LES FONDEMENTS DE L'ANALYSE STATISTIQUE IMPLICATIVE

REGIS GRAS[1]

RESUME : *Partie de situations de didactique des mathématiques, la méthode implicative se développe au fil des problèmes rencontrés et des questions posées. Son objectif majeur vise la structuration de données croisant individus et variables, l'extraction de règles entre les variables et, à partir de la contingence de ces règles, l'explication et donc la prévision dans divers domaines : psychologie, sociologie, biologie, etc.. C'est ainsi que les concepts d'intensité d'implication, de cohésion de classes, d'implication-inclusion, de significativité de niveaux hiérarchiques, de contribution de variables supplémentaires, etc., sont fondés. De la même façon, au traitement de variables binaires s'ajoutent ceux des variables modales, fréquentielles et, récemment, de variables-intervalles.*

MOTS-CLES : *implication statistique, implication-inclusion, entropie, cohésion, graphe implicatif, arbre hiérarchique, niveau significatif, variable binaire, variable modale, variable fréquentielle, variable-intervalle, règle, méta-règle, contribution, typicalité.*

SUMMARY : FOUNDATIONS OF IMPLICATIVE STATISTICAL ANALYSIS
Implicative analysis arose to solve practical problems in mathematical didactics and developed from there as problems were encountered and questions posed. Its main aim is to interpret the structure of data, both specific and variable, the determination of patterns among the variables, and, following from these patterns, to provide analysis of and hence predictions about various disciplines: psychology, sociology, biology, etc. In this way it is possible to provide a foundation for the concepts of: strength of implication, class cohesion,implication-inclusion, significance of hierarchical levels and contribution of supplementary variables. In the same way the concepts of ordinal, discrete and, more recently, continuous variables are able to be added to an analysis of binary variables.

KEYWORDS : *statistical implication, implication-inclusion, entropy, cohesion, implication graph, hierarchical tree, significance level,binary variable, ordinal variables, discrete variables, continuous variables, rules, meta-rules, contribution, typicality.*

[1]Professeur Emérite à l'Ecole Polytechnique de l'Université de Nantes, La Chantrerie, BP 50609 44306 Nantes cedex 03, e-mail : regisgra@club-internet.fr

INTRODUCTION

Différentes approches théoriques ont été adoptées pour modéliser l'extraction et la représentation de règles d'inférence imprécises (ou partielles) entre variables binaires (ou attributs ou caractères) décrivant une population d'individus (ou sujets ou objets). Mais les situations de départ et la nature des données ne modifient pas la problématique fondamentale. Il s'agit de découvrir des règles non symétriques pour modéliser des relations du type *"si a alors presque b"*. C'est, par exemple, l'option des réseaux bayesiens (par exemple : Amarger S., Pearl J. 1988) ou des treillis de Galois (par exemple : Simon A. 2000). Mais le plus souvent, la probabilité conditionnelle est le moteur de la définition de l'association, même si l'indice de cette association retenu est de type multivarié (par exemple : Bernard J.-M. 1999). De plus et à notre connaissance, d'une part, le plus souvent les développements s'arrêtent à la proposition d'un indice d'implication partielle pour des données binaires, d'autre part, cette notion n'est pas étendue à l'extraction et la représentation hiérarchique de méta-règles, ni à la recherche des sujets et catégories de sujets responsables des associations. Nous proposons ici ces prolongements après avoir rappelé le paradigme fondateur.

L'INTENSITE D'IMPLICATION REVISITEE

RAPPEL DE LA PROBLEMATIQUE DE L'IMPLICATION STATISTIQUE DANS LE CAS BINAIRE

Une population E d'objets ou de sujets est croisée avec des variables (caractères, critères, réussites, ...) que l'on interroge de la façon suivante : "*dans quelle mesure peut-on considérer que le fait de relever de la variable a implique celui de relever de la variable b ? Autrement dit, les sujets ont-ils tendance à être b si l'on sait qu'ils sont a ?"*. Dans les situations naturelles, humaines ou sciences de la vie, où les théorèmes (si a alors b) au sens mathématique du terme ne peuvent être établis du fait des exceptions qui les entachent, il est important pour le chercheur et le praticien de "*fouiller dans ses données*" afin de dégager cependant des règles assez fiables (des sortes de "théorèmes partiels") pour pouvoir conjecturer une genèse, décrire, structurer une population et faire l'hypothèse d'une certaine stabilité à des fins prédictives. Mais cette fouille exige la mise au point de méthodes pour la guider et pour la dégager du tâtonnement et de l'empirisme.

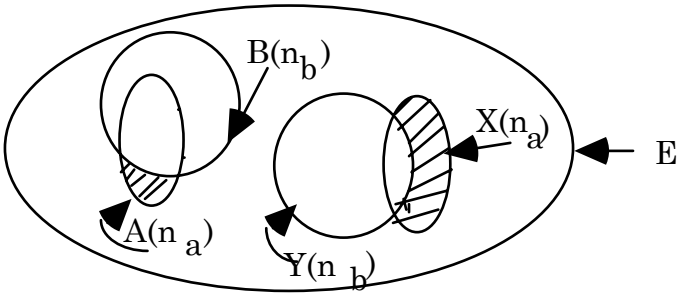
Pour cela, à l'instar de la méthode de mesure de la similarité de I.C. Lerman (Lerman I.C. 1970, 1981), nous définissons (Gras R. 1979, 1996) la mesure de la relation implicative a => b à partir de l'invraisemblance de l'apparition, dans les données, du nombre de cas qui l'infirment, c'est-à-dire pour lesquels a est vérifié sans que b ne le soit. Cette mesure est relativisée au nombre de données vérifiant respectivement a et non b. Elle quantifie "*l'étonnement"* de l'expert devant le nombre invraisemblablement petit de contre-exemples eu égard à une indépendance présumée et aux effectifs en jeu.

Précisons. Un ensemble fini V de v variables est donné : a,b,c,...Dans la situation paradigmatique classique, il s'agit des performances (réussite-échec) à des items d'un questionnaire. A un ensemble fini E de n sujets x, on associe, par abus d'écriture, les fonctions du type : $x \rightarrow a(x)$ où a(x) = 1 (ou a(x) = vrai) si x satisfait ou possède le caractère a et 0 (ou a(x) = faux) sinon. En intelligence artificielle, on dira que x est un exemple ou une instance pour a si a(x) = 1 et un contre-exemple dans le cas contraire.

La règle "a => b" est logiquement vraie si pour tout x, b(x) n'est nul que dans le cas où a(x) l'est aussi ; autrement dit si l'ensemble A des x pour lesquels a(x)=1 est contenu dans l'ensemble B des x pour lesquels b(x)=1. Cependant, cette inclusion stricte n'est qu'exceptionnellement observée dans la réalité. Dans le cas d'un questionnaire de connaissances, on pourrait en effet observer quelques rares élèves réussissant un item a et ne réussissant pas l'item b, sans que soit contestée la *tendance* à avoir b quand on a a.

Relativement aux cardinaux de E (soit n), mais aussi de A (soit n_a) et B (soit n_b), c'est donc le "poids" des contre-exemples (soit $n_{a\wedge\bar{b}}$) qu'il faudra prendre en compte pour accepter statistiquement de conserver ou non la **quasi-implication** ou la **quasi-règle** "a => b".

Pour mathématiser cette quasi-règle, nous considérons, comme le fait I.C. Lerman pour la similarité, deux parties quelconques X et Y de E, choisies aléatoirement et indépendamment (absence de lien a priori entre ces deux parties) et de mêmes cardinaux respectifs que A et B. Soit $\overline{Y}$ et $\overline{B}$ les complémentaires respectifs de Y et de B dans E de même cardinal $n_{\bar{b}} = n\text{-}n_b$.



Les parties hachurées correspondent aux contre-exemples de l'implication a => b.

Figure 1

Nous dirons alors :

DEFINITION 1: a=>b est *admissible au niveau de confiance1-* α si et seulement si

$$\Pr\left[\mathrm{card}\left(X \cap \overline{Y}\right) \leq \mathrm{card}\left(A \cap \overline{B}\right)\right] \leq \alpha$$

Il est établi (Lerman I.C., 1981) que, pour un certain processus de tirage :

la variable aléatoire $\mathrm{Card}(X \cap \overline{Y})$ suit la loi de Poisson de paramètre $\frac{n_a n_{\bar{b}}}{n}$.

Dans le cas où $n_{\bar{b}} \neq 0$, nous réduisons et centrons cette variable de Poisson en la variable :

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

Dans la réalisation expérimentale, la valeur observée de $Q(a,\bar{b})$ est $q(a,\bar{b})$

DEFINITION 2: $q(a,\bar{b}) = \dfrac{n_{a\wedge\bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}}$ est appelé <u>indice d'implication,</u> nombre que nous retenons comme indicateur de la non-implication de a sur b.

Dans les cas légitimant convenablement l'approximation (par exemple, $\frac{n_a n_{\bar{b}}}{n} \geq 3$), la variable $Q(a,\bar{b})$ suit approximativement la loi normale centrée réduite. <u>L'intensité</u>

d'implication, qualité de l'admissibilité de a => b, pour $n_a \leq n_b$ et $n_b \neq n$, est alors définie à partir de l'indice $q(a,\overline{b})$ par :

DEFINITION 3 : Dans le cas où $n_b \neq n$, l'intensité d'implication de a sur b est :

$$\varphi(a,b)=1-\Pr\left[Q(a,\overline{b})\leq q(a,\overline{b})\right]=\frac{1}{\sqrt{2\pi}}\int_{q(a,\overline{b})}^{\infty} e^{-\frac{t^2}{2}}\,dt$$

Par suite, la définition de l'implication statistique devient :
DEFINITION 4 : L'implication a => b est admissible au niveau de confiance 1-α si et seulement si : $\varphi(a,b) \geq 1-\alpha$

Rappelons que cette modélisation de la quasi-implication mesure l'étonnement de constater la petitesse des contre-exemples en regard du nombre surprenant des instances de l'implication. Par conséquent, si la règle est triviale, comme dans le cas où B est très grand ou coïncide avec E, cet étonnement devient petit. Nous démontrons (Gras R., 1996) d'ailleurs que cette trivialité se traduit par une intensité d'implication très faible, voire nulle :

Si, n_a étant fixé et A étant inclus dans B, n_b tend vers n (B "croît" vers E), alors $\varphi(a,b)$ tend vers 0.

Remarque 1: D'autres modélisations, autres que celle de Poisson, sont possibles. Citons :
* *une modélisation binomiale* : considérant les variables duales $\mathrm{card}(A\cap\overline{Y})$ et $\mathrm{card}(X\cap\overline{B})$, où X et Y sont des parties choisies de façon indépendante dans E et respectant les propriétés cardinales respectives de A et B, tout élément de E, par exemple, a la probabilité $\frac{n_a}{n}\frac{n_{\overline{b}}}{n}$ d'appartenir à $A\cap\overline{Y}$. Par suite :

$$\Pr\left[\mathrm{card}(A\cap\overline{Y})=k\right]=C_n^k\left(\frac{n_a n_{\overline{b}}}{n^2}\right)^k\left(1-\frac{n_a n_{\overline{b}}}{n^2}\right)^{n-k}=\Pr\left[\mathrm{card}(X\cap\overline{B})=k\right]$$

* *une modélisation hypergéométrique* : on peut le voir rapidement en considérant encore les variables aléatoires $\mathrm{card}(A\cap\overline{Y})$ et $\mathrm{card}(X\cap\overline{B})$ où X et Y possèdent les mêmes propriétés cardinales respectives que A et B. On a, en effet :

$$\Pr[\mathrm{card}(A\cap\overline{Y})=k]=\frac{C_{n_a}^k C_{n-na}^{n-nb-k}}{C_n^{n-nb}}=\frac{n_a!\,n_{\overline{a}}!\,n_b!\,n_{\overline{b}}!}{k!\,n!\,(n_a-k)!\,(n_{\overline{b}}-k)!\,(n_{\overline{b}}-n_a-k)!}$$

$$=\frac{C_{n-n_b}^k C_{n_b}^{n_a-k}}{C_n^{n_a}}=\Pr[\mathrm{card}(X\cap\overline{B})=k]$$

Si la modélisation binomiale reste compatible avec la sémantique de l'implication, relation binaire non symétrique, il n'en est plus de même pour la modélisation hypergéométrique. Aussi, nous ne retiendrons que le modèle de Poisson et le modèle binomial.

Remarque 2 : La quasi-implication, d'indice $q(a,\bar{b})$ non symétrique, ne coïncide pas avec le coefficient de corrélation $\rho(a,b)$ qui est symétrique et qui rend compte de la liaison entre les variables a et b. En effet, nous montrons que si $q(a,\bar{b}) \neq 0$ alors $\frac{\rho(a,b)}{q(a,\bar{b})} = -\sqrt{\frac{n}{n_b n_{\bar{a}}}}$

Remarque 3 : Nous pouvons définir des conjonctions de variables du type "a et b" ou "(a et b) ou c..." afin de modéliser les phénomènes qui relèvent de concepts comme il est fait en apprentissage ou en intelligence artificielle. Les calculs associés restent compatibles avec la logique des propositions reliées par des connecteurs.

Remarque 4 : Contrairement à l'indice de Loevinger (Loevinger J. 1947) et à la probabilité conditionnelle (Pr[B/A]) et tous ses dérivés, l'intensité d'implication varie avec la dilatation des ensembles E, A et B, résiste aux bruits, c'est-à-dire en particulier au voisinage de 0 pour $n_{a\wedge\bar{b}}$, ce qui ne peut que rendre statistiquement crédible la relation que nous voulons modéliser.

CAS DES VARIABLES MODALES ET FREQUENTIELLES

Dans la suite de nos travaux, nous étendons la notion d'implication statistique à des variables autres que binaires. C'est le cas des variables modales qui sont associées à des phénomènes où les valeurs a(x) sont des nombres de l'intervalle [0,1] et qui décrivent des degrés d'appartenance ou de satisfaction comme en logique floue, par exemple, "peut-être", "un peu", "quelquefois", etc. C'est aussi le cas des variables fréquentielles qui sont associées à des phénomènes où les valeurs de a(x) sont des réels positifs quelconques .

J.B. Lagrange (Lagrange J.B. 1998) a démontré que, dans le cas modal,

- si a(x) et $\bar{b}(x)$ sont les valeurs prises en x par les variables modales a et $\bar{b}$, avec $\bar{b}(x)=1-b(x)$
- si s_a^2 et $s_{\bar{b}}^2$ sont les variances empiriques des variables a et $\bar{b}$

alors l'indice d'implication, qu'il dénomme *indice de propension*, devient :

DEFINITION 5 : $q(a,\bar{b}) = \dfrac{\sum_{x\in E} a(x)\bar{b}(x) - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{(n^2 s_a^2 + n_a^2)((n^2 s_{\bar{b}}^2 + n_{\bar{b}}^2)}{n^3}}}$ est l'indice de propension de variables modales

Il prouve également que cet indice coïncide avec l'indice défini précédemment dans le cas binaire si le nombre de modalités de a et de b est justement 2, car dans ce cas : $n^2 s_a^2 + n_a^2 = n n_a$, $n^2 s_{\bar{b}}^2 + n_{\bar{b}}^2 = n n_{\bar{b}}$ et $\sum_{x\in E} a(x)\bar{b}(x) = n_{a\wedge\bar{b}}$.

Cette solution apportée au cas modal est aussi applicable au cas des *variables fréquentielles,* voire *des variables numériques positives,* à condition d'avoir normalisé les valeurs observées sur les variables, telles que a et b, la normalisation dans [0,1] étant faite à partir du maximum de la valeur prise respectivement par a et b sur l'ensemble E.

CAS DES VARIABLES-INTERVALLES

Situation fondamentale

Deux variables réelles a et b prennent un certain nombre de valeurs sur 2 intervalles finis $[a_1, a_2]$ et $[b_1, b_2]$. Soit A (resp. B) l'ensemble des valeurs de a (resp. b) observées sur $[a_1, a_2]$ (resp. $[b_1, b_2]$. Par exemple, a représente les poids d'un ensemble de n sujets et b les tailles de ces mêmes sujets.

Deux problèmes se posent :

1° peut-on définir des sous-intervalles adjacents de $[a_1, a_2]$ (resp. $[b_1, b_2]$.) afin que la partition la plus fine obtenue respecte au mieux la distribution des valeurs observées dans $[a_1, a_2]$ (resp. $[b_1, b_2]$.) ?

2° peut-on trouver les partitions respectives de $[a_1, a_2]$ et $[b_1, b_2]$ constituées de réunions des sous-intervalles adjacents précédents, partitions qui maximisent l'intensité d'implication moyenne des sous-intervalles de l'un sur des sous-intervalles sur l'autre appartenant à ces partitions ?

Nous allons tenter de répondre à ces deux questions dans le cadre de notre problématique en faisant choix des critères à optimiser pour satisfaire l'optimalité attendue dans chaque cas. A la première question, de nombreuses solutions ont été apportées dans d'autres cadres (par exemple, par Lahanier-Reuter D. 1998).

Premier problème

On va s'intéresser à l'intervalle $[a_1, a_2]$ en le supposant muni d'une partition initiale triviale de sous-intervalles de même longueur, mais pas nécessairement de même distribution des fréquences observées sur ces sous-intervalles.

Notons $P_0 = \{A_{01}, A_{02},..., A_{0p}\}$, cette partition en p sous-intervalles. On cherche à obtenir une partition P_q^* de $[a_1, a_2]$ en p sous-intervalles $A_{q1}, A_{q2},..., A_{qp}$ de telle façon qu'au sein de chaque sous-intervalle on ait une bonne homogénéité statistique (faible inertie intra-classe) et que ces sous-intervalles présentent une bonne hétérogénéité mutuelle (forte inertie inter-classe). On sait que si l'un des critères est vérifié l'autre l'est nécessairement. (théorème de Koenig-Huyghens).

Pour ce faire, on adoptera une méthode directement inspirée de la méthode des nuées dynamiques conçue par Edwin Diday (Diday E., 1972) et adaptée à la situation présente. Pour cela, on cherche à minimiser une certaine fonction W, définie sur l'ensemble G des points réels de $[a_1, a_2]$ et l'ensemble des partitions P de $[a_1, a_2]$ en p sous-intervalles $A_{i,}$, de la façon suivante :

$$W(G, P) = \sum_{i=1}^{p} D(G_i, A_i) \text{ avec } D(G_i, A_i) = \sum_{x \in A_i} (G_i - x)^2 \text{ pour tout i= 1,2,...,p.}$$

Ainsi, si G est le barycentre des valeurs observées dans A, si G_i est le barycentre des valeurs observées dans $A_{i,}$ alors W(G, P) est l'inertie intra-classe de A et $D(G_i, A_i)$ est l'inertie de A_i .

1ère étape

On part de la partition $P_0 = \{A_{01}, A_{02},..., A_{0p}\}$. On choisit ce que E.Diday appelle les *noyaux*, en nombre p, dans $[a_1, a_2]$. Ces noyaux sont choisis confondus avec les barycentres respectifs, tels que G_{1i} , des sous-intervalles A_{0i} des valeurs qui y sont observées. Soit G_1 leur barycentre.

On cherche alors la partition $P_1 = \{A_{11}, A_{12},..., A_{1p}\}$ telle que pour tout i :

$$A_{1i} = \left\{ x \in A \,/\, \forall j \quad (G_{1i} - x)^2 \le (G_{1j} - x)^2 \right\}$$

Cela revient à constituer A_{1i} à l'aide des points qui sont les plus proches du barycentre de A_{0i}. Cela revient aussi à reporter dans les sous-intervalles voisins les points qui sont plus proches de leurs propres barycentres respectifs. En cas d'égalité, on affecte x au sous-intervalle de plus petit indice.

On démontre alors que $W(G_1, P_1) \le W(G_1, P_0)$ (voir plus loin)

2ème étape

On dispose de la partition : $P_1 = \{A_{11}, A_{12},..., A_{1p}\}$

On choisit p noyaux G_{2i} qui minimisent respectivement les quantités $D(y, A_{1i})$, c'est-à-dire : pour tout y , valeur observée dans A, $D(G_{2i}, A_{1i}) \le D(y, A_{1i}) = \sum_{x \in A_{1i}} (y - x)^2$, qui est l'inertie de A_{1i} autour de y à un coefficient près. Par conséquent, le noyau G_{2i} est le barycentre de A_{1i} .

On obtient donc une nouvelle suite de noyaux : $\{G_{21}, G_{22}, ..., G_{2p}\}$ dont G_2 est le barycentre et l'on est ramené au procédé de l'étape précédente.

On détermine en effet la partition $P_2 = \{A_{21}, A_{22},..., A_{2p}\}$ telle que :

$$A_{2i} = \left\{ x / \forall j (G_{2i} - x)_2 \le (G_{2j} - x)_2 \right\}$$

On obtient de même : $W(G_2, P_2) \le W(G_2, P_1)$

kème étape

On dispose de la partition : $P_{k-1} = \{A_{(k-1)1}, A_{(k-1)2},..., A_{(k-1)p}\}$

On choisit p noyaux G_{ki} qui minimisent respectivement les quantités $D(y, A_{(k-1)i})$, c'est-à-dire tels que:

pour tout y , valeur observée dans A, $D(G_{2i}, A_{(k-1)i}) \le D(y, A_{(k-1)i}) = \sum_{x \in A(k-1)i} (y - x)^2$

Le noyau G_{ki} est le barycentre de $A_{(k-1)i}$. On obtient ainsi une nouvelle suite de noyaux : $\{G_{k1}, G_{k2}, ..., G_{kp}\}$, dont G_k est le barycentre.

On détermine la partition $P_k = \{A_{k1}, A_{k2},..., A_{kp}\}$ telle que :

$$A_{ki} = \left\{ x \,/\, \forall j \quad (G_{ki} - x)^2 \le (G_{kj} - x)^2 \right\}$$

On obtient encore : $W(G_k, P_k) \le W(G_k, P_{k-1})$.

Le processus est fini. En effet, la suite $W(G_k, P_k)$ est non croissante. Si elle devient stationnaire, elle a convergé vers son minimum, les noyaux et les partitions étant inchangées du fait que la somme $\sum_{i=1}^{p} D(G_i, A_i)$ est positive et constituée d'éléments non croissants.

Exemple

Supposons que nous ayons observé les poids suivants de 17 individus :

$$A = \{ \underbrace{54, 55, 55, 57, 57, 58}_{A_{01}}, \underbrace{60, 66, 67, 67, 68}_{A_{02}}, \underbrace{70, 71, 74, 78, 79, 79}_{A_{03}} \}$$

1ère étape :

On choisit

* la partition $P_0 = \{[50, 60[, [60, 70[, [70, 80]\}$

* et les noyaux : $G_{11} = 56$; $G_{12} = 65{,}25$; $G_{13} = 75{,}17$

On calcule les valeurs $(G_{11} - x)^2$ pour tous les x de A, puis de la même façon tous les $(G_{12} - x)^2$ et les $(G_{13}- x)^2$. On associe à G_{11} les valeurs de x qui minimisent les expressions $(G_{11}- x)^2$, soit les nombres : 54, 55, 55, 57, 57, 58, 60.

Donc $A_{11} = \{54, 55, 55, 57, 57, 58, 60\}$.

On obtient de même : $A_{12} = \{66, 67, 67, 68, 70\}$, puis $A_{13} = \{71, 74, 78, 79, 79\}$

2ème étape

On choisit pour noyaux les barycentres respectifs des A_{1i} , pour i=1,2,3.

Soit $G_{21} = 56{,}57$; $G_{22} = 67{,}75$; $G_{23} = 76{,}2$

On obtient sans peine : $A_{21} = \{54, 55, 55, 57, 57, 58, 60\}$.

$A_{22} = \{66, 67, 67, 68, 70, 71\}$, puis $A_{23} = \{74, 78, 79, 79\}$

3ème étape

On choisit pour noyaux les barycentres respectifs des A_{2i} , pour i=1,2,3.

Soit $G_{31} = 56{,}57$; $G_{32} = 68{,}4$ $G_{33} = 77{,}5$

On obtient : $A_{31} = A_{21}$; $A_{32} = A_{22}$; $A_{33} = A_{23}$

Le processus a donc convergé et l'algorithme s'arrête sur la dernière partition.

Remarque

Afin d'évaluer la qualité de la partition obtenue, on calcule le rapport :

$$\tau = \frac{\text{Inertie}_{\text{inter-classe}}}{\text{Inertie}_{\text{totale}}}$$

Or l'inertie inter-classe de A pour la partition $P_3 = \{A_{31}, A_{32}, A_{33}\}$ est :

$\sum_{i=1,2,3} m(G_{3i})\left(G_{3i} - G\right)^2 = 47{,}433$ où $m(G_{3i})$ est l'effectif des x dans A_{3i} et G est le barycentre de A.

L'inertie totale est $\sum_{x \in A} (G - x)^2 = 49{,}82$, soit le taux τ excellent de 0,95.

Deuxième problème[2]

On suppose maintenant que les intervalles $[a_1, a_2]$ et $[b_1, b_2]$ sont munis de partitions optimales P et Q, respectivement, au sens des nuées dynamiques. Soit p et q les nombres respectifs de sous-intervalles composant P et Q. A partir de ces deux partitions, il est possible d'engendrer 2^{p-1} et 2^{q-1} partitions obtenues par réunions itérées de sous-intervalles adjacents respectivement de P et de Q [3].

---

[2] D. Lahanier-Reuter en propose une autre approche dans sa thèse (cf infra)

[3] Il suffit de considérer l'arborescence dont A_1 est la racine, puis de le réunir ou non à A_2 qui lui-même sera ou non réuni à A_3, etc. Il y a donc 2^{p-1} branches dans cette arborescence.

On calcule les intensités d'implication respectives de chaque sous-intervalle réuni ou non à un autre de la première partition sur chaque sous-intervalle réuni ou non à un autre de la seconde, puis les valeurs des intensités des implications réciproques.

Il y a donc au total $2.2^{p-1}.2^{q-1}$ familles d'intensités d'implication, chacune d'entre elles nécessitant le calcul de tous les éléments d'une partition de $[a_1, a_2]$ sur tous les éléments d'une des partitions de $[b_1, b_2]$et réciproquement.

On choisit comme *critère d'optimalité* la moyenne géométrique des intensités d'implication, moyenne associée à chaque couple de partitions d'éléments réunis ou non définies inductivement. On note les deux maxima obtenus (implication directe et sa réciproque) et on retient les deux partitions associées en déclarant que l'implication de la variable-intervalle a sur la variable-intervalle b est optimale lorsque l'intervalle $[a_1, a_2]$ admet la partition correspondant au premier maximum et que l'implication réciproque optimale est satisfaite pour la partition de $[b_1, b_2]$ correspondant au deuxième maximum.

Remarque

1° Il n'existe pas de relation d'ordre entre $\varphi(a,b)$ et $\varphi(a,b\vee c)$ connaissant $\varphi(a,b)$ et $\varphi(a,c)$.

En effet, on peut avoir $\varphi(a,b) < \varphi(a,b\vee c)$ et $\varphi(a,c) < \varphi(a,b\vee c)$.

exemple1 : $n = 100$; $n_a = 16$; $n_b = 35$; $n_c = 30$; $n_{a\wedge\bar{b}} = 10$; $n_{a\wedge\bar{c}} = 8$; $n_{a\wedge(\bar{b\vee c})} = 2$

Mais on peut également avoir : $\varphi(a,b) > \varphi(a,b\vee c)$ et $\varphi(a,c) > \varphi(a,b\vee c)$

exemple 2 : $n = 100$; $n_a = 30$; $n_b = 50$; $n_c = 49$; $n_{a\wedge\bar{b}} = 15$; $n_{a\wedge\bar{c}} = 16$; $n_{a\wedge(\overline{b\vee c})} = 1$

2° De même, il n'existe pas a priori de relation entre $\varphi(a,c)$, $\varphi(b,c)$ et $\varphi(a\vee b,c)$

En effet, on peut avoir, $\varphi(a,c) < \varphi(a\vee b,c)$ et $\varphi(b,c) < \varphi(a\vee b,c)$

exemple 3 : $n = 100$; $n_a = 20 = n_b$; $n_c = 35$; $n_{a\wedge\bar{c}} = 16 = n_{b\wedge\bar{c}}$; $n_{a\vee b,\bar{c}} = 10$

Mais on peut avoir également : $\varphi(a,c) > \varphi(a\vee b,c)$ et $\varphi(b,c) > \varphi(a\vee b,c)$

exemple 4 : $n = 100$; $n_a = 20 = n_b$; $n_c = 48$; $n_{a\wedge\bar{c}} = 11 = n_{b\wedge\bar{c}}$; $n_{a\vee b,\bar{c}} = 22$

Ceci montre que l'algorithme de recherche de l'optimum des implications au cours des réunions successives doit fonctionner "jusqu'au bout", c'est-à-dire lorsque toutes les réunions ont été produites et estimées quant à leur puissance implicative.

Décroissance de la fonction W

Quelques notations

Soit A l'ensemble des valeurs observées dans l'intervalle $[a_1, a_2]$ et $L= \{N_1, N_2, ...,N_p\}$ un ensemble de parties de A. N_i est appelé $i^{ème}$ *noyau* de L. Ces noyaux sont choisis de telle façon que card N_i soit le même pour tout i.

Soit $P = \{A_1, A_2,..., A_p\}$ une partition de A en p classes.

On pose $W(L,P) = \sum_{i=1}^{p} D(N_i, A_i) = \sum_{i=1}^{i=p} \sum_{x\in A_i, y\in N_i} d(x,y)$ [4]

$D(N_i, A_i)$ est une sorte de mesure de dissemblance entre le noyau N_i et la classe A_i.

Le problème des nuées dynamiques vise à minimiser W(L,P) par la construction d'un ensemble convenable de p noyaux dans L^* et d'une partition P^* en p classes.

---

[4] d(x,y) peut être égal à $(x-y)^2$ comme nous l'avons choisi dans le premier problème

Algorithme des nuées dynamiques

Les noyaux $L_0 = \{N_{01}, N_{02}, ...,N_{0p}\}$ sont donnés (ou choisis arbitrairement).

On définit la partition $P_1 = \{A_{11}, A_{12},..., A_{1p}\}$ qui s'en déduit par :

$$A_{1i} = \{x \in A / \forall j \quad d(N_{0i}, x) \leq d(N_{0j}, x)\}.$$

En cas d'égalité, on affecte x à la classe d'indice plus petit.

On pose $P_1 = f(L_0)$.

On construit alors les noyaux de $L_1 = \{N_{11}, N_{12}, ...,N_{1p}\}$ par le procédé :

$$N_{1k} = \{x \in A / d(x, A_{1k}) = \inf_j d(x, A_{1j})\}$$

On note $L_1 = g(P_1)$ et on itère l'algorithme.

PROPOSITION : W(L,P) décroît à chaque itération, i.e. :

a) $\forall P$, $W(L, f(L)) \leq W(L, P)$

b) $\forall L$, $W(g(P), P) \leq W(L, P)$

Preuve :

a) Soit la partition quelconque au cours de l'algorithme $P = \{A_1, A_2,..., A_p\}$ et soit $f(L)= Q = \{B_1, B_2, ...,B_p\}$ la partition obtenue à l'aide des noyaux $L = \{N_1, N_2, ...,N_p\}$ définis à partir de P.

$$\text{Alors } W(L,P) = \sum_{i=1}^{p} D(N_i, A_i) = \sum_{i=1}^{i=p} \sum_{x \in A_i} d(N_i, x) = \sum_{j=1}^{j=p} \sum_{x \in Nj} d(x, A_j)$$

$$\text{et } W(L,Q) = \sum_{j=1}^{j=p} \sum_{y \in Bj} d(N_j, y)$$

Soit $x \in A$. Pour tout i et tout x tel que $x \in A_i \cap B_i$, alors x a la même contribution aux sommes W(L,P) et W(L,Q). Par contre, si $x \in A_i$ et $x \in B_j$, alors $d(N_j, x) \leq d(N_i, x)$

car par construction : $B_j = \{x \in A / \forall i \ d(N_j, x) \leq d(N_i, x)\}$

Par suite, pour tout $x \in A$, les contributions de x à W(L, Q) sont inférieures ou égales à celles de W(L, P), soit encore $\forall P$, $W(L, f(L)) \leq W(L, P)$.

b) Soit $g(P) = \{O_1, O_2, ...,O_p\}$ un ensemble de p noyaux obtenus à la suite de la partition $P = \{A_1, A_2,..., A_p\}$, elle-même dérivée de $L = \{N_1, N_2, ...,N_p\}$ et définis ainsi:

$$\forall k, O_k = \{x \in A / d(x, A_k) = \inf_j d(x, A_j)\}.$$

$$\text{Par suite, } \forall j, \sum_{x \in Oj} d(x, A_j) \leq \sum_{x \in Nj} d(x, A_j) \text{ d'où } \forall L, W(g(P), P) \leq W(L, P)$$

Remarque : En suivant la problématique de E.Diday et ses collaborateurs, si les valeurs prises selon les sujets par les variables a et b sont de nature symbolique, en l'occurrence des intervalles de R^+, il est possible d'étendre les algorithmes ci-dessus. Par exemple, à la variable a sont associés des intervalles de poids et à la variable b des intervalles de tailles, intervalles dus à une imprécision des mesures. Effectuant la réunion des intervalles I_x et J_x décrits par les sujets x de E selon respectivement chacune des variables a et b, on obtient deux intervalles I et J recouvrant toutes les valeurs possibles de a et de b. Sur chacun d'eux on peut définir une partition en un certain nombre d'intervalles respectant comme plus haut un certain critère d'optimalité. Pour cela, les intersections des intervalles tels que I_x et J_x avec

ces partitions seront munies d'une distribution prenant en compte les étendues des parties communes. Cette distribution peut être uniforme ou d'un autre type discret ou continu. Mais ainsi, nous sommes ramenés à la recherche de règles entre deux ensembles de variables-intervalles qui prennent comme précédemment leurs valeurs sur [0,1] et sur lesquelles on pourra chercher les implications optimales.

L'IMPLICATION-INCLUSION

Deux raisons nous ont conduits à améliorer le modèle réalisé par l'intensité d'implication:

- lorsque la taille des échantillons traités, et en particulier celui de E, croît (de l'ordre du millier et plus), l'intensité $\varphi(a,b)$ a tendance à ne plus être suffisamment discriminante car ses valeurs peuvent être très voisines de 1, alors que l'inclusion dont elle cherche à modéliser la qualité, est loin d'être satisfaite (phénomène signalé dans (Bodin A. 1997) qui traite des grandes populations d'élèves à travers des enquêtes internationales) ;
- le modèle de la quasi-implication précédent retient essentiellement la mesure de la force de la règle a => b. Or, la prise en compte d'une puissance concomitante de non b => non a (contraposée de l'implication) est indispensable pour renforcer l'affirmation d'une bonne qualité de la relation quasi-implicative de a sur b[5]. En même temps, elle pourrait permettre de corriger la difficulté évoquée ci-dessus (si A et B sont petits par rapport à E, leurs complémentaires seront importants et réciproquement).

La solution que nous apportons utilise à la fois l'intensité d'implication et un autre indice qui rend compte de la dissymétrie entre les situations S_1 = (a et b) et S'_1 (a et non b) (resp. S_2 = non a et non b) et S'_2 = (a et non b)) en faveur de la première nommée. La faiblesse relative des instances qui contredisent la règle et sa contraposée est ainsi fondamentale. D'ailleurs, le nombre de contre-exemples $n_{a \wedge \bar{b}}$ à a=>b est celui à la contraposée. C'est donc au concept d'entropie de Shannon que nous faisons référence :

$$H(b/a) = -\frac{n_{a\wedge b}}{n_a}\log_2\frac{n_{a\wedge b}}{n_a} - \frac{n_{a\wedge \bar{b}}}{n_a}\log_2\frac{n_{a\wedge \bar{b}}}{n_a} ,$$

entropie conditionnelle relative aux cases (a et b) et (a et non b) lorsque a est réalisée

$$H(\bar{a}/\bar{b}) = -\frac{n_{a\wedge \bar{b}}}{n_{\bar{b}}}\log_2\frac{n_{a\wedge \bar{b}}}{n_{\bar{b}}} - \frac{n_{\bar{a}\wedge \bar{b}}}{n_{\bar{b}}}\log_2\frac{n_{\bar{a}\wedge \bar{b}}}{n_{\bar{b}}}$$

entropie conditionnelle relative aux cases (non a et non b) et (a et non b) lorsque non b est réalisée

Ces entropies, à valeurs dans [0,1], devraient donc être simultanément faibles et donc les dissymétries entre les situations S_1 et S'_1 (resp. S_2 et S'_2) devraient être simultanément fortes si l'on souhaite disposer d'un bon critère d'inclusion de A dans B. En effet, les entropies représentent l'incertitude moyenne des expériences qui consistent à observer si b est réalisé (resp. si non a est réalisé) lorsque l'on a observé a (resp. non b). Le complément à 1 de cette incertitude représente donc l'information moyenne recueillie par la réalisation de ces expériences. Plus cette information est importante, plus forte est la garantie de la qualité de l'implication et de sa contraposée. Nous devons maintenant adapter ce critère numérique entropique au modèle attendu dans les différentes situations cardinales.

Pour que le modèle ait la signification attendue, il doit satisfaire, selon nous, les contraintes suivantes :

---

[5] Ce phénomène est signalé par Y.Kodratoff dans son article publié dans ces Actes.

1° il devra intégrer les valeurs de l'entropie et, pour les contraster, par exemple, intégrer ces valeurs au carré,

2° comme ce carré varie de 0 à 1, afin de dénoter le déséquilibre et donc l'inclusion, afin de s'opposer à l'entropie, la valeur retenue sera le complément à 1 de son carré tant que le nombre de contre-exemples sera inférieur à la moitié des observations de a (resp. de non b). Au delà de ces valeurs, les implications n'ayant plus de sens inclusif, on affectera au critère la valeur 0,

3° afin de prendre en compte les deux informations propres à a=>b et non b=> non a, le produit rendra compte de la qualité simultanée des valeurs retenues. Le produit a la propriété de s'annuler dès que l'un de ses termes s'annule, i.e. dès que cette qualité s'efface,

4° enfin, le produit ayant une dimension 4 par rapport à l'entropie, sa racine quatrième sera de la même dimension.

Posons $\alpha = \frac{n_a}{n}$, la fréquence de a et $\bar{\beta} = \frac{n_{\bar{b}}}{n}$, la fréquence de non b. Notons, en fonction de la fréquence $t = \frac{n_{a \wedge \bar{b}}}{n}$ de contre-exemples, les deux termes significatifs des qualités respectives de l'implication et sa contraposée :

$$h_1(t) = H(b/a) = -(1-\frac{t}{\alpha})\log_2(1-\frac{t}{\alpha}) - \frac{t}{\alpha}\log_2\frac{t}{\alpha} \text{ si } t \in [0, \frac{\alpha}{2}[\text{ et } h_1(t) = 1 \text{ si } t \in [\frac{\alpha}{2}, a]$$

$$h_2(t) = H(\bar{a}/\bar{b}) = -(1-\frac{t}{\bar{\beta}})\log_2(1-\frac{t}{\bar{\beta}}) - \frac{t}{\bar{\beta}}\log_2\frac{t}{\bar{\beta}} \text{ si } t \in [0, \frac{\bar{\beta}}{2}[\text{ et } h_2(t) = 1 \text{ si } t \in [\frac{\bar{\beta}}{2}, \bar{\beta}]$$

D'où la définition permettant de déterminer le critère entropique :

DEFINITION 6 : L' indice d'inclusion de A, support de a, dans B, support de b, est le nombre :

$$i(a,b) = [(1-h_1^2(t))(1-h_2^2(t))]^{1/4}$$

qui intègre l'information délivrée par la réalisation d'un faible nombre de contre-exemples, d'une part à la règle a => b et, d'autre part, à la règle non b => non a

L'intensité d'implication-inclusion *(ou intensité entropique)* est le nombre :

$$\psi(a,b) = (i(a,b).\varphi(a,b))^{\frac{1}{2}}$$

qui intègre à la fois l'étonnement statistique et la qualité inclusive.

La fonction ψ suivant la variable t admet une représentation qui a la forme indiquée par la figure 2, pour n_a et n_b fixés. On remarquera la différence de comportement de la fonction par rapport à la probabilité conditionnelle P(B/A), indice fondamental des autres modélisations de la mesure des règles, par exemple chez Agrawal et son école. Outre son caractère linéaire, donc peu nuancé, cette probabilité conduit à une mesure qui décroît trop vite dès les premiers contre-exemples et résiste ensuite trop longuement lorsque ceux-ci deviennent importants.

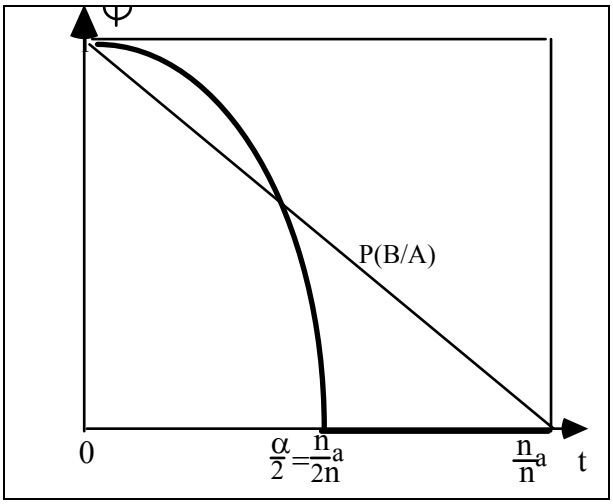


figure 2

On constate que cette représentation de fonction continue de t traduit les propriétés attendues du critère d'inclusion :

* "réaction" lente aux premiers contre-exemples (résistance au bruit),
* "accélération" du rejet de l'inclusion au voisinage de l'équilibre soit $\frac{n_a}{2n}$,
* rejet au-delà de $\frac{n_a}{2n}$ ce que n'assurait pas l'intensité d'implication.

Exemple 1

	b	$\bar{b}$	marge
a	200	400	600
$\bar{a}$	600	2800	3400
marge	800	3200	4000

L'intensité d'implication est 0,9999 ($q(a,\bar{b}) = -3,65$)
Les valeurs entropiques de l'expérience sont $h_1 = 0 = h_2$
La valeur du coefficient modérateur est donc : $i(a,b) = 0$
Par suite $\psi(a,b) = 0$ alors que P(B/A) = 0,33333

Ainsi, les fonctions "entropiques" "modérent" l'intensité d'implication dans ce cas où justement l'inclusion est médiocre.

Exemple 2

	b	$\bar{b}$	marge
a	400	200	600
$\bar{a}$	1000	2400	3400
marge	1400	2600	4000

L'intensité d'implication est 1 ($q(a,\bar{b}) = -8,43$)
Les valeurs entropiques de l'expérience sont $h_1 = 0,918$ $h_2 = 0,391$
La valeur du coefficient modérateur est donc : $i(a,b) = 0,6035$
Par suite $\psi(a,b) = 0,777$ alors que P(B/A) = 0,66666

Remarque

La correspondance entre $\varphi(a,b)$ et $\psi(a,b)$ n'est pas monotone comme le montre le deuxième exemple suivant :

	b	$\bar{b}$	marge
a	40	20	60
$\bar{a}$	60	280	340
marge	100	300	400

L'intensité d'implication est inférieure à la précédente car : $q(a,\bar{b}) = -6,47$. Les valeurs entropiques sont : $h_1 = 0,918$, $h_2 = 0,353$
La valeur du coefficient modérateur est : $i(a,b) = 0,608$
Donc $\psi(a,b) = 0,78$ alors que P(B/A) = 0,66666

Ainsi, alors que $\varphi(a,b)$ a décru du 1er au 2ème exemples, i(a,b) a crû de même que $\psi(a,b)$. En revanche, la situation contraire est la plus fréquente. Notons que, dans les deux cas, la probabilité conditionnelle ne change pas.

GRAPHE D'IMPLICATION

La relation définie par l'implication statistique, si elle est réflexive et non symétrique, n'est pas transitive bien évidemment. Or nous voulons qu'elle modélise la relation d'ordre partiel

entre deux variables (les réussites dans notre exemple initial). Par convention, si a => b et si b => c, nous accepterons la fermeture transitive a => c seulement si ψ (a,c) $\geq$ 0,5, c'est-à-dire si la relation implicative de a sur c est meilleure que la neutralité.

Par exemple, supposons qu'entre les 7 variables *a, b, c, d, e* et *f* existent, au seuil supérieur à 0,5, les relations suivantes: *e => c, a, f, b* ; *c => a, f* ;*b=> a, f* ; *g => d, f* ; *a=> f*. On pourra alors traduire cet ensemble de relations par le graphe suivant[6] :

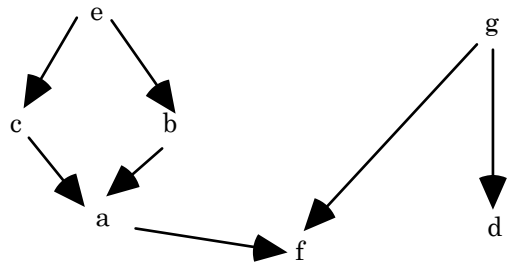


figure 3

Notons que ce graphe n'est pas un treillis puisque, par exemple la variable a n'implique pas la variable (a ou non a) dont le support est E. A fortiori, ce ne peut être un treillis de Galois.

Remarque :

A l'instar de la définition 6 d'un indice d'implication-inclusion, A. Bodin (Bodin A. 1997) propose un test d'hypothèse afin de valider ou réfuter une hypothèse d'inclusion de l'ensemble A dans l'ensemble B. Ce test est intégré au menu du logiciel C.H.I.C. et conduit au graphe dit d'implication-inclusion établi sur la base d'un seuil donné par l'utilisateur. On peut noter quelques divergences entre ce graphe et le graphe classique d'implication, en particulier, comme l'avait dénoncé A.Bodin, lorsque les populations en jeu deviennent importantes. Par contre, les divergences sont plus faibles lorsque le graphe d'implication est construit à l'aide de l'"intensité entropique"que nous avons construite dans le paragraphe précédent.

IMPLICATION ENTRE REGLES ET META-REGLES

Une implication entre classes de variables ne prend véritablement son sens qu'à condition qu'à l'intérieur de chaque classe de variables dont on examine la relation avec d'autres, existe une certaine "**cohésion**" entre les variables qui la constituent. On souhaite ainsi que le "flux" implicatif d'une classe ***A*** sur une classe ***B*** soit nourri d'un "flux" interne à ***A*** et alimente un "flux" interne à ***B*** (ce mot *flux* est choisi pour sa connotation métaphorique hydraulique ou thermodynamique). Pour cela, le concept d'entropie H permettant de rendre compte du désordre entre des variables, nous définissons la cohésion entre deux variables par :

DEFINITION 7 : La cohésion de la classe (a,b) est le nombre c(a,b) tel que :

. si $p = \max(\psi(a,b), \psi(b,a))$ et $H = -p\log_2 p-(1-p)\log_2(1-p)$, alors $c(a,b) = \sqrt{1 - H^2}$

. si p = 1, alors c(a,b) = 1

[6]Les traitements automatiques des calculs et des graphiques sont exécutés à l'aide du logiciel C.H.I.C. (Classification Hiérarchique Implicative et Cohésitive) disponible sous Windows 95. Ce logiciel, à partir d'une première version établie par R.Gras, révisée sous Pascal par S.Ag Almouloud, est maintenant développé par R.Couturier et constamment étendu par lui aux nouveaux concepts et nouveaux algorithmes.

. si $p \leq 0{,}5$, alors $c(a,b) = 0$

Intuitivement, la cohésion mesure le déséquilibre des occurrences des événements $a \wedge b$ et $a \wedge \bar{b}$ en faveur du premier.

DEFINITION 8 : La cohésion de la classe de variables $\boldsymbol{A} = (a_1,....,a_r)$ est alors définie par extension :

$$C(\underline{\boldsymbol{A}}) = \left[\prod_{i\in\{1,...,r-1\}} \prod_{j\in\{2,...,r\},j>i} c(ai,aj) \right]^{\frac{2}{r(r-1)}}$$

C'est la moyenne géométrique des cohésions de classes à 2 éléments qui, en tant que telle, s'annule dès que l'une des cohésions en jeu s'annule.

Enfin nous pouvons modéliser l'implication statistique d'une classe de variables sur une autre classe en exigeant du modèle qu'il intègre les informations suivantes :

- les cohésions respectives des 2 classes,
- une intensité d'implication extrémale des éléments d'une classe sur les éléments de l'autre,
- les cardinaux respectifs des 2 classes.

Chacune de ces informations crédite l'indice que nous retiendrons si :

- l'indice croît avec les cohésions de chaque classe et s'annule lorsque la cohésion de l'une d'entre elles est nulle,
- l'indice croît avec la liaison extrémale (minimale si l'on vise un degré d'exigence élevé, maximale si l'on recherche une souplesse réaliste),
- l'indice décroît avec les cardinaux des classes, eu égard à la prise en compte d'une liaison maximale.

Par suite, notant $\underline{\boldsymbol{A}}$ et $\underline{\boldsymbol{B}}$ des classes de variables d'éléments génériques a_i et b_j, puis $C(\underline{\boldsymbol{A}})$ et $C(\underline{\boldsymbol{B}})$ leurs cohésions respectives, l'intensité d'implication de $\underline{\boldsymbol{A}}$ sur $\underline{\boldsymbol{B}}$ est donnée par :

DEFINITION 9 : L'intensité d'implication de $\underline{\boldsymbol{A}}$ sur $\underline{\boldsymbol{B}}$ est :

$$\psi(\underline{\boldsymbol{A}},\underline{\boldsymbol{B}}) = \left[\sup_{i\in\{1,...,r\},j\in\{1,..,s\}} \psi(a_i,b_j) \right]^{rs} . \left[C(\underline{\boldsymbol{A}}).C(\underline{\boldsymbol{B}})\right]^{1/2}$$

On pourra constater que cet indice satisfait les contraintes sémantiques déclarées ci-dessus. Définissant à partir de cet indice une méthode de classification descendante classique par un critère de cohésion décroissante, on obtiendra par exemple des arbres comme celui-ci :

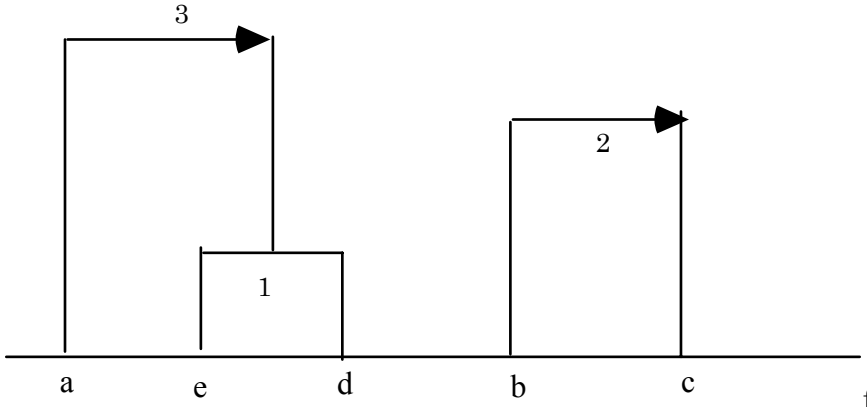


figure 4

Des interprétations de telles méta-règles sont plus complexes, comme par exemple, illustrée par la figure 4, la méta-règle $a \Rightarrow (e \Rightarrow d)$. Mais le mathématicien pourra faire la comparaison entre de tels assemblages et le fonctionnement interne d'une théorie mathématique : une propriété ou un axiome entraîne un théorème, un théorème entraîne un autre théorème, etc.. De plus, quelques méta-règles sont réductibles à des assemblages aisément interprétables. Par exemple, $a \Rightarrow (e \Rightarrow d)$ se ramène logiquement à $a \wedge e \Rightarrow d$.

NIVEAUX SIGNIFICATIFS D'UNE HIERARCHIE COHESITIVE

Etant 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 alors de façon comparable à celle adoptée primitivement par I.C.Lerman (Lerman I.C. 1981) et relativement à la hiérarchie de similarité, mais en reconditionnant son approche.

PREORDRE COHESITIF

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

DEFINITION 10 : On appelle <u>préordre initial et global cohésitif</u> sur VxV (ou préordonnance), le préordre Ω induit par l'application cohésion c sur VxV.

Soit $G(\Omega)$ son graphe dans VxV. D'après les deux premiers paragraphes, il s'ensuit que:

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

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

Soit $S_{\prod k}$ l'ensemble des couples séparés à ce niveau et $R_{\prod k}$ 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 (...(...a,...))...b)...).

L'ensemble $G(\Omega) \cap [S_{\prod k} \times R_{\prod k}]$ 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 $G(\Omega) \cap [S_{\prod k} \times R_{\prod k}]$.

Comme il a été fait dans le premier paragraphe pour le cardinal de $A \cap \overline{B}$, nous associons au cardinal de $G(\Omega) \cap [S_{\prod k} \times R_{\prod k}]$ l'indice aléatoire $card[G(\Omega^*) \cap [S_{\prod k} \times R_{\prod k}]]$ 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 card[$S\prod_k$x $R\prod_k$] et pour variance card[$S\prod_k$x $R\prod_k$] card[G(Ω)].

Soit s(Ω, k) l'indice centré réduit obtenu :

$$\frac{(\mathrm{card}[G(\Omega^*)\cap[S\Pi_k x R\Pi_k]]-1/2\,\mathrm{card}[S\Pi_k x R\Pi_k])}{(\mathrm{card}[S\Pi_k x R\Pi_k]\,\mathrm{card}[G(\Omega)])^{1/2}}.$$

DEFINITION 11 : On appelle noeud significatif tout noeud correspondant à un maximum local de s(Ω, k) au cours de la constitution de la hiérarchie implicative. Nous dirons dans ce cas que la partition $\prod_k$ est en résonance partielle avec Ω.

Si, de plus, G(Ω)«[$S\prod_k$ x $R\prod_k$] = $S\prod_k$x $R\prod_k$, nous dirons que la partition $\prod_k$ est en *résonance totale* avec Ω.

Le logiciel d'analyse de données C.H.I.C. permet le traitement complet de données quantitatives, ainsi que la sortie du graphe d'implication et de la hiérarchie implicative en mentionnant les noeuds significatifs.

TYPICALITE ET CONTRIBUTION DES SUJETS ET DES VARIABLES SUPPLEMENTAIRES

Nous introduisons la notion de variable supplémentaire en analyse implicative à l'instar de la même notion définie en analyse factorielle, c'est-à-dire variable extrinsèque, descripteur par exemple, n'intervenant pas directement dans les liaisons exprimées par la classification entre les variables dites principales de V, donc n'intervenant pas dans la structure de cet ensemble sous la forme graphe ou hiérarchie. Par exemple, une variable supplémentaire pourra représenter une catégorie de sujets (âge, sexe, catégorie socio-professionnelle, etc.).

A un niveau quelconque de la hiérarchie se forme une classe **C** de cohésion non nulle. Notre objectif, particulièrement dans le cas d'un noeud significatif, est de définir un critère permettant d'identifier un ou des sujets, puis la catégorie de sujets, ou tout autre variable supplémentaire(âge, sexe, catégorie socio-professionnelle, etc.), contribuant le plus à la constitution de cette classe. Le comportement de ces sujets sera ainsi en harmonie avec le comportement statistique à l'origine de la classe. Une approche comparable est faite conjointement pour étudier la typicalité et la contribution des sujets et des variables supplémentaires à la constitution d'un arc ou d'un chemin du graphe implicatif.

PUISSANCE IMPLICATIVE DE CLASSE ET DE CHEMIN

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***=>***B*** au sens du paragraphe "Implication entre règles et méta-règles"

DEFINITION 12 : Le couple (a,b) tel que: $\forall i \in \boldsymbol{A}$, $\forall j \in \boldsymbol{B}$ $\psi(a,b) \geq \psi(i,j)$ 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 $\psi(a,b)$ 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≤k) sous-classes (**C** comprise), il y a g couples génériques à l'origine de **C** et g intensités maximales d'implication $\psi_1, \psi_2, ..., \psi_g$, qui leur correspondent.

Dans le cas d'un chemin **C** fermé transitivement (chaque arc de la fermeture admet une intensité d'implication au moins égale à 0.50), composé de g nœuds, **C** présente g(g-1)/2 arcs

transitifs. A chacun de ces arcs, par exemple (a,b), on associe, comme pour une classe, l'intensité d'implication $\psi(a,b)$, que l'on dira encore générique.

DEFINITION 13 : Le vecteur $\psi_1, \psi_2, ..., \psi_g$, élément de $[0,1]^g$, est appelé vecteur puissance implicative de **C**, traduisant une force implicative interne à **C**.

PUISSANCE IMPLICATIVE D'UN SUJET SUR UNE CLASSE OU UN CHEMIN ET DISTANCE A CETTE CLASSE OU A CE CHEMIN

Un sujet x quelconque respecte ou non l'implication du couple générique d'une classe ou d'un arc de chemin avec un ordre de qualité comparable. Associant logique formelle et considération sémantique, nous poserons, par exemple et en fonction des valeurs prises par a et b en x:

$\psi_x(a, \bar{b})=1$ si a=1 ou 0 et b=1; $\psi_x(a, \bar{b})=0$ si a=1 et b=0 ; $\psi_x(a, \bar{b})=p$ si a=b=0 avec $p \in]0,1]$. Dans nos premières expériences, nous choisissions p=.5, valeur neutre. Dans le logiciel CHIC, le calcul des typicalités et des contributions se fait cependant en modulant ces valeurs afin de mieux prendre ne compte la sémantique des valeurs attribuées par x à a et à b.

Ainsi, à x, nous pouvons associer n nombres $\psi_{x,1}$, $\psi_{x,2}, ..., \psi_{x,g}$ correspondant aux valeurs prises en x par les g implications génériques de la classe ou du chemin **C**.

DEFINITION 14 : Le vecteur $(\psi_{x,1}, \psi_{x,2}, ..., \psi_{x,g})$, élément de $[0,1]^g$, est appelé vecteur puissance implicative de x. Le sujet x_t, peut-être fictif, dont toutes les composantes du vecteur puissance sont égales à 1, est appelé sujet idéal théorique de **C**.

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

DEFINITION 15 : On appelle distance implicative d'un sujet x à la classe ou au chemin **C** le nombre:

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

Ce nombre n'est autre que la distance dite du χ^2 entre les deux distributions $\{1-\psi_i\}_i$ et $\{1-\psi_{x,i}\}_i$ qui expriment les écarts entre les implications génériques empiriques et l'implication stricte. C'est pour cette raison que nous avons choisi le mot **typicalité** que nous allons définir plus bas**.** Si pour un i, $\psi_i = 1$, nous poserons, par convention, $\psi_{x,i} = 1$. 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 sujets x de E. Ainsi, si le dénominateur s'annule, il en est de même du numérateur, sauf exception, et l'on pourra de toute façon attribuer la valeur o au quotient.

TYPICALITE ET CONTRIBUTION D'UN SUJET ET D'UNE VARIABLE SUPPLEMENTAIRE A UNE CLASSE OU A UN CHEMIN

TYPICALITE

Nous définirons la typicalité à partir de la "distorsion" du sujet considéré par rapport au sujet idéal théorique, tout en remarquant qu'il peut exister des sujets réels dont la distance à **C** soit inférieure à la distance à cette même classe ou à ce même chemin du sujet idéal théorique. La typicalité d'une catégorie de sujets ou d'une variable supplémentaire G s'en déduira.

DEFINITION 16 : La typicalité de x à **C** est : $\gamma(x,\mathbf{C})=\frac{d(x_t,\mathbf{C})}{d(x,\mathbf{C})}$

et celle de G est : $\gamma(G,\mathbf{C})=\frac{1}{cardG}\sum_{x\in G}\gamma(x,G)$

Ces typicalités peuvent être infinies (pour des configurations contenant des x à distance nulle de **C**) mais, en particulier, supérieures à 1 pour certains sujets. Afin de donner au chercheur le moyen de savoir ou de vérifier rapidement si telle catégorie de sujets qui l'intéresse est statistiquement déterminante dans la constitution d'une classe implicative ou d'un chemin transitif, un algorithme a été élaboré en s'appuyant sur les deux notions suivantes: groupe optimal et catégorie déterminante.

DEFINITION 17 : Soit E la population étudiée. Un groupe optimal d'une classe implicative ou d'un chemin **C,** groupe noté GO(**C**), est le sous-ensemble de E qui accorde à **C** une typicalité plus grande que le complémentaire de GO(**C**) et qui forme avec celui-ci une partition en deux groupes maximisant la variance inter-classe de la série statistique des typicalités individuelles. Une telle partition est dite *significative*.

L'existence de ce groupe optimal est démontrée dans (Gras R. et Ratsimba-Rajohn H. 1997). 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., chaque sous-groupe optimal.

Considérons une partition $\{G_i\}_i$ de E. Cette partition peut être définie par une variable supplémentaire correspondant par exemple à un descripteur de E. Soit X_i une partie aléatoire de E ayant le même cardinal que G_i, et Z_i la variable aléatoire Card $(X_i \cap GO(\mathbf{C}))$. Z_i suit une loi binomiale de paramètres : card G_i et card GO(**C**) / card E.

DEFINITION 18 : On appelle catégorie la plus typique de la classe implicative ou du chemin **C**, la catégorie qui minimise l'ensemble $\{p_i\}_i$ des probabilités p_i telles que:

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

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 ou d'un chemin ayant été donnée par l'expert, 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, mais au moyen d'une modélisation et de concepts différents.

D'ailleurs, nous pouvons remarquer que nous pouvons associer au sous-groupe optimal une variable binaire correspondant à la fonction indicatrice de ce sous-ensemble de E. De la même façon, nous pouvons également associer à la catégorie G_i ou bien à la variable supplémentaire correspondante, une variable binaire dont l'indice de similarité $s=\frac{n_{a\wedge b}-\frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}$, au sens de I.C. Lerman, vérifie : $p_i = Pr[S \geq s]$, S étant la valeur aléatoire dont s est la réalisation. Ainsi, minimiser l'ensemble des probabilités $\{p_i\}_i$ revient à maximiser l'indice de similarité entre les variables binaires, indicatrices de sous-ensembles,

associées respectivement l'une au groupe optimal GO(**C**) et les autres aux différentes catégories $\{G_i\}_i$.

Cette remarque permet d'étendre efficacement la notion de variable supplémentaire la plus typique à des variables numériques, prenant leurs valeurs sur [0,1]. Il suffit dans ce cas d'extraire la plus forte des valeurs de similarité entre la variable binaire indicatrice définie par le groupe optimal et les différentes variables numériques placées en supplémentaire, l'indice étant calculé selon le principe retenu en analyse implicative pour les variables numériques (cf. Définition 5). Nous savons que sa restriction au cas binaire coïncide avec sa valeur s dans le cas où les deux variables sont binaires.

Ainsi il est possible de dégager à la fois les individus et les groupes d'individus typiques d'une liaison ou d'un ensemble (classe ou chemin) de liaisons. Ce sont donc ceux qui sont le plus en accord avec la qualité de ces liaisons au sein de la population E considérée. Si par exemple la liaison entre les variables a et b est quantifiée par le nombre $\psi(a,b)= 0{,}92$, les individus x qui attribuent à cette liaison la valeur $\psi_x(a,b) = 0{,}90$ sont plus typiques que ceux qui lui attribuent la valeur 0,98. La nuance entre cette notion et celle de contribution que nous allons définir prend tout son sens dans l'étude des variables modales ou numériques.

CONTRIBUTION

Cette notion se distingue de la précédente par l'examen de la responsabilité des individus, puis des variables supplémentaires, qui peuvent en être des descripteurs, *à l'existence* d'une liaison d'une règle ou d'une méta-règle entre variables principales. Supposons, en effet, que deux variables a et b (resp. plusieurs variables sur un chemin du graphe ou bien deux classes de la hiérarchie) soient réunies par un arc sur un graphe à un certain seuil (resp. en un chemin transitif **C** du graphe ou bien en une classe **C** dans une hiérarchie à un certain niveau). Connaissant la valeur $\psi_{x,i}$ attribuée par l'individu à la règle i : a=>b (resp. règle i du chemin **C** ou bien de la classe **C** constituée de g règles génériques) supposée admissible,

DEFINITION 19 : On appelle *distance de x à (a,b) ou à **C*** :

$$d(x,\mathbf{C}) = \left\lceil \frac{1}{g}\sum_{i=1}^{i=g}\left[1-\psi_{x,i}\right]^2 \right\rceil \text{ où g = 1 dans le cas de l'arc (a,b)}$$

On appelle *contribution de x à **C*** le nombre : $\gamma(x,\mathbf{C}) = 1\text{-}d(x,\mathbf{C})$

Cette contribution a pour maximum 1 dans le cas où l'individu x a donné la valeur 1 à toutes les règles i. Ceci permet de concilier sémantique et définition formelle. La suite des définitions et des algorithmes de calcul (contribution d'une catégorie ou d'une variable supplémentaire G, groupe optimal d'individus, catégorie ou variable supplémentaire la plus contributive) se transpose immédiatement à partir des principes de la typicalité. Mais dans les situations réelles, nous observons la nuance entre les deux concepts ce qui enrichit l'information exploitable par l'utilisateur. Cependant, <u>le concept de contribution est plus volontiers retenu pour l'interprétation.</u>

EN CONCLUSION, les applications de la méthode ont d'ores et déjà donné des résultats très satisfaisants, et non seulement à la discipline où elle a pris naissance mais aussi dans d'autres domaines de l'éducation ou de recherche scientifique plus générale, comme le montrent les autres communications présentées au cours des journées qui lui étaient consacrées. Le plus

souvent, les interprétations s'appuient complémentairement sur l'analyse de similarités ou/et sur les méthodes factorielles tout en apportant des informations qui lui sont spécifiques en raison de son caractère non symétrique. Les analyses bénéficient efficacement du logiciel C.H.I.C., qui permet, avec une certaine convivialité, tous les traitements algorithmiques et graphiques des méthodes évoquées dans cet article. Son développement suit régulièrement toutes les nouvelles avancées de la théorie de l'implication statistique. Citons, à cet égard, quelques pistes de recherche actuelle : données incomplètes, contraction de l'ensemble des variables, variables à valeurs intervalles, etc.

BIBLIOGRAPHIE

[1] AGRAWAL R., IMIELINSKY T., SWAMI A., "Mining association rules between sets of items in large databases", *Proc. of the ACM SIGMOD'93,* (1993)

[2] AMARGER S., DUBOIS D., PRADE H., "Imprecise quantifiers and conditional probabilities" - *in Symbolic and quantitative approaches to uncertainty (R. KRUSE, P. SIEGEL), Springer-Verlag,* (1991), 33-37.

[3] BAILLEUL M., *Analyse statistique implicative: variables modales et contribution des sujets. Application à la modélisation de l'enseignant dans le système didactique*, Thèse de l'Université de Rennes 1, juin 1994.

[4] BAILLEUL M. et GRAS R., "L'implication statistique entre variables modales", *Mathématique, Informatique et Sciences Humaines, E.H.E.S.S. Paris, n°128,* (1995), 41-57

[5] BERNARD J.-M. et POITRENAUD S., "L'analyse implicative bayesienne d'un questionnaire binaire : quasi-implications et treillis de Galois simplifié", *Mathématiques, Informatique et Sciences Humaines, n° 147,* (1999), 25-46

[6] BODIN A., "Modèles sous-jacents à l'analyse implicative et outils complémentaires". *Prépublication IRMAR. n°97-32,* (1997)

[7] BODIN A. et GRAS R., 1999 : "Analyse du préquestionnaire enseignants avant EVAPM-Terminales"*, Bulletin n° 425 de l'Association des Professeurs de Mathématiques de l'Enseignement Public*, *Paris* (1999), 772-786

[8] COUTURIER R. et GRAS R., "Introduction de variables supplémentaires dans une hiérarchie de classes et application à CHIC", *Actes des 7èmes Rencontres de la Société Francophone de Classification, Nancy,* (15-17 septembre 1999), 87-92

[9] DIDAY E., *Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes,* Thèse d'Etat, Université de Paris VI, (1972)

[10]GANASCIA J.G., *AGAPE et CHARADE : deux techniques d'apprentissage symbolique appliquées à la construction de bases de connaissances*, Thèse d'Etat, Université de Paris Sud, (1987)

[11] GRAS R., *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*, Thèse d'Etat, Université de Rennes I, (1979)

[12] GRAS R. et LARHER A., L'implication statistique, une nouvelle méthode d'analyse de données, *Mathématique, Informatique et Sciences Humaines, E.H.E.S.S. Paris, n°120* (1992), 5-31

[13] GRAS R. et RATSIMBA-RAJOHN H. "Analyse non symétrique de données par l'implication statistique". *RAIRO-Recherche Opérationnelle, 30-3, AFCET, Paris*, (1996), 217-232

[14] GRAS R., BRIAND H. et PETER P. "Structuration sets with implication intensity", *Proceedings of the International Conference on Ordinal and Symbolic Data Analysis - OSDA 95, E.Diday, Y.Chevallier, Otto Opitz, Eds., Springer, Paris* (1996), 147-156

[15] GRAS R. et coll., *L'implication Statistique,* Collection Associée à "Recherches en Didactique des Mathématiques", La Pensée Sauvage, Grenoble, (1996)

[16] GRAS R., BRIAND H., PETER P., PHILIPPE J., 1997] - "Implicative statistical analysis", *Proceedings of International Congress I.F.C.S., 96, Kobé, Springer-Verlag, Tokyo* (1997), 412-419

[17] GRAS R. et PETER P., "From a cognitive complexity problem to an implicative model", *Actes de l'Intensive Programme Socrates/Erasmus 1998/1999, University of Cyprus, in A multidimensional approach to learning in mathematics and sciences, A.Gagatsis Ed., Intercollege Press Cyprus, 491-500, Nicosia,* (1999), 491-500

[18] GRAS R., KUNTZ P., COUTURIER R. et GUILLET F.. « Une version entropique de l'intensité d'implication pour les corpus volumineux ». *Extraction des Connaissances et Apprentissage (ECA)*, vol. 1, n° 1-2, 69-80. Hermès Science Publication.

[19] LAGRANGE J.B.,"Analyse implicative d'un ensemble de variables numériques ; application au traitement d'un questionnaire à réponses modales ordonnées", *Revue de Statistique Appliquée., Institut Henri Poincaré*, Paris, (1998, 71-93

[20] LAHANIER-REUTER D.,*Etude de conceptions du hasard : approche épistémologique, didactique et expérimentale en milieu universitaire*, Thèse de l'Université de Rennes 1, 1998

[21] LERMAN I.C., *Les bases de la classification hiérarchique*, Gauthier-Villars, Paris, (1970)

[22] LERMAN I.C., *Classification et analyse ordinale des données*, Dunod, Paris, (1981).

[23] LERMAN I.C., GRAS R., ROSTAM H., "Elaboration et évaluation d'un indice d'implication pour des données binaires", *I et II, Mathématiques et Sciences Humaines ,* (1981), n° 74,, 5-35 et n° 75, 5-47

[24] LOEVINGER J., "A systematic approach to the construction and evaluation of tests of abilities", *Psychological Monographs, 61, n° 4,* (1947)

[25] PEARL J., *Probabilistic Reasoning in intelligent systems,* San Mateo, CA, Morgan Kaufmann., (1988)

[26] SIMON A., *Outils classificatoires par objets pour l'extraction de connaissances dans des bases de données,* Thèse de l'Université de Nancy 1, (2000)