Les fondements de l'analyse statistique implicative

Conférence par Régis Gras

Ecole Polytechnique de l'Université de Nantes

Résumé

Partie de situations de didactique des mathématiques où il s'agissait de hiérarchiser des problèmes en fonction des difficultés ressenties par les élèves, la méthode implicative se développe au fil des problèmes qu'elle rencontre, des questions qu'ils posent. Problèmes sur lesquels elle pense pouvoir agir afin de structurer et permettre, à partir de la contingence de règles, d'expliquer et donc de prévoir dans différents domaines : la psychologie, la sociologie, la biologie, entre autres. C'est ainsi que les concepts d'intensité d'implication, de cohésion de classes, de mesure d'implication-inclusion, de significativité de niveaux de hiérarchie, de contribution de variables supplémentaires, etc., ont été développés. C'est aussi pour y répondre qu'au traitement de variables binaires se sont ajoutés ceux des variables modales, fréquentielles et, plus récemment, de variables-intervalles.

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[1].

§ 1 L'intensité d'implication statistique revisitée

1-1 Rappelons la **problématique de l'implication statistique dans le cas binaire.**

Une population E d'objets ou sujets est croisée avec des variables (caractères, critères, réussites, ...) d'un ensemble 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 objets 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 régles assez fiables pour pouvoir décrire, structurer une population et en conjecturer 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

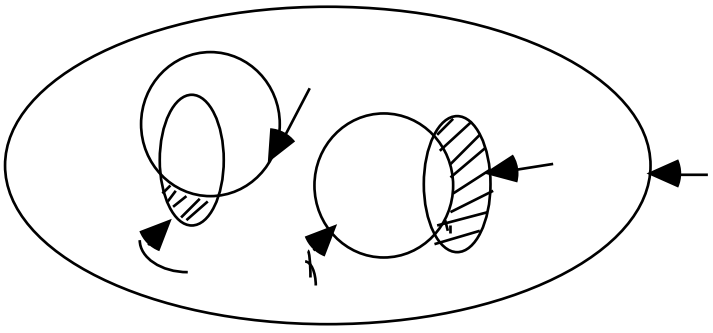
[1]La version pour Windows 95 est développée par Raphaël COUTURIER

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 -> 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 $n_{a\ \bar{b}}$) qu'il faudra prendre en compte pour accepter statistiquement de conserver ou non la **quasi-implication** ou la **quasi-règle** "a => b".

Pour mathématiser cela, 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 cardinal $n_{\bar{b}}$.



Les parties hachurées correspondent à la non-satisfaction de l'implication de a sur b.

Nous dirons alors :

Définition 1

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

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

Nous démontrons (LERMAN I.C. GRAS R., ROSTAM H., 1981) que, pour un certain processus de tirage

La variable aléatoire 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 :

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

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

Définition 2

$$q(a,\bar{b}) = \frac{n_{a\ \bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}}$$

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 d'implication</u>, 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 $q(a,\bar{b})$ par :

Définition 3

Dans le cas où $n_b \neq n$, l'<u>intensité d'implication</u> de a sur b est :

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

Par suite, la définition de l'implication devient :

Définition 4

L'implication a => b sera <u>admissible au niveau de confiance 1 - α</u>, si et seulement si

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

Rappelons que cette modélisation de la quasi-implication mesure <u>l'étonnement de constater la petitesse des contre-exemples en regard du nombre surprenant des instances de l'implication</u>. 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 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 $\phi(a,\bar{b})$ tend vers 0.

Remarque 1: D'autres modélisations, autres que celle de Poisson, sont possibles. Citons :
* *une modélisation binomiale* : considérant les variables duales $\text{card}(A \cap \overline{Y})$ et $\text{card}(X \cap \overline{B})$, 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 $\frac{n_a}{n}\frac{n_{\bar{b}}}{n}$ d'appartenir à $A \cap \overline{Y}$. Par suite :

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

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

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

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

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 (1942) 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, ce qui ne peut que rendre statistiquement crédible la relation que nous voulons modéliser.

1-2 Cas des variables modales et fréquentielles

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. 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 (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 :

Défintion 5

$$q(a,\bar{b}) = \frac{\sum_{x \in E} a(x)\bar{b}(x) - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{(n^2 s_a^2 + n_a^2)((n^2 s_{\bar{b}}^2 + n_{\bar{b}}^2)}{n}}}$$ est l'indice de propension de variable modale

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 : $n^2 s_a^2 + n_a^2 = nn_a$, $n^2 s_{\bar{b}}^2 + n_{\bar{b}}^2 = n n_{\bar{b}}$ et $\sum_{x \in E} a(x)\bar{b}(x) .= \frac{n_a n_{\bar{b}}}{n}$.

Cette solution apportée au cas modal est auusi 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.

1-3 Cas des variables-intervalles

1-3-1 Situation fondamentale

Deux variables réelles a et b prennent un certain nombre de valeurs sur 2 intervalles finis $[\alpha_1, \alpha_2]$ et $[\beta_1, \beta_2]$. Soit A (resp. B) l'ensemble des valeurs de a (resp. b) observées sur $[\alpha_1, \alpha_2]$ (resp. $[\beta_1, \beta_2]$. 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 $[\alpha_1, \alpha_2]$ (resp. $[\beta_1, \beta_2]$.) afin que la partition la plus fine obtenue respecte au mieux la distribution des valeurs observées dans $[\alpha_1, \alpha_2]$ (resp. $[\beta_1, \beta_2]$.) ?

2° peut-on trouver les partitions respectives de $[\alpha_1, \alpha_2]$ et $[\beta_1, \beta_2]$.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 en faisant choix des critères à optimiser pour satisfaire l'optimalité attendue dans chaque cas.

1-3-2 Premier problème

On va s'intéresser à l'intervalle $[\alpha_1, \alpha_2]$ 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 $[\alpha_1, \alpha_2]$ 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[2] 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 $[\alpha_1, \alpha_2]$ et l'ensemble des partitions P de $[\alpha_1, \alpha_2]$ en p sous-intervalles $A_{i,}$, de la façon suivante :

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

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 noyaux, en nombre p, dans $[\alpha_1, \alpha_2]$. 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 :

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

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

[2]Diday (E), *Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes,* thèse de doctorat d'état, Université de Paris VI, 1972

proches de leurs propres barycentres respectifs. En cas d'égalité, on affecte x au sous-intervalle de plus petit indice.

On démontre alors que $W(G_1, P_1) \leq W(G_1, P_0)$ (cf

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, $D(G_{2i}, A_{1i}) \leq D(y, A_{1i}) = \sum_{x \; A_{1i}} (y-x)^2$, 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 :

$$A_{2i} = \left\{x / \quad j \quad (G_{2i} - x)^2 \leq (G_{2j} - x)^2\right\}$$

On obtient de même : $W(G_2, P_2) \leq W(G_2, P_1)$ (cf proposition et sa preuve plus loin)

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:

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

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 :

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

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

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= {54, 55, 55, 57, 57, 58, 60, 66, 67, 67, 68, 70, 71, 74, 78, 79, 79}

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 :

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

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.

1-3-3 Deuxième problème

On suppose maintenant que les intervalles $[\alpha_1, \alpha_2]$ et $[\beta_1, \beta_2]$ 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].

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 $[\alpha_1, \alpha_2]$ sur tous les éléments d'une des partitions de $[\beta_1, \beta_2]$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 $[\alpha_1, \alpha_2]$ admet la partition correspondant au premier maximum et que l'implication réciproque optimale est satisfaite pour la partition de $[\beta_1, \beta_2]$ correspondant au deuxième maximum.

Remarque

1° Il n'existe pas de relation d'ordre entre $\phi(a, \bar{b})$ et $\phi(a, b\overline{\ \ }c)$ connaissant $\phi(a, \bar{b})$ et $\phi(a, \bar{c})$.

En effet, on peut avoir $\phi(a, \bar{b}) < \phi(a, b\overline{\ \ }c)$ et $\phi(a, \bar{c}) < \phi(a, b\overline{\ \ }c)$.
exemple1 : n = 100 ; n_a = 16 ; n_b = 35 ; n_c = 30 ; $n_{a\ \bar{b}}$ = 10 ; $n_{a\ \bar{c}}$ =8 ; $n_{a\ (b\overline{\ \ }c)}$ = 2

Mais on peut également avoir : $\phi(a, \bar{b}) > \phi(a, b\overline{\ \ }c)$ et $\phi(a, \bar{c}) > \phi(a, b\overline{\ \ }c)$
exemple 2 : n = 100 ; n_a = 30 ; n_b = 50 ; n_c = 49 ; $n_{a\ \bar{b}}$ = 15 ; $n_{a\ \bar{c}}$ =16 ; $n_{a\ (b\overline{\ \ }c)}$ = 1

2° De même, il n'existe pas a priori de relation entre $\phi(a, \bar{c})$, $\phi(b, \bar{c})$ et $\phi(a\ \ b, \bar{c})$

En effet, on peut avoir, $\phi(a, \bar{c}) < \phi(a\ \ b, \bar{c})$ et $\phi(b, \bar{c}) < \phi(a\ \ b, \bar{c})$
exemple 3 : n = 100 ; n_a = 20 = n_b ; n_c = 35 ; $n_{a\ \bar{c}}$ =16 = $\phi(b, \bar{c})$; $n_{a\ b,\bar{c}}$ = 10

Mais on peut avoir également : $\phi(a, \bar{c}) > \phi(a\ \ b, \bar{c})$ et $\phi(b, \bar{c}) > \phi(a\ \ b, \bar{c})$
exemple 4 : n = 100 ; n_a = 20 = n_b ; n_c = 48 ; $n_{a\ \bar{c}}$ =11 = $\phi(b, \bar{c})$; $n_{a\ b,\bar{c}}$ = 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.

[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.

Décroissance de la fonction W

Quelques notations

Soit A l'ensemble des valeurs observées dans l'intervalle $[\alpha_1, \alpha_2]$ 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 $W(L,P) = \sum_{i=1}^{p} D(N_i,A_i) = \sum_{i=1}^{i=p} \sum_{x\ A_i,y\ N_i} d(x,y)$ [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.

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 :

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

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 :

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

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

Proposition :

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

a) P, W(L, f(L)) ≤ W(L, P)

b) L, W(g(P), P) ≤ 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.

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

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

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

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

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

Par suite, pour tout x A, les contributions de x à W(L, Q) sont inférieures ou égales à celles de W(L, P), soit encore 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:

$$k, O_k = \left\{x \quad A / d(x,A_k) = \inf_j d(x, A_j)\right\}.$$

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

1-4 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 $\phi(a,\bar{b})$ a tendance à ne plus être suffisamment discriminante car ses valeurs peuvent être voisines de 1, alors que l'inclusion dont elle cherche à modéliser la qualité, est loin d'être satisfaite (phénomène signalé par A.Bodin 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. 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 qualité de l'information fournie par la faiblesse relative des instances qui contredisent la règle et sa contraposée. C'est donc au concept d'entropie de Shannon que nous faisons référence :

$$H(b/a = 1) = -\frac{n_{a\ b}}{n_a} \log_2 \frac{n_{a\ b}}{n_a} - \frac{n_{a\ \bar{b}}}{n_a} \log_2 \frac{n_{a\ \bar{b}}}{n_a} ,$$

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

$$H(\bar{a} / \bar{b}=1) = - \frac{n_{a\ \bar{b}}}{n_{\bar{b}}} \log_2 \frac{n_{a\ \bar{b}}}{n_{\bar{b}}} - \frac{n_{\bar{a}\ \bar{b}}}{n_{\bar{b}}} \log_2 \frac{n_{\bar{a}\ \bar{b}}}{n_{\bar{b}}}$$

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

Ces entropies devraient donc être simultanément petites si l'on souhaite disposer d'un bon critère d'inclusion de A dans B. Cependant avec une moyenne arithmétique faible de ces

deux entropies, nous disposons d'un bon critère d'inclusion, qu'il nous faut maintenant adapter au modèle attendu dans les différentes situations cardinales.

En fait, afin que soient respectés les nombres d'instances respectives de a par rapport à b et de non b par rapport à non a, l'inclusion de A dans B aura d'autant plus de sens que seront faibles dans l'observation les deux fonctions h_1 et h_2 définies respectivement à partir de H(b/a=1) et H($\bar{a}$ / $\bar{b}$=1), par :

$$h_1(a,b) = \frac{1}{2}\left(-\frac{n_{a\,b}}{n_a}\log_2\frac{n_{a\,b}}{n_a} - \frac{n_{a\,\bar{b}}}{n_a}\log_2\frac{n_{a\,\bar{b}}}{n_a}\right)\mathbf{1}_{\left[0;\frac{n_a}{2n}\right]}(t) +$$

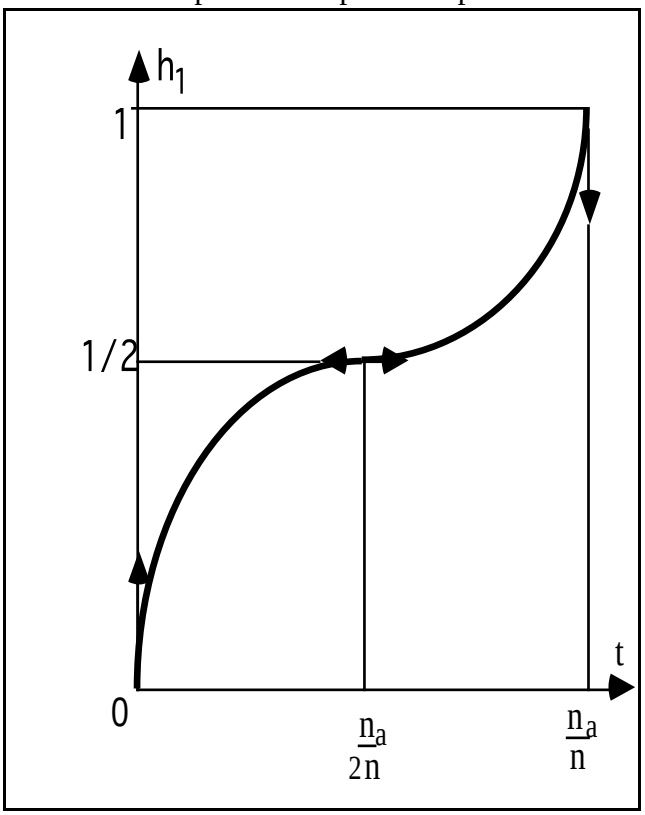
$$\frac{1}{2}\left(2 + \frac{n_{a\,b}}{n_a}\log_2\frac{n_{a\,b}}{n_a} + \frac{n_{a\,\bar{b}}}{n_a}\log_2\frac{n_{a\,\bar{b}}}{n_a}\right)\mathbf{1}_{\left[\frac{n_a}{2n};\frac{n_a}{n}\right]}(t)$$

$$h_2(a,b) = \frac{1}{2}\left(-\frac{n_{a\,\bar{b}}}{n_{\bar{b}}}\log_2\frac{n_{a\,\bar{b}}}{n_{\bar{b}}} - \frac{n_{\bar{a}\,\bar{b}}}{n_{\bar{b}}}\log_2\frac{n_{\bar{a}\,\bar{b}}}{n_{\bar{b}}}\right)\mathbf{1}_{\left[0,\frac{n_{\bar{b}}}{n}\right]}(t) +$$

$$\frac{1}{2}\left(2 + \frac{n_{a\,\bar{b}}}{n_{\bar{b}}}\log_2\frac{n_{a\,\bar{b}}}{n_{\bar{b}}} + \frac{n_{\bar{a}\,\bar{b}}}{n_{\bar{b}}}\log_2\frac{n_{\bar{a}\,\bar{b}}}{n_{\bar{b}}}\right)\mathbf{1}_{\left[\frac{n_{\bar{b}}}{2n};\frac{n_{\bar{b}}}{n}\right]}(t)$$

où l'on note $t = \frac{n_{a\,\bar{b}}}{n}$, et par exemple, $\mathbf{1}_{\left[0;\frac{n_a}{2n}\right]}$ la fonction indicatrice sur l'intervalle $\left[0;\frac{n_a}{2n}\right]$ de la variable t, fréquence des contre-exemples, pour n_a et n_b fixés.

Représentons par exemple la fonction h_1 de t :



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

* "réaction" rapide aux premiers contre-exemples,
* "ralentissement" du rejet de l'inclusion au voisinage de l'équilibre soit $\frac{n_a}{2n}$,
* rejet de plus en plus accentué au-delà de $\frac{n_a}{2n}$ ce que n'assurait pas l'intensité d'implication.

En définitive, nous retenons :

Définition 6

comme <u>indice d'inclusion</u> le nombre :

$$i(a,b) = 1 - \frac{1}{2}(h_1(a,b) + h_2(a,b))$$

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

comme <u>intensité d'implication-inclusion</u> *(ou intensité entropique)* le nombre :

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

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

<u>Exemple</u>

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

L'intensité d'implication est 0,9999 ($q(a,\bar{b}) = -3,65$)

L'entropie moyenne de l'expérience est égale à : 0,406 ($h_1 = 0,541$, $h_2 = 0,272$). La valeur du coefficient modérateur est : $i(a,b) = 0,594$

Donc $\psi(a,b) = 0,77$

Ainsi, l'entropie "modére" l'intensité d'implication dans ce cas où justement l'inclusion est médiocre, tout en prendant en compte la qualité de la contraposée, comme ci-dessus. Si chacun des cardinaux du tableau précédent est divisé par 100, le coefficient i(a,b) ne change pas alors que l'intensité devient 0,602.

<u>Remarque</u>

La correspondance entre $\phi(a,\bar{b})$ et i(a,b) n'est pas monotone comme le montre le deuxième exemple suivant :

	b	$\bar{b}$	marge
a	250	450	700
$\bar{a}$	600	2750	3350
marge	850	3200	4050

L'intensité d'implication est supérieure à la précédente: $q(a,\bar{b}) = -4,38$

L'entropie moyenne de l'expérience est égale à : 0,411 ($h_1 = 0,530$, $h_2 = 0,2929$). La valeur du coefficient modérateur est : $i(a,b) = 0,588$

Donc $\psi(a,b) = 0,767$

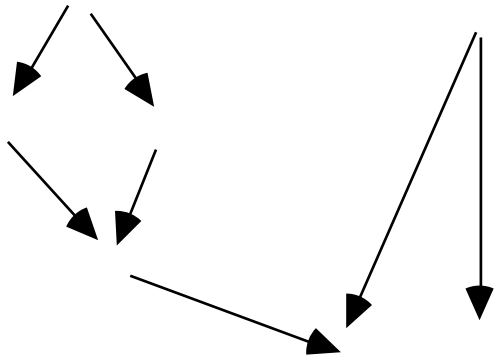
Ainsi, alors que $\phi(a,\bar{b})$ a crû du 1er au 2ème exemples, i(a,b) a décru. En revanche, la situation contraire est la plus fréquente.

1-5 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 $\phi(a,\bar{c}) \geq 0,5$, 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 :



§ 2 Implication entre règles et méta-règles

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 conotation 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 :

Définition 7

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

. si $p = \max(\varphi(a,\bar{b}), \varphi(b,\bar{a}))$ 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

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

Définition 8

La cohésion de la classe de variables ***A***= $(a_1,.....,a_r)$ est alors définie par extension :

$$C(\mathbf{A}) = \left(\prod_{i \in \{1,...,r-1\},\ j \in \{2,...,r\},\ j>i} c(a_i,a_j) \right)^{\frac{2}{r(r-1)}}$$

C'est la moyenne géométrique des cohésions de classes à 2 éléments

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.

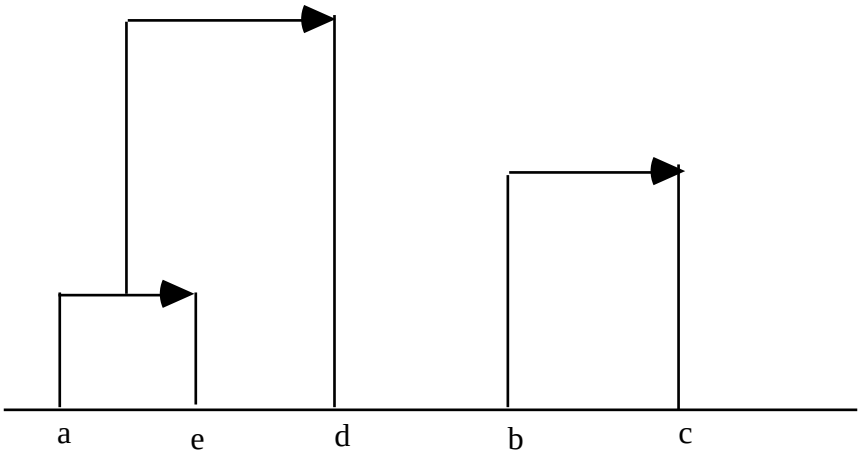
Par suite, notant $\underline{\boldsymbol{A}}$ et $\underline{\boldsymbol{B}}$ des classes de variables d'éléments génériques a_i et b_j, puis $C(\underline{\boldsymbol{A}})$ et $C(\underline{\boldsymbol{B}})$ leurs cohésions respectives, l'intensité d'implication de $\underline{\boldsymbol{A}}$ sur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

Définition 9

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

$$\psi(\underline{A},\underline{B}) = \left[\sup_{\{1,\ldots,r\},j\ \{1,\ldots,s\}} \phi(a_i,\bar{b}_j)^{rs}\right] . \left[C(\underline{A}).C(\underline{B})\right]^{1/2}$$

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 :



§ 3 Niveaux significatifs d'une hiérarchie implicative

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 (cf LERMAN I.C. 1981) 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 $\{a_1, a_2, ..., a_m\}$ et l'ensemble des couples (a,b) de VxV tels que a≠b. Il existe m(m-1) tels couples auxquels on a associé leurs cohésions c(a,b) respectives.

Définition 10:

On appelle préordre initial et global cohésitif sur VxV (ou préordonnance), le préordre Ω induit par l'application cohésion c sur VxV.

Soit G(Ω) son graphe dans VxV. D'après les **§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 $\phi(a,\bar{b}) \leq 0{,}5$,

* 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 $\prod_k$ 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(Ω)) 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 pour le cardinal de $A \cap \bar{B}$, associons (cf. LERMAN I.C. 1981]) au cardinal de $G(\Omega) \cap [\, S_{\prod_k} \times R_{\prod_k}]$ l'indice aléatoire $\mathrm{card}[G(\Omega^*) \cap [S_{\prod_k} \times R_{\prod_k}]]$ 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[$S\Pi_k$x $R\Pi_k$] et pour variance card[$S\Pi_k$x $R\Pi_k$] card[G(Ω)].

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

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

Définition 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 Π_k est en résonance partielle avec Ω.

Si, de plus, $G(\Omega) \cap [\, S\Pi_k x R\Pi_k] = S\Pi_k x R\Pi_k$, nous dirons que la partition Π_k est en résonance totale avec Ω.

Le logiciel d'analyse de données C.H.I.C. 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.

§ 4 Contribution des sujets et des variables supplémentaires

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-profesionnelle, 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 a été faite conjointement pour étudier la contribution des variables supplémentaires à la constitution d'un arc ou d'un chemin du graphe implicatif.

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**=>**B** au sens du **§ 2.**

Définition 12

Le couple (a,b) tel que: $\forall i \in A, \forall j \in \mathbf{B}\ \phi(a,\bar{b}) \geq \phi(i,\bar{j})$ 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 $\phi(a,\bar{b})$ 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 $\phi_1, \phi_2, ..., \phi_g$ qui leur correspondent.

Définition 13

Le vecteur $(\phi_1, \phi_2, ..., \phi_g)$, 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 sujet sur une classe et distance à cette classe

Un sujet 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:

$\phi_X(a,\bar{b})=1$ si a=1 ou 0 et b=1; $\phi_X(a,\bar{b})=0$ si a=1 et b=0 ; $\phi_X(a,\bar{b})=p$ si a=b=0 avec $p \in]0,1]$. Le plus souvent, nous choisissons p=.5, valeur neutre.

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

Définition 14

Le vecteur $(\phi_{x,1}, \phi_{x,2}, ..., \phi_{x,g})$, 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 χ^2 afin d'accentuer les effets de fortes implications génériques.

Définition 15

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

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

Ce nombre n'est autre que la distance dite du χ^2 entre les deux distributions $\{1-\phi_i\}_i$ et $\{1-\phi_{x,i}\}_i$ qui expriment les écarts entre les implications génériques empiriques et l'implication stricte. Si pour un i, $\phi_i = 1$, nous poserons, par convention, $\phi_{x,i} = 1$. 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, et l'on pourra attribuer la valeur o au quotient.

4.3. Contribution d'un sujet et d'une variable supplémentaire à une classe

Nous la définirons 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 la classe **C** soit inférieure à la distance à cette même classe du sujet idéal théorique. La contribution d'une catégorie de sujets ou d'une variable supplémentaire G s'en déduira.

Définition 16

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

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

Ces contributions peuvent être infinies (pour des configurations contenant des x à distance nulle de **C**) mais, en particulier, supérieures à 1 pour certains 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, un algorithme a été élaboré en s'appuyant sur les deux notions suivantes : groupe optimal et catégorie déterminante.

Définition 17

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

Définition 18

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:

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

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 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.

En conclusion, les applications de la méthode ont d'ores et déjà donné de très bons résultats et non seulement à la discipline où elle a pris naissance (la didactique des mathématiques) mais aussi dans d'autres domaines de l'Education, en psychologie, en sociologie, en biologie, etc. Les analyses bénéficient du logiciel C.H.I.C., développé sous Windows 95 par R.COUTURIER, logiciel qui permet, avec une certaine convivialité, tous les traitements des méthodes évoquées dans cet article. Son développement suit régulièrement toutes les nouvelles avancées de la théorie de l'implication statistique.

Bibliographie :

[AG ALMOULOUD S., 1992] - L'ordinateur, outil d'aide à l'apprentissage de la démonstration et traitement de données didactiques, *Thèse de doctorat de l'Université de Rennes I.*

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

[BAILLEUL M., 1994] - 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.*

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

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

[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*, 772-786, Paris

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

[DIDAY E., 1972] *Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes,* thèse de doctorat d'état, Université de Paris VI

[GANASCIA J.G., 1991] - CHARADE : Apprentissages de bases de connaissances dans *"Induction symbolique -numérique à partir de données", Ed. KODRATOFF et DIDAY, CEPADUES, 1991.*

[GRAS R., 1979] - 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.*

[GRAS R. et RATSIMBA-RAJOHN H. 1996] - Analyse non symétrique de données par l'implication statistique. *RAIRO-Recherche Opérationnelle, 30-1, AFCET, Paris.*

[GRAS R. et PETER P. 1996] - 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*

[GRAS R. et als 1996] - *L'implication Statistique,* Ouvrage de 320 pages dans la Collection Associée à "Recherches en Didactique des Mathématiques", La Pensée Sauvage, Grenoble

[GRAS R., BODIN A., COUTURIER R., GUILLET F., 1996] - *Apprentissage automatique et implication : mise en oeuvre sur un espace d'apprentissage en ressources humaines* et *Analyse d'une épreuve de concours par la méthode implicative. Présentation interactive.* Communications lors des "Quatrièmes Journées de la Société Francophone de classification, Vannes 19-20 Septembre 1996

[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*

[GRAS R. 1997] : *Metodologia d'analisi di indagini,* in Quaterni di Ricerca in Didattica, Palerme n°7, 99-109

[GRAS R. et PETER P. 1999] : 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*

[LAGRANGE J.B., 1998] - Analyse implicative d'un ensemble de variables numériques ; application au traitement d'un questionnaire à réponses modales ordonnées, *Revue de Statistique Appliquée., I.H.P. Paris*

[LARHER A., 1991] - Implication statistique et applications à l'analyse de démarches de preuve mathématique, *Thèse de l'Université de Rennes I.*

[LERMAN I.C., 1970] - *Les bases de la classification hiérarchique,* chap. 1, Gauthier-Villars, Paris.

[LERMAN I.C., 1981] - *Classification et analyse ordinale des données,* Dunod, 1981.

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

[LERMAN I.C., 19984, 1995] - Rôle de l'inférence statistique dans une approche de l'analyse classificatoire des données, *Actes du Colloque "Méthodes d'analyses statsitiques multidimensionnelles en didactique des mathématiques", I.U.F.M. de Caen 27-29 Janvier 1995, Ed. R.Gras, IRMAR, Rennes 1*

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

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

[RATSIMBA-RAJOHN H., 1992] - Contribution à l'étude de la hiérarchie implicative. Application à l'analyse de la gestion didactique des phénomènes d'ostension et de contradictions. *Thèse de doctorat de l'Université de Rennes I.*

[TOTOHASINA A., 1992] - Méthode implicative en analyse de données et application à l'analyse de conceptions d'étudiants sur la notion de probabilité conditionnelle. *Thèse de doctorat de l'Université de Rennes I.*