Typicalité et contribution des sujets et des variables supplémentaires en Analyse Statistique Implicative

Régis Gras *, Jérôme David*, Jean-Claude Régnier**, Fabrice Guillet*

* LINA– Ecole Polytechnique de l'Université de Nantes
La Chantrerie BP 60601 44306 Nantes cedex
regisgra@club-internet.fr, jerome.david, fabrice.guillet@polytech.univ-nantes.fr
http://www.sciences.univ-nantes.fr/lina/
**EA 3727 Savoirs, Diversité et Professionnalisation, Lyon 2
86, rue Pasteur 69365 Lyon cedex 07
Jean-claude.regnier@univ-lyon2.fr

Résumé. L'analyse statistique implicative traite des tableaux sujets x variables afin d'extraire règles et métarègles statistiques entre les variables. L'article interroge les structures obtenues représentées par graphe et hiérarchie orientés afin de dégager la responsabilité des sujets ou des groupes de sujets (variables supplémentaires) dans la constitution des chemins du graphe ou des classes de la hiérarchie. On distingue les concepts de typicalité pour signifier la proximité des sujets avec le comportement moyen de la population envers les règles statistiques extraites, puis de contribution pour quantifier le rôle qu'auraient les sujets par rapport aux règles strictes associées. Un exemple de données réelles, traité à l'aide du logiciel CHIC, illustre et montre l'intérêt de ces deux concepts.

1 Introduction

Les données traitées par l'analyse statistique implicative (en abrégé : A.S.I.) se présentent sous forme de tableaux numériques croisant une population E de sujets, ou individus ou objets, associé chacun à une ligne, et un ensemble V de variables simples ou conjointes (attributs binaires, variables numériques, rang, intervalle) chacune associée à une colonne. A l'intersection de la ligne x et de la colonne j figure la valeur prise par le sujet x selon la variable j. La finalité première de l'A.S.I. vise à dégager de V ou de l'ensemble de toutes les conjonctions d'éléments de V[1], des règles d'association non symétrique, contrairement à la similarité, sur une base statistique, du type : « si la variable ou une conjonction de variables a est observée sur E alors la variable b a tendance à être observée », règle notée $a \Rightarrow b$. Une mesure de qualité, non symétrique, de telles règles[2] est définie par :

[1] Dorénavant nous continuerons à noter V, pour éviter des notations excessives, aussi bien l'ensemble des variables que celui de toutes les conjonctions de ses éléments.

[2] D'autres mesures existent comme celle d'(Agrawal et al.,1993) basée sur les deux paramètres : support (fréquence de a et b) et confiance (fréquence conditionnelle de b sachant a)

l'intensité d'implication, notée φ, qui prend ses valeurs dans [0,1], construite selon un modèle probabiliste sur la base du nombre de contre-exemples à la règle et des occurrences en jeu[3] (Gras, 1979 et al. 1996b), ou, dans le cas des tableaux de grande taille, *l'intensité d'implication-inclusion*, notée ψ, à valeurs également dans [0,1], qui intègre en outre la qualité de la contraposée de la règle et l'entropie des expériences associées (Gras, 2000), (Gras et al. 2001), (Blanchard et al., 2003). Cette version de l'intensité d'implication permet de mieux cerner la notion et la recherche de causalité entre variables.

En fait, deux structures de V, résumant l'ensemble des règles d'association, sont obtenues à partir de ces règles et conduisent à deux types de représentation :

un *graphe* dit *implicatif,* orienté, sans cycle, pondéré par une mesure de qualité des règles ; les nœuds sont les variables ; un arc du graphe représente une règle, par exemple, $a \Rightarrow b$. Il est fermé transitivement dès lors que la mesure de l'intensité entre deux nœuds quelconques d'un de ses chemins est au moins égale à 0,5 (FIG.1). Dans ce premier exemple, $(M \Rightarrow F \Rightarrow OP5 \Rightarrow OP4 \Rightarrow OP6)$ est un chemin,

une *hiérarchie*, dite *cohésitive*, orientée, représentant une classe de règles dites généralisées, La hiérarchie est indicée par une ultramétrique (Gras et al, 2005), mesure de qualité de classe de règles, dénommée *cohésion* (FIG.2). Dans ce second exemple, sur les mêmes données, $(OP2 \Rightarrow (OP5 \Rightarrow OP4)) \Rightarrow OP6$ est une classe orientée, encore notée (OP2,(OP5,OP4)),OP6), voire plus simplement (OP2, OP5, OP4, OP6) si la hiérarchie est connue.*(On remarquera la différence informative et structurelle entre ces deux images).*

Algorithmes et représentations, à un seuil choisi par le chercheur, sont implémentés dans le logiciel de traitement de données appelé CHIC (Couturier, 2000) .(cf. FIG 1 et FIG. 2).

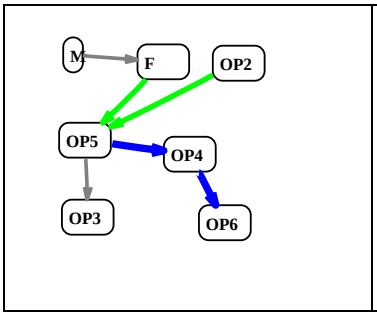
Nous introduisons la notion de variable supplémentaire en A.S.I. à l'instar de la même notion définie en analyse factorielle (Benzecri, 1973). Il s'agit d'une variable extrinsèque, un descripteur par exemple, n'intervenant pas directement dans les liaisons exprimées par la classification entre les variables dites principales de V. Elle n'intervient donc pas dans la représentation de la structure de cet ensemble, qu'il s'agisse du graphe ou de la hiérarchie. Par exemple, une variable supplémentaire pourra représenter une catégorie de sujets (âge, sexe, attitude, catégorie socio-professionnelle, etc.).

Au cours de l'analyse, à 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 de la hiérarchie, est de définir un critère permettant d'identifier un ou des sujets, puis la catégorie de sujets, ou tout autre variable supplémentaire,

---

[3] La règle $a \Rightarrow b$ est dite *admissible* au niveau de confiance 1-α si la probabilité que le nombre de contre-exemples dans les observations soit supérieur au nombre de contre-exemples attendus sous l'hypothèse *H0* d'indépendance entre *a* et *b* est faible, c'est-à-dire si Prob $(Q(a,b) \leq n_{a \wedge \neg b}) \leq \alpha$ où *Q(a,b)* est le nombre aléatoire de contre-exemples à l'implication (cf. l'algorithme de la vraisemblance du lien de I.C. Lerman (Lerman, 1981a). Ce critère d'admissibilité est comparable à celui du philosophe des sciences H. Atlan dans « A tort et à raison. Intercritique de la science et du mythe », Seuil, 1986. Il écrit : « … [en accord avec Jung] si la fréquence des coïncidences n'excède pas de façon significative la probabilité qu'on peut leur calculer en les attribuant au seul hasard à l'exclusion de relations causales cachées, nous n'avons certes aucune raison de supposer l'existence de telles relations ». .

La distribution de la variable aléatoire *Q(a,b)* dépend des hypothèses de tirage : par exemple, une loi hypergéométrique ou une loi binomiale, ou une loi de Poisson (Lerman et al., 1981b)



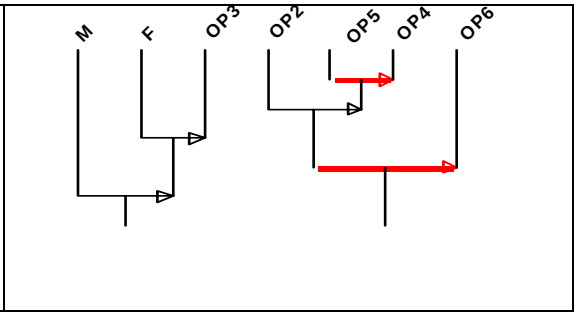


FIG.1 *Graphe implicatif à 7 variables* FIG.2 *Hiérarchie cohésitive à 7 variables*

- ou bien plus ou moins *typiques* du comportement moyen de la population ; en d'autres termes, le comportement de ces sujets sera ainsi en harmonie avec le comportement statistique de la population à l'origine de la classe **C,**

- ou bien *contribuant* le plus à la constitution de **C** ; en d'autres termes, plus ou moins responsables de l'agrégation conduisant à **C**.

Une approche comparable est faite 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[4].

2 Puissance implicative de classe et de chemin

2.1 Couples génériques

L'idée directrice suivie consiste à porter notre attention sur les « lignes de force », (ou, selon une autre métaphore : les « lignes de crête ») des associations, plutôt que de les retenir avec le risque afférent d'être submergé par leur nombre et contraint par les bruits qui les accompagnent. Plaçons-nous à un niveau k de la hiérarchie où viennent de se réunir, pour former **C**, deux classes $\underline{\boldsymbol{A}}$ et $\underline{\boldsymbol{B}}$ telles que $\underline{\boldsymbol{A}} \Rightarrow \underline{\boldsymbol{B}}$. Dans la FIG. 2, au niveau 2, on aurait $\underline{\boldsymbol{A}}$ = OP2 et $\underline{\boldsymbol{B}}$ = (OP5, OP4). Au niveau 4, on lirait : $\underline{\boldsymbol{A}}$ = (OP2, (OP5, OP4)) et $\underline{\boldsymbol{B}}$ = OP6.

Définition 1 : Etant donné les intensités d'implication $\psi(i, j)$ [5], le couple (a,b) tel que: $\forall\, i \in \underline{\boldsymbol{A}}, \forall\, j \in \underline{\boldsymbol{B}} \quad \psi(a, b) \geq \psi(i, j)$ est appelé **couple générique** de **C**[6]. Le nombre $\psi(a, b)$ est appelé **intensité 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

[4] Le travail présenté ici diffère de celui de (Gras et al., 1996a) par la distinction de ces deux notions. Pour l'étude de la responsabilité du sujet dans la similarité, voir par ex. (Lerman, 1981a).

[5] Nous convoquons l'intensité Ψ mais toute la suite est valable avec l'intensité dite classique φ .

6 C'est ce couple, généralement unique, qui intervient par le sup. dans le calcul de l'implication de $\underline{\boldsymbol{A}}$ sur $\underline{\boldsymbol{B}}$ (Gras et al, 1996b).

génériques à l'origine de **C** et g intensités maximales d'implication notées $\psi_1, \psi_2, ..., \psi_g$, qui leur correspondent.

Dans le cas d'un chemin **C,** du graphe implicatif**,** chemin 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 ex. (a,b), on associe, comme pour une classe, l'intensité d'implication $\psi(a,b)$, que l'on dira encore générique.

Définition 2 : 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**. Ce vecteur a la propriété, en ne retenant que les lignes de force (ou de crête) de **C,** de représenter une sorte de « flux » implicatif au sein de la classe.

2.2 Puissance implicative d'un sujet sur une classe ou sur un chemin du graphe et distance à cette classe ou à 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 noterons $\psi x(a,b)$ cette qualité de respect en x de l'implication $a \Rightarrow b$, par exemple et en fonction des valeurs prises en a et b par x:

$\psi x(a,b)=1$ si a=1 ou 0 et b=1; $\psi x(a,b)=0$ si a=1 et b=0 ; $\psi x(a,b)=p$ si a=b=0 avec $p \in]0,1]$. Dans nos premières expériences, nous choisissions p=0.5, valeur neutre7.

Ainsi, à x, nous pouvons associer g nombres $\psi x,1$, $\psi x,2,...,\psi x,g$ correspondant aux g valeurs respectivement prises par x selon les g règles génériques de la classe ou du chemin C.

Définition 3 : Le vecteur $(\psi x,1, \psi x,2,...,\psi x,g)$ est appelé vecteur contingent générique de x ou puissance implicative de x sur C. Le sujet théorique xt qui admettrait $(\psi 1, \psi 2,....,\psi g)$ comme vecteur contingent générique est appelé sujet typique optimal

En effet, on peut interpréter ce vecteur comme étant celui d'un individu « typique » des règles génériques puisque les valeurs prises par ce sujet selon ces règles sont exactement celles de l'ensemble de la population. Ce sujet, image conforme de E, n'existe pas réellement en général. Dans ces conditions, on peut munir l'espace des puissances $[0,1]^g$ d'une métrique afin d'obtenir un contraste accentuant les effets de fortes intensités génériques ou, réciproquement, minorant les effets d'une faible intensité générique.

Définition 4 : On appelle **distance de typicalité d'un sujet quelconque x à la classe ou au chemin C** le nombre: $d(x,\mathbf{C})= \left[\frac{1}{g} \sum_{i=1}^{i=g} \frac{[\psi_i - \psi_{x,i}]^2}{1-\psi_i} \right]^{\frac{1}{2}}$

7 Dans le logiciel CHIC, le calcul des typicalités (et des contributions) se fait cependant en modulant ces valeurs, à l'aide d'une fonction ad hoc, afin de mieux prendre en compte la sémantique des valeurs attribuées par x à a et à b. Par exemple, pour a=0 et b=1, la fonction prend, dans CHIC, la valeur 0.682.

Ce nombre, qui vérifie formellement les 3 axiomes d'une distance, n'est autre également que la distance du type χ^2 entre les deux distributions $\{1-\psi_i\}_i$ et $\{1-\psi_{x,i}\}_i$ qui expriment les écarts entre les implications génériques contingentes et l'implication stricte. Elle exprime, aussi et en particulier, l'écart observé sur les règles génériques entre le sujet considéré x et le sujet théorique typique optimal, écart nuancé par ces intensités. C'est pour cette raison que nous avons choisi le mot **typicalité** pour quantifier le comportement de x selon les règles génériques. Nous allons le préciser plus loin**.** Lorsque $\psi_i = 1$, une légère correction sur cette valeur permet d'éviter la division par zéro (par exemple, prendre $\psi_i = 0.99999999$) ce qui ne change pas fondamentalement la distance.

Remarque : Une classe C étant donnée, on peut définir une structure d'espace métrique sur E par la donnée de la distance indicée par C entre deux sujets quelconques de E, distance qui mesure la différence de comportement des sujets x et y à l'égard de C :

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

On voit alors que la distance de typicalité donnée plus haut n'est que la spécification de d_C aux sujets respectivement x et x_t. La distance d_C permet de conférer à E une **C**-structure topologique discrète. Cette topologie est équivalente à celle qui serait définie sur l'ensemble des vecteurs contingents $(\psi_{x,1}, \psi_{x,2}, \ldots, \psi_{x,g})$, sous-ensemble d'un espace vectoriel normé de dimension g et de norme : $\|\vec{x} - \vec{y}\| = d_C(x,y)$. L'opérateur symétrique associé à la forme quadratique qui conduit à cette distance, a pour matrice la matrice diagonale d'éléments $[g(1-\psi_i]^{-1}$ pour $i=1,\ldots,g$. Il est bien évident que la somme de deux tels vecteurs n'a qu'un sens théorique, c'est-à-dire hors du contexte dans lequel nous travaillons en A.S.I..

Une application intéressante peut consister à déterminer le ou les sujets appartenant à une boule de diamètre donné et de centre l'un des sujets pré-désignés, comme par exemple, l'individu optimal. En prolongement de cette approche métrique, le problème de complétion des données manquantes pourrait y puiser une solution originale.

3 Typicalité, spécificité et contribution d'un sujet et d'une variable supplémentaire à une classe ou à un chemin

3.1 Typicalité

Nous définirons la mesure de typicalité à partir du rapport entre la distance de typicalité relative au sujet considéré et la distance à **C** la plus grande dans l'ensemble des sujets. Cette distance maximale est celle des sujets y dont les $\psi_{y,i}$ sont tous nuls ou très faibles. Ces sujets sont donc les sujets les plus opposés aux règles génériques. La typicalité d'un sujet sera alors d'autant plus grande qu'il s'écartera de ces mêmes sujets, donc qu'il aura un comportement

comparable à celui du sujet théorique optimal. La typicalité d'une catégorie de sujets ou d'une variable supplémentaire G[8] s'en déduira :

Définition 5 : La **typicalité de x à C** est :

$\gamma(x, C) = 1 - \frac{d(x,C)}{\max_{y \in E} (d(y,C))}$ et **celle de G** est : $\gamma(G, \mathbf{C}) = \frac{1}{\text{card}G} \sum_{x \in G} \gamma(x, G)$

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 que l'on définit ci-dessous : groupe optimal et catégorie déterminante.

Définition 6 : 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 des sujets les constituant. Une telle partition est dite *significative*.

L'existence de ce groupe optimal est démontrée dans (Gras R. et al., 1996a et b). 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.

En effet, 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 à deux ou plus modalités binaires, par exemple des catégories socio-professionnelles. Soit X_i une partie aléatoire de E ayant le même cardinal que G_i et Z_i la variable aléatoire Card $(X_i \cap GO(\mathbf{C}))$. Selon un modèle équiprobable, Z_i suit une loi binomiale de paramètres : card G_i et card (GO(**C**)) / card E qui est la fréquence du groupe optimal de la classe ou du chemin **C**.

Définition 7 : On appelle **variable supplémentaire ou 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 $(G_i \cap GO(\mathbf{C}) < Z_i)$].

Ainsi, établir que G_j est la catégorie la plus typique revient à déceler, parmi les catégories, celle dont le nombre de sujets appartenant en même temps au groupe optimal est le plus étonnamment grand *eu égard à son cardinal*. Une catégorie G_0 est dite **déterminante au risque ou au seuil** α si la probabilité associée p_0 est inférieure à α. Autrement dit, le risque de se tromper en affirmant cette propriété est donc au plus égale à α.

Par suite, 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, celle correspondant au risque minimum. Cette approche est comparable à celle de (Lerman,1981a) pour l'analyse des similarités, mais au moyen d'une modélisation et de concepts différents.

[8] Les deux mots « catégorie » et « variable supplémentaire » seront utilisés indifféremment, le premier ayant une charge sémantique plus forte que le second.

D'ailleurs, nous pouvons remarquer qu'il est possible d'associer au 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 alors 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. Nous savons que sa restriction au cas binaire coïncide avec sa valeur s dans le cas où les 2 variables sont binaires.

Ainsi en résumé, il est possible de dégager à la fois les individus et les groupes d'individus typiques d'une règle ou d'un ensemble (classe ou chemin) de règles généralisées. 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 ex. la liaison entre les variables a et b est quantifiée par $\psi(a,b) = 0{,}92$, les individus x qui lui attribuent la valeur $\psi_x(a,b) = 0{,}90$ sont plus typiques que ceux qui lui attribuent la valeur 0,98. Ceux-ci sont à une distance plus grande que les premiers pour le comportement statistique de la population. La nuance entre cette notion et celle de contribution définie dans **2.3** prend tout son sens dans l'étude des variables modales ou numériques.

3.2 Spécificité

Si à chaque classe ou chemin $\mathbf{C_j}$ on peut associer au moins un groupe typique, il est pertinent de mettre en évidence le couple (variable supplémentaire $\mathbf{G_i}$, classe ou chemin $\mathbf{C_j}$) remarquable quant à l'optimalité de sa conjugaison. D'où la définition :

Définition 8 : La variable supplémentaire $\mathbf{G_i}$ étant donnée, le couple ($\mathbf{G_i}$, $\mathbf{C_j}$) est dit **mutuellement spécifique** lorsque $\mathbf{G_i}$ est la variable la plus spécifique de la règle associée à $\mathbf{C_j}$ et lorsque la probabilité (le risque) p_i^k de G_i par rapport aux autres classes de la hiérarchie ou aux chemins $\mathbf{C_k}$ du graphe implicatif est supérieure à un seuil β (à la discrétion de l'utilisateur).

Une analyse étant donnée, il peut exister 0 ou plusieurs couples mutuellement spécifiques. Ce ou ces couples offrent l'intérêt de faire porter l'attention de l'expert sur les plus fortes associations prenant origine dans une variable supplémentaire.

Définition 9 : De la même façon, un individu x étant donné, le couple (x, $\mathbf{C_j}$) est **mutuellement spécifique** lorsqu'il appartient au groupe optimal relatif à la règle associée à $\mathbf{C_j}$ et que sa typicalité à $\mathbf{C_j}$ est maximale par rapport à toutes ses autres typicalités aux classes de la hiérarchie cohésitive ou aux chemins du graphe implicatif.

3.3 Contribution

Cette notion se distingue de la précédente, ce que nous ne faisions pas en 1996, par l'examen de la responsabilité des individus, puis des variables supplémentaires -qui peuvent en être des descripteurs- *à l'existence* d'une règle ou d'une règle généralisée 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 : $a \Rightarrow b$ (resp. règle i du chemin **C** ou bien de la classe **C** constituée de g règles génériques) supposée admissible, on donne la

Définition 10 : On appelle **distance de contribution de x à (a,b) ou à C :**

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

Cette distance, de type euclidien, mesure l'écart entre le vecteur **contingent générique** de x, ($\psi_{x,1}$, $\psi_{x,2}$,...,$\psi_{x,g}$), et le vecteur à g composantes (1,1,1,…, ,1). Ce dernier est le vecteur d'un *sujet théorique optimal qui satisferait strictement toutes les règles génériques.* C'est en ce sens que typicalité et contribution sont distinctes.

Remarque : A l'instar de ce que nous avons fait pour la typicalité, nous pouvons définir sur E une topologie discrète d'espace normé dont la norme est associée à la distance entre deux sujets quelconques suivante : $d_C(x,y) = \left[\frac{1}{g} \sum_{i=1}^{i=g} \left[\psi_{x,i} - \psi_{y,i} \right]^2 \right]^{\frac{1}{2}}$.

Définition 11 : On appelle **contribution de x à C** le nombre : $\gamma(x,\mathbf{C}) = 1-d(x,\mathbf{C})$.

Cette définition est la restriction de celle de la typicalité au cas où, cette fois, on compare le sujet x aux « pires » sujets par rapport aux règles génériques : leur comportement s'oppose à l'implication de chaque règle (1 pour la prémisse et 0 pour la conclusion). 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. En effet, plus la différence est importante, plus le sujet observé a un comportement voisin de celui du sujet théorique optimal et plus il s'éloigne de ceux qui réfutent les règles génériques : on peut donc dire qu'en contribuant à l'émergence de la classe, ils en sont responsables.

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 , couple mutuellement spécifique) se transpose immédiatement à partir des principes de la typicalité et la spécificité. Mais dans les situations réelles, nous observons la nuance entre les deux concepts ce qui enrichit l'information exploitable par l'utilisateur. Cependant, *le concept de contribution est plus volontiers retenu pour l'interprétation dans une perspective inductive*.

4 Application

Dans le cadre d'une enquête de l'Association des Professeurs de Mathématiques de l'Enseignement Public (APMEP) auprès de professeurs de mathématiques de classes terminales (séries scientifiques S et ES, littéraires LI et technologiques TE sont les variables supplémentaires), nous avons recueilli et analysé (Bodin et al., 1999) les réponses de 311 professeurs, à des classements (de 1 à 6) portant sur quinze objectifs qu'ils assignent à leur enseignement (A, B, C, …O)[9] et sur leurs opinions relatives à dix phrases susceptibles d'être communément énoncées (OP1, OP2,..OPX)[10]. La variable PER donne la possibilité d'énoncer les objectifs jugés non pertinents. Les 26 variables correspondantes ne sont pas binaires, sauf PER, mais ordinales (valeurs (1, 0.8, 0.6, 0.4, 0.2, 0.1, 0) pour les objectifs et (1, 0.5, 0) pour les opinions). Ainsi l'analyse intègre l'intensité des attitudes, d'un choix prioritaire d'un objectif à un choix plus secondaire, voire non retenu.

Les occurrences des 26 variables sont les suivantes :
A : 105.70 B : 8.80 C : 9.70 D : 140.00 E : 21.80 F : 138.70 G : 19.50 H : 44.80
I I : 83.10 J : 108.40 K : 77.60 L : 4.60 M : 90.20 N : 66.60 O : 33.20
OP1 : 81.50 OP2 : 147.50 OP3 : 242.50 OP4 : 229.00 OP5 : 190.00 OP6 : 240.00 OP7 : 200.00 OP8: 165.00 OP9 : 98.00 OPX : 207.00 PER : 254.

Les occurrences des variables supplémentaires sont :
S(cientifique) : 155 ES(économique et sociale) : 68 LI(ttéraire) : 22 TE(chnologique) : 66. La hiérarchie cohésitive obtenue par CHIC à partir d'un nombre réduit des variables, afin de conserver les niveaux les plus significatifs, est donnée par la FIG. 3.

Considérons la classe **C**= [E $\Rightarrow$ (OP8 $\Rightarrow$ OP7)] $\Rightarrow$ OPX. Son sens, analysé plus en détail dans (Bodin et al., 1999), est fortement marqué par l'importance accordée à l'imagination et à la recherche personnelle, par les enseignants d'accord avec ces objectifs et ces opinions, La variable la plus **typique** pour cette classe est S (série Scientifique) avec un risque de : 0.00393.

En effet, 116 des enseignants de S parmi les 155 de cette série qui ont répondu au sondage, figurent dans le groupe optimal **(GO)** de cardinal 201 relatif à **C**. Soit X une partie aléatoire de même cardinal (155) que S et Z la variable aléatoire égale au cardinal de l'intersection de X et du groupe optimal **GO**. Selon un modèle équiprobable de distribution des enseignants, Z suit la loi binomiale de paramètres 155 et 201/311 soit 0.656. La probabilité pour que Z soit plus grande que 116 est le risque annoncé, soit 0.00393**.** Mais pour **S**, c'est le couple (**S, (OP8, OP7**)) qui est mutuellement spécifique au seuil $\beta = 2.10^{-5}$.

[9] Par exemple, E symbolise l'objectif : « développement de l'imagination et de la créativité »

[10] Par exemple, OP4 symbolise : « Pour corriger, j'aime bien un barème très détaillé sur les résultats à obtenir »

On retrouve une telle mutuelle spécificité pour **TE** avec le couple **(TE, (B,K))** à un seuil 5.10^{-7} nous confirmant, sans surprise, que les enseignants des sections techniques **(TE)** considèrent que les mathématiques doivent être utiles à la vie professionnelle (**B**) et, en conséquence, aux autres disciplines (**K**) et y sont les plus attachés.

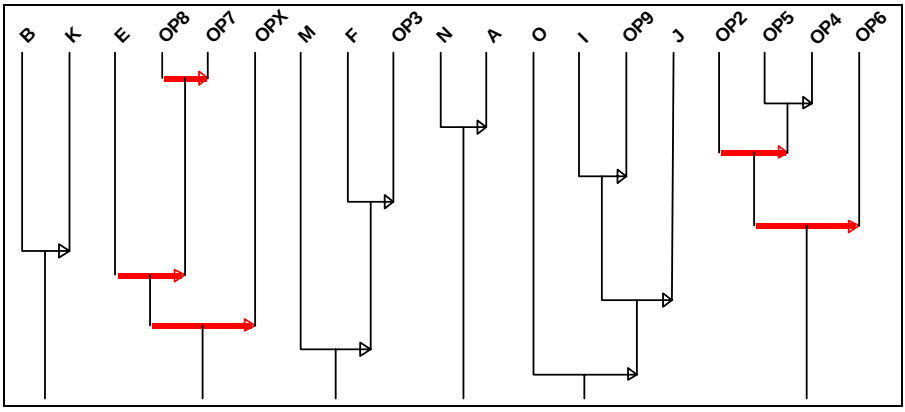


FIG. 3 - *Hiérarchie cohésitive significative*

Les calculs de **contribution** à la classe **C** montrent que, cette fois, 111 enseignants sur les 311 sondés, participent au groupe optimal. Le nombre d'enseignants de **S** a diminué (il passe de 116 à 67) et, surtout, sa proportion est bien moindre que précédemment dans le **GO**. Ceci se ressent dans le seuil qui est 0.0251, soit un risque 6 fois plus élevé que pour la typicalité. Ce sont les enseignants sondés de **S** qui sont les plus typiques, c'est-à-dire « conformes » au comportement général de la population elle-même sondée. Mais ils sont moins contributeurs dans les relations strictes entre les 4 variables constituant **C**. Cette remarque nous montre les nuances apportées par les deux concepts : typicalité et contribution

Certaines liaisons apparues et commentées ci-dessus se retrouvent dans le graphe de la FIG. 4. Les contributions calculées dans CHIC montrent encore que les enseignants de la série S contribuent le plus au chemin : **E** $\Rightarrow$ **OP8** $\Rightarrow$ **OP7** $\Rightarrow$ **OPX** avec un risque d'erreur de 0.00746, la transitivité le long de ce chemin étant assurée au niveau 0.75.

5 Conclusion

Les applications de la méthode A.S.I. ont d'ores et déjà donné des résultats très satisfaisants, non seulement dans la discipline où elle a pris naissance, la didactique des mathématiques, mais aussi dans d'autres domaines de l'éducation ou de recherche scientifique différente (biologie, économie,…) comme l'a montré la 3ème Rencontre Interna nationale ASI 3 de Palerme en octobre dernier. Le plus souvent, les interprétations des experts s'appuient complémentairement sur l'analyse de similarités ou/et sur les méthodes factorielles, tout en obtenant des informations qui sont spécifiques de l'A.S.I. en raison de son caractère non

symétrique. Mais ces méthodes visent un objectif commun : l'accès à la signification d'un *tout* non réduit à la somme des significations de la somme de ses *parties*.[11]

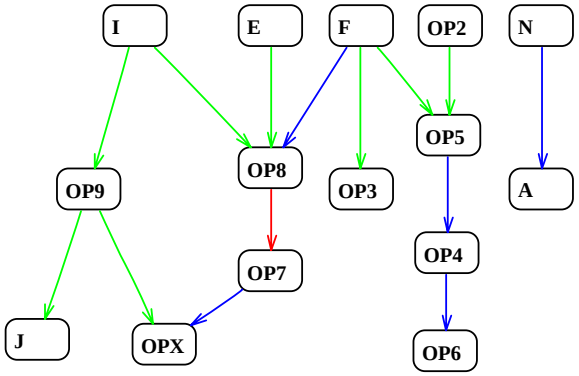


FIG. 4 - *Graphe implicatif au niveau de confiance 0.90*

Les analyses bénéficient efficacement du logiciel C.H.I.C., qui permet, avec une certaine convivialité, tous les traitements algorithmiques et graphiques des questions évoquées dans cet article. Son développement suit régulièrement toutes les nouvelles avancées de la théorie de l'implication statistique. Ses fonctions respectives de révélateur et d'analyseur qui semblent opérer avec bonheur dans de multiples domaines nous promettent encore d'intéressantes perspectives théoriques et appliquées.

Références

Agrawal, R., T. Imielinski. Et A. Swami (1993). Mining association rules between sets of items in large databases. *In the1993 ACM SIGMOD international conference on management of data, ACM Press*

Benzecri, J.P. (1973). L'analyse des données (vol 1), *Dunod, Paris*

Bodin, A. et R. Gras,. (1999). Analyse du préquestionnaire enseignants avant EVAPM-Terminales. *Bulletin de l'Association des Professeurs de Mathématiques, n° 425*, 772-786

Blanchard, J., P. Kuntz, F. Guillet., R. Gras, (2003). Implication intensity: from the basic statistical definition to the entropic version , *Statistical Data Mining and Knowledge Discovery, Chapman & Hall/CRC, Washington*, 473-485

[11] C'est bien ce qu'affirme, au même titre, le philosophe L. Sève : « …dans le passage non additif, non linéaire des parties au tout, il y a *apparition de propriétés* qui ne sont d'aucune manière *précontenues* dans les parties et qui ne peuvent donc s'expliquer par elles » (« Emergence, complexité et dialectique », Odile Jacob, mai 2005).

Couturier, R. (2000) Traitement de l'analyse statistique implicative dans CHIC, *Actes des Journées Fouille des données par la méthode d'analyse implicative, IUFM Caen*, 33-50

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, Rennes 1*

Gras, R. et H. Ratsimba-Rajohn. (1996a). Analyse non symétrique de données par l'implication statistique*. RAIRO-Recherche Opérationnelle, 30(3)*, 217-232

Gras, R., S. Ag Almouloud, M. Bailleul , A. Larher., M. Polo, H. Ratsimba-Rajohn et A. Totohasina, (1996b*). L'implication Statistique, La Pensée Sauvage, Grenoble*

Gras, R. (2000). I fondamenti dell'analisi statistica implicativa, *Quaderni di ricerca in didattica, GRIM Palerme, 9* (http://math.unipa.it/~grim/Rgras_pubb_postPhD.htm)

Gras, R., P. Kuntz. H. Briand (2001).Les fondements de l'analyse statistique implicative et quel- ques prolongements pour la fouille de données, *Math et Sc. Humaines, n° 154-155*, 9-29

Gras, R, P. Kuntz.et J.C. Régnier (2004). Significativité des niveaux d'une hiérarchie orientée, *Classification et fouille de données, RNTI-C-1, Cépaduès- Editions*, 39-50

Gras, R. et P. Kuntz (2005). Discovering R-rules with a directed hierarchy, *Soft Computing, A fusion of Foundations, Methodologies and Applications,Vol. 1*

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

Lerman, I.C., R. Gras et H. Rostam, Elaboration et évaluation d'un indice d'implication pour données binaires , *Mathématiques et Sc. Humaines, n°74*, 5-35.

Régnier J.C. et R. Gras (2005). Statistique de rangs et analyse statistique implicative, *Revue de Statistique Appliquée,* LIII (1), 5-38

Sebag M. et Schoenauer (1991), Un réseau de règles d'apprentissage, *Induction symbolique-numérique à partir de données*, *Cépaduès Editions*

Summary

Implicative statistical analysis deals with tables of variable x individuals in order to extract statistical rules and meta-rules between variables. These rules are used to build an implicative graph or an oriented hierarchy. This paper presents two concepts, which explain respectively, using the previous structures, the *typicality* and the *contribution* of individuals (called additional variables) towards a path in the implicative graph or a cluster in the oriented hierarchy. The typicality measures the proximity between individuals and the average behaviour of the population towards a given statistical rule. The contribution measures the relation of individuals with an associated logical rule. We show the interest of these notions of typicality and contribution through an example with the help of the CHIC software program.