_k}] = S_{\Pi_k} \times R_{\Pi_k}$, nous dirons que la partition Π_k est en *résonance totale* avec Ω .

Le logiciel d'analyse de données C.H.I.C. (cf. [13]) permet le traitement complet de données quantitatives, ainsi que la sortie du graphe d'implication et de la hiérarchie implicative en mentionnant les nœuds significatifs.

4. CONTRIBUTION DES INDIVIDUS (OU OBJETS) DANS LE CAS BINAIRE

A un niveau quelconque de la hiérarchie se forme une classe C de cohésion non nulle. Notre objectif, particulièrement dans le cas d'un nœud significatif, est de définir un critère permettant d'identifier un ou des individus, puis, dans le paragraphe suivant, la catégorie d'individus, contribuant le plus à la constitution de cette classe. Le comportement de ces individus sera en harmonie avec le comportement statistique à l'origine de la classe.

4.1. Puissance implicative d'une classe

Plaçons-nous à un niveau k de la hiérarchie où viennent de se réunir, pour former C , deux classes A et B telles que $A \Rightarrow B$ au sens du 2.3.

DÉFINITION 7 : Le couple (a, b) tel que : $\forall i \in A, \forall j \in B \varphi(a, \bar{b}) \geq \varphi(i, \bar{j})$ est appelé *couple générique* de C . C'est ce couple, généralement unique, qui intervient par le sup. dans le calcul de l'implication de A sur B . Le nombre $\varphi(a, \bar{b})$ est appelé *implication générique* de C .

Mais, dans chaque sous-classe de C , existe également un couple générique. Précisément, si C est constituée de g ($g \leq k$) sous-classes (C comprise), il y a g couples génériques à l'origine de C et g intensités maximales d'implication $\varphi_1, \varphi_2, \dots, \varphi_g$ qui leur correspondent.

DÉFINITION 8 : Le vecteur $(\varphi_1, \varphi_2, \dots, \varphi_g)$, élément de $[0, 1]^g$, est appelé *vecteur puissance implicative de C*, traduisant une force implicative interne à C.

4.2. Puissance implicative d'un individu sur une classe et distance à cette classe

Un individu x quelconque respecte ou non l'implication du couple générique d'une classe. Associant logique formelle et considération sémantique, nous poserons, en fonction des valeurs prises par a et b en x :

$$\begin{aligned} \varphi_x(a, \bar{b}) &= 1 \quad \text{si } a = 1 \text{ ou } 0 \quad \text{et } b = 1; \\ \varphi_x(a, \bar{b}) &= 0 \quad \text{si } a = 1 \text{ et } b = 0; \quad \varphi_x(a, \bar{b}) = p \quad \text{si } a = b = 0 \end{aligned}$$

avec $p \in]0, 1]$. Le plus souvent, nous choisissons $p = .5$, valeur neutre.

Ainsi, à x , nous pouvons associer g nombres $(\varphi_{x,1}, \varphi_{x,2}, \dots, \varphi_{x,g})$ correspondant aux valeurs prises en x par les g implications génériques de la classe C.

DÉFINITION 9 : Le vecteur $(\varphi_{x,1}, \varphi_{x,2}, \dots, \varphi_{x,g})$, élément de $[0, 1]^g$, est appelé *vecteur puissance implicative de x*. L'individu x_t , peut-être fictif, dont toutes les composantes du vecteur puissance sont égales à 1 est appelé *individu idéal théorique de C*.

Dans ces conditions, on peut munir l'espace des puissances $[0, 1]^g$ d'une métrique du type χ^2 afin d'accentuer les effets de fortes implications génériques.

DÉFINITION 10 : On appelle *distance implicative d'un individu x à la classe C* le nombre :

$$d(x, \mathbf{C}) = \left[\frac{1}{g} \sum_{i=1}^{i=g} \frac{[\varphi_i - \varphi_{x,i}]^2}{1 - \varphi_i} \right]^{\frac{1}{2}}$$

Ce nombre n'est autre que la distance du χ^2 entre les deux distributions $\{1 - \varphi_i\}_i$ et $\{1 - \varphi_{x,i}\}_i$ qui expriment les écarts entre les implications génériques empiriques et l'implication stricte. Si pour un i , $\varphi_i = 1$,

nous poserons, par convention, $\varphi_{x,i} = 1$. Cette convention ne se fait pas contre nature puisque, dans ce cas, l'implication générique est maximale et significative d'une excellente liaison implicative entre ses deux termes, vérifiée par tous les individus x de E . Ainsi, si le dénominateur s'annule, il en est de même du numérateur, et l'on pourra attribuer la valeur 0 au quotient.

4.3. Contribution d'un individu et d'une catégorie d'individus à une classe

Nous la définirons à partir de la « distorsion » de l'individu considéré par rapport à l'individu idéal théorique, tout en remarquant qu'il peut exister des individus réels dont la distance à la classe C soit inférieure à la distance à cette même classe de l'individu idéal théorique. La contribution d'une catégorie G d'individus s'en déduira.

DÉFINITION 11 : La contribution de x à C est : $\gamma(x, C) = \frac{d(x_t, C)}{d(x, C)}$ et celle de G est : $\gamma(G, C) = \frac{1}{\text{card } G} \sum_{x \in G} \gamma(x, G)$.

Ces contributions peuvent être infinies (pour des configurations contenant des x à distance nulle de C) mais, en particulier, supérieures à 1 pour certains individus.

Afin de donner au chercheur le moyen de savoir ou de vérifier rapidement si telle catégorie de sujets-objets qui l'intéresse est statistiquement déterminante dans la constitution d'une classe implicative, un algorithme a été élaboré en s'appuyant sur les deux notions suivantes : groupe optimal et catégorie déterminante.

DÉFINITION 12 : Soit E la population étudiée. Un *groupe optimal d'une classe implicative* C , noté $GO(C)$, est le sous-ensemble de E qui accorde à cette classe une contribution plus grande que son complémentaire et qui forme avec celui-ci une partition en deux classes maximisant la variance inter-classe de la série statistique des contributions individuelles. Une telle partition est dite *significative*.

L'existence de ce groupe optimal est démontrée dans [12]. Les propriétés utilisées sont aussi celles qui le sont pour établir l'algorithme sur lequel se basent les modules des programmes informatiques qui construisent automatiquement dans C.H.I.C. (cf. [13]) chaque sous-groupe optimal.

Considérons une partition $\{G_i\}_i$ de E , X_i une partie aléatoire de E ayant le même cardinal que G_i , et Z_i la variable aléatoire $\text{Card}(X_i \cap GO(C))$. Z_i suit une loi binomiale de paramètres : $\text{card } G_i$ et $\text{card } GO(C)/\text{card } E$.

DÉFINITION 13 : On appelle *catégorie la plus contributive à la constitution de la classe implicative C* , la catégorie qui minimise l'ensemble $\{p_i\}_i$ des probabilités p_i telles que :

$$\forall i, \quad p_i = \text{Prob}[\text{card } G_i \cap GO(C) < Z_i]$$

Une catégorie G_0 est dite *déterminante au seuil α* si la probabilité associée p_0 est inférieure à α .

Ainsi, la signification d'une classe ayant été donnée par le chercheur, il lui associera la sous-population la plus porteuse de ce sens. Cette approche est comparable à celle de I.-C. Lerman pour l'analyse des similarités.

5. EXEMPLE

Considérons, au sein d'une population de 23 individus, un petit sondage au moyen de six questions à réponse « oui » ou « non ». On rassemble, dans le tableau ci-dessous, les réponses recueillies en notant 1 pour « oui » et 0 pour « non » en fonction de l'opinion des individus I001, I002, . . . , I023 par rapport aux questions 1, 2, . . . , 6. Lorsque les individus, classés de façon non exclusive dans trois catégories notées X , Y et Z (par exemple : âge, sexe, C.S.P.), y sont présents, on marque cette apparence par le chiffre 1, et 0 dans le cas contraire.

Le premier objectif est de savoir si, dans la population, le fait de répondre « oui » à une question entraîne statistiquement une réponse « oui » à une autre question; de façon plus large, on peut se demander également si la réponse « oui » à un ensemble de questions est suivie préférentiellement de cette réponse dans un autre ensemble de questions.

TABLEAU 1

	1	2	3	4	5	6	X	Y	Z
I001	1	1	1	0	1	0	0	0	1
I002	0	0	1	0	1	0	1	1	0
I003	1	0	0	0	1	0	0	0	1
I004	0	0	1	0	1	1	1	1	0
I005	1	1	1	1	1	1	0	0	1
I006	1	1	0	0	0	0	0	0	1
I007	1	0	0	0	0	0	0	0	1
I008	1	0	0	0	1	0	0	0	1
I009	0	0	1	0	1	0	1	1	0
I010	0	0	1	0	1	0	1	1	0
I011	0	0	0	0	1	0	0	0	0
I012	0	0	0	0	0	0	0	0	0
I013	0	0	1	0	1	0	1	1	1
I014	0	0	1	0	1	1	1	1	0
I015	1	0	0	1	1	0	0	0	1
I016	0	0	0	0	1	0	0	0	0
I017	0	0	1	0	1	0	1	1	0
I018	1	1	0	0	1	0	0	0	1
I019	1	0	1	1	0	0	1	0	1
I020	0	0	0	0	0	0	0	0	0
I021	0	0	0	0	1	0	0	0	0
I022	1	1	0	0	1	0	0	0	1
I023	0	0	0	0	0	0	1	1	0

La matrice des implications 2 à 2 permet de satisfaire la première attente et de représenter sous la forme de graphes les structurations successives des 6 questions à des seuils différents.

TABLEAU 2

$U \Rightarrow V$	1	2	3	4	5	6
1	1.000000	0.000000	0.285385	0.000000	0.404286	0.000000
2	0.953623	1.000000	0.458802	0.000000	0.605066	0.000000
3	0.285385	0.000000	1.000000	0.000000	0.840373	0.000000
4	0.903567	0.589788	0.703405	1.000000	0.402944	0.646862
5	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
6	0.407601	0.589788	0.903567	0.646862	0.811825	1.000000

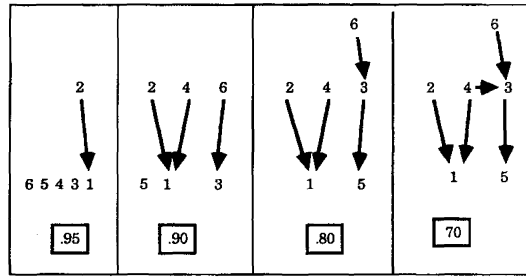
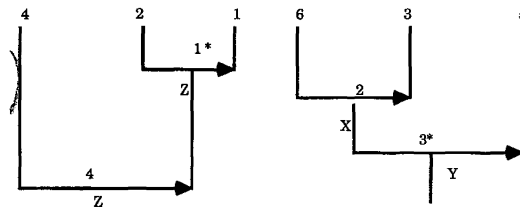


Figure 3.

La hiérarchie, définie à partir des cohésions des classes, permet de son côté de rendre compte de façon plus dynamique des relations implicatives entre classes de questions et, ainsi, de répondre à la seconde attente.



* classe d'implication significative
 X, Y, Z : catégories déterminantes par niveau.

Figure 4.

Sur cet arbre (fig. 4), nous repérons les niveaux où la partition de l'ensemble des questions est significatif. A ces niveaux, la classification est en accord, en résonance partielle avec le préordre initial Ω défini sur l'ensemble des couples :

- niveau 1 : 10 accords avec Ω (niveau *significatif*)
- niveau 2 : 9 accords (*non significatif*)
- niveau 3 : 12 accords (*significatif*)
- niveau 4 : 5 accords (*non significatif*).

Nous pouvons rechercher les groupes d'individus qui ont le plus contribué à la formation de la classe constituée à chacun des niveaux.

TABLEAU 3
Groupe Optimal par Niveau

1	2	3	4
I001	I001	I001	I001
I005	I005	I005	I005
I019	I019	I002	I019
I003	I002	I004	I003
I006	I004	I009	I007
I008	I010	I013	I008
I015	I013	I014	I015
I018	I014	I017	I018
I022	I017		I022

Le tableau ci-dessus y répond en désignant le groupe optimal associé à chaque classe. Enfin, si nous souhaitons savoir, avec le minimum de risque d'erreur, laquelle des 3 catégories *a priori* X , Y ou Z a contribué le plus à la formation de chaque classe, utilisant la procédure indiquée dans le paragraphe 4-3, nous pouvons y répondre selon la suite de décisions données par le tableau suivant :

TABLEAU 4

Niveau 1 : catég. Z	Risque : 0,001600	Niveau 2 : catég. X	Risque : 0,007100
Niveau 3 : catég. Y	Risque : 0,007400	Niveau 4 : catég. Z	Risque : 0,001600

Cette méthode a été utilisée dans la thèse de H. Ratsimba-Rajohn (*cf.* [12]) pour mettre en évidence un fait didactique intéressant. Une étude est conduite avec 104 élèves répartis sur deux ans dans deux classes par année. Trois enseignants se partagent l'encadrement de ces classes. L'analyse se fonde sur un test mathématique et montre des regroupements ordonnés de réponses des élèves en classes implicatives, significativement différentes d'une année à l'autre et d'un enseignant à un autre. Les résultats fournis par l'étude implicative sont corroborés par des observations directes. Ils montrent l'effet d'une certaine obsolescence des situations didactiques, due sans doute à la médiocre gestion par les enseignants d'un processus didactique appelé « ostension ». Il n'était possible de détecter ce phénomène que par une étude non symétrique de comportement de réponse.

CONCLUSION

Nous voulons souligner la richesse informative d'une telle approche que permet l'implication statistique. Les structures obtenues, différant

sensiblement de celles que fournit une similarité symétrique, permettent d'en dégager une dynamique, voire des genèses, des généralisations qui intéressent non seulement les psychologues et les didacticiens pour lesquels a été conçue la méthode, mais également des chercheurs en Intelligence Artificielle dont la problématique se centre sur l'élaboration de bases de connaissances. Certains développements ultérieurs devraient contribuer à ce type d'étude.

RÉFÉRENCES

1. S. ACID, L. M. DE CAMPOS, A. GONZALEZ, R. MOLINA et N. PEREZ DE LA BLANCA, *Learning with Castle in Symbolic and quantitative approaches to uncertainty*, Springer-Verlag, 1991, p. 99-106.
2. E. DIDAY, *Towards a statistical theory of intentions for knowledge analysis*, Rapport de recherche 1494, I.N.R.I.A., Rocquencourt, 1991.
3. J. G. GANASCIA, Charade : Apprentissage de bases de connaissances, *Induction symbolique numérique à partir de données*, CEPADUES, 1991.
4. R. GRAS, Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Thèse d'État Mathématiques et Applications, Université de Rennes 1, 1979.
5. R. GRAS et A. LARHER, L'implication statistique, une nouvelle méthode d'analyse de données, *Mathématiques, Informatique et Sciences Humaines*, 1993, 120, p. 5-31.
6. R. GRAS, A. TOTOHASINA, S. AG ALMOULOU, H. RATSIMBA-RAJOHN et M. BAILLEUL, *la méthode implicative en didactique. Applications*, Actes du Colloque « 20 ans de Didactique des Mathématiques en France », La Pensée Sauvage, Grenoble, 1994.
7. A. LARHER, Implication statistique et applications à l'analyse de démarches de preuve mathématique, Thèse de l'Université de Rennes 1, Mathématiques et Applications, 1991.
8. I. C. LERMAN, *Classification et analyse ordinale des données*, Dunod, Paris, 1981.
9. I. C. LERMAN, R. GRAS et H. ROSTAM, Elaboration et évaluation d'un indice d'implication pour des données binaires, *Mathématiques et Sciences Humaines*, 1981, 74, p. 5-35 et 75, p. 5-47.
10. J. LOEVINGER, A systematic approach to the construction and evaluation of tests of ability, *Psychological Monographs*, 1947, 61, n° 4.
11. J. PEARL, *Probabilistic Reasoning in intelligent systems*, Morgan Kaufmann, San Mateo, 1988.
12. H. RATSIMBA-RAJOHN, Contribution à l'étude de la hiérarchie implicative, application à l'analyse de la gestion didactique des phénomènes d'ostension et de contradiction, Thèse de l'Université de Rennes 1, Mathématiques et Applications, 1992.
13. C.H.I.C. ou Classification Hiérarchique Implicative et Cohésitive : logiciel d'analyse de données, IRMAR, Université de Rennes.