UN MARIAGE ARRANGÉ ENTRE L'IMPLICATION ET LA CONFIANCE ?

Régis GRAS[1], Raphaël COUTURIER[2], Pablo GREGORI[3]

RÉSUMÉ

La plupart des indices d'association entre variables binaires utilisent la fréquence conditionnelle qu'ils appellent confiance ou expressions algébriques d'instanciations pour décider d'une liaison entre deux variables et en apprécier la qualité. En Analyse Statistique Implicative, une autre mesure, l'intensité d'implication, vise le même objectif en se limitant à l'implication tout en s'appuyant plutôt sur la probabilité d'apparition des contre-exemples à ce type de liaison. Dans cet article nous comparons ces deux mesures en montrant qu'elles sont étrangères mais possèdent des relations analytiques intéressantes. De ce fait, nous concevons et expérimentons une nouvelle mesure de qualité d'implication en deux approches qui associe confiance et intensité. Nous montrons l'intérêt présenté par cette combinaison pour intégrer la contraposée de l'implication, condition nécessaire pour faire jouer à ce nouvel indice une fonction d'analyse causale.

***Mots-clés* :** confiance, fréquence conditionnelle, intensité d'implication, intensité entropique, implifiance, règle, quasi-règle, contraposée

[1] École Polytechnique de l'Université de Nantes, Équipe DUKE Data User Knowledge, Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR 6241, : regisgra@club-internet.fr

[2] FEMTO-ST, Université de Franche-Comté, Belfort, raphael.couturier@univ-fcomte.fr

[3] Dept. Matemàtiques, Universitat Jaume I, Campus del Riu Sec, Castelló E-12071 (SPAIN), gregori@mat.uji.es , http://www3.uji.es/~gregori

ABSTRACT

Most of the indexes used with association rules are based on the conditional probability, also called confidence, to measure the quality of a rule. With Statistical Implicative Analysis, another measure, the implication intensity, focuses on the same objective using only the implication. It is built with the probability of appearance of counter-examples of a rule. In this paper, we compare those two measures, showing they are different but also have interesting analytical properties. Hence, a new measure based on those two approaches, confidence and implication is proposed. The interest of this combination is to integrate the contrapositive of the implication, in order to give this new index a causal property.

Keywords : confidence, conditional probability, implication intensity, entropic intensity, implifiance (implidence), rule, quasi-rule, contrapositive

1 Introduction

1.1 Motivation

Nous nous plaçons dans le cadre du croisement d'un ensemble de sujets et d'un ensemble de variables binaires selon lesquelles les sujets prennent leurs valeurs. De nombreux auteurs (voir Références) ont établi des indices qui permettent de mettre en évidence des relations implicatives entre des variables binaires telles que, par exemple, a et b. Parmi ces indices et très souvent, ils font référence à la confiance, fréquence conditionnelle de b sachant a et établissent leurs indices d'association sur celle-ci. Dans cet article, nous établirons une relation entre l'intensité d'implication et la confiance, relation qui soulignera à la fois la différence entre ces deux indicateurs et leur éventuelle covariation. *De plus, pour des raisons sémantico-statistiques, nous les combinerons pour construire un nouvel indice composite d'implication statistique qui aura la vertu d'associer la tendance inclusive donnée par la confiance et l'étonnement statistique par l'intensité d'implication.*

Nous présentons également une nouvelle approche de la modélisation en A.S.I. pour des variables binaires prenant en compte autant l'implication directe que sa contraposée. Elle vise à se substituer à la modélisation entropique jusqu'alors utilisée et qui présente un caractère jugé trop ad-hoc par les familiers de l'A.S.I. Elle va donc être construite *contre une connaissance antérieure* comme le dit G. Bachelard dans (Bachelard G. 1967). Cependant, cette précédente modélisation était loin de déplaire aux utilisateurs qui en appréciaient la capacité à accepter plus facilement la grande taille de l'échantillon des sujets considéré. D'où son intérêt pour ce que l'on appelle les « big data ». De plus, prenant en compte la contraposée, elle semble remplir le rôle d'extracteur de relations causales, ce que ne permet pas ou moins finement l'implication dite classique mesurée par l'intensité d'implication. En effet, celle-ci conduit à la même valeur d'intensité autant pour l'implication directe que pour sa contraposée. Elle ne permet pas ainsi d'intégrer l'information nuancée apportée par la contraposée. Rappelons par contre qu'en cas d'implication logique stricte les deux formes

implicatives sont équivalentes. Ce qui n'est pas le cas de la quasi-implication, objet central de nos recherches. Ainsi, cette nouvelle modélisation de l'implication entre deux variables binaires, tout en intégrant à la fois l'implication directe et sa contraposée, mais également la confiance, va adopter une méthodologie comparable à celle utilisée dans le cas de l'implication mesurée par l'intensité d'implication définie par R. Gras (Gras, 1979) à la base des premiers fondements de l'A.S.I.

1.2 Remarque épistémologique importante

Il pourrait être objecté à l'égard des auteurs de cet article que ce nouveau concept soit seulement le fruit d'un mariage arrangé entre deux partenaires étrangers l'un à l'autre, regardant dans des directions différentes. Or, l'un et l'autre alimentent l'objectif que nous nous sommes donné sachant que nous traitons essentiellement des problèmes qui relèvent du réel comme en sciences humaines ou relativement à des phénomènes scientifiques non entièrement modélisés (médecine, biologie, etc.) : rendre compte de l'implication en serrant au plus près la notion logique de celle-ci sachant que leurs célibats respectifs ne le permettent pas. Nous prétendons, en plagiant G. Vergnaud (Vergnaud, 2007) s'exprimant au sujet de l'esthétisme, qu'il n'y a pas d'analyse de données sans **confiance**. Mais il n'y pas non plus d'analyse de données sans **surprise**[4], ni sans **correction d'échelle** comme il est fait en analyse en composants principales, par exemple, où une division par $\sqrt{n}$ s'impose dans la détermination de la distance entre sujets. Accoupler confiance et surprise n'a rien de plus artificiel que celui auquel sont soumis les physiciens de l'acoustique, de la thermodynamique, de l'électricité, de la relativité, etc.

Par exemple, sur le problème de chute des corps, l'observateur relève expérimentalement l'influence de certaines variables qui expriment que la résistance de l'air est proportionnelle au carré de la vitesse V, proportionnelle à la surface S du maître-couple du solide, à la masse spécifique r_0 de l'air, à un coefficient de forme C. Tout ceci que l'on résume par la formule : $R=kCr_0SV^2$. C'est une relation construite sur

[4] C'est aussi ce qu'affirme René Thom (« Paraboles et catastrophes », 1980, p.130) : « …*le problème n'est pas de décrire la réalité, le problème consiste bien plus à repérer en elle ce qui a de sens pour nous, ce qui est surprenant dans l'ensemble des faits*. *Si les faits ne nous surprennent pas, ils n'apportent aucun élément nouveau pour la compréhension de l'univers : autant donc les ignorer* » et plus loin : « *... ce qui n'est pas possible si l'on ne dispose pas déjà d'une théorie* ».

la base de mesures empiriques qui conduisent à la satisfaire. Il en est de même de la formule expérimentale des gaz parfaits : pV=Cte et toutes les règles énoncées sous le terme de principe[5].

Bien entendu, d'autres phénomènes physiques seront plongés par essence dans un modèle mathématique déduit (par ex. les phénomènes vibratoires). Et d'autres seront extraits de l'expérimentation et de la mise en évidence des variables intervenantes. Nous pourrons dire que notre modèle est bon lorsque les résultats expérimentaux que nous mènerons avec satisferont à la fois l'explicabilité, le bon sens et l'observation mais aussi auront une capacité prédictive. D'autres modèles concurrents du concept implicatif de notre étude présente existent et nous avons mis en évidence leurs adéquation ou leurs limites vis-à-vis des variables requises ou de leur pondération (Gras et Couturier, 2010). De toute façon, quel qu'il soit, le modèle utilisé est bon s'il rend compte au mieux de la réalité jusqu'à sa substitution éventuelle par un autre plus performant. Or c'est cette performance que nous voulons atteindre (cf. aussi note 2).

2 Relation entre confiance et intensité d'implication.

2.1 Rappels sur l'intensité d'implication

Relativement aux cardinaux de E (soit n), espace des sujets, de A (soit n_a), sujets satisfaisant a et de B (soit n_b), sujets satisfaisant b, c'est le poids des contre-exemples (soit $n_{a\wedge\bar{b}}$), i.e. des sujets satisfaisant a et non b, qu'il faut donc prendre en compte pour accepter statistiquement de conserver ou non la **quasi-implication** ou **quasi-règle** $a \Rightarrow b$. Ainsi, c'est à partir de la dialectique entre les exemples et les contre-exemples que la règle apparaît comme le dépassement de la contradiction.

Pour formaliser cette quasi-règle, nous formulons l'hypothèse que a et b sont indépendantes. Puis nous considérons, comme le fait I.C. Lerman dans (Lerman, 1981) pour la similarité, deux parties quelconques X et Y de E, choisies aléatoirement et

[5] « On nomme **principe physique** une loi physique apparente, qu'aucune expérience n'a invalidée jusque là bien qu'elle n'ait pas été démontrée, et joue un rôle voisin de celui d'un postulat en mathématiques ». (Wikipedia).
« Parfois un principe peut être démontré à partir d'un ou plusieurs autres, à charge au physicien de choisir le principe de base pour ses raisonnements : par exemple, le principe de moindre action est équivalent au principe fondamental de la dynamique associé au principe de d'Alembert » (Wikipedia).

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 ensembles complémentaires respectifs de Y et de B dans E de même cardinal $n_{\bar{b}} = n\text{-}n_b$. Soit α un réel quelconque de l'intervalle [0,1].

Définition 1: la quasi-règle $a \Rightarrow b$ est *admissible avec l'intensité* 1-α si et seulement si $\Pr\left[\text{Card}(X \cap \overline{Y}) \leq \text{card}(A \cap \overline{B})\right] \leq \alpha$.

Définition 2: On appelle intensité d'implication de la quasi-règle $a \Rightarrow b$, pour la modélisation précédente, le nombre $\varphi(a,b) = 1 - \Pr\left[\text{Card}(X \cap \overline{Y}) \leq \text{Card}(A \cap \overline{B})\right]$ si $n_b \neq n$ et $\varphi(a,b) = 0$ si $n_b = n$.

On démontre (Gras R, 1995, puis autre preuve en 2009) que, pour une certaine modélisation de tirage de X et de Y, $\text{Card}(X \cap \overline{Y})$ suit une loi de Poisson de paramètre $\lambda = \frac{n_a n_{\bar{b}}}{n}$.

L'intensité d'implication est une valeur probabiliste, contrairement aux indices implicatifs les plus usités, qui fonde la décision de retenir ou non une relation de quasi-implication entre les variables binaires a et b. Intuitivement, elle représente une sorte d'étonnement statistique (une *surprise* ?) que le nombre de contre-exemples à $a \Rightarrow b$ soit petit alors qu'elles sont supposées indépendantes.

2.2 La confiance en tant que réalisation d'une variable aléatoire

Avec les mêmes notations, dans un corpus de données binaires, la confiance c est la fréquence conditionnelle pour qu'un sujet satisfasse à b sachant qu'il satisfait à a soit

$$c = \text{Fr}[b|a] = \text{Fr}[a \wedge b|a] = [(\text{card } A \cap B) / \text{card } A] = \frac{n_{a\wedge b}}{n_a} = \frac{n_{a\wedge b}/n}{n_a/n} = 1 - \frac{n_{a\wedge \bar{b}}}{n_a}$$

La variable aléatoire card $(X \cap Y) = N_{a\wedge b}$ est réalisée par card $(A \cap B)$ c'est-à-dire $n_{a\wedge b}$. Aussi, comme nous l'avons fait pour $N_{a\wedge \bar{b}}$ (Gras et Régnier, 2013) et selon une modélisation comparable, nous démontrerions que $N_{a\wedge b}$ suit approximativement, sous l'hypothèse d'indépendance, une loi de Poisson de paramètre estimé $\frac{n_a n_b}{n}$.

Considérons c comme la réalisation d'une variable aléatoire C fonction uniquement de card $(X \cap Y)$ définie par : $C = \frac{N_{a\wedge b}}{n_a}$. Cette confiance aléatoire conditionnelle prend ses valeurs réelles sur [0,1] et a pour loi la loi image, c'est-à-dire transportée de celle de $N_{a\wedge b}$. Ses valeurs sont donc des fractions des valeurs entières dont le numérateur est Poissonnien.

D'où $\Pr[C \geq \alpha] = \Pr[1 - \frac{N_{a\wedge \bar{b}}}{N_a} \geq \alpha] = \Pr[\frac{N_{a\wedge \bar{b}}}{N_a} \leq 1-\alpha]$

Or dans la contingence, la variable aléatoire N_a prend la valeur fixée et égale à n_a.

Par suite, $\Pr[C \geq \alpha] = \Pr[N_{a\wedge \bar{b}} \leq n_a(1-\alpha)]$.

2.3 Comparaison intensité d'implication et confiance

Rappelons que $Q(a,\bar{b}) = \dfrac{\mathrm{Card}(X \cap \bar{Y}) - \dfrac{n_a n_{\bar{b}}}{n}}{\sqrt{\dfrac{n_a n_{\bar{b}}}{n}}}$ est la variable aléatoire gaussienne centrée réduite limite de la variable de Poisson $\mathrm{Card}(X \cap \bar{Y})$ ou $N_{a\wedge\bar{b}}$, nombre aléatoire de contre-exemples à l'implication. D'où l'on tire :

$$\Pr[N_{a\wedge\bar{b}} \leq n_a(1-\alpha)] = \Pr\left[Q(a,\bar{b}).\sqrt{\frac{n_a n_{\bar{b}}}{n}} + \frac{n_a n_{\bar{b}}}{n} \leq n_a(1-\alpha)\right].$$

Posons $r(a,\bar{b}) = \dfrac{n_a(1-\alpha) - \dfrac{n_a n_{\bar{b}}}{n}}{\dfrac{n_a n_{\bar{b}}}{n}}$. Alors $\Pr[C \geq \alpha] = \dfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{r(a,\bar{b})} e^{-\frac{t^2}{2}} dt$

On vérifie bien que $\Pr[C \geq \alpha]$ est fonction décroissante de α.

A titre de comparaison, rappelons que $\varphi(a,b) = 1 - \Pr\left[Q(a,\bar{b}) \leq q(a,\bar{b})\right] = \dfrac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$. Posons : $\dfrac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt = \alpha$.

Alors : si $\dfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{r(a,\bar{b})} e^{-\frac{t^2}{2}} dt \geq \dfrac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$ la confiance est plus grande que l'intensité d'implication, sinon, c'est cette dernière qui l'emporte.

Or, avec les mêmes notations, pour α donné, $[\varphi(a,b) \geq \alpha]$ qui est aussi équivalent à $\Pr[N_{a\wedge\bar{b}} \geq n_{a\wedge\bar{b}}] \geq \alpha$ nous rappelle que l'évènement $[\varphi(a,b) \geq \alpha]$ est de moins en moins probable lorsque α croît. Nous y reviendrons plus loin.

Autrement dit, encore, partant de l'inégalité : $n_{a\wedge\bar{b}} \leq N_{a\wedge\bar{b}} \leq n_a\,(1-\alpha)$, pour un seuil comparable, si l'inégalité de gauche est satisfaite l'implication est de bonne qualité. Si celle de droite l'est, ce sera la confiance.

Les deux concepts (confiance et intensité d'implication) répondent donc à des principes relativement distincts mais non contradictoires : la confiance Fr[b|a] est fondée sur la subordination de la variable b à la variable a alors que l'intensité d'implication se fonde sur les contre-exemples à la relation de sujétion de b par a.

Ainsi, si l'on souhaite une bonne qualité d'implication (par ex. α= 0.95), il est nécessaire que le nombre de contre-exemples obtenus par hasard soit plus grand que celui de la contingence. En revanche, si l'on veut que la confiance C ait des chances d'être grande (par ex. β = 0.95), il est nécessaire que le nombre d'exemples que donnerait le seul hasard ne soit pas plus grand que 0.05 n_a. On note alors que la probabilité de la variable confiance ne dépend plus de n_b, contrairement à l'intensité d'implication.

Rappelons, à ce sujet, que nous avons prouvé que lorsque le nombre de sujets vérifiant b tend vers n ou est égal à n, la relation implicative devient triviale, banale alors que la confiance sera maximale. Nous avons également montré que la confiance ne varie pas dans toute dilatation de E et des exemples de A et de B. Ce qui nous a conduits à refuser le critère « fréquence conditionnelle » comme seul critère de révélation d'une relation implicative. Elle est cependant, le fondement de la notion de réseau Bayesien qui permet d'organiser un ensemble de variables selon un graphe, la transition d'un nœud du graphe au suivant se faisant sur la base de la fréquence conditionnelle du premier sur le second.

Remarque prospective

Il serait alors possible, dans le cadre d'une recherche, de définir un autre indice probabiliste comme l'est φ(a,b), pour évaluer la qualité de l'implication de a sur b.

3 Lemme

Le comportement asymptotique de φ(a,b), intensité d'implication, est celui d'une variable uniforme sur l'intervalle [0 ,1].

Ce lemme a été démontré dans Gras et al (1996), mais sera rappelé et illustré par un exemple numérique dans l'annexe. Nous utiliserons ce résultat dans la suite de cet article.

4 Indicateur de proximité asymptotique entre confiance et intensité d'implication

Ainsi, $\Pr[\varphi(a,b)\leq\alpha]\sim\alpha$ ou, de façon équivalente vu la continuité de la mesure, $\Pr[\varphi(a,b)\geq\alpha] \sim 1-\alpha$. Or : $\Pr[C\geq\alpha]= 1 - \frac{1}{\sqrt{2\pi}} \int_{r(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$

Dans ces conditions, le rapport $\Pr[C\geq\alpha]/\Pr[\varphi(a,b)\geq\alpha] \sim \frac{\Pr[C\geq\alpha]}{1-\alpha}$ est un bon indicateur de satisfaction entre confiance et intensité d'implication : plus grand que 1, la confiance est alors meilleure que l'intensité ; inférieure à 1, c'est l'intensité qui est plus forte. Une recherche ultérieure pourrait s'appuyer sur cet indicateur.

5 Une première approche d'une intensité d'implication intégrant la contraposée

Nous présentons cette approche pour des raisons didactiques : partant d'une intuition, un développement théorique se construit mais son débouché conduit à un problème que les auteurs avaient dénoncé et déjà tenté de résoudre. Cela illustre cependant la démarche du chercheur dont les errements peuvent mener à des impasses.

L'implication $a \Rightarrow b$, où $n_a \leq n_b$, pour des variables binaires, est d'autant plus réalisée que $n_{a\wedge b}$ est peu différent de n_a. , c'est-à-dire que l'inclusion de A dans B est vérifiée.

La contraposée de cette implication, $\bar{b} \Rightarrow \bar{a}$, est elle-aussi d'autant plus réalisée que l'inclusion de $\bar{B}$ dans $\bar{A}$ est validée, c'est-à-dire que $n_{a\wedge b} + n_{\bar{a}\wedge\bar{b}}$.est peu différent de $n_a + n_{\bar{b}}$, condition nécessaire à la satisfaction de bonnes qualités d'implication tant directe que contraposée.

	b	$\bar{\mathbf{b}}$	**Totaux**
a	$n_{a\wedge b}$	$n_{a\wedge\bar{b}}$	$\mathbf{n_a}$
$\bar{\mathbf{a}}$	$n_{\bar{a}\wedge b}$	$n_{\bar{a}\wedge\bar{b}}$	$\mathbf{n_{\bar{a}}}$
Totaux	$\mathbf{n_b}$	$\mathbf{n_{\bar{b}}}$	**n**

Tab. 1

Soit $N_{a\wedge b}$ et $N_{\bar{a}\wedge\bar{b}}$ les variables aléatoires qui représentent les cardinaux des sous-ensembles aléatoires $X \cap Y$ et $\overline{X} \cap \overline{Y}$ de E respectivement associés à $A \cap B$ et $\overline{A} \cap \overline{B}$ et de mêmes cardinaux. Posons $T = N_{a\wedge b} + N_{\bar{a}\wedge\bar{b}}$ et $t = n_{a\wedge b} + n_{\bar{a}\wedge\bar{b}}$ sa réalisation dans l'expérience de tirage uniforme.

Définition 3

On appelle **intensité d'implication inclusive** associée à $a \Rightarrow b$, dans l'hypothèse d'indépendance a priori de a et b, la probabilité :

$K(a,b) = Pr[T \leq t]$ ou explicitement : $K(a,b) = Pr[N_{a\wedge b} + N_{\bar{a}\wedge\bar{b}} \leq n_{a\wedge b} + n_{\bar{a}\wedge\bar{b}}]$

Cette intensité inclusive est la probabilité pour que, si les variables a et b n'avaient aucun lien entre elles, le hasard des inclusions de A dans B et de leurs complémentaires inversés conduise globalement à plus d'exemples que ceux qui ont été observés. Autrement dit, plus les quasi-inclusions de A dans B et de $\overline{B}$ dans $\overline{A}$ seront proches des inclusions strictes, plus le nombre des seuls exemples satisfaisant les inclusions et dues au hasard sera réduit.

Par suite : $Pr[T \leq t] = \sum_{k=0}^{k= n_{a\wedge b} + n_{\bar{a}\wedge\bar{b}}} Pr[N_{a\wedge b} + N_{\bar{a}\wedge\bar{b}} = k]$

Choix du modèle hypergéométrique de T.

Adoptant une distribution uniforme sur l'ensemble des sujets la loi de T est hypergéométrique. Par suite :

$$Pr[T \leq t] = \sum_{k=0}^{k= n_{a\wedge b} + n_{\bar{a}\wedge\bar{b}}} . \sum_{l=0}^{l=n_{a\wedge b}} Pr[N_{a\wedge b}=l,\ N_{\bar{a}\wedge\bar{b}} = k - l]$$

$$= \sum_{k=0}^{k= n_{a\wedge b} + n_{\bar{a}\wedge\bar{b}}} \sum_{l=0}^{l=n_{a\wedge b}} \frac{C^{l}_{n_{a\wedge b}} . C^{k}_{n_{a\wedge b}-l}}{C^{l}_{n_{a\wedge b}}}$$

Les valeurs de ce nouvel indice d'implication inclusif utiliseront les moyenne et variance de la loi hypergéométrique et, de ce fait, pourront, dans la plupart des cas, être ramenés à ceux d'une loi binomiale, voire gaussienne comme il est fait dans le cas classique. Une étude ultérieure approfondie serait intéressante en partant de la loi de T.

Remarque

Une modélisation différente pourrait consister à définir séparément les lois de probabilité des variables $N_{a\wedge b}$ et $N_{\bar{a}\wedge\bar{b}}$ en comparant leurs variations aux réalisations respectives à $n_{a\wedge b}$ et $n_{\bar{a}\wedge\bar{b}}$. Cependant, tenant compte des simulations effectuées, nous abandonnerons cette démarche qui, bien qu'intégrant de façon naturelle l'information due à la contraposée, conduit aux mêmes réserves que l'implication classique, à savoir

une valeur de l'intensité proche de 1 dès que les occurrences sont grandes. Aussi, nous nous tournons vers une deuxième approche.

6 Une deuxième approche d'une intensité d'implication intégrant la contraposée[6]

6.1 Motivation sémantique et épistémologique

Puisque implication et confiance admettent des racines communes (cf. § 3), nous associerons, sans agir contre nature, ces deux concepts, à la façon du physicien expérimentaliste comme il a été dit dans l'introduction. Les valeurs de la confiance relative respectivement aux règles $a \Rightarrow b$ et $\bar{b} \Rightarrow \bar{a}$ sont, en utilisant les mêmes notations que dans 2.1 :

$$C_1(a,b) = \text{Fr}[Y \mid X] = (\text{[card } X \cap Y]) / (\text{card } X) = \frac{n_{a \wedge b}/n}{n_a/n} = \frac{n_{a \wedge b}}{n_a}$$

$$C_2(\bar{b},\bar{a}) = \text{Fr}[\bar{X} \mid \bar{Y}] = (\text{[card } \bar{X} \cap \bar{Y}]) / (\text{card } \bar{Y}) = \frac{n_{\bar{a} \wedge \bar{b}}/n}{n_{\bar{b}}/n} = \frac{n_{\bar{a} \wedge \bar{b}}}{n_{\bar{b}}}$$

Or, compte tenu de leur définition, les deux intensités d'implication $\varphi(a,b)$ et $\varphi(\bar{b},\bar{a})$ sont égales. De ce fait, l'intensité ne nuance pas l'étonnement statistique de l'implication et de sa contraposée. En revanche, les deux confiances sont différentes en général. C_1 et C_2 sont d'excellents indicateurs des inclusions partielles : $A \subset B$ et $\bar{B} \subset \bar{A}$. Mais contrairement à $\varphi(a,b)$ et $\varphi(\bar{b},\bar{a})$, ils sont invariants dans toute homothétie globale des ensembles en jeu. Par suite, ils ne permettront pas de déceler des relations « surprenantes » eu égard à l'échantillon observé de taille n. C'est pour cette raison que nous leur associerons l'intensité d'implication classique,

Nous utiliserons cependant les nuances inclusives apportées par ces indicateurs d'inclusion, pour affecter l'intensité d'implication classique de coefficients permettant à la fois :

- d'intégrer les informations de l'implication directe et de sa contraposée

[6] Ne craignons pas le changement que nous introduisons car comme l'affirme G.Bachelard : « *Accéder à la science, c'est, spirituellement rajeunir, c'est accepter une mutation brusque qui doit contredire un passé* ».(Extrait de « La Formation de l'esprit scientifique », 1938).

- d'obtenir une meilleure discrimination des valeurs critiques des contre-exemples pour des grands échantillons où la valeur 1 apparaît comme point d'accumulation de l'intensité d'implication,
- et aussi de crédibiliser le caractère inclusif de A dans B.

Ainsi, en associant les deux indices, nous devrions obtenir une mesure qui intègre à la fois la recherche des implications surprenantes (à la charge de l'intensité d'implication classique) et la crédibilité de l'inclusion conditionnelle de A dans B et de $\overline{B}$ dans $\overline{A}$ (à la charge des confiances).

Notre objectif est alors de définir une fonction de C_1 et C_2 qui permette de satisfaire les critères précédents tout en majorant l'effet implicatif dont rend compte l'intensité d'implication afin d'obtenir une relation de la forme : $\varphi(a,b).\ f(C_1, C_2)$. Une fonction du type : $f(C_1.C_2) = [C_1(a,b).C_2(\bar{b}, \bar{a})]^r$ avec $r < ½$ satisfait nos exigences. Nous choisirons arbitrairement $r = ¼$ et examinerons, à travers nos expériences si cette valeur est satisfaisante pour restituer une information riche, plausible et susceptible de prédictabilité dans un cas causal. Quoi qu'il en soit, cette valeur petite de r affecte le pouvoir de la confiance en lui réduisant son effet d'évidence intrinsèque. De ce fait, il préserve la propriété de « surprise », « d'étonnement statistique » que recèle l'intensité d'implication. Ce qui nous satisfait.

6.2 La corbeille de la mariée.

Définition 4

On appelle **implifiance**[7] la mesure de l'implication statistique qui prend en compte l'implication directe et sa contraposée ainsi que la confiance en chacune de ces deux formes inclusives. Sa valeur est :

$$\Phi(a,b) = \varphi(a,b).\ [C_1(a,b).C_2(\bar{b}, \bar{a})]^{1/4}$$

Par exemple, si l'on extrait une règle dont l'implifiance est égale à 0.95, son intensité d'implication est au minimum égale à 0.95 et chacune des confiances C_1 et C_2 est au moins égale à 0.81. Si l'implifiance est égale à 0.90, les minimas respectifs sont 0.90 et 0.66, ce qui préserve la plausibilité de la règle.

Conclusion partielle

---

[7] Comme son nom le laisse penser, ce nouveau mot est la contraction de « **Impli**cation » et « Con**fiance** » puisque le concept associé est la combinaison des deux concepts évoqués ici.

Ainsi, même si les covariations de l'intensité d'implication et la confiance ne sont pas anarchiques, si elles n'obéissent pas ensemble à une référence en termes de probabilité, même si nous nous trouvons devant le choix arbitraire d'une définition composite d'où est partiellement exclue la part de contrôle de l'utilisateur de cette mesure, elles porteront les traces pondérées de deux indices majeurs pour évaluer la qualité implicative.

Quelques propriétés

- P_1 : si $n_a = n_{\bar{b}}$, alors $C_1 = C_2$. Nous avons dans ce cas, une même confiance en l'implication directe et sa contraposée ;
- P_2 : $\frac{1-C_1}{1-C_2} = \frac{n_{\bar{b}}}{n_a}$. Si ce dernier rapport reste constant, les deux confiances varient dans le même sens ;
- P_3 : en exprimant $C_1 = \frac{n_{a\wedge b}}{n_a} = 1 - \frac{n_{a\wedge\bar{b}}}{n_a}$ et $C_2 = \frac{n_{\bar{a}\wedge\bar{b}}}{n_{\bar{b}}} = 1 - \frac{n_{a\wedge\bar{b}}}{n_{\bar{b}}}$ en fonction du nombre de contre-exemples à l'implication directe, on obtient les dérivées partielles de ces confiances par rapport à ce paramètre :
  * $\frac{\partial C_1}{\partial n_{a\wedge\bar{b}}} = -\frac{1}{n_a}$ donc C_1 décroît quand les contre-exemples augmentent et d'autant plus vite que n_a est petit ,
  * $\frac{\partial C_2}{\partial n_{a\wedge\bar{b}}} = -\frac{1}{n_{\bar{b}}}$ donc C_2 décroît quand les contre-exemples augmentent et d'autant plus vite que n_b est grand, les autres paramètres étant constants.
  Ces résultats sont compatibles avec la formule établie dans (Gras et Régnier, 2003) : $\frac{\partial q}{\partial n_{a\wedge\bar{b}}} = \frac{1}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} = \frac{1}{\sqrt{\frac{n_a(n-n_b)}{n}}} > 0$, qui assurent ainsi qu'intensité d'implication et confiance varient plutôt dans le même sens en accord avec le point 2.3.

Remarque

Dans notre première approche, le cadre de la modélisation imposait la nature binaire des variables en jeu. Dans cette seconde approche, les deux notions de confiance, détachée de la sémantique forte en environnement binaire, gardent des propriétés fonctionnelles ce qui autorise la définition d'implifiance dans les cas où les variables sont de nature quelconque (numériques, modales, floues, variables intervalles, etc.).

7 Quelques comparaisons

7.1 Comparaisons entre l'intensité d'implication et l'implifiance

Voici une représentation (Fig. 1) de l'intensité d'implication (II) et de l'implifiance (IC) (en ordonnée) pour des valeurs de n = 100, n_a = 40, de n_b = 50 et où l'on fait varier $n_{a\wedge\bar{b}}$ de 0 à 40 en abscisses. Sur cette même représentation, nous faisons figurer, dans un objectif de comparaison, comment varient les 3 modes de valuation de la qualité de l'implication II, IC et IE (intensité entropique utilisée précédemment). Cette dernière sera également présente dans les figures 2,3, 4, 5, 6, 7, 8 et 9 où nous faisons varier les paramètres fondamentaux : n, n_a et n_b.

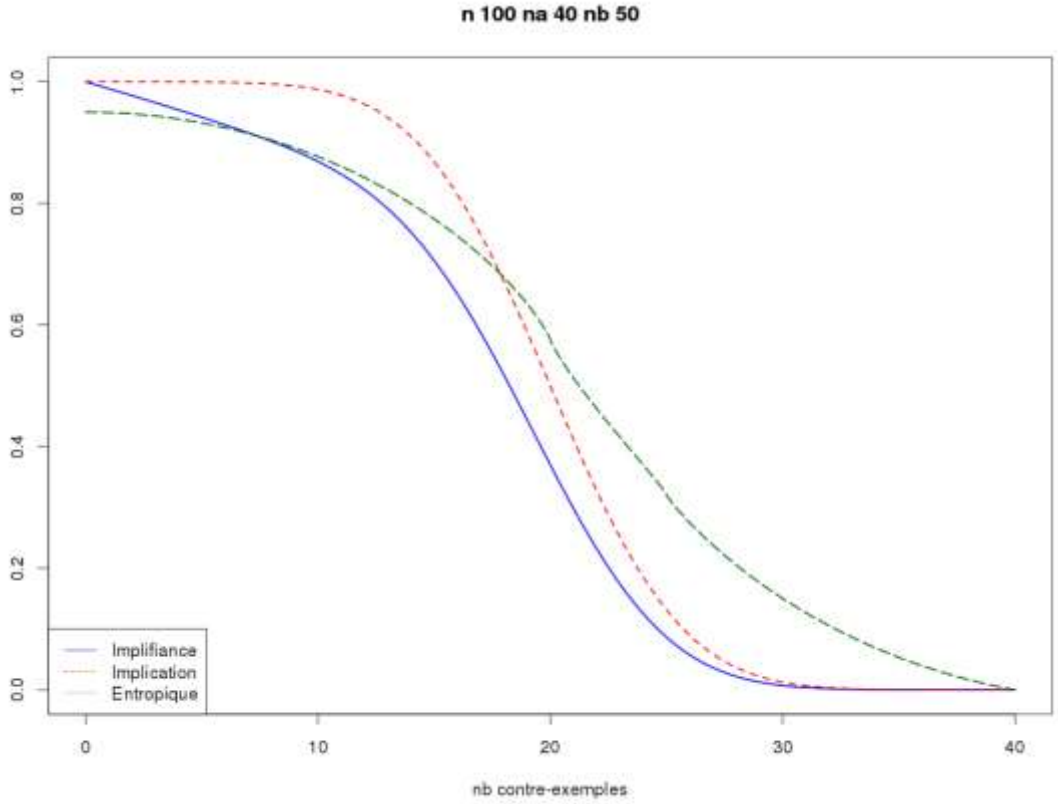

Figure 1

On remarque immédiatement, ce qui était attendu compte tenu de sa définition, que l'implifiance est toujours inférieure à l'intensité d'implication classique. Une valeur de 0.95 pour l'intensité d'implication pourra être accompagnée d'une valeur de l'implifiance compris entre 0.85 et 0.90. Que l'usager ne s'en inquiète pas. Par exemple, des confiances C_1 et C_2 supérieures à 0.50 conduisent à une implifiance supérieure à 0.67 alors que l'intensité est 0.95. De plus, l'IC n'ayant pas de dérivée nulle à l'origine décroît plus vite que l'II mais cette décroissance est lente avant de s'accélérer à n/4, voisinage du point d'inflexion de la courbe. En revanche, comme nous le savons, l'II décroît très lentement au début de la croissance des contre-exemples mais cette lenteur est préjudiciable pour des valeurs de n grandes car elle devient alors peu discriminante comme nous le voyons par une homothétie x100, dans la Fig. 2 où cette fois n = 10000, n_a = 4000, de n_b = 5000 et où l'on fait varier $n_{a\wedge\bar{b}}$ de 0 à 4000. L'II (intensité d'implication) ne varie que très peu de la valeur maximum 1 jusqu'à $n_{a\wedge\bar{b}}$ = 2000, puis plonge brusquement plus vite, ce qui la valorise, que IE, l'intensité entropique.

Remarquons que l'IC est plus fidèle à l'II que ne l'était la version précédente de la mesure de qualité entropique.

D'autres situations avec n= 10 et n_a = 1, n = 100 et n_a = 10 (Fig. 3), n = 1000 et n_a = 1000 montrent que le rejet de l'implication, en raison de la faiblesse de la valeur de IC, n'intervient pas trop tôt et seulement au voisinage du cardinal de $X \cap \overline{Y} = n_a/2$.

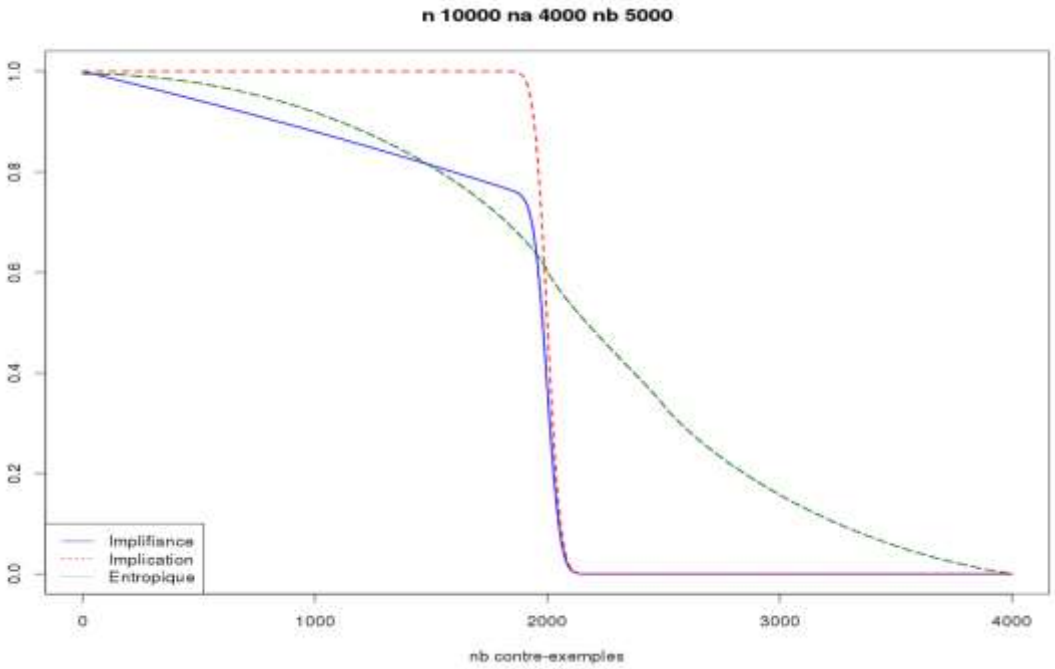

Figure 2

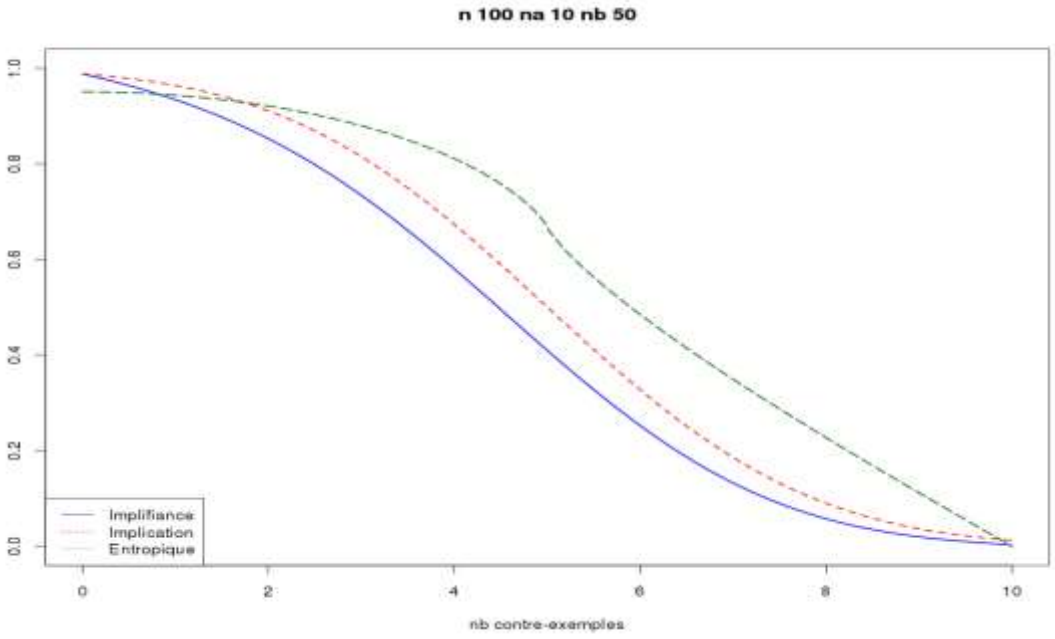

Figure 3

Lors de l'examen des spécificités de l'ASI disposant de l'intensité φ(a,b) comme outil de mesure implicative, nous avions souligné la propriété suivante : au voisinage de $N_{a\wedge\bar{b}} = 0$, l'intensité d'implication varie peu et reste très voisine de 1, comme le montrent les courbes (cf Fig.1 et Fig. 2), dont la dérivée à l'origine est presque nulle. Cette propriété a la vertu de permettre une certaine résistance de la fonction implicative, donc également prédictive, avant de céder à de trop forts écarts autour de 0. Certes, ceci évite le rejet brutal de l'implication pour quelques contre-exemples qui pourraient n'être

qu'accidentels ou dus à des erreurs de mesure. Mais le prix à payer pour garantir cette sagesse se facture en des valeurs de l'intensité difficilement discriminantes comme nous venons de le voir. Aussi, l'implifiance semble associer harmonieusement cette réserve à l'égard des écarts admissibles à l'implication et cependant la mise en garde du chercheur envers l'accumulation de contre-exemples devient insoutenable.

7.2 Comparaisons du comportement de l'implifiance par rapport à celui des confiances C_1 et C_2

Une autre propriété de l'intensité d'implication II a été soulignée. Ses variations en fonction de la variable $N_{a\wedge\bar{b}}$ ne sont pas linéaires contrairement à la seule confiance des multiples et autres indices d'association. Or, nous avons dénoncé l'inadéquation de la linéarité en tant que modélisante dans le fonctionnement de la pensée. La philosophie structuraliste nous rappelle que le « tout est plus riche que la somme de ses parties ». « *Autrement dit, 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 ne peuvent donc s'expliquer par elles* » (Sève L., 2005). Ce phénomène qui conduit bien souvent à l'émergence d'une propriété du tout nécessite analytiquement un caractère non dérivable ou à points d'inflexion de la fonction qui mesure le phénomène. D'où cette mise à l'écart de la linéarité. En effet, en un point d'inflexion de l'implifiance, celui-ci marque le passage de la phase d'accélération de la décroissance de l'implifiance à sa phase de décélération. Les courbes qui décrivent les variations de l'implifiance montrent bien ce caractère non linéaire qui tient à la présence multiplicative de l'intensité d'implication dans sa définition.

En revanche, comme le montrent les courbes représentant C_1 et C_2 des figures 4a, 5a et 6a, la linéarité de celles-ci ne respecte pas la philosophie dont s'arroge la nouvelle mesure implicative. Rappelons que C_1 mesure la fréquence de b sachant a et que C_2 mesure celle de non a sachant non b. Ces deux fréquences ne sont qu'exceptionnellement égales et se présentent avec une valuation indifférente : tantôt C_1 est au-dessus de C_2 et tantôt le contraire.

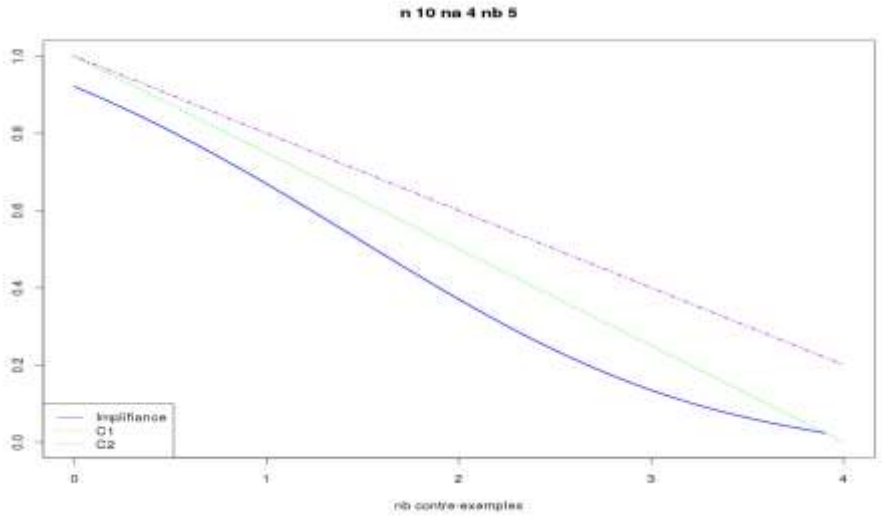

Figure 4a

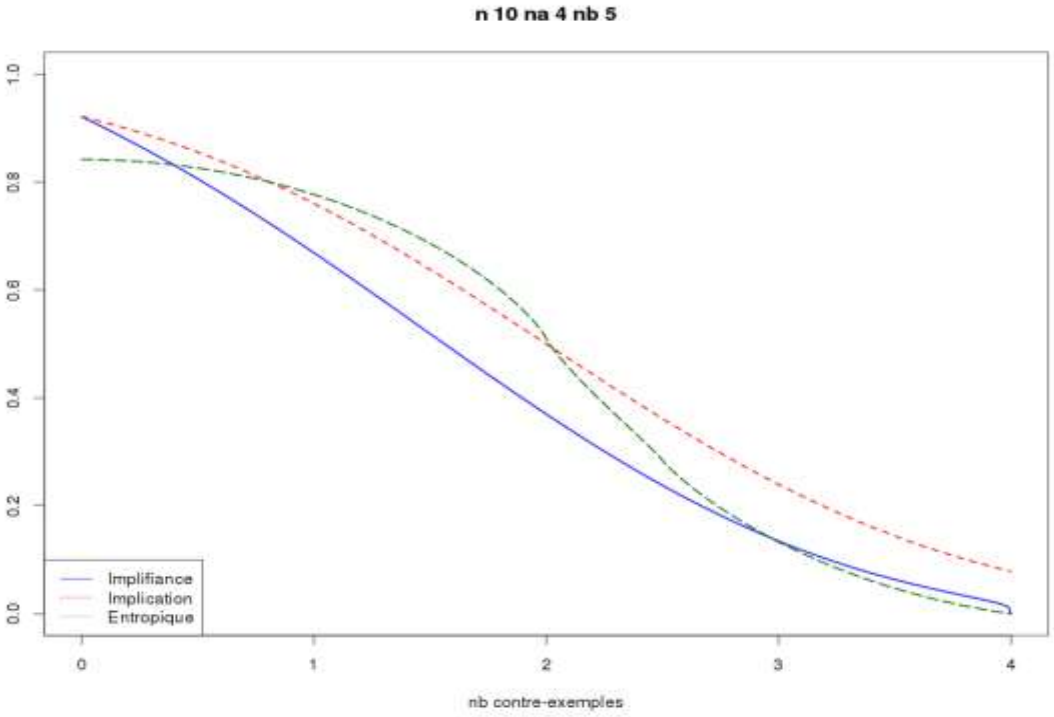

Figure 4

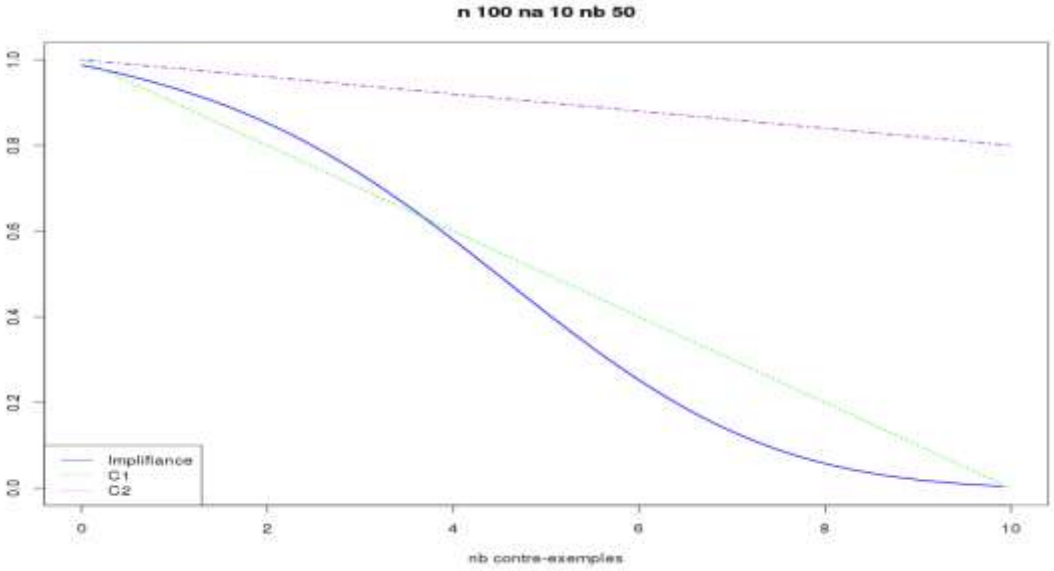

Figure 5a

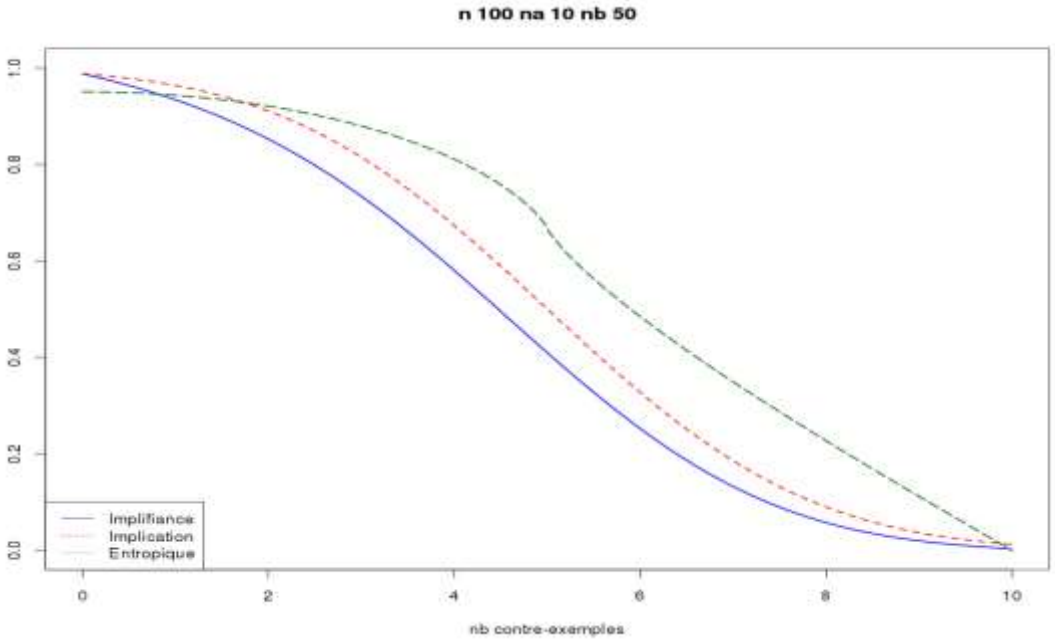

Figure 5

En revanche, la figure 6a montre clairement l'effet « lissage » produit par les coefficients des deux facteurs de confiance mais sans en supporter le défaut majeur de linéarité sur l'ensemble des nombres de contre-exemples : décroissance de l'implifiance de 0 à $n_a/2$, puis plongeon de celle-ci lorsque la règle implicative associée devient insoutenable. Cet effet « lissage » est répercuté sur l'implifiance comme on le voit sur la figure 6.

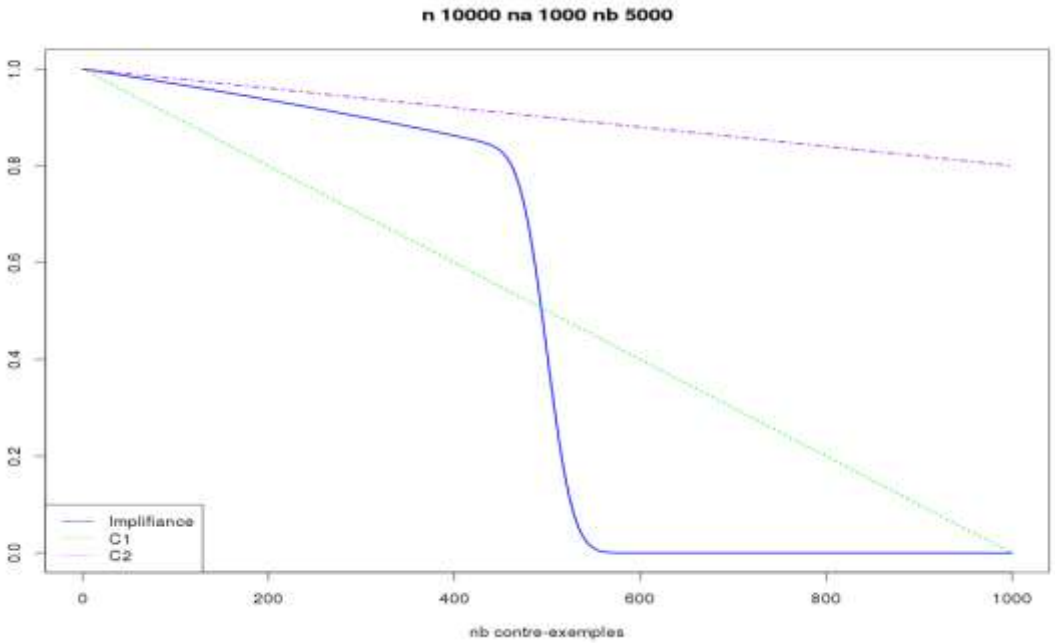

Figure 6a

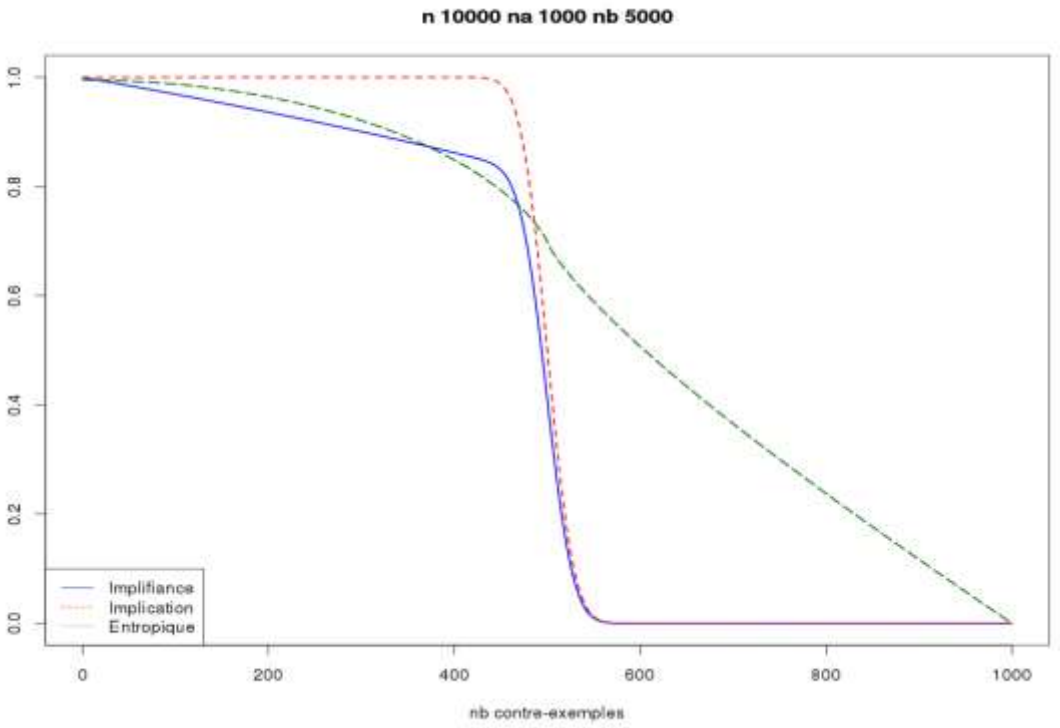

Figure 6

7.3 Comparaisons entre l'intensité d'implication, l'implifiance et l'intensité entropique

En réaction au comportement de l'intensité d'implication II pour les corpus volumineux, insuffisamment discriminante sur une grande plage de contre-exemples, nous avions défini un nouvel indice pour mesurer la qualité des règles implicatives. La définition s'appuyait sur la notion d'entropie conditionnelle des évènements associés à l'implication directe et sa contraposée. Comme nous l'avons dit dans l'introduction, il lui fut reproché son côté quelque peu ad-hoc ainsi que la nécessité de modifier l'entropie à partir de $n_a/2$ pour lui ôter sa propriété de symétrie incompatible avec la philosophie implicative. Autre grief : comme l'II la décroissance de l'intensité entropique semblait insuffisante pour rendre compte de l'admissibilité du caractère implicatif des règles évaluées. Ceci nous a conduits à la nouvelle mesure envisagée dans cet article : l'implifiance.

Les courbes 7, 8 et 9 qui suivent, obtenues également dans la situation où le nombre de sujets est important : n =10000, illustrent la comparaison entre les 3 mesures : l'intensité d'implication classique, l'intensité entropique et l'implifiance.

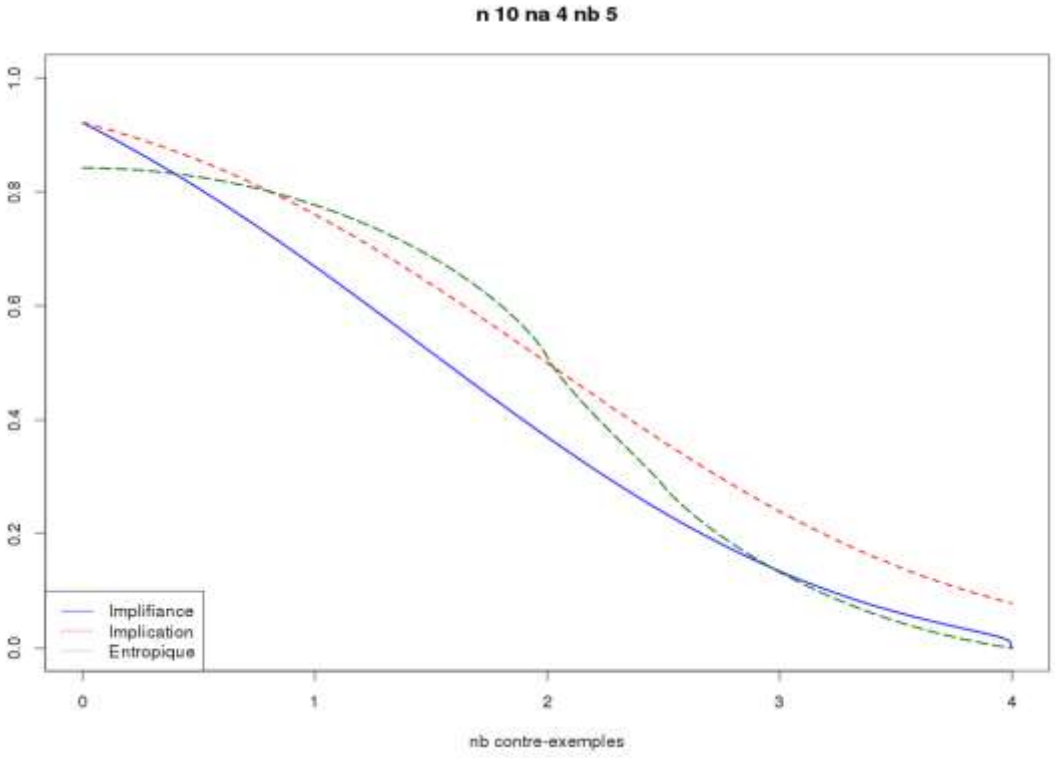

Figure 7

On remarquera, en effet, la plus grande résistance au rejet de la règle implicative dès que le nombre de contre-exemples devient trop importants eu égard aux valeurs de n_a et n_b, et particulièrement selon la courbe 9 des grandes valeurs des données.

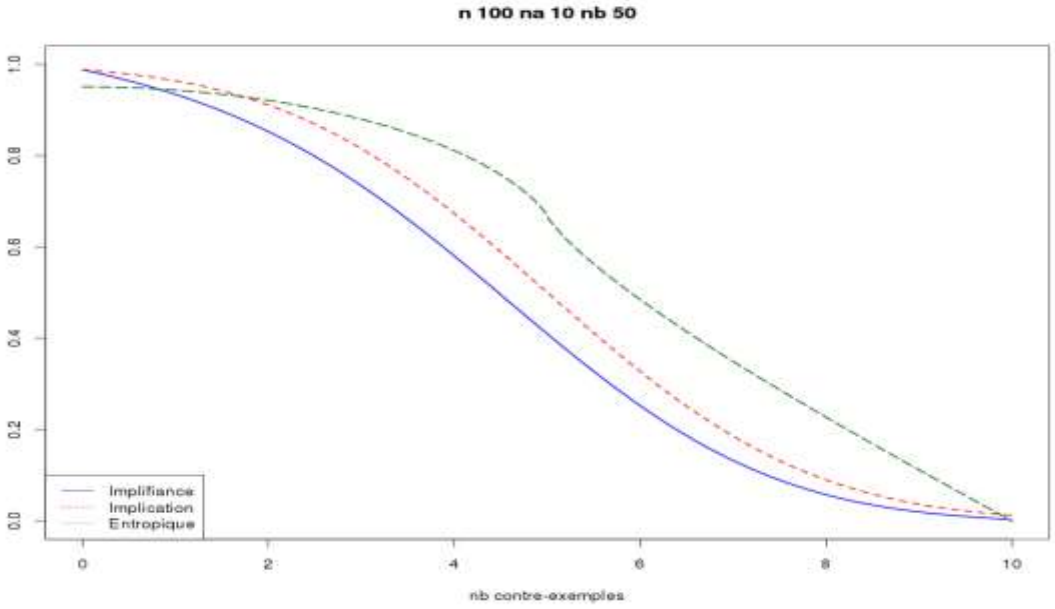

Figure 8

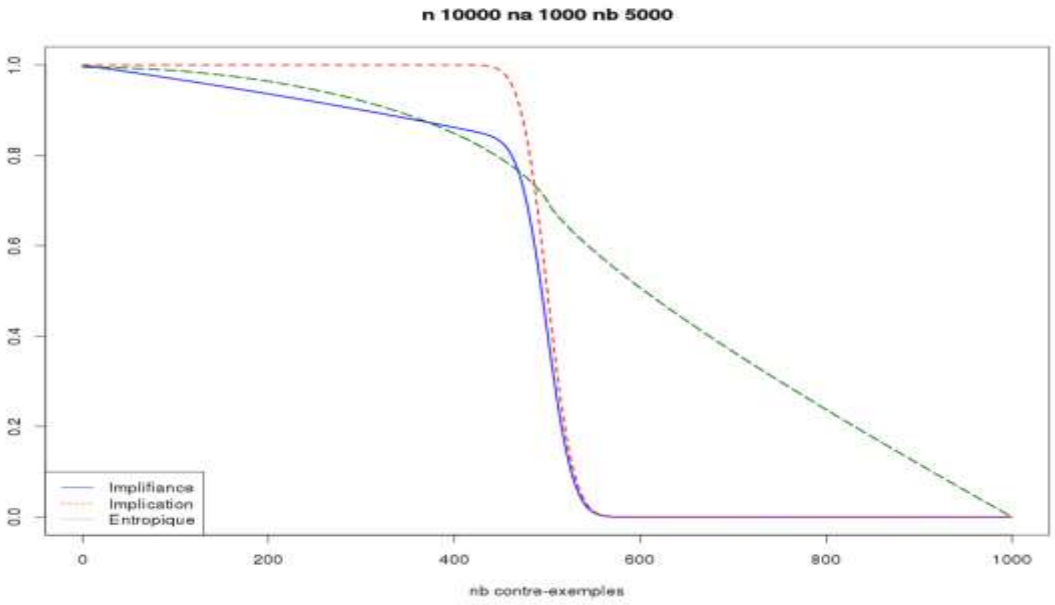

Figure 9

Comme nous l'avons déjà souligné, nous notons, avec ces 3 exemples 7, 8 et 9, une plus grande affinité entre l'intensité II et l'implifiance IC qu'entre II et l'intensité entropique IE. La fig. 9 est significative à cet égard où IE tarde à décroître avec le nombre croissant de contre-exemples. Ceci n'empêche pas l'implifiance de restituer à l'II les nuances de « confiance » que lui confèrent les deux composantes de la fréquence conditionnelle C_1 et C_2. Ceci confirme le contenu du § 3 où nous avions montré la relation entre intensité d'implication et confiance.

8 Conclusion

Nous proposons dans cet article la construction d'une nouvelle mesure qui permette d'évaluer la qualité de règles implicatives sur la base de l'importance relative des contre-exemples à l'implication. *Nous poursuivons donc la consolidation de notre édifice* ***non-réductionniste*** *de la mise en évidence et de représentation de relations présumées causales, cachées au sein d'un large ensemble de variables, édifice intégrant*

aussi la dualité sujets-variables, édifice qui fait de l'ASI un modèle original et fécond. D'ailleurs, nous avons pu laisser ouvertes quelques pistes de recherche ultérieure.

Cette mesure, *l'implifiance*, comme l'intensité entropique jusqu'alors utilisée en complément de l'intensité d'implication, a la vertu de prendre en compte la contraposée de l'implication faute de quoi la fonction de recherche causale nous semblerait imparfaite. *Par rapport à l'intensité entropique, elle satisfait le principe philosophique du rasoir d'Occam*[8] *car sa définition est moins complexe*. Elle intègre cette fois et en outre, la notion de confiance ou fréquence conditionnelle afin de limiter à l'examen des règles celles qui présentent un caractère de liaison conditionnelle entre la prémisse et la conclusion. Tout comme ses mesures aînées, cette implifiance subira les assauts des données réelles *(« Expérience, source unique de vérité »* écrit Henri Poincaré dans « Science et hypothèse ») pour pouvoir prétendre être un bon prédicteur de relation causale et un bon outil d'extraction de pépites de connaissance. Et de toute façon, la qualité de la prévision restera entachée des incertitudes inhérentes au choix du modèle probabiliste. Car, comme le dit avec humour le physicien Niels Bohr : « *La prévision est un art difficile, surtout quand elle concerne l'avenir »*.

Références

[1] Agrawal R., T. Imielinsky and A. Swami (1993), Mining association rules between sets of items in large databases, *Proc. of the ACM SIGMOD'93*, 207-216.

[2] Amarger S., D. Dubois and H. Prade (1991), Imprecise quantifiers and conditional probabilities, In *Symbolic and quantitative approaches to* uncertainty (R. KRUSE, P. SIEGEL), Springer-Verlag, 33-37.

[3] Bachelard G. (1967), *La Formation de l'esprit scientifique*, Paris, 5e édition, Librairie philosophique J. Vrin.

[4] Briand H., M. Sebag, R. Gras et F.Guillet (2004), *Mesures de Qualité pour la Fouille de Données*, H.Briand, M.Sebag, R.Gras et F.Guillet eds, RNTI-E-1, Cépaduès, 2004.

[8] *Ce principe est généralement retenu par les scientifiques qui affirment qu'en présence de plusieurs théories en charge de la même réalité, il est préférable de choisir la plus simple.*

[5] Brin S., R. Motwani and C. Silverstein (1997), Beyond market baskets: generalizing association rules to correlations, *Proc. Of ACM SIGMOD Conf. On Management of Data SIGMOD'97,* 265-276.

[6] Couturier, R. (2008). CHIC: cohesive Hierarchical Implicative Classification, In *Statistical implicative analysis*. Volume 127 of Studies in Computational Intelligence, Springer Verlag, p. 41–54.

[7] Fayyad U., G. Piatetsky-Shapiro and P. Smyth (1996), From Data Mining to Knowledge Discovery. In *Advances In Knowledge Discovery and Data Mining*, Fayyad U., Piatetsky-Shapiro G., Smyth P, and Uthurusamy R. eds, AAAI/MIT Press, 1-31.

[8] Frawley W., G. Piatetski-Shapiro and C. Matheus (1992), Knowledge discovery in databases: an overview. *AI Magazine*. 14(3), 57-70.

[9] 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 1.

[10] Gras R., S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn et A. Totohasina (1996), *L'implication Statistique*, Collection Associée à Recherches en Didactique des Mathématiques, Grenoble : La Pensée Sauvage.

[11] Gras R., R. Couturier, J. Blanchard, H. Briand, P. Kuntz et P. Peter (2004), Quelques critères pour une mesure de qualité de règles d'association. Un exemple : l'implication statistique, *Mesures de qualité pour la fouille de données, RNTI-E-1, Cépaduès –Editions*, 3-32.

[12] Gras R. et R. Couturier (2010), Spécificités de l'Analyse Statistique Implicative (A.S.I.) par rapport à d'autres mesures de qualité de règles d'association, *Quaderni di Ricerca in Didattica - GRIM (ISSN on-line 1592-4424)*, Eds : J.C. Régnier, R.Gras, F.Spagnolo, B. Di Paola, Université de Palerme, *p.19-57.*

[13] Gras R. et J.-C. Régnier (2013), Fondements théoriques de l'Analyse Statistique Implicative, *Méthode exploratoire et confirmatoire à la recherche de causalités*, sous la direction de Gras R., eds Gras R., Régnier J.-C., Marinica C., Guillet F., Cépaduès Editions, 522 pages, ISBN 978.2.36493.056.8, p 25-186.

[14] Gras R., J.-C. Régnier et F. Guillet (2009), *Analyse Statistique implicative.Une méthode d'analyse de données pour la recherche de causalités*, sous la direction de Régis Gras, réd. invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse.

[15] Gras R., J.-C. Régnier, C. Marinica et F. Guillet (2013), *L'analyse statistique implicative, Méthode exploratoire et confirmatoire à la recherche de causalités*, Cépaduès Editions, 522 pages, ISBN 978.2.36493.056.8.

[16] Gras R., E. Suzuki, F. Guillet and F. Spagnolo (2008), *Statistical Implicative Analysis, Theory and Applications*, R.Gras, E. Suzuki, F. Guillet, F. Spagnolo, eds, Springer.

[17] Guillet F. et H. Hamilton (2007), *Quality Measures in Data Mining* , F.Guillet et H.Hamilton eds, Springer.

[18] Hipp J., U. Guntzer and J. Nakhaeizadeh (2000), Mining association rules: Deriving a superior algorithm by analyzing today's approach , *Proc. of 4th Eur. Conf. on Principles of Data Mining and Knowledge Discovery*, *Lect. N. in Art. Int. 1910*, 160-168.

[19] Lallich S., O. Teytaud et E. Prudhomme (2007), Association Rule Interestingness: Measure and Statistical Validation, *F.Guillet and H. J.Hamilton eds, Studies in Computational Intelligence 43, Springer*, p. 251-275.

[20] Lenca P. et S. Lallich (2011), *Le choix d'une bonne mesure de qualité, condition du succès d'un processus de fouille de données*, Atelier Data Mining, Applications, Cas d'Etudes et Success Stories, Extraction et Gestion des Connaissances, *5-8.*

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

[22] Orús P., L. Zamora et P. Gregori (2009), *Teoria y Aplicaciones del Analisis Estadistico Implicativo,* Eds: P. Orús, L. Zamora y P. Gregori, Universitat Jaume I Castellon (Espagne), ISBN : 978-84-692-3925-4.

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

[24] Régnier J.-C., M. Bailleul et R. Gras (2012), *L'Analyse Statistique Implicative : de l'exploratoire au confirmatoire.* Eds : J.C. Régnier, Marc Bailleul, Régis Gras, Université de Caen, ISBN : 978-2-7466-5256-9, 2012.

[25] Saporta G. (2006), *Probabilités, Analyse de Données et statistique*, Paris : Ed. Technip.

[26] Schektman Y., J. Trejos et M. Troupe (1992), Un générateur de règles floues à partir de bases de données volumineuse, *Actes des 3èmes Journées "Symboliques-Numériques", mai 1992*, Paris.

[27] Sève L. (2005), *Emergence, complexité et dialectique*, Odile Jacob, Paris.

[28] Vaillant B., S. Lallich and P. Lenca (2008), On the behaviour of the generalisations of the intensity of implication: a data-driven comparative study, *Statistical Implicative Analysis, Theory and Applications*, R.Gras, E. Suzuki, F. Guillet, F. Spagnolo, Studies in Computational Intelligence, 127, Springer-Verlag Berlin Heidelberg.

[29] Vergnaud G. (2007), *Activités humaines et conceptualisation*, Presses Universitaires du Mirail, p.29.

[30] Xuan-Hiep Huynh, F. Guillet, J. Blanchard, P. Kuntz, H. Briand et R. Gras (2007), A Graph-based Clustering Approach to Evaluate Interestinguess Measures: A Tool and and a Comparative Study, *F.Guillet and H. J.Hamilton eds, Studies in Computational Intelligence 43, Springer*, p. 25-50.

ANNEXE

Lemme

Le comportement asymptotique de φ(a,b), intensité d'implication, est celui d'une variable uniforme sur l'intervalle [0 ,1].

Notons $p=\varphi(a,b)$, p étant une probabilité, ne peut être considérée formellement comme une variable aléatoire qu'en changeant le modèle probabiliste jusqu'alors adopté. En effet, l'univers des possibles Ω introduit pour l'étude de l'étonnement à l'observation de la valeur $q(a,\bar{b})$, suppose que les valeurs également observées n_a et n_b soient des réalisations des cardinaux des variables aléatoires indépendantes X et Y. Si l'on veut alors mesurer un "étonnement" au sujet de la valeur de la cohésion, il faut considérer un autre univers des possibles Ω' sur lequel p suivrait une certaine loi de probabilité. Le problème revient alors à choisir la "bonne" loi de probabilité que suivrait p.

Nous allons la calculer pour n suffisamment grand, de telle façon que l'on puisse se considérer dans le cas asymptotique. Pour cela, on suppose que la valeur observée $n_{a\wedge\bar{b}}$ est la réalisation d'une variable aléatoire qui n'est plus Card $(X\cap\overline{Y})$, mais Card $(A\cap\overline{B})$, où A et B sont des sous-ensembles aléatoires indépendants dans E et fonctions d'éventualités $\omega\in\Omega'\neq\Omega$.

On obtient dans ces nouvelles conditions:

$$\text{Prob}\,[\,p\le\alpha\,] = \text{Prob}_{\Omega'}\,[\,\text{Prob}_{\Omega}\,(\,\text{Card}\,(X\cap\overline{Y})\le n_{a\wedge\bar{b}}\,)\le\alpha\,]$$

$$= \text{Prob}_{\Omega'}\,\{\,\omega' \mid \text{Prob}_{\Omega}\,\{\,\omega \mid \text{Card}\,(X\cap\overline{Y})(\omega)\le \text{Card}\,(A\cap\overline{B})(\omega')\,\}\le\alpha\}$$

$$= \text{Prob}_{\Omega'}\,\{\,\omega' \mid \text{Prob}_{\Omega}\,\{\,\omega \mid N(\omega)\le N'(\omega')\}\le\alpha\}\}$$

où N et N' admettent des lois identiques, par exemple des lois normales centrées réduites et dans ce cas :

$$\text{Prob}\,[\,p\le\alpha\,] = \text{Prob}_{\Omega'}\,\{\,\omega' \mid \int_{-\infty}^{N'(\omega')} e^{-\frac{t^2}{2}}\,dt\le\alpha\}$$

$$= \text{Prob}_{\Omega'}\,\{\,\omega' \mid N'(\omega')\le F^{-1}\,(\alpha)\}, \text{ où F est la fonction de répartition de la loi normale,}$$

$$= \int_{0}^{F^{-1}(\alpha)} e^{\frac{-t^2}{2}}\,dt = F(F^{-1}(\alpha))$$

$$= \alpha \qquad (1)$$

Ainsi, la loi de p est uniforme sur l'intervalle [0,1]. Cela signifie que, si n est très grand, un tirage au hasard de deux variables a et b étant donné,

l'évènement $[[\varphi(a,b)\leq\alpha]=\alpha]$ ou $[\varphi(a,b) \geq \alpha]=1-\alpha]$ est presque sûr. De plus, la loi de l'intensité d'implication, en tant que variable aléatoire, tend vers une loi uniforme sur [0,1].

Exemple numérique fictif : 11 variables, 1081 objets (des véhicules)

On dispose de 11x10/2 soit 55 couples envisageables pour définir une règle de type a $\Rightarrow$ b respectant $n_a \leq n_b$. Le tableau des intensités d'implication donne des résultats n'invalidant pas la relation (1) :

- 5 intensités supérieures ou égales à 0.95 (1-a = 0.05) pour 5% x 55 = 2.75 intensités attendues d'après (1) ;
- 6 intensités supérieures ou égales à 0.90 (1-a = 0.10) pour 10% x 55 = 5.5 intensités attendues d'après (1) ;
- 8 intensités supérieures ou égales à 0.80 (1-a = 0.20) pour 20% x 55 = 11.5 intensités attendues d'après (1) ;

Ce qui statistiquement et pour ces trois mesures représente 19 résultats observés pour 19.75 attendus confortant notre lemme.